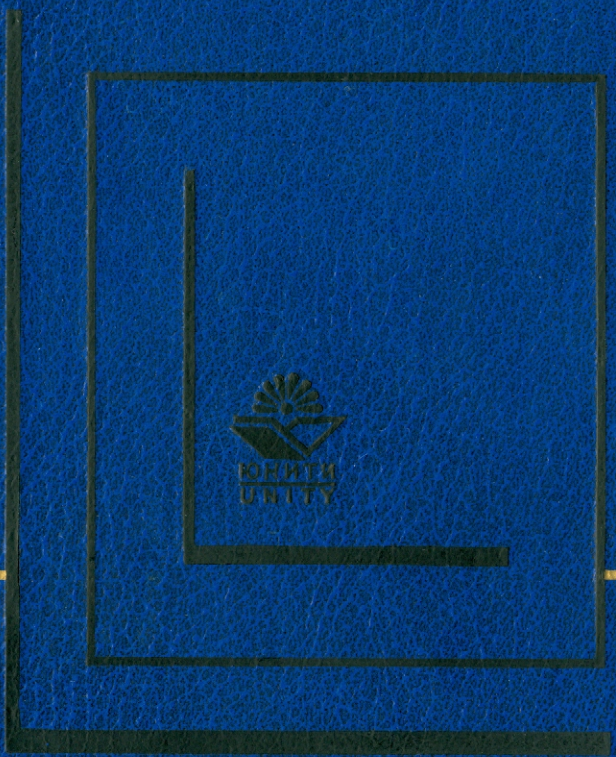
Многомерный статистический анализ в экономике



Многомерный статистический анализ в экономике

1573

L. A. SOSHNIKOVA
V. N. TAMASHEVICH
G.UEBE
M. SHEFER

MULTIDIMENSIONAL STATISTICAL ANALYSIS IN ECONOMICS

Edited by Prof. ***V. N. Tamashevich***

Textbook

ЮНИТИ
UNITY

Moscow • 1999

Л. А. СОШНИКОВА
В. Н. ТАМАШЕВИЧ
Г.УЕБЕ
М. ШЕФЕР

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ В ЭКОНОМИКЕ

Под редакцией профессора ***В. Н. Тамашевича***

Рекомендовано Министерством образования Российской Федерации в качестве учебного пособия для студентов высших учебных заведений

ЮНИТИ
UNITY

Москва • 1999

ББК 65.051я73
У32

Рецензенты:
кафедра статистики экономического факультета Московского государственного университета им. М.В. Ломоносова (зав. кафедрой д-р экон. наук Ю.Н. Иванов) и канд. физ.-мат. наук, проф. Г.М. Булдык

Главный редактор издательства *Н.Д. Эриашвили*

Сошникова Л.А., Тамашевич В.Н., Уебе Г., Шефер М.
У32 Многомерный статистический анализ в экономике: Учеб. пособие для вузов/Под ред. проф. В.Н. Тамашевича. — М.: ЮНИТИ-ДАНА, 1999. — 598 с.
ISBN 5-238-00099-5.

Достаточно полно представлены теоретические основы и важнейшие методы многомерной статистики, открывающей для исследователя широкие возможности моделирования сложных реальных процессов, явлений и визуализации данных.

Знание методов многомерной статистики сегодня необходимо не только для аналитической работы, но и для понимания новейших теорий по обработке данных массовых наблюдений. Участие в составе авторского коллектива преподавателей университета Бундесвера (г. Гамбург) позволило включить в пособие последние результаты глубоких теоретических исследований в области статистики и учесть опыт ее преподавания в ведущих высших учебных заведениях Западной Европы.

Приведено большое число примеров решения конкретных задач из области экономики.

Для студентов и преподавателей экономических вузов и специальностей, пособие будет также полезно специалистам в области медицины, инженерного дела, психологии человека и другим практическим работникам в качестве справочника по многомерной статистике.

ББК 65.051я73

ISBN 5-238-00099-5

© Л.А. Сошникова, В.Н. Тамашевич, Г. Уебе, М. Шефер, 1999
© ООО "ИЗДАТЕЛЬСТВО ЮНИТИ-ДАНА", 1999.
Воспроизведение всей книги или любой ее части запрещается без письменного разрешения издательства

Оглавление

Принятые в учебном пособии обозначения 9

Предисловие 11

Глава 1. Задачи и методы многомерного статистического анализа (МСА) 15

1.1. Теоретические основы МСА, его место в социально-экономических исследованиях 15
1.2. Методы МСА 22
1.3. Многомерное признаковое пространство. Особенности обработки многомерных статистических данных 27

Глава 2. Элементы математики в МСА 32

2.1. Основы аналитической геометрии 32
2.2. Элементы матричной алгебры 41

Глава 3. Случайные величины. Законы распределения и плотность вероятности 89

3.1. Случайные величины и их распределения 89
3.2. Некоторые виды параметрических распределений 96
3.3. Непрерывные распределения 103
3.4. Математическое ожидание и дисперсия 115
3.5. Двумерные и многомерные случайные величины 123
3.6. Статистические методы точечного оценивания 139

Глава 4. Проверка статистических гипотез 172

4.1. Статистические гипотезы в анализе данных 172

4.2. Проверка гипотез о равенстве вектора средних значений постоянному вектору 176
4.3. Проверка гипотез о равенстве двух векторов средних значений 183
4.4. Проверка гипотез о равенстве ковариационных матриц 189

Глава 5. Робастное статистическое оценивание 196

5.1. Грубые ошибки и методы их выявления в статистической совокупности данных 196
5.2. Методы исчисления устойчивых статистических оценок: Пуанкаре, Винзора, Хубера 200

Глава 6. Многомерный регрессионный анализ 214

6.1. Введение в множественный корреляционно-регрессионный анализ 214
6.2. Линейная регрессия — классический случай зависимости двух переменных *X* и *Y* 223
6.3. Свойства статистических оценок параметров регрессионной модели 232
6.4. Статистическое оценивание методом наименьших квадратов — обобщения на случай матричного представления линейной регрессии 239
6.5. Нелинейные регрессионные модели 304

Глава 7. Факторный анализ 333

7.1. Сущность методов факторного анализа и их классификация 333
7.2. Фундаментальная теорема факторного анализа Тэрстоуна 340
7.3. Общий алгоритм и теоретические проблемы факторного анализа 341
7.4. Метод главных компонент 347
7.5. Разложение дисперсии в факторном анализе 368
7.6. Метод главных факторов 372
7.7. Метод максимального правдоподобия 379

7.8. Вращение пространства общих факторов 385
7.9. Статистическая оценка надежности решений методами главных компонент и факторного анализа 393

Глава 8. Многомерное шкалирование 401

8.1. Многомерное шкалирование в статистических исследованиях 401
8.2. Представление и первичная обработка статистических данных в многомерном шкалировании 407
8.3. Классическая модель многомерного шкалирования Торгерсона 415
8.4. Неметрические методы многомерного шкалирования 421
8.5. Модели поиска индивидуальных различий 432
8.6. Анализ предпочтений 445

Глава 9. Кластерный анализ 468

9.1. Общая характеристика методов кластерного анализа 468
9.2. Меры сходства 471
9.3. Иерархический кластерный анализ 474
9.4. Метод k-средних 486
9.5. Метод поиска сгущений 493
9.6. Критерии качества классификации 497

Глава 10. Дискриминантный анализ 507

10.1. Основные положения дискриминантного анализа 507
10.2. Дискриминантные функции и их геометрическая интерпретация 509
10.3. Расчет коэффициентов дискриминантной функции 511
10.4. Классификация при наличии двух обучающих выборок 513
10.5. Классификация при наличии *k* обучающих выборок 516
10.6. Взаимосвязь между дискриминантными переменными и дискриминантными функциями 519

Глава 11. Метод канонических корреляций 526
11.1. Сущность и теоретические основы метода 526
11.2. Подготовка информации и вычисления канонических корреляций 528
11.3. Оценка значимости канонических корреляций 534
11.4. Экономическая интерпретация результатов канонического анализа 536
Глава 12. Компьютерный анализ многомерных статистических данных 542
12.1. Характеристика и особенности построения пакетов STATGRAPHICS и DSTAT 542
12.2. Реализация методов многомерного статистического анализа в пакетах STATGRAPHICS и DSTAT 550
Библиографический список 577
Приложения 584
Предметный указатель 592

Принятые в учебном пособии обозначения

R^n	— n-мерное евклидово пространство
(a, b)	— скалярное произведение векторов a и b
x, y	— векторы (элементы) линейного пространства
$(x_{0-\delta}; x_{0+\delta})$	— δ-окрестность точки x_0
A, B	— матрицы
I_n	— единичная матрица размерностью $n \times n$
rg (A)	— ранг матрицы
A^{-1}, A^*	— обратная и транспонированная матрицы
A^+	— псевдообратная матрица: $A = AA^+A$, применяется при решении систем линейных уравнений
$X, \xi, \lambda, \zeta \ldots$	— случайные величины
var (X)	— дисперсия случайной величины
$E(X)$	— математическое ожидание случайной величины
cov (X)	— ковариация случайной величины X
diag (A)	— диагональная матрица
$\in$	— знак принадлежности множеству
$\notin$	— знак непринадлежности множеству
$\subset$	— знак включения в множество
$\cup$	— знак объединения множеств (событий)
$\cap$	— знак пересечения множеств (событий)
$\rightarrow$	— знак логического следования
$\Leftrightarrow$	— знак равносильности
$:=$	— знак присвоения (придать значение)
$\therefore$	— знак следования логического вывода
■	— знак завершения доказательства
$\wedge$	— знак конъюнкции («и»)
$\vee$	— знак дизъюнкции («или»)
$\forall$	— квантор общности, соответствует словам: «для любого», «для всех», «все»
$\exists$	— квантор существования, соответствует словам: «имеется», «найдется», «существует»

θ — параметрическая величина

$N(\mu, \sigma^2)$ — нормальный закон распределения с параметрами случайной величины: математическое ожидание μ и дисперсия σ^2

$X \sim N(\mu, \sigma^2)$ — подчинение X (эквивалентность) закону нормального распределения

$\frac{\partial \varphi}{\partial x}, \frac{\partial^2 \varphi}{\partial y \partial z}$ — частные производные функции $\varphi = f(x, y, z)$

$p\lim$ — предел по вероятности уровня p

$\dim V$ — размерность векторного пространства V

$\| x \|$ — норма вектора x

Предисловие

Методы многомерного статистического анализа (МСА), представленные в учебном пособии, сегодня называют интеллектуальным инструментарием исследователя. Они составляют неотъемлемую часть фундаментальных курсов университетского образования и активно используются в аналитической практике в странах с передовой экономикой.

Постоянно возрастающий интерес к МСА объясняется прежде всего его широкими возможностями в отображении и моделировании реальных явлений и процессов, изначально имеющих, как известно, многопризнаковую природу. Кроме того, без базовых знаний по обработке многомерных данных просто не могут развиваться современные математика и статистика. Все новейшие разработки, посвященные проблемам приложения нечетких множеств, моделирования катастроф, распознавания образов, сценарного прогнозирования и т.д., предполагают многомерное представление наблюдаемых объектов.

МСА в теоретическом плане представляет собой дальнейшее развитие традиционной одномерной статистики, его отличают трудоемкие алгоритмы реализации вычислительных процедур, практически всегда рассчитанные на привлечение технических средств, и сложная интерпретируемость аналитических результатов. Это требует от пользователя достаточно глубокой подготовки как в области математической статистики, так и в области, в которой проводятся конкретные исследования: экономики, медицины и т.п.

Исторически многомерный статистический анализ можно рассматривать как одно из новых самостоятельных направлений развития статистической теории. Его появление в начале XX в. связывают с публикациями работ К. Пирсона, в которых были изложены концептуальные основы построения алгоритмов сжатия статистических данных. Первые теоретические разработки

МСА, включая и работы К. Пирсона, проводились с учетом потребностей аналитической практики в психологии, социологии, медицине. Наиболее активно формирование теории МСА происходило уже в 40—60-е годы, в это время область ее приложения значительно расширилась (военная промышленность, геодезия), захватывая и экономику.

В СССР изучение и использование методов многомерной статистики долгое время (до конца 50-х — начала 60-х годов) сдерживалось идеологическим неприятием формальной статистики вообще. Применение этих методов начиналось с военной промышленности, позже они вошли в аналитическую практику исследований в здравоохранении, экономике, других отраслях.

В предлагаемом учебном пособии авторы ставили своей целью комплексно и доступно, в том числе для читателей, не имеющих специального математического образования, представить классические методы МСА. Изложение материала построено по общепринятой логичной схеме: от простого к более сложному. Открывают учебное пособие главы (2, 3), посвященные теоретическим основам МСА, где приводятся наиболее важные фрагменты из аналитической геометрии и матричной алгебры.

В последующих главах (4, 5) кратко изложены основные положения многомерной математической статистики (законы распределения, методы оценивания многомерных случайных величин и проверки многомерных гипотез, приемы робастного оценивания случайных величин), а затем представлены сами методы МСА: множественного корреляционно-регрессионного анализа, многомерного шкалирования, кластер-анализа и т.д. (гл. 6—11). Чтобы облегчить восприятие теоретического материала и иметь возможность контроля за его усвоением, эти главы дополняются историческими справками о появлении и развитии методов, примерами их приложения в конкретных экономических исследованиях, а завершают каждую главу вопросы и задачи для самостоятельной работы.

В заключение приводится глава (12) с рекомендациями по реализации методов МСА при помощи стандартных пакетов прикладных программ.

Авторы стремились избегать чрезмерного математизирования. В большинстве глав приводятся только самые необходимые для понимания логической конструкции описываемого метода формулы и доказательства. Несколько выделяются на общем

фоне гл. 2, 3, 6, подготовленные немецкими авторами: г-ном проф. д-ром Г. Уебе и г-ном д-ром М. Шефером. В них изложение материала в большей мере ориентировано на формализованное и доказательное представление. Этот материал интересен и весьма полезен для читателя тем, что в нем обобщается опыт большой исследовательской и педагогической работы преподавателей одного из самых авторитетных вузов Германии — Университета Бундесвера. Сам материал доступен даже для читателя, не искушенного в математике, в силу логичности его построения и поистине мастерского изложения. Наконец, именно здесь можно увидеть и ощутить самые тонкие материи теоретической статистики.

В работе над пособием авторы ориентировались на требования профессиональной подготовки в экономическом вузе. Но в общем книга с облегченным изложением основ МСА, достаточно большим перечнем методов и примерами их практического приложения может служить справочником как для студентов, так и для практиков самых различных специальностей: медиков, инженеров и т.д.

Белорусские авторы выражают искреннюю благодарность *проф. Г. Уебе* и *д-ру М. Шеферу* за участие в подготовке рукописи. Доверительное отношение со стороны немецких коллег, их бескорыстная и пунктуальная работа позволили подготовить книгу в полном объеме к изданию. Благодаря усилиям г-на Г. Уебе, участию администрации университета Бундесвера в течение работы над рукописью, проходившей без сторонней финансовой поддержки, удавалось находить возможности для встреч и продолжения сотрудничества. Для нас совместная работа, основанная на дружеских и доброжелательных отношениях, была не только полезной, но и весьма приятной. Надеемся, что установленные контакты помогут нам и в будущем вместе приходить к интересным идеям и добиваться их реализации.

Свою признательность мы выражаем также Министру статистики и анализа Республики Беларусь *В.И. Зиновскому*, заведующему кафедрой статистики БГЭУ проф. *И.Е. Теслюку*, известному специалисту в области программирования и создания компьютерных систем, автору большого числа книг по языкам программирования *А.Н. Вальвачеву*. На разных этапах более чем четырехлетней работы над рукописью их советы, поддержка, идеи, редакционные замечания помогали авторам преодолевать со-

мнения, позволяли устранять многие неточности и наконец довести работу до завершения.

Большая техническая помощь при подготовке рукописи в машинописном виде была оказана сотрудниками белорусского Института статистики *Н.А. Курец, Е.Е. Судник.*

Для нас работа над проблемами МСА была во многом новой, мы часто использовали информацию из монографий, только что опубликованных, или источников, не переведенных на русский язык. Наверняка какие-то важные моменты могли оказаться упущенными, где-то недостаточно четко проведена систематизация или недостаточно нагляден пример. Мы будем рады, если эта книга вызовет интерес и найдет отклик у читателей. Отзывы и пожелания помогут найти более интересные и прагматичные направления для будущей работы.

В.Н. Тамашевич
Л.А. Сошникова

Авторы учебного пособия:

Л.А. Сошникова — канд. экон. наук, доцент Белорусского государственного экономического университета (гл. 9, 10, 11, 12)

Готц Уебе — проф., д-р Университета Бундесвера (г. Гамбург), директор Института статистики и эконометрии Университета Бундесвера (гл. 6)

Мартин Шефер — д-р Университета Бундесвера (г. Гамбург), преподаватель Института статистики и эконометрии Университета Бундесвера (гл. 2, 3)

В.Н. Тамашевич — канд. экон. наук, директор Института статистики при Министерстве статистики и анализа Республики Беларусь (гл. 1, 4, 5, 7, 8, перевод с немецкого гл. 2, 3, 6)

Глава 1
Задачи и методы многомерного статистического анализа (МСА)

1.1. Теоретические основы МСА, его место в социально-экономических исследованиях

МСА следует рассматривать как логическое развитие методов традиционной статистики, обобщенных в курсе общей теории статистики. Принципиальное отличие заключается в том, что объекты, социальные и экономические явления рассматриваются здесь с учетом не одного-двух, а одновременно некоторого множества признаков. Это позволяет добиваться в исследованиях полноты теоретического описания наблюдаемых объектов и объективности последующих выводов. Действительно, если судить о человеке только по уровню его заработной платы, или заработной плате и уровню образования, то наши выводы будут ограниченны и неточны. Другое дело, если мы воспользуемся набором признаков, представляющих состояние здоровья, социальное положение, уровень профессиональной подготовки и т.д. Совместное исследование значений этих признаков позволит моделировать образ субъекта и реально оценивать его поведенческую реактивность. Подобные примеры можно привести и из области экономики, социологии, политики. Так, если на предприятии имеется высокий уровень производительности труда, то это вовсе не обязательно означает, что оно работает устойчиво, имеет достаточное финансовое обеспечение и может выступать надежным партнером. В данном случае для достоверной оценки дополнительно необходимы характеристики ликвидности средств предприятия, структуры капитала, эффективности вложений капитала и т.п.

Реально изучаемые объекты и явления имеют практически всегда многопризнаковую природу, надежное отображение их в

экономико-математических моделях возможно при условии учета комплекса присущих им наиболее существенных характеристик.

Переходя к определению МСА, отметим, что это сформировавшаяся самостоятельная область теоретической статистики. *Это совокупность глубоко формализованных статистических методов, базирующихся на представлении исходной информации в многомерном геометрическом пространстве и позволяющих определять неявные (латентные), но объективно существующие закономерности в организационной структуре и тенденциях развития изучаемых социально-экономических явлений и процессов.*

Для МСА как самостоятельной области науки характерны следующие особенности:

- Методы МСА в отличие от классической статистики появились сравнительно недавно: факторный анализ — на рубеже XIX и XX вв., многомерное шкалирование — в конце 30-х — в 40-х годах нашего столетия, кластер-анализ — 10—20-е годы и т.д. Основная часть методов еще находится в стадии активной разработки, область их применения четко не разграничена. Отсутствуют строгие рекомендации по приложению этих методов в решении большого числа конкретных ситуационных задач.
- Для методов МСА характерны, как правило, глубокая формализация, сложная логико-математическая конструкция. Работа с этими методами требует углубленных знаний в области как экономической теории, так и математики. Недостаток в уровне подготовки исследователей обычно проявляется в некорректном приложении методов или в ошибочной интерпретации аналитических результатов.
- Применение методов МСА требует творческого подхода к решению аналитических задач. В данном случае это требование к исследователю значительно сильнее, чем при работе с методами общей теории статистики. Во-первых, методы МСА весьма многообразны и многочисленны. Для решения даже одного типа задач здесь существуют десятки и сотни различных приемов: в кластер-анализе насчитывается более 200 различных подходов и методов; в факторном анализе, многомерном шкалировании — десятки различных методов и т.д. Чтобы правильно выбрать тот или иной метод или комплекс методов для последовательного решения поставленной проблемы, естественно, необходимы профессионализм и хорошая интуиция.

Во-вторых, творческий подход и профессиональная подготовка приобретают особенно важное значение при интерпретации аналитических результатов, часто неоднозначных, когда экономические, социологического плана выводы должны отвечать логической схеме сложных математических расчетов (противоречия в экономических и математических выводах свидетельствуют о некорректности решения задачи или некорректности интерпретации аналитических результатов). Например, в МСА часто используется понятие латентного (скрытого) признака. Этот признак обобщает несколько элементарных признаков, известных из общей теории статистики, таких, как производительность труда, уровень квалификации, стаж работы и т.п. Названия латентного признака первоначально не существует, оно должно быть определено исследователем по комбинации элементарных признаков в латентном. В каждой отдельной задаче приходится иметь дело с особенными латентными признаками, и их нельзя ввести в какой-либо справочник или каталог. Именно исследователь каждый раз решает вопросы, стоит ли оставлять в анализе выделенный латентный признак (насколько он значим?) и как он должен быть назван.

15 73

- В МСА обрабатываются многомерные (многопризнаковые) совокупности данных. Число признаков (или размерность совокупности) при этом может быть любым — от 1 до 100 и более, но обычно более двух, и максимально ограничивается 20—33 признаками. Существует точка зрения, что для описания реальных объектов достаточно 33 измерений, сверх этого — уже излишняя информация. Такой подход оправдывает себя довольно часто, но не может считаться законом, в конечном счете все определяется условиями задачи и целями исследования.
- Практическое применение методов МСА требует обязательного использования вычислительной техники. Можно сказать, что эти методы в силу сложности и трудоемкости нереализуемы без технических средств. Широкое распространение МСА в исследованиях началось именно с появлением первых ЭВМ.

Идеи МСА не являются открытием XX столетия. Еще Аристотелем в III в. до н. э. был предложен в сущности многомерный подход при классификации предметов по их сходству и различиям. В новейшей истории, в XVIII — до начала XX вв. сама возможность многопараметрического описания объектов, явлений, процессов в научных исследованиях становится осо-

бенно привлекательной. Ее активно разрабатывают французский ботаник М. Адамсон (60-е годы XVIII в.) при идентификации растений; английский естествоиспытатель Ч. Дарвин (60-е годы XIX в.) — в своей селекции видов и при определении факторов эволюции органического мира; Д.И. Менделеев (60—70-е годы XIX в.) — при систематизации качественных характеристик химических элементов. Не только в естествознании, но и в экономике, статистике многомерные подходы становятся популярны. Уже во второй половине XIX — начале XX вв. русскими земскими статистиками были сделаны успешные попытки многопризнаковых классификаций крестьянских хозяйств (А.П. Шликевич, С.А. Харизоменов, Н.Ф. Анненский, позднее — А.И. Хрящева), многими экономистами России и Западной Европы — классификаций промышленных предприятий, проведения анализа капитализации экономики и т.д. Здесь приведены далеко не все, а лишь наиболее известные исторические факты. Несомненно, что идеи многомерной оценки явлений и процессов эксплуатировались значительно шире. Все дело в том, что именно многомерный подход позволяет адекватно оценивать сложную природу естественных процессов и исследователю было необходимо обращаться к нему.

История МСА как науки с собственной теоретической базой и опытом экспериментальных исследований открывается в начале XX столетия. Собственно это «открытие» связывают с появлением в 1901 и 1904 гг. научных статей английских ученых К. Пирсона и Ч. Спирмена, посвященных теории факторного анализа. Первоначально методы МСА разрабатывались и широко применялись для исследований в области психологии и биологии, а позже — в медицине, военной промышленности, техническом проектировании и, наконец, в экономике. В числе наиболее ярких имен ученых, заложивших фундамент теории МСА: Л.Л. Терстоун, Л.Р. Такер, Р. Хорст, К. Холзингер, С. Барт, Г. Томсон, Т. Келли, Г. Кайзер, Д. Максвелл, С.Р. Рао, Г. Харманн (факторный анализ), Р. Трионон, Р. Льюис, Р. Сокал, Дж. Снит, Р. Сибсон, У. Уильямс, Т. Танимото, М. Жамбю (кластерный анализ), Дж. Б. Краскал, Р.В. Хемминг, Л. Гуттман (многомерное шкалирование), Р. Фишер, Т.В. Хейк, В.Р. Клекка, А. Барр (дискриминантный анализ), С. Райт (путевой анализ), Л. Заде (теория размытых множеств), Л. Гудман, П. Лазарсфельд, О. Андерсон, П. Махаланобис, С. Уилкс (многомерная математическая статистика). Внимательно рассматривая ретроспективу развития МСА, можно заметить, что период его становления и наиболее активного развития фундаментальной теории пришелся на 20—50-е годы, а в числе ученых-теоретиков МСА преобладают представители американской и английской математических школ:

Американская школа — факторный анализ, многомерное шкалирование, концепции новейшей статистической теории: размытых множеств, путевого анализа и т.п., многомерная математическая статистика;

Английская школа — факторный анализ; дискриминантный анализ, многомерный корреляционно-регрессионный анализ, многомерная математическая статистика;

Французская школа — кластер-анализ.

В советской статистике методы МСА получили распространение значительно позже, во второй половине 60-х и в 70-х годах. Такое опоздание было обусловлено длительным периодом неприятия формальной статистики и чрезмерной политизацией отечественной статистики вообще, когда внимание прежде всего уделялось проведению сплошных обследований и представлению, а не анализу данных, причем сами данные, даже в ущерб объективности, должны были соответствовать идеологическим концепциям государства. Известные трудности, однако, не помешали ученым СССР увидеть перспективу практического приложения методов МСА и внести значительный вклад в теоретические разработки. В этом заслуга таких известных отечественных статистиков, как А.Я. Боярский, С.А. Айвазян, П.Ф. Андрукович, А.М. Дубров, А.А. Френкель, И.И. Елисеева, А.И. Орлов, И.С. Енюков, Б.Г. Миркин, И.Д. Мандель, Л.Д. Мешалкин, В.М. Бухштабер, В.С. Мхитарян и многих других.

В настоящее время работа по освоению методов МСА и внедрению их в аналитическую практику становится особенно актуальной для экономистов и статистиков при переходе страны к рыночной экономике. Обусловлено это следующим.

- Переход к рыночной экономике сопровождается коренными изменениями структуры и связей народного хозяйства, нарастанием негативных кризисных явлений. Отслеживание и адекватная реакция на эти явления возможны при наличии системы объективного отражения и оперативной передачи достаточно полной микро- и макроуровневой информации — это позволяет решать именно МСА.

• Создание разнообразных форм собственности, появление большого числа новых предприятий, в том числе малого размера, заставляют, с одной стороны, обращаться к рациональным методам статистических обследований (выборкам, многомерным классификациям, устойчивому оцениванию и т.п.), а с другой — выдвигает повышенные требования к достоверности и оперативности информации. Частному предпринимателю уже не нужно «поддерживать идеологию», ему необходимы объективная картина развития бизнеса и обоснованный перспективный прогноз. При этом также предпочтительны многомерные данные, позволяющие комплексно описывать процессы и явления и реально их моделировать.

• В современных условиях особенно значимыми становятся исследования комплексов показателей, представляющих различные сферы общественной жизни: экономику, политику, экологию, социальную жизнь и т.д. Сегодня уже недостаточно просто планировать на каком-либо предприятии рост объемов производства или производительности труда. Необходимо одновременно оценивать, как это отразится, например, на экологической ситуации, социальном положении работников и т.п. Для размещения инвестиций необходимо знание не только экономического положения, но и реальной оценки уровней социальной, политической стабильности, экологичности размещаемых производственных объектов и т.д.

• Многомерные методы длительное время широко используются в практической статистике передовых стран Европы, Америки, Азии, где созданы и функционируют технологичные системы обработки и передачи многомерных данных в компьютерных сетях. Статистика страны с переходной экономикой включается в международные информационные системы интеграция при этом не может осуществляться без специалистов, обладающих адекватными знаниями по современным технологиям сбора и обработки больших массивов данных.

• Наконец, знания в области МСА представляются необходимыми для овладения новейшими методологическими разработками в области теоретической статистики, не говоря о ее развитии. Это ступень, на которой формируются базовые знания современной статистики; следуя за классической общей теорией статистики, она предваряет нетрадиционную статистику, статистику на размытых множествах, статистику катастроф и другие новейшие отрасли статистики.

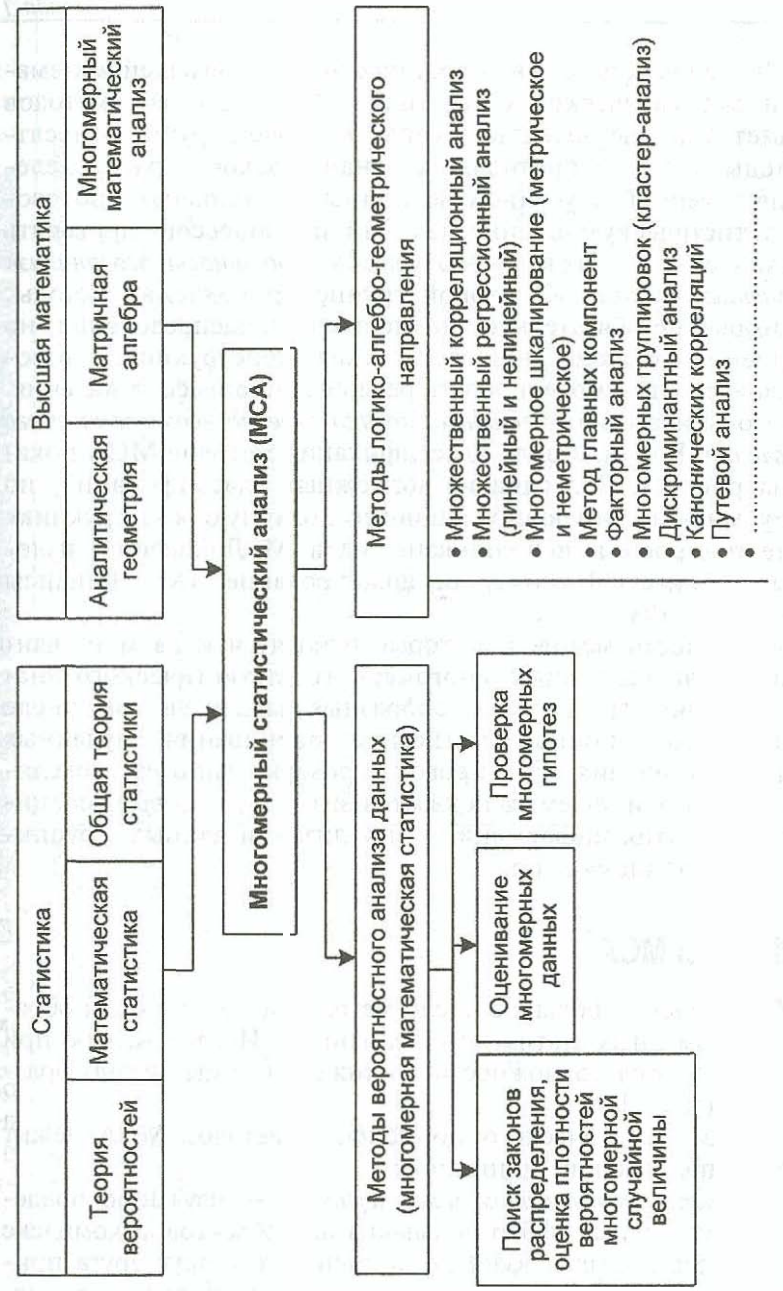


Рис. 1.1. Многомерный статистический анализ — наука о статистических методах

МСА основывается на теоретической базе высшей математики и математической статистики. Множество его методов разбивается на две большие группы. К первой группе относятся методы, которые предполагают знание законов распределения многомерной случайной величины и позволяют производить статистическую оценку явлений и процессов, проверять статистические гипотезы — это *методы вероятностного анализа многомерных данных*. Ко второй группе принадлежат методы, для которых не обязательно знание законов распределения, но существенна рациональная логическая конструкция, позволяющая адекватно моделировать реальные процессы и явления. Эти методы называют *методами логико-алгебро-геометрического направления*. В общем виде классификация методов МСА показана на рис. 1.1. Это одна из возможных классификаций, по нашему мнению, имеющих наиболее логичную конструкцию. При ее построении использованы идеи М. Дэйвисона, изложенные в книге «Многомерное шкалирование» (М.: Финансы и статистика, 1987).

Совокупность методов, которые относятся к двум названным выше направлениям многомерного статистического анализа, позволяют решать разнообразные задачи, в том числе присущие традиционной статистике: оценивания случайных величин, построения группировок, проверки гипотез, моделирования связей изучаемых показателей и т.п., и новые, специфические: сжатия информации, визуализации данных, группировки с «обучением» и т.д.

1.2. Методы МСА

МСА обобщает большое число методов и приемов для обработки многомерных статистических данных. Исследователю при этом открываются возможности достижения самых разнообразных целей (табл. 1.1).

В основе практического применения методов МСА лежит ряд основополагающих принципов:

1) *эффекта существенной многомерности* — изучению подлежит не произвольный набор признаков или объектов, а комплекс органично связанных и взаимно дополняющих друг друга признаков, которые позволяют полно и всесторонне оценивать явление (процесс). Признаки при оптимальном подборе не повторяют

Таблица 1.1. **Основные методы многомерного статистического анализа**

Метод	*Сокращенное обозначение*	*Тип задачи*	*Комментарии*
1	2	3	4
Статистического оценивания многомерной случайной величины	*СО МСВ*	Оценка параметров многомерной совокупности	Определение: многомерной средней, матрицы ковариаций, вероятностных оценок, робастное оценивание и т.п.
Проверки многомерных гипотез	*ПМГ*	Проверка гипотез о равенстве параметров многомерных совокупностей и соответствии некоторому закону распределения	—
Множественный корреляционно-регрессионный анализ	*МКРА*	Измерение и моделирование связей изучаемых признаков или объектов	—
Многомерное шкалирование	*МШ*	Визуализация данных, моделирование сложных систем	Представление данных в теоретическом пространстве, описание процессов и явлений, которые ввиду своей сложности или нестабильности не поддаются моделированию традиционными методами
Главных компонент	*МГК*	Сжатие данных	Сведение множества элементарных признаков к небольшому числу значимых «обобщенных признаков» и выявление латентных факторов

Продолжение табл. 1.1

1	2	3	4
Факторный анализ	*ФА*	*— " —*	Эта же задача может решаться относительно не только признаков, но и объектов
Многомерной группировки (кластерный анализ)	*КА*	*— " —* Группировка многомерных объектов (или признаков)	*— " —* —
Дискриминантный анализ	*ДА*	Группировка с «обучением»	Поиск эталонных групп, расклассификация новых объектов по известным эталонным группам
Канонических корреляций	*МКК*	Сжатие данных и моделирование связей обобщенных признаков	Устанавливается форма связи комплексов (наборов) зависимых переменных (Y_k) с независимыми факторными переменными (X_j). Y и X могут быть обобщенными признаками (см. ФА)
Многомерный дисперсионный анализ	*МДА*	Оценка и исследование дисперсий комплексов признаков	—
Многомерный ковариационный анализ	*МКА*	Оценка зависимости вариации результативного признака от факторного	Предполагает предварительную классификацию данных и поиск регрессионных связей для каждого класса. Затем вычисляются и анализируются оценки ковариаций (T_{xx}, T_{yy}, T_{yx}).

отдельных качественных характеристик, они рациональны по числу и четко структурированы по уровням представления явлений (процессов). Например, размеры предприятий можно характеризовать довольно большим числом признаков, но, ограничивая рамки исследования и допуская определенную грубость в выводах, вполне можно остановиться на трех важнейших признаках: объем производственных фондов, средняя численность работников, оборот капитала. Данную систему признаков неверно было бы дополнять оценками стоимости или мощности оборудования, числом инженерно-технических работников — это признаки более низкого уровня, их информация обобщена уже данными об объеме производственных фондов и средней численности всех работников;

2) *лаконичного описания наблюдаемых многомерных объектов.* Под этим понимается необходимость максимально сжатого и строго структурированного представления информации. В определенной мере такому требованию отвечает матричная форма записи, она экономична, остается доступной для прочтения и по ней можно легко определять структуру данных. Исходные данные часто обобщаются в виде матрицы значений признаков (в табл. 1.2), или симметрической матрицы с данными сравнений объектов (признаков), как это показано в табл. 1.3.

В качестве сравнительных характеристик (табл. 1.3) выступают величины соотношений некоторых количественных оценок или величины связей, теоретических расстояний между объектами;

3) *максимального использования «обучения» в настройке математических моделей*, т.е. использование информации, позволяющей наиболее точно идентифицировать изучаемые объекты, соотнести их с классом хорошо изученных явлений.

Использование обучающей информации значительно повышает достоверность статистических выводов, дает возможность рационально строить расчеты;

4) *оптимизационной формулировки задач МСА.* Имеется в виду рациональный выбор из всего арсенала методов МСА одного или нескольких дополняющих друг друга методов, которые при минимуме вычислительной работы позволили бы получить аналитические результаты с хорошей интерпретируемостью, достаточно полно и достоверно представляющие изучаемые явления (процессы). С целью упрощения выводов и оперативной проверки

Таблица 1.2. **Матрица значений аналитических признаков (X_j)**

Объект	X_1	X_2	X_3	X_4	...	X_m
n_1	x_{11}	x_{12}	x_{13}	x_{14}	...	x_{1m}
n_2	x_{21}	x_{22}	x_{23}	x_{24}	...	x_{25}
n_3	x_{31}	x_{32}	x_{33}	x_{34}	...	x_{35}
...	...	...	...	...	...	...
n_n	x_{n1}	x_{n2}	x_{n3}	x_{n4}	...	x_{n5}

Таблица 1.3. **Матрица теоретических расстояний между объектами (n_j)**

Объект	Объект				
	n_1	n_2	n_3	...	n_n
n_1	c_{11}	c_{12}	c_{13}	...	c_{1n}
n_2	c_{21}	c_{22}	c_{23}	...	c_{2n}
n_3	c_{31}	c_{32}	c_{33}	...	c_{3n}
...	...	...	...	...	...
n_n	c_{n1}	c_{n2}	c_{n3}	...	c_{nn}

их на адекватность реальным процессам методы МСА могут также применяться в комплексе с традиционными методами статистики. В качестве примера назовем здесь группировку объектов по значениям какого-либо одного обобщенного признака (латентного фактора) или индексный анализ по предварительно полученным данным факторного или дискриминантного анализа и т.д.

При реализации МСА можно выделить следующие основные этапы исследовательской работы:

- формулировка задачи исследования на предметно-содержательном уровне, определение объемов входной и выходной информации, формы выходных данных;
- определение последовательности обработки входной информации методами МСА. При этом ограничивается сам набор методов и уточняется порядок (чередование) их работы;
- сбор и систематизация исходной информации для последующей ее машинной обработки;

- предварительный анализ данных: их однородности, соответствия некоторым статистическим гипотезам, подчинения известным законам распределения, содержания грубых ошибок и т.д.;
- с учетом предыдущего этапа уточняется математическая постановка задачи и определяется возможность применения ранее отобранных методов МСА, в случае необходимости набор методов изменяется;
- проведение вычислений. Из-за трудоемкости методов МСА практически всегда эта работа планируется и выполняется при помощи вычислительной техники;
- результаты анализа сводятся и оцениваются на адекватность при помощи статистических критериев. Устанавливается непротиворечивость математических результатов и экономических выводов, оценивается степень интерпретируемости «выходных» данных;
- результаты исследования обобщаются в наглядных таблицах и на графиках, интерпретируются, формулируются окончательные выводы, даются практические рекомендации.

На практике все перечисленные этапы не обязательно присутствуют и четко разграничены. Некоторые из них могут объединяться или исключаться. Знание всех этапов, тем не менее, позволяет оптимально планировать реализацию методов МСА и учитывать предстоящие объемы работы.

1.3. Многомерное признаковое пространство. Особенности обработки многомерных статистических данных

Методы МСА базируются на геометрическом представлении данных. Наблюдаемые объекты располагаются в теоретическом пространстве размерностью, соответствующей числу признаков (элементарных или латентных), которыми они характеризуются. Можно предположить частные случаи признакового пространства: с нулевой размерностью — объекты не имеют характеристик; с единичной размерностью (одномерное признаковое пространство) — объекты отражаются значениями одного какого-либо признака; многомерное пространство — объекты

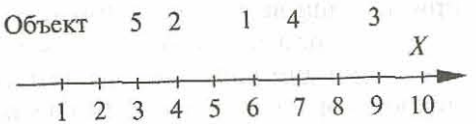
представлены значениями двух и более, до некоторого числа m, признаков (m-мерное признаковое пространство).

Рассмотрим простой пример, когда пять промышленных предприятий последовательно характеризуются значениями одного, двух и трех признаков:

I. Одномерное признаковое пространство

Предприятие № п/п	*Среднечасовой уровень выработки одного рабочего, долл. США (X)*
1	6
2	4
3	9
4	7
5	3

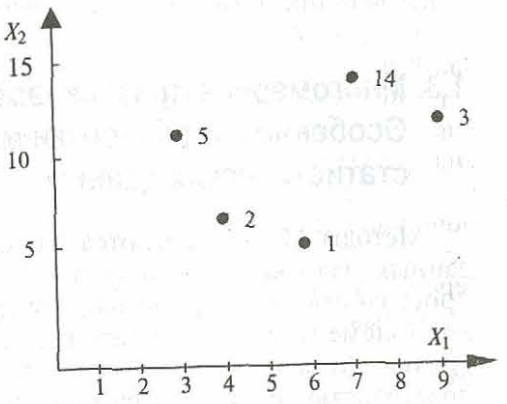
Его можно представить в виде одной градуированной шкалы:



II. Двумерное признаковое пространство.

Наблюдаемые объекты геометрически представляются на плоскости в двумерной (декартовой) системе координат:

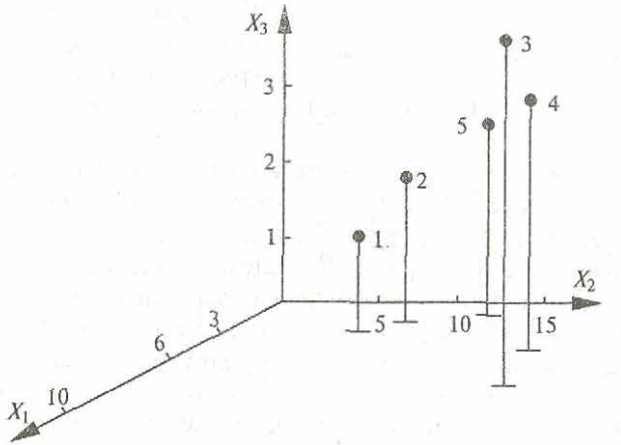
Предприятие № п/п	*Среднечасовой уровень выработки рабочего, долл. США (X_1)*	*Средний стаж работы на предприятии рабочего, лет (X_2)*
1	6	5
2	4	7
3	9	12
4	7	14
5	3	11



III. Трехмерное признаковое пространство.

Наблюдаемые объекты представляются в трехмерной системе координат:

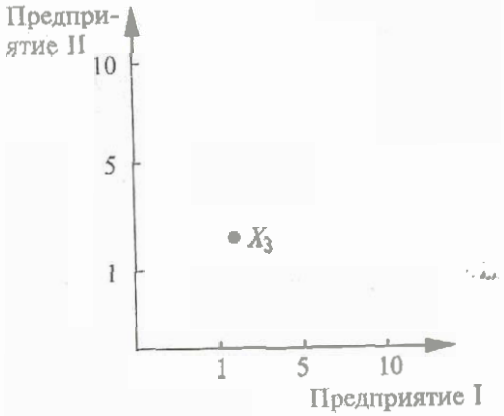
Предприятие № п/п	*Среднечасовой уровень выработки рабочего, долл. США (X_1)*	*Средний стаж работы рабочего на предприятии, лет (X_2)*	*Средний уровень квалификации работников по тарифному разряду (X_3)*
1	6	5	1,2
2	4	7	1,9
3	9	12	3,5
4	7	14	2,7
5	3	11	2,8



Характеристика и пространственное представление наблюдаемых объектов: предприятий, территорий, групп населения и т.д. по значениям признаков — это наиболее распространенная и привычная форма организации статистических данных. Однако в многомерной статистике возможны и достаточно часто встречаются случаи с другой организацией данных, когда оценочные признаки сами выступают в качестве наблюдаемых объектов и помещаются в теоретическое пространство предприятий, территориальных единиц и т.п.

Изменим в предыдущем примере исходное условие: пусть требуется характеристика признаков по предприятиям. Для наглядности отберем два первых предприятия и покажем возможность размещения их на координатной плоскости признаковых значений[1]:

[1] Вопросы приведения признаков к одному основанию в данном случае намеренно не рассматриваются.



Для работы с теоретическим пространством признаков или объектов применяются специальные обозначения. Выше уже использовались привычные символы для записи объектов (n_i, $i = \overline{1,n}$) и признаков (X_j, $j = \overline{1,m}$). Размерность пространства описывается при помощи сокращения французского слова dimension — размерность, как dim U, т.е. размерность пространства U. Частными случаями многомерного пространства являются 0-мерное пространство — пространство, не содержащее признаков, или объектов, dim $U_0 = 0$, и одномерное пространство dim $U_1 = 1$. В общем случае dim $U_m = m$.

В многомерном пространстве признаки или объекты имеют определенные количественные характеристики. Все принимаемые значения признаков (объектов) представляют собой множества вещественных чисел, и это множество обозначают символом R^m, где m по-прежнему указывает размерность пространства. Часто просто говорят R^m — пространство с размерностью m.

В аналитической работе при обращении к многомерному пространству признаков (объектов) принимаются во внимание следующие особенности:

• в m-мерном пространстве сохраняют силу принципиальные положения, аксиомы обычной евклидовой геометрии. Например, в прямоугольной системе координат углы между всеми парами осей составляют 90°; параллельные прямые, плоскости или гиперплоскости не пересекаются, если квадрат расстояния между двумя точками в двумерном пространстве (R^2) определяется по известной формуле Пифагора: $c^2 = a_1^2 + a_2^2$, то в многомерном пространстве — аналогичным образом:

$c^2 = a_1^2 + a_2^2 + a_3^2 + ... + a_m^2$ и т.д.;

• пространство, размерность которого превышает три, уже не может быть представлено визуально, и все задачи в этом случае решаются при помощи абстрактной логики и алгебраических методов;

• в многомерном анализе, как правило, используется большое число признаков, разнородных по своей природе. В связи с этим на первом этапе исследований обычно возникает проблема приведения всех анализируемых признаков к одному основанию — *сопоставимому виду*. Подобные проблемы решают нормированием данных, что геометрически означает изменение масштаба и другие преобразования координатной системы;

• обработка m-мерных совокупностей включает, как правило, большое число сложных и трудоемких для выполнения арифметических операций, поэтому осуществляется на основе пошаговых алгоритмов. Конечный результат при этом достигается последовательным решением отдельных, более мелких задач на вычисления;

• при работе с m-мерными данными совместно используются чисто математические, абстрактные методы и методы экономико-статистического анализа, ориентированные на конкретные сферы приложения. Следует обращать внимание на непротиворечивость результатов, получаемых различными методами. Возникающие противоречия указывают на нарушения логики решения экономической задачи и становятся источником ошибочных выводов.

Вопросы

1. Что является предметом МСА?
2. В чем заключаются особенности методов МСА и их отличие от методов классической статистики?
3. Перечислите методы МСА и задачи, решаемые при помощи этих методов.
4. Назовите основные исторические этапы развития методов МСА.
5. Что является методологической и теоретической основой МСА?
6. Сформулируйте понятие признакового пространства. Приведите примеры одномерного, двумерного и многомерного признакового пространства.

Глава 2

Элементы математики в МСА

2.1. Основы аналитической геометрии

2.1.1. Введение

Понятие вектора вначале появилось в физике и использовалось для представления величин силы, скорости, ускорения[1]. Приложение силы в некоторой точке, например, определялось отрезком (вектором), направление которого указывало на направление действия силы, а длина — на величину этой силы. Соответствующим образом векторное отображение использовалось и для представления скорости и ускорения элементарных частиц. Умножение вектора на реальное число указывало при этом на величину положительного или отрицательного прироста силы (скорости, ускорения). Результатом умножения были так же сила, скорость, ускорение.

Несколько сложнее было представление операции сложения сил (скоростей, ускорений): в случае разнонаправленности их действия изменялись не только величина, но и направление результирующего вектора. Вспомним тривиальный пример сложения сил по так называемому правилу параллелограмма.

Выполнение операций сложения векторов, как и умножения, не изменяло природы анализируемых признаков, их результатами всегда оставались сила (скорость, ускорение), только с измененной величиной.

Позднее понятие вектора нашло широкое поле приложения и в экономике. Часто векторное представление используется здесь для характеристики цен и товарных потоков. В отличие от

[1] Понятие вектора впервые дано в работах немецкого математика Г. Гроссмана и ирландского математика У. Гамильтона в XIX в. — *Прим. пер.*

физики в экономике обычно направление вектора не имеет принципиального значения, более важными видятся его компонентный состав, общая математическая структура, обусловливающие форму визуализации и определенный набор эконометрических приемов для аналитической обработки данных.

2.1.2. Основные определения

Обозначим V любое непустое множество объектов, а R — множество действительных чисел. Ясно, что объекты из V могут объединяться, аналитически это представляется при помощи операции сложения (+), говорят еще, что объекты аддитивны. Сложение, или аддитивность, объектов имеет следующие основные свойства:

пусть $x, y, z \in V$, тогда

$(x+y)+z = x+(y+z)$ — *ассоциативный закон*; (2.1)

если имеется элемент $0 \in V$, такой, что $x+0=0+x=x$, то его называют *нейтральным* или *нуль-элементом* в V; (2.2)

для каждого $x \in V$ всегда имеется элемент $x^{-1} \in V$, с которым $x+x^{-1}=x^{-1}+x=0$. Такой элемент (x^{-1}) называют *обратным* к x; (2.3)

$x+y=y+x$ — *коммутативный закон*. (2.4)

Взаимодействие и изменение объектов в V может кроме операции сложения определяться операцией умножения векторов на некоторое действительное число или скалярным произведением ($\cdot$), имеющим следующие свойства:

$\alpha \cdot (\beta \cdot x) = (\alpha \cdot \beta) \cdot x$ (ассоциативный закон); (2.5)

$1 \cdot x = x \cdot 1 = x$ (нейтральный элемент); (2.6)

$\alpha \cdot (x + y) = \alpha \cdot x + \beta \cdot y$ (первый дистрибутивный закон); (2.7)

$(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$ (второй дистрибутивный закон). (2.8)

Множество V, в котором определены операции сложения (+) и умножения на скаляр ($\cdot$), а выполнение этих операций удовлетворяет свойством (2.1) — (2.4) и (2.5) — (2.8), называют *векторным пространством*. Элементы множества V есть векторы.

2.1.3. Арифметическое векторное пространство

Пусть имеются $x=(x_1,x_2,...,x_n)$ и $y=(y_1,y_2,...,y_n)$, $n\in N$ и $x_i,y_i\in R$ $(i=1,2,...,n)$ — два действительных n-вариантных числа, обладающих аддитивностью, тогда операцию покомпонентного сложения векторов можем записать:

$$x+y=(x_1+y_1,x_2+y_2,...,x_n+y_n),$$

по аналогии скалярное произведение будет

$$\alpha_s x=(\alpha x_1,\alpha x_2,...,\alpha x_n),\ \alpha\in R$$

и векторное пространство для x:

$$V=\{x\mid x=(x_1,x_2,...,x_n),\ n\in N,\ x_i\in R,\ i=1,...,n\}.$$

Векторное пространство V называют арифметическим векторным пространством и обозначают R^n. Такое пространство включает подмножества (системы) из n действительных чисел и сохраняет силу следующих утверждений:

- в V действуют свойства (2.1) — (2.4), так как действительные числа обладают аддитивностью и ассоциативностью, и свойства (2.5) — (2.8) — согласно правилам выполнения арифметических операций над действительными числами;
- нулевой элемент в V имеет n нулевых компонент $(0,0,...,0)$;
- для $x=(x_1,x_2,...,x_n)$, $x\in V$, всегда существует отрицательный вектор $-x=(-x_1,-x_2,...,-x_n)$.

Особым случаем для пространства V является $R^1=R$, т. е. пространство одного вектора или пространство над простым множеством действительных чисел.

2.1.4. Непрерывные функции

Пусть $V=\{f:[a,b]\to R\mid f$ непрерывна на интервале $[a,b]\}$.

Аддитивность непрерывных функций и их скалярное произведение определяются соответствующими равенствами:

$(f+g)(x)=f(x)+g(x)$ для $f,g\in V$; $(\alpha_s f)(x)=\alpha f(x)$.

Для суммы непрерывных функций имеют силу свойства (2.1) — (2.4). Это следует непосредственно из сохранения правил выполнения арифметических действий над действительными числами и из того, что сумма непрерывных функций есть также непрерывная функция. В качестве нейтрального элемента здесь выбирается нуль-функция:

$0:[a,b]\to R$, с $0(x)=0$ для всех $x\in[a,b]$, тогда $(f+0)(x)=f(x)+0(x)$.

Результат обращения функции будет $-f(x)$, при этом $(f+(-f))(x)=f(x)-f(x)=0=0(x)$ для всех $x\in[a,b]$.

Свойства (2.5) — (2.8) для скалярного произведения функций также определяются правилами выполнения арифметических действий над действительными числами и той особенностью, что в результате многократного изменения непрерывной функции получают вновь непрерывную функцию.

Один, не очень серьезный, пример на множества

Этот пример показывает, что понятие аддитивности имеет важное значение и может применяться весьма широко.

Пусть имеем множество V {кошка, собака, мышка}. Элементы этого множества принадлежат трем различным видам и находятся в следующей взаимосвязи:

а) если собака встретит кошку, то она съест кошку и превратится в мышку;

б) если собака встретит собаку, то одна собака съест другую собаку и превратится в кошку;

в) если собака встретит мышку, то она съест мышку и останется собакой;

г) если кошка встретит кошку, то одна из них съест другую и превратится в собаку;

д) **если кошка встретит** мышку, то она съест мышку и останется **кошкой**;

е) если мышка встретит мышку, то одна мышка съест другую и останется мышкой.

Легко заметить, что в этом замечательном зверинце количество корма для животных каждого вида определяется межвидовой иерархией. Правила «поедания» (а) — (е) можно рассматривать в конкретном векторном пространстве V, обладающем свойством аддитивности:

$$C+K=M;\ C+C=K;\ C+M=C;\ K+K=C;\ K+M=K;\ M+M=M.$$

Следуя аксиоматичному правилу (2.1), можем записать:

$$(C+C)+K=K+K=C,\qquad (C+M)+M=C+M=C,$$
$$C+(C+K)=C+M=C,\qquad C+(M+M)=C+M=C,$$
$$(C+K)+K=M+K=K,\qquad (K+K)+M=C+M=C,$$
$$C+(K+K)=C+C=K,\qquad K+(K+M)=K+K=C,$$
$$(C+C)+M=K+M=K,\qquad (K+M)+M=K+M=K,$$
$$C+(C+M)=C+C=K,\qquad K+(M+M)=K+M=K.$$

Относительно других свойств аддитивности отметим:

(2.2) — нейтральным элементом в V является мышка, так как каждый из зверей, съедающий мышку, остается неизменным, т.е. сохраняет свою видовую принадлежность;

(2.3) — обратным элементом для собаки является кошка, а для кошки — собака. Для мышки (нейтрального элемента) обратным элементом множества будет тоже мышка;

(2.4) — это аксиоматичное правило также сохраняется, так как результат встречи зверей не зависит от того, в какой последовательности она происходит, скажем, собака встречает кошку или кошка встречает собаку.

Очевидно, что в примере действуют также свойства (2.5) и (2.6), кроме того

$$\alpha \dot{S} (C+K)=\alpha \dot{S} M \quad \text{кратное } \alpha \text{ — увеличение корма свидетельствует об увеличении популяции мышей,} \quad (2.7)$$

$$(\alpha+\beta) \dot{S} (C+K)=\alpha \dot{S} C+\beta \dot{S} C. \quad (2.8)$$

2.1.5. Понятие подпространства

Подмножество U векторного пространства V называют векторным подпространством V, если в U как самостоятельном векторном пространстве из V, выполнимы операции сложения и скалярного произведения векторов.

Пусть U — векторное подпространство V, тогда существующий в V нулевой вектор $0=(0,0,..,0)$ является определенным нуль-вектором и для U. Подобный вывод следует из равенства: $x+0=x$ для всех $x \in U$.

Принимая $0 \in U$, заметим, что любое подпространство в V есть непустое подмножество V.

Требование выполнения операций сложения и скалярного произведения векторов в V, а также в U позволяет сформулировать основной критерий существования векторного подпространства: подмножество U векторного пространства V будет подпространством V, если выполняются одновременно следующих два условия:

1) U — непустое множество

2) Для всех $x,y \in U$, $\alpha \in R$, действует $x+y \in U$ и $\alpha x \in U$, т. е. U замкнуто относительно операций сложения векторов и умножения вектора на число.

$x \in U$

□ *Доказательство:*

а. Пусть U есть векторное подпространство в V. Тогда если $U \neq \emptyset$, то U содержит по крайней мере нулевой вектор и U замкнуто относительно сложения и скалярного произведения векторов в соответствии со свойствами векторного пространства.

б. Пусть $U \neq \emptyset$ и замкнуто относительно сложения и скалярного произведения векторов. Тогда, с учетом, что $U \subset V$, для всех элементов U имеют силу свойства (2.1) — (2.4) и (2.5) — (2.8). Чтобы при этом показать принадлежность подпространства U к пространству V, необходимо подтвердить наличие свойств (2.2) и (2.3). Но так как $U \neq \emptyset$, можем записать $x \in U$, одновременно принимая условие (2): $0 \cdot x = 0 \in U$, а это означает, что вектор $\mathbf{0}$ есть нулевой вектор и, кроме того, нейтральный элемент в U, т. е. действует свойство (2.2). Далее, для каждого $x \in U$ существует $(-1) \dot{S} x = -x \in U$, другими словами, для каждого $x \in U$ имеется обратный элемент $(-x)$ в U, т.е. действует свойство (2.3). ■

2.1.6. Образующие системы

Линейные комбинации. Пусть V есть векторное пространство и имеются $x_1,...,x_n$ — элементы V. Тогда вектор вида:

$$y=\sum_{i=1}^{n} \alpha_i x_i, \ \alpha_i \in R, \ n \in N,$$

называют линейной комбинацией векторов $x_i \in V$. Говорят также, что y представлен линейной комбинацией векторов x_i.

Пример 2.1. В двумерном пространстве R^2 имеется вектор $y=\begin{pmatrix}1\\6\end{pmatrix}$, представляющий собой линейную комбинацию трех других векторов

$$x_1=\begin{pmatrix}1\\2\end{pmatrix}, \quad x_2=\begin{pmatrix}0\\1\end{pmatrix} \quad \text{и} \quad x_3=\begin{pmatrix}0\\2\end{pmatrix}, \quad \text{или} \quad y=x_1-2x_2+3x_3, \quad \text{а также}$$

$y=x_1+4x_2$.

Очевидно, что, кроме двух приведенных, существует еще достаточно большое число других возможных линейных комбинаций векторов x_1,x_2,x_3 для y, а также других линейных комби-

наций векторов x_1,x_2,x_3 в R^2 для любого элемента, кроме y. Принимая это во внимание, введем понятие линейной оболочки.

Линейной оболочкой некоторого векторного пространства называют множество векторов: $x_1,...,x_n \in V$. Линейную оболочку обозначают: $[x_1,...,x_n]$.

Утверждение: если имеется некоторое подмножество M векторного пространства V, $M \subset V$ и $[M]$ — линейная оболочка для векторов из M, то $[M]$ есть подпространство V.

□ *Доказательство*:

При условии, что $M \neq \emptyset$, а также $[M] \neq \emptyset$, для каждой из линейных комбинаций:

$$y = \sum_{i=1}^{n} \alpha_i x_i, \alpha_i \in R \quad \text{с} \quad x_i \in M \quad \text{и некоторым определенным} \quad \lambda \in R$$

имеет силу $\lambda y = \sum_{i=1}^{p} (\lambda\alpha_i) x_i \in [M]$, очевидно, что λy с $\beta_i := \lambda\alpha_i$ тоже будет линейной комбинацией векторов x_i.

Одновременно можем записать расширение числа линейных комбинаций (z) для векторов из M:

$$z = \sum_{i=1}^{n} \gamma_i x_i, \ \gamma_i \in R,$$

$$y + z = \sum_{i=1}^{n} \alpha_i x_i + \sum_{i=1}^{n} \gamma_i x_i = \sum_{i=1}^{n} (\alpha_i + \gamma_i) x_i.$$

В конечном счете вновь получаем линейную комбинацию $x_i \in M$. Таким образом, $[M]$ согласно критерию существования векторного подпространства является подпространством в V. ■

Линейная оболочка обладает следующими свойствами:

1) $M \subset [M]$;
2) $M_1 \subset M_2 \Rightarrow [M_1] \subset [M_2]$;
3) $[M] = M \Leftrightarrow M$ есть подпространство V;
4) $[[M]] = [M]$.

Определение: Пусть U — векторное подпространство V, тогда если имеется такое множество M, что $[M] = U$, то его называют *образующим множеством* U.

Обычно всегда $M = U$, и целесообразно рассматривать возможно наименьшее образующее множество подпространства

U. Образующее множество M векторного подпространства U называют минимальным, если не существует другого действительного подмножества $M' \subset M$, для которого имело бы силу $[M'] = [M] = U$.

Линейная зависимость. Линейной оболочке для множества векторов $x_1,x_2,...,x_n$ принадлежат все линейные комбинации $x_1,x_2,...,x_n$, в том числе и нулевой вектор:

$$0 = \sum_{i=1}^{n} 0 \cdot x_i .$$

Так как все коэффициенты подобной линейной комбинации — нули, то говорят о тривиальном представлении нулевого вектора. Последнее, нетрудно заметить, возможно всегда.

Напротив, нетривиальное представление нулевого вектора при заданных $x_1,x_2,...,x_n$, когда все $x_i \neq 0$, возможно не всегда:

$$0 = \sum_{i=1}^{n} \alpha_i x_i, \ \alpha_i \in R, \text{ и не обязательно все } \alpha_i = 0.$$

Если векторы $x_1,x_2,...,x_n$ позволяют нетривиально представить нулевой вектор, то их называют линейно-зависимыми. Соответственно линейно независимыми векторы $x_1,x_2,...,x_n$ будут при условии:

$$0 = \sum_{i=1}^{n} \alpha_i x_i \text{, только когда } \alpha_1 = \alpha_2 = ... = \alpha_n = 0.$$

Пример 2.2. Рассмотрим три приведенных ранее вектора:

$$x_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad x_3 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Все три вектора, а также пара векторов x_2,x_3 линейно-зависимы. Векторы x_1,x_2 и x_1,x_3 являются попарно линейно независимыми.

Для любой системы векторов правомерны следующие высказывания:

1) два вектора всегда линейно-зависимы, если они взаимно пропорциональны;

2) векторы $x_1,x_2,...,x_n$ будут линейно-зависимыми, если из них хотя бы один вектор представляет линейную комбинацию других векторов;

3) если в системе векторов $x_1, x_2, ..., x_n$ существует нулевой вектор, то все эти векторы линейно-зависимы;

4) $p+i$ линейных комбинаций из p числа (любых) векторов — всегда будут линейно-зависимыми.

Базис и размерность. Минимальное образующее множество векторного пространства V называют его *базисом*.

Пример 2.3. а) Зададим базис арифметического векторного пространства R^n системой так называемых собственных векторов:

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad ..., \quad x_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Очевидно, можно любой предполагаемый вектор $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ представить в виде линейной комбинации x_i, т. е.

$$y = \sum_{i=1}^{n} \alpha_i x_i, \ \alpha_i = y_i.$$

б) Рассмотрим векторное пространство всех полиномов до полинома высшей степени $n \in N$:

$$V = \left\{ p_n : R \to R,\ p_n(x) = b_0 + b_1 x + b_2 x^2 + ... + b_n x^n \mid \alpha_i \in R,\ i = 1,...,n \right\},$$

базисом для V будет множество $B \in \left\{ x^i \mid i \in \{0,1,2,...,n\} \right\}$.

Свойства базиса. 1. Базис B для пространства V представляет линейно независимую систему векторов.

Предположим, что B — это система линейно-зависимых векторов, т.е. имеется по крайней мере один вектор $x_p \in B$, который можно представить как линейную комбинацию других векторов базиса. Но тогда $\left[B \setminus \{x_p\}\right] = [B]$ и B не есть минимальное образующее множество. Последнее означает, что любое образующее множество M пространства V, включающее линейно независимые векторы, является базисом для V.

2. Если B есть базис $V \neq 0$, то любой вектор $y \in V$ может быть разложен по этому базису и представлен в виде:

$$y = \sum_{i=1}^{n} \alpha_i x_i, \quad x_i \in B \quad (i = 1,...,n).$$

И, наоборот, если имеется любой вектор $y \in V \neq 0$, который может быть представлен линейной комбинацией векторов из подмножества B и $B \subset U$, то B — это базис V.

Пусть векторное пространство $V \neq 0$ и B есть базис для V, тогда размерностью V ($\dim V$) называют число векторов базиса (B).

Если имеется U — некоторое подмножество V, то действуют следующие правила:

$\dim U \leq \dim V$,

$\dim U = \dim V \Leftrightarrow U = V$.

Пример 2.4. а) Для трехмерного пространства (R^3) базис будет иметь форму:

$$B = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

и размерность, равную трем: $\dim R^3 = 3$.

б) Вернемся к векторному пространству всех полиномов, высшая степень полинома равна n. Размерность этого пространства будет $n+1$, так как его базис

$$B = \left\{ 1, x, x^2, ..., x^n \right\}.$$

2.2. Элементы матричной алгебры

2.2.1. Трансформация базиса

Пусть V — это n-мерное векторное пространство, а $B = \{x_1, ..., x_n\}$ и $B' = \{x'_1, ..., x'_n\}$ — два базиса V. Тогда можем x'_i — элементы из базиса B' представить в виде линейной комбинации векторов из B:

$$x'_i = \sum_{j=1}^{n} \alpha_{ij} x_j, \ i = 1,...,n; \ \alpha_{ij} \in R, \tag{2.9}$$

и наоборот, x_j — элементы базиса B записываются как линейная комбинация векторов из B':

$$x_j = \sum_{i=1}^{n} \beta_{ji} x'_i, \ j = 1,...,n; \ \beta_{ji} \in R. \tag{2.10}$$

Подставим (2.10) в (2.9) и получим[1]:

$$x_i' = \sum_{k=1}^{n}\left(\sum_{j=1}^{n}\alpha_{ij}\beta_{jk}\right)x_k',$$

где $\sum_{j=1}^{n}\alpha_{ij}\beta_{jk} = \delta_{ik} = \begin{cases}0, & i \neq k\\ 1, & i = k.\end{cases}$ (2.11)

Аналогичный результат получают и после подстановки (2.9) в (2.10):

$$x_j' = \sum_{k=1}^{n}\left(\sum_{i=1}^{n}\beta_{ji}\alpha_{ik}\right)x_k \quad \text{и} \quad \sum_{i=1}^{n}\beta_{ji}\alpha_{ik} = \delta_{jk}. \qquad (2.12)$$

Как видим, уравнения вида (2.9) и (2.10) позволяют заменить базис B на B', и наоборот.

Множества коэффициентов α_{ij} и β_{ji} в развернутом виде представляются элементами квадратных матриц A и B, которые называют также матрицами линейного оператора:

$$\begin{pmatrix}\alpha_{11} & \alpha_{12} & \dots & \alpha_{1n}\\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n}\\ \vdots & \vdots & \vdots & \vdots\\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn}\end{pmatrix} = A \quad \text{и} \quad \begin{pmatrix}\beta_{11} & \beta_{12} & \dots & \beta_{1n}\\ \beta_{21} & \beta_{22} & \dots & \beta_{2n}\\ \vdots & \vdots & \vdots & \vdots\\ \beta_{n1} & \beta_{n2} & \dots & \beta_{nn}\end{pmatrix} = B$$

В матрице A, например, строки состоят из элементов $\alpha_{i1},\dots,\alpha_{in}$, а столбцы — из элементов $\alpha_{1j},\alpha_{2j},\dots,\alpha_{nj}$.

В дальнейшем мы рассмотрим понятие матрицы и различные ее формы.

Пример 2.5. Имеется векторное пространство V с размерностью, равной 2, т. е. $\dim V = 2$; для этого пространства существует базис $B = \{x_1, x_2\}$, векторы которого позволяют переходить к другим базисам V, скажем

$$x_1' = x_1 - x_2,$$

$$x_2' = x_1 + x_2.$$

Матрица линейного оператора для приведенной пары уравнений принимает вид:

$$A = \begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}.$$

[1] После подстановки индексация x изменяется с j на k. — *Прим. автора.*

Система уравнений, соответствующая равенству (2.10):

$$x_1 = \frac{1}{2}x_1' + \frac{1}{2}x_2';$$

$$x_2 = -\frac{1}{2}x_2' + \frac{1}{2}x_2',$$

и матрица $B = \begin{pmatrix}\frac{1}{2} & \frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{pmatrix}$.

2.2.2. Расчеты в матрицах

Матрица представляет собой обычную таблицу, удобную для систематизации количественных данных, не более того. Порядок выполнения действий над матрицами предполагает, однако, что столбцы и строки матрицы — это не простые наборы чисел, а определенные системы векторов.

Соответственно представлениям о векторном пространстве рассмотрим операции сложения и скалярного произведения матриц. Примем обозначения пусть в любой матрице будет n строк и m столбцов. Когда $n = m$, матрица имеет равное число строк и столбцов и называется ***квадратной***, а если $n \neq m$, матрица *прямоугольная*. Совокупность элементов матрицы образует множество, поэтому для любой из них можем записать: $A \in R^{n,m}$. При выполнении операции сложения необходимо, чтобы слагаемые матрицы имели одинаковую размерность, скажем имеется $A \in R^{n,m}$ и $B \in R^{n,m}$:

$$A = \begin{pmatrix}a_{11} & a_{12} & \dots\\ a_{21} & a_{22} & \dots & a_{2m}\\ \vdots & \vdots & \vdots & \vdots\\ a_{n1} & a_{n2} & \dots & a_{nm}\end{pmatrix}, \quad B = \begin{pmatrix}b_{11} & b_{12} & \dots & b_{1m}\\ b_{21} & b_{22} & \dots & b_{2m}\\ \vdots & \vdots & \vdots & \vdots\\ b_{n1} & b_{n2} & \dots & b_{nm}\end{pmatrix}.$$

Сумма этих матриц образуется покомпонентным сложением элементов $(a_{ij} + b_{ij})$:

$$A + B = \begin{pmatrix}a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1m}+b_{1m}\\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2m}+b_{2m}\\ \vdots & \vdots & \vdots & \vdots\\ a_{n1}+b_{n1} & a_{n2}+b_{n2} & \dots & a_{nm}+b_{nm}\end{pmatrix}.$$

Скалярное произведение матрицы и любого действительного числа (α) также находится путем умножения элементов исходной матрицы на скаляр α:

$$\alpha \cdot A := \begin{pmatrix} \alpha a_{11} & \alpha a_{12} & \dots & \alpha a_{1m} \\ \alpha a_{21} & \alpha a_{22} & \dots & \alpha a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha a_{n1} & \alpha a_{n2} & \dots & \alpha a_{nm} \end{pmatrix}.$$

При помощи двух описанных операций — сложения и скалярного произведения — получают все возможные матрицы ($n \times m$) определенного векторного пространства. Докажем, что в таком векторном пространстве действуют известные правила (2.1) — (2.4) и (2.5) — (2.8).

□ *Доказательство:*

Выполнение аксиоматичных правил (2.1) и (2.4) следует автоматически из свойств ассоциативности и коммутативности операции сложения действительных чисел. (2.2) также выполнимо с нулевой матрицей, а сама 0-матрица при этом выступает как нейтральный элемент:

$$0 = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \text{ — нейтральный элемент.}$$

Правило (2.3) выполняется, так как элементы любой матрицы $-A$ представляют простую инверсию элементов A: знаки у элементов A в $-A$ изменены на противоположные.

Очевидно также, что и (2.5) — (2.8) сохраняют свою силу. ■

Один из базисов рассматриваемого векторного пространства образует система:

$$\left\{ \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \right\}.$$

Эта система состоит из $n \times m$ элементов. Ясно, что размерность самого векторного пространства для всех (n,m)-матриц будет равна произведению $n \times m$.

Остается рассмотреть операцию перемножения матриц. Она производится уже не покомпонентным способом. Пусть имеется

A — матрица размерностью $n \times m$ и B — матрица размерностью ($m \times p$), тогда произведение $A \cdot B$ найдем следующим образом:

$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj} \quad (i = 1,2,\dots,n;\ j = 1,2,\dots,p).$$

Результатом перемножения A и B является матрица C размерностью $n \times p$. Операция умножения матриц выполнима при условии согласованности перемножаемых матриц по числу строк и столбцов: число столбцов в умножаемой матрице должно быть равно числу строк в матрице-множителе. Порядок выполнения операции умножения матриц предопределяет ее некоммутативность, т. е. $AB \neq BA$.

Пример 2.6. Обратимся к примеру из области экономики и рассмотрим процесс товарного производства. Данный процесс характеризуется последовательным прохождением этапов: подготовка производства (сырье, материалы, техника, рабочая сила), выпуск промежуточного и конечного продуктов. Обозначим вектор факторов производства $v' = (v_1, v_2, v_3)$, промежуточный продукт — $z' = (z_1, z_2)$, конечный продукт — $e' = (e_1, e_2)$. Представим схематично общий процесс производства (рис. 2.1).

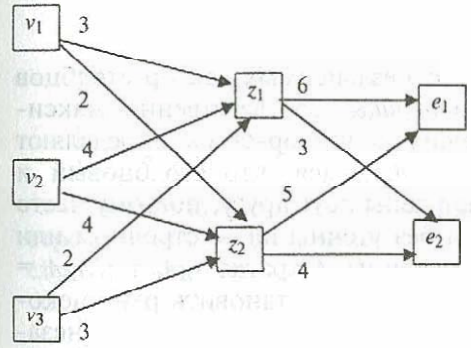


Рис. 2.1

На рисунке цифры над стрелками указывают конкретные объемные величины факторов производства и произведенного продукта. В конечном счете из производства получают конечного продукта: $e_1 = 3$ и $e_2 = 4$ единицы. Определим, какие входные характеристики должен содержать вектор $v' = (v_1, v_2, v_3)$, чтобы был обеспечен выпуск конечного продукта объемом e_1 и e_2.

Пусть имеем:

$$A = \begin{pmatrix} 6 & 3 \\ 5 & 4 \end{pmatrix} \text{ и } B = \begin{pmatrix} 3 & 2 \\ 4 & 4 \\ 2 & 3 \end{pmatrix}, \text{ а также } e = \begin{pmatrix} 3 \\ 4 \end{pmatrix} \text{ (рис. 2.1),}$$

тогда из произведения матриц найдем:

$$z = Ae = \begin{pmatrix} 30 \\ 31 \end{pmatrix} \text{ и } v = BAe = \begin{pmatrix} 3 & 2 \\ 4 & 4 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 30 \\ 31 \end{pmatrix} = \begin{pmatrix} 152 \\ 44 \\ 153 \end{pmatrix}.$$

2.2.3. Ранг матрицы и ее элементарные преобразования

Будем рассматривать строки и столбцы матрицы как некоторые векторы, тогда можем сказать, что матрица размерностью $n \times m$ состоит из n строчных векторов, компонентами которых являются данные m столбцов или, наоборот, матрица с числом векторов m, каждый с компонентами n из строчных данных. Другими словами, любая матрица A объединяет m столбцовых и n строчных векторов:

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}, \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix}, \ldots, \begin{pmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{pmatrix} \quad \begin{matrix} (a_{11}a_{12}\ldots a_{1m}), \\ \text{или } (a_{21}a_{22}\ldots a_{2m}), \\ (a_{n1}a_{n2}\ldots a_{nm}). \end{matrix}$$

Максимальное число линейно независимых вектор-столбцов называют *столбцовым рангом матрицы*, соответственно максимальное число линейно независимых вектор-строк определяют как *строчный ранг A*. Далее мы убедимся, что столбцовый и строчный ранги матрицы всегда равны друг другу, поэтому часто говорят просто о ранге матрицы без уточнения — строчном или столбцовом. Записывают ранг матрицы A кратко $\mathrm{rg}A$, т.е. $\mathrm{rg}A$ = = ранг матрицы A. Чтобы количественно установить ранг некоторой матрицы A, необходимо определить число линейно независимых вектор-строк или число линейно независимых вектор-столбцов. Мы ограничимся действиями над строками матрицы и будем использовать следующие возможные для них элементарные преобразования:

1) замена i-й строки в матрице i-й строкой, умноженной на некоторое действительное число α;

2) замена i-й строки в матрице строкой, полученной в результате суммирования i-й и j-й строк, $i \neq j$, $i, j \in \{1,\ldots, n\}$.

При помощи преобразований 1) и 2) любая матрица приводится к виду, при котором все линейно-зависимые вектор-строки обнаруживают себя как нулевые векторы.

Пример 2.7. Определим ранг матрицы A:

$$A = \begin{pmatrix} 1 & 1 & -3 & 1 \\ 2 & 1 & 1 & -1 \\ 0 & 2 & -13 & 1 \\ 2 & -1 & 14 & -2 \end{pmatrix}.$$

Произведем двойное вычитание первой строки из второй и четвертой строк матрицы A, первая и третья строки матрицы остаются при этом без изменения:

$$\begin{pmatrix} 1 & 1 & -3 & 1 \\ 0 & -1 & 7 & -3 \\ 0 & 2 & -13 & 1 \\ 2 & -3 & 20 & -4 \end{pmatrix}.$$

Последовательно сложим вторую строку с первой и умножим вторую строку на (-1):

$$\begin{pmatrix} 1 & 0 & 4 & -2 \\ 0 & 1 & -7 & 3 \\ 0 & 2 & -13 & 1 \\ 2 & -3 & 20 & -4 \end{pmatrix}.$$

Вновь произведем двойное вычитание, на этот раз из третьей строки второй, и, кроме того, троекратное сложение второй строки с четвертой:

$$\begin{pmatrix} 1 & 0 & 4 & -2 \\ 0 & 1 & -7 & 3 \\ 0 & 0 & 1 & -5 \\ 0 & 0 & -1 & 5 \end{pmatrix}.$$

Сложим третью строку с четвертой:

$$\begin{pmatrix} 1 & 0 & 4 & -2 \\ 0 & 1 & -7 & 3 \\ 0 & 0 & 1 & -5 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Очевидно, что при помощи элементарных преобразований другие строки матрицы A, кроме четвертой, не могут быть обращены в ноль. Следовательно, число линейно независимых векторов равно трем и соответственно $\operatorname{rg} A = 3$.

Покажем теперь, что столбцовый и строчный ранги любой матрицы равны между собой. Пусть имеем матрицу A со столбцовым рангом z и строчным рангом s. Посредством элементарных преобразований получим из матрицы A некоторую трансформированную матрицу A^* со строчным рангом z^* и столбцовым рангом s^* следующего вида:

$$A^* = \begin{pmatrix}
0 & \dots & 0 & 1 & * & \dots & * & 0 & * & \dots & * & 0 & & 0 & * & \dots & * \\
\vdots & & \vdots & 0 & 0 & \dots & 0 & 1 & * & \dots & * & 0 & & 0 & \vdots & & \vdots \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & 0 & 0 & \dots & 0 & 1 & & \vdots & \vdots & & \vdots \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & 0 & \ddots & \vdots & \vdots & & \vdots \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & 0 & * & \dots & * \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & 1 & * & \dots & * \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & 0 & 0 & \dots & 0 \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\
0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & & 0 & 0 & \dots & 0
\end{pmatrix}.$$

В матрице A^* z' — число тех строк, которые не равны нулю, т.е. z' — это число линейно независимых векторов в A^*, и должно выполняться правило: $z' = z$, так как линейная оболочка для вектор-строк в A идентична линейной оболочке для вектор-строк A^*, а значит и размерность обеих линейных оболочек одинакова.

Для A^* допустим, что $z^* = s^*$ и осуществим обратный переход от A^* к A. При этом, естественно, все действия со строками будут происходить как бы в обратном порядке. По аналогии с приведенной выше аргументацией тогда устанавливается равенство $s^* = s$ и соответственно $z = s$. Но последнее означает, что определение ранга матрицы может производиться при помощи преобразований строк или столбцов, причем в результате этих взаимозаменяемых действий будет получен один и тот же результат. В общем каждую матрицу ранга r представляется возможным в ходе выполнения элементарных действий над строками (столбцами) привести к простой форме:

$$\begin{pmatrix}
1 & & & & & \dots & 0 \\
0 & 1 & 0 & & & \dots & 0 \\
0 & 0 & \ddots & 0 & & \dots & 0 \\
0 & 0 & 0 & 1 & 0 & \dots & 0 \\
0 & 0 & 0 & 0 & 0 & & 0 \\
\vdots & & & & & \ddots & \vdots \\
0 & & & & & \dots & 0
\end{pmatrix} = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix},$$ число единиц в матрице равно r.

2.2.4. Система линейных уравнений

Определения. Систему линейных уравнений в общем виде можно записать:

$$a_{11}x_1 + a_{12}x_m + \dots + a_{1m}x_m = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2m}x_m = b_2$$

$$a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nm}x_m = b_n,$$

т.е. система состоит из n линейных уравнений и m переменных $(x_1, \dots, x_m)$, высшая степень для всех переменных — первая.

При помощи операции перемножения матриц данную систему уравнений можно переписать в сокращенном виде:

$Ax = b$ при $A \in R^{n,m}$, $x \in R^m$, $b \in R^n$.

Здесь коэффициенты a_{ij} и свободные члены (правосторонние переменные) — это действительные числа. В случае, если b является нулевым вектором, систему линейных уравнений называют гомогенной, во всех других случаях — негомогенной[1]. Решением такой линейной системы уравнений будет вектор:

$$x^* = \begin{pmatrix} x_1^* \\ \vdots \\ x_m^* \end{pmatrix},$$

после его подстановки в систему все равенства выполняются. При этом x^* представляется решающим вектором, а множество $L = \{x^* \in R^m \mid Ax^* = b\}$ — решающим множеством, в свою очередь L есть подмножество в R^m.

[1] В отечественной литературе обычно такие системы линейных уравнений называют однородными и неоднородными. — *Прим. пер.*

Остановимся подробнее на понятии решающего множества и с этой целью рассмотрим три следующих простых случая:

I. Система не имеет решения:

$3x_1 - x_2 = 4,$

$6x_1 - 2x_2 = 7.$

Если найти решение первого уравнения для x_2 и подставить результат во второе уравнение, получим $8 = 7$. Это означает, что решающее множество является пустым ($L = \emptyset$) и система линейных уравнений не имеет решения.

II. Существует бесконечное множество решений:

$3x_1 - x_2 = 4,$

$6x_1 - 2x_2 = 8.$

Решая данную систему уравнений, как показано в случае I, всегда будем иметь $8 = 8$, т.е. система линейных уравнений обладает бесконечным множеством решений. При этом можем записать:

$x_2 = 3x_1 - 4, x_1 \in R,$

и множество решений будет

$$L = \left\{ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in R^2 \mid \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -4 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \lambda \in R \right\}.$$

В этом случае говорят, что решение представляется полем действительных чисел уровня, равного трем, проходящим через точку $(0, -4)$.

III. Имеется одно решение:

$3x_1 - x_2 = 4,$

$x_1 - x_2 = 0.$

Здесь после определения $x_2 = 3x_1 - 4$ и подстановки x_2 во второе уравнение получим $x_1 = 1$ и соответственно $x_2 = -1$. Решение одно и оно четко определено:

$$L = \left\{ \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\}.$$

Рассмотренные случаи вплотную подводят к вопросу о том, имеют ли анализируемые системы уравнений решение, и в случае, когда решение есть, — к вопросу о структуре решающего множества. Ответы на эти вопросы могут быть найдены при помощи элементарных преобразований так называемой

расширенной матрицы коэффициентов системы линейных уравнений[1]:

$$(A \vdots b) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} & : & b_1 \\ a_{21} & a_{22} & \dots & a_{2m} & : & b_2 \\ \vdots & \vdots & & \vdots & : & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} & : & b_n \end{pmatrix}.$$

Система линейных уравнений $Ax = b$ будет иметь решение, если ранг матрицы A равен рангу расширенной матрицы $(A \vdots b)$. В частности, это означает, что гомогенные системы линейных уравнений всегда имеют решения благодаря существованию тривиального решения $x = 0$.

2.2.5. Решение системы линейных уравнений

Допустим, что располагаем системой линейных уравнений:

$Ax = b, A \in R^{n,m}, x = R^m, b \in R^n,$

и эта система имеет решение. Тогда следует рассмотреть два возможных случая, рассмотренных выше:

- система имеет одно определенное решение;
- система имеет бесконечное множество решений.

Вначале рассмотрим гомогенную систему линейных уравнений $Ax = 0$, она обладает тривиальным решением $x = 0$ при условии, что $n \geq m$ и $\operatorname{rg} A = m$. При заданных условиях для негомогенной системы уравнений решением будет $x^* = b^*$, где b^* — правосторонние члены расширенной матрицы коэффициентов линейных уравнений после проведения с ней элементарных преобразований:

$$A^* = \begin{pmatrix} 1 & 0 & & \dots & 0 & : & b_1^* \\ 0 & 1 & 0 & \dots & 0 & : & b_2^* \\ \vdots & & \ddots & & \vdots & : & \vdots \\ 0 & \dots & 0 & \dots & 1 & : & b_m^* \\ 0 & \dots & 0 & \dots & 0 & : & 0 \\ & & & & \vdots & : & \vdots \\ 0 & \dots & 0 & \dots & 0 & : & 0 \end{pmatrix} = \begin{pmatrix} I_r & b^* \\ 0 & 0 \end{pmatrix}.$$

[1] В отечественной литературе: расширенной матрицы системы. — *Прим. пер.*

Уравнения в A^* начиная с $m+1$ до n не имеют никакого значения для решения системы. Это линейные комбинации строчных векторов:

$x'_1 = (1, 0, 0, \ldots, 0), \ldots, x'_m = (0, \ldots, 0, 1),$

которые в ходе элементарных преобразований обращаются в нулевые строки.

При $n \geq m$ и $\operatorname{rg} A = p < m$, или $n > m$, гомогенная система линейных уравнений имеет бесконечное множество решений, так как в матрице A в этом случае всегда имеются вектор-столбцы, линейно-зависимые от p — ее линейно независимых вектор-столбцов. Решение негомогенной системы получают при этом путем обобщения решения, найденного для гомогенной системы, когда после проведения элементарных преобразований расширенная матрица коэффициентов линейных уравнений дополняется правосторонними элементами b_i^*, представляющими определенные суммарные значения:

$$A^* = \begin{pmatrix} 1 & 0 & & \ldots & 0 & * & \ldots & * & : & b_1^* \\ 0 & 1 & 0 & \ldots & 0 & * & \ldots & * & : & b_2^* \\ \vdots & & \ddots & & \vdots & \vdots & & \vdots & : & \vdots \\ \vdots & & & \ldots & 0 & * & \ldots & * & : & b_{p-1}^* \\ 0 & \ldots & 0 & \ldots & 1 & * & \ldots & * & : & b_p^* \\ 0 & \ldots & \ldots & \ldots & 0 & 0 & \ldots & 0 & : & 0 \\ \vdots & & & & \vdots & & \ddots & \vdots & : & \vdots \\ 0 & \ldots & 0 & \ldots & 0 & & 0 & 0 & : & 0 \end{pmatrix} = \begin{pmatrix} I_p & B & b^* \\ 0 & 0 & 0 \end{pmatrix}, \quad p < m.$$

Обобщая основные положения, касающиеся поиска решений для систем линейных уравнений, отметим, что системы $Ax = b$, содержащие n уравнений и m переменных, имеют решение, когда ранг матрицы A равен рангу расширенной матрицы $(A : b)$. Если же система уравнений имеет решение, то это решение будет единственным при условии, что принадлежащая исходной гомогенная система $Ax = 0$ обладает только одним тривиальным решением $x = 0$. Во всех других случаях система уравнений имеет бесконечное множество решений, которые следует рассматривать как суммы решений гомогенной системы и одного определенного решения негомогенной системы уравнений.

Пример 2.8. Для рассмотренных выше трех случаев с различными системами линейных уравнений покажем элементарные преобразования расширенной матрицы $(A : b)$:

$$\begin{pmatrix} 3 & -1 & 4 \\ 6 & -2 & 7 \end{pmatrix} \to \begin{pmatrix} 3 & -1 & 4 \\ 0 & 0 & -1 \end{pmatrix} \to \begin{pmatrix} 1 & -1/3 & 4/3 \\ 0 & 0 & -1 \end{pmatrix} \quad \text{(случай I)},$$

$$\begin{pmatrix} 1 & -1 & 4 \\ 6 & -2 & 8 \end{pmatrix} \to \begin{pmatrix} 3 & -1 & 4 \\ 0 & 0 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & -1/3 & 4/3 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{(случай II)},$$

$$\begin{pmatrix} 3 & -1 & 4 \\ 1 & 1 & 0 \end{pmatrix} \to \begin{pmatrix} 4 & 0 & 4 \\ 1 & 1 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{pmatrix} \quad \text{(случай III)}.$$

Пример 2.9. Возьмем в качестве матрицы коэффициентов линейных уравнений матрицу A

$$A = \begin{pmatrix} 1 & 1 & -3 & 1 \\ 2 & 1 & 1 & -1 \\ 0 & 2 & -13 & 1 \\ 2 & -1 & 14 & -2 \end{pmatrix}$$

и дополним ее вектором значений свободных членов (правосторонних элементов):

$$b = \begin{pmatrix} 1 \\ 0 \\ -1 \\ \alpha \end{pmatrix}.$$

Получим:

$$\begin{aligned} x_1 + x_2 - 3x_3 + x_4 &= 1, \\ 2x_1 + x_2 + x_3 - x_4 &= 0, \\ 2x_2 - 13x_3 + x_4 &= -1, \\ 2x_1 - x_2 + 14x_3 - 2x_4 &= \alpha,\ \alpha \in \boldsymbol{R}. \end{aligned}$$

В приведенной системе уравнений $n = m = 4$. Производя элементарные преобразования, построим следующие расширенные матрицы коэффициентов линейных уравнений:

а)
$$\begin{pmatrix} 1 & 1 & -3 & 1 & : & 1 \\ 2 & 1 & 1 & -1 & : & 0 \\ 0 & 2 & -13 & 1 & : & -1 \\ 2 & -1 & 14 & 2 & : & \alpha \end{pmatrix} \begin{matrix} \\ \text{II} - 2 \cdot \text{I}, \\ \\ \text{IV} - 2 \cdot \text{I}, \end{matrix}$$

б)
$$\begin{pmatrix} 1 & 1 & -3 & 1 & : & 1 \\ 0 & -1 & 7 & -3 & : & -2 \\ 0 & 2 & -13 & 1 & : & -1 \\ 0 & -3 & 20 & -4 & : & \alpha - 2 \end{pmatrix} \begin{matrix} \text{I} + \text{II}, \\ (-1)\text{II}, \\ \text{III} + 2 \cdot \text{II}, \\ \text{IV} - 3 \cdot \text{II}, \end{matrix}$$

в) $\begin{pmatrix} 1 & 0 & 4 & -2 & : & -1 \\ 0 & 1 & -7 & 3 & : & 2 \\ 0 & 0 & 1 & -5 & : & -5 \\ 0 & 0 & -1 & 5 & : & \alpha+4 \end{pmatrix}$ $\begin{matrix} \text{I}-4\cdot\text{III}, \\ \text{II}+7\cdot\text{I}, \\ \\ \text{IV}+\text{III}, \end{matrix}$ г) $\begin{pmatrix} 1 & 0 & 0 & 18 & : & 19 \\ 0 & 1 & 0 & -32 & : & -33 \\ 0 & 0 & 1 & -5 & : & -5 \\ 0 & 0 & 0 & 0 & : & \alpha-1 \end{pmatrix}$.

Для представления алгоритма элементарных преобразований используется запись вида III + 2 II, что означает следующее: элементы второй строки матрицы умножаются на два, затем складываются с элементами третьей строки и заменяют исходные третьей строки.

Очевидно, что приведенная матрица имеет ранг $\operatorname{rg}A = 3$. Соответственно и расширенная матрица приобретает такой же ранг, одновременно выбирается $\alpha = 1$.

При заданных $p = 3 < 4 = n = m$ для системы уравнений имеется бесконечное множество решений. В то же время, принимая, например, $x_4 = 0$, можно получить конкретное решение для негомогенной системы:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 19 \\ -33 \\ -5 \\ 0 \end{pmatrix}.$$

В общем для гомогенной системы множество решений формируется при свободном выборе значений для одной из переменных. В нашем случае любые значения принимает x_4, тогда $x_1 = -18x_4$, $x_2 = 32x_4$, $x_3 = 5x_4$. Формально все множество решений системы при этом записывается:

$$L = \{x \in R^4 / \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 19 \\ -33 \\ -5 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} -18 \\ 32 \\ 5 \\ 1 \end{pmatrix}, \beta \in R\},$$

здесь роль x_4 выполняет параметр β.

Пример 2.10. Решение системы уравнений при условии $n < m$:

$$\begin{aligned} x_1 + 7x_2 + x_3 + x_4 &= 10, \\ x_2 + x_4 &= 2, \\ x_1 + x_3 &= 2, \end{aligned}$$

Произведем элементарные преобразования:

а) $\begin{pmatrix} 1 & 7 & 1 & 1 & : & 10 \\ 0 & 1 & 0 & 1 & : & 2 \\ 1 & 0 & 1 & 0 & : & 2 \end{pmatrix}$ $\text{I}-7\cdot\text{II};$ $\begin{pmatrix} 1 & 0 & 1 & -6 & : & -4 \\ 0 & 1 & 0 & 1 & : & 2 \\ 0 & 0 & 0 & 0 & : & 0 \end{pmatrix}$ $\begin{matrix} \text{I}+\text{III} \\ \text{II}-\text{III}/6 \\ \text{III}/6; \end{matrix}$

б) $\begin{pmatrix} 1 & 0 & 1 & -6 & : & -4 \\ 0 & 1 & 0 & 1 & : & 2 \\ 1 & 0 & 1 & 0 & : & 2 \end{pmatrix}$ $\text{III}-\text{I};$ г) $\begin{pmatrix} 1 & 0 & 1 & 0 & : & 2 \\ 0 & 1 & 0 & 0 & : & 1 \\ 0 & 0 & 0 & 1 & : & 1 \end{pmatrix}$.

Очевидно, $\operatorname{rg}A = 3$ и равен рангу расширенной матрицы. В результате получаем решение линейной системы уравнений: x_1 находится в зависимости от x_3, т.е.

$x_1 = 2 - x_3$, $x_2 = 1$; $x_3 = 1$.

В данном случае для определения четырех переменных мы имели три уравнения, т.е. заданы три условия на входе. Допуская, что система вообще обладает решением, вначале было необходимо найти по крайней мере одну из переменных: $x_1 = 2 - x_3$, $x_3 \in R$. После этого множество решений системы можно показать, используя уравнения связи векторов:

$$L = \{x \in R^4 / \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} + \beta \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \beta \in R\}.$$

▷***Пример 2.11.*** Решение системы уравнений при условии $n > m$:

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 1, \\ 4x_1 + 5x_2 + 6x_3 &= 2, \\ 7x_1 + 8x_2 + 9x_3 &= 3, \\ 5x_1 + 7x_2 + 9x_3 &= 4. \end{aligned}$$

В примере заданы четыре уравнения (входных условия) для определения трех переменных. При помощи уже известных нам элементарных преобразований попытаемся определить неизвестные переменные x_j.

Приведенная система уравнений не имеет решения, так как $\operatorname{rg} A = 3$, а $\operatorname{rg}(A : b) = 4$ и четвертое уравнение противоречит другим уравнениям системы, т.е. $L = 0$.

2.2.6. Квадратные матрицы

Выше показано, что множество $(n \times m)$-матриц, обладающих свойствами аддитивности и скалярного произведения, образует векторное пространство. В дальнейшем из всего множества матриц будем рассматривать квадратные матрицы и основные алгебраические действия над ними. В квадратной матрице, как известно, число строк равно числу столбцов.

2.2.7. Транспонированные матрицы

Для любой $(n \times m)$-матрицы A, с необязательным выполнением равенства $n = m$ можно найти транспонированную матрицу (ее обозначают A^T, или A'). В транспонированной матрице A' строки матрицы A заменены столбцами, или, что то же самое, столбцы заменены строками. В случае, когда $n = m$, транспонированная матрица представляет собой зеркальное отображение элементов матрицы A относительно ее главной диагонали.

Выполнение арифметических действий над транспонированными матрицами подчиняется следующим правилам. Пусть имеются две матрицы A и B размерностью $n \times m$ и третья матрица размерностью $n \times p$, тогда:

$$(A + B)' = A' + B', \tag{2.13}$$

$$(CA)' = A'C', \tag{2.14}$$

$$\operatorname{rg} A = \operatorname{rg} A'. \tag{2.15}$$

Правило (2.13), очевидно, не требует доказательства своей правомерности. Правило (2.14) получают исходя из определения произведения матриц:

$$CA = \sum_{k=1}^{n} c_{ik} a_{ki} \qquad (i = 1,\ldots,p;\ j = 1,\ldots,m)$$

и, следовательно,

$$(CA)' = \sum_{k=1}^{n} a_{jk} c_{ki} = A'C' \qquad (j = 1,\ldots,m;\ i = 1,\ldots p).$$

Правило (2.15) вытекает непосредственно из аксиоматичного положения о том, что строчный ранг матрицы всегда равен ее столбцовому рангу.

2.2.8. Структура множества квадратных матриц

Будем рассматривать множество $M^{(n,\,n)}$ всех $(n \times n)$-матриц, включая матрицы, представляющие всегда определенные результаты произведений квадратных матриц. Векторное пространство над множеством $M^{(n,\,n)}$ обладает рядом примечательных свойств:

пусть A, B и C — матрицы и одновременно элементы $M^{(n,\,n)}$, тогда

$(AB)\,C = A\,(BC)$

$(\lambda A)\,B = \lambda(AB) = A\,(\lambda B)$ — свойство ассоциативности (2.16)

$A\,(B + C) = AB + AC$

$(B + C)\,A = BA + CA$ — свойство дистрибутивности (2.17)

$I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$ — единичная матрица, выступает как нейтральный элемент, имеет силу равенство $A \cdot I = I \cdot A = A$. (2.18)

Для произведения квадратных матриц, как и в общем случае, свойство коммутативности не действует: $AB \neq BA$.

Наконец, для квадратных матриц выполняется операция обращения, т.е. осуществляется поиск элемента A^{-1} такого, что

$A \cdot A^{-1} = A^{-1} \cdot A = I$.

Эта операция более подробно рассмотрена ниже.

Замечание. Свойства (2.16) *и* (2.17) *сохраняют свою силу*, поскольку операция произведения квадратных матриц всегда определена.

2.2.9. Обращение матрицы

Пусть имеется некоторая матрица A размерностью $(n \times n)$ и I_n — единичная матрица той же размерности. Попытаемся ответить на вопрос, может ли быть найдена четко определенная $(n \times n)$-матрица A_1, для которой выполняется $A \cdot A^{-1} = I_n$? Запишем элементы приведенного уравнения более подробно:

$$A \cdot A^{-1} = \left(a_{ij}\right), \qquad i = 1,\ldots,n;\quad j = 1,\ldots,n;$$

$$A^{-1} = \left(x_{jk}\right), \qquad j = 1,\ldots,n;\quad k = 1,\ldots,n.$$

Принимая во внимание, что $A \cdot A^{-1} = I$, можем записать

$$A \cdot A^{-1} = \sum_{j=1}^{n} a_{ij} x_{jk} = \sigma_{ik} = \begin{cases} 1, & \text{если} \quad i = k \\ 0, & \text{если} \quad i \neq k. \end{cases}$$

Последняя запись означает, что для некоторого установленного числа k существует следующая система линейных уравнений:

$$Ax_k = e_k \text{ с } e_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow k\text{-я строка матрицы } A.$$

Данная система из n линейных уравнений решается с матрицей коэффициентов A, в расширении которой располагаются элементы $b = e_k$.

Алгоритм, допускающий одновременное решение системы n линейных уравнений и приводящий к обращению исходной матрицы, покажем на примере:

Пример 2.12. Задача с обращением исходной матрицы:

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & -1 & 4 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 & 0 & 1 \end{array}\right) \quad \begin{matrix} \text{II} - 2 \cdot \text{I} \\ \text{III} - \text{I} \end{matrix}$$

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -5 & -2 & -2 & 1 & 0 \\ 0 & -2 & -1 & -1 & 0 & 1 \end{array}\right) \quad \begin{matrix} \text{I} + \text{III} \\ \text{II} - 3 \cdot \text{III} \end{matrix}$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & -3 \\ 0 & -2 & -1 & -1 & 0 & 1 \end{array}\right) \quad \text{III} + 2 \cdot \text{II}$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & -3 \\ 0 & 0 & 1 & 1 & 2 & -5 \end{array}\right) \quad \begin{matrix} \text{I} - 2 \cdot \text{III} \\ \text{II} - \text{III} \end{matrix}$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -2 & -4 & 11 \\ 0 & 1 & 0 & 0 & -1 & 2 \\ 0 & 0 & 1 & 1 & 2 & -5 \end{array}\right).$$

При решении задачи преобразования матрицы на третьем шаге алгоритма произведены с учетом аксиоматичного положения о равенстве строчного и столбцового рангов. На завершающем шаге алгоритма получаем матрицу, в расширении которой систематизированы элементы обратной матрицы A^{-1}:

$$A^{-1} = \begin{pmatrix} -2 & -4 & 11 \\ 0 & -1 & 2 \\ 1 & 2 & -5 \end{pmatrix}.$$

Очевидно, что предпосылкой для успешного поиска единственного решения представленной выше системы линейных уравнений было изначальное соблюдение равенства $\operatorname{rg} A = n$.

Обобщим теоретические выводы относительно действия обращения матрицы: для имеющейся полноранговой матрицы A ($\operatorname{rg} A = n$) можно найти единственную матрицу A^{-1}, с которой выполняется условие: $A \cdot A^{-1} = A^{-1} \cdot A = I$.

Матрицу A^{-1} получают расширением исходной матрицы A путем введения соответствующей единичной матрицы и последующим проведением элементарных преобразований в расширенной матрице. В ходе преобразований получают также расширенную матрицу, в ее левой части находится единичная матрица, а в правой — элементы обращенной матрицы A^{-1}.

В качестве наглядного примера решения экономической задачи при помощи разнообразных операций над матрицами, рассмотренных выше, представим общую схему построения межотраслевого баланса, разработанную известным экономистом В. Леонтьевым, ее еще называют таблицей «затраты — выпуск» (рис. 2.2). Такая таблица состоит из четырех квадрантов, четвертый квадрант остается свободным, три других квадранта формируются в виде матриц. Назначение каждого из заполняемых квадрантов определяется содержащейся в нем матрицей.

Первый квадрант: матрица межотраслевых потоков ($X = x_{ij}$, $i, j = 1, 2, \ldots, k$; X — квадратная матрица). Элементы матрицы x_{ij} представляют объемы производства и поставки продукции из i-й отрасли в j-ю отрасль.

S_1 — суммирующий столбец, отражающий общие объемы производства в каждой из k отраслей экономики, т.е. k — строки. Значения S_1 легко исчисляются с единичным вектором $e_k := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$, его число строк также равно k: $S_1 = X \cdot e_k$.

	Промежуточный спрос — производственная сфера 1 2 ..... k	$\sum_{1}^{k}$	Конечный спрос 1 2 ..... m	$\sum_{1}^{m}$	$\sum_{1}^{k}\sum_{1}^{m}$
Промежуточные затраты производственных отраслей 1 2 . . . . . k	I квадрант X Центральная матрица межотраслевых потоков (хозяйственных операций)	S_1	II квадрант Y Матрица конечного спроса	$S_2 = y$	$S_3 = x$
$\sum_{1}^{k}$	S_4'		S_5'		
Первичные затраты (платежи за факторы производства и импорт — добавленная стоимость)* 1 2 . . . . . n	III квадрант Z Матрица первичных затрат	S_6	IV квадрант		
$\sum_{1}^{n}$	$S_7' = Z'$				
$\sum_{1}^{k}\sum_{1}^{n}$	$S_8' = X'$				

*Рис. 2.2. **Схема построения таблицы «затраты — выпуск» В. Леонтьева (базовая схема межотраслевого баланса)**.*

Соответствующим образом j-й столбец первого квадранта показывает поступление продукции в каждую j-ю отрасль из какой-либо другой, или своей же отрасли (внутреннее потребление). Строчный вектор S_4 получают суммированием всех поступлений продукции по k отраслям:

$S_4' = e_k' \cdot X, , (e_k, k), (k, k) \to (e_k, k)$.

Второй квадрант: матрица конечного спроса

$Y = (y_{is})$, $i = 1,..., k$; $S = 1,..., m$,

где y_{is} — продукция для конечного потребления, поставленная из отрасли i в S-ю группу конечного спроса.

Во втором квадранте формируются следующие показатели сумм:

$S_2 = Y \cdot e_k$ — общий объем поставки продукции каждой из k производящих отраслей своим потребителям (для промежуточного потребления);

$S_5' = e_m' Y$ — объем совокупного спроса в каждой из m групп конечных потребителей;

$S_3 = S_1 + S_2$ — общий объем поставляемой продукции каждой из k отраслей, представляющий суммарный итог поставок для промежуточного и конечного потребления: $S_3 = S_1 + S_2 = X \cdot e_k' + Y \cdot e_m' =: x$. Компоненты вектора S_3 отражают общие потребления производственных товаров и услуг — $x' = (x_1, x_2, ..., x_k)$.

Третий квадрант: $Z = (z_{rj})$, $j = 1,..., k$; $r = 1,..., n$,

где z_{rj} — количественная характеристика стоимости факторов производства и импорта (добавленная стоимость), т.е. каждая r-я строка матрицы Z отражает потребление r-го производственного фактора в каждой из k отраслей.

Суммарные итоги квадранта:

$S_6 = Z \cdot e_k$ — общее использование каждого из производственных факторов;

$S_1 = e_k' \cdot Z$ — общее использование производственных факторов в каждой из j отраслей производящей сферы.

Уже известный нам, но только транспонированный вектор x' может быть получен при помощи матрицы Z:

$x' = S_4' + S_1' = X \cdot e_k' + Z \cdot e_k' = (x_1, ..., x_n)$.

Таким образом для каждой k-й отрасли находится общее потребление производственных товаров (услуг) и факторов производства. При этом для каждой из k отраслей производящей сферы сохраняется равенство сумм, отражающих потоки продукции на входе и выходе, и действует равенство: $x = (x')'$.

Построение таблицы «затраты — выпуск» осуществляется одним из двух методов:

1) методом «выпуска», когда заполняются строки матриц таблицы,
2) методом «затрат», когда заполняются столбцы матриц таблицы.

Метод «выпуска»

При формировании матриц X и Y выпуск товаров и услуг по отраслям производственной сферы определяется на основе данных статистики сбыта.

Метод «затрат»

Матрицы X и Z заполняются данными о затратах, произведенных в отдельных отраслях, на основе сведений статистики движения сырья и материалов, а также статистики затрат и статистики товарных знаков.

На практике оба метода используют вместе, в комбинации, учитывая возможность получения тех или иных данных.

Информация из таблицы «затраты — выпуск» широко используется для:

- изучения структуры центральной матрицы X;
- исчисления (приблизительного оценивания) структурных коэффициентов прямых затрат:

$$a_{ij} = \frac{x_{ij}}{x_{\cdot j}},\ i = 1,...,k,\ \text{при установленном } j \in \{1, ..., k\}.$$

Так как во всех отраслях в произведенный продукт, кроме промежуточного потребления (x_{ij}), входят элементы добавленной стоимости, действует неравенство:

$$\sum_{i=1}^{k} a_{ij} < 1,$$

a_{ij} называют также производственными коэффициентами прямых затрат. Коэффициенты первичных затрат c_{rj} получают по данным матрицы Z:

$$c_{rj} = \frac{z_{rj}}{x_{\cdot j}},\ r = 1,...,n,\ \text{при установленном } j \in \{1, ..., k\}.$$

При обобщении данных расчетов коэффициентов a_{ij} и c_{rj} будем иметь матрицы:

$A = (a_{ij})$ — матрицу коэффициентов прямых затрат;

$C = (c_{rj})$ — матрицу коэффициентов первичных затрат.

Кроме этого, по данным первого квадранта таблицы «затраты — выпуск» исчисляют коэффициенты полного выпуска, обобщающие относительные величины промежуточного и конечного потребления:

$$b_{ij} = \frac{x_{ij}}{x_{i\cdot}},\ \text{при установленном } j \in \{1,..., k\},$$

для коэффициентов b_{ij} также действует неравенство

$$\sum_{j=1}^{k} b_{ij} < 1.$$

Множество коэффициентов b_{ij} образует матрицу коэффициентов распределения: $B = (b_{ij})$.

Наконец данные матрицы X и Y позволяют определять коэффициенты «конечного спроса — выпуска»:

$$d_{is} = \frac{y_{is}}{x_{i\cdot}},\ s = 1,...,m,\ \text{при установленном } i \in \{1,...,k\}.$$

Матрица $D = (d_{is})\ c \sum_{j=1}^{k} b_{ij} + \sum_{s=1}^{m} d_{is} = 1$ для всех $i \in \{1,...,k\}$.

Исчисляемые в рамках межотраслевого баланса коэффициенты открывают возможность построения аналитических моделей «затраты — выпуск».

Аналитическая модель затрат. Вектор конечного спроса y^*, как известно, определяется соответствующими объемами прямых (x_{ij}) и первичных (z^*_{rj}) затрат, именно они составляют величину продукта, используемого затем для удовлетворения конечного спроса. Таким образом, для каждой i-й строки матриц X и Y, $i \in \{1,..., k\}$, можем записать:

$$\sum_{j=1}^{k} x_{ij} + \sum_{s=1}^{m} y_{is} = \sum_{j=1}^{k} a_{ij} x_{\cdot j} + y_{i\cdot},$$

или уравнение в матричном виде

$$x = Ax + y \Leftrightarrow x - Ax \Leftrightarrow (I\text{-}A)\ x = y \Rightarrow x^* = (I\text{–}A)^{-1} y^*.$$

Матрицу вида $(I-A)$ называют матрицей Леонтьева или матрицей технологических затрат, обратная ей матрица $V = (I-A)^{-1}$ — это матрица коэффициентов межотраслевых связей (или просто обратная матрица Леонтьева). Компоненты матрицы V, т.е. v_{ij}, показывают, на сколько следует увеличить производство продукции в каждой из i отраслей, чтобы обеспечить дополнительную единицу конечного спроса в продукции j-й отрасли. Другими словами, коэффициенты v_{ij} — это мультипликаторные величины.

Аналитическая модель выпуска. Модель имеет вид:

$$x_{.j} = \sum_{i=1}^{k} x_{ij} + \sum_{r=1}^{m} z_{rj} = \sum_{i=1}^{k} b_{ij} + x_{i.} + z_{.j},$$

или в матричной форме

$$x' = x' \cdot B + z' \Leftrightarrow x^{*\prime} = z^{*\prime} - x'B = z' \Leftrightarrow x' (I-B) = z' \Leftrightarrow x^{*\prime} = z^{*\prime} (I-B)^{-1}.$$

Последняя из этих записей означает, что на основе данных о величине первичных затрат ($z^{*\prime}$) могут определяться прямые затраты $x^{*\prime}$, и те, и другие затратные элементы участвуют в формировании стоимости продукта, удовлетворяющего конечный спрос (Y^*).

2.2.10. Определители

Определители матрицы и их свойства. Здесь будем рассматривать только квадратные матрицы $(n \times n)$, причем матрицы, обобщающие множества действительных чисел.

Определителем квадратной матрицы A называют некоторое число, полученное путем преобразования ее элементов по определенному правилу.

Определитель матрицы det: $= M(n, n) \to R$ обладает следующими основными свойствами:

определитель матрицы представляет линейную форму каждой из ее i строк, т.е. $\det A = d_j a_{ij}$, $j = 1,..., m$; (2.19)

$\operatorname{rg} A < n \Leftrightarrow \det A = 0$; (2.20)

$\det I = 1$. (2.21)

Кроме приведенной формы записи определителя $\det A$, существуют одинаково правомерные другие формы: ΔA, $|A|$ и т.д., для удобства выберем и будем чаще использовать $\det A$ и $|A|$.

Первое свойство определителя означает следующее: пусть имеется $(n \times n)$-матрица A:

$$A_i = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{i1} & \dots & a_{in} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix},$$

произведем замену одной из строк матрицы A, скажем, элементы i-й строки заменим вектором $(a^*_{i1}, ..., a^*_{in})$:

$$A^*_i = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{i-1,1} & \dots & a_{i-1,n} \\ a^*_{i1} & \dots & a^*_{in} \\ a_{i+1,1} & \dots & a_{i+1,n} \\ \vdots & & \vdots \\ a_{n1} & & a_{nn} \end{pmatrix}.$$

Далее обозначим B матрицу, которая получена из исходной матрицы A заменой в ней одной (i-й) строки, т.е. заменой элементов этой строки $a_{i1}, ..., a_{in}$ новыми значениями $a^*_{i1}, ..., a^*_{in}$:

$$B = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{i-1,1} & \dots & a_{i-1,n} \\ a_{il} + a^*_{il} & \dots & a_{in} + a^*_{in} \\ a_{i+1,1} & \dots & a_{i+1,n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}.$$

Согласно первому свойству определителя матрица для $|B|$ действует:

$$|B| = |A_i| + |A^*_i| \qquad (i = 1,..., n).$$

Предположим, что замена i-й строки матрицы A произведена несколько иным образом, а именно: $a^*_{ij} = \alpha \cdot a_{ij}$, и построена матрица C, учитывающая эту замену:

$$C = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{i-1,1} & \dots & a_{i-1,n} \\ \alpha \cdot a_{i1} & \dots & \alpha \cdot a_{in} \\ a_{i+1,1} & \dots & a_{i+1,n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}, \quad \alpha \in R \quad (i = 1,\dots, n).$$

Тогда по первому свойству определяется:

$C = \alpha \,|\, A_i \,|$ $\qquad (i = 1,\dots, n).$

Само по себе определение детерминанта матрицы как det: $M^{(n,\,n)} \to R$ не дает оснований для формирования алгоритма его исчисления. Это становится возможным с использованием уже приведенных выше трех основных свойств детерминанта и следующих правил:

а) если имеется матрица A^*, полученная из A в результате многократного прибавления элементов одной ее строки к элементам другой, то $|A| = |A^*|$;

б) если матрица A^* получена из A заменой одной строки на другую, то $|A| = -|A^*|$;

в) если A' — это транспонированная матрица A, то $|A| = |A'|$;

г) для всех матриц A, $B \in B^{(n,\,n)}$ имеет силу $|AB| = |A| \cdot |B|$;

д) $|\alpha A| = \alpha^n |A|$;

е) если матрица A имеет хотя бы одну нулевую вектор-строку, то $|A| = 0$;

и) если элементы двух строк (две вектор-строки) матрицы A находятся в линейной зависимости, то $|A| = 0$.

Перечисленные свойства детерминанта связаны так или иначе с преобразованием строк произвольной матрицы A. Понятно, что эти же свойства сохраняют свою силу при проведении аналогичных преобразований не со строками, а со столбцами матрицы A.

Определение детерминанта. Расчет детерминанта для некоторой матрицы $A \in M^{(n,\,n)}$ потребует введения еще одной матрицы A_{ij} — это такая матрица A, из которой удалены i-я строка и j-й столбец, т.е. матрица A_{ij} имеет размерность $(n-1) \times (n-1)$. Имея A и A_{ij}, можем записать:

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \,|A_{ij}\,|.$$

Такую формулу называют Лапласовым разложением детерминанта по j-му столбцу. Ее применение на практике позволяет сжимать любую квадратную матрицу до минимального размера, с включением всего нескольких чисел, а затем по простому алгоритму вычислять сам детерминант.

Пример 2.13. Пусть имеется (4×4)-матрица A:

$$A = \begin{pmatrix} 2 & 1 & 4 & -1 \\ 1 & 0 & 3 & 2 \\ 1 & -1 & 0 & 3 \\ 4 & 2 & 1 & 0 \end{pmatrix}.$$

Произведем разложение детерминанта по элементам второго столбца. Выбор столбца в данном случае объясняется тем, что он содержит по крайней мере один нуль:

$$\det|A| = (-1)^{(1+2)} \cdot 1 \cdot \det\begin{pmatrix} 1 & 3 & 2 \\ 1 & 0 & 3 \\ 4 & 1 & 0 \end{pmatrix} + (-1)^{(2+2)} \cdot 0 \cdot \det\begin{pmatrix} 2 & 4 & -1 \\ 1 & 0 & 3 \\ 4 & 1 & 0 \end{pmatrix} +$$

$$+ (-1)^{(3+2)} \cdot (-1) \cdot \det\begin{pmatrix} 2 & 4 & -1 \\ 1 & 3 & 2 \\ 4 & 1 & 0 \end{pmatrix} + (-1)^{(4+2)} \cdot 2 \cdot \det\begin{pmatrix} 2 & 4 & -1 \\ 1 & 3 & 2 \\ 1 & 0 & 3 \end{pmatrix}.$$

Полученные детерминанты размерности (3×3) могут быть подвергнуты дополнительному разложению. Обратим внимание, что столбцы, по элементам которых осуществляется разложение, целесообразно выбирать с наибольшим числом нулевых элементов. В данном случае, как видно, разложение первого и четвертого детерминантов второго уровня следует производить по элементам второго столбца, а третьего детерминанта — по элементам третьего столбца:

$$\det A = -(-1)^{(1+2)} \cdot 3 \cdot \det\begin{pmatrix} 1 & 3 \\ 4 & 0 \end{pmatrix} - (-1)^{(2+2)} \cdot 0 \cdot \det\begin{pmatrix} 1 & 2 \\ 4 & 0 \end{pmatrix} - (-1)^{(3+2)} \cdot 1 \cdot \det\begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} +$$

$$+ (-1)^{(1+3)} \cdot (-1) \cdot \det\begin{pmatrix} 1 & 3 \\ 4 & 1 \end{pmatrix} + (-1)^{(2+3)} \cdot 2 \cdot \det\begin{pmatrix} 2 & 4 \\ 4 & 1 \end{pmatrix} + (-1)^{(3+3)} \cdot 0 \cdot \det\begin{pmatrix} 2 & 4 \\ 1 & 3 \end{pmatrix} +$$

$$+ (-1)^{(1+2)} \cdot 4 \cdot \det\begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} + (-1)^{(2+2)} \cdot 3 \cdot \det\begin{pmatrix} 2 & -1 \\ 1 & 3 \end{pmatrix} + (-1)^{(3+2)} \cdot 0 \cdot \det\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}.$$

Продолжая разложение детерминантов (третий уровень), получим $|A|$:

$$\det A = +3[(-1)^{(1+2)}\cdot 3\cdot 4+(-1)^{(1+1)}\cdot 0\cdot 1\ ]+1[(-1)^{(1+1)}\cdot 1\cdot 3+(-1)^{(2+1)}\cdot 1\cdot 2\ ]-$$
$$-1[(-1)^{(1+1)}\cdot 1\cdot 1+(-1)^{(2+1)}\cdot 4\cdot 3\ ]-2[(-1)^{(1+1)}\cdot 2\cdot 1+(-1)^{(2+1)}\cdot 4\cdot 4]-$$
$$-8[(-1)^{(1+1)}\cdot 1\cdot 3+(-1)^{(2+1)}\cdot 1\cdot 2\ ]+6[(-1)^{(1+1)}\cdot 2\cdot 3+(-1)^{(2+1)}\cdot 1\cdot(-1)\ ]=$$
$$=-36+1+11+28-8+42=38.$$

Для случаев, когда $n \le 3$, с целью определения детерминанта матрицы могут применяться более экономные методы по сравнению с рассмотренным в примере выше. Скажем:

$n = 2$ и $A=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, тогда $\det A = a_{11}a_{22} - a_{12}a_{21}$.

Если $n = 3$ и $A=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$, расчет детерминанта может производиться по так называемому *правилу Сарруса*:

$$\det A = a_{11}a_{22}a_{33}+a_{12}a_{23}a_{31}+a_{13}a_{21}a_{32}-(a_{13}a_{22}a_{31}+a_{12}a_{21}a_{33}+a_{11}a_{23}a_{32}).$$

В данном случае детерминант A представляет собой сумму произведений элементов, расположенных на главной диагонали, за вычетом сумм произведений элементов, принадлежащих побочной диагонали. Естественно, детерминанты матрицы с $n > 3$ могут подвергаться разложению до тех пор, пока не станет $n = 3$, а затем определяться по упрощенному правилу Сарруса.

Линейные системы уравнений и детерминанты. Свойства определителя, в частности (2.19), позволяют делать вывод о том, что элементарные преобразования матрицы не изменяют его величины. Кроме того, следует признать правомерным утверждение: $\det A = 0 \Leftrightarrow \operatorname{rg}(A) < n$.

С учетом приведенных высказываний для линейной системы уравнений: $Ax = b$, $A \in R^{n,\,n}$; $x, b \in R^n$, существует единственное решение при условии равенства числа уравнений системы числу ее неизвестных переменных (x), т.е. когда $\operatorname{rg} A = n$ и $\det A \neq 0$. Решающий вектор x^* тогда определяется по формуле: $x^* = A^{-1}b$. Отдельные компоненты вектора x^* могут быть исчислены без

полной обратной матрицы A^{-1}, при этом любая k-я компонента находится по формуле:

$$x'_k = \sum_{i=1}^{n} \tilde{a}_{ki} b_i,$$

где $\tilde{a}_{ki}$ — элементы обратной матрицы A^{-1}.

Элементы $\tilde{a}_{ki}$ легко определяются при помощи матрицы A_{ij}:

$$\tilde{a}_{ki} = \frac{(-1)^{i+j} \det A_{ij}}{\det A},$$

соответственно можно сделать запись для x'_k:

$$x'_k = \frac{1}{\det A} \sum_{i=1}^{n} b_i (-1)^{i+k} \det A_{ki}.$$

Следуя основным свойствам детерминанта (2.19) и (2.20), выражение, стоящее под знаком суммы, можно представить также определителем:

$$D_k := \sum_{i=1}^{n} b_i (-1)^{i+k} \det A_{ki} = \det \begin{pmatrix} a_{11} & \dots & a_{1,k-1}\ b_1\ a_{1,k+1} & \dots & a_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{n,k-1}\ b_n\ a_{n,k+1} & \dots & a_{nn} \end{pmatrix}.$$

Обобщая теоретические выводы по оцениванию вектора x^*, приходим к так называемому *правилу Крамера*: если имеется некоторая система линейных уравнений: $Ax = b$, $A \in R^n$, n; $x, b \in R^n$, и эта система имеет единственное решение при условии $\det A \neq 0$, то это решение находится последовательным исчислением компонентов решающего вектора по формуле:

$$x'_k = \frac{D_k}{\det A};\ k = 1,\dots,\ n.$$

Пример 2.14. Пусть имеется система линейных уравнений:

$$x_1 + x_2 - x_3 = 0,$$
$$x_1 \quad + x_3 = 1,$$
$$2x_1 - x_2 \quad = 2.$$

Следуя правилу Крамера:

$$\det A = \det \begin{pmatrix} 1 & 1 & -1 \\ 1 & 0 & 1 \\ 2 & -1 & 0 \end{pmatrix} = 2+1+1 = 4 \neq 0 \text{ и}$$

$$D_1 = \det\begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 1 \\ 2 & -1 & 0 \end{pmatrix} = 2+1=3;$$

$$D_2 = \det\begin{pmatrix} 1 & 0 & -1 \\ 1 & 1 & 1 \\ 2 & 2 & 0 \end{pmatrix} = -2+2-2=-2;$$

$$D_3 = \det\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 2 & -1 & 2 \end{pmatrix} = 2-2+1=1.$$

Система имеет единственное решение: $x^* = \frac{1}{4}\begin{pmatrix} 3 \\ -2 \\ 1 \end{pmatrix}$.

2.2.11. Собственные числа и собственные векторы

Пусть по-прежнему имеется матрица A размерностью $(n \times n)$ и некоторое действительное число $\lambda \in R$. Рассмотрим уравнение вида:

$Ax = \lambda x$, где $A \in R^{(n,n)}$, $x \neq 0$, $\lambda \in R$.

Число λ, как и вектор x, является решением приведенного уравнения, при этом λ и x принадлежат друг другу, и λ называют *собственным числом*, а x — *собственным вектором* матрицы A. Как собственное число, так и собственный вектор для одной матрицы представляют не единственное решение, их может быть несколько. В частности, всегда допускается в качестве одного из решений нулевое значение собственного числа: $\lambda = 0$.

Если существует λ — собственное число матрицы A и E_λ — это множество всех принадлежащих этому числу λ собственных векторов, то $E_\lambda \cup \{0\}$ есть векторное пространство собственного числа λ.

Для определения алгоритма исчисления собственных чисел и собственных векторов рассмотрим вытекающую из приведенного выше уравнения систему уравнений вида: $Ax - \lambda x = 0$ или эквивалентное ей выражение $(A - \lambda I)\,x = 0$. Это по сути гомогенная система уравнений, которая по меньшей мере имеет одно тривиальное решение $x^* = 0$; напомним, что это решение представляется некоторым собственным вектором.

Собственные векторы для матрицы A имеются и тогда, когда линейная система уравнений обладает нетривиальными решениями. Например, хорошо известен случай: $\mathrm{rg}\,(A - \lambda I) < n$.

Задача определения собственных векторов сводится к оценке ранга матрицы $(A - \lambda I)$ и последующему решению приведенной выше системы линейных уравнений.

В п. 2.2.10 мы установили, что

$\mathrm{rg}\,(A - \lambda I) < n \Leftrightarrow \det\,(A - \lambda I) = 0$,

следовательно λ точно воспроизводит собственное число матрицы A при условии, что $\det\,(A - \lambda I) = 0$. Оценкой данного детерминанта является полином n-й степени. Такой полином называют *характеристическим полиномом матрицы* A, собственные числа — это исходные (нулевые) точки полинома — одно из основных положений алгебры. Матрица размерностью $(n \times n)$ всегда имеет n собственных чисел.

Пример 2.15. Пусть имеем матрицу A:

$A = \begin{pmatrix} 2 & -2 & -4 & -2 \\ -8 & 5 & 1 & -8 \\ 0 & 3 & -1 & 0 \\ -2 & -1 & 5 & 2 \end{pmatrix}$; тогда ее определитель будет:

$$\det(A-\lambda I) = \det\begin{pmatrix} 2-\lambda & -2 & -4 & -2 \\ -8 & 5-\lambda & 1 & -8 \\ 0 & 3 & -1-\lambda & 0 \\ -2 & -1 & 5 & 2-\lambda \end{pmatrix}.$$

Произведем разложение детерминанта по элементам его первого столбца:

$$\det(A-\lambda I) = (2-\lambda)\cdot(-1)^{(1+1)}\cdot\det\begin{pmatrix} 5-\lambda & 1 & -8 \\ 3 & -1-\lambda & 0 \\ -1 & 5 & 2-\lambda \end{pmatrix} -$$

$$-8\cdot(-1)^{(2+1)}\cdot\det\begin{pmatrix} -2 & -4 & -2 \\ 3 & -1-\lambda & 0 \\ -1 & 5 & 2-\lambda \end{pmatrix} - 2\cdot(-1)^{(4+1)}\cdot\det\begin{pmatrix} -2 & -4 & -2 \\ 5-\lambda & 1 & -8 \\ 3 & -1-\lambda & 0 \end{pmatrix} =$$

$$= (2-\lambda)\cdot[(5-\lambda)(-1-\lambda)(2-\lambda) - 120 + 8(1+\lambda) - 3(2-\lambda)] +$$

$$+8\cdot[2(1+\lambda)(2-\lambda) - 30 + 2(1+\lambda) + 12(2-\lambda)] +$$

$$+2\cdot[96 + 2(5-\lambda)(1+\lambda) + 6 + 16(1+\lambda)] = \ldots =$$

$$= \lambda^4 - 8\lambda^3 - 16\lambda^2 + 128\lambda = 0 \Rightarrow \lambda = 0, \text{ или } \lambda^3 - 8\lambda^2 - 16\lambda + 128 = 0.$$

В ходе решения или даже простым предположением получим результат: $\lambda_{2,3} = \pm 4$.

Отсутствующие исходные точки полинома можно определить при помощи факторного расположения последнего:

$\lambda^3 - 8\lambda^2 - 16\lambda + 128 = (\lambda - 4)(\lambda + 4) \Rightarrow \lambda - \lambda_4$.

Последующим делением полинома на $(\lambda^2 - 16)$ получим значение λ_4:

$(\lambda^3 - 8\lambda^2 - 16\lambda + 128) : (\lambda^2 - 16) = \lambda - 8$, откуда $\lambda_4 = 8$.

Запишем все множество собственных чисел для матрицы A: $\{-4, 0, 4, 8\}$. Собственное пространство для $\lambda = -4$ получим, решая следующую систему линейных уравнений:

$$\begin{aligned} 6x_1 - 2x_2 - 4x_3 - 2x_4 &= 0 \\ -8x_1 + 9x_2 + x_3 - 8x_4 &= 0 \\ 3x_2 + 3x_3 &= 0 \\ -2x_1 - x_2 + 5x_3 + 6x_4 &= 0, \end{aligned}$$

и $E_{-4} = \{\alpha \begin{pmatrix} 0 \\ -1 \\ 1 \\ -1 \end{pmatrix} \mid \alpha \in R\}$.

Аналогичным образом получим векторные пространства для $\lambda = 0$, $\lambda = 4$ и $\lambda = 8$:

$$E_0 = \{\alpha \begin{pmatrix} 4 \\ 1 \\ 3 \\ -3 \end{pmatrix} \mid \alpha \in R\}; \; E_4 = \{\alpha \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix} \mid \alpha \in R\} \text{ и } E_8 = \{\alpha \begin{pmatrix} 2 \\ -3 \\ -1 \\ -1 \end{pmatrix} \mid \alpha \in R\}.$$

Вопросы поиска собственных чисел и собственных векторов весьма многогранны и имеют широкое поле приложения. Им уделяется значительное внимание не только в математике, но и в физике и, как мы еще не раз убедимся, в экономической теории.

2.2.12. Евклидово векторное пространство

Введем метрические понятия: расстояния, угла наклона для наблюдаемых объектов в пространстве, ортогональности. Здесь нам вновь понадобится расширенное определение для произведения векторов, включая так называемое скалярное произведение.

Скалярное произведение. Пусть V, как и прежде, это векторное пространство над полем действительных чисел R. Обозначим S скалярное произведение векторных пространств:

$$S = V \cdot V \to R,$$

или при записи в векторной форме:

$$(x, y) \to S(x, y).$$

Скалярное произведение векторов обладает следующими основными свойствами:

$S(x, y) = S(y, x)$ для всех $x, y \in V$ (симметричность); (2.22)

$S(x, y)$ — линейно относительно x и y (билинейность); (2.23)

$S(x, x) > 0$ для $x \in V \setminus \{0\}$ (положительная определенность). (2.24)

Векторное пространство над полем действительных чисел, в котором определена операция скалярного произведения векторов, называется *евклидовым векторным пространством*.

Пример 2.16. Предположим, что в n-мерном векторном пространстве R^n имеем:

$$S: \begin{cases} R^n \times R^n \to R \\ ((x_1,\ldots,x_n),(y_1,\ldots,y_n)) \to \sum_{i=1}^{n} x_i y_i. \end{cases}$$

Можем утверждать, что скалярное произведение векторов (S) будет положительно определенной величиной. Данный вывод для S следует из уже известных свойств его симметричности, билинейности и положительной определенности

$$(S(x, x) = \sum_{i=1}^{n} x_i^2 > 0 \text{ для всех } x \neq 0).$$

Пример 2.17. В векторном пространстве $C(I)$ для непрерывной функции действительных чисел, замкнутой на интервале $I = [a, b]$, имеем скалярное произведение:

$$S: \begin{cases} C \times C \to R \\ (g,h) \to \int_a^b g(t)h(t)dt. \end{cases}$$

Основываясь на элементарных правилах интегрального исчисления, можем заключить, что скалярное произведение S в

данном случае симметрично, билинейно и положительно определено, так как для всех $h \in C$ действует правило:

$$S(h, h) = \int_a^b h^2(t)dt \geq 0,$$

а из $S(g, g) = \int_a^b g^2(t)dt = 0$ следует, что $g(t) = 0$ для всех $t \in [a, b]$.

С другой стороны, если бы было $g(t_0) \neq 0$ для всех $t_0 \in [a, b]$, то с учетом непрерывности функции g следовало бы определить окрестность $U = U(t_0)$ для точки (t_0) такую, в которой $g(t) \neq 0$ для всех $t \in U(t_0)$, но тогда также и $\int_{U(t_0)} g^2(t)\,dt \neq 0$, что противоречит исходной посылке задачи.

Замечания. 1. Результат скалярного произведения не следует смешивать с умножением векторов на скаляр, о котором говорилось в п. 2.1.2. Умножение на скаляр связывает определенным образом вектор $x \in V$ с некоторым числом $\alpha \in R$, в то время как скалярное произведение связывает друг с другом два вектора. Кроме того, скалярное произведение выражается всегда некоторым действительным числом, не обязательно являющимся элементом исходного векторного пространства V, и тем самым как бы выходит за рамки обычных внутренних связей V.

2. С учетом положительной определенности скалярного произведения следует вывод о правомерности выражения:

$$S(x, x) = 0 \Leftrightarrow x = 0.$$

В последующем будем использовать при записи скалярного произведения обычную форму представления операции перемножения векторов:

$S(x, y) := xy$.

Отметим, что в евклидовом векторном пространстве (V) имеет силу неравенство Коши—Шварца: для всех $x, y \in V$ имеет силу: $(x, y)^2 \leq x^2y^2$. Неравенство превращается в равенство, когда векторы x и y линейно-зависимы. Докажем правомерность неравенства Коши—Шварца в пространстве V.

а) пусть x и y линейно независимы. Тогда при условии, что $x \neq 0$, всегда $x^2 > 0$ и $\lambda x + y \neq 0$ для всех $\alpha \in R$.

С учетом положительной определенности скалярного произведения имеем:

$$0 < (\lambda x + y)^2 = x^2 (\lambda + \frac{xy}{x^2})^2 + \frac{1}{x^2}(x^2y^2 - (xy)^2),$$

и для $\lambda = -\frac{xy}{x^2}$ следует $0 < x^2y^2 - (xy)^2$;

б) предположим, векторы x и y — линейно-зависимы. В этом случае можем записать: $y = \alpha x$ $(\alpha \in R)$, и так как скалярное произведение линейно относительно обоих аргументов (билинейно), то:

$$(xy)^2 = (x\lambda x)^2 = \lambda^2 (xx)^2 = \lambda^2 x^2 x^2 = x^2\lambda^2 x^2 = x^2y^2.$$

2.2.13. Норма вектора

Скалярное произведение позволяет для любого вектора получить некоторое число, интерпретируемое как длина этого вектора.

Пусть V — евклидово векторное пространство, тогда *норма вектора* этого пространства может быть формально представлена следующей записью:

$$N: \begin{cases} V \to R \\ x \to N(x) = \|x\|. \end{cases}$$

Норма вектора обладает следующими свойствами:

$\|x\| > 0$ для всех $x \in V \setminus \{0\}$, $\|x\| = 0 \Leftrightarrow x = 0$ (положительная определенность); (2.25)

$\|\lambda x\| = |\lambda| \|x\|$ для $\lambda \in R$, $x \in V$ (гомогенность); (2.26)

$\|x + y\| \leq \|x\| + \|y\|$ для $x, y \in V$ (неравенство треугольника). (2.27)

В евклидовом векторном пространстве существует возможность всегда найти такую норму вектора, для которой будет выполняться равенство:

$$\|x\| = +\sqrt{x^2} = +\sqrt{xx},$$

т.е. в данном случае мы имеем дело со скалярным произведением вектора x на самого себя, извлечение корня квадратного из такого произведения всегда будет давать положительный результат.

Перечисленные выше свойства нормы вектора (2.25) — (2.27) всегда сохраняют свою силу, в частности с учетом того, что для (2.25) действует $\|x\| = +\sqrt{x^2} > 0$ для всех $x \neq 0$;

(2.26) подтверждается тем, что $\|\lambda x\| = \sqrt{\lambda^2 x^2} = |\lambda| \sqrt{x^2}$;

(2.27) — вывод получаем на основе неравенства Коши—Шварца:
$\|x + y\|^2 = (x + y)^2 = x^2 + 2xy + y^2 \le \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2$.

Иллюстрации теоретического материала: 1. В векторном пространстве R^n норма задается равенством: $\|x\| = +\sqrt{\sum_{i=1}^{n} x_i^2}$.

2. Норма всех непрерывных функций g, существующих в векторном пространстве $C([a, b])$, на интервале $[a, b]$ определяется как корень квадратный из определенного интеграла, построенного для квадратов функций g:

$$\|g\| = +\sqrt{\int_a^b g^2(t)dt}.$$

Заметим, что собственно норма, приведенная для функций в векторном пространстве $C([a, b])$, вовсе не обязательно определяется посредством скалярного произведения, она, например, может быть выражена определенной максимизирующей величиной:

$$\|f\| = \max_{x \in [a,b]} |f(x)|.$$

Вопросы выбора, обоснования преимуществ той или иной нормы векторного пространства требуют самостоятельного, более детального анализа, и в рамках данной работы не рассматриваются.

Замечание. Норму вектора всегда можно получить как результат скалярного произведения при условии, что это скалярное произведение определено в существующем векторном пространстве.

Определение 1. Векторное пространство, в котором определена норма, называют *нормированным векторным пространством.*

Линейное пространство U нормированного векторного пространства V будет также нормированным пространством, если его элементы нормированы и выступают как подмножество элементов V. При этом норму, определенную в U, называют *индуцированной нормой пространства V.*

Обратное утверждение, что нормированное векторное пространство V всегда воспроизводит евклидово векторное пространство, не является правомерным. На самом деле, чтобы при помощи равенства $\|x\| = \sqrt{x^2}$ получить скалярное произведение,

необходимо выполнение так называемого равенства параллелограмма:

$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$ для всех $x, y \in V$.

2.2.14. Угол, образованный двумя векторами

Согласно неравенству Коши—Шварца для двух векторов x и y в одном векторном пространстве V, в котором $x \neq 0$ и $y \neq 0$, действует правило:

$$\frac{(xy)^2}{x^2y^2} \le 1, \quad -1 \le \frac{xy}{\|x\|\|y\|} \le 1.$$

При этом всегда имеется некоторое число $\varphi(x, y)$, такое, что $\cos\varphi(x, y) = \frac{xy}{\|x\|\|y\|}$, $\varphi(x, y) \in [0,\pi]$. Число $\varphi(x, y)$ определяется всегда однозначно и называется *углом между векторами* x и y. Очевидно, $\cos\varphi(x, y) = 0 \Leftrightarrow xy = 0$, т.е. если $xy = 0$, то $\varphi(x, y)$ — прямой угол. Обратим внимание, что скалярное произведение двух векторов x и y может быть также равно нулю и в случае, когда $x \neq 0$, $y \neq 0$, что невозможно, если мы имеем дело с обычными числами (скалярами).

Рассматривая скалярное произведение векторов и угол, образованный парой векторов, мы подходим к понятию ортогональности.

Определение 2. Два вектора x и $y \in V$, $x \neq 0$, $y \neq 0$ называют *ортогональными*, если их скалярное произведение равно нулю: $xy = 0$.

Ортогональные векторы часто являются наиболее удобными для организации базиса векторного пространства (например, обычная декартова система координат — *примеч. пер.*). Конструктивно подобный базис задается множеством ортогональных векторов. При этом непустое подмножество M евклидова векторного пространства V называют ортогональной системой, если $0 \notin M$, но $xy = 0$ для каждой пары векторов $x, y \in M$, другими словами, если все векторы из M попарно ортогональны.

Векторы ортогональной системы $(x_1,..., x_r)$ линейно независимы. На самом деле неравенство $\sum_{i=1}^{r} \alpha_i x_i = 0$ предполагает, что для любого $p \in \{1,..., r\}$, с учетом взаимной ортогональности всех

пар векторов ($x_i x_j = 0$ для $i \neq j$) правомерна запись: $\left(\sum_{i=1}^{r} \alpha_i x_i\right) x_p = \alpha_p x_p^2 = 0$, но при условии $x_p^2 \neq 0$, тогда однозначно следует $\alpha_p = 0$.

Подобным образом можно конструктивно представить n-мерное евклидово векторное пространство как ортогональную систему, включающую n элементов, а точнее n полярно ортогональных векторов. Можно сказать, что такая n-элементная ортогональная система M из n-мерного векторного пространства V является его базисом.

Рассмотрим, каким образом любой базис для V может быть преобразован в ортогональную систему. Пусть $b_1, \ldots, b_k$ — система из k линейно независимых векторов пространства V. Выберем из множества этих векторов один, первый вектор $b_1 = c_1$ для конструируемой ортогональной системы. На следующем шаге примем, что уже имеется $i - 1$ $(i \leq k)$ ортогональных векторов $c_1, \ldots, c_{i-1}$, чья линейная оболочка совпадает с линейной оболочкой векторов $b_1, \ldots, b_{i-1}$. Следующий ортогональный вектор c_i определим по формуле: $c_i = b_i + \lambda_1 c_1 + \ldots + \lambda_{i-1} c_{i-1}$; здесь $\lambda_1, \ldots, \lambda_{i-1}$ находится из системы уравнений, имеющей при $c_1^2 \neq 0, \ldots, c_{i-1}^2 \neq 0$ единственное решение:

$$c_1 c_i = c_1 b_i + \lambda_1 c_1^2 = 0,$$

$$\ldots$$

$$c_{i-1} c_i = c_{i-1} b_i + \lambda_{i-1} c_{i-1}^2 = 0.$$

Полученный в результате вектор c_i будет ортогонален по отношению к уже имеющимся векторам $c_1, \ldots, c_{i-1}$. Более того, $c_i \neq 0$, так как в противном случае, если учесть свойства скалярного произведения векторов, b_i есть линейная комбинация из $c_1, \ldots, c_{i-1}$, и тем самым подмножество векторов $b_1, \ldots, b_{i-1}$ приходит в противоречие с исходным условием линейной независимости векторов $b_1, \ldots, b_k$.

Таким образом может быть получен каждый из множества ортогонализированных базисов евклидова векторного пространства. При этом одна из ортогональных систем $\{c_1, \ldots, c_n\}$ некоторого n-мерного векторного пространства V называется ортогональным базисом V. Особенное значение имеет ортогональный базис $\{e_1, \ldots, e_n\}$, он наиболее часто встречается в различных задачах МСА. Чтобы получить единичный вектор e_i, компоненты исходного вектора c_i нормируют:

$$\frac{c_i}{\|c_i\|} = e_i.$$

Пример 2.18. Пусть $V = R^3$ — трехмерное арифметическое векторное пространство с базисом $\{(1, 2, 0), (0, 3, 4), (2, 0, 1)\}$ (см. 2.3.1). При $b_1' = (1, 2, 0) = c_1$ имеем

$$c_2 = b_2 + \lambda_1 c_1 = \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix} + \lambda_1 \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix},$$

с произведением $c_1 c_2$:

$$c_1 c_2 = c_1 b_2 + \lambda_1 c_1^2 = 6 + \lambda_1 5 = 0 \Rightarrow \lambda_1 = \frac{6}{5},$$

откуда $c_2 = \begin{pmatrix} -\frac{6}{5} \\ \frac{3}{5} \\ 4 \end{pmatrix}$.

Далее для c_3 имеем $c_3 = b_3 + \lambda_1 c_1 + \lambda_2 c_2$ с произведением $c_1 c_3$:

$$c_1 c_3 = c_1 b_3 + \lambda_1 c_1^2 = 2 + \lambda_1 5 = 0 \Rightarrow \lambda_1 = -\frac{2}{5};$$

$$c_2 c_3 = c_2 b_3 + \lambda_2 c_2^2 = \frac{8}{5} + \lambda_2 \frac{89}{5} = 0 \Rightarrow \lambda_2 = -\frac{8}{89}.$$

Следовательно, $c_3 = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} -\frac{2}{5} \\ -\frac{4}{5} \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{48}{445} \\ -\frac{24}{445} \\ -\frac{32}{89} \end{pmatrix} = \begin{pmatrix} \frac{152}{89} \\ -\frac{76}{89} \\ \frac{57}{89} \end{pmatrix}$.

Полученные результаты легко перепроверить:

$$c_1 c_2 = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} -\frac{6}{5} \\ \frac{3}{5} \\ 4 \end{pmatrix} = 0, \quad c_1 c_3 = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} -\frac{152}{89} \\ -\frac{76}{89} \\ \frac{57}{89} \end{pmatrix} = 0, \quad c_2 c_3 = \begin{pmatrix} -\frac{6}{5} \\ \frac{3}{5} \\ 4 \end{pmatrix} \cdot \begin{pmatrix} \frac{152}{89} \\ -\frac{76}{89} \\ \frac{57}{89} \end{pmatrix} = 0,$$

соответственно

$$\|c_1\| = \sqrt{5},\ \|c_2\| = \sqrt{\frac{89}{5}},\ \|c_3\| = \sqrt{\frac{32129}{7921}} = 2{,}014.$$

В результате получим ортогональный базис:

$$e_1 = \frac{c_1}{\|c_1\|} = \begin{pmatrix} 0{,}4472 \\ 0{,}8944 \\ 0 \end{pmatrix},\ e_2 = \frac{c_2}{\|c_2\|} = \begin{pmatrix} -0{,}2844 \\ 0{,}1422 \\ 0{,}9481 \end{pmatrix},\ e_3 = \frac{c_3}{\|c_3\|} = \begin{pmatrix} 0{,}848 \\ -0{,}424 \\ 0{,}318 \end{pmatrix}.$$

Замечание. Приведенный метод ортогонализации не позволяет получать всегда однозначно определенный ортогональный базис, его результаты зависят от первоначального выбора базиса. Скажем, имеется два различных ортогональных базиса некоторого n-мерного евклидова векторного пространства V: $\{e_1, ..., e_n\}$ и $\{\bar{e}_1, ..., \bar{e}_n\}$. При условии, что

$$e_i e_j = \bar{e}_i \bar{e}_j = \delta_{ij} = \begin{cases} 1, \text{ если } i = j \\ 0, \text{ если } i \neq j \end{cases},$$

т.е. e и $\bar{e}$ — ортонормированные векторы, и

$$\bar{e}_i = \sum_{j=1}^{n} \alpha_{ij} e_j \qquad (i = 1, ..., n),$$

т.е. каждый вектор из V представляется линейной комбинацией векторов из базиса e_j ($j = 1, ..., n$), имеет силу

$$\delta_{ik} = \bar{e}_i \bar{e}_k = \sum_{j,m=1}^{n} a_{ij} a_{km} e_j e_m = \sum_{j,m=1}^{n} a_{ij} a_{km} \delta_{jm} = \sum_{j=1}^{n} a_{ij} a_{ki},$$

т.е. имеется некоторая матрица A, которая может один ортонормированный базис переводить в другой. Такая матрица A будет ортонормирована и для нее выполняется $AA' = I$. Последнее означает, что вектор-строки матрицы A образуют ортонормированный базис R^n. Кроме того, для A действует $A' = A^{-1}$, откуда следует $A^{-1}A = A'A = I$, а значит и вектор-столбцы A также образуют ортонормированный базис R^n. Обобщая сказанное, отметим, что условием ортонормированности матрицы A размерностью $(n \times n)$ является равенство: $A^{-1} = A'$.

Пример 2.19. Имеется ортонормированная матрица (2×2):

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Из ортонормированности векторов $\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$ и $\begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix}$ следует, что $a_{1j}^2 + a_{2j}^2 = 1$ для $j = 1, 2$ и $|a_{ij}| \leq 1$. Дополняя условие, введем величину ζ, $\zeta \in [0, 2\pi]$.

Пусть $a_{11} = \cos\zeta$, $a_{21} = \sin\zeta$. Тогда согласно теореме Пифагора $\cos^2\zeta + \sin^2\zeta = 1$.

Принимая, что $a_{i1}^2 + a_{i2}^2 = 1$ с $i = 1, 2$, можем записать:

$$a_{12}^2 = 1 - a_{11}^2 = 1 - \cos^2\xi = \sin^2\xi,$$
$$a_{22}^2 = 1 - a_{21}^2 = 1 - \sin^2\xi = \cos^2\xi,$$
$$a_{12} = \pm\sin\zeta \text{ и } a_{22} = \pm\cos\zeta.$$

С учетом ортогональности вектор-столбцов матрицы A $a_{11}a_{12} + a_{21}a_{22} = 0$, как и (± 1) (± 1) $\sin\zeta\ \cos\zeta = 0$. Сама матрица A с обозначениями, принятыми для a_{11} и a_{21}, оставаясь ортонормированной, принимает вид:

$$A = \begin{pmatrix} \cos\zeta & -\sin\zeta \\ \sin\zeta & \cos\zeta \end{pmatrix}, \text{ или } \bar{A} = \begin{pmatrix} \cos\zeta & \sin\zeta \\ \sin\zeta & -\cos\zeta \end{pmatrix}.$$

При условии, что $\sin\zeta\ \cos\zeta = 0$ и значения ζ определяются множеством $\zeta \in \left\{0, \frac{\pi}{2}, \pi, \frac{3}{2}\pi\right\}$, получим по крайней мере две из возможных восьми матриц: $\begin{pmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{pmatrix}$ или $\begin{pmatrix} 0 & \pm 1 \\ \pm 1 & 0 \end{pmatrix}$, которые являются производными уже известных матриц A, $\bar{A}$.

2.2.15. Квадратичные формы

Выше в главе скалярное произведение рассматривалось как симметрическая билинейная и положительно определенная форма $(V \times V)$ над полем действительных чисел. При дальнейшем рассмотрении требование положительной определенности опускается и теоретические выводы относительно векторного пространства расширяются при менее строгих исходных посылках.

Симметрические билинейные формы. Пусть имеется векторное пространство V над полем действительных чисел R и функ-

ция $f: \begin{cases} V \times V \to R \\ (x, y) \to f(x, y) \end{cases}$, линейная по каждому из своих двух аргументов.

Назовем f билинейной формой. Симметрической билинейной формой функция f будет при условии, что $f(x, y) = f(y, x)$ для всех $(x, y) \in V \times V$.

Обозначим $\{y_1, ..., y_n\}$ базис некоторого n-мерного векторного пространства V и $(\zeta_{11}, ..., \zeta_{n1})^t = x_1$, как и $(\zeta_{21}, ..., \zeta_{n2})^t = x_2$ вектора из V, тогда билинейную форму $f(x_1, x_2)$ в развернутом виде можем записать:

$$f(x_1, x_2) = \sum_{i=1}^{n}\sum_{j=1}^{n}\zeta_{i1}\zeta_{j2}f(y_i, y_j).$$

В n-мерном векторном пространстве V с базисом $\{y_1, ..., y_n\}$ каждая из существующих билинейных форм f задается квадратичной матрицей

$$B = (f(y_i, y_j)) \quad (i = 1, ..., n; j = 1, ..., n).$$

Введем обозначение $f(y_i, y_j) = s_{ij}$, тогда билинейную $f(x_1, x_2)$ можем переписать так:

$$f(x_1, x_2) = \sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\zeta_{i1}\zeta_{j2}.$$

В данном случае следует обратный по отношению к предыдущему вывод о том, что каждая квадратичная матрица B определяется посредством билинейной формы, существующей в векторном пространстве V и задаваемой уравнением относительно $f(x_1, x_2)$. Матрицу B еще называют матрицей коэффициентов билинейной формы f относительно базиса $\{y_1, ..., y_n\}$.

Квадратичная форма и ее матрица. Введем понятие: *квадратичной формой* называют функцию, определенную на множестве действительных чисел: $q(x) = f(x, x)$, $x \in V$, где f — симметрическая билинейная форма.

Квадратичная форма может быть получена из каждой симметрической $(n \times n)$-матрицы как результат произведения:

$$q_A(x) = x'Ax = \sum_{i=1}^{n}\sum_{j}^{n}a_{ij}x_i x_j. \qquad (2.28)$$

q_A называют квадратичной формой, полученной с матрицей A (A — это матрица квадратичной формы, а ее элементы a_{ij} — коэффициенты квадратичной формы). По определенной симметрической матрице всегда получают также определенную квадратичную форму. Другими словами, для каждой квадратичной формы имеется своя четко заданная симметрическая матрица, для которой выполняется равенство (2.28).

Если имеем n-мерное векторное пространство и A — симметрическую матрицу размерностью $(n \times n)$, то квадратичная форма будет: $q_A(x) = x'Ax$, $x \in V$, т.е. квадратический полином уровня n, обладающий свойствами: $q_A(\lambda x) = \lambda^2 q_A(x)$ для $\lambda \in R$.

Матрица квадратичной формы A будет:

- положительно определена, если $q_A(x) > 0$, $x \neq 0$;
- положительно полуопределена, если $q_A(x) \geq 0$, для всех x;
- отрицательно определена, если $q_A(x) < 0$, $x \neq 0$;
- отрицательно полуопределена, если $q_A(x) \leq 0$, для всех x,

и остается неопределенной, если имеет место какой-либо другой, не подпадающий под четыре приведенных выше, случай, т.е. если два вектора $y, z \in V$ имеют квадратичные формы: $q_A(y) < 0$ и $q_A(z) > 0$.

Пример 2.20. Имеется матрица A размерностью (2×2), т.е. $n = 2$:

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

и квадратичная форма с A:

$$q_A(x) = (x_1\ x_2)\begin{pmatrix} a & b \\ b & c \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (x_1\ \ x_2)\begin{pmatrix} ax_1 + bx_2 \\ bx_1 + cx_2 \end{pmatrix} = ax_1^2 + 2bx_1x_2 + cx_2^2.$$

Сформулируем основные правила для $\det A$ и $q_A(x)$:

$\det A > 0 \Rightarrow q_A(x)$ — положительно определена, если $a > 0$, отрицательно определена, если $a < 0$;

$\det A = 0 \Rightarrow q_A(x)$ — положительно полуопределена, если $a > 0$, или $a = 0$, $c \leq 0$, отрицательно полуопределена, если $a < 0$, или $a = 0$, $c \geq 0$;

$\det A < 0 \Rightarrow q_A(x)$ — неопределенна.

Покажем справедливость выдвинутых правил относительно определенности $\det A$. При условии, что $a \neq 0$, квадратичную форму матрицы A легко представить как полином второй степени:

$$q_A(x) = ax_1^2 + 2bx_1x_2 + cx_2^2,$$

после введения квадратичного дополнения получим:

$$a\, q_A(x) = (ax_1 + bx_2)^2 + (\det A)\, x_2^2,$$

Отсюда

$$q_A(x) = \frac{1}{a}(ax_1 + bx_2)^2 + \frac{ac - b^2}{a} x_2^2$$

или

$$q_A(x) = a\left(x_1 + \frac{b}{a}x_2\right)^2 + \frac{ac - b^2}{a} x_2^2. \qquad (2.29)$$

Из уравнения (2.29) видно, что $q_A(x) > 0$, если $a > 0$ и $\frac{ac - b^2}{a} > 0$. Квадратичная форма также будет положительно определенной: ($q_A(x) > 0$) с $ac - b^2 = \det A > 0$. Аналогичным образом легко устанавливается, что $q_A(x) < 0$, если $a < 0$ и $\det A > 0$.

В случае, когда $\det A = 0$, равенство (2.29) преобразуется в:

$$q_A(x) = a\left(x_1 + \frac{b}{a}x_2\right)^2,$$

и, следовательно, $q_A(x)$ положительно полуопределена, если $a > 0$, и отрицательно полуопределена, если $a < 0$.

Последний из представленных случаев, когда $\det A < 0$, очевидно, означает, что при заданных a, b и c квадратичная форма $q_A(x)$ независимо от значений x сама может принимать большие или меньшие значения и, таким образом, является величиной неидентифицируемой.

Выводы[1]

В главе 2, написанной д-ром Мартином Шефером, приводятся основополагающие сведения из двух больших разделов математики: аналитической геометрии и матричной алгебры. Эти сведения — важный элемент подготовки специалистов, занимающихся многомерным статистическим анализом. Необходимость математических знаний обусловливается конструктивными особенностями методов МСА, предполагающих в подавляющем большинстве случаев пространственное представление наблюдаемых многомерных объектов и рациональное описание этих объектов средствами формального языка, в том числе при помощи векторов и матриц.

В §2.1 последовательно рассматриваются понятия вектора, векторного пространства, функции, подпространства, линейной оболочки и образующей системы. §2.2 начинается определениями трансформации базиса векторного пространства, матрицы, линейного оператора, затем освещаются технические вопросы исчисления ранга и детерминанты матрицы, решения систем линейных уравнений, поиска собственных векторов и собственных чисел матрицы, построения и использования в анализе квадратичной формы.

В изложении основ математики автор не выходит за рамки евклидовой геометрии и понятия линейной связи. Таким образом, глава представляет хорошо разработанную область математической науки и ее материал доступен для восприятия широкого круга читателей. В связи с этим отметим, что новейшие разработки теории МСА нередко опираются на менее известные и более сложные математические методы, использующие в своей основе предположения о нелинейности признаковых связей и расширенные представления о теоретическом пространстве (унитарное, афинное пространства и т.п.). Однако эти методы в настоящей книге не рассматриваются. С одной стороны, в силу сложности и объемности исходного теоретического материала они выступают предметом для самостоятельного изучения, с другой стороны, корректная и эффективная работа с этими методами обусловливается знаниями классической теории МСА.

[1] Параграф подготовлен В.Н. Тамашевичем.

Вопросы и задачи[1]

1. Проверьте:

а) являются ли векторы a и b линейно независимыми:

$$a = \begin{pmatrix} 4 \\ 12 \\ 1 \\ 7 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ -2 \\ 1 \\ 5 \end{pmatrix};$$

[1] Подготовлены Л.А. Сошниковой и В.Н. Тамашевичем.

б) являются ли векторы $a_1' = (1, 3, 1, 3)$, $a_2' = (2, 1, 1, 2)$ и $a_3' = (3, -1, 1, 1)$ линейно зависимыми.

2. Покажите, что векторы $a_1' = (1, 0, 0)$, $a_2' = (0, 1, 0)$, $a_3' = (0, 0, 1)$ образуют ортонормированное векторное пространство.

3. Найдите скалярное произведение векторов a и b:

$$a' = (2\ 3\ 8),\quad b = \begin{pmatrix} 4 \\ 7 \\ 16 \end{pmatrix}.$$

4. Найдите произведение матриц A, B и C, если

$$A = \begin{pmatrix} 4 & 3 \\ 7 & 5 \end{pmatrix};\ B = \begin{pmatrix} -28 & 43 \\ 38 & -12 \end{pmatrix};\ C = \begin{pmatrix} -7 & 3 \\ 2 & 4 \end{pmatrix}.$$

5. Определите матрицу показателей стоимости продукции по ее видам и сортам, если имеются матрица A — объемов выпуска продукции по видам и сортам и матрица B — уровней цен на каждый вид и сорт продукции:

$$\text{вид}\downarrow\ \text{сорт}\rightarrow \quad A = \begin{pmatrix} 50 & 26 & 120 \\ 70 & 9 & 45 \\ 84 & 32 & 225 \end{pmatrix},\qquad \text{вид}\downarrow\ \text{сорт}\rightarrow \quad B = \begin{pmatrix} 5 & 7 & 12 \\ 25 & 30 & 58 \\ 20 & 22 & 26 \end{pmatrix}.$$

6. Вычислите определители матриц A и B:

$$A = \begin{pmatrix} 1 & 2 & -1 \\ 2 & -3 & -1 \\ 4 & -1 & -5 \end{pmatrix};\ B = \begin{pmatrix} -1 & 2 & 3 & 4 \\ 2 & -3 & 4 & 5 \\ -3 & 4 & 4 & 2 \\ 4 & 3 & -2 & 3 \end{pmatrix}.$$

7. Определите, имеет ли матрица A обратную матрицу (A^{-1}), если имеет, то вычислите ее:

$$A = \begin{pmatrix} 5 & -8 & -5 \\ -4 & 7 & -1 \\ -6 & 5 & 9 \end{pmatrix}.$$

8. Найдите ранг матрицы:

$$A = \begin{pmatrix} 2 & 5 & 6 \\ 4 & -1 & 5 \\ 2 & -6 & -1 \end{pmatrix}.$$

9. Определите максимальное число линейно независимых строк матрицы A:

$$A = \begin{pmatrix} 2 & 0 & 3 & 2 & 1 \\ 1 & 3 & 1 & 7 & 5 \\ 0 & 3 & -5 & -3 & 3 \\ 2 & 3 & -2 & 2 & 4 \end{pmatrix}.$$

10. Используя метод Гаусса, решите систему уравнений:

$$\begin{cases} x_1 + 2x_2 - x_3 = 9, \\ 2x_1 - 3x_2 + x_3 = 3, \\ 4x_1 + x_2 - x_3 = 16. \end{cases}$$

11. По приведенным в таблице фрагментарным данным межотраслевого баланса за отчетный период (усл. ден. ед.) вычислите объем валового выпуска каждой отрасли при условии, что:

а) конечный спрос энергетической отрасли увеличивается вдвое, а машиностроения сохраняется на прежнем уровне;

б) конечный спрос энергетической отрасли уменьшится в 1,8 раза, а машиностроения возрастет на 20%.

Отрасль		*Промежуточное потребление*		*Конечный спрос*	*Валовой выпуск*
		Энергетика	*Машиностроение*		
Производство	Энергетика	10	15	75	100
	Машиностроение	22	25	53	100

12. В базисе e_1, e_2, e_3 заданы векторы $a_1' = (1, 1, 0)$, $a_2' = (1, -1, 1)$ и $a_3' = (-3, 5, -6)$. Покажите, что векторы a_1, a_2, a_3 также образуют свой базис.

13. Вектор $b = (4, -4, 5)$ задан в базисе e_1, e_2, e_3. Выразите этот вектор в базисе a_1, a_2, a_3, который задан векторами: $a_1' = (1, 1, 0)$, $a_2' = (1, -1, 1)$ и $a_3' = (-3, 5, -6)$.

14. Пусть в пространстве R^3 линейный оператор A в базисе e_1, e_2, e_3 задан матрицей:

$$A = \begin{pmatrix} 3 & 2 & 4 \\ -1 & 5 & 6 \\ 1 & 8 & 2 \end{pmatrix}.$$

Найдите образ ($y = A(x)$) вектора $x = 4e_1 - 3e_2 + e_3$.

15. В базисе e_1, e_2 линейный оператор преобразования A имеет матрицу:

$$A = \begin{pmatrix} 17 & 6 \\ 6 & 8 \end{pmatrix}.$$

Найдите матрицу оператора A в базисе $e_1^* = e_1 + 2e_2$ и $e_2^* = -2e_1 + e_2$.

16. Найдите собственные числа и собственные векторы линейного оператора A, заданного матрицей:

$$A = \begin{pmatrix} 1 & 4 \\ 7 & 1 \end{pmatrix}.$$

17. Задана квадратичная форма: $q(x_1, x_2) = 2x_1^2 + 4x_1x_2 - 3x_2^2$. Найдите производную для нее квадратичную форму $q'(y_1, y_2)$, представляющую линейные преобразования вида $x_1 = 2y_1 - 3y_2$; $x_2 = y_1 + y_2$.

18. Найдите собственные числа и собственные векторы линейного оператора, заданного матрицей:

$$A = \begin{pmatrix} 5 & 4 \\ -1 & -3 \end{pmatrix}.$$

19. Запишите в матричном виде квадратичную форму: $q(x) = 2x_1^2 + 3x_2^2 - x_3^2 + 4x_1x_2 - 6x_1x_3 + 10x_2x_3$.

Глава 3

Случайные величины. Законы распределения и плотность вероятностей

3.1. Случайные величины и их распределения

Исследователю нередко приходится сталкиваться с серьезными трудностями при выявлении, адекватном описании закономерностей, лежащих в основе развития стохастических явлений и процессов. Эти трудности в значительной мере преодолеваются, если понятие вероятности использовать относительно не только непосредственно наблюдаемых событий, но и множества всех возможных элементарных результатов их наступления или ненаступления, представленных некоторыми действительными числами. При этом получают как бы новое, в отличие от эмпирического, теоретическое пространство событий, с элементами, также являющимися действительными числами. Понятие вероятности в этом теоретическом пространстве (R) определяется соответственно понятию вероятности одномерной или многомерной случайной величины.

Случайная величина. *Случайной величиной* X называют множество элементарных событий, каждое из которых на интервале в поле действительных чисел R определяется с некоторой известной вероятностью $P(\Omega)$:

$$X: \Omega \to R.$$

Замечание. Если некоторое множество Ω используется для описания случайного (стохастического) процесса, то результаты последнего отражаются четко установленным рядом действительных чисел, представляющим состав определенной случайной величины.

Пример 3.1. Во время теннисного турнира игрок B участвует в трех играх. Победа в каждой из них засчитывается при условии

выигрыша двух партий. Представим пространство с возможными результатами игр B:

$\Omega = \{(0, 0, 0),$
$(0, 0, 1), (0, 1, 0), (1, 0, 0),$
$(0, 1, 1), (1, 0, 1), (1, 1, 0), (2, 0, 0), (0, 2, 0), (0, 0, 2),$
$(1, 1, 1), (0, 1, 2), (1, 0, 2), (1, 2, 0), (0, 2, 1), (2, 0, 1), (2, 1, 0),$
$(0, 2, 2), (2, 0, 2), (2, 2, 0), (2, 1, 1), (1, 2, 1), (1, 1, 2),$
$(2, 2, 1), (2, 1, 2), (1, 2, 2),$
$(2, 2, 2)\}$.

Определим вероятное число выигрышей игрока B в турах по три игры со случайными исходами ($X: \Omega \to R$):

$X(0, 0, 0) = 0$
$X(0, 0, 1) = X(0, 1, 0) = X(1, 0, 0) = 1$
$X(0, 1, 1) = X(1, 0, 1) = X(1, 1, 0) = X(2, 0, 0) = X(0, 2, 0) = X(0, 0, 2) = 2$
$X(1, 1, 1) = X(0, 1, 2) = X(1, 0, 2) = X(1, 2, 0) = X(0, 2, 1) = X(2, 0, 1) = X(2, 1, 0) = 3$
$X(0, 2, 2) = X(2, 0, 2) = X(2, 2, 0) = X(2, 1, 1) = X(1, 2, 1) = X(1, 1, 2) = 4$
$X(2, 2, 1) = X(2, 1, 2) = X(1, 2, 2) = 5$
$X(2, 2, 2) = 6$.

Функциональное значение x, принимаемое случайной величиной X, есть некоторое число $x \in R$, появляющееся с определенным уровнем вероятности $X^{-1}(x)$ и представляющее одно из множества событий $A \subseteq \Omega$. Число $x = X(A)$ называют *реализацией случайной величины X*.

Распределение случайной величины. *Функцией распределения случайной величины X* называют функцию F: $R \to [0, 1]$, определенную на интервале $(-\infty, x]$ с вероятностью

$P(\{\omega \in \Omega \mid X(\omega) \in (-\infty, x]\})$.

Замечание. Функция F на всем множестве значений x является *монотонно возрастающей*, она не может принимать отрицательных или превышающих единицу значений.

Функция плотности вероятностей. Пусть W есть множество с некоторым, бесконечно большим, числом элементов, тогда функцию типа:

f: $R \to [0, 1]$, где $f(x) = P(X = x) = P(\{\omega \mid X(\omega) = x\})$

называют *функцией плотности вероятностей* случайной величины X, а X — дискретной случайной величиной.

Распределение дискретной случайной величины. Для дискретной случайной величины X с ранжированными значениями

$x_1 \le x_2 \le \ldots \le x_k$ функция распределения записывается следующим образом:

$$F_k := F(x_k) := P[X \le x_k] = \sum_{i=1}^{k} p_i = \sum_{i=1}^{k} P(X = x_i).$$

Функция распределения случайной величины X принимает значения:

а) $0 \le F_k \le 1$,
б) $0 = F_1$,
в) $1 = F_k$ — монотонно неубывающая функция (монотонно возрастающая функция).

Доказательство положений а) — в) для F_k следует непосредственно из свойств функций плотности вероятностей. В том, что распределение F_k характеризуется функцией ступенчатого вида, не убывающей справа, убедимся, построив график распределения случайной величины X (рис. 3.1).

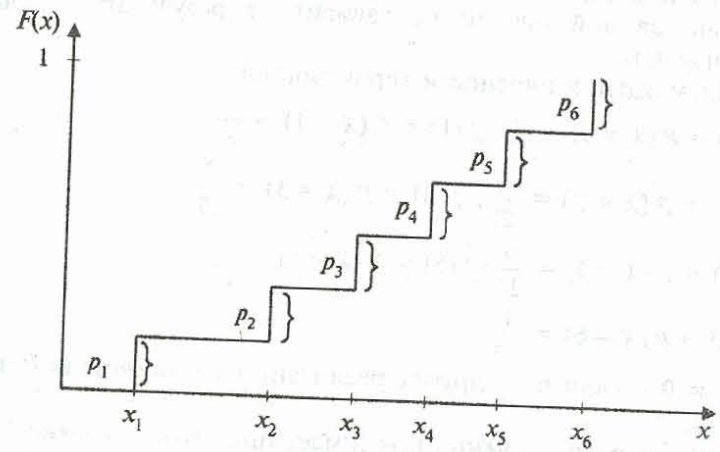


Рис. 3.1. Распределение случайной величины X

Замечание. Если Ω — некоторое конечное множество, то имеются также и реализации $x_1, x_2, \ldots, x_k$, и функция распределения X, соответствующие эмпирической функции распределения. Область определения для функции распределения заключена в интервале от нуля до единицы:

$F_{1-} := F(x_1) := P[X < x_1] = 0$,
$F_{k+} := F(x_{k+1}) := P[x < x_k + d] = 1$ $(d > 0)$.

Понятие распределения случайной величины позволяет определять вероятность появления ее различных реализаций на любом интервале:

$P[u \le X \le o] = F_o - F_u$.

Обратим внимание на существование различных случаев интервальной определенности вероятностных характеристик:

$P[u < X < o] = F_{o-1} - F_u$,

$P[u \le X < o] = F_{o-1} - F_{u-1}$,

$P[u < X \le o] = F_o - F_u$,

$P[u \le X \le o] = F_o - F_{u-1}$.

Соответственно функции распределения случайной величины определяется и ее плотность вероятностей:

$p_k = P[X = x_k] = F_k - F_{k-1} \quad (k = 2, 3, ...)$

$p_1 = P[X = x_1] = F_1$.

Пример 3.2. В продолжение примера 3.1 рассмотрим случай, когда игрок *B* выигрывает одну партию, причем выигрыш или проигрыш данной партии не зависит от результатов игры в других партиях.

Найдем оценки плотности вероятностей:

$$f(0) = P(X=0) = \frac{1}{27}; f(1) = P(X=1) = \frac{3}{27};$$

$$f(2) = P(X=2) = \frac{6}{27}; f(3) = P(X=3) = \frac{7}{27};$$

$$f(4) = P(X=4) = \frac{6}{27}; f(5) = P(X=5) = \frac{3}{27};$$

$$f(6) = P(X=6) = \frac{1}{27}$$

и $f(x) = 0$ — для всех других реализаций случайной величины $X (x \in R)$.

Функция распределения принимает при этом значения:

$$F(0) = P(X \le 0) = \frac{1}{27}; F(1) = P(X \le 1) = f(0) + f(1) = \frac{4}{27};$$

$$F(2) = P(X \le 2) = \frac{10}{27}; F(3) = P(X \le 3) = \frac{17}{27};$$

$$F(4) = P(X \le 4) = \frac{23}{27}; F(5) = P(X \le 5) = \frac{26}{27}; F(6) = P(X \le 6) = 1.$$

Определение плотности вероятностей. Если функция распределения F случайной величины X не дифференцируема в некоторой точке из всего сколь угодно большого их множества

$(x_i; i = 1, 2,...)$, то по крайней мере в этой точке существует дискретная функция $f(x)$, представляющая производную от $F(x)$:

$$f(x) = F'(x) = \frac{dF(x)}{dx} \quad \text{для } x \ne x_i \ (i = 1, 2, ...).$$

Такая функция $f(x)$ называется *функцией плотности вероятностей* случайной величины X, которая рассматривается как непрерывная величина.

Замечание. Так как F — монотонно возрастающая (не убывающая) функция, то очевидно, что функция плотности вероятностей будет всегда *положительно определенной*:

$f(x) \ge 0$ для всех $x \in R$.

Принимая, что $\lim\limits_{x \to \infty} F(x) = 1$, можем записать

$$\int_R f(x)\,dx = 1.$$

И, наоборот, функция распределения случайной величины в свою очередь может быть получена из функции плотности распределения:

$$F(x) := P[X \le x] = \int_{-\infty}^{x} f(u)du.$$

Как мы видели выше, вероятность реализации случайной величины X определяется интервалом существования последней:

$P[u \le X \le o] = F(o) - F(u)$.

При предельном сокращении интервала существования X вероятность реализации обращается в нуль:

$\lim\limits_{u \to 0} P[u \le X \le o] = 0$ и $\lim\limits_{0 \to u} P[u \le X \le o] = 0$.

Другими словами, вероятность реализации некоторого точно определенного значения X равна нулю.

Пример 3.3. Пусть существует функция распределения F: $R \to [0, 1]$

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{3}x^2, & 0 \le x < 1 \\ \frac{1}{3}, & 1 \le x < 2 \\ \frac{2}{3}x - 1, & 2 \le x < 3 \\ 1, & x \ge 3, \end{cases}$$

т.е. $F(x)$ — частично дифференцируемая функция распределения для случайной величины X. График этой функции имеет вид (рис. 3.2).

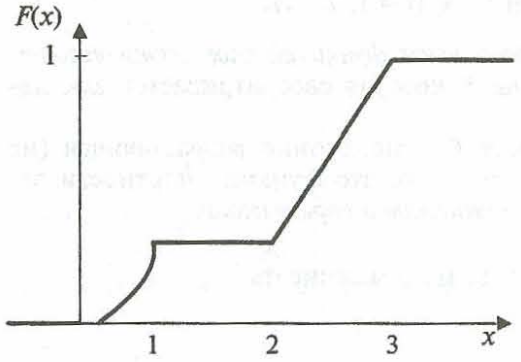


Рис. 3.2

Очевидно, что в точках $x_1 = 1$, $x_2 = 2$ и $x_3 = 3$ функция $F(x)$ — не дифференцируема. Найдем для $F(x)$ первую производную и тем самым перенесем $F(x)$ на другие опорные точки. В результате получим частично непрерывную функцию плотности вероятностей $f: \boldsymbol{R} \to \boldsymbol{R}$:

$$f(x) = \begin{cases} 0, & x < 0 \\ \frac{2}{3}x, & 0 \le x < 1 \\ 0, & 1 \le x < 2 \\ \frac{2}{3}, & 2 \le x < 3 \\ 0, & x \ge 3. \end{cases}$$

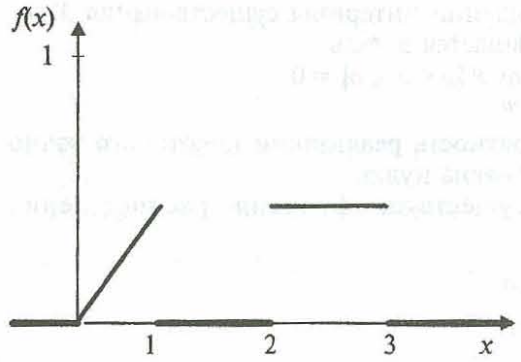


Рис. 3.3

Новая функция $f(x)$ в точках недифференцируемости $F(x)$, а именно $x_1 = 1$, $x_2 = 2$ и $x_3 = 3$, дискретна (рис. 3.3):

Пример 3.4. Общий случай для любого распределения с графиком любого вида на интервале $[a, b]$. Пусть имеется некоторая функция плотности распределения

$f: [0, 1] \to \boldsymbol{R}$ и $f(x) = \frac{3}{2}(1 - x^2)$, $0 \le x \le 1$.

График функции $f(x)$ принимает вид:

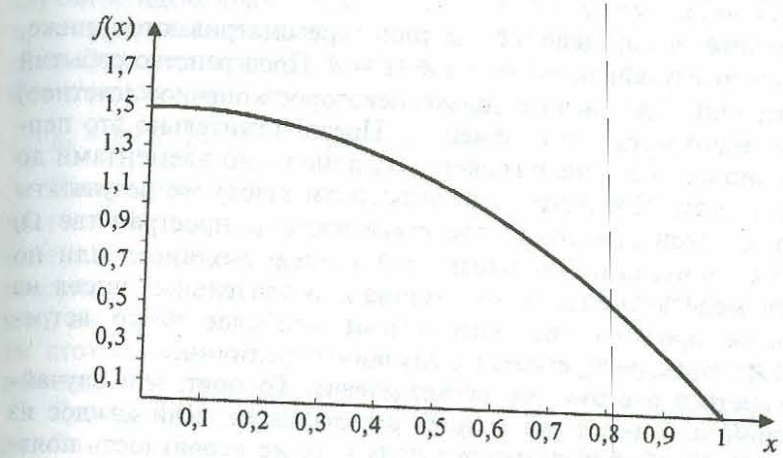


Рис. 3.4

Функция распределения, принадлежащая функции плотности $f(x)$, $F: [0, 1] \to [0, 1]$, $F(x) = \frac{3}{2}x\left(1 - \frac{x^2}{3}\right)$, $0 \le x \le 1$, и имеет график:

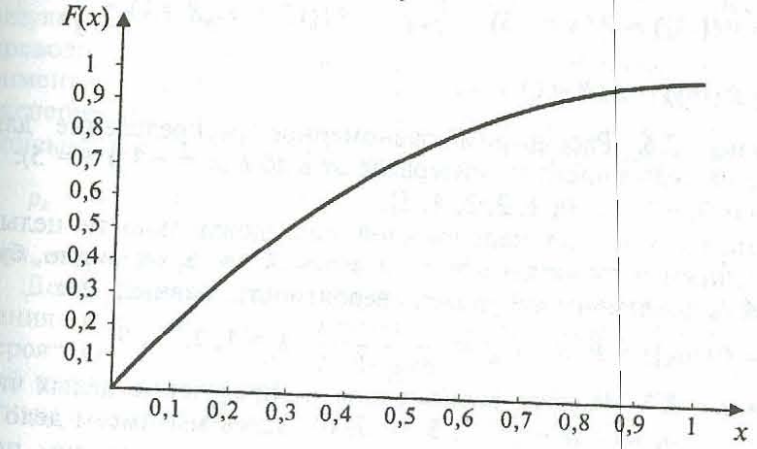


Рис. 3.5

3.2. Некоторые виды параметрических распределений

В основе распределений, которые рассматриваются ниже, всегда лежит случайная величина X: $\Omega \to R$. Пространство событий содержит при этом вначале только некоторое конечное (счетное) число из всего множества элементов. Предположительно это первоначально счетное число может быть дополнено элементами до бесконечности. Последнее в сущности означает, что результаты случайных экспериментов, представленные в пространстве Ω, являются элементами множества действительных чисел или по крайней мере элементами множества действительных чисел на некотором пространстве. Рассмотрим наиболее часто встречающиеся виды распределения случайных величин.

Дискретное равномерное распределение. Говорят, что случайная величина X имеет *равномерное распределение*, если каждое из множества N событий сохраняет одну и ту же вероятность появления, т.е.

$\Omega = \{\omega_1, \omega_2, ..., \omega_N\}$, и

$$P_k = P(\{\omega_k\}) = P(X = k) = \frac{1}{N}, \; k = 1, 2, ..., N.$$

Пример 3.5. Имеется равномерное распределение $\Omega = \{-3, 1, 6\}$. Вероятность реализации каждого элемента из Ω равна 1/3:

$$p_1 = P(\{-3\}) = P(X = -3) = \frac{1}{3}, \; p_2 = P(\{1\}) = P(X = 1) = \frac{1}{3},$$

$$p_3 = P(\{6\}) = P(X = 6) = \frac{1}{3}.$$

Пример 3.6. Рассмотрим равномерное распределение для множества целых чисел на интервале от a до b ($a = -3$ и $b = 5$):

$\Omega = \{-3, -2, -1, 0, 1, 2, 3, 4, 5\}$.

В этом случае для непрерывной последовательности целых чисел с нижней границей $a = -3$ и верхней $b = 5$, очевидно, будем иметь поэлементный уровень вероятности, равный 1/9:

$$p_k = P(\{\omega_k\}) = P(X = \omega_k) = \frac{1}{b-a+1} = \frac{1}{9}, \; k = 1, 2, ..., 9.$$

Пример 3.7. Имеется равномерное распределение целых чисел от 1 до 6, или $\Omega = \{1, 2, 3, 4, 5, 6\}$. Здесь мы имеем дело с дискретным равномерным распределением, описывающим последовательность натуральных чисел без пропусков, на интервале

с нижней границей $a = 1$ и верхней $b = 6$. Это так называемый случай с идеальным игральным кубиком, на сторонах которого проставлены числа от 1 до 6. Естественно, как и в предыдущих примерах, вероятность появления каждого из натуральных чисел будет одной и той же:

$$p_k = P(\omega_k\}) = P(X = \omega_k) = \frac{1}{b-a+1} = \frac{1}{6}, \; k = 1, 2, 3, 4, 5, 6.$$

Распределение Бернулли. Если пространство событий заключает в себя только два элемента $\Omega = \{\omega_1, \omega_2\}$ и каждое событие наступает с вероятностью $P(\omega_1) = p$, $P(\omega_2) = 1 - p$, $0 < p < 1$, то говорят, что случайная величина имеет *распределение Бернулли*, или, в сокращенной записи, $X \sim B(1, p)$. Всевозможные результаты стохастического процесса в этом случае представляются при помощи двух чисел: 0 и 1.

Пример 3.8. Приведем некоторые из многочисленных примеров бинарной значимости случайной величины из практики:

	0	1
1. Загрузка компьютерного чипса	Включено	Выключено
2. Положение выключателя	Мальчик	Девочка
3. Пол ребенка	Орел	Решка
4. Результат подбрасывания монеты		

Случай, когда вероятность наступления каждого из двух возможных событий одинакова ($p = 1/2$), характеризуется как случай (модель) с идеальной монетой.

Биномиальное распределение. Если случайный эксперимент, результаты которого подчиняются распределению Бернулли, проводится не один, а n раз, причем каждый новый из n экспериментов независим от предыдущих, то в бинарных результатах экспериментов $\Omega = \{\omega_1, \omega_2\} = \{0, 1\}$ частота появления единиц описывается биномиальным законом плотности распределения:

$$p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \; 0 < p < 1, \; k = 0, 1, 2, ..., n$$

или, в сокращенной форме записи, $X \sim B(n, p)$

Доказательство тому, что биномиальная плотность распределения чисел согласуется с тремя основополагающими аксиомами вероятности, легко найти в теории биномов. На самом деле случай плотности распределения Бернулли является, очевидно, частным случаем биномиального распределения при $n = 1$:

$$p_0 = \binom{1}{0} p^0 (1-p)^{1-0} = 1 - p, \; p_1 = \binom{1}{1} p^1 (1-p)^{1-1} = p.$$

Закон биномиальной плотности распределения можно наглядно представить в виде дерева (рис. 3.6). Пусть левые ветви такого дерева отражают случайный процесс появления единиц, а правые — нулей. С учетом взаимной независимости проводимых экспериментов, следуя вниз по ветвям, можно получить мультипликативные вероятностные результаты p и $(1-p)$. В конце дерева определяется точное число появившихся единиц:

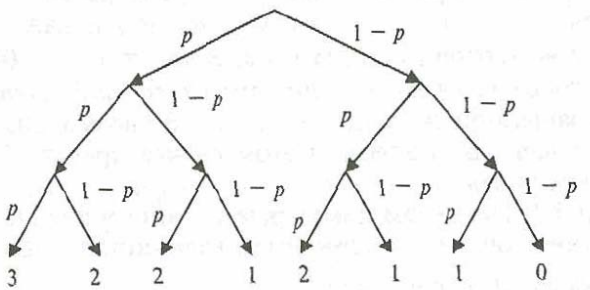


Рис. 3.6

Особенное теоретическое значение в подобного рода экспериментах приобретают случаи совершенно симметрического (идеального) распределения при $p = 1/2$.

Пример 3.9. Идеальная монета подбрасывается некоторое определенное число раз. Определим вероятность числа раз появления орла, равного в общем числу появлений решки:

$$P(X=n)=\binom{2n}{n}\left(\frac{1}{2}\right)^{n}\left(\frac{1}{2}\right)^{2n-n}=\binom{2n}{n}\left(\frac{1}{2}\right)^{2n}=\frac{(2n)!}{n!\,n!}\left(\frac{1}{2}\right)^{2n}.$$

Для $2n = 4$ — числа экспериментов и возможных случайных результатов: A — орел, Z — решка, X — число выпадения орла в четырех подбрасываниях, имеем пространство событий Ω:

		ZAAZ		
		ZAZA		
	ZAAA	*AZAZ*	*ZZZA*	
	AZAA	*ZZAA*	*ZZAZ*	
	AAZA	*AZZA*	*ZAZZ*	
AAAA	*AAAZ*	*AAZZ*	*AZZZ*	*ZZZZ*
$X=4$	$X=2$	$X=2$	$X=1$	$X=0$

При этом $P(X=2)=\binom{4}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{4-2}=\binom{4}{2}\left(\frac{1}{2}\right)^{4}=\frac{4!}{2!2!}\left(\frac{1}{2}\right)^{4}=\frac{3}{8}.$

Геометрическое распределение. В отличие от двух предыдущих примеров распределений пространство событий данного распределения не конечно, а предположительно включает бесконечное число результатов:

$\Omega = \{\omega_1, \omega_2, ..., \omega_{n-1}, \omega_n, ...\} = \{0, 1, ..., n-2, n-1, ...\}$.

Плотность геометрического распределения при этом определяется по формуле:

$p_k = P(X=k) = p(1-p)^k, \quad 0 < p < 1, \quad k = 0, 1, 2, ...,$

а сама случайная величина X называется в этом случае геометрически распределенной.

Наглядно плотность вероятностей геометрически распределенной случайной величины представляется, по аналогии с биномиальным распределением, в виде дерева событий (рис. 3.7), ветви которого слева отражают случайный процесс наступления события, а справа — его ненаступления:

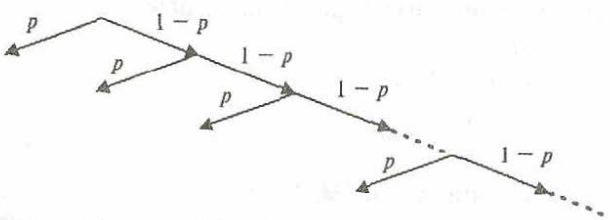


Рис. 3.7

Геометрическая плотность распределения p_k описывает вероятность появления положительного результата (наступления события) после получения первых k отрицательных результатов. Аналогично биномиальной плотности распределения геометрическая плотность распределения также укажет на вероятность появления первой единицы, однако с тем решающим отличием, что эксперимент здесь продолжается именно до появления этой первой единицы. Теоретически эксперимент может длиться бесконечно долго.

Геометрическое распределение, очевидно, можно представить как сумму вероятностных характеристик:

$$\sum_{i=0}^{k} p_i = \sum_{i=0}^{k} p(1-p)^k = p\frac{1-p^{k+1}}{1-(1-p)} = 1-p^{k+1}.$$

Формальная запись геометрически распределенной случайной величины имеет вид: $X \sim NB(1, p)$.

Распределение Паскаля. Представляет собой обобщение геометрического распределения, его также определяют как отрицательное (негативное) биномиальное распределение, записывается: $X \sim NB(r, p, r = 1, 2, 3, ...)$.

В данном случае имеют дело с вероятностью того, что $r(r \geq 1)$ — число положительных результатов опытов — появляется после k отрицательных опытов (ненаступления события):

$P\{k$ отрицательных результатов перед r-м положительным результатом, $r \geq 1\} = P(X = k) = \binom{r+k-1}{k} p^r (1-p)^k$;

$k = 0, 1, 2, ...; r = 1, 2, ...$

Для частного случая $r = 1$ имеет место геометрическое распределение плотности вероятностей $X \sim NB(1, p)$.

Гипергеометрическое распределение. Для натуральных чисел M, N, n, k при $0 \leq k \leq m = \min\{n, M\}$, $M \leq N$ с пространством событий $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ плотность гипергеометрического распределения задается величиной вероятности, равной:

$$p_k = P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, 2, ..., n.$$

В сокращенной записи: $X \sim H(M, N, n)$.

Моделью данного распределения может служить эксперимент с урной, в которой находятся N шаров, в том числе M шаров имеют определенный цвет, а оставшиеся $N - M$ шаров — какой-либо другой цвет (скажем, красный и голубой цвета). Из урны шары выбираются случайным образом и без возврата; после отбора n шаров ($n < m$), при этом получаются следующие результаты (табл. 3.1).

Исходя из равенства $\binom{N}{n}$ сумме различных комбинаций результатов эксперимента, показанных в правом крайнем столбце, следует признать существование плотности распределения случайной величины, для которой правомерна запись:

$$0 \leq p_k \leq 1, \quad \sum_{k=0}^{M} p_k = 1.$$

Таблица 3.1

Красные шары	*Голубые шары*	*Различные комбинации результатов эксперимента*
0	$n - 0$	$\binom{M}{0}\binom{N-M}{n}$
1	$n - 1$	$\binom{M}{1}\binom{N-M}{n-1}$
2	$n - 2$	$\binom{M}{2}\binom{N-M}{n-2}$
3	$n - 3$	$\binom{M}{3}\binom{N-M}{n-3}$
.	.	.
.	.	.
k	$n-k$	$\binom{M}{k}\binom{N-M}{n-k}$
.	.	.
.	.	.
.	.	.
n	0	$\binom{M}{n}\binom{N-M}{0}$

Пример 3.10. Обратимся к проблеме выигрыша в лотто «6 из 49». На бланке лотерейного билета записано 49 чисел ($N = 49$), из них следует указать 6 чисел ($M = 6$), которые впоследствии могут оказаться выигрышными.

Пусть на билете зачеркнуты какие-либо 6 чисел ($n = 6$), тогда p_k будет вероятностью попадания из n зачеркнутых k выигрышных чисел. Обычно вероятность для $k \geq 3$ остается весьма малой, в этом можно убедиться по данным табл. 3.2. Хотя игры-лотто весьма распространены в Европе и Северной Америке, по данным таблицы видно, что скорее они иррациональны и предполагают некоторый психологический настрой (как, например, стать миллионером), нежели обоснованный научный расчет.

Определим вероятность правильного выбора чисел на карте лотто:

Таблица 3.2

k	p_k
0	$p_0 = \binom{6}{0}\binom{43}{6}/\binom{49}{6} = 6\,096\,454/13\,983\,816 = 0{,}436$
1	$p_1 = \binom{6}{1}\binom{43}{5}/\binom{49}{6} = 5\,775\,588/13\,983\,816 = 0{,}413$
2	$p_2 = \binom{6}{2}\binom{43}{4}/\binom{49}{6} = 1\,851\,150/13\,983\,816 = 0{,}1324$
3	$p_3 = \binom{6}{3}\binom{43}{3}/\binom{49}{6} = 246\,820/13\,983\,816 = 0{,}017$
4	$p_4 = \binom{6}{4}\binom{43}{2}/\binom{49}{6} = 13\,545/13\,983\,816 = 0{,}001$
5	$p_5 = \binom{6}{5}\binom{43}{1}/\binom{49}{6} = 258/13\,983\,816 = 0{,}00001845$
6	$p_6 = \binom{6}{6}\binom{43}{0}/\binom{49}{6} = 1/13\,983\,816 = 0{,}000000072$

Пример 3.11. Рассмотрим задачу статистического оценивания случайной величины по данным некоторой генеральной совокупности. Ниже будет показано, что такая задача предполагает решение с определением неизвестных параметров распределения, например, определение неизвестного параметра N для распределения $H(M, N, n)$.

В случае гипергеометрического распределения часто возникает ситуация, когда параметр M известен, а параметр N нет, например, при определении величины популяции вида птиц, которому угрожает вымирание. Через некоторый промежуток времени после того, как с исследовательской целью M птиц отловлено, окольцовано и вновь отпущено на свободу, предполагается проведение контрольных мероприятий. Птицы данного вида, на этот раз их число n, вновь отлавливаются и проверяются по признаку наличия кольца, таких может быть k птиц. Полученные данные позволяют установить общее число птиц в популяции — N. Так как плотность распределения $p_k = p_k(N)$ представляет одно из всех N значений унимодальной функции, запишем равенство соотношений:

$$p_k(N) / p_k(N-1) = [(N-M)(N-n)] / [N-M-n-k].$$

При максимальном значении N наибольшая из оценок плотности вероятностей для наблюдаемой выборки $\{M, n, k\}$ равна единице:

$$p_k(N) / p_k(N-1) = 1,$$

откуда следует условие $M / N = k / n$, т.е. как в генеральной совокупности, так и в выборке удельные веса наблюдений, обладающих признаками «наличие кольца» и «нет кольца», распределяются равномерно.

Из последнего равенства получим:

$$N = Mn / k.$$

Зададим конкретные значения: пусть первоначально окольцовано 100 птиц, при повторной выборке проверено 50 птиц, из которых 20 оказалось с кольцами. Тогда общее число птиц данной популяции, согласно нашей формуле, будет $N = 250$.

Распределение Пуассона $X \sim P(\lambda)$. Пространство описываемых событий содержит бесконечное множество элементов:

$$\Omega\{\omega_1, \omega_2, \ldots, \omega_n, \omega_{n+1}, \ldots\}.$$

Функция плотности распределения случайной величины принимает вид:

$$p_k = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ 0 < \lambda \in \boldsymbol{R},\ k = 0, 1, 2, \ldots$$

При этом возможен переход к простой рекуррентной формуле вычисления вероятностных характеристик:

$$p_k = p_{k-1}\frac{\lambda}{k},\ k = 1, 2, \ldots, \text{ при } p_0 = e^{-\lambda}.$$

Интерпретировать распределение Пуассона можно как оценку вероятности очень редких событий. В прошлом толчком для разработки теории этого распределения послужило предположение о возможной в течение года смерти кавалериста от удара копытом его же лошади. Сегодня эмпирически более важным представляется вероятность несчастного случая, скажем, на шоссе или появление помех при телефонных переговорах в некоторый заранее определенный период времени.

3.3. Непрерывные распределения

В основе этого класса распределений всегда лежит непрерывная случайная величина: $X: \Omega \to \boldsymbol{R}$.

Непрерывное равномерное распределение $X \sim U(a, b)$

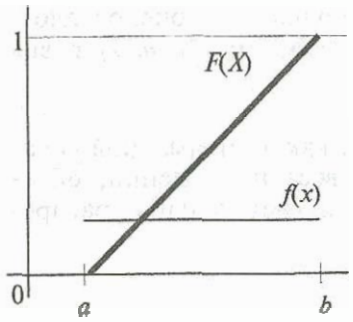


Рис. 3.8. График непрерывного равномерного распределения $U(a, b)$

Пусть имеются некоторые значения $a, b \in R$. Тогда случайную величину, имеющую плотность распределения f: $[a, b] \to R$ и $f(x) := \frac{1}{b-a}$, $-\infty < a \le x \le b < \infty$, **называют** *равномерно распределенной на интервале* $[a, b]$.

Функция распределения в этом случае, очевидно, будет точно такой же, как и для дискретной случайной величины (3.8):

$$F(x) = \frac{x-a}{b-a}, \quad -\infty < a \le x \le b < \infty.$$

Треугольное распределение. Пусть имеются $a, b, c \in R$. Тогда *треугольным распределением на интервале* $[a, c]$ называется случайная величина с плотностью распределения $f: [a, b] \in R$ и

$$f(x) := \begin{cases} \frac{ah}{a-b} + \frac{hx}{b-a}, & 0 \le a \le x \le b, \\ \frac{ch}{c-b} + \frac{hx}{b-c}, & b \le x \le c. \end{cases}$$

Константные значения a, b, c и h выбираются таким образом, чтобы получить треугольник площадью, равной единице. Например, $a = 0$, $b = 0{,}5$, $c = 1$ и $h = 2$, т.е.:

$$f(x) := \begin{cases} 0+4x, & 0 \le x \le 1/2, \\ 4-4x, & 1/2 \le x \le c. \end{cases}$$

Для данной плотности вероятностей соответственно можем записать функцию распределения:

$$F(x) = \begin{cases} 0, & x < 0 \\ 2x^2, & 0 \le x \le 1/2 \\ 4x - 2x^2 - 1, & 1/2 < x \le 1. \end{cases}$$

Распределение Парето. Пусть имеется $a, b \in R$, тогда случайная величина X называется *распределенной по Парето* при условии, что для нее существуют:

функция плотности вероятностей вида:

$$f := [a, b] \to R, \quad f(x) := \frac{a}{b}\left(\frac{b}{x}\right)^{a+1} \quad \text{при } 0 < a,\ 0 < b < x,$$

и функция распределения

$$F(x) = \begin{cases} 0 & \text{при } x \le b \\ 1 - \left(\frac{b}{x}\right)^a & \text{при } 0 < a,\ 0 < b < x. \end{cases}$$

То, что $f(x)$ может быть в данном случае получено из $F(x)$, и наоборот, следует непосредственно из связи первой производной $f(x)$ и $F(x)$, трансформирующих величину $u = \frac{1}{x}$ и соответственно описывающих плотность распределения и само распределение случайной величины X.

Типичным примером данного вида распределения является распределение доходов населения (рис. 3.9).

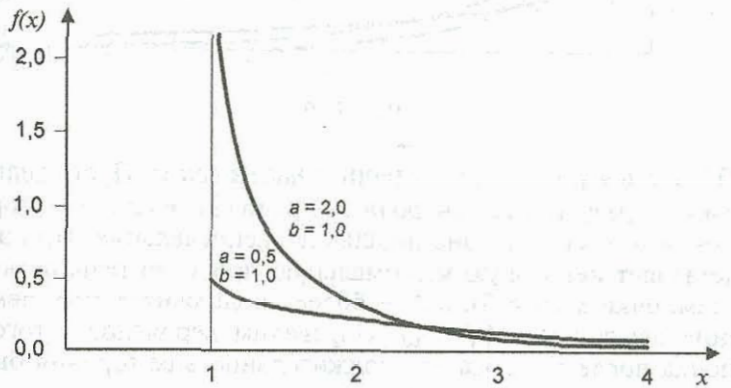


Рис. 3.9

Экспоненциальное распределение $X \sim P(1, \lambda)$. Пусть имеется $a, b \in R$. Тогда случайная величина X с плотностью распределения

$$f(x) := \begin{cases} 0 & \text{для } x \le 0 \\ \lambda e^{-\lambda x} & \text{для } x > 0 \end{cases}$$

и функцией распределения:

$$F(x) := \begin{cases} 0 & \text{для } x \le 0 \\ 1 - e^{-\lambda x} & \text{для } x > 0 \end{cases}$$

называется *экспоненциально распределенной с параметром* λ. Причем параметр λ выполняет условие:

$$0 < \lambda = r(x) := \frac{f(x)}{1 - F(x)} = \text{const}.$$

График плотностей экспоненциального распределения случайной величины с заданными параметрами $\lambda = 1$, $\lambda = 3$ и $\lambda = 5$ имеет вид, показанный на рис. 3.10.

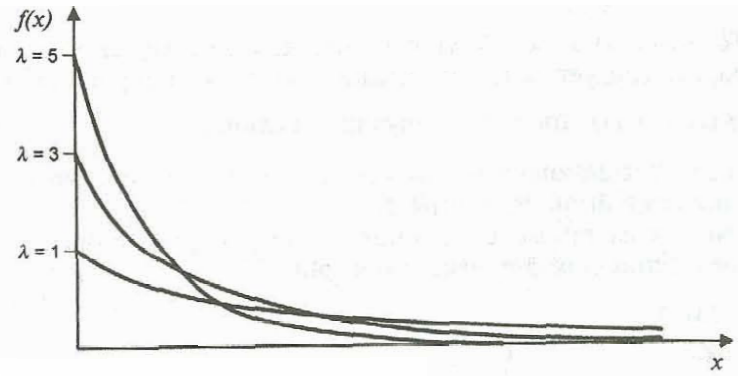


Рис. 3.10

Обратимся к примеру из теории надежности. Пусть величина X отражает результаты контроля за продолжительностью горения лампочки и имеет экспоненциальное распределение. При этом A представляет некоторую минимальную продолжительность горения лампочки t_0 ($t_0 > 0$), а B — более продолжительное, чем t_0 — горение лампочки t_1 ($t_1 > t_0$). Определим вероятность того, что лампочка после того, как продолжительность ее горения была t_0, будет гореть еще до периода t_1.

Уровень вероятности вычислим по известной формуле:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

результата A: $P(A) = e^{-\lambda t_0}$, для результата $A \cap B$: $P(A \cap B) = e^{-\lambda t_1}$, так как t_1 можно получить только зная t.

После подстановки функциональных значений $P(A)$ и $P(A \cap B)$ получим:

$$P(B|A) = \frac{e^{-\lambda t_1}}{e^{-\lambda t_0}} = e^{-\lambda t_1 + \lambda t_0} = e^{-\lambda(t_1 - t_0)}.$$

Подобный результат обычно характеризуется как *экспоненциальное распределение* без фиксации (запоминания) временного периода. В данном случае речь идет только о развитии событий

в различные временные периоды, но различия самих временных периодов как таковых не рассматриваются.

Распределение Эрланга $X \sim P(n, \lambda)$. В качестве обобщения экспоненциального распределения рассмотрим двупараметрическое унимодельное распределение Эрланга. Случайная величина называется *распределенной по Эрлангу* с параметрами λ и n, если она имеет плотность распределения:

$f: R \to R$,

$$f(x) := \begin{cases} 0 & \text{для } x \le 0 \\ x^{n-1}\lambda^n \dfrac{e^{-\lambda x}}{(n-1)!} & \text{для } x > 0 \end{cases}$$

с $n \in N$ и $\lambda > 0$. Значение $n = 1$ трансформирует распределение Эрланга вновь в экспоненциальное распределение. Параметрические значения $\lambda = 3$ и $n = 3$ или $n = 6$ определяют форму распределения Эрланга, как это показано на рис. 3.11.

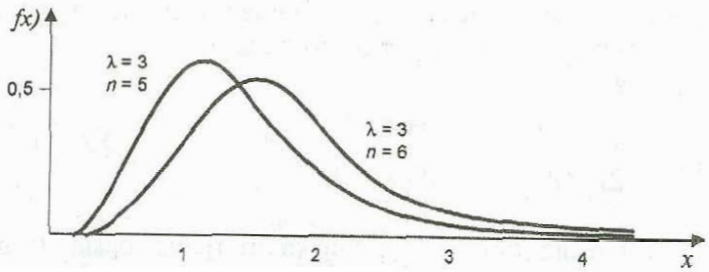


Рис. 3.11

Распределение Вейбулла. Еще одним обобщением экспоненциального распределения является двупараметрическое унимодельное распределение Вейбулла.

Случайная величина имеет распределение Вейбулла с параметрами λ и n, если ее плотность описывает функция вида:

$f: R \to R$,

$$f(x) := \begin{cases} 0 & \text{для } x \le 0 \\ nx^{n-1}\lambda e^{-\lambda x^n} & \text{для } x > 0 \end{cases}$$

при $n \in N$ и $\lambda > 0$.

Распределение Вейбулла, как и предыдущее, при $n = 1$ обращается в экспоненциальное, а при заданных параметрах $\lambda = 0{,}25$ и $n = 2$ принимает вид, показанный на рис. 3.12:

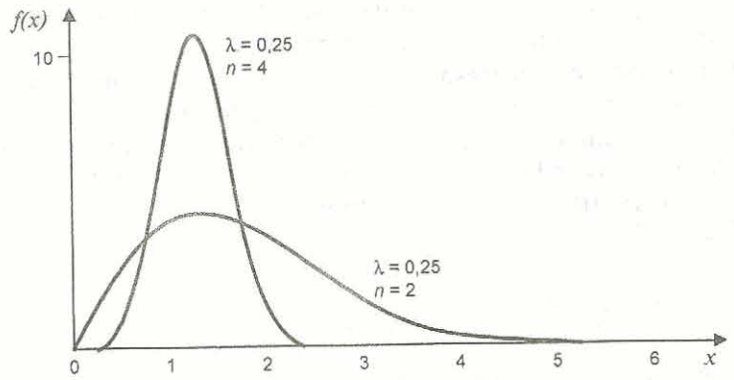


Рис. 3.12

Гиперэкспоненциальное распределение. Может рассматриваться как третье обобщение экспоненциального распределения. Данный вид распределения предполагает наличие случайной величины, имеющей плотность распределения:

$$f: R \to R,$$

$$f(x) := \begin{cases} 0 & \text{для } x \le 0 \\ \sum_{i=1}^{n} a_i \lambda_i e^{-\lambda_i x} & \text{для } x > 0 \end{cases} \quad \text{с} \quad \lambda_i > 0 \quad \text{и} \quad \sum_{i}^{n} a_i = 1,\ n \in N.$$

Как и прежде, при $n = 1$ происходит трансформация в обычное экспоненциальное распределение.

Двустороннее экспоненциальное распределение. Это четвертое обобщение экспоненциального распределения.

Двустороннее экспоненциальное распределение с параметром λ имеет случайная величина с плотностью распределения:

$$f: R \to R,$$

$$f(x) := \begin{cases} \frac{\lambda}{2} e^{+\lambda x} & \text{для } x \le 0 \\ \frac{\lambda}{2} e^{-\lambda x} & \text{для } x > 0 \end{cases} \quad \lambda > 0.$$

На рис. 3.13 представлен график плотности распределения случайной величины с двусторонним экспоненциальным распределением при $\lambda = 1$, $\lambda = 2$ и $\lambda = 3$:

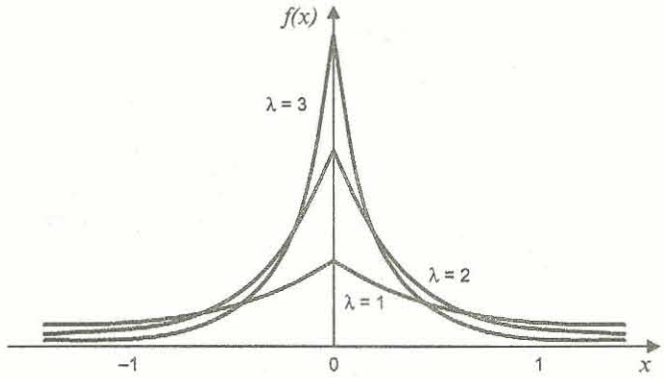


Рис. 3.13

Область практического приложения данного вида распределения иллюстрируют примеры статистического вероятностного оценивания:

- средней продолжительности «жизни» технических приборов;
- затрат времени на обслуживание клиентов в банке или магазине;
- вероятность умереть в младенческом возрасте.

Нормальное распределение (распределение Гаусса) — $X \sim N(\mu, \sigma^2)$. Случайную величину с плотностью распределения:

$$f: R \to R, \quad f(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

называют нормально распределенной с параметрами $\mu \in R$, $\sigma^2 > 0$.

Отметим следующие особенности графического изображения плотности вероятностей нормальной случайной величины:

1) значения f распределяются симметрично, влево и вправо от экстремальной параметрической оценки μ, т.е. $f'(\mu) = 0$;

2) функция f является строго монотонно возрастающей на интервале $[-\infty, \mu]$ и строго убывающей на интервале $[\mu, +\infty]$;

3) функция f имеет две точки перегиба — в $\mu \pm \sigma$.

В общем виде график нормального распределения напоминает колокол (рис. 3.14):

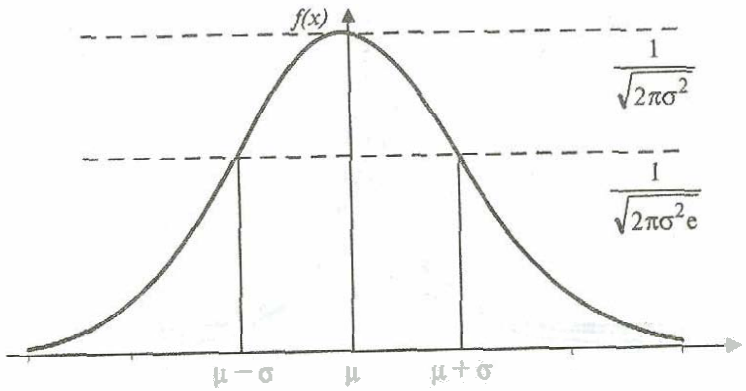


Рис. 3.14

характеристики плотности вероятностей нормального распределения табулированы. В связи с этим отметим, что график нормального распределения не имеет каких-либо закрытых областей, т.е. для него не существует некоторой определенной интегральной функции $F(x)$. Тем не менее представляется возможным произвести количественную оценку площади графика, находящегося под кривой функции $f(x)$. Именно это позволяет строить таблицы для значений $F(x)$ и быстро находить в них интересующие исследователя величины.

В теории вероятностей предлагается одна из нормального распределения — полунормальное распределение, описывающее плотность вероятностей:

$$f: [-\infty, \mu] \to R, \quad f(x) := \frac{2}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{}}, \quad x \le$$

при этом значения $f(x)$ по сравнению с нормальным распределением ограничиваются половиной ординатной оси.

Логнормальное распределение. Это распределение родственно нормальному, предполагает наличие случайной величины с плотностью вероятностей:

$$f: R \to R, \quad f(x) := \frac{1}{2\pi\sigma^2} \cdot \frac{1}{x} \cdot e^{-\frac{(\ln x - a)^2}{2\sigma^2}} \quad x > 0, \sigma^2 > 0, \ a \in R.$$

Логнормальное распределение имеет два параметра: a и σ^2. Приведенная функция плотности вероятностей в данном случае выполняет ту же роль, что и распределение Парето в распределении случайной величины.

Распределение Коши. Имеет параметры k и m, используется для случайной величины с плотностью вероятностей:

$$f: R \to R, \quad f(x) := \frac{k}{\pi(k^2 + (x-m)^2)}, \ 0 < k, \ -\infty < x < \infty.$$

Максимальное значение функции плотности вероятностей достигается при $x = m$ (рис. 3.15):

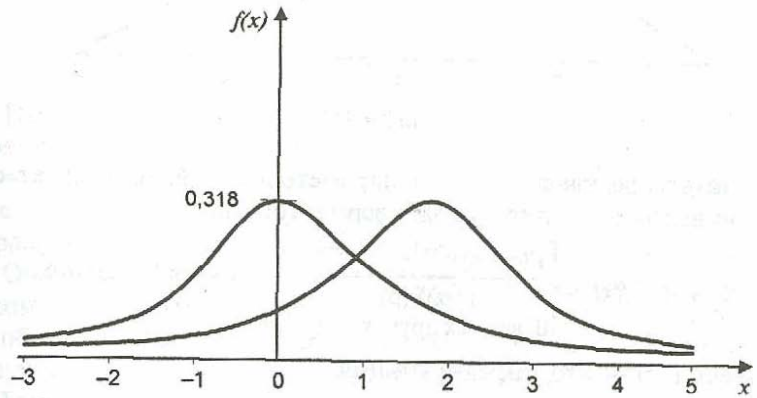


Рис. 3.15

***t*-распределение (распределение Стьюдента).** Случайная величина с плотностью вероятностей

$$f: R \to R, \quad f(t) := \frac{\Gamma\left(\frac{N+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)\sqrt{1+\left(\frac{t^2}{n}\right)^{n+1}}}, \ -\infty < t < \infty$$

имеет t-распределение с $\nu = n$ параметрическим числом степеней свободы.

При $n=1$ и $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ распределение Стьюдента (t-распределение) трансформируется в распределение Коши с $k = 1$.

График t-распределения с числом степеней свободы $n = 5$ и $n = 9$ имеет следующий вид (рис. 3.16):

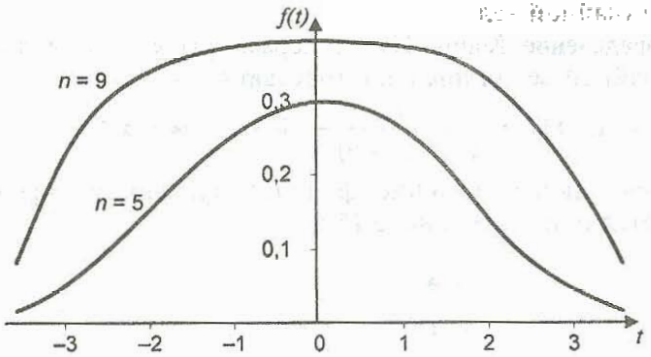


Рис. 3.16

Бета-распределение. Имеет параметры α и β, описывает случайную величину с плотностью вероятностей:

$$f: R \to R, \quad f(x) := \begin{cases} \dfrac{\Gamma(\alpha+\beta)x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}, & 0<x<1, \quad 0<\alpha,\beta, \\ 0 \text{ во всех других случаях.} \end{cases}$$

Здесь $\Gamma(z)$ — это гамма-функция:

$$\Gamma(z) := \int_0^\infty x^{z-1}e^{-x}dx, \; 0<z, \; 0\le x<$$

Бета-распределение представляется одним из наиболее важных, именно оно лежит в основе целого ряда других представленных ниже распределений.

***F*-распределение Фишера.** Имеет два параметрических числа степеней свободы ν_1 и ν_2, используется для случайной величины с плотностью вероятностей:

$$f: [0, \infty) \to R, \quad f(x) := \frac{\Gamma\left(\dfrac{\nu_1+\nu_2}{2}\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}x^{\nu_1/2-1}}{\Gamma\left(\dfrac{\nu_1}{2}\right)\Gamma\left(\dfrac{\nu_2}{2}\right)(\nu_1+\nu_2 x)^{(\nu_1+\nu_2)/2}},$$

где $0 < \nu_1, \nu_2$ — целые числа и $0 \le x < \infty$.

График распределения Фишера при $\nu_1 = \nu_2 = 10$; $\nu_1 = \nu_2 = 6$ и $\nu_1 = \nu_2 = 4$ имеет вид, показанный на рис. 3.17.

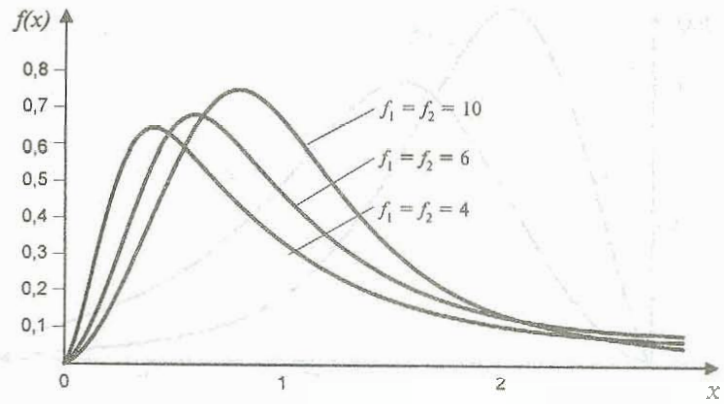


Рис. 3.17

При $\nu_1 = 1$, $\nu_2 = \nu$ и $x = t^2$ плотность вероятностей Фишера переходит в t-плотность Стьюдента, при $a =: \nu_1/2$, $b =: \nu_2/2$, $F := (\nu_1/\nu_2)[x/(1-x)]$ и целочисленных значениях ν_1, ν_2 F-распределение Фишера может рассматриваться как особый случай Бета-распределения.

Обратим внимание также на связь распределения Фишера с биномиальным распределением. В этом случае происходит как бы обмен ролями у величин и параметров: a и b — целочисленные значения величины, а x — это параметр p.

Гамма-распределение [$\Gamma(\alpha,\beta)$]. Имеет два параметра α и β, применяется для случайной величины с плотностью распределения:

$$f: (0, \infty) \to R, \quad f(x) := \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}, \; 0<\alpha,\beta; \; 0<x<\infty.$$

Хи-квадрат (χ^2)-распределение. В краткой записи $\chi^2(\gamma)$ имеет один параметр γ, который характеризуется так же, как число степеней свободы. Предполагается наличие случайной величины с плотностью вероятностей:

$$f: [0, \infty) \to R, \quad f(x) := \frac{x^{\gamma/2-1}e^{x/2}}{2^{\gamma/2}(\gamma/2-1)!}, \; 0<\gamma; \; 0<x<\infty.$$

Распределение χ^2 при $\gamma = 6$ и $\gamma = 10$ показано на рис. 3.18.

Сравнивая χ^2-распределение с гамма-плотностью распределения вероятностей, можем записать:

$\chi^2(\gamma) = \Gamma(\gamma/2,2)$, $\chi^2(2\gamma) = \Gamma(\alpha,2)$.

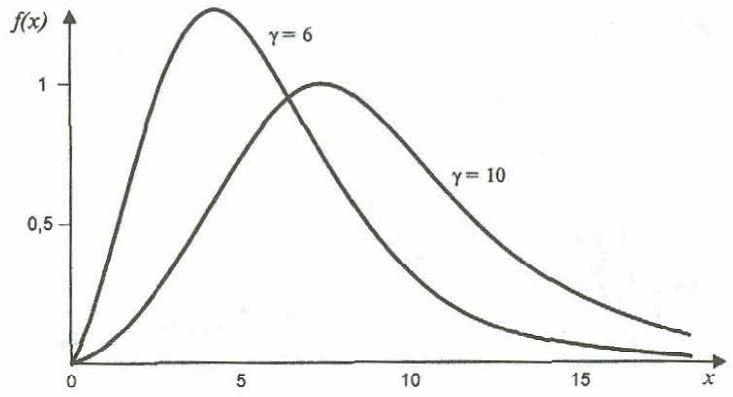


Рис. 3.18

Распределение χ^2 может быть получено также из распределения F-Фишера и соответственно бета-распределения. В этом случае устанавливается, что $x := \chi^2 = x/\gamma$, $\gamma := f_1$ и $f_2 \to \infty$.

Распределение Вейбулла-гамма. Имеет три параметра: b, d, k и используется для случайной величины с плотностью вероятностей:

$$f(x) := \begin{cases} \dfrac{bkdx^{b-1}}{(x^b + d)^{k+1}}, & x \geq 0, \quad 0 < b,d,k, \\ 0, & x < 0. \end{cases}$$

Для распределения Вейбулла-гамма можно выделить два особых случая. Первый случай — с так называемой плотностью Бёрра ($d = 1$):

$$f(x) := \begin{cases} \dfrac{bkx^{b-1}}{(x^b)^{k+1}}, & x \geq 0, \quad 0 < b,k, \\ 0, & x < 0. \end{cases}$$

Второй случай предполагает Парето-плотность вероятностей, $d = b = 1$ и $y := 1 + x$:

$$f: \mathbf{R} \to \mathbf{R}, \quad f(y) := \begin{cases} \dfrac{k}{y^{k+1}}, & y \geq 1, \quad 0 < k, \\ 0, & y < 1. \end{cases}$$

Общая схема взаимосвязей непрерывных статистических распределений представлена на рис. 3.19.

*Рис. 3.19. **Взаимосвязи непрерывных статистических распределений***

3.4. Математическое ожидание и дисперсия

В предыдущих параграфах мы видели, что в рамках дескриптивной статистики эмпирические распределения представляются с помощью параметрических величин, определяющих их положение и степень рассеяния в координатной системе. Похожим образом распределения случайных величин оказывается возможным грубо описывать некоторыми числами, задающими положение и степень рассеяния самих случайных величин. В этом случае роль показателей уровня вероятности берут на себя характеристики относительной частоты появления того или иного события.

Математическое ожидание. а) Пусть X — дискретная случайная величина с плотностью вероятностей $p_k = P(X = x_k)$ $(k \in K)$,

где K — индекс некоторого множества, включающего элементы пространства событий Ω, $\Omega = \{\omega_k \mid k \in K\}$, тогда число

$$E(x) := \sum_{k \in K} p_k x_k$$

называется математическим ожиданием X.

б) Пусть X — непрерывная случайная величина, имеющая функцию плотности вероятностей f: $D \to R$, тогда число

$$E(x) := \int_D x f(x) dx$$

есть математическое ожидание X.

Пример 3.12. Рассмотрим случай с подбрасыванием игрального шестигранного кубика:

$\Omega = \{k \in K \mid K = \{1, 2, 3, 4, 5, 6\}\} = \{1, 2, 3, 4, 5, 6\}$,

$p_k = \frac{1}{6} (k = 1, 2, 3, 4, 5, 6)$,

математическое ожидание случайной величины X будет

$$E(x) = \sum_{k=1}^{6} p_k x_k = \frac{1}{6} \sum_{k=1}^{6} k = 3{,}5.$$

Пример 3.13. Имеется случайная величина X, X: $\Omega \to R$, при этом $\Omega = \{k \in K \mid K = \{1, 2, ..., N\}\} = \{1, 2, 3, ..., N\}$ и $p_k = ak$ $(k = 1, 2, ..., N)$.

Определим математическое ожидание для X, приняв во внимание свойства статистической оценки уровня вероятности:

$$1 = \sum_{k=1}^{N} p_k = a \sum_{k=1}^{N} k = \frac{aN(N+1)}{2} \Rightarrow a = \frac{2}{N(N+1)},$$

следовательно,

$$E(X) = \sum_{k=1}^{N} p_k x_k = a \sum_{k=1}^{N} k^2 = \frac{2N+1}{3}.$$

Пример 3.14. Пусть имеется X — случайная величина с функцией плотности вероятностей:

$f(x) = ax^b$, $0 < x < b$, т. е. $D = (0, b)$.

Из условия, что $\int_D f(x) dx = 1$, получаем

$$1 = \int_0^b f(x) dx = \frac{ab^{b+1}}{b+1} \Rightarrow a = \frac{b+1}{b^{b+1}},$$

$$E(X) = \int_D x f(x) dx = \int_0^b \left(\frac{b+1}{b^{b+1}}\right) x^{b+1} dx = \frac{b(b+1)}{b+2}.$$

Пример 3.15. Случайная величина X имеет экспоненциальное распределение с параметром λ, ее математическое ожидание будет:

$$E(x) = \int_D x f(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Подобный вывод базируется на справедливости следующих утверждений:

$e^{-\lambda x} dx = du$ и $v = x$, а также $-\frac{1}{\lambda} e^{-\lambda x} = u$ и $dv = dx$.

В данном случае следует обратить внимание, что

$$\left(-xe^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x}\right) \to 0 \text{ при } x \to \infty.$$

Замечание. О трансформации случайной величины.

Пусть g: $R \to R$ — интегрируемая функция, тогда

$E[g(X)] = \sum_k g(x_k) p_k$, если X — дискретная величина, и

$E[g(X)] = \int_D g(x) f(x) dx$, если X — непрерывная величина.

В частности, это можно увидеть на примере простой линейной зависимости двух величин X и Y: $Y = a + bX$ $(a, b \in R)$: $E(Y) = a + b[E(X)]$.

Неравенство Маркова. Пусть X — случайная величина, принимающая только неотрицательные значения, тогда можем записать следующее неравенство Маркова:

$$P[X \geq c] \leq \frac{E(x)}{c}, \; c > 0.$$

□ *Доказательство*: Обратимся к случаю с дискретной случайной величиной:

$$E(X) = \sum_k p_k x_k = \sum_{k: x_k < c} p_k x_k + \sum_{k: x_k \geq c} p_k x_k \geq \sum_{k: x_k \geq c} p_k x_k = c \sum_{k: x_k \geq c} p_k = cP(X \geq c). \blacksquare$$

Если X непрерывна, доказательство существования неравенства Маркова строится аналогичным образом.

Дисперсия. а) Если X — случайная величина с плотностью вероятностей $p_k = P(X = x_k)$, то ее дисперсией называют число (var X):

$$\text{var}(X) := \sum_{k \in K} p_k [x_k - E(X)]^2.$$

б) Если случайная величина X непрерывна и имеет плотность вероятностей $f(x)$, $x \in D$, то ее дисперсией будет:

$$\operatorname{var}(X) := \int_D [x - E(x)]^2 f(x)dx.$$

Дисперсия может быть представлена при помощи математического ожидания случайной величины как разность:

$$\operatorname{var}(X) = E(X^2) - [E(X)]^2.$$

Это равенство часто используется как удобная форма расчета дисперсии.

□ *Доказательство*: Для дискретной случайной величины X действует:

$$\operatorname{var}(X) = \sum_{k\in K} p_k [x_k - E(X)]^2 = \sum_{k\in K} p_k \left[x_k^2 - 2x_k E(X) + [E(X)]^2\right] =$$

$$= \sum_{k\in K} x_k^2 p_k - 2E(X) \sum_{k\in K} x_k p_k + [E(X)]^2 \sum_{k\in K} p_k = E(X)^2 - [E(X)]^2.$$

Если случайная величина X непрерывна, то соответствующим образом, как доказано выше, получим:

$$\operatorname{var}(X) = \int_D [x - E(X)]^2 f(x)dx = \int_D \left[x - 2xE(X) + [E(X)]^2\right] f(x)dx =$$

$$= \int_D x^2 f(x)dx - 2E(X) \int_D x f(x)\ldots dx + [E(X)]^2 \int_D f(x)dx =$$

$$= E(X^2) - [E(X)]^2. \blacksquare$$

Пример 3.16. (Продолжение примера 3.12). Вернемся к примеру с подбрасыванием идеального игрального шестигранного кубика:

$$\operatorname{var}(X) = \sum_{k=1}^{6} p_k [x_k - E(X)]^2 = \sum_{k=1}^{6} \frac{1}{6}\left(k - \frac{7}{2}\right)^2 = \frac{35}{12}.$$

Пример 3.17. (Продолжение примера 3.13)

$$\operatorname{var}(X) = \sum_{k=1}^{N} p_k [x_k - E(X)]^2 = \sum_{k=1}^{N} p_k x_k^2 - E(X)^2 = \frac{(K^2 + K - 2)}{18}.$$

Замечание. Линейная трансформация случайной величины. Если X — случайная величина и существует зависимость вида: $Y = a + bX$, $a, b \in R$, т.е. X линейно трансформируема, то действует равенство:

$$\operatorname{var}(Y) = b^2 \operatorname{var}(X).$$

□ *Доказательство*: Пусть случайная величина X дискретна, тогда

$$\operatorname{var}(Y) = \sum_{k\in K} p_k [y_k - E(Y)]^2 = \sum_{k\in K} p_k [a + bx_k - a - bE(X)]^2 =$$

$$= b^2 \sum_{k\in K} p_k [x_k - E(X)]^2 = b^2 \operatorname{var}(X).$$

В случае, если X — непрерывная случайная величина, то и Y непрерывна и дисперсию получим в результате следующих формальных преобразований:

$$\operatorname{var}(Y) = \int_D [y - E(Y)]^2 f(y)dy = \int_D [a + bx - a - bE(X)]^2 f(x)dx =$$

$$= \int_D [bx - bE(X)]^2 f(x)dx = \int_D b^2 [x - E(X)]^2 f(x)dx = b^2 \operatorname{var}(X). \blacksquare$$

Рассмотрим примеры, позволяющие перейти к обобщению понятий математического ожидания и дисперсии.

Пример 3.18. Определим математическое ожидание и дисперсию для случайной величины, имеющей распределение Бернулли:

$$E(X) = 1p + 0(1 - p) = p,$$

$$(X) = (1 - p)^2 p + (0 - p)^2 (1 - p) = p(1 - p) = pq \ (q := 1 - p).$$

Пример 3.19. Для случайной величины, подчиняющейся биномиальному закону распределения, математическое ожидание и дисперсия будут:

$$E(X) = np \text{ и } \operatorname{var}(X) = np(1 - p) = npq.$$

Формулы для определения математического ожидания и дисперсии при этом получают непосредственно преобразованием биномиальных коэффициентов:

$$E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^k \text{ и } \operatorname{var}(X) = \sum_{k=0}^{n} (k - np)^2 \binom{n}{k} p^k (1-p)^k,$$

откуда для математического ожидания имеем:

$$E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^k = \sum_{k=0}^{n} \frac{n!k}{(n-k)!k!} p^k (1-p)^k =$$

$$= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{(n-1)-(k-1)} =$$

$$= np \sum_{j=0}^{m} \binom{m}{j} p^j (1-p)^{m-j} = np, \ m := n-1, \ j := k-1.$$

Пример 3.20. При условии, что случайная величина X имеет распределение Пуассона, математическое ожидание и дисперсия находятся весьма просто:

$$E(X)=\sum_{k=0}^{\infty} kp_k=\sum_{k=0}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!}=\lambda\sum_{k=0}^{\infty}=\lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=\lambda.$$

Дисперсию найдем, используя связи моментов:

$$E[X(X-1)]=\sum_{k=0}^{\infty}k(k-1)p_k=\sum_{k=0}^{\infty}k(k-1)\frac{\lambda^k e^{-\lambda}}{k!}=\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}e^{-\lambda}}{(k-2)!}=$$

$$=\lambda^2 e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=\lambda^2,$$

откуда

$$=E[X(X-1)]-E(X)[E(X)-1]=\lambda^2-\lambda(\lambda-1)=\lambda.$$

Математическое ожидание и дисперсию для плотности вероятностей Пуассона определяют аналогично, как и для самой случайной величины, имеющей распределение Пуассона.

Пример 3.21. Покажем пример плотности вероятностей, для которой не существует математического

$$p_j=P[X=j]=\frac{1}{j(j+1)},\quad j=1,2...$$

Приведенное выражение действительно задает плотность вероятностей, так как

$$\sum_{j=1}^{n}\frac{1}{j(j+1)}=\frac{n}{n+1}\ \text{и}\ \lim_{n\to\infty}\frac{n}{n+1}=1.$$

При определении математического ожидания случайной величины X имеем:

$$E(X)=\sum_{j=1}^{n}p_j j=\sum_{j=1}^{n}\frac{j}{j(j+1)}=\sum_{j=1}^{n}\frac{1}{j+1}.$$

Введем некоторые определенные значения:

$$\sum_{k=2^j+1}^{2^{j+1}}\frac{1}{k}\ge\frac{1}{2},\quad j=0,1,2,...$$

Эти значения формально при условии $\left(2^{j+1}-2^j=2^j\right)$ представляются минимальными $\frac{2^{j+1}-2^j}{2^{j+1}}=\frac{1}{2}$, $j=0,1,2...$, или больше минимальных. Но тогда

$$\lim_{n\to\infty}\sum_{j=0}^{n}\sum_{k=2^j+1}^{2^{j+1}}\frac{1}{k}\ge\lim_{n\to\infty}\sum_{j=0}^{n}\frac{1}{2}=\lim_{n\to\infty}\frac{n}{2},$$

т.е. $E(X)$ бесконечно велико.

Пример 3.22. Приведем без доказательства оценки математического ожидания и дисперсии для нормального распределения случайной величины:

$E(X)=\mu$, $\text{var}(X)=\sigma^2$.

Выводы для μ и σ^2 получают разнообразными статистическими способами, широко представленными в теоретической литературе.

Утверждение. Случайная величина X, симметрично распределенная вокруг некоторой точки S, имеет математическое ожидание, равное S.

□ ***Доказательство***: Предположим, что X — дискретная величина, но тогда для каждого из ее значений $S+x_k=:y_k$ плотность вероятностей будет $p_k=P[X=y_k]$, то же для значений $S-x_k=y_k$ плотность вероятностей $p_k=p_l=P[X=y_l]$ — тем самым доказательство исчерпывается.

Если X непрерывна, то с учетом заданного условия можно записать равенство функций плотности вероятностей: $f(S+x)=f(S-x)$, что также является достаточным доказательством выдвинутого утверждения. ■

Коэффициент вариации. Пусть X — некоторая случайная величина, дискретная или непрерывная, имеющая математическое ожидан $E(X)$ и дисперсию $\text{var}(X)$, отличные от нуля, тогда ее коэффициентом вариации $V(X)$ выступает отношение:

$$V(X):=\frac{\pm\sqrt{\text{var}(X)}}{E(X)}.$$

Например, для биномиального распределения коэффициент вариации будет:

$$V(X):=\sqrt{\frac{1-p}{np}},$$

для геометрического распределения:

$$V(X):=\frac{1}{\sqrt{n-p}},$$

для распределения Пуассона:

$$V(X):=\frac{1}{\sqrt{\lambda}}.$$

Коэффициент вариации для стандартизованной случайной величины. Пусть имеется нормально распределенная величина с математическим ожиданием $E(X) =: \mu$, дисперсией и средним квадратическим отклонением, соответственно равными: $\mathrm{var}(X) =: \sigma^2$, $\sqrt{\mathrm{var}(X)} = \sigma$. Стандартизируя значения X по формуле $Y := \dfrac{X-\mu}{\sigma}$, имеем:

$$E(Y) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}(E(X)-\mu) = \frac{1}{\sigma}(\mu-\mu) = 0,$$

$$\mathrm{var}(Y) = \mathrm{var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{var}(X-\mu) = \frac{1}{\sigma^2}\mathrm{var}(X) = \frac{1}{\sigma^2}\cdot\sigma^2 = 1.$$

Далее систематизируем наиболее распространенные виды статистических распределений и покажем для каждого из них алгоритмы вычислений математического ожидания и дисперсии (табл. 3.2).

Таблица 3.2. **Виды и параметры статистических распределений**

Статистическое распределение	*Математическое ожидание*	*Дисперсия*
1	*2*	*3*
Биномиальное $B(n; p)$	np	$np\,(1-p)$
Геометрическое $NB(1, p)$	$\dfrac{1-p}{p}$	$\dfrac{1-p}{p^2}$
Паскаля $NB(r, p)$	$\dfrac{r(1-p)}{p}$	$\dfrac{r(1-p)}{p^2}$
Гипергеометрическое $H(M, N, n)$	$\dfrac{Mn}{N}$	$\dfrac{M(N-M)(N-n)n}{N^2(N-1)}$
Пуассона $P(\lambda)$	λ	λ
Непрерывное равномерное $U(a, b)$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Парето	$\dfrac{ab}{a-1}$	$\dfrac{ab^2}{(a-1)^2(a-2)^2}$, $a \neq 1, a \neq 2$

Продолжение табл.3.2

1	*2*	*3*
Экспоненциальное $P(1, \lambda)$	$1/\lambda$	$1/\lambda^2$
Эрланга $P(n, \lambda)$	n/λ	n/λ^2
Вейбулла	$\dfrac{\Gamma\left(\frac{1+n}{n}\right)}{\sqrt[n]{\lambda}}$	$\dfrac{\Gamma\left(\frac{2+n}{n}\right)-\left(\Gamma\left(\frac{1+n}{n}\right)\right)^2}{\sqrt[n]{\lambda^2}}$
Двустороннее экспоненциальное	0	$\dfrac{1}{\lambda^2}$
Нормальное $N(\mu,\sigma)$	μ	σ^2
Логнормальное	$e^a + \dfrac{\sigma^2}{2}$	$e^{2a+\sigma^2}(e^{e^2}-1)$
Коши	нет	нет
t-Стьюдента	0	$\dfrac{n}{n-2}$, $n>2$
Бета	$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
F-Фишера	$\dfrac{f_1}{f_1-2}$	$\dfrac{2f_1^2(f_2+f_1-2)}{f_2(f_1-2)^2(f_1-4)}$
Гамма $G(\alpha, \beta)$	$\alpha\,\beta$	$\alpha\,\beta^2$
Кси-квадрат $G(\alpha,2)=\chi^2(g)$	g	$2g$

3.5. Двумерные и многомерные случайные величины

Обратимся к проблемам статистического изучения дву- и многомерных случайных величин. Наиболее важные и сложные вопросы при этом — оценка силы и определение мерного масштаба взаимосвязей таких величин.

Дискретные случайные величины. Пусть имеются X и Y — две дискретные случайные величины с некоторой ограниченной общей областью их оценивания, тогда показатели вероятности

совместных реализаций определенных значений этих величин могут быть обобщены в так называемой таблице контингенций (табл.3.3).

Таблица 3.3. **Контингенции для двумерной дискретной случайной величины** (X, Y)

X \ Y	y_1	y_2	...	y_{m_2}	
x_1	p_{11}	p_{12}	...	p_{1m_2}	$p_{1\cdot}$
x_2	p_{21}	p_{22}	...	p_{2m_2}	$p_{2\cdot}$
. . .					
x_{m_1}	$p_{m_1 1}$	$p_{m_1 2}$	...	$p_{m_1 m_2}$	$p_{m_1\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot m_2}$	$p_{\cdot\cdot}$

В клетках таблицы, окаймленных жирными линиями, представлены величины, которые указывают на совместную вероятность появления конкретных значений двумерной случайной величины (X, Y), а в клетках, расположенных за жирными линиями, т.е. в клетках последнего столбца и последней строки, — показатели условной вероятности, существующие раздельно для значений каждой из случайных величин X и Y. Напомним о статистическом анализе условных вероятностей одномерных случайных величин X и Y, о которых говорилось выше в пп. 3.2.1—3.2.4; для нового понятия совместной вероятности введем обозначение p_{ij}, при этом $0 \le p_{ij} \le 1$, $\Sigma\Sigma p_{ij} = 1$ $(i = 1, 2, ..., m; j = 1, 2, ..., n)$ с целью упрощения записи примем, что $m_1 =: m$ и $m_2 =: n$.

Пример 3.23. Условное и совместное распределение. Покажем, что различные совместные распределения двух случайных величин X и Y могут приводить к образованию идентичных условных распределений.

Первое совместное распределение:

$P[X = -1, Y = 0] = P[X = 0, Y = 1] = P[X = 0, Y = -1] = 1/4.$

Условные распределения при этом:

$P[X = -1] = 1/4$
$P[X = 0] = 1/2$
$P[X = 1] = 1/4$

и согласно условию симметричности

$P[Y = -1] = 1/4$
$P[Y = 0] = 1/2$
$P[Y = 1] = 1/4.$

Второе совместное распределение:

$P[X = -1, Y = 1] = P[X = -1, Y = -1] = P[X = 1, Y = 1] = P[X = 1, Y = -1] = 1/8$

и

$P[X = 0, Y = 0] = 1/2.$

Условные распределения здесь такие же, как и в предыдущем случае:

$P[X = -1] = 1/4$
$P[X = 0] = 1/2$
$P[X = 1] = 1/4$

и соответственно условию симметричности

$P[Y = -1] = 1/4$
$P[Y = 0] = 1/2$
$P[Y = 1] = 1/4.$

Для наглядности проиллюстрируем пример рисунками, показывающими распределения совместных вероятностей (рис. 3.20 и 3.21):

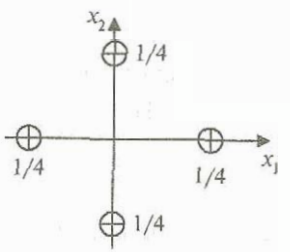


Рис. 3.20

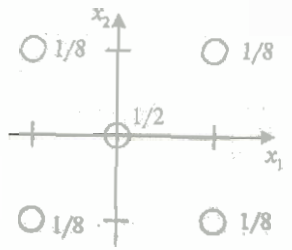


Рис. 3.21

Таким образом, можно утверждать, что для совместной вероятности существуют и могут быть найдены соответствующие условные вероятности. Обратное утверждение, что некоторые условные вероятности задают совместную определенную вероятность, как правило, не действует. Исключением выступает

только случай с независимыми величинами, допускающими образование совместных вероятностей из условных.

Анализ многомерных случайных величин предполагает введение параметрических оценок по аналогии с математическим ожиданием и дисперсией, используемых для обычных одномерных величин. Исходными и весьма важными в числе таких оценок являются показатели ковариации.

Ковариацией случайных величин X и Y $[\mathrm{cov}(X,Y)]$ является некоторое число:

$$\mathrm{cov}(X,Y) := \sum_{j=1}\sum_{i=1}[x_i - E(X)][y_i - E(Y)]p_{ij}.$$

Для исчисления характеристик ковариации может применяться формула, значительно упрощающая расчеты:

$$\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y).$$

□ *Доказательство*:

$$\mathrm{cov}(X,Y) = \sum_{j=1}^{n}\sum_{i=1}^{m}[x_i - E(X)][y_j - E(Y)]p_{ij} =$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{m}[x_i y_j - x_i E(Y) - y_j E(X) + E(X)E(Y)]p_{ij} =$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{m}x_i y_j p_{ij} - E(Y)\sum_{i=1}^{m}x_i p_{i\cdot} - E(X)\sum_{j=1}^{n}y_j p_{\cdot j} + E(X)E(Y) =$$

$$= E(XY) - E(X)E(Y).$$

Полученное новое выражение $E(XY)$ можно рассматривать как простое математическое ожидание (см. §3.4):

$$E(XY) := \sum_{j=1}^{n}\sum_{i=1}^{m}x_i \cdot y_{\cdot j}\, p_{ij}. \blacksquare$$

Определение независимости случайных величин. Две случайные величины X и Y называются независимыми, если для них имеет силу:

$$p_{ij} = P(X = x_i,\ Y = y_j) = P(X = x_i)P(Y = y_j) = p_{i\cdot}.p_{\cdot j}.$$

Ковариация двух независимых случайных величин равна нулю: $\mathrm{cov}(X,Y) = 0$.

□ *Доказательство*: Само доказательство в данном случае вытекает непосредственно из определения независимости

случайных величин. Для независимых величин X и Y действует, как показано выше, $p_{ij} = p_{i\cdot}.p_{\cdot j}$, откуда следует

$$E(XY) = \left[\sum_{i=1}^{m}x_i.p_{i\cdot}\right]\left[\sum_{j=1}^{n}y_{\cdot j}\,p_{\cdot j}\right] = E(X)E(Y). \blacksquare$$

Замечание. $\mathrm{cov}(X,Y) = 0 \Rightarrow$ независимость X и Y. Обратный вывод о том, что если $\mathrm{cov}(X,Y) = 0$, то X и Y — независимые величины, неправомерен. В этом легко убедиться на последующих примерах, представляющих ковариацию и связи независимых величин.

Пример 3.24. Имеются данные распределения двух случайных величин X и Y:

X \ Y	−2	−1	1	2	
1	0	1/4	1/4	0	1/2
4	1/4	0	0	1/4	1/2
	1/4	1/4	1/4	1/4	

При этом $E(X) = 5/2$, $\mathrm{var}(X) = 9/4$,

$E(Y) = 0$, $\mathrm{var}(Y) = 5/2$.

$E(XY) = 0$,

т.е. $\mathrm{cov}(X,Y) = 0$, хотя X и Y, как видно из данных таблицы, — не независимые величины.

Пример 3.25. Пусть имеются две независимые и идентично распределенные случайные величины U и V, которые, скажем, представляют результаты подбрасываний двух игральных кубиков. Тогда можем записать:

$X := U + V$ — сумма результатов подбрасывания кубиков,

$Y := U - V$ — разность результатов подбрасывания кубиков.

Для показателей ковариации X и Y имеем следующий набор равенств:

$$E(X)=E(U)+E(V);\quad E(Y)=E(U)-E(V)=0;$$
$$E(X,Y)=E[(U+V)(U-V)]=E(U^2-V^2)=E(U^2)-E(V^2)=0;$$
$$\operatorname{cov}(X,Y)=E(XY)-E(X)E(Y)=0-E(X)\cdot 0=0.$$

По условию U и V — две независимые величины. Величины X и Y конструктивно построены таким образом, что находятся в тесной связи друг с другом, но в то же время их ковариация равна нулю.

Понятие коэффициента корреляции ρ. Коэффициентом корреляции называют оценку ковариации двух стандартизованных случайных величин (X^*, Y^*):

$$\rho:=\operatorname{cov}(X^*,Y^*)=\sum_{j=1}^{n}\sum_{i=1}^{m}\left[x_i^*-E\left(X^*\right)\right]\left[y_j^*-E\left(Y^*\right)\right]p_{ij}.$$

Коэффициент корреляции отражает линейные связи, для него имеют силу:

а) $-1\le r\le 1$,

б) $r=\pm 1\Leftrightarrow Y=a+bX\ (a,b\in R)$.

□ *Доказательство:*

Для свойства а)

Предположим, X и Y — две любые случайные величины, имеющие совместную плотность вероятностей и определенные величины математических ожиданий $E(X), E(Y)$ и дисперсий $\operatorname{var}(X), \operatorname{var}(Y)$. Тогда для дисперсии суммы $Z := X + Y$ действует:

$$0\le \operatorname{var}(Z)=E[Z-E(Z)]^2=E[X+Y-E(X+Y)]^2=$$
$$=E\{[X-E(X)]+[Y-E(Y)]\}^2=E\{[X-E(X)]^2+[Y-E(Y)]^2+$$
$$+2[X-E(X)][Y-E(Y)]\}=\operatorname{var}(X)+\operatorname{var}(Y)+2\operatorname{cov}(X,Y). \qquad (3.1)$$

Чтобы показать справедливость *а)*, перейдем к рассмотрению стандартизованных величин X^* и Y^*, запишем для них:

$$0\le \operatorname{var}(X^*+Y^*)=1+1+2\rho=2(1+\rho), \qquad (3.2)$$

для дисперсии разности $Z := X - Y$ или $Z := Y - X$ соответственно будет:

$$0\le \operatorname{var}(X^*-Y^*)=1+1-2\rho=2(1-\rho). \qquad (3.3)$$

Из (3.2) и (3.3) непосредственно следует правомерность установленных нижней и верхней границ колебаний коэффициента корреляции.

Для свойства б):

Пусть имеются две любые случайные величины X и Y, они линейно взаимосвязаны, т.е.

$Y = a + bx$ (где a, b — параметры), тогда

$E(Y) = a + bE(X)$ и $\operatorname{var}(Y) = b^2 \operatorname{var}(X)$.

Для стандартизованных величин X^* и Y^* соответственно получим:

$$X^*=\frac{X-E(X)}{\sigma(X)},\quad \sigma(X)=\sqrt{\operatorname{var}(X)}. \qquad (3.4)$$

$$Y^*=\frac{Y-E(Y)}{\sigma(Y)},\quad \sigma(Y)=\sqrt{\operatorname{var}(Y)}. \qquad (3.5)$$

Следовательно, коэффициент корреляции ρ для случайных величин X и Y при условии их полной линейной зависимости будет равен единице:

$$\rho=\operatorname{cov}(X^*,Y^*)=E\left\{\left[X^*-E\left(X^*\right)\right]\left[Y^*-E\left(Y^*\right)\right]\right\}=$$
$$=E(X^*Y^*)=E\left[\frac{[X-E(X)]}{\sigma(X)}\frac{[Y-E(Y)]}{\sigma(Y)}\right]=$$
$$=E\left[\frac{X-E(X)}{\sigma(X)}\cdot\frac{a+bX-a-bE(X)}{b\sigma(X)}\right]=$$
$$=E\left[\frac{X-E(X)}{\sigma(X)}\cdot\frac{X-E(X)}{\sigma(X)}\right]=\frac{1}{\sigma^2(X)}E\left[(X-E(X))^2\right]=\frac{\sigma^2(X)}{\sigma^2(X)}=1.$$

Аналогично доказывается, что может быть $\rho=-1$.

Доказательство для свойства *б)* строится от обратного: если коэффициент корреляции двух случайных величин равен единице ($\rho=1$), то эти две величины находятся в полной линейной связи. Начнем с аксиоматичного утверждения: $\operatorname{var}(X^*-Y^*) = 0$. Это означает, что разность двух стандартизованных случайных величин есть некоторая постоянная величина: $X^* - Y^* = k$. При подробной записи X^* и Y^* имеем:

$$\frac{X-E(X)}{\sigma(X)}-\frac{Y-E(Y)}{\sigma(Y)}=k$$

или $X=\left[\frac{-\sigma(X)}{\sigma(Y)}E(Y)+k\sigma(X)+E(X)\right]+\frac{\sigma(X)}{\sigma(Y)}Y=a+bY.$

Соответствующим образом существование полной линейной зависимости X и Y доказывается при значениях $\rho=-1$ и $\mathrm{var}(X+Y)=0$. ■

Приведенные доказательства сохраняют силу вне зависимости от единиц измерения случайных величин.

Пример 3.26. Пусть известна вероятностная характеристика двумерной случайной величины (X, Y): $P[X=0,Y=2]=1/2$, $P[X=1,Y=1]=1/3$, $P[X=2,Y=0]=1/6$, очевидно при этом: $\rho=-1$ и $Y=2-X$. **Отметим, что при оценке коэффициентов корреляции** ρ используются все промежуточные значения величины $\mathrm{cov}(X^*,Y^*)$, а именно: $E(X), E(Y), E(X^2), E(Y^2), \sigma(X), \sigma(Y), E(XY)$.

Понятие определенности. Мерой определенности представляется величина $\rho^2=:R^2$, очевидно $0\le R^2\le 1$. Для независимых случайных величин $\rho=0$ и $R^2=0$. При этом ρ выступает как эмпирический коэффициент корреляции, его статистической оценкой будет величина r:

$$r(X,Y):=\sum_{i=1}^{n}\frac{(x_i-\bar{x})(y_i-\bar{y})}{s(X)s(Y)},$$

где $\{x_i\}$ и $\{y_i\}$ — два множества количественных данных наблюдений, $i=1, 2,..., n$;

$s(X)$ и $s(Y)$ — оценки эмпирических величин стандартных квадратических отклонений:

$$s^2(X):=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2;\ s^2(Y):=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

При определении коэффициента корреляций может использоваться и другая формула, которая дает полное представление обо всех вычислительных процедурах для значений двух элементарных признаков X и Y:

$$r(X,Y)=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{\left[n\sum x_i^2-\left(\sum x_i\right)^2\right]\left[n\sum y_i^2-\left(\sum y_i\right)^2\right]}}.$$

Говоря о коэффициентах корреляции (измерителях стохастической определенности), необходимо сделать некоторые замечания. Дело в том, что практическая статистика нередко злоупотребляет теоретическими выводами о линейной связи. Так, если

r по величине приближается к значению ± 1, часто говорят о тесной корреляции признаков (X и Y), соответственно предполагая их тесную взаимную причинную обусловленность. Однако подобное заключение вне оценки логики и формы связи признаков, как правило, служит примером всего лишь бессмысленной, или ложной, корреляции. На самом деле при $r\to\pm 1$ можно говорить о тесной взаимосвязи X и Y, если по крайней мере подтверждается линейность связи этих признаков. С другой стороны, даже анализ на линейность не обязательно свидетельствует о надежности корреляционных оценок, как в случае, если r^2 принимает малые значения, или, скажем, принимаются некорректные обобщения для специфических совокупностей данных (данпанельных обследований, данные опросов населения и т.п.). Обычно низкие значения r^2 указывают на то, что метод моделирования, в частности линейный подход, выбран для анализа явлений неверно.

Двумерная непрерывная случайная величина X, Y. Вопросы теории распределения непрерывной двумерной величины представляются важными, если учесть чисто технические сложности статистического анализа. Как известно, двумерная случайная величина предполагает не суммирование, а интегрирование значений изучаемых признаков. Достаточно эффективным приемом для исследователя при этом может быть простое распространение на случай с непрерывной величиной результатов анализа двумерной дискретной величины посредством интегрирования последней. Следует только принимать во внимание, что такое распространение аналитических результатов не всегда допустимо, о чем уже говорилось, например, выше в данной главе относительно распределения одномерной случайной величины.

Определение 1. *Совместным распределением случайных величин* X и Y называют интегральную функцию вида:

$$P[X\le x \quad \text{и} \quad Y\le y]=F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)\,du\,dv.$$

Первая производная от $F(x, y)$, или $f(x, y)$, есть функция совместной плотности вероятностей.

Интерпретация функций распределения и плотности вероятностей двумерной величины производится аналогично, как и для одномерной случайной величины.

Определение 2. ***Раздельной (граничной) плотностью вероятностей двумерной случайной величины*** называется соответствующая интегральная функция:

$$\int_D f(x,y)dy = f(x,.) =: f(x),$$

$$\int_D f(x,y)dx = f(.,y) =: g(y).$$

Определение 3. ***Ковариацией*** случайных величин X и Y называют характеристику связи этих величин, имеющую следующее формальное выражение:

$$\operatorname{cov}(X,Y) := \iint_{DD} [x - E(X)][y - E(Y)] f(x,y)\,\mathrm{d}x\mathrm{d}y.$$

Как в случае с дискретными величинами, может быть предложено также другое формальное представление ковариации:

$$\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y),$$

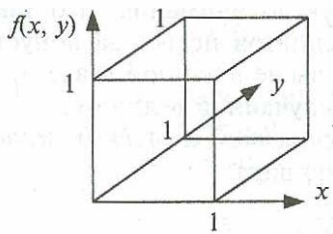
где $E(XY) := \iint_{DD} xy\, f(x,y)\ \mathrm{d}x\mathrm{d}y.$

Определение 4. Две случайные величины X и Y называются ***независимыми***, если для них имеет силу:

$$f(x,y) = f(x)g(y),$$

т.е. совместная плотность вероятностей есть произведение условных вероятностей.

Пример 3.27. Равномерное распределение в кубе со стороной единичной длины (рис. 3.22):



$f(x,y) = 1, \quad (x,y) \in ([0,1]\times[0,1]),$
$F(x,y) = yx; \quad (x,y) \in \{[0,1]\times[0,1]\}.$

Рис. 3.22

З а м е ч а н и е. Ковариация независимых случайных величин. Если случайные величины X и Y независимы, то их ковариация равна нулю.

□ *Д о к а з а т е л ь с т в о*: Следуя определению независимости случайных величин, можем записать

$$f(x,y) = f(x)\cdot g(y),$$

значит

$$E(X,Y) = \iint_{DD} xy\, f(x,y)\,\mathrm{d}x\mathrm{d}y = \iint_{DD} xy\, f(x)g(y)\,\mathrm{d}x\mathrm{d}y =$$

$$= \int_D xf(x)\,\mathrm{d}x \int_D yg(y)\,\mathrm{d}y = E(X)E(Y).$$

Как и в случае с дискретными величинами, здесь сохраняют силу отношения

{Независимость} $\Rightarrow$ {Ковариация = 0}, но не наоборот,

{Независимость} $\not\Leftarrow$ {Ковариация = 0}. ■

Определение коэффициента корреляции ρ. *Коэффициент корреляции* ρ есть ковариация двух стандартизованных случайных величин X^* и Y^*:

$$\rho = \operatorname{cov}(X^*,Y^*) = \iint_{DD} [x^* - E(X^*)][y^* - E(Y^*)] f^*(x,y)\mathrm{d}x\mathrm{d}y,$$

где $f^*(x,y)$ — функция совместной плотности вероятностей случайных величин X^* и Y^*.

З а м е ч а н и е. Коэффициент корреляции и линейность. Для коэффициентов корреляции, характеризующих связи непрерывных величин, сохраняют силу свойства:

а) $-1 \le \rho \le 1,$

б) $\rho = \pm 1 \Leftrightarrow Y = a + bX \quad (a,b \in \boldsymbol{R}).$

Определение. *Показатель определенности* есть $\rho^2 =: R^2, \ 0 \le R^2 \le 1.$ Для независимых случайных величин X и Y будет $\rho = 0$ и $R^2 = 0.$

Двумерное нормальное распределение. Рассмотрим вначале случай многомерного случайного распределения. Вектор $X = \boldsymbol{R}^m$ назовем многомерной нормально распределенной случайной величиной: $X \sim N(\mu,\Sigma)$ с математическим ожиданием, равным $E(X) = \mu \in \boldsymbol{R}^m$, а также матрицей ковариаций $\operatorname{cov}(X) = \Sigma \in \boldsymbol{R}^{m,m}$ и функцией плотности вероятностей

$$f(x) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{m/2}} \exp\left[-\frac{(x-\mu)^{1/2}\Sigma^{-1}(x-\mu)}{2}\right].$$

Теперь распределение двумерной случайной величины будем рассматривать как частный случай распределения многомерной величины при $m = 2$ $(\mu \ne 0)$:

$$X \sim N(\mu,\Sigma), \ \mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \ X = \begin{pmatrix}X_1\\ X_2\end{pmatrix}, \ \Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}, \ m = 2.$$

Если случайные величины X_1 и X_2 взаимно не коррелированны и соответственно $\sigma_{12}=\sigma_{21}=0, \sigma_{11}=:\sigma_1^2, \sigma_{22}=:\sigma_2^2$, функция совместной плотности вероятностей будет:

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1{}^2}+\frac{(x_2-\mu_2)^2}{\sigma_2{}^2}\right)\right],$$

$$(x_1,x_2)\in\{(-\infty,+\infty)\times(-\infty,+\infty)\}. \qquad (3.6)$$

Особый случай: $m=2$, $\mu=0$. Пусть $\sigma_{12}=\sigma_{21}=0$; $\sigma_{11}=\sigma_{22}=1$, $\mu_1=\mu_2=0$, тогда имеем дело со случайной величиной (условие $\mu_1=\mu_2=0$ значительно упрощает ее параметрическое представление), имеющей стандартизованные значения. Если эта величина нормально распределена, то функцию плотности вероятностей для нее можем записать так:

$$f(x_1,x_2)=\frac{1}{2\pi}e^{-\left(\frac{x_1^2}{2}+\frac{x_2^2}{2}\right)}$$

или, вводя обозначения $x:=x_1$ и $y:=x_2$, получим

$$f(x,y)=\frac{1}{2\pi}e^{-\left(\frac{x^2}{2}+\frac{y^2}{2}\right)}.$$

Для этой случайной величины функции условных вероятностей принимают вид:

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{x_j^2}{2}}\ (j=1,2) \text{ или } \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \text{ и } \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}};$$

другими словами, имеются две функции условных вероятностей для каждой из двух одномерных нормально распределенных случайных величин.

Ковариация X и Y при заданных условиях выражается единичной матрицей: $\operatorname{cov}(X,Y)=\Sigma=I_2$.

Кроме показанной выше, существует другая возможность представления бивариантного распределения: при помощи коэффициента корреляции ρ (возможно в ходе репараметризации параметров распределения). При этом, как видно из табл. 3.4, для обоих способов представления статистического распределения случайной величины прослеживается взаимосвязь параметрических оценок:

Таблица 3.4

Параметрические оценки	*Первый способ представления*	*Второй способ представления*
Математическое ожидание	μ_1,μ_2	μ_1,μ_2
Дисперсия	σ_{11},σ_{22}	$\sigma_1^2,\ \sigma_2^2$
Ковариация, Корреляция	$\sigma_{12}=\sigma_{21}$	$\rho=\dfrac{\sigma_{12}}{\sqrt{\sigma_{11}}\cdot\sqrt{\sigma_{22}}}$

При втором способе представления распределений (посредством перехода к ρ — параметрической оценке) функция плотности вероятностей будет:

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot\exp\left[-\frac{1}{2}\left(\frac{1}{\sqrt{1-\rho^2}}\right)\cdot\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-\right.\right.$$

$$\left.\left.-2\rho\frac{x_1-\mu_1}{\sigma_1}\cdot\frac{x_2-\mu_2}{\sigma_2}+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right].$$

Без доказательства можно привести несколько основных положений относительно вариативности и связей случайных величин X_1 и X_2:

1) X_1 и X_2 — вариативны, нормально распределены и соответственно имеют параметры — $X_1\sim N(\mu_1,\sigma_1^2)$ и $X_2\sim N(\mu_2,\sigma_2^2)$;

2) условные плотности вероятностей, т.е. $f(x_i|x_j)$ для X_1 и X_2 также имеют нормальное распределение с параметрами:

$$\mu=\mu_i+\frac{\sigma_i}{\sigma_j}\rho(x_j-\mu_j),\ \sigma^2=\sigma_i^2(1-\rho^2),\ i,j=1,2;$$

3) каждая из линейных комбинаций X_1 и X_2, скажем $Z=aX_1+bX_2$, вариативна относительно значений X_1 и X_2 и нормально распределена:

$$Z\sim N(a\mu_1+b\mu_2, a\sigma_1^2+b\sigma_2^2);$$

4) случайная величина Z, образованная из X_1 и X_2 как $Z=\frac{1}{1-\rho^2}\left(X_1^*-X_2^*\right)^2$, или в более подробной записи:

$$Z=\frac{1}{1-r^2}\left[\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2-2\rho\frac{X_1-\mu}{\sigma_1}\cdot\frac{X_2-\mu_2}{\sigma_2}+\left(\frac{X_2-\mu_2}{\sigma_2}\right)^2\right]$$

имеет экспоненциальное распределение с параметром $\lambda=\frac{1}{2}$. Особый случай этого распределения, при $\rho=0$, обнаруживает связь с положением 1), функция плотности вероятностей при этом будет:

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2}\cdot\exp\left[-\frac{1}{2}\cdot\left[\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{X_2-\mu_2}{\sigma_2}\right)^2\right]\right].$$

Сумма двух независимых нормально распределенных случайных величин. Пусть X и Y — две независимые нормально распределенные случайные величины:

$$X\sim N(\mu_1,\sigma_{11})\ \text{и}\ Y\sim N(\mu_2,\sigma_{22}),$$

тогда сумма $Z=X+Y$ имеет нормальное распределение $Z:=(X+Y)\sim N(\mu_1+\mu_2,\sigma_{11}+\sigma_{22})$. Плотность вероятностей величины Z также распределена нормально.

Сумма двух зависимых нормально распределенных случайных величин. Пусть X и Y — две зависимые нормально распределенные случайные величины, другими словами, имеем бивариантную нормально распределенную случайную величину, тогда сумма $Z=X+Y$ также нормально распределена:

$$Z:=(X+Y)\sim N(\mu_1+\mu_2,\sigma_{11}+\sigma_{22}+2\sigma_{12}),$$

где $\sigma_{12}=\rho\cdot\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$ и $\rho=\frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$.

(Здесь σ_{12} — общая ковариация $\sigma_{12}=\sigma_{21}$, а ρ — коэффициент корреляции, $-1\le\rho\le1$).

***m*-мерное нормальное распределение.** Случайный вектор $\begin{pmatrix}x_1\\x_2\\\vdots\\x_m\end{pmatrix}=x\in R^m$ имеет нормальное распределение при условии,

каждая из линейных комбинаций $\sum_{i=1}^{m}a_ix_i=a'x:=Z\in R$ вариативна относительно x и нормально распределена, т. е. $Z\sim N(\mu,\sigma^2)$.

Заметим, что из m вариативных нормально распределенных случайных величин всегда можно образовать одну m-мерную нормально распределенную случайную величину. В данном случае действует и обратное утверждение.

Линейные трансформации. Пусть имеем m-мерный нормально распределенный вектор $b\in R^m$, заданную матрицу $A\in R^{n,m}$ и заданный вектор $b\in R^n$, тогда $y=b+AX$ также имеет нормальное распределение. В частности, если x — единичный вектор, то каждая i-я его компонента нормально распределена и $b=0$, $A=(000...010...0)\in R^{1,m}$.

Ковариация и независимость. С точки зрения теории сотношение ковариации и независимости представляется весьма важным. Пусть по-прежнему x — нормально распределенный случайный вектор:

$$\begin{pmatrix}x_1\\x_2\end{pmatrix}=x\in R^{mn},\ x_1\in R^m\ \ x_2\in R^n;$$

здесь x_1 и x_2 имеют совместное нормальное распределение с математическим ожиданием $\mu=\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}$ и матрицей ковариаций $\Sigma=\begin{pmatrix}\sigma_{11}&\sigma_{12}\\\sigma_{21}&\sigma_{22}\end{pmatrix}$.

Независимость и ковариация. Два вектора x^1 и x^2 взаимно независимы тогда и только тогда, когда $\sigma_{12}=\sigma_{21}=0$.

Прямое утверждение, что независимость случайных величин обусловливает их нулевую ковариацию, всегда сохраняет силу. Этого, однако, нельзя сказать об обратном утверждении, что нулевая ковариация означает независимость анализируемых случайных величин. Последнее как раз не всегда правомерно, так же как, например, нельзя считать достаточным для утверждения о нормальности распределения вектора нормальное распределение его части. Для более наглядной демонстрации сказанного обратимся к примеру с линейной трансформацией случайных величин.

Пример 3.28. Исчезающая ковариация в случае зависимости двух случайных величин возникает в ходе их линейной трансформации.

1. Пусть x — нормально распределенный случайный вектор:

$$x \in R^m,\ X \sim N(\mu, \sigma^2 I_m),$$

и $A_1 \in R^{p,m}$, $A_2 \in R^{q,m}$, $A_1 A_2' = 0 \in R^{p,q}$, $Y := A_1 X$, $Y \in R^p$,

$Z := A_2 x$, $Z \in R^q$.

При заданном условии ковариация линейно-зависимых величин Y и Z будет равна нулю, в чем легко убедиться:

$$\operatorname{cov}(Y,Z) = E((A_1 X - E(A_1 X))(A_2 X - E(A_2 X))') =$$

$$= A_1 E(x - E(x - E(X))E(x - E(X))' A_2' =$$

$$= A_1 \operatorname{cov}(X) A_2' = \sigma^2 A_1 A_2' = 0.$$

2. Пусть имеется нормально распределенная величина $Y \sim N(0,1)$ и некоторая случайная независимая от Y величина b с распределением Бернулли, принимающая значения: +1 или −1 с одинаковой вероятностью, равной 1/2, при этом $Z := bY$. В данном случае можем утверждать, что имеем дело с двумя нормально распределенными величинами: Y — по исходному условию и Z — так как эта величина представляет собой смесь, состоящую наполовину из значений $+Y$ и наполовину из значений $-Y$. Более того, Z и Y — отнюдь не независимые величины, что следует из состава Z и принятой формы связи $Z := bY$. Но в то же время ковариация Z и Y остается равной нулю в соответствии с принятым в примере условием независимости Y и b:

$$\operatorname{cov}(Y,Z) = E(Y \cdot b(Y) - E(Y) \cdot E(bY) = E(Y \cdot b(Y)) - 0 \cdot 0 = E(bY^2) = 0.$$

Построение стандартных таблиц с количественными данными статистических распределений. Построение таких таблиц значительно упрощает практическое приложение статистических законов распределения; при этом отпадает необходимость в громоздких вычислениях в каждом отдельном случае. Достаточно найти нужную таблицу и на пересечении соответствующих графы и строки выделить необходимую оценку, например, плотности вероятностей или др.

Для нормального распределения, наиболее часто применяемого в исследовательской практике, определение количественных характеристик, аккумулируемых в стандартной таблице,

обычно производится при помощи приближенных оценок из распределения Бёрра.

Количественная аппроксимация нормального распределения распределением Бёрра предполагает выполнение следующего равенства:

$$\Phi(x) \approx f(x) = 1 - (1 + x^c)^{-k},\quad x \ge 0,\ c > 0,\ k > 0.$$

Например:

$$f(x) = \begin{cases} 1 - \left[1 + (0{,}644693 + 0{,}161984 \cdot x)^{4{,}874}\right]^{-6{,}158} & -3{,}97998 \le x \\ 0 & x < -3{,}97998\,. \end{cases}$$

Аппроксимация нормального распределения (плотности распределения и самого распределения) распределением Бёрра достигает достаточно высокого уровня, ошибка при этом не превышает 1%. Для наглядности покажем аппроксимирующие свойства нормального распределения графически:

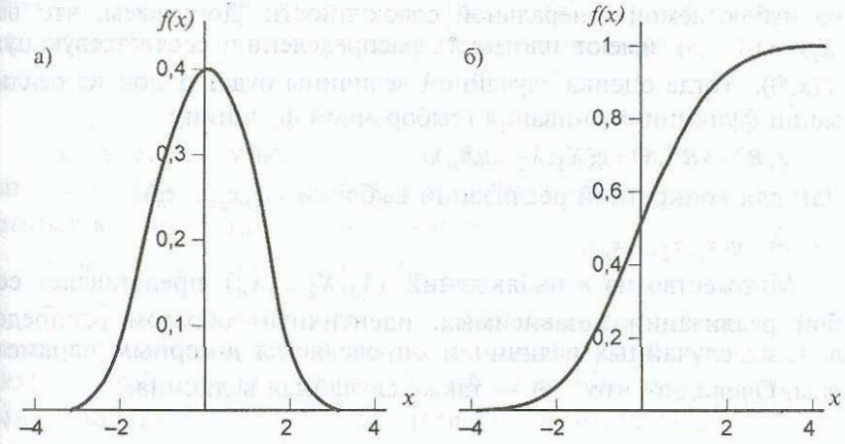


*Рис. 3.23. **Статистические распределения: нормальное (а) и Бёрра (б)***

3.6. Статистические методы точечного оценивания

Истинные параметры статистического распределения для исследователя, как правило, неизвестны. Поэтому их пытаются определить посредством приближенного оценивания по данным неко-

торой выборочной совокупности. Эта задача решается таким образом, чтобы вычисляемые параметрические оценки максимально соответствовали истинным, но неизвестным характеристикам генеральной совокупности. Так как количественные оценки параметров не обязательно полностью согласуются с эмпирическими величинами, в них предполагается наличие определенной ошибки — *ошибки оценивания*. При этом исходят из того, что изучаемый признак в генеральной совокупности подчиняется определенному закону распределения и оценка его параметров становится возможной по данным части генеральной (т.е. выборочной) совокупности, имеющей такой же, как генеральная, закон распределения. Рассмотрим это положение с формальной точки зрения.

Пусть имеется $f(x,\Theta)$ — функция плотности распределения для некоторой случайной величины X с неизвестными параметрами $\Theta \in R^p$ (для упрощения выводов предположим, что $p = 1$, хотя в общем случае $p \geq 1$). И пусть $X_1, X_2, ..., X_n$ — случайные выборки из наблюдаемой генеральной совокупности. Допускаем, что все $X_i\ (i = 1,2,...,n)$ имеют плотность распределения, соответствующую $f(x,\Theta)$. Тогда оценка случайной величины будет одной из реализаций функции оценивания (выборочной функции):

$$g: R^n \to R^p,\ \hat{\Theta} = g(X_1, X_2, ..., X_n),$$

или для конкретной реализации выборки $(x_1, x_2, ..., x_n)$:

$$\hat{\Theta} = g(x_1, x_2, ..., x_n),$$

Множество из n наблюдений $(X_1, X_2, ..., X_n)$ представляет собой реализации независимых, идентичным образом распределенных случайных величин и определяется p-мерным параметром. Очевидно, что $\hat{\Theta}$ — также случайная величина.

3.6.1. Принцип максимального правдоподобия Фишера

Пусть $(X_1, X_2, ..., X_n)$ — случайные выборки, т.е. все $X_i\ (i = 1,2,...,n)$ — независимые и идентичным образом распределенные случайные величины, имеющие плотность распределения $f(x;\Theta)$, с неизвестным параметром Θ. Тогда функцией максимального правдоподобия называют функцию вида:

$$L = \prod_{i=1}^{n} f(x_i|\Theta), \qquad (3.7)$$

логарифм этой функции

$$L = \sum_{i=1}^{n} \ln f(x_i|\Theta) — \qquad (3.8)$$

логарифмическая функция максимального правдоподобия.

Функция максимального правдоподобия максимизирует количественную оценку $\hat{\theta}$ для $\hat{\Theta}$ — оценки истинного параметра Θ. При этом оценка $\hat{\Theta}$ выбирается таким образом, что реализация функции (3.7) или эквивалентной ей функции (3.8) будет иметь наибольшее значение.

Переход от функции максимального правдоподобия к ее логарифму осуществляется на практике довольно часто и позволяет значительно упростить процедуру поиска оптимальной оценки (максимума), причем, как увидим ниже, результаты оценивания в ходе логарифмирования не претерпевают изменений.

Оптимизация при помощи монотонной трансформации. Как и прежде, будем считать $p = 1$. Пусть $f(z)$ — любая функция $R \to R$ и пусть для этой функции в окрестности $U(z_0)$ оптимальная точка будет z_0:

$$f(z_0) \geq f(z), \quad z \in U(z_0).$$

Тогда можем утверждать, что при монотонной трансформации функции $f(z)$ положение ее оптимальной точки z_0 с окрестностью $U(z_0)$ остается неизменным:

$$g: R \to R,\ g(f(z_0)) \geq g(z)),\ z \in U(z_0).$$

□ *Доказательство:* Экстремальные значения функции, как известно, определяются приравниванием первой производной этой функции нулю: $f'(z) = 0$. Для монотонно трансформируемой функции $f(z)$ соответственно будем иметь:

$$\frac{dg}{dz} = \frac{dg}{df}\frac{df}{dz},\ \frac{df}{dz} = 0 \Rightarrow \frac{dg}{dz} = 0 \quad \text{и} \quad \frac{d^2g}{dz^2} = \frac{d}{dz}\left(\frac{dg}{df}\frac{df}{dz}\right) = \left(\frac{dg}{df}\right)\left(\frac{d^2f}{dz^2}\right).$$

Учитывая монотонность функции g, знак для второй производной, а следовательно и заключение о том, достигнут максимум $f(z)$ или нет, принимается исключительно по данным самой функции $f(z)$. Особым случаем монотонной трансформации функции максимального правдоподобия является ее логарифмирование. ■

Пример 3.29. Оценивание методом максимального правдоподобия в случае экспоненциального распределения случайной величины.

В данном случае плотность вероятностей:

$$f(x) := \begin{cases} 0 & \text{для} \quad x \le 0, \\ \lambda e^{-\lambda x} & \text{для} \quad x > 0, \end{cases}$$

и функция максимального правдоподобия для одной из случайных выборок $(X_1, X_2, ..., X_n)$ величины X:

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda x_i},$$

а после логарифмирования: $L = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$.

Введем условие достижения экстремума: $\frac{dL}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0.$

Найдем вторую производную: $\frac{d^2 L}{d\lambda^2} = -\frac{n}{\lambda^2} < 0$.

Таким образом, решение, полученное методом максимального правдоподобия, принимает вид:

$$\hat{\lambda} = \overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i,$$

т.е. $\hat{\lambda}$ — максимизирующая оценка для λ.

Пример 3.30. Оценивание методом максимального правдоподобия случайной величины, имеющей нормальное распределение Гаусса.

Запишем функцию плотности вероятностей для нормального распределения:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Эта функция имеет два параметра: $\mu \in R$ и $\sigma^2 > 0$. Для одной из случайных выборок $(X_1, X_2, ..., X_n)$ функция максимального правдоподобия:

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

а ее логарифм:

$$L = -n \ln(\sqrt{2\pi}) - \frac{n}{2} \ln \sigma^2 - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}. \qquad (3.9)$$

Известно, что приравнивая первую производную функции максимального правдоподобия (3.9) нулю, выдвигают условие достижения L экстремума и получают возможность формального определения параметров нормально распределенной случайной величины (μ и σ^2). В то же время производная $\frac{dL}{d\Theta} = 0$ как бы исчезает, не позволяет установить, действительно ли в некоторой определенной точке дифференцируемая функция достигает своего максимума. Этот вопрос решается уже при нахождении вторых производных.

Найдем первые производные для L:

$$\frac{\partial L}{\partial \mu} = \frac{\sum_{i=1}^{n} (x_i - \mu)}{\sigma^2} = 0, \qquad (3.10)$$

$$\frac{\partial^2 L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2(\sigma^2)^2} = 0. \qquad (3.11)$$

Нормальное уравнение (3.10) решается относительно μ при любом $\sigma^2 > 0$, нормальное уравнение (3.11) позволяет определить σ^2 после того, как стала уже известной оценка математического ожидания μ:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{X}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 = s^2.$$

Теперь определим вторые производные функции L:

$$\frac{\partial^2 L}{\partial \mu \, \partial \mu} = -\frac{n}{2} < 0; \quad \frac{\partial^2 L}{\partial \mu \, \partial \sigma^2} = -\frac{\sum_{i=1}^{n} (x_i - \mu)}{(\sigma^2)^2} = 0;$$

$$\frac{\partial^2 L}{\partial \mu \, \partial \sigma^2} = -\frac{\sum_{i=1}^{n} (x_i - \mu)}{(\sigma^2)^2} = 0;$$

$$\frac{\partial^2 L}{\partial\sigma^2\partial\sigma^2}=\frac{n}{2\sigma^4}-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^6}=\frac{n}{2\sigma^4}-\frac{n}{\sigma^4}=-\frac{n}{2\sigma^4}<0.$$

Построив матрицу Гессе, видим, что ее детерминантный критерий действительно указывает на достижение функцией L с μ и σ^2 максимума:

$$H=\frac{-n}{\sigma^2}\begin{pmatrix}1 & 0\\ 0 & \frac{1}{2\sigma^2}\end{pmatrix};\ \det(H)=-\frac{n}{\sigma^2}\frac{1}{2\sigma^2}<0.$$

Статистическое оценивание двумерного (бивариантного) нормального распределения. Запишем функцию максимального правдоподобия для случая, когда $m=2$ и $\mu=0$:

$$\mathrm{L}=\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)^n\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_{1i}^2}{\sigma_1^2}+\frac{X_{2i}^2}{\sigma_2^2}\right)\right\}=$$

$$=\left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^n\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{X_{1i}^2}{\sigma_1^2}\right\}\cdot\left(\frac{1}{\sigma_2\sqrt{2\pi}}\right)^n\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{X_{2i}^2}{\sigma_2^2}\right\}=\mathrm{L}_1\cdot\mathrm{L}_2,$$

или

$$\mathrm{L}_k=\left(\frac{1}{\sigma_k\sqrt{2\pi}}\right)^n\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{X_{ki}^2}{\sigma_k^2}\right\}\quad(k=1,2).$$

Необходимое условие для максимизации функции L ($L=\ln\mathrm{L}=\ln\mathrm{L}_1+\ln\mathrm{L}_2$):

$$\frac{\partial\mathrm{L}}{\partial\sigma_1}=0\Rightarrow\sigma_1^2=\frac{\sum X_1^2}{n},\ \frac{\partial\mathrm{L}}{\partial\sigma_2}=0\Rightarrow\sigma_2^2=\frac{\sum X_2^2}{n},$$

т.е. для m-мерной нормально распределенной случайной величины X оценка дисперсии:

$$\hat{\sigma}_k^2=\frac{\sum X_k^2}{n}\qquad(k=1,2,\ldots,m).$$

Обобщим выводы, сделанные на случай с независимыми переменными. Пусть $\mu_1\neq0$ и $\mu_2\neq0$, тогда функция плотности вероятностей принимает вид:

$$f(X_1,X_2)=\frac{1}{2\pi\sigma_1\sigma_2}\cdot\exp\left[\frac{(X_1-\mu_1)^2}{\sigma_1^2}+\frac{(X_2-\mu_2)^2}{\sigma_2^2}\right]$$

и соответственно следует факторизация функции максимального правдоподобия:

$$\mathrm{L}=\left[\left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^n\exp\left(-\Sigma\frac{X_1^2}{2\sigma_1^2}\right)\right]\cdot\left[\left(\frac{1}{\sigma_2\sqrt{2\pi}}\right)^n\exp\left(-\Sigma\frac{X_2^2}{2\sigma_2^2}\right)\right]=\mathrm{L}_1\cdot\mathrm{L}_2.$$

Последнее уравнение после дифференцирования по μ и σ^2 позволяет получить оценки параметрических величин:

$$\hat{\mu}_1=\frac{\sum X_{1i}}{n}=\bar{X}_1;\qquad\hat{\mu}_2=\frac{\sum X_{2i}}{n}=\bar{X}_2;$$

$$\hat{\sigma}_1^2=\frac{\sum(X_{1i}-\mu_1)^2}{n};\quad\hat{\sigma}_2^2=\frac{\sum(X_{2i}-\mu_2)^2}{n}.$$

Очевидно, что проблемы статистического оценивания двумерных величин при условии зависимости или независимости переменных X_1 и X_2 имеют прямую связь с уже рассмотренными ранее проблемами оценивания одномерных нормальных распределений.

Функция плотности вероятностей в общем для m-мерной случайной величины с независимыми переменными может быть записана следующим образом:

$$f(x)=\frac{|\Sigma|^{-1/2}}{(2\pi)^{m/2}}\exp\left[-\frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{2}\right],\ \mu\in R^m,\ \Sigma\in R^{m,m},$$

для одномерной случайной величины эта же функция:

$$f(x)=\frac{|\sigma^2|^{-1/2}}{(2\pi)^{1/2}}\exp\left[-\frac{(x-\mu)(\sigma^2)^{-1}(x-\mu)}{2}\right],\ \mu\in R^1,\ \Sigma=\sigma\in R^{1,1},$$

для двумерного случая в подробной записи:

$$f(x)=f(x_1,x_2)=$$

$$=\frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}-\rho^2}}\exp\left[-\frac{1}{2}(x_1-\mu_1,x_2-\mu_2)\begin{pmatrix}\sigma_{11} & \rho\\ \rho & \sigma_{22}\end{pmatrix}^{-1}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}\right],$$

где $\begin{pmatrix}\sigma_{11} & \rho\\ \rho & \sigma_{22}\end{pmatrix}$ — упрощенная запись обычной матрицы ковариаций:

$$\Sigma=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}=\begin{pmatrix}\sigma_{11} & \rho\\ \rho & \sigma_{22}\end{pmatrix}$$

здесь $\sigma_{12}=\sigma_{21}$ обозначаются символом ρ.

В двумерном случае, как видим, возникает проблема статистического оценивания пяти параметров: $\Theta=\{\mu_1,\mu_2,\sigma_{11},\sigma_{22},\rho\}$. Использование функции плотности вероятностей в оценивании Θ предполагает определение следующих дисперсионных характеристик:

• корня квадратного из определителя матрицы ковариаций (считается, что детерминант $|\Sigma|$ положительно определен)

а) $|\Sigma|^{\frac{1}{2}}=\frac{1}{\sqrt{|\Sigma|}}$, $|\Sigma|=\sigma_{11}\sigma_{22}-\rho^2 \therefore \frac{1}{\sqrt{|\Sigma|}}=\frac{1}{\sqrt{\sigma_{11}\sigma_{22}-\rho^2}}$;

• обратной матрицы ковариаций — положительно определенной:

б) $\Sigma^{-1}=\frac{1}{\sigma_{11}\sigma_{22}-\rho^2}\begin{pmatrix}\sigma_{22} & -\rho\\ -\rho & \sigma_{11}\end{pmatrix}=D\begin{pmatrix}\sigma_{22} & -\rho\\ -\rho & \sigma_{11}\end{pmatrix}$, $D:=\frac{1}{\sigma_{11}\sigma_{22}-\rho^2}$

• показателя дисперсии по выборочным наблюдениям

в) $S:=(x_1-\mu_1,x_2-\mu_2)\begin{pmatrix}\sigma_{22} & -\rho\\ -\rho & \sigma_{11}\end{pmatrix}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}=$

$=\sigma_{22}(x_1-\mu_1)^2+\sigma_{11}(x_2-\mu_2)^2-2\rho(x_1-\mu_1)(x_2-\mu_2)$.

Для нахождения всех пяти параметрических оценок Θ воспользуемся уже известным методом максимального правдоподобия (при $m=2$) и соответственно логарифмом максимизирующей функции L. Обратим внимание, что в последующих выводах знак $\sum$ применяется только как знак суммы $\left(\sum_{i=1}^{n}\right)$ и не связан с представлением матрицы ковариаций. Логарифм функции максимального правдоподобия в данном случае принимает вид:

$$L=-\frac{n}{2}\ln(\sigma_{11}\sigma_{22}-\rho^2)-\frac{D}{2}\sum[\sigma_{22}(x_1-\mu_1)^2+\sigma_{11}(x_2-\mu_2)^2-$$
$$-2\rho(x_1-\mu_1)(x_2-\mu_2)]=-\frac{n}{2}\ln(|\Sigma|)-\frac{1}{2}D\sum S.$$

Чтобы найти оценки двух математических ожиданий для $X'=(X_1X_2)$, продифференцируем функцию L последовательно по μ_1 и μ_2:

$$\frac{\partial L}{\partial\mu_1}=0=-\frac{D}{2}[-2\sum\sigma_{22}(x_1-\mu_1)+2\rho(x_2-\mu_2)],$$
$$\frac{\partial L}{\partial\mu_2}=0=-\frac{D}{2}[-\sum 2\sigma_{11}(x_2-\mu_2)+2\rho(x_1-\mu_1)]$$

или запишем в виде нормальных уравнений:

$$-\sigma_{22}\sum(x_1-\mu_1)+\rho\sum(x_2-\mu_2)=0, \quad (3.12)$$
$$-\sigma_{11}\sum(x_2-\mu_2)+\rho\sum(x_1-\mu_1)=0. \quad (3.13)$$

Уравнения (3.12), (3.13) представляются для определения μ_1, μ_2 совершенно равноценными, при этом всегда

$$\hat{\mu}_1=\frac{\sum X_{1i}}{n}=\bar{X}_1, \quad \hat{\mu}_2=\frac{\sum X_{2i}}{n}=\bar{X}_2.$$

Для нахождения трех оставшихся оценок Θ также произведем дифференцирование функции L, но уже по соответствующим величинам $\sigma_{11},\sigma_{22},\rho$:

$$\frac{\partial L}{\partial\rho}=0,\ \frac{\partial L}{\partial\sigma_{11}}=0,\ \frac{\partial L}{\partial\sigma_{22}}=0;$$
$$\frac{\partial L}{\partial\rho}=\frac{n}{2}\frac{2\rho}{\sigma_{11}\sigma_{22}-\rho^2}-\frac{1}{2}\left(\frac{\partial D}{\partial\rho}\cdot S+D\cdot\frac{\partial S}{\partial\rho}\right)=0;$$
$$\frac{\partial L}{\partial\sigma_{11}}=-\frac{n}{2}\frac{\sigma_{22}}{\sigma_{11}\sigma_{22}-\rho^2}-\frac{1}{2}\left(\frac{\partial D}{\partial\sigma_{11}}\cdot S+D\cdot\frac{\partial S}{\partial\sigma_{11}}\right)=0;$$
$$\frac{\partial L}{\partial\sigma_{22}}=-\frac{n}{2}\frac{\sigma_{11}}{\sigma_{11}\sigma_{22}-\rho^2}-\frac{1}{2}\left(\frac{\partial D}{\partial\sigma_{22}}\cdot S+D\cdot\frac{\partial S}{\partial\sigma_{22}}\right)=0.$$

Кроме этого нам понадобятся частные производные по дисперсионным показателям D и S:

$$\frac{\partial D}{\partial\rho}=-(\sigma_{11}\sigma_{22}-\rho^2)^{-2}(-2\rho)=\frac{2\rho}{(\sigma_{11}\sigma_{22}-\rho^2)^2};$$
$$\frac{\partial D}{\partial\sigma_{ii}}=-(\sigma_{11}\sigma_{22}-\rho^2)^{-2}\cdot\sigma_{jj},\ \text{если } i=1, j=2 \text{ и если } i=2, j=1, \text{ т.е}$$
$$\frac{\partial D}{\partial\sigma_{11}}=-\frac{\sigma_{22}}{(\sigma_{11}\sigma_{22}-\rho^2)^2},\ \frac{D}{\partial\sigma_2}=-\frac{\sigma_{11}}{(\sigma_{11}\sigma_{22}-\rho^2)^2};$$
$$\frac{\partial S}{\partial\sigma_{11}}=\sum_{i=1}^{n}(x_{2i}-\mu_2)^2,\ \frac{\partial S}{\partial\sigma_{22}}=\sum_{i=1}^{n}(x_{1i}-\mu_i)^2,\ \frac{\partial S}{\partial\rho}=-2\sum_{i=1}^{n}(x_{1i}-\mu_1)(x_{2i}-\mu_2).$$

Система нормальных уравнений для дифференциалов $\frac{dD}{\partial\rho}$; $\frac{\partial D}{\partial\sigma_{11}}; \frac{\partial S}{\partial\sigma_{22}}$ принимает вид:

$$0=\frac{n\rho}{\sigma_{11}\sigma_{22}-\rho^2}-\frac{1}{2}\frac{\partial D}{\partial\rho}\sum S-\frac{1}{2}D\sum\frac{\partial S}{\partial\rho}, \quad (3.14)$$

$$0=\frac{-n\sigma_{22}}{\sigma_{11}\sigma_{22}-\rho^2}-\frac{\partial D}{\partial\sigma_{11}}\sum S-D\sum\frac{\partial S}{\partial\sigma_{11}}, \quad (3.15)$$

$$0=\frac{-n\sigma_{11}}{\sigma_{11}\sigma_{22}-\rho^2}-\frac{\partial D}{\partial\sigma_{22}}\sum S-D\sum\frac{\partial S}{\partial\sigma_{22}}. \quad (3.16)$$

При установленных μ_1 и μ_2 решение системы нелинейных уравнений (3.14) — (3.16) приводит к получению наилучших оценок $\hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\rho}$. Это, как можно заметить, обычные показатели дисперсии и ковариации:

$$\hat{\sigma}_{11}=\frac{\sum(x_{12}-\bar{x}_1)^2}{n};\ \hat{\sigma}_{22}=\frac{\sum(x_{22}-\bar{x}_2)^2}{n};\ \hat{\sigma}_{12}=\frac{\sum(x_{12}-\bar{x}_1)^2(x_{22}-\bar{x}_2)^2}{n}.$$

Замечание 1. Факторизация функции максимального правдоподобия. Введем обозначение: пусть $d:=\left|\sum\right|=\sigma_{11}\sigma_{22}-\rho$. Тогда функцию максимального правдоподобия можем переписать в виде:

$$L=-\frac{n}{2}\ln d-\frac{\sigma_{22}}{2d}\sum(x_1-\mu_1)^2-\frac{\sigma_{11}}{2d}\sum(x_2-\mu_2)^2+\frac{2\rho}{2d}\sum(x_1-\mu_1)(x_2-\mu_2),$$

откуда, например, методом моментов получаем

$$\sigma_{11}=E(x_1-\mu_1)^2,\ \sigma_{22}=E(x_2-\mu_2)^2,\ \sigma_{12}=E(x_1-\mu_1)(x_2-\mu_2).$$

Следовательно,

$$L=-\frac{n}{2}\ln d-\frac{\sigma_{22}\sigma_{11}n}{2d}-\frac{\sigma_{11}\sigma_{22}n}{2d}+\frac{2\rho}{2d}\sigma_{12}n=$$

$$=-\frac{n}{2}\left(\ln d+\frac{1}{d}(2\sigma_{22}\sigma_{11}-2\sigma_{12}\sigma_{21})\right)=-\frac{n}{2}(\ln d+2).$$

Таким образом, запись функции максимального правдоподобия может быть существенно сокращена и сведена по сути к формальной увязке только двух, заранее известных, константных величин (n и d). Важным представляется, что редуцированная запись L сохраняет силу для любой выборки и любой совокупности выборочных данных.

Замечание 2. Проверка нормальных уравнений. Покажем справедливость утверждения, что максимально правдоподобные статистические оценки будут удовлетворять некоторым уравнениям (3.17)—(3.19), приведенным ниже. С этой целью обратимся к частным производным, для которых имеют силу следующие равенства:

$$\frac{\partial D}{\partial\rho}=\frac{2\rho}{d^2};\ \frac{\partial D}{\partial\sigma_{11}}=-\frac{\sigma_{22}}{d^2};\ \frac{\partial D}{\partial\sigma_{22}}=-\frac{\sigma_{11}}{d^2};$$

$$\sum\frac{\partial S}{\partial\sigma_{11}}=n\sigma_{22};\ \sum\frac{\partial S}{\partial\sigma_{22}}=n\sigma_{11};\ \sum\frac{\partial S}{\partial\rho}=-2n\sigma_{12};$$

$$\sum S=\sigma_{22}n\sigma_{11}+\sigma_{11}n\sigma_{22}-2\rho n\sigma_{12}n(2\sigma_{11}\sigma_{22}-2\sigma_{12}\sigma_{21})=2nd.$$

Соответственно построим систему нормальных уравнений:

$$\frac{n\rho}{d}-\frac{1}{2}\cdot\frac{2\rho}{d^2}\cdot 2nd-\frac{1}{2}\cdot\frac{1}{d}\cdot(-2n\sigma_{12})=0\ (\sigma_{12}=\rho\ \text{по определению}), \quad (3.17)$$

$$\frac{n\rho}{d}-\frac{2n\rho}{d}+\frac{n\sigma_{12}}{d}=0\Leftrightarrow\frac{n}{d}(\rho-2\rho+\rho)=0;$$

$$-\frac{n\sigma_{22}}{d}-\left(-\frac{\sigma_{22}}{d^2}\right)\cdot 2nd-\frac{1}{d}\cdot(n\sigma_{22})=0, \quad (3.18)$$

$$\frac{n}{d}(-\sigma_{22}+2\sigma_{22}-\sigma_{22})=0;$$

$$-\frac{n\sigma_{11}}{d}-\left(-\frac{\sigma_{11}}{d^2}\right)\cdot 2nd-\frac{1}{d}\cdot(n\sigma_{11})=0, \quad (3.19)$$

$$\frac{n}{d}(-\sigma_{11}+2\sigma_{11}-\sigma_{11})=0.$$

Примеры смеси нормально распределенных случайных величин. Предположим, что имеются две нормально распределенные случайные и независимые величины X и Y. При условии $N(0, 1)$ можно рассматривать следующие варианты суммирования этих величин.

1. Первая вариация дисперсии. Пусть имеется

$$Z:=0{,}8\cdot X+0{,}2\cdot Y,\ X\sim N\left(\frac{1}{0{,}8+0{,}2\alpha^2}\right),\ Y\sim N\left(\frac{\alpha^2}{0{,}8+0{,}2\alpha^2}\right),\ \frac{1}{8}\le\alpha\le 1.$$

Тогда принимая, что величина Z, как X и Y, нормально распределена, и $Z \sim N(0, 1)$ для любых α, можем построить два графика смесей (рис. 3.24 а,б):

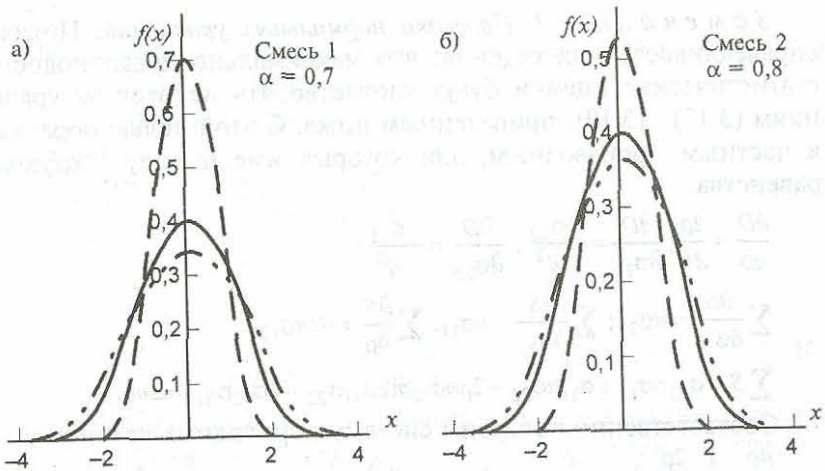


Рис. 3.24

2. Вторая вариация дисперсии. Пусть имеется:

$$Z := 0{,}7 \cdot X + 0{,}3 \cdot Y,\ X \sim N\left(\frac{\mu}{0{,}7}, 1\right),\ Y \sim N\left(-\frac{\mu}{0{,}3}, 1\right),\ 0 \le \mu \le 1.$$

Здесь также Z — нормально распределенная величина и $Z \sim N(0, 1)$ для любых μ. Построим два графика смесей: для $\mu = 0{,}9$ и $\mu = 0{,}7$ (рис. 3.25 а,б).

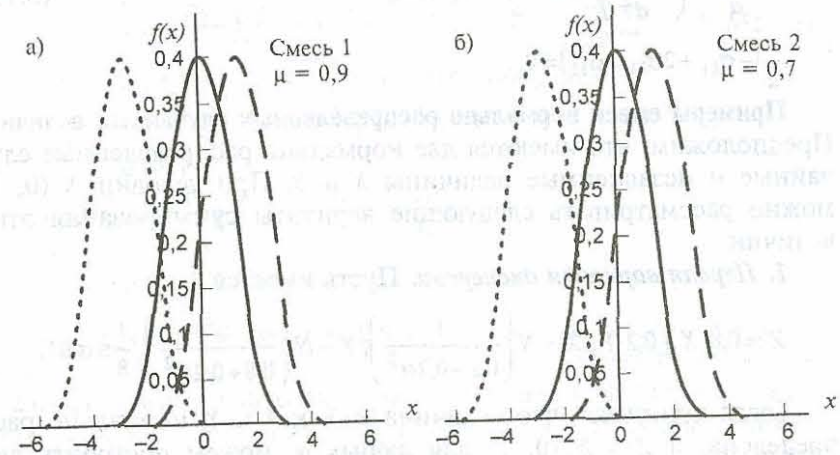


Рис. 3.25

Двумерное ненормальное распределение с нормальными раздельными плотностями вероятностей[1]. *Утверждение.* Двумерная случайная величина, не подчиняющаяся нормальному закону распределения с плотностью вероятностей $f(x, y)$, может иметь нормально распределенные раздельные плотности вероятностей:

$$f(x,y) = 2\delta(x \cdot y) \cdot g(x) \cdot g(y), \delta(x \cdot y) = \begin{cases} 1 & \text{для } xy > 0, \\ 0 & \text{для } xy \le 0, \end{cases}$$

где g — функция плотности вероятностей, имеющая нормальное распределение $N(0, 1)$.

□ *Доказательство*: Совместная плотность вероятностей двумерной величины может рассматриваться как результат интегрирования раздельных плотностей вероятностей. При этом для одной из двух переменных, например X, правомерно записать:

$$\int_{-\infty}^{+\infty} f(x,y)\,dy = 2\int_{-\infty}^{+\infty} \delta(x \cdot y) \cdot g(x) \cdot g(y)\,dy =$$

$$= 2g(x) - \int_{-\infty}^{+\infty} \delta(x \cdot y) \cdot g(y)\,dy = 2g(x) \cdot \frac{1}{2}.$$

Появление в формальной записи значения 1/2 объясняется тем, что функции $g(y)$ и $g(x)$ согласно δ-условию интегрируют на интервале $[0; \infty)$. Но тогда X и Y — симметрично распределенные величины, что и исчерпывает наше доказательство. ■

Пример 3.31. Оценивание методом максимального правдоподобия распределений: биномиального и Бернулли.

Для распределения Бернулли функция плотности вероятностей имеет вид:

$$p_x = \binom{1}{x} p^x (1-p)^{1-x};\ x = 0,1; \qquad (3.20)$$

при этом некоторая случайная выборка

$$\{x_1, x_2, \ldots, x_n | x_i \in \{0,1\}\ \ i = 1,2,\ldots,n\}$$

позволяет построить следующую функцию максимального правдоподобия:

$$L = \prod_{i=1}^{n} \binom{1}{x_i} p^{x_i} (1-p)^{1-x_i};$$

[1] В отечественной литературе встречаются термины: раздельные, граничные или маргинальные. (*Прим. пер.*)

логарифм этой функции:

$$L=\sum_{i=1}^{n}\left(\ln\binom{1}{x_i}+x_i\ln p+(1-x_i)\ln(1-p)\right) \qquad (3.21)$$

или, учитывая последующую оптимизацию, в приближении к исходной функции L можем записать:

$$L=\sum_{i=1}^{n}x_i\ln p+(1-x_i)\ln(1-p).$$

Введем первое условие достижения L экстремальных значений:

$$\frac{dL}{dp}=\frac{\sum_{i=1}^{n}x_i}{p}-\frac{\sum_{i=1}^{n}(1-x_i)}{1-p}=0,$$

откуда непосредственно находим решение: $\hat{p}=\overline{X}$.

Таким образом, полученные алгебраические результаты представляются соотношением числа испытаний, другими словами, удельным весом успешных испытаний (их квотой).

Дальнейшее исследование экстремумов функции максимального правдоподобия требует определения второй производной:

$$\frac{d^2L}{dp^2}=-\frac{\sum_{i=1}^{n}x_i}{p^2}-\frac{\sum_{i=1}^{n}(1-x_i)}{(1-p)^2}<0,$$

откуда следует: $\hat{p}$ задает максимум оценивающей функции L, именно это и требовалось доказать.

Выводы, полученные при анализе распределения Бернулли, непосредственно распространяются и на случай биномиального распределения. Исходная функция плотности вероятностей при этом имеет вид:

$$p_x=\binom{K}{x}p^x(1-p)^{K-x},\quad x=0,1,2,\ldots,K. \qquad (3.20')$$

Функция максимального правдоподобия для некоторой случайной выборки $\{x_1,x_2,\ldots,x_n | x_i\in\{0,1,2,\ldots,K\};\ i=1,2,\ldots,n\}$ будет:

$$L=\sum_{i=1}^{n}\left(\ln\binom{K}{x_i}+x_i\ln p+(K-x_i)\ln(1-p)\right). \qquad (3.21')$$

В отличие от распределения Бернулли здесь область определения функции плотности вероятностей не $\{0,1\}$, а $\{0,1,2,\ldots,K\}$.

При оптимизации решающей функции L получаем $\hat{p}=\frac{\overline{X}}{K}$.

3.6.2. Метод моментов

Достаточно часто в случаях, когда при оценивании статистического распределения становится невозможным использование метода максимального правдоподобия, положительного результата добиваются при помощи метода моментов. В статистической теории уже давно доказано, что несмотря на то, что оценки, полученные методом моментов, по своим свойствам уступают максимально правдоподобным оценкам, в отличие от последних практически всегда могут быть найдены для параметрического распределения с заданной плотностью вероятностей. На практике это означает, что решение может быть получено даже тогда, когда не представляется возможным статистически описать все особенности структуры наблюдаемой совокупности данных, или когда имеющееся формальное выражение функции плотности вероятностей ставится исследователем под сомнение в силу своей специфичности.

Основной принцип оценивания статистического распределения методом моментов. Пусть имеется множество из n случайных независимых величин $(X_1, X_2, .., X_n)$, подчиняющихся одному закону распределения. Плотность вероятностей для такого множества определяется p — числом параметров: $\Theta=(\Theta_1,\Theta_2,\ldots,\Theta_p)$, что в свою очередь позволяет говорить об $(X_1, X_2, \ldots, X_n)$ как о параметрически определенной случайной выборке. Структура последней поддерживается существованием некоторых моментов, в числе их математическое ожидание: $E(X^r)$ $(r=1,2,3,\ldots)$. При $r=1$, $r=2$ структура данных выборки обусловливается существованием $E(X)$ и var (X). Впрочем, как показывает распределение Коши, существование именно математического ожидания не всегда обязательно.

Эмпирические моменты являются выборочными функциями. Принимая это положение за исходное, обратимся непосредственно к рассмотрению проблем оценивания методом моментов. Процесс оценивания предполагает получение некоторой количе-

ственной характеристики $\hat{\theta}$ для неизвестных параметров случайной выборки $(X_1, X_2, ..., X_n)$. С этой целью априори допускается равенство эмпирических и теоретических моментов, и сама оценка $\hat{\theta}$ выбирается таким образом, чтобы, например, для плотности двупараметрического нормального распределения иметь:

$$E(X)=g_1(\Theta)=g_1(\hat{\Theta})=\bar{X}, \qquad (3.22)$$

$$\mathrm{var}(X)=E(X^2)-E(X)^2=g_2(\hat{\Theta})=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\bar{X}^2. \qquad (3.23)$$

Подобный подход характеризуется как оценивание методом моментов по Пирсону.

Пример 3.32. Ситуация с оцениванием нормального распределения видится особенно простой, так как в этом случае оценки, полученные методом моментов и методом максимального правдоподобия, полностью совпадают:

$$\mu=E(X)=\bar{X}=\hat{\mu},\quad \sigma^2=\mathrm{var}(X)=s^2=\hat{\sigma}^2.$$

Пример 3.33. Оценивание методом моментов экспоненциального распределения. Для экспоненциального распределения имеем:

$$f(X)=\begin{cases}0 & x\le 0,\\ \lambda e^{-\lambda x} & \text{для } x>0;\end{cases}$$

$E(X)=1/\lambda$ и $\mathrm{var}(X)=1/\lambda^2$.

Оценка X, полученная методом моментов, будет:

$$\hat{\lambda}=\frac{1}{\bar{X}}.$$

3.6.3. Основные свойства статистических оценок

Представив выше различные методы точечного оценивания случайных величин, обратимся к вопросам критериального выбора наилучших параметрических оценок.

Понятие несмещенности и состоятельности математического ожидания. Пусть имеем случайную выборку достаточного объема $X=(X_1, X_2, ..., X_n)$ и соответствующую для X_i функцию плотности вероятностей $f(\Theta)$, $\Theta\in R^K$; кроме того, пусть $\hat{\Theta}_k=f_k(X)$ $(k=1,2,...,K)$ — выборочная функция для неизвестного параметра

Θ_k. Тогда $\hat{\Theta}_k$ называют *несмещенной* и *состоятельной оценкой*, если $E(\hat{\Theta}_k)=\Theta_k$ $(k=1,2,...,K)$.

При этом существующая гипотетически величина $\mathrm{bias}(\hat{\Theta}_k)=E(\hat{\Theta}_k)-\Theta_k$ $(k=1,2,...,K)$ характеризуется как результат смещения параметрической оценки.

Пример 3.34. Несмещённость оценки средней. Пусть имеем $\hat{\mu}=\bar{X}$, тогда $E(\bar{X})=\mu$. Этот вывод непосредственно следует из правила по определению математического ожидания.

Пример 3.35. Несмещённость оценки, полученной методом моментов для распределения $U(0,b)$. Пусть случайная величина X имеет равномерное распределение $U(0;b)$ и $2\bar{X}=\hat{b}$, тогда имеет силу равенство:

$$E(\hat{b})=2E(\bar{X})=2\frac{b}{2}=b.$$

Смещенность оценки дисперсии. Пусть имеем случайную величину X с любым распределением, но с математическим ожиданием, равным μ, и дисперсией σ^2. Тогда для оценки

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad E(\hat{\sigma}^2)=E(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2)=\frac{n-1}{n}\sigma^2, \qquad (3.24)$$

т.е. оценка $\hat{\sigma}^2$ смещенная. Доказательство выдвинутого положения с целью упрощения может быть проведено первоначально при условии, что $n=2$, а затем индуктивным способом распространено на общий случай, когда $n\ge 2$.

Устранение смещённости дисперсионной оценки достигается заменой (3.24) на модифицированную формулу:

$$\tilde{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2. \qquad (3.25)$$

В формуле (3.25) оценка дисперсии уже несмещенная.

Существует также и другой способ избежать смещенности дисперсионной оценки — увеличение размера выборки, т.е. необходимо асимптотическое предположение, что $n\to\infty$.

Определение: *Асимптотическая несмещенность.*

Пусть некоторая оценка $\hat{\Theta}:=f_k(X)$ $(k=1,2,...,K)$ ного параметра Θ_k смещенная. Это означает, что:

$$E(\hat{\Theta}_k)\ne\Theta_k \quad (k=1,2,...,K).$$

Эту же оценку называют *асимптотически несмещенной*, если ее смещенность исчезает при условии $n\to\infty$, т.е. действуют равенства:

$$\lim_{n\to\infty} E(\hat{\Theta}_k)=\Theta_k \quad (k=1,2,...,K) \tag{3.26}$$

и соответственно $\lim_{n\to\infty} \text{bias}(\hat{\Theta}_k)=0.$ (3.26')

Пример 3.36. Асимптотическая нормальность оценки дисперсии.

Пусть X — выборка любой величины с любым распределением и имеющая характеристику дисперсии σ^2, тогда

$$\lim_{n\to\infty} E(\hat{\sigma}^2)=\lim_{n\to\infty} E(s^2)=\lim_{n\to\infty}\frac{n-1}{n}\sigma^2=\sigma^2.$$

Эффективность статистических оценок. ***Определение:*** Оценку $\hat{\Theta}$ называют по отношению к оценке $\tilde{\Theta}$ *более эффективной*, если при прочих равных условиях выполняется неравенство:

$$\text{var}(\hat{\Theta})\le\text{var}(\tilde{\Theta}).$$

С введенным понятием эффективности тесно взаимосвязано понятие наилучшей оценки, т.е. имеющей наименьшую дисперсию в некотором классе оценок (например, в классе линейных несмещенных оценок).

Пример 3.37. Оценка средней.

Сравним три оценки средней величины:

$$\hat{\mu}=\overline{X},\ E(\overline{X})=\mu,$$

$$\tilde{\mu}=a\left(\frac{1}{n-1}\sum_{i=1}^{n-1}X_i\right)+(1-a)X_n,\ 0<a<1,\ E(\tilde{\mu})=\mu,$$

$$\overline{\mu}=\sum_{i=1}^{n}a_iX_i,\ 0<a_i<1,\quad \sum_{i=1}^{n}a_i=1,$$

при этом даже для $a_i\neq\frac{1}{n}$ будет выполняться равенство: $E(\overline{\mu})=\mu$.

Для дисперсии случайной величины X имеем:

$$\text{var}(\hat{\mu})=\frac{\sigma^2}{n},$$

$$\text{var}(\tilde{\mu})=\sigma^2\left[\frac{na^2}{n-1}+(1-2a)\right],$$

$$\text{var}(\overline{\mu})=\sigma^2\left[\sum_{i=1}^{n}a_i^2\right]>\text{var}(\hat{\mu})=\frac{\sigma^2}{n}.$$

Определение: Асимптотическая эффективность.

Оценку $\hat{\Theta}$ называют асимптотически более эффективной относительно другой оценки $\tilde{\Theta}$, если

$$\lim_{n\to\infty}\text{var}(\hat{\Theta})\le\lim_{n\to\infty}(\tilde{\Theta}).$$

Средняя квадратическая ошибка. ***Определение:*** Величина

$$MSE:=E[(\hat{\Theta}_k-\Theta_k)^2]\quad (k=1,2,...,K),$$

или $MSE:=E[(\hat{\Theta}-\Theta)^2]$ для $K=1$, характеризуется как средний *квадрат ошибки* ("Mean Square Error") оценки $\hat{\Theta}_k$ (или просто $\hat{\Theta}$).

Определение: Разложение дисперсии MSE.

Запишем *MSE* как

$$MSE=E[(\hat{\Theta}-\Theta)^2]=E\{[[\hat{\Theta}-E(\hat{\Theta})]-[\Theta-E(\hat{\Theta})]]^2\}=$$

$$=E[\hat{\Theta}-E(\hat{\Theta})]^2+\{E[\Theta-E(\hat{\Theta})]\}^2+2E[\hat{\Theta}-E(\hat{\Theta})]E[\Theta-E(\hat{\Theta})]\}=$$

$$=\text{var}(\hat{\Theta})+\text{bias}(\hat{\Theta})^2.$$

Подобная запись представляет разложение дисперсии среднего квадратического отклонения. Очевидно, что средняя квадратическая ошибка тогда и только тогда равна нулю, когда равны нулю дисперсия и смещение статистической оценки.

Замечание. Условие равенства нулю для дисперсии статистической оценки целесообразно вводить при необходимости нивелирования средней квадратической ошибки. В сущности такое условие означает, что сама статистическая оценка перестает быть случайной величиной и устанавливается детерминистическим путем. В аналитической практике известен по крайней мере один из примеров, когда логично требование устранения *MSE* — когда мы имеем дело не с выборочной, а с генеральной совокупностью.

Состоятельность статистических оценок. ***Определение:*** *Сходимость по вероятности.* Некоторый ряд, состоящий из независимых, идентично распределенных случайных величин, называют сходящимся по вероятности к x, если выполняется равенство:

$$\lim_{n\to\infty}P[|X_n-x|\ge k]=0.$$

Определение. *Состоятельность.* Статистическая оценка $\hat{\Theta}$ называется состоятельной, если по вероятности сходится к истинной оценке Θ, т.е.:

$$\lim_{n\to\infty} P\left[\left|\hat{\Theta}-\Theta\right|\geq c\right]=0, \quad c>0.$$

Замечание: Вывод о состоятельности может быть сделан относительно оценки ($\hat{\Theta}$) самой случайной величины, но не относительно ее среднего значения. На деле это означает, что $\hat{\Theta}$ не должна отклоняться от Θ, или что вероятность такого отклонения очень мала (см. неравенство Чебышева).

Для несмещенных и состоятельных оценок действует общее правило: при увеличивающемся числе наблюдений их дисперсия уменьшается и постепенно с увеличивающейся плотностью концентрируется вокруг оцениваемого параметра. Проиллюстрируем это графически (рис. 3.26):

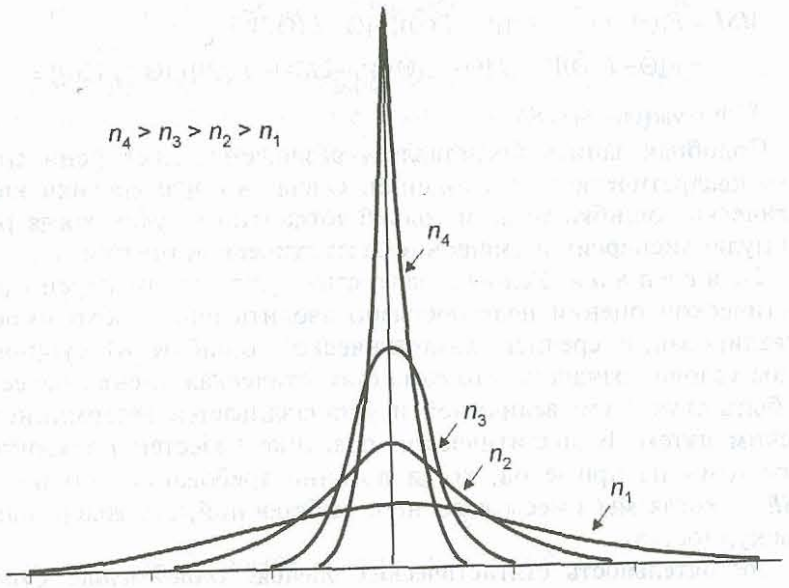


Рис. 3.26

В статистике для состоятельной оценки используется следующая сокращенная запись: $p\lim\hat{\Theta}=\Theta$.

Замечание. *Состоятельность статистической оценки* представляет собой асимптотический результат, полученный при условии: $n\to\infty$. Это, однако, не исключает получение подобных же результатов и для относительно малых выборок, скажем, когда $n>30$.

Пример 3.38. *Оценка средней.*

Оценка средней является состоятельной оценивающей функцией для μ:

$$E(\bar{X})=\mu, \quad \lim_{n\to\infty}\operatorname{var}(\bar{X})=\lim_{n\to\infty}\frac{\sigma^2}{n}=0,$$

следовательно $0\leq\lim_{n\to\infty}P(\left|\bar{X}-E(\bar{X})\right|>c)\leq\lim_{n\to\infty}\frac{\operatorname{var}(\bar{X})}{c^2}=0.$

Пример 3.39. Пусть имеется некоторая оценка $\hat{\Theta}$ для двухточечного распределения:

$$P[\hat{\Theta}=0]=1-\frac{1}{n} \text{ и } P[\hat{\Theta}=n]=\frac{1}{n}[E(\hat{\Theta})=1;\ \operatorname{var}(\hat{\Theta})=n-1].$$

При условии состоятельности оценки $\hat{\Theta}$ можем записать:

$$\lim P[\left|\hat{\Theta}-\Theta\right|]=0, \text{ или } p\lim\hat{\Theta}=\Theta=0,$$

и, учитывая возможность существования некоторого, сколь угодно малого положительного числа (e):

$$1\geq P[\left|\hat{\Theta}\right|<e]\geq 1-\frac{1}{n}.$$

Другими словами, сама оценка $\hat{\Theta}$ может быть сколь угодно малой.

Отметим, что в данном случае мы не имеем дело с известным неравенством Чебышева. Последнее для нашей задачи не несет ничего рационального, так как при $n\to\infty$ дисперсия становится величиной неопределенной:

$$P[\left|\hat{\Theta}-E(\hat{\Theta})\right|\leq k]\geq 1-\frac{\operatorname{var}(\hat{\Theta})}{k^2},$$

$$P[\left|\hat{\Theta}-1\right|\leq k]\geq 1-\frac{n-1}{k^2}\to-\infty \quad (\text{для } n\to\infty).$$

В заключение сделаем обзор некоторых взаимосвязанных свойств статистической оценки $\hat{\Theta}$ для неизвестного параметра Θ (рис. 3.27):

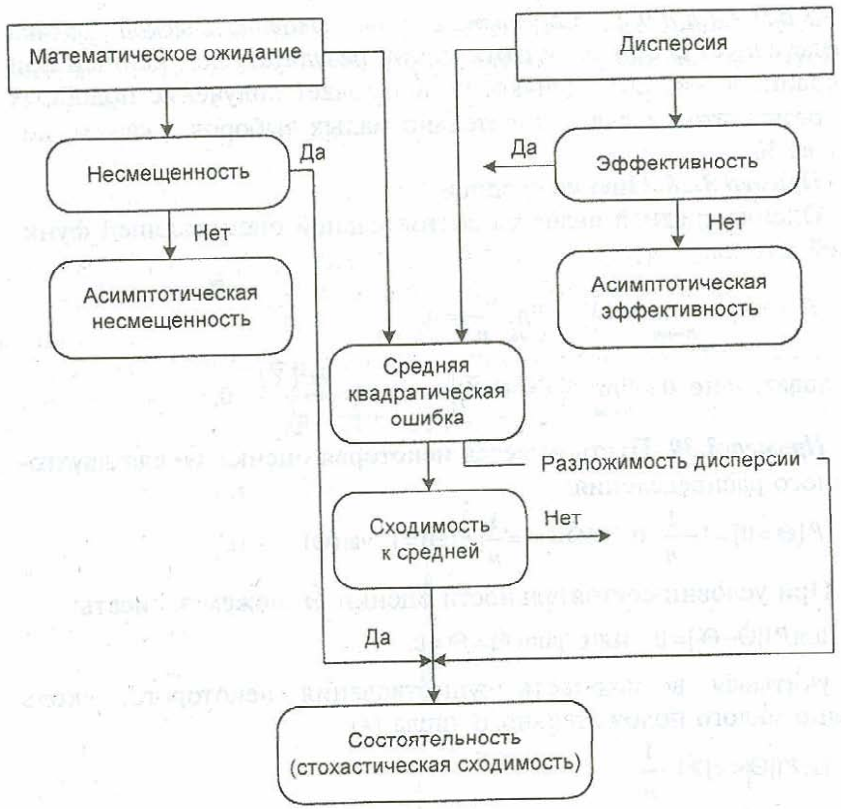


*Рис. 3.27. **Обзор некоторых важнейших свойств статистической оценки для неизвестного параметра** Θ:*

→ возможные переходы

Выводы[1]

Наличие данной главы в составе пособия по многомерному анализу объясняется большим набором методов, в которых исследователю приходится сталкиваться с изучением характера

[1] Параграф подготовлен Л.А. Сошниковой и В.Н. Тамашевичем

распределения случайных величин. В соответствии с законом распределения оценивают математическое ожидание и дисперсию случайной величины $X = (X_1, X_2 \ldots X_m)$, проверяют различные статистические гипотезы. Знание характера распределения случайной величины позволяет исследователю корректно подходить к применению тех или иных методов МСА. В частности, корреляционный анализ применим лишь в том случае, если две случайные величины (Y и X) нормально распределены. При этом, если гипотеза о нормальности распределения не подтверждается, считается, что парный коэффициент корреляции не дает адекватную оценку степени тесноты связи. Более того, проверка значимости парных и частных коэффициентов корреляции осуществляется при тех же предпосылках нормальности их распределения. Подобные предположения используются и в методе канонических корреляций. В регрессионном анализе предположение о нормальности распределения выдвигается (и должно быть проверено!) при проверке значимости отдельных коэффициентов регрессии по t-критерию Стьюдента, а также при оценке коррелированности остатков по критерию Дарбина—Уотсона. Можно привести примеры и других видов распределений, встречающихся (и проверяемых) в ходе реализации методов многомерного анализа: в пошаговой регрессии при включении или исключении какого-либо факторного признака используются величины F-включения и F-исключения, имеющие распределение Фишера; при анализе таблиц сопряженности для проверки гипотезы о наличии связи между двумя признаками используется величина, имеющая χ^2-распределение; в факторном анализе используется предположение о нормальном распределении значений общих факторов (главных компонент) и т. д.

Распределения подразделяются на два больших класса: дискретные и непрерывные. Дискретные распределения отражают прерывность значений случайной величины, среди них наиболее известны биномиальное, гипергеометрическое распределения, распределение Пуассона.

Непрерывные распределения составляют более многочисленный по сравнению с дискретными класс распределений. Они представляют случайные величины, в значениях которых априори нет пропусков (разрывов). К непрерывным относит-

ся и нормальное распределение, которое широко известно и особенно часто встречается в теоретических разработках и аналитической практике. Доказано, что с помощью нормального распределения можно описать подавляющее большинство реально происходящих процессов. Кроме того, с увеличением числа наблюдений нередко другие виды распределений принимают вид нормального. С учетом своих свойств, хорошей разработанности и сравнительно простой формальной структуры нормальное распределение чаще других применяется в многомерной статистике. В классе непрерывных распределений, кроме нормального, известно большое число других видов распределений: гамма-распределение, распределение Стьюдента, распределения Вейбулла и Рэлея, распределение Парето и т. д.

В зависимости от того, по каким данным строятся распределения (наблюденным или вычисленным), их называют эмпирическими или теоретическими.

Важно отметить, что бывают одномерные и многомерные распределения. Многомерные распределения учитывают значения нескольких признаков одновременно и при числе аналитических признаков, превышающем 3, не поддаются графической интерпретации.

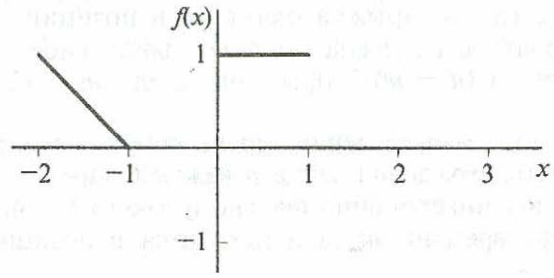
В этой главе автор рассматривает большое число разнообразных распределений, показывает особенности их практического приложения и взаимные связи. Эти распределения широко известны и успешно используются при решении многочисленных задач в экономике, социологии и т. д. Для исследователя выбор и область приложения законов распределения в анализе будут определяться прежде всего уровнем его профессиональной подготовки и возможностями располагаемых статистических пакетов программ для ЭВМ. Это объясняется трудоемкостью операций по вычислению функций и плотностей вероятностей статистических распределений. В настоящее время в специальных статистических пакетах программ для ЭВМ реализованы практически все наиболее распространенные семейства дискретных и непрерывных распределений.

Вопросы и задачи

1. Имеется функция плотности вероятностей:

$$f(x)=\begin{cases} -x-1 & -2\le x\le -1 \\ 1/2 & 0\le x\le 1 \\ 0 & \text{для других значений } x. \end{cases}$$

График функции $f(x)$ имеет вид:



Определите соответствующую $f(x)$ функцию распределения случайной величины.

2. Продолжительность работы элементов электрического аппарата часто описывается случайной величиной со следующей функцией плотности вероятностей:

$$f(x)=\begin{cases} \dfrac{m}{2x_0}x^{m-1}e^{\frac{-x^m}{2x_0}}, & x\ge 0 \\ 0, & x<0. \end{cases}$$

а) Найдите функцию распределения X;

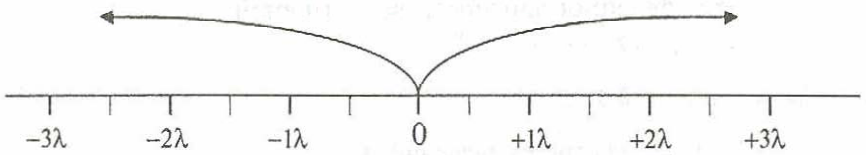
б) Определите показатели вероятности:

$P(x\le 0{,}75)$ и $P(0{,}25\le x\le 0{,}75)$.

В анализе используйте параметрические значения $x_0 = 1$ и $m = 2$.

3. Обратимся к примеру из жизни животных, скажем, зайцев. Пусть в начальное время (время-нуль) заяц находится в исходной позиции (позиции-нуль). В момент времени kt, т.е. по прошествии определенного времени начиная от времени-

нуль, заяц прыгает на расстояние, измеряемое в λ-единицах, вперед или назад с одной и той же вероятностью, равной 1/2:



Пусть при этом m и n — целые числа. Какова будет в этом случае вероятность того, что заяц через какое-то время (nt) после своего последнего прыжка окажется в позиции $m\lambda$? Очевидно, эта вероятность должна учитывать возможность $(m+n)/2$ прыжков вперед и $(n-m)/2$ прыжков назад: $(m+n)/2-(n-m)/2=m$.

Покажите, что распределение вероятностей для определения случайного местонахождения зайца в каждый определенный момент времени подчиняется биномиальному закону и при этом:

$$P\{\text{в момент времени } nt \text{ заяц находится в позиции } m\lambda\} = \left(\frac{n}{\frac{n+m}{2}}\right)\frac{1}{2^n}=\left(\frac{n}{\frac{n-m}{2}}\right)\frac{1}{2^n}.$$

4. Урезание распределения. Пусть случайная величина X имеет любое распределение на интервале $[a_1, b_1]$ и A — такое вероятностное событие, при котором X получает на отрезке интервала $[a_2, b_2]$ ограничение $a_1 \le a_2$, $b_2 \le b_1$, например, для нормального распределения:

$N(0,\sigma^2)$ $a_1=-\infty$, $b_1=+\infty$ и $a_2=-\sigma$, $b_2=+\sigma$.

Покажите, что если $F(x)$ есть функция распределения на интервале $[a_1, b_1]$, то $F(x|A)$ — функция распределения на интервале $[a_2, b_2]$ и

$$F(x|A)=\frac{F(x)-F(a_2)}{F(b_2)-F(a_2)}.$$

5. Пусть величина U имеет нормальное распределение $N(\mu,\sigma^2)$. Покажите, что при этом действует равенство:

$$E(e^{a\mu})=e^{a\mu}+\frac{1}{2}\sigma^2.$$

Этот случай имеет в эконометрике весьма важное значение, например, при оценивании логлинейных уравнений. Он также демонстрирует действие простого правила линейной трансформации.

6. Коэффициенты вариации $v(x)$ имеют большое прикладное значение в теории надежности, где в зависимости от значений, которые они приобретают (большие, малые или равные единице), производится классификация распределений случайных величин. Важнейшим с точки зрения теории при этом представляется распределение с $v(x)=1$, например, экспоненциальное или логнормальное с параметром $(\sigma^2=\ln 2)$ распределения.

Покажите, что Гамма-распределение и распределение Вейбулла также имеют $v(x)=1$.

7. Пусть известны три координатные точки функции плотности вероятностей:

$$P[X=n/3]=1/n$$

$$P[X=n^2/3]=1/n^2$$

$$P[X=n^2/3(n^2-n-1)]=(n^2-n-1)/n^2.$$

Покажите, что для любого n математическое ожидание: $E(x)=1$, а дисперсия при $n\to\infty$ не существует.

8. Имеется дискретная случайная величина:

$$P\left[X_n=2^{n-1}\right]=\left(\frac{1}{2}\right)^n,\quad n=1,2,...$$

а) Покажите, что X — случайная величина.

б) Найдите математическое ожидание для заданной случайной величины P.

Замечание. Проблема, с которой сталкиваются при решении данной задачи, носит название *санкт-петербургского парадокса.*

9. Имеется следующая таблица с характеристиками контингенций:

X \ Y	0	1	2	3	Σ
1	2q	0	0	q	3q
2	6q	6q	6q	0	18q
3	0	6q	0	0	6q
Σ	8q	12q	6q	q	27q

$q=1/27$.

а) Определите уровень ковариации переменных X и Y.

б) Являются ли X и Y независимыми переменными?

10. Имеется функция распределения вида:

$F(x,y)=y^{a_1}x^{a_2}$, $(x,y)\in\{[0,1]\times[0,1]\}$.

При каких значениях параметров a_1 и a_2 F становится функцией распределения над квадратом с единичной стороной?

11. Покажите, что функция $f(x,y)=3y$, $0<x<y<1$ есть функция плотности вероятностей.

Сделайте обобщения для случая с функцией $f(x,y)=\alpha y$, $0<x<y<\beta$.

Как взаимосвязаны константные величины α и β?

12. Имеется функция:

$$f(x,y)=\frac{(1-\rho)\rho^{Y}\mu e^{-\mu x}(\mu x)^{Y}}{Y!}$$

где $0\le x$; $y=0,1,2,\ldots$; $0<\rho<1$; $\mu>0$.

Покажите, что данная функция есть функция плотности вероятностей и что условные (граничные) плотности вероятностей имеют геометрическое (с параметром $\rho=\lambda/\mu$, т.е. величиной, кратной математическому ожиданию случайной величины X) и экспоненциальное (с параметром μ) распределения.

Указание: при решении задачи следует обратить внимание, что функция плотности вероятностей $f(x,y)$ включает одну непрерывную, а другую дискретную переменные. Используйте

расходящийся ряд для e^z и произведите многократное интегрирование этого ряда с частной производной $u'=e^{-\mu x}$.

Приведенная в задаче функция плотности вероятностей играет важную роль в теории очередей. При этом для случайных величин предполагаются следующие эмпирические значения: X — время ожидания в очереди (системе), Y — число человек (покупателей) в очереди (системе).

13. Имеется функция плотности вероятностей:

$f(x,y)=xe^{-x(1+y)}$, $0<x$, $0<y$.

Покажите, что в данном случае получим раздельные (граничные) функции:

плотности вероятностей:

$f(x)=e^{-x}$ и $g(y)=1/(1+y)^2$,

распределения:

$F(x)=1-e^{-x}$ и $G(y)=y/(1+y)$.

14. Имеются следующие функции плотности вероятностей:

$f_1(x,y)=4xy$ $\quad (x,y)\in([0,1]\times[0,1])$,

$f_2(x,y)=\lambda_1\lambda_2 e^{-\lambda_1 x-\lambda_2 y}$ $\quad (x,y)\in([0,\infty)\times[0,\infty))$,

$f_3(x,y)=\frac{6}{7}(x^2+\frac{xy}{2})$ $\quad (x,y)\in\{(0,1)\times(0,2)\}$.

Соответственно приведенным функциям плотности вероятностей определите совместные и раздельные функции распределений.

15. Данная задача наглядно показывает, что даже когда характеристика связи $R^2=0$, между переменными X и Y может существовать взаимосвязь, другими словами, что малые уровни показателя определенности (R^2) еще ничего не говорят о фактической взаимосвязи признаков вида: $P[Y=X^2]=1$.

Пусть имеются данные наблюдений, описываемые параболой:

$\{X_t\}=\{-4,-3,-2,-1,0,1,2,3,4\}$,

$\{Y_t\}=\{16,9,4,1,0,1,4,9,16\}$.

Следует показать, что при этом $R^2=0$.

16. Случайные величины X и Y имеют совместную функцию плотности вероятности:

$$f(x,y)=\begin{cases}\frac{1}{16}x, & y=1\\ \frac{2}{16}x, & y=2\\ \frac{3}{16}x, & y=3\\ 0, & \text{для других значений } y\end{cases}, \qquad x\in\{1,2\}.$$

а) Постройте таблицу (таблицу контингенций) для определенных значений функции f.

б) Ответьте на вопрос о правомерности равенства:
$f(1,1)=f_x(1)\cdot f_y(1)$.

в) Являются ли переменные X и Y независимыми?

г) Вычислите совместное математическое ожидание $E(X, Y)$.

17. Покажите, что для величины ρ из равенства:

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot\exp\left[-\frac{1}{2}\left(\frac{1}{\sqrt{1-\rho^2}}\right)\cdot\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-\right.\right.$$

$$\left.\left.-2\rho\frac{x_1-\mu_1}{\sigma_1}\cdot\frac{x_2-\mu_2}{\sigma_2}+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right]$$

имеет силу $\rho=\dfrac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$.

18. Для дискуссионного обсуждения предлагается вопрос о форме статистического распределения, представляющего смесь двух раздельных нормальных распределений:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}\left[0{,}5\times\exp\left[-\frac{(x+\mu)^2}{2\sigma^2}\right]+0{,}5\times\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\right].$$

Следует найти также математическое ожидание и дисперсию случайной величины X.

Решение: В данном случае математическое ожидание будет равно нулю, а дисперсия — единице. Такой результат становится очевидным на рис. 3.28:

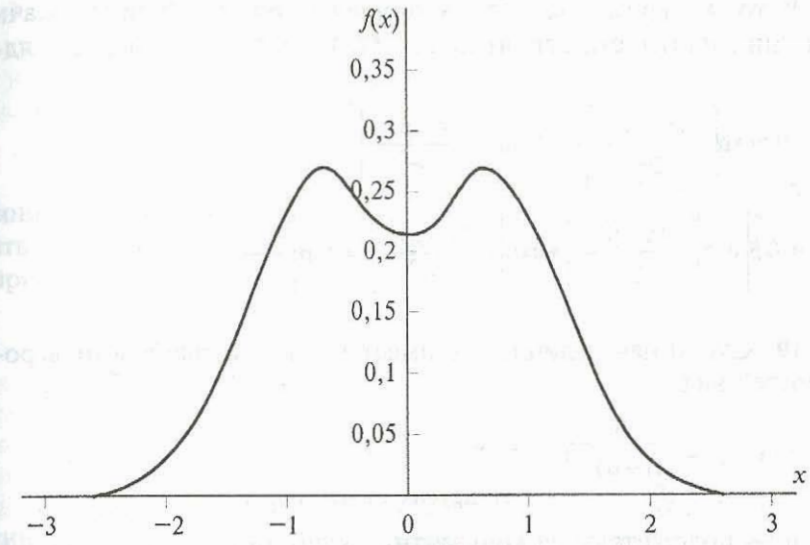


Рис. 3.28

Функцию $f(x)$ нельзя считать простым результатом объединения двух независимых плотностей вероятностей. Для ядра (центра распределения) имеем:

$$\exp\left[-\frac{(x+\mu)^2}{2\sigma^2}\right] \text{ и } \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

и соответственно подходящими функциями здесь могли бы быть:

$$g(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x+\mu)^2}{2\sigma^2}\right], \qquad X\sim N(-\mu,\sigma^2),$$

$$h(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad X\sim N(\mu,\sigma^2),$$

т.е. сумма принадлежащих ядру плотностей вероятностей $Z\sim N(0,2\sigma^2)$ должна была бы выражаться следующей функцией плотности вероятностей:

$$f^*(x)=\frac{1}{\sqrt{4\pi\sigma^2}}\exp\left[-\frac{x^2}{4\sigma^2}\right].$$

В то же время исходя из приведенной в условии задачи функции плотности вероятностей $f(x)$ мы получим область ядра:

$$\left[0{,}5\exp\left[-\frac{(x+\mu)^2}{2\sigma^2}\right]+0{,}5\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\right]=$$

$$=0{,}5\left[\exp\left[-\frac{(x+\mu)^2}{2\sigma^2}\right]+\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\right]\neq\exp\left[-\frac{x^2}{4\sigma^2}\right].$$

19. Случайная величина X имеет функцию плотности вероятностей вида:

$$P(X=x)=\begin{cases}\dfrac{\mu^x}{(1+\mu)^{x+1}}, & x=0,1,2,\ldots\\ 0, & \text{при других значениях } x,\end{cases}$$

где μ — положительная константная величина.

При условии, что имеются данные выборочного наблюдения (2, 3, 0, 1), найдите методом максимального правдоподобия оценку для μ.

20. Случайная величина X имеет следующую функцию плотности вероятностей:

$$f(x)=\begin{cases}2bx\mathrm{e}^{-bx^2}, & x\geq 0\\ 0, & \text{при других значениях } x.\end{cases}$$

Выборочной совокупностью данных с числом наблюдений $n=4$ и плотностью вероятностей, функционально описываемой $f(x)$, задаются следующие значения X:

$$(x_1,x_2,x_3,x_4)=\left(2,\frac{3}{2},\frac{4}{3},1\right).$$

Методом максимального правдоподобия произведите оценивание параметра b.

21. Произведите для случайной величины X, принимающей два значения и имеющей соответственно две различные плотности вероятностей, оценивание математического ожидания и дисперсии:

$X=0$ и $X=T$

$$f(x)=1-\frac{1}{T};\quad f(x)=\frac{1}{T}.$$

(*Решение*: $E(X)=1$, $\mathrm{var}(X)=T-1$)

22. Пусть случайная величина X имеет распределение Пуассона с параметром λ. По данным выборки тогда из всего множества различных оценок (статистик) λ можно по крайней мере привести следующие три:

$$f_1(X):=\frac{X_1}{4}+\frac{X_2}{2}+\frac{X_3}{4},$$

$$f_2(X):=\frac{X_1}{8}+\frac{X_2}{8}+\frac{X_3}{4},$$

$$f_3(X):=\frac{X_1}{3}+\frac{X_2}{3}+\frac{X_3}{3}.$$

Обсудите, попытайтесь доказать несмещенность и эффективность трех приведенных оценок (f_1, f_2, f_3). Почему вообще в данном случае употребимо понятие наилучших оценок?

Глава 4

Проверка статистических гипотез

4.1. Статистические гипотезы в анализе данных

Статистические гипотезы служат инструментом проверки выдвигаемых теоретических предположений. Предположения могут быть сделаны относительно параметров статистического распределения (в случае нормального распределения — математического ожидания μ или дисперсии σ^2), тогда гипотезу называют *параметрической*, или относительно распределения случайной величины (подчинение ее нормальному закону распределения, закону Пуассона и т. д.) — в этом случае проверяемую гипотезу называют *непараметрической*.

Проверка статистической гипотезы предполагает наличие выборочной совокупности данных, которая параметрически (своей функцией распределения) сравнивается с генеральной совокупностью или другой выборкой. Скажем, производится сравнение среднего уровня урожайности сельскохозяйственных культур, среднего уровня заболеваемости населения и т. д. какой-либо из областей республики (выборочная совокупность) со средним уровнем по всей республике (генеральная совокупность).

При проверке статистических гипотез используется понятие нулевой (прямой) и альтернативной (обратной) гипотез. *Прямая гипотеза* (H_0) является основной и обычно содержит утверждение об отсутствии различий между сравниваемыми величинами. *Альтернативная гипотеза* (H_1) представляется конкурирующей по отношению к нулевой и принимается после того, как отвергнута основная. Приведем примеры статистических гипотез относительно параметров нормально распределенной одномерной и многомерной случайных величин:

	Нулевые гипотезы	*Альтернативные гипотезы*
Одномерная случайная величина	$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$, или $\mu < \mu_0$, $\mu > \mu_0$
	$H_0 : \sigma^2 = \sigma_0^2$	$H_1 : \sigma \neq \sigma_0$, или $\sigma^2 < \sigma_0^2$, $\sigma^2 > \sigma_0^2$
Многомерная случайная величина	$H_0 : \mu_j = \mu_{0j}$	$H_1 : \mu_j = \mu_{0j}$
	$H_0 : \Sigma = \Sigma_0$	$H_1 : \Sigma = \Sigma_0$

В зависимости от выдвигаемых предположений параметрические гипотезы подразделяют на простые и сложные. *Простая гипотеза* содержит только одно предположение относительно оцениваемого параметра: $H_0 : a = 0$, или $H_0 : a = 1$ и т. д. *Сложная гипотеза* состоит из конечного или бесконечного числа простых гипотез. Например, $H_0 : a > 4$ означает, что могут быть $H_0 : a = 5$; $H_0 : a = 5{,}5$ и т. п., т.е. здесь гипотеза состоит из бесконечного набора гипотез вида $H_0 : a = a_i$, где $i = \overline{4{,}00 \ldots 01}; \infty$.

Суждения об истинности или ложности статистической гипотезы строятся на основе критериальной (тестовой) проверки. Существуют статистические критерии, отражающие результаты сравнений и принимающие наблюденные значения по выборочным данным, а также критерии критических значений, установленные теоретическим путем. В ходе сопоставления критериальных величин выясняется: можно принимать или следует отвергнуть нулевую гипотезу. Если наблюденное значение критерия не превышает критического, то по крайней мере теоретически отсутствуют основания, чтобы отвергнуть прямую (нулевую) гипотезу. В противоположном случае целесообразно предположить справедливость альтернативной гипотезы H_1.

Критериальная проверка статистических гипотез допускает определенную вероятность ошибки в выводах. При этом разделяют вероятность ошибки первого рода (α) — отвергнуть нулевую гипотезу, когда она справедлива, и второго рода (β) — принять нулевую гипотезу, когда она ложна. Графически вероятностные оценки ошибки представляют некоторые области плотности распределения значений статистического критерия. Область с малой вероятностью попадания критериальных значений α характеризуется как критическая (пороговая), а область,

остающаяся за вычетом критической, т.е. $1-\alpha$ — как область допустимых значений[1] (рис. 4.1):

Рис. 4.1

В экономических исследованиях из двух вероятностных характеристик ошибок, которые допускаются при проверке гипотез (α и β), обычно используется α — вероятность ошибки первого рода. Наиболее распространенными в практике значениями α являются: 0,01; 0,05; 0,1, что соответственно указывает на вероятность получения достоверного вывода ($1-\alpha$), равную: 0,99; 0,95; 0,90.

Критериальная проверка гипотезы в сущности означает сравнительную характеристику параметрических значений или самого распределения случайной величины. Так, значения нормально распределенной величины, расположенные в непосредственной близости от центра распределения (математического ожидания μ) и априорно имеющие более высокую степень (вероятность) принадлежности изучаемой совокупности данных, соответственно будут принимать критериальные значения, подтверждающие несущественность их отклонений от μ.

[1] Статистический критерий не обязательно имеет нормальное распределение.

Схематично графически все множество значений нормально распределенной случайной величины можно показать следующим образом (рис. 4.2):

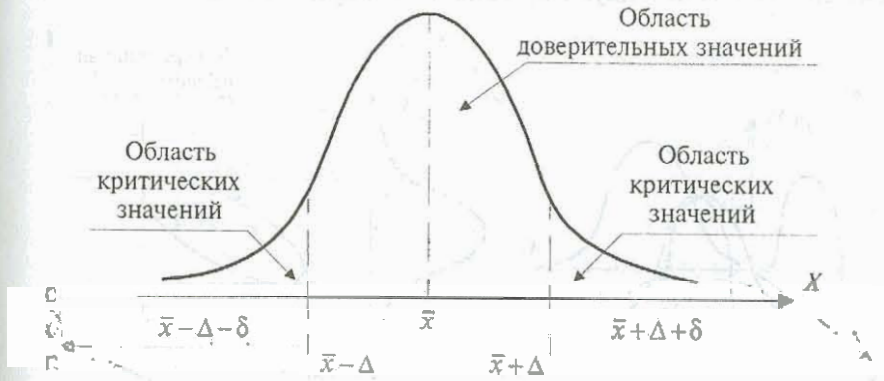


Рис. 4.2

Если мы имеем дело с многомерной случайной величиной, то доверительная область будет представлять собой зону пересечения доверительных областей всех составляющих ее одномерных величин. Это хорошо видно на примере двумерной случайной величины с нормальным распределением (рис. 4.3):

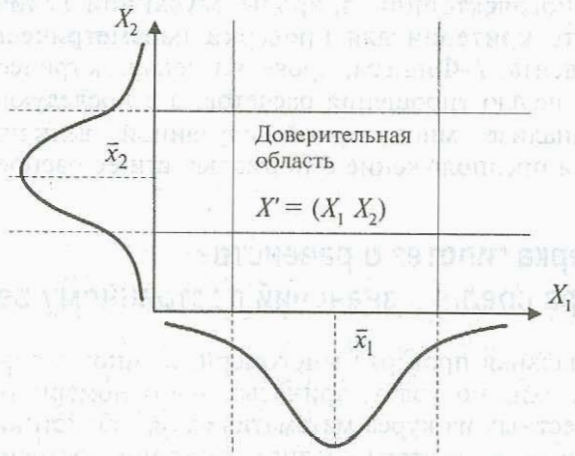


Рис. 4.3

Заметим, что приведенный выше рисунок является упрощенным. На самом деле область двумерной величины будет принимать форму не простого прямоугольника, а эллипса — результат наложения двух нормальных распределений:

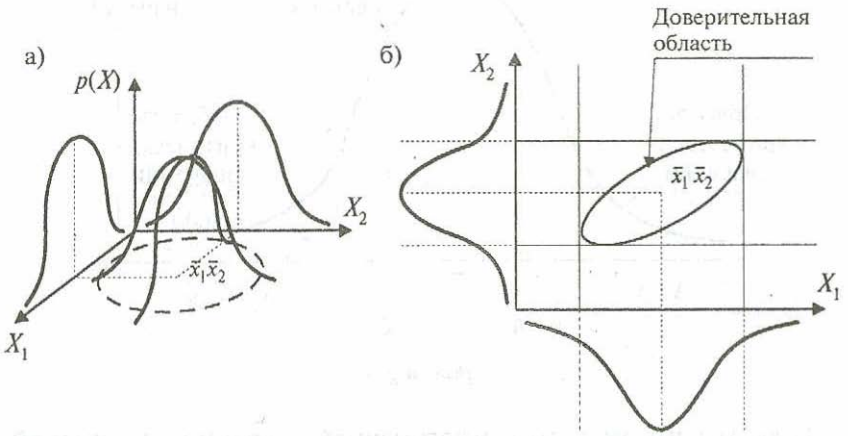


Рис. 4.4

В многомерном случае используются те же статистические критерии соответствия, что и в одномерном, но они изменяются с учетом многовекторной природы случайной величины. Чаще всего это критерии для проверки параметрических гипотез: t-Стьюдента, F-Фишера, проверки непараметрических гипотез — χ^2. С целью упрощения расчетов, а в последующем и выводов в анализе многомерной случайной величины часто принимается предположение о нормальности ее распределения.

4.2. Проверка гипотез о равенстве вектора средних значений постоянному вектору

Критериальная проверка многомерных гипотез основывается на теоретических подходах, принятых для одномерного случая и хорошо известных из курса математической статистики. Так, гипотеза о равенстве вектора средних значений постоянному вектору, когда число анализируемых признаков $m = 1$ и значения

случайной величины распределены нормально, оценивается по t-критерию Стьюдента:

$$t_н = \frac{\bar{x}-\mu}{s}\sqrt{n},$$

где $\bar{x}$ — среднее значение случайной величины $X' = (x_1, x_2, \ldots x_n)$;

μ — математическое ожидание (некоторое стандартное или заданное значение);

s — среднее квадратическое отклонение, оцененное по выборочным данным X;

n — объем выборочной совокупности, т.е. данных, участв проверке.

Наблюденное значение t-критерия сравнивается с критическим $t_{кр}$ при заданном $\alpha/2$ или α — уровне значимости и числе степеней свободы $\nu = n-1$. Гипотеза о равенстве $\bar{x} = \mu$ ($H_0 : \bar{x} = \mu$) подтверждается, когда $t_н < t_{кр}$, в других случаях допускается существенность различий $\bar{x}$ и μ.

Пример 4.1. На склад магазина поступила большая партия апельсинового сока. По стандарту содержание натурального сока в упаковке, рассчитанной на 1 л продукции, должно составлять 80%, остальное — консерванты и пищевые добавки. При проверке 49 упаковок оказалось, что средний процент содержания натурального сока фактически составляет 75% при среднем квадратическом отклонении $s = 4\%$. Проверим гипотезу на соответствие полученной продукции стандарту качества.

Запишем прямую и альтернативную гипотезы по условию задачи:

$H_0 : \bar{x} = 80$; $H_1 : \bar{x} \neq 80$.

Рассчитаем значение t-статистики:

$$t_н = \frac{\bar{x}-\mu}{s}\sqrt{n} = \frac{75-80}{4}\cdot\sqrt{49} = -8{,}75.$$

При $\alpha = 0{,}10$ по таблицам t-критерия Стьюдента найдем критическое значение: $t_{0{,}05;\ 48} = 1{,}678$, т. е. $t_н > t_{кр}$. Следовательно, мы не можем считать партию сока отвечающей норме. С вероятностью допустить ошибку в выводах равной 0,05 следует признать существенность различий качества сока, отвечающего стандарту и попавшего в выборку.

С учетом значений $t_{кр}$ могут быть рассчитаны доверительные интервалы для $\bar{x}$: $\bar{x} \pm s/\sqrt{n} \cdot t_{\alpha/2,n-1}$.

Для нашего примера

$$\frac{s}{\sqrt{n}} \times t_{\alpha/2,n-1} = \frac{4}{\sqrt{49}} \times 1{,}678 = 0{,}960 \text{ и}$$

$75 - 0{,}96 \le \bar{x} \le 75 + 0{,}96$, т.е. $74{,}04\% \le \bar{x} \le 75{,}96\%$.

Таким образом, доверительные интервалы для значений интересующего нас признака X, определяющие границы колеблемости $\bar{x}$ в генеральной совокупности, не покрывают заданного значения ($\mu = 80\%$). Подтверждается ранее сделанный вывод о несоответствии уровня качества сока в поставленной партии стандарту.

В многомерном случае имеем дело уже с m — числом выборочных средних, т.е. вектором средних значений: $\bar{X}' = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)$. Вектор $\bar{X}$ сравнивается с постоянным вектором $\mu' = (\mu_1, \mu_2, \ldots, \mu_m)$. Прямая гипотеза имеет вид $H_0: \bar{X}' = \mu'$ при альтернативной $H_1: \bar{X}' \neq \mu'$.

Согласно двум приведенным гипотезам все $\bar{x}_j$ соответствуют или существенно отличаются от μ_j.

При построении критерия для проверки многомерной гипотезы воспользуемся известной формулой t-статистики:

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n}.$$

Возведем в квадрат правую и левую части равенства:

$$t^2 = n(\bar{x} - \mu)'(s^2)^{-1}(\bar{x} - \mu).$$

Воспользуемся характеристиками многомерной случайной величины:

- вместо значения $(\bar{x} - \mu)$ примем вектор $(\bar{X} - \mu)$, где $\bar{X}$ и μ — в свою очередь векторы средних и постоянных значений;
- вместо s^2 — ковариационную матрицу Σ и получим:

$$T^2_{\text{н}} = n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu),$$

где $\Sigma = \frac{1}{n-1}(K'K)$ — ковариационная матрица и K — матрица с центрированными данными: $k_{ij} = \|x_{ij} - \bar{x}_{ij}\|$.

В анализе обычно ковариационная матрица по генеральной совокупности данных неизвестна и вместо нее используют ковариационную матрицу по выборочной совокупности данных (S);

тогда $T^2_{\text{н}} = n(\bar{X} - \mu)' S^{-1}(\bar{X} -$ T^2-критерий известен как критерий Хотеллинга.

Наблюденное значение $T^2_{\text{н}}$ сопоставляется с критическим, исчисляемым при заранее заданном уровне вероятности α — допустить ошибку в выводах и числе степеней свободы $\nu_1 = m$ и $\nu_2 = n - m$:

$$T^2_{\alpha,m,n-m} = \frac{m(n-1)}{n-m} F_{\alpha,m,n-m}.$$

В формуле $F_{\alpha,\ m,\ n-m}$ — табличное значение F-критерия Фишера для известных $\nu_1 = m$ и $\nu_2 = n - m$. Многомерная гипотеза подтверждается при $T^2_{\text{н}} < T^2_{\alpha,m,n-m}$ и не может быть принята, если $T^2_{\text{н}} > T^2_{\alpha,m,n-m}$.

Приведенная выше формула T^2-критерия Хотеллинга является общей и рассчитана на проверку гипотезы сразу по всему числу m анализируемых признаков. Однако реально, даже отрицании гипотезы $H_0: \bar{X}_j = \mu_j$, значения одних признаков могут существенно отличаться от некоторых постоянных значений, а другие — несущественно. Возникает необходимость проверки гипотезы по каждому отдельному признаку или нескольким признакам ($k < m$) при условии нивелирования значений остальных признаков, представляющих одно и то же явление или процесс. Для решения подобной задачи можно использовать частный критерий Хотеллинга T^2_j, наблюденное значение которого оценивается по формуле:

$$T^2_{\text{нj}} = \frac{n(c'_j(\bar{x}_j - \mu_j))^2}{c'_j S c_j},$$

где c_j — специальный вектор, нивелирующий значения всех признаков, кроме одного или нескольких, участвующих в проверке статистической гипотезы. Компоненты вектора c_j — нули и единицы, единицы указывают на признак или признаки, по значениям которых осуществляется проверка гипотезы. Скажем, анализируются данные по четырем признакам, для проверки гипотезы используется один третий признак (X_3), тогда $c'_3 = (0\ 0\ 1\ 0)$.

Доверительные интервалы для многомерной случайной величины в общем определяются по значительно более сложным

алгоритмам, чем в случае проверки одномерной гипотезы. Здесь следует выделить три возможных различных подхода:

Первый подход применим при расчете только частного критерия Хотеллинга. Исследователь решает максимально упростить задачу расчета доверительных интервалов и абстрагируется от существования всех других признаков, кроме одного. Решение будет заведомо грубым, так как не учитывает ковариации анализируемых признаков и сводится к расчетам по известной простой формуле:

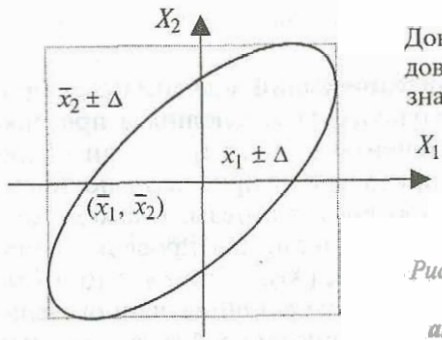
$$x \pm \frac{s}{\sqrt{n}} t_{\alpha/2,n-1}.$$

Второй подход применим при условии также выделения и анализа значений только одного из комплекса признаков, но при этом принимаются во внимание многомерность случайной величины и соответственно параметры многомерной статистической совокупности. Доверительные интервалы определяются здесь с учетом ограничивающих значений F-критерия Фишера:

$$c_j' \bar{x} \pm \left(\frac{m(n-1)}{n(n-m)} \times c_j' S c_j \times F_{\alpha,m,n-m} \right)^{1/2}.$$

Третий подход, когда определяются не отдельные интервалы, а доверительная область, охватывающая одновременно допустимые значения всех анализируемых признаков, представляющих некоторую многомерную случайную величину.

Выше (§1.3) показано, что такая доверительная область описывается эллипсоидом или при $m > 2$ — эллипсом. При числе анализируемых признаков, равном двум ($m = 2$), доверительную область можно легко показать графически (рис. 4.5).



Доверительная область и доверительные интервалы для значений каждого из признаков X_j

Рис. 4.5. Совместная доверительная область для значений двух аналитических признаков X_1 и X_2

В общем случае ($m > 2$) совместная доверительная область ограничивается поверхностью, задаваемой уравнением:

$$(\bar{X} - \mu)' \Sigma^{-1} (\bar{X} - \mu) = \frac{m(n-1)}{n(n-m)} F_{\alpha,m,n-m}.$$

Данное уравнение задает эллипсоид с центром $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)$. Нетрудно заметить, что при $m = 1$ и извлечении квадратного корня из левой и правой частей уравнения мы возвращаемся к простейшей формуле:

$$\bar{x} \pm \mu = \pm \frac{s}{\sqrt{n}} t_{\alpha/2,n-1}.$$

Решение уравнения с одновременным поиском доверительных интервалов для всех m анализируемых признаков сводится к последовательному определению собственных чисел, собственных векторов, интервальных величин и приведению их к натуральному масштабу.

Пример 4.2. Для предприятий, торгующих продуктами питания в административном районе, установлены нормативные экономические показатели эффективности деятельности: уровень рентабельности товарооборота — 20% и средняя оборачиваемость товарных запасов — 12 дн. Более низкие значения показателей рентабельности и скорости оборота запасов означают нарушение ритмичности товарно-денежных операций и опасное снижение конкурентоспособности предприятия.

С целью оперативного контроля результатов коммерческой деятельности в одной из торговых фирм района проведен анализ эффективности торговых операций за последние 10 месяцев и получены следующие данные:

Месяц	*Рентабельность товарооборота, %*	*Продолжительность оборота товарных запасов, дн.*
01	14	19
02	12	15
03	16	19
04	14	17
05	15	24
06	18	12
07	22	10
08	20	15
09	13	18
10	9	20
Среднее значение	15,3	16,9

Оценим существенность различий экономических показателей торговой фирмы и нормативных. Уровень α зададим равным 0,05.

Решение:

1. Определим параметры многомерной совокупности данных.

Вектор средних величин: $\overline{X'} = (15{,}3\ \ 16{,}9)$;

Ковариационная матрица:

$$S = \frac{1}{n-1}(K'K) =$$

$$= \frac{1}{9} \times \begin{pmatrix} 14-15{,}3 & 12-15{,}3 & 16-15{,}3 & 14-15{,}3 & 15-15{,}3 & \dots & 9-15{,}3 \\ 19-16{,}9 & 15-16{,}9 & 19-16{,}9 & 17-16{,}9 & 24-16{,}9 & \dots & 20-16{,}9 \end{pmatrix} \times$$

$$\times \begin{pmatrix} 14-15{,}3 & 19-16{,}9 \\ 12-15{,}3 & 15-16{,}9 \\ 16-15{,}3 & 19-16{,}9 \\ 14-15{,}3 & 17-16{,}9 \\ 15-15{,}3 & 24-16{,}9 \\ \vdots & \vdots \\ 9-15{,}3 & 20-16{,}9 \end{pmatrix} = \frac{1}{9} \begin{pmatrix} 134{,}1 & -87{,}7 \\ -87{,}7 & 148{,}9 \end{pmatrix} = \begin{pmatrix} 14{,}90 & -9{,}74 \\ -9{,}74 & 16{,}50 \end{pmatrix}.$$

Обратная ковариационная матрица будет:

$$S^{-1} = \frac{1}{|S|} adjS = \begin{pmatrix} 0{,}0987 & -0{,}0645 \\ -0{,}0645 & 0{,}1093 \end{pmatrix}.$$

2. Рассчитаем T^2-критерий Хотеллинга:

$$T^2_{\text{н}} = n(\overline{X} - \mu)' S^{-1} (\overline{X} - \mu) = 10 \times (15{,}3 - 20 \quad 16{,}9 - 12) \times$$

$$\times \begin{pmatrix} 0{,}0987 & -0{,}0645 \\ -0{,}0645 & 0{,}1093 \end{pmatrix} \begin{pmatrix} 15{,}3 & -20 \\ 16{,}9 & -12 \end{pmatrix} = 28{,}1.$$

3. Найдем критическое значение T^2-критерия:

$$T^2_{\text{кр}} = \frac{m(n-1)}{n-m} F_{0{,}05;2;8} = \frac{2(10-1)}{10-2} \times 4{,}459 = 10{,}03.$$

Как видим, наблюденное значение T^2-критерия значительно больше критического (28,1 > 10,0). Следует сделать вывод о существенности различий фактических значений экономических показателей, оцененных по выборочной совокупности предприятий, и значений, принятых за нормативные.

Теперь посмотрим, величиной какого из признаков определяется существенность различий. Рассчитаем частные значения T^2-критерия Хотеллинга:

$$T^2_{\text{н}j} = \frac{n(c'_j(\overline{x}_j - \mu))^2}{c'Sc}.$$

Для значений первого признака — рентабельности товарооборота — получим

$$T^2_{\text{н}1} = \frac{10 \cdot \left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} -4{,}8 \\ 4{,}9 \end{pmatrix} \right)^2}{\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 14.90 & -9{,}74 \\ -9{,}74 & 16{,}50 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}} = \frac{230{,}4}{14{,}9} = 15{,}46.$$

Для значений второго признака — продолжительности оборота товарных запасов — частный критерий Хотеллинга:

$$T^2_{\text{н}2} = \frac{10 \left(\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} -4{,}8 \\ 4{,}9 \end{pmatrix} \right)^2}{\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 14.90 & -9{,}74 \\ -9{,}74 & 16{,}50 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}} = \frac{240{,}1}{16{,}5} = 14{,}55.$$

По значениям обоих оцененных признаков торговая фирма не может быть отнесена к числу устойчиво работающих предприятий.

4.3. Проверка гипотез о равенстве двух векторов средних значений

Для одномерной нормально распределенной совокупности данных проверка гипотезы о равенстве двух средних величин: $H_0 : \bar{x}_1 = \bar{x}_2$ осуществляется с использованием известных t-распределения и t-критерия Стьюдента. Наблюденное значение $t_{\text{н}}$ исчисляется по формуле:

$$t_{\text{н}} = \frac{(\bar{x}_1 - \bar{x}_2)\sqrt{n_1 n_2}}{s_* \sqrt{n_1 + n_2}},$$

где $\bar{x}_1$, $\bar{x}_2$ — средние для двух выборочных совокупностей значения анализируемого признака;

n_1, n_2 — объемы выборочных совокупностей;

s_* — корень квадратный из объединенной дисперсии двух выборочных совокупностей:

$$s_* = \sqrt{s_*^2} = \left[\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}\right]^{1/2}.$$

Наблюденное значение t-критерия сравнивается с критическим (табличным), которое определяется с $\nu = n_1+n_2-2$ — числом степеней свободы и при заданном уровне значимости α.

Доверительные интервалы для разности средних значений $\Delta\bar{x} = (\bar{x}_1 - \bar{x}_2)$ определяются с учетом величины объединенной дисперсии s_* и критического значения t-критерия ($t_{\alpha/2,\nu}$):

$$\bar{x} \pm s_{\Delta\bar{x}} \times t_{\alpha/2,\nu},$$

здесь $s_{\Delta\bar{x}}$ — корень квадратный из дисперсии разности двух выборочных средних значений ($\overline{x_1}$ — $\overline{x_2}$). Дисперсия разности средних исчисляется по формуле:

$$s_{\Delta\bar{x}}^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{n_1+n_2}{n_1 n_2} s_*^2.$$

Пример 4.3. В фермерском хозяйстве апробируется новый вид удобрения. Чтобы узнать его эффективность, под опытные посевы зерновых выделено 40 делянок. На первом опытном участке — 15 делянках — вносились старые удобрения и полученный урожай оценивался средней величиной 18 ц/га. На 25 делянках — втором участке — были внесены новые удобрения и средний уровень урожайности оценивался 20 ц/га. Средние квадратические отклонения в уровне урожайности по первому и второму опытным участкам соответственно составили 2 и 3 ц/га.

С вероятностью допустить ошибку в выводах 0,05 следует ответить на вопрос: случайно ли расхождение в уровне урожайности на опытных участках и действительно ли новые удобрения существенным образом повышают урожайность зерновых?

Решение:

1. Исчислим объединенную дисперсию по данным двух опытных участков с посевами зерновых:

$$s_*^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)} = \frac{(15-1)4+(25-1)\times 9}{(15-1)+(25-1)} = 7{,}158,$$

т. е. $s_* = \sqrt{7{,}158} = 2{,}675$.

2. Определим наблюденное значение t-критерия:

$$t_н = \frac{(\bar{x}_1-\bar{x}_2)\sqrt{n_1 n_2}}{s_*\sqrt{n_1+n_2}} = \frac{(18-20)\sqrt{15\times 25}}{2{,}675\sqrt{15+25}} = -2{,}287.$$

Наблюденное значение t-критерия по абсолютной величине ($t_н$ =2,287) превышает t критическое ($t_{0,05;38}$ = 1,687). Нулевую гипотезу о равенстве средних значений $\overline{x_1}$ и $\overline{x_2}$ следует отвергнуть, тем самым признаем, что уровни урожайности зерновых культур на двух опытных участках различаются и зависят от вида вносимых удобрений.

3. Найдем доверительный интервал для разности средних значений ($\Delta\bar{x}$):

$$\bar{x} \pm s_{\Delta\bar{x}} \times t_{\alpha/2;\nu};\ \bar{x} \pm \left(\frac{15+25}{15\times 25}7{,}158\right)^{1/2} \times 2{,}026 = 1{,}773,$$

или $-3{,}773 \le \bar{x} \le -0{,}227$.

Предположительно для генеральной совокупности отклонение двух одинаковых по величине средних значений будет равным нулю. Нуль не попадает в интервал для $\bar{x}$ (−3,773 ÷ −0,227), т.е. $\Delta\bar{x} \neq \Delta\mu$ и вывод о существенности различий $\overline{x_1}$ и $\overline{x_2}$ подтверждается.

В многомерном статистическом анализе проверяется гипотеза о равенстве векторов средних значений:

$$H_0: (\bar{X}_{11}\bar{X}_{12}\bar{X}_{13}...\bar{X}_{1m}) = \bar{X}_{21}\bar{X}_{22}\bar{X}_{23}...\bar{X}_{2m},$$

$$H_1: (\bar{X}_{11}\bar{X}_{12}\bar{X}_{13}...\bar{X}_{1m}) \neq \bar{X}_{21}\bar{X}_{22}\bar{X}_{23}...\bar{X}_{2m},$$

или в векторной форме:

$$H_0: \bar{X}_1 = \bar{X}_2;\quad H_1: \bar{X}_1 \neq \bar{X}_2.$$

Построим многомерный T^2-критерий, используя в качестве исходной соответствующую формулу одномерного t-критерия:

$$t_н = \frac{(\bar{x}_1-\bar{x}_2)\sqrt{n_1 n_2}}{s_*\sqrt{n_1+n_2}}.$$

Возведем в квадрат обе части равенства и перегруппируем элементы в его правой части, получим

$$t_н^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1-\bar{x}_2)\left(s_*^2\right)^{-1}(\bar{x}_1-\bar{x}_2).$$

Последняя формула может служить для расчета многомерного T^2-критерия, если в нее ввести многомерные параметры:

$$T_{\text{н}}^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{X}_1 - \bar{X}_2)' S_*^{-1}(\bar{X}_1 - \bar{X}_2),$$

здесь $\bar{X}_1$, $\bar{X}_2$ — векторы средних значений;

S_*^{-1} — матрица, обратная объединенной ковариационной матрице: $S_* = \frac{1}{n_1 + n_2 - 2}(K_1'K_1 + K_2'K_2)$ по выборочной совокупности данных, где K — матрица центрированных значений с элементами $k_{ij} = \left\| x_{ij} - \bar{x}_j \right\|$.

Критические значения для T^2 находятся по специальной формуле, включающей F-критерий Фишера, определяемый при заданном уровне значимости α и с числом степеней свободы $\nu_1 = m$, $\nu_2 = n_1 + n_2 - m - 1$:

$$T_{\alpha,m,n_1+n_2-m-1}^2 = \frac{(n_1 + n_2 - 2)m}{n_1 + n_2 - m - 1} \times F_{\alpha,m,n_1+n_2-m-1}.$$

При $T_{\text{н}}^2 < T_{\alpha,m,n_1+n_2-m-1}^2$ нулевая гипотеза H_0: $\bar{X}_1 = \bar{X}_2$ принимается, и вывод о равенстве векторов средних значений следует с вероятностью $(1 - \alpha)$. Если же $T_{\text{н}}^2 > T_{\alpha,m,n_1+n_2-m-1}^2$, то гипотеза о равенстве векторов средних значений не может считаться достоверной и отвергается.

При этом также существует возможность расчета частных критериев T_j^2 для сравнений одного или нескольких средних значений из каждой выборочной совокупности:

$$T_{\text{н}j}^2 = \frac{n_1 n_2 (c_j'(\bar{X}_1 - \bar{X}_2))^2}{(n_1 + n_2) c_j' S_* c_j},$$

где c_j — вектор, нивелирующий средние значения, не участвующие в сравнении, $1 \le j \le m$.

Для частных оценок различий средних значений критические величины определяются формулой:

$$T_{\alpha,j,n_1+n_2-j-1}^2 = \frac{(n_1 + n_2 - 2)j}{n_1 + n_2 - j - 1} \times F_{\alpha,j,n_1+n_2-j-1}.$$

Доверительная область для векторных разностей $(\Delta\bar{X} = \bar{X}_1 - \bar{X}_2)$ задается уравнением эллипсоида:

$$(\Delta\bar{X} - \Delta\mu)' S_*^{-1}(\Delta\bar{X} - \Delta\mu) = \frac{n_1 + n_2}{n_1 n_2} \times \frac{(n_1 + n_2 - 2)m}{n_1 + n_2 - m - 1} \times F_{\alpha,m,n_1+n_2-m-1},$$

где $(\Delta\bar{X} - \Delta\mu)$ — вектор, представляющий разности отклонений средних значений по выборочной $(\Delta\bar{X})$ и генеральной $(\Delta\mu)$ совокупностям. Подтверждение гипотезы H_0: $\bar{X}_1 = \bar{X}_2$, или H_0: $\Delta X = 0$ в сущности означает признание правомерности другой гипотезы H_0: $\Delta\mu = 0$.

Когда число признаков, участвующих в анализе, ограничивается, $1 \le j < m$, и используется критерий T^2 частного вида, доверительная область определяется уравнением, содержащим вектор c, нивелирующий по выбору исследователя значения отдельных признаков:

$$(c'\Delta\bar{X} - c'\Delta\mu)'(c'S_*c)^{-1}(c'\Delta\bar{X} - c'\Delta\mu) = \frac{n_1 + n_2}{n_1 n_2} \times \frac{(n_1 + n_2 - 2)j}{n_1 + n_2 - j - 1} \times F_{\alpha,j,n_1+n_2-j-1}.$$

Пример 4.4. С целью оценки воздействия состояния окружающей среды на здоровье населения обследованы два административных района. В первом районе, с низким уровнем техногенной

Населенный пункт	*Первый район*		*Населенный пункт*	*Второй район*	
	Младенческая смертность (X_1)	*Заболеваемость злокачественными новообразованиями (X_2)*		*Младенческая смертность (X_1)*	*Заболеваемость злокачественными новообразованиями (X_2)*
1	8	206	1	15	215
2	11	210	2	16	212
3	12	212	3	14	214
4	10	216	4	18	225
5	9	184	5	22	230
6	14	201	6	12	207
7	6	165	7	17	256
8	12	195	8	14	236
			9	20	302
			10	24	220
			11	10	214
			12	18	198
$\bar{X}$	10,3	198,6	$\bar{X}$	16,8	225,2

нагрузки, проверено 8 крупных населенных пунктов, во втором, имеющем крупные химические и нефтехимические предприятия и соответственно высокий уровень техногенной нагрузки — 12 пунктов. По данным обследований населенных пунктов, приведенным ниже, следует определить при $\alpha = 0{,}01$ существенность различий двух районов по X_1 — уровню младенческой смертности (‰) и X_2 — уровню заболеваемости населения злокачественными новообразованиями (на 100 000 чел. населения):

Решение:

1. Определим исходные векторы и ковариационную матрицу, необходимые в последующем для расчета T^2-критерия:

$$X_1' = (10{,}3 \quad 198{,}6); \qquad X_2' = (16{,}8 \quad 225{,}2),$$

$$S_* = \frac{1}{n_1 + n_2 - 2}(K_1' K_1 + K_2' K_2) = \frac{1}{8+12-2}\begin{pmatrix} 45{,}52 & 174{,}74 \\ 174{,}74 & 2027{,}88 \end{pmatrix} +$$

$$+\begin{pmatrix} 173{,}68 & 443{,}32 \\ 443{,}32 & 8759{,}68 \end{pmatrix} = \frac{1}{18}\begin{pmatrix} 219{,}20 & 618{,}06 \\ 618{,}06 & 10787{,}56 \end{pmatrix} = \begin{pmatrix} 12{,}2 & 34{,}3 \\ 34{,}3 & 599{,}3 \end{pmatrix},$$

тогда $S_*^{-1} = \begin{pmatrix} 0{,}0977 & 0{,}0056 \\ 0{,}0056 & 0{,}0020 \end{pmatrix}$.

2. Теперь можно рассчитать наблюденное значение общего T^2-критерия Хоттелинга:

$$T_н^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\bar{X}_1 - \bar{X}_2\right)' S_*^{-1}\left(\bar{X}_1 - \bar{X}_2\right) =$$

$$= \frac{96}{20}\ (-6{,}5 \quad 26{,}6)\begin{pmatrix} 0{,}0977 & 0{,}0056 \\ 0{,}0056 & 0{,}0020 \end{pmatrix}\begin{pmatrix} -6{,}5 \\ -26{,}6 \end{pmatrix} = 36{,}0.$$

3. Найдем критическое значение T^2-критерия и сравним значения $T_{кр}^2$ и $T_н^2$:

$$T_{кр}^2 = \frac{(n_1 + n_2 - 2)m}{n_1 + n_2 - m - 1} \times F_{0{,}1;2;17} = \frac{36}{17} \times 6{,}11 = 12{,}94.$$

Наблюденное значение T^2-критерия более чем в два раза превышает критическое и следует сделать вывод о существенных различиях условий проживания населения в первом и втором территориальных районах. Очевидно, наличие крупной химической промышленности во втором районе обусловливает существенные негативные отклонения значений показателей младенческой смертности и заболеваемости злокачественными новообразованиями.

В дальнейшем, как и в предыдущем параграфе, с использованием частных критериев T_j^2 может быть оценена существенность отклонений по каждому из анализируемых признаков $(\bar{X}_1, \bar{X}_2)$.

4.4. Проверка гипотез о равенстве ковариационных матриц

Сравнение ковариационных матриц, отражающих взаимосвязи изучаемых признаков, открывает возможность дополнить и уточнить гипотетические предположения относительно самих признаков. Это приобретает особенное значение, если принять во внимание, что даже специфические индивидуальные признаковые характеристики могут совпадать случайно.

В социальных и экономических исследованиях существует множество задач, требующих идентификации признаковых связей. Особенно часто они возникают при классификации наблюдаемых объектов, распознавании образов и т.п., например, при оценке кредитоспособности клиентов банков, группировке предприятий по уровню устойчивости финансового положения или при оценке эффективности производственной и коммерческой деятельности. Кроме этого, изучение взаимосвязей показателей представляет самостоятельный интерес при решении многих аналитических вопросов. Наконец, сами решения многомерными методами статистики большинства задач изначально предполагают равенство ковариационных матриц различных выборочных совокупностей.

На практике учет ковариаций (корреляций) изучаемого комплекса признаков и проверка равенства ковариационных матриц значительно снижают возможность появления ошибки в выводах. Это происходит из-за весьма малой вероятности случайного совпадения одновременно большого числа сложных характеристик признаковых связей. Наглядными здесь могут быть примеры из области медицины. Так, часто встречаются случаи, когда один или несколько симптомов (признаков) совпадают, указывая на определенную болезнь, но на самом деле заболевание

может быть иного рода. Диагностика заболевания становится гораздо более точной, если характеристику состояния какого-либо органа (например, печени) дополнить статистическими оценками связей с характеристиками состояния других органов (сердца, почек, центральной нервной системы и т.д.). Окончательный вывод о заболевании позволит сделать проверка гипотезы о существенности различий признаков, дополненная проверкой равенства ковариационных матриц для двух групп людей: с устанавливаемым и уже известным диагнозом.

Для одномерных выборочных совокупностей
тезы об однородности дисперсий осуществляется при помощи критерия Бартлетта:

$$\chi^2_н = \frac{2{,}303}{c}(n-k)\lg s_*^2 - \sum_{j=1}^{k}\left((n_j-1)\lg s_j^2\right)$$

при $c = 1+\frac{1}{3(k-1)}\left(\sum\frac{1}{n_j-1}-\frac{1}{n-k}\right)$,

где k — число нормально распределенных выборочных совокупностей;

n_j — объемы каждой из k выборок, $j=\overline{1,k}$;

n — общий объем всех выборочных совокупностей $n=\sum n_j$;

s_j^2 — дисперсия признака в j-й выборочной совокупности, $j=\overline{1,k}$;

s_*^2 — объединенная (средняя) по выборкам дисперсия,

$$s_*^2 = \frac{1}{n-k}\sum_{j=1}^{k}(n_j-1)s_j^2 .$$

Для χ^2-статистики критические значения находят по квантилей χ^2-распределения по заданному уровню значимости α и числу степеней $\nu = k - 1$. Нулевая гипотеза о равенстве дисперсий отклоняется, если $\chi^2_н \ge \chi^2_{\alpha,\nu}$, и принимается, когда $\chi^2_н \le \chi^2_{\alpha,\nu}$.

В многомерном анализе формула расчета статистики χ^2 преобразуется с учетом того, что сравниваются ковариационные матрицы двух m-мерных выборочных совокупностей и вместо скаляров используются многопараметрические оценки: векторы и матрицы. Критерий приобретает вид: $W_н = b(-2\ln\nu_1)$,

где параметры b и $-2\ln\nu_1$ определяются по формулам:

$$b = 1-\left(\sum_{j=1}^{2}\frac{1}{n_j-1}-\frac{1}{\sum_{j=1}^{2}(n_j-1)}\right)\left(\frac{2m^2+3m-1}{6(m+1)}\right),$$

$$-2\ln\nu_1 = \left(\sum_{j=1}^{2}(n_j-1)\right)\times\ln|S_*| - \sum_{j=1}^{2}\left((n_j-1)\ln|S_j|\right),$$

где m — число признаков, представляющих многомерную выборочную совокупность.

Величина многомерного W-критерия сравнивается с $\chi^2_{\alpha,\nu}$ — табличными значениями и $\nu = n_1 + n_2 - 2$.

Пример 4.5. Произведем расчет W-критерия по уже известным данным ковариационных матриц (по данным примера 4.4 из § 4.3):

$$S_1 = \frac{1}{7}\begin{pmatrix}45{,}52 & 174{,}74\\ 174{,}74 & 2027{,}88\end{pmatrix};\quad S_2 = \frac{1}{11}\begin{pmatrix}173{,}68 & 443{,}32\\ 443{,}32 & 8759{,}68\end{pmatrix};$$

$$S_* = \begin{pmatrix}12{,}2 & 34{,}3\\ 34{,}3 & 599{,}3\end{pmatrix}.$$

и $n_1 = 8$; $n_2 = 12$; $|S_1| = 1260{,}0$; $|S_2| = 10949{,}96$; $|S_*| = 6134{,}97$.

Чтобы упростить вычисления W-критерия Бартлетта, произведем предварительные расчеты параметров b и $-2\ln\nu_1$:

$$b = 1-\left(\frac{1}{7}+\frac{1}{11}-\frac{1}{18}\right)\left(\frac{2\times4+3\times2-1}{6(2+1)}\right) = 0{,}871;$$

$$-2\ln\nu_1 = 18\ln 6134{,}97-(\ln 1260{,}0+11\ln 10949{,}96) = 13{,}724.$$

Остается вычислить наблюденное значение W-критерия:

$$W_н = 0{,}871\times13{,}724 = 11{,}954.$$

Критическое значение W-критерия найдем по таблицам χ^2-распределения при $\alpha = 0{,}05$ и числе степеней свободы $m\,(m+1)/2$ или $\nu = (2\times3):2 = 3$; $\chi^2_{0{,}05;3} = 7{,}815$. Так как $W > \chi^2_{0{,}05;3}$, мы отвергаем нулевую гипотезу о равенстве ковариационных матриц (S_1 и S_2) и считаем, что при заданном уровне значимости $\alpha = 0{,}05$ их различие существенно.

Выводы

Проверка статистических гипотез проводится с целью оценки соответствия распределения случайной величины известному закону распределения, а также соответствия параметров статистической совокупности заранее заданным величинам или параметрам другой совокупности. Параметрами многомерной совокупности, подчиняющейся нормальному закону распределения, выступают вектор средних значений и матрица ковариаций.

Суждения об истинности или ложности проверяемой гипотезы строятся на основе вычисления наблюденных значений специальных критериев и последующего их сравнения с табличными (критическими) значениями. Качество проверки задается уровнем α — вероятности допустить ошибку в выводах.

При проверке статистических гипотез для многомерных случайных величин возможны различные подходы: использование общего и частных критериев с выделением и анализом различных комплексов признаков, определение доверительной области, охватывающей одновременно допустимые значения всех признаков многомерной случайной величины, и т.д.

Вопросы и задачи

1. Что понимают под статистической гипотезой и какие характерные признаки для нее существуют?

2. Что означает: простая и сложная, одномерная и многомерная статистические гипотезы?

3. Какого рода ошибки могут допускаться и чем определяется достоверность выводов при проверке статистических гипотез?

4. Покажите графически доверительную область для двумерной случайной величины: $X' = (X_1\ X_2)$, если X_1 — уровень оплатоемкости продукции — принимает параметрические значения — $\overline{x_1} = 0{,}20$ с предельно допустимой колеблемостью $(\overline{x_1} - x_{i1}) \le 0{,}5$, а X_2 — уровень энергоемкости продукции при $\overline{x_2} = 0{,}8$ и $(\overline{x_2} - x_{i2}) \le 0{,}1$.

5. Повторите решение задачи №4, при условии трехмерной случайной величины X, к уже известным признакам X_1 и X_2 добавляется характеристика X_3 — удельный вес производственных налогов в стоимости продукции: $\overline{x_3} = 0{,}45$ при $(\overline{x_3} - x_{i3}) \le 0{,}25$.

6. Будет ли статистической гипотеза:

а) о равенстве двух ковариационных матриц, представляющих связи признаков здоровья людей в двух группах: не имеющих хронической заболеваемости и с заболеваемостью;

б) о несущественности различий характеров двух человек? И если решать этот же вопрос относительно макроэкономического положения двух стран?

в) о том, что многомерная случайная величина $X' = (X_1\ X_2\ X_3)$, представляющая совокупность предприятий, подчиняется нормальному закону распределения?

Приведите собственные примеры многомерных статистических гипотез.

7. На частном предприятии с численностью работников 70 чел. средний уровень выработки одного рабочего составляет 500 деталей в день при среднеквадратическом отклонении 16 дет. С целью корректировки среднего нормативного уровня выработки в сторону повышения администрацией была отобрана группа наиболее подготовленных рабочих и по ним установлен средний уровень выработки, равный 580 дет. Можно ли вводить новый нормативный уровень выработки на предприятии, не опасаясь конфликтов с рабочими? Решите задачу при заданном уровне значимости $\alpha = 0{,}05$.

8. Чтобы оценить производственную эффективность предложенной к внедрению технологии, проведена проверка качества продукции, выпущенной на старой и новой автоматических линиях, при этом получены следующие данные об удельном весе продукции высшего качества, %:

Партия №	*Старая линия*	*Партия №*	*Новая линия*
1	2	3	4
1	58	1	74
2	62	2	59
3	51	3	69

Продолжение

1	2	3	4
4	67	4	78
5	41	5	82
6	53	6	75
		7	86
		8	63
Средний уровень	55,3	Средний уровень	73,3

При $\alpha = 0{,}001$ следует установить, действительно ли новая линия, налаженная на передовую технологию, позволяет получать более высокий уровень качества продукции?

9. Для оценки существенности воздействия состояния окружающей среды на здоровье людей в районе с неблагоприятной экологической обстановкой проведены медицинские обследования 12 отобранных случайных групп населения:

Половозрастная группа населения	*Средний уровень продолжительности жизни, лет*	*Заболеваемость онкологическими болезнями, на 100 000 жителей*	*Уровень младенческой смертности, %*
1	64	590	18
2	58	604	17
3	67	598	15
4	66	610	17
5	71	690	14
6	56	540	21
7	58	624	18
8	62	670	16
9	64	656	14
10	61	711	15
11	63	630	16
12	68	705	11

Известно, что средний по республике уровень продолжительности жизни составляет 69 лет, заболеваемости онкологическими болезнями — 580 случаев на 100 000 жителей, уровень младенческой смертности — 12‰. При $\alpha = 0{,}020$ определите, действительно ли экологические условия района оказывают существенное негативное влияние на уровень здоровья населения. После проверки гипотезы по всем трем характерным признакам повторите проверку по каждому из признаков и сформулируйте выводы.

10. Проверьте существенность различий уровня эффективности работы предприятий двух отраслей «А» и «Б» по следующим данным ($\alpha = 0{,}1$):

Отрасль «А»			*Отрасль «Б»*		
Предприятия №	*Рентабельность производства, % (X_1)*	*Среднегодовая выработка на одного работника, тыс. долл. США*	*Предприятие №*	*Рентабельность производства, % (X_1)*	*Среднегодовая выработка на одного работника, тыс. долл. США*
1	14	3,6	1	4	2,8
2	18	4,4	2	7	2,6
3	12	4,2	3	12	4,1
4	16	3,9	4	6	2,3
5	11	3,4	5	8	3,5
6	9	2,8	6	11	3,8
			7	5	2,2
			8	11	3,7
Средняя величина	13,3	3,72	—	8,0	3,12

11. Оцените существенность различий двух рынков сбыта легковых автомобилей, если на первом рынке средний уровень реализационной цены автомобиля составляет 15 тыс. долл., а экспертная оценка качества обслуживания (по 5 балльной системе) — 3,4 балла, на втором рынке соответственно: 18 тыс. долл. и 4,2 балла. Пусть $\alpha = 0{,}05$, объединенная ковариационная матрица имеет вид:

$$S_* = \begin{pmatrix} 9{,}4 & 0{,}28 \\ 0{,}28 & 2{,}0 \end{pmatrix}.$$

12. Проверьте предположение о равенстве двух ковариационных матриц, представляющих связи экономических показателей за два различных периода времени:

$$S_1 = \begin{pmatrix} 6050 & 1820 \\ 1820 & 3690 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 8400 & 2060 \\ 2060 & 4170 \end{pmatrix}.$$

Глава 5

Робастное статистическое оценивание

5.1. Грубые ошибки и методы их выявления в статистической совокупности данных

При исследовании статистических совокупностей часто приходится иметь дело с данными, отклоняющимися от основного массива, т.е. с ошибками, или выбросами. Приведем простой пример: на десяти предприятиях отрасли легкой промышленности произведены контрольные расчеты уровня рентабельности производства по итогам работы в первом полугодии и получены следующие результаты:

Предприятие	1	2	3	4	5	6	7	8	9	10
Уровень рентабельности продукции, %	15,4	13,2	18,3	47,1	12,0	16,3	65,2	17,4	11,0	12,9

В приведенных данных имеются два значения: 47,1 и 65,2, которые значительно больше всех других значений, покрываемых интервалом [11,0; 18,3]. При выявлении подобных «выбросов» возникают серьезные вопросы: являются ли отклоняющиеся данные действительно ошибками (например, регистрации) или это реальные значения и как получить адекватные оценки для параметров изучаемой совокупности. Решением подобных вопросов занимается специальный раздел статистики — *робастное (устойчивое) оценивание.*

Методы робастного оценивания — это статистические методы, которые позволяют получать достаточно надежные оценки статистической совокупности с учетом неявности закона ее распределения и наличия существенных отклонений в значениях

данных. У истоков развития методов робастного оценивания стояли американский статистик Д. Тьюки и швейцарский математик П. Хубер.

При решении задач робастного оценивания выделяют два типа данных, засоряющих статистическую совокупность. К первому типу относят данные, несущественно отличающиеся от значений, которые наиболее часто встречаются в изучаемой совокупности. Эти данные не вызывают значительных искажений в аналитических результатах и могут обрабатываться обычными методами статистического оценивания.

Второй тип данных — резко выделяющиеся на фоне изучаемой совокупности, их называют «засорением» или «грубыми ошибками», они оказывают сильное искажающее воздействие на аналитические результаты. Эти данные должны подвергаться специальной обработке.

В практике устойчивого оценивания различают следующие основные причины появления грубых ошибок:

- Специфические особенности отдельных элементов изучаемой совокупности. Как правило, они приводят к появлению случайных, или «нормальных» («обычных») отклонений.
- Неправильное причисление элементов к исследуемой совокупности, например, ошибки группировки, ошибки при организации наблюдения и т. п.
- Грубые ошибки при регистрации и обработке данных.

Если грубые ошибки являются результатом неправильных причислений элементов или ошибок регистрации (§ 2,3), то их появление и уровни непредсказуемы, а распределение может значительно отклоняться от гипотетического распределения основного массива статистических данных.

При обработке «грубых» ошибок (засорений) легко выделить два основных подхода. Первый ориентирован на устранение из выборочной совокупности ошибок и оценку параметров по оставшимся «истинным» значениям. Второй подход предполагает в каждом случае с грубой ошибкой выделение истинных значений признака и собственно ошибки $x = x_{ист} + \xi$; при этом осуществляется модификация данных таким образом, чтобы искажающий элемент ξ получил нормальное распределение с нулевым математическим ожиданием. Тогда для некоторого множества грубых ошибок вариативной величины x сумма $\Sigma\,\xi$ приближается к ну-

лю, а оценки $\hat{x}$ — к истинным значениям параметров выборочной совокупности.

Алгоритм обработки «засорений» включает последовательное выполнение шагов:

1) распознавание ошибок в данных;
2) выбор метода и проведение робастного оценивания данных;
3) критериальная или логическая проверка и интерпретация результатов устойчивого оценивания.

Выявление грубых ошибок и оценка степени засорения выборки возможны при визуальном анализе данных или проверке статистической гипотезы на наличие ошибки. Во втором случае предусматривается расчет специальных статистических критериев.

Простой формальный прием для обнаружения грубых ошибок основывается на расчете T- критерия Граббса:

$$T_{\text{н}} = \frac{x - \bar{x}}{s},$$

где $\bar{x}$ — выборочная средняя. Ее оценка предпочтительна по истинным данным, в противном случае $\bar{x} = \sum_{1}^{n} \frac{x}{n}$;

s — выборочное среднеквадратическое отклонение случайной величины. Для s также предпочтительна оценка по истинным данным, в противном случае расчет производится, как и в первом случае, по данным всей выборочной совокупности.

Наблюденные значения T-критерия сравнивают с пороговыми, заданными соответствующим распределением. Проверяемые признаковые значения относят к классу выбросов, если $T_{\text{н}} > T_{\text{кр}}$ ($T_{\text{кр}} = T_{\alpha,h}$). Если $T_{\text{н}} < T_{\text{кр}}$, то считается, что эти значения несущественно отличаются от других данных и не будут давать сильного искажающего эффекта.

Критерий Граббса прост и легко применим в анализе, но как установлено, имеет существенные недостатки. В частности, исследователи обращают внимание на его недостаточную точность (часто дает весьма грубые оценки) и, кроме того, он «нечувствителен» к маскирующим эффектам, когда выбросы группируются достаточно близко друг от друга в отдаленности от основной массы наблюдений.

Более точными по сравнению со статистикой Граббса оценками грубых ошибок признаются L- и E-критерии, предложенные американскими статистиками Г. Титьеном и Г. Муром:

1. L-критерий исчисляется для выявления грубых ошибок в верхней части ранжированного ряда данных:

$$L = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x}_k)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

где x_i — выборка i наблюдений по какому-либо одному, j-му признаку;

n — объем выборки;

k — число наблюдений с резко отклоняющимися значениями признака;

$\bar{x}$ — общая для выборочной совокупности данных средняя величина;

$\bar{x}_k$ — средняя, которую рассчитывают по $n - k$ наблюдениям, остающимися после отбрасывания k грубых ошибок «сверху» ранжированного ряда данных: $\bar{x}_k = \frac{\sum_{i=1}^{n-k} x_i}{n-k}$.

2. L'-критерий применяется для выявления грубых ошибок в данных, расположенных в нижней части ранжированного ряда данных:

$$L' = \frac{\sum_{i=k+1}^{n} (x_i - \bar{x}_k)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

где $\bar{x}$ — средняя, рассчитанная по $n - k$ наблюдениям, остающимся после отбрасывания k грубых ошибок «снизу»:

$$\bar{x}_k = \frac{\sum_{i=k+1}^{n-k} x_i}{n-k}.$$

3. E-критерий используется, когда в выборке имеются предположительно грубые ошибки с наибольшими и наименьшими

значениями, т.е. расположенные в верхней и нижней частях ранжированного ряда данных:

$$E=\frac{\sum_{i=k+1}^{n-k'}(x_i-\bar{x}_{k'})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

где $\bar{x}_{k'}$ — средняя, рассчитанная по «истинным» данным после отбрасывания из выборки наименьших (k) и наибольших (k') значений засоряющих совокупность данных: $\bar{x}_{k'}=\frac{\sum_{i=k+1}^{n-k'}x_i}{n-(k+k')}$.

Все три критерия L, L' и E имеют табулированные критические значения для заданного уровня значимости α при известном объеме выборки n и предполагаемом числе ошибок k. Если наблюденные значения критериев оказываются меньше пороговых $C_{\alpha,k}$, то ошибки в данных, подвергаемые проверке, признаются грубыми, существенно отклоняющимися от основного массива данных. При L, L', $E>C_{\alpha,k}$ данные гипотетически предполагаются типичными для изучаемой выборочной совокупности.

5.2. Методы исчисления устойчивых статистических оценок: Пуанкаре, Винзора, Хубера

После обнаружения выбросов в данных решается задача оценивания параметров выборочной совокупности. При этом, как выше уже сказано, используются два основных подхода: экстремальные значения (грубые ошибки) отбрасываются либо модифицируются.

Наиболее простыми представляются оценки по усеченной совокупности данных, остающейся после отбрасывания грубых ошибок. Американский статистик Пуанкаре предложил следующую формулу для расчета средней по усеченной совокупности (урезанную среднюю):

$$T(\alpha)=\frac{1}{n-2k}\sum_{i=k+1}^{n-k}x_i.$$

В формуле k — число грубых ошибок, $k\le\alpha n$ — целая часть от произведения αn, где n — объем выборочной совокупности, а α — некоторая функция величины засорения выборки ξ. Значения α находят по специальным таблицам (см. табл. 5.5). Обычно α колеблется в пределах от нуля до 0,5.

Другой подход демонстрирует оценка Винзора, она предполагает замену признаковых значений, засоряющих выборку, на модифицированные (винзорированные) значения с устраненными или уменьшенными ошибками.

Средняя по Винзору определяется также с известным заранее уровнем α ($0<\alpha<1/2$) по формуле:

$$W(\alpha)=\frac{1}{n}\left(\sum_{i=k+2}^{n-k-1}x_i+k(x_{k+1}+x_{n-k})\right).$$

По аналогии с оценками $T(\alpha)$ и $W(\alpha)$, т.е. соответственно по усеченной совокупности, или винзорированным данным, могут быть найдены не только средние величины, но и другие оценки параметров статистической совокупности, например, вариации, моды, медианы и т.п.

Приемы робастного оценивания Пуанкаре и Винзора дают хорошие результаты на выборках с симметричным распределением засорений, когда грубые ошибки группируются примерно на одном расстоянии от центра в нижней и верхней частях статистической совокупности.

Наряду с уже названными методами робастного оценивания, широкое распространение имеет ставший классическим подход Хубера. Он напоминает процедуры для последовательного «улучшения» данных по Винзору. При этом используется некоторая исходная величина k, определяемая с учетом степени «засорения» статистической совокупности ξ и определяющая шаг модификации резко отличающихся наблюдений (см. табл. 5.6).

Оценка средней величины по методу Хубера производится по формуле:

$$\hat{\theta}=\frac{1}{n}\left(\sum_{|x_i-\theta|<k}x_i+(n_2-n_1)k\right),$$

где $\hat{\theta}$ — устойчивая оценка, определяется при помощи итеративных процедур;

k — величина, которая допускается в качестве отклонения от центра совокупности, принимает постоянные значения

с учетом удельного веса грубых ошибок в совокупности данных ξ;

n_1 — численность группы наблюдений из совокупности, отличающихся наименьшими значениями: $x_i < \theta - k$, или значения в интервале $(-\infty; \theta - k)$;

n_2 — численность группы наблюдений из совокупности, отличающихся наибольшими значениями: $x_i < \theta + k$, или значения в интервале $(\theta + k; \infty)$.

При расчетах по приведенной выше формуле в качестве начальной оценки θ может приниматься обычная средняя арифметическая или медиана, оцененная по выборке. Затем на каждой итерации производится разделение выборочной совокупности на три части. В одну часть попадают «истинные» признаковые значения, которые остаются без изменения ($|x_i - \theta| < k$). В две другие части совокупности (для $x_i > \theta + k$ и $x_i < \theta - k$) попадают «ошибки», они не исключаются из рассмотрения, а заменяются соответственно на величины $x_i - k$ и $x_i + k$. По «истинным» и модифицированным данным каждый раз определяется новая оценка средней θ и итерация возобновляется. Итерации повторяются до тех пор, пока все наблюдения не оказываются в интервале «истинных» значений: $|x_i - \theta| < k$.

Оценка $\hat{\theta}$, найденная по методу Хубера, представляется достаточно эффективной, но быстро теряет оптимальные свойства с увеличением засорения выборки (ростом ξ).

Пример 5.1. По 20 городам с различным уровнем насыщенности промышленного производства имеются данные о заболеваемости верхних дыхательных путей у детей, на 100 000 детей:

Город	*Уровень заболеваемости верхних дыхательных путей*	*Город*	*Уровень заболеваемости верхних дыхательных путей*
1	1258,7	11	1307,2
2	1060,2	12	1078,5
3	1186,5	13	4020,1
4	1263,7	14	1285,8
5	100,9	15	919,3
6	3610,7	16	1389,4
7	1037,5	17	1161,7
8	1507,9	18	1015,8
9	1291,9	19	124,2
10	1403,3	20	965,8

На основе приведенных данных найдем обычные оценки средней и дисперсии и устойчивые оценки, учитывающие наличие в данных грубых ошибок.

В исходной совокупности выделяются значения: 100,9; 124,2 и 3610,7; 4020,1. Можно предположить, что эти данные записаны неверно, взяты из другой графы отчетности или, наконец, представляют города с резко отличающимися от основной совокупности своими экологическими характеристиками. Проверим эти данные на «засорение», применив критерий Граббса:

$$T_1 = \frac{100{,}9 - 1349{,}5}{897{,}8} = 1{,}391;\quad T_3 = \frac{3610{,}7 - 1349{,}5}{897{,}8} = 2{,}519;$$

$$T_2 = \frac{124{,}2 - 1349{,}5}{897{,}8} = 1{,}365;\quad T_4 = \frac{4020{,}1 - 1349{,}5}{897{,}8} = 2{,}975.$$

В расчетах использованы обычные оценки: $\bar{x} = 1349{,}5$ ($\bar{x} = \sum x/n$) и $\sigma = 897{,}8$ ($\sigma = \left(\sum_1^n (x - \bar{x})^2 / n\right)^{1/2}$).

Сравнивая наблюденные значения T-критерия с критическими для $\alpha = 0{,}10$ (см. табл. 5.2), к грубым ошибкам следует отнести только два значения: 3610,7 и 4020,1, для них $T_н > T_{кр}$, $T_{кр} = 2{,}447$. Тем не менее очевидно, что два оставшихся значения: 100,9 и 124,2 также значительно отличаются от основного массива данных. Уточним результаты проверки при помощи более чувствительного E-критерия Титьена и Мура; при этом предположительно отнесем к числу ошибок все четыре значения выделяющиеся в данных:

$$E_н = \frac{\sum_{i=5}^{18} (x_i - \bar{x}_{k'})^2}{\sum_1^{20} (x_i - \bar{x})^2} = \frac{(919{,}3 - 1195{,}8)^2 + (965{,}8 - 1195{,}8)^2 + \ldots + (1507{,}9 - 1195{,}8)^2}{(100{,}9 - 1349{,}5)^2 + (124{,}4 - 1349{,}5)^2 + \ldots + (4020{,}1 - 1349{,}5)^2} =$$

$$= \frac{4364627}{16119810{,}6} = 0{,}027.$$

Здесь $\bar{x}_{k'} = \frac{\sum_{i=3}^{18} x_i}{n - k'} = \frac{919{,}3 + 965{,}8 + \ldots + 1507{,}9}{16} = 1195{,}8.$

При критических значениях $E_{0,05}$ = 0,027 (табл. 5.4) все четыре значения, подозреваемые на грубые ошибки, следует действительно признать «засорением» совокупности, так как $E_н$ значительно меньше $E_{0,05}$, или 0,027 < 0,221.

Для проведения расчетов устойчивых статистических оценок построим специальную таблицу с данными, систематизированными соответствующим образом (табл. 5.1).

Таблица 5.1. **Систематизированные данные об уровне заболеваемости верхних дыхательных путей у детей, проживающих в 20 различных городах**

№ п/п	*Номер города*	*Уровень заболеваемости верхних дыхательных путей, на 100 000 детей*	*Усеченная совокупность данных*	*Винзорированные данные*
1	5	100,9	—	919,3
2	19	124,2	—	919,3
3	15	919,3	919,3	919,3
4	20	965,8	965,8	965,8
5	18	1015,8	1015,8	1015,8
6	7	1037,5	1037,5	1037,5
7	2	1060,2	1060,2	1060,2
8	12	1078,5	1078,5	1078,5
9	17	1161,7	1161,7	1161,7
10	3	1186,5	1186,5	1186,5
11	1	1258,7	1258,7	1258,7
12	4	1263,7	1263,7	1263,7
13	14	1285,8	1258,8	1285,8
14	9	1291,9	1261,9	1291,9
15	11	1307,2	1307,2	1307,2
16	16	1389,4	1389,4	1389,4
17	10	1403,3	1403,3	1403,3
18	8	1507,9	1507,9	1507,9
19	6	3610,7	—	1507,9
20	13	4020,1	—	1507,9
Итого		26989,1	19133,2	23987,6
Средняя величина ($\bar{x}$)		1349,5	1195,8	1199,4

В табл. 5.1 наряду с исходными данными приведены результаты расчетов, необходимых для исчисления устойчивых средних:

- средняя арифметическая простая:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{100,9 + 124,2 + 919,3 + \ldots + 4020,1}{20} = 1349,5;$$

- средняя, исчисленная по усеченной совокупности данных (средняя по Пуанкаре):

$$T(\alpha) = \frac{1}{n-2k}\sum_{k+1}^{n-k} x_i = \frac{919,3 + 965,8 + 1015,8 + \ldots + 1507,9}{16} = 1195,8;$$

- средняя, исчисленная по винзорированным данным:

$$W(\alpha) = \frac{1}{n}\left(\sum_{k+2}^{n-k-1} x_i + k(x_{k+1} + x_{h-k})\right) =$$

$$= \frac{1}{20}\left[(919,3 + 965,8 + \ldots + 1507) + 2(919,3 + 1507,9)\right] = 1199,4.$$

Как видим, средние значения, рассчитанные по формулам устойчивых оценок Пуанкаре и Винзора, близки по величине, но отличаются от обычной средней примерно на 11%.

По упорядоченным данным табл. 5.1 могут быть рассчитаны и другие статистические оценки: дисперсия, среднее квадратическое отклонение и т. д.

Расчет оценки Хубера в отличие от уже рассмотренных алгоритмов поиска оценок по Пуанкаре и Винзору предусматривает итеративно повторяющиеся вычислительные процедуры.

На первом шаге в качестве исходной оценки θ выберем простую арифметическую среднюю: $\theta = \bar{x} = 1349,5$. При числе ошибок k в выборочной совокупности, равном 4, найдем значение параметра $k(\xi)$: $k(\xi) = 0,862$.

Для последующего улучшения θ-оценки разобъем всю имеющуюся совокупность данных на три класса. В первый класс войдут значения, незначительно отличающиеся от предварительной оценки θ (истинные), во второй класс — значения, существенно меньшие величины θ, и в третий класс — значения, существенно превышающие θ. Затем соответствующим образом модифицируем x_i, если $x_i > \theta + k$, или $x_i < \theta - k$:

	I класс $\lvert x_i-\theta\rvert<k$	II класс $x_i>\theta+k$ $(x_i>1350{,}4)$	III класс $x_i<\theta-k$ $x_i<1348{,}6$
Исходные значения	—	1389,4 3610,7 1403,3 4020,1 1507,9	100,9 1015,8 1161,7 1285,8 124,2 1037,5 1186,5 1291,9 919,3 1060,2 1258,7 1307,2 965,8 1078,5 263,7
Модифицированные значения	—	1388,5 3609,8 1402,4 4019,2 1507,0	101,8 1016,7 1162,6 1286,7 125,1 1038,4 1187,4 1292,8 920,2 1061,1 1259,6 1308,1 966,7 1079,4 1264,6

Рассчитаем оценку θ по данным, модифицированным первый раз $\hat{\theta}_1$:

$$\hat{\theta}_1=\frac{1}{n}\left(\sum_{|x_i-\theta|<k}x_i+(n_2-n_1)k\right).$$

Возобновим итерацию по данным, модифицированным на предыдущем шаге:

	I класс $\lvert x_i-\theta\rvert<k$ $1349{,}0<x_i<1350{,}8$	II класс $x_i>\theta+k$ $(x_i>1350{,}4)$	III класс $x_i<\theta-k$ $x_i<1348{,}6$
Исходные значения	—	1388,5 3609,8 1402,4 4019,2 1507,0	101,8 1016,7 1162,6 1286,7 125,1 1038,4 1187,4 1292,8 920,2 1061,1 1259,6 1308,1 966,7 1079,4 1264,6
Модифицированные значения	—	1387,6 3608,9 1401,5 4018,3 1506,1	102,7 1017,6 1163,5 1287,6 126,0 1039,3 1188,3 1293,7 921,1 1062,0 1260,5 1309,0 967,6 1080,3 1265,5

второй итерации оценка θ будет:

$$\hat{\theta}_2=\frac{1}{20}(26997{,}7+(15-15)0{,}862)=1350{,}3 \text{ и т. д.}$$

Очевидно, что для данных, имеющих большой разброс значений, число итераций будет достаточно велико.

В многомерном случае «засорением» совокупности данных уже будут не отдельные значения, а вектор значений, представляющий аномальный объект.

Чтобы удостовериться, что многомерное наблюдение является действительно выбросом, обычно используют расстояние Махаланобиса:

$$d^m=\left(X-\bar{X}\right)'\sum{}^{-1}\left(X-\bar{X}\right),$$

где X — вектор признаковых значений, подозреваемых на «выброс»;

$\bar{X}$ — вектор средних значений для многомерной совокупности данных;

Σ — матрица ковариаций.

Критерий F для проверки гипотезы о существенности отклонения случайного вектора X строится следующим образом:

$$F_{\text{н}}=\frac{(n-m)n}{2(n-1)m}\left(X-\bar{X}\right)'\sum{}^{-1}\left(X-\bar{X}\right).$$

Для F-критерия существуют числа $\nu_1=m$ и $\nu_2=n-m-1$ степеней свободы. При заданном уровне значимости α, если $F_{\text{н}}>F_{\alpha,\nu_1,\nu_2}$, проверяемое наблюдение действительно признается аномальным. В противном случае, т. е. когда $F_{\text{н}}\le F_{\alpha,\nu_1,\nu_2}$, отклонение случайного вектора от вектора средних значений считается приемлемым, а гипотеза о «засорении» совокупности отбрасывается.

В случае значительного засорения многомерная совокупность подвергается проверке итеративным способом:

а) одно из наблюдений, которое предположительно является «засорением», подвергается проверке. Если предположение оправдывается, «выброс» устраняется из выборки;

б) по усеченной совокупности многомерных объектов определяется новый вектор средних значений;

в) проверке подвергается следующий объект, повторяются шаги а и б, и т. д.

К выявленным грубым ошибкам в многомерной совокупности можно применять уже известные для одномерного случая приемы обработки данных: их устранение, или винзорирование.

Выводы

При обнаружении «засорения», или «грубых ошибок», в совокупности данных, т. е. значений, резко отличающихся от медианных, используются принципы проверки статистических гипотез. Наиболее простыми и распространенными являются методы поиска ошибок Граббса, Титьена и Мура. Если в статистической совокупности действительно выявлены «грубые ошибки», то для уменьшения их влияния на аналитические результаты рекомендуется применение специальных приемов обработки данных. Их сущность сводится к одному из двух решений: устранению из совокупности аномальных наблюдений, т. е. усечению совокупности, или модификации резко отличающихся значений с целью уменьшения ошибок в данных. Первое решение представлено в подходе Пуанкаре, второе — в подходах Хубера и Винзора, при этом само изменение данных, направленное на минимизацию ошибки в них, определяется как винзорирование.

Получение устойчивых характеристик статистической совокупности носит название робастного оценивания.

Проверка статистических гипотез и робастное оценивание в экономических исследованиях используются часто как самостоятельные статистические приемы в решении задач оценки качества товаров, оценки адекватности заданным производственным, технологическим, экологическим условиям и т. п. Представленные методы в комплексе с другими статистическими методами позволяют предварительно анализировать наблюденные значения характерных признаков, выявить в них несоответствия и грубые ошибки, провести модификацию данных, повышающую гомогенность изучаемой совокупности.

Вопросы и задачи

1. Что понимается под грубыми ошибками и каковы причины их проявления в статистической совокупности?

2. Какие существуют подходы при обработке грубых ошибок?

3. Назовите основные методы устойчивого оценивания параметров выборочной совокупности.

4. На предприятии за 14 месяцев собраны данные об удельном весе брака в общем объеме продукции:

Месяц	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Удельный вес брака продукции, %	3	4	2	3	14	2	3	8	95	5	3	4	2	4

Используя приемы Граббса, Титьена и Мура, определите наличие грубых ошибок в совокупности данных.

5. Имеются сведения о размере прибыли, млн. руб. (X_1) и объеме основных производственных фондов, млрд. руб. (X_2) по 16 предприятиям:

Предприятие	X_1	X_2	*Предприятие*	X_1	X_2
1	15	9	9	22	0
2	20	12	10	7	94
3	380	14	11	18	8
4	−90	68	12	14	6
5	24	15	13	−120	14
6	10	6	14	450	−20
7	8	11	15	16	12
8	11	4	16	9	18

Рассчитайте обычную и устойчивую средние, используя методы Пуанкаре и Винзора, сравните полученные результаты.

При расчетах примите во внимание наличие двух ошибок «снизу» и «сверху» ранжированного ряда данных.

6. По 10 наблюдаемым объектам имеются данные о средней концентрации загрязняющих веществ в воздухе:

Объект	*Пыль, мг/м³*	*Окись углерода, мг/м³*	*Объект*	*Пыль, мг/м³*	*Окись углерода, мг/м³*
1	0,12	1,2	6	0,27	2,8
2	0,10	1,6	7	0,18	1,1
3	0,16	1,8	8	0,10	1,3
4	0,14	1,7	9	0,16	1,4
5	0,20	3,0	10	0,21	2,9

Используя критерии Титьена и Мура, определите наличие грубых ошибок по каждой переменной и рассчитайте устойчивые средние по методу Пуанкаре.

Табулированные значения статистических критериев, применяемых при обработке совокупностей данных с грубыми ошибками

Таблица 5.2. **Процентные точки критерия Смирнова—Граббса (T)**

№ n/n	Доверительная вероятность (1 — α)			№ n/n	Доверительная вероятность (1 — α)		
	0,9	0,95	0,99		0,9	0,95	0,99
3	1,412	1,414	1,414	27	2,749	2,913	3,239
4	1,689	1,710	1,728	28	2,764	2,929	3,258
5	1,869	1,917	1,972	29	2,778	2,944	3,275
6	1,996	2,067	2,161	30	2,792	2,958	3,291
7	2,093	2,182	2,310	31	2,805	2,972	3,307
8	2,172	2,273	2,431	32	2,818	2,985	3,322
9	2,238	2,349	2,532	33	2,830	2,998	3,337
10	2,294	2,414	2,616	34	2,842	3,010	3,351
11	2,343	2,470	2,689	35	2,853	3,022	3,364
12	2,387	2,519	2,753	36	2,864	3,033	3,377
13	2,426	2,563	2,809	37	2,874	3,044	3,389
14	2,461	2,602	2,859	38	2,885	3,055	3,401
15	2,494	2,638	2,905	39	2,894	3,065	3,413
16	2,523	2,670	2,946	40	2,904	3,075	3,424
17	2,551	2,701	2,983	41	2,913	3,084	3,435
18	2,577	2,728	3,017	42	2,922	3,094	3,445
19	2,601	2,754	3,049	43	2,931	3,103	3,455
20	2,623	2,779	3,079	44	2,940	3,112	3,465
21	2,644	2,801	3,106	45	2,948	3,120	3,474
22	2,664	2,823	3,132	46	2,956	3,129	3,483
23	2,683	2,843	3,156	47	2,964	3,137	3,492
24	2,701	2,862	3,179	48	2,972	3,145	3,501
25	2,718	2,880	3,200	49	2,980	3,152	3,510
26	2,734	2,897	3,220	50	2,987	3,160	3,518

Таблица 5.3. **Критические значения C_α-оценки для L- и L′-критериев Титьена и Мура ($\alpha = 0{,}05$)**

№	1	2	3	4	5	6	7	8	9	10
3	0,003									
4	051	0,001								
5	125	018								
6	203	055	0,010							
7	273	106	032							
8	326	146	064	0,022						
9	372	194	099	045						
10	418	233	129	070	0,034					
11	0,454	0,270	0,162	0,098	0,054					
12	489	305	196	125	076	0,042				
13	517	337	224	150	098	060				
14	540	363	250	174	122	079	0,050			
15	556	387	276	197	140	097	066			
16	575	410	300	219	159	115	082	0,055		
17	594	427	322	240	181	136	100	072		
18	608	447	337	259	200	154	116	086	0,062	
19	624	462	354	277	209	168	130	099	074	
20	639	484	377	299	238	188	150	115	088	0,066
25	696	550	450	374	312	262	222	184	154	126
30	730	599	506	434	376	327	283	245	212	183
35	762	642	554	482	424	376	334	297	264	235
40	784	672	588	523	468	421	378	342	310	280
45	802	696	618	556	502	456	417	382	350	320
50	820	722	646	588	535	490	450	414	383	356

Таблица 5.4. **Значения C_α-оценки для E-критерия Титьена и Мура ($\alpha = 0{,}05$)**

№	1	2	3	4	5	6	7	8	9	10
3	0,001									
4	025	0,001								
5	081	010								
6	146	034	0,004							
7	208	065	016							
8	265	099	034	0,010						
9	314	137	057	021						
10	356	172	083	037	0,014					
11	386	204	107	055	026					
12	424	234	133	073	039	0,018				
13	455	262	156	092	053	028				
14	0,484	0,293	0,179	0,112	0,068	0,039	0,021			
15	509	317	206	134	084	052	030			
16	526	340	227	153	102	067	041	0,024		
17	544	362	248	170	116	078	050	032		
18	562	382	267	187	132	091	062	041	0,026	
19	581	398	287	203	146	105	074	050	033	
20	597	416	302	221	163	119	085	059	041	0,028
25	652	493	381	298	236	186	146	114	089	068
30	698	549	443	364	298	246	203	116	137	112
35	732	596	495	417	351	298	254	214	181	164
40	758	629	534	458	395	343	297	259	223	195
45	778	658	567	492	433	381	337	299	263	233
50	797	684	599	529	468	417	373	334	299	268

Таблица 5.5. **Значения α для расчета устойчивых оценок $T(\alpha)$ — Пуанкаре и $W(\alpha)$ — Винзора**

ξ	α	ξ	α
0	0	0,20	0,194
0,001	0,004	0,25	0,222
0,002	0,008	0,3	0,247
0,005	0,015	0,4	0,291
0,01	0,026	0,5	0,332
0,02	0,043	0,65	0,386
0,05	0,081	0,80	0,436
0,10	0,127	1	0,500
0,15	0,164		

Таблица 5.6. **Значения $k = f(\xi)$ для расчета устойчивой оценки Хубера**

ξ	k	ξ	k
0	0	0,20	0,862
0,001	2,630	0,25	0,766
0,002	2,435	0,3	0,685
0,005	2,160	0,4	0,550
0,01	1,945	0,5	0,436
0,02	1,717	0,65	0,291
0,05	1,399	0,80	0,162
0,10	1,140	1	0
0,15	0,980		

Глава 6

Множественный регрессионный анализ

6.1. Введение в множественный корреляционно-регрессионный анализ

Корреляционно-регрессионный анализ, как известно, является одним из наиболее широко распространенных и гибких приемов обработки статистических данных. Его появление связывают с именем английского исследователя Фрэнсиса Гальтона, предложившего в 1795 г. теоретические основы регрессионного метода, а в 1801 г. рассчитавшего с его помощью траекторию полета планеты Церера. Известны также имена Огюста Браве, Густава Теодора Фехнера, Фрэнсиса Эджворта, высказывавших в середине—конце XIX в. первые идеи о количественном измерении связей явлений. В разное время над теорией анализа работали известные в области теоретической статистики ученые: Карл Фридрих Гаусс, Андриан Мари Лежандр, Карл Пирсон и др.

Наиболее простой формой корреляционно-регрессионного анализа являются парная корреляция и парная регрессия. Многомерный анализ, как известно, отличают процедуры обработки множественных характеристик, комплексно представляющих взаимосвязанные признаки (объекты). При этом в множественном регрессионном анализе:

- исследуется зависимость результативной величины — отклика (y) от нескольких независимых переменных — предикторов (X_j), т.е.

$$y = f(X_1, X_2, X_3, \ldots, X_m) + \xi;$$

- выделяется понятие чистой регрессии — зависимости между некоторыми парами предикторов из их множества при условии нивелирования действия остальных предикторов;

- учитывается возможность наличия тесных связей (когда коэффициент корреляции превышает уровень 0,7—0,8) между парами предикторов, искажающих конечные результаты регрессионного анализа отклика. Это явление носит название *мультиколлинеарности*, устраняется оно, как правило, одним из двух способов: один из пары предикторов, подверженных мультиколлинеарности, выводится из модели или заменяется другим новым предиктором — новым факторным признаком;
- существует необходимость установления определенного соотношения между числом наблюдаемых объектов и числом предикторов. Корректное проведение анализа требует обычно, чтобы это соотношение было 6—8 к 1;
- принимается во внимание, что при числе предикторов, превышающем два, графическое изображение результатов регрессионного анализа становится невозможным и все выводы формируются в ходе формального решения аналитической задачи;
- в связи с тем, что в множественном корреляционно-регрессионном анализе (МКРА) определяется большое число параметров, проверке на достоверность подлежат не только регрессионная модель в целом, но и каждый из ее параметров, а также всевозможные парные и частные коэффициенты корреляции.

Приведем основные положения теории МКРА.

Парная корреляция. Коэффициенты парной корреляции используются для измерения силы линейных связей различных пар признаков из их множества. При этом учитывается, что связь каждой пары признаков находится под воздействием связей всех других признаков между собой и с признаками из данной пары.

Для множества признаков объектов матрицу парных корреляций R получают в ходе следующих преобразований матрицы исходных данных X:

$$X \to Z \to Z'Z \to \frac{1}{n} Z'Z = R,$$

где Z — матрица стандартизованных значений, ее элементы получают из x_{ij} как $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$ и $Z = \|z_{ij}\|$.

Частная корреляция. Коэффициенты частной корреляции также представляют линейные связи признаков, но при этом во внимание принимается чистая связь пары признаков при условии, что связи всех других признаков с признаками из данной пары не действуют, нивелированы. Элементы матрицы коэффи-

циентов частной корреляции r_{ij} можно получить по данным известной матрицы парных корреляций R:

$$r_{ij.} = -\frac{A_{ij}}{(A_{ii}A_{jj})^{1/2}},$$

где A_{ij}, A_{ii} и A_{jj} — алгебраические дополнения к соответствующим элементам матрицы парных корреляций R.

Знак коэффициенту частной корреляции присваивается согласно знаку соответствующего коэффициента регрессии в линейной модели.

Коэффициент множественной корреляции R_0 представляет собой численную характеристику силы связи отклика со всеми предикторами. Если известна матрица парных корреляций R, то

$$R_0 = \left(1 - \frac{|R|}{|R_j|}\right)^{1/2},$$

где $|R|$ — определитель матрицы парных корреляций;

$|R_j|$ — минор к матрице парных корреляций R. В матрице R вычеркиваются строка и столбец, представляющие характеристики связи с j-м признаком, выступающим в качестве отклика.

Коэффициент множественной детерминации R_0^2 — численная характеристика доли вариации признака, объясненной вариацией всех предикторов:

$$R_0^2 = 1 - \frac{|R|}{|R_j|} = (R_0)^2.$$

Коэффициенты множественной корреляции и детерминации представляют собой оценки силы линейных связей изучаемых признаков.

Коэффициент неопределенности — численная характеристика доли вариации отклика, не поддающейся объяснению вариацией предикторов: $R_н^2 = 1 - R_0^2$.

Регрессионные модели используются для представления формы связи изучаемых признаков.

Наиболее простым классом регрессионных моделей являются линейные:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m + \xi.$$

Вектор параметров такой модели находят при условии минимизации ее ошибки ξ. С использованием метода наименьших квадратов (*МНК*) легко выводится формула для определения множества параметрических значений b_j:

- выдвигаем *МНК*-требование:

$\sum \xi^2 = \sum (y - \hat{y})^2 \to 0$;

- перепишем *МНК*-условие, заменив $\hat{y}$ на произведение матрицы X и вектора B, т.е. $\hat{y} = XB$, а множество значений y_i представим вектором Y и получим

$(Y - XB)'(Y - XB) \to 0$;

- выполним операцию умножения и продифференцируем полученное выражение относительно параметра B:

$Y'Y - 2B'X'Y + B'X'XB \to 0$; $\frac{d(\xi^2)}{dB} = -2X'Y + 2BX'X = 0$, откуда

$-X'Y = -BX'X$ и $B = (X'X)^{-1}X'Y$.

При определении вектора B матрица исходных данных может принимать вид X_1 или X_2:

$$X_1 = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix}.$$

Если в анализе используется матрица вида X_1, то в ходе решения регрессионного уравнения находят все b_j, кроме b_0, а затем b_0 вычисляют как разность:

$b_0 = \bar{y} - \bar{X}'B^*$,

где $\bar{y}$ — среднее значение отклика;

$\bar{X}'$ — вектор средних значений предикторов $\bar{X}' = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m)$;

B^* — неполный параметрический вектор, $B'^* = (b_1, b_2, \ldots, b_m)$.

Если в анализе используется матрица вида X_2, то одновременно находят все множество параметрических оценок, т.е. полный вектор B.

Регрессионное уравнение $\hat{y} = XB$ называют уравнением в натуральном масштабе. Его коэффициенты показывают, на сколь-

ко натуральных единиц изменится отклик при изменении значений соответствующего предиктора на одну единицу.

Кроме регрессии в натуральном масштабе, может быть построена регрессия в стандартизованном виде:

$$\hat{Z}_1 = \beta_2 Z_2 + \beta_3 Z_3 + \ldots + \beta_m Z_m, \quad z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j};$$

здесь коэффициенты регрессии показывают, на сколько средних квадратических отклонений изменится отклик при изменении соответствующего предиктора на одно среднее квадратическое отклонение. Построение регрессии в стандартизованном виде предполагает решение системы нормальных уравнений:

$$r_{12} = \beta_2 + \beta_3 r_{32} + \ldots + \beta_m r_{m2},$$

$$r_{13} = \beta_2 r_{23} + \beta_3 + \ldots + \beta_m r_{m3},$$

. . . . . .

$$r_{1m} = \beta_2 r_{2m} + \beta_3 r_{3m} + \ldots + \beta_m.$$

Вектор значений β-коэффициентов определяется при известной матрице парных корреляций R просто:

$$\beta_j = \frac{|R_i|}{|R|},$$

где $|R|$ — определитель матрицы парных корреляций или определитель матрицы системы нормальных уравнений;

$|R_i|$ — определитель матрицы системы, в которой столбец, включающий неизвестные параметры, заменяется свободными членами системы.

От стандартизованных коэффициентов регрессии всегда можно осуществить переход к коэффициентам в натуральном масштабе:

$$b_j = \beta_j \frac{\sigma_1}{\sigma_j},$$

где σ_1 — среднее квадратическое отклонение значений отклика Y;

σ_j — среднее квадратическое отклонение значений соответствующего предиктора x_j.

На основе значений β_j рассчитывают частные и множественный коэффициенты детерминации:

- частные коэффициенты детерминации: $r_j^2 = \beta_j r_{1j}$;
- коэффициент множественной детерминации: $R_0^2 = \sum \beta_j r_{1j}$.

Хотя линейная регрессия — наиболее удобная и простая форма описания взаимодействия изучаемых признаков (объектов), она далеко не всегда является достаточно надежной моделью для реально происходящих явлений и процессов. В исследованиях поэтому нередко используются нелинейные регрессионные модели, а чтобы избежать сложностей с определением параметрических оценок и интерпретацией регрессионных коэффициентов, такие модели стараются привести к линейному виду и находить их решение по хорошо известной формуле: $B = (X'X)^{-1}X'Y$. В табл. 6.1 приведены наиболее распространенные нелинейные модели, а также показана возможность их преобразования в линейные.

Надежность решений, полученных методами корреляционного и регрессионного анализа. Корреляционно-регрессионный анализ логично завершается оценкой достоверности полученной модели и ее параметрических характеристик, а затем интерпретацией результатов.

В общем числе критериальных оценок надежности *МКРА* выделим следующие группы.

- Статистические оценки надежности регрессионной модели в целом:

а) коэффициенты множественной детерминации и корреляции. Допустимые значения для коэффициента множественной детерминации следующие:

0,01—0,09 — связь отклика и предикторов слабая, теоретически подтверждена недостаточно;

0,09—0,49 — связь средняя;

0,49—1,00 — связь достаточно сильная, использование регрессионной модели в анализе теоретически обоснованно;

б) MSE — средний квадрат модельной ошибки, $MSE = \frac{1}{n}\sum(y - \hat{y})^2$.

Наилучшей считается модель с минимальным значением величины MSE;

Таблица 6.1. **Избранные нелинейные регрессионные модели и способы их приведения к линейному виду**

Нелинейная модель	*Преобразование исходных данных для приведения модели к линейному виду*	*Описание явлений, процессов*
Полиномиальная: $\hat{y}_x = a_0 + a_1 x_1 + a_2 x_2^2 + \ldots + a_m x_m^m$	$Y^* = Y;\ X^* = \begin{pmatrix} x_{11} & x_{12}^2 & \ldots & x_{1m}^m \\ x_{12} & x_{22}^2 & \ldots & x_{2m}^m \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n}^2 & \ldots & x_{nm}^m \end{pmatrix}$	Гибкая модель для описания разнообразных процессов, меняющих направления своего развития (имеющих точки перегиба)
Линейно-логарифмическая: $\hat{y} = ax_1^{a_1} x_2^{a_2} \ldots x_m^{a_m}$	$Y^* = \lg(Y);\ X^* = \left\| a_j \lg x_{ij} \right\|$	Модель с быстрой реакцией на изменения в данных, описывает процессы с этапами ускорения (замедления)
Экспоненциальная: $\hat{y} = e^{a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m}$	$Y^* = \ln(Y);\ X^* = X$	Процессы, имеющие этапы ускоренного (замедленного) развития
Сложная экспоненциальная: $\hat{y} = 1/(1 + e^{a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m})$	$Y^* = \ln(Y^{-1} - e);\ X^* = X$	— " —
Обратная: $\hat{y} = 1/(a_0 + a_1 x_1 + a_2 x_2^2 + \ldots + a_m x_m^m)$	$Y^* = Y^{-1};\ X^* = X$	Процессы, имеющие всплески в развитии, обусловленные постепенным накоплением количественных изменений в прошлом

в) MAPE — коэффициент аппроксимации, или средняя относительная величина модельной ошибки, $\text{MAPE} = \frac{1}{n}\sum \frac{|y - \hat{y}|}{\hat{y}} \cdot 100$.

Данный критерий принимает известные пороговые значения:

Оценка MAPE, %	*Характеристика качества регрессионной модели*
< 10	Высокая точность
10-20	Хорошая точность
20-50	Удовлетворительная точность
> 50	Неудовлетворительная точность

г) *F*-критерий Фишера,

$$F_{\text{н}} = \frac{(XB)'(XB)/(m+1)}{(Y - XB)'(Y - XB)/(n-m-1)}.$$

Наблюденные значения F-критерия $F_{\text{н}}$ сравнивают с критическими (табличными) при заданном уровне значимости α и числе степеней свободы $\nu_1 = m+1$ и $\nu_2 = n-m-1$. Надежность регрессионной модели подтверждается при условии, что $F_{\text{н}} > F_{\text{табл}}$.

• Статистическая оценка надежности коэффициентов регрессии:

производится при помощи *t*-критерия Стьюдента. Для некоторого *j*-го параметра регрессионной модели наблюденное значение *t*-критерия определяется по формуле:

$$t_{\text{н}(j)} = b_j / \hat{S}_j,$$

где $\hat{S}_j$ — средняя ошибка *j*-го коэффициента регрессии $\hat{S}_j = (S_{\text{ост}}^2 \cdot c_{jj})^{1/2}$, дисперсионная характеристика $\hat{S}_{\text{ост}}^2 = (Y - XB)'(Y - XB)/(n-m-1)$, а c_{jj} — это соответствующие (b_j) диагональные элементы матрицы $(X'X)^{-1}$, при условии, что X — матрица, расширенная за счет введения в нее единичного вектора, т.е. $X = X_2$ (см. с.217).

Наблюденное значение *t*-критерия также сравнивается с табличным при заданном α и числе степеней свободы $\nu = n - m - 2$. Значимость коэффициента регрессии подтверждается, когда $t_{\text{н}} > t_{\text{табл}}$. При этом могут устанавливаться допустимые пределы колеблемости статистической оценки *j*-го параметра:

$\hat{b}_j - t_{\text{табл}}\hat{S}_j \le b_j \le \hat{b}_j + t_{\text{табл}}\hat{S}_j$.

• Статистические оценки достоверности коэффициентов корреляции:

частные и парные коэффициенты корреляции проверяются при помощи t-критерия Стьюдента:

$$t_{\text{н}} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-m-2}.$$

При известном пороговом значении t-критерия ($t_{a,n-m-2}$) значимость коэффициентов корреляции подтверждается, если $t_{\text{н}} > t_{\text{табл}}$.

Коэффициент множественной детерминации (корреляции) оценивается с использованием F-критерия Снедекора:

$$F_{\text{н}} = \frac{(n-m)R_0^2}{(m-1)(1-R_0^2)}.$$

Значимость R_0^2 считается подтвержденной при $F_{\text{н}} > F_{\text{табл}}$. $F_{\text{табл}}$ находят по таблицам при известном уровне α и числе степеней свободы $\nu_1 = m-1$; $\nu_2 = n-m-1$.

В главе изложение материала построено по принципу: от простого к более сложному. Вначале рассматривается простейший случай линейной зависимости двух переменных Y и X, здесь Y выступает как зависимая переменная, определяемая факторной переменной X и случайной ошибкой U: $Y = b_0 + b_1 X + U$.

Y характеризуется так же, как линейная функция объясняющей величины X и ошибки регрессии, или как величина, определяемая систематической величиной X и стохастически аддитивной случайной величиной U. В §6.2 показано решение исходной задачи: как на основе известных данных выборочного обследования Y_i и X_i ($i = 1,2,...,n$) произвести статистическое оценивание неизвестных параметров регрессионного уравнения b_0, b_1, а также дисперсии ошибки σ^2.

В §6.3 представлены методы проверки статистических гипотез, с помощью которых оцениваются качество регрессионных моделей и параметрические значения: b_0, b_1, σ^2.

В §6.4 теоретические положения обобщаются для случаев с любым числом факторных переменных X, т.е. когда систематическая часть объединяет m независимых переменных $X: X_j$, $j = 1,2,...,m$.

Завершающий §6.5 посвящен вопросам распространения выводов теории линейной регрессии на нелинейные регрессионные модели.

В конце главы читателю предлагаются задачи и иллюстративный материал, которые должны помочь более глубокому пониманию и овладению техникой вычислений при использовании методов регрессионного анализа на практике.

6.2. Линейная регрессия — классический случай зависимости двух переменных X и Y

В регрессионном анализе в качестве исходной рассматривается линейная модель вида:

$$Y_i = b_0 + b_1 x_i + U_i, \quad U_i \sim N(0,\sigma^2), \quad i = 1,2,...,n, \tag{6.1}$$

где U_i — случайное слагаемое или ошибка модели,

$U_i = Y_i - b_0 - b_1 X_i$.

Функция плотности вероятностей для случайного слагаемого будет:

$$f(U_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{U_i^2}{2\sigma^2}} \quad \text{или} \quad f(U_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - b_0 - b_1 X_i)^2}{2\sigma^2}}.$$

Для определения трех неизвестных параметров регрессионной модели $b_0, b_1 \in R$ и σ^2 по данным некоторой случайной выборки может быть использована эффективная функция максимального правдоподобия:

$$L(b_0, b_1, \sigma^2) = L = \prod_{i=1}^{n} f(U_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - b_0 - b_1 X_i)^2}{2\sigma^2}}.$$

Эта же функция после логарифмирования принимает вид:

$$L(b_0, b_1, \sigma^2) = L = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2\right\}. \tag{6.2}$$

Символом $\sum$ здесь и в последующем обозначается обычная сумма $\sum_{i=1}^{n}$.

Необходимым условием для параметрической оценки методом максимального правдоподобия является решение системы нормальных уравнений:

$$(1)\quad \frac{\partial L}{\partial b_0}=\frac{1}{\sigma^2}\left\{\sum(Y_i-b_0-b_1X_i)\right\},$$

$$(2)\quad \frac{\partial L}{\partial b_1}=\frac{1}{\sigma^2}\left\{\sum(Y_i-b_0-b_1X_i)X_i\right\},$$

$$(3)\quad \frac{\partial L}{\partial \sigma^2}=\frac{-n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\left\{\sum(Y_i-b_0-b_1X_i)^2\right\},$$

Решая уравнения (1), (2) при любых $\sigma^2>0$, будем иметь:

$$\hat{b}_1=\frac{\overline{YX}-\overline{Y}\cdot\overline{X}}{\overline{XX}-\overline{X}\cdot\overline{X}} \text{ и } \hat{b}_0=\overline{Y}-\hat{b}_1\overline{X}.$$

Средние значения переменных $\overline{X},\overline{Y},\overline{XY}$ находят по известным формулам:

$$\overline{X}=\frac{1}{n}\sum X_i,\ \overline{Y}=\frac{1}{n}\sum Y_i,\ \overline{XX}=\frac{1}{n}\sum X_i^2 \text{ и } \overline{YX}=\frac{1}{n}\sum Y_iX_i.$$

Решение дифференциального уравнения (3) по $\hat{\sigma}^2$ позволяет при заданных модельных параметрах $\hat{b}_0$ и $\hat{b}_1$ получить оценку ошибки регрессии:

$$\hat{\sigma}^2=\frac{\sum(Y_i-\hat{b}_0-\hat{b}_1X_i)(Y_i-\hat{b}_0-\hat{b}_1X_i)}{n}=\frac{\sum\hat{U}_i^2}{n},$$

$$\hat{U}_i:=Y_i-\hat{b}_0-\hat{b}_1X_i.$$

С определением второй производной для L-функции появляется возможность, используя матрицу Гессе для функции максимального правдоподобия[1], обобщить алгоритм статистического оценивания параметров регрессионной модели.

[1] О матрице Гессе см.: *Пшеничный Б.Н., Данилин Ю.М.* Численные методы в экстремальных задачах. — М., 1975; *Фадеев Д.К., Фадеева В.Н.* Вычислительные методы в линейной алгебре. 2-е изд. — М., 1963. (*Прим. пер.*)

Пусть $H=\dfrac{\partial L^2}{\partial\theta_k\,\partial\theta_l}$ $(k,l=1,2,3,\ \theta'=(b_0,b_1,\sigma^2))$.

Введем новые обозначения для нормальных уравнений:

$$N_1:=\sum(Y_i-b_0-b_1X_i)=0;$$

$$N_2:=\sum(Y_i-b_0-b_1X_i)X_i=0.$$

Теперь можем построить саму матрицу Гессе для вторых производных:

$$\left(\begin{array}{cc|c} \frac{n}{\sigma^2} & \frac{\sum X_i}{\sigma^2} & -\frac{N_1}{(\sigma^2)^2} \\ \frac{\sum X_i}{\sigma^2} & \frac{\sum X_i^2}{\sigma^2} & -\frac{N_2}{(\sigma^2)^2} \\ \hline -\frac{N_1}{(\sigma^2)^2} & -\frac{N_2}{(\sigma^2)^2} & \frac{n}{(\sigma^2)^2}-\frac{1}{(\sigma^2)^3}\sum\hat{U}_i^2 \end{array}\right)=:\frac{1}{\sigma^2}\left(\begin{array}{cc|c} M & & 0 \\ \hline & & \\ 0 & & \frac{n}{\sigma^2}-\frac{1}{(\sigma^2)^2}\hat{U}'\hat{U} \end{array}\right)=H.$$

Матрица H отвечает известному критерию главных миноров для существующего максимума. Матрица M в H характеризуется как матрица моментов:

$$M=\begin{pmatrix} n & n\overline{X} \\ n\overline{X} & n\overline{XX} \end{pmatrix}=n\begin{pmatrix} 1 & \overline{X} \\ \overline{X} & \overline{XX} \end{pmatrix}.$$

Оценивание параметров регрессионной модели по центрированным данным. Исходные данные, используемые в регрессионном анализе, с целью упрощения вычислений могут быть центрированы, т.е. вместо x_{ij} используются отклонения от соответствующих средних значений. Обозначим центрированные величины строчными буквами y_i и x_i:

$$x_i=X_i-\overline{X};\ y_i=Y_i-\overline{Y}.$$

В этом случае формулы для определения модельных параметров преобразуются:

$$\tilde{b}_1=\frac{\sum x_iy_i}{\sum x_i^2},\ \tilde{b}_0=\overline{Y}-\tilde{b}_1\overline{X},\ \tilde{\sigma}^2=\frac{\sum(Y_i-\tilde{b}_0-\tilde{b}_1X_i)}{n}=\frac{\sum\hat{U}_i^2}{n},$$

$\hat{U}_i = Y_i - \tilde{b}_0 - \tilde{b}_1 X_i$.

Введенные при записи параметрических данных символы ~ над буквой соответствуют обозначениям, принятым в эконометрике, на самом деле они могут быть опущены, так как гипотетически $\tilde{b}_0 = \hat{b}_0$, $\tilde{b}_1 = \hat{b}_1$, $\tilde{\sigma}^2 = \hat{\sigma}^2$.

Простейшим доказательством правомерности формул для нахождения значений b_0, b_1 и σ^2 служит соблюдение в регрессионном уравнении равенства его левой и правой частей, а также выбор самого решения уравнения методом максимального правдоподобия:

$$Y_i - \overline{Y} = b_0 - b_0 + b_1 X_i - b_1 \overline{X} + U_i,\ U_i \sim N(0,\sigma^2) \quad (i = 1, 2, ..., n) \qquad (6.1')$$

$$y_i = b_1 x_i + U_i,\ U_i \sim N(0,\sigma^2) \quad (i = 1, 2, ..., n) \qquad (6.1'')$$

Использование метода максимального правдоподобия в данном случае означает переход к анализу функции L с двумя неизвестными параметрами (b_1 и σ^2):

$$L(b_1,\sigma^2) = L = \prod_{i=1}^{n} f(U_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - b_1 x_i)^2}{2\sigma^2}}.$$

Эта же функция максимального правдоподобия после логарифмирования запишется:

$$L(b_1,\sigma^2) = L = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left\{\sum (y_i - b_1 x_i)^2\right\}. \qquad (6.2')$$

Статистическое оценивание параметров регрессионной модели методом максимального правдоподобия предполагает построение системы нормальных уравнений:

$$(2')\ \frac{\partial L}{\partial b_1} = \frac{1}{\sigma^2}\left\{\sum (y_i - b_1 x_i) x_i\right\}.$$

$$(3')\ \frac{\partial L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left\{\sum (y_i - b_1 x_i)^2\right\}.$$

Решение уравнений (2') и (3') относительно $\tilde{b}_1$ и $\tilde{\sigma}^2$ приводит к результатам, соответствующим анализу функции L для случая с тремя неизвестными параметрами: $\hat{b}_0$, $\hat{b}_1$ и $\hat{\sigma}^2$. При этом оценка b_0 может производиться при помощи полученного выше дифференциального уравнения (1).

Первый особый случай: регрессионная модель не содержит параметра b_0. Если в исходной регрессионной модели отсутствует свободный член b_0, то задача параметрического оценивания сводится к определению b_1 и σ^2, соответственно как и в случае с центрированными данными:

$Y_i = b_1 X_i + U_i$, $U_i \sim N(0,\sigma^2)$, $(i = 1, 2, ..., n)$.

Решение такого уравнения показано выше:

$$\hat{b}_1 = \frac{\sum X_i Y_i}{\sum X_i^2},\ \hat{\sigma}^2 = \frac{\sum (Y_i - \hat{b}_1 X_i)^2}{n} = \frac{\sum \hat{U}_i^2}{n},\ \hat{U}_i := Y_i - \hat{b}_1 X_i.$$

Имея дело со значениями центрированных переменных, следует учитывать, что их сумма всегда равна нулю:

$\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_i = 0$, тогда как в общем случае, как правило, $\sum_{i=1}^{n} Y_i \neq 0$ и $\sum_{i=1}^{n} X_i \neq 0$.

Второй особый случай: регрессионная модель содержит один коэффициент регрессии b_1. Если регрессионная модель содержит только один из двух коэффициентов регрессии b_1 и не включает b_0, то ее решение сводится к задаче так называемого оценивания средней:

$$Y_i = b_0 + U_i,\ U_i \sim N(0,\sigma^2) \quad (i = 1, 2, ..., n). \qquad (6.1'')$$

В этом случае параметр $\hat{b}_0$ есть среднее значение случайной величины Y и, определяя среднюю для $Y(\mu)$, соответственно находим b_0. Покажем это: пусть случайная величина Y имеет плотность вероятности:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Для случайной выборки $\{Y_1, Y_2, ..., Y_n\}$ функция максимального правдоподобия имеет вид:

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \qquad L = -n\ln\left(\sqrt{2\pi}\right) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Экстремальные значения функции L найдем решением дифференциальных уравнений:

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{n} \frac{(y_i - \mu)}{\sigma^2} = 0,$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2(\sigma^2)^2} = 0.$$

Получаем статистические оценки:

$$\hat{\mu} = \hat{b}_0 = \sum_{i=1}^{n} \frac{y_i}{n} = \bar{Y} \text{ и } \hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{b}_0)^2}{n} = \frac{\sum \hat{U}_i^2}{n}, \ \hat{U}_i := Y_i - \hat{b}_0.$$

Оценивание методом наименьших квадратов. Для нормальных уравнений (1), (2) (см. с. 224) в качестве альтернативных могут быть предложены решения, базирующиеся на значительно более простых исходных посылках. Например, не обязательно предположение об известном законе распределения изучаемых случайных величин X и Y и соответствии этого закона нормальному. Для исходной модели: $Y_i = b_0 + b_1 X_i + U_i$ достаточно принять, что ненаблюдаемая ошибка U_i является любой случайной величиной с математическим ожиданием, равным нулю:

$$E(U_i) = 0, \ E(U_i U_s) = \begin{cases} \operatorname{var}(U_i) = \sigma^2, & s = i \\ \operatorname{cov}(U_i, U_s) = 0, & s \neq i \end{cases} \quad s, i = 1, 2, .., n.$$

По выборочным данным с оцененными параметрами $\hat{b}_0$ и $\hat{b}_1$ ошибка регрессионной модели будет вполне определенной величиной: $\hat{U}_i := Y_i - \hat{b}_0 - \hat{b}_1 X_i$.

При этом остается выдвинуть условие минимизации ошибки, что фактически означает выдвижение принципа наилучшей адаптации модели к эмпирическим наблюдениям. Метод, реализующий такой принцип, называется *методом наименьших квадратов*:

$$Q(b_0, b_1) := \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 \to \underset{b_0, b_1}{\operatorname{Min}},$$

или $U_i := Y_i - b_0 - b_1 X_i$, $Q(b_0, b_1) := \sum_{i=1}^{n} U_i^2 \to \underset{b_0, b_1}{\operatorname{Min}}$.

Условие минимизации компоненты U_i в регрессионной модели означает правомерность построения дифференциальных уравнений для первой производной гипотетической функции:

$$\frac{\partial Q}{\partial b_0} = -2\left\{\sum (Y_i - b_0 - b_1 X_i)\right\} = 0;$$

$$\frac{\partial Q}{\partial b_1} = -2\left\{\sum (Y_i - b_0 - b_1 X_i) X_i\right\} = 0.$$

Для второй производной гипотетической функции $Q(b_0, b_1)$ построим матрицу Гессе H:

$$H = \begin{pmatrix} \frac{\partial^2 Q}{\partial b_0 \partial b_0} & \frac{\partial^2 Q}{\partial b_0 \partial b_1} \\ \frac{\partial^2 Q}{\partial b_1 \partial b_0} & \frac{\partial^2 Q}{\partial b_1 \partial b_1} \end{pmatrix} = \begin{pmatrix} 2n & 2\sum X_i \\ 2\sum X_i & 2\sum X_i X_i \end{pmatrix},$$

откуда

$$H = 2M := 2\begin{pmatrix} n & n\bar{X} \\ n\bar{X} & n\overline{XX} \end{pmatrix} = 2n\begin{pmatrix} 1 & \bar{X} \\ \bar{X} & \overline{XX} \end{pmatrix}.$$

С учетом правила нахождения детерминанта минимизация функциональных величин Q достигается при следующих значениях элементов матрицы H:

$$h_{11} = 1 > 0; \ h_{22} = \overline{XX} > 0; \ h_{11} h_{22} - h_{21} h_{12} = \overline{XX} - \bar{X}^2 < 0.$$

Последнее неравенство называют неравенством Шварца. Зная вторую производную функции Q, статистическую оценку дисперсии ошибки при известных параметрах b_0, b_1 можно установить достаточно просто:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2}{n}.$$

Таким образом, статистическое оценивание параметров нормальной регрессионной модели методом наименьших квадратов и методом максимального правдоподобия дает одни и те же результаты.

Метод моментов. Для нормальных уравнений вида 1, 2 (см. с. 224), построенных для линейной регрессионной модели, решения могут быть найдены не только методом максимального правдоподобия или наименьших квадратов, но также методом моментов.

Оставим в силе нестрогие посылки метода наименьших квадратов (МНК-метода) о существовании некоторого любого

закона распределения случайных величин X и Y, и пусть имеются оба первых момента для ошибки модели U, причем объясняющая величина X не коррелирует с U, т.е.

$E(X_i U_s)=0$ для всех i и s $(i,s=1,2,...,n)$.

Тогда имеем исходное регрессионное уравнение вида:

$Y_i = b_0 + b_1 X_i + U_i$,

после перемножения левой и правой частей данного уравнения на X_i имеем:

$X_i Y_i = b_0 X_i + b_1 X_i X_i + U_i X_i$.

В первом регрессионном уравнении просуммируем все наблюдения и заменим эмпирические моменты теоретическими:

$$\frac{1}{n}\sum Y_i = b_0 + b_1 \frac{1}{n}\sum X_i + \frac{1}{n}\sum U_i ;$$

$$E(\overline{Y}) = b_0 + b_1 E(\overline{X}) + E(\overline{U}) \approx \overline{Y} = b_0 + b_1 \overline{X} .$$

Для второго регрессионного уравнения получим аналогичные результаты:

$$\frac{1}{n}\sum X_i Y_i = b_0 \frac{1}{n}\sum X_i + b_1 \frac{1}{n}\sum X_i X_i + \frac{1}{n}\sum U_i X_i ;$$

$$E(\overline{XY}) = b_0 E(\overline{X}) + b_1 E(\overline{XX}) + E(\overline{XU}) \approx \overline{XY} = b_0 \overline{X} + b_1 \overline{XX} .$$

Тем самым оба первых регрессионных уравнения получают новую форму записи — через математические ожидания случайных величин. Теперь последовательно найдем статистические оценки параметров b_0, b_1, σ^2:

$$\hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X} ,\ \hat{b}_1 = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{\overline{YX} - \overline{Y}\ \overline{X}}{\overline{XX} - \overline{X}\ \overline{X}} ;\ \hat{\sigma}^2 = \frac{1}{n}\sum (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2 .$$

Обратная регрессия. Многообразие существующих подходов в оценивании параметров регрессионной модели позволяет рассматривать случай, когда вместо стохастической определенности, предполагающей нормальное распределение и причинную обусловленность случайных величин (X *и* U *определяют* Y), согласно *принципу лезвия Оккама* постулируется просто, насколько это возможно, линейная связь[1]. При этом становится возможным анализ так называемой обратной регрессии.

[1] Принцип лезвия Оккама — принцип минимальной достаточности. (*Прим. пер.*)

Пусть имеет место исходная регрессионная модель:

$Y_i = b_0 + b_1 X_i + U_i$, $E(U_i)=0$, $\mathrm{var}(U_i) = E(U_i^2) = \sigma^2$,

$\mathrm{cov}(U_i, U_s) = E(U_i U_s) = 0$, $i \neq s$.

Очевидно, что данную модель можно переписать с учетом замены зависимой переменной с Y на X:

$$X_i = b_0' + b_1' Y_i + V_i ,\ E(V_i)=0 ,\ \mathrm{var}(V_i) = E(V_i^2) = \rho^2 , \qquad (6.3)$$

$\mathrm{cov}(V_i, V_s) = E(V_i V_s) = 0$, $i \neq s$.

Модель вида (6.3) будет обратной относительно исходной модели (6.1).

При условии использования данных, отклоняющихся от средних значений (т. е. центрированных данных) $x_i := X_i - \overline{X}$, $y_i := Y_i - \overline{Y}$, или в векторной записи

$x' = (x_1, x_2, ..., x_n)$, $y' = (y_1, y_2, ..., y_n)$, с $\hat{U}' = (\hat{U}_1, \hat{U}_2, ..., \hat{U}_n)$,

$\hat{V}' = (\hat{V}_1, \hat{V}_2, ..., \hat{V}_n)$,

и при следующих определениях для оценок случайных величин, параметров и ошибки:

$\hat{y} := \hat{b}_1 x$, $\hat{x} := \hat{b}_1' \hat{y}$, $\hat{U} := \hat{y} - y$, $\hat{V} := \hat{x} - x$

найдем модельные параметры для обычной регрессии:

$$\hat{b}_1 = \frac{y'x}{x'x} ,\ \hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X} ,\ \hat{\sigma}^2 = \frac{\hat{U}'\hat{U}}{n}$$

и соответственно для обратной регрессии:

$$\hat{b}_1' = \frac{x'y}{y'y} ,\ \hat{b}_0' = \overline{X} - \hat{b}_1' \overline{Y} ,\ \hat{\rho}^2 = \frac{\hat{V}'\hat{V}}{n} .$$

Так как в регрессионном анализе используются многочисленные приемы статистических исчислений, представляется важной апробация его алгоритмов вручную. Отметим, что практически на всех больших и малых вычислительных машинах имеется программное обеспечение, реализующее разнообразные методы регрессии. В то же время на небольших массивах данных расчеты могут быть проведены и без технических средств.

Пример 6.1. Пусть имеются данные случайной выборки: $\{(Y_i, X_i)\}$, $i = 1, 2, 3, 4$ и $\{(60, 10), (75, 20), (85, 30), (100, 40)\}$.

Определим неизвестные параметры для обычной регрессии:

$$\hat{b}_1 = \frac{y'x}{x'x} = \frac{650}{500} = 1{,}3 ,\ \hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X} = 80 - 1{,}3 \cdot 25 = 47{,}5 ,$$

$$\hat{\sigma}^2 = \frac{\hat{U}'\hat{U}}{n} = \frac{5}{4} = 1{,}25$$

и для обратной регрессии

$$\hat{b}_1' = \frac{x'y}{y'y} = \frac{650}{850} = 0{,}765\,,\quad \hat{b}_0' = \bar{X} - \hat{b}_1'\bar{Y} = 25 - 0{,}765 \cdot 80 = -36{,}2\,,$$

$$\hat{\rho}^2 = \frac{\hat{V}'\hat{V}}{n} = \frac{3{,}06}{4} = 0{,}735\,.$$

6.3. Свойства статистических оценок параметров регрессионной модели

При формировании суждений о статистических оценках линейной регрессии весьма важным представляется подход Гаусса — Маркова.

Подход Гаусса — Маркова. *Утверждение*: статистические оценки параметров $\hat{b}_0$ и $\hat{b}_1$ являются наилучшими, линейными и несмещенными (правило НЛН, или по общепринятой аббревиатуре — BLU).

Прежде всего отметим, что $\hat{b}_0$ и $\hat{b}_1$ — наилучшие оценки, так как алгоритм их вычислений определяется при помощи оптимизирующих методов: максимального правдоподобия и наименьших квадратов.

Свойство 1: *Линейность статистической оценки* $\hat{b}_1$

Оценка $\hat{b}_1$ есть линейная функция наблюдаемой величины Y:

$$\hat{b}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} - \frac{\bar{Y}\sum x_i}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2},$$

так как $\sum x_i = 0$ (центрированные значения).

Введем обозначение: $w_i = \dfrac{x_i}{\sum x_i^2}$, тогда $\hat{b}_1 = \sum w_i x_i$.

При этом имеем: $\sum w_i = 0$, так как $\sum x_i = 0$, $\sum w_i^2 = \dfrac{1}{\sum x_i^2}$, $\sum w_i x_i = 1$, $\sum w_i \bar{x}_i = 1$.

Имея в качестве исходной модель вида: $Y_i = b_0 + b_1 X_i$ и принимая во внимание, что $\sum w_i = 0$, $\sum w_i x_i = 1$, запишем

$$\hat{b}_1 = b_0 \sum w_i + b_1 \sum w_i X_i + \sum w_i U_i = b_1 + \sum U_i w_i\,,$$

т.е. оценка $\hat{b}_1$ является линейной функцией от зависимой переменной Y_i и ошибки модели U_i.

Свойство 2: *Оптимальность математического ожидания оценки* $\hat{b}_1$

$$E(\hat{b}_1) = b_1 + E\left(\sum w_i U_i\right) = b_1 + \sum w_i E(U_i) = b_1\,.$$

Свойство 3: *Оценка* $\hat{b}_1$ *имеет наименьшую дисперсию*

$$\begin{aligned}\operatorname{var}(\hat{b}_1) &= E\left\{\left[\hat{b}_1 - E(\hat{b}_1)\right]^2\right\} = E\left\{\left[\hat{b}_1 - b_1\right]^2\right\} = E\left\{\sum w_i U_i\right\}^2 = \\ &= E(w_1^2 U_1^2 + w_1 w_2 U_1 U_2 + \ldots + w_1 w_n U_1 U_n + \\ &+ w_2 w_1 U_2 U_1 + w_2^2 U_2^2 + \ldots + w_2 w_n U_2 U_n + \\ &\vdots \\ &+ w_n w_1 U_n U_1 + w_n w_2 U_n U_2 + \ldots + w_n^2 U_n^2) = \\ &= E\left[\sum_{i=1}^{n} w_i^2 U_i^2\right] = \sum_{i=1}^{n} w_i^2 E[U_i^2] = \sigma^2 \sum_{i=1}^{n} w_i^2\,.\end{aligned}$$

Предположим, что имеется некоторая линейная оценка $\tilde{b}_1$, отклоняющаяся от $\hat{b}_1$ таким образом, что весовое значение w_i необходимо скорректировать на величину d_i:

$$\tilde{b}_1 = \sum (w_i + d_i) Y_i = \sum c_i Y_i\,,\ \text{где}\ c_i = w_i + d_i\,.$$

При этом для математического ожидания $\tilde{b}_1$ остается в силе:

$$\begin{aligned}E(\tilde{b}_1) &= E\left\{\sum c_i (b_0 + b_1 X_i + U_i)\right\} = b_0 \sum c_i + b_1 \sum c_i X_i + \sum c_i E(U_i) = \\ &= b_0 \sum c_i + b_1 \sum c_i X_i = b_1\,,\end{aligned}$$

и весовые коэффициенты c_i и d_i сохраняют свои свойства:

$$\sum c_i = 0;\ \sum c_i X_i = 1 \ \text{и}\ \sum d_i = 0;\ \sum d_i X_i = 1\,.$$

Следуя определению x_i как центрированной величины, имеем:

$$\sum d_i X_i = \sum d_i (x_i + \bar{X}) = \sum d_i x_i + \bar{X}\sum d_i = \sum d_i x_i .$$

Таким образом, для дисперсии $\tilde{b}_1$ можем записать:

$$\mathrm{var}(\tilde{b}_1) = E\left\{\left[\tilde{b}_1 - E(\tilde{b}_1)\right]^2\right\} = E\left\{\left[\tilde{b}_1 - b_1\right]^2\right\}$$ — с учетом несмещенности;

$$\mathrm{var}(\tilde{b}_1) = E\left[\sum c_i U_i\right]^2$$ — с учетом линейности;

$$\mathrm{var}(\tilde{b}_1) = E\left\{\left[\sum (w_i + d_i)U_i\right]^2\right\} = E\left\{\left[\sum w_i U_i + \sum d_i U_i\right]^2\right\}$$ — принимая во внимание, что $c_i = w_i + d_i$.

Из приведенных равенств следует, что дисперсия параметра $\tilde{b}_1$ может рассматриваться как некоторый линейный оператор математического ожидания над суммами значений определенных случайных величин и имеющий стохастическую природу:

$E(U_i) = 0$, $E(U_i U_s) = \sigma^2$, если $i = s$, и $E(U_i U_s) = 0$, если $i \neq s$, таким образом, получаем

$$E\left\{\left[\sum w_i U_i + \sum d_i U_i\right]^2\right\} = \sigma^2 \sum w_i^2 + \sigma^2 \sum d_i^2 = \mathrm{var}(\hat{b}_1) + \sigma^2 \sum d_i^2 .$$

Доказательство свойств для параметра b_0 строится аналогично.

Смещенность оценки дисперсии ошибки регрессионной модели. *Утверждение*: статистическая оценка

$\hat{\sigma}^2 = \frac{1}{n}\sum (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2$ не совпадает с истинными значениями σ^2.

□ *Доказательство*:

Из $\hat{U}_i = Y_i - \hat{Y}_i = Y_i - \hat{b}_0 - \hat{b}_1 X_i$ и $Y_i = b_0 + b_1 X_i + U_i$, $U_i \sim \boldsymbol{N}(0, \sigma^2)$, следует $\hat{U}_i = b_0 - \hat{b}_0 + (b_1 - \hat{b}_1)X_i + U_i$.

С учетом того, что $\bar{Y} = b_0 + b_1 \bar{X} + \bar{U}$ и $\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$, получим $b_0 - \hat{b}_0 + (b_1 - \hat{b}_1)\bar{X} = \bar{U}$.

Тогда $\hat{U}_i = (b_1 - \hat{b}_1)(X_i - \bar{X}) + U_i - \bar{U}$ и

$$\hat{U}_i^2 = (b_1 - \hat{b}_1)^2 (X_i - \bar{X})^2 + (U_i - \bar{U})^2 + 2(b_1 - \hat{b}_1)(X_i - \bar{X})(U_i - \bar{U});$$

$$\sum \hat{U}_i^2 = (b_1 - \hat{b}_1)^2 \sum (X_i - \bar{X})^2 + \sum (U_i - \bar{U})^2 + 2(b_1 - \hat{b}_1)\sum (X_i - \bar{X})(U_i - \bar{U}) =$$

$$= (b_1 - \hat{b}_1)^2 \sum (X_i - \bar{X})^2 + \sum (U_i - \bar{U})^2 + 2(b_1 - \hat{b}_1)\left[\sum U_i (X_i - \bar{X}) - \sum \bar{U}(X_i - \bar{X})\right].$$

Так как $\sum U_i (X_i - \bar{X}) = -(b_1 - \hat{b}_1)\sum (X_i - \bar{X})^2$ и $\sum \bar{U}(X_i - \bar{X}) = 0$, будем иметь

$$\sum \hat{U}_i^2 = (b_1 - \hat{b}_1)^2 \sum (X_i - \bar{X})^2 + \sum (U_i - \bar{U})^2 - 2(b_1 - \hat{b}_1)^2 \sum (X_i - \bar{X})^2 =$$

$$= \sum (U_i - \bar{U})^2 - (b_1 - \hat{b}_1)^2 \sum (X_i - \bar{X})^2 = \hat{\sigma}^2 .$$

Чтобы перепроверить предположение о смещенности оценки $\hat{\sigma}^2$, найдем ее математическое ожидание:

$$E(\hat{\sigma}^2) = E\left(\sum \hat{U}_i^2\right) = \sum E(U_i - \bar{U})^2 - E\left[(b_1 - \hat{b}_1)^2 \sum (X_i - \bar{X})^2\right] =$$

$$= \sum E(U_i^2) + \sum E(\bar{U}^2) - 2\sum E(U_i \bar{U}) - \sigma^2 =$$

$$= n\sigma^2 + nE(\bar{U}^2) - \frac{2}{n}\sum E\left(U_i^2 + \sum_{i \neq s} U_i U_s\right) - \sigma^2 =$$

$$= n\sigma^2 + nE\left[\frac{1}{n^2}\sum \left(U_i^2 + \sum_{i \neq s} U_i U_s\right)\right] - \frac{2}{n}\sum \left[E(U_i^2) - \sum_{i \neq s} E(U_i U_s)\right] - \sigma^2 =$$

$$= n\sigma^2 - \frac{1}{n}\sum \left[E(U_i^2) - \sum_{i \neq s} E(U_1 U_2)\right] - \sigma^2 = n\sigma^2 - \frac{1}{n} n\sigma^2 - 0 - \sigma^2 = (n-2)\sigma^2 .$$

Очевидно, что несмещенную оценку σ^2 можно рассчитать по формулам:

$$\tilde{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{U}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2 .$$ ■

Распределение и критериальная проверка надежности статистических оценок параметров регрессионной модели. В соответствии с подходом Гаусса — Маркова при оценке параметров регрессионной модели выдвигаются следующие предположения:

$$\hat{b}_0 \sim \boldsymbol{N}\left(b_0, \frac{\sigma^2 \sum X_i^2}{n\sum x_i^2}\right), \quad \hat{b}_1 \sim \boldsymbol{N}\left(b_1, \frac{\sigma^2}{\sum x_i^2}\right).$$

Так как σ^2 остается величиной неизвестной, произведем замену и вместо дисперсии ошибки будем рассматривать оценки ее математического ожидания, которые можно получить из анализа распределения χ^2-статистики, а также двух t-статистических распределений Стьюдента, каждое из которых имеет число степеней свободы $n-2$:

$$t(b_0) := \left(\hat{b}_0 - b_0\right)\frac{\sqrt{n\sum x_i^2}}{\hat{\sigma}\sqrt{\sum X_i^2}} \sim t_{n-2} \text{ и } t(b_1) := \left(\hat{b}_1 - b_1\right)\frac{\sqrt{\sum x_i^2}}{\hat{\sigma}} \sim t_{n-2},$$

$$z := \sum \frac{\hat{U}_i^2}{\sigma^2},\ \hat{U}_i := Y_i - \hat{b}_0 - \hat{b}_1 X_i \text{ или } \hat{U}_i := y_i - \hat{b}_1 X_i,\ z \sim \chi^2_{n-2}.$$

Симметричные доверительные интервалы с доверительной вероятностью $1-\alpha$. Известный закон распределения статистических оценок параметров b_0, b_1 и σ^2 позволяет легко для них устанавливать доверительные интервалы и критериальные характеристики надежности. При допустимом уровне ошибки в выводах, равном α, и уровне доверительной вероятности $1-\alpha$ будем иметь доверительные интервалы:

- для параметра b_0

$$\hat{b}_0 \pm t_{\alpha/2}\hat{\sigma}\frac{\sqrt{\sum X_i^2}}{\sqrt{n\sum x_i^2}} = \left[\hat{b}_0 + t_{\alpha/2}\frac{\hat{\sigma}\sqrt{\sum X_i^2}}{\sqrt{n\sum x_i^2}},\ \hat{b}_0 - t_{\alpha/2}\frac{\hat{\sigma}\sqrt{\sum X_i^2}}{\sqrt{n\sum x_i^2}}\right];$$

- для параметра b_1

$$\hat{b}_1 \pm t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n\sum x_i^2}} = \left[\hat{b}_1 + t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{\sum x_i^2}},\ \hat{b}_1 - t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}\right];$$

- для дисперсии ошибки регрессионной модели σ^2:

$$\left[\frac{\hat{\sigma}^2(n-2)}{\chi^2_{n-2,1-\alpha/2}}, \frac{\hat{\sigma}^2(n-2)}{\chi^2_{n-2,\alpha/2}}\right].$$

Статистические гипотезы для b_0, b_1 **и** σ^2. Знание представленных выше статистических критериев и предполагаемый уровень вероятности допустить ошибку в выводах (α) позволяют сформулировать следующие решающие правила:

Симметричные гипотезы: Семейство гипотез имеет вид:

H_0: $\tilde{b}_0 = b_0$ против H_1: $\tilde{b}_0 \neq b_0$;

H_0: $\tilde{b}_1 = b_1$ против H_1: $\tilde{b}_1 \neq b_1$;

H_0: $\sigma^2 = \sigma_0^2$ против H_1: $\sigma^2 \neq \sigma_0^2$.

Гипотезы этого же семейства, записанные для существующих доверительных интервалов, будут:

$$P_{H_0}\left[-t_{n-2,1-\frac{\alpha}{2}} \le t(b_0) \le t_{n-2,1-\frac{\alpha}{2}}\right] = 1-\alpha;$$

$$P_{H_0}\left[-t_{n-2,1-\frac{\alpha}{2}} \le t(b_1) \le t_{n-2,1-\frac{\alpha}{2}}\right] = 1-\alpha;$$

$$P_{H_0}\left[\chi^2_{n-2,\frac{\alpha}{2}} \le \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \le \chi^2_{n-2,1-\frac{\alpha}{2}}\right] = 1-\alpha.$$

Схема проверки статистических гипотез во всех случаях остается общей:

Левосторонняя критическая область (критическая область 1)	≤	Статистика (критерий)	≤	Правосторонняя критическая область (критическая область 2)

Односторонние статистические гипотезы:

H_0: $b_0 \ge \tilde{b}_0$ против H_a: $b_0 < \tilde{b}_0$;

H_0: $b_1 \ge \tilde{b}_1$ против H_a: $b_1 < \tilde{b}_1$;

H_0: $\sigma^2 \ge \sigma_0^2$ против H_a: $\sigma^2 < \sigma_0^2$, и

H_0: $b_0 \le \tilde{b}_0$ против H_a: $b_0 > \tilde{b}_0$;

H_0: $b_1 \le \tilde{b}_1$ против H_a: $b_1 > \tilde{b}_1$;

H_0: $\sigma^2 \le \sigma_0^2$ против H_a: $\sigma^2 > \sigma_0^2$.

Гипотезы этого семейства для определенных доверительных интервалов будут:

$$P_{H_0}\left[t_{n-2,\alpha} \le t(b_0)\right] = 1-\alpha;$$

$$P_{H_0}\left[t_{n-2,\alpha} \le t(b_1)\right] = 1-\alpha;$$

$$P_{H_0}\left[\chi^2_{n-2,\alpha} \le \frac{(n-2)\hat{\sigma}^2}{\sigma^2}\right] = 1-\alpha;$$

$$P_{H_0}\left[t_{n-2,1-\alpha} \le t(b_0)\right] = 1-\alpha;$$

$$P_{H_0}\left[t_{n-2,1-\alpha} \le t(b_1)\right] = 1-\alpha;$$

$$P_{H_0}\left[\chi^2_{n-2,1-\alpha} \le \frac{(n-2)\hat{\sigma}^2}{\sigma^2}\right] = 1-\alpha\,.$$

Схема проверки статистических гипотез:

Доверительная область ≤ Статистика (критерий).

Примеры определения доверительной области и критериальной проверки статистических гипотез. По данным примера 6.1 проведем проверку параметров регрессионной модели в соответствии с изложенными выше правилами

Пример 6.2. Определим доверительный интервал для параметра b_0 при заданном уровне значимости $\alpha = 0{,}05$ ($\frac{\alpha}{2} = 0{,}025$ и $1-\frac{\alpha}{2} = 0{,}975$):

$$K_{b_0} = \left[\hat{b}_0 - t_{2,0{,}975}\hat{\sigma}_{\hat{b}_0},\ \hat{b}_0 + t_{2,0{,}975}\hat{\sigma}_{\hat{b}_0}\right] =$$

$$= \left[47{,}5 - 4{,}3\cdot 1{,}0897,\ 47{,}5 + 4{,}3\cdot 1{,}0897\right] = \left[42{,}8142,\ 52{,}1858\right],$$

здесь $\hat{\sigma}^2_{\hat{b}_0} = \dfrac{\hat{\sigma}^2\sum X_i^2}{n\sum\left(X_i - \bar{X}\right)^2} = \dfrac{1{,}25\cdot 3000}{4\cdot 500} = 1{,}875$, $\hat{\sigma}_{\hat{b}_0} = 1{,}0897$, $t_{2,\,0{,}975} = 4{,}3$.

Пример 6.3. Найдем доверительный интервал для дисперсионной оценки σ^2 при $\alpha = 0{,}10$ ($\frac{\alpha}{2} = 0.05$ и $1-\frac{\alpha}{2} = 0{,}95$):

$$K_{\sigma^2} = \left[\frac{\hat{\sigma}^2(n-2)}{\chi^2_{n-2,1-\alpha/2}}, \frac{\hat{\sigma}^2(n-2)}{\chi^2_{n-2,\alpha/2}}\right] = \left[\frac{2\cdot 1{,}25}{\chi^2_{2;0{,}95}}, \frac{2\cdot 1{,}25}{\chi^2_{2;0{,}05}}\right] =$$

$$= \left[\frac{2{,}5}{5{,}99}, \frac{2{,}5}{1{,}03}\right] = \left[0{,}4174,\ 2{,}4272\right]$$

Пример 6.4. Проверим при $\alpha = 0{,}01$ гипотезу $H_0 : b_1 = 0$ с альтернативной гипотезой $H_1 : b_1 \ne 0$. Согласно правилу проверки нулевая гипотеза отбрасывается, если:

$$H_0 \Leftrightarrow \frac{\left|\hat{b}_1 - \tilde{b}_1\right|}{\hat{\sigma}_{\hat{b}_1}} > t_{n-2,1-\alpha/2} \Leftrightarrow \frac{\left|\hat{b}_1\right|}{\hat{\sigma}_{\hat{b}_1}} > t_{2,0{,}995} \Leftrightarrow \frac{1{,}3}{0{,}05} > 9{,}92\,,$$

здесь $\hat{\sigma}^2_{\hat{b}_1} = \dfrac{\hat{\sigma}^2}{\sum\left(X_i - \bar{X}\right)^2} = \dfrac{1{,}25}{500} = 0{,}0025$, $\hat{\sigma}_{\hat{b}_1} = 0{,}05$, $t_{2,0{,}995} = 9{,}92$.

Нулевая гипотеза отвергается, так как наблюденное значение t-критерия больше критического, $t_н > t_{кр}$.

Пример 6.5. Проверим при $\alpha = 0{,}05$ гипотезу $H_0 : b_1 > 1{,}5$ с альтернативной гипотезой $H_1 : b_1 \le 1{,}5$.

Согласно правилу проверки нулевая гипотеза не принимается, если $H_0 \Leftrightarrow \dfrac{\hat{b}_1 - \tilde{b}_1}{\hat{\sigma}_{\hat{b}_1}} < t_{n-2,\alpha}$. В нашем случае имеем: $t_{2,0{,}95} = -2{,}92$, а наблюденная величина критерия $\dfrac{\hat{b}_1 - \tilde{b}_1}{\hat{\sigma}_{\hat{b}_1}} = \dfrac{1{,}3-1{,}5}{0{,}05} = -4$ и нулевая гипотеза отвергается, так как $t_н < t_{кр}$. Можно констатировать, что $b_1 = 1{,}3$ существенно меньше $b_1 = 1{,}5$.

6.4. Статистическое оценивание методом наименьших квадратов — обобщения на случай матричного представления линейной регрессии

6.4.1. Применение МНК при решении линейных регрессионных моделей

Обобщим ситуации, проанализированные в §6.2, 6.3. Вместо уравнения (6.1) рассмотрим случай, когда число экзогенных переменных (факторных признаков) равно двум или больше двух, т.е. $m \ge 2$:

$$y_i = \sum_{j=1}^{m} X_{ij} b_j + u_i\,,\quad E(u_i) = 0\,,\quad E(u_i u_s) = \sigma^2 v_{is}\,,\quad i,s = (1, 2, \ldots, n).$$

Последняя формальная запись регрессионной модели (6.1) при помощи простых преобразований может быть легко специализирована или, напротив, обобщена с целью отображения большого числа встречаемых в реальности разнообразных явлений и процессов. При этом всегда будем иметь $m \ge 1$ параметров (коэффициентов), подлежащих оцениванию, в систематической части модели регрессии. В простейшем случае оценивается (на-

пример, методом наименьших квадратов) только m параметров b_j, в других, более сложных, случаях, учитывающих стохастическую природу регрессии, дополнительно определяются случайная компонента σ^2, представляющая величину ошибки модели (u), и параметр случайной компоненты ν. При решении перечисленных задач целесообразно использовать матричное представление данных. Возникающие на начальном этапе при переходе к матрицам некоторые сложности, связанные с введением новых обозначений (например, вместо Y_i используется обозначение y_i), в последующем не будут препятствием для понимания теоретического материала. Легко убедиться, что матричная запись сравнительно проста, существенно экономичнее обычной, более наглядна. Кроме того, работа с матрицей приобщает исследователей к общепринятой в научном мире символике и накопленному интернациональному опыту по систематизации и обработке больших массивов статистических данных.

Матричное представление регрессии. Линейная регрессионная модель (6.1), отображающая в том числе и линейность своих параметров, записывается в матричном виде следующим образом:

$$y = Xb + u,\ E(u) = 0,\ \operatorname{cov}(u) = E(uu') = \sigma^2 V; \qquad (6.4)$$

$$y \in \boldsymbol{R}^n,\ X \in \boldsymbol{R}^{n,m},\ \operatorname{rg}(X) \le m \le n,\ b \in \boldsymbol{R}^m,\ u \in \boldsymbol{R}^n,\ 0 \in \boldsymbol{R}^n,$$

$$V \in \boldsymbol{R}^{n,n},\ \sigma^2 \in \boldsymbol{R}_{++}$$

и $V \subset P$ есть неединичная, положительно определенная матрица.

В формуле (6.4) приняты обозначения:

$$y := \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix},\ X := \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1m} \\ X_{21} & X_{22} & \ldots & X_{2m} \\ X_{31} & X_{32} & \ldots & X_{3m} \\ & \ldots & & \\ X_{n1} & X_{n2} & \ldots & X_{nm} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{pmatrix},\ b := \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{pmatrix},\ u := \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \cdot \\ \cdot \\ \cdot \\ u_m \end{pmatrix}.$$

Выше рассматривался частный случай многомерной регрессионной модели, когда

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix},\ X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ X_{n1} & X_{n2} \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_n \end{pmatrix},\ b := \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix};\ u := \begin{pmatrix} U_1 \\ U_2 \\ \cdot \\ \cdot \\ \cdot \\ U_n \end{pmatrix}$$

Различие регрессионных моделей может в значительной степени определяться различием матриц ковариаций V.

Расширенные представления о ковариации остатков регрессионной модели. С учетом возможного появления остатков (ошибок) регрессионной модели (U) логично предположить существующие между ними взаимосвязи. Последние находят отображение в матрице ковариаций ($\operatorname{cov} U$). Состав элементов $\operatorname{cov} U$ позволяет выделять дополнительно ряд стандартного вида моделей: гомоскедастичных, гетероскедастичных, с эквивалентной корреляцией остатков и авторегрессионные. Рассмотрим эти модели. Чтобы упростить ход рассуждений, предварительно примем, что общее число остатков (ошибок) модели (k) равно числу наблюдений (n): $k = n$.

Гомоскедастичная модель. *Гомоскедастичностью* называется случай постоянства дисперсии случайной компоненты. Предполагается, что величина U не меняется от наблюдения к наблюдению. При этом ковариации остатков модели для различных наблюдений не рассматриваются, считается, что все $\sigma_{ij} = 0$:

$$\{y \in \boldsymbol{R}^n,\ X \in \boldsymbol{R}^{n,m},\ m \le n,\ b \in \boldsymbol{R}^m,\ E(u) = 0,$$

$$\operatorname{cov}(u) = E(uu') = V = \sigma^2 I_n,\ \sigma^2 > 0\},$$

$$V = \begin{pmatrix} E(u_1^2) & E(u_1u_2) & \ldots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \ldots & E(u_2u_n) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ E(u_nu_1) & E(u_nu_2) & \ldots & E(u_n^2) \end{pmatrix} = \sigma^2 I_n = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & & 0 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix}.$$

Здесь V — скалярная матрица. Общее число параметрических оценок модели будет: $l = m + 1$, т. е. m оценок для регрес-

сионных коэффициентов b_j и оценка дисперсии σ^2. Гипотетическое предположение о равенстве дисперсий остатков и соответствующий ему алгоритм решения модели являются классическими и в регрессионном анализе используются чаще всего. Для поиска модельных параметров при этом может применяться метод наименьших квадратов без каких-либо модификаций.

Гетероскедастичная модель. *Гетероскедастичностью* называется явление, когда дисперсии случайной компоненты U по наблюдениям принимают различные величины. Ковариаций остатков модели, как и в предыдущем случае, не предполагается ($\sigma_{ij}^2 = 0$):

$$\{y \in R^n,\ X \in R^{n,m},\ m \le n,\ b \in R^m,\ E(u) = 0,$$
$$\operatorname{cov}(u) = E(uu') = V = \operatorname{diag}\{\sigma_{ii}\}\};$$

$$\operatorname{cov}(u) = V = \begin{pmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & \sigma_{nn} \end{pmatrix} = \operatorname{diag}(\sigma_{ij}), \quad \sigma_{ii} > 0 \ (i = 1, 2, \dots, n),$$

где V — неотрицательная диагональная матрица.

С учетом различия дисперсий остатков оценке подлежит $l = m + n$ параметров — по числу коэффициентов регрессии m и дисперсий остатков модели σ_{ii}.

Модель с эквивалентной корреляцией остатков. Данный случай дополняет понятие гомоскедастичности. Он допускает возможность коррелированности остатков регрессионной модели $\sigma_{ij} \neq 0$, но одновременно предполагается, что уровни связей остатков одинаковы и равны некоторому числу a, а с целью соблюдения требования положительной определенности матрицы ковариаций остатков налагается ограничение $0 < a < 1$:

$$\{y = Xb + u,\ E(u) = 0,\ \operatorname{cov}(u) = E(uu') = V = \sigma^2 A,$$
$$A = (1-a)I + aE = (1-a)I_n + aE_{n,n},\ 0 < a < 1\};$$

$$A = \begin{pmatrix} 1 & a & a & \dots & a \\ a & 1 & a & \dots & a \\ a & a & 1 & \dots & a \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a & a & a & \dots & 1 \end{pmatrix},\ a \in R,\ e = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix},\ e \in R^n,\ ee' =: E_{n,n} \in R^{n,n}.$$

Число параметров, оцениваемых в регрессионной модели: $l = m + 2$.

Авторегрессионная модель. Имеет вид:

$y = Xb + u \Leftrightarrow y_i = X_i b + u_i$, где $i = 1, 2, \dots, n$, или

$$u_i = \rho u_{i-1} + v_i,\ |\rho| < 1,\ (i = 1, 2, \dots, n),\ E(v_i) = 0,\ E(v_i v_s) = \begin{cases} \sigma^2, & i = s \\ 0, & i \neq s. \end{cases}$$

Авторегрессионная модель представляет случай коррелированности наблюдений, например, во времени (последующее событие часто зависит от совершения предыдущего).

Согласно элементарным рассуждениям матрица ковариаций для остатков модели, учитывающая их авторегрессионные связи, строится следующим образом:

$$\operatorname{cov}(u) = V = E(uu') = \frac{\sigma^2}{1-\rho^2} A,\ A = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}.$$

Статистические оценки модельных параметров, полученные методом наименьших квадратов (МНК-оценки). Выше нами принималось исходное предположение только о нормальности изучаемой регрессии. Дополним его условием оценивания параметров регрессии методом наименьших квадратов. Такой подход наиболее ясно представляет основополагающие идеи регрессионного анализа.

Оценивание методом наименьших квадратов, как собственно и проективное оценивание методом моментов, осуществляется

на основе важнейшего предположения о гомоскедастичности стохастической структуры регрессионной модели:

$$y = Xb + u,\ E(u) = 0,\ \operatorname{cov}(u) = E(uu') = \sigma^2 I \text{ или}$$

$$y_i = X_i b + u_i\ \ (i = 1, 2, ..., n),\ E(u_i) = 0,\ E(u_i u_s) = \begin{cases} \sigma^2, i = s \\ 0, i \neq s. \end{cases}$$

Количественная определенность МНК-оценок не предполагает даже существование математического ожидания и ковариаций остатков регрессии, гипотетически они полагаются равными нулю.

Алгоритм оценивания методом МНК при использовании матричной формы записи выглядит весьма кратким и включает только два шага:

Шаг 1. Определение основополагающего требования метода наименьших квадратов — минимизации остатков регрессионной модели:

$$L := \sum_{i=1}^{n} \left(y_i - \sum X_{ij} b_j \right)^2 \to \min_b \Leftrightarrow L = (y - Xb)'(y - Xb) \to \min_b .$$

Шаг 2. Оптимизация оценок параметров регрессионной модели.

Процесс статистического оценивания предусматривает выполнение ряда условий, необходимых для достижения в решении задачи экстремума. Эти условия в общем определяются четырьмя простыми правилами дифференцирования матриц[1]:

Пусть $f(x)$ — любая функция:

дифференцирование постоянного вектора a

$$f(x) = a,\ a \in \boldsymbol{R}^n,\ \frac{\partial a}{\partial x} = 0,\ a \in \boldsymbol{R}^n,\ 0 \in \boldsymbol{R}^n; \qquad (6.5)$$

дифференцирование линейной функции вектора a

$$f(x) = a'x = x'a,\ a \in \boldsymbol{R}^n,\ x \in \boldsymbol{R}^n,\ \frac{\partial (a'x)}{\partial x} = \frac{\partial (x'a)}{\partial x} = a',\ a \in \boldsymbol{R}^n; \qquad (6.6)$$

дифференцирование линейной функции матрицы A с правосторонним умножением на вектор x

$$f(x) = Ax,\ A \in \boldsymbol{R}^{m,n}, x \in \boldsymbol{R}^n,\ \frac{\partial Ax}{\partial x} = A,\ A \in \boldsymbol{R}^{m,n}; \qquad (6.7)$$

[1] См., например, Dhrymes. 1978 Die Ableitungsregeln für Matrizen (Драймес П. Дж. Правила дифференцирования матриц. Нью-Йорк, 1978). *Прим. пер.*

дифференцирование симметрической квадратичной функции матрицы A

$$f(x) = x'Ax,\ A \in \boldsymbol{R}^{n,n},\ A = A',\ x \in \boldsymbol{R}^n,$$

$$\frac{\partial x'Ax}{\partial x} = 2x'A,\ A \in \boldsymbol{R}^{n,n},\ x \in \boldsymbol{R}^n. \qquad (6.8)$$

С учетом действия четырех правил дифференцирования (6.5) — (6.8), применяя метод наименьших квадратов, получим наилучшие оценки $\hat{b}_j$ для параметров регрессионной модели b_j:

$$\frac{\partial L}{\partial b} = -2y'X + 2b'X'X = 0 \Rightarrow \hat{b} = (X'X)^{-1} X'y.$$

МНК-оценки как матричные проекции. Одной из предпосылок реализации метода наименьших квадратов является нестрогое предположение о МНК-оценке как проекции. Известно, что проекция $\hat{y} \in \boldsymbol{R}^n$ вектора $y \in \boldsymbol{R}^n$ на матрицу $X \in \boldsymbol{R}^{n,m}$ характеризуется двумя важнейшими свойствами: линейностью и ортогональностью. Эти же свойства постулируются и для оценок $\hat{b} \in \boldsymbol{R}^m$:

$\hat{y} = X\hat{b},\ \hat{b} \in \boldsymbol{R}^m$ (линейность),

$(y - \hat{y}) \perp X \Leftrightarrow \hat{y}'X = y'X$ (ортогональность).

Принимая во внимание шаги (1) и (2) алгоритма оценивания, получим ***МНК*-оценки**: из $\hat{b}'X'X = y'X$ следует $\hat{b} = (X'X)^{-1}X'y$. Последнее равенство позволяет рассматривать сумму остатков регрессионной модели с точки зрения возможности ее разложения, точнее говоря, разложению может быть подвергнута сумма квадратов ошибок модели:

$$(y - Xb)'(y - Xb) = \hat{s}^2 n + (b - \hat{b})' X'X (b - \hat{b}),$$

$$\hat{b} := (X'X)^{-1} X'y,\ \hat{s}^2 := \frac{1}{n}(y - X\hat{b})'(y - X\hat{b}).$$

Тем самым устанавливается, что параметрические оценки $\hat{s}^2$ и $\hat{b}$ являются минимизирующим решением для функции наименьших квадратов.

Оценивание методом наименьших квадратов и линейное программирование. Описанные выше шаги (1) и (2) алгоритма метода наименьших квадратов могут рассматриваться как некоторые

фрагменты общей теории квадратично-линейного программирования:

Фрагмент 1

МНК-оценка для b в $y = Xb + u$ есть $\hat{b} = (X'X)^{-1}X'y$.

□ *Доказательство:* Исходное условие

$$L = (y - Xb)'(y - Xb) \to \min_b,$$

тогда

$$L = (y - Xb)'(y - Xb) = y'y - y'Xb - b'X'y + b'X'Xb = y'y - 2y'Xb + b'X'Xb;$$

$$\frac{\partial L}{\partial b} = 0 - 2y'X + 2b'X'X = 0 \Leftrightarrow y'X = b'X'X \Leftrightarrow X'y =$$

$$= X'Xb \Rightarrow \hat{b} = (X'X)^{-1}X'y. \blacksquare$$

Результат, как видим, полностью соответствует решению методом наименьших квадратов.

В случае с линейным программированием вывод опосредуется предположением, что ранг матрицы исходных данных X равен числу анализируемых признаков m, т. е. все признаки, определяющие размерность выборочной совокупности данных, линейно независимы. В действительности такая теоретическая посылка представляется слабой.

Обобщения для метода моментов как инструментария статистического оценивания. Теория метода наименьших квадратов позволяет делать широкие обобщения и рассматривать моменты в качестве особенных инструментальных приемов оценивания регрессии. Будем исходить из модельного равенства, записанного в матричной форме:

$$y = Xb + u.$$

Умножим обе части равенства слева на X', затем сократим запись, сохранив детерминированную часть модели и опустив случайную компоненту $X'u$. После этого поиск параметрического вектора b сводится к решению простой алгебраической задачи оценивания методом моментов:

$$X'y = X'Xb \Rightarrow \hat{b} = (X'X)^{-1}X'y.$$

Идея обобщения сводится к операции сквозного перемножения модели на некоторую (любую) матрицу Z и определению

ожидаемой статистической оценки таким образом, чтобы нивелировать влияние случайных возмущений:

$$Z'y = Z'Xb \Rightarrow \tilde{b} = (Z'X)^{-1}Z'y.$$

Это так называемое обычное оценивание при помощи инструментальных признаков. При сравнении с исходной моделью следует обратить внимание, что здесь мы имеем дело не с некоторой экзогенной, а с дополнительной величиной, выступающей как некий новый инструментарий.

Пример 6.6. Статистическое оценивание параметров уравнения линейной регрессии

Чтобы легче было производить расчеты, используем уже полученные выше результаты решения примера 6.1, и будем строить только те матрицы, которые необходимы для определения параметров b_j:

по условию имеем

$$y = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} 60 \\ 75 \\ 85 \\ 100 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \\ \cdot & \cdot \\ \cdot & \cdot \\ X_{n1} & X_{n2} \end{pmatrix} = \begin{pmatrix} 1 & 10 \\ 1 & 20 \\ 1 & 30 \\ 1 & 40 \end{pmatrix},$$

$$b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}, \quad u = \begin{pmatrix} U_1 \\ U_2 \\ \cdot \\ \cdot \\ \cdot \\ U_n \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix},$$

следовательно,

$$X'X = \begin{pmatrix} 4 & 100 \\ 100 & 3000 \end{pmatrix}, \quad (X'X)^{-1} = \begin{pmatrix} 1{,}5 & -0{,}05 \\ -0{,}05 & 0{,}002 \end{pmatrix}, \quad X'y = \begin{pmatrix} 320 \\ 8650 \end{pmatrix},$$

$$\hat{b} = (X'X)^{-1}X'y = \begin{pmatrix} 47{,}5 \\ 1{,}30 \end{pmatrix}.$$

Некоторые особые случаи регрессионной модели, не включающей константную величину b_0 (Приемы оценивания средних величин для простейшей модели с одним уравнением). Теоретические выводы, полученные в предыдущих параграфах главы, могут быть наглядно представлены на простом примере регрессионной модели с одним объясняющим (факторным) признаком ($m=1$). При этом легко обнаруживается связь с другими методами оценивания, например, с методом максимального правдоподобия, а также появляется возможность продемонстрировать различные подходы к оцениванию средних величин для элементарной модели с одним уравнением.

Случай 0: Модель с одним факторным признаком

Модель общего вида

$$\left\{y = Xb + u,\quad y \in R^n,\ X \in R^{n,m},\ E(u)=0,\quad \operatorname{cov}(u)=\sigma^2 I\right\}$$

сужается до

$$\left\{y = Xb+u, b=b_1,\ y \in R^n,\ X \in R^{n,1} \Leftrightarrow X := x \in R^n,\ E(u)=0,\ \operatorname{cov}(u)=\sigma^2 I\right\}.$$

Методом наименьших квадратов получают оценку:

$$\hat{b}_1 = \hat{\mu} = (x'x)^{-1}x'y = \frac{x'y}{x'x} = \frac{y'x}{x'x}.$$

МНК-оценки как проекции. Альтернативную оценку для модели в случае 0 особенно легко представить как проекцию. Так, проекция $\hat{y} \in R^n$ вектора $y \in R^n$ на вектор $x \in R^n$ характеризуется двумя основными свойствами:

а) $\hat{y} = \hat{b}_1,\ \hat{b}_1 \in R$ — линейность,

б) $(y-\hat{y}) \perp x \Leftrightarrow \hat{y}'x = y'x$ — ортогональность.

Посылка а) в б) дает результат: $\hat{y}'x = \hat{b}_1 x'x \therefore \hat{b}_1 = \frac{y'x}{x'x}$.

Оценка $\hat{b}_1 = \frac{y'x}{x'x}$ есть не что иное, как МНК-оценка. В особом случае, если модель в правой части имеет только одну константу b_0, в силу вступает условие равенства вектора значений факторного признака x единичному вектору: $x = e$, где $e' = (1\,1\ldots1)$ — единичный вектор и тогда $\hat{b}_1 = \hat{\mu} = \hat{y} = \frac{e'y}{e'e}$.

Приведенная спецификация рассматривается в теории статистики как простое усредненное оценивание и одновременно как

важнейший канонический случай. В дальнейшем это станет очевидным при сравнениях полученной оценки с максимально правдоподобной оценкой.

Случай 1: *Оценивание одной средней величины и дисперсии — канонический случай* (Нормальная регрессия с одним параметром).

В регрессионной модели обычно предполагается нормальное распределение остатков

$$y_i = X_i b_i + u_i = \mu_i + u_i,\ \text{при}\ u_i \sim N(0,\sigma^2)$$

(выдвинутое условие о нормальности распределения σ^2 не обязательно для МНК-метода, см. выше).

$$E(u_i) = 0 \neq E(y_i) = \mu;\quad E(u_i u_s) = E(y_i y_s) = \begin{cases}\sigma^2 & i = s\\ 0 & i \neq s\end{cases}.$$

Функция плотности вероятности нормально распределенной случайной величины записывается как:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Следуя правилам метода максимального правдоподобия, оценку параметров нормально распределенной величины получим при непосредственном дифференцировании функции $L(\mu,\sigma^2)$:

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right) =$$

$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)\sigma^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right),$$

при условии достижения экстремумов: $\frac{\partial L}{\partial \mu} = 0$ и $\frac{\partial L}{\partial \sigma^2} = 0$ или, что то же самое, для прологарифмированной функции максимального правдоподобия: $\frac{\partial L}{\partial \mu} = 0$ и $\frac{\partial L}{\partial \sigma^2} = 0$. Решая систему дифференциальных уравнений, получим:

• оценку средней и смещенную оценку дисперсии остатков

$$\hat{\mu} := \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{и} \quad \hat{\sigma}^2 := s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 ,$$

• скорректированную (несмещенную) оценку дисперсии остатков

$$\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Принимая $\hat{\sigma}^2 := s^2$ и $\hat{\mu} := \bar{x}$, сделаем соответствующую подстановку и перепишем функцию максимального правдоподобия:

$$L(\hat{\mu},\hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left(-\frac{n-1}{2}\right).$$

Полученные оценки s^2 и $\bar{x}$ максимизируют функцию $L(\hat{\mu},\hat{\sigma}^2)$:

$$\frac{\partial^2 L}{\partial^2 \hat{\sigma}^2} = -\frac{n}{2}(2\pi\hat{\sigma}^2)^{-\frac{n}{2}-1} \exp\left(-\frac{n-1}{2}\right) < 0 .$$

Обобщения для регрессионной модели, имеющей $m>1$ параметров. Параметрические оценки, полученные методами максимального правдоподобия и наименьших квадратов, полностью согласуются. Такой теоретический результат следует при распространении приведенных выше выводов на регрессионные модели с числом параметров $m>1$.

Обратимся еще раз к функции максимального правдоподобия и запишем ее логарифм:

$$L(b,\sigma^2) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-Xb)'(y-Xb)}{2\sigma^2}} =$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{(y-Xb)'(y-Xb)}{2\sigma^2}\right);$$

$$L(b,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-Xb)'(y-Xb).$$

Последовательным дифференцированием функции $L(b,\sigma^2)$ по b и σ^2 получим результаты, совпадающие с МНК-оценками:

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \hat{b} = (X'X)^{-1}X'y \quad \text{и} \quad \frac{\partial L}{\partial \sigma^2} = 0 \Rightarrow \hat{\sigma}^2 = s^2 = \frac{1}{n}(y - X\hat{b})'(y - X\hat{b}).$$

Вывод о согласованности оценок, найденных методами МНК и максимального правдоподобия, не является универсальным, подтверждаемым во всех возможных случаях. В этом можно убедиться, рассмотрев следующие три варианта усредненного оценивания.

***Случай 2:** n — средних и одна дисперсия.*

Если каждое наблюдение характеризуется константной величиной, но остается предположение о соответствии ее распределения нормальному закону, метод максимального правдоподобия оказывается неприменим, это пример так называемой перепараметризации [116, с. 114].

Пусть имеем модель вида:

$$\left\{ y \in \boldsymbol{R}^n, X \in \boldsymbol{R}^{n,m}, X = I_n = I_m, m = n, \mu' = (\mu_1 \mu_2 ... \mu_n), \mu \in \boldsymbol{R}^m, \right.$$

$$\left. E(u) = 0, \operatorname{cov}(u) = \sigma^2 I \right\}, \quad y_i = X_i b_i + u_i = \mu_i + u_i \quad \text{и} \quad u_i \sim \boldsymbol{N}(0,\sigma^2),$$

$$E(u_i) = 0 \neq E(y_i) = \mu; \quad E(u_i u_s) = E(y_i y_s) = \begin{cases} \sigma^2 & i = s \\ 0 & i \neq s. \end{cases}$$

Используя метод наименьших квадратов, имеем

$\hat{b} = \hat{\mu} = (X'X)^{-1}X'y = (I'I)^{-1}I'y = \bar{y} = \hat{\mu}$, $m = n$ — линейность оценки,

$E(\hat{\mu}) = E(y) = \mu$ — оптимальность оценки,

$$\operatorname{cov}(\hat{\mu}) = \operatorname{cov}(y) = E\left\{(y-\mu)(y-\mu)'\right\} = E(uu') = \sigma^2 I_m = \sigma^2 I_n .$$

В то же время для оценки $\sigma^2 > 0$ в силе остается необходимое условие:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 = 0 \Rightarrow \hat{\sigma}^2 = 0 ,$$

т.е. оценивание параметра становится невозможным.

Метод максимального правдоподобия приводит к аналогичным результатам:

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu_i)^2} ,$$

после логарифмирования $\ln L =: L$ выдвигаем необходимые условия

$$\frac{\partial L}{\partial \mu_i} = 0 \Rightarrow \hat{\mu}_i = y_i , \qquad (6.9)$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = 0 \Rightarrow n\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \tag{6.10}$$

$$= (y-\hat{\mu})'(y-\hat{\mu})$$

Поясним возникающую проблему перепараметризации. Общее число параметров в данной ситуации составляет $n+1$ (!), их оценивание для исследователя становится весьма непростой задачей, если учесть, что функция максимального правдоподобия гипотетически имеет вид:

$$\mathrm{L}(\mu,\sigma^2) = L = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum (y_i - \mu_i)^2\right],$$

и дает оценки

$$y_i = \hat{\mu}_i,\ i = 1,2,..,n \ \text{ и } \ \hat{\sigma}^2 = \frac{(y-\hat{\mu})'(y-\hat{\mu})}{n}.$$

Подстановка оценки $\hat{\mu}$ в формулу для $\hat{\sigma}^2$ приводит к равенству $\hat{\sigma}^2 = 0$, а это бессмысленный результат, поскольку действует предположение $0 < \sigma^2 < \infty$.

Объяснить подобный парадокс можно, рассматривая плотность вероятности функции максимального правдоподобия. После подстановки $\hat{\mu}_i$ и σ^2 в функцию L получим:

$$\mathrm{L} = \left(\frac{1}{(y-\mu)'(y-\mu)}\right)^n \exp\left(\frac{1}{2}\right), \ \lim_{y\to\mu} \mathrm{L} \to \infty.$$

т.е. функция максимального правдоподобия дискретна и стремится к бесконечности.

Случай 3: *Оценка одной средней и n — дисперсий.*

Предлагаемая спецификация может рассматриваться как случай, в некоторой мере противоположный предыдущему:

в модели $y = Xb + u$ имеем

$$\left\{y \in \boldsymbol{R}^n,\ X \in \boldsymbol{R}^{n,1},\ X \in e_n,\ m = 1,\ \mu \in \boldsymbol{R},\ E(u) = 0,\ \mathrm{cov}(u) = \mathrm{diag}\left(\sigma_i^2\right)\right\},$$

$y_i = X_i b_i + u_i = \mu + u_i$ (и $u_i \sim \boldsymbol{N}\left(0,\sigma_i^2\right)$ не обязательно, см. выше),

$$E(u_i) = 0 \neq E(y_i) = \mu; \quad E(u_i u_s) = E(y_i y_s) = \begin{cases} \sigma_i^2 & i = s \\ 0 & i \neq s. \end{cases}$$

Здесь возникает как бы обратная задача — оценивание одной средней и n дисперсий против n средних и одной дисперсии, как это было в случае 2, т. е. мы имеем дело с гетероскедастичностью:

$$y_i = \mu + u_i,\ u_i \sim \boldsymbol{N}\left(0,\sigma_i^2\right),\ i = 1, 2, ..., n.$$

Основная трудность в оценивании параметров регрессии при этом обусловливается невозможностью применения метода максимального правдоподобия:

$$\mathrm{L}(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \mathrm{e}^{-\frac{1}{2\sigma_i^2}(y_i-\mu)^2},$$

после подстановки $\ln \mathrm{L} =: L$ и дифференцирования будет

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{n} \frac{y_i - \mu}{\sigma_i^2} = 0, \tag{6.11}$$

$$\frac{\partial L}{\partial \sigma_i^2} = -\frac{1}{\sigma_i^2} + \frac{1}{\sigma_i^4}(y_i - \mu)^2 = 0 \Rightarrow \hat{\sigma}_i^2 = (y_i - \hat{\mu})^2, \tag{6.12}$$

подставляя результат из (6.12) в (6.11), получим:

$$f(\hat{\mu}) := \sum_{i=1}^{n} \frac{1}{y_i - \hat{\mu}} = 0.$$

Последнее уравнение относительно μ вообще не имеет решения и, кроме того, функция $f(\hat{\mu})$ дискретна.

Иллюстрация. Пусть имеется три наблюдения $y = \{-1,0,1\}$, тогда

$$f(\mu) = \frac{1}{-1-\mu} + \frac{1}{0-\mu} + \frac{1}{1-\mu} = \frac{1-\mu^3}{\mu(1+\mu)(1-\mu)}.$$

На интервале $[-2; 2]$ функция $f(\mu)$ принимает вид (рис. 6.1).

На графике в двух точках пересечения с осью абсцисс, где $y = 0$, параметр μ принимает значение $\mu_{1,2} = \pm\frac{1}{\sqrt{3}} = \pm 0{,}57735$.

Случай 4: *n оценок средних при известных дисперсиях.*

Рассмотрим в заключение еще один случай, когда для модели $y = Xb + u$ принимаются следующие предположения:

$$\left\{y \in \boldsymbol{R}^n,\ X \in \boldsymbol{R}^{n,1},\ X = e_n,\ m = 1,\ \mu \in \boldsymbol{R},\ E(u) = 0,\ \mathrm{cov}(u) = \mathrm{diag}\left(\sigma_i^2\right)\right\}.$$

При условии, что значения всех дисперсий σ_i^2 известны, переходим к так называемому варианту оценивания по Айткину

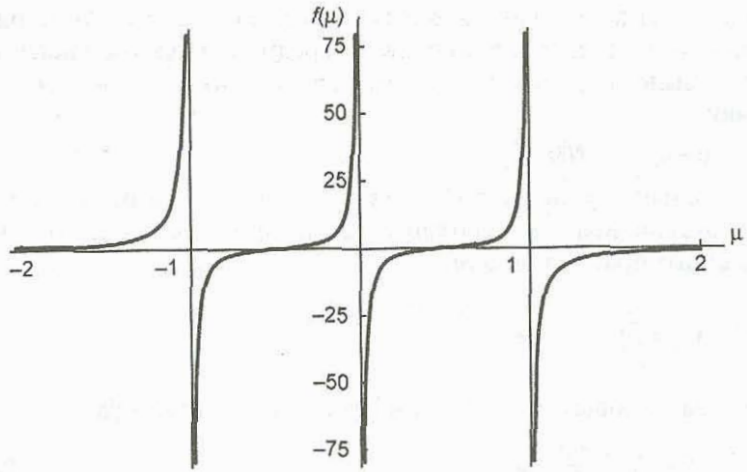


Рис. 6.1

(подробнее см. пп. 6.5.3, 6.5.4). При этом вместо исходной модели вида $y_i = \mu + u_i$ оценивание производится по трансформированной модели:

$$z_i = x_i\mu + v_i,\ z_i := \frac{y_i}{\sigma_i},\ x_i := \frac{1}{\sigma_i},\ v_i := \frac{u_i}{\sigma_i}.$$

Метод наименьших квадратов при этом позволяет получить результаты:

$$E(u_i) = E(v_i) = 0,\ \operatorname{var}(u_i) = \sigma_i^2,\ \operatorname{var}(v_i) = 1 = \sigma^2.$$

Отсюда соответственно будем иметь оценку средней:

$$\hat{\mu} = \frac{1}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}\sum_{i=1}^{n}\frac{y_i}{\sigma_i^2}.$$

Предложенный теоретический вывод представляется весьма важным. Следует учитывать, однако, что практика далеко не всегда подпадает под известные правила. Например, в каждый момент времени может производиться более одного наблюдения, скажем, некоторое равное число наблюдений более одного. Такое условие приводит к модели с усредненным оцениванием при многократно повторяющихся сериях наблюдений — для каждого σ_i^2.

К рассмотренной двухпризнаковой модели, представляющей связи величин X и Y, относится понятие парной корреляции, в последующем остановимся на этом подробнее[1].

Количественная определенность корреляционной связи в регрессии с двумя параметрами. Регрессионная модель, имеющая два параметра, имеет стандартный вид:

$$Y_i = b_1 + b_2X_i + u_i \Leftrightarrow Y_i = b_0 + b_1X_i + u_i \Leftrightarrow$$

$$\Leftrightarrow Y_i \in \boldsymbol{R},\ X_i \in \boldsymbol{R},\ b_1, b_2 \in \boldsymbol{R};\ u_i \in \boldsymbol{R}, Y_i \in \boldsymbol{R},\ X_i \in \boldsymbol{R};\ b_0 := b_1;\ b_1 := b_2;$$

$$\left\{y_i = bx_i + u_i,\ y_i := Y_i - \bar{Y},\ x_i := X_i - \bar{X}\right\} \Leftrightarrow \{y = bx + u\},$$

$$y_i \in \boldsymbol{R}, x_i \in \boldsymbol{R}, b \in \boldsymbol{R}, u_i \in \boldsymbol{R};\ y \in \boldsymbol{R}^n, x \in \boldsymbol{R}^n, u \in \boldsymbol{R}^n.$$

Статистические оценки параметров $\hat{b} \in \boldsymbol{R}$ (величин $b \in \boldsymbol{R}$) определяются на основе известных теоретических выводов о том, что:

$$\hat{b} = \frac{y'x}{x'x},\ \hat{u} := y - \hat{y},\ \hat{y} := x\hat{b} = x\frac{y'x}{x'x},\ y'x = \sum_i y_ix_i,\ x'x = \sum_s x_sx_s\ .$$

Существующая линейная связь x и y при этом имеет количественную определенность и может быть измерена:

$$r^2 := \frac{\hat{y}'\hat{y}}{y'y} = \frac{\hat{b}x'x\hat{b}}{y'y} = \hat{b}^2\frac{x'x}{y'y} = \frac{\frac{x'y}{x'x}x'x\frac{y'x}{x'x}}{y'y} = \frac{(x'y)^2}{x'xy'y} \Leftrightarrow \frac{R^2}{\hat{b}^2} = \frac{y'x}{y'y} = \frac{x'y}{y'y}.$$

Замечание 1. Переход к обратной регрессии.

Обратная регрессия подразумевает отображение связи признаков в модели вида: $x_i = b^0y_i + v_i$, соответственно имеем:

$$\hat{b}^0 = \frac{y'x}{y'y},\ \hat{v} := x - \hat{x},\ \hat{x} := x\hat{b}^0 = x\frac{y'x}{y'y},\ y'x = \sum_i y_ix_i,\ y'y = \sum_s y_sy_s\ ,$$

$$r^2 := \frac{\hat{x}'\hat{x}}{x'x} = \frac{\hat{b}^0y'y\hat{b}^0}{x'x} = \hat{b}^{0^2}\frac{y'y}{x'x} = \frac{\frac{y'x}{y'y}y'y\frac{x'y}{y'y}}{x'x} = \frac{(x'y)^2}{x'x\cdot y'y} \Leftrightarrow \frac{r^2}{\hat{b}^{0^2}} = \frac{y'x}{x'x} = \frac{x'y}{x'x}.$$

[1] В отечественной литературе связь только двух изучаемых признаков обычно определяется как простая (простая корреляция). Связь двух признаков из их комплекса при условии действия связей всех пар признаков в комплексе называется парной (парная корреляция). Наконец, если выделяется связь двух признаков из их некоторого множества при условии, что все другие парные связи признаков этого множества нивелированы, то такие связи называются частными (частная корреляция). — *Прим. пер.*

Замечание 2. Определение корреляции.

Количественная определенность связей x и y, y и x полностью согласуется:

$$r^2 := \hat{b}^2 \frac{x'y}{y'y} = \hat{b}^{0^2} \frac{x'y}{x'x} \text{ и } \left\{\hat{b} = \frac{1}{\hat{b}^0}, \hat{b}^0 = \frac{1}{\hat{b}}\right\} \Leftrightarrow r^2 = 1,$$

$$\rho = |r| = \sqrt{r^2}, \ \rho = \frac{x'y}{\sqrt{x'x} \cdot \sqrt{y'y}}, \ -1 \le \rho \le 1.$$

ρ называют корреляцией признаков y и x, имеющих центрированные значения.

Замечание 3. Определение парной регрессии.

Для отображения допустимой линейной формы связи двух любых центрированных признаков $x_j \in \boldsymbol{R}^n$ и $x_k \in \boldsymbol{R}^n$ могут быть построены регрессионные модели:

$$x_{ji} = b^0 x_{ki} + u_{jk,i} \text{ и } x_{ki} = b^0 x_{ji} + v_{kj,i}.$$

Соответственно можем записать тождество мер связей для пары признаков: $r^2_{jk} \equiv r^2_{kj}$.

Замечание 4. Количественная определенность парной корреляции.

Для двух признаков $x_j \in \boldsymbol{R}^n$ и $x_k \in \boldsymbol{R}^n$ имеем:

$$\frac{r^2_{jk}}{\hat{b}^2_{jk}} = \frac{x'_k x_j}{x'_j x_j} \text{ или } \frac{r^2_{kj}}{\hat{b}^{0^2}_{kj}} = \frac{x'_k x_j}{x'_k x_k}.$$

В предыдущем случае просто принималось: $x_j = y$ и $x_k = x$ или $x_j = x$, а $x_k = y$.

Определение. *Парная корреляция двух признаков с центрированными значениями* $x_j \in \boldsymbol{R}^n$ и $x_k \in \boldsymbol{R}^n$ есть

$$\sqrt{r^2_{jk}} = \sqrt{r^2_{kj}} =: \rho_{kj} = \frac{x'_k x_j}{\sqrt{x'_k x_k}\sqrt{x'_j x_j}}.$$

Замечание 5. Корреляция и корреляционная матрица.

При обобщении всех имеющихся показателей парных корреляций получают корреляционную матрицу (матрицу парных корреляций R). Она в сущности может рассматриваться как простая трансформация матрицы моментов $X'X$, где X — матрица с центрированными данными:

$$R := \begin{pmatrix} \frac{x'_1 x_1}{\sqrt{x'_1 x_1}\sqrt{x'_1 x_1}} & \frac{x'_1 x_2}{\sqrt{x'_1 x_1}\sqrt{x'_2 x_2}} & \cdots & \frac{x'_1 x_k}{\sqrt{x'_1 x_1}\sqrt{x'_k x_k}} \\ \frac{x'_2 x_1}{\sqrt{x'_2 x_2}\sqrt{x'_1 x_1}} & \frac{x'_2 x_2}{\sqrt{x'_2 x_2}\sqrt{x'_2 x_2}} & \cdots & \frac{x'_2 x_m}{\sqrt{x'_2 x_2}\sqrt{x'_k x_k}} \\ \vdots & \vdots & & \vdots \\ \frac{x'_m x_1}{\sqrt{x'_k x_k}\sqrt{x'_1 x_1}} & \frac{x'_k x_2}{\sqrt{x'_k x_k}\sqrt{x'_2 x_2}} & \cdots & \frac{x'_k x_k}{\sqrt{x'_k x_k}\sqrt{x'_k x_k}} \end{pmatrix} =$$

$$= \operatorname{diag}\left(\frac{1}{\sqrt{x'_k x_k}}\right) \begin{pmatrix} x'_1 x_1 & x'_1 x_2 & \cdots & x'_1 x_k \\ x'_2 x_1 & x'_2 x_2 & \cdots & x'_2 x_k \\ \vdots & \vdots & & \vdots \\ x'_k x_1 & x'_k x_2 & \cdots & x'_k x_k \end{pmatrix} \operatorname{diag}\left(\frac{1}{\sqrt{x'_k x_k}}\right) = D(X'X)D'.$$

Уточнение. Парная корреляция, исключающая третий фактор[1].

Допустим, имеется не два, а три коррелирующих признака с центрированными значениями: $x_j \in \boldsymbol{R}^n$, $x_k \in \boldsymbol{R}^n$ и $x_i \in \boldsymbol{R}^n$. Это то же, как если бы были x, y и некоторая третья случайная величина z [114, с.181—186]. Тогда коэффициент парной корреляции, исключающий влияние третьего признака (z), запишется как

$$\rho_{x,y|z} := \frac{\rho_{xy} - \rho_{xz}\rho_{zy}}{\sqrt{1-\rho^2_{xz}}\sqrt{1-\rho^2_{zy}}}$$

в отличие от обычного парного коэффициента корреляции[2]

$$\rho_{kj} := \frac{x'_k x_j}{\sqrt{x'_k x_k}\sqrt{x'_j x_j}}, \ x_k, x_j \in \{x, y, z\}.$$

Величины двух приведенных коэффициентов корреляции будут тем более близки, чем в меньшей мере проявляется линейная связь двух переменных x и y с третьей переменной z, т.е. если эта связь слабая или вовсе отсутствует. Это определяется так называемым неравенством треугольника:

---

[1] Частная корреляция. — *Прим. пер.*

[2] ρ — обозначение, принятое обычно для характеристики связи признаков в генеральной совокупности, для выборки принимается символ r. Так как в данном случае не принципиально использование именно ρ или r, сохранены авторские обозначения коэффициентов корреляции. — *Прим. пер.*

В другой ситуации, при совершенной мультиколлинеарности, когда $x = b^0 z$ и $y = bz$, числитель коэффициента $\rho_{x,y/z}$ обращается в нуль:

$$\rho_{xy} - \rho_{xz}\rho_{zy} = \frac{x'y}{\sqrt{x'x \cdot y'y}} - \frac{x'z}{\sqrt{x'x} \cdot \sqrt{z'z}} \cdot \frac{z'y}{\sqrt{z'z} \cdot \sqrt{y'y}} =$$

$$= \frac{x'y}{\sqrt{x'x \cdot y'y}} \left(1 - \frac{x'z}{\sqrt{z'z}} \cdot \frac{z'y}{\sqrt{z'z}} \cdot \frac{1}{x'y}\right) = \rho_{xy}\left(1 - \frac{b^0 z'z}{z'z} \cdot \frac{bz'z}{b^0 bz'z}\right) = \rho_{xy} \cdot 0,$$

одновременно, в силу некоррелированности x и y, появляется необходимость в делении ρ_{xy} на нуль, что невозможно. Таким образом, случай частной корреляции двух переменных при условии полной связи каждого из них с третьей переменной с точки зрения количественной определенности рассматриваться не может.

Приведем примеры для иллюстрации теоретических выводов.

Пример 6.7. Пусть имеются некоторые гипотетические данные четырех наблюдений:

x	y	z
0,5	1	1
5	4	2
7	9	3
15	16	4

В ходе расчетов получим:

$\rho_{xy} = 0{,}9835$, $\rho_{xz} = 0{,}9692$, $\rho_{yz} = 0{,}9844$,

а также $\rho_{x,y/z} = 0{,}6773$, $\rho_{x,z/y} = 0{,}0362$, $\rho_{y,z/x} = 0{,}6992$.

Обобщая аналитические результаты, построим матрицы по значениям всех трех переменных:

ρ_{ij}	x	y	z
x	1	0,9835	0,9692
y	0,9835	1	0,9844
z	0,9692	0,9844	1

$\rho_{ij/k}$	x	y	z
x	1	0,6773	0,0362
y	0,6773	1	0,6992
z	0,0362	0,6992	1

,

ρ_{ij} / $\rho_{ij/k}$	x	y	z
x	1	0,9835	0,9692
y	0,6773	1	0,9844
z	0,0362	0,6992	1

Обратим внимание, что достаточно сильная парная корреляция x и z ($\rho_{xz} = 0{,}9692$) наблюдается при малой величине частной корреляции двух переменных, исключающей воздействие третьей переменной y: $\rho_{x,z/y} = 0{,}0362$. Объяснение подобной закономерности находится просто, при более детальном анализе связей трех любых переменных и исчисленных по их значениям коэффициентов частной корреляции: $\rho_{x_1,x_2/x_3}; \rho_{x_1,x_3/x_2}; \rho_{x_2,x_3/x_1}$.

Наличие корреляции переменных предполагает существование линейных моделей, отображающих форму их связей:

$X_1 = \hat{a}_{1,3} + \hat{b}_{1,3}X_3 + v_{1,3}$ или $v_{1,3} := X_1 - \hat{a}_{1,3} - \hat{b}_{1,3}X_3 = x_1 - \hat{b}_{1,3}x_3$;

$X_2 = \hat{a}_{2,3} + \hat{b}_{2,3}X_3 + v_{2,3}$ или $v_{2,3} := X_2 - \hat{a}_{2,3} - \hat{b}_{2,3}X_3 = x_2 - \hat{b}_{2,3}x_3$ и

$$\rho_{x_1,x_2/x_3} = \frac{v'_{1,3}v_{2,3}}{\sqrt{v'_{1,3}v_{1,3}}\sqrt{v'_{2,3}v_{2,3}}} = \rho_{v_1,v_2};$$

$X_1 = \hat{a}_{1,2} + \hat{b}_{1,2}X_2 + v_{1,2}$ или $v_{1,2} := X_1 - \hat{a}_{1,2} - \hat{b}_{1,2}X_2 = x_1 - \hat{b}_{1,2}x_2$;

$X_3 = \hat{a}_{3,2} + \hat{b}_{3,2}X_2 + v_{3,2}$ или $v_{3,2} := X_3 - \hat{a}_{3,2} - \hat{b}_{3,2}X_2 = x_3 - \hat{b}_{3,2}x_2$ и

$$\rho_{x_1,x_3/x_2} = \frac{v'_{1,2}v_{3,2}}{\sqrt{v'_{1,2}v_{1,2}}\sqrt{v'_{3,2}v_{3,2}}} = \rho_{v_1,v_3};$$

$X_2 = \hat{a}_{2,1} + \hat{b}_{2,1}X_1 + v_{2,1}$ или $v_{2,1} := X_2 - \hat{a}_{2,1} - \hat{b}_{2,1}X_1 = x_2 - \hat{b}_{2,1}x_1$;

$X_3 = \hat{a}_{3,1} + \hat{b}_{3,1}X_1 + v_{3,1}$ или $v_{3,1} := X_3 - \hat{a}_{3,1} - \hat{b}_{3,1}X_1 = x_3 - \hat{b}_{3,1}x_1$ и

$$\rho_{x_2,x_3/x_1}=\frac{v'_{2,1}v_{3,1}}{\sqrt{v'_{2,1}v_{2,1}}\sqrt{v'_{3,1}v_{3,1}}}=\rho_{v_2,v_3}.$$

Согласно логике расчетов оценка ошибки регрессионной модели v есть центрированная величина и из $\rho_{k,j}:=\rho_{x_k,x_j}$ следует $\rho_{x_i,x_j/x_k}=\rho_{v_i,v_j}$, т.е. частный коэффициент корреляции представляет корреляцию ошибок соответствующей пары линейных регрессионных моделей:

$$v_{1,3}=:v_1=x_1-\hat{b}_{1,3}x_3=x-\frac{x'z}{z'z}z \quad \text{и} \quad v_{2,3}=:v_2=x_2-\hat{b}_{2,3}x_3=y-\frac{y'z}{z'z}z.$$

В заключение запишем основные выводы:

$$v'_1v_1=x'x-\frac{(x'z)^2}{z'z}=\frac{x'x\cdot z'z-x'z\cdot x'z}{z'z};$$

$$v'_2v_2=x'x-\frac{(y'z)^2}{z'z}=\frac{x'x\cdot z'z-y'z\cdot y'z}{z'z};$$

$$v'_2v_1=v'_1v_2=\frac{x'y\cdot z'z-x'z\cdot y'z}{z'z};$$

$$\therefore \rho_{x_1,x_2|x_3}=\rho_{x,y|z}=\frac{x'y\cdot z'z-x'z\cdot y'z}{\sqrt{y'y\cdot z'z-y'z\cdot y'z}\cdot\sqrt{x'x\cdot z'z-x'z\cdot x'z}}.$$

Пример 6.8. Имеются данные:

x	2	3	3	4	5	5	6	6	6	7	7	8	8
y	2	2	5	3	3	6	4	5	7	7	8	7	9
z	1	2	4	5	3	5	6	7	5	7	9	8	10

Построим матрицу парных и частных коэффициентов корреляций:

$\rho_{ij/k}$ \ ρ_{ij}	x	y	z
x	1	0,208	0,673
y	0,832	1	0,488
z	0,908	0,869	1

Свойства статистических оценок. НЛН-свойства по Гауссу—Маркову[1]. *Утверждение.* Оценка параметра регрессионной модели $\hat{b}$, полученная методом наименьших квадратов, является наилучшей линейной и несмещенной (НЛН).

□ *Доказательство:* Согласно методу наименьших квадратов $\hat{b}=(X'X)^{-1}X'y$. Зная, что $y=Xb+u$, можем равенство относительно $\hat{b}$ переписать в виде:

$$\hat{b}=(X'X)^{-1}X'(Xb+u)=(X'X)^{-1}X'Xb+(X'X)^{-1}X'u=b+(X'X)^{-1}X'u,$$

отсюда следуют выводы о линейности и несмещенности оценки $\hat{b}$:

$$E(\hat{b})=b+(X'X)^{-1}X'E(u).$$

Так как X — величина нестохастическая и не коррелирует с u, имеем:

$$\therefore E(\hat{b})=b.$$

Определим матрицу ковариаций для оценок параметров регрессионой модели $\hat{b}$:

$$E\left[(\hat{b}-E\hat{b})(\hat{b}-E\hat{b})'\right]=E\left[(\hat{b}-b)(\hat{b}-b)'\right]=E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right]=$$

$$=(X'X)^{-1}X'E(uu')X(X'X)^{-1}=\sigma^2(X'X)^{-1}X'X(X'X)^{-1}=\sigma^2(X'X)^{-1}.$$

То, что оценка, найденная при помощи метода наименьших квадратов, обладает наименьшей дисперсией относительно других линейных несмещенных оценок, подтверждается ее сопоставлением с некоторой гипотетической величиной b^*.

Пусть $b^*=Hy$, тогда

$b^*=Hy=HXb+Hu$ и $E(b^*)=HXb$.

Здесь, по предположению, H — это МНК-оценка плюс некоторая матрица константных значений C, определяющая отличие H от $(X'X)^{-1}X'$:

$$H=(X'X)^{-1}X'+C.$$

Несмещенность статистической оценки требует, чтобы $HX=I$, значит

$$HX=(X'X)^{-1}(X'X)+CX=I+CX \quad \therefore HX=I \Leftrightarrow CX=0.$$

[1] НЛН — наилучшая, линейная, несмещенная оценка статистической величины (BLU — bester, linearer und unverzerter Schätzer). — *Прим. пер.*

Положительная определенность ковариационной матрицы, построенной исходя из условий существования b^*, становится очевидной при разложении этой матрицы на составляющие, т. е. декомпозиции ее элементов: дисперсий и ковариаций:

$$E\left[(b^*-b)(b^*-b)'\right]=E(Huu'H')=E\left(\left[(X'X)^{-1}X'+C\right]uu'\left[X(X'X)^{-1}+C'\right]\right)=$$
$$=\sigma^2\left[(X'X)^{-1}X'X(X'X)^{-1}+(X'X)^{-1}X'C'+CX(X'X)^{-1}+CC'\right]=$$
$$=\sigma^2(X'X)^{-1}+\sigma^2CC'.$$

$CX=0$, как установлено ранее, $X'X$ и CC' положительно определены в силу собственных конструктивных особенностей. Но если CC' — положительно определенная матрица, дисперсия b^* всегда будет больше или равна дисперсии $\hat{b}$, так как разность двух ковариационных матриц даст также положительно определенную матрицу. ■

Замечание 1. Эффективность: Оценка $\hat{b}$ по сравнению с любой другой является линейной и несмещенной оценкой для b. МНК-оценка есть также НЛН-оценка и полностью отвечает ее свойствам — она наилучшая, линейная и несмещенная.

Замечание 2. При переходе от регрессионной модели с двумя параметрами ($m=2$) к модели с любым числом параметров более двух ($m=1,2,..$) теоретические подходы к доказательству свойств статистических оценок и формированию выводов об их надежности по Гауссу—Маркову сохраняют свою силу.

Распределения статистических оценок. Методом наименьших квадратов получают $\hat{b}=(X'X)^{-1}X'y$. Это означает, что оценка дисперсии остатков модели может быть исчислена по отклонениям:

$$\hat{\sigma}^2=\frac{\hat{u}'\hat{u}}{n-m},\quad \hat{y}=y(X'X)^{-1}X'y,\quad \hat{u}:=y-\hat{y}.$$

Оценки $\hat{b}$ и σ^2 при этом статистически независимы и в основном описываются следующими двумя видами распределений:

t-распределение параметра $\hat{b}$:

$$t(b_i)=\frac{\hat{b}_i-b_i}{\hat{\sigma}\sqrt{S^{i,i}}}\sim t_{n-m},\quad i=1,2,..,\ m \text{ с числом степеней свободы: } v=n-m;$$

χ^2-распределение параметра $\hat{\sigma}^2$:

$$\frac{\hat{u}'\hat{u}}{\hat{\sigma}^2}\sim\chi^2_{n-m}.$$

Здесь $S^{i,i}$ — i-е диагональные элементы обратной матрицы моментов, т.е. $(X'X)^{-1}$.

Параметр $\hat{b}$ имеет $n-m$ — число степеней свободы и подчиняется закону t-распределения, параметр $\hat{\sigma}^2$ — закону χ^2-распределения.

Отметим, что теоретические выводы сделаны на основе простейшей двухпараметрической линейной модели ($m=2$), одновременно осуществлен переход от двух показателей кросс-произведений $S^{i,j}$ ($i=1,2$ для $m=2$) к обратной матрице моментов (X и к оценке дисперсии $\hat{\sigma}^2$ как средней характеристике ковариации модельных остатков.

Пример 6.9. Обратимся вновь к данным из §6.3

$$y=\begin{pmatrix}60\\75\\85\\100\end{pmatrix},\quad X=\begin{pmatrix}1&10\\1&20\\1&30\\1&40\end{pmatrix}\rightarrow(X'X)^{-1}=\begin{pmatrix}1{,}5&-0{,}05\\-0{,}05&0{,}002\end{pmatrix},\quad \hat{b}=\begin{pmatrix}47{,}5\\1{,}3\end{pmatrix},$$

$\hat{\sigma}^2=1{,}25.$

Рассчитаем оценки регрессии при условии расширения матрицы предикторов, т.е. перейдем от $m=2$ к случаю, когда $m=3$:

$$y=\begin{pmatrix}60\\75\\85\\100\end{pmatrix},\ X=\begin{pmatrix}1&10\\1&20\\1&30\\1&40\end{pmatrix}\rightarrow y=\begin{pmatrix}60\\75\\85\\100\end{pmatrix},\ X=\begin{pmatrix}1&10&2\\1&20&3\\1&30&5\\1&40&2\end{pmatrix}\rightarrow$$

$$(X'X)^{-1}=\begin{pmatrix}2{,}57759&-0{,}0413793&-0{,}431034\\-0{,}0413793&0{,}00206897&-0{,}00344828\\-0{,}431034&-0{,}00344828&0{,}172414\end{pmatrix},\ \hat{b}=\begin{pmatrix}48{,}7931\\1{,}31034\\-0{,}517241\end{pmatrix},$$

$\hat{\sigma}^2=0{,}862069.$

Как видим, алгоритм решения задачи по построению регрессионной модели в общем не изменился. Соответственно этому обычным способом можем также проверить гипотезу о существовании некоторой другой экзогенной величины, незначительно отличающейся от оценки параметра $\hat{b}_2$. Пусть $\alpha = 0{,}01$, прямая гипотеза выдвигается для равенства: $b_2 = 0$, или $H_0 : b_2 = 0$, против альтернативной гипотезы $H_1 : b_2 \neq 0$. Согласно правилу проверки статистических гипотез нулевая гипотеза отбрасывается, если

$$H_0 \Leftrightarrow \frac{|\hat{b}_2 - b_{20}|}{\hat{\sigma}_{b_2}} > t_{n-3,1-\alpha/2} \Leftrightarrow \frac{|\hat{b}_2|}{\hat{\sigma}_{b_2}} > t_{1;0,995} \Leftrightarrow$$

$$\Leftrightarrow \frac{1{,}31034}{\sqrt{0{,}00206897}\sqrt{0{,}862069}} \approx 31{,}03? \leqslant ? t_{1;0,995} = 63{,}7.$$

Тестовое неравенство не выполняется, соответственно прямая гипотеза H_0 не может отбрасываться и следует признать, что значение $b_2 = 1{,}31034$ при заданных $\alpha = 0{,}01$, $m = 4$ и $m = 3$ несущественно отличается от $\hat{b}_2 = 0$.

Трансформации регрессионной модели. Для рассмотренных случаев статистического оценивания регрессии важно отметить, что на практике довольно часто встречаются специфичные типы моделей, для которых принципиальные подходы Гаусса—Маркова становятся недействительны. Далее при изложении теоретических вопросов о трансформации регрессионной модели и репараметризации ее ограничений этой проблеме уделяется больше внимания.

Оценивание методом наименьших квадратов с обобщениями (трансформация и оценивание по Айткину, репараметризация данных). Оценивание по Айткину требует нетрадиционных формулировок и введения новых обозначений. Вместо $y = Xb + u$ будем рассматривать $Py = PXb + Pu$, где P задает исходное условие оценивания для метода наименьших квадратов с обобщениями $P'P = V^{-1}$. При этом оценка параметра b представляется следующим образом:

$$\hat{b} = (X'P'PX)^{-1} X'P'Py = \left(X'V^{-1}X\right)^{-1} X'V^{-1}y.$$

Чтобы наглядно показать значение метода оценивания по Айткину, в последующем тексте приведены примеры, а также

объясняются общие принципы трансформации данных, которые не отвечают GLS-условиям (условиям оценивания методом наименьших квадратов с обобщениями)[1].

Линейная трансформация и оценивание. При линейных преобразованиях данных соответствующим образом трансформируется и рассмотренный метод статистического оценивания. Так, при переходе от $y = Xb + u$ к $Py = PXb + Pu$ с P — некоторой величиной, полученной при трансформации данных и принимающей ранговые значения, формальная оценка параметров регрессии изменяет свой вид: $\hat{b} = (X'P'PX)^{-1} X'P'Py$.

Обратим внимание, что после трансформации данных, а затем введения GLS-условия ограничения, заложенные в подходе Гаусса—Маркова, становятся необязательными.

Из большого числа существующих приемов трансформации данных приведем два простейших: первый предполагает предварительное умножение матрицы исходных данных на некоторую диагональную матрицу, второй — центрирование исходных данных. Эти виды трансформации имеют важное теоретическое значение и более подробно рассматриваются ниже.

Переход к центрированным данным, дешкалирование статистической величины. Пусть для обычной линейной модели с одним уравнением:

$$y = Xb + u,\ E(u) = 0,\ \operatorname{cov}(u) = E(uu') = V,\ u, y \in \boldsymbol{R}^n,\ X \in \boldsymbol{R}^{n,m},\ b \in \boldsymbol{R}^m$$

имеем центрированные данные, т. е. все значения признаков уменьшены на величину средней. При этом могут рассматриваться три различных случая:

1. Новые шкальные значения (другой масштаб) получает объясняемая величина y:

$$z_i = y_i - \bar{y},\ i = 1, 2, ..., n,\ \bar{y} := \sum_{i=1} \frac{y_i}{n}.$$

2. Новые шкальные значения приобретает объясняющая (экзогенная) величина X:

$$x_{ij} := X_{ij} - \bar{X}_j,\ i = 1,2,...,n,\ \bar{X}_j := \bar{X}_{.j} := \sum_{i=1}^{n} \frac{X_{ij}}{n},\ j = 1,2,...,m.$$

[1] GLS-условие (generalized least squares) — условие для метода наименьших квадратов с обобщениями. (*Прим. пер.*)

3. Новые шкальные значения по отклонениям от средних получает статистическая величина u:

$$v_i := u_i - \bar{u},\ i = 1, 2, ..., n,\ \bar{u} := \sum_{i=1}^{n} \frac{u_i}{n}\ .$$

Третий случай имеет только теоретическое значение, так как величина модельной ошибки u реально не наблюдается.

Использование в качестве исходных данных центрированных значений признаков позволяет наглядно продемонстрировать процесс последовательной трансформации данных с использованием некоторой определенной матрицы M:

$$M := I_n - \frac{ee'}{n},\ M \cdot M = M^2 = M,\ Me = 0,\ e'M = 0,\ e' = (11...1),\ \mathrm{rg}(M) \le n-1.$$

Конструктивно матрица M есть идемпотентная и, кроме того, ортогональная по отношению к собственным векторам e.

Использование M для центрирования данных представляет одну из основных областей приложения идемпотентных матриц и одновременно особый случай *проекционной* матрицы:

$$A = I_n - \frac{aa'}{a'a},\ a \in \boldsymbol{R}^n\ (a = e \text{ или } a \ne e).$$

При этом сами исходные данные для построения регрессионной модели записываются весьма просто:

$$z' = (z_1 z_2 ... z_n),\ y' = (y_1 y_2 ... y_n),\ e' = (111...1),\ z, y, e \in \boldsymbol{R}^n$$

и для трех случаев центрирования исходных данных имеем:

$$z = y - e\bar{y} \Leftrightarrow z = My \qquad e'z = 0 \in \boldsymbol{R},$$

$$x_j = X_j - \bar{X}_j e \Leftrightarrow x_j = MX_j \qquad e'x_j = 0 \in \boldsymbol{R},\ j = 1,2,...,m \Leftrightarrow$$

$$x = MX \qquad \Leftrightarrow e'x = 0 \in \boldsymbol{R}^{1,m},$$

$$v = u - \bar{u}e \Leftrightarrow v = Mu \qquad e'v = 0.$$

Как видим, суммы значений новых, центрированных, признаков как бы исчезают, становятся равными нулю. Причем равенство нулю выступает правилом и может служить элементарной проверкой корректности выполнения операции по трансформации данных.

Для особого случая, когда в матрице исходных данных первый вектор есть величина постоянная (см. выше для одновременного определения параметров регрессионного уравнения b_j

и константы b_0 в матрице исходных данных первый признак X_1 представляется единичным вектором), имеем:

$$X_1 := X_{\cdot 1} := b_0 e,\ b_0 \in \boldsymbol{R} \setminus \{0\} \quad \text{и} \quad x_1 = MX_1 = Mb_0 e = b_0 Me = 0 \in \boldsymbol{R}^n.$$

Другими словами, при центрировании данных первый вектор (единичный) трансформируется в нуль-вектор и из модели устраняется. Если же в модели для центрированных данных константа b_0 все-таки появляется в ходе расчетов, то ее следует рассматривать с учетом двух последующих замечаний.

Замечание 1. Обычная регрессионная модель, включающая свободный член (константу b_0), при ее параметризации может быть представлена с учетом векторной записи матрицы исходных данных следующим образом:

$$y = \sum_{j=1}^{m} X_j b_j + u \Leftrightarrow y = (X_1 X_2 ... X_m) \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{pmatrix} + u,\ E(u) = 0,\ Cov(u) = \sigma^2 I_n.$$

Если вектор-столбец X_1 в матрице X представляет значения признака, дополнительно введенного для определения константы b_0, т.е. является единичным вектором, то получим привычное представление модели:

$$y = eb_1 + X_0 b_0 + u,\ b_0' := (b_2 b_3 ... b_m),\ X := (e\ \ X_0),$$

и после трансформации посредством центрирования с использованием матрицы M будем иметь:

$$z = xb_0 + v,\ E(v) = 0,\ \mathrm{cov}(v) = E(vv') =: Q \in \boldsymbol{R}^{nn},\ \mathrm{rg}Q = n-1,$$

где $v, z \in \boldsymbol{R}^n$; $x \in \boldsymbol{R}^{n,m-1}$, $\mathrm{rg}\, x \le m-1$, $b_0' := (b_2, b_3, ..., b_m)$, $b_0 \in \boldsymbol{R}^{m-1}$.

Из приведенных выше аксиоматических положений следует, что трансформированная матрица ковариаций Q является единичной и имеет искаженное ранговое число, также равное единице. С учетом ортогональности e и M можем записать:

$$E(v) = 0,\ \mathrm{cov}(v) = M\,\mathrm{cov}(u)M' = \sigma^2 M.$$

Таким образом, если в случае центрирования исходных данных появляется константа b_0, то это фактически означает нару-

шение гипотетического распределения случайной величины u. Альтернативный подход, позволяющий устранить такие нарушения, рассматривается в замечании 2.

Замечание 2. Введем требование, что первоначально трансформация распространяется только на систематические величины y и X, а после трансформации исходных данных — также и на случайную величину u. Тогда модель приобретает вид:

$$z = xb_0 + v,\ E(v) = 0,\ \operatorname{cov}(v) = \sigma^2 I_n =: Q \in \mathbf{R}^{n,n},\ \operatorname{rg} Q = n,$$

$$v, z \in \mathbf{R}^n, x \in \mathbf{R}^{n,m-1},\ \operatorname{rg} x \le m-1;\ b_0 \in \mathbf{R}^{m-1}.$$

Обратим внимание, что модели в замечании 1 и замечании 2 различаются.

6.4.2. Репараметризация вектора модельных параметров

Кроме центрирования данных, другим весьма существенным моментом, обусловливающим решение проблем трансформирования данных и реформирования моделей, является репараметризация модельных параметров. Остановимся на ней более подробно.

Пусть в обычной линейной модели вектор b трансформирован

$$y = Xb + u,\ E(u) = 0,\ \operatorname{cov}(u) = \sigma^2 I; \qquad (6.13)$$

$$b = Z\gamma,\ b \in \mathbf{R}^m,\ \gamma \in \mathbf{R}^l, z \in \mathbf{R}^{m,l},\ b \to \gamma : \mathbf{R}^m,\ l \le m. \qquad (6.14)$$

Сделаем подстановку из (6.14) в (6.13) и получим:

$$y = (XZ)\gamma + u \Leftrightarrow y = \tilde{X}\gamma + u. \qquad (6.13)'$$

Используя метод наименьших квадратов до и после трансформации, будем иметь:

$\hat{b} = (X'X)^{-1}X'y$, для (6.13),

$\hat{\gamma} = (Z'X'XZ)^{-1}Z'Xy$ и $\tilde{b} := Z(Z'X'XZ)^{-1}Z'X'y = Z\hat{\gamma}$ для (6.13)′.

При этом закономерно

$\operatorname{rg} Z \le \operatorname{rg} X$, $\operatorname{rg}(Z'X'XZ) \le \operatorname{Min}(\operatorname{rg}(X), \operatorname{rg}(Z))$,

т.е. основанием для репараметризации является сокращение числа модельных параметров. Соответственно трансформационная модель (6.14) предполагает оценивание меньшего числа параметров (γ вместо $b, l \le m$).

Иллюстрация 1: Лаг треугольника: увеличение или сокращение?

$$\lambda_j = \frac{j}{m+1},\ j = 1,2,\ldots,m \Leftrightarrow \left\{\frac{1}{m+1}, \frac{2}{m+1}, \ldots, \frac{m-1}{m+1}, \frac{m}{m+1}\right\},$$

$$\lambda_j = \frac{m+1-j}{m+1},\ j = 1,2,\ldots,m \Leftrightarrow \left\{\frac{m}{m+1}, \frac{m-1}{m+1}, \ldots, \frac{2}{m+1}, \frac{1}{m+1}\right\},$$

$$\therefore y = Xb + u \Leftrightarrow y_i = \sum_{j=1}^{m} X_{i-j} b_j + u_i \Leftrightarrow y_i = \left(\sum_{j=1}^{m} X_{i-j}\lambda_j\right)\gamma + u_i,\ b_j = \gamma\lambda_j \Leftrightarrow$$

$$\Leftrightarrow y = XZ\gamma + u \Leftrightarrow y_i = \tilde{X}_i\gamma + u_i,\ \tilde{X}_i := \sum_{j=1}^{m} X_{i-j}\lambda_j,\ X \in \mathbf{R}^{n,m} \to XZ = \tilde{X} \in \mathbf{R}^{n,1}.$$

Иллюстрация 2: Лаг Альмона: лаговый подход при задании m опорных точек. Пусть имеем $(i_1, i_2, \ldots, i_m)$ — некоторые известные реальные числа, не равные нулю, причем эти числа являются не обязательно равновесными при задании опорных точек. Матрица репараметризации тогда может принимать вид:

$$\mathbf{R}^m \to \mathbf{R}^l,\ Z := \begin{pmatrix} 1 & i_1 & i_1^2 & \ldots & i_1^{l-1} \\ 1 & i_2 & i_2^2 & \ldots & i_2^{l-1} \\ 1 & i_3 & i_3^2 & \ldots & i_3^{l-1} \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ 1 & i_m & i_m^2 & \ldots & i_m^{l-1} \end{pmatrix} \in \mathbf{R}^{m,l} \Leftrightarrow b_k = \sum_{j=0}^{l-1} i_k^j \gamma_k \Leftrightarrow b = Z\gamma.$$

Тем самым модель $y = Xb + u$, редуцированная по числу параметров, записывается как:

$$y = XZ\gamma + u \Leftrightarrow y = \tilde{X}\gamma + u,\ X \in \mathbf{R}^{m,n} \to \tilde{X} \in \mathbf{R}^{l,n},\ b \in \mathbf{R}^m \to b_0 \in \mathbf{R}^l,\ l < m.$$

Иллюстрация 2.1 (дополнение): Лаг Альмона с равными интервалами.

Пусть в числе экзогенных переменных имеются переменные с различной скоростью изменения (например, с замедлением) их значений:

$$y = Xb + u \Leftrightarrow y_i = \sum_{j=0}^{l} X_{i-j} b_j + u_i,\ m = l+1,\ b_j = \sum_{s=0}^{r} \gamma_s j^s,\ r \le l \Leftrightarrow b = Z\gamma,$$

$$Z := \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 4 & \dots & 2^r \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 1 & l & l^2 & \dots & l^r \end{pmatrix}, \quad y_i = \sum_{j=0}^{l}\sum_{s=0}^{r} X_{i-j}\gamma_s j^s + u_i = \sum_{s=0}^{r}\left(\sum_{j=0}^{l} j^s X_{i-j}\right)\gamma_s + u_i,$$

$$y_i = \sum_{s=0}^{r} \tilde{X}_s\gamma_s + u_i \Leftrightarrow y = (XZ)\gamma + u \Leftrightarrow y = \tilde{X}\gamma + u.$$

При условии, что число опорных точек $\{j\} = \{i_j\}$, $j = 1, 2, \dots, r$, общее количество модельных параметров, подлежащих оцениванию, может быть сокращено с $m = l+1$ до $r+1$.

Иллюстрация 3: Распределенный лаг в модели Койка, лаг бесконечной величины.

Предположим, следует произвести оценивание линейной модели с одним уравнением:

$$y_i = \sum_{j=0}^{\infty} a_j x_{i-j} + u_i \Leftrightarrow y_i = a_0 x_i + a_1 x_{i-1} + a_2 x_{i-2} + a_3 x_{i-3} + \dots + u_i, \qquad (6.15)$$

$$b_1 = a_0,\ b_2 = a_1, \dots, b_j = a_j, \dots,\ j \to \infty.$$

Подобная модель, имеющая бесконечное число параметров, в общем не подлежит статистическому оцениванию. В то же время при помощи подхода репараметризации Койка число коэффициентов регрессии может быть сокращено весьма существенно, скажем до двух, а затем, оценено:

$$a_j = bc^j \ (j = 0, 1, 2, 3, \dots),\ b > 0,\ 0 < c < 1 \Leftrightarrow a_0 = b,\ a_1 = bc,\ a_2 = bc^2,$$

$$a_3 = bc^3 \dots (j = 0, 1, 2, 3, \dots),$$

$$y_i = b\sum_{j=0}^{\infty} c^j x_{i-j} + u_i. \qquad (6.15')$$

При лаговой инвариантности из (6.15) следует

$$y_i = bx_i + bcx_{i-1} + bc^2 x_{i-2} + bc^3 x_{i-3} \cdots,$$

$$cy_{i-1} = bcx_{i-j} + bc^2 x_{i-2} + bc^3 x_{i-3} \cdots + cu_{i-1}$$

и в общем виде

$$y_i - cy_{i-1} = b\sum_{j=0}^{\infty} c^j x_{i-j} - bc\sum_{j=0}^{\infty} c^j x_{i-1-j} + u_i - cu_{i-1} = bx_i + u_i - cu_{i-1}. \qquad (6.16)$$

Сжатие модели с любым исходным числом параметров до двухпараметрической модели носит название метода Койка или трансформации Койка:

$$y_i = bx_i + cy_{i-1} + u_i - cu_{i-1} \Leftrightarrow y_i = b_1 x_1 + b_2 y_{i-1} + v_i,\ v_i := u_i - cu_{i-1}. \qquad (6.16')$$

Собственно это обычная модель с одним уравнением, включающим два параметра, для оценивания которых могут использоваться различные методы. Например, приложение методов GLS-оценивания в данном случае требует использования матрицы ковариаций остатков модели u, x считается величиной, линейно независимой от u и одновременно гомоскедастичной:

$$E(u_i) = 0,\ E(u_i u_r) = \begin{cases} \sigma^2, & i = r \\ 0, & i \neq r. \end{cases}$$

Исходя из предположения о стохастичности u, можем для v записать:

$$E(v_i) = 0,\ E(v_i v_r) = \begin{cases} E(u_i - cu_{i-1})^2 = \sigma^2(1+c^2) & i = r \\ E[(u_i - cu_{i-1})(u_r - cu_{r-1})] = \begin{cases} -c\sigma^2, & i = r \pm 1 \\ 0, & \text{в других случаях.} \end{cases} \end{cases}$$

При этом ковариационная матрица $\mathrm{cov}(v) = \sigma^2 V$ в развернутом виде будет:

$$V = \begin{pmatrix} 1+c^2 & -c & 0 & \dots & 0 & 0 \\ -c & 1+c^2 & -c & \dots & 0 & 0 \\ 0 & -c & 1+c^2 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & -c & 0 \\ 0 & 0 & 0 & \dots & 1+c^2 & -c \\ 0 & 0 & 0 & \dots & -c & 1+c^2 \end{pmatrix} = \tau\left(1+c^2, 1+c^2, 1+c^2; -c\right).$$

В дальнейшем доказательство нашего исходного предположения получим в результате непосредственного обобщения теоретических выводов, например, для $E(v_i v_{i-1})$, если объясняющие признаки x_i и y_{i-1} коррелируют с ошибкой модели, то:

$$E(v_i y_{i-1}) = E\left[(u_i - cu_{i-1})\left(b\sum_{j=0}^{\infty} c^j x_{i-1-j} + u_{i-1}\right)\right] = E(-cu_{i-1}u_{i-1}) = -c\sigma^2 .$$

Аналогично можно получить и другие показатели.

Иллюстрация 4: Оценивание с лагом: процесс сжатия формы регрессии (*VAR*/ *p*/ -процесс), реализуемый для простейшей линейной модели.

а) Специализация модели процесса редуцирования (сжатия) регрессии, или *AR*/ *p*/ -процесса, имеет вид:

$$y_i = r_1 y_{i-1} + r_2 y_{i-2} + r_3 y_{i-3} + \ldots + r_p y_{i-p} + u_i,\ E(u_i) = 0,\ \mathrm{var}(u_i) = \sigma^2 .$$

В сущности данная модель может рассматриваться как гомоскедастичная, для которой существуют два различных случая:

$$\left\{z = y \in \boldsymbol{R}^n,\ X \in \boldsymbol{R}^{n,p},\ b = r \in \boldsymbol{R}^p,\ E(v) = 0,\ E(vv') = V = \sigma^2 I\right\} \text{ и}$$
$$\left\{z = y \in \boldsymbol{R}^{n-p},\ X \in \boldsymbol{R}^{n-p,p},\ b = r \in \boldsymbol{R}^p,\ E(v) = 0,\ V = \sigma^2 I_{n-p}\right\}.$$

Принимая во внимание область определения переменных и учитывая принятые ранее обозначения, построим производную модель первого порядка для *VAR* -процесса (*VAR*/1/ -процесса):

$$z_i = Rz_{i-1} + v_i ,$$

$$z_i = \begin{pmatrix} y_i \\ y_{i-1} \\ y_{i-2} \\ \cdot \\ \cdot \\ \cdot \\ y_{i-p+2} \\ y_{i-p+1} \end{pmatrix},\ R = \begin{pmatrix} r_1 & r_2 & r_3 & & r_{p-2} & r_{p-1} & r_p \\ 1 & 0 & 0 & & 0 & 0 & 0 \\ 0 & 1 & 0 & & 0 & 0 & 0 \\ & \cdot & & & & \cdot & \\ & \cdot & & \ldots & & \cdot & \\ & \cdot & & & & \cdot & \\ 0 & 0 & 0 & & 1 & 0 & 0 \\ 0 & 0 & 0 & & 0 & 1 & 0 \end{pmatrix},\ v_i = \begin{pmatrix} u_i \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{pmatrix},\ r = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ \cdot \\ \cdot \\ \cdot \\ r_{p-1} \\ r_p \end{pmatrix}.$$

Приведенная модификационная модель определяется эндогенным вектором z и матрицей исходных данных X. На основе

исходной модели можно получить следующие важные теоретические выводы:

1) задавая величине p стартовые оценки $(y_{i-1}, y_{i-2}, y_{i-3}, \ldots, y_{i-p})$, будем иметь:

$$z = \begin{pmatrix} y_i \\ y_{i+1} \\ y_{i+2} \\ \cdot \\ \cdot \\ \cdot \\ y_{i+n-1} \end{pmatrix},\ X = \begin{pmatrix} y_{i-1} & y_{i-2} & y_{i-3} & \cdots & y_{i-p+1} & y_{i-p} \\ y_i & y_{i-1} & y_{i-2} & \cdots & y_{i-p+2} & y_{i-p+1} \\ y_{i+1} & y_i & y_{i-1} & \cdots & y_{i-p+3} & y_{i-p+2} \\ & & & \cdot & & \\ & & & \cdot & & \\ & & & \cdot & & \\ y_{i+n-2} & y_{i+n-3} & y_{i+n-4} & \cdots & y_{i+n-p} & y_{i+n-p-1} \end{pmatrix};$$

2) допуская замену стартовых оценок значениями эндогенной переменной $(y_i, y_{i+1}, y_{i+2}, \ldots, y_{i+p-1})$, перейдем к сокращенной в сравнении с предыдущим случаем модели для тех же z и X:

$$z = \begin{pmatrix} y_{i+p} \\ y_{i+p+1} \\ y_{i+p+2} \\ \cdot \\ \cdot \\ \cdot \\ y_{i+n-1} \end{pmatrix},\ X = \begin{pmatrix} y_{i+p-1} & y_{i+p-2} & y_{i+p-3} & \cdots & y_{i+1} & y_i \\ y_{i+p} & y_{i+p-1} & y_{i+p-2} & \cdots & y_{i+2} & y_{i+1} \\ y_{i+p+1} & y_{i+p} & y_{i+p-1} & \cdots & y_{i+3} & y_{i+2} \\ & & & \cdot & & \\ & & & \cdot & & \\ & & & \cdot & & \\ y_{i+n-2} & y_{i+n-3} & y_{i+n-4} & \cdots & y_{i+n-p} & y_{i+n-p-1} \end{pmatrix},$$

$$\therefore\ y = Xb + u,\quad y = z \in \boldsymbol{R}^{n-p},\quad X \in \boldsymbol{R}^{n-p,p},\quad b = r \in \boldsymbol{R}^p .$$

6.4.3. Параметрические ограничения для регрессионной модели

Обзор случаев с параметрическими ограничениями. Параметрические ограничения с учетом аксиоматических положений оценивания по Айткину имеют силу и для обычных, и для модифицированных форм регрессии. Основополагающим при этом является предположение априори о линейных ограничениях для

модельных параметров. Будем различать следующие случаи ограничений:

1) нуль ограничения,
2) гомогенные детерминированные ограничения[1],
3) общие линейные детерминированные ограничения,
4) общие линейные стохастические ограничения.

Введение ограничений и их учет предопределяют изменения форм регрессий и существенно увеличивают потенциальное число вариантов оценивания методом наименьших квадратов. В качестве одного из таких вариантов, например, можно рассматривать случай 4, принадлежащие ему алгоритмы позволяют имеющуюся некоторую стохастическую систему как, собственно и модели, представленные в п. 6.4.6, привести к стандартному виду (см. п. 6.4.1). Теоретической основой подобной трансформации выступает подход Гаусса—Маркова, определяющий статистические свойства параметрических оценок $\hat{b}$ (п. 6.4.1, с. 264). Все это еще раз указывает на широкий спектр практического приложения метода наименьших квадратов.

Исследование ограничений в регрессионной модели в последующем подводит нас к весьма важным заключениям относительно GLS-трансформаций и возможности расширенной демонстрации подхода Гаусса—Маркова.

С целью упростить изложение материала процесс параметрического оценивания при помощи различных модификаций метода наименьших квадратов, обусловленных накладываемыми ограничениями, рассматривается главным образом без описания структуры стохастических величин (u,v). При этом, однако, сохраняется предположение о существовании остатков модели, как и оценок их математического ожидания $E(u)$ и ковариации $\mathrm{cov}(u)$.

Линейные параметрические оценки при нуль-ограничениях. Репараметризация урезанной модели. В линейной регрессионной модели посредством наложения нуль-ограничения соответствующая компонента параметрического вектора обращается в нуль, например, $b_2=0$, и

$$b:=\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad b\in R^m, \quad b_1\in R^{j_1}, \quad b_2\in R^{j_2}, \quad j_1+j_2=m;$$

[1] Детерминированные, т.е. основывающиеся на жестких, функциональных связях. (*Прим. пер.*)

$$\{b_2=0,\ 0\in R^{j_2}\}\Leftrightarrow\{Rb=0\}\Leftrightarrow\left\{\left(0_{j_2,j_1}\ I_{j_2}\right)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}=0,\ 0\in R^{j_2}\right\}.$$

Наиболее просто и естественно рассматривать нуль-ограничение (нуль-коэффициент) как результат элиминирования соответствующих компонентов параметрического вектора b столбца в матрице исходных данных X. Сокращенная (урезанная или редуциро регрессионная модель принимает вид:

$$y=X_1b_1+u.$$

Решение модели опосредуется процессом репараметризации:

$$y=XS'Sb+u=ZSb+u,\ S=\left(I_{j_1}0_{j_1,j_2}\right),\ S'S=\begin{pmatrix} I \\ 0 \end{pmatrix}(I\quad 0)=\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix},$$

$$Z:=XS'=\left(X_1|X_2\right)\begin{pmatrix} I \\ 0 \end{pmatrix}=X_1,\ SX'XS'=Z'Z=\begin{pmatrix} X_1'X_1 & 0 \\ 0 & 0 \end{pmatrix}=:M,$$

$$M^+=\begin{pmatrix} \left(X_1'X_1\right)^{-1} & 0 \\ 0 & 0 \end{pmatrix},\ Sb=\left(I_{j_1}\quad 0_{j_1,j_2}\right)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}=b_1,$$

$$Z'y=SX'y=\left(I\quad 0\right)\begin{pmatrix} X_1' \\ X_2' \end{pmatrix}y=X_1'y;$$

$$y=X_1b_1+u\Leftrightarrow y=XS'Sb+u=ZSb+u.$$

Для урезанной модели $y=X_1b_1+u$ получим оценки:

$$\tilde{b}=(Z'Z)^+Z'y=\begin{pmatrix} \tilde{b}_1 \\ 0 \end{pmatrix}\therefore\tilde{b}_1=(X_1'X_1)^{-1}X_1'y \quad \text{вместо} \quad \hat{b}=(X'X)^{-1}X'y.$$

Параметрические линейные оценки для модели с гомогенными ограничениями, отличными от нуля. Для модели вида:

1) $y=Xb+u_1,\ E(u_1)=0,\ \mathrm{cov}(u_1)=V=\sigma^2I$

найдем оценки параметра b при условии, что

$r=Rb, r=0, R\neq I, S$ (нуль-ограничение);

2) $X\in R^{n,m}, \mathrm{rg}(X)=m$ (m — число независимых столбцов матрицы X);

3) $R\ R^{q,m},\ \mathrm{rg}(R)=q<m$ (q — число независимых строк матрицы R).

С учетом введенных гомогенных ограничений для параметрического вектора получим сокращенную форму модели. Пусть

$$(R_1 \quad R_2)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = 0$$

и либо $b_1 = -R_1^{-1}R_2 b_2$ (R_1 — неединичная матрица),

либо $b_2 = -R_2^{-1}R_1 b_1$ (R_2 — неединичная матрица).

Модель 1) в таком случае принимает вид:

$$y = \left(X_1 - X_2 R_2^{-1} R_1\right) b_1 + u \Leftrightarrow y = \dot{X}_1 b_1 + u . \qquad (6.17)$$

Обобщая выводы, отметим, что при условии:

$$(R_1 \quad R_2)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = 0 \therefore b_1 = -R_1^{-1}R_2 b_2 \text{ или } b_2 = -R_2^{-1}R_1 b_1 ;$$

$$y = \left(X_1 - X_2 R_2^{-1} R_1\right) b_1 + u \Leftrightarrow y = \dot{X}_1 b_1 + u,\ \dot{X} := X_1 - X_2 R_2^{-1} R_1,$$

будем иметь обычную МНК-оценку:

$$\tilde{b}_1 = \left(\dot{X}_1{}' \dot{X}_1\right)^{-1} \dot{X}_1{}' y \ .$$

Любые негомогенные ограничения

1) $y = Xb + u_1, E(u_1) = 0,\ \operatorname{cov}(u_1) = V = \sigma^2 I,$

2) $r = Rb,\ r \neq 0,\ R \neq I, S$

3) $X \in \boldsymbol{R}^{n,m}, \operatorname{rg}(X) = m$ (m — число независимых столбцов матрицы X);

4) $R \in \boldsymbol{R}^{q,m}, \operatorname{rg}(R) = q < m$ (q — число независимых строк матрицы R).

При помощи соответствующих замен, как и в предыдущем случае с гомогенными ограничениями, перейдем к сокращенной форме регрессионной модели и если R — некоторая неединичная матрица, например, R_2, то получим[1]:

$$r = (R_1 \quad R_2)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \Leftrightarrow b_2 = R_2^{-1}(r - R_1 b_1)\text{, откуда следует}$$

$$y = (X_1 \quad X_2)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + u \Leftrightarrow \left(y - X_2 R_2^{-1} r\right) = \left(X_1 - X_2 R_2^{-1} R_1\right) b_1 + u \Leftrightarrow$$

$$\Leftrightarrow \dot{y} = \dot{X}_1 b_1 + u. \qquad (6.18)$$

[1] Здесь матрица ограничений R не имеет ничего общего с характеристикой связи R^2 (*Прим. авт.*)

Смешанная регрессия. Альтернативным и в то же время обобщающим подходом для рассмотренных выше задач с ограничениями является следующий. Будем исходить из условия существования некоторой модели с некоррелированными остатками, включающей более одного уравнения:

$$y = Xb + u_1,\ E(u_1) = 0,\ \operatorname{cov}(u_1) = \sigma^2 \Omega_1 , \qquad (6.19)$$

$\operatorname{rg}(X) = m$ (m — число независимых столбцов в матрице X).

Эта же модель после трансформации будет

$$r = Rb + u_2,\ E(u_2) = 0,\ \operatorname{cov}(u_2) = \sigma^2 \Omega_2 , \qquad (6.20)$$

$\operatorname{rg}(R) = q < m$ (q — число независимых строк матрицы R),

$E\left(u_1 u_2{}'\right) = \Omega_{12} = 0$ (некоррелированные остатки моделей (6.19) и (6.20)).

Такая модель представляет более высокий уровень обобщения, поскольку в ней допускаются ограничения с ошибками, соответственно в ней присутствует стохастическая величина u_2. Обобщенная форма модели при этом записывается следующим образом:

$$\tilde{y} = \begin{pmatrix} y \\ r \end{pmatrix},\ \tilde{X} := \begin{pmatrix} X \\ R \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \\ R_1 & R_2 \end{pmatrix},\ \tilde{X} \in R^{n+q,m};$$

$$u := \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},\ \Omega := \begin{pmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{pmatrix} = \operatorname{diag}(\Omega_s),\ s = 1,2 ;$$

$$\tilde{y} = \tilde{X} b + u,\ E(u) = 0,\ \operatorname{cov}(u) = \sigma^2 \Omega .$$

Это обычная линейная модель или так называемая модель смешанной регрессии. Введение дополнительных условий означает в сущности ее трансформацию.

Пусть:

$$P = \begin{pmatrix} I_n & -X_2 R_2^{-1} \\ 0 & I_q \end{pmatrix},\ P^{-1} = \begin{pmatrix} I_n & X_2 R_2^{-1} \\ 0 & I_q \end{pmatrix},$$

тогда следует $\dot{y} = \dot{X} b + \dot{u}$ с

$$P\tilde{y} = \begin{pmatrix} I_n & -X_2 R_2^{-1} \\ 0 & I_q \end{pmatrix}\begin{pmatrix} y \\ r \end{pmatrix} = \begin{pmatrix} y - X_2 R_2^{-1} r \\ r \end{pmatrix} = \begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} \dot{y}_1 \\ r \end{pmatrix} = \dot{y} ,$$

$$P\tilde{X}=\begin{pmatrix} I_n & -X_2R_2^{-1}\\ 0 & I_q\end{pmatrix}\begin{pmatrix} X_1 & X_2\\ R_1 & R_2\end{pmatrix}=\begin{pmatrix} X_1-X_2R_2^{-1}R_1 & 0\\ R_1 & R_2\end{pmatrix}=:\begin{pmatrix}\dot{X}_{11} & 0\\ \dot{X}_{21} & \dot{X}_{22}\end{pmatrix}=\dot{X},$$

$$Pu=\begin{pmatrix} I_n & -X_2R_2^{-1}\\ 0 & I_q\end{pmatrix}\begin{pmatrix} u_1\\ u_2\end{pmatrix}=\begin{pmatrix} u_1-X_2R_2^{-1}u_2\\ u_2\end{pmatrix}=\dot{u},$$

$$\therefore\ P\left(\tilde{y}=\tilde{X}b+u\right)\Leftrightarrow \dot{y}=\dot{X}_1b_1+\dot{u}_1.$$

Иллюстрация. Обобщение данных.

Покажем пример смешанной модели при обобщении данных двух выборочных совокупностей:

$$\begin{pmatrix} y_1\\ y_2\end{pmatrix}=\begin{pmatrix} X_1 & 0\\ 0 & X_2\end{pmatrix}\begin{pmatrix} b_1\\ b_2\end{pmatrix}+\begin{pmatrix} u_1\\ u_2\end{pmatrix}\quad\begin{matrix}(y_1,X_1)\\ (y_2,X_2)\end{matrix}\quad\begin{matrix}\text{(данные временного ряда),}\\ \text{(выборочные данные из временного ряда),}\end{matrix}$$

$$\Leftrightarrow y=Xb+u,\ E(u)=0,\ cov(u)=\sigma^2V=\sigma^2\begin{pmatrix} V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix}.$$

Результаты решения подобной модели, т.е. параметры b_1 и b_2 — обычно неравные, причем это не зависит от того, каким образом произведена выборка: поперечным, продольным срезом или по типу панельной выборки.

З а м е ч а н и е. Замена детерминированных ограничений.

В общем случае система регрессионных уравнений представляется смешанной моделью, но при условии, что X_2 — случайная величина, модельный остаток нивелируется и устраняется из модели:

$$Pu=\begin{pmatrix} I_n & -X_2R_2^{-1}\\ 0 & I_q\end{pmatrix}\begin{pmatrix} u_1\\ 0\end{pmatrix}=\begin{pmatrix} u_1\\ 0\end{pmatrix}=\dot{u}.$$

В результате прямого расширения МНК-оценки $\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{y}$ получим:

$$\tilde{b}=\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{y}=(X'X+R'R)^{-1}(X'y+R'r).$$

Применение МНК-оценивания после соответствующих замен и сжатия модели приводит к следующим выводам:

$$\dot{y}=\dot{X}b+\dot{u}\Leftrightarrow \dot{y}_1=\dot{X}_{11}b_1+u_1,\ b_1\in\boldsymbol{R}^{m-q}\ \text{и}$$

$$\left\{\dot{y}_2=R_1b_1+R_2b_2\right\}\Leftrightarrow\left\{r=\begin{pmatrix} R_1 & R_2\end{pmatrix}\begin{pmatrix} b_1\\ b_2\end{pmatrix}\Leftrightarrow b_2=R_2^{-1}\left(r-R_1b_1\right)\right\};$$

$$\tilde{b}_1=\left(\dot{X}_{11}'\dot{X}_{11}\right)^{-1}\dot{X}_{11}'\dot{y}_1.$$

Статистическую оценку $\tilde{b}_1$ для b_1 получают из обычной урезанной линейной модели. Оценку $\tilde{b}_2$ для b_2 соответственно получают на основе изменения формы регрессии посредством введения в нее ограничения после определения первого параметрического вектора:

$$\therefore\tilde{b}_2:=R_2^{-1}\left(r-R_1\tilde{b}_1\right).$$

З а м е ч а н и е. Обзор особых случаев введения ограничений.

Запишем спецификации для различного вида ограничений:

- Нуль-ограничения —

  $R_1=0$, $R_2=I$ и $r=0$, как и $u_2\equiv 0$.
- Гомогенные ненулевые ограничения —

  R_1 — любая матрица, R_2 — неединичная матрица и $r=0$; $u_2\equiv 0$.
- Обобщенный случай негомогенных ограничений —

  R_1 — любая матрица, R_2 — неединичная матрица и $r\neq 0$; $u_2\equiv 0$.
- Смешанная модель —

  R_1 — любая матрица, R_2 — неединичная матрица и $r\neq 0$; $u_2\neq 0$.

Вместо ограничений могут рассматриваться также возможности замещения переменных и стандартизации регрессионной формы.

Линейные оценки при дополнительных детерминированных условиях. Для линейной гомоскедастичной модели с одним уравнением ($\Omega_1=I,\Omega_2\equiv 0$) вида

$$y=Xb+u,\ E(u)=0,\ \operatorname{cov}(u)=V=\sigma^2I,\ r=Rb,\ r\neq 0, \tag{6.21}$$

методом наименьших квадратов получим оценку

$$\tilde{b}=\hat{b}+(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}\left(r-R\hat{b}\right);\ \hat{b}:=(X'X)^{-1}X'y. \tag{6.22}$$

□ *Д о к а з а т е л ь с т в о*: По определению, метод наименьших квадратов минимизирует ошибку модели:

$$L_1=\left(y-Xb\right)'\left(y-Xb\right)\to\min_b,\ \text{при этом}\ r=Rb.$$

Подобный результат может быть достигнут также при помощи функции Лагранжа:

$$L_2 = (y - Xb)'(y - Xb) - 2\lambda'(r - Rb) \to \underset{b,\lambda}{\text{Min}} \Leftrightarrow$$

$$\Leftrightarrow L_2 = y'y - 2y'Xb + b'X'Xb - 2\lambda'r + 2\lambda'Rb \to \underset{b,\lambda}{\text{Min}}.$$

Решение для функции Лагранжа следует из предположения о существовании экстремумов:

$$\frac{\partial L_2}{\partial b} = 0 = -2y'X + 2b'X'X + 2\lambda'R,$$

$$\frac{\partial L_2}{\partial \lambda} = 0 = -2(r - Rb)';$$

$$\therefore \tilde{b} = (X'X)^{-1}X'y - (X'X)^{-1}R'\lambda = \hat{b} - (X'X)^{-1}R'\lambda,$$

$$r = R\tilde{b} \Leftrightarrow \lambda = -\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{b} - r\right) = -S\left(R\hat{b} - r\right), \quad S := \left(R(X'X)^{-1}R'\right)^{-1},$$

$$\tilde{b} = \hat{b} - (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{b} - r\right) = \hat{b} + (X'X)^{-1}R'S\left(r - R\hat{b}\right) \blacksquare$$

GLS-оценка при дополнительных детерминированных условиях и с любой неединичной матрицей ковариаций. Для обобщенной формы линейной модели с одним уравнением:

$$y = Xb + u,\ E(u) = 0,\ \operatorname{cov}(u) = \sigma^2 V,\ r = Rb \tag{6.23}$$

методом наименьших квадратов можно получить следующую параметрическую оценку:

$$\tilde{b} = \hat{b} + (X'V^{-1}X)^{-1}R'\left[R(X'V^{-1}X)^{-1}R'\right]^{-1}(r - R\hat{b}),$$

$$\hat{b} := (X'V^{-1}X)^{-1}X'y. \tag{6.24}$$

Замечание. Параметрическая оценка в случае гомоскедастичности модели будет:

$$\hat{b} = \left((X_1' \quad X_2')\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right)^{-1}(X_1' \quad X_2')\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (X_1'X_1 + X_2'X_2)^{-1}(X_1'y_1 + X_2'y_2),$$

т.е. она эквивалентна оценке для модели смешанной регрессии

$$\hat{b} = \left((X' \quad R')\begin{pmatrix} X \\ R \end{pmatrix}\right)^{-1}(X'R')\begin{pmatrix} y \\ r \end{pmatrix} = (X'X + R'R)^{-1}(X'y + R'r),$$

или то же, но в привычной форме записи регрессионной модели

$$\tilde{y} = \tilde{X}b + u \Leftrightarrow \begin{pmatrix} y \\ r \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix}b + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \Leftrightarrow \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}b + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

GLS-оценки для особых случаев гомогенных детерминированных ограничений. Линейная модель с одним уравнением является частным случаем модели, содержащей систему уравнений:

$$y = Xb + u,\ E(u) = 0,\ \operatorname{cov}(u) = \sigma^2 I,\ y, u \in \mathbf{R}^n,\ X \in \mathbf{R}^{n,m},\ \operatorname{rg}(X) = m,\ b \in \mathbf{R}^m$$

и $Rb = r = 0,\ 0 \equiv r \in \mathbf{R}^q,\ R \in \mathbf{R}^{q,m},\ q \le m.$

Оценивание параметров модели общего вида с любым числом уравнений производится специфичным образом:

$$\tilde{b} = \left(I - (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R\right)\hat{b} = \left(I - (X'X)^{-1}R'SR\right)\hat{b} =$$

$$= \hat{b} - (X'X)^{-1}R'SR\hat{b}, S' = S = \left(R(X'X)^{-1}R'\right)^{-1}.$$

При отсутствии ограничений для общей регрессионной модели, очевидно, будем иметь неизменную оценку: $\{\tilde{b} \mid 0 \equiv r, 0 \equiv R\} = \hat{b} := (X'X)^{-1}X'y$ так же, как и после введения условия $R\tilde{b} = 0$, когда сама оценка начинает выполнять роль ограничения и появляется возможность непосредственно проводить вычисления:

$$E(\tilde{b}) = b,\ \operatorname{cov}(\tilde{b}) = \sigma^2\left[I - (X'X)^{-1}R'SR\right](X'X)^{-1}.$$

Иллюстрация. Отдельные примеры распространенных случаев гомогенных и негомогенных линейных ограничений.

Пусть имеем:

$$Rb = r,\ b \in \mathbf{R}^m,\ r \in \mathbf{R}^q,\ R \in \mathbf{R}^{q,m},\ q \le m.$$

Выделим несколько возможных случаев:

1. *Все коэффициенты регрессии принимают некоторые определенные значения:* $b = b_0$,

$$r = b_0;\ R = I_m,\ b_0\begin{cases} = 0 \\ \ne 0 \end{cases},\quad q < m.$$

2. *Некоторые коэффициенты регрессии принимают определенные значения:*

$$b_j = b_{j0},\ j \in \{1, 2, \ldots, m\},$$

$$r = b_0;\ R = (I_q \mid 0);\ q < m.$$

Наиболее важным здесь выступает частный случай, когда $b_j := 0$, т.е. имеем дело с нулевым коэффициентом регрессии или нуль-ограничением.

3. *Нуль-ограничения.*

3.1. *Имеется подмножество нулевых коэффициентов регрессии.*

Допустим, что из модели удалено некоторое число объясняющих переменных:

$b_1 \neq 0,\ b_2 = 0,\ R := (0 I_q);\ r := 0 \in R^q$ (нулевой вектор),

при этом отдельные модельные параметры неоправданно обращаются в нуль. Данный случай связывается с проблемой идентификации линейной модели, для которой по возможности предполагается определение значений отсутствующих параметров на основе представленных объясняющих переменных.

Проанализируем решение подобной задачи. Следуя подходу Гаусса—Маркова, можем записать:

$$y = Xb + u = (X_1 | X_2)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + u,\ E(u) = 0,\ \operatorname{cov}(u) = \sigma^2 I_n .$$

Вместо $\hat{b} = (X'X)^{-1}X'y$ используем $\hat{\hat{b}}_1 = (X_1'X_1)^{-1}X_1'y$, установим также, что $\hat{b}_2 := 0$, тогда имеем:

$$E\left(\hat{\hat{b}}_1\right) = \left\{(X_1'X_1)^{-1}X_1'(X_1b_1 + X_2b_2 + u)\right\} = b_1 + (X_1'X_1)^{-1}X_1'X_2b_2 .$$

Последнее означает смещение частной оценки $\hat{\hat{b}}_1$, поскольку $b_2 \neq 0$:

$$\operatorname{cov}\left(\hat{\hat{b}}_1\right) = E\left\{(AX_2b_2 + Au)(AX_2b_2 + Au)'\right\} \quad \text{с } A := (X_1'X_1)^{-1}X_1' =$$

$$= AX_2b_2b_2'X_2'A' + E(AX_2b_2u'A') + E(Aub_2'X_2'A') + E(Auu'A) =$$

$$= AX_2b_2b_2'X_2'A' + AE(uu')A = AX_2b_2b_2'X_2'A' + \sigma^2(X_1'X_1)^{-1} .$$

Вывод: поскольку $b_2 \neq 0$, $\hat{\hat{b}}_1$ не может быть эффективной оценкой.

3.2. *Иерархия статистических гипотез.*

На основе одного имеющегося уравнения регрессии, при последовательном введении альтернативных нуль-ограничений, можно получать множество модифицированных моделей. Покажем это на примере трехпараметрической модели:

$y_i = b_0 + b_1X_{1i} + b_2X_{2i} + u_i$ (исходная модель)

$y_i = b_0 + b_1X_{1i} + u_i$ (двупараметрическая модификация)

$y_i = b_0 + b_2X_{2i} + u_i$ (двупараметрическая модификация)

$y_i = b_1X_{1i} + b_2X_{2i} + u_i$ (модель без константы)

$y_i = b_0 + u_i$ (оценивание средней величины)

$y_i = b_1X_{1i} + u_i$ (однопараметрическая модификация)

$y_i = b_2X_{2i} + u_i$ (однопараметрическая модификация)

$y_i = +u_i$ (чистый стохастический процесс)

4. *Ненулевые ограничения.*

Введем условие существования для коэффициентов регрессии с заданными весами некоторых суммарных значений:

$$\sum_{j=1}^{m} c_jb_j = r,\ r \in R \Leftrightarrow Rb = c'b = r;\ c = e \vee c \neq e ;$$

$$R := (c_1c_2...c_m) =: c',\ c_j \neq 0,\ j \in \{1, 2, ..., m\};\ r = 0 \text{ или } r \neq 0 .$$

Используя известный метод Лагранжа, а также принимая во внимание коммутативные свойства скаляров, произведем прямую корректировку параметрической оценки $\hat{b} = (X'X)^{-1}X'y$ и получим:

$$\tilde{b} = (X'X)^{-1}\left(X'y + \frac{r - c'\hat{b}}{s}c\right),\ \left(R(X'X)^{-1}R'\right)^{-1} = \left(c'(X'X)^{-1}c\right)^{-1} =: \frac{1}{s} > 0.$$

4.1. *Оценивание производственной функции Кобба—Дугласа.*

Нелинейная модель сокращенной производственной функции (производственной функции Кобба—Дугласа с заданной эластичностью шкалы признаковых значений и мультиплицированной ошибкой) после ее логарифмирования и замены обозначений трансформируется в обычную линейную модель:

$$Q_i = \alpha_0K_i^{\alpha_1}L_i^{\alpha_2}e^{u_i},\ \alpha_1 + \alpha_2 = r > 0 \Leftrightarrow$$

$$\Leftrightarrow \ln Q_i = \ln\alpha_0 + \alpha_1 \ln K_i + \alpha_2 \ln L_i + u_i .$$

Пусть $y_i = \ln Q_i;\ X_{1i} = 1;\ X_{2i} = \ln K_i;\ X_{3i} = \ln L_i, b_1 := \ln\alpha_0;\ b_2 := \alpha_1;\ b_3 := \alpha_2$, тогда

$$y_i = \sum_{j=1}^{3} X_{ji}b_j + u_i \Leftrightarrow y = Xb + u,\ b_2 + b_3 = r \Leftrightarrow Rb = r;\ R = (0, 1, 1),$$

$$R \in R^{1,3}, r > 0.$$

Здесь r представляет эластичность шкалы признаковых значений.

Очевидно, что в модели заложена возможность расширения числа независимых производственных факторов (k,l), т.е. модель для Q может быть с $m>2$ числом факторных признаков $(k,l,\ldots)$.

5. *Условие существования для пары регрессионных коэффициентов определенного соотношения.*

С точки зрения практики весьма важным представляется случай линейных ограничений при установленной заранее величине соотношения между парами модельных параметров. Как показывают последующие рассуждения, задание таких соотношений означает в сущности введение обычных линейных гомогенных ограничений типа $Rb=0$:

$$\frac{b_k}{b_l}=a \Leftrightarrow b_k=ab_l \Leftrightarrow b_k-ab_l=0 \Leftrightarrow \{r:=0;\ R=(0\ldots\underset{\uparrow}{0}1\underset{k\text{-я}}{0}\ldots 0\underset{\uparrow}{-a}\underset{l\text{-я}}{0}\ldots 0)\}.$$

и позиции

6.4.4. Ограничения для модели с мультиколлинеарностью

Параметрические ограничения и нарушения ранга матрицы исходных данных. При введении детерминированных ограничений можно получить еще один класс моделей, представляющих значительный интерес с точки зрения их изучения. Таким моделям и посвящен данный раздел.

До сих пор решение задачи параметрического оценивания производилось при условии существования матрицы $(X'X)^{-1}$. Эта предпосылка в свою очередь базировалась на предположении о том, что матрица исходных данных X является полноранговой, т.е. $\operatorname{rg}(X)=m$. Ограничения, накладываемые на регрессию, приводили к сокращению регрессионной модели до $(m-q)$ параметров, которые могли бы быть оценены статистически. При этом само оценивание всегда принципиально оставалось возможным.

Для класса моделей, представленных в данном разделе, ограничения используются уже с целью обеспечить процесс оценивания, который изначально нереализуем. Роль ограничений меняется. Они становятся необходимыми, для того чтобы устранить соответственно конструктивным особенностям матрицы данных X нарушения в ее ранге. По терминологии эконометрики явление, которое мы будем исследовать, носит название *мультиколлинеарности.*

По определению матрица моментов $A:=X'X$ называется *мультиколлинеарной*, если для $X\in\mathbf{R}^{n,m}$, $n>m$, $\operatorname{rg}(A)=\operatorname{rg}(X'X)=\operatorname{rg}(X)<m$.

Иллюстрация: Регрессионная модель с одним уравнением, подверженная мультиколлинеарности из-за пропорциональности значений изучаемых факторных признаков ($m=2$).

Для случая с двумя объясняющими переменными ($m=2$), пусть имеется константа и одна переменная. Запишем исходную матрицу X:

$$X=\begin{pmatrix}X_{11}&X_{12}\\X_{21}&X_{22}\\\cdot&\cdot\\\cdot&\cdot\\\cdot&\cdot\\X_{n1}&X_{n2}\end{pmatrix}=\begin{pmatrix}1&X_{12}\\1&X_{22}\\\cdot&\cdot\\\cdot&\cdot\\\cdot&\cdot\\1&X_{n2}\end{pmatrix}=\begin{pmatrix}1&X_1\\1&X_2\\\cdot&\cdot\\\cdot&\cdot\\\cdot&\cdot\\1&X_n\end{pmatrix}=(e|x).$$

Мультиколлинеарность в данном случае означает, что $X_{i2}=p\forall i$ (то же и в случае, если имеем дело с двумя константами, когда, например, и вторая переменная — тоже константа). Это приводит к неидентифицируемости модели и действительно:

$$X'X=\begin{pmatrix}1&1&\ldots&1\\X_1&X_2&\ldots&X_n\end{pmatrix}\begin{pmatrix}1&X_1\\1&X_2\\\cdot&\cdot\\\cdot&\cdot\\\cdot&\cdot\\1&X_n\end{pmatrix}=n\begin{pmatrix}1&\bar X\\\bar X&\overline{XX}\end{pmatrix}=n\begin{pmatrix}1&p\\p&p^2\end{pmatrix},$$

$$|X'X|=n\begin{vmatrix}1&p\\p&p^2\end{vmatrix}=n\begin{vmatrix}1&0\\p&0\end{vmatrix}=n\begin{vmatrix}1&p\\0&0\end{vmatrix}=n\begin{vmatrix}1&0\\0&0\end{vmatrix}=0,$$

$$\left(\bar X=\sum_{i=1}^{n}\frac{1}{n}X_i,\ \overline{XX}=\sum_{i=1}^{n}\frac{1}{n}X_i^2,\ s:=\overline{XX}-\bar X^2=|X'X|\right).$$

Приведенное обоснование несколько расширяет наше представление о двухпризнаковой модели следующим аксиоматическим положением: если имеется $X_i\neq\bar X$ по меньшей мере

наблюдений, то объясняющая переменная в окрестности константы должна быть вариативна. Одновременно с помощью неравенства Шварца ($\Leftrightarrow s>0$) постулируется нормальность произведения $X'X$:

$$(X'X)^{-1}=\begin{pmatrix} n & n\overline{X} \\ n\overline{X} & n\overline{XX} \end{pmatrix}^{-1}=\frac{1}{ns}\begin{pmatrix} \overline{XX} & -\overline{X} \\ -\overline{X} & 1 \end{pmatrix}, s>0.$$

Чтобы преодолеть нарушения в статистическом оценивании модельных параметров регрессии с мультиколлинеарностью, необходимо использовать определенные ограничения. Последние вводятся одновременно с дополнением ранга матрицы исходных данных. Покажем это на примерах с количественными данными, которых для матрицы моментов вначале решается вопрос корректировки ранга, а затем на основе уже скорректированных данных — поиска допустимой оценки параметра b.

Пример 6.10. Ранговые дополнения.

а) $y=Xb+u, X=\begin{pmatrix} e & e & 0 \\ e & 0 & e \end{pmatrix}$,

$$X'=\begin{pmatrix} e' & e' \\ e' & 0' \\ 0' & e' \end{pmatrix}=\begin{pmatrix} 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 \end{pmatrix}, \operatorname{rg}(X)=2,$$

$Rb=r, R=(0\ \ 1\ \ 1), r=0$.

$$X'X+R'R=\begin{pmatrix} 2n & n & n \\ n & n+1 & 1 \\ n & 1 & n+1 \end{pmatrix}, \operatorname{rg}(X'X+R'R)=3,$$

$$X'X=n\begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \operatorname{rg}(X'X)=2,\ R'R=\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \operatorname{rg}(R'R)=1,$$

$$M:=(X'X+R'R)^{-1}=\frac{1}{4n}\begin{pmatrix} n+2 & -n & -n \\ -n & n+2 & n-2 \\ -n & n-2 & n+2 \end{pmatrix}.$$

б) Пусть имеется

$$y=Xb+u,\ X=\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \operatorname{rg}(X)=2.$$

Первый из альтернативных подходов к устранению дефектов в ранге матрицы предполагает введение гомогенных ограничений, например

$$r=Rb,\ R=\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},\ r=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}=0 \Leftrightarrow b_1=0 \wedge b_2=0 \wedge b_3=-b_4.$$

Отсюда следует: $y_i=(X_{i4}-X_{i3})b_4+u_i \therefore y_i=u_i (i=1,2,\ldots,n)$. Здесь b_3 и b_4 — по-прежнему любые величины, в так называемом чистом линейном стохастическом процессе они не играют никакой роли и вообще не подлежат оцениванию.

Гомогенные ограничения имеют широкие возможности для решения разнообразных задач регрессионного анализа, и это наглядно демонстрирует второй, альтернативный, подход их приложения:

$$r=Rb,\ R=\begin{pmatrix} 1 & 1 & 2 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & a & 0 \end{pmatrix},\ a\neq 1 \wedge a\neq 2,\ r=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}=0 \Leftrightarrow$$

$$\Leftrightarrow b_1=0 \wedge b_2=-ab_3 \wedge b_4=(a-2)b_3.$$

В дальнейшем решается тривиальная задача по оценке параметров простой модели с одним уравнением при условии, что анализируются данные четырех наблюдений: исходная матрица R — полноранговая и существует только один модельный параметр:

$$y_i=(-aX_{i2}+X_{i3}+(a-2)X_{i4})b_3+u_i \Leftrightarrow y_i=Z_i b_3+u_i \quad (i=1,2,\ldots,n),$$

$$Z_i:=-aX_{i2}+X_{i3}+(a-2)X_{i4}.$$

Третьим альтернативным подходом к устранению ранговых нарушений в регрессии является введение негомогенных огра-

ничений. Например, пусть матрица R останется неизменной, но $r \neq 0$:

$$r = Rb, \quad R = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \neq 0.$$

После введения дополнительных условий: о числе анализируемых наблюдений (четыре), полноранговости R и существовании единственного параметра устанавливаются значения всех трех коэффициентов регрессии. Оцениваемая модель при этом представляется трансформированной в простейшую, с одним уравнением:

$$(y_i - X_{i3}) = (X_{i4} - X_{i3})b_4 + u_i \Leftrightarrow z_i = Z_i\alpha + u_i \quad (i = 1, 2, \ldots, n),$$

$z_i := y_i - X_{i3}$, $Z_i := X_{i4} - X_{i3}$, $\alpha := b_4$.

Таким образом, нарушения в рангах матрицы устранимы с введением некоторого числа q ограничений. Тем самым реализуется возможность перехода от необращаемой матрицы $X'X$ к инверсионной сумме матриц $X'X + R'R$. Метод наименьших квадратов, который в общем случае обеспечивает поиск m параметров регрессионного уравнения, при этом используется в сокращенном виде, для нахождения $m-q$ параметров (по числу столбцов с нарушениями ранга, устраняемых при помощи дополнительных условий). В последующем мы увидим также, что решения задач «с нарушениями ранга матрицы исходных данных» дают основания делать теоретические обобщения для метода ранговых дополнений.

Обобщения для метода ранговых дополнений. ***Определение.*** *Матрица* $M = (X'X + R'R)^{-1}$ *является неединичной и представляет своего рода обобщение для алгебраической операции обращения мультипликативной матрицы вида* $X'X$.

□ *Доказательство*: справедливость приведенного высказывания подтверждается результатом произведения: $(X'X)M(X'X) = X'X$.

Дополнением рангов добиваются оценок, которые соответствуют оценкам, полученным при помощи метода моментов; так, после предварительного умножения модели $y = Xb + u$ вначале на X', а затем на $(X'X + R'R)^{-1}$ получают

$$\begin{aligned} & y = Xb + u \Rightarrow X'y && = X'Xb + X'u \Rightarrow \\ \Rightarrow\ & (X'X + R'R)^{-1}X'y && = (X'X + R'R)^{-1}X'Xb + (X'X + R'R)^{-1}X'u \Rightarrow \\ \Rightarrow\ & X'X(X'X + R'R)^{-1}X'y && = X'X(X'X + R'R)^{-1}X'Xb + X'X(X'X + R'R)^{-1}X'u = \\ & && = X'Xb + X'X(X'X + R'R)^{-1}X'u. \end{aligned}$$

∴ В то же время при использовании метода моментов $X'X(X'X + R'R)^{-1}X'y = X'Xb \Leftrightarrow X'X\left[(X'X + R'R)^{-1}X'y - b\right] = 0$, причем $\hat{b} = X'X(X'X + R'R)^{-1}X'y$ является возможно даже несмещенной оценкой. ■

Метод ранговых дополнений с обобщениями для случая, когда $\operatorname{rg}(X) < m$. Запишем для регрессионной модели аналитическую систему в целом:

$$\tilde{y} = \begin{pmatrix} y \\ r \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} X \\ R \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ 0 \end{pmatrix} \Leftrightarrow \tilde{y} = \tilde{X}b + u.$$

Используя метод моментов в оценивании модели, умножим ее левую и правую части на $\tilde{X}'$, получим:

$$\tilde{X}'\tilde{y} = \tilde{X}'\tilde{X}b + \tilde{X}'u.$$

При решении последнего уравнения относительно параметра b случайная величина u изначально не определена и может быть отброшена как аддитивный стохастический линейный элемент модели, тогда получим:

$$\tilde{X}'\tilde{y} = \tilde{X}'\tilde{X}b. \qquad (6.25)$$

Если матрица X имеет нарушения ранга, то параметр b для модели $y = Xb + u$ в принципе не может быть оценен. Однако, если вводятся ранговые дополнения, то на основании того же уравнения (6.25) найдем:

$$\hat{b} = (X'X + R'R)^{-1}(X'y + R'r).$$

Матрица, обратная $X'X + R'R$, в отличие от $X'X$ существует, а значит нарушения ранга исходной матрицы с введением ограничений устраняются.

Ридж-оценивание. При помощи рассуждений, аналогичных приведенным выше, можно также сделать простейшее обоснование для так называемого ридж-оценивания.

Пусть, как и раньше, $X'X$ неинверсионна из-за нарушения ранга X. Введем корректирующий элемент αI. Скорректирован-

ная матрица $X'X+\alpha I$ будет всегда обращаемой и даст возможность получить параметрическую оценку вида:

$$\tilde{b}=(X'X+\alpha I)^{-1}X'y.$$

GLS-оценка смешанной модели

$$\dot{y}=\dot{X}b+u,\ E(u)=0,\ \mathrm{cov}(u)=\sigma^2V,$$

где $\dot{y}:=\begin{pmatrix}y\\r\end{pmatrix}$, $\dot{X}:=\begin{pmatrix}X\\R\end{pmatrix}=\begin{pmatrix}X_1 & X_2\\R_1 & R_2\end{pmatrix}$, $u:=\begin{pmatrix}u_1\\u_2\end{pmatrix}$, $V:=\begin{pmatrix}V_1 & 0\\0 & V_2\end{pmatrix}=\mathrm{diag}(V_s)$,

$s=1,2$.

В соответствии с подходом Айткина имеем:

$$\tilde{b}=\left(\dot{X}'V^{-1}\dot{X}\right)^{-1}\dot{X}'V^{-1}\dot{y}.$$

Пример 6.11. Пусть имеется любая гомоскедастичная модель. Тогда в особом случае, при $V_s=\sigma_s^2 I\ (s=1,2)$, будет:

$$V^{-1}=\begin{pmatrix}I_n/\sigma_1^2 & 0\\0 & I_q/\sigma_2^2\end{pmatrix}\ \text{и}\quad \tilde{b}=\left(\frac{X'X}{\sigma_1^2}+\frac{R'R}{\sigma_2^2}\right)^{-1}\left(\frac{X'y}{\sigma_1^2}+\frac{R'r}{\sigma_2^2}\right).$$

6.4.5. Регрессионная модель ANOVA

ANOVA (analysis of variance model) — регрессионный анализ дисперсионной модели. В этом анализе матрица исходных данных X состоит исключительно из нулей или единиц. Все модели типа ANOVA подразделяются на два класса: с одинаково большим и неодинаково большим числом наблюдений [103, с. 61—120; 112, с. 27—28; 113, с. 197). Исходная модель имеет общий вид:

$$y_{ijt}=\mu+a_i+b_j+c_{ij}+u_{ijt}\quad \begin{pmatrix}i=1,2,...,G;\ j=1,2,...,T_i,\\ \sum_{i=1}^{G}T_i=\tau;\ t=1,2,..,n_{ij}\end{pmatrix}.$$

Параметрический вектор включает компоненты:

$$\beta=\begin{pmatrix}\mu\\a\\b\\c\end{pmatrix},\ \beta\in R^{1+G+T+GT},\ K=1+G+T+GT.$$

Для того чтобы в дальнейшем упростить представление матриц, примем часто используемое в теории равенство $T_i:=T$. Символами i и j обозначены значения двух факторных признаков, которые либо обнаруживают свое воздействие на отклик (y_{ijt}), либо нет. Обозначение t принято для представления определенного класса наблюдений (выборочного элемента) типа $\{i,j\}$, всего имеется неравных множеств наблюдений каждого типа $n_{ij}\geq 1$. Оцениванию в модели подлежат параметры: μ — общая средняя (или общий эффект — grand mean); a_i и b_j — средние значения специфических признаков (или факторных эффектов) информационного блока i или j; c_{ij} — общая величина (общий или интегрированный эффект, индивидуальная константа).

Пусть имеем определенный собственный вектор:

$$e_s'=(111...1),\ e_s\in R^s,\ s\in\{G,T,GT\},$$

в этом случае можно рассматривать три эквивалентных представления ANOVA-модели:

- *скалярный тип*:
  $y_{ij}=\mu+a_i+b_j+c_{ij}+u_{ij}\quad \left(i=1,2,...,G;\ j=1,2,...,T\right)$;
- *тип модели для уравнения (признака) относительно i:*
  $y_{i\cdot}=\mu e_T+a_i e_T+b+c_{i\cdot}+u_{i\cdot}$;
- *тип модели для уравнения (признака) относительно j:*
  $y_{\cdot j}=\mu e_G+a+b_j e_G+c_{\cdot j}+u_{\cdot j}$.

В перечисленных моделях приняты обозначения:

$$\mu e=\begin{pmatrix}\mu\\\mu\\\cdot\\\cdot\\\cdot\\\mu\end{pmatrix}\in R^{GT},\ y_{i\cdot}=\begin{pmatrix}y_{i1}\\y_{i2}\\\cdot\\\cdot\\\cdot\\y_{iT}\end{pmatrix}\in R^{T},\ y_{\cdot j}=\begin{pmatrix}y_{1j}\\y_{2j}\\\cdot\\\cdot\\\cdot\\y_{Gj}\end{pmatrix}\in R^{G},\ y=\begin{pmatrix}y_1\\y_2\\\cdot\\\cdot\\\cdot\\y_G\end{pmatrix}\in R^{GT},$$

$$u_{i\cdot}=\begin{pmatrix}u_{i1}\\u_{i2}\\\cdot\\\cdot\\\cdot\\u_{iT}\end{pmatrix}\in R^{T},\quad u_{\cdot j}=\begin{pmatrix}u_{1j}\\u_{2j}\\\cdot\\\cdot\\\cdot\\u_{Gj}\end{pmatrix}\in R^{G},\ u=\begin{pmatrix}u_1\\u_2\\\cdot\\\cdot\\\cdot\\u_G\end{pmatrix}\in R^{GT},\ a=\begin{pmatrix}a_1\\a_2\\\cdot\\\cdot\\\cdot\\a_G\end{pmatrix}\in R^{G},$$

$$b=\begin{pmatrix}b_1\\b_2\\\cdot\\\cdot\\\cdot\\b_T\end{pmatrix}\in R^T,\ c_{i\cdot}=\begin{pmatrix}c_{i1}\\c_{i2}\\\cdot\\\cdot\\\cdot\\c_{iT}\end{pmatrix}\in R^T,\ c_{\cdot j}=\begin{pmatrix}c_{1j}\\c_{2j}\\\cdot\\\cdot\\\cdot\\c_{Gj}\end{pmatrix}\in R^G,\ c=\begin{pmatrix}c_1\\c_2\\\cdot\\\cdot\\\cdot\\c_G\end{pmatrix}\in R^{GT}.$$

Очевидно, что модель при $K>GT$ является перепараметризованной с целью дальнейшей ее идентификации. Подробно такая модель записывается следующим образом:

$$\begin{pmatrix}y_{11}\\y_{12}\\\vdots\\y_{1T}\\\vdots\\y_{G1}\\y_{G2}\\\vdots\\y_{GT}\end{pmatrix}=\begin{pmatrix}\mu\\\mu\\\cdot\\\cdot\\\cdot\\\mu\end{pmatrix}+\begin{pmatrix}a_1&0&\dots&0\\a_1&0&\dots&0\\&&\vdots&\\a_1&0&\dots&0\\&&\vdots&\\0&0&\dots&a_G\\0&0&\dots&a_G\\&&\vdots&\\0&0&\dots&a_G\end{pmatrix}\begin{pmatrix}1\\1\\\cdot\\\cdot\\\cdot\\1\end{pmatrix}+\begin{pmatrix}b_1&0&\dots&0\\0&b_2&\dots&0\\&&\vdots&\\0&0&\dots&b_T\\&&\vdots&\\b_1&0&\dots&0\\0&b_2&\dots&0\\&&\vdots&\\0&0&\dots&b_T\end{pmatrix}\begin{pmatrix}1\\1\\\cdot\\\cdot\\\cdot\\1\end{pmatrix}+\begin{pmatrix}c_{11}\\c_{12}\\\vdots\\c_{1T}\\\vdots\\c_{G1}\\c_{G2}\\\vdots\\c_{GT}\end{pmatrix}+\begin{pmatrix}u_{11}\\u_{12}\\\vdots\\u_{1T}\\\vdots\\u_{G1}\\u_{G2}\\\vdots\\u_{GT}\end{pmatrix}.$$

Приведенная модель является обычной, линейной. Здесь мы не будем говорить собственно об анализе факторов и его возможностях. Теоретически более важным представляется вопрос о многообразии линейных форм регрессионной модели ANOVA. С этой точки зрения рассмотрим более подробно некоторые особенные случаи, демонстрирующие возможности построения практически необозримого числа специализированных моделей.

О необходимости использования ограничений для модели ANOVA. Естественно полагать, что отдельные коэффициенты регрессии в модели могут присутствовать, оказаться незначимыми, принимать нулевые значения (выше нами рассматривались нуль-ограничения для параметра b). Соответственно с учетом числа комбинаций значимых коэффициентов в модели уже можно предположить достаточно большое число особых случаев для

модели ANOVA. В то же время сопровождающие различные модельные спецификации нарушения рангов изначально делают невозможным параметрическое оценивание. Как известно, выход заключается при этом во введении дополнительных ограничений. Наиболее часто встречаются четыре следующих ограничения (нуль-ограничения для сумм с симметрическим расположением):

$$\sum_i a_i=\sum_t b_t=\sum_i c_{it}=\sum_t c_{it}=0.$$

ANOVA — **модель простой классификации.**

Иллюстрация 1 (one way layout). Рассмотрим модель с дополнением рангов при условии $b\equiv 0,\ c\equiv 0$. ANOVA — модель простой классификации — представляет такой класс моделей, в котором регрессионные уравнения содержат только один признак:

$$y_{ij}=\mu+a_i+u_{ij}\ (i=1,2,\dots,G;\quad j=1,2,\dots,n_i);$$

$$y=e_n\mu+\begin{pmatrix}e_{n_1}&0&\dots&0\\0&e_{n_2}&&0\\\cdot&&\cdot&\cdot\\\cdot&&\cdot&\cdot\\\cdot&&\cdot&\cdot\\0&0&\dots&e_{n_G}\end{pmatrix}a+u;\quad n=\sum_{i=1}^{G}n_i;\quad e_s=\begin{pmatrix}1\\1\\\cdot\\\cdot\\\cdot\\1\end{pmatrix}\in R^s\ (s=n_1,n_2,\dots,n_G,n);$$

$$y=X\beta+u,\ X=(X_1X_2)=(e_nX_2)\in R^{n,1+G},\ \mathrm{rg}(X)=G<n,\ b=\begin{pmatrix}\mu\\a\end{pmatrix}\in R^{1+G}.$$

При устранении ранговых нарушений различие в значениях n_i не играет никакой роли, так как это происходит при условии $n_i\equiv 1\forall i$, или $n_i\geq 1,\ i=1,2,\dots,G$:

$$\sum_{i=1}^{G}a_i=0\Leftrightarrow e'a=0\Leftrightarrow R\beta=r,\ e'=(111\dots1)\in R^{1,G},\ R:=(0e')\in R^{1,G+1},\ r:=0,$$

или

$$\sum_{i=1}^{G}n_ia_i=0\Leftrightarrow n'a=0\Leftrightarrow R\beta=r,\ R:=(0n')=(0n_1n_2..n_G)\in R^{1,1+G},\ r:=0.$$

(По данной проблеме составлены практические задания № 14, 15, 21).

Оценивание в модели ANOVA средних величин.

Иллюстрация 2. Рассмотрим вопросы о ранговых дополнениях для модели ANOVA при условии, что $\mu\equiv 0$, a или $b\equiv 0,\ c\equiv 0$.

Оценивание средней по значениям признака первого фактора i $(\mu \equiv 0,\ b \equiv 0,\ c \equiv 0)$:

$$K = G < GT \ (i = 1,2,...,G;\quad t = 1,2,...,T);$$

$y_{it} = a_i + u_{it} \Leftrightarrow y_t = a_i e_T + u_i \Leftrightarrow y_t = \mu e_T + u_i$; т.е. $\mu = a_i$.

Оценивание средней по значениям признака второго ра j $(\mu \equiv 0,\ a \equiv 0,\ c \equiv 0)$:

$$K = T < GT (i = 1, 2, ..., G; t = 1, 2, ..., T);$$

$y_{it} = b_t + u_{it} \Leftrightarrow y_t = b_t e_G + u_t \Leftrightarrow y_t = \mu e_G + u_t$, т.е. здесь $\mu = b_t$.

Оба приведенных случая не идентичны, в то же время можно рассматривать их связь и идентичность с моделями, отображающими динамические процессы.

Иллюстрация 3. **Динамические процессы и их моделирование.** Ранговые дополнения для ANOVA-модели при условии $b \equiv 0$ или $a \equiv 0,\ c \equiv 0$.

Модель эквивалентной корреляции с одной константой

В частном случае, когда присутствуют только одна экзогенная величина и константа, общая модель эквивалентной корреляции с одним уравнением сокращается и переходит в модель типа "оценивания средней":

Пусть в матрице исходных данных X t-я строка, или вектор X_t, сжимается до единственного значения, равного единице, т.е. $X_t = 1$, тогда:

$$y = X\beta + u,\ y \in R^T,\ X = e \in R^T,\ \beta \in R;$$

$$E(u) = 0;\ E(uu') = \operatorname{cov}(u) = V = \sigma^2[(1-a)I + aee'],\ 0 < a < 1;$$

$$y_t = X_t\beta + u_t \ (t = 1,2,...,T) \Leftrightarrow y_t = \mu + u_t \Leftrightarrow y = \mu e + u \text{ с}$$

$$E(u_t) = 0 \neq E(y_t) = \mu;\ E(u_t u_s) = E(y_t y_s) = \begin{cases} \sigma^2, & t = s, \\ a\sigma^2, & t \neq s. \end{cases}$$

Подобного типа модели возникают при отображении специфических стохастических процессов, показанных ниже.

- *Модель для временных рядов с двумя ошибками.*

При условии

$$\{ y \in R^T,\ X = e \in R^T,\ \mu \in R,\ E(u) = 0,\ V = \sigma^2[(1-a)I + aee'],\ 0 < a < 1 \}$$

сама модель принимает вид $y_{ts} = \mu + u_t + v_s$,

при этом $E(u_t) = 0$, $\operatorname{var}(u_t) = E(u_t u_t) = \sigma^2$, $E(u_t u_s) = 0$ для $t \neq s$;

$E(v_s) = 0$, $\operatorname{var}(v_s) = E(v_s v_s) = \rho^2$, $E(v_t v_s) = 0$ для $t \neq s$;

$E(u_t v_s) = 0$ (остатки модели представляют собой независимые величины, не коррелирующие между собой, т.е. u и v $\forall t,s$).

Здесь случайная величина ставится в зависимость и определяется двумя другими случайными величинами. Первая из них u — зависима от наблюдений, другая величина v — в общем независима от наблюдений и определяется самим анализируемым признаком:

$$E(y_t) = \mu,\ E(y_t - \mu)(y_s - \mu) = \begin{cases} \sigma^2 + \rho^2, & t = s, \\ \rho^2, & t \neq s \end{cases}$$

в матричной форме:

$$E(y_t - \mu)(y_s - \mu) = \begin{pmatrix} \sigma^2 + \rho^2 & \rho^2 & \dots & \rho^2 \\ \rho^2 & \sigma^2 + \rho^2 & \dots & \rho^2 \\ \vdots & \vdots & \vdots & \vdots \\ \rho^2 & \rho^2 & \dots & \sigma^2 + \rho^2 \end{pmatrix} = \sigma^2 I_n + \rho^2 E_n.$$

- *Модель факторного анализа с оцениванием средней.*

Конструктивная схема построения данной модели базируется также на модели с дисперсионными компонентами, в сущности это всего лишь другая форма записи последней:

$y_{ts} = \mu + a_t + w_s$,

$E(w_s) = 0$, $E(a_t) = 0$, $\operatorname{var}(w_s) = \sigma^2$, $\operatorname{var}(a_t) = \rho^2$, $E(a_t w_s) = 0 \forall t,s$,

где t — индекс анализируемого признака,

s — индекс для s-го наблюдения по t-признаку,

μ — неизвестная общая средняя,

α_t — стохастическая величина отклонения от общей средней,

w_s — стохастическая спонтанная вариативность.

6.4.6. Значение подхода Гаусса — Маркова в оценивании параметров регрессионной модели

С целью показать значение подхода Гаусса — Маркова остановимся подробнее на ряде примеров с количественными данными. В первом производятся оценивание данных по Айткину и сценарное моделирование параметрических оценок.

Пример 6.12. Гетероскедастичная модель. Пусть $m=1$, $n=2$, а случайная величина u задается двумерным дискретным распределением со следующими численными характеристиками:

u_2 \ u_1	-1	1
-2	0,25	0,25
2	0,25	0,25

При помощи двух первых моментов найдем среднюю и дисперсию для модельных остатков:

$$E(u_1)=E(u_2)=0,\ E(u)=0,\ u=\begin{pmatrix}u_1\\u_2\end{pmatrix},$$

$$E(uu')=E\begin{pmatrix}u_1u_1 & u_1u_2\\ u_2u_1 & u_2u_2\end{pmatrix}=\begin{pmatrix}E(u_1u_1) & E(u_1u_2)\\ E(u_2u_1) & E(u_2u_2)\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & 4\end{pmatrix},$$

это гетероскедастичная модель с уравнениями:

$y_1=X_1b+u_1$,

$y_2=X_2b+u_2$.

При условии, что двумерная экзогенная величина $\{X_1X_2\}$, как и истинный параметр b, известна, скажем $X_1=1$, $X_2=2$ и $b=2$ для каждой комбинации $\{u_1u_2\}$, можно рассматривать некоторую соответствующую комбинацию $\{Y_1Y_2\}$. Сведем в таблице четыре различные комбинации входных и результативных показателей:

$b=2$	$u_1=-1, u_2=-2$	$u_1=-1, u_2=2$	$u_1=1, u_2=-2$	$u_1=1, u_2=2$
$X_1=1, X_2=2$	$Y_1=1, Y_2=2$	$Y_1=1, Y_2=6$	$Y_1=3, Y_2=2$	$Y_1=3, Y_2=6$
$\sum_{t=1}^{T}X_tY_t$	5	13	7	15
$\sum_{t=1}^{T}X_t^2$	5	5	5	5
$\hat{b}$	$\frac{5}{5}=1$	$\frac{13}{5}=2,6$	$\frac{7}{5}=1,4$	$\frac{15}{5}=3$

Приведенные в таблице параметрические оценки вычислены по известной формуле:

$$\hat{b}=\frac{\sum X_iY_i}{\sum X_i^2}.$$

Уровни вероятности появления каждой из параметрических оценок $(\hat{b})$ остаются равными и принимают значение 0,25:

$\hat{b}$	1,0	1,4	2,6	3
Вероятность появления параметрической оценки	0,25	0,25	0,25	0,25

Таким образом можем определить математическое ожидание для $\hat{b}$ ($E(\hat{b})=b$):

$E(\hat{b})=0,25\cdot1+0,25\cdot1,4+0,25\cdot2,6+0,25\cdot3=2=b$;

$\mathrm{var}(\hat{b})=0,25\cdot(1-2)^2+0,25\cdot(1,4-2)^2+0,25\cdot(2,6-2)^2+0,25\cdot(3-2)^2=0,68$.

Пример 6.13. Вернемся к рассмотрению гомоскедастичной модели.

Чтобы наглядно показать свойства GLS-оценки, обратимся к трансформации по Айткину и осуществим ее для второго уравнения следующей модели:

$Y_1=X_1b+u_1,\ E(u_1)=0,\ \mathrm{var}(u_1)=\sigma_1^2=1$;

$Y_2=X_2b+u_2,\ E(u_2)=0,\ \mathrm{var}(u_2)=\sigma_2^2=4$;

умножим левую и правую части второго уравнения на $\frac{1}{2}=\frac{1}{\sigma_2}$ и получим

$$\frac{1}{2}Y_2=\frac{1}{2}X_2b+\frac{1}{2}u_2,\ E\left(\frac{1}{2}u_2\right)=0,\ \mathrm{var}\left(\frac{1}{2}u_2\right)=\left(\frac{1}{2}\right)^2\sigma_2^2=1.$$

Построим таблицу с гипотетическими данными гомоскедастичной модели:

u_2 \ u_1	-1	1
-1	0,25	0,25
1	0,25	0,25

$$E(u_1)=E(u_2)=0,\ E(u)=0,\ u=\begin{pmatrix}u_1\\u_2\end{pmatrix},\ E(uu')=E\begin{pmatrix}u_1u_1 & u_1u_2\\u_2u_1 & u_2u_2\end{pmatrix}=$$

$$=\begin{pmatrix}E(u_1u_1) & E(u_1u_2)\\E(u_2u_1) & E(u_2u_2)\end{pmatrix}=I_2.$$

Соответствующим образом, как и в предыдущем примере, получим плотность распределения параметрических оценок b:

$\tilde{b}$	1,0	2	3
Вероятность появления параметрической оценки	0,25		0,25

Теперь можем с учетом плотности рассеяния параметрических оценок вычислить математическое ожидание и дисперсию параметра b ($E(\tilde{b})=E(\hat{b})=b$ и $\operatorname{var}\tilde{b}$):

$E(\tilde{b})=0{,}25\cdot1+0{,}50\cdot2+0{,}25\cdot3=2=b$;

$\operatorname{var}(\tilde{b})=0{,}25\cdot(1-2)^2+0{,}50\cdot(2-2)^2+0{,}25\cdot(3-2)^2=0{,}50$.

Полученные оценки являются эффективными по Айткину. Сходный результат для рассматриваемой модели может быть достигнут при авторегрессионном оценивании.

Пример 6.14. Авторегрессионная модель.

Будем исходить из следующей плотности распределения случайных величин:

u_2 \ u_1	−1	1
−1	0,125	0,375
1	0,375	0,125

Математическое ожидание модельных остатков, как обычно, полагается равным нулю:

$$\hat{E}(u_1)=E(u_2)=0,\ \ E(u)=0,\ u=\begin{pmatrix}u_1\\u_2\end{pmatrix},$$

$$E(uu')=E\begin{pmatrix}u_1u_1 & u_1u_2\\u_2u_1 & u_2u_2\end{pmatrix}=\begin{pmatrix}E(u_1u_1) & E(u_1u_2)\\E(u_2u_1) & E(u_2u_2)\end{pmatrix}=\begin{pmatrix}\rho^0 & \rho^1\\\rho^1 & \rho^0\end{pmatrix}=$$

$$=\begin{pmatrix}1 & \rho\\\rho & 1\end{pmatrix}=\begin{pmatrix}1 & -0{,}5\\-0{,}5 & 1\end{pmatrix},\ \rho=-0{,}5.$$

После умножения левой и правой частей модели на P получим:

$$P=\begin{pmatrix}1 & 0\\\dfrac{-\rho}{\sqrt{1-\rho^2}} & \dfrac{1}{\sqrt{1-\rho^2}}\end{pmatrix}=\begin{pmatrix}1 & 0\\\dfrac{1}{\sqrt{3}} & \dfrac{2}{\sqrt{3}}\end{pmatrix}=\begin{pmatrix}1 & 0\\0{,}57735 & 1{,}1547\end{pmatrix}.$$

Имеем: $PY=PXb+Pu\Leftrightarrow Y^*=X^*b+v$ с $E(v_1)=E(v_2)=0$ и

$$E(v)=0,\ v=\begin{pmatrix}v_1\\v_2\end{pmatrix},\ \begin{matrix}v_1=u_1\\v_2=\dfrac{-\rho u_1+u_2}{\sqrt{1-\rho^2}}\end{matrix},\ E(vv')=E\begin{pmatrix}v_1v_1 & v_1v_2\\v_2v_1 & v_2v_2\end{pmatrix}=PE(uu')P'=I_2.$$

В результате получим трансформированные данные:

$b=2$	$u_1=-1,\ u_2=-1$	$u_1=-1,\ u_2=1$	$u_1=1,\ u_2=-1$	$u_1=1,\ u_2=1$
	$v_1=-1$ $v_2=-1{,}73205$	$v_1=-1$ $v_2=0{,}57735$	$v_1=1$ $v_2=-0{,}57735$	$v_1=1$ $v_2=1{,}73205$
$X_1=1,\ X_2=2$	$Y_1=1,\ Y_2=3$	$Y_1=1,\ Y_2=5$	$Y_1=3,\ Y_2=3$	$Y_1=3,\ Y_2=5$
$X_1=1,$ $X_2=2{,}88675$	$Y_1=1,$ $Y_2=4{,}04145$	$Y_1=1,$ $Y_2=6{,}35085$	$Y_1=3,$ $Y_2=5{,}19615$	$Y_1=3,$ $Y_2=7{,}50555$
$\sum_{t=1}^{T}X_tY_t$	$7\rightarrow$ $12{,}\overline{6}$	$11\rightarrow$ $19{,}\overline{3}$	$9\rightarrow$ $17{,}99999$	$13\rightarrow$ $24{,}\overline{6}$
$\sum_{t=1}^{T}X_t^2$	$5\rightarrow$ $9{,}\overline{3}$	$5\rightarrow$ $9{,}\overline{3}$	$5\rightarrow$ $9{,}\overline{3}$	$5\rightarrow$ $9{,}\overline{3}$
$\hat{b}$	$\frac{7}{5}=1{,}4$	$\frac{11}{5}=2{,}2$	$\frac{9}{5}=1{,}8$	$\frac{13}{5}=2{,}6$
b	$\frac{12{,}66}{9{,}33}=1{,}35714$	$\frac{19{,}33}{9{,}33}=2{,}07143$	$\frac{17{,}99}{9{,}33}=1{,}92857$	$\frac{24{,}66}{9{,}33}=2{,}64286$

В таблице переходы $a\rightarrow b$ означают трансформационные изменения данных.

Произведем сортировку вероятностных характеристик появления параметрических оценок, полученных до и после трансформации данных:

$\hat{b}$	1,4	1,8	2,2	2,6
Вероятность появления параметрической оценки до трансформации модели	0,125	0,375	0,375	0,125
$\tilde{b}$	1,375	1,928	2,072	2,643
Вероятность появления параметрической оценки после трансформации модели	0,125	0,375	0,375	0,125

Обе оценки $\hat{b}$ и $\tilde{b}$ — несмещенные. Дисперсионные показатели указывают на повышение эффективности параметрической оценки после трансформации регрессионной модели. Действительно,

$$\operatorname{var}(\hat{b}) = 0{,}125\cdot(1{,}4-2)^2 + 0{,}375\cdot(1{,}8-2)^2 + 0{,}375\cdot(2{,}2-2)^2 + 0{,}125\cdot(2{,}6-2)^2 = 0{,}12$$

против

$$\operatorname{var}(\tilde{b}) = 0{,}125\cdot(1{,}375-2)^2 + 0{,}375\cdot(1{,}928-2)^2 + 0{,}375\cdot(2{,}072-2)^2 + 0{,}125\cdot(2{,}643-2)^2 = 0{,}107248.$$

Разложение ковариационной матрицы. В ходе оценивания по Айткину обобщаются отдельные показатели ковариации модельных остатков. Рассмотрим на примере различных моделей, каким образом решается обратная задача, а именно разложение матрицы ковариации остатков. Результаты показанных решений могут затем использоваться в трансформационных преобразованиях моделей:

- *Гетероскедастичная модель*

$$V = \begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{nn} \end{pmatrix},\quad V^{-1} = \begin{pmatrix} 1/\sigma_{11} & 0 & \ldots & 0 \\ 0 & 1/\sigma_{22} & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/\sigma_{nn} \end{pmatrix} = P'P,$$

$$P = \begin{pmatrix} 1/\sqrt{\sigma_{11}} & 0 & \ldots & 0 \\ 0 & 1/\sqrt{\sigma_{22}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/\sqrt{\sigma_{nn}} \end{pmatrix}.$$

- *Модель с эквивалентной корреляцией остатков:*

$$V = \begin{pmatrix} 1 & a & a & \ldots & a \\ a & 1 & a & \ldots & a \\ a & a & 1 & \ldots & a \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & a & a & \ldots & 1 \end{pmatrix} = (1-a)I_n + aE_n,\ a \in \boldsymbol{R},\ e = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},\ e \in \boldsymbol{R}^n,$$

$$ee' =: E_n \in \boldsymbol{R}^{n,n},$$

$$V^{-1} = \alpha I_n + \beta E_n,\ \alpha = \frac{1}{1-a},\ \beta = \frac{1}{a-1}\frac{a}{1+a(n-1)},$$

$$V^{-1} = P'P,\ P = \delta_1 I_n - \delta_2 E_n,\ \delta_1 = \sqrt{\alpha},\ \delta_2 = \frac{1}{n}\left(\sqrt{\alpha} \pm \sqrt{\alpha + n\beta}\right).$$

- *Авторегрессионная модель*

$$V = \begin{pmatrix} 1 & b & b^2 & \ldots & b^{n-1} \\ b & 1 & b & \ldots & b^{n-2} \\ \cdot & \cdot & \cdot & \vdots & \cdot \\ b^{n-1} & b^{n-2} & b^{n-3} & \ldots & 1 \end{pmatrix},\ V^{-1} = \tau\ \left(1, 1+b^2, 1; -b, n\right),$$

$$V^{-1} = \frac{1}{1-b^2}\begin{pmatrix} 1 & -b & 0 & \ldots & 0 & 0 & 0 \\ -b & 1+b^2 & -b & \ldots & 0 & 0 & 0 \\ 0 & -b & 1+b^2 & \ldots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \ddots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1+b^2 & -b & 0 \\ 0 & 0 & 0 & \ldots & -b & 1+b^2 & -b \\ 0 & 0 & 0 & \ldots & 0 & -b & 1 \end{pmatrix} = \frac{1}{1-b^2}P'P,$$

$$P = \begin{pmatrix} \sqrt{1-b^2} & 0 & 0 & \ldots & 0 & 0 & 0 \\ -b & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -b & 1 & \ddots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \ddots & 0 & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 & 0 & 0 \\ 0 & 0 & 0 & \ldots & -b & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & -b & 1 \end{pmatrix}.$$

- *Модель Койка* ($a \neq \pm 1$)

$$V = \begin{pmatrix} 1+a^2 & -a & 0 & \cdots & 0 & 0 \\ -a & 1+a^2 & -a & \cdots & 0 & 0 \\ 0 & -a & 1+a^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+a^2 & -a \\ 0 & 0 & 0 & \cdots & -a & 1+a^2 \end{pmatrix},$$

$$\Delta\Delta' = \begin{pmatrix} -a & 1 & 0 & 0 \\ 0 & -a & 1 & 0 \\ 0 & 0 & -a & 1 \end{pmatrix} \begin{pmatrix} -a & 0 & 0 \\ 1 & -a & 0 \\ 0 & 1 & -a \\ 0 & 0 & 1 \end{pmatrix} =: W(a) = \begin{pmatrix} 1+a^2 & -a & 0 \\ -a & 1+a^2 & -a \\ 0 & -a & 1+a^2 \end{pmatrix} =$$

$$= \tau\left(1+a^2, 1+a^2, 1+a^2; -a\right),$$

$$V^{-1} = A^{-1} = \frac{1}{C_n} \begin{pmatrix} C_0a^0C_{n-1} & C_0a^1C_{n-2} & C_0a^2C_{n-3} & \cdots & C_0a^{n-1}C_0 \\ C_0a^1C_{n-2} & C_1a^0C_{n-2} & C_1a^1C_{n-3} & \cdots & C_1a^{n-2}C_0 \\ C_0a^2C_{n-3} & C_1a^1C_{n-3} & C_2a^0C_{n-3} & \cdots & C_2a^{n-3}C_0 \\ \cdot & \cdot & \cdot & & \cdot \\ C_0a^{n-1}C_0 & C_1a^{n-2}C_0 & C_2a^{n-3}C_0 & \cdots & C_{n-1}a^0C_0 \end{pmatrix},$$

$$C_i = \frac{1-a^{2(i+1)}}{1-a^2} = \sum_{h=0}^{i} \left(a^2\right)^h \qquad (i = 0,1,2,\ldots,n).$$

С целью трансформации регрессионной модели для матрицы P не задается условия, что $P'P = V^{-1}$, одновременно становится возможной реализация GLS-трансформации. В данном случае допустима итеративная процедура определения величины a и непосредственного исчисления параметрических оценок по Айткину.

Обобщения для GLS-оценивания (Оценивание любого числа m параметров, средней и ковариации в линейной модели). Пусть V — любая, не обязательно неединичная, положительно определенная матрица для некоторой линейной регрессионной модели:

$$y = Xb + u,\ E(u) = 0,\ \operatorname{cov}(u) = E(uu') = V \in \boldsymbol{R}^{n,n},$$

$$y \in \boldsymbol{R}^n,\ X \in \boldsymbol{R}^{n,m},\ \operatorname{rg}(X) \leq m,\ b \in \boldsymbol{R}^m,\ u \in \boldsymbol{R}^n.$$

Параметрическая GLS-оценка тогда будет

$$\hat{b} := \left(XV^+X\right)^{-1} XV^+y = b + \left(XV^+X\right)^{-1} XV^+u.$$

Оценка параметра $\hat{b}$ линейна относительно y и u, ее математическое ожидание приближается к истинному значению b:

$$E(\hat{b}) := E\left\{\left(XV^+X\right)^{-1} XV^+y\right\} = E\left\{\left(XV^+X\right)^{-1} XV^+(Xb+u)\right\} =$$

$$= \left(XV^+X\right)^{-1} XV^+Xb + \left(XV^+X\right)^{-1} XV^+XE(u) = b.$$

Характеристика ковариации для параметрической оценки b, используемая в целях трансформации регрессионной модели, будет:

$$\operatorname{cov}(\hat{b}) := E\left\{\left(\hat{b}-b\right)\left(\hat{b}-b\right)'\right\} = \left(XV^+X\right)^{-1} XV^+E(uu')V^+X\left(XV^{-1}X\right)^{-1} =$$

$$= \sigma^2\left(XV^+X\right)^{-1} XV^+VV^+X\left(XV^{-1}X\right)^{-1} = \sigma^2\left(XV^+X\right)^{-1}.$$

При этом становятся возможными трансформационные преобразования по Айткину в соответствии с требованиями GLS-подхода. Трансформационная матрица M выбирается таким образом, чтобы $M'M = W^{-1}$ с W — оцененной матрицей ковариации остатков: $\hat{V} =: W$. Оценивание модельных параметров производится с учетом состава матрицы V и типа регрессионной модели:

- *гетероскедастичная модель* — матрица V содержит дисперсионные показатели $\{\sigma_{ii}\}$;
- *модель с эквивалентной корреляцией остатков* — матрица V включает значения некоторой постоянной величины a. Существует различие оценок дисперсии и ковариаций;
- *авторегрессионная модель* — матрица V обобщает автокоррелирующие значения величины b;
- *модель Койка* — матрица V обобщает значения постоянной величины a.

Для каждого из перечисленных типов моделей оценивание параметров специфично и существенно отличается от других типов.

После оценивания модельных параметров может осуществляться разложение матрицы ковариаций, только вместо неизвестной, определяемой на первом этапе матрицы V ($V^{-1} = M'M$), те-

перь оцениванию подлежит матрица W, также обеспечивающая линейность и эффективность параметрической оценки $(\hat{b})$ —

$$\hat{b} := \left(X'W^{-1}X\right)^{-1}X'W^{-1}y = b + \left(X'W^{-1}X\right)^{-1}X'W^{-1}u;$$

$$E(\hat{b}) = E\left\{\left(X'W^{-1}X\right)^{-1}X'W^{-1}y\right\} = \left(X'W^{-1}X\right)^{-1}X'W^{-1}Xb = b.$$

Эффективность параметрической оценки характеризуется величиной остаточной ковариации:

$$\operatorname{cov}(\hat{b}) = E\left\{(\hat{b}-b)(\hat{b}-b)'\right\} = \left(X'W^{-1}X\right)^{-1}X'W^{-1}E(uu')V^{-1}X\left(X'W^{-1}X\right)^{-1} =$$

$$= \sigma^2\left(X'W^{-1}X\right)^{-1}X'W^{-1}VW^{-1}X\left(X'W^{-1}X\right)^{-1} \neq \sigma^2\left(X'V^{-1}X\right)^{-1}.$$

Другими словами, $W \neq V$, но при достаточно большом числе наблюдений оценки ковариации в матрицах W и V приближаются к истинным значениям, и тогда правомерным становится равенство $\hat{V} = W$, или, при более точной записи, $p\lim_{n\to\infty} W_n = V$.

6.5. Нелинейные регрессионные модели

Для адекватного представления реальных процессов и явлений часто требуются нелинейные регрессионные модели. Как правило, их отличают более сложные процедуры поиска и интерпретации параметрических оценок. В значительной мере эти задачи упрощаются при переходе от нелинейных к линейным модельным формам. Продемонстрированные нами широкие возможности линейной модели с одним уравнением показывают, что нелинейные модели в общем могут подвергаться оцениванию и анализироваться как линейные.

Непосредственная трансформация в линейную модель. Переход от нелинейных связей к линейным может осуществляться на основе простейших трансформационных преобразований модельных параметров или самих изучаемых признаков. В последующем это позволяет производить оценивание параметров с помощью хорошо известного метода наименьших квадратов. Покажем это на примерах моделей, когда $m = 1,2,3$ (табл. 6.2 — с целью упростить представление данных в таблице характеристики остатков модели опущены).

Модель кривой динамического роста (полиномиальная регрессия)[1]. Исходные условия: $y \in R^T$, $X \in R^{T,K}$, $K = P+1$, $\beta \in R^K$.

Таблица 6.2. **Трансформационные переходы от нелинейных к линейным формам регрессионных моделей**

Тип модели	*Параметрические ограничения*	*Трансформационные переходы*	*Замечания*
1	2	3	4
0. $y_i = a + b\cdot g(x_i) + c\cdot h(x_i) + u_i$, $g,h: R \to R$	$a,b,c \in R$	$x_2 := g(x)$, $x_3 := h(x)$	
1. $w_i = a + bv_i + cv_i^2$ 1.0. $w_i = a + cv_i^2$	$a,b,c \in R$ $(b \equiv 0)$	$y := w$, $x := v^2$	Полиномиальная регрессия
1.1. $w_i = a + bi + ci^2$	$a,b,c \in R$	$y := w$, $x := i$, $z := i^2$	Полиномиальный тренд
1.2. $y = \frac{a}{x} + b\log x$			Логарифмическая обратная модель
1.3. $w_i = a + \frac{b}{v_i^k}$, $k = 1,2,\ldots$, $v_i \neq 0$	$a,b, \in R, k \in N$	$y := w$, $x := \frac{1}{v_i}$	Общая инверсионная модель
1.4. $y_i = \frac{b}{i}$	$v_i = i$, $k = 1$		Инверсионный тренд
1.5. $w_i = a + b\cdot i + c\cdot\sin\left(\frac{2\pi i}{12}\right)$	$a,b,c \in R$	$y := w$, $x := i$, $z := \sin(.)$	Тригонометрическая функция
1.6. $w_i = a + bv_i^c$	$c = \frac{1}{2}, 1, 2, 3, \ldots$		См. 1.0

[1] Здесь и далее в § 6.6 в основном приводятся известные эконометрические модели, для них сохраняются оригинальная символика и индексация переменных. (*Прим. пер.*)

Продолжение табл. 6.2

1	2	3	4
1.7. $w_i = bv_i^c$	$b,c \in \mathbf{R}, c \neq 1$, например: $c = \frac{1}{2}, 2, 3, \ldots$	$y := \ln w$, $x := \ln v_i$ $A := \ln b$	См. 1.6 при условии $a \equiv 0$
1.8. $y = a + b \log v_i$			Полулогарифмическая модель
2. $w_i = ae^{bv_i}$	$a,b \in \mathbf{R}$	$y := \ln w$, $x := v_i$	Модель типа log–log
2.1. $w_i = ae^{bi}$	$a,b \in \mathbf{R}$	$y := \ln w$, $x := v_i := i$	Модель постоянного темпа роста
2.2. $w_i = a(1+b)^i$		$y := \ln w$, $x := i$	
2.3. $y_i = A + Bi$		$A := \ln a$, $B := \ln(1+b)$	
3. $w_i = ab^i c^{v_i}$	$a,b,c \in \mathbf{R}$	$y := \ln w, x := i$, $z := v_i$	
3.1. $y = A + B \cdot i + Cv_i$		$A := \ln a$, $B := \ln(1+b)$ $C := \ln c$	

(π — константа тригонометрической периодичности $2\pi r$, r — радиус)

В объясняющую часть модели здесь входит матрица исходных данных X, известная как специальная матрица Вандермонда:

$$X = \begin{pmatrix} 1 & t_1^1 & t_1^2 & \ldots & t_1^P \\ 1 & t_2^1 & t_2^2 & \ldots & t_2^P \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_T^1 & t_T^2 & \ldots & t_T^P \end{pmatrix} \in \mathbf{R}^{T,P+1} \qquad (\therefore m := P+1).$$

В общем виде динамическую модель порядка P можно записать:

$$y_t = \sum_{p=0}^{P} \beta_p t_t^p + u_t .$$

Чтобы расширить представление о модели, рассмотрим для нее ряд особых случаев:

$P := 0$ — константная величина плюс ошибка модели (т. е. модель с оцениванием средней);

$P := 1$ — простая линейная функция плюс ошибка;

$P := 2$ — квадратичная функция плюс ошибка (стохастический квадратичный тренд).

Обобщение динамической полиномиальной модели предполагает возможность ее построения на любых переменных:

$$y_t = \sum_{p=0}^{P} \beta_p (X_t - x)^p + u_t , \quad x_t = X_t - x,$$

здесь x — величина, принимающая некоторые определенные значения, скажем, в случае с центрированными данными $x = \bar{X}$, тогда

$$y = X_{p,x} b + u \text{ и } X_{p,x} = \begin{pmatrix} 1 & X_1 - x & \ldots & (X_1 - x)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n - x & \ldots & (X_n - x)^p \end{pmatrix}.$$

Последняя модель характеризуется так же, как полиномиальная регрессия.

Некоторые тривиальные виды трансформаций в линейном оценивании

Оценивание после введения в модель новой переменной (модель темпов роста). Линеаризация нелинейной модели часто осуществляется посредством введения новых переменных, представляющих переменные исходной модели.

Рассмотрим показатели скорости изменения явлений во времени — темпы роста:

$$y_{it} := \frac{\tilde{y}_{it} - \tilde{y}_{it-1}}{\tilde{y}_{it-1}} \text{ или } y_{it} := \frac{\tilde{y}_{it} - \tilde{y}_{it-1}}{\tilde{y}_{it}},$$

или в более общем виде $y_{it} := \frac{\tilde{y}_{it} - \tilde{y}_{it-1}}{\tilde{y}_{kt-1}}$, $k \neq 1$.

На основе взаимосвязанных темпов роста построим логнормальную модель. Пусть $z_t = y_{tt}$, тогда правомерно перейти к производственной функции Кобба — Дугласа, оценивание которой хорошо известно из экономической теории ($f: R^n \to R$):

$f(z) = a_0 \prod z_j^{a_j} \Phi u$ (u — случайная величина, остаток).

Гипотетически пренебрегаем симметрично распределенной ошибкой Φu и получим ряд Тейлора для разложения f по z_0:

$$f(z) \approx f(z_0) + \sum_{j=1}^{n} \frac{\partial f(z_0)}{\partial z_j}(z_j - z_{0j}),$$ (z_0 — исходный пункт развития явления во времени)

$$\frac{f(z)-f(z_0)}{f(z_0)} = \sum_{j=1}^{n} \frac{\partial f(z_0)}{\partial z_j}\frac{(z_j - z_{0j})}{f(z_0)} = \sum_{j=1}^{n} \frac{a_j f(z_0)}{z_j}\frac{(z_j - z_{0j})}{f(z_0)} =$$

$$= \sum_{j=1}^{n} \frac{a_j (z_j - z_{0j})}{z_j}$$

$$\therefore y_{0t} = \sum_{j=1}^{n} a_j y_{jt},\ y_{0t} := \frac{f(z)-f(z_0)}{f(z_0)},\ y_{jt} := \frac{z_j - z_{0j}}{z_j}.$$

Коэффициенты a_j в данном случае интерпретируются как коэффициенты эластичности:

$$a_j = \frac{\partial f(z_0) z_j}{f(z_0)\partial z_j} = \frac{\partial f}{f}\cdot\frac{z_j}{\partial z_j}.$$

Область практического приложения производственной функции Кобба — Дугласа в экономической теории достаточно широка. Кроме линеаризации, в той или иной форме она используется при выводе формул Мейда, построении закона Вердурна, построении экономической модели общего равновесия.

Оценивание дискретной функции Торнквиста. Обратимся непосредственно к проблеме линейного оценивания. Например, при помощи МНК-метода можно определить a, b, c — параметры следующей функции:

$$y_t = a\frac{x_t - c}{x_t + b} \quad (t = 1, 2, ..., T), \tag{6.26}$$

здесь $\{y_t, x_t\}$ — наблюдаемые величины.

В ходе преобразований (6.26) получаем обычную линейную регрессионную модель:

$$y_t = a\frac{x_t - c}{x_t + b} \Leftrightarrow (x_t + b)y_t = a(x_t - c) \Leftrightarrow x_t y_t = -ac - by_t + ax_t \Leftrightarrow$$

$$\Leftrightarrow Y_t = \beta_1 + \beta_2 y_t + \beta_3 x_t.$$

Таким образом, параметры модели идентифицируются и легко вычисляются по известной формуле: $\beta = (X'X)^{-1}X'y$. Данный подход позволяет решать аналогичные задачи по линеаризации для самых разнообразных форм связи.

Итеративное линейное оценивание билинейной модели

Итеративная линейная модель. Расширенное представление о нелинейной модели связывается с допущением изменения числа ее параметров. При этом фиксация значений одного или нескольких параметров (β_j) делает возможным переход от нелинейной модели к линейной.

Пусть в регрессионной модели вида:

$y_t = f(x_t; \beta) + u_t$ $(t = 1,2,...,T)$, $\beta' = (\beta_1 \beta_2, ..., \beta_K)$, $K \in N$

или после трансформации

$y_t = X_t(\gamma)\beta + u_t$,

где оцениванию уже подлежит некоторое число γ из всех параметрических оценок β.

Обратим внимание, что после трансформации модель становится линейной, т.е. можно сказать, что линейность в данном случае обусловливается параметрическим числом γ. При этом часто K весьма мало, скажем, $K \le 5$. В английском языке подобная модифицированная модель определяется как частично линейная модель, сепарабельная нелинейная модель, дихотомическая модель[1]. Типичный пример такой модели:

$$y_t = \alpha + \beta e^{\gamma x_t} + u_t \Leftrightarrow y_t = \alpha + x(\gamma)_t \beta + u_t \text{ и } x(\gamma)_t := e^{\gamma x_t}.$$

Линейное оценивание α и β в ней производится после предварительного итеративного задания γ.

Итеративный подход. Предположим, вместо обычной линейной модели

$y = X\beta + u$, $E(u)$, $\operatorname{cov}(u) = V$

[1] Partially linear model, separable nonlinear model, dichotome model. (*Прим. пер.*)

существует форма связи:

$y = X(\beta)\beta + u$, или $y(\beta) = X\beta + u$, или $y(\beta) = X(\beta)\beta + u$.

Одно из обоих числовых полей данных (y, X) или одновременно оба определяют некоторую функцию, оценивание которой ведет к получению параметрического вектора β. При этом оценивание, приводящее нелинейную модель к линейной форме, строится по следующей итеративной схеме:

1) задание исходного значения b, b — оценка параметра β;

2) определение начального итерационного индекса для решения линейной модели $i := 0$;

3) оценивание линейной модели $i := i + 1$ (переход на следующий итерационный круг):

$$X^{(i)} := X\left(b^{(i-1)}\right),\ y^i := y\left(b^{(i-1)}\right),\ b^{(i)} := \left(X^{i\,\prime} X^i\right)^{-1} X^{i\,\prime} y;$$

4) проверка по критерию конвергенции и принятие решения о прерывании итераций или их продолжении, т.е. о переходе к шагу (3).

Иллюстрация статистического оценивания модельных параметров итеративным способом. В табл. 6.3 показаны некоторые важнейшие случаи практического приложения итеративного подхода, случайная величина ошибки u_t во внимание в данном случае не принимается.

Пример 6.15. Оценивание CES-функции.

Исходная функция имеет вид:

$$Q = \left(\sum_{k=1}^{K} \beta_k v_k^{\alpha}\right)^{1/\alpha},\ \alpha \neq 0.$$

После возведения левой и правой частей функциональной модели в степень α становится возможным провести линейное оценивание:

$$Q^{\alpha} = \sum_{k=1}^{K} \beta_k v_k^{\alpha} + u \Leftrightarrow Q_t^{\alpha} = \sum_{k=1}^{K} \beta_k v_{kt}^{\alpha} + u_t.$$

Здесь величина, определяемая итерационным способом, — это α.

Пример 6.16. Оценивание авторегрессионной модели. Итерационный поиск двух параметров.

Таблица 6.3. **Примеры оценивания модельных параметров итеративным способом**

Тип модели	*Параметр*	*Параметр, определяемый итеративным способом*
1. $w_t = ab^t + c^{v_t}$	$a,b,c \in \boldsymbol{R}$	c
2. $y_t = \alpha + \beta_1 x_t + \beta_2 x_t^{\gamma}$	$\alpha,\beta_1,\beta_2 \in R$	γ
2.1. $y_t = \alpha + \beta_1 x_t + \beta_2\beta_3 x_t^2$	$\alpha,\beta_1,\beta_2,\beta_3 \in R$	$\beta_1 \vee \beta_2$
2.2. $w_t^a = b + cv_t^d$	$a,b,c,d \in \boldsymbol{R}$	d
3. $w_t = a + bv_t + ce^{dt}$	$a,b,c,d \in \boldsymbol{R}$	d
3.1. $y_t = \alpha + \beta e^{\gamma x_t}$	$\alpha,\beta,\gamma \in \boldsymbol{R}$	$\alpha \vee \gamma$
4. $w_t = a + \dfrac{b}{v_t - c},\ v_t \neq c$	$a,b,c \in \boldsymbol{R}$	c
5. $y_t = \alpha + \beta_1(x_t - \gamma) + \beta_2(x_t - \gamma)^2$	$\alpha,\beta_1,\beta_2,\gamma \in R$	γ
5.1. $w_t = a\left(v_t^b + t^b\right)^c$	$c = 1$, см. п. 1.	b
6. $y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 \exp(\beta_4 x_{3t})$	$\beta_1,\beta_2,\beta_3,\beta_4 \in R$	β_4
6.1. $y_t = \alpha + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 \exp(\beta_4 x_t)$	$\alpha,\beta_1,\beta_2,\beta_3,\beta_4 \in R$	β_4

Рассмотрим линейную авторегрессионную модель, включающую две случайные величины — u и v:

(1) $y_t = X_t\beta + u_t,\ u_t = \rho u_{t-1} + v_t$.

Очевидно, что исходная модель может быть модифицирована:

$\rho y_{t-1} = \rho X_{t-1}\beta + \rho u_{t-1}$ (с учетом временной инвариантности модели),

$y_t - \rho y_{t-1} = (X_t - \rho X_{t-1})\beta + (u_t - \rho u_{t-1})$ (модификация модели элементарным вычитанием),

$y_t - \rho y_{t-1} = (X_t - \rho X_{t-1})\beta + v_t \Leftrightarrow$

(2) $\Leftrightarrow y_t = \rho y_{t-1} + (X_t - \rho X_{t-1})\beta + v_t$.

Из уравнения (2) следует как бы двойная итерационная процедура:

Шаг 1. Зададим для ρ_k, $k := -1$ (стартовый итерационный индекс).

Шаг 2. Пусть $k := k+1$.

Оценка $\hat{\beta} \therefore$ (МНК-метод, первое приложение для k-й итерации)

$z_t := y_t - \rho_k y_{t-1}$, $w_t := x_t - \rho_k x_{t-1}$,

$$z_t = w_t\beta + v_t \therefore \hat{\beta} = \frac{\sum_{t=2}^{n} w_t z_t}{\sum_{t=2}^{n} w_t^2}.$$

Шаг 3. На основе линейной модели (2) определим новое ρ_{k+1}. Исходные данные для этого представляются «значениями новых переменных» (наблюденными оценками остатков $\hat{u}_t$ и $\hat{u}_{t-1}$): $\hat{y}_t := X_t\hat{\beta}$, $y_t - \hat{y}_t =: \hat{u}_t$.

Шаг 4. Оценка $\hat{\rho} \therefore$ (МНК-метод, второе приложение для k-й итерации)

$$\hat{u}_t = \rho\,\hat{u}_{t-1} + \hat{v}_t \therefore \hat{\rho} = \frac{\sum_{t=2}^{n} \hat{u}_t\hat{u}_{t-1}}{\sum_{t=2}^{n} \hat{u}_{t-1}^2}.$$

Шаг 5. Возвращение к шагу 1 до тех пор, пока показатель конвергенции не окажется меньше некоторого порогового значения:

$|\hat{\rho}_k - \hat{\rho}_{k-1}| < \varepsilon$.

Если неравенство выполняется, алгоритм завершен, если неравенство не выполняется, происходит возобновление алгоритма в результате возвращения к шагу 1.

Как видим, итерации последовательно выполняются для двух оцениваемых величин — ρ и β [105, с. 159–165].

Отметим, что в модели Койка похожим образом, т.е. в ходе итераций, одновременно оптимизируются параметр a и связанный с ним второй параметр b.

Пример 6.17. Логит — трансформация: линейное оценивание для логистической модели.

Логистическая регрессия представляется нелинейной трехпараметрической моделью, на графике это s-образная кривая,

описывающая текущее развитие процесса под воздействием фактора времени t (рис. 6.2):

$$y_t = \frac{a}{1 + be^{-ct}},\ 0 < a,\ a \geq y_t,\ 0 < c, \Leftrightarrow a, b, c \in \boldsymbol{R}_{++}.$$

Исходная модель логистической регрессии с целью перехода к линейному оцениванию может быть модифицирована, например, следующим образом:

$$Y_t := \ln\left(\frac{a}{y_t} - 1\right) = \ln b - ct =: b_1 + b_2 t.$$

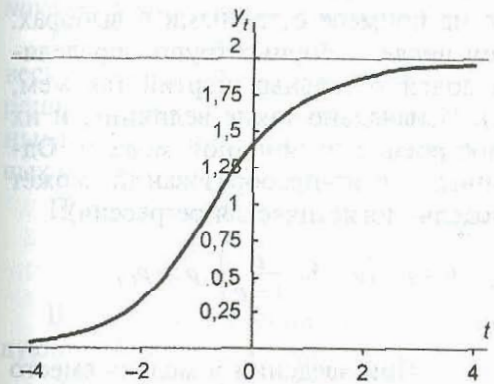


Рис. 6.2. График логистической регрессии при $a = 2$, $b = 0{,}5$, $c = 1$

Трансформационный прием, при котором в анализе осуществляется переход к величине $\ln\left(\frac{a}{y_t} - 1\right)$, называют *логит-трансформацией* (или логит-оцениванием). При дополнительной замене — $ct \sim X_t' b$ получают обобщенную линейную форму логистической модели:

$$Y_t := \ln\left(\frac{a}{y_t} - 1\right) = \ln b + X_t'\beta =: \beta_0 + X_t'\beta,$$ или, после исключения константы: $Y_t = X_t'\beta$.

Как видим, получена обычная линейная модель. Итеративным способом задавая значения a (а также введением аддитивной случайной величины u_t), добиваются оптимизации решающей модели и адекватных теоретических значений для y_t.

Существует и другой вариант логит-трансформации. Если объясняющая эндогенная переменная принимает только два — значения 0 и 1, как это, скажем, бывает при обработке данных о политических выборах, покупках автомобилей, наложении штрафов, то все наблюдения разделяются также на две части, соответственно представленные двумя величинами 1 и 0:

$p_t = [Z_t = 1]$ — событие наступило,

$1 - p_t = [Z_t = 0]$ — событие не наступило.

При этом допустимо для случайной величины p_t предположить действие закона распределения Бернулли, тогда математическим ожиданием для y_t будет $E(y_t) = p_t$, а дисперсией — $\mathrm{var}(y_t) = (1 - p_t)p_t$ и модель для p_t (наблюдаемой доли случаев наступления события), запишется как линейная: $p_t = X_1\beta_1 + X_2\beta_2 + u_t$.

Показанная последовательность преобразования линейной модели универсальна для широкого круга практических задач. Остановимся более подробно на примере с данными о выборах. Удельный вес партий в общем числе выборных групп определяется как взвешенная сумма долей отдельных партий (скажем, партий рабочих и служащих). Изначально такие величины и их взаимосвязи предполагают построение нелинейной модели. Однако в ходе модификационных логит-преобразований может быть получена и линейная модель (логистическая регрессия):

$$p = f(y),\ y = X\beta,\ f(y) = \frac{e^z}{1+e^z},\ y = f^{-1}(p) = \ln\left(\frac{p}{1-p}\right),\ p := p_t,$$

$$\therefore y_t = X_t\beta.$$

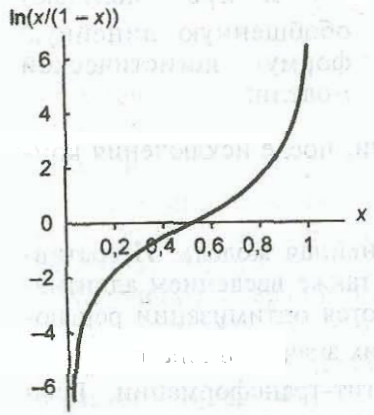


Рис. 6.3

При введении в модель вместо p наблюдаемой доли $y_t = \ln\left(\frac{a_t}{1-a_t}\right)$ и одновременном дополнении ее аддитивной случайной величиной получаем самую обычную линейную модель, только с ограничением $0 \le p_t \le 1$:

$f(y) = y = X\beta$.

Рис. 6.3 демонстрирует трансформационные преобразования и переход от p к y:

Замечания к линеаризации. *Замечание 1. Стохастическая спецификация.*

Обратим внимание, что само преобразование модели в линейную оправдано только при учете стохастической величины

u_t как аддитивной переменной (шума), в некоторых случаях величина u_t также подвергается оцениванию. В связи с этим будем различать две основные стохастические спецификации для нелинейной модели:

$$y_t = f(x_t;\theta)\Phi u_t, \text{ где } \Phi u_t \begin{cases} +u_t \\ \cdot e^{u_t} \end{cases}, \text{ и}$$

$$\therefore \text{либо } y_t = f(x_t;\theta) + u_t, \text{ либо } y_t = f(x_t;\theta)e^{u_t} \Leftrightarrow \ln y_t = \ln f(x_t;\theta) + u_t.$$

Замечание 2. Выбор зависимых величин и введение ограниченного числа новых обозначений.

Как показывает пример, приведенный ниже, часто бывает весьма важным решить вопрос о выборе зависимых величин, т.е. решить, какие из величин будут эндогенными, а какие экзогенными. Параллельно возникает задача корректного введения новых обозначений.

Пусть $X = AB^Y u \Leftrightarrow \ln X = \ln A + Y\ln B + \ln u,\ u > 0,\ A > 0,\ B > 0 \Leftrightarrow$

$$\Leftrightarrow Y = a + b\ln X + v,\ a := -\frac{\ln A}{\ln B},\ b := -\frac{1}{\ln B},\ v := -\frac{\ln u}{\ln B}.$$

В качестве альтернативных можно рассматривать также следующие преобразования:

$$X = AB^Y e^u \Leftrightarrow \ln X = \ln A + Y\ln B + u,\ u \in R,\ A > 0,\ B > 0 \Leftrightarrow$$

$$\Leftrightarrow Y = a + b\ln X + v,\ a := -\frac{\ln A}{\ln B}, b := -\frac{1}{\ln B},\ v := -\frac{u}{\ln B}.$$

Выводы[1]

Регрессионный анализ как статистический метод предназначен для моделирования стохастических процессов.

Вероятностный характер природы наблюдаемых и описываемых с его помощью объектов требует поиска по возможности наиболее простой теоретической формы представления признаковых связей и статистической оценки надежности как самих моделей, так и модельных параметров. С этой точки зрения особое значение приобретают линейные регрессионные модели, а

[1] Параграф подготовлен В.Н. Тамашевичем.

также исходное предположение о нормальности распределения параметрических оценок.

Линейные модели отличаются простой интерпретируемостью и хорошо разработанными приемами оценивания коэффициентов регрессии. Обычно для них все три наиболее распространенных метода статистического оценивания — максимального правдоподобия, наименьших квадратов и моментов — дают оптимальные решения и соответственно приводят к оценкам, обладающим линейностью, эффективностью, несмещенностью. Надежность получаемых оценок при этом обусловливается теоретическими посылками используемых подходов: Гаусса—Маркова — о нормальности распределения и оптимальности параметрических оценок и Айткина — об оптимальности оценок трансформационной модели.

Принимая во внимание, что линейные регрессионные модели не могут с одинаково высокой степенью достоверности описывать многообразные процессы, происходящие в реальности, их дополняет большой класс нелинейных моделей. Для последних, однако, с учетом их сложности и специфичности приемов параметрического оценивания предпочтительно остается приведение к простой линейной форме.

В общем многообразие видов регрессионных моделей порождается формой связи изучаемых признаков (линейной или нелинейной) и представлениями о распределении остатков (ошибки, шума) модели. Кроме того, модели более высокого уровня включают не одно, а систему регрессионных уравнений. Поиск решений для множества моделей приводит исследователя к задаче преобразования этих моделей и получения форм с хорошо известными и реализуемыми алгоритмами оценивания, как, например, в описанном выше случае с нелинейными моделями. Реформирование моделей производится при помощи трансформационных изменений переменных (отклика, предикторов) или введением особых ограничений на признаковые или параметрические значения.

Благодаря своей разработанности и гибкости метод регрессионного анализа в настоящее время широко распространен в аналитической практике. Он становится также неотъемлемой частью или обычным логическим дополнением многих методов многомерной статистики: в факторном, дискриминантном анализе, методе канонических корреляций, многомерном шкалировании, кластерном анализе и т. д.

Дальнейшее развитие теории регрессионного анализа прежде всего видится в разработке новых нелинейных форм, позволяющих с высокой степенью адекватности описывать реальные процессы; расклассификации многочисленных регрессионных моделей и методов их решения, ориентированной на конкретные группы исследовательских задач; определении перспектив использования регрессионного анализа в сочетании с другими методами статистического анализа.

Вопросы и задачи

1. Определите оценки $\hat{b}_0, \hat{b}_1$ и $\hat{\sigma}^2$ по следующим данным трех наблюдений:

$$y = \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} \in R^3, \quad X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \in R^{3,2}.$$

2. Имеется:

$$y = \begin{pmatrix} 16 \\ 10 \\ 7 \\ 0 \\ 2 \\ 1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & -1 & -1 & 0 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}.$$

Определите параметрический вектор $\hat{b} = (X'X)^{-1}X'y$ и покажите, что для матрицы исходных данных X и вектора y, имеющих компоненты в виде целочисленных величин, значения $\hat{b}_j$ будет всегда также целыми числами.

3. Моделирование циклических колебаний. Найдите параметрические оценки для следующих двух простых регрессионных моделей:

$y_t = a + a(-1)^t + u_t$, $t = 1,2,...,T$ (T может быть четко определено либо оставаться неопределенным),

$y_t = a + b\sin(t) + u_t$, $t = 1,2,...,T$.

4. Пусть

$$X = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \\ \\ 1 \\ n \end{pmatrix} \in R^n \text{ в } y = Xb + u.$$

Покажите, что при этом параметрическая МНК-оценка $\hat{b}$ будет равна:

$$\hat{b} = \frac{\sum_{i=1}^{n-1} y_i + n y_n}{n - 1 + n^2}.$$

5. Если в исходной регрессионной модели вместо константы имеется какая-либо любая переменная и эта переменная принимает только единичные значения, то эту модель можно представить в виде, пригодном для оценивания ее параметров известными методами, например:

$y_i = b_1 X_{1i} + b_2 X_{2i} + u_i$, $u_i \sim N(0, \sigma^2)$ $(i = 1,2,\ldots,n)$, или

$u_i = Y_i - b_1 X_{1i} - b_2 X_{2i}$.

Для оценивания b_1, b_2 и σ^2 во втором случае может быть использована функция максимального правдоподобия (константа заменяется любой переменной).

Оцените параметры регрессионного уравнения по данным трех наблюдений:

$$y = \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} \in R^3,\ X = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \in R^{3,2} \text{ и } y = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \in R^3,\ X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \in R^{3,2}.$$

6. Обратная регрессия. По данным задачи **1** постройте модель обратной регрессии и произведите оценивание ее параметров $(\hat{b}_0, \hat{b}_1, \hat{\sigma}^2)$.

7. Задача с конечным лагом. Пусть имеем линейную модель вида:

$$y_t = \sum_{s=0}^{n} X_{t-s} b_s + u_t,\ n = 1,2,\ldots,$$

где $n=4$ и $b_s = w$ — константа, подлежащая оцениванию.

а) Задайте обычную линейную модель типа $y = Xb + u$ и запишите для нее данные (четыре числовых поля — для y, X, b, u).

б) Интерпретируйте полученную модель и ее параметры, попытайтесь определить область приложения данной модели.

в) Каким образом можно оценить параметр w? Произведите оценивание и найдите $\hat{w}$.

Указание: ответ для в) следует из а).

8. Построение прогноза. Для получения прогнозных данных в имеющейся регрессионной модели задаются некоторые значения экзогенной величины X_T. Какими свойствами при этом будут обладать прогнозные значения y_T, если:

$$y_T = b_0 + b_1 X_T + u_T \quad .$$

Теоретическое решение: фактически параметры b_0 и b_1 неизвестны. С целью прогноза предварительно следует найти оценки $\hat{b}_0$ и $\hat{b}_1$ и подставить их в исходную модель:

$$\hat{y}_T = \hat{b}_0 + \hat{b}_1 X_T.$$

Свойства прогнозных значений $\hat{y}_T$ определяются оценками его математического ожидания и дисперсии:

$$E(\hat{y}_T) = E(\hat{b}_0 + \hat{b}_1 X_T) = b_0 + b_1 X_T,$$

$$\operatorname{var}(\hat{y}_T) = E\left[\hat{y}_T - E(\hat{b}_0 + \hat{b}_1 X_T)\right]^2 = E\left[(\hat{b}_0 - b_0) + (\hat{b}_1 - b_1) X_T\right]^2 =$$

$$= E\left[(\hat{b}_0 - b_0)^2 + E(\hat{b}_1 - b_1)^2 X_T^2\right] + 2E\left[(\hat{b}_0 - b_0)(\hat{b}_1 - b_1) X_T\right] =$$

$$= \operatorname{var}(\hat{b}_0) + \operatorname{var}(\hat{b}_1) X_T^2 + 2\operatorname{cov}(\hat{b}_0, \hat{b}_1) X_T.$$

9. Обобщенный случай квантимального оценивания. Интересной областью применения метода наименьших квадратов является оценивание параметров распределения Парето. Полученные при этом дискретные параметрические оценки открывают возможность сравнения эмпирических и теоретических распределений данных.

Проанализируем функцию вида: $\hat{F}(x_t) := 1 - \left(\frac{b}{x_t}\right)^a$.

После логарифмирования и добавления мультипликативной величины ошибки модели будем иметь:

$$1 - \hat{F}(x_t) := \left(\frac{b}{x_t}\right)^a e^{U_t} \Leftrightarrow y_t := \ln\left[1 - \hat{F}(x_t)\right] = a\ln b - a\ln x_t + U_t .$$

Последняя линейная форма регрессионного уравнения позволяет при помощи метода наименьших квадратов определить интересующие нас параметры a и b. Однако вначале введем новые обозначения и приведем аналитическое уравнение к классическому виду:

$y_t = b_1 + b_2 \ln x_t + u_t, b_1 := a\ln b,\ b_2 := -a,\ u_t := U_t$.

Оценивание b_1 и b_2 проводится хорошо известным нам образом. В свою очередь при известных $\hat{b}_1$ и $\hat{b}_2$ легко вычисляются параметрические оценки a и b:

$$\hat{a} := -\hat{b}_2,\ \ \hat{b} := \exp\left(\frac{\hat{b}_1}{\hat{b}_2}\right).$$

Решение задачи в значительной степени упрощается, если заранее известны оба параметра b_1 и b_2.

10. Покажем на примере, что применение метода наименьших квадратов становится весьма проблематичным, если регрессионная модель состоит из более чем одного уравнения. Пусть имеется: $q = b_0 y + u$; $y = b_1 q + v$, где, например, q — количество продуктов питания, y — уровень доходов населения.

Обе переменные имеют центрированные значения, т.е. $e'q = 0, e'y = 0,\ e, q, y \in R^T$. Оценка параметра b_0 методом наименьших квадратов есть $\hat{b}_0$. Следует доказать, что при этом имеет силу равенство:

$$p\lim \hat{b}_0 = p\lim \frac{\sum_t q_t y_t}{\sum_s y_s y_s} = b_0 + \frac{E\left[\sum_t y_t u_t\right]}{E\left[\sum_s y_s y_s\right]} = b_0 + (1 - b_0 b_1)\frac{b_1\sigma_u^2 + \sigma_{uv}}{b_1^2\sigma_u^2 + 2b_1\sigma_{uv} + \sigma_v^2}.$$

В порядке дискуссии обсудите особый случай, когда $\sigma_{uv} = 0$.

11. Существование двух эмпирических функций. Оцените две следующие функции:

а) функция уровня урожайности сельскохозяйственных культур:

$y_i = b_0 + b_1 f_i + b_2 f_i^2 + u_i,\ b_0 > 0,\ b_1 > 0,\ b_2 < 0$,

где y — урожайность (с одного гектара);

f — количество удобрений, внесенных на один гектар;

б) функциональная модель с двумя уравнениями, описывающими реально существующие зависимости объема производства y и уровня потребительских цен p в стране с небольшой территорией:

- уравнение (1) для отклика «объем производства» y имеет вид:

$$d\ln y = -\frac{\delta_p\varepsilon_w}{\varepsilon_p + \delta_p} d\ln w + \left(\frac{\delta_p^f\varepsilon_p - \delta_p\varepsilon_p^f}{\varepsilon_p + \delta_p}\right) d\ln p_f + \frac{\delta_e\varepsilon_p - \delta_p\varepsilon_e}{\varepsilon_p + \delta_p} d\ln e + \frac{\delta_m\varepsilon_p}{\varepsilon_p + \delta_p} d\ln m + \frac{\delta_g\varepsilon_p}{\varepsilon_p + \delta_p} d\ln g + \frac{\delta_x\varepsilon_p}{\varepsilon_p + \delta_p} d\ln x_f - \frac{\delta_p\varepsilon_z}{\varepsilon_p + \delta_p} d\ln z ; \quad (1)$$

- уравнение (2) для отклика «уравнение цен» p:

$$d\ln p = \frac{\varepsilon_w}{\varepsilon_p + \delta_p} d\ln w + \left(\frac{\delta_p^f - \varepsilon_p^f}{\varepsilon_p + \delta_p}\right) d\ln p_f + \left(\frac{\delta_e + \varepsilon_e}{\varepsilon_p + \delta_p}\right) d\ln e + \frac{\delta_m}{\varepsilon_p + \delta_p} d\ln m + \frac{\delta_g}{\varepsilon_p + \delta_p} d\ln g + \left(\frac{\delta_x}{\varepsilon_p + \delta_p}\right) d\ln x_f + \frac{\varepsilon_z}{\varepsilon_p + \delta_p} d\ln z . \quad (2)$$

Здесь приняты обозначения: δ и ε — показатели эластичности; w — ставка заработной платы; p_f — средняя цена импорта; e — обменный курс для отечественной денежной единицы; m — объем реального импорта; g — реальные расходы государства; x_f — объем реального экспорта; z — прочие определенные величины; d — процентное изменение.

Из общего числа переменных две первые — δ и ε — подлежат оцениванию, все другие $(w, p_f, e, m, g, x_f, z)$ — это объясняющие экзогенные величины.

12. Докажите, что если некоторый k-й столбец в матрице исходных данных X представляет постоянную величину (константные значения):

$$X_k = a \cdot e,\ e' = (1\ \ 1\ \ \ldots\ \ 1),\ e \in R^n,\ a \neq 0,\ a \in R,$$

то для него имеет силу равенство:

$$(X'X)^{-1} X'e = \frac{1}{a} e_k,$$

где e_k — k-й единичный вектор.

Указание: решение становится очевидным, если рассматривать линейную модель:

$$y = Xb + u \ \text{ с } \ y = e \ \text{ и } \ X \in R^{n,m}, \qquad (1)$$

X — любая матрица: $X = (X_1 X_2 \ldots X_k \ldots X_m)$, $X_k = a \cdot e$. Равенство (1) выполняется, когда k-й единичный вектор представляется параметрической оценкой: $b := e_k$, т.е. $b_j = 0$, $j \neq k$, b_k

13. Выборочная величина. Пусть имеется выборка с одним наблюдением, произведенная из генеральной совокупности и подчиняющаяся закону нормального распределения: $X \sim N(\mu, \sigma^2)$. Обозначим это наблюдение x. Функция максимального правдоподобия для x запишется:

$$L(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right].$$

Условием, необходимым для получения оптимальной оценки функции максимального правдоподобия, является ее дифференцирование:

$$\frac{\partial L}{\partial \mu} = 0 = \frac{-(x-\mu)}{\sigma^2} \Rightarrow \hat{\mu} = x, \quad \frac{\partial L}{\partial \sigma^2} = 0 = -\frac{1}{2\sigma^2} + \frac{(x-\hat{\mu})^2}{2\sigma^4} \Rightarrow \hat{\sigma}^2 = 0.$$

Нулевое значение, полученное для дисперсии, — бессмысленный результат. Число наблюдений, равное единице, очевидно, слишком мало, чтобы надежно определить два параметра регрессионной модели.

14. Ранговые дополнения. Рассмотрим линейную регрессионную модель из области биологии. Первоначально имеются данные о весе растений некоторого вида, причем шесть изученных растений расклассифицированы по трем типовым группам:

Тип	*Вес*	*Средний вес*
1	101, 105, 94	100
2	84, 88	86
3	32	32
В целом для популяции		84

Основным проблемным вопросом здесь является определение влияния типа растения на его вес. Суть задачи требует использования ANOVA-модели вида $y_{ij} = \mu + \alpha_i + u_{ij}$, где μ — величина, представляющая тип растения, например, среднее значение изучаемого признака — вес для данной популяции; α_i — величина, характеризующая степень влияния на значение изучаемого признака i-го типа растения; u_{ij} — пусть будет некоррелированной гомоскедастичной ошибкой модели. Индекс j изменяется от 1 до n_i как число наблюдений для i-го типа.

Соответственно каждому наблюдению можем построить свое регрессионное уравнение связи:

$$\begin{array}{llll}
y_{11} = \mu + \alpha_1 + u_{11} & y_1 = b_1 + b_2 + u_1, \\
y_{12} = \mu + \alpha_1 + u_{12} & y_2 = b_1 + b_2 + u_2, \\
y_{13} = \mu + \alpha_1 + u_{13} & y_3 = b_1 + b_2 + u_3, \\
y_{21} = \mu_1 + \alpha_2 + u_{21} & y_4 = b_1 + b_3 + u_4, \\
y_{22} = \mu + \alpha_2 + u_{22} & y_5 = b_1 + b_3 + u_5, \\
y_{31} = \mu + \alpha_3 + u_{31} & y_6 = b_1 + b_4 + u_6.
\end{array} \quad \Leftrightarrow$$

Все уравнения укладываются в одну регрессионную модель, включающую три независимые объясняющие переменных, четыре параметра и данные шести наблюдений. Такая модель без введения дополнительных ограничений не имеет решени относительно b, так как имеют место нарушения ранга матрицы X.

Запишем модель в матричном виде, а именно $y = Xb + u$, или

$(\mu \ \alpha_1 \ \alpha_2 \ \alpha_3)$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 101 \\ 105 \\ 94 \\ 84 \\ 88 \\ 32 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad rg(X) = 3, \ Xz = 0, \ z := \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix}, \quad b = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}.$$

Ответьте на вопросы:

1) Какой метод оценивания следует избрать для определения четырех параметров: $\mu, \alpha_1, \alpha_2, \alpha_3$? Покажите алгоритм решения задачи параметрического оценивания. Какие суждения могут быть высказаны относительно статистической оцениваемости модельных параметров в данном случае?

2) Как может измениться алгоритм решения, если дополнительно ввести следующие параметрические ограничения:

2а) $Rb = r, R = (1 \ 1 \ 1 \ -2), 0 = r \Leftrightarrow \mu + \alpha_1 + \alpha_2 = 2a_3$, или в виде альтернативы

2б) $Rb = r, R = (0 \ 1 \ 1 \ 0), 0 \neq r = 220 \Leftrightarrow b_1 + b_2 = 220.$

Ограничения 2а) и 2б) позволяют получать обычную полноранговую модель, т.е. носят идентифицирующий характер.

3) Какое воздействие окажут ограничения 2а) и 2б) на оценивание модельных параметров?

15. Неполный ранг. Имеется линейная модель со следующими значениями эндогенной переменной y и матрицы исходных данных X:

$$y = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 3 \\ 3 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Каким образом может быть оценен параметр b в модели вида: $y = Xb + u$?

16. Оценивание при введении дополнительных условий. Имеется $y = Xb + u, r = Rb$, при

$$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \quad r = \begin{pmatrix} 7 \\ 4 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

Определите оценки $\hat{b}$ для $y = Xb + u$.

Ответ: $\hat{b}' = \left(\frac{23}{4}, -\frac{1}{4}, \frac{3}{4}\right)$.

17. Гетероскедастичность. Имеются модель вида:

$y = \mu + u \Leftrightarrow y_i = \mu + u_i \ (i = 1,2,3,\ldots,n)$

и два особых случая для этой модели:

Случай 1: $E(u_i \mid x_i) = 0$, но $\mathrm{var}(u_i \mid x_i) = \sigma^2 x_i^2, \ x_i > 0$;

Случай 2: Ошибка модели имеет нормальное распределение с математическим ожиданием, равным нулю, и дисперсией $\mathrm{var}(u_i \mid x_i) = \sigma^2 \left(1 + (\gamma x_i)^2\right)$.

Принимая во внимание оба приведенных случая, ответьте на следующие вопросы:

а) Покажите, что σ^2 и γ^2 могут быть вещественно оценены методом наименьших квадратов.

б) Для одной выборки данных y и x произведите оценивание μ.

в) Насколько могут различаться параметрические оценки в обоих случаях при их оценивании методом наименьших квадратов?

г) Охарактеризуйте эффективность полученных статистических оценок.

д) Определите основные свойства (например, вещественную структуру) для s^2, если

$$s^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2 .$$

е) Обсудите оценку: $\mathrm{var}(\bar{y}) = s^2(X'X)^{-1} = \frac{s^2}{n}$.

Указание: в результате трансформации данных и достижения гетероскедастичности осуществляется переход к обычной гомоскедастичной модели для средней величины или, что то же самое, линейной модели с переменными, которые могут быть оценены методом наименьших квадратов, причем оценки обладают BLU-свойствами. Тем самым получим основания для ответов на все вопросы:

$$\frac{y_i}{x_i} = \mu\frac{1}{x_i} + \frac{u_i}{x_i} \Leftrightarrow z_i = \mu v_i + w_i,\ E(w_i) = 0,\ \mathrm{var}(w_i) = \sigma^2,\ \hat{\mu} = \frac{\sum_{i=1}^{n} v_i z_i}{\sum_{i=1}^{n} v_i v_i}.$$

Для случая 2 требуется предварительное введение альтернативной величины γ. С этой целью исходную модель умножают на дробь $\frac{1}{\sqrt{1+\gamma^2 x^2}}$ и получают модифицированную модель с одним параметром, обладающим BLU-свойствами.

18. Используя неравенство Шварца:

- покажите, что

  1) $(a_1 + a_2 + \ldots + a_n)^2 \le n\left(a_1^2 + a_2^2 + \ldots + a_n^2\right)$;

- докажите, что для существующих векторов x и y с любой размерностью выполняются неравенства:

  2) $(x'y)^2 \le (x'x)(y'y)$ и

  3) $\sqrt{(x+y)'(x+y)} \le \sqrt{x'x} + \sqrt{y'y}$.

19. Случай с двумя выборками. Произведите оценивание параметров b_0 и b_1 для следующей регрессионной модели:

$$y_i = b_0 x_i + b_1(1 - x_i) + u_i \quad (i = 1,2,\ldots,n),$$
$$x_i = 1\ (i = 1,2,\ldots,k),$$
$$x_i = 0\ (i = k+1,\ k+2,\ldots,2k = n),$$
$\therefore$ при этом

$$\begin{aligned} y_1 &= b_0 + u_1 & y_{k+1} &= b_1 + u_{k+1} \\ y_2 &= b_0 + u_2 & y_{k+2} &= b_1 + u_{k+1} \\ &\vdots & &\vdots \\ y_k &= b_0 + u_k & y_n &= b_1 + u_n. \end{aligned}$$

20. Линейная модель для группы наблюдений. Охарактеризуйте следующую исходную модель:

$$y_{ij} = \mu_i + u_{ij} \quad (i = 1,2,\ldots,n,\ j = 1,2,\ldots,J_i).$$

21. Необходимость введения ограничений — трехмерный случай. Следует проанализировать матрицу данных для линейной модели с расширенной записью стохастической спецификации (ошибки модели):

$$y_{ijk} = \mu + a_i + \beta_i + \gamma_k + \delta_{ij} + \varepsilon_{ik} + \eta_{jk} + \psi_{ijk}.$$

При этом имеет силу

$$\sum_i a_i = 0,\ \sum_j \beta_j = 0,\ \sum_k \gamma_k = 0,\ \sum_i \delta_{ij} = \sum_j \delta_{ij} = 0,\ \sum_i \varepsilon_{ik} = \sum_k \varepsilon_{ik} = 0,$$

$$\sum_j \eta_{jk} = \sum_k \eta_{jk} = 0 \text{ и } \sum_i \psi_{ijk} = \sum_j \psi_{ijk} = \sum_k \psi_{ijk} = 0.$$

В уравнениях все величины, обозначенные греческими буквами, — это параметры, подлежащие оцениванию. Величина y — наблюдаемая. Необходимость указанных ограничений следует из единичности неограниченной модели, данные ограничения являются достаточными для приведенной нормальной модели.

Указание: примите для рассмотрения упрощенную модель, например, когда $|i| = |j| = |k| = 2$. Для этого случая имеем:

ijk	μ	α_i		β_i		γ_i		δ_{ij}				ε_{ik}				η_{jk}			
		α_1	α_2	β_1	β_2	γ_1	γ_2	δ_1	δ_1	δ_2	δ_2	ε_1	ε_1	ε_2	ε_{22}	η_1	η_1	η_2	η_2
111	1	1		1		1		1				1				1			
211	1		1	1		1				1				1		1			
121	1	1			1	1			1			1						1	
221	1		1		1	1					1			1				1	
112	1	1		1			1	1					1				1		
212	1		1	1			1			1					1	1			
122	1	1			1		1		1				1						1
222	1		1		1		1				1				1				1

22. Линейное оценивание нелинейных регрессионных связей. Обозначим u случайную величину (ошибку модели); символами a, b и c — параметры регрессионной модели, Y и X — переменные.

а) Покажите, что представленные ниже связи Y и X описываются кривой Гомпертца, или логистической функцией:

1) $Y = uab^{c^X}$,

2) $\frac{1}{Y} = a + bc^X + u$.

б) Покажите, что для обеих функций применимы приемы линейного оценивания.

23. Линейное оценивание нелинейных связей. Оцените параметры:

$\theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ для модели вида:

$$y_i = \theta_1 \exp\left(-\frac{x_i}{\theta_2}\right) + \theta_3 + \theta_4 x_i + u_i \quad (i = 1,2,\ldots,n).$$

Указание: в данном случае следует использовать итеративное оценивание.

24. Линейное оценивание нелинейных регрессионных связей.

а) Оцените параметры $\theta = \{\theta_1, \theta_2, \theta_3\}$ модели

$y_t = f(x_t;\theta) \oplus u_t \quad (t = 1,2,\ldots,n)$,

$\frac{1}{2} v_t^c f(x;\theta) = \frac{\theta_1\theta_3 x_1}{1+\theta_1 x_1 + \theta_2 x_2}$, $\theta' = (\theta_1 \quad \theta_2 \quad \theta_3)$, $\theta \in \boldsymbol{R}^3$.

б) Оцените параметры $\delta = \{\delta_0, \delta_1, \delta_2, \delta_3\}$ нелинейной (?) перепараметризованной модели:

$$y = f(x;\theta) = \frac{\delta_1 \cdot x_1}{\delta_0 + \delta_2 \cdot x_1 + \delta_3 x_2} \Leftrightarrow (b_0 + b_1 x_{1t} + b_2 x_{2t}) y_t = (\delta_1 \cdot x_t) \oplus u_t .$$

В решении задачи используйте данные пяти наблюдений:

y_t	x_{1t}	x_{2t}
0,126	1,0	1,0
0,219	2,0	1,0
0,076	1,0	2,0
0,126	2,0	2,0
0,186	0,1	0,0

25. Имеется модель:

$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_3 \exp(\theta_4 x_{3t}) + u_t \quad (i = 1,2,\ldots,n = 30)$,

$(x_{1t}, x_{2t}) = (1,1)\,(i = 1,2,\ldots,15)$, $(x_{1t}, x_{2t}) = (1,1)\,(i = 16,17,\ldots,30)$,

(x_{3t}) случайные числа $(i = 1,2,\ldots,30)$.

Каким образом вы оценили бы параметры этой модели $\{\theta_k, k = 1,2,\ldots,30\}$? Покажите решение соответственно заданным значениям l и n.

26. Имеется модель:

$y_t = (\alpha + x_t\beta)(\gamma + x_t\delta) + u_t \quad (t = 1,2,\ldots,n)$.

Обсудите возможные подходы к оцениванию данной функции.

1) Будут ли различаться приемы оценивания, если величина ошибки u вводится перед трансформацией модели, после нее?

2) Покажите подходы, которые позволяют имеющуюся модель непосредственно привести к линейному виду, и решающие эту же задачу, но после аппроксимации модели методом Тэйлора.

27. Калибровочная модель. Обсудите представленный ниже двумя уравнениями класс моделей и решите вопрос о возможности линейного оценивания на основе данных, приведенных в таблице.

Регрессионные модели:

$$y_t = \alpha + \frac{\beta}{1+\gamma x_t^\delta} + u_t \quad (t = 1,2,\ldots,n) \quad \text{и} \quad y_t = f(x_t) = \frac{1}{(\alpha + \beta x_t^\gamma + \delta x_t^2)^\varepsilon}.$$

Таблица содержит определенный набор вариантов для параметризации моделей, символом ! в ней обозначены данные параметров, не равные нулю.

α	β	γ	δ	ε
!	!	1	0	1
!	!	!	!	1
!	!	!	0	$1/\theta$
!	!	1	0	$1/\theta$
!	!	1	0	1

28. Обсудите в дискуссии стандартную модель логистической функции, а также возможность ее преобразования в линейную модель:

$$y_t = y_{t-1}\left(1 + a\left(1 - \frac{y_t}{x_t}\right)\right) \Leftrightarrow y = f(a),$$

здесь a — неизвестный параметр.

Ход ваших рассуждений покажите на конкретном примере с данными, выбранными произвольно. Это могут быть данные о биологических популяциях или о товарах длительного пользования (телевизорах, видеомагнитофонах, проигрывателях и т. д.), или о заболеваемости населения и т. п.

Указание: См. табл. 6.3, а также теоретический материал в § 6.6:

$$y = f(a) \Leftrightarrow \frac{y_t - y_{t-1}}{y_{t-1}} = a\left(1 - \frac{y_t}{x_t}\right).$$

29. Переменные, принимающие альтернативные значения 0 и 1. Используя логистическую регрессию, покажите возможность решения вопроса о вынесении смертного приговора для взломщика сейфов и похитителя большой суммы денег. Пусть p — вероятность вынесения смертного приговора и имеется по крайней мере две экзогенные величины с альтернативными значениями, представляющими первого обвиняемого $\{X_1' = (0,1)\}$ и второго обвиняемого $\{X_2' = (0,1)\}$.

30. Частная модель сбережений и инвестиций. Типичным и достаточно простым примером линейной модели является формальное описание регрессионной зависимости сумм инвестиционных вложений и сбережений (I — инвестиции, Y — доходы населения, S — сбережения):

$$\frac{I_t}{Y_t} = a + b\frac{S_t}{Y_t} + u_t \ (a \approx 0;\ b \approx 0{,}4).$$

Произведите оценивание такой функции по данным любой выбранной вами страны.

31. Произведите оценивание уровня безработицы u, используя понятие AR-процесса. Исходная модель отражает основной тренд для уровня безработицы:

$$u_t = \beta_1 + \beta_2 u_{t-1} + \beta_3 t + \beta_4 t^2 + v_t,\ v_t \approx N(0, \sigma^2).$$

32. Уравнение для заработной платы. Приведенное ниже регрессионное уравнение (нидерландское уравнение) представьте в матричной форме как линейную модель вида: $y = Xb + u$.

Нидерландское уравнение относительно отклика — уровня заработной платы по Ж.Ж. Граафланду и Ж.П. Вербруггену (J.J. Graafland and J.P. Verbruggen, Macro against sectoral wage equations for the Netherlands, Research Memorandum 103, Central Planning Bureau, The Hague 1993):

$$\log w_i = \delta_1 \log py_i + \delta_2 \log h_i + \delta_3 \log pc + \delta_4 \log h + \delta_5 \log(1 - \tau_i) +$$
$$+ \delta_6 \log(1 - \tau_e) + \delta_7 \log l_i + \delta_8 \log ur + \delta_9 \log rp_i + \delta_{10} \log \hat{w}_{wl} + \delta_{11},$$

$\delta_1, \delta_2, \delta_7 \ldots \delta_{10} > 0$, а $\delta_3 \ldots \delta_6$ по знаку + или − не определены,

δ_i, $i = 1,2,\ldots,11$ — параметры, подлежащие оцениванию;

здесь w_i — уровень заработной платы в определенном секторе экономики, py_i — цена за единицу продукта в секторе экономики, h_i — уровень производительности труда в секторе экономики, pc — уровень потребительских цен, h — уровень производительности труда на макроуровне; τ_i — ставка взносов работодателей на социальное страхование, τ_e — ставка налогов и взносов на социальное страхование для работающих по найму, l_i — уровень занятости в секторе экономики, ur — уровень безработицы на макроуровне, rp_i — флуктуационное соотношение для сектора экономики, $\hat{w}_{wl}$ — средняя ставка заработной платы в секторе экономики.

33. Постройте линейную регрессионную модель роста валового национального продукта в Германии за период 1815—1996 гг.

Указания:

1) В качестве исходной выберите регрессионную зависимость вида:

$$Y_t = Y_0(1 + g)^t ab.$$

2) С целью характеристики влияния на экономическое развитие нарушений в отраслевой структуре в 1871, 1914—1918, 1918—1923, 1930—1933, 1939—1945, 1948, 1990 гг. используйте соответствующие коэффициенты и формы матрицы исходных данных. Обобщите ваши рассуждения относительно модели для структурных нарушений в экономике.

3) Обсудите вопрос о том, каким образом могли бы быть получены необходимые для модельных расчетов исходные данные.

4) Письменно обоснуйте свои ответы.

Глава 7

Факторный анализ

7.1. Сущность методов факторного анализа и их классификация

В современной статистике под *факторным анализом* понимают совокупность методов, которые на основе реально существующих связей признаков (или объектов) позволяют выявлять латентные обобщающие характеристики организационной структуры и механизма развития изучаемых явлений и процессов.

Понятие латентности в определении ключевое. Оно означает неявность характеристик, раскрываемых при помощи методов факторного анализа. Вначале мы имеем дело с набором элементарных признаков X_j, их взаимодействие предполагает наличие определенных причин, особенных условий, т.е. существование некоторых скрытых факторов. Последние устанавливаются в результате обобщения элементарных признаков и выступают как интегрированные характеристики, или признаки, но более высокого уровня. Естественно, что коррелировать могут не только тривиальные признаки X_j, но и сами наблюдаемые объекты N_i, поэтому поиск латентных факторов теоретически возможен как по признаковым, так и по объектным данным. Рассмотрим несколько примеров.

Предположим, n наблюдаемых объектов (автомобилей) оценивается в двумерном признаковом пространстве R^2 с координатными осями: X_1 — стоимость автомобиля и X_2 — длительность рабочего ресурса мотора. При условии коррелированности X_1 и X_2 в системе координат появляется направленное и достаточно плотное скопление точек, формально отображаемое новыми осями (F_1 и F_2). Характерная особенность F_1 и F_2 заключается в том, что они проходят через плотные скопления точек и в свою

очередь коррелируют с X_1 и X_2. Максимальное число новых осей F_r будет равно числу элементарных признаков (рис. 7.1 а, б).

Допуская линейную зависимость F_r от X_j, можем записать:

$F_1 = a_1 x_1 + a_2 x_2$ и $F_2 = a_1 x_1 + a_2 x_2$.

Интерпретируем оси F_r: пусть F_1 — экономичность автомобиля, F_2 — его надежность в эксплуатации. Суждение об F_1 и F_2 базируется на оценке структуры латентных факторов, т.е. оценке весов X_1 и X_2 в F_r, а именно по значениям коэффициентов a_j.

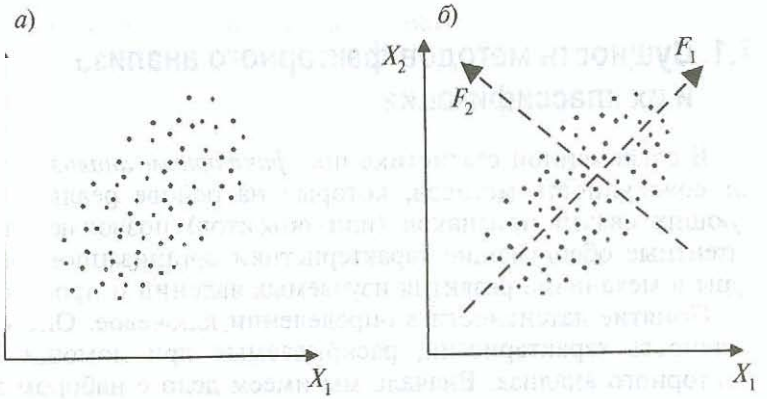


*Рис. 7.1. **Геометрическое представление n наблюдаемых объектов в тривиальном пространстве элементарных признаков (а) и латентных факторов (б)***

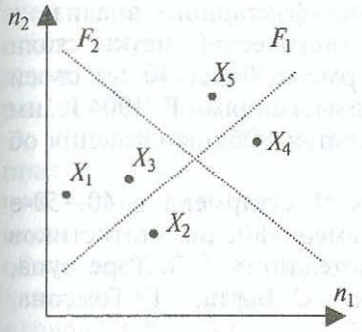
Если объекты характеризуются достаточно большим числом элементарных признаков ($m > 3$), то логично и другое предположение — о существовании плотных скоплений точек (признаков) в пространстве n объектов. При этом новые оси обобщают уже не признаки X_j, а объекты n_i, соответственно и латентные факторы F_r будут распознаны по составу наблюдаемых объектов:

$F_r = c_1 n_1 + c_2 n_2 + ... + c_N n_N$,

где c_i — вес объекта n_i в факторе F_r.

Гипотетически легко представить следствием такого анализа, скажем, выявление классифицирующих факторов: F_1 — промышленность, F_2 — сельское хозяйство и т. п.

На рис. 7.2 один из примеров распределения значений признаков (X_1 — стоимость автомобиля, X_2 — длительность рабочего ресурса мотора, X_3 — время набора максимальной скорости после старта, X_4 — количество потребляемого бензина на 100 км пути, X_5 — дальность тормозного пути) в координатном пространстве для двух видов автомобилей: n_1 — «Вольво» и n_2 — «Фольксваген».



Распределение признаков X_j в объектном пространстве R^n строится по координатным данным транспонированной исходной матрицы X_1:

	X_j	n_1	n_2	...	n_N
$X' =$	X_1	X_{11}	X_{21}	...	X_{N1}
	X_2	X_{12}	X_{22}	...	X_{N2}
	X_3	X_{13}	X_{23}	...	X_{N3}
	X_m	X_{1m}	X_{2m}	...	X_{Nm}

*Рис. 7.2. **Распределение элементарных признаков в координатном пространстве n_1 и n_2 — двух видов легковых автомобилей***

В зависимости от того, какой из рассмотренных выше тип корреляционной связи — элементарных признаков или наблюдаемых объектов — исследуется в факторном анализе, различают ***R*** и ***Q*** — технические приемы обработки данных.

Название ***R***-техники носит объемный анализ данных по m признакам, в результате него получают r линейных комбинаций (групп) признаков ($F_r = f(X_j)$; $r = 1 \div m$). Анализ по данным о близости (связи) n наблюдаемых объектов называется ***Q***-техникой и позволяет определять r линейных комбинаций (групп) объектов:

($F = f(n_i)$; $i = 1 \div N$).

В настоящее время на практике более 90% задач решается при помощи ***R***-техники.

Факторный анализ имеет сравнительно небольшую, но насыщенную историю. Исследователям из-за трудоемкости вычислительных процедур приходилось тратить годы не только на разработку самих аналитических методов, но и на демонстрацию их приложения.

Появления факторного анализа связывают с выходом в свет в 1901 г. статьи английского ученого К. Пирсона «Переход по линиям и плоскостям к визуализированным системам точек в про-

странстве», в которой была высказана идея построения главных осей. В сущности это была идея метода главных компонент, позволяющего полно отображать реальное координатное пространство в теоретическом пространстве, оси которого обобщают значения исходных элементарных признаков.

Определяющими для формирования факторного анализа в самостоятельный большой раздел статистической науки стали работы британского психолога Ч. Спирмена, более 40 лет своей жизни посвятившего теоретическим изысканиям. В 1904 г. им была опубликована фундаментальная статья «Общие сведения об объективных решениях и измерениях».

На основе теоретических посылок Ч. Спирмена в 40—50-е годы появились глубокие разработки американских статистиков и математиков: Л. Гуттмана, Г. Хотеллинга, Л. Тэрстоуна, К. Хользингера, С. Рао, английских: С. Барта, Г. Томсона, Д. Лоули, А. Максвелла и др.

В 60-е годы арсенал методов факторного анализа значительно пополнился новыми, оптимизирующими решениями Г. Кайзера, Р. Йорескога и др. В 1960 г. выходит труд американского ученого Г. Хармана «Современный факторный анализ» — уникальный по своей глубине и охвату разнообразных методов и, что важно, доступный для ознакомления самому широкому кругу читателей.

Толчком для развития методов факторного анализа изначально послужили задачи и проблемы из области психологии и почти все первые работы по вопросам факторного анализа были напечатаны в журналах: «Психология», «Психометрика» и т. п. Позже методы факторного анализа стали активно применяться в социологических исследованиях, медицине, затем в военной промышленности, экономике. Сегодня уже трудно назвать области научных исследований, где в той или иной мере эти методы не использовались бы. Непременным условием расширения сферы их приложения остается компьютеризация аналитической практики.

Набор методов факторного анализа в настоящее время достаточно велик, насчитывает десятки различных подходов и приемов обработки данных. Чтобы в исследованиях ориентироваться на правильный выбор методов, необходимо представлять их особенности.

Разделим все методы факторного анализа на несколько классификационных групп:

1. Метод главных компонент (Г. Хотеллинг). Строго говоря, его не относят к факторному анализу, хотя он имеет с ним много общего. Специфическим является, во-первых, то, что в ходе вычислительных процедур одновременно получают все главные компоненты и их число первоначально равно числу элементарных признаков; во-вторых, постулируется возможность полного разложения дисперсии элементарных признаков, другими словами, ее полное объяснение через латентные факторы (обобщенные признаки).

2. Методы факторного анализа. Дисперсия элементарных признаков здесь объясняется не в полном объеме, признается, что часть дисперсии остается нераспознанной как характерность. Факторы обычно выделяются последовательно: первый, объясняющий наибольшую долю вариации элементарных признаков, затем второй, объясняющий меньшую, вторую после первого латентного фактора часть дисперсии, третий и т.д. Процесс выделения факторов может быть прерван на любом шаге, если принято решение о достаточности доли объясненной дисперсии элементарных признаков или с учетом интерпретируемости латентных факторов.

Методы факторного анализа целесообразно разделить дополнительно на два класса: упрощенные и современные аппроксимирующие методы.

Простые методы факторного анализа в основном связаны с начальными теоретическими разработками. Они имеют ограниченные возможности в выделении латентных факторов и аппроксимации факторных решений. В числе этих методов следует назвать:

- *однофакторную модель* Ч. Спирмена. Она позволяет выделить только один генеральный латентный и один характерный факторы. Для возможно существующих других латентных факторов делается предположение об их незначимости;
- *бифакторную модель* Г. Хользингера. Допускает влияние на вариацию элементарных признаков не одного, а нескольких латентных факторов (обычно двух) и одного характерного фактора;
- *центроидный метод* Л. Тэрстоуна. В нем корреляции между переменными рассматриваются как пучок векторов, а латентный фактор геометрически представляется как уравно-

вешивающий вектор, проходящий через центр этого пучка. Метод позволяет выделять несколько латентных и характерные факторы, впервые появляется возможность соотносить факторное решение с исходными данными, т.е. в простейшем виде решать задачу аппроксимации.

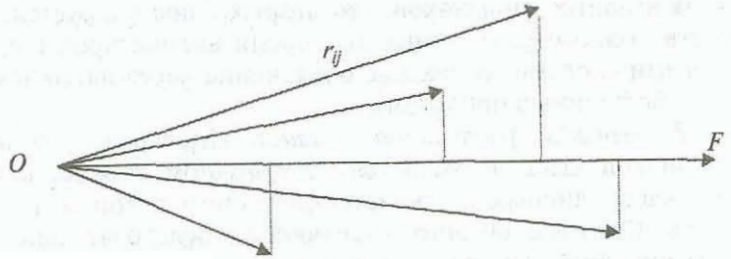


*Рис. 7.3. **Геометрическое представление парных корреляций элементарных признаков и латентного фактора в центроидном методе***

Современные аппроксимирующие методы часто предполагают, что первое, приближенное решение уже найдено каким-либо из способов, последующими шагами это решение оптимизируется. Методы отличаются сложностью вычислений. К этим методам относятся:

- *групповой метод* Л. Гуттмана и П. Хорста. Решение базируется на предварительно отобранных каким-либо образом группах элементарных признаков;
- *метод главных факторов* Г. Томсона. Наиболее близок методу главных компонент, отличие заключается в предположении о существовании характерностей;
- *метод максимального правдоподобия* (Д. Лоули), *минимальных остатков* (Г. Харман), *α-факторного анализа* (Г. Кайзер и И. Кэффри), *канонического факторного анализа* (К. Рао), все *оптимизирующие*. Позволяют последовательно улучшить предварительно найденные решения на основе использования статистических приемов оценивания случайной величины или статистических критериев, предполагают большой объем трудоемких вычислений. Наиболее перспективным и удобным для работы в этой группе признается метод максимального правдоподобия.

На рис. 7.4 представлена схематично общая классификация методов факторного анализа.

*Рис. 7.4. **Общая классификация методов факторного анализа***

Основной задачей, которую решают разнообразными методами факторного анализа, включая и метод главных компонент, является сжатие информации, переход от множества значений по m элементарным признакам с объемом информации $n \times m$ к ограниченному множеству элементов матрицы факторного отображения ($m \times r$) или матрицы значений латентных факторов для каждого наблюдаемого объекта размерностью $n \times r$, причем обычно $r < m$.

Методы факторного анализа позволяют также визуализировать структуру изучаемых явлений и процессов, а это значит определять их состояние и прогнозировать развитие. Наконец, данные факторного анализа дают основания для идентификации объекта, т.е. решения задачи распознавания образа.

Методы факторного анализа обладают свойствами, весьма привлекательными для их использования в составе других статистических методов, наиболее часто в корреляционно-регрессионном анализе, кластерном анализе, многомерном шкалировании и др.

7.2. Фундаментальная теорема факторного анализа Тэрстоуна

Изучение факторных воздействий предполагает выявление взаимосвязей характерных признаков. Для многомерных объектов показателями связи являются оценки дисперсии и коэффициенты ковариации, которые обобщаются в матрице ковариаций Σ (по выборочным данным — матрица S). Когда исходные значения признаков нормированы, т.е. имеем $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$, матрица ковариаций, как известно, переходит в матрицу парных корреляций: $S = R = \frac{1}{n} Z'Z$.

Симметрическая матрица R имеет собственную систему координат в пространстве R^m, где m — число анализируемых признаков. Допуская преобразования координатной системы в систему пространства латентных факторов, можно записать Z_{ij} в виде линейной комбинации новых координат:

$$Z_{ij} = a_{j1} f_{1i} + a_{j2} f_{2i} + \ldots + a_{jr} f_{ri},$$

или в матричной форме: $Z = AF$.

Воспользуемся возможностью подстановки в уравнение для R вместо Z произведения матриц AF и получим

$$R = \frac{1}{n} AF\,(AF)' = \frac{1}{n} AFF'A'.$$

Изменив место расположения скаляра $1/n$, выделим произведение $\frac{1}{n} FF'$, результат произведения интерпретируется как матрица корреляций C, определяемая для латентных факторов F_r. После замены $1/n\ FF'$ на C запишем: $R = ACA'$.

В предположении, что факторы F_r некоррелированы, т.е. $C = E$, где E — единичная матрица, приходим к равенству: $R = AA'$.

Л.Л. Тэрстоуном равенства типа: $R = ACA'$ и $R = AA'$ названы *фундаментальной факторной теоремой*, A — здесь матрица факторного отображения, а ее элементы a_{jr} — величины факторных нагрузок. Суть теоремы — в возможности воспроизведения исходной корреляционной матрицы R через матрицу факторного отображения A. При $C = E$ связь матричных элементов r и a записывается в виде уравнения: $r_{ij} = a_{i1}\, a_{j1} + a_{i2}\, a_{j2} + \ldots + a_{ir}\, a_{jr}$.

Другими словами, корреляция пары характерных признаков r_{ij} опосредуется корреляцией каждого из признаков с некоторыми латентными факторами F_r. Латентные факторы определяют само существование связи i-го и j-го коррелирующих признаков. Если $C = E$, латентные факторы неортогональны и матрица корреляций R отображается в A с учетом их взаимодействия:

$$R = A \times C \times A'$$

$$\begin{pmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{21} & 1 & \ldots & r_{2m} \\ r_{m1} & r_{m2} & \ldots & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1r} \\ a_{21} & a_{22} & \ldots & a_{2r} \\ a_{m1} & a_{m2} & \ldots & a_{mr} \end{pmatrix} \cdot \begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1r} \\ c_{21} & c_{22} & \ldots & c_{2r} \\ c_{r1} & c_{r2} & \ldots & c_{rr} \end{pmatrix} \cdot \begin{pmatrix} a_{11} & a_{21} & \ldots & a_{m1} \\ a_{12} & a_{22} & \ldots & a_{m2} \\ a_{1r} & a_{2r} & \ldots & a_{mr} \end{pmatrix}.$$

Равенства Тэрстоуна допускаются гипотетически. Реально AA' и ACA' будут далеко не всегда в точности воспроизводить R. По крайней мере это объясняется двумя причинами. Во-первых, в факторном анализе, позволяющем эффективно объяснять общую дисперсию данных, r — число латентных (обобщенных) признаков, как правило, значительно меньше числа исходных признаков m. И, во-вторых, в матрице A объединяются теоретические оценки факторных нагрузок $\hat{a}_{ij}$. С учетом различий математических методов и специфичности вычислительных процедур следует допустить, что $\hat{a}_{ij}$ не абсолютно истинны.

Таким образом, можно ожидать, что воспроизведенная из AA' или ACA' матрица корреляций R^+ будет отлична от R. Как следствие, на главной диагонали R^+ располагаются величины, обычно не равные, а меньшие единицы. На практике значения r^+_{ij} принимают за общности, т.е. характеристики части дисперсии, поддавшейся объяснению через латентные факторы F_r, а $1 - a^+_{ij}$ — специфичность, т.е. необъясненная часть дисперсии. По степени расхождения R^+ и R судят о достаточности числа выделенных латентных факторов и адекватности аналитических выводов.

7.3. Общий алгоритм и теоретические проблемы факторного анализа

Методы факторного анализа при всем их многообразии имеют общий алгоритм решения, представленный на рис. 7.5. Начинаясь построением матрицы исходных данных X, этот алго-

ритм завершается получением матриц факторного отображения и значений факторов A и F. С учетом принятых обозначений, где n — число наблюдений, m — число аналитических признаков X, r — число значимых обобщенных признаков (латентных факторов), на схеме показана размерность матриц данных для каждого алгоритмического шага.

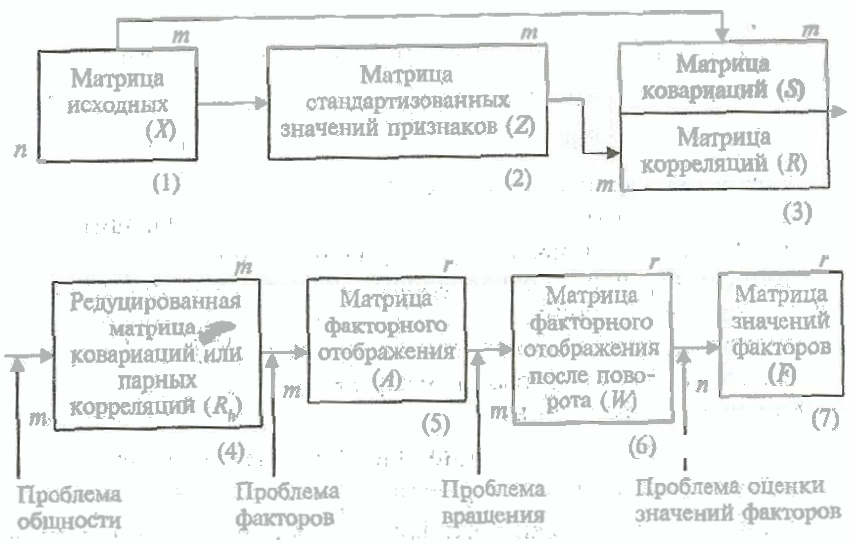


Рис. 7.5. Алгоритмическая схема реализации методов факторного анализа

Первые шаги алгоритма 1—3 не вызывают каких-либо затруднений. Переход от матрицы исходных данных X к матрице стандартизованных данных Z осуществляется после пересчета всех элементов x_{ij} по формуле: $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$.

Как известно, для стандартизованных значений z_{ij} математическое ожидание равно нулю ($E(z) = 0$) и дисперсия $D(z) = 1$.

На следующем шаге простым перемножением скаляра $\frac{1}{n}$ и матриц Z' и Z получаем матрицу парных корреляций: $R = \frac{1}{n} Z'Z$. Шаг 2 может быть опущен и тогда последующее факторное ре-

шение находят не по матрице R, а по матрице ковариаций: $S = \frac{1}{n} X'X$, в последнем случае желательно, чтобы анализируемые признаки X имели одни и те же единицы измерения[1].

Выполнение четвертого шага алгоритма обусловливается решением первой проблемы — построения редуцированной матрицы корреляций (ковариаций). Проблема актуальна именно для методов факторного анализа, так как в методе главных компонент принимается, что всю вариацию исходных признаков полностью объясняют латентные факторы, при этом матрица парных корреляций R размерности $m \times m$ остается и как редуцированная R_h, в которой все общности $\sum h_j^2 = 1$:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} \\ r_{21} & 1 & r_{23} & \dots & r_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \dots & 1 \end{pmatrix}.$$

В факторном анализе матрица корреляций R преобразуется в R_h:

$$R_h = \begin{pmatrix} h_1^2 & r_{12} & r_{13} & \dots & r_{1m} \\ r_{21} & h_2^2 & r_{23} & \dots & r_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \dots & h_m^2 \end{pmatrix}$$

с <1, т.е. вариация признаков (X_j, $j = \overline{1,m}$) может быть объяснена не на 100%, а несколько меньше, с учетом существования их нераскрываемой характерности.

Существуют достаточно простые методы поиска общностей h_j^2:

- *метод наибольшей корреляции.* На главной диагонали с положительным знаком записывается наибольший по величине коэффициент корреляции;

[1] Смещенность дисперсионных оценок авторами принимается несущественной, ее учет в данном случае может вызывать нарушения элементарной логики оценок взаимосвязей признаков, например, на главной диагонали матрицы R появляются величины, большие единицы.

- *метод Барта.* По каждому столбцу матрицы R вначале находят среднее значение коэффициентов корреляции $\bar{r}_j$, затем, если $\bar{r}_j$ сравнительно велико, за общность принимается значение, которое несколько выше наибольшего в столбце коэффициента корреляции и, если $\bar{r}_j$ — сравнительно малое значение, общность будет несколько меньше наибольшего в столбце коэффициента корреляции;
- *метод триад.* Общности для каждого j-го столбца R вычисляют по формуле: $n_j^2 = \frac{r_{ik} r_{il}}{r_{kl}}$, где r_{ik} и r_{il} — коэффициенты корреляции, наибольшие в столбце;
- *метод малого центроида.* Для каждой переменной j строится корреляционная матрица размерности 4 × 4. Включая саму переменную в эту матрицу, записывают оценки корреляции трех других переменных, особенно тесно связанных с первой. По данным малой матрицы корреляций и рассчитывают общности:

$$h_j^2 = \frac{\left(\sum_i r_{i1}\right)^2}{\sum_{ij} r_{ij}},$$

где $\sum_i r_{i1}$ — сумма элементов первого столбца; $\sum_{i,j} r_{ij}$ — сумма всех элементов матрицы 4 x 4.

Вторая проблема возникает на этапе построения матрицы отображения A и заключается в выборе оптимального метода для поиска весовых коэффициентов a_{ij} элементов матрицы A. Напомним, что наилучшие решения обычно находят при помощи современных методов факторного анализа: главных факторов, максимального правдоподобия и др. В общем случае выделенные факторы не обязательно ортогональны ($R = ACA'$) и тогда векторы (столбцы) матрицы A будут линейно-зависимыми.

Выполнение шага 6 алгоритма и решение проблемы вращения пространства общих факторов не обязательно. Потребность в этом возникает, когда пространственное расположение общих факторов F_r нелогично или трудно поддается интерпретации.

Возможность появления алогичных первых результатов анализа объясняется не определяемым четко и не задаваемым положением факторных осей в пространстве, или грубо говоря, отсутствием изначально какой-либо пространственной привязки для осей F_r.

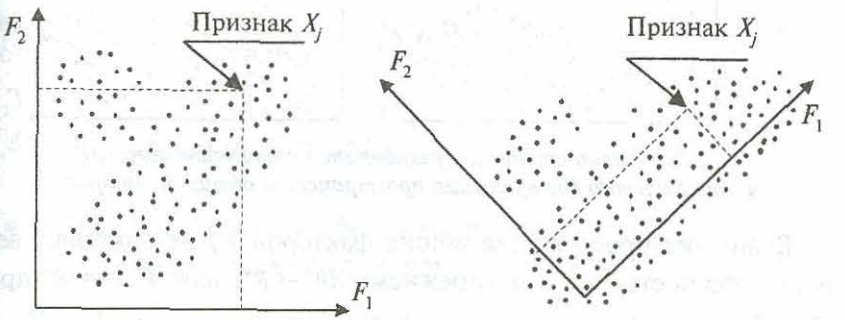


Рис. 7.6. Исходные признаки X_j в пространстве общих факторов F_1, F_2

На рис. 7.6 показаны два различных положения в пространстве факторных осей (F_1 и F_2). Легко заметить, что изменение положения F_1, F_2 одновременно приводит к изменению координат исходных признаков X_j. Цель поворота — преобразование координат (факторных нагрузок) таким образом, чтобы факторобразующие признаки имели наибольшие нагрузки, близкие к единице ($|a_{ir}| \to 1$), а остальные признаки — минимальные значения, близкие к нулю, т.е. добиваются экономичного описания данных.

Повороты осей могут быть ортогональными и косоугольными (рис. 7.7). Предпочтительно, хотя и более трудно выполнимо и интерпретируемо, косоугольное вращение, при этом, как видно из рис. 7.7б, значительно повышаются возможности оптимального отображения сгущений признаков в пространстве R^F.

На рис. 7.7а ось F' после поворота F, очевидно, займет более рациональное положение, но из-за жесткости осевой конструкции положение F_2 удаляется от оптимального; на рис. 7.7б косоугольным вращением ($\alpha \neq 90°$) приходят к оптимизации положения сразу обеих осей F_1 и F_2.

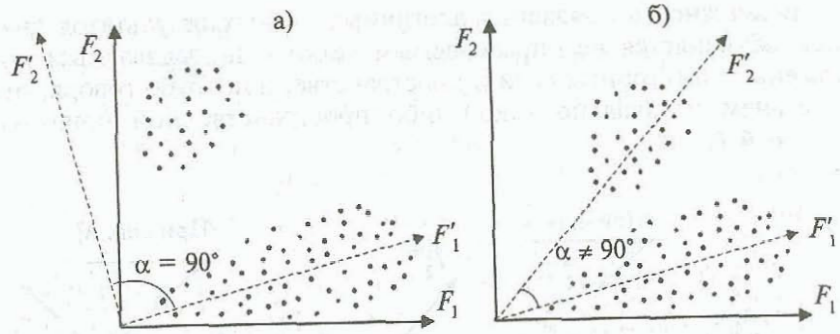


Рис. 7.7. Гипотетические результаты ортогонального (а) и косоугольного (б) вращения пространства общих факторов

Вращение пространства общих факторов F_r) не изменяет величин общностей h_j^2 и по-прежнему $AA' = R^+$, или $ACA' = R^+$ при $R^+ \to R$.

На заключительном этапе алгоритма рассчитывают матрицу значений факторов F, ее элементы — это факторные значения f_{ir} для каждой единицы наблюдения. Тем самым определяем положение n объектов в пространстве R^F с r — числом факторных осей (рис. 7.8).

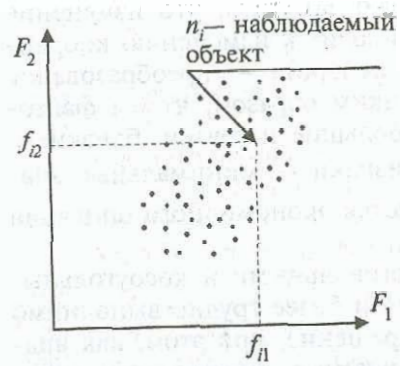


Рис. 7.8. Отображение наблюдаемых объектов n_i в пространстве двух общих факторов F_1 и F_2

Так как число общих факторов F_r, как правило, значительно меньше числа исходных признаков X_j, матрица F имеет размерность $n \times r$ в отличие от исходной матрицы X размерностью $n \times m$. Проблема заключается в выборе методики перехода от матрицы исходных данных Z (или X) по известной матрице факторного отображения A к матрице F, ее решают обычно одним из двух методов: вращением пространства исходных признаков (R^2), а значит посредством алгебраического произведения отображающих матриц $Z(X)$ и A, т.е. $F = f(Z, A)$, где f — функция линейной функциональной формы связи, во-вторых, F находят с помощью

множественного регрессионного анализа: $F = (B, Z, R, A) + \xi$ (линейная стохастическая форма связи), где B — матрица коэффициентов регрессии размерностью $r \times m$, ее элементы β_{jr} — регрессионные коэффициенты каждого из факторов F_r по переменной X_j.

Алгоритмы факторного анализа отличаются, как видим, трудоемкостью, их полное выполнение возможно при условии использования технических средств.

7.4. Метод главных компонент

7.4.1. Общая математическая модель метода главных компонент

Из числа методов, позволяющих обобщать значения элементарных признаков, метод главных компонент выделяется простой логической конструкцией и в то же время на его примере становятся понятными общая идея и целевые установки многочисленных методов факторного анализа.

Метод главных компонент дает возможность по m — числу исходных признаков выделить m главных компонент, или обобщенных признаков. Пространство главных компонент ортогонально.

Математическая модель главных компонент базируется на логичном допущении, что значения множества взаимосвязанных признаков порождают некоторый общий результат. Предположив линейную форму связи признаков X_j, запишем в матричной форме уравнение зависимости результата F от X: $F = XB$, где B — вектор параметрических значений линейного уравнения связи. Условием выполнения такого равенства является соответствие дисперсий, т.е. $D(X) = D(XB)$. Поскольку X — многомерная случайная величина, ее дисперсионная оценка — это ковариационная матрица S. Постоянная величина B выносится за знак дисперсии и возводится в квадрат, получаем: $D(F) = B'SB$.

Поиск главных компонент сводится к задаче последовательного выделения первой главной компоненты F_1, обладающей максимальной дисперсией, второй главной компоненты, имею-

щей вторую по величине дисперсию, и т.д. Подобная задача имеет решение при условии введения ограничений. Пусть

$B'B = b_1^2 + b_2^2 + ... + b_m^2 = 1$.

При $B'B = 1$ максимизируем $B'SB$, используя метод множителей Лагранжа:

$r = B'SB - \lambda(B'B - 1)$ и $\frac{\partial r}{\partial B} = 2SB - 2\lambda B = 0$, откуда $SB - \lambda B = 0$.

Следовательно, получим $|S - \lambda E|B = 0$ и характеристическое уравнение для поиска λ_j будет: $|S - \lambda E| = 0$.

Из множества значений характеристических чисел λ_j относительно первого, наибольшего λ_1 находим вектор B_1 значений для первой главной компоненты F_1, для второго по величине характеристического числа λ_2 — вектор значений второй компоненты B_2 и т.д. до λ_m и B_m для F_m при m — исходном числе анализируемых признаков. Здесь B — векторы величин, представляющих координаты главных компонент F_r в пространстве признаков R^X, они же характеристики силы связи r-й главной компоненты и j-го признака X_j.

Если исходную матрицу данных X предварительно стандартизировать, то матрица ковариаций S перейдет в матрицу парных корреляций R, и вектор B будет собственным вектором по стандартизованным данным U. Решающее уравнение в матричной форме принимает вид: $(R - \lambda E)U = 0$.

Результаты применения метода главных компонент представляются данными матрицы отображения A. Возможна итоговая запись зависимости значений исходных признаков от значений главных компонент:

$$Z = AF' \quad \text{или} \quad z_{ij} = a_{j1}f_{i1} + a_{j2}f_{2i} + ... + a_{jr}f_{ri} \qquad (7.1)$$

либо зависимости значений главных компонент от значений элементарных признаков

$$F = A^{-1}Z', \text{ или } f_{ri} = \frac{1}{\lambda_r}(a_{1r}z_{i1} + a_{2r}z_{i2} + ... + a_{mr}z_{im}). \qquad (7.2)$$

В уравнениях (7.1) и (7.2) приняты обозначения:

z_{ij} — значение j стандартизованной переменной по i-му объекту наблюдения;

f_{ri} — r-я главная компонента F_r по i-му объекту наблюдения;

a_{jr} — весовой коэффициент r-й главной компоненты для j-й переменной, оценка частного коэффициента корреляции для F_r и X_j (элементы j-й строки матрицы A);

a_{mr} — весовые коэффициенты (характеристики силы связи) m элементарн признаков ($j = \overline{1,m}$) для r-й главной компоненты.

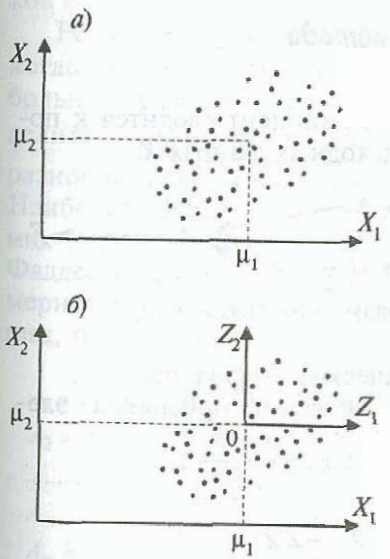
Уравнения (7.2) относительно F (главных компонент) являются производными от (7.1). Покажем это.

Известно, что алгебраически $A = V\Lambda^{1/2}$: умножим обе части матричного уравнения слева на $(V\Lambda^{1/2})$ и легко убедимся, что $A'A = \Lambda$. Далее, имеем $Z = AF$, умножим обе части этого уравнения на A', затем на $(A'A)^{-1}$ и получим $F = (A'A)^{-1}A'Z$, или $F = \Lambda^{-1}A'Z$, т.е.

$$F_r = \frac{1}{\lambda_r}(a_{1r}z_{i1} + a_{2r}z_{i2} + ... + a_{mr}z_{im}),$$

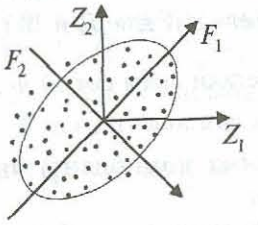
где $a_{1r}, a_{2r}, ..., a_{mr}$ — элементы r-го столбца для r-й главной компоненты матрицы факторного отображения A.

В упрощенном виде, для двумерной случайной величины, процедуру выделения главных компонент можно показать геометрически (рис. 7.9).



а) Первоначально имеется некоторое эмпирическое распределение данных в двумерном признаковом пространстве с центром (μ_1; μ_2)

б) Центрированием и стандартизацией данных исходное пространство признаков сжимается и система координат переносится в центр распределения данных



в) Решением матричного уравнения $(R-\lambda E)V=0$ находят параметры эллипса, описывающего эмпирическое распределение объектов в нормированном признаковом пространстве R^Z, соответственно устанавливается положение главных компонент (осей), обобщающих вариацию признаков Z_1 и Z_2

*Рис. 7.9. **Геометрическая интерпретация алгоритма метода главных компонент***

На рисунке видно, что задача выделения главных компонент сводится к поэтапному решению классических вопросов аналитической геометрии: изменению масштаба пространства, повороту координатной системы, координатному отображению векторов в старой системе координат и новой, после поворота. Рис. 7.9в позволяет видеть возможность отображения Z в F, и наоборот, F в Z, что записано несколько выше в виде функциональных линейных уравнений связи (7.1) и (7.2). Первоначально число главных компонент равно числу исходных элементарных признаков m.

7.4.2. Вычислительные процедуры метода главных компонент

Решение задачи методом главных компонент сводится к поэтапному преобразованию матрицы исходных данных X:

где X — матрица исходных данных, размерностью $n \times m$;
n — число объектов наблюдения;
m — число элементарных аналитических признаков;
Z — матрица стандартизованных значений признаков, элементы матрицы вычисляют по формуле: $z_{ij}=\dfrac{x_{ij}-\bar{x}_j}{\sigma_j}$;

R — матрица парных корреляций: $R=\dfrac{1}{n}Z'Z$.

Если предварительная стандартизация данных не проводилась, то на данном шаге получают матрицу $S=\dfrac{1}{n}X'X$, элементы матрицы X для расчета S будут центрированными величинами: $x_{ij}=x_{ij}-\bar{x}_j$;

Λ — диагональная матрица собственных (характеристических) чисел:

$$\Lambda=\begin{pmatrix}\lambda_1 & 0 & 0 & \dots & 0\\ 0 & \lambda_2 & 0 & \dots & 0\\ 0 & 0 & \lambda_3 & \dots & 0\\ 0 & 0 & 0 & \dots & \lambda_m\end{pmatrix}.$$

Множество значений λ_j находят решением характеристического уравнения $|R-\lambda E|=0$. λ_j — это характеристики вариации, точнее, показатели дисперсии каждой главной компоненты. Суммарное значение $\sum\lambda_j$ равняется сумме дисперсий элементарных признаков X_j. При условии стандартизации исходных данных, когда $D(z_{ij})=1$, $\sum\lambda_j$ равно числу элементарных признаков m.

Решение характеристического уравнения относительно λ, когда число признаков m достаточно велико и матрица R большой размерности, вызывает трудности при расчете определителя $|R|$. Они успешно преодолеваются с применением разнообразных математических методов матричной алгебры. Наиболее эффективен и легко поддается алгоритмизации среди них метод, базирующийся на рекуррентных соотношениях Фаддеева. Если A — некоторая симметрическая матрица размерностью $m \times m$, то ее определитель находится по следу матриц, производных из A:

$A_1=A$	$P_1=t_r A_1$	$B_1=A_1-P_1E$
$A_2=AB_1$	$P_2=1/2\,t_r A_2$	$B_2=A_2-P_2E$
.....	.....	.....
$A_{m-1}=AB_{m-2}$	$P_{m-1}=1/(m-1)\,t_r A_{m-1}$	$B_{m-1}=A_{m-1}-P_{m-1}E$
$A_m=AB_{m-1}$	$P_m=1/m\,t_r A_m$	$B_m=A_m-P_mE,\ B_m=0.$

На заключительном этапе расчетов P_m и есть определитель матрицы A ($P_m = |A|$). Для проверки вычислений может использоваться условие: $B_m = 0$.

После вычислений рекуррентных соотношений записывается характеристический многочлен: $P_m(\lambda) = \lambda^m - P_1\lambda^{m-1} - P_2\lambda^{m-2} - \ldots - P_m$.

Значения λ находят после того, как характеристический многочлен приравнивают нулю, получают характеристическое уравнение и решают его относительно характеристических корней λ_j:

V — матрица нормированных собственных (характеристических) векторов. Число векторов V_j первоначально равно m, т.е. $j = \overline{1,m}$. Получают V_j преобразованием ненормированных собственных векторов U:

$$V_j = \frac{U_j}{|U_j|},$$

где $|U_j|$ — норма вектора U, т.е. $|U_j| = \left(u_{1j}^2 + u_{2j}^2 + \ldots + u_{mj}^2\right)^{1/2}$.

Необходимость повторного, после получения матрицы Z, нормирования пространства теперь уже обобщенных признаков $\boldsymbol{R}^F$ объясняется механическим появлением в ходе предыдущих расчетов результатов, искажающих нормированное пространство.

В свою очередь собственные векторы U_j находят из матричного уравнения: $(R - \lambda E)U = 0$. Реально это означает решение m систем линейных уравнений для каждого λ_j при $j = \overline{1,m}$. В общем виде система уравнений имеет вид:

$$\begin{aligned}
&(1-\lambda_j)u_{1j} + r_{12}u_{2j} + r_{13}u_{3j} + \ldots + r_{1m}u_{mj} = 0,\\
&r_{21}u_{1j} + (1-\lambda_j)u_{2j} + r_{23}u_{3j} + \ldots + r_{2m}u_{mj} = 0\\
&r_{31}u_{1j} + r_{32}u_{2j} + (1-\lambda_j)u_{3j} + \ldots + r_{3m}u_{mj} = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&r_{m1}u_{1j} + r_{m2}u_{2j} + r_{m3}u_{3j} + \ldots + (1-\lambda_j)u_{mj} = 0.
\end{aligned}$$

Приведенная система объединяет однородные линейные уравнения, и так как число ее уравнений равняется числу неизвестных u_{mj}, имеет бесконечное множество решений. Конкретные значения собственных векторов при этом можно найти, задавая произвольно по крайней мере величину одной компоненты каждого вектора и обычно, чтобы не усложнять расчетов, ее приравнивают единице;

A — матрица факторного отображения, ее элементы a_{ij} — весовые коэффициенты. Вначале A имеет размерность $m \times m$ — по числу элементарных признаков X_j, затем в анализе остается r наиболее значащих компонент, $r \le m$. Вычисляют матрицу A по известным данным матрицы собственных чисел Λ и нормированных собственных векторов V по формуле $A = V\Lambda^{1/2}$;

F — матрица значений главных компонент размерностью $r \times n$, $F = A^{-1}Z'$ или $F = \Lambda^{-1}A'Z'$, или $F = \Lambda^{-1/2}V'Z'$.

Матрица F в общем виде записывается:

	n_1	n_2	…	n_n
F_1	f_{11}	f_{12}	…	f_{1n}
F_2	f_{21}	f_{22}	…	f_{2n}
…	…	…	…	…
F_3	f_{r1}	f_{r2}	…	f_m

Объект

Главная компонента

Пример 7.1. Совокупность из четырех промышленных предприятий оценена по трем характерным признакам: выработке на одного среднегодового работника X_1, уровню рентабельности X_2 и уровню фондоотдачи X_3.

В результате предварительных аналитических расчетов по исходным данным X получена матрица парных корреляций:

$$R = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}.$$

Используя алгоритм метода главных компонент, найдем собственные числа и собственные векторы матрицы R и построим матрицы с аналитическими результатами (A и F):

1. По рекуррентным соотношениям Фаддеева исчислим определитель матрицы парных корреляций $|R|$.

Первый шаг: $R = A$ и $A = A_1$, тогда $P_1 = t_r A_1 = 1 + 1 + 1 = 3$,

$$B_1 = A_1 - P_1 E = \begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & -2 & 0{,}439 \\ 0{,}154 & 0{,}439 & -2 \end{pmatrix}.$$

Второй шаг:

$$= AB_1 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix} \cdot \begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & -2 & 0{,}439 \\ 0{,}154 & 0{,}439 & -2 \end{pmatrix} =$$

$$= \begin{pmatrix} -1{,}638 & -0{,}513 & 0{,}101 \\ -0{,}513 & -1{,}469 & -0{,}350 \\ 0{,}101 & -0{,}350 & -1{,}783 \end{pmatrix};$$

$$P_2 = (1/2) t_r A_2 = (1/2)(-1{,}638) + (-1{,}469) + (-1{,}783) = -2{,}445,$$

$$B_2 = A_2 - P_2 E = \begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix}.$$

Третий шаг:

$$A_3 = AB_2 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}524 & 0 & 0 \\ 0 & 0{,}524 & 0 \\ 0 & 0 & 0{,}524 \end{pmatrix};$$

$P_3 = 1/3\ (3 \cdot 0{,}524) = 0{,}524, \quad B_3 = A_3 - P_3 E = 0.$

В итоге $|R| = 0{,}524$ и $B_3 = 0$. Обратим внимание, что в ходе расчетов все промежуточные матрицы A_j и B_j — симметрические.

2. Построим характеристическое уравнение:

$$\lambda^3 - 3\lambda^2 + 2{,}445\lambda - 0{,}524 = 0,$$

откуда $\lambda_1 = 1{,}798$; $\lambda_2 = 0{,}875$ и $\lambda_3 = 0{,}327$.

Таким образом, наши исходные элементарные признаки X_1, X_2, X_3 могут быть обобщены значениями трех главных компонент, причем первая главная компонента F_1 объяснит примерно

60% всей вариации X_j (1,798/3 = 0,599), вторая главная компонента F_2 объяснит 29,2% — меньшую часть по сравнению с F_1 общей дисперсии (0,875/3 = 0,292), наконец, третья главная компонента F_3 охватывает оставшуюся, еще не объясненную вариацию входных признаков — 10,9% (0,327/3= 0,109). Все главные компоненты F_1, F_2, F_3 объясняют вариацию X_1, X_2, X_3 полностью, на 100% (59,9 + 29,2 + 10,9).

Собственные векторы матрицы парных корреляций R найдем решением трех систем линейных уравнений соответственно для $\lambda_1 = 1{,}798$; $\lambda_2 = 0{,}875$ и $\lambda_3 = 0{,}327$.

Для определения области решений в каждой системе будем задавать одному из неизвестных признаков u_{3j} значение, равное единице.

Первая система уравнений, для $\lambda_1 = 1{,}798$:

$$\begin{array}{l|l} (1 - 1{,}798)\ u_{11} + 0{,}581\ u_{21} + 0{,}154\ u_{31} = 0 & u_{11} = 1{,}262 \\ 0{,}581\ u_{11} + (1 - 1{,}798)\ u_{21} + 0{,}439\ u_{31} = 0 & u_{21} = 1{,}469 \\ 0{,}154\ u_{11} + 0{,}439\ u_{21} + (1 - 1{,}798)\ u_{31} = 0 & u_{31} = 1{,}000. \end{array}$$

Вторая система уравнений, для $\lambda_2 = 0{,}875$:

$$\begin{array}{l|l} (1 - 0{,}875)\ u_{12} + 0{,}581\ u_{22} + 0{,}154\ u_{32} = 0 & u_{12} = 0{,}144 \\ 0{,}581\ u_{12} + (1 - 0{,}875)\ u_{22} + 0{,}439\ u_{32} = 0 & u_{22} = 0{,}234 \\ 0{,}154\ u_{12} + 0{,}439\ u_{22} + (1 - 1{,}875)\ u_{32} = 0 & u_{32} = 1{,}000. \end{array}$$

Третья система уравнений, для $\lambda_3 = 0{,}327$:

$$\begin{array}{l|l} (1 - 0{,}327)\ u_{13} + 0{,}581\ u_{23} + 0{,}154\ u_{33} = 0 & u_{13} = 1{,}307 \\ 0{,}581\ u_{13} + (1 - 0{,}327)\ u_{23} + 0{,}439\ u_{33} = 0 & u_{23} = 1{,}779 \\ 0{,}154\ u_{13} + 0{,}439\ u_{23} + (1 - 0{,}327)\ u_{33} = 0 & u_{33} = 1{,}000. \end{array}$$

Матрица собственных векторов принимает вид:

$$U = \begin{pmatrix} 1{,}262 & -0{,}144 & 1{,}307 \\ 1{,}469 & -0{,}234 & -1{,}779 \\ 1{,}000 & 1{,}000 & 1{,}000 \end{pmatrix}.$$

Пронормируем векторы U_j, т.е. найдем $V_j = U_j / |U_j|$ и получим матрицу нормированных значений собственных векторов:

$$V = \begin{pmatrix} 0{,}579 & -0{,}139 & 0{,}539 \\ 0{,}674 & -0{,}225 & -0{,}734 \\ 0{,}459 & 0{,}964 & 0{,}413 \end{pmatrix},$$

так как V — матрица, отображающая ортонормированное пространство, в общем должно выполняться условие: $V'V = E$.

Матрицу факторного отображения (A) получим из матричного уравнения $A = V\Lambda^{1/2}$.

$$A = \begin{pmatrix} 0{,}579 & -0{,}139 & 0{,}539 \\ 0{,}674 & -0{,}225 & -0{,}734 \\ 0{,}459 & 0{,}964 & 0{,}413 \end{pmatrix} \cdot \begin{pmatrix} \sqrt{1{,}798} & & \\ & \sqrt{0{,}875} & \\ & & \sqrt{0{,}327} \end{pmatrix} =$$

	F_1	F_2	F_3
X_1	0,776	−0,130	0,308
X_2	0,904	−0,210	−0,420
X_3	0,616	0,902	0,236

Матрица A содержит частные коэффициенты корреляции, представляющие связи исходных признаков X_j и главных компонент F_r. Соответственно все элементы a_{ij} могут варьировать в пределах от −1 до +1.

Из равенства $A'A = \Lambda$ следует условие $\sum_j a_{jr}^2 = \lambda_r$. Проверим, как оно выполняется на исчисленных данных матрицы A:

$$\sum_j a_{j1}^2 = 0{,}776^2 + 0{,}904^2 + 0{,}616^2 = 1{,}798;$$

$$\sum_j a_{j2}^2 = 0{,}875; \quad \sum_j a_{j3}^2 = 0{,}327.$$

Теперь запишем системы линейных уравнений зависимости элементарных признаков Z_j и главных компонент, или обобщенных признаков F_r; максимально $r = j = 3$:

$$Z_1 = 0{,}776F_1 - 0{,}130F_2 + 0{,}308F_3,$$
$$Z_2 = 0{,}904F_1 - 0{,}210F_2 - 0{,}420F_3,$$
$$Z_3 = 0{,}616F_1 + 0{,}902F_2 + 0{,}236F_3, \text{ и}$$

$$F_1 = 1/1{,}798\ (\ 0{,}776Z_1 + 0{,}904Z_2 + 0{,}616Z_3),$$
$$F_2 = 1/0{,}875\ (-0{,}130Z_1 - 0{,}210Z_2 + 0{,}902Z_3),$$
$$F_3 = 1/0{,}375\ (\ 0{,}308Z_1 - 0{,}420Z_2 + 0{,}236Z_3).$$

На завершающем шаге алгоритма исчислим значения главных компонент для всех наблюдаемых объектов и построим матрицу F:

$F = A^{-1}Z'$ и матрица Z известна из условия задачи,

$$F = \begin{pmatrix} 0{,}542 & 0{,}507 & 0{,}196 \\ -0{,}776 & -0{,}010 & 0{,}994 \\ 1{,}554 & -1{,}283 & -0{,}075 \end{pmatrix} \cdot \begin{pmatrix} -0{,}971 & -0{,}868 & 1{,}478 & 0{,}361 \\ 0{,}549 & -1{,}684 & 0{,}882 & 0{,}253 \\ 0{,}076 & -1{,}069 & -0{,}534 & 1{,}527 \end{pmatrix} =$$

$$= \begin{pmatrix} -0{,}233 & -1{,}533 & 1{,}143 & 0{,}623 \\ 0{,}823 & -0{,}372 & -1{,}687 & 1{,}236 \\ -2{,}218 & 0{,}892 & 1{,}205 & 0{,}121 \end{pmatrix}.$$

Более привычной формой записи $n \times r$ значений главных компонент является транспонированная матрица F:

Fr — главные компоненты →

		F_1	F_2	F_3
$F' =$	n_1	−0,233	0,823	−0,218
	n_2	−1,533	−0,372	0,892
	n_3	1,143	−1,687	1,205
n_i — объекты ↓	n_4	0,623	1,236	0,121
	$\sum f_{ri}$	0	0	0

Центр распределения значений главных компонент F_2 находится в точке (0, 0, ..., 0), как это показано на рис. 7.9. Отсюда следует правило равенства суммы элементов каждого столбца матрицы F' нулю. В примере это правило выдерживается.

Далее аналитические выводы по результатам расчетов следуют уже после принятия решения о числе значащих признаков Z_j и главных компонент F_r и определения названий главным компонентам.

7.4.3. Оценка уровня информативности и поиск названий для главных компонент

Алгебраическими преобразованиями матрицы исходных данных X выделяют главные компоненты F и устанавливают их пространственное местоположение. Задачи распознавания главных компонент, определения для них названий решают затем субъективно на основе весовых коэффициентов a_{jr} из матрицы отображения A.

Для каждой главной компоненты F множество значений a_{jr} условно разбивается на четыре подмножества с нечеткими границами:

— подмножество незначимых весовых коэффициентов;

W_2 — подмножество значимых весовых коэффициентов;

W_3 — подмножество значимых весовых коэффициентов, не участвующих в формировании названия главной компоненты;

$W_2 — W_3$ — подмножество значимых весовых коэффициентов, участвующих в формировании названия.

Дополнительное выделение подмножества W_3 объясняется стремлением к более простой структуре главной компоненты, всегда легче поддающейся интерпретации. На своих границах подмножество W_3 имеет критические значения: в $a_{кр.1}$ — максимальное число признаков, объясняющих главную компоненту, в $a_{кр.2}$ — минимальное число объясняющих признаков. При тщательном анализе критические границы W_3 могут устанавливаться по статистическому критерию t, как в случае обычной проверки значимости коэффициентов корреляции.

Общий состав множества весовых коэффициентов представлен на рис. 7.10.

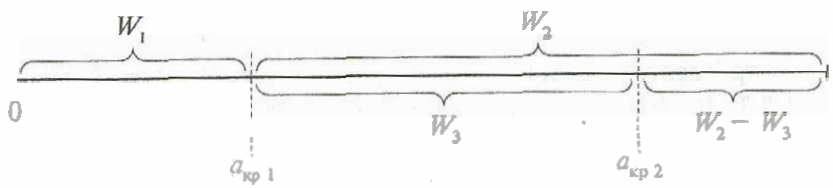


Рис. 7.10. Состав множества весовых коэффициентов a_{jr}, $j=\overline{1,m}$, для r-й главной компоненты

Подтверждение значимости признаков (X_j или Z_j), участвующих в формировании названия главной компоненты, можно получить расчетным путем при определении коэффициента информативности:

$$K_и = \frac{\sum_j a_{jr}^2 \{W_2 - W_3\}}{\sum_{j=1}^{m} a_{jr}^2}$$

Набор объясняющих признаков считается удовлетворительным, если значения $K_и$ лежат в пределах 0,75 — 0,95.

Рассмотрим пример гипотетических данных:

Исходные элементарные признаки X_j	*Главные компоненты* F_1	F_2
X_1 — уровень выработки на одного среднегодового работника	$a_{11} = 0,9$	$a_{12} = 0,1$
X_2 — уровень фондоотдачи	$a_{21} = 0,8$	$a_{22} = 0,4$
X_3 — размер оборотных производственных средств	$a_{31} = 0,1$	$a_{32} = 0,8$
X_4 — размер затрат на выпуск единицы товарной продукции	$a_{41} = 0,8$	$a_{42} = 0,3$
X_5 — численность промышленно-производственного персонала	$a_{51} = 0,3$	$a_{52} = 0,7$
X_6 — рентабельность продукции	$a_{61} = 0,7$	$a_{62} = 0,2$
X_7 — уровень энерговооруженности труда	$a_{71} = 0,2$	$a_{72} = 0,6$

Выделим для первой главной компоненты F_1 подмножества весовых коэффициентов на основе простой визуальной оценки аналитических результатов:

$$W_1 = \begin{Bmatrix} a_{31} = 0,1 \\ a_{71} = 0,2 \end{Bmatrix}; \qquad W_2 = \begin{Bmatrix} a_{11} = 0,9 & a_{51} = 0,3 \\ a_{21} = 0,8 & a_{61} = 0,7 \\ a_{41} = 0,8 \end{Bmatrix};$$

$$W_3 = a_{51} = 0,3; \qquad W_2 - W_3 = \begin{Bmatrix} a_{11} = 0,9 & a_{41} = 0,8 \\ a_{21} = 0,8 & a_{61} = 0,7 \end{Bmatrix}.$$

Пограничные значения для подмножества W_3 будут: $a_{кр.1} = 0,2$ и $a_{кр.2} = 0,7$. В решающее подмножество $W_2 — W_3$ вошли элементарные признаки: X_1, X_2, X_4, X_6, все они представляют характеристики эффективности производственной деятельности. Назовем F_1 *эффективность производства*. Значение коэффициента информативности дает основания утверждать, что состав подмножества $W_2 — W_3$ для главной компоненты F_1 достаточно надежен:

$$K_{и1} = \frac{0,81+0,64+0,64+0,49}{0,81+0,64+0,01+0,64+0,09+0,49+0,04} = 0,948,$$

т.е. значениями признаков X_1, X_2, X_4, X_6 состав главной компоненты F_1 определяется более чем на 94%.

Для второй главной компоненты F_2:

$$W_1 = \begin{Bmatrix} a_{12} = 0{,}1 \\ a_{62} = 0{,}2 \end{Bmatrix} \qquad W_{22} = \begin{Bmatrix} a_{22} = 0{,}4 \\ a_{32} = 0{,}8 \\ a_{62} = 0{,}6 \end{Bmatrix} \begin{Bmatrix} a_{42} = 0{,}3 \\ a_{52} = 0{,}7 \end{Bmatrix};$$

$$W_{22} = \begin{Bmatrix} a_{22} = 0{,}4 \\ a_{42} = 0{,}3 \end{Bmatrix} \qquad W_2 - W_3 = \begin{Bmatrix} a_{32} = 0{,}8 \\ a_{52} = 0{,}7 \\ a_{71} = 0{,}6 \end{Bmatrix} \text{ и } \qquad \begin{Bmatrix} a_{\text{кр.1}} = 0{,}2 \\ a_{\text{кр.2}} = 0{,}6 \end{Bmatrix}.$$

Название главной компоненты F_2 определяется наличием в ее структуре значимых признаков X_3, X_5, X_7, т.е. F_2 — это *размер производственных ресурсов*. Коэффициент информативности подтверждает существенный состав и этой главной компоненты:

$$K_{\text{и}2} = \frac{0{,}64+0{,}49+0{,}36}{0{,}01+0{,}16+0{,}64+0{,}09+0{,}49+0{,}04+0{,}36} = 0{,}832.$$

Отбор значащих элементарных признаков при определении названия главной компоненты производится прежде всего по абсолютной величине весового коэффициента a_{jr}, знак коэффициента приобретает значение в последующем, при логическом объяснении состава и установлении его непротиворечивости названию главной компоненты.

Обратимся к примеру из п. 7.4.2. В составе первой главной компоненты все исходные признаки значимы, наиболее существенно влияние признака X_2 (рентабельность производства), корректным названием для F_1 будет «эффективность производства». Во второй главной компоненте значением весового коэффициента выделяется признак X_3 (уровень фондоотдачи), он на 93% определяет состав F_2, назовем компоненту «эффективность использования основных производственных средств». Третья главная компонента в анализе может не рассматриваться. Если первые две главные компоненты объясняют 78,5% общей вариации X_1, X_2, X_3, то F_3 — только 10.9%. Таким образом, для выводов при минимальной потере информативности остаются:

$$F_1 = \frac{1}{1{,}798}(0{,}776Z_1 + 0{,}904Z_2 + 0{,}616Z_3) \text{ и } F_2 = \frac{1}{0{,}875}0{,}902Z_3.$$

Главные компоненты в свою очередь определяют значения элементарных признаков исходной совокупности данных. Вариация признака X_1 в большей мере определяется вариацией F_1, т.е. эффективность использования живого труда обусловливается общей эффективностью производства, или общей эффективной организацией производства. То же касается и рентабельности производства X_2, которая также в основном определяется колебаниями F_1. Значения третьего признака X_3 находятся под влиянием F_1 и F_2. В общем правомерна аналитическая запись:

$$Z_1 = 0{,}776F_1; \quad Z_2 = 0{,}904F_2; \quad Z_3 = 0{,}616F_1 + 0{,}902F_2.$$

Приведенный условный пример демонстрирует вычислительные процедуры метода главных компонент и построен на матрице исходных данных малой размерности. Дефицит информации вызывает необходимость определенных допущений при интерпретации аналитических результатов. Однако общий ход рассуждений соответствует действительности и остается для исследователя показательным.

Обратим внимание на знаки весовых коэффициентов a_{jr}, они могут серьезно затруднять выводы. Так, видно, что вариация признака X_3 находится под положительным влиянием главных компонент F_1, F_2 и отрицательным — F_3. При условии существенной значимости F_3 следовало бы объяснить наличие отрицательной связи и подтвердить соответствие структуры названию главной компоненты. Собственно одновременно решается вопрос соответствия результатов алгебраических преобразований логике экономических выводов.

Нелогичность знаков весовых коэффициентов, как и нечеткая структура главных компонент, когда все весовые коэффициенты имеют близкие по величине значения, становится причиной слабой интерпретируемости или даже неинтерпретируемости главных компонент. Классически эта проблема решается удалением из анализа малозначащих главных компонент, объясняющих незначительную долю вариации исходных признаков (т.е. имеющих малое λ_r, часто $\lambda_r < 1$), или вращением пространства главных компонент.

7.4.4. Использование метода главных компонент в системе других статистических методов

Метод главных компонент имеет не только самостоятельное значение в анализе, но и широко используется с другими мето-

дами статистики, наиболее часто с методами группировок и корреляционно-регрессионного анализа. *Группировка* осуществима как для объектов наблюдения, так и для элементарных признаков X_j. Важно, что в качестве группировочного выступают обобщенные признаки, значительно сжимающие исходную информацию. Возможность расчленения совокупности объектов (признаков) на группы становится очевидной при геометрическом представлении данных компонентного анализа (рис. 7.11).

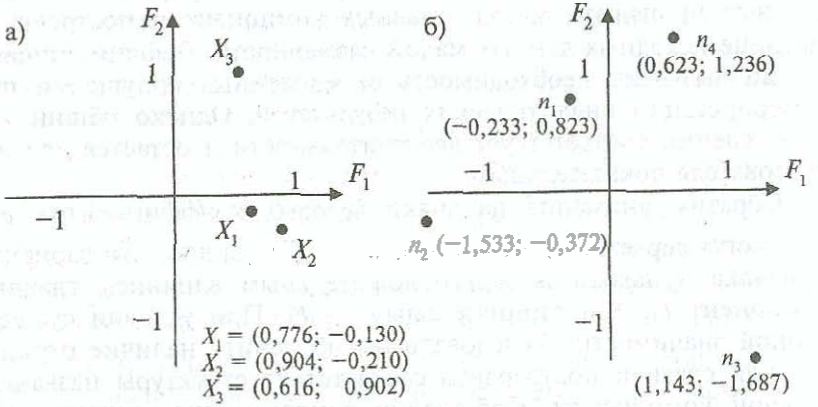


*Рис. 7.11. **Эмпирические распределения в пространстве двух главных компонент*** (по данным примера из п. 7.4.2):
а) распределение элементарных признаков по данным весовых коэффициентов a_{jr} из матрицы факторного отображения A;
б) распределение четырех промышленных предприятий по данным матрицы значений главных факторов F.

Когда число наблюдаемых единиц достаточно велико, легко допустить образование в пространстве главных компонент сгущений похожих объектов (рис. 7.12), что, собственно, является предпосылкой их группировки. Графики распределений обычно строятся для пространства двух каких-либо главных компонент (графики-биплоты), более сложный — вариант-отображение распределений в пространстве трех главных компонент (графики-триплоты).

Группировка может проводиться по одной (табл. 7.1) или сразу нескольким главным компонентам, в последнем случае

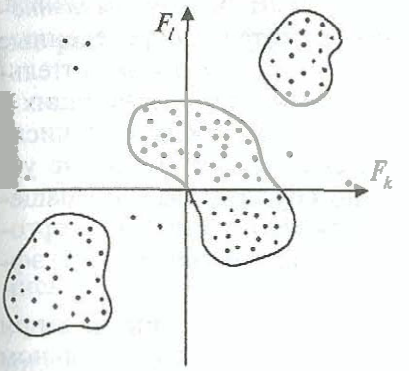


*Рис. 7.12. **Распределение наблюдаемых объектов в пространстве двух главных компонент по гипотетическим данным***

разбиение совокупности требует привлечения методов кластерного анализа (см. гл. 9).

Пример 7.2. Совокупность из 50 сельскохозяйственных предприятий анализировалась по значениям признаков: качеству сельскохозяйственных угодий X_1, климатическим условиям зоны территориального расположения X_2, величине нагрузки техногенного фактора X_3, уровню развития инфраструктуры X_4. После проведения анализа методом компонент осуществлена группировка предприятий по значениям первой главной компоненты F_1, объясняющей более 79% общей дисперсии элементарных признаков и имеющей название «условия сельскохозяйственного производства» (табл. 7.1).

Таблица 7.1. **Группировка сельскохозяйственных предприятий по качеству условий сельскохозяйственного производства (по значениям F_1 — главной компоненты)**

Группы предприятий по качеству условий для ведения сельскохозяйственного производства (F_1)	*Число сельскохозяйственных предприятий*	*Уровень выработки на одного среднегодового работника, стоим. ед.*	*Фондоотдача, стоим. ед.*	*Валовой выпуск в расчете на 100 га сельскохозяйственных угодий, стоим. ед.*
До 0,2	5	10450	0,75	123509
0,2 — 0,6	21	11670	0,81	151080
0,6 — 1,0	12	12010	0,83	150970
1,0 — 1,8	8	19100	1,05	173730
1,8 и более	4	18540	1,10	181600
Итого	50	13427	0,87	154362

Применение метода главных компонент в *корреляционно-регрессионном анализе* также дает исследователю определенные преимущества. Во-первых, появляется возможность значительного увеличения числа элементарных признаков, участвующих в анализе, при условии введения в регрессию небольшого числа только значащих главных компонент. Это, тем не менее, не усложняет самой модели и одновременно обусловливает сокращение доли необъясненной дисперсии отклика. Во-вторых, ортогональность главных компонент предотвращает проявление эффекта мультиколлинеарности.

Линейное уравнение регрессии на главных компонентах, при условии, что значения отклика (y) измерены в натуральном масштабе, записывается:

$$\hat{y}_F = \bar{y} + y_{i1}F_1 + y_{i2}F_2 + \ldots + y_{ir}F_r, \text{ или } \hat{y}_F = \bar{y} + \sum_{k=1}^{r} y_{ik}F_k = \bar{y} + yF_r',$$

где $\bar{y}$ — среднее значение зависимой переменной как оценка свободного члена уравнения;

y — вектор оценок коэффициентов регрессии при главных компонентах. Его находят решением известного матричного уравнения, минимизизирующего сумму квадратов отклонений: $\Sigma(\hat{y} - yF')^2 : \hat{y} = (F'F)^{-1}F'y$;

F — матрица значений главных компонент обычного вида размерностью $n \times r$.

Коэффициенты y_{ir} — это некоторые условные единицы, имеющие один масштаб измерения.

Уравнение регрессии на главных компонентах эквивалентно регрессии на стандартизованных значениях признаков:

$$\bar{y} + yF' = \bar{y} + Z'\beta,$$

где $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$;

β — вектор стандартизованных коэффициентов регрессии.

При построении регрессионной модели возникает вопрос об оптимальном составе главных компонент. На практике рекомендуется первоначально получить модель с учетом всех m главных компонент, затем с учетом вариации оценки надежности регрессионной модели и колебаний регрессионных коэффициентов число главных компонент может быть уменьшено. Незначимые для регрессии главные компоненты устанавливаются просто, по

величине собственных чисел λ_k или в ходе проверки параметров регрессии по t- или F-критериям:

$$t_н = \frac{|y_k|\sqrt{n \cdot \lambda_k}}{\sigma_y}, \text{ при } t_{кр} = t_{\alpha/2,\, n-m-1};$$

$$F_н = \frac{\sigma_y^2}{y_k^2 \cdot n \cdot \lambda_k}, \text{ при } F_{кр} = F_{\alpha/2,\, m,\, n-m-1}.$$

Компонента исключается из регрессии, когда собственное число λ_k мало, менее 75—90% и одновременно несущественно участие k-й компоненты в формировании результата $\hat{y}$, или при низких наблюденных значениях критериев t и F.

Пример 7.3. В результате статистического наблюдения за экологической обстановкой и уровнем заболеваемости населения в семи городах с различным уровнем техногенной нагрузки на окружающую среду получены данные (табл. 7.2).

Таблица 7.2. **Уровень заболеваемости населения и характеристики экологической обстановки в семи городах**

Показатель	*Город А*	*Город Б*	*Город В*	*Город Г*	*Город Д*	*Город Е*	*Город И*
Уровень заболеваемости взрослого населения злокачественными новообразованиями, на 1000 человек (y)	3,60	1,19	2,87	5,40	0,47	5,60	2,54
Средняя концентрация загрязняющих веществ в атмосферном воздухе (мг/м^3):							
пыль (X_1)	0,14	0,10	0,25	0,27	0,22	0,16	0,21
сернистый ангидрид (X_2)	0,005	0,004	0,005	0,010	0,070	0,012	0,030
окись углерода (X_3)	1,6	1,2	2,4	1,7	3,0	1,8	1,1
двуокись азота (X_4)	0,02	0,04	0,05	0,04	0,08	0,06	0,04
Сброс загрязненных вод в водоемы коммунальными организациями, млн. м^3 (X_5)	7578	87474	38496	329000	7200	1093102	8212

После анализа данных экологической обстановки (из табл. 7.2) по элементарным признакам (X_1 — X_5) методом главных компонент получены следующие результаты.

Собственные числа

λ_1	λ_2	λ_3	λ_4	λ_5
2,611	1,154	0,801	0,344	0,090

Матрица факторного отображения (A)

	F_1	F_2	F_3	F_4	F_5
X_1	0,568	−0,263	0,762	0,176	0,007
X_2	0,840	−0,133	−0,382	0,317	−0,147
X_3	0,887	−0,010	0,046	−0,454	−0,094
X_4	0,891	0,361	−0,159	0,045	0,222
X_5	−0,054	0,968	0,214	0,070	−0,102

Пообъектные значения главных компонент и зависимой переменной

Город	*У*	F_1	F_2	F_3	F_4	F_5
А	3,60	−1,05	−0,69	−0,44	−0,91	−1,78
Б	1,19	−1,05	0,02	−1,30	−0,26	1,48
В	2,87	0,39	−0,53	1,15	−1,38	1,14
Г	5,40	−0,06	−0,15	1,64	0,58	−0,32
Д	0,47	2,13	−0,29	−1,05	−0,03	−0,35
Е	5,60	−0,05	2,37	0,09	0,03	−0,30
И	2,54	−0,31	−0,73	−0,08	1,97	0,13
Среднее значение:	3,095	—	—	—	—	—

Коэффициенты корреляции независимой переменной и главных компонент

	F_1	F_2	F_3	F_4	F_5
У	−0,341	0,514	0,697	0,059	−0,366

Параметры регрессионных уравнений определялись по индивидуальным значениям главных компонент методом пошаговой регрессии, первое уравнение включало все пять выделенных компонент, затем каждый раз одна главная компонента, с наименьшим значением λ_k, исключалась. Выводы о качестве регрессионных уравнений следуют по данным статистических критериев: t-оценок Стьюдента существенности предикторов, F-критерия адекватности модели, R и R^2 — множественных коэффициентов корреляции и детерминации. Результаты регрессионного анализа сведены в табл. 7.3. По данным таблицы можно заключить, что

Таблица 7.3. Результаты регрессионного анализа данных, обработанных методом главных компонент

Параметры регрессионного уравнения	*Регрессионная модель*									
	$\hat{y}_1$		$\hat{y}_2$		$\hat{y}_3$		$\hat{y}_4$		$\hat{y}_5$	
	Параметр	*Критерий*	*Параметр*	*Критерий*	*Параметр*	*Критерий*	*Параметр*	*Критерий*	*Параметр*	*Критерий*
$\hat{y}_0$ ($\bar{y}$)	3,095	—	3,095	—	3,095	—	3,096	—	3,096	—
$\hat{y}_1$	−0,606	10,81	−0,606	1,32	−0,607	1,58	−0,614	0,87	−0,614	0,81
$\hat{y}_2$	0,926	26,54	0,927	2,01	0,925	2,41	0,925	1,31	—	—
$\hat{y}_3$	1,250	22,32	1,252	2,71	1,252	3,27	—	—	—	—
$\hat{y}_4$	0,095	1,70	0,120	0,26	—	—	—	—	—	—
$\hat{y}_5$	−0,651	11,60	—	—	—	—	—	—	—	—
F-критерий	206,36		3,31		6,36		1,23		1,52	
R	0,9995	—	0,932	—	0,930	—	0,616	—	0,341	—
R^2	0,9990		0,869		0,864		0,380		0,116	

в общем величины самих регрессионных коэффициентов с уменьшением числа предикторов колеблются незначительно. Но критерии надежности модели и ее параметров отличаются существенно и для $\alpha = 0{,}05$ четыре уравнения, кроме первого, по F-критерию нельзя считать адекватными реальным данным, с некоторой натяжкой исключение допустимо сделать для $\hat{y}_3$. Так же резко снижаются и значения t-критериев надежности регрессионных коэффициентов. Следует обратить внимание, что выводы о малой значимости главной компоненты F_4 по $\lambda_4 = 0{,}344$ подтверждаются и ее низкой t-оценкой в регрессионной модели. Для компоненты F_5 с $\lambda_5 = 0{,}09$ регрессионный критерий $t = 11{,}6$ указывает на ее существенность в уравнении, при решении вопроса об исключении F_5 следовало бы провести более глубокую логическую и критериальную проверку. В целом же на примере первой регрессионной модели ($\hat{y}_1$) видно, что с помощью обобщенных факторных признаков (главных компонент) реально построить хорошую линейную регрессию с высокой адекватностью и значительной объясняющей способностью ($R_1^2 = 0{,}999$).

Продемонстрированные примеры практического приложения метода главных компонент в кластерном и корреляционно-регрессионном анализе представляют в общем возможности и методов факторного анализа.

7.5. Разложение дисперсии в факторном анализе

Методы факторного анализа базируются на реальной оценке самой возможности исследования явлений. Здесь заранее принимается, что общими факторами объяснить вариацию элементарных признаков на 100% нельзя, некоторая часть вариации остается скрытой характерностью изучаемого явления. Стандартизованное значение признака по i-му наблюдаемому объекту при этом представляется линейной зависимостью вида:

$$z_{ij} = a_{j1}f_{1i} + a_{j2}f_{2i} + \ldots + a_{jm}f_{mi} + a_j d_{ji},$$

где d_{ji} — латентный признак, обобщающий нераскрываемую, характерную вариацию элементарного признака;

a_j — весовой коэффициент при характерном факторе.

Уравнение функциональной зависим z_{ij} от F_r и D_j перепишем, введя обобщающий знак суммирования:

$$z_{ij} = \sum_{r=1}^{m} a_{jr}f_{ri} + a_j d_{ji}.$$

Как известно, дисперсия стандартизованной величины равна единице: $D(z_{ij}) = 1$, но одновременно $D(z_{ij}) = \frac{1}{n}\sum_{i=1}\left(\sum_{r=1}^{m} a_{jr}f_{ri} + a_j d_{ji}\right)^2$.

Возведем в квадрат правую часть дисперсионного равенства:

$$D(z_{ij}) = \frac{1}{n}\left[a_{j1}^2\sum_{i=1}^{n} f_{1i}^2 + a_{j2}^2\sum_{i=1}^{n} f_{2i}^2 + \ldots + a_{jm}^2\sum_{i=1}^{n} f_{mi}^2 + d_j^2\sum_{j=1}^{n} u_{ji}^2 + \right.$$

$$+2\left(a_{j1}a_{j2}\sum_{i=1}^{n} f_{1i}f_{2i} + a_{j1}a_{j3}\sum_{i=1}^{n} f_{1i}f_{3i} + \ldots + a_{j(m-1)}a_{jm}\sum_{i=1}^{n} f_{(m-1)i}f_{mi} + \right.$$

$$\left.\left. + a_{jm}a_j\sum_{i=1}^{n} f_{mi}d_{ji}\right)\right].$$

Здесь запись произведений упрощается с учетом аксиом ортонормированного пространства, в котором:

$$\frac{1}{n}\sum_{i=1}^{n} f_{ri}^2 = 1, \quad \frac{1}{n}a_{jr}^2\sum_{i=1}^{n} f_{ri} = a_{jr}^2, \quad \frac{a_{jk}a_{jl}}{n}\sum_{i=1}^{n} f_{ki}f_{li} = a_{jk}a_{jl}r_{f_k f_l},$$

где r — коэффициент корреляции.

Таким образом, можем записать:

$$D(z_{ij}) = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2 + d_j^2 + 2\left(a_{j1}a_{j2}r_{f_1f_2} + a_{j1}a_{j3}r_{f_1f_3} + \ldots + a_{jm}a_j r_{f_m d_{ji}}\right).$$

При условии линейной независимости общих факторов: $r_{f_k f_l} = 0$, в последнем равенстве выражение в скобках обращается в ноль, в итоге дисперсия стандартизованной величины раскладывается по оценкам дисперсии общих факторов и характерности:

$$D(z_{ij}) = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2 + d_j^2 = 1.$$

Метод главных компонент может рассматриваться как частный от приведенного выше случай, когда

$$d_j^2 = 0 \text{ и } D(z_{ij}) = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2 = 1.$$

Назовем $\sum_{r=1}^{m} a_{jr}^2$ общностью, это доля общей дисперсии, которая поддается объяснению через общие факторы. Обозначим ее

h_j^2. Характерность — доля дисперсии, не объясненной общими факторами, или вклад в общую вариацию признака X_j некоторого характерного, скрытого фактора. Примем для характерности обозначение d_j^2.

Методы факторного анализа позволяют в изучении дисперсии пойти дальше и расчленить вариацию характерного признака, дополнительно выделив:

специфичность (b_j^2) — долю дисперсии, обусловленную вариабельной спецификой признака X_j;

ненадежность (l_j^2) — долю дисперсии, обусловленную несовершенством измерений;

надежность (c_j^2) — долю дисперсии характерного фактора без учета ошибки, или дополнение дисперсии ошибки l_j^2 до полной дисперсии признака X_j.

Взаимосвязь показателей состава дисперсии элементарного признака X_j укладывается в общую схему (рис. 7.13):

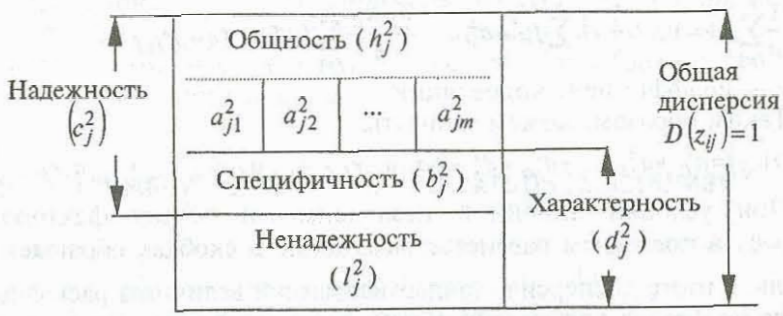


Рис. 7.13. Состав дисперсии элементарного признака X_j в факторном анализе

Рис. 7.13 позволяет установить алгебраические связи показателей состава дисперсии:

Полная дисперсия $h_j^2 + d_j^2 = h_j^2 + b_j^2 + l_j^2 = 1$

Общность $h_j^2 = 1 - d_j^2$

Характерность $d_j^2 = 1 - h_j^2 = b_j^2 + l_j^2$

Специфичность $b_j^2 = d_j^2 - l_j^2$

Ненадежность (дисперсия ошибки) $l_j^2 = 1 - h_j^2 - b_j^2$

Надежность $c_j^2 = h_j^2 + b_j^2 = 1 - l_j^2$.

Соответственно уровню разложения дисперсии можно записать матричные линейные уравнения относительно элементарных признаков и воспроизведенной матрицы парных корреляций (R^+):

Метод главных компонент $Z = AF;\ R^+ = AA';$

Методы факторного анализа

а) при условии ортогональности общих факторов и выделении только характерного фактора $Z_j = AF + a_j D_j;$ $R^+ = AA' + U^2.$

б) при условии косоугольного решения и выделении только характерного фактора $Z_j = AF + a_j D_j;$ $R^+ = ACA' + U^2;$

в) при условии косоугольного решения и расчленении характерности на специфичность и ненадежность $Z_j = AF + a_j D_j + d_{j+1} C;$ $R^+ = ACA' + D^2 + C^2.$

Пример записи матричного уравнения $R = AA' + U^2$ в развернутом виде:

AA'	D^2	h_j^2	$D(Z)$
$\begin{pmatrix} 0{,}84 & 0{,}04 \\ 0{,}68 & 0{,}01 \\ 0{,}56 & 0{,}08 \\ 0{,}16 & 0{,}25 \end{pmatrix}$	$\begin{pmatrix} 0{,}12 & 0 & 0 & 0 \\ 0 & 0{,}31 & 0 & 0 \\ 0 & 0 & 0{,}36 & 0 \\ 0 & 0 & 0 & 0{,}59 \end{pmatrix}$	$\begin{matrix} 0{,}88 \\ 0{,}69 \\ 0{,}64 \\ 0{,}41 \end{matrix}$	$\begin{matrix} 1{,}00 \\ 1{,}00 \\ 1{,}00 \\ 1{,}00 \end{matrix}$

Вариант в), с полным расчленением дисперсии признака, используется достаточно редко, так как предполагает выполнение сложных алгоритмов вычислительных процедур.

Анализ состава дисперсии позволяет дополнить факторный анализ важными выводами о степени специфичности изучаемого явления и информативности статистических данных.

7.6. Метод главных факторов

Метод главных факторов можно рассматривать как развитие метода главных компонент. Основное отличие заключается в использовании редуцированной корреляционной матрицы R_h, на главной диагонали которой расположены уже не единицы, а характеристики общности h_j^2.

Согласно классической модели факторного анализа, уравнение для определения коэффициентов при общих факторах F_r записывается в виде:

$$Z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jm}F_m + a_jD_l,$$

или в матричной форме:

$$Z_j = AF + a_jD_J,$$

где D_j — характерный фактор.

Решение уравнения при условии максимизации сумм:

$\sum_{j=1}^{m} a_{j1}^2 = \lambda_1$ — первый максимум в части описанной дисперсии элементарных признаков ($D(Z_j)$);

$\sum_{j=1}^{m} a_{j2}^2 = \lambda_2$ — второй максимум, относительно оставшейся после λ_1 дисперсии и т.д.,

сводится к определению собственных значений и собственных векторов симметрической матрицы R из равенства $(R - \lambda E)U = 0$. В §7.4, посвященном методу главных компонент, показано, что при известных значениях λ_j и U_j коэффициенты a_{jr} можно рассчитать по формуле:

$$A = V \cdot \Lambda^{1/2},$$

где V — матрица нормированных векторов U_j: $V_j = \dfrac{U_j}{|U_j|}$.

С учетом этого для вычислений a_{jr} выводится общая формула:

$$a_{jr} = \frac{u_{jr}\sqrt{\lambda_r}}{\sqrt{u_{1r}^2 + u_{2r}^2 + \ldots + u_{mr}^2}}.$$

В исследовательской практике существуют различные приемы, способы нахождения параметров модели главных факторов

λ, U_j, a_{jr}, условно их легко разделить на две большие группы. Первая группа ориентирована на алгоритм метода главных компонент и в сущности повторяет его, единственное отличие в том, что вычисления проводятся по данным редуцированной корреляционной матрицы, а не обычной матрицы парных корреляций. В этом случае сразу получают все m значений собственных чисел λ_j и m собственных векторов. В факторном анализе, однако, такой подход используется редко, он считается менее адекватным традиционным целям факторного анализа и менее экономичным.

Во второй группе объединяются приемы, которые позволяют последовательно, начиная с первого, устанавливать значения собственных чисел и собственных векторов. Последующие шаги выполняются после предварительной проверки на достаточность информативности уже выделенных главных факторов. Такой подход можно считать классическим в факторном анализе. Наиболее представительным здесь является метод, разработанный Г. Хотеллингом. Он позволяет сравнительно быстро выделить небольшое число общих факторов, учитывающих почти всю суммарную общность. Метод предполагает итеративное решение. На первом этапе редуцированную матрицу парных корреляций многократно возводят в квадрат, чем добиваются сходимости к первому собственному значению λ_1, затем вычисляют соответствующие значения собственного вектора U_1 и факторные нагрузки a_{j1}. В завершение этапа находят произведения векторов $A_1 \times A_1' = R^+$, где R^+ — воспроизведенная корреляционная матрица. Остаточная матрица парных корреляций будет $R - R^+$. Если при проверке разность $R - R^+$ существенна, переходят ко второму этапу, и описанная выше итерация повторяется, но относительно второго собственного числа — λ_2, вычисляемого по данным матрицы остаточных коэффициентов корреляции, и т.д.

Итерации повторяют до тех пор, пока разность $R - R^+$ не становится достаточно малой и тогда алгоритм завершается.

Пример 7.4. Воспользуемся данными матрицы парных корреляций из п. 7.4.2

$$R = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}.$$

К редуцированной корреляционной матрице перейдем, приняв характерности на уровне наибольших коэффициентов корреляции в каждом столбце, с поправками Барта:

$$R_h = \begin{pmatrix} 0{,}670 & 0{,}581 & 0{,}154 \\ 0{,}581 & 0{,}630 & 0{,}439 \\ 0{,}154 & 0{,}439 & 0{,}420 \end{pmatrix}.$$

Для редуцированной матрицы выполняется операция возведения в степень: $R^2 = R'R$. Процедура повторяется до тех пор, пока α — оценки матрицы R до и после возведения в степень — не перестанут существенно различаться, т.е. $d = (\alpha_{(i)} - \alpha_{(i-1)})$ и d должны быть минимальны, меньше некоторого заранее заданного порогового уровня. Оценки α — это приближения факторного отображения, $\alpha = p_i / p_{\max}$, где p — скаляр $\left(p_i^{(i+1)} = R_h^{(1)} \cdot S^{(1)}\right)$, соответствующий величине суммы коэффициентов корреляции по каждой строке ($S = \sum_j r_{ij}$), значения величин p и S взаимно контролируются.

Будем матрицу R_h возводить в степень, одновременно вычисляя оценки α, S и p, результаты расчетов сведем в табл. 7.4 — 7.7.

После первого возведения в квадрат матрицы R_h значения d-характеристики еще достаточно велики. Продолжим операцию умножения матриц, отметив про себя, что мы имеем всякий раз дело с симметрической матрицей, и расчеты можно значительно сократить, вычисляя только элементы по главной диагонали и наддиагональные. Кроме того, можем пропустить R_h^3 и сразу найти R_h^4 из произведения $R_h^2 \times R_h^2$ (табл. 7.6).

Таблица 7.4. **Исходная редуцированная корреляционная матрица**

Признак	R_h			$S_i^{(1)} = \sum_j r_{ij}$	$\alpha^{(1)} = \frac{S_i^{(1)}}{S_{\max}}$
	X_1	X_2	X_3		
X_1	0,670	0,581	0,154	1,405	0,851
X_2	0,581	0,630	0,.439	1,650	1,000
X_3	0,154	0,439	0,420	1,013	0,614

Таблица 7.5. **Первый цикл итерации: возведение в квадрат корреляционной матрицы**

Признак	$R_h^2 = R_h' R_h$			$S^{(2)} = \sum_j r_{ij}$	$P_i^{(2)} = R_h S^{(1)}$	$\alpha_i^{(2)} = P_i / P_{\max}$	$d = \lvert\alpha^{(2)} - \alpha^{(1)}\rvert$
	X_1	X_2	X_3				
X_1	0,811	0,823	0,423	2,057	2,056	0,894	0,043
X_2	0,823	0,928	0,550	2,301	2,300	1,000	0,000
X_3	0,423	0,550	0,393	1,366	1,365	0,593	0,021

Таблица 7.6. **Второй цикл итерации: корреляционная матрица в четвертой степени**

Признак	$R_h^4 = R_h^2 R_h^2$			$S_i^{(3)}$	$P_i^{(3)} = R_h^2 S^{(2)}$	$\alpha^{(3)} = P_i / P_{\max}$	$d = \lvert\alpha^{(3)} - \alpha^{(2)}\rvert$
	X_1	X_2	X_3				
X_1	1,514	1,664	0,962	4,140	4,140	0,904	0,010
X_2	1,664	1,836	1,074	4,574	4,579	1,000	0,000
X_3	0,962	1,074	0,635	2,671	2,673	0,585	0,008

После возведения корреляционной матрицы в четвертую степень d-разности резко уменьшились, вычислим R_h^8 и завершим итерацию.

Таблица 7.7. **Третий цикл итерации: корреляционная матрица в восьмой степени**

Признак	$R_h^8 = R_h^4 R_h^4$			$S_i^{(4)}$	$P_i^{(4)} = R_h^4 S^{(3)}$	$\alpha^{(4)} = P_i / P_{\max}$	$d = \lvert\alpha^{(4)} - \alpha^{(3)}\rvert$
	X_1	X_2	X_3				
X_1	5,986	6,607	3,854	16,447	16,448	0,906	0,002
X_2	6,607	7,293	4,255	18,155	18,156	1,000	0,000
X_3	3,854	4,255	2,481	10,590	10,591	0,583	0,002

Оценки S и p подтверждают правильность проведенных вычислений, их максимальное расхождение после выполнения трех циклов первой итерации не превысило пяти тысячных, таким образом, оценки компонент первого собственного вектора мож-

но считать достоверными. Собственный вектор — это ненормированные значения $\alpha^{(4)}$, т.е. $\alpha_1^{(4)} = U_1$.

Перейдем к определению нагрузок первого главного фактора (табл. 7.8).

Таблица 7.8. **Вычисление нагрузок первого главного фактора**

Признак	$\alpha_1^{(4)} = U_1$ (из табл. 7.7)	$\beta_1 = R_h.\alpha_1^{(4)}$	$A = \dfrac{U_1\sqrt{\lambda_1}}{(\sum U_{1i}^2)^{1/2}}$
X_1	0,906	1,278	0,732
X_2	1,000	1,412	0,808
X_3	0,583	0,823	0,471

В табл. 7.8 собственное число λ_1 — это наибольшая величина вектора β_1. Компоненты a_{j1} легко рассчитать по известному $\lambda_1 = 1{,}412$ и отношению

$$\frac{U_1\sqrt{\lambda_1}}{(\sum U_{1i}^2)^{1/2}} = \frac{\sqrt{1{,}412}}{(0{,}821+1{,}000+0{,}340)^{1/2}} = 0{,}808.$$

Итогами первой итерации будут: первое собственное число $\lambda_1 = 1{,}412$ и вектор факторных нагрузок: $A_1' = (0{,}732 \quad 0{,}808 \quad 0{,}471)$. Проверим выполнение требования:

$\sum a_{j1}^2 = \lambda_1$, или $0{,}732^2 + 0{,}808^2 + 0{,}471^2 = 1{,}412$.

Остается определить воспроизведенную матрицу парных корреляций (R^+_h) и решить вопрос о необходимости выполнения второй итерации с поиском второго собственного числа λ_2 и вектора факторных нагрузок A_2.

Воспроизведенная корреляционная матрица по только одному, первому, вектору факторного отображения будет $R_h^+ = AA'$:

$$R_h^+ = \begin{pmatrix}0{,}732\\0{,}808\\0{,}471\end{pmatrix}\cdot(0{,}732\ 0{,}808\ 0{,}471) = \begin{pmatrix}0{,}536 & 0{,}591 & 0{,}345\\0{,}591 & 0{,}653 & 0{,}381\\0{,}345 & 0{,}381 & 0{,}222\end{pmatrix}.$$

Разность матриц $R_h - R_h^+$ покажет остаточную, не объясненную первым главным фактором, корреляцию и поможет ответить на вопрос о целесообразности выделения второго главного фактора:

$$R_1 = R_h - R_h^+ = \begin{pmatrix}0{,}670 & 0{,}581 & 0{,}154\\0{,}581 & 0{,}630 & 0{,}439\\0{,}154 & 0{,}439 & 0{,}420\end{pmatrix} - \begin{pmatrix}0{,}536 & 0{,}591 & 0{,}345\\0{,}591 & 0{,}653 & 0{,}381\\0{,}345 & 0{,}381 & 0{,}222\end{pmatrix} =$$

$$= \begin{pmatrix}0{,}134 & -0{,}010 & -0{,}191\\-0{,}010 & -0{,}023 & 0{,}058\\-0{,}191 & 0{,}058 & 0{,}198\end{pmatrix}.$$

Матрица первых остаточных коэффициентов корреляции содержит еще достаточно большие величины и вполне допускает оценку второго главного фактора. Последующее выполнение второй итерации аналогично первой, только вычисления производятся на данных матрицы остатков R_1.

В табл. 7.9 получены более грубые оценки элементов собственного вектора, чем в первом случае (табл. 7.8). Тем не менее, итерация прервана с учетом, что элементы квадрируемой матрицы резко снижают свою значимость и теряют наглядность, в то же время заданная условность примера допускает различную степень приближения аналитических результатов, вычисляемых по уже знакомому алгоритму.

Таблица 7.9. **Исходная корреляционная матрица и матрица парных корреляций в четвертой степени используемая для расчетов собственного вектора U_2**

Признак	R_1			$S^{(1)}$	$\alpha^{(1)}$
	X_1	X_2	X_3		
X_1	0,134	−0,010	−0,191	−0,067	−1,000
X_2	−0,010	−0,023	0,058	0,025	0,373
X_3	−0,191	0,058	0,198	0,065	0,970

Признак	$R_1^4 = R_1^2 R_1^2$			$S^{(3)}$	$P^{(3)}$	$\alpha^{(3)}$	$d = \lvert\alpha^{(3)} - \alpha^{(2)}\rvert$
	X_1	X_2	X_3				
X_1	0,00723	−0,00149	−0,00870	−0,00296	−0,00297	−0,830	0,036
X_2	−0,00149	0,00032	0,00179	0,00062	0,00061	0,170	0,027
X_3	−0,00870	0,00179	0,01051	0,00360	0,00358	1,000	0,000

По данным табл. 7.9 определим нагрузки второго главного фактора (табл. 7.10).

Таблица 7.10. **Вычисление нагрузок второго главного фактора**

Признак	$\alpha_2^{(3)} = U_2$	$\beta_2 = R_1\alpha_2^{(3)}$	$A_2 = \dfrac{U_2\sqrt{\lambda_2}}{(\sum U_{2i}^2)^{1/2}}$
X_1	−0,830	−0,304	−0,383
X_2	0,170	0,062	0,079
X_3	1,000	0,366	0,462
—	$\sum_i U_{2i}^2 = 1{,}718$	$\lambda_2 = 0{,}366$	$\sum_j a_{jr}^2 = 0{,}366$

Матрицу вторых остаточных коэффициентов корреляции R_2 находят из разности $R_1 - A_2 \times A_2'$:

$$R_2 = \begin{pmatrix} 0{,}134 & -0{,}010 & -0{,}191 \\ -0{,}010 & -0{,}023 & 0{,}058 \\ -0{,}191 & 0{,}058 & 0{,}198 \end{pmatrix} \cdot \begin{pmatrix} -0{,}383 \\ 0{,}079 \\ 0{,}462 \end{pmatrix} \cdot (-0{,}383\ 0{,}079\ 0{,}462) =$$

$$= \begin{pmatrix} -0{,}013 & 0{,}020 & -0{,}014 \\ 0{,}020 & -0{,}029 & 0{,}022 \\ 0{,}014 & 0{,}022 & -0{,}015 \end{pmatrix}.$$

Элементы матрицы R_2 имеют малые значения и выделение третьего главного фактора не видится необходимым. Такой же вывод следует и по данным собственных чисел λ, а именно: при помощи первого и второго главных факторов полностью воспроизведена общность ($t_r R_h = 1{,}720$) и почти на 60% удалось объяснить вариацию элементарных признаков X_1, X_2, X_3 $\left(\dfrac{1{,}412+0{,}366}{3}\right)$.

В заключение обобщим итоги решения задачи методом главных факторов, выделяя общности и характерности для каждого из элементарных X_j признаков (табл. 7.11).

Таблица 7.11. **Факторные нагрузки, общности и характерности, вычисленные методом главных факторов**

Признак	*Главный фактор (факторные нагрузки)*		*Общность* $h_j^2 = \sum_r a_{jr}^2$	*Характерность* d_j^2
	F_1	F_2		
X_1	0,732	−0,383	0,683 (0,670)	0,317
X_2	0,808	0,079	0,660 (0,630)	0,340
X_3	0,471	0,462	0,435 (0,420)	0,565
—	$\sum_j a_{j1}^2 = 1{,}412$	$\sum_j a_{j2}^2 = 0{,}366$	$\sum h_j^2 = 1{,}778$	$\sum d_j^2 = 1{,}222$

В табл. 7.11 уровни общностей несколько отличаются от данных исходной матрицы R_h (приведенных в скобках). Это могло произойти из-за допускаемой грубости итеративных решений и округлений. В скобках приведены первоначально принятые величины общностей для каждого признака. Общности и характерности совместно полностью представляют дисперсии признаков, равные единице, в общем:

$$\sum h_j^2 + \sum d_j^2 = 3.$$

Алгоритм Хотеллинга, как видим, дает ценные преимущества: он легко укладывается в общую схему для реализации на ПЭВМ, одновременно освобождая исследователя от громоздких и сложных вычислений определителя корреляционной матрицы и поиска корней характеристического уравнения степени m. Последовательное выделение общих факторов позволяет прерывать решающий алгоритм и опускать ненужные расчеты малозначащих собственных чисел и векторов, экономя время и затраты исследователя.

7.7. Метод максимального правдоподобия

Это один из наиболее мощных методов в статистическом оценивании. С его помощью поиск наилучших параметров статистической совокупности осуществляется совместным оцениванием n — числа независимых наблюдений $(x_1, x_2, \ldots, x_n)$ из произведения отдельных функций плотности вероятностей: $L = \prod_{i=1}^{n} f(x_i, \theta_1, \theta_2)$. Для нормально распределенной одномерной

случайной величины функция максимального правдоподобия принимает вид:

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\frac{(x_i-\bar{x})^2}{\sigma^2}}$$

и приводит к хорошо знакомым результатам — $\theta_1 = \mu$ и $\theta_2 = \sigma^2$.

В работе Д. Лоули [59] показано, что в факторном анализе метод максимального правдоподобия применим для разработки вычислительных процедур и установления оптимальных оценок симметрической матрицы парных корреляций R, а именно: матрицы A — факторного отображения и D^2 — диагональной матрицы характерностей. Результаты решений отличаются определенностью. Пространственное положение ортогональной системы векторов факторных нагрузок сразу представляется наилучшим образом, так что не возникает необходимости ее последующего вращения и перебора вариантов пространственной ориентации с целью улучшения оценок матрицы A. В последующем, впрочем, допустимо переходить к косоугольной или какой-либо другой системе факторов.

Общая модель факторного анализа остается в данном случае действующей, описывая гипотетическую зависимость элементарного признака X_j от некоторого набора латентных (общих) факторов:

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jr}F_r + a_j D_i ,$$

здесь значения j-го элементарного признака X_{ij} могут быть величинами (натуральными или стандартизованными), F_r и D_j при $j = \overline{1,m}$ соответственно общие и характерный факторы, по предположению они независимы и нормально распределены.

Оценкой вариации многомерной случайной величины X_j, как известно, является ковариационная матрица $\sum$, при имеющихся истинных оценках матриц факторного отображения A и характерностей D^2 можем записать $\sum = AA' + D^2$. На самом деле всегда есть некоторое приближение оценок к $\sum = \hat{A}\hat{A}' + \hat{D}^2$.

Исследователями, работавшими над методом максимального правдоподобия в факторном анализе (Д. Лоули, А. Максвелл, К. Йореског, У.Дж. Хеммерл и др.), было показано, что оценки факторных нагрузок пропорциональны стандартному отклонению соответствующего параметра X_j и оценочные уравнения в

общем не зависят от единиц измерения x_{ij}. Это приводит к выводу о возможности построения оценочных уравнений для a_{jr} не только на коэффициентах матрицы $\sum$, но и на коэффициентах матрицы R в генеральной совокупности (матрицы ρ). Конечные результаты по данным матриц $\sum$ и ρ идентичны, что означает правомерность записи:

$$\rho = \hat{A}\hat{A}' + \hat{D}^2 = AA' + D,$$

причем $\hat{A} = \rho R^{-1}\hat{A}$ и $D^2 = E - AA'$, с произведением AA' — строго диагональной матрицей. Допуская $\rho = R$, дальнейшие рассуждения можно упростить. Опустим также, чтобы облегчить запись формул, непринципиальные для наших выводов символы ^. Будем иметь:

$$R = AA' + D^2.$$

Переход к матричному уравнению, представляющему алгоритм вычислений оптимальных параметров A и D, осуществляется умножением левой и правой частей предыдущего уравнения на $A' D^{-2}$, тогда:

$$A'D^{-2}R = (A'D^{-2}A + E)\,A'.$$

Вводя обозначение: $A'D^{-2}A = J$, имеем:

$$A'D^{-2}R = (E + J)\,A', \text{ или } AD^{-2}R - A = JA'.$$

Последнее матричное уравнение позволяет построить алгоритм для поиска улучшенных оценок факторных нагрузок. Зная, что $A' = J^{-1}(AD^{-2}R - A)$, получаем

$$A' = [(A'D^{-2}A)^2]^{-1/2}(AD^{-2}R - A) =$$
$$= [A'D^{-2}(RD^{-2}A - A)]^{-1/2}(AD^{-2}R - A).$$

Собственно в алгоритме итеративной вычислительной процедуры совместно используются два равенства:

$$JA' = AD^{-2}R - A; \quad D^2 = E - \operatorname{diag} A'A;$$

и дополнительное условие, что $A'R^{-1}A$ — диагональная матрица. Вычисления проводятся по известным заранее некоторым приближенным оценкам матрицы факторного отображения A, в которой каждый вектор-столбец есть подмножество элементов $A_r' = \{a_{1r}a_{2r}\ldots a_{mr}\}$.

Обозначим улучшенную оценку a_{jr} через l_{jr}, множество l_{jr} образует матрицу оптимизированных значений факторных нагрузок L. Векторы максимально правдоподобных оценок L_r находят по схеме:

$$L_1 = \left(RD^{-2}A_1 - A_1\right) \Big/ \sqrt{A_1' D^{-2}\left(RD^{-2}A_1 - A_1\right)};$$

$$L_2 = \left(RD^{-2}A_2 - A_2 - L_1 L_1' D^{-2} A_2\right) \Big/ \sqrt{A_2' D^{-2}\left(RD^{-2}A_2 - A_2 - L_1 L_1' D^{-2} A_2\right)};$$

$$\vdots$$

$$L_r = \left(RD^{-2}A_r - A_r - L_1 L_1' D^{-2} A_r - L_2 L_2' D^{-2} A_r - \ldots - L_{r-1} L_{r-1}' D^{-2} A_r\right) \Big/$$

$$\Big/ \sqrt{A_r' D^{-2}\left(RD^{-2}A_r - A_r - L_1 L_1' D^{-2} A_r - L_2 L_2' D^{-2} A_r - \ldots - L_{r-1} L_{r-1}' D^{-2} A_r\right)}.$$

Матрица улучшенных оценок характерностей G^2 вычисляется из равенства: $G^2 = E - \mathrm{diag} LL'$.

Итерации с определением новых векторных оценок L_r повторяются каждый раз на основе уже улучшенных оценок L_r и G^2 из предыдущей итерации до тех пор, пока факторные нагрузки не перестанут существенно изменяться, например, величины L_r по данным последней и предыдущей итерации не будут расходиться менее чем на 0,01.

Пример 7.5. По данным редуцированной корреляционной матрицы R_h, построенной для шести элементарных признаков X_j, центроидным методом выделено три общих фактора F_r и установлены характерности D^2 [66, с. 71, 81]:

$$R_h = \begin{pmatrix} 0{,}400 & 0{,}299 & 0{,}400 & 0{,}297 & 0{,}116 & 0{,}232 \\ 0{,}299 & 0{,}568 & 0{,}568 & 0{,}534 & 0{,}432 & 0{,}154 \\ 0{,}400 & 0{,}568 & 0{,}568 & 0{,}487 & 0{,}436 & 0{,}071 \\ 0{,}297 & 0{,}534 & 0{,}487 & 0{,}545 & 0{,}545 & 0{,}092 \\ 0{,}116 & 0{,}432 & 0{,}436 & 0{,}545 & 0{,}545 & 0{,}016 \\ 0{,}232 & 0{,}154 & 0{,}071 & 0{,}092 & 0{,}016 & 0{,}232 \end{pmatrix};$$

$$A = \begin{pmatrix} 0{,}500 & 0{,}365 & 0{,}145 \\ 0{,}733 & -0{,}118 & 0{,}075 \\ 0{,}726 & -0{,}095 & 0{,}312 \\ 0{,}717 & -0{,}220 & -0{,}155 \\ 0{,}590 & -0{,}404 & -0{,}194 \\ 0{,}219 & 0{,}365 & -0{,}099 \end{pmatrix}; \qquad D^2 = \begin{pmatrix} 0{,}596 \\ 0{,}443 \\ 0{,}367 \\ 0{,}414 \\ 0{,}451 \\ 0{,}809 \end{pmatrix},$$

$\lambda_1 = 2{,}226;\ \lambda_2 = 0{,}500;\ \lambda_3 = 0{,}246.$

Используя метод максимального правдоподобия, выполним первую итерацию по улучшению оценок факторных нагрузок и характерностей. Все расчеты для наглядности представим в таблице (табл. 7.12). Там же дается формальная запись всех выпол-

Таблица 7.12. Расчет максимально правдоподобных оценок для трех главных факторов и характерностей признаков X_j

(Первая итерация)

№ п/п	Исходные матричные элементы и выполняемые операции	Параметр элементарного признака X_j					
		X_1	X_2	X_3	X_4	X_5	X_6
А	1	2	3	4	5	6	7
1	a_{j1}	0,500	0,733	0,726	0,717	0,590	0,219
2	a_{j2}	0,365	−0,118	−0,095	−0,220	−0,404	0,365
3	a_{j3}	0,145	0,075	0,312	−0,155	−0,194	−0,099
4	$a_j^2 = 1 - a_{j1}^2 - a_{j2}^2 - a_{j3}^2$	0,596	0,443	0,367	0,414	0,451	0,809
5	$a_{j1}/a_j^2 = K_1$	0,838	1,655	1,978	1,732	1,308	0,271
6	$a_{j2}/a_j^2 = K_2$	0,612	−0,266	−0,259	−0,531	−0,896	0,451
7	$a_j/a_j^2 = K_3$	0,243	0,169	0,850	−0,374	−0,430	−0,122
8	$R_h K_1$	2,350	3,845	3,831	3,778	3,335	0,832
9	$R_h K_1 - a_{j1} = M_1$	1,754	3,112	3,105	3,061	2,745	0,613
10	$L_1 = M_1 / \sqrt{K_1 M_1'}$	0,376	0,666	0,665	0,655	0,588	0,131

Продолжение табл. 7.12

(Вторая итерация)

A	1	2	3	4	5	6	7
11	$R_h K_2$	−0,095	−0,717	−0,671	−0,822	−0,927	0,125
12	$R_h K_2 - a_{j2} = M_2^{(1)}$	−0,460	−0,599	−0,576	−0,602	−0,523	−0,240
13	$M_2^{(1)} - (K_2 L_1') = M_2$	−0,108	0,024	0,046	0,010	0,027	−0,118
14	$L_2 = M_2 / \sqrt{K_2 M_2'}$	0,403	−0,089	−0,171	−0,037	−0,101	0,440
15	$R_h K_3$	0,299	0,247	0,298	0,127	0,032	0,073
16	$R_h K_3 - a_{j3} = M_3^{(2)}$	0,154	0,172	−0,014	0,282	0,226	0,172
17	$M_3^{(2)} - (K_3 L_1)\ L_1 = M_3^{(1)}$	0,054	0,003	−0,183	0,116	0,077	0,139
18	$M_3^{(1)} - (K_3 L_2)\ L_2 = M_3$	0,078	−0,002	−0,193	0,114	0,071	0,165
19	$L_3 = M_3 / \sqrt{K_3 M_3'}$	−0,159	0,004	0,395	−0,233	−0,145	−0,337
20	$g_j^2 = 1 - l_{j1}^2 - l_{j2}^2 - l_{j3}^2$	0,672	0,549	0,373	0,516	0,623	0,675

няемых шагов, идентичных алгоритму, задаваемому матричным уравнением: $AD^2R - A = JA'$.

Обратите внимание, что при выполнении итерации в табл. 7.12 многократно повторяются операции умножения матриц и векторов, при этом R_{kj} — умножение исходной корреляционной матрицы R_h на вектор K_j с компонентами a_{jr}/a_j^2; $K_j M_j'$ и $K_j L_j$ — это перемножение векторов, порождающее скаляр; $M_j / \sqrt{K_j M_j'}$ и $(K_j L_{j-1}) L_j$ — операции деления на скаляр и соответственно умножения скаляра на вектор.

В ходе вычислений L_2 и L_3 были получены результаты, требующие операции зеркального отображения векторов в пространстве и соответствующей смены знаков у величин M_j (см. табл. 7.12, шаги 13—14 и 18—19). Эта простая операция позволила продолжить вычисления максимально правдоподобных оценок без определения мнимого пространства, сохраняя в общем логическую структуру латентных факторов.

Проведенные вычисления только первой итерации показывают трудоемкость реализации метода. В действительности таких итераций выполняются десятки, что конечно же требует привлечения технических средств.

7.8. Вращение пространства общих факторов

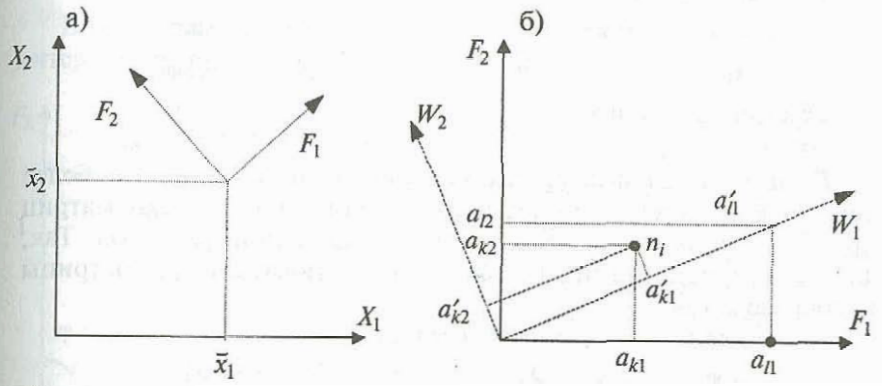
Задача вращения общих факторов решается с целью улучшения их интерпретируемости. Если факторные нагрузки a_{jr} в структуре фактора имеют более-менее равномерное распределение, поиск названия этого фактора затрудняется из-за неявности его особенностей, структурных акцентов. И, наоборот, простая структура фактора, в которой несколько элементарных признаков X_j очевидно доминируют над другими по своей значимости, позволяет определить его название и место в конкретном анализе легко и достаточно надежно. Покажем это на простом примере.

Предположим, имеются данные по результатам факторного анализа (методом главных факторов) до и после вращения пространства главных факторов:

Признак (X_j)	Общие факторы			
	до вращения		после вращения	
	F_1	F_2	W_1	W_2
Обеспеченность основными средствами в расчете на одного среднегодового работника (X_1)	0,463	0,796	0,920	0,087
Уровень энерговооруженности труда (X_2)	0,614	0,411	0,892	0,299
Фондоотдача (X_3)	0,496	0,545	0,410	0,754
Рентабельность продукции (X_4)	0,572	0,378	0,089	0,868
Выработка продукции на одного работника в среднем за год (X_5)	0,723	0,671	0,265	0,570
Уровень реализуемости товарной продукции (X_6)	0,581	0,511	0,372	0,461

Как видим, после вращения в структуре первого фактора наиболее существенные нагрузки сместились к значениям признаков X_1 (обеспеченность труда основными средствами) и X_2 (уровень энерговооруженности труда), в $\lambda_1 = 2{,}025$ их удельный вес теперь занимает более 80%, этот фактор логично назвать «оснащенностью труда средствами производства». В структуре второго фактора наиболее значимыми стали элементарные признаки: X_4 (рентабельность продукции), X_3 (фондоотдача), X_5 (выработка продукции на одного работника) и в меньшей степени X_6 (реализуемость товарной продукции). Они объясняют 95,1% ($\lambda_2 = 1{,}954$) дисперсии второго главного фактора и определяют его название как — эффективность производственной и коммерческой деятельности.

С теоретической точки зрения вращение общих факторов есть уход от неопределенности пространственного расположения факторной координатной системы. Методы аналитической геометрии дают возможность строго устанавливать центр распределения любой m-мерной совокупности данных, но в то же время неоднозначны относительно поворота системы координат и направления координатных осей. При этом решение оптимизируется при условии, что факторные оси проходят через наиболее плотные скопления точек, так что большинство точек распределения падает на оси или располагается в непосредственной близости от начала координатных осей, приобретая наименьшие значения координат (рис. 7.13). При повороте факторного пространства сохраняет силу равенство: $R = AA'$, т.е. $AA' = WW'$.



Центр нормально распределенных данных всегда можно определить, он будет в точке, описываемой вектором средних величин: $\bar{X} = (\bar{x}_1 \bar{x}_2 ... \bar{x}_m)$

Вращение факторного пространства изменяет координаты точек в нем. На рисунке в новой системе координат a'_{k2} значительно меньше a_{k2}, a_{l2} обратилась в нуль, одновременно выросли значения $a'_{k1} > a_{k1}$ и $a'_{l1} > a_{l1}$, в общем $R = AA'$ для $A = \| a_{jr} \|$ и $A = \| a'_{jr} \|$ и $R^+ \rightarrow R$

Рис. 7.13

Как уже сказано выше, вращение общих факторов может быть ортогональным, когда взаимодействие факторов исключается, и косоугольным, порождающим корреляционные связи латентных факторов. При выборе каждого из двух типов вращения требуется последовательное решение таких вопросов:

- Какие вычислительные процедуры следует выполнить для вращения факторного пространства?

- До каких пор (сколько раз) должна повторяться операция вращения?
- Какой угол лучше установить для поворота?

Наиболее простым является ортогональное вращение. Оно производится умножением матрицы факторных нагрузок на некоторую ортогональную матрицу T, задающую угол поворота, размерностью $r \times r$ по числу общих факторов. Поворот может задаваться по или против часовой стрелки, например для матрицы факторных нагрузок A с числом общих факторов $r = 2$:

$$T = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix};$$

Вращение по часовой стрелке

$$T = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

Вращение против часовой стрелки

Если матрица факторных нагрузок содержит данные более чем по двум латентным факторам, строится несколько матриц преобразования T для всех парных комбинаций факторов. Так, для трехмерной матрицы A будут использоваться три матрицы преобразования T:

вращение против часовой стрелки:

$$T_{12} = \begin{pmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}; \quad T_{13} = \begin{pmatrix} \cos\varphi & 0 & -\sin\varphi \\ 0 & 1 & 0 \\ \sin\varphi & 0 & \cos\varphi \end{pmatrix},$$

$$T_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{pmatrix}.$$

Полная матрица преобразования для трехмерного случая будет: $T = T_{12} \times T_{13} \times T_{23}$. Этот подход обобщается на случай, когда число общих факторов $r > 3$, для четырехмерной матрицы A полная матрица преобразования T будет произведением уже шести матриц вращения, для всех пар общих факторов:

$T = T_{12} \times T_{13} \times T_{14} \times T_{23} \times T_{24} \times T_{34}$ и т. д.

При условии ортогонального вращения всегда $T' T = E$.

Пример 7.6. Выполним однократный поворот двумерной матрицы факторных нагрузок, полученной методом главных компонент (п. 7.4.2). Пусть угол поворота будет равен 30° ($\angle\varphi = 30°$):

$$A = \begin{pmatrix} 0{,}776 & -0{,}130 \\ 0{,}904 & -0{,}210 \\ 0{,}616 & 0{,}902 \end{pmatrix}; \qquad T = \begin{pmatrix} \cos 30° & \sin 30° \\ -\sin 30° & \cos 30° \end{pmatrix}.$$

Вращение матрицы факторных нагрузок:

$$W = A \times T = \begin{pmatrix} 0{,}776 & -0{,}130 \\ 0{,}904 & -0{,}210 \\ 0{,}616 & 0{,}902 \end{pmatrix} \times \begin{pmatrix} 0{,}866 & 0{,}500 \\ -0{,}500 & 0{,}866 \end{pmatrix} = \begin{pmatrix} 0{,}737 & 0{,}275 \\ 0{,}888 & 0{,}270 \\ 0{,}082 & 1{,}089 \end{pmatrix}.$$

В результате вращения получена матрица факторных нагрузок, интерпретируемая проще исходной. В составе первой главной компоненты заметно определяющее значение признаков: X_1, X_2, во второй — X_3. Одновременно выполняется требование $AA' = WW'$. После перемножения нормированное пространство искажено, т.е. появляются элементы, большие единицы, и не обязательно $\sum w_{jr}^2 = \lambda_j$. Чтобы вернуться к принятым пространственным соотношениям, достаточно нормализовать матрицу W по известному правилу: векторы W —матрицы факторного отображения после поворота $W_j = \dfrac{W_j}{|W_j|}$.

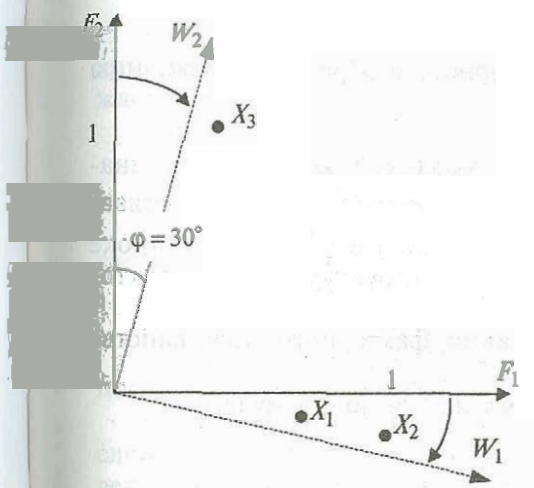


Рис. 7.14. Расположение элементарных признаков в пространстве главных компонент до и после его поворота по часовой стрелке на 30°

Построим график с исходным положением пространства главных компонент до и после поворота. В пространстве главных компонент точками показаны элементарные признаки X_1, X_2, X_3 (рис. 7.14).

Косоугольное вращение матрицы факторных нагрузок может проводиться поочередным вращением каждого фактора на определенный угол или одновременным вращением всех факторов посредством умножения исходной матрицы нагрузок A на неор-

тогональную матрицу поворота, предусматривающую корреляционные связи латентных факторов.

Вопрос достаточности числа поворотов пространства факторов может решаться субъективно, построением графиков распределений наблюдаемых объектов в пространстве повернутых факторов — это наиболее трудоемкий путь, или аналитическими средствами, с использованием специальных критериев для оценки структуры общих факторов. Все критерии конструктивно базируются на представлении величины дисперсии факторных нагрузок как меры сложности структуры факторов. Дисперсия рассчитывается по формуле:

$$D(F) = \frac{1}{r}\sum_{k=1}^{r}(a_{jk}^2 - \bar{a}_{jk}^2)^2,$$

где a_{jk} — элементы матрицы факторного отображения, величины факторных нагрузок;

r — число общих факторов.

Величина дисперсии будет максимальной, когда одно из значений квадратов нагрузок равно общности h_j^2, а в методе главных компонент — единице, и все остальные элементы в строке нулевые. Название критериев качества структуры общих факторов определяет название самого метода вращения.

В ортогональном преобразовании факторного пространства наиболее популярны критерии:

• *Квартимакс* — критерий вычисляется по формуле:

$$q = \sum_{j=1}^{m}\frac{\sum_{k=1}^{r}(a_{jk})^4 - \left(\sum_{k=1}^{r}a_{jk}^2\right)^2}{r^2}.$$

Метод «квартимакс» предусматривает вращение факторных осей таким образом, чтобы величины факторных нагрузок максимизировали q, одновременно учитывается качество структуры всех участвующих в анализе общих факторов.

• *Варимакс* — рассчитывается V_r критерий качества структуры каждого фактора:

$$V_r = \frac{m\sum_{j=1}^{m}a_{jr}^4 - \left(\sum_{j=1}^{m}a_{jr}^2\right)^2}{m^2}.$$

При помощи метода «варимакс» достигают максимального упрощения в описании столбцов матрицы факторного отображения. Возможно раздельное улучшение структуры факторов. Наилучшим здесь также будет максимальное значение критерия, как и в предыдущем случае.

Если в анализе используется косоугольное вращение пространства факторов, то наиболее часто пользуются критериями облимакс, квартимин, облимин.

• *Облимакс* — критерий имеет вид:

$$K = \sum_{j=1}^{m}\sum_{k=1}^{r}a_{jk}^4 \Bigg/ \left(\sum_{j=1}^{m}\sum_{k=1}^{r}a_{jk}^2\right)^2.$$

Решение методом «облимакс» направлено на максимизацию критерия K, оценивается улучшение элементов структуры a_{jk} каждого общего фактора и по всем r факторам.

• *Квартимин* — метод, минимизирующий N-критерий:

$$N = \sum_{f,q=1}^{r}\sum_{j=1}^{m}\left(\sum_{k=1}^{r}a_{jk}p_{kf}\right)^2\left(\sum_{k=1}^{r}a_{jk}p_{kq}\right)^2,$$

где a_{jk} — как и раньше, элементы матрицы факторных нагрузок;

p_{kf} и p_{kq} — элементы матрицы косоугольного преобразования P,

f, g — векторы, задающие вращение k-го общего фактора;

$a_{jk}p_{kf}$ и $a_{jk}p_{kq}$ — элементы матрицы факторного отображения после косоугольного поворота.

• *Облимин* — метод, использующий B-критерий достаточно сложной конструкции, являющейся производной взвешенной величиной от двух других критериев качества косоугольного вращения: квартимина и коваримина —

$$B = \sum_{f,q=1}^{r}\left[m\sum_{j=1}^{m}(a_{jk}p_{kf})^2(a_{jk}p_{kq})^2 - \frac{\beta}{\alpha+\beta}\sum_{j=1}^{m}(a_{jk}p_{kf})^2\sum_{j=1}^{m}(a_{jk}p_{kf})^2\right],$$

где α и β — специальные параметры, определяемые в интервале от нуля до единицы и задающие степень косоугольности системы факторов.

Предположительно (по Кэроллу) часто наилучшие результаты получают при соотношении: $\frac{\beta}{\alpha+\beta} = 0{,}5$.

B-критерий при поиске оптимальной структуры факторов минимизируется.

Как видим, два последних приведенных критерия — квартимин и облимин — построены так, что одновременно в расчет принимаются старые значения координат элементарных признаков a_{jk}, значения элементов, задающих вращение латентных факторов p_{kf} и новые координаты признаков как результат умножения двух предыдущих значений a_{jk} p_{kf}. Оптимизируется структура каждого фактора по составу элементарных признаков (X_j, $j = \overline{1,m}$) и затем осуществляется обобщение на все общие факторы вращаемого пространства ($W_{k(l)}$ k, $l = \overline{1,r}$).

Применение любого из критериев качества структуры общих факторов означает, что после каждого поворота факторного пространства проводятся расчет критерия и соответствующая оценка качества структуры факторов. Вращение завершается, когда критерий, достигнув максимального (минимального) значения, на следующем шаге алгоритма начинает отклоняться от оптимума.

Выбор угла вращения системы факторов представляется наиболее тонким вопросом при улучшении факторной структуры. Часто его решают субъективно, например, после построения графического изображения разброса точек, отображающих расположение элементарных признаков в пространстве общих факторов. Успех решения вопроса при этом во многом будет определяться опытом самого исследователя. Могут применяться и аналитические подходы, например, при ортогональном вращении тангенс наилучшего угла для поворота факторных осей рекомендуется находить по формуле [90]:

$$\operatorname{tg} 4\varphi = \frac{2\sum_{j=1}^{m}(2a_{jk}a_{jl})(a_{jk}^2 - a_{jl}^2)}{\sum_{j=1}^{m}(a_{jk}^2 - a_{jl}^2)^2 - (2a_{jk}a_{jl})^2},$$

где k и l — пара факторов (F_k и F_l) из матрицы факторного отображения A (k, $l = \overline{1,r}$).

Вращение общих факторов дает возможность максимально сократить разброс субъективных суждений относительно названий (образов) общих факторов, неизменно присутствующий в факторном анализе. Тем самым методы факторного анализа из группы в большей степени теоретических методов статистики, ненадежных для аналитических выводов, как бы переходят в

другую группу методов, гибких и позволяющих достаточно достоверно судить о явлениях и процессах.

7.9. Статистическая оценка надежности решений методами главных компонент и факторного анализа

Корректное решение задач при помощи методов факторного анализа предполагает подтверждение значимости исходной матрицы парных корреляций (ковариаций) и достаточности числа обобщенных факторных признаков в анализе.

Значимость корреляционной матрицы подвергается проверке, если принять во внимание, что незначимые корреляционные (ковариационные) связи элементарных признаков не дают оснований вообще для поиска обобщенных признаков. В этом случае все вычислительные собственные числа ($\lambda_1, \lambda_2, ..., \lambda_m$) будут близки единице, и обобщенные факторы по составу становятся очень похожими на элементарные признаки. Проведение факторного анализа теряет смысл.

Проверка значимости матрицы парных корреляций осуществляется при помощи критерия Уилкса — χ^2, его наблюденное значение оценивается по формуле:

$$\chi^2_{\text{н}} = -\left(n - \frac{1}{6}(2m+5)\right)\ln|R|,$$

где R — матрица парных корреляций;

n, m — соответственно число наблюдаемых объектов и число элементарных признаков в анализе.

Расчет χ^2-критерия не представляет трудностей, так как уже на начальной стадии исследования, после получения собственных чисел λ_j определитель $|R|$ легко находится из произведения $\prod_{j=1}^{m}\lambda_j$, т.е. $|R| = \lambda_1 \cdot \lambda_2 \cdot \ldots \cdot \lambda_m$. Наблюденное значение критерия Уилкса сравнивается с табличным, для χ^2-распределения при заданном уровне параметра α и числе степеней свободы $\nu = 1/2\, m\,(m-1)$. Значимость корреляционной матрицы подтверждается при $\chi^2_{\text{н}} > \chi^2_{\alpha,\nu}$.

Критерий χ^2 несколько иной конструкции используется и в оценке достаточности числа выделенных общих признаков (факторов). Для метода главных компонент расчет $\chi^2_н$-критерия осуществляется по формуле Бартлетта:

$$\chi^2_н = -\left(n - \frac{1}{6}(2m+5) - \frac{2}{3}r\right)\ln R_{m-r},$$

здесь r — число оставленных в анализе главных компонент F_r;

$$R_{m-r} = \frac{|R|}{\lambda_1 \cdot \lambda_2 \cdot \ldots \cdot \lambda_r \left(\frac{m-\lambda_1-\lambda_2-\ldots-\lambda_r}{m-r}\right)^{m-r}}.$$

Число степеней свободы для χ^2 будет $\nu = 1/2\,((m-r)-m-r-1)$, наблюденное значение сравнивается с табличным $\chi^2_{\alpha,\nu}$ и, если $\chi^2_н < \chi^2_{\alpha,\nu}$, принимается предположение о том, что выделенные r главных компонент достаточно полно представляют дисперсию m элементарных признаков и остальные главные компоненты: F_{r+1}, F_{r+2}, ..., F_m могут в анализе не рассматриваться из-за незначительного уровня их информативности. Если $\chi^2_н > \chi^2_{\alpha,\nu}$, то в анализе для формирования выводов должны быть введены дополнительно другие главные компоненты.

В факторном анализе для проверки гипотезы о достаточности числа обобщенных признаков (факторов) используется χ^2-критерий Лоули, имеющий аналогичную, как и в предыдущем случае, смысловую нагрузку:

$$\chi^2_н = (n-1)\ln\frac{|R^+|}{|R|},$$

где $|R^+|$ и $|R|$ — определители матриц парных корреляций: воспроизведенной $(R^+ = AA')$ и исходной.

Критическое значение для χ^2-критерия находят по таблицам при заданном уровне значимости α и числе степеней свободы $\nu = 1/2\left((m-r)^2 - m - r\right)$. Предположение о достаточном числе общих факторов подтверждается, когда $\chi^2_н < \chi^2_{\alpha,\nu}$.

Нередко приложение различных методов факторного анализа на одном массиве данных приводит к весьма отличным друг от друга решениям. При этом возникают вопросы: имеются ли в различных решениях близкие по составу общие факторы и действительно ли существенным образом различаются результаты факторных решений или они идентичны? Соответственно ответам на перечисленные вопросы исследователь в дальнейшем останавливается на любом из имеющихся факторных решений или выбирает из них наилучшее. Для оценки сходства различных факторных решений может использоваться коэффициент конгруэнтности:

$$\varphi_{pq} = \frac{\sum_{j=1}^{m} a_{j(p_1)} a_{j(q_2)}}{\left(\sum_{j=1}^{m} a^2_{j(p_1)} \sum_{j=1}^{m} a^2_{j(q_2)}\right)^{1/2}},$$

где p и g — сопоставляемые общие факторы соответственно в первом и втором факторных решениях;

$a_{j(p_1)}$ и $a_{j(q_2)}$ — коэффициенты факторных нагрузок на j-й признак ($j = \overline{1,m}$) в p- и q-факторах при первом и втором факторных решениях.

Коэффициент φ_{pq} изменяется в пределах от -1 до $+1$. Его значения, близкие к -1, указывают на полную и обратную связь факторов, $+1$ — связь полную и прямую, наконец, при $\varphi_{pq} \to 0$ следует предположение об отсутствии связи факторов.

Для оценки адекватности факторной модели в целом может использоваться подход, описанный Харманом [90]. К сожалению, он не содержит рекомендаций по поводу пороговых уровней, скажем неадекватности, но в сравнении с другими приемами значительно легче при реализации и базируется на простой средней оценке расхождений исходных и воспроизведенных коэффициентов корреляции:

$$\sum\sum_{j \neq i}\left(r_{ij} - r^+_{ij}\right)^2 \Big/ m(m-1).$$

Средний квадрат отклонений Хармана исчисляется по всем, кроме диагональных, коэффициентам корреляции. Из нескольких моделей факторного анализа, естественно, будет лучше та, для которой средняя сумма квадратов отклонений окажется наименьшей.

Выводы

Методы главных компонент и факторного анализа базируются на общей идее, что связи элементарных признаков (X_1, X_2, ..., X_m) — это результат воздействия сравнительно небольшого числа неявных, т.е. латентных, факторов (F_1, F_2, ..., F_r). Метод главных компонент гипотетически позволяет объяснить всю, на 100%, вариацию X_j через латентные факторы. Классические методы факторного анализа допускают наличие нескрываемой характерности признаков и изначально ориентированы на объяснение не всей, а только определенной части вариации элементарных признаков, остающейся за вычетом характерности.

Поскольку число латентных факторов r обычно значительно меньше числа элементарных признаков m, основной задачей факторного анализа считаются сжатие анализируемого признакового пространства и переход от массива исходных данных размерностью $n \times m$ к аналитическим данным в матрицах факторного отображения ($m \times r$) и значений общих факторов ($n \times r$) при $r < m$.

Сложной и всегда с неоднозначным решением остается задача определения достаточности числа и интерпретации общих факторов. Результат принимается субъективно, в определенной мере интерпретация облегчается возможностью вращения факторного пространства R^F и получения более простой структуры факторов.

Наиболее часто встречающейся, тривиальной является ортогональная система факторов. Возможно также первичное получение, или в результате вращения линейно-зависимых факторов, не ортогональной, а косоугольной их системы. В этом случае общие факторы могут рассматриваться как признаки первого уровня, и для них вновь решаема задача факторизации, с выделением теперь уже надфакторов или факторов второго уровня. Реальный подход в анализе позволяет выделять общие факторы до третьего уровня.

В истории развития факторного анализа обнаруживаются три периода: формирования его принципиальной теоретической базы — от начала нашего века и до 30-х годов, активной разработки разнообразных подходов и методов простого факторного анализа (однофакторного, бифакторного, центроидного и т.д.) — 30—50-е годы и, наконец, появления гибких, адаптационных методов факторного анализа (максимального правдоподобия, минимальных остатков, альфа-факторного анализа и т.п.) — с середины 60-х годов. В настоящее время теория факторного анализа продолжает развиваться. Перспективными считаются направления разработки нелинейных моделей факторного анализа и методов, которые отвечали бы интуитивному представлению исследователя о факторах. Ярким представителем в современном факторном анализе является конфирматорный метод. Теоретически он базируется на предположении о существовании некоторой заранее определенной, принимаемой исследователем в качестве гипотезы, структуры ожидаемой матрицы факторных нагрузок. Структурная гипотеза затем подвергается статистической проверке и нагрузки уточняются. Постепенно, используя вращение системы факторов, исследователь приходит к матрице нагрузок, адекватно представляющей структуру исходной корреляционной матрицы.

Вопросы и задачи

1. В чем состоит различие понятий «общий фактор» и «элементарный признак»?

2. Какие преимущества получает исследователь с переходом от анализа признаков к анализу общих факторов?

3. Как определить достаточное число факторов для характеристики изучаемого явления или процесса?

4. Какие особенности должны быть учтены, если для проведения анализа выбирается метод главных компонент или один из таких методов факторного анализа, как однофакторный, бифакторный, центроидный, или адаптационных методов (максимального правдоподобия, минимальных остатков)?

5. Что подразумевается под ортогональным и косоугольным факторными решениями?

6. Какие алгоритмические шаги выполняются при реализации:
а) метода главных компонент?
б) методов факторного анализа?

7. Чем объясняется возможность факторного отображения множества элементарных признаков?

8. Что означает «простая структура» фактора? Почему и какими средствами добиваются «простой структуры»?

9. Имеются данные обследования десяти промышленных предприятий по двум характерным признакам: численности работников, тыс. чел. (X_1) и размеру получаемой прибыли, млн. руб. (X_2):

Предприятие	X_1	X_2	*Предприятие*	X_1	X_2
1	2,8	38,5	6	3,4	29,6
2	4,7	36,2	7	8,0	26,7
3	10,9	49,4	8	2,3	47,6
4	1,5	23,8	9	11,4	56,0
5	6,4	58,7	10	0,5	69,7

Представьте результаты статистического обследования в двумерной системе координат, предположительно покажите векторы, которые обобщали бы значения признаков X_1, X_2.

10. Для корреляционной матрицы R четвертого порядка вычислите, используя рекуррентные соотношения Фаддеева, определитель, постройте характеристическое уравнение:

$$R = \begin{pmatrix} 1 & & & \\ 0{,}400 & 1 & & \\ 0{,}530 & -0{,}025 & 1 & \\ 0{,}480 & 0{,}050 & 0{,}080 & 1 \end{pmatrix}.$$

11. По данным опроса практиков-экономистов построена матрица корреляционной зависимости R характерных признаков: X_1 — уровень оплаты труда; X_2 — возраст; X_3 — трудовой стаж:

Признак	X_1	X_2	X_3
X_1	1	−0,388	0,665
X_2	−0,388	1	0,740
X_3	0,665	0,740	1

Проведите анализ данных матрицы парных корреляций R методом главных компонент, определите уровень информативности каждой из главных компонент и ее признаковый состав. Покажите распределение элементарных признаков в пространстве двух первых главных компонент (F_1 и F_2).

12. Осуществите известными способами переход от обычной к редуцированной матрице парных корреляций, если имеется следующая исходная матрица парных корреляций R:

$$R = \begin{pmatrix} 1 & 0{,}740 & 0{,}565 & 0{,}875 & 0{,}362 & 0{,}730 \\ 0{,}740 & 1 & 0{,}478 & 0{,}690 & 0{,}802 & 0{,}558 \\ 0{,}565 & 0{,}478 & 1 & 0{,}672 & 0{,}475 & 0{,}650 \\ 0{,}875 & 0{,}690 & 0{,}672 & 1 & 0{,}525 & 0{,}820 \\ 0{,}362 & 0{,}802 & 0{,}475 & 0{,}525 & 1 & 0{,}539 \\ 0{,}730 & 0{,}558 & 0{,}650 & 0{,}820 & 0{,}539 & 1 \end{pmatrix}.$$

13. По известной матрице факторных нагрузок A воспроизведите матрицу парных корреляций R^+:

$$A = \begin{pmatrix} 0{,}76 & 0{,}42 \\ 0{,}45 & 0{,}21 \\ 0{,}65 & 0{,}37 \\ 0{,}38 & 0{,}19 \end{pmatrix}.$$

14. Заполните таблицу недостающими данными:

Компоненты дисперсии	*Условное обозначение*	*Варианты задач*					
		1	2	3	4	5	6
Общность	h^2	0,84	—	0,75	—	—	—
Характерность	d^2	—	0,40	—	—	—	0,18
Специфичность	b^2	—	—	—	0,24	0,30	—
Дисперсия ошибки	l^2	—	0,20	0,10	—	0,25	—
Надежность	v^2	0,90	—	0,90	0,60	—	0,84

15. Используя данные задачи **11**, найдите факторное решение методом главных факторов (для получения собственных чисел достаточно возведения матрицы парных корреляций в четвертую степень).

16. Имеется матрица значений общих факторов по шести типам промышленных товаров:

$$F = \begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ n_5 \\ n_6 \end{array} \begin{array}{c} \begin{array}{cc} F_1 & F_2 \end{array} \\ \begin{pmatrix} -2{,}245 & -0{,}711 \\ 0{,}499 & 1{,}680 \\ -0{,}402 & 0{,}060 \\ 1{,}155 & 0{,}524 \\ 1{,}048 & 0{,}675 \\ -0{,}055 & -2{,}228 \end{pmatrix} \end{array}.$$

Известны названия главных факторов: F_1 — качество товара; F_2 — уровень обслуживания покупателя.

Покажите распределение товаров в пространстве двух общих факторов. Можно ли выделить однородные по F_1 и F_2 группы товаров?

Постройте линейную регрессионную модель, в которой y — это удельный вес продаж каждого вида товара из всей партии в течение первых 15 дней реализации:

$$Y' = (0{,}10 \quad 0{,}42 \quad 0{,}13 \quad 0{,}82 \quad 0{,}63 \quad 0{,}21),$$

факторные признаки регрессионной модели: F_1 и F_2.

17. Имеются матрица факторных нагрузок A и матрица ортогонального вращения T, с целью улучшения структуры общих факторов выполните поворот системы векторов матрицы A.

$$A = \begin{pmatrix} 0{,}86 & -0{,}25 \\ 0{,}61 & -0{,}31 \\ 0{,}44 & 0{,}51 \\ -0{,}47 & -0{,}28 \\ 0{,}38 & 0{,}08 \end{pmatrix}; \qquad T = \begin{pmatrix} 0{,}574 & 0{,}819 \\ -0{,}819 & 0{,}574 \end{pmatrix}.$$

Ответьте на дополнительные вопросы:

а) Матрица T — матрица ортогонального или косоугольного вращения?

б) Само вращение задается против или по часовой стрелке? Какой принят угол поворота?

Глава 8

Многомерное шкалирование

8.1. Многомерное шкалирование в статистических исследованиях

Методы многомерного шкалирования (МШ) разрабатывались и применяются в практике для исследований сложных явлений и процессов, не поддающихся непосредственному описанию или моделированию. В основу теории многомерного шкалирования положена идея о возможности развертывания наблюдаемых объектов в некотором теоретическом пространстве, адекватно отображающем реальность.

В отличие от других статистических методов поиск координатного пространства в МШ осуществляется не по значениям самих характеризующих объекты признаков, а по данным, представляющим различия, или, наоборот, сходство этих объектов. Основным источником данных здесь являются в одних случаях эксперты, субъективно воспринимающие и оценивающие относительное расположение объектов наблюдения в реальных условиях, в других — результаты прямой регистрации сведений о состоянии и поведении объектов. Тривиально и больше распространено экспертное оценивание.

Цель аналитической работы с данными — определение местонахождения объекта в «пространстве восприятия (субъектов)» и создание его образа. Имеется в виду, что непосредственно о самом объекте даже по значениям некоторого набора признаков нельзя судить достаточно надежно или полно. В то же время эксперты или просто наблюдатели еще до проведения аналитических расчетов видят, интуитивно чувствуют различия изучаемых объектов. Неосознанные, нечеткие представления об объектах должны быть конкретизированы и это осуществимо в теоретиче-

ском «пространстве восприятия», построенном по субъективным оценкам. В этом представляемом пространстве проявляют себя латентные факторы, становится очевидным действие этих факторов на пространственное расположение объекта, измеримо расстояние между объектами.

Рис. 8.1. Конфигурация городов в двумерном пространстве

Чтобы сущность и цели МШ стали более понятными, рассмотрим несколько графических примеров.

На рис. 8.1 показан простейший пример шкалирования. Десять белорусских городов представлены в двумерном пространстве с координатными осями: юг—север и запад—восток. Географические оси, как известно, материально не существуют, но они отображают реальность. Конфигурация городов позволяет создать определенный образ каждого из них, представив его территориальное расположение, климатические особенности, связи с другими городами и т.д.

На рис. 8.2 в пространстве двух шкал: трудоемкость и капиталоемкость производства, расположены отдельные предприятия различных отраслей. Такая информация могла бы быть полезна, например, при выборе вида предпринимательской деятельности. Очевидно, что в сельскохозяйственном производстве (D) следует считаться с необходимостью больших затрат живого труда. Для производства экологической очистки отходов (B) необходимы крупные первоначальные вложения финансовых средств, высокий уровень автоматизации процессов переработки отходов позволяет впоследствии иметь сравнительно низкие затраты на рабочую силу. Объекты A и C — соответственно предприятия торговли и услуг по ремонту бытовой техники, предполагают примерно равные затраты капитала и рабочей силы. В торговой сфере, однако, начало деятельности может обусловливаться незначительными производственными затратами, тогда как оказание услуг по ремонту техники в любом случае требует сравнительно больших вложений денежных средств и наличия квалифицированной рабочей силы.

Рис. 8.2. Конфигурация промышленных предприятий в двумерном стимульном пространстве

Рис. 8.3. Конфигурация претендентов на должность менеджера фирмы в трехмерном пространстве

Рис. 8.3 после обсуждения рис. 8.1 и 8.2 становится понятен. Очевидно, при приеме на работу наиболее предпочтительны шансы для кандидатуры 5, субъект занимает достаточно стабильное положение в обществе, имеет высокий уровень профессиональной подготовки, обладает неплохим здоровьем, близким по показателям к среднему уровню. В общем расположение кандидатов в трехмерном теоретическом пространстве позволяет сделать выводы о достоинствах и недостатках каждого из них, получить точные характеристики по значениям трех обобщающих признаков (шкал), определить различия кандидатов по всем трем, двум или только по одному общему признаку в зависимости от выбора исследователя.

Следует обратить внимание, что на рисунках во всех случаях мы имели дело с обобщающими латентными признаками. Размерность пространства не превышает трех. Последнее требование необязательно, однако к нему в МШ прибегают часто, чтобы предусмотреть визуализацию аналитических данных.

В МШ имеются особые теоретические подходы, немало специфических понятий, обозначений, их логика становится понятной, если обратиться к истории.

Мысль о возможности шкалирования наблюдаемых объектов появилась довольно давно. Еще И. Ньютон в 1704 г. показал,

что спектральные цвета можно представить на окружности. Различия цветов воспроизводились при этом расстояниями между точками окружности. Позже, в XIX — начале XX вв. были сделаны попытки наглядного отображения различий звуковых сигналов (1846 г., Дробиш) и звуковых ощущений (1916 г., Хемминг). Первые теоретические обоснования методов и алгоритмов МШ были даны американскими учеными: М.В. Ричардсоном и Л.Л. Терстоуном (20—30-е годы), их работы проводились в рамках исследований в области психологии человека.

В 50—60-е годы разработка теории МШ велась особенно активно и широко, самый живой интерес был проявлен со стороны представителей других наук: математики, социологии, экономики и др. В это же время выделилось метрическое и неметрическое многомерное шкалирование. *Метрическое* МШ основывалось на использований количественных признаковых характеристик объектов, для его реализации У. Торгерсоном был предложен один из наиболее известных сегодня алгоритмов. *Неметрическое* МШ предназначалось для обработки неколичественных, ранговых (или порядковых) данных. Заслуга разработки методов этой группы принадлежит ученым: Р. Шепарду и Дж. Краскалу.

В начале 70-х годов появляются первые работы русских авторов: А.Ю. Терехиной, С.А. Клигера, В.С. Каменского, Г.А. Сатарова.

В настоящее время методы МШ продолжают бурно развиваться в разделах математики и теории анализа данных. Они остаются наиболее гибкими для отображения реальных процессов, но их применение предъявляет самые высокие требования к интеллектуальному потенциалу исследователя, так как его знания, творческие способности во многом определяют оптимальный выбор и успешную реализацию вычислительных процедур МШ, корректность интерпретации аналитических результатов.

Традиционно в МШ применяются понятия, которые не привычны или вовсе не встречаются в разделах классической статистики:

Стимул — имеет самое широкое толкование, подразумевают-ся некоторые свойства объекта, его признак, качественная или количественная, но непосредственно не измеряемая характеристика изучаемого объекта. Можно сказать, что это некоторый образ объекта.

В аналитической практике стимул — это обычно объект, обладающий определенным набором присущих ему характеристик.

Встречаемое понятие стимульного пространства означает теоретическое пространство с погружением в него стимулов.

Шкала — ось теоретического пространства, которая является носителем значений обобщенного признака (фактора). Координатные оси определяются на завершающем этапе исследований, как правило, известными методами главных компонент или факторного анализа. Одновременное шкалирование — обычно весьма абстрактное примитивное представление реальных объектов: как, например, систематизация данных о работниках по состоянию их здоровья или промышленных предприятий — по их размеру. Действительный мир явлений более сложен и разнообразен, для его объективного отображения требуется многомерное пространство шкал — *многомерное шкалирование.*

Суждения в нашем примере окажутся более правильными и ценными, если для работников будут известны дополнительно оценки их социального положения, интеллектуальных способностей и т.д., для предприятий — уровень эффективности их работы, развитие коммуникационных связей, уровень организационной работы и т.д.

В определении шкал обычно возникают проблемы: поиска правильного «наименования», т.е. идентификации шкалы, принятия решения о достаточности числа шкал в анализе и интерпретации пространственного расположения объектов в шкальном пространстве — их разрешение полностью находится в компетенции исследователя.

Как это чаще всего принято в МШ, будем обозначать номер шкалы символом k (т.е. шкала X_k).

Эксперт — понятие не нуждается в подробном разъяснении. Это субъект S, который с высоким профессионализмом дает суждение о некотором объекте, явлении, процессе. В нашем случае возможен и менее строгий подход, эксперт — также простой наблюдатель, регистрирующий поведение объекта.

Предпочтение — суждение об объекте с точки зрения его близости представляемому идеалу. Величина предпочтений может возрастать или уменьшаться.

Стресс-формулы — формулы, применяемые для оценки соответствия эмпирических и теоретических ранговых данных.

К настоящему времени известным становится все большее число методов МШ и чтобы правильно пользоваться ими на практике, необходима их общая классификация. Выше уже ска-

зано, что в зависимости от характера исходных данных все методы МШ легко разделяются на две большие группы: метрические и неметрические. Возможна также другая их группировка, с учетом того, *что* выступает объектом исследования:

- *анализ стимулов* — изучение и построение образа объекта (объекта в обычном понимании слова: предприятия, субъекта, комплекса, признака);
- *анализ индивидуальных различий* — изучение особенностей субъективного восприятия стимулов;
- *анализ предпочтений* — изучение стимулов относительно представлений об идеале. Здесь возможно совмещение решаемых задач с анализом стимулов и индивидуальных различий;
- *анализ идеальных точек* — поиск и описание идеального положения стимулов.

При всем многообразии анализа в МШ имеются общие формальные модели как теоретическая основа для разработки алгоритмов вычислительных процедур. Существует два основных типа таких моделей: дистанционные и векторные. В *дистанционных моделях* исходные различия описываются расстояниями в теоретическом шкальном пространстве, при этом в большинстве случаев для оценки расстояний используют евклидову метрику:

$$\hat{d}_{ij} = \left(\sum_k (x_{ik} - x_{jk})^2 \right)^{1/2}.$$

В *векторных моделях* присутствуют меры близостей или связей — величины, обратные характеристикам различий, они аппроксимируются скалярными произведениями векторов, соединяющих начало координат с точками расположения стимулов:

$$\hat{b}_{ij} = \sum_k x_{ik} x_{jk},$$

где x_{ik}, x_{jk} — координатные значения стимулов в шкальном пространстве $\boldsymbol{R}^k$;

$\hat{d}_{ij}$, $\hat{b}_{ij}$ — теоретически оцененные меры различия и близости стимулов.

В ходе анализа методами МШ исследователь решает вопросы, важные для обобщения многомерных данных:

- поиска и интерпретации латентных переменных — общих факторов;

- сжатия исходного массива данных;
- визуализации геометрической конфигурации наблюдаемых объектов в координатном пространстве латентных признаков.

8.2. Представление и первичная обработка статистических данных в многомерном шкалировании

В МШ исследовательская работа уже на начальном этапе, при организации статистического наблюдения и регистрации данных, проводится особым образом. Основным источником исходной информации здесь выступают эксперты или просто субъекты, оценивающие предъявляемые им некоторые стимулы. Поступающие сведения обобщаются в квадратного или прямоугольного вида матрицы с названиями, соответствующими методике получения данных и их характеру. В последующем определяются различия (или сходства) наблюдаемых объектов.

Матрица условных вероятностей, или *матрица идентификаций*, определяется относительными данными «по узнаванию стимулов». Строки такой матрицы представляют собой перечень стимулов, предъявляемых для оценки, столбцы — стимулов, распознанных экспертами. Данные матрицы, таким образом, с одной стороны, позволяют выявить ошибки в распознавании стимулов, а с другой, идентифицировать их, если сами стимулы нечетки и необходимо внести определенность в их классификацию. На рис. 8.4 показано, как могут выглядеть матрицы подобного типа. В случае а) предположительно экспертами (скажем, профессиональными экономистами) оценивались четыре рынка автомобилей (А, В, С, Д), расположенных в разных городах и имеющих различные условия торговли: А — широкий выбор автомобилей (по маркам, уровню цен и т.д.), разнообразные сервисные услуги, в целом наиболее благоприятные условия для совершения покупок; В — рынок, специализированный на автомобилях одной марки, но предоставляющий разнообразные услуги, в общем имеет благоприятные условия для покупателя; С — ограниченный выбор автомобилей, суженный круг услуг, незначительные расхождения в ценах на автомобили, Д — рынок с определенно неблагоприятными условиями для покупате-

ля, где к продаже представлено небольшое количество автомобилей, сервисные услуги ограничены и оказываются неквалифицированными работниками, имеется жесткая схема оплаты за покупку, скажем, только наличным расчетом. Свои ответы эксперты должны были дать по ранее сложившимся представлениям после чтения прессы и общения с покупателями, им назывался рынок и его следовало отнести к классу А, В, С или Д.

	А	В	С	Д
А	0,70	0,25	0,05	0,00
В	0,30	0,50	0,15	0,05
С	0,10	0,40	0,40	0,10
Д	0,02	0,03	0,20	0,75

а)

	E	F	G	H	I
E	0,20	0,60	0,10	0,02	0,08
F	0,65	0,20	0,05	0,10	0,00
G	0,10	0,05	0,70	0,00	0,15
H	0,00	0,10	0,00	0,85	0,05
I	0,05	0,05	0,40	0,20	0,30

б)

Рис. 8.4. Матрицы условных вероятностей с результатами экспертного оценивания автомобильных рынков и производственных фирм

По данным рис.8.4а легко установить удельный вес неправильных ответов по каждому рынку, т.е. долю «нераспознанных объектов». Не менее важный вывод следует о степени сходства наблюдаемых рынков. Так, очевидно, что эксперты выделяют схожие по условиям сбыта рынки А и Б и почти не смешивают их с рынками С и Д. Рынок Д, имеющий характерные негативные черты, наиболее четко представляется экспертам и при его узнавании допускается наименьшее число ошибок.

В случае рис. 8.4б эксперты должны определить фирмы, эффективно осуществляющие деятельность при наиболее рациональном использовании трудовых ресурсов E, а также с высокой отдачей материальных ресурсов F и капитала G, с оптимальной структурой управления H, активно использующей новшества НТП Y. Понятно, что успешно работающей фирме в той или иной мере должны быть присущи все перечисленные выше характеристики (E, F, G, H, Y), что затрудняет четкое выделение «типа» с преобладающим влиянием какого-либо из признаков. В данном случае матрица условных вероятностей поможет при

идентификации предполагаемых типов фирм и в то же время укажет на реально существующие связи между признаками, представляющими различные стороны одного процесса, а именно эффективности производства.

В числе матриц условных вероятностей можно выделить *матрицы перехода*. Свое название они получили от того, что их данные отражают трансформационные процессы, протекающие во времени или пространстве. Строки матрицы перехода дают представление об изучаемом явлении в момент начала, столбцы — в момент окончания чего-либо, или то же, но при переходе от одной к другой территориальной единице и т. п.

Предположим, для суждения о новом налоговом законе проводилось два выборочных опроса руководителей предприятий: до и через год после введения нового закона. Участникам опроса предложено выбрать предпочтительные варианты дальнейших тактических шагов для органов государственного управления:

K — налоговый закон должен оставаться в силе;

L — закон требует уточнения основополагающих положений;

M — закон наносит существенный ущерб предпринимательской инициативе и должен быть отменен;

N — закон должен быть приведен в соответствие налоговому законодательству развитых промышленных стран Западной Европы.

Гипотетическая матрица перехода в таком случае может принять вид:

i \ *j*	K	L	M	N
K	0,30	0,50	0,05	0,15
L	0,35	0,45	0,10	0,10
M	0,40	0,35	0,10	0,15
N	0,40	0,45	0,05	0,10

Рис. 8.5. Матрица перехода с оценками перспективности введения нового закона о налогах

Каждый элемент матрицы p_{ij} представляет долю руководителей предприятий в начале года, до введения закона, высказавших мнение *i*, а по завершению года — *j*, т.е. для первой

строки, например, 30% опрошенных по-прежнему придерживаются своего мнения, что принимаемый закон достаточно совершенен и должен оставаться в силе; 50% хотя и поддерживают закон в целом, как и прежде, но теперь видят необходимость в уточнении и конкретизации его положений, а 5% изменили свое мнение на противоположное и утверждают, что новый закон может нанести ощутимый ущерб экономике. Данные матрицы перехода делают очевидными анкетные вопросы, необходимость работы по которым признает абсолютное большинство экспертов, в нашем случае это пункт L. Хотя и в меньшей мере, но также обращается внимание на важность приведения закона в соответствие с аналогами индустриально развитых стран Западной Европы, через год после принятия закона эту точку зрения поддерживают от 10 до 15% экспертов.

Наконец, представляется возможным выделить пары стимулов, наиболее похожих друг на друга по частоте обращения к ним экспертов. Это K и L, затем K и N и M и N.

В приведенных выше трех примерах матриц условных вероятностей можно отметить следующие конструктивные особенности: все матрицы квадратные, итоги по строкам представляют общее количество экспертных оценок, т.е. 100%, или 1,0 и, что важно, эти матрицы несимметрические (в них $p_{ij} \neq p_{ji}$ [1]). Последнее видится правомерным и отвечает представлению экспертов о стимулах и их взаимосвязях. Скажем, нет ничего необычного в том, что разбросом мнений экспертов допускается «вероятность» 0,35 перехода от полного отрицания нового налогового закона к компромиссному решению по его доработке и принятию, а вероятность решения, что предложенный к доработке закон нужно будет заблокировать, — лишь 0,05. Не отрицая в общем логичности таких данных, напомним, что расхождения вероятностных показателей и их асимметричное расположение в матрице могут порождаться также и обычными ошибками при распознавании стимулов.

В любом случае асимметричные данные методами МШ не обрабатываются и не могут в последующем быть размещены в k-мерном шкальном пространстве. Поэтому существуют простые

[1] В общем это неравенство выдерживается совершенно необязательно для матриц условных вероятностей, формально они вполне могут быть симметрическими. Дело здесь скорее в технологии проведения экспертного оценивания.

методы преобразования асимметричных матриц условных вероятностей в симметрические. Для этого от характеристик сходств p_{ij} переходят к новым мерам σ_{ij}, где σ_{ij} — например, сумма элементов матрицы условных вероятностей, одинаково удаленных от главной диагонали ($\sigma_{ij} = p_{ij} + p_{ji}$)[1]. Одновременно на главной диагонали значения p_{jj} нивелируются и ставятся прочерки, тем самым данные полностью приводятся в соответствие аксиоматичному представлению о расстоянии между парой объектов в пространстве, когда $d(a, a) = 0$, $d(a, b) > 0$ и $d(a, c) + d(c, b) \geq d(a, b)$.

Построим симметрическую матрицу на примере данных матрицы условных вероятностей (см. рис. 8.4 б). Чтобы новые вероятностные оценки не превысили предельного в этом случае значения 1,00, исчислим их как простые средние величины:

0,20	0,60	0,10	0,02	0,08
0,65	0,20	0,05	0,10	0,00
0,10	0,05	0,70	0,00	0,15
0,00	0,10	0,00	0,85	0,05
0,05	0,05	0,40	0,20	0,30

$$\delta_{ij} = \frac{p_{ij} + p_{ji}}{2} \rightarrow \Delta =$$

—	0,63	0,10	0,01	0,06
0,63	—	0,05	0,10	0,03
0,10	0,05	—	0,00	0,23
0,01	0,10	0,00	—	0,13
0,06	0,03	0,23	0,13	—

Нужно признать, что использование матриц условных вероятностей и необходимость последующих преобразований с целью симметризации данных в момент перехода от элементов p_{ij} к b_{ij} связаны с неизбежной потерей части информации или ее некоторым искажением, отступлением от реальных асимметричных соотношений.

Другой вид матриц для обобщения исходных данных, широко применяемых в МШ, — это *матрица совместных вероятностей*. Название говорит само за себя, в такой матрице элементы p_{ij} характеризуют удельный вес взаимодействий стимулов i и j. Таким образом, здесь строки и столбцы содержат согласованные значения, а сама матрица будет симметрической.

[1] Очевидно, могут быть предложены и другие методы для определения δ_{ij}, например, со взвешиванием элементов p_{ij} на величины, отражающие значимость соответствующих элементов на главной диагонали, или $\delta_{ij} = \max(p_{ii}, p_{jj})$ и т.д.

Разнообразие приемов, приводящих к построению матрицы совместных вероятностей, велико и, пожалуй, соответствует многообразию исследовательских целей и индивидуальных профессиональных подходов в решении конкретных задач сводки статистических данных. Одним из примеров может служить изучение взаимосвязи показателей экологичности работы промышленных предприятий и заболеваемости населения. В результате опроса экспертов (прямой регистрации данных) о вредности 50 различных производств могут быть получены оценки частоты встречаемости некоторого вида заболевания i и выброса какого-либо из вредных веществ j во внешнюю среду.

Наиболее простым для восприятия и в общем тривиальным приемом обобщения исходных данных в МШ является построение матриц *мер различия профилей*. Под профилем V понимается простой набор количественных признаков объекта. Матрица, как правило, имеет прямоугольную форму и строится таким образом, что ее строки — это перечень объектов наблюдения, столбцы — характерные признаки. Набор количественно определенных данных по j признакам для i-го объекта строки будет профилем оценок объекта i. В табл. 8.1 а) и б) приведены примеры матриц мер различия профилей двух вариантов: в одном аналитические признаки — это статистические показатели результатов производственной и коммерческой деятельности, в другом — временные интервалы.

Таблица 8.1. **а) Данные об обороте капитала, уровне доходности деятельности и численности работников четырех крупнейших компаний Европы, за 1993 г.**

Признак / *Объект*	*Оборот капитала, млрд. долл. США*	*Прибыль, млн. долл. США*	*Количество работников, тыс. чел.*
МАН (Германия)	11,8	429,6	63,4
СЕБ (Франция)	1,5	88,8	10,1
Даниска (Дания)	2,0	124,6	11,5
Нокиа (Финляндия)	3,1	−17,1	26,8
Средние значения	5,1	214,3	28,3

б) Динамика уровня рентабельности продукции предприятий по отраслям промышленности Республики Беларусь

Признак / *Отрасль*	*1980*	*1985*	*1990*	*1993*
Электроэнергетика	33,1	22,7	18,6	−1,3
Химическая промышленность	51,9	30,3	34,4	39,1
Топливная промышленность	25,1	15,7	19,5	39,1
Промышленность строительных материалов	8,5	17,1	15,6	16,4
Легкая промышленность	12,0	11,0	10,6	35,4

Из табл. 8.1 а) видно, что профили могут представлять ряды оценок различной физической природы, в этом случае, как известно, на начальном этапе исследований решается задача по стандартизации исходных данных. В МШ наиболее часто используются варианты стандартизации:

$$z_{ij} = \frac{v_{ij} - \bar{v}_j}{\delta_j}, \quad z_{ij} = v_{ij} / \bar{v}_j, \quad z_{ij} = \log v_{ij}.$$

Для определения матрицы различий Δ по стандартизованным данным находят меры различий профилей, другими словами, определяют расстояния между профилями. При этом из большого числа метрических формул особенно часто используется евклидово расстояние:

$$\delta_{ij}{}^{E} = \left(\sum_k (v_{ik} - v_{jk})^2 \right)^{1/2}$$

или величина, равная квадрату евклидова расстояния:

$$\delta_{ij}{}^{E} = \sum_k (v_{ik} - v_{jk})^2 .$$

Но возможно использование и других известных метрик: city-block, Минковского и др.

Следует обратить внимание, что в таблице мер различия профилей в качестве объекта для исследований методами МШ могут выступать учетные единицы, расположенные в строках (i-е объекты), или столбцах (признаки j-х наблюдаемых объектов). Это означает, что можно построить стимульное пространство

для развертывания в нем четырех европейских компаний, и наоборот, признаковое пространство на главных факторах (или компонентах) с погружением в него самих аналитических признаков — все зависит от конкретных целей исследователя.

Чтобы показать расчет матрицы различий Δ для мер профилей, остановимся на постановке задачи, обычной в МШ, когда различия объектов определяют по значениям аналитических признаков и на завершающем этапе получают пространственное расположение объектов. Вычисления проведем по данным табл. 8.1 а, предварительно стандартизировав их по варианту $z_{ij} = v_{ij} / \bar{v}_j$:

$z_{11} = 2{,}56;$ $z_{21} = 2{,}74;$
$z_{12} = 0{,}33;$ $z_{22} = 0{,}56;$
$z_{13} = 0{,}43;$ ..... и т.д.
$z_{14} = 0{,}67;$

Исчисленные стандартизованные значения анализируемых признаков представим в матрице:

$$Z = \begin{pmatrix} 2{,}56 & 2{,}74 & 2{,}27 \\ 0{,}33 & 0{,}56 & 0{,}36 \\ 0{,}43 & 0{,}80 & 0{,}41 \\ 0{,}67 & -0{,}11 & 0{,}96 \end{pmatrix}.$$

Для построения матрицы различий Δ остается рассчитать расстояния между всеми парами наблюдаемых объектов и свести результаты теперь уже в матрице симметричного вида. В качестве меры расстояния выберем тривиальную метрику Евклида d^E:

$$\delta_{1,2} = \left(\sum_k (v_{1k} - v_{2k})^2 \right)^{1/2} = \left[(2{,}56 - 0{,}33)^2 + (2{,}74 - 0{,}56)^2 + \right.$$

$$\left. + (2{,}27 - 0{,}36)^2 \right]^{1/2} = 3{,}657;$$

$$\delta_{1,3} = \left((2{,}56 - 0{,}43)^2 + (2{,}74 - 0{,}80)^2 + (2{,}27 - 0{,}41)^2 \right)^{1/2} = 3{,}429 \ldots$$

Упорядочив исчисленные оценки расстояний между всеми парами объектов, получим матрицу различий

$$\Delta = \begin{pmatrix} 0 & 3{,}657 & 3{,}429 & 3{,}662 \\ & 0 & 0{,}265 & 0{,}961 \\ & & 0 & 0{,}116 \\ & & & 0 \end{pmatrix}.$$

По элементам матрицы Δ можно заметить, что компания МАН по своим параметрам резко отличается от других компаний и имеет с ними наименьшее сходство. Наибольшим сходством отличается пара компаний: СЕБ и Даниска.

После того как становится известна матрица различий, приступают к выполнению шагов алгоритма многомерного шкалирования.

Таким образом, в МШ имеются две возможности для общего представления входной информации. В первом случае используют матрицу вероятности, при этом необходимо следить, чтобы на завершающем этапе работы с исходными данными была получена симметрическая матрица. Если это требование не выдерживается, то осуществляется такое преобразование элементов имеющейся матрицы, чтобы с наименьшей потерей информативности и наименьшей степенью отступления от реального распределения данных добиться их симметричности.

Во втором случае объекты с набором характерных признаков представляются как меры различий профилей. Обычно это связывается с дополнительным решением задач нормирования данных и исчисления мер различий (сходств) между парами наблюдаемых объектов. Одной из центральных проблем здесь становится правильный выбор методики нормирования и метрики расстояния. Результатом же всегда будет симметрическая матрица различий Δ, готовая для обработки методами МШ.

8.3. Классическая модель многомерного шкалирования Торгерсона

Американским статистиком У. Торгерсоном в начале 50-х годов предложен один из первых алгоритмов МШ, впоследствии известный как метрический метод Торгерсона. С его помощью легко представить общую логику МШ. Имея простую формальную конструкцию, он отличается наглядностью и легко воспринимается.

Теоретически метод Торгерсона базируется на жестких гипотетических предположениях:

- в некотором определенном шкальном пространстве X расстояния между наблюдаемыми объектами соответствуют величинам, характеризующим их различия, т.е. $\delta_{ij} = d_{ij}$;

• сами расстояния между объектами в теоретическом пространстве достаточно точно описываются метрикой Евклида:

$$\delta_{ij} = d_{ij} = \left(\sum_k (x_{ik} - x_{jk})^2 \right)^{1/2};$$

• в шкальном пространстве X средние значения координат стимулов по каждой оси равны нулю, нуль — исходная точка отсчета:

$$\sum_i x_{ik} = \sum_j x_{jk} = 0 .$$

Алгоритм Торгерсона минимизирует меру соответствия:

$$F = \sum_{i,j} (\delta^*_{ji} - \sum_k x_{ik} x_{jk})^2 \to \min ,$$

т.е. сумма квадратов разностей центрированных величин — характеристик различий объектов и расстояний между объектами в некотором теоретически определенном нормированном шкальном пространстве X, — должна быть минимальной (требование метода наименьших квадратов). При наличии исходной *матрицы различий* Δ *с элементами* δ_{ij} алгоритм реализуется последовательным выполнением следующих шагов.

Шаг 1. Приняв $\sum_i x_{ik} = \sum_j x_{jk} = 0$ и $\delta_{ij} = d_{ij}$, следует полагать, что существуют адаптивные реальным характеристикам различий величины σ^*_{ij}, для которых выполнялось бы аналогичное условие: $\sum_i \delta^*_{ji} = \sum_j \delta^*_{jk} = 0$. Значения δ^*_{ij} находят по формуле:

$$\delta^*_{ij} = -\frac{1}{2}\left(\delta^2_{ij} - \delta^2_{i\cdot} - \delta^2_{\cdot j} + \delta^2_{\cdot\cdot}\right),$$

где $\delta^2_{i\cdot}$ — средняя для характеристик различий в j-х столбцах i-й строки, возведенных в квадрат: $\delta^2_{i\cdot} = \frac{1}{J}\sum_j \delta^2_{ij}$;

$\delta^2_{\cdot j}$ — средняя для характеристик различий в i-х строках j-го столбца, возведенных в квадрат: $\delta^2_{\cdot j} = \frac{1}{I}\sum_i \delta^2_{ij}$;

$\delta^2_{\cdot\cdot}$ — средняя величина для квадратов характеристик различий матрицы Δ: $\delta^2_{\cdot\cdot} = \frac{1}{IJ}\sum_i \sum_j \delta^2_{ij}$.

Матрица, все элементы которой δ^*_{ij}, называется матрицей с двойным центрированием Δ^*, средние значения элементов каждой ее строки и каждого столбца равны нулю.

Шаг 2. По Торгерсону, для исчисленных значений δ^*_{ij}, если

$$\delta_{ij} = d_{ij} = \left(\sum_k (x_{ik} - x_{jk})^2 \right)^{1/2},$$

имеет силу равенство:

$\delta^*_{ij} = \sum_k x_{ik} x_{jk}$, или в матричном виде: $\Delta^* = XX'$,

где X — матрица координат стимулов размерности $I \times K$.

Определение матрицы X, как видно, тождественно решению задачи поиска собственных векторов, при этом могут использоваться: метод главных компонент или методы факторного анализа, рассмотренные в гл. 7. Векторы матрицы X — это шкалы (оси) полученного теоретического пространства стимулов.

На заключительном этапе решаются вопросы определения оптимальной размерности этого пространства и интерпретируемости аналитических результатов.

Размерность (или число шкал) должна быть такой, чтобы сохранялась большая часть (чаще всего не менее 80—90%) информативности стимулов, но при этом могут быть использованы и другие критерии. Так, для продолжения анализа иногда оставляют такое число шкал, которое позволяет визуализировать его результаты, т.е. две или три шкалы. В других случаях оставляют только те шкалы, векторные значения которых интерпретируемы, или шкалы, для которых оцененное собственное значение λ не меньше единицы (или некоторого порогового значения ζ, или, наконец, такое число шкал, при котором наилучшим образом воспроизводятся реальные различия объектов, когда $d_{ij} \to \delta_{ij}$. Во всех перечисленных случаях конечное число шкал, как правило, невелико.

Интерпретируемость результатов МШ определяется компонентным составом векторов, представляющих шкалы. Если значения на шкалах не поддаются логичному объяснению, осущест-

вляется поворот шкального пространства. Для этого, как известно, можно использовать также различные приемы: вращения вручную, при помощи аналитических методов – варимакса, квартимакса (ортогонального вращения) или квартимина, облимакса (косоугольного вращения) и т.д.

Для демонстрации алгоритма Торгерсона используем уже имеющиеся данные из табл. 8.1 а) и полученную на их основе матрицу различий.

На первом шаге исчислим значения центрированных элементов и преобразуем матрицу различий Δ в матрицу с двойным центрированием Δ^*. С этой целью предварительно следует все элементы матрицы Δ возвести в квадрат и рассчитать константные величины $\delta^2_{..}$, $\delta^2_{i.}$, $\delta^2_{.j}$:

$$\Delta = \begin{pmatrix} 0 & 3{,}657 & 3{,}429 & 3{,}662 \\ & 0 & 0{,}265 & 0{,}961 \\ & & 0 & 0{,}116 \\ & & & 0 \end{pmatrix}; \quad \Delta^2 = \begin{pmatrix} 0 & 13{,}374 & 11{,}758 & 13{,}410 \\ 13{,}374 & 0 & 0{,}070 & 0{,}924 \\ 11{,}758 & 0{,}070 & 0 & 1{,}245 \\ 13{,}410 & 0{,}924 & 1{,}245 & 0 \end{pmatrix};$$

$$\delta^2_{..} = \frac{81{,}562}{16} = 5{,}098.$$

При расчете $\delta^2_{i.}$ и $\delta^2_{.j}$ следует обратить внимание, что $\delta^2_{i.} = \delta^2_{.j}$, если $i = j$, т.е. в данном примере нужно рассчитать четыре средних значения δ^2 для столбцов (строк):

$$\delta^2_{.1} = \delta^2_{1.} = \frac{38{,}542}{4} = 9{,}635; \quad \delta^2_{.2} = \delta^2_{2.} = \frac{14{,}368}{4} = 3{,}529;$$

$$\delta^2_{.3} = \delta^2_{3.} = \frac{13{,}073}{4} = 3{,}268; \quad \delta^2_{.4} = \delta^2_{4.} = \frac{15{,}578}{4} = 3{,}895.$$

Исчислим центрированные элементы первого столбца:

$$\delta^*_{11} = -\frac{1}{2}\left(\delta^2_{11} - \delta^2_{1.} - \delta^2_{.1} + \delta^2_{..}\right) = -\frac{1}{2}(0 - 9{,}635 - 9{,}635 + 5{,}098) = 7{,}086;$$

$$\delta^*_{21} = -\frac{1}{2}\left(\delta^2_{21} - \delta^2_{2.} - \delta^2_{.2} + \delta^2_{..}\right) = -\frac{1}{2}(13{,}374 - 3{,}592 - 9{,}635 + 5{,}098) = -2{,}622$$

и т. д.

Результаты вычислений сведем в симметрической матрице с двойным центрированием (матрице скалярных произведений). Правильность расчетов элементов этой матрицы легко проверить — их суммы по каждой строке (столбцу) должны быть равны нулю, допускаются незначительные отклонения за счет округлений:

$$\Delta^* = \begin{pmatrix} 7{,}086 & -2{,}663 & -1{,}976 & -2{,}489 \\ -2{,}622 & 1{,}043 & 0{,}846 & 0{,}733 \\ -1{,}976 & 0{,}846 & 0{,}719 & 0{,}410 \\ -2{,}489 & 0{,}733 & 0{,}410 & 1{,}346 \end{pmatrix}.$$

Для известной матрицы скалярных произведений по Торгерсону действует равенство: $\Delta^* = XX'$, из которого находят X-матрицу координат стимулов.

Определим X методом главных факторов. С учетом уровня информативности в анализе остаются два первых общих фактора (две координатные оси, шкалы) — они объясняют более 98% вариации величин, характеризующих различия стимулов:

$$X = \begin{matrix} X_1 & X_2 \end{matrix} \begin{pmatrix} -1{,}000 & -0{,}022 \\ 0{,}998 & -0{,}059 \\ 0{,}990 & -0{,}138 \\ 0{,}975 & 0{,}177 \end{pmatrix}.$$

Зная координаты стимулов, легко показать их пространственное расположение (рис. 8.6).

Характер распределения фирм в двумерном шкальном пространстве (рис. 8.6) позволяет достаточно просто решить вопрос интерпретируемости аналитических результатов. По оси X_1 на значительном удалении от других находится фирма МАН. Данные табл. 8.1а объясняют такую удаленность существенными расхождениями в размерах фирм. Фирма МАН — крупнейшая по всем своим параметрам, фирмы СЕБ, Нокиа и Даниска — значительно меньших размеров, в то же время имеют и меньшие различия между собой, об-

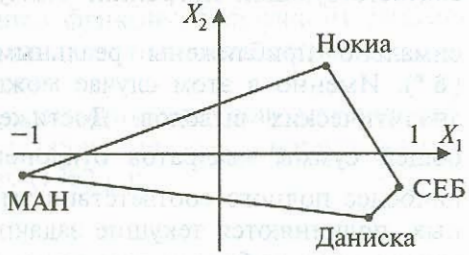


Рис. 8.6. Двумерная конфигурация шкал, построенная по данным четырех европейских фирм

разуя на другом конце первой координатной оси заметное сгущение. Думается, было бы правомерно назвать X_1 осью размера фирм.

Для оси X_2 корректным будет название «эффективности деятельности». По этой оси — на одном полюсе фирмы СЕБ, Даниска, МАН, а на другом — Нокиа, явно уступающая первой группе по характеристикам оборота капитала и массы прибыли на одного работника. То же и относительно массы прибыли на единицу капитала в обороте.

Вопросы вращения координатного пространства в задаче не рассматриваются, так как подробно представлены в гл. 7, кроме того, для интерпретации расчетных результатов не возникает острой необходимости в координатных преобразованиях. Можно предположить, что при осуществлении поворота были бы получены более логичные результаты, например, по оси X_1: фирмы распределились в направлении от мелких к крупным, а не наоборот, и фирмы, имеющие большие размеры и успешнее действующие, перешли в положительно определенные плоскости системы координат, поменявшись местами с явно более слабыми фирмами.

Остается отметить, что при реализации алгоритма Торгерсона минимизацией F-меры соответствия добиваются, чтобы теоретические величины $\hat{\delta}^*$, исчисляемые как суммы произведений соответствующих координат стимулов $\hat{\delta}^*_{ij} = \sum_k x_{ik} x_{jk}$, были максимально приближены реальным скалярным произведениям (δ^*). Именно в этом случае можно надеяться на адекватность аналитических выводов. Достижению наименьшей величины общей суммы квадратов отклонений $\sum(\delta^* - \hat{\delta}^*)^2$, а значит, и наиболее полного соответствия эмпирических и расчетных данных, подчиняются текущие задачи алгоритма МШ: нормирования данных, выбора метода поиска координатного пространства и установления достаточного числа координатных осей, поворота системы координат.

Выше уже сказано, что алгоритм Торгерсона имеет жесткие допущения: $\delta_{ij} = d_{ij}$. Однако добиться выполнения этого равенства в действительности сложно. Существуют метрические модели с менее жесткими требованиями, позволяющие расширить

сферу приложения подхода Торгерсона. Это модели Кэролла, Рамсея и др., они допускают отклонения в эмпирических и теоретических данных о различиях стимулов, делая правомерной запись:

$$\delta_{ij} = d_{ij} + \varsigma,$$

где ς — аддитивная константа, или пороговая величина, определяющая разность $\delta_{ij} - d_{ij}$.

При решении таких моделей вначале оценивают величину ς, находят толерантные отклонения ($\delta_{ij} - \hat{\varsigma}$), а затем реализуют алгоритм Торгерсона.

8.4. Неметрические методы многомерного шкалирования

Методы неметрического МШ применяют для обработки ранговых (порядковых) данных. Решающим условием, обеспечивающим адекватность аналитических выводов, здесь становится соответствие монотонных связей эмпирических и теоретических данных, т.е. если реально существует порядковая зависимость $\delta_{ij} < \delta_{tl}$, то в определяемом шкальном пространстве соответственно должно быть $d_{ij} < d_{tl}$. Вид монотонности заранее неизвестен и методом проб подбирается функция, наилучшим образом описывающая эмпирические данные: линейная, степенная, показательная или логарифмическая.

Отобрав в качестве меры расстояния евклидову метрику (d^E). можно записать равенство, задающее алгоритм поиска шкального пространства по Шепарду (1962 г.):

$$\delta_{ij} = f(d_{ij}) = f\left(\sum_k (x_{ik} - x_{jk})^2\right)^{1/2}.$$

где f — произвольная монотонная функция. Если, например, f — линейная функция, приведенное равенство можно переписать в виде:

$$\delta_{ij} = f(d_{ij}) = a_0 + a_1\left(\sum_k (x_{ik} - x_{jk})^2\right)^{1/2}.$$

Более общий случай предполагает оценку различий объектов в m-мерном пространстве Минковского (подход Дж. Краскала, 1964 г.), тогда:

$$\delta_{ij} = f(d_{ij}) = f\left(\sum_k (x_{ik} - x_{jk})^m\right)^{1/m}.$$

Универсальная модель неметрического МШ, построенная на метрике Минковского, легко позволяет перейти к другим моделям:

- с евклидовой метрикой, при $m = 2$;
- с метрикой доминирования, при $m \to \infty$. Модель имеет вид:

$$\delta_{ij} = f(d_{ij}) = f\left(\max_k |x_{ik} - x_{jk}|\right),$$

т.е. расстояние между стимулами i и j определяется разностью координат только по одной оси, по которой величина разности $|x_{ik} - x_{jk}|$ максимальна;

- с метрикой города (city-block , или l_1-норма). Для этого случая предположение о монотонности данных формально записывается следующим образом:

$$\delta_{ij} = f(d_{ij}) = f\left(\sum_k |x_{ik} - x_{jk}|\right).$$

Вне зависимости от выбора базовой модели для описания различий объектов методы неметрического МШ реализуются в последовательности, как это показано на рис. 8.7.

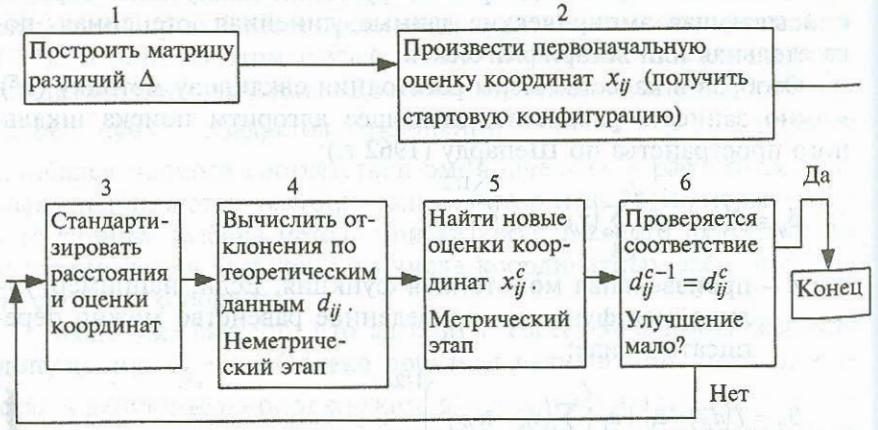


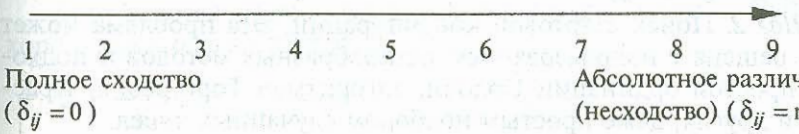
Рис. 8.7. ***Схема алгоритма неметрического МШ***

Остановимся на основных алгоритмических шагах неметрического МШ:

Шаг 1. Получение матрицы различий, содержащей ранговые данные — характеристики непохожести анализируемых объектов.

Существуют различные приемы получения исходных ранговых данных, наиболее распространены в анализе из них следующие:

- *метод последовательной рандомизации*, его сущность в последовательно проводимом делении совокупности наблюдаемых объектов на группы. При первом делении появляются две группы — пары похожих объектов и пары непохожих объектов. Затем в каждой группе соответственно находят пары с наиболее и менее похожими объектами и т. д. На заключительном этапе получают $n\,(n-1)/2$ пар, ранжированных по степени сходства (или наоборот — «непохожести»);
- *метод исходной (якорной) точки*, из общего числа n объектов на первом шаге отбирают один и его положение в совокупности принимается за исходное относительно других объектов. Степень сходства всех прочих объектов с первым (якорным) оценивается экспертами с присвоением ранга. На следующем шаге якорным становится другой, следующий из совокупности объект. И так для всех объектов. В общем получают $n\,(n-1)$ ранговых оценок парных сходств, по которым легко строится матрица различий Δ;
- *метод рейтинговой оценки.* Экспертам предлагается шкала с некоторым числом делений (обычно 7–9), позволяющих оценивать каждую пару объектов по степени их сходства, например, как это показано на рисунке:



По результатам экспертного оценивания получают $n\,(n-1)/2$ пар объектов, упорядоченных по ранговым характеристикам сходств.

Предположим, что одним из перечисленных методов установлены ранговые оценки для пяти государств, бывших республик СССР, с учетом их экономического и политического положения в 1994 г. Результаты экспертного оценивания после их

обобщения могли бы быть представлены, например, как показано в табл. 8.2[1].

Данные табл. 8.2 подтверждают, что для пяти наблюдаемых объектов будет получено именно 10, то есть $n(n-1)/2$ ранговых оценок.

Таблица 8.2. **Порядковые характеристики различий пяти государств с учетом их экономического и политического положения Δ**

	Армения	*Беларусь*	*Россия*	*Таджикистан*	*Литва*
Армения[1]	—	10	9	3	7
Беларусь	10	—	1	5	2
Россия	9	1	—	4	6
Таджикистан	3	5	4	—	8
Литва	7	2	6	8	—

Для следующего алгоритмического шага данные о различиях пяти стран можно оставить в первоначальном виде или преобразовать их в количественные. Другими словами, возможна их оцифровка. В своей книге М. Дэйвисон [32, с. 107] описывает надежный и одновременно простой прием перехода к матрице с количественными характеристиками различий: вначале на ранговых данных строится матрица корреляций R, оценки различий, т.е. элементы матрицы Δ, определяются затем с учетом имеющихся величин парных коэффициентов корреляций r_{ij} по формуле: $\delta_{ij} = (1 - r_{ij})^{1/2}$.

Шаг 2. Поиск стартовой конфигурации. Эта проблема может быть решена с использованием разнообразных методов и подходов: простой ординации Орлочи, алгоритмов Торгерсона, Краскала и других, даже простым подбором случайных чисел.

В примере по данным табл. 8.2 первые приблизительные оценки координат вычислены методом главных компонент. Получены нестандартизованные характеристики по первым двум координатным осям, объясняющим более 98% общей дисперсии значений стимульных признаков (табл. 8.3).

[1] Максимальная ранговая оценка различий: 10.

Координатные оценки стимулов позволяют дать названия каждой из шкал. По оси X_1 на одном конце наибольшую факторную нагрузку имеют Армения и Таджикистан, на другом — Беларусь и Литва; очевидно, что эта ось вытянута в направлении Юг-Север. Аналогичного рода рассуждения приводят к мысли, что ось X_2 определяется направлением Восток-Запад.

Таблица 8.3. **Стартовая конфигурация для неметрического шкалирования пяти государств, республик бывшего СССР**

Стимул	*Первая координатная ось* X_1	*Вторая координатная ось* X_2
Армения	−0,974	0,217
Беларусь	0,958	−0,254
Россия	0,701	−0,710
Таджикистан	−0,690	−0,697
Литва	0,772	0,610

Шаг 3. Стандартизация оценок координат и расстояний. Стандартизация проводится с целью сохранить пропорции ортонормированного стимульного пространства и избежать вырожденных решений, когда пространство стимулов сжимается до размеров точки и анализ не дает сколько-нибудь значимых результатов. Например, когда несколько стимулов получают одинаковые оценки координат или их координатные оценки близки нулю, т.е. расположены вблизи начала системы координат. Стимулы в таком теоретическом пространстве шкал как бы сливаются и становятся неразличимы для исследователя.

Допускаются различные варианты стандартизации. Выберем хорошо знакомый способ:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\delta_j}.$$

Стандартизовав координаты стимулов и рассчитав по ним расстояния между стимулами, получим элементы матрицы стандартизованных оценок расстояний. Величины расстояний в пространстве шкал X_1, X_2 будем оценивать по формуле евклидовой метрики:

$$d_{ij}^{E} = \left(\sum_k (x_{ik} - x_{jk})^2 \right)^{1/2}.$$

Стандартизованные оценки, результаты вычислений сведем в табл. 8.4.

Таблица 8.4. **Нестандартизованные и стандартизованные оценки координат и расстояний для пяти государств**

Нестандартизованные оценки

координат			*расстояний* d_{ij}				
	X_1	X_2					
Армения	−0,974	0,217	0	1,989	1,914	0,962	1,789
Беларусь	−0,958	−0,254	1,989	0	0,523	1,703	0,884
Россия	0,701	−0,710	1,914	0,523	0	1,391	1,322
Таджикистан	−0,690	−0,697	0,962	1,703	1,391	0	1,961
Литва	0,772	0,610	1,789	0,884	1,322	1,961	0

Стандартизованные оценки

координат $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\delta_j}$			*расстояний* d_{ij} (*рассчитано по стандартизованным оценкам координат* X_1, X_2)				
	X_1	X_2					
Армения	−1,384	0,724	0	2,532	2,700	1,759	2,269
Беларусь	−0,988	−0,164	2,532	0	0,918	2,190	1,646
Россия	0,672	−1,026	2,700	0,918	0	1,708	2,507
Таджикистан	−1,036	−1,000	1,759	2,190	1,708	0	3,051
Литва	0,760	1,466	2,269	1,646	2,507	3,051	0

Шаг 4. Неметрический этап. Алгоритмический шаг предназначен для упорядочения оценок расстояний между стимулами.

В теоретическом пространстве шкал X_k монотонность исходных данных может нарушаться (рис. 8.8). Корректировка теоретических величин расстояний d_{ij} производится при неизменных оценках координат стимулов и таким образом, чтобы восстановить общую тенденцию к возрастанию в исходных данных о различиях.

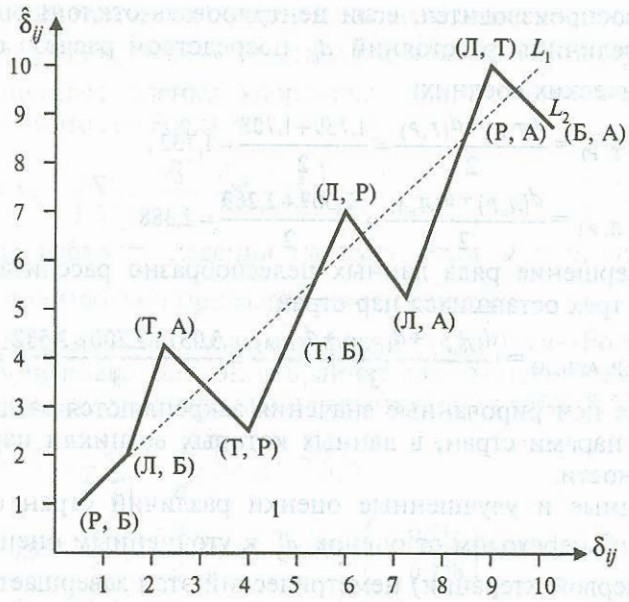


Рис. 8.8. ***Отношения ранговых порядков стимулов по исходным и теоретическим данным на первой итерации***

Рис. 8.8 построен по данным рассматриваемого примера и наглядно показывает возникшее несоответствие в изменении исходных и теоретических ранговых оценок (табл. 8.5). По оси δ_{ij} отложены фактические значения характеристик различий, по оси $\hat{\delta}_{ij}$ — значения, принимаемые в теоретическом пространстве шкал X_1, X_2. Линия L_1 — прямая монотонной функции равномерно возрастающих оценок δ_{ij}, линия L_2 построена с учетом отклонений эмпирических ранговых оценок от теоретических. Прописными буквами обозначены пары стран.

Графическое изображение несоответствий ранговых оценок можно получить и несколько иным образом, если по оси y вместо ранговой теоретической величины различий $\hat{\delta}_{ij}$ откладывать количественно определенные значения расстояний между объектами d_{ij}. Такой рисунок носит название диаграммы Шепарда.

В данном примере улучшить оценки расстояний достаточно просто: монотонность равномерно возрастающих теоретических

данных воспроизводится, если центрировать отклоняющиеся от прямой величины расстояний d_{ij} посредством расчета обычных арифметических средних:

$$d_{(\text{Т. А; Т, Р})} = \frac{d_{(Т,А)} + d_{(Т,Р)}}{2} = \frac{1{,}759 + 1{,}708}{2} = 1{,}733,$$

$$d_{(\text{Л, Р; Л, А})} = \frac{d_{(Л,Р)} + d_{(Л,А)}}{2} = \frac{2{,}507 + 2{,}269}{2} = 2{,}388.$$

В завершение ряда данных целесообразно рассчитать среднюю для трех оставшихся пар стран:

$$d_{(\text{Л, Р; Р, А; Б, А})} = \frac{d_{(Л,Р)} + d_{(Р,А)} + d_{(В,А)}}{3} = \frac{3{,}051 + 2{,}700 + 2{,}532}{3} = 2{,}761.$$

Новые центрированные значения закрепляются за двумя соседними парами стран, в данных которых возникли нарушения монотонности.

Исходные и улучшенные оценки различий стран сведем в табл. 8.5. С переходом от оценок d_{ij}^{c} к уточненным оценкам d_{ij}^{c+1} (с+1 — первой итерации) неметрический этап завершается.

Таблица 8.5. **Исходные ранговые оценки различий стран и величины расстояний между ними в теоретическом пространстве шкал — первичные и уточненные**

Исходный ранговый порядок (δ_{ij})	*Стимул*	*Стимул*	*Стандартизованные расстояния* (d_{ij})	*Ранговый порядок стимулов в пространстве шкал* X_1, X_2 ($\hat{\delta}_{ij}$)	*Улучшенные оценки расстояний* (d_{ij}^{1})
1	Россия	Беларусь	0,918	1	0,918
2	Литва	Беларусь	1,646	2	1,646
3	Таджикистан	Армения	1,759	4	1,733
4	Таджикистан	Россия	1,708	3	1,733
5	Таджикистан	Беларусь	2,190	5	2,190
6	Литва	Россия	2,507	7	2,388
7	Литва	Армения	2,269	6	2,388
8	Литва	Таджикистан	3,051	10	2,761
9	Россия	Армения	2,700	9	2,761
10	Беларусь	Армения	2,532	8	2,761

Шаг 5. Метрический этап. На данном этапе имеющимся исходным и уточненным величинам расстояний (d_{ij}^{c} и d_{ij}^{c+1}) находят уточненные оценки координат. Для расчетов используют формулу Лингоса—Роскама:

$$x_{ik}^{c+1} = x_{ik} - \frac{1}{j}\sum_{j}\left(1 - \frac{d_{ij}^{c+1}}{d_{ij}^{c}}\right)\left(x_{ik}^{c} - x_{jk}^{c}\right).$$

Чтобы избежать деления на нуль, если $d_{ij} = 0$, отношение $d_{ij}^{c+1} / d_{ij}^{c}$ произвольно приравнивается единице.

Посмотрим, как применить формулу Лингоса—Роскама при вычислении новых оценок координат для стимула Беларусь (исходные данные, участвующие в расчетах, см. в табл. 8.4 и 8.5):

$$X_{\text{Б1}}^{1} = X_{\text{Б1}}^{0} - \frac{1}{5}\sum_{1}^{5}\left(1 - \frac{d^{1}}{d^{0}}\right)\left(X_{\text{Б1}}^{1} - X_{j1}^{0}\right) =$$

$$= 0{,}988 - \frac{1}{5}\left[\left(1 - \frac{0}{0}\right)(0{,}988 - 0{,}988) + \left(1 - \frac{0{,}918}{0{,}918}\right)(0{,}988 - 0{,}672) + \right.$$

$$\left. + \left(1 - \frac{2{,}190}{2{,}190}\right)(0{,}988 - 1{,}036) + \left(1 - \frac{2{,}761}{2{,}532}\right)(0{,}988 - (-1{,}384))\right] = 1{,}030;$$

$$X_{\text{Б2}}^{1} = X_{\text{Б2}}^{0} - \frac{1}{5}\sum_{1}^{5}\left(1 - \frac{d^{1}}{d^{0}}\right)\left(X_{\text{Б2}}^{1} - X_{j2}^{0}\right) =$$

$$= -0{,}164 - \frac{1}{5}\left[\left(1 - \frac{0}{0}\right)(-0{,}164 - (-0{,}164)) + \left(1 - \frac{0{,}918}{0{,}918}\right)(-0{,}164 - (-1{,}026)) + \right.$$

$$+ \left(1 - \frac{1{,}646}{1{,}646}\right)(-0{,}164 - 1{,}466) + \left(1 - \frac{2{,}190}{2{,}190}\right)(-0{,}164 - (-1{,}000)) +$$

$$\left. + \left(1 - \frac{2{,}761}{2{,}532}\right)(-0{,}164 - 0{,}724)\right] = -0{,}180,$$

т.е. новые координаты стимула Республика Беларусь будут: $X^{1} = (1{,}030 \;\; -0{,}180)$ в отличие от начальных координат $X^{0} = (0{,}988 \;\; -0{,}164)$.

Подобные расчеты проводятся для всех участвующих в анализе объектов, после этого уже по новым оценкам координат (x_{ik}^{c+1}) находят расстояния между стимулами в теоретическом пространстве (d^{c+1}) и первая итерация заканчивается, остается только оценить качество ее результатов.

Шаг 6. Оценка соответствий монотонных ранговых эмпирических и теоретических данных. Собственно проверке на монотонность подлежат теоретические данные d_{ij}^{c} и d_{ij}^{c+1}, рассматривается степень их улучшения на прошлой итерации. Если улучшение существенно, итерация возобновляется после стандартизации полученных на шаге 5 оценок координат и расстояний, если же улучшение мало, итерации заканчиваются, и приступают к интерпретации итогов анализа.

Оценивание соответствий теоретических результатов эмпирическим данным осуществляется при помощи специальных стресс-формул или коэффициента отчуждения:

Стресс-формулы Краскала

$$S_1 = \left(\frac{\sum_{ij} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{ij} \hat{d}_{ij}^2} \right)^{1/2}, \qquad S_2 = \left(\frac{\sum_{ij} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{ij} \left(\hat{d}_{ij} - \hat{d}_{..} \right)^2} \right)^{1/2};$$

Стресс-формулы Юнга

$$SS_1 = \left(\frac{\sum_{ij} \left(d_{ij}^2 - \hat{d}_{ij}^2 \right)^2}{\sum_{ij} \hat{d}_{ij}^4} \right)^{1/2}, \qquad SS_2 = \left(\frac{\sum_{ij} \left(d_{ij}^2 - \hat{d}_{ij}^2 \right)^2}{\sum_{ij} \left(\hat{d}_{ij}^2 - \hat{d}_{..}^2 \right)^2} \right)^{1/2}.$$

Коэффициент отчуждения Гуттмана

$$k = \left(1 - \mu^2 \right)^{1/2}, \text{ где } \mu = \frac{\sum d_{ij} \hat{d}_{ij}}{\left[\left(\sum_{ij} d_{ij}^2 \right) \left(\sum_{ij} \hat{d}_{ij}^2 \right) \right]^{1/2}}.$$

Во всех перечисленных формулах символами d и $\hat{d}$ обозначены величины расстояний: исходные и уточненные, после выполнения определенного шага алгоритма, или завершения итерации, $\hat{d}_{..}$ — среднее арифметическое всех оцененных расстояний:

$$\hat{d}_{..} = \frac{1}{ij} \sum_{ij} d_{ij}.$$

Расчет стресс-формул продемонстрируем на данных табл. 8.5. Выбрав S_1 и S_2 Дж. Краскала, посмотрим, насколько улучшены

оценки d^1 по сравнению с оценками d^0 (табл. 8.6). Задачу интерпретации величин, исчисленных по стресс-формулам, облегчают известные заранее стандартные характеристики (табл. 8.7).

Таблица 8.6. **Проверка на существенность улучшения теоретических оценок расстояний с использованием стресс-формул Дж. Краскала**

Стимул	*Стимул*	*Исходная ранговая оценка*	d_{ij}^0	d_{ij}^1	$(d_{ij}^0)^2$	$(d_{ij}^0 - d_{ij}^1)^2$	$(d_{ij}^1 - d_{..})^2$
Россия	Беларусь	1	0,918	0,918	0,843	0	1,464
Литва	Беларусь	2	1,646	1,646	2,709	0	0,232
Таджикистан	Армения	3	1,759	1,733	3,094	0,0007	0,156
Таджикистан	Россия	4	1,708	1,733	2,917	0,0007	0,156
Таджикистан	Беларусь	5	2,190	2,190	4,796	0	0,038
Литва	Россия	6	2,507	2,388	6,285	0,0142	0,068
Литва	Армения	7	2,269	2,388	5,148	0,0142	0,068
Литва	Таджикистан	8	3,051	2,761	9,309	0,0841	0,401
Россия	Армения	9	2,700	2,761	7,290	0,0037	0,401
Беларусь	Армения	10	2,532	2,761	6,411	0,0524	0,401
$\sum$	—		21,280	21,279	48,802	0,1700	3,385

$$d_{..}^1 = \frac{21{,}280}{10} = 2{,}128$$

$$S_1 = \sqrt{0{,}17 / 48{,}820} = 0{,}059$$

$$S_1 = \sqrt{0{,}17 / 3{,}385} = 0{,}224$$

Таблица 8.7. **Содержательная оценка величин, исчисленных по стресс-формулам S_1 и S_2 (Дж. Краскала)**

Степень соответствия	*Для формулы*	
	S_1	S_2
Низкая	0,2	0,4
Удовлетворительная	0,1	0,2
Хорошая	0,05	0,1
Отличная	0,025	0,15
Превосходная	0	0

Согласно данным табл. 8.7 значения критериев S_1 и S_2, рассчитанные в табл. 8.6, дают основание судить о результатах нашего решения как удовлетворительных. В прикладном анализе, думается, исследователем была бы предпринята при этом попытка продолжить итерации и найти более адекватные оценки координат стимулов и расстояний.

Обобщая материал § 8.4, отметим, что в рамках методов неметрического МШ решаются схожие с метрическим МШ задачи: оценки координат стимулов и расстояний между стимулами, вращения системы координат, интерпретации аналитических результатов. В то же время заметны и отличия. Неметрическое МШ имеет более сложные алгоритмы, включающие: поиск стартовой конфигурации, неметрический этап — для корректировки распределения теоретических оценок расстояний и, наконец, метрический этап — для уточнения оценок координат стимулов. Итеративная реализация алгоритма неметрического МШ строится таким образом, чтобы предупредить появление вырожденных решений и существенные расхождения функциональных монотонных связей эмпирических и теоретических данных. В его алгоритмах проблемными остаются вопросы: подборки вида монотонной функции, отвечающей фактическому распределению характеристик различий стимулов, неизвестной заранее, и, как прежде, задача интерпретируемости итогов анализа.

8.5. Модели поиска индивидуальных различий

Рассмотренные выше методы метрического и неметрического многомерного шкалирования могут применяться для координатного описания только самих стимулов. Но в исследованиях не менее важно иметь представление и о различиях источников информации. В конечном счете пространственное положение стимулов объясняется не только их «непохожестью», но и расхождениями суждений о них, или различием приемов оценивания, получения данных. Действительно, если данные получают посредством анкетирования или экспертного оценивания, то они нередко существенно различаются в силу особенностей поведения и склонностей субъектов, выступающих в роли экспертов, когда же ведется прямая регистрация сведений о явлениях, процессах, свой отпечаток налагают особенности наблюдаемых объ-

ектов, условия, в которых они находятся (климатические, экологические) и т.д.

В сущности задача моделирования индивидуальных различий сводится к реализации алгоритма для нахождения шкал и представления в координатном пространстве как стимулов, так и субъектов, их оценивающих.

Координатами субъектов при этом служат значения весовых коэффициентов ω_{ks}, характеризующие уровень значимости координатной оси k для субъекта s.

На рис. 8.9 показано гипотетическое распределение субъектов (экспертов) в двумерном пространстве шкал, определяющих экономичность производства. Расположение субъектов задается значениями весовых коэффициентов ω_{ks}.

По данным рис.8.9 можно видеть, что, например, субъект 1 в определении эффективности производства примерно равное значение придает характеристикам ресурсоемкости и трудоемкости производства. Субъект 2 считает, что эффективность в наибольшей мере определяется ресурсоемкостью производства, весовой коэффициент для этого общего признака почти в 2 раза превышает оценку значимости по шкале «трудоемкость производства». Наконец, субъект 3 находит, что определяющим для эффективности производства является именно характеристика результатов использования живого труда.

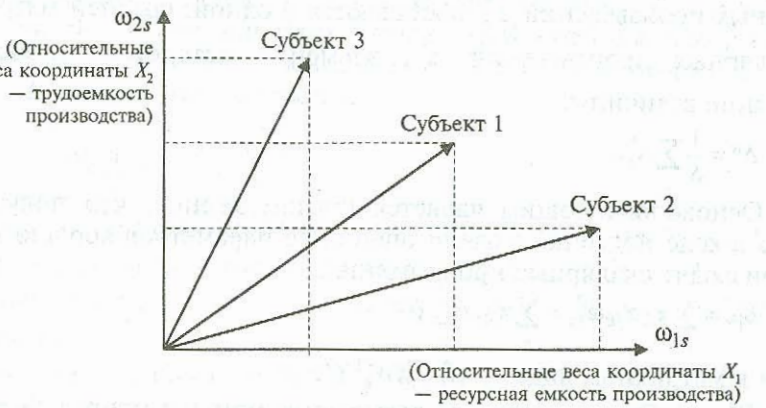


*Рис. 8.9. **Расположение трех субъектов в двумерном шкальном пространстве процесса эффективности производства***

В моделировании индивидуальных различий существует два основных подхода. Первый подход базируется на предположении о независимости координатных осей и объединяет так называемые модели индивидуального шкалирования — Кэррола, Чанга, Хорана и др. (теоретические работы 1968—1970 гг.). Второй подход допускает, что субъекты различаются не только весами координат, но и силой взаимодействия координатных осей (стимулов). Его модели были разработаны в основном в 1972—1980 гг. Наиболее представительной здесь является *трехмодальная модель Такера.*

Алгоритмы вычислений для различных моделей индивидуальных различий включают следующие общие шаги:

Шаг 1. Построение матриц различий стимулов Δ_s для каждого из субъектов.

Шаг 2. Построение S матриц скалярных произведений Δ_s^*.

С учетом того, что анализируются матрицы различий субъектов Δ_s, формулы для определения матриц скалярных произведений запишутся в следующем виде:

$$\delta_{ijs}^* = -\frac{1}{2}(\delta_{ijs}^2 - \delta_{i\cdot s}^2 - \delta_{\cdot js}^2 + \delta_{\cdot\cdot s}^2),$$

где $\delta_{i\cdot s}^2 = \frac{1}{j}\sum_j \delta_{ijs}^2$; $\delta_{\cdot js}^2 = \frac{1}{i}\sum \delta_{ijs}^2$; $\delta_{\cdot\cdot}^2 = \frac{1}{ij}\sum_i\sum_j \delta_{ijs}^2$.

Затем, при поиске стартовой конфигурации, S матриц скалярных произведений Δ_s^* обобщаются в одной, средней матрице скалярных произведений Δ^*, элементы которой — простые средние величины:

$$\Delta^* = \frac{1}{S}\sum_s \Delta_s^*.$$

Основополагающим является предположение, что полученные в ходе подгонки модели оценки ее параметров хорошо воспроизводят скалярные произведения:

$$\delta_{ijs}^* = \sum_k x_{ik}x_{jk}\omega_{ks}^2 = \sum_k x_{iks}x_{jks},$$

или в матричном виде — $\Delta_s^* = XW_s^2X'$.

Шаг 3. Поиск одним из возможных методов стартовой конфигурации (определение матрицы X^0, где 0 указывает на начальную итерацию).

Шаг 4. Оценка весовых коэффициентов ω_{ks}. Множество значений ω_{ks} образует матрицу W с данными по k координатным осям и s субъектам, т.е. для конкретного субъекта s в W — диагональной матрице имеется некоторый элемент ω_{ks}, представляющий его суждение о k-м общем признаке (k-й шкале).

Шаг 5. Оценка координат стимулов, построение матрицы X размерности $j \times k$ — по числу j стимулов (строк) и k координатных осей (шкал).

Шаг 6. Проверка качества полученного решения методом наименьших квадратов:

$$F = \sum_{i,j,s}\left(\delta_{ijs}^* - \hat{\delta}_{ijss}^*\right)^2 \to 0,$$

где δ_{ijs}^* и $\hat{\delta}_{ijs}^*$ — скалярные произведения по исходным и теоретическим данным.

Если квадрат разности между фактическими и теоретическими скалярными произведениями наименьший или меньше некоторого заранее известного порогового значения, то полученная конфигурация X^0 и матрица оценок весов W считаются наилучшими, и алгоритм завершен. Если же значения критерия F неудовлетворительны, оптимизирующие шаги 4—6 повторяются.

Остановимся подробнее на важнейших моделях индивидуальных различий.

Взвешенная евклидова модель — модель первого типа рассчитана на получение линейно независимой системы координат (шкал). Конструктивно основывается на использовании взвешенной евклидовой метрики:

$$d_{ijs} = \left(\omega_{ks}^2\sum_k\left(x_{ik} - x_{jk}\right)^2\right)^{1/2},$$

где ω_{ks}^2 — квадрат величины ω_{ks}, представляющей вес (важность, значимость) k-той шкалы для субъекта s.

Значение ω_{ks} линейно связывается с координатами стимула i субъекта s:

$x_{iks} = x_{ik}\,\omega_{ks}$, или в матричном виде: $X_s = XW_s$.

Очевидно, что при прочих равных величинах координат стимулов увеличение ω_{ks} означает и большее различие между стимулами i и j.

При реализации алгоритма анализа индивидуальных различий решаются задачи оценки координат стимулов, оценки величин ω_{ks} и их оптимизации.

Для примера возьмем гипотетические матрицы различий, это могут быть, скажем, результаты оценки двумя субъектами уровня экологичности производства до и после проведения природоохранных мероприятий в трех административных районах (табл. 8.8).

Таблица 8.8. **Исходные матрицы различий по результатам экспертного оценивания двумя субъектами и исчисленные по ним матрицы скалярных произведений**

Матрицы различий

$$\Delta_1 = \begin{array}{c} \text{Стимул} \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{pmatrix} - & 0{,}80 & 0{,}40 \\ 0{,}80 & - & 0{,}60 \\ 0{,}40 & 0{,}60 & - \end{pmatrix} \end{array} \qquad \Delta_2 = \begin{array}{c} \text{Стимул} \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{pmatrix} - & 0{,}70 & 0{,}50 \\ 0{,}70 & - & 0{,}30 \\ 0{,}50 & 0{,}30 & - \end{pmatrix} \end{array}$$

Матрицы скалярных произведений

$$\Delta_1^* = \begin{array}{c} \text{Стимул} \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{pmatrix} 0{,}14 & -0{,}15 & 0{,}01 \\ -0{,}15 & 0{,}20 & -0{,}05 \\ 0{,}01 & -0{,}05 & 0{,}04 \end{pmatrix} \end{array} \qquad \Delta_2^* = \begin{array}{c} \text{Стимул} \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{pmatrix} 0{,}15 & -0{,}12 & -0{,}03 \\ -0{,}70 & 0{,}10 & 0{,}02 \\ -0{,}03 & 0{,}02 & 0{,}01 \end{pmatrix} \end{array}$$

Для индивидуальных матриц скалярных произведений рассчитаем одну общую, т.е. среднюю матрицу скалярных произведений, воспользовавшись для этого тривиальной формулой средней:

$$\Delta^* = \frac{1}{S}\sum_s \Delta_s^* ,$$

в нашем случае

$$\Delta^* = \frac{1}{2}(\Delta_1^* + \Delta_2^*) = \begin{pmatrix} -0{,}15 & -0{,}14 & -0{,}01 \\ -0{,}14 & 0{,}15 & -0{,}01 \\ -0{,}01 & -0{,}01 & 0{,}02 \end{pmatrix}.$$

Первые приближенные оценки координат стимулов получим методом главных компонент. С учетом того, что первые шкалы

X_1 и X_2 объясняют более 99% вариации стимульных значений, запишем двумерную матрицу стартовой конфигурации:

$$X^0 = \begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{pmatrix} -0{,}998 & -0{,}596 \\ +0{,}998 & -0{,}596 \\ 0{,}000 & +1{,}000 \end{pmatrix} \end{array}.$$

На первом шаге итерации последовательно вычисляют величины ω^2 — матрицы весов с элементами ω_{ks}^2, а также новых оцененных матриц: $\hat{X}$ — координат стимулов, $\hat{\Delta}$ — расстояний и $\hat{\Delta}^*$ — скалярных произведений.

Для нахождения матрицы W^2 следует построить две исходные матрицы: S — объединенную матрицу скалярных произведений субъектов и B — координат для сочетающихся пар стимулов. Первая матрица S имеет i^2 столбцов — по числу всех возможных парных комбинаций стимулов и s строк — по числу субъектов:

Субъект \ Пары стимулов	n_1n_1	n_1n_2	n_1n_3	n_2n_1	n_2n_2	n_2n_3	n_3n_1	n_3n_2	n_3n_3
Субъект 1	0,14	−0,15	0,01	−0,15	0,20	−0,05	0,01	−0,05	0,04
Субъект 2	0,15	−0,12	−0,03	−0,12	0,10	0,02	−0,03	0,02	0,01

Элементы матрицы B находят перемножением текущих оценок координат из каждой пары стимулов: $b_{k(i,j)} = x_{ik}x_{jk}$.

Таким образом, сама матрица имеет также i^2 столбцов и k строк — по числу координатных осей (шкал) в анализе. Сделаем несколько примерных расчетов значений элементов $b_{k(i,j)}$ для первой шкалы:

$$b_1(n_1n_1) = -0{,}998\times(-0{,}998) = 0{,}996 \qquad b_1(n_2n_1) = 0{,}998\times(-0{,}998) = -0{,}996$$
$$b_1(n_1n_2) = -0{,}998\times 0{,}998 = -0{,}996 \qquad b_1(n_2n_2) = 0{,}998\times 0{,}998 = 0{,}996$$
$$b_1(n_1n_3) = -0{,}998\times 0{,}000 = 0{,}000 \qquad b_1(n_2n_3) = 0{,}998\times 0{,}000 = 0{,}000.$$

В общем матрица B принимает вид:

Координата \ Пары стимулов	n_1n_1	n_1n_2	n_1n_3	n_2n_1	n_2n_2	n_2n_3	n_3n_1	n_3n_2	n_3n_3
$B=$ Координата X_1	0,996	−0,996	0	−0,996	0,996	0	0	0	0
Координата X_2	0,355	0,355	−0,596	0,355	0,355	−0,596	−0,596	−0,596	1,000

Матрица W^2, т.е. наилучшие на первой итерации оценки весов, определяется из уравнения $S = BW^2$ методом наименьших квадратов:

$W^2 = (BB')^{-1}BS'$.

Построим матрицу W^2:

Координата \ Субъект	s_1	s_2
$W^2 =$ X_1	0,160	0,166
X_2	0,026	0,009

Исчислив корни квадратные из каждого значения w_{ks}^2, построим матрицу весовых значений W:

$$W = \begin{pmatrix} 0{,}40 & 0{,}41 \\ 0{,}16 & 0{,}09 \end{pmatrix}.$$

Данные матрицы W показывают, что оба эксперта одинаково большое значение придают первой шкале, второй субъект — даже несколько большее, чем первый; значение, которое придается второй шкале, существенно меньше, приблизительно в 3—5 раз.

После того как определена матрица W, становится возможным найти новые оценки координат стимулов. Вычисления строятся таким образом, чтобы минимизировать сумму квадратов разностей: $\sum(S-BX)^2$, т.е. $X = SB'(BB')^{-1}$. При этом матрицы S и B принимают иной вид, чем это было раньше. Элементы матрицы S — по-прежнему скалярные произведения, но все возможные пары стимулов представлены здесь отдельно для каждого из субъектов, в общем матрица S — результат простого объединения матриц скалярных произведений:

Стимул \ Субъект	s_1			s_2		
	n_1	n_2	n_3	n_1	n_2	n_3
n_1	0,14	−0,15	0,01	0,15	−0,12	−0,03
$S = n_2$	−0,15	0,20	−0,05	−0,12	0,10	0,02
n_3	0,01	−0,05	0,04	−0,13	0,02	0,01

Матрица B аккумулирует координатные значения стимулов, исчисленные с поправкой на величину весовых коэффициентов ω_{ks}^2, которые, как известно, являются субъективными оценками значимости координатных осей: $b_k(s_j) = \omega_{ks}^2 x_{jk}$.

Примерные расчеты:

$b_1(s_{1,1}) = 0{,}160 \times (-0{,}998) = -0{,}160$; $b_1(s_{2,1}) = 0{,}166 \times (-0{,}998) = -0{,}166$;

$b_1(s_{1,2}) = 0{,}160 \times 0{,}998 = 0{,}160$; $b_1(s_{2,2}) = 0{,}166 \times 0{,}998 = 0{,}166$;

$b_1(s_{1,3}) = 0{,}160 \times 0{,}000 = 0$; $b_1(s_{2,3}) = 0{,}166 \times 0{,}000 = 0$ и т.д.

Матрица B принимает вид:

Стимул \ Субъект	s_1			s_2		
	n_1	n_2	n_3	n_1	n_2	n_3
$B=$ Координата 1	−0,160	0,160	0	−0,166	0,166	0
Координата 2	−0,015	−0,015	0,026	−0,005	−0,005	0,009

Теперь могут быть найдены новые, улучшенные оценки координат стимулов:

Стимул \ Координата	X_1	X_2
$\hat{X}' = SB'(BB')^{-1} =$ n_1	−0,857	−0,015
n_2	0,869	−1,354
n_3	−0,008	1,369

В имеющемся (нестандартизованном) шкальном пространстве легко определить расположение стимулов с учетом мнения

каждого из субъектов. Для этого достаточно координатные значения стимула x_{ik} перемножить на соответствующие им величины субъективных весов ω_{ks} (рис. 8.10).

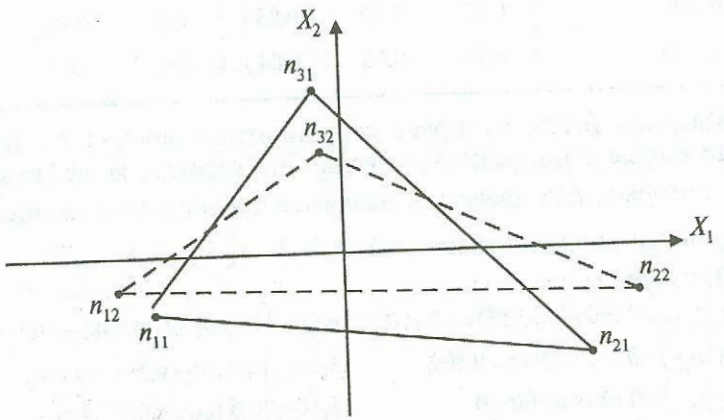


*Рис. 8.10. **Шкальное пространство и расположение в нем стимулов по данным экспертного оценивания двух субъектов***

Координаты стимулов:

для s_1 — n_1 (−0,343; −0,024), n_2 (0,347; −0,217), n_3 (−0,001; 0,219);

для s_2 — n_1 (−0,351; −0,001), n_2 (0,356; −0,122), n_3 (−0,003; 0,123).

Заметим, что пространственное расположение стимулов в двумерном пространстве субъектов позволяет определить различия между стимулами, как, собственно, и между самими субъектами. Нас в данном случае интересуют различия стимулов, их характеристиками будут величины расстояний d_{ijs}:

$$d_{ijs} = \left(\sum_k \omega_{ks}^2 \left(\hat{x}_{ik} - \hat{x}_{jk} \right)^2 \right)^{1/2}.$$

В табл. 8.9 сведены результаты расчетов расстояний и скалярных произведений для новых оцененных координат стимулов, чтобы можно было наглядно видеть результаты первого шага подгонки евклидовой модели индивидуальных различий. В этой же таблице приведены исходные данные о различиях, скалярные произведения и координаты стимулов.

Таблица 8.9. **Матрицы расстояний, скалярных произведений и координат стимулов по данным экспертного оценивания двумя субъектами**

По исходным данным	*После выполнения первого шага подгонки евклидовой модели*

Матрицы расстояний

$$\Delta_1 = \begin{pmatrix} - & 0{,}80 & 0{,}40 \\ 0{,}80 & - & 0{,}60 \\ 0{,}40 & 0{,}60 & - \end{pmatrix} \qquad \hat{\Delta}_1 = \begin{pmatrix} - & 0{,}72 & 0{,}40 \\ 0{,}72 & - & 0{,}56 \\ 0{,}40 & 0{,}56 & - \end{pmatrix}$$

$$\Delta_2 = \begin{pmatrix} - & 0{,}70 & 0{,}50 \\ 0{,}70 & - & 0{,}30 \\ 0{,}50 & 0{,}30 & - \end{pmatrix} \qquad \hat{\Delta}_2 = \begin{pmatrix} - & 0{,}71 & 0{,}37 \\ 0{,}71 & - & 0{,}44 \\ 0{,}37 & 0{,}44 & - \end{pmatrix}$$

Матрицы скалярных произведений

$$\Delta_1^* = \begin{pmatrix} 0{,}14 & -0{,}15 & 0{,}01 \\ -0{,}15 & 0{,}20 & -0{,}05 \\ 0{,}01 & -0{,}05 & 0{,}04 \end{pmatrix} \qquad \Delta_1^* = \begin{pmatrix} 0{,}12 & -0{,}12 & 0 \\ -0{,}12 & 0{,}17 & -0{,}05 \\ 0 & -0{,}05 & 0{,}05 \end{pmatrix}$$

$$\Delta_1^* = \begin{pmatrix} 0{,}15 & -0{,}12 & -0{,}03 \\ -0{,}12 & 0{,}10 & 0{,}02 \\ -0{,}03 & 0{,}02 & 0{,}01 \end{pmatrix} \qquad \Delta_2^* = \begin{pmatrix} 0{,}12 & -0{,}12 & 0 \\ -0{,}12 & 0{,}14 & -0{,}02 \\ 0 & -0{,}02 & 0{,}02 \end{pmatrix}$$

Матрицы координат стимулов

$$X = \begin{pmatrix} -0{,}998 & -0{,}596 \\ 0{,}998 & -0{,}596 \\ 0 & 1{,}000 \end{pmatrix}$$

Нестандартизованные оценки:

$$\hat{X}^1 = \begin{pmatrix} -0{,}857 & -0{,}015 \\ 0{,}869 & -1{,}354 \\ -0{,}008 & 1{,}369 \end{pmatrix}$$

Стандартизованные оценки:

$$\hat{X}^{1c} = \begin{pmatrix} -1{,}219 & -0{,}015 \\ 1{,}232 & -1{,}354 \\ -0{,}013 & 1{,}369 \end{pmatrix}$$

Из табл. 8.9 видно, что оцененные и исходные данные в общем согласованы, и в то же время первый шаг подгонки не дает еще достаточного приближения теоретических оценок расстояний реальным.

После получения новых матриц $\hat{\Delta}$, $\hat{\Delta}^*$ и $\hat{X}$ первый шаг итерации считается выполненным. Его сущность оценивается по критерию F:

$$F = \sum_{ijs} \left(\delta^*_{ijs} - \hat{\delta}^*_{ijs}\right)^2, \text{ где } \hat{\delta}^*_{ijs} = \sum_k x_{ik} x_{jk} \omega^2_{ks}.$$

Цель итеративного алгоритма состоит в минимизации значений критерия F, т.е. сумма квадратов разностей между фактическими и оцененными скалярными произведениями должна быть наименьшей. Итерации повторяются до тех пор, пока при переходе к последующей итерации величина F не станет незначительной, например, меньшей 0,001. В противном случае оценки $\hat{x}_{ik}$ стандартизируются, за исходные принимаются оцененные на предыдущем шаге матрицы: $\hat{\Delta}^{-1}$, $\hat{\Delta}^{*-1}$, $\hat{X}^{-1}$ и итерации с последовательным вычислением матриц $\hat{W}$, $\hat{X}$, $\hat{\Delta}$ и $\hat{\Delta}^*$ возобновляются.

После прекращения итераций значения оценок координат стимулов стандартизируются так, чтобы их дисперсия по каждой координатной оси была равна 1,00 и последний раз оцениваются субъективные веса (ω_{ks}).

Таблица 8.10. **Расчет F-критерия качества итеративных оценок скалярных произведений**

$\delta^*_{ij1} - \hat{\delta}^*_{ij1}$	$(\delta^*_{ij1} - \hat{\delta}^*_{ij1})^2$	$\delta^*_{ij2} - \hat{\delta}^*_{ij2}$	$(\delta^*_{ij2} - \hat{\delta}^*_{ij2})^2$	$\sum_s (\delta^*_{ijs} - \hat{\delta}^*_{ijs})^2$
0,14 − 0,12= 0,02	$0{,}02^2 = 0{,}0004$	0,03	0,0009	0,0013
−0,15 − (−0,12)= 0,03	$(-0{,}03)^2 = 0{,}0009$	0	0	0,0009
0,01 − 0= 0,01	$0{,}01^2 = 0{,}0001$	−0,03	0,0009	0,0010
−0,15 − (0,12)= −0,03	$(-0{,}03)^2 = 0{,}0009$	0	0	0,0009
0,20 − 0,17 = 0,03	$0{,}03^2 = 0{,}0009$	0,04	0,0016	0,0025
−0,05 − (−0,05) = 0	0	0,04	0,0016	0,0016
0,01 − 0 = 0,01	$0{,}01^2 = 0{,}0001$	0,03	0,0009	0,0010
−0,05 − (−0,05) = 0	0	0,04	0,0016	0,0016
0,04 − 0,05 = 0,01	$0{,}01^2 = 0{,}0001$	0,03	0,0009	0,0010
$\sum$	–	–	–	0,0128

Стандартизацию вообще считают полезной не только на завершающем этапе расчетов, но и в ходе реализации алгоритма подгонки евклидовой модели. Ее осуществление для матриц скалярных произведений перед каждой итерацией позволяет предотвращать чрезмерное влияние данных какого-либо одного субъекта на итоговые аналитические результаты. Напомним, что во всех случаях вариантом нормирования данных, приводящим к стандартизованным значениям дисперсии с суммой $\sum \sigma^2 = 1$, может быть следующий:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}.$$

Трехмодальная модель — модель второго типа. Алгоритм подгонки этой модели хотя и мало отличается от рассмотренного выше подробно случая с взвешенной евклидовой моделью, предполагает проведение значительно более сложных вычислений.

Во взвешенной евклидовой модели субъективные координаты стимулов находят с учетом общей матрицы координат стимулов и матрицы субъективных весов: $X_s = XW_s$, в трехмодальной модели добавляется еще один элемент — матрица ортогональных (как правило) преобразований с вектор-строками единичной длины T_s, модель принимает вид: $X_s = XW_sT_s$. Таким образом, взвешенная евклидова модель может рассматриваться как частный случай трехмодальной модели, когда T_s — единичная матрица ($T_s = J$).

При помощи матрицы T_s обнаруживаются субъективные оценки взаимодействия координат стимулов. Так, произведение матрицы T_s на саму себя дает матрицу корреляций: $T_sT_s' = R_s$, ее элементы $r_{kk's}$ характеризуют силу связей разностей координат по всем возможным парам координатных осей, при этом знак $r_{kk's}$ указывает на направление взаимодействия, а абсолютная величина коэффициента корреляции $r_{kk'}$ — на силу взаимодействия разностей координат. Проще говоря, погружая стимулы в координатное пространство, будем учитывать величину различий стимулов, субъективные оценки значимости шкал ω_{ks} и дополнительно оценки связей латентных факторов, определяющих расположение стимулов в гипотетическом пространстве.

Теоретически различие между стимулами по субъективным оценкам представляется формулой:

$$d_{ijs} = \left(\sum_k \omega_{ks}^2 (x_{ik} - x_{jk})^2 + \sum_{kk'} \omega_{ks}\omega_{k's} r_{kk's} (x_{ik} - x_{jk})(x_{ik'} - x_{jk'}) \right)^{1/2}.$$

Модель упрощается и переходит к форме взвешенной евклидовой при $r_{kk's} = 0$, когда по субъективным оценкам взаимодействие координатных значений стимулов отсутствует. В ходе итераций для подгонки модели оцениваются матрицы: X, Δ и Δ^*, W, T.

Рассмотрим упрощенный пример определения координат стимулов для каждого субъекта, если известны параметры трехмодальной модели. Вычисления проведем по уже имеющимся расчетным данным матриц X и предполагаемым данным матриц T_1 и T_2: $X_1 = XW_1T_1$ и $X_2 = XW_2T_2$:

$$\underset{}{X_1} = \begin{pmatrix} 0{,}396 & -0{,}488 \\ 0{,}269 & 0{,}492 \\ 0{,}007 & -0{,}005 \end{pmatrix} = \underset{X}{\begin{pmatrix} -1{,}219 & -0{,}015 \\ 1{,}232 & -1{,}354 \\ -0{,}013 & 1{,}369 \end{pmatrix}} \cdot \underset{W_s}{\begin{pmatrix} 0{,}40 & 0 \\ 0 & 0{,}16 \end{pmatrix}} \cdot \underset{T_s}{\begin{pmatrix} 0{,}81 & 1{,}00 \\ 0{,}60 & 0 \end{pmatrix}},$$

$$X_2 = \begin{pmatrix} 0{,}500 & -0{,}376 \\ 0{,}505 & 0{,}257 \\ 0{,}005 & -0{,}077 \end{pmatrix} = \begin{pmatrix} -1{,}219 & -0{,}015 \\ 1{,}232 & -1{,}354 \\ -0{,}013 & 1{,}369 \end{pmatrix} \cdot \begin{pmatrix} 0{,}41 & 0 \\ 0 & 0{,}09 \end{pmatrix} \cdot \begin{pmatrix} 1{,}00 & 0{,}75 \\ 0 & 0{,}66 \end{pmatrix}.$$

Как видно, взаимодействие координат стимулов способно существенно изменить пространственное расположение и конфигурацию последних.

При рассмотрении моделей индивидуальных различий мы не обращались к вопросу о рациональном числе координатных осей для пространства стимулов. Этот вопрос можно решать в соответствии с рекомендациями, приведенными в гл. 7, или расчетным путем с использованием стресс-формул, оценивающих степень расхождений между исходными и теоретическими величинами различий стимулов:

$$SS_1 = \left(\frac{1}{S} \sum_s \frac{\sum_{ij} (\delta_{ijs}^2 - d_{ijs}^2)^2}{\sum_{ij} (d_{ijs}^2)^2} \right)^{\frac{1}{2}}.$$

Этот же критерий допустимо использовать при определении оптимальной размерности стимульного пространства каждого из субъектов:

$$SS_1 = \left(\frac{\sum_{ij} (\delta_{ijs}^2 - d_{ijs}^2)^2}{\sum_{ij} (d_{ijs}^2)^2} \right)^{\frac{1}{2}}.$$

Здесь δ_{ijs}^2 и d_{ijs}^2 — фактическая и оцененная в теоретическом пространстве величины различий стимулов i и j для субъекта s. Наилучшее число координатных осей минимизирует значения критерия SS_1.

В ходе анализа индивидуальных различий часто получают цифровой материал, не просто поддающийся интерпретации. Последний этап работы исследователя, на котором формируются выводы и делаются заключения, будет значительно облегчен, если предварительно решаются задачи оптимизации координатного стимульного пространства посредством его ортогонального или косоугольного вращения, упорядочения стимулов, построения группировок для стимулов и субъектов.

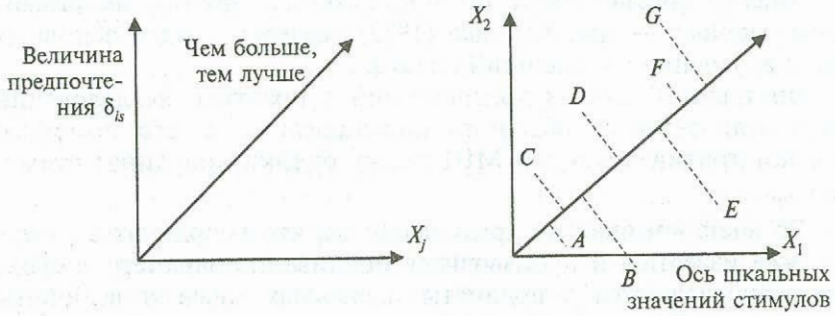
8.6. Анализ предпочтений

Одни и те же явления, процессы, как известно, субъектами оцениваются индивидуально, особенным образом. В конечном счете это проявляется в значениях весовых коэффициентов для различных шкал, о чем говорилось в предыдущем параграфе. Но что же порождает расхождение мнений и оценок, в чем кроется их причина? Существенную, если не главную, роль играют здесь представления об идеале и удаленности от идеала оцениваемого объекта. Изучение предпочтений, т.е. пространственного расположения стимулов относительно идеальных точек, становится логичным завершением исследований индивидуальных различий.

Упоминаемые в анализе предпочтений идеальные представления субъектов — наиболее тривиальная, часто встречающаяся, но не единственная проблематика. Идеальные характеристики (параметры) существуют и должны исследоваться также вне суждений субъектов, они свойственны непосредственно наблюдаемым объектам в силу специфичности их строения, взаимо-

тонного изменения оценки предпочтения: если $a_{ks} > 0$, предпочтение возрастает, при $a_{ks} < 0$ — уменьшается;

$\hat{\delta}_{is}$ — оцененные величины предпочтений для субъекта s.



а) Линейная функция предпочтений векторной модели

б) Вектор предпочтений как линия, результирующая разброс индивидуальных представлений об идеале; *A, B, C, D, F, G* — стимулы

Рис. 8.11

Параметры модели a_0 и a_{ks} определяются методами множественной регрессии с минимизацией квадрата ошибки предпочтений (ε_δ^2), т.е. по правилу наименьших квадратов:

$$\sum(\delta_{is} - \hat{\delta}_{is})^2 \to 0.$$

В неметрическом многомерном шкалировании минимизируется различие эмпирических и теоретических монотонно изменяющихся ранговых оценок предпочтений.

Статистическая оценка надежности векторной модели осуществляется при помощи коэффициентов множественной корреляции (R_V) и F-критерия с числом степеней свободы: K — числа шкал и $Y - K - 1$, где Y — число стимулов.

Исходная векторная модель для δ_{is} может быть модифицирована с целью визуализации идеальной точки, при этом осуществим переход от неопределенности положения: «чем больше» к оценке координат этой точки.

Модель получает вид:

$$\hat{\delta}_{is} = a_0 + \sum_k a_{ks} x_{ik} + a_{ks+1} q_k,$$

где $q_k = \sum_1^k x_{ik}^2$ — сумма квадратов координат стимула i. Координаты идеальной точки x_{sk} тогда находят по формуле:

$$x_{sk} = \frac{-a_{ks}}{2a_{ks+1}}, \quad k = \overline{1, K}.$$

Рассмотрим пример. Предположим, двумя экспертами, с точки зрения выбора места размещения новых объектов сельскохозяйственного производства, по данным об экологическом состоянии и уровне развития инфраструктуры определены предпочтения для шести областей Республики Беларусь. В сводной таблице обобщаются данные о рангах предпочтения каждого субъекта, соответственно координаты стимулов и исчисленные при помощи векторной модели величины предпочтений ($\hat{\delta}_{is}$) (табл. 8.11)[1].

Параметры векторной модели определяются из матричного уравнения:

$$\hat{\Delta}_s = (X_s' X_s)^{-1} X_s' \Delta_s,$$

где Δ_s и $\hat{\Delta}_s$ — эмпирический и оцененный векторы данных о предпочтениях субъекта s,

X — матрица координат стимулов.

Для вычисления значений $\hat{\delta}_{iA}$ и $\hat{\delta}_{iB}$ (см. табл. 8.8) имеем регрессионные уравнения:

$$\hat{\delta}_{iA} = 8{,}20 - 4{,}57 x_{i1} - 2{,}42 x_{i2} \text{ с } R_V^2 = 0{,}872 \text{ и } F = 4{,}78,$$

$$\hat{\delta}_{iB} = 7{,}67 - 4{,}13 x_{i1} - 2{,}05 x_{i2} \text{ с } R_V^2 = 0{,}962 \text{ и } F = 4{,}78.$$

Результаты регрессионного анализа достаточно надежны: даже без проверки на адекватность моделей видно, что найденные оценки предпочтений $\hat{\delta}_{is}$ для первого субъекта полностью совпадают с эмпирическими ранговыми данными, для второго субъекта имеются расхождения, но они несущественны.

[1] Данные могут быть не обязательно ранговыми, но и количественными.

Таблица 8.11. **Величины предпочтений, координаты стимулов и теоретические оценки предпочтений для условий организации сельскохозяйственного производства в шести областях Республики Беларусь (по данным двух экспертов)**[1]

№ п/п	Объект (стимул i)	Ранг предпочтения (δ_{is})		Координата стимулов (x_{ik})		Оценки предпочтения ($\hat{\delta}_{is}$)	
		субъектом A (δ_{iA})	субъектом B (δ_{iB})	(X_{i1})	(X_{i2})	субъекта A ($\hat{\delta}_{iA}$)	субъекта B ($\hat{\delta}_{iB}$)
1	Брестская обл.	2	1	1,15	0,30	2,44	2,30
2	Витебская обл.	4	3	0,49	0,87	4,65	3,87
3	Гомельская обл.	6	6	0,65	−0,15	5,59	5,30
4	Гродненская обл.	3	4	0,78	0,45	3,55	3,53
5	Минская обл.	1	2	1,10	1,04	0,65	1,00
6	Могилевская обл	5	5	0,20	0,90	5,11	5,00

[1] Ранг 1 - высшая оценка предпочтения.

Обратим внимание, что величины $\hat{\delta}_{is}$ позволяют сопоставлять количественно не только наблюдаемые объекты, но и сами субъекты, по оцененной силе их предпочтений. Так, Гомельская область для двух экспертов есть наименее предпочитаемая, она занимает последнее (6-е) место в ряду оцениваемых территориальных единиц и по исходным, и по теоретическим данным. Первому эксперту А эта область в большей мере кажется неудобной для развития сельскохозяйственного производства ($\hat{\delta}_{\text{Гом.А}} = 5{,}59$ и $\hat{\delta}_{\text{Гом.В}} = 5{,}30$), его негативная реакция сильнее, чем у эксперта В примерно на 5% ($\hat{\delta}_A / \hat{\delta}_B$).

В завершение примера дадим визуальное представление аналитических результатов. Предпочтение экспертов А и В покажем в трехмерной системе координат: X_1, X_2, $\hat{\delta}_{is}$ (рис.8.12). А чтобы изображение было более логичным, легче воспринималось, сделаем так, что большая величина предпочтений будет иметь соответственно и большую координату по оси $\hat{\delta}$, для этого примем вместо оценок $\hat{\delta}_{is}$ масштабированные оценки $1/\hat{\delta}_{is}$.

В общем на рис. 8.12 заметно, что экспертом В чаще предлагаются более оптимистические оценки, исключение отмечается только по Минской области. Легко узнаваемы шесть областей по степени предпочтительности экспертами: наиболее благоприятны условия для производства в Минской области, она имеет довольно значительный отрыв по величине $\hat{\delta}$ от всех других областей; затем довольно плотной группой идут Брестская, Гродненская и Витебская области. Наконец, особняком остается Гомельская область с наихудшими оценками перспектив для размещения производственных объектов.

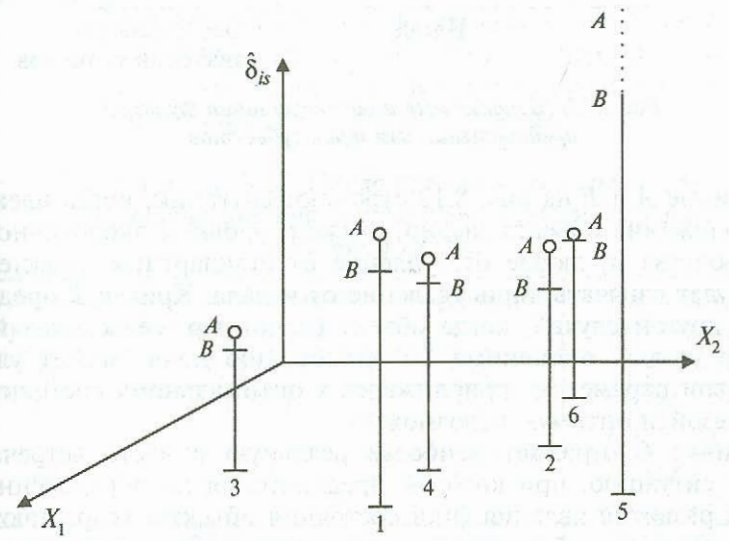


Рис. 8.12. ***Предпочтение двух экспертов А и В при выборе административной области (по данным табл. 8.8, гр. 1) для перспективного ведения сельскохозяйственного производства***

В отличие от векторной три других типа моделей — простая евклидова (*E*), взвешенная (*W*) и обобщенная (*G*) евклидовы — принадлежат к классу нелинейных. Для практических исследований они часто более предпочтительны, так как обладают меньшей инерционностью и повышенными адаптационными свойствами. Формально для описания зависимости предпочтений и шкальных значений в дистанционной модели может быть использована монотонная функция (кривые *A* и *B*) или одновершинная немонотонная функция (кривая *C*), как показано на рис. 8.13.

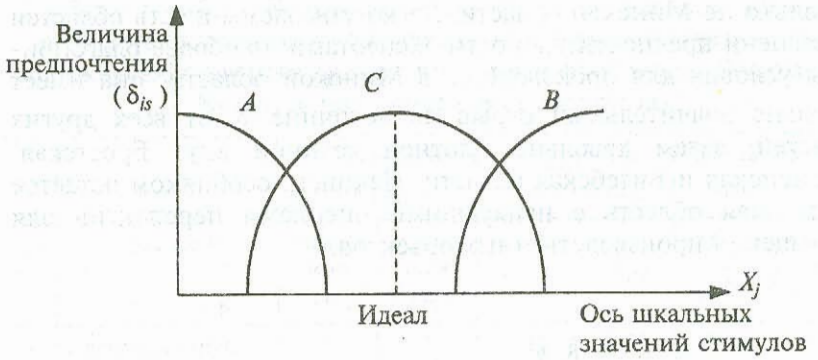


Рис.8.13. Монотонные и одновершинная функции предпочтения для трех субъектов

Кривые *A* и *B* на рис. 8.13 отражают ситуации, когда идеальное положение объекта задано, скажем, уровнем экологичности производства *A*, любое отступление от стандартных характеристик будет означать лишь удаление от идеала. Кривая *B* предполагает другой случай, когда объект (например, сельскохозяйственные угодья, оцененные по плодородию почв) может улучшать свои параметры, приближаясь к оптимальному состоянию, но превзойти оптимум невозможно.

Кривая *C* отражает наиболее реальную и часто встречающуюся ситуацию, при которой предпочтения до определенного уровня развития явления (или состояния объекта) возрастают, а затем монотонно убывают. Скажем, спрос на кредитные средства может расти до некоторого известного уровня кредитной ставки, но превышение этого уровня постепенно сокращает число фирм, желающих воспользоваться кредитом, то же с размером налоговых ставок и уровнем деловой активности и т. д.

С учетом того, что основные алгоритмические шаги вычислительных процедур для дистанционных моделей в теории анализа предпочтений не претерпевают существенных изменений и рассматривались в предыдущих главах, обратим внимание только на общие вопросы приложения этих моделей в исследовательской практике.

В табл. 8.13 с целью сравнения приводятся четыре основных типа внешних моделей, включая векторную. Видно, что все модели содержат аддитивную константу a_0, выполняющую роль

балансира для оценок модельных параметров. Адекватность моделей проверяется при помощи двух основных критериев: R^2 и F, соответственно можно выбрать оптимальный тип модели для конкретного анализа:

$R_{V(E,W,G)}$ — множественный коэффициент детерминации, служит мерой соответствия эмпирических и теоретических оценок предпочтений δ_{is}.

$F_{V(E,W,G)}$ — как и в регрессионном анализе, оценка адекватности модели в целом. Критическое значение F-критерия находится по таблицам Фишера при заданном α-уровне значимости и соответствующем числе степеней свободы ν_1 и ν_2 (табл. 8.12).

Таблица 8.12. **Число степеней свободы (ν_1, ν_2) для F-критерия при оценке адекватности моделей внешнего анализа**

F-критерий	*Число степеней свободы (ν_1, ν_2)*	*F-критерий*	*Число степеней свободы (ν_1; ν_2)*
F_V	K, $J-K-1$	F_{W3}	$K-1$, $J-2K-1$
F_{E1}	$K+1$, $J-K-1$	F_{G1}	$1/2K\,(K+3)$ $J-1/2K\,(K+3)-1$
F_{E2}	1, $J-K-2$	F_{G2}	$1/2K\,(K+1)$ $J-1/2K\,(K+3)-1$
F_{W1}	$2K$, $J-2K-1$	F_{G3}	$1/2K\,(K+1)-1$ $J-1/2K\,(K+3)-1$
F_{W2}	K, $J-2K-1$	F_{G4}	$1/2K\,(K-1)$ $J-1/2K\,(K+3)-1$

Евклидовы модели типов *E*, *W*, *G* в табл. 8.13 характерно различаются оценками весов W^2 координатных осей и включением метрической разности: $(x_{sk}-x_{sk})^2$. Все эти типы моделей позволяют устанавливать координаты идеальных точек, но различаются они способностью учитывать разную значимость и силу взаимодействия координатных осей. По простой евклидовой модели с весовым коэффициентом w_s^2 весовая нагрузка шкал (координатных осей) оценивается так, что для одного

субъекта все w_s^2 равны, т.е., считается, что для него все координатные оси одинаково важны. Посредством несложных преобразований простую евклидову модель приводят к векторному виду:

$$\hat{\delta}_{is} = a_0 + w_s^2 \sum_k x_{ik} + \sum_k a_{ks} x_{ik}, \text{ где } a_0 = \sum_k w_s^2 x_{sk}^2 + a_0^* \text{ и } a_{ks} = -2w_s^2 x_{sk}.$$

При $w_s = 0$ эта формула тождественна записи векторной модели, поэтому задачи поиска весового коэффициента w_s^2 и оценки пространственного положения идеальной точки имеют регрессионное решение:

$$x_{sk} = \frac{a_{ks}}{-2w_s^2},$$

где a_{ks} — коэффициент регрессии.

Вывод об общности с векторной моделью распространяется и на два последующих случая, алгебраически доказывается, что взвешенная и обобщенная евклидовы модели — это разные уровни ее интегрирования. При этом взвешенная евклидова модель сводится при упрощении к векторной или простой евклидовой модели, обобщенная евклидова в свою очередь включает векторную, простую и взвешенную евклидовы модели. Это означает, что множественная регрессия остается универсальным приемом для поиска модельных параметров и координат идеальной точки.

Взвешенная евклидова модель содержит весовой коэффициент w_{ks}^2 и в отличие от предыдущего случая позволяет оценить значимость каждой координатной оси отдельно. После возведения в квадрат выражения в скобках и элементарных перестановок (табл. 8.13) будем иметь расширенную запись модели:

$$\hat{\delta}_{is} = a_0^* + \sum_k w_{ks}^2 x_{ik}^2 - 2\sum_k w_{ks}^2 x_{sk} x_{ik} + \sum_k w_{ks}^2 x_{sk}^2 \quad \text{или}$$

$$\hat{\delta}_{is} = a_0 + \sum_k w_{ks}^2 x_{ik}^2 + \sum_k a_{ks} x_{ik}, \text{ если } a_0 = \sum_k w_{ks}^2 x_{sk}^2 + a_0^*, \text{ и } a_{ks} = -2w_{ks}^2 x_{sk}.$$

В рамках модели координаты идеальной точки находят по формуле:

$$x_{sk} = \frac{a_{ks}}{-2w_{ks}^2}.$$

Обобщенная евклидова модель в дополнение к моделям типов E и W учитывает силу взаимодействия координатных осей

Таблица 8.13. **Виды моделей внешнего анализа и критериев для поиска наилучшей модели**

Модель	*Условное обозначение*	*Вид модели*	*Критерий для отбора наилучшей модели*
Векторная	V	$\hat{\delta}_{is} = a_0 + \sum_k a_{ks} x_{ik}$	$R_V^2,\ F_V = \frac{R_V^2}{1-R_V^2} \cdot \frac{J-K-1}{K}$
Простая евклидова	E	$\hat{\delta}_{is} = a_0^* + \sum_k w_s^2 (x_{ik} - x_{sk})^2$	$R_E^2,\ F_{E1} = \frac{R_E^2}{1-R_E^2} \cdot \frac{J-K-1}{K+1},\ F_{E2} = \frac{R_E^2 - R_V^2}{1-R_E^2} \cdot (J-K-2)$
Взвешенная евклидова	W	$\hat{\delta}_{is} = a_0^* + \sum_k w_{ks}^2 (x_{ik} - x_{sk})^2$	$R_W^2,\ F_{W1} = \frac{R_W^2}{1-R_W^2} \cdot \frac{J-2K-1}{2K}$ $F_{W2} = \frac{R_W^2 - R_V^2}{1-R_W^2} \cdot \frac{J-2K-1}{K},\ F_{W3} = \frac{R_W^2 - R_E^2}{1-R_W^2} \cdot \frac{J-2K-1}{K-1}$
Обобщенная евклидова	G	$\hat{\delta}_{is} = a_0^* + \sum_k w_{ks}^2 (x_{iks} - x_{sks})^2$ в матричном виде уравнение для поиска координат стимулов для субъекта s (X_s): $X_s = X w_s T_s$	$R_G^2,\ F_{G1} = \frac{R_G^2}{1-R_G^2} \cdot \frac{J-1/2K(K+3)-1}{1/2K(K+3)}$ $F_{G2} = \frac{R_G^2 - R_V^2}{1-R_G^2} \cdot \frac{J-1/2K(K+3)-1}{1/2K(K+1)}$ $F_{G3} = \frac{R_G^2 - R_E^2}{1-R_G^2} \cdot \frac{J-1/2K(K+3)-1}{1/2K(K+1)-1}$ $F_{G4} = \frac{R_G^2 - R_W^2}{1-R_G^2} \cdot \frac{J-1/2K(K+3)-1}{1/2K(K-1)}$

по данным субъекта s. В развернутом виде модель записывается:

$$\hat{\delta}_{is} = a_0 + \sum w_{ks}^2 (x_{ik} - x_{sk})^2 + \sum_k w_{ks} w_{k's} r_{kk's} (x_{ik} - x_{sk})(x_{ik'} - x_{sk'}) \text{ при } k' \neq k .$$

Согласно модели показатель силы взаимодействия координатных осей $r_{kk's}$ вычисляют по формуле:

$$r_{kk's} = \frac{a_{kk's}}{w_{ks} w_{k's}}; \quad a_{kk's} = 2 w_{ks} w_{k's} r_{kk's},$$

здесь k и k' — различные координатные оси,

$a_{kk's}$ — коэффициент регрессии при независимой переменной, представляющей взаимодействие осей.

Координаты идеальной точки находят из матричного уравнения: $X_s = -\frac{1}{2} R_s^{-1} A_s$, здесь X_s — вектор координат идеальной точки с компонентами x_{sk}; R_s — симметрическая матрица взаимодействия координатных осей, на ее главной диагонали находятся элементы ω_{ks}^2, а все другие элементы — это произведения: $w_{ks} w_{k's} r_{kk's}$, или, проще, $\frac{1}{2} a_{kk's}$. A_s — вектор с компонентами a_{ks} — регрессионными коэффициентами для переменных x_{ik}.

В практике набор сложных формул при решении евклидовых моделей не столь существенно затрудняет работу. Возможность использования хорошо известных алгоритмов регрессионного анализа требует лишь предварительной правильной подготовки данных для обработки.

В табл. 8.14 показано, каким образом задаются входные данные и какие результаты после обработки этих данных методами множественной регрессии получают. Вычисления повторяются для каждого субъекта s.

По данным примера (табл. 8.11) сделаны расчеты входной информации для последующей оценки регрессионных весов (табл. 8.15).

После выполнения алгоритма многошаговой регрессии получены величины параметров (a_{ks}, $a_{kk's}$) и значений весовых коэффициентов (w_s^2, w_{ks}^2), по ним исчислены координаты идеальных точек субъектов, все результаты анализа сведены в табл. 8.16. Чтобы особенности упомянутых в этом разделе моделей стали

Таблица 8.14. **Входные данные и параметрические оценки евклидовых моделей типа *E, W, G* после проведения регрессионного анализа**

Входные данные	*Выходные данные: параметрические оценки (регрессионные веса)*		
	Простая евклидова (E)	*Взвешенная евклидова (W)*	*Обобщенная евклидова (G)*
Отклик (зависимая переменная) δ_{is}			
Предикторы			
x_{ik}	a_{ks}	a_{ks}	a_{ks}
$\sum_k x_{ik}^2$	w_s^2	—	—
x_{ik}^2	—	w_{ks}^2	w_{ks}^2
$x_{ik'}^2$	—	$w_{k's}^2$	$w_{k's}^2$
$x_{ik}\ x_{ik'}$	—	—	$a_{kk's}$

Таблица 8.15. **Расчетные входные данные для оценки регрессионных весов (для двух субъектов)**

Стимул (область)	*Зависимая переменная*		*Независимые переменные*					
	Оценки предпочтений		*Координаты стимулов (x)*		x_{i1}^2	x_{i2}^2	$x_{i1}^2 + x_{i2}^2$	$x_{i1} x_{i2}$
	δ_A	δ_B	x_{i1}	x_{i2}				
Брестская	2	1	1,15	0,30	1,322	0,090	1,412	0,345
Витебская	4	3	0,49	0,87	0,240	0,757	0,997	0,426
Гомельская	6	6	0,65	−0,15	0,422	0,223	0,645	−0,098
Гродненская	3	4	0,78	0,45	0,608	0,202	0,810	0,351
Минская	1	2	1,10	1,04	1,210	1,082	2,292	1,144
Могилевская	5	5	0,20	0,90	0,040	0,810	0,850	0,180

Таблица 8.16. **Регрессионные и весовые коэффициенты, оценки координат идеальных точек и значения критериев надежности решений для четырех типов внешних моделей анализа предпочтений (для двух экспертов)**

Показатель	Координатная ось (X_k)	Тип внешней модели									
		Векторная (V)		Простая евклидова (E)			Взвешенная евклидова (W)			Обобщенная евклидова (G)	
		Субъект		Субъект			Субъект			Субъект	
		A	B	A	B		A	B		A	B
Регрессионный коэффициент (a_{ks})	I	−4,57	−4,13	−6,49	−6,61		−4,42	−0,163		9,67	1,10
	II	−2,42	−2,05	−3,70	−3,70		−3,74	−3,807		5,96	−8,76
Субъективные веса (w_s^2)	I, II	—	—	1,45	1,88		-	-		—	—
Субъективные веса координат (w_{ks}^2)	I	—	—	—	—		2,77	−2,58		−5,17	−7,36
	II	—	—	—	—		1,72	2,69		4,98	4,77
Субъективная оценка взаимодействия шкал ($a_{kk's}$)	I, II	—	—	—	—		—	—		−8,30	−5,15
Координаты идеальной точки (x_{sk})	I	—	—	2,24	1,76		0,80	−0,03		−0,85	−0,16
	II	—	—	1,27	0,98		1,09	0,71		−0,11	−0,82
Коэффициент детерминации (R^2)	—	0,872	0,962	0,989	0,806		0,996	0,867		0,994	0,858
F-критерий	—	4,78	4,78	61,8	2,77		56,6	1,636		34,2	1,210

очевидны, таблица дополнена данными расчетов также по векторной модели и критериальными оценками надежности решений: R^2 и F (F — обычный F-критерий, применяемый для оценки адекватности регрессионной модели).

По данным табл. 8.16 более высокими адаптационными свойствами обладают модели для субъекта А, в их числе наилучшие — простая и взвешенная эвклидовы, для них $R^2 > 0{,}980$ и $F > 56$. Отрицательные значения коэффициентов регрессии a_{ks} свидетельствуют, что предпочтения субъекта в общем представляются монотонно убывающей функцией. В двумерной системе координат более серьезное значение им отводится оси $X_1(w_{1A}^2 w_{2A}^2)$, наконец, идеальная точка приближается к стимулу «Минская область», сравним: $X'_{kA} = (2{,}24\ 1{,}27)$ и $X'_{iA} = (1{,}10\ 1{,}04)$, т.е. наиболее благоприятные условия для организации производства видятся на территории, близкой по своим параметрам Минской области, но по шкале (признаку) X_1 требуется существенное улучшение.

Для субъекта В наилучшей является векторная модель ($R^2 = 0{,}962$ и $F = 4{,}78$), а для всех других типов моделей значения показателей адекватности значительно снижаются. Выводы для субъекта В можно предложить в последовательности, как и для А.

Обратим внимание, что в случаях со взвешенной и обобщенной евклидовыми моделями появляются отрицательные оценки субъективных весов координат (w_{ks}^2). Это допустимо, и по Кэроллу объясняется появлением антиидеальной точки X_{ks}, чем меньше стимул (i-область) похож на «антиидеал», тем больше он нравится субъекту.

Отрицательный знак оценки взаимодействия шкал не вызывает противоречий, это результат разнонаправленности действия X_1 и X_2 (не забудем, X_1, X_2 — некоторые обобщенные признаки), положительный знак свидетельствовал бы о том, что X_1, X_2 действуют в одном направлении.

Расчеты координат идеальных точек для субъектов А и В идентичны, поэтому покажем, как получены табличные значения x_{ks} только для одного субъекта А:

- Простая евклидова модель

$$x_{A1} = \frac{a_{1A}}{-2w_A^2} = \frac{-6{,}49}{-2 \cdot 1{,}45} = 2{,}24; \qquad x_{A2} = \frac{a_{2A}}{-2w_A^2} = \frac{-3{,}70}{-2 \cdot 1{,}45} = 1{,}27,$$

и вектор координат идеальной точки $X'_A = (2{,}24;\ 1{,}27)$.

- Взвешенная евклидова модель

$$x_{A1} = \frac{a_{1A}}{-2w_{1A}^2} = \frac{-4{,}42}{-2\cdot 2{,}77} = 0{,}80; \quad x_{A2} = \frac{a_{2A}}{-2w_{2A}^2} = \frac{-3{,}74}{-2\cdot 1{,}72} = 1{,}09$$

и $X'_A = (0.80;\ 1.09)$.

- Обобщенная евклидова модель

$$X_A = -\frac{1}{2}R_A^{-1}A_A; \quad X_A = -\frac{1}{2}\begin{pmatrix} -5{,}17 & -4{,}15 \\ -4{,}15 & 4{,}98 \end{pmatrix}^{-1} \times \begin{pmatrix} 9{,}67 \\ 5{,}96 \end{pmatrix} = \begin{pmatrix} -0{,}848 \\ -0{,}109 \end{pmatrix},$$

здесь антиидеальная точка имеет координаты $X'_A = (-0{,}848;\ -0{,}109)$.

Используя значения регрессионных весов для обобщенной евклидовой модели, дополнительно можно исчислить характеристику коррелированности осей k и k', т.е. показатель $r_{kk's}$:

$$r_{kk's} = \frac{a_{kk's}}{2w_{ks}w_{k's}} = \frac{-8{,}30}{2(5{,}17\cdot 4{,}98)^{1/2}} = -0{,}818\,.$$

Приведенные расчеты подтверждают важность правильного выбора модели в анализе и достижения адекватности расчетных модельных параметров. Как видим, аналитические результаты существенно различаются: оценки регрессионных и весовых коэффициентов в моделях имеют значительные расхождения, координаты идеальных точек трансформируются и переходят в координаты «антиидеала».

Таким образом, МШ данных о предпочтениях включает два этапа, соответственно этому и сам анализ разделяется на внутренний и внешний. На первом этапе выдвигается тривиальная задача поиска координат стимулов, решаемая методами метрического или неметрического шкалирования при условии адекватности теоретических пространственных оценок предпочтений эмпирическим. Проверка качества координатных оценок на этом этапе обычно осуществляется при помощи метода наименьших квадратов или известных стресс-формул. На втором этапе анализа, т.е. во внешнем анализе, устанавливаются субъективные параметры моделей предпочтений: весовые и регрессионные коэффициенты, координаты идеальных точек.

Для поиска адекватных решений имеется широкий набор моделей, сводящихся к основным четырем типам: векторная, простая, взвешенная и обобщенная евклидовы модели, их параметры могут быть найдены методами множественной регрессии. Соответствие модельных значений фактическим статистически оценивается при помощи коэффициента детерминации (R^2) и F — дисперсионного критерия Фишера.

Анализ предпочтений органично сочетается с другими статистическими методами: многомерных группировок, дискриминантного анализа и т.д. Представляя преимущества группировки субъектов по предпочтениям, можно сказать, например, о выделении перспективных групп покупателей, товара. Сгущения идеальных точек помогут найти оптимальные параметры реальных объектов. Если же в координатном пространстве шкал идеальные объекты пространственно удалены от реальных, то следует вывод о необходимости совершенствования последних, что будет означать необходимость разработки, скажем, новой продукции: средств производства, потребительских товаров, или принятие качественно нового управляющего решения и т. п.

Как видим, анализ предпочтений позволяет решать широкий спектр практических задач. Процессы и явления из области экономики, политики, социальной жизни моделируются таким образом, что исследователь определяет перспективы развития и желаемые (идеальные) результаты.

Выводы

Многомерное шкалирование представляет собой совокупность методов, которые по характеристикам различий позволяют погружать изучаемые объекты и субъектов-экспертов в некоторые теоретические пространства с размерностью, как правило, допускающей визуализацию аналитических результатов. Пространственное расположение объектов, по предварительному условию, адекватно реальной ситуации. С учетом экспертных мнений это расположение можно идеализировать и тогда, когда исследователь обоснованно решает задачу поиска оптимальных параметров для изучаемых объектов или выдвигает гипотезу о необходимости конструирования нового образца и последующей замены существующего того или иного объекта.

Арсенал МШ достаточно богат и сложным вопросом остается правильный выбор метода в конкретном анализе. Классификация методов прежде всего обусловлена исходной информацией.

По наличию в ней количественных или порядковых характеристик МШ соответственно разделяют на метрическое и неметрическое.

В основе МШ лежат модели двух классов: линейные (векторная) и нелинейные (евклидова — простая, взвешенная и т.д.). Алгоритмически методы, как и модели МШ, имеют общие этапы реализации: подготовки данных о различиях наблюдаемых объектов, поиска и оптимизации теоретического координатного пространства, оценки качества решения и интерпретации аналитических результатов.

Проведение анализа методами МШ подразумевает выполнение сложных итеративных процедур и практически всегда привлечения технических средств. Важнейшим требованием являются профессиональные навыки работы с другими статистическими методами, так как МШ часто комплексно используется с методами главных компонент и факторного анализа, корреляционно-регрессионного анализа, кластер-анализа и др.

Теория МШ — одна из новейших в статистике — сформировалась в 50—60-е годы нашего века и в настоящее время продолжает активно разрабатываться. Широкий спрос и быстрый рост сферы практического приложения позволяют оценивать перспективы для нее весьма оптимистично.

Вопросы и задачи

1. Какие особенности в сравнении с другими статистическими методами отличают методы многомерного шкалирования? Приведите собственные примеры одномерного и многомерного шкалирования.

2. Перечислите основные виды многомерного шкалирования. Для решения каких типов задач они применимы?

3. Покажите основные формальные модели, лежащие в основе алгоритмов многомерного шкалирования, в чем их различие?

4. Известны две матрицы исходных данных: (а) — о предпочтениях студентами профессий при поступлении в университет и после его окончания; (б) — о предпочтениях экспертами акций

пяти белорусских компаний в момент выпуска ценных бумаг и спустя один год:

а)

Стимул \ Стимул	М	Б	Э	П
Медицина (М)	0,30	0,20	0,40	0,10
Биология (Б)	0,20	0,30	0,10	0,40
Экономика (Э)	0,05	0,10	0,70	0,15
Педагогика (П)	0,02	0,08	0,30	0,60

б)

Стимул \ Стимул	Амкодор	Белмед	Горизонт	Ранар	Субинвест
Амкодор	0,60	0,10	0,18	0,12	0,00
Белмед	0,35	0,30	0,20	0,05	0,10
Горизонт	0,20	0,20	0,40	0,10	0,10
Ранар	0,35	0,20	0,23	0,15	0,07
Субинвест	0,42	0,24	0,18	0,06	0,10

Постройте симметрические матрицы сходств (Δ), поясните смысл данных в исходных и преобразованных матрицах.

5. На рис. 8.14 в двумерном пространстве представлены пять стран: Беларусь, США, Франция, Великобритания, Россия (данные за 1993 г.). Оси координат: X_1 — размер ВВП, тыс. дол. США в расчете на душу населения и X_2 — средняя продолжительность жизни населения, лет:

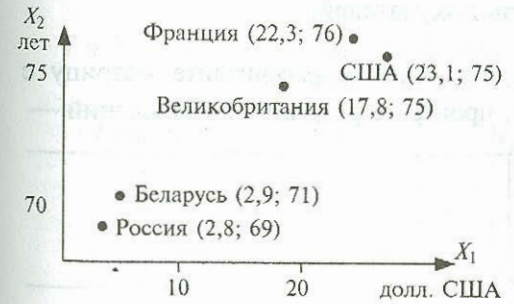


Рис. 8.14. Расположение пяти стран мира в двумерном пространстве, по данным за 1993 г.

а) Оцените расстояния между всеми парами стран.

б) Каким образом будут проранжированы страны, если игнорировать координатную ось X_1 — размер ВВП на душу населения, и какие страны будут при этом наиболее близки по своим характеристикам Республике Беларусь? Изменятся ли ранги стран, если игнорируется ось X_2 — средняя продолжительность жизни?

6. Имеются экспертные ранговые оценки экономичности X_1, надежности X_2 и долговечности X_3 четырех марок автомобилей: «Жигули», «Опель», «Вольво», «Ситроен» (1 — оценка высшего уровня) (рис. 8.15):

Рис. 8.15. ***Четыре типа автомобилей в трехмерном координатном пространстве по экспертным оценкам уровня их экономичности, надежности и долговечности***

а) Постройте симметрическую матрицу различий автомобилей четырех марок.

б) Принимая во внимание, что имеются две частные оценки покупателями идеальных автомобилей: Владимиром — (1,2; 2,0; 1,1) и Виктором — (1,6; 2,4; 2,8), определите модели автомобилей, которые предпочтет каждый из субъектов, и по степени различия этих моделей, расположению идеальных точек оцените величину расхождения вкусов покупателей.

7. По данным матрицы различий Δ рассчитайте матрицу с двойным центрированием Δ^*, проверьте результаты вычислений:

$$\Delta = \begin{pmatrix} 0 & 0{,}070 & 1{,}40 & 1{,}80 \\ 0{,}70 & 0 & 0{,}95 & 0{,}64 \\ 1{,}40 & 0{,}95 & 0 & 1{,}92 \\ 1{,}80 & 0{,}64 & 1{,}92 & 0 \end{pmatrix}.$$

8. Известны статистические данные шести белорусских акционерных обществ (по данным за 1993 г., табл. 8.17)

Таблица 8.17. **Статистические данные о текущих активах, затратах и финансовых обязательствах акционерных обществ, тыс. руб.**

Акционерное общество	*Текущие активы*	*Затраты*	*Краткосрочные финансовые обязательства*
«Стройинвест»	14377	10000	29767
«Минская биржа»	10354	957	1949
Торговый дом «Приднепровье»	24954	21745	22512
Белорусская биржа «Нефтепродукты и полимеры»	12139	292	4998
Западная биржа «Недвижимость»	16943	485	7491

а) Постройте матрицу различий Δ и матрицу с двойным центрированием Δ^*.

б) Используя метод главных компонент, определите координатное пространство стимулов, покажите на рисунке расположение объектов в пространстве двух первых координатных осей (шкал).

в) Назовите первые шкалы, прокомментируйте полученные аналитические результаты.

9. По имеющимся ранговым данным сравнительной попарной оценки качества четырех видов пищевого продукта (табл. 8.18) и полученным координатам стимулов (табл. 8.19) определите степень соответствия теоретических и эмпирических данных, по стресс-формулам Юнга SS_1 и SS_2 выполните первый шаг по улучшению координат стимулов.

Таблица 8.18. **Матрица данных о различиях четырех видов пищевых продуктов по качеству**

Продукт	А	Б	С	Д
А	—			
Б	2	—		
С	3	1	—	
Д	6	4	5	—

Таблица 8.19. **Координаты стимулов на стартовой конфигурации**

Продукт	*Координаты*	
	X_1	X_2
А	3	3
Б	1	6
С	2	4
Д	8	5

10. По данным опроса двух субъектов, идентифицирующих показатели эффективности коммерческой деятельности торговых предприятий, получены следующие матрицы различий:

$$\Delta_1 = \begin{pmatrix} 0{,}00 & 0{,}15 & 0{,}75 \\ 0{,}15 & 0{,}00 & 0{,}35 \\ 0{,}75 & 0{,}35 & 0{,}00 \end{pmatrix} \quad \text{и} \quad \Delta_2 = \begin{pmatrix} 0{,}00 & 0{,}25 & 0{,}60 \\ 0{,}25 & 0{,}00 & 0{,}40 \\ 0{,}60 & 0{,}40 & 0{,}00 \end{pmatrix}.$$

Необходимо последовательно произвести расчеты:

а) матриц скалярных произведений (Δ^*_1 и Δ^*_2);

б) средней матрицы скалярных произведений (Δ^*);

в) матрицы с данными «стартовой конфигурации», используя метод главных компонент. При этом решите вопрос о числе шкал, которое было бы предпочтительным для дальнейшего анализа индивидуальных различий.

11. Сведения экспертов обобщены в матрицах различий:

$$\Delta_1 = \begin{pmatrix} 0 & 2 & 3 \\ 2 & 0 & 5 \\ 3 & 5 & 0 \end{pmatrix} \text{ и } \Delta_2 = \begin{pmatrix} 0 & 1 & 5 \\ 1 & 0 & 4 \\ 5 & 4 & 0 \end{pmatrix}.$$

Соответственно им рассчитаны матрицы скалярных произведений, средняя матрица скалярных произведений и матрица данных «стартовой конфигурации»:

$$\Delta^*_1 = \begin{pmatrix} 0{,}11 & 0{,}77 & -0{,}88 \\ 0{,}77 & 5{,}44 & -6{,}22 \\ -0{,}88 & -6{,}22 & 7{,}11 \end{pmatrix}, \quad \Delta^*_2 = \begin{pmatrix} 4 & 2 & -6 \\ 2 & 1 & -3 \\ -6 & -3 & 9 \end{pmatrix},$$

$$\Delta^* = \frac{1}{2}(\Delta^*_1 + \Delta^*_2) = \begin{pmatrix} 2{,}06 & 1{,}39 & -3{,}45 \\ 1{,}39 & 3{,}22 & -4{,}61 \\ -3{,}45 & -4{,}61 & 8{,}06 \end{pmatrix},$$

Стимул \ *Шкала*		X_1	X_2
	С	−0,982	0,177
X =	Д	−0,990	−0,159
	Е	1,000	0,017

Используя приведенные выше данные, сделайте первый шаг подгонки взвешенной евклидовой модели, для шага (+1) постройте матрицы оценок расстояний (Δ_{1+}, Δ_{2+}) и скалярных произведений (Δ^*_{1+}, Δ^*_{2+}).

12. Используя векторную модель Такера, произведите оценку предпочтений фирменной марки видеомагнитофона. В ходе обследований состояния локального рынка видеоаппаратуры имеются ранговые данные о предпочтениях эксперта *s* и матрица координат стимулов *X*:

Марка видеомагнитофона	*Ранг предпочтения*	*Координаты стимулов*	
		X_1	X_2
Филиппс	5	9	8
Орион	1	2	5
Сони	7	14	11
Шнайдер	3	6	4
Шарп	6	12	7
Нокиа	4	8	6
Айва	2	5	10

После проведения модельных расчетов найдите и интерпретируйте характеристики адекватности модели по R^2-, *R*-, *F*-критериям. Эмпирические и теоретические данные о предпочтениях представьте графически, определите характер функции предпочтения.

Глава 9

Кластерный анализ

9.1. Общая характеристика методов кластерного анализа

Кластерный анализ — это совокупность методов, позволяющих классифицировать многомерные наблюдения, каждое из которых описывается набором исходных переменных X_1, X_2, ..., X_m. Целью кластерного анализа является образование групп схожих между собой объектов, которые принято называть кластерами. Слово *кластер* английского происхождения (*cluster*), переводится как *сгусток*, *пучок*, *группа*. Родственные понятия, используемые в литературе, — *класс*, *таксон*, *сгущение*.

В отличие от комбинационных группировок кластерный анализ приводит к разбиению на группы с учетом всех группировочных признаков одновременно. Например, если каждый наблюдаемый объект характеризуется двумя признаками X_1 и X_2, то при выполнении комбинационной группировки вся совокупность объектов будет разбита на группы по X_1, а затем внутри каждой выделенной группы будут образованы подгруппы по X_2. Такой подход получил название *монотетического*. Определить принадлежность каждого объекта к той или иной группе можно, последовательно сравнивая его значения X_1 и X_2 с границами выделенных групп. Образование группы в этом случае всегда связано с указанием ее границ по каждому группировочному признаку отдельно.

В кластерном анализе используется иной принцип образования групп, так называемый *политетический* подход. Все группировочные признаки одновременно участвуют в группировке, т.е. они учитываются все сразу при отнесении наблюдения в ту или иную группу. При этом, как правило, не указаны четкие границы каждой группы, а также неизвестно заранее, сколько же групп целесообразно выделить в исследуемой совокупности.

Кластерный анализ — одно из направлений статистического исследования. Особо важное место он занимает в тех отраслях науки, которые связаны с изучением массовых явлений и процессов. Необходимость развития методов кластерного анализа и их использования продиктована прежде всего тем, что они помогают построить научно обоснованные классификации, выявить внутренние связи между единицами наблюдаемой совокупности. Кроме того, методы кластерного анализа могут использоваться с целью сжатия информации, что является важным фактором в условиях постоянного увеличения и усложнения потоков статистических данных.

Первые публикации по кластерному анализу появились в конце 30-х годов нашего столетия, но активное развитие этих методов и их широкое использование началось в конце 60-х — начале 70-х годов [61]. В дальнейшем это направление многомерного анализа очень интенсивно развивалось. Появились новые методы, новые модификации уже известных алгоритмов, существенно расширилась область применения кластерного анализа. Если первоначально методы многомерной классификации использовались в психологии, археологии, биологии, то сейчас они стали активно применяться в социологии, экономике, статистике, в исторических исследованиях. Особенно расширилось их использование в связи с появлением и развитием ЭВМ и, в частности, персональных компьютеров. Это связано прежде всего с трудоемкостью обработки больших массивов информации (вычисление и обращение матриц больших размерностей).

Методы кластерного анализа позволяют решать следующие задачи:

- проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов. Решение такой задачи, как правило, приводит к углублению знаний о совокупности классифицируемых объектов;
- проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов, т.е. поиск существующей структуры;
- построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности и попытаться привнести в нее структуру.

Методы кластерного анализа можно разделить на две большие группы: агломеративные (объединяющие) и дивизимные (разделяющие). *Агломеративные методы* последовательно объединяют отдельные объекты в группы (кластеры), а *дивизимные методы* расчленяют группы на отдельные объекты. В свою очередь каждый метод как объединяющего, так и разделяющего типа может быть реализован при помощи различных алгоритмов. Отдельные примеры агломеративных и дивизимных алгоритмов рассмотрены в § 9.3. В частности, наиболее подробно описан самый доступный для понимания иерархический агломеративный кластерный анализ. Следует заметить, что как агломеративные, так и дивизимные алгоритмы трудоемки и их сложно использовать для больших совокупностей. Кроме того, результаты работы таких алгоритмов (их графическое изображение) трудно поддаются визуальному анализу.

В кластерном анализе существуют также методы, которые трудно отнести к первой или ко второй группе. Например, итеративные методы, в частности, метод k-средних и метод поиска сгущений, описанные в § 9.4 и 9.5. Их характерная особенность в том, что кластеры формируются исходя из задаваемых условий разбиения (параметров), которые в процессе работы алгоритма могут быть изменены пользователем для достижения желаемого качества разбиения. Итеративные методы относятся к быстродействующим, что позволяет использовать их для обработки больших массивов исходной информации.

В отличие от агломеративных и дивизимных методов итеративные алгоритмы могут привести к образованию пересекающихся кластеров, когда один объект может одновременно принадлежать нескольким кластерам.

Достаточно подробный обзор и систематизация различных методов кластерного анализа приводятся в работе [61].

Для удобства записи формализованных алгоритмов кластерного анализа введем следующие условные обозначения:

$X_1, X_2..., X_n$ — совокупность объектов наблюдения;

$X_i = (X_{i1}, X_{i2}, ..., X_{im})$ — i-е многомерное наблюдение в m-мерном пространстве признаков ($i = 1, 2, ..., n$);

d_{il} — расстояние между i-м и l-м объектами;

z_{ij} — нормированные значения исходных переменных;

D — матрица расстояний между объектами.

9.2. Меры сходства

Для проведения классификации необходимо ввести понятие сходства объектов по наблюдаемым переменным. В каждый кластер (класс, таксон) должны попасть объекты, имеющие сходные характеристики.

В кластерном анализе для количественной оценки сходства вводится понятие *метрики*. Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается k признаками, то он может быть представлен как точка в k-мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние. В кластерном анализе используются различные меры расстояния между объектами:

1) евклидово расстояние: $d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$,

2) взвешенное евклидово расстояние: $d_{ij} = \sqrt{\sum_{k=1}^{m} w_k (x_{ik} - x_{jk})^2}$,

3) расстояние *city-block*: $d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$,

4) расстояние *Минковского*: $d_{ij} = \left(\sum_{k=1}^{m} |x_{ik} - x_{jk}|^p \right)^{1/p}$,

5) расстояние *Махаланобиса*:

$d_{ij} = (X_i - X_j)' S_*^{-1} (X_i - X_j)$,

где d_{ij} — расстояние между i-м и j-м объектами;

x_{il}, x_{jl} — значения l-й переменной соответственно у i-го и j-го объектов;

X_i, X_j — векторы значений переменных у i-го и j-го объектов;

S_* — общая ковариационная матрица;

w_k — вес, приписываемый k-й переменной.

Пример 9.1. Определим сходство между предприятиями, если каждое из них характеризуется тремя признаками: X_1 — производство продукции, млрд. руб., X_2 — стоимость основных производственных фондов, млрд. руб.; X_3 — фонд заработной платы промышленно-производственного персонала, млрд. руб. (табл. 9.1, 9.2).

Таблица 9.1. **Матрица исходных данных**

№ п/п	X_1	X_2	X_3
1	32,5	40,3	3,5
2	38,4	46,8	4,3
3	16,7	25,7	2,0
4	42,3	44,0	4,5

Таблица 9.2. **Матрица евклидовых расстояний**

№ п/п	1	2	3	4
1	0	8,81	21,55	10,36
2		0	30,35	30,48
3			0	31,57
4				0

Оценка сходства между объектами сильно зависит от абсолютного значения признака и от степени его вариации в совокупности. Чтобы устранить подобное влияние на процедуру классификации, можно значения исходных переменных нормировать одним из следующих способов:

1. $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$, 2. $z_{ij} = \dfrac{x_{ij}}{x_{\max j}}$, 3. $z_{ij} = \dfrac{x_{ij}}{\bar{x}_j}$, 4. $z_{ij} = \dfrac{x_{ij}}{x_{\min j}}$.

Продемонстрируем на нашем примере, как скажется нормирование исходных переменных на мерах сходства между объектами. Заменим x_{ij} новыми значениями z_{ij}, полученными по формуле: $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$, и построим матрицу стандартизованных значений признаков и новую матрицу расстояний (табл. 9.3, 9.4).

В первой матрице расстояний (табл. 9.2) самыми «близкими» были объекты n_1 и n_2 (d_{12} = 8,81), а самыми «дальними» — объекты n_3 и n_4 (d_{34}= 31,57). После нормирования значений исходных переменных самыми «близкими» стали объекты n_2 и n_4 (d_{24} = 0,56), а самыми «дальними» — объекты n_2 и n_3 (d_{23} = 13,2) (табл. 9.4).

Таблица 9.3. **Матрица стандартизованных значений признаков**

№ п/п	Z_1	Z_2	Z_3
1	0,00205	0,13530	−0,10204
2	0,60718	0,93481	0,71429
3	−1,61846	−1,66052	−1,63215
4	1,00718	0,59041	0,.91837

Таблица 9.4. **Матрица расстояний**

№ п/п	1	2	3	4
1	0	1,29	2,86	1,50
2		0	13,20	0,56
3			0	4,30
4				0

В качестве меры сходства отдельных переменных используются парные коэффициенты корреляции Пирсона. Если исходные переменные являются альтернативными признаками, т.е. принимают только два значения, то в качестве меры сходства можно использовать коэффициенты ассоциативности.

Вопрос о придании переменным соответствующих весов должен решаться после проведения исследователем тщательного анализа изучаемой совокупности и социально-экономической сущности классифицирующих переменных. Веса задаются пропорционально степени важности переменных. Например, если для классификации предприятий используются переменные: X_1 — прибыль предприятия, X_2 — выработка продукции на одного работающего, X_3 — среднегодовая стоимость основных производственных фондов, то можно переменным задать веса пропорционально их степени важности для эффективности работы предприятия:

$w_{x_1} = 0{,}6$; $w_{x_2} = 0{,}3$; $w_{x_3} = 0{,}1$.

Тогда евклидово расстояние будет определяться по формуле:

$$d_{ij} = \sqrt{0{,}6\cdot(x_{i1} - \bar{x}_{j1})^2 + 0{,}3\cdot(x_{i2} - x_{j2})^2 + 0{,}1\cdot(x_{i3} - x_{j3})^2}.$$

Выбор меры расстояния и весов для классифицирующих переменных — очень важный этап кластерного анализа, так как от этих процедур зависят состав и количество формируемых кластеров, а также степень сходства объектов внутри кластеров.

Если алгоритм кластеризации основан на измерении сходства между переменными, то в качестве мер сходства могут быть использованы:

- линейные коэффициенты корреляции;
- коэффициенты ранговой корреляции;
- коэффициенты контингенции и т. д.

В зависимости от типов исходных переменных выбирается один из видов показателей, характеризующих близость между ними.

9.3. Иерархический кластерный анализ

Из всех методов кластерного анализа, перечисленных в § 9.1, самыми распространенными являются иерархические агломеративные методы. Сущность этих методов заключается в том, что на первом шаге каждый объект выборки рассматривается как отдельный кластер. Процесс объединения кластеров происходит последовательно: на основании матрицы расстояний или матрицы сходства объединяются наиболее близкие объекты. Если матрица сходства первоначально имеет размерность $m \times m$, то полностью процесс кластеризации завершается за $m-1$ шагов, в итоге все объекты будут объединены в один кластер. Последовательность объединения легко поддается геометрической интерпретации и может быть представлена в виде графа-дерева (дендрограммы). На дендрограмме указываются номера объединяемых объектов и расстояние (или иная мера сходства), при котором произошло объединение (рис. 9.1).

Дендрограмма на рисунке показывает, что в данном случае на первом шаге были объединены в один кластер объекты n_2 и n_3. Расстояние между ними 0,15. На втором шаге к ним присоединился объект n_1. Расстояние от первого объекта до кластера, содержащего объекты n_2 и n_3, было 0,3 и т. д.

Множество методов иерархического кластерного анализа различается не только используемыми мерами сходства (различия), но и алгоритмами классификации. Из них наиболее распространены метод одиночной связи, метод полных связей, метод средней связи, метод Уорда.

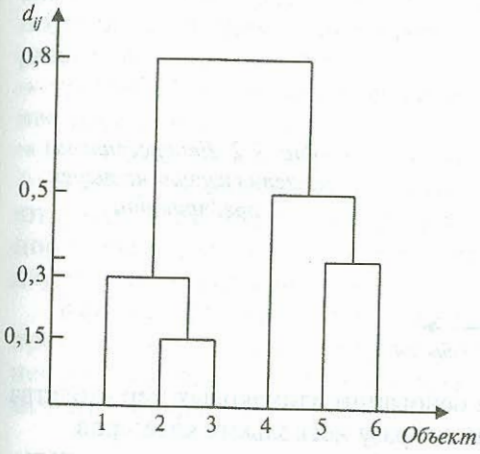


Рис. 9.1. Пример дендрограммы иерархического агломеративного кластерного анализа

Метод одиночной связи. Алгоритм образования кластеров следующий: на основании матрицы сходства (различия) определяются два наиболее схожих или близких объекта, они и образуют первый кластер. На следующем шаге выбирается объект, который будет включен в этот кластер. Таким объектом будет тот, который имеет наибольшее сходство хотя бы с одним из объектов, уже включенных в кластер. Например, имеется матрица евклидовых расстояний между объектами:

№ п/п	1	2	3	4
1	0	2,06	4,03	2,50
2		0	2,24	4,12
3			0	6,32
4				0

В первый кластер будут включены первый и второй объекты, так как расстояние между ними минимальное ($d_{12} = 2{,}06$). На следующем шаге к этому кластеру будет подключен третий объект, так как расстояние $d_{23} = \min\{d_{13}, d_{23}, d_{14}, d_{24}\}$. На последнем шаге в кластер будет включен четвертый объект. Графически это будет выглядеть следующим образом (рис. 9.2):

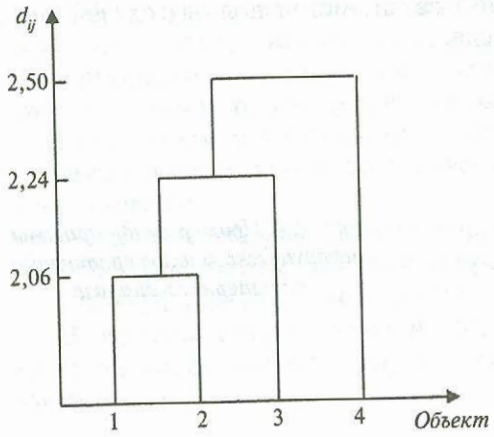
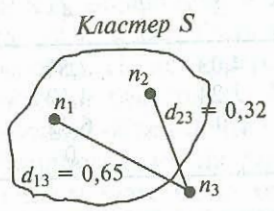


Рис. 9.2. Дендрограмма кластеризации четырех предприятий

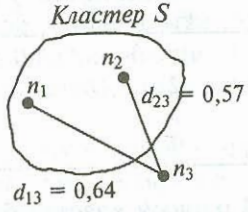
При совпадении данных на основании одинаковых мер сходства (различия) будет идти образование сразу нескольких кластеров.

К достоинствам этого метода следует отнести нечувствительность алгоритма к преобразованиям исходных переменных и его простоту. Недостатками являются необходимость постоянного хранения матрицы сходства и невозможность определения по результатам кластеризации, сколько же кластеров можно образовать в исследуемой совокупности объектов.

Метод полных связей. Включение нового объекта в кластер происходит только в том случае, если расстояние между объектами не меньше некоторого заданного уровня. На рис. 9.3 изображены два случая:



а) если задано предельное расстояние 0,3, то третий объект не будет включен в кластер S, так как $d_{13} > 0,3$ и $d_{23} > 0,3$



б) если задано предельное расстояние 0,7, то третий объект будет включен в кластер S, так как $d_{13} < 0,7$ и $d_{23} < 0,7$

Рис. 9.3. Определение состава кластера при различных уровнях сходства наблюдаемых объектов

Метод средней связи. Для решения вопроса о включении нового объекта в уже существующий кластер вычисляется среднее значение меры сходства, которое затем сравнивается с заданным пороговым уровнем. Для примера на рис. 9.3 в случае *а*) среднее расстояние будет равно $(d_{13} + d_{23})/2 = (0,65 + 0,32)/2 = 0,485 > 0,3$ — третий объект не будет включен в кластер S; в случае *б*) среднее расстояние будет равно $(d_{13} + d_{23}) / 2 = (0,64 + +0,57) / 2 = = 0,605 < 0,7$, значит третий объект будет включен в кластер S.

Если речь идет об объединении двух кластеров, то вычисляют расстояние между их центрами и сравнивают ее с заданным пороговым значением. Рассмотрим геометрический пример с двумя кластерами (рис. 9.4).

Каждый кластер содержит по три объекта. Чтобы решить вопрос об объединении этих двух кластеров, нужно определить их центры тяжести и расстояние между ними. Звездочками (*) отмечены центры тяжести:

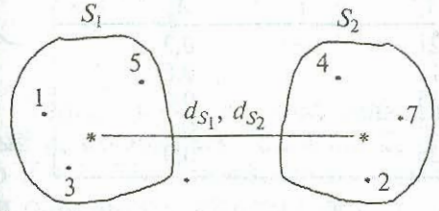


Рис. 9.4. Объединение двух кластеров по методу средней связи

Если расстояние между центрами ($d_{S_1;S_2}$) будет меньше заданного уровня, то кластеры S_1 и S_2 будут объединены в один.

Метод Уорда. Данный метод предполагает, что на первом шаге каждый кластер состоит из одного объекта. Первоначально объединяются два ближайших кластера. Для них определяются средние значения каждого признака и рассчитывается сумма квадратов отклонений V_k:

$$V_k = \sum_{i=1}^{n_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{jk})^2 ,$$

где k — номер кластера,

i — номер объекта,

j — номер признака,

p — количество признаков, характеризующих каждый объект,

n_k — количество объектов в k-м кластере.

В дальнейшем на каждом шаге работы алгоритма объединяются те объекты или кластеры, которые дают наименьшее приращение величины V_k. Метод Уорда приводит к образованию кластеров приблизительно равных размеров с минимальной внутрикластерной вариацией. В итоге все объекты оказываются объединенными в один кластер.

Рассмотрим пример работы алгоритма метода Уорда. Пусть имеются пять объектов, каждый из которых характеризуется двумя признаками (табл. 9.5). Одним из недостатков метода является необходимость перебора всех возможных вариантов включения новых объектов в кластер. При больших объемах исходной совокупности это приводит к значительным затратам машинного времени и требует больших объемов памяти для вычислений.

Таблица 9.5. **Матрица исходных данных по пяти наблюдаемым объектам**

Номер объекта	X_1	X_2
1	21	0,3
2	18	0,8
3	15	0,6
4	13	0,7
5	25	0,9

Рассчитаем на основе данных табл. 9.5 матрицу расстояний между объектами, используя евклидову метрику:

$$\Delta = \begin{pmatrix} 0 & 3{,}0414 & 6{,}0075 & 8{,}0099 & 4{,}0447 \\ & 0 & 3{,}0066 & 5{,}0010 & 7{,}0007 \\ & & 0 & 2{,}0025 & 10{,}0045 \\ & & & 0 & 12{,}0017 \\ & & & & 0 \end{pmatrix}$$

Самые близкие объекты n_3 и n_4 объединяются в один кластер. Для нового кластера определяем сумму квадратов отклонений V_k.

В данном случае присвоим этому кластеру номер S_3, т.е. $k = 3$, и рассчитаем величину V_k:

$$V_k = \sum_{i=1}^{2}\sum_{j=1}^{2}(x_{ij} - \bar{x}_{j3})^2,$$

где $\bar{x}_{j3}$ — среднее значение j-го признака в кластере S_3.

$\bar{x}_{13} = 14$, $\bar{x}_{23} = 0{,}65$,

$$V_3 = \left[(15-14)^2 + (0{,}6-0{,}65)^2\right] + \left[(13-14)^2 + (0{,}7-0{,}65)^2\right] = 2{,}005.$$

Теперь нужно решить вопрос о том, какой новый объект может быть на следующем шаге присоединен к третьему кластеру или какие кластеры можно объединить.

Если объединить первый и второй объекты, тогда $V = 4{,}625$; для пары n_1 и n_5 $V = 8{,}0081$; для пары n_2 и n_5 $V = 12{,}25$.

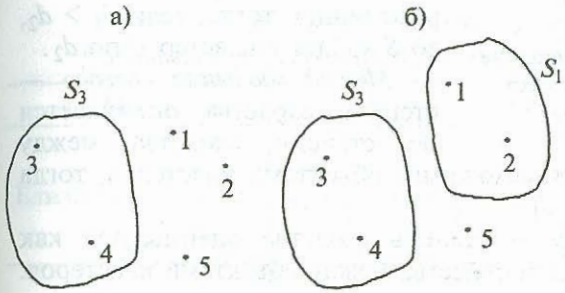


*Рис. 9.5. **Кластеризация наблюдаемых объектов***

Теперь пробуем присоединять к кластеру S_3 поочередно каждый из оставшихся объектов. Если объединить кластеры S_1 и S_3, то $V = 34{,}75$; если объединить кластеры S_2 и S_3, то $V = 12{,}69$; если объединить кластеры S_5 и S_3, то $V = 82{,}71$. Мы видим, что минимальное значение сумма квадратов отклонений принимает в случае, если будут объединены объекты n_1 и n_2. После этого объединения получаем ситуацию, изображенную на рис. 9.5 б).

Алгоритм иерархического кластерного анализа можно представить в виде последовательности процедур:

Шаг 1. Значения исходных переменных нормируются одним из способов, указанных в 9.2.

Шаг 2. Рассчитывается матрица расстояний или матрица мер сходства.

Шаг 3. Находится пара самых близких кластеров. По выбранному алгоритму объединяются эти два кластера. Новому кластеру присваивается меньший из номеров объединяемых кластеров.

Шаг 4. Процедуры 2, 3 и 4 повторяются до тех пор, пока все объекты не будут объединены в один кластер или до достижения заданного «порога» сходства.

Мера сходства для объединения двух кластеров в п. 3 определяется четырьмя методами:

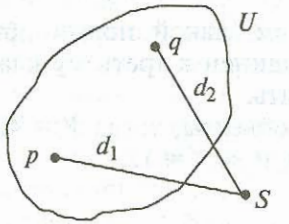


*Рис. 9.6. **Определение расстояния между кластерами методом «ближайшего соседа»***

Метод «ближайшего соседа» — степень сходства оценивается по степени сходства между наиболее схожими (ближайшими) объектами этих кластеров (рис. 9.6).

Пусть d_1 и d_2 — евклидовы расстояния, тогда, если $d_1 > d_2$, то S войдет в кластер U по d_2.

Метод «дальнего соседа» — степень сходства оценивается по степени сходства между наиболее отдаленными (несхожими) объектами кластеров, тогда S войдет в кластер U по d_1.

Метод средней связи — степень сходства оценивается как средняя величина степеней сходства между объектами кластеров. В этом случае S войдет в кластер U по $d = \frac{1}{2}(d_1 + d_2)$.

Метод медианной связи — расстояние между любым кластером S и новым кластером, который получился в результате объединения кластеров p и q, определяется как расстояние от центра кластера S до середины отрезка, соединяющего центры кластеров p и q (рис.9.7).

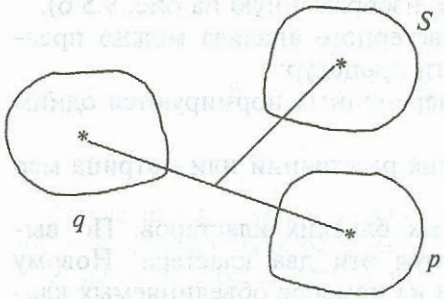


*Рис. 9.7. **Определение расстояний между кластерами методом медианной связи***

Если процесс объединения в иерархическом кластерном анализе проводится вручную, то оценка сходства может быть дана либо визуально по матрице сходств (расстояний) D, либо на основании значений d_{US}:

$$d_{US} = \alpha_p d_{ps} + \alpha_q d_{qs} + \beta d_{pq} + \gamma \left| d_{ps} - d_{qs} \right|,$$

где d_{ps}, d_{qs}, d_{pq} — расстояния между соответствующими кластерами;

$\alpha_p, \alpha_q, \beta, \gamma$ — параметры, учитывающие особенности конкретного алгоритма кластеризации данных; их значения приведены в табл. 9.6.

В формулах табл. 9.6 n_p, n_q — количество объектов в соответствующих кластерах.

Таблица 9.6. **Константные значения параметров, учитывающие особенности метода кластеризации данных**

Метод	*Параметры*
Ближайшего соседа	$\alpha_p = \frac{1}{2}$, $\alpha_q = \frac{1}{2}$, $\beta = 0$, $\begin{cases} \gamma = -\frac{1}{2} & \text{— расстояние} \\ \gamma = \frac{1}{2} & \text{— подобие} \end{cases}$
Дальнего соседа	$\alpha_p = \alpha_q = \frac{1}{2}$, $\beta = 0$, $\begin{cases} \gamma = -\frac{1}{2} & \text{— расстояние} \\ \gamma = -\frac{1}{2} & \text{— подобие} \end{cases}$
Средней связи	$\alpha_p = \frac{n_p}{n_p + n_q}$, $\alpha_q = \frac{n_q}{n_p + n_q}$, $\beta = \gamma = 0$
Центроидный	$\alpha_p = \frac{n_p}{n_p + n_q}$, $\alpha_q = \frac{n_q}{n_p + n_q}$, $\beta = -\frac{n_p}{n_p + n_q} \cdot \frac{n_q}{n_p + n_q}$.
Медианной связи	$\alpha_p = \alpha_q = \frac{1}{2}$, $\beta = -0{,}25$, $\gamma = 0$

Использование различных алгоритмов объединения в иерархических агломеративных методах приводит к различным кластерным структурам и сильно влияет на качество проведения кластеризации. Поэтому алгоритм должен выбираться с учетом имеющихся сведений о существующей структуре совокупности наблюдаемых объектов или с учетом требований оптимизации математических критериев.

Если алгоритм иерархического кластерного анализа реализуется в программе для ЭВМ, то оценка сходства между кластерами дается только на основании значений d_{US}.

Рассмотрим алгоритм иерархического кластерного анализа на примере.

Пример 9.2. Необходимо провести классификацию пяти предприятий, каждое из которых характеризуется тремя переменными: X_1 — среднегодовая стоимость основных производственных фондов, млрд. руб.; X_2 — материальные затраты на 1 руб. произведенной продукции, коп.; X_3 — объем произведенной продукции, млрд. руб. Значения переменных приведены в табл. 9.7.

Таблица 9.7. **Исходные данные**

Номер предприятия	X_1	X_2	X_3
1	120,0	94,0	164,0
2	85,0	75,2	92,0
3	145,0	81,0	120,0
4	78,0	76,8	86,0
5	70,0	75,9	104,0
Среднее значение $(\bar{x}_j)$	99,6	80,6	113,2
Среднее квадратическое отклонение (σ)	28,4	10,9	27,9

Перед тем как вычислять матрицу расстояний, нормируем исходные данные по формуле:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}.$$

Матрица значений нормированных переменных будет:

$$Z = \begin{pmatrix} 0{,}718 & 1{,}229 & 1{,}821 \\ -0{,}514 & -2{,}238 & -0{,}760 \\ 1{,}598 & 0{,}037 & 0{,}244 \\ -0{,}760 & -0{,}349 & -0{,}975 \\ -1{,}042 & -0{,}431 & -0{,}330 \end{pmatrix}.$$

Классификацию проведем при помощи иерархического агломеративного метода. Для построения матрицы расстояний воспользуемся евклидовым расстоянием. Тогда, например, расстояние между первым и вторым объектами:

$$d_{12} = \left\{[0{,}718-(-0{,}514)]^2 + [1{,}229-(-2{,}238)]^2 + [1{,}821-(-0{,}760)]^2\right\}^{1/2} = 4{,}49.$$

Первоначальная матрица расстояний D_0 характеризует расстояния между отдельными объектами, каждый из которых на первом шаге является отдельным кластером:

$$D_0 = \begin{pmatrix} 0 & 4{,}49 & 2{,}16 & 3{,}53 & 3{,}24 \\ & 0 & 3{,}26 & 1{,}92 & 1{,}93 \\ & & 0 & 2{,}68 & 2{,}74 \\ & & & 0 & 0{,}71 \\ & & & & 0 \end{pmatrix}.$$

Как видно по элементам матрицы D_0, наиболее близкими являются объекты n_4 и n_5 ($d_{45} = 0{,}71$). Объединим их в один кластер и присвоим ему номер S_4. Пересчитаем расстояния всех оставшихся объектов (кластеров) до кластера S_4, получим новую матрицу расстояний D_1:

$$D_1 = \begin{pmatrix} 0 & 4{,}49 & 2{,}16 & 3{,}53 \\ & 0 & 3{,}26 & 1{,}92 \\ & & 0 & 2{,}74 \\ & & & 0 \end{pmatrix}.$$

В матрице D_1 расстояния между кластерами определены по алгоритму «дальнего соседа». Тогда расстояние между объектом n_1 и кластером S_4 равно:

$$d_{S_1,S_4} = \max\{d_{14}, d_{15}\} = \max\{3{,}53; 3{,}24\} = 3{,}53 \text{ и т. д.}$$

В матрице D_1 опять находим самые близкие кластеры. Это будут S_2 и S_4, поскольку $d_{24} = 1{,}93$. Следовательно, на этом шаге объединяем кластеры S_2 и S_4; получим новый кластер, содержащий объекты n_2, n_4, n_5. Присвоим ему номер S_2. Теперь имеем три кластера $S_1\{1\}, S_2\{2,4,5\}, S_3\{3\}$. Пересчитываем расстояния d_{12} и d_{23}, получаем матрицу D_2:

$$d_{12} = \max\{d_{1,2}, d_{1,IV}\} = \max\{4{,}49; 3{,}53\} = 4{,}49.$$

$$d_{23} = \max\{d_{2,3}, d_{3,IV}\} = \max\{3{,}26; 2{,}74\} = 3{,}26.$$

$$D_2 = \begin{pmatrix} 0 & 4{,}49 & 2{,}16 \\ & 0 & 3{,}26 \\ & & 0 \end{pmatrix}.$$

Судя по матрице D_2, на следующем шаге объединяем кластеры S_1 и S_3 ($d_{13} = 2{,}16$) в один кластер и присвоим ему номер S_1. Теперь имеем только два кластера:

$$\left.\begin{array}{l} S_1 \text{ кластер (объекты } n_1, n_3) \\ S_2 \text{ кластер (объекты } n_2, n_4, n_5) \end{array}\right\} \quad d_{12} = \max\{d_{1,2}, d_{3,2}\} = \max\{4{,}49; 3{,}26\} = 4{,}49.$$

$$D_3 = \begin{pmatrix} 0 & 4{,}49 \\ & 0 \end{pmatrix}.$$

И наконец, на последнем шаге объединяем кластеры S_1 и S_2 на расстоянии 4,49. Представим результаты классификации в виде дендрограммы (рис. 9.8). Дендрограмма свидетельствует о том, что кластер S_2 более однороден по составу входящих объектов, так как в нем объединение происходило при меньших расстояниях, чем в кластере S_1.

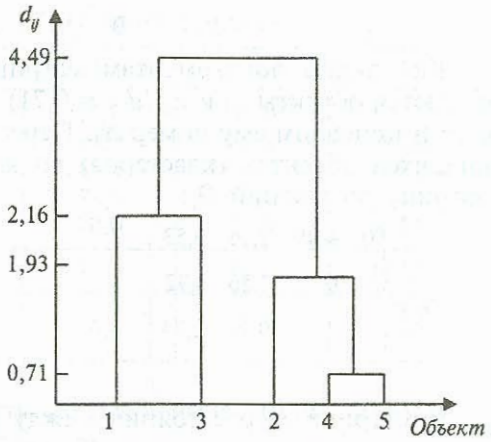


Рис. 9.8. Дендрограмма кластеризации пяти объектов

Кроме рассмотренных агломеративных методов иерархического кластерного анализа существуют методы, противоположные им по логическому построению процедур классификации. Они называются *иерархические дивизимные методы*. Основной исходной посылкой дивизимных методов является то, что первоначально все объекты принадлежат одному кластеру (классу). В процессе классификации по определенным правилам постепенно от этого кластера отделяются группы схожих

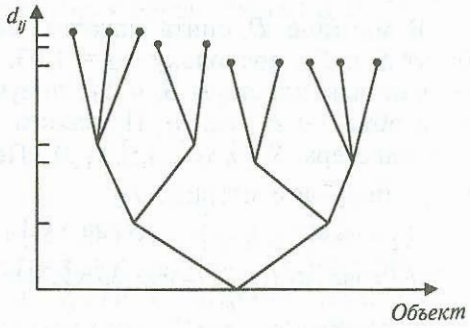


Рис. 9.9. Дендрограмма иерархического дивизимного алгоритма

между собой объектов. Таким образом, на каждом шаге количество кластеров возрастает, а мера расстояния между кластерами уменьшается. Дендрограмма для дивизимных иерархических методов изображена на рис. 9.9.

Пример 9.3. Пусть дана следующая матрица расстояний между объектами:

$$D = \begin{pmatrix} 0 & 4{,}49 & 2{,}16 & 3{,}53 & 3{,}24 \\ & 0 & 3{,}26 & 1{,}92 & 1{,}93 \\ & & 0 & 2{,}68 & 2{,}74 \\ & & & 0 & 0{,}71 \\ & & & & 0 \end{pmatrix}.$$

Проведем классификацию по дивизимному алгоритму.

Наиболее удаленными являются объекты n_1 и n_2; оценим расстояния оставшихся объектов до первого и второго объектов:

$d_{31} < d_{32}$ — объект n_3 ближе к n_1,
$d_{41} > d_{42}$ — объект n_4 ближе к n_2,
$d_{51} > d_{52}$ — объект n_5 ближе к n_2.

Итак, получили теперь два кластера: $S_1\{1,3\}$ и $S_2\{2,4,5\}$. В каждом из них анализируем расстояния между объектами, и на очередном шаге происходит разделение того кластера, где достигается максимум расстояния между объектами.

$d_{13} = 2{,}16$, $d_{25} = 1{,}93$, $d_{24} = 1{,}92$, $d_{45} = 0{,}71$.

Наибольшее расстояние $d_{13} = 2{,}16$, следовательно, объекты n_1 и n_3 выделяем в отдельные кластеры. В кластере $S_2\{2,4,5\}$ ищем максимальное расстояние $\max\{d_{24}, d_{25}, d_{45}\} = 1{,}93$. На следующем шаге из этого кластера выделяем объект n_2, и, наконец, на последнем шаге разделяем кластер $S_4\{4,5\}$ на два кластера на расстоянии 0,71.

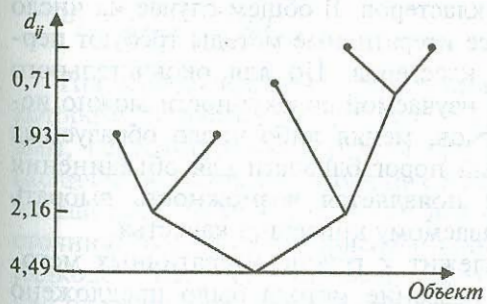


Рис. 9.10. Дендрограмма кластеризации объектов по иерархическому дивизимному методу

Дендрограмма процесса классификации по иерархическому дивизимному методу представлена на рис. 9.10.

Как видно из этого примера, дивизимный алгоритм не требует пересчета матрицы расстояний на каждом шаге классификации в отличие от агломеративных методов, что способствует снижению трудоемкости расчетов.

9.4. Метод *k*-средних

Наряду с иерархическими методами классификации, рассмотренными в § 9.3, существует многочисленная группа так называемых *итеративных методов* кластерного анализа. Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т. д.). Итеративные методы в большей степени, чем иерархические, требуют от пользователя интуиции при выборе типа классификационных процедур и задания начальных условий разбиения, так как большинство этих методов очень чувствительны к изменению задаваемых параметров. Например, выбранное случайным образом число кластеров может не только сильно увеличить трудоемкость процесса классификации, но и привести к образованию «размытых» или мало наполняемых кластеров. Поэтому целесообразно сначала провести классификацию по одному из иерархических методов или на основании экспертных оценок, а затем уже подбирать начальное разбиение и статистический критерий для работы итерационного алгоритма. Как и в иерархическом кластерном анализе, в итерационных методах существует проблема определения числа кластеров. В общем случае их число может быть неизвестно. Не все итеративные методы требуют первоначального задания числа кластеров. Но для окончательного решения вопроса о структуре изучаемой совокупности можно испробовать несколько алгоритмов, меняя либо число образуемых кластеров, либо установленный порог близости для объединения объектов в кластеры. Тогда появляется возможность выбрать наилучшее разбиение по задаваемому критерию качества.

Метод *k*-средних принадлежит к группе итеративных методов эталонного типа. Само название метода было предложено Дж. Мак-Куином в 1967 г. [3].

В отличие от иерархических процедур метод *k*-средних не требует вычисления и хранения матрицы расстояний или сходств

между объектами. Алгоритм этого метода предполагает использование только исходных значений переменных. Для начала процедуры классификации должны быть заданы k случайно выбранных объектов, которые будут служить эталонами, т.е. центрами кластеров. Считается, что алгоритмы эталонного типа удобные и быстродействующие. В этом случае важную роль играет выбор начальных условий, которые влияют на длительность процесса классификации и на его результаты.

Метод *k*-средних удобен для обработки больших статистических совокупностей. В исследованиях Хартигана, Болла, В.Н. Елкиной, Н.Г. Загоруйко предлагаются различные модификации этого метода.

Рассмотрим математическое описание алгоритма метода *k*-средних (Мак-Куина), данное в [3, с.115].

Пусть имеется n наблюдений, каждое из которых характеризуется p признаками $X_1, X_2, \ldots, X_p$. Эти наблюдения необходимо разбить на k кластеров. Для начала из n точек исследуемой совокупности отбираются случайным образом или задаются исследователем исходя из каких-либо априорных соображений k точек (объектов). Эти точки принимаются за эталоны. Каждому эталону присваивается порядковый номер, который одновременно является и номером кластера. На первом шаге из оставшихся $(n - k)$ объектов извлекается точка X_i с координатами $(x_{i1}, x_{i2}, \ldots, x_{ip})$ и проверяется, к какому из эталонов (центров) она находится ближе всего. Для этого используется одна из метрик, например, евклидово расстояние:

$$d_{il} = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{lj}\right)^2} .$$

Проверяемый объект присоединяется к тому центру (эталону), которому соответствует $\min d_{il}$ $(l = 1, \ldots, k)$. Эталон заменяется новым, пересчитанным с учетом присоединенной точки, и вес его (количество объектов, входящих в данный кластер) увеличивается на единицу. Если встречаются два или более минимальных расстояния, то i-й объект присоединяют к центру с наименьшим порядковым номером. На следующем шаге выбираем точку X_{i+1} и для нее повторяются все процедуры. Таким образом, через $(n - k)$ шагов все точки (объекты) совокупности окажутся отнесенными к одному из k кластеров, но на этом процесс разбиения не заканчивается. Для того чтобы добиться устойчивости разбиения по тому

же правилу, все точки X_1, X_2, ..., X_n опять подсоединяются к полученным кластерам, при этом веса продолжают накапливаться. Новое разбиение сравнивается с предыдущим. Если они совпадают, то работа алгоритма завершается. В противном случае цикл повторяется. Окончательное разбиение имеет центры тяжести, которые не совпадают с эталонами, их можно обозначить C_1, C_2,..., C_k. При этом каждая точка X_i (i = 1, 2, ..., n) будет относиться к такому кластеру (классу) l, для которого

$$d(x_j, c_l) = \min_{1 \le j \le R} d(x_j, C_j).$$

Возможны две модификации метода k-средних. Первая предполагает пересчет центра тяжести кластера после каждого изменения его состава, а вторая — лишь после того, как будет завершен просмотр всех данных. В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется. Рассмотрим работу итеративного алгоритма метода k-средних на примере.

Пример 9.4. Пусть имеются шесть объектов, которые необходимо разбить на три класса (кластера) при помощи метода k-средних. Каждый из объектов описывается тремя переменными X_1, X_2, X_3. Исходные значения этих переменных представлены в табл. 9.9.

Таблица 9.9. **Исходные данные**

Номер объекта	X_1	X_2	X_3
1	0,10	10	5,0
2	0,80	14	2,0
3	0,40	12	3,0
4	0,18	11	4,0
5	0,25	13	3,2
6	0,67	15	2,4

В качестве эталонов возьмем первые три объекта (k = 3). Согласно выбранному правилу классификации запишем исходные значения эталонов и весов:

$$\left.\begin{aligned} E_1^0 &= X_1 = (0{,}10 \quad 10 \quad 50), \quad w_1^0 = 1 \\ E_2^0 &= X_2 = (0{,}80 \quad 14 \quad 2{,}0), \quad w_2^0 = 1 \\ E_3^0 &= X_3 = (0{,}40 \quad 12 \quad 3{,}0), \quad w_3^0 = 1 \end{aligned}\right\} \text{— нулевая итерация.}$$

На *первом шаге* берем четвертый объект и определяем его расстояние до каждого из эталонов по евклидовой метрике:

$$d_{41} = \sqrt{(0{,}18-0{,}10)^2 + (11-10)^2 + (4{,}0-5{,}0)^2} = 1{,}416,$$

$$d_{42} = \sqrt{(0{,}18-0{,}80)^2 + (11-14)^2 + (4{,}0-5{,}0)^2} = 3{,}222,$$

$$d_{43} = \sqrt{(0{,}18-0{,}40)^2 + (11-12)^2 + (4{,}0-3{,}0)^2} = 1{,}431.$$

Следовательно, рассматриваемый объект должен быть присоединен к первому эталону и первый эталон будет пересчитан, а второй и третий не меняются:

$$E_1^1 = \frac{w_1^0 \cdot E_1^0 + X_4}{w_1^0 + 1},$$

$W_1^1 = W_1^0 + 1 = 2$, $E_2^1 = E_2^0$, $E_3^1 = E_3^0$, $W_2^1 = W_2^0$, $W_3^1 = W_3^0$,

где X_4 — вектор значений переменных для четвертого объекта,

E_1 — пересчитанное значение эталона;

$$E_1^1 = \left(\frac{0{,}10+0{,}18}{2}; \frac{10+11}{2}; \frac{5+4}{2}\right) = (0{,}14;\ 10{,}5;\ 4{,}5).$$

На *втором шаге* проверяем, к какому эталону ближе всего находится пятый объект:

$$d_{51} = \sqrt{(0{,}25-0{,}14)^2 + (13-10{,}5)^2 + (3{,}2-4{,}5)^2} = 2{,}820,$$

$$d_{52} = \sqrt{(0{,}25-0{,}80)^2 + (13-14)^2 + (3{,}2-2{,}0)^2} = 1{,}656,$$

$$d_{53} = \sqrt{(0{,}25-0{,}40)^2 + (13-12)^2 + (3{,}2-3{,}0)^2} = 1{,}031$$

Так как $d_{53} = \min\{d_{5,1}, d_{5,2}, d_{5,3}\}$, следовательно, пятый объект присоединяется к третьему эталону, этот эталон пересчитывается и вес его увеличивается:

$$E_3^2 = \left(\frac{0{,}40+0{,}25}{2}, \frac{13+12}{2}, \frac{3{,}0+3{,}2}{2}\right) = (0{,}325;\ 12{,}5;\ 3{,}1).$$

$W_3^2 = W_3^1 + 1 = 2$; $E_1^2 = E_1^1$; $E_2^2 = E_2^1$; $W_1^2 = W_1^1$; $W_2^2 = W_2^1$.

На *третьем шаге* все рассуждения повторяем для шестого объекта:

$$d_{61} = \sqrt{(0{,}67-0{,}14)^2 + (15-10{,}5)^2 + (2{,}4-4{,}5)^2} = 4{,}994,$$

$$d_{62} = \sqrt{(0{,}67-0{,}80)^2 + (15-14)^2 + (2{,}4-2{,}0)^2} = 1{,}085,$$

$$d_{63} = \sqrt{(0{,}67-0{,}325)^2 + (15-12{,}5)^2 + (2{,}4-3{,}1)^2} = 2{,}619.$$

Пересчитываем второй эталон и его вес:

$$E_2^3 = \left(\frac{0{,}80+0{,}67}{2}, \frac{14+15}{2}, \frac{2{,}0+2{,}4}{2}\right) = (0{,}735;\ 14{,}5;\ 2{,}2),$$

$W_2^3 = W_2^2 + 1 = 2$; $E_1^3 = E_1^2$; $E_3^2 = E_3^1$; $W_1^3 = W_1^2$; $W_3^2 = W_3^1$.

После того как просмотрены все объекты, кроме первых трех, процесс «зацикливается», т.е. по тому же правилу осуществляются просмотр и присоединение к соответствующему эталону каждого из шести объектов. При этом происходит пересчет эталонов и продолжается наращивание их весов. Результаты расчетов, начиная с четвертой итерации, представлены в табл. 9.10.

Итак, на этом процесс завершается, так как последующее разбиение (интерации 16—21) дали такой же результат, как и предыдущее разбиение (итерации 10—15).

Образованы три кластера: S_1 {1}, S_2 {2, 6}, S_3 {3, 4, 5}. Вычисляем центры тяжести полученных кластеров, причем в общем случае эти центры не совпадают с эталонами:

C_1 = (0,10; 10; 5,0) — центр 1 кластера,
C_2 = (0,735; 14,5; 2,2) — центр 2 кластера,
C_3 = (0,277; 12,00; 3,4) — центр 3 кластера.

После этого строится окончательное разбиение: каждая многомерная точка относится к тому кластеру, центр которого ближе всех к этой точке.

Для нашего примера определяем поочередно расстояния всех точек (X_1, X_2, X_3, X_4, X_5, X_6) до центров трех кластеров (табл. 9.11).

Как видно из табл. 9.11, подтверждается полученное разбиение на три кластера: S_1 {1}, S_2 {2,6}, S_3 {3,4,5}. На этом алгоритм завершается.

Таблица 9.10. **Параметрические данные кластеризации объектов методом *k*-средних**

Номер итерации	Эталоны и их веса		
	1	2	3
1	2	3	4
4	$E_1^4 = (0{,}127;\ 7;\ 4{,}7)$ $W_1^4 = 3$	$E_2^4 = (0{,}735;\ 14{,}5;\ 2{,}2)$ $W_2^4 = 2$	$E_3^4 = (0,\ 325;\ 12{,}5;\ 3{,}1)$ $W_3^4 = 2$
5	$E_1^5 = (0{,}127;\ 7;\ 4{,}7)$ $W_1^5 = 3$	$E_2^5 = (0{,}757;\ 14{,}33;\ 2{,}133)$ $W_2^5 = 3$	$E_3^5 = (0{,}325;\ 12{,}5;\ 3{,}1)$ $W_3^5 = 2$
6	$E_1^6 = (0{,}127;\ 7;\ 4{,}7)$ $W_1^6 = 3$	$E_2^6 = (0{,}757;\ 14{,}33;\ 2{,}133)$ $W_2^6 = 3$	$E_3^6 = (0{,}35;\ 12{,}33;\ 3{,}07)$ $W_3^6 = 3$
7	$E_1^7 = (0{,}140;\ 8;\ 4{,}3)$ $W_1^7 = 4$	$E_2^7 = (0{,}757;\ 14{,}33;\ 2{,}133)$ $W_2^7 = 3$	$E_3^7 = (0{,}35;\ 12{,}33;\ 3{,}07)$ $W_3^7 = 3$
8	$E_1^8 = (0{,}14;\ 8;\ 4{,}3)$ $W_1^8 = 4$	$E_2^8 = (0{,}735;\ 14{,}5;\ 2{,}2)$ $W_2^8 = 4$	$E_3^8 = (0{,}35;\ 12{,}33;\ 3{,}07)$ $W_3^8 = 3$
9	$E_1^9 = (0{,}132;\ 8{,}4;\ 4{,}44)$ $W_1^9 = 5$	$E_2^9 = (0{,}735;\ 14{,}5;\ 2{,}2)$ $W_2^9 = 4$	$E_3^9 = (0{,}35;\ 12{,}33;\ 3{,}07)$ $W_3^9 = 3$
10	$E_1^{10} = (0{,}126;\ 8{,}72;\ 4{,}55)$ $W_1^{10} = 6$	$E_2^{10} = (0{,}735;\ 14{,}5;\ 2{,}2)$ $W_2^{10} = 4$	$E_3^{10} = 0{,}35;\ 12{,}33;\ 3{,}07)$ $W_3^{10} = 3$
11	$E_1^{11} = (0{,}132;\ 8{,}4;\ 4{,}44)$ $W_1^{11} = 6$	$E_2^{11} = (0{,}748;\ 14{,}4;\ 2{,}16)$ $W_2^{11} = 5$	$E_3^{11} = (0{,}35;\ 12{,}33;\ 3{,}07)$ $W_3^{11} = 3$
12	$E_1^{12} = (0{,}132;\ 8{,}4;\ 4{,}44)$ $W_1^{12} = 6$	$E_2^{12} = (0{,}748;\ 14{,}4;\ 2{,}16)$ $W_2^{12} = 5$	$E_3^{12} = (0{,}36;\ 12{,}25;\ 3{,}05)$ $W_3^{12} = 4$

Продолжение табл. 9.10

1	2	3	4
13	E_1^{13} = (0,132; 8,4; 4,44) W_1^{13} = 6	E_2^{13} = (0,748; 14,4; 2,16) W_2^{13} = 5	E_3^{13} = (0,324; 12; 3,24) W_3^{13} = 5
14	E_1^{14} = (0,132;8,4; 4,44) W_1^{14} = 6	E_2^{14} = (0,748; 14,4; 2,16) W_2^{14} = 5	E_3^{14} = (0,312; 12,17; 3,23) W_3^{14} = 6
15	E_1^{15} = (0,132; 8,4; 4,44) W_1^{15} = 6	E_2^{15} = (0,735; 14,5; 2,2) W_2^{15} = 6	E_3^{15} = (0,312; 12,17; 3,23) W_3^{15} = 6
16	E_1^{16} = (0,122; 8,9; 4,61) W_1^{16} = 7	E_2^{16} = (0,735; 14,5; 2,2) W_2^{16} = 6	E_3^{16} = (0,312; 12,17; 3,23) W_3^{16} = 6
17	E_1^{17} = (0,122; 8,9; 4,61) W_1^{17} = 7	E_2^{17} = (0,744; 14,43; 2,17) W_2^{17} = 7	E_3^{17} = (0,312; 12,17; 3,23) W_3^{17} = 6
18	E_1^{18} = (0,122; 8,9; 4,61) W_1^{18} = 7	E_2^{18} = (0,744; 14,43; 2,17) W_2^{18} = 7	E_3^{18} = (0,324; 12,14; 3,20) W_3^{18} = 7
19	E_1^{19} = (0,122; 8,9; 4,61) W_1^{19} = 7	E_2^{19} = (0,744; 14,43; 2,17) W_2^{19} = 7	E_3^{19} = (0,306; 12,00; 3,3) W_3^{19} = 8
20	E_1^{20} = (0,122; 8,9; 4,61) W_1^{20} = 7	E_2^{20} = (0,744; 14,43; 2,17) W_2^{20} = 7	E_3^{20} = (0,3; 12,11; 3,29) W_3^{20} = 8
21	E_1^{21} = (0,122; 8,9; 4,61) W_1^{21} = 7	E_2^{20} = (0,735; 14,5; 2,2) W_2^{20} = 8	E_3^{21} = (0,3; 12,11; 3,29) W_3^{21} = 9
Номер объекта	1	2, 6	3, 4, 5

Если было проведено несколько разбиений заданной совокупности, то необходимо оценить качество каждого разбиения на основании одного из выбранных критериев (см. § 9.6), чтобы прийти к окончательному решению.

Таблица 9.11. **Расстояния до центров классов**

Центры кластеров	*Объекты*					
	1	2	3	4	5	6
C_1	0	5,049	2,844	1,416	3,502	5,664
C_2	5,338	0,542	2,646	3,939	1,867	0,542
C_3	5,920	2,497	0,418	1,169	1,020	3,187

Рассмотренный метод *k*-средних допускает в качестве исходного разбиения использовать группировку, полученную одним из методов иерархического кластерного анализа. Такой подход можно рекомендовать для сокращения времени обработки в том случае, когда совокупность объектов достаточно велика и пользователь затрудняется указать количество образуемых кластеров.

Вычислительные процедуры большинства итеративных методов классификации сводятся к выполнению следующих шагов:

Шаг 1. Выбор числа кластеров, на которые должна быть разбита совокупность, задание первоначального разбиения объектов и определение центров тяжести кластеров.

Шаг 2. В соответствии с выбранными мерами сходства определение нового состава каждого кластера.

Шаг 3. После полного просмотра всех объектов и распределения их по кластерам осуществляется пересчет центров тяжести кластеров.

Шаг 4. Процедуры 2 и 3 повторяются до тех пор, пока следующая итерация не даст такой же состав кластеров, что и предыдущая.

9.5. Метод поиска сгущений

Одним из итеративных методов классификации, не требующих задания числа кластеров, является *метод поиска сгущений.*

В теории и на практике существует несколько различных модификаций этого метода [61]. Каждая модификация отличается задаваемым начальным состоянием и критериями завершения классификации. Остановимся подробно на одном из алгоритмов поиска сгущений, который получил название «форель». Подробно этот метод описан в работах [38,61]. Суть итеративного алгоритма типа «форель» заключается в применении гиперсферы заданного радиуса, которая перемещается в пространстве классификационных признаков с целью поиска локальных сгущений точек. Рассмотрим схему данного алгоритма в общем виде и на конкретном примере.

Метод поиска сгущений требует вычисления матрицы расстояний (или матрицы мер сходства) между объектами. Затем выбирается объект, который является первоначальным центром первого кластера. Выбор такого объекта может быть произвольным, а может основываться на предварительном анализе точек и их окрестностей. При использовании второго подхода можно значительно сократить число итераций, приводящих к распределению всех точек по кластерам.

Выбранная точка принимается за центр гиперсферы заданного радиуса R. Определяется совокупность точек, попавших внутрь этой сферы, и для них вычисляются координаты центра (вектор средних значений признаков). Далее вновь рассматриваем гиперсферу такого же радиуса, но с новым центром, и для совокупности попавших в нее точек опять рассчитываем вектор средних значений, принимаем его за новый центр сферы и т.д. Когда очередной пересчет координат центра сферы приводит к такому же результату, как и на предыдущем шаге, перемещение сферы прекращается, а точки, попавшие в нее, образуют кластер и из дальнейшего процесса кластеризации исключаются. Для всех оставшихся точек процедуры повторяются, т.е. опять выбирается произвольный объект, который является первоначальным центром сферы радиуса R, и т.д.

В работе [22] доказана сходимость этого алгоритма для любых начальных точек и для совокупностей различного объема.

Таким образом, работа алгоритма завершается за конечное число шагов и все точки оказываются распределенными по кластерам. Число образовавшихся кластеров заранее неизвестно и сильно зависит от выбора радиуса сферы. Некоторые модификации алгоритма позволяют разделить совокупность на заданное

число кластеров путем последовательного изменения радиуса сферы.

Для оценки устойчивости полученного разбиения целесообразно повторить процесс кластеризации несколько раз для различных значений радиуса сферы, изменяя каждый раз радиус на небольшую величину.

Существует несколько способов выбора радиуса сферы. Если d_{lk} — расстояние между l-м и k-м объектами, то в качестве нижней границы радиуса $R_н$ выбирают $R_н = \min\{d(X_l, X_k)\}$, а верхняя граница радиуса $R_в$ может быть определена как $R_в = \max\{d(X_l, X_k)\}$.

Если начинать работу алгоритма с величины $R = \min d(X_l, X_k) + \delta$ и при каждом его повторении изменять δ на небольшую величину, то можно выявить значения радиусов, которые приводят к образованию одного и того же числа кластеров, т.е. к устойчивому разбиению.

Рассмотрим пример группировки объектов по методу поиска сгущений.

Пример 9.5. По семи предприятиям имеются следующие данные о результатах работы за отчетный период:

Таблица 9.12. **Исходные данные**

Номер предприятия	*Выпуск продукции на одного работающего, млрд. руб.*	*Прибыль от реализации продукции, млрд. руб.*
1	51	28
2	63	39
3	48	29
4	39	37
5	30	18
6	58	36
7	61	55

Необходимо выделить первый кластер методом поиска сгущений.

Прежде всего рассчитаем матрицу расстояний между объектами, воспользовавшись евклидовой метрикой, и выберем ради-

ус сферы. Матрица расстояний на первом шаге будет выглядеть следующим образом:

$$D = \begin{pmatrix} 0 & 1{,}628 & 0{,}316 & 1{,}500 & 1{,}562 & 1{,}063 & 2{,}879 \\ & 0 & 1{,}803 & 2{,}408 & 3{,}911 & 0{,}583 & 1{,}612 \\ & & 0 & 1{,}204 & 2{,}109 & 1{,}221 & 2{,}907 \\ & & & 0 & 2{,}102 & 1{,}903 & 2{,}842 \\ & & & & 0 & 3{,}329 & 4{,}827 \\ & & & & & 0 & 1{,}923 \\ & & & & & & 0 \end{pmatrix}.$$

Радиус сферы следует выбирать в пределах от 0,316 до 4,827. В качестве центра сферы возьмем первый объект и радиус сферы $R = 1{,}2$. Тогда в эту сферу попадут все точки (объекты), расстояние которых до объекта n_1 меньше 1,2. Это будут объекты n_3 и n_6. Для этих трех точек определяем координаты центра тяжести $\bar{x} = (5{,}2;\ 3{,}1)$. Принимаем точку с координатами (5,2; 3,1) за новый центр сферы и рассчитываем расстояния от этой точки до всех остальных:

$d_{1,\bar{x}} = 0{,}316;$ $d_{4,\bar{x}} = 1{,}432;$ $d_{7,\bar{x}} = 2{,}563.$

$d_{2,\bar{x}} = 1{,}360;$ $d_{5,\bar{x}} = 2{,}555;$

$d_{3,\bar{x}} = 0{,}447;$ $d_{6,\bar{x}} = 0{,}781;$

Мы видим, что в сферу с центром (5,2; 3,1) опять попали те же объекты: n_1, n_2 и n_6. Следовательно, центр тяжести не меняется, т.е. первый кластер уже создан и в дальнейших процедурах объекты n_1, n_3 и n_6 не участвуют.

Для поиска второго кластера в любую из оставшихся точек помещаем центр сферы и повторяем все процедуры, т.е. перемещаем сферу радиуса $R = 1{,}2$ до тех пор, пока центр кластера не останется неизменным после очередного пересчета.

Кроме рассмотренного выше алгоритма поиска сгущений типа «форель», существует очень схожий с ним алгоритм взаимного поглощения, который также является итерационным и использует идею гиперсферы [31]. Суть его заключается в том, что для каждой многомерной величины X_i определяется свой радиус R_i, например следующим образом:

$$R_i = \max_l d(X_i, X_l) - \delta,$$

где — $d(X_i, X_l)$ — расстояние от i-й до l-й точки;

δ — некоторая выбираемая величина, постоянная для всех точек.

Сферы с радиусами R_i строятся с центрами в точках X_i ($i = 1, \ldots, n$). Области пересечения, содержащие центры этих сфер, называются областью взаимного поглощения. А совокупность центров сфер, попавших в эту область, называется кластером.

Графически для двумерных величин это можно представить так, как показано на рис 9.12.

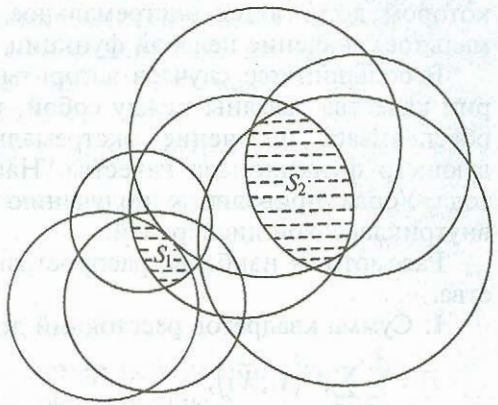


Рис. 9.12. Области взаимного поглощения и два образованных ими кластера

Области взаимного поглощения на рисунке заштрихованы и ясно видны группы точек, попавших в два кластера S_1 и S_2. Изменяя величину δ, можно повторить разбиение несколько раз. В качестве окончательного решения задачи следует выбирать вариант разбиения, сохраняющийся при нескольких значениях δ, как наиболее устойчивый.

9.6. Критерии качества классификации

При использовании различных методов кластерного анализа для одной и той же совокупности могут быть получены различные варианты разбиения. Существенное влияние на характеристики кластерной структуры оказывают: во-первых, набор признаков, по которым осуществляется классификация, во-вторых, тип выбранного алгоритма. Например, иерархические и итеративные методы приводят к образованию различного числа кластеров. При этом сами кластеры различаются и по составу, и по степени близости объектов. Выбор меры сходства также влияет на результат разбиения. Если используются методы с эталонными алгоритмами, например, метод k-средних, то задаваемые начальные условия разбиения в значительной степени определяют конечный результат разбиения.

После завершения процедур классификации необходимо оценить полученные результаты. Для этой цели используется некоторая мера качества классификации, которую принято называть *функционалом* или *критерием качества*. Наилучшим по выбранному функционалу следует считать такое разбиение, при котором достигается экстремальное (минимальное или максимальное) значение целевой функции — функционала качества.

В большинстве случаев алгоритмы классификации и критерии качества связаны между собой, т.е. определенный алгоритм обеспечивает получение экстремального значения соответствующего функционала качества. Например, использование метода Уорда приводит к получению кластеров с минимальной внутриклассовой дисперсией.

Рассмотрим наиболее распространенные функционалы качества.

1. Сумма квадратов расстояний до центров классов:

$$F_1 = \sum_{l=1}^{k} \sum_{i \in S_l} d^2\left(X_i, \overline{X}_l\right),$$

где l — номер кластера (l = 1,2,..., k), $\overline{X}$ — центр l-го кластера, X_l — вектор значений переменных для i-го объекта, входящего в l-й кластер, $d\left(X_i, \overline{X}_l\right)$ — расстояние между i-м объектом и центром l-го кластера.

При использовании этого критерия стремятся получить такое разбиение совокупности объектов на k кластеров, при котором значение F_1 было бы минимальным.

2. Сумма внутриклассовых расстояний между объектами:

$$F_2 = \sum_{l=1}^{k} \sum_{i,j \in S_l} d_{ij}^2 .$$

В этом случае наилучшим следует считать такое разбиение, при котором достигается минимальное значение F_2, т.е. получены кластеры большой «плотности». Объекты, попавшие в один кластер, близки между собой по значениям тех переменных, которые использовались для классификации.

3. Суммарная внутриклассовая дисперсия:

$$F_3 = \sum_{l=1}^{k} \sum_{j=1}^{p} \sigma_{lj}^2 ,$$

где σ_{lj}^2 — дисперсия j-й переменной в кластере S_l.

В данном случае разбиение, при котором сумма внутриклассовых (внутригрупповых) дисперсий будет минимальной, следует считать оптимальным. Существует несколько алгоритмов кластерного анализа, обеспечивающих оптимальное разбиение с точки зрения функционала F_3. Например, итерационный алгоритм, включающий следующие вычислительные процедуры:

а) в качестве начального разбиения задается разбиение совокупности объектов на k кластеров. Оно может быть получено одним из иерархических методов;

б) для каждого кластера S_l определяется центр $\overline{X}_l = \left(\bar{x}_{1l}, \bar{x}_{2l}, \ldots, \bar{x}_{pl}\right)$. Каждая координата центра вычисляется следующим образом:

$$\bar{x}_{jl} = \frac{1}{n_l} \sum_{i=1}^{n_l} x_{ij},$$

где i — номер объекта, j — номер переменной, l — номер кластера, n_l — количество объектов в кластере S_l;

в) все объекты исходной совокупности распределяются по кластерам в зависимости от их расстояния до центров этих кластеров, т.е. i-й объект будет включен в кластер S_l в том случае, если его расстояние до центра этого кластера

$$d_{i\bar{x}l} = \min_{q=1\ldots k} \left\{ d_{i\bar{x}q} \right\}.$$

После распределения объектов по k кластерам сравнивают первоначальный состав этих кластеров с вновь полученным. Если обнаруживается несовпадение, тогда работа алгоритма продолжается, повторяются процедуры (*б*) и (*в*). Локальный экстремум достигается в том случае, если совпадают результаты последующей и предыдущей группировок. Следует заметить, что для другого начального разбиения оптимальное значение функционала F_3 будет отличаться. На принципе минимизации внутрикластерной дисперсии основаны алгоритмы метода k-средних и метода Уорда.

Рассмотрим на примере расчет функционала F_3. Предположим, что шесть объектов наблюдения (см. табл. 9.9) распределены по методу k-средних на три кластера следующим образом:

кластер S_1 — объект n_1,
кластер S_2 — объекты n_2, n_6,
кластер S_3 — объекты n_3, n_4, n_5.

Рассчитаем дисперсии для каждой переменной в каждом кластере (σ_{ij}^2):

$\sigma_{11}^2 = 0$; $\sigma_{21}^2 = 0{,}004225$; $\sigma_{31}^2 = 0{,}008422$;

$\sigma_{12}^2 = 0$; $\sigma_{22}^2 = 0{,}25$; $\sigma_{32}^2 = 0{,}666(6)$;

$\sigma_{13}^2 = 0$; $\sigma_{23}^2 = 0{,}20$; $\sigma_{33}^2 = 0{,}1866(6)$.

$\sum_{j=1}^{3}\sigma_{1j}^2 = 0$; $\sum_{j=1}^{3}\sigma_{2j}^2 = 0{,}454225$; $\sum_{j=1}^{3}\sigma_{3j}^2 = 0{,}8617$.

Тогда суммарная дисперсия всех переменных по трем кластерам будет равна:

$$F_3 = \sum_{j=1}^{3}\sigma_{1j}^2 + \sum_{j=1}^{3}\sigma_{2j}^2 + \sum_{j=1}^{3}\sigma_{3j}^2 = 1{,}3159.$$

Проведем классификацию тех же шести объектов по методу иерархического кластерного анализа, используя алгоритм «дальнего соседа». Получим разбиение на три кластера:

кластер S_1 — объекты n_1, n_4,

кластер S_2 — объекты n_2, n_6,

кластер S_3 — объекты n_3, n_5.

Суммарная дисперсия всех переменных по трем кластерам будет равна:

$$F_3 = \sum_{j=1}^{3}\sigma_{1j}^2 + \sum_{j=1}^{3}\sigma_{2j}^2 + \sum_{j=1}^{3}\sigma_{3j}^2 = 0{,}5016 + 0{,}4542 + 0{,}8617 = 1{,}8175.$$

Если судить по суммарной дисперсии трех переменных, то разбиение по методу k-средних оказалось лучше, чем по иерархическому методу.

В данном примере функционал качества F_3 выступает мерой однородности всех кластеров в целом. Аналогичны по смыслу критерии F_1 и F_2.

Если оценивать качество разбиения по степени удаленности кластеров друг от друга, то можно использовать функционал F_4 — средние межклассовые расстояния:

$$F_4 = \sum_{\substack{i\in S_q \\ j\in S_l}} d_{ij} \Big/ \sum_{l<q} n_l n_q - \max.$$

Кроме названных функционалов качество классификации можно также оценивать и при помощи критерия Хотеллинга

T^2 для проверки гипотезы о равенстве векторов средних для многомерных совокупностей:

$$T^2 = \frac{n_l n_q}{n_l + n_q}(\bar{X}_l - \bar{X}_q)' S_*^{-1}(\bar{X}_l - \bar{X}_q).$$

Судить о качестве разбиения позволяют и некоторые простейшие приемы. Например, сравнение средних значений признаков в отдельных кластерах (группах) со средними значениями в целом по всей совокупности объектов. Если отличие групповых средних от общего среднего значения существенное, то это может являться признаком хорошего разбиения. Оценка существенности различий может быть выполнена с помощью t-критерия Стьюдента.

Результаты многомерной классификации можно оценивать и по типу образовавшихся кластеров. Считается, что чем больше среди них кластеров типа сгущения или «сильных» кластеров, тем лучше качество разбиения [50, с. 188].

Перечисленные выше способы оценки качества разбиения предполагают чисто формальный подход и являются для исследователя только вспомогательными средствами. Основная роль принадлежит содержательному анализу результатов классификации. Выбор лучшего варианта разбиения облегчается в значительной мере серьезной подготовительной работой, в частности, выбором признаков, характеризующих классифицируемые объекты. В зависимости от количества признаков, их взаимосвязи, выбранного масштаба измерения подбирается наиболее подходящий алгоритм классификации, задаются начальные параметры разбиения. Все это облегчает интерпретацию результатов разбиения и позволяет судить о его качестве с точки зрения поставленной задачи.

Выводы

Методы кластерного анализа представляют собой очень эффективное средство статистической обработки данных, позволяющее решать следующие задачи:

1) изучение структуры совокупности с целью выделения групп объектов, схожих между собой по нескольким признакам;
2) проверка выдвигаемых предположений о наличии структуры в изучаемой совокупности, а также построение новых классификаций;

3) снижение размерности признакового пространства перед проведением корреляционно-регрессионного анализа без потери информации о взаимосвязи между переменными.

Классификация в иерархических алгоритмах осуществляется за один просмотр исходных данных. Алгоритм обработки не предусматривает повторного просмотра с целью улучшения качества уже образованных кластеров. Фактически все методы кластерного анализа в той или иной мере содержат определенные допущения, что приводит к появлению различных вариантов классификации даже в рамках одного метода. Большинство иерархических методов приводит к образованию непересекающихся кластеров. В итеративных методах задание параметров разбиения может базироваться на результатах иерархического кластерного анализа.

Выбор алгоритма классификации во многом зависит от принимаемого критерия качества разбиения на классы. Существенное влияние на выбор алгоритма и на качество классификации оказывают перечень и структура признаков классификации. Для устранения влияния масштаба классификационных признаков на результаты кластерного анализа необходимо предварительно данные нормировать (стандартизовать). Если в числе классификационных признаков встречаются качественные характеристики или квазиколичественные (например, кодовые обозначения объектов), то необходимо произвести оцифровку этих признаков по определенному правилу, т.е. заменить значения признаков числовыми метками.

Результатом проведения кластерного анализа является получение групп сходных объектов. Различные методы кластерного анализа позволяют получать кластеры, различающиеся по размеру и по форме.

Вопросы и задачи

1. В чем состоит принципиальное отличие методов многомерных классификаций от комбинационных группировок?

2. Перечислите задачи, решаемые при помощи методов кластерного анализа.

3. Назовите две основные группы методов кластерного анализа и укажите их сходство и различие.

4. Какие меры сходства используются при проведении многомерной классификации?

5. Какие методы кластерного анализа могут привести к образованию пересекающихся кластеров?

6. Перечислите основные вычислительные процедуры метода поиска сгущений.

7. Как оценивается качество полученного разбиения совокупности на кластеры?

8. По результатам статистического наблюдения получены следующие данные о населении за 1993 г.:

Государство	*Коэффициент младенческой смертности, %*	*Ожидаемая продолжительность жизни при рождении, лет*	*Государство*	*Коэффициент младенческой смертности, %*	*Ожидаемая продолжительность жизни при рождении, лет*
Азербайджан	28,2	70,5	Таджикистан	42,7	70,5
Армения	17,1	70,4	Туркмения	45,9	65,8
Беларусь	12,5	70,3	Узбекистан	35,6	69,3
Казахстан	28,4	68,6	Украина	14,9	70,5
Кыргызстан	31,8	68,8	Грузия	18,3	72,6
Молдова	21,5	67,7	Латвия	15,9	69,2
Россия	20,0	69,0	Литва	15,2	70,7
			Эстония	15,7	70,0

Проведите классификацию государств, используя методы: *k*-средних и агломеративные с алгоритмами «ближайшего» и «дальнего соседа». Результаты классификации представьте в виде дендрограммы.

9. На основании приведенных ниже данных произведите группировку магазинов по площади торгового зала и по товарообороту:

Порядковый номер магазина	Площадь торгового зала, м²	Товарооборот, млрд. руб.	Порядковый номер магазина	Площадь торгового зала, м²	Товарооборот, млрд. руб.
1	110	525	11	200	600
2	100	546	12	180	867
3	90	501	13	200	464
4	90	427	14	170	512
5	95	683	15	180	600
6	120	644	16	155	695
7	130	615	17	205	662
8	100	580	18	180	480
9	100	700	19	160	570
10	130	532	20	150	564

Для классификации магазинов воспользуйтесь методом поиска сгущений. По результатам разбиений рассчитайте сумму внутриклассовых дисперсий. Для выполнения всех необходимых расчетов воспользуйтесь ПЭВМ и пакетом *CLUSTAN*.

10. Используя приведенные ниже исходные данные, проведите иерархический кластерный анализ по алгоритмам «ближайшего соседа» и «дальнего соседа». Рассчитайте для каждого разбиения сумму квадратов расстояний до центров классов (функционал F_1).

Определите, какой из двух алгоритмов приводит к лучшим результатам.

Порядковый номер рабочего	Стаж работы, лет	Выполнение норм выработки, %	Средняя месячная заработная плата, тыс. руб.
1	6	108	3785
2	5	101	2413
3	11	126	5840
4	7	109	4600
5	18	110	6020
6	13	160	6100

Примечание. Исходные данные предварительно нормируйте по способу $z_{ij} = x_{ij} / \bar{x}_j$. Для определения расстояний воспользуйтесь метрикой «евклидово расстояние».

11. На основании комплекса показателей, приведенных в таблице, проведите классификацию предприятий по методу поиска сгущений. Проанализируйте полученные результаты и оцените степень однородности образованных кластеров на основании внутрикластерных дисперсий:

Номер предприятия	Удельный вес продукции, изготовленной по новым технологиям, %	Коэффициент использования сырья и материалов, %	Удельный вес продукции высшей категории качества, %
1	45,0	60,3	15,7
2	38,5	64,1	11,8
3	55.3	88,9	14,0
4	24,0	44,3	16,1
5	37,8	45,7	12,5
6	30,5	71,0	11,4
7	50,0	90,6	12,8
8	25,5	68,9	15,7
9	36,1	72,3	13,0
10	41,7	66,4	16,5

Расчеты выполните на ПЭВМ.

12. По приведенным ниже данным для шести предприятий проведите иерархическую агломеративную группировку, используя алгоритмы «ближайшего соседа» и «дальнего соседа». Результаты каждой группировки изобразите в виде дендрограммы и оцените качество полученных разбиений при помощи функционала F_1:

Номер предприятия	X_1	X_2	X_3
1	570	19,4	48,4
2	860	21,6	50,2
3	1150	28,8	64,7
4	610	20,5	54,4
5	502	23,3	53,8
6	670	17,8	36,2
Итого	4362	131,4	307,7

Примечание. В качестве меры сходства используйте расстояние city-block.

Здесь X_1 — численность работников; X_2 — фондовооруженность труда одного работника, млн. руб.; X_3 — энерговооруженность труда одного рабочего, кВт.

13. Из годовых отчетов сельскохозяйственных предприятий известны следующие данные по трем признакам: X_1 — уровень

энерговооруженности труда, л.с.; X_2 — фондоотдача основных производственных фондов сельскохозяйственного назначения, руб.; X_3 — урожайность зерновых культур, ц / га.

Наименование хозяйства	X_1	X_2	X_3
ОАО «Полесье»	38	0,68	18,3
Колхоз «Рассвет»	35	0,62	20,7
Колхоз «Ударник»	27	0,56	14,2
ОАО «Ждановичи»	48	0,72	22,4

Проведите стандартизацию исходных данных различными способами и определите, как это влияет на результаты классификации.

14. В результате проведенной классификации десяти городов по трем признакам получено следующее разбиение на три группы:

Номер группы и номера объектов	*Значения группировочных признаков (x_{ij})*			
1 группа {4,5}	$x_{41} = 76$ $x_{42} = 16,5$ $x_{43} = 10,5$	$x_{51} = 75$ $x_{52} = 16,3$ $x_{53} = 11,0$		
2 группа {1,2,6,7}	$x_{11} = 71$ $x_{12} = 17,1$ $x_{13} = 12,2$	$x_{21} = 72$ $x_{22} = 17,0$ $x_{23} = 12,0$	$x_{61} = 70$ $x_{62} = 17,6$ $x_{63} = 12,8$	$x_{71} = 73$ $x_{72} = 17,8$ $x_{73} = 13,1$
3 группа {3,8,9,10}	$x_{31} = 74$ $x_{32} = 18,0$ $x_{33} = 13,0$	$x_{81} = 75$ $x_{82} = 17,7$ $x_{83} = 13,2$	$x_{91} = 72$ $x_{92} = 17,5$ $x_{93} = 13,4$	$x_{101} = 76$ $x_{102} = 18$ $x_{103} = 13,5$

Здесь X_1 — средняя продолжительность жизни, лет; X_2 — коэффициент рождаемости, ‰; X_3 — коэффициент младенческой смертности, ‰.

Проведите классификацию этих же городов по методу поиска сгущений и сравните качество полученных разбиений при помощи функционала F_1.

Глава 10

Дискриминантный анализ

10.1. Основные положения дискриминантного анализа

Дискриминантный анализ — это раздел математической статистики, содержанием которого является разработка методов решения задач различения (дискриминации) объектов наблюдения по определенным признакам. Например, разбиение совокупности предприятий на несколько однородных групп по значениям каких-либо показателей производственно-хозяйственной деятельности.

Методы дискриминантного анализа находят применение в различных областях: медицине, социологии, психологии, экономике и т.д. При наблюдении больших статистических совокупностей часто появляется необходимость разделить неоднородную совокупность на однородные группы (классы). Такое расчленение в дальнейшем при проведении статистического анализа дает лучшие результаты моделирования зависимостей между отдельными признаками.

Дискриминантный анализ оказывается очень удобным и при обработке результатов тестирования отдельных лиц. Например, при выборе кандидатов на определенную должность можно всех опрашиваемых претендентов разделить на две группы: «подходит» и «не подходит».

Можно привести еще один пример применения дискриминантного анализа в экономике. Для оценки финансового состояния своих клиентов при выдаче им кредита банк классифицирует их на надежных и ненадежных по ряду признаков. Таким образом, в тех случаях, когда возникает необходимость отнесения того или иного объекта к одному из реально существующих или выделенных определенным способом классов, можно воспользоваться дискриминантным анализом.

Аппарат дискриминантного анализа разрабатывался многими учеными-специалистами, начиная с конца 50-х годов XX в. Дискриминантным анализом, как и другими методами многомерной статистики, занимались П.Ч. Махаланобис, Р. Фишер, Г. Хотеллинг и другие видные ученые.

Все процедуры дискриминантного анализа можно разбить на две группы и рассматривать их как совершенно самостоятельные методы. Первая группа процедур позволяет интерпретировать различия между существующими классами, вторая — проводить классификацию новых объектов в тех случаях, когда неизвестно заранее, к какому из существующих классов они принадлежат.

Пусть имеется множество единиц наблюдения — генеральная совокупность. Каждая единица наблюдения характеризуется несколькими признаками (переменными) x_{ij} — значение j-й переменной у i-го объекта $i = 1, ..., N$; $j = 1, ..., p$.

Предположим, что все множество объектов разбито на несколько подмножеств (два и более). Из каждого подмножества взята выборка объемом n_k, где k — номер подмножества (класса), $k = 1, ..., q$.

Признаки, которые используются для того, чтобы отличать один класс (подмножество) от другого, называются *дискриминантными переменными*. Каждая из этих переменных должна измеряться либо по интервальной шкале, либо по шкале отношений. Интервальная шкала позволяет количественно описать различия между свойствами объектов. Для задания шкалы устанавливаются произвольная точка отсчета и единица измерения. Примерами таких шкал являются календарное время, шкалы температур и т. п. В качестве оценки положения центра используются средняя величина, мода и медиана.

Шкала отношений — частный случай интервальной шкалы. Она позволяет соотнести количественные характеристики какого-либо свойства у разных объектов, например, стаж работы, заработная плата, величина налога.

Теоретически число дискриминантных переменных не ограничено, но на практике их выбор должен осуществляться на основании логического анализа исходной информации и одного из критериев, о котором речь пойдет немного ниже. Число объектов наблюдения должно превышать число дискриминантных переменных, как минимум, на два, т. е. $p < N$. Дискриминантные переменные должны быть линейно независимыми. Еще одним

предположением при дискриминантном анализе является нормальность закона распределения многомерной величины, т.е. каждая из дискриминантных переменных внутри каждого из рассматриваемых классов должна быть подчинена нормальному закону распределения. В случае, когда реальная картина в выборочных совокупностях отличается от выдвинутых предпосылок, следует решать вопрос о целесообразности использования процедур дискриминантного анализа для классификации новых наблюдений, так как в этом случае затрудняются расчеты каждого критерия классификации.

10.2. Дискриминантные функции и их геометрическая интерпретация

Перед тем как приступить к рассмотрению алгоритма дискриминантного анализа, обратимся к его геометрической интерпретации.

На рис. 10.1 изображены объекты, принадлежащие двум различным множествам M_1 и M_2. Каждый объект характеризуется в данном случае двумя переменными x_1 и x_2. Если рассматривать проекции объектов (точек) на каждую ось, то эти множества пересекаются, т.е. по каждой переменной отдельно некоторые объекты обоих множеств имеют сходные характеристики. Чтобы наилучшим образом разделить два рассматриваемых множества, нужно построить соответствующую линейную комбинацию переменных x_1 и x_2. Для двумерного пространства эта задача сводится к определению новой системы координат. Причем новые оси L и C должны быть расположены таким образом, чтобы проекции объектов, принадлежащих разным множествам на ось L, были максимально разделены. Ось C перпендикулярна оси L и разделяет два «облака» точек наилучшим образом, т.е. чтобы множества оказались по разные стороны от этой прямой. При этом вероятность ошибки классификации должна быть минимальной. Сформулированные условия должны быть учтены при определении коэффициентов a_1 и a_2 следующей функции:

$$f(x) = a_1x_1 + a_2x_2. \qquad (10.1)$$

Функция $f(x)$ называется *канонической дискриминантной функцией*, а величины x_1 и x_2 — дискриминантными переменными.

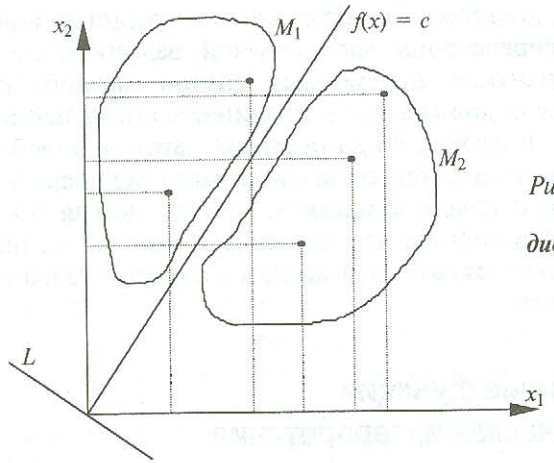


Рис. 10.1. Геометрическая интерпретация дискриминантной функции и дискриминантных переменных

Обозначим x_{ij} — среднее значение j-го признака у объектов i-го множества (класса). Тогда для множества M_1 среднее значение функции $f_1(x)$ будет равно:

$\overline{f_1}(x) = a_1\bar{x}_{11} + a_2\bar{x}_{12}$;

для множества M_2 среднее значение функции $f_2(x)$ равно:

$\overline{f_2}(x) = a_1\bar{x}_{21} + a_2\bar{x}_{22}$.

Геометрическая интерпретация этих функций — две параллельные прямые, проходящие через центры классов (множеств) (рис. 10.2).

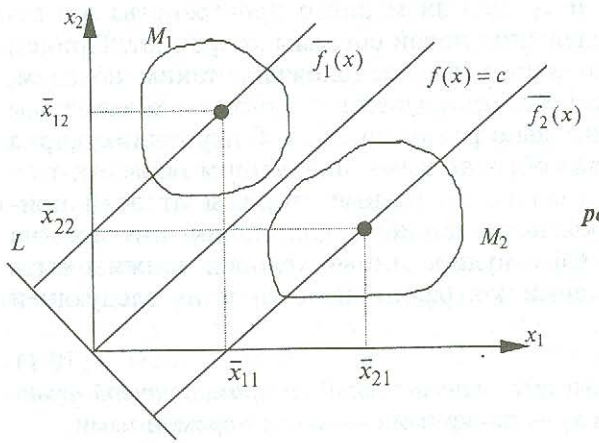


Рис. 10.2. Центры разделяемых множеств и константа дискриминации

Дискриминантная функция может быть как линейной, так и нелинейной. Выбор ее вида зависит от геометрического расположения разделяемых классов в пространстве дискриминантных переменных. Для упрощения выкладок в дальнейшем рассматривается линейная дискриминантная функция.

10.3. Расчет коэффициентов дискриминантной функции

Коэффициенты дискриминантной функции a_i определяются таким образом, чтобы $\overline{f}_1(x)$ и $\overline{f}_2(x)$ как можно больше различались между собой, т.е. чтобы для двух множеств (классов) было максимальным выражение

$$\overline{f}_1(x) - \overline{f}_2(x) = \sum_{i=1}^{n_1} a_1 x_{1i} - \sum_{i=1}^{n_2} a_i x_{2i}\,. \qquad (10.2)$$

Тогда можно записать следующее:

$$f_{kt}(x) - \overline{f}_k(x) = a_1(x_{1kt} - \bar{x}_{1k}) + a_2(x_{2kt} - \bar{x}_{2k}) + \cdots + a_p(x_{pkt} - \bar{x}_{pk})\,, \qquad (10.3)$$

где k — номер группы;

p — число переменных, характеризующих каждое наблюдение.

Обозначим дискриминантную функцию $f_k(x)$ как Y_{kt} (k — номер группы, t — номер наблюдения в группе). Внутригрупповая вариация может быть измерена суммой квадратов отклонений:

$$\sum_{t=1}^{n_k} (Y_{kt} - \overline{Y}_k)^2.$$

По обеим группам это будет выглядеть следующим образом:

$$\sum_{k=1}^{2}\sum_{t=1}^{n_k} (Y_{kt} - \overline{Y}_k)^2 = \sum_{k=1}^{2}\sum_{t=1}^{n_k} \left[a_1(x_{1kt} - \bar{x}_{1k}) + a_2(x_{2kt} - \bar{x}_{2k}) \cdots + a_p(x_{pkt} - \bar{x}_{pk})\right]^2.$$

В матричной форме это выражение может быть записано так:

$$\sum_{i=1}^{2}\sum_{t=1}^{n_k} (Y_{kt} - \overline{Y}_k)^2 = A'(X_1'X_1 + X_2'X_2)A, \qquad (10.4)$$

где A — вектор коэффициентов дискриминантной функции;

X_1' — транспонированная матрица отклонений наблюдаемых значений исходных переменных от их средних величин в первой группе

$$X_1' = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n1} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n1} \\ \cdots & \cdot & \cdot & \cdot \\ x_{p,1} & x_{p,2} & \cdots & x_{p,n1} \end{pmatrix};$$

$$X_2' = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n2} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n2} \\ \cdot\cdot & \cdot\cdot & \cdot & \cdot \\ x_{p,1} & x_{p,2} & \cdots & x_{p,n2} \end{pmatrix},$$

X_2' — аналогичная матрица для второй группы.

Объединенная ковариационная матрица S_* определяется так:

$$S_* = \frac{1}{n_1 + n_2 - 2}(X_1'X_1 + X_2'X_2),$$

следовательно, выражение (10.4) дает оценку внутригрупповой вариации и его можно записать в виде:

$$\sum_{k=1}^{2}\sum_{t=1}^{n_k}(Y_{kt} - \overline{Y}_k)^2 = A'[(n_1 + n_2 - 2)\cdot S_*]A. \qquad (10.5)$$

Межгрупповая вариация может быть измерена как

$$(\overline{Y}_1 - \overline{Y}_2)^2 = A'(\overline{X}_1 - \overline{X}_2)\cdot(\overline{X}_1 - \overline{X}_2)'A.$$

При нахождении коэффициентов дискриминантной функции a_j следует исходить из того, что для рассматриваемых объектов внутригрупповая вариация должна быть минимальной, а межгрупповая вариация — максимальной. В этом случае мы достигнем наилучшего разделения двух групп, т.е. необходимо, чтобы величина F была максимальной:

$$F = \frac{A'(\overline{X}_1 - \overline{X}_2)\cdot(\overline{X}_1 - \overline{X}_2)'A}{A'[(n_1 + n_2 - 2)S_*]A} - \max. \qquad (10.6)$$

В точке, где функция F достигает максимума, частные производные по a_j будут равны нулю. Если вычислить частные производные

$$\frac{dF}{da_1}; \frac{dF}{da_2}; \ldots; \frac{dF}{da_p}$$

и приравнять их нулю, то после преобразований получим выражение:

$$A = S_*^{-1}(\overline{X}_1 - \overline{X}_2). \qquad (10.7)$$

Из этой формулы и определяется вектор коэффициентов дискриминантной функции (A).

Полученные значения коэффициентов подставляют в формулу (10.1) и для каждого объекта в обеих группах (множествах) вычисляют дискриминантные функции, затем находят среднее значение для каждой группы. Таким образом, каждое i-е наблюдение, которое первоначально описывалось m переменными, будет как бы помещено в одномерное пространство, т.е. ему будет соответствовать одно значение дискриминантной функции, следовательно, размерность признакового пространства снижается.

10.4. Классификация при наличии двух обучающих выборок

Перед тем как приступить непосредственно к процедуре классификации, нужно определить границу, разделяющую в частном случае две рассматриваемые группы. Такой величиной может быть значение функции, равноудаленное от $\overline{f}_1$ и $\overline{f}_2$, т.е.

$$C = \frac{1}{2}(\overline{f}_1 + \overline{f}_2).$$

Величина C называется *константой дискриминации*.

На рис. 10.1 видно, что объекты, расположенные над прямой $f(x) = a_1x_1 + a_2x_2 + \cdots + a_px_p = C$, находятся ближе к центру множества M_1 и, следовательно, могут быть отнесены к первой группе, а объекты, расположенные ниже этой прямой, ближе к центру второго множества, т.е. относятся ко второй группе. Если граница между группами выбрана так, как сказано выше, то суммарная вероятность ошибочной классификации минимальная.

Рассмотрим пример использования дискриминантного анализа для проведения многомерной классификации объектов. При этом в качестве обучающих будем использовать сначала две выборки, принадлежащие двум классам, а затем обобщим алгоритм классификации на случай k классов.

Пример 10.1. Имеются данные по двум группам промышленных предприятий машиностроительного комплекса:
X_1 — фондоотдача основных производственных фондов, руб.;
X_2 — затраты на рубль произведенной продукции, коп.;
X_3 — затраты сырья и материалов на один рубль продукции, коп.

	Номер предприятия	X_1	X_2	X_3
1-я группа	1	0,50	94,0	8,50
	2	0,67	75,4	8,79
	3	0,68	85,2	9,10
	4	0,55	98,8	8,47
2-я группа	5	1,52	81,5	4,95
	6	1,20	93,8	6,95
	7	1,46	86,5	4,70

Необходимо провести классификацию четырех новых предприятий, имеющих следующие значения исходных переменных:
1-е предприятие: $x_1 = 1{,}07$, $x_2 = 93{,}5$, $x_3 = 5{,}30$,
2-е предприятие: $x_1 = 0{,}99$, $x_2 = 84{,}0$, $x_3 = 4{,}85$,
3-е предприятие: $x_1 = 0{,}70$, $x_2 = 76{,}8$, $x_3 = 3{,}50$,
4-е предприятие: $x_1 = 1{,}24$, $x_2 = 88{,}0$, $x_3 = 4{,}95$.

Для удобства запишем значения исходных переменных для каждой группы предприятий в виде матриц X_1 и X_2:

$$X_1 = \begin{pmatrix} 0{,}50 & 94{,}0 & 8{,}50 \\ 0{,}67 & 75{,}4 & 8{,}79 \\ 0{,}68 & 85{,}2 & 9{,}10 \\ 0{,}55 & 98{,}8 & 8{,}47 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1{,}52 & 81{,}5 & 4{,}95 \\ 1{,}20 & 93{,}8 & 6{,}95 \\ 1{,}46 & 86{,}5 & 4{,}70 \end{pmatrix}.$$

Рассчитаем среднее значение каждой переменной в отдельных группах для определения положения центров этих групп:

I гр. $\bar{x}_{11} = 0{,}60$, $\bar{x}_{21} = 88{,}4$, $\bar{x}_{31} = 8{,}72$

II гр. $\bar{x}_{12} = 1{,}39$, $\bar{x}_{22} = 87{,}3$, $\bar{x}_{32} = 5{,}53$.

Дискриминантная функция $f(x)$ в данном случае имеет вид:

$$f(x) = a_1x_1 + a_2x_2 + a_3x_3. \qquad (10.8)$$

Коэффициенты a_1, a_2 и a_3 вычисляются по формуле:

$$A = S_*^{-1}(\bar{X}_1 - \bar{X}_2),$$

где $\bar{X}_1, \bar{X}_2$ — векторы средних в первой и второй группах;
A — вектор коэффициентов;
S_* — матрица, обратная совместной ковариационной матрице.

Для определения совместной ковариационной матрицы S_* нужно рассчитать матрицы S_1 и S_2. Каждый элемент этих матриц представляет собой разность между соответствующим значением исходной переменной x_{ij} и средним значением этой переменной в данной группе $\bar{x}_{ik}$ (k — номер группы):

$$S_1 = \begin{pmatrix} 0{,}0238 & -2{,}2460 & 0{,}0698 \\ -2{,}2460 & 318{,}76 & -5{,}958 \\ 0{,}0698 & -5{,}958 & 0{,}2602 \end{pmatrix}; \quad S_2 = \begin{pmatrix} 0{,}0579 & -2{,}0450 & -0{,}4033 \\ -2{,}0450 & 76{,}530 & 13{,}2580 \\ -0{,}4033 & 13{,}258 & 3{,}0417 \end{pmatrix}.$$

Тогда совместная ковариационная матрица S_* будет равна:

$$S_* = \frac{1}{n_1 + n_2 - 2}(S_1 + S_2),$$

где n_1, n_2 — число объектов 1-й и 2-й группы;

$$S_* = \frac{1}{(4+3-2)} \begin{pmatrix} 0{,}0817 & -4{,}291 & -0{,}3335 \\ -4{,}291 & 395{,}290 & 7{,}300 \\ -0{,}3335 & 7{,}300 & 3{,}3019 \end{pmatrix} = \begin{pmatrix} 0{,}01634 & -0{,}8582 & -0{,}0667 \\ -0{,}8582 & 79{,}058 & 1{,}460 \\ -0{,}0667 & 1{,}460 & 0{,}6604 \end{pmatrix}.$$

Обратная матрица S_*^{-1} будет равна:

$$S_*^{-1} = \begin{pmatrix} 339{,}970 & -3{,}190 & 27{,}290 \\ -3{,}190 & 0{,}043 & -0{,}227 \\ 27{,}290 & -0{,}227 & 8{,}380 \end{pmatrix}.$$

Отсюда находим вектор коэффициентов дискриминантной функции по формуле:

$$A = S_*^{-1}(\bar{X}_1 - \bar{X}_2) = \begin{pmatrix} 339{,}970 & -3{,}190 & 27{,}290 \\ -3{,}190 & 0{,}043 & -0{,}227 \\ 27{,}290 & -0{,}227 & 8{,}380 \end{pmatrix} \cdot \begin{pmatrix} -0{,}79 \\ 1{,}10 \\ 3{,}19 \end{pmatrix} = \begin{pmatrix} -185{,}03 \\ 1{,}84 \\ 4{,}92 \end{pmatrix},$$

т.е. $a_1 = -185{,}03$, $a_2 = 1{,}84$, $a_3 = 4{,}92$.

Подставим полученные значения коэффициентов в формулу (10.8) и рассчитаем значения дискриминантной функции для каждого объекта:

Для 1-го множества:

$$\begin{cases} f_{11} = 0{,}5(-185{,}03) + 94 \times 1{,}84 + 8{,}5 \times 4{,}92 = 122{,}265, \\ f_{12} = 0{,}67(-185{,}03) + 75{,}4 \times 1{,}84 + 8{,}79 \times 4{,}92 = 58{,}0127, \\ f_{13} = 0{,}68(-185{,}03) + 85{,}2 \times 1{,}84 + 9{,}1 \times 4{,}92 = 75{,}7196, \\ f_{14} = 0{,}55(-185{,}03) + 98{,}8 \cdot 1{,}84 + 8{,}47 \times 4{,}92 = 121{,}6979, \end{cases}$$

$$\bar{f}_1 = 94{,}4238.$$

Для 2-го множества:

$$\begin{cases} f_{21} = 1{,}52 - 185{,}03 + 81{,}5 \times 1{,}84 + 4{,}95 \times 4{,}92 = -106{,}9316, \\ f_{22} = 1{,}20 - 185{,}03 + 93{,}8 \times 1{,}84 + 6{,}95 \times 4{,}92 = -15{,}25, \\ f_{23} = 1{,}46 - 185{,}03 + 86{,}5 \times 1{,}84 + 4{,}7 \times 4{,}92 = -87{,}8598, \end{cases}$$

$$\bar{f}_2 = -70{,}0138.$$

Тогда константа дискриминации C будет равна:

$$C = \frac{1}{2}(94{,}4238 - 70{,}0138) = 12{,}205.$$

После получения константы дискриминации можно проверить правильность распределения объектов в уже существующих двух классах, а также провести классификацию новых объектов.

Рассмотрим, например, объекты с номерами 1, 2, 3, 4. Для того чтобы отнести эти объекты к одному из двух множеств, рассчитаем для них значения дискриминантных функций (по трем переменным):

$f_1 = -185{,}03 \times 1{,}07 + 1{,}84 \times 93{,}5 + 4{,}92 \times 5{,}30 = 0{,}1339,$
$f_2 = -185{,}03 \times 0{,}99 + 1{,}84 \times 84{,}0 + 4{,}92 \times 4{,}85 = -4{,}7577,$
$f_3 = -185{,}03 \times 0{,}70 + 1{,}84 \times 76{,}8 + 4{,}92 \times 3{,}50 = 29{,}0110,$
$f_4 = -185{,}03 \times 1{,}24 + 1{,}84 \times 88{,}0 + 4{,}92 \times 4{,}95 = -43{,}1632.$

Таким образом, объекты 1, 2 и 4 относятся ко второму классу, а объект 3 относится к первому классу, так как $f_1 < c$, $f_2 < c$, $f_3 > c$, $f_4 < c$.

10.5. Классификация при наличии k обучающих выборок

При необходимости можно проводить разбиение множества объектов на k классов (при $k > 2$). В этом случае нужно рассчитать k дискриминантных функций, так как классы будут отделяться друг от друга индивидуальными разделяющими поверхностями. На рис. 10.3 показан случай с тремя множествами и тремя дискриминантными переменными:

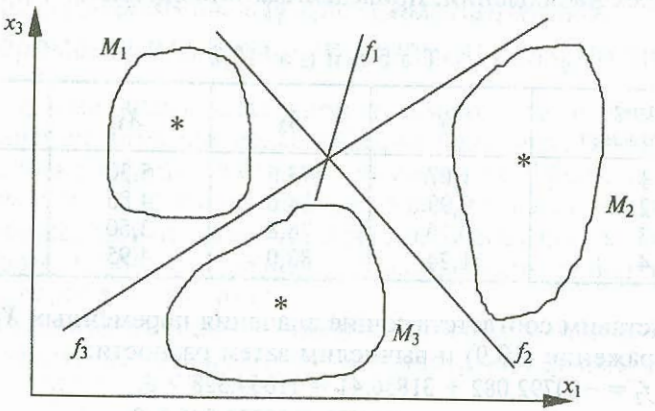


Рис. 10.3. ***Три класса объектов и разделяющие их прямые***
f_1 — первая, f_2 — вторая и f_3 — третья дискриминантные функции.

Пример 10.2. Рассмотрим случай, когда существует три класса (множества) объектов. Для этого к двум классам из предыдущего примера добавим еще один. В этом случае будем иметь уже три матрицы исходных данных:

$$X_1 = \begin{pmatrix} 0{,}50 & 94{,}0 & 8{,}50 & 6707 \\ 0{,}67 & 75{,}4 & 8{,}79 & 5037 \\ 0{,}68 & 85{,}2 & 9{,}10 & 3695 \\ 0{,}55 & 98{,}8 & 8{,}47 & 6815 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1{,}52 & 81{,}5 & 4{,}95 & 3211 \\ 1{,}20 & 93{,}8 & 6{,}95 & 2890 \\ 1{,}46 & 86{,}5 & 4{,}70 & 2935 \end{pmatrix},$$

$$X_3 = \begin{pmatrix} 1{,}70 & 80{,}0 & 4{,}5 & 3510 \\ 1{,}65 & 85{,}0 & 4{,}8 & 2900 \\ 1{,}49 & 78{,}5 & 4{,}1 & 2850 \end{pmatrix}.$$

Если в процессе дискриминации используются все четыре переменные (X_1, X_2, X_3, X_4), то для каждого класса дискриминантные функции имеют вид:

$$f_1 = 613{,}6X_1 + 5{,}482X_2 + 37{,}53X_3 - 8{,}286X_4 + 22460,$$
$$f_2 = 657{,}2X_1 + 6{,}11X_2 + 33{,}82X_3 - 8{,}377X_4 + 11800, \quad (10.9)$$
$$f_3 = 625{,}0X_1 + 5{,}778X_2 + 31{,}53X_3 - 7{,}692X_4 + 11060.$$

Определим теперь, к какому классу можно отнести каждое из четырех наблюдений, приведенных в табл. 10.2:

Таблица 10.2

Номер наблюдения	X_1	X_2	X_3	X_4
1	1,07	93,5	5,30	5385
2	0,99	84,0	4,85	5225
3	0,70	76,8	3,50	5190
4	1,24	88,0	4,95	6280

Подставим соответствующие значения переменных X_1, X_2, X_3, X_4 в выражение (10.9) и вычислим затем разности:

$f_1 - f_2 = -20792{,}082 + 31856{,}41 = 11064{,}328 > 0,$

$f_1 - f_3 = -20792{,}082 + 40016{,}428 = 19224{,}346 > 0.$

Следовательно, наблюдение 1 в табл. 10.2 относится к первому классу. Аналогичные расчеты показывают, что и остальные три наблюдения следует отнести тоже к первому классу.

Чтобы показать влияние числа дискриминантных переменных на результаты классификации, изменим условие последнего примера. Будем использовать для расчета дискриминантных функций только три переменные: X_1, X_2 и X_3. В этом случае выражения для дискриминантных функций будут иметь вид:

$$f_1 = 36{,}93X_1 + 1{,}288X_2 + 8{,}644X_3 - 105{,}6,$$
$$f_2 = 29{,}29X_1 + 2{,}043X_2 + 5{,}617X_3 - 125{,}1,$$
$$f_3 = 25{,}65X_1 + 2{,}211X_2 + 4{,}579X_3 - 120{,}6.$$

Подставив в эти выражения значения исходных переменных для классифицируемых объектов, нетрудно убедиться, что все они попадают в третий класс, так как

$f_1 - f_2 = -26{,}87 < 0,$

$f_1 - f_3 = -37{,}68 < 0,$

$f_2 - f_3 = -10{,}809 < 0.$

Таким образом, мы видим, что изменение числа переменных сильно влияет на результат дискриминантного анализа. Чтобы судить о целесообразности включения (удаления) дискриминантной переменной, обычно используют специальные статистические критерии, позволяющие оценить значимость ухудшения или улучшения разбиения после включения (удаления) каждой из отобранных переменных [34].

10.6. Взаимосвязь между дискриминантными переменными и дискриминантными функциями

Для оценки вклада отдельной переменной в значение дискриминантной функции целесообразно пользоваться стандартизованными коэффициентами дискриминантной функции. Стандартизованные коэффициенты можно рассчитать двумя путями:

- стандартизовать значения исходных переменных таким образом, чтобы их средние значения были равны нулю, а дисперсии — единице;
- вычислить стандартизованные коэффициенты исходя из значений коэффициентов в нестандартной форме:

$$b_j = a_j \sqrt{\frac{W_{jj}}{p-m}},$$

где p — общее число исходных переменных, m — число групп, W_{jj} — элементы матрицы ковариаций:

$$W_{jj} = \sum_{k=1}^{m}\sum_{i=1}^{n_k}(x_{ikj} - \bar{x}_{kj})(x_{ikj} - \bar{x}_{kj}),$$

где i — номер наблюдения, j — номер переменной, k — номер класса, n_k — количество объектов в k-м классе.

Стандартизованные коэффициенты применяют в тех случаях, когда нужно определить, какая из используемых переменных вносит наибольший вклад в величину дискриминантной функции. В примере с двумя классами, рассмотренном выше, дискриминантная функция имела вид:

$$f = -185{,}03X_1 + 1{,}84X_2 + 4{,}92X_3.$$

Следовательно, наибольший вклад в величину дискриминантной функции вносит переменная X_1.

Определим значения стандартизованных коэффициентов и запишем новое значение дискриминантной функции:

$$f' = -211{,}32Z_1 - 99{,}26Z_2 - 12{,}48Z_3,$$

где $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_{x_j}}$.

Стандартизованные коэффициенты дискриминантной функции тоже показывают определяющее влияние первой переменной на величину дискриминантной функции.

Помимо определения вклада каждой исходной переменной в дискриминантную функцию, можно проанализировать и степень корреляционной зависимости между ними.

Для оценки тесноты связи между отдельными переменными и дискриминантными функциями служат коэффициенты корреляции, которые называются структурными коэффициентами. По величине структурных коэффициентов судят о связи между переменными и дискриминантными функциями. Структурные коэффициенты позволяют также в случае необходимости присвоить *имя* каждой функции. Они могут быть рассчитаны в целом по всей совокупности объектов (R) и для каждого класса отдельно (Rx_jf_k).

Покажем на примере 10.1 расчет структурных коэффициентов в целом для трех классов. Исходные данные для расчета коэффициентов представлены в табл. 10.3. Вычисленные структурные коэффициенты (Rx_jf) имеют следующие значения:

$R_{x1f} = 0{,}650$ $R_{x2f} = -0{,}576$ $R_{x3f} = -0{,}506$ $R_{x4f} = -0{,}951$
$R_{x1f1} = -0{,}036$ $R_{x2f1} = 0{,}486$ $R_{x3f1} = -0{,}211$ $R_{x4f1} = 0{,}217$
$R_{x1f2} = -0{,}728$ $R_{x2f2} = 0{,}878$ $R_{x3f2} = 0{,}511$ $R_{x4f2} = -0{,}998$
$R_{x1f3} = -0{,}713$ $R_{x2f3} = 0{,}258$ $R_{x3f3} = -0{,}122$ $R_{x4f3} = -0{,}998.$

Таблица 10.3

Номер наблюдения	X_1	X_2	X_3	X_4	X_5
1	0,50	94,0	8,50	6707	−31973,089
2	0,67	75,4	8,79	5037	−18122,238
3	0,68	85,2	9,10	3695	−6930,930
4	0,55	98,8	8,47	6815	−32812,109
5	1,52	81,5	4,95	3211	−13434,229
6	1,20	93,8	6,95	2890	−10812,723
7	1,46	86,5	4,70	2935	−11139,514
8	1,70	80,0	4,50	3510	−14272,295
9	1,65	85,0	4,80	2900	−9573,076
10	1,49	78,5	4,10	2850	−9348,104

Если рассматривать абсолютные значения структурных коэффициентов, видно, например, что наибольшая зависимость функций f_1 наблюдается от переменной X_2, а функций f_2 и f_3 — от переменной X_4.

Различные знаки у структурных коэффициентов можно интерпретировать следующим образом. Исходные переменные, имеющие различное направление связи с дискриминантной функцией, т.е. положительные или отрицательные структурные коэффициенты, будут ориентировать объекты в различных направлениях, удаляя или приближая их к центрам соответствующих классов. Из данного примера видно, что переменная X_1 и функция f_1 имеют коэффициент -0,036. Это значит, что при увеличении значений X_1 функция f_1 уменьшается. Допустим, все разности $(f_1 - f_l) > 0$ ($l = 2, \ldots, k$) для i-го наблюдения, значит его следует отнести к первому классу. Если у классифицируемых объектов значения переменной X_1 будут возрастать, то значения функции f_1 для этих объектов будут уменьшаться, что приведет к отдалению их от центра первого класса. В конце концов X_1 достигнет у p-го объекта «критического» значения, которому будет соответствовать неравенство $(f_1 - f_l) < 0$, т.е. i-й объект уже не попадет в первый класс. Аналогичные рассуждения проводятся и для положительных структурных коэффициентов.

Выводы

Дискриминантный анализ так же, как и кластерный анализ, относится к методам многомерной классификации, но при этом базируется на иных предпосылках. Основное отличие заключается в том, что в ходе дискриминантного анализа новые кластеры не образуются, а формулируется правило, по которому новые единицы совокупности относятся к одному из уже существующих множеств (классов). Основанием для отнесения каждой единицы совокупности к определенному множеству служит величина дискриминантной функции, рассчитанная по соответствующим значениям дискриминантных переменных.

Основными проблемами дискриминантного анализа являются, во-первых, определение набора дискриминантных переменных, во-вторых, выбор вида дискриминантной функции. Суще-

ствуют различные критерии последовательного отбора переменных, позволяющих получить наилучшее различение множеств [50, с.123—132]. Можно также воспользоваться алгоритмом пошагового дискриминантного анализа, который в литературе подробно описан [11, с.344—354]. После уточнения оптимального набора дискриминантных переменных исследователю предстоит решить вопрос о выборе вида дискриминантной функции, т.е. выбрать вид разделяющей поверхности. Чаще всего на практике применяют линейный дискриминантный анализ. В этом случае дискриминантная функция представляет собой либо прямую, либо плоскость (гиперплоскость).

Линейная дискриминантная функция не всегда подходит в качестве описания разделяющей поверхности между множествами. Например, в тех случаях, когда различаемые множества не являются выпуклыми, правомерно предположить, что дискриминантная функция, приводящая к наименьшим ошибкам классификации, не может быть линейной.

Если множества, используемые в качестве обучающих выборок, близко расположены друг к другу, то возрастает вероятность ошибочной классификации новых объектов, особенно в тех случаях, когда классифицируемый объект сильно удален от центров обоих множеств. Складывается ситуация, при которой распознавание объекта затруднено. Одним из возможных выходов в таком случае является пересмотр набора дискриминантных переменных.

В качестве дискриминантных переменных могут выступать не только исходные (наблюдаемые) признаки, но и главные компоненты или главные факторы.

Дискриминантный анализ можно использовать как метод прогнозирования (предсказания) поведения наблюдаемых единиц статистической совокупности на основе имеющихся стереотипов поведения аналогичных объектов, входящих в состав объективно существующих или сформированных по определенному принципу множеств (обучающих выборок).

Вопросы и задачи

1. В чем заключаются сущность дискриминантного анализа и его отличие от других методов многомерной классификации?

2. Как определяется количество дискриминантных функций?

3. Сформулируйте правило дискриминации.

4. В чем состоит сложность отбора дискриминантных переменных?

5. Приведите графический пример, когда дискриминантная функция будет нелинейной.

6. Поясните, в каких случаях затруднена правильная классификация новых объектов.

7. Из годовых отчетов двух групп промышленных предприятий получены следующие данные, характеризующие их деятельность:

Группы предприятий	*Среднегодовая стоимость основных производственных фондов, млрд. руб.* (X_1)	*Среднесписочная численность работающих, тыс. чел.* (X_2)	*Объем произведенной продукции, млрд. руб.* (X_3)
Первая группа	170,5 200,0 186,4 154,2	10,0 18,2 15,8 10,3	250,95 380,60 300,20 280,36
Вторая группа	60,6 90,8 100,4	9,0 9,7 8,3	100,50 147,60 194,30

Проверьте, можно ли отнести ко второй группе предприятие, у которого рассматриваемые переменные имеют значения: $x_1 = 98{,}8$; $x_2 = 11{,}5$; $x_3 = 146{,}0$.

8. По 15 предприятиям имеются следующие данные, характеризующие их производственно-хозяйственную деятельность:

Номер предприятия	Фондовооруженность труда, млн. руб./ чел. (X_1)	Фондоотдача активной части ОПФ, руб./ руб., (X_2)	Удельный вес рабочих в составе ППП (X_3)
1	4,80	1,94	0,73
2	4,85	1,85	0,71
3	3,75	1,59	0,79
4	6,35	1,24	0,71
5	8,70	0,75	0,68
6	7,30	1,15	0,72
7	6,40	1,26	0,70
8	5,90	1,43	0,75
9	6,00	1,28	0,63
10	8,95	0,95	0,76
11	7,41	1,18	0,69
12	4,71	1,90	0,71
13	5,03	1,81	0,72
14	6,94	1,29	0,73
15	7,85	0,98	0,70

Используя любой алгоритм кластерного анализа, сформируйте из первых десяти наблюдений две обучающие выборки. На основании полученных выборок проведите классификацию пяти оставшихся предприятий. Дайте экономическую интерпретацию результатов дискриминантного анализа.

9. Используя специальный статистический пакет «*DSTAT*» для ПЭВМ, выполните классификацию пяти промышленных объектов, характеризующихся четырьмя переменными:

X_1 — производительность труда работающего, млн. руб.,

X_2 — доля рабочих в общей численности работающих,

X_3 — рентабельность продукции, %,

X_4 — коэффициент использования сырья и материалов.

Исходные значения переменных приведены в таблице:

Номер объекта	X_1	X_2	X_3	X_4
1	8,4	0,62	7,5	81,5
2	9,1	0,78	10,0	94,0
3	5,5	0,73	6,1	74,0
4	4,3	0,65	6,0	70,8
5	9,7	0,70	11,0	92,5

Классификацию проведите при помощи дискриминантного анализа, используя в качестве обучающих выборок следующие группы объектов:

1 группа

Номер объекта	X_1	X_2	X_3	X_4
1	4,0	0,63	6,0	80,0
2	4,9	0,60	6,3	78,6
3	6,1	0,61	7,0	75,9
4	5,3	0,62	7,1	74,0
5	5,8	0,60	6,8	81,5

2 группа

Номер объекта	X_1	X_2	X_3	X_4
1	8,7	0,70	9,0	90,7
2	10,3	0,78	10,5	94,6
3	11,6	0,75	10,9	94,0
4	10,8	0,77	11,0	92,5

Поясните полученные результаты дискриминантного анализа, дайте экономическую интерпретацию выявленных различий между группами на основании расчета выборочных характеристик.

Глава 11

Метод канонических корреляций

11.1. Сущность и теоретические основы метода

Метод канонических корреляций относится к статистическим методам анализа связей между массовыми общественными явлениями и процессами. В экономико-статистических исследованиях часто возникает необходимость выявить на основании эмпирических данных зависимость основных результативных показателей производственно-хозяйственной деятельности от большого числа факторов, определяющих уровень этих показателей. Если рассматривается зависимость между одним результативным показателем Y и одним фактором X, то речь идет о парной корреляции. Когда имеется несколько переменных X и одна переменная Y, проводится множественный корреляционный анализ для установления и измерения степени связи между переменными. Каноническая корреляция — это распространение парной корреляции на случай, когда имеется несколько результативных показателей Y и несколько факторов X.

Например, эффективность работы предприятий оценивается такими результативными показателями, как производительность труда, фондоотдача основных фондов, прибыль, рентабельность и др. В этом случае факторами являются следующие показатели: численность работающих, стоимость основных фондов, оборачиваемость оборотных средств, удельный вес потерь от брака, трудоемкость единицы продукции, коэффициент сменности работы оборудования и т. п. Метод канонических корреляций дает возможность одновременно анализировать взаимосвязь нескольких выходных показателей и большого числа определяющих факторов. При этом не требуется отсутствия корреляции как в группе результативных показателей, так и в группе факторных. Алгоритм расчетов метода канонических корреляций строится таким образом, что исходные переменные заменяются их линейными комбинациями, которые являются линейно независимыми. В то же время обеспечивается высокая степень связи между линейными комбинациями факторов и линейными комбинациями исследуемых выходных показателей.

Основная цель применения этого метода в экономическом анализе состоит прежде всего в поиске максимальных корреляционных связей между группами исходных переменных: показателями-факторами и результативными качественными показателями. Кроме того, метод канонических корреляций дает возможность сократить объем исходных данных за счет отсева малозначимых факторов.

В каноническом анализе матрица значений исходных переменных разбита на две части:

Номер наблюдения	X_1	X_2	....	X_g	Y_1	Y_2...	Y_p
1	x_{11}	x_{12}	....	x_{1g}	y_{11}	y_{12}...	y_{1p}
2	x_{21}	x_{22}	....	x_{2g}	y_{21}	y_{22}...	y_{2p}
3	x_{31}	x_{32}	....	x_{3g}	y_{31}	y_{32}...	y_{3p}
...	...	...	...	...	...	...	...
n	x_{n1}	x_{n2}	....	x_{ng}	y_{n1}	y_{n2}...	y_{np}

$X_1, X_2, \ldots, Xg$ — переменные факторы;

$Y_1, Y_2, \ldots, Y_p$ — результативные показатели.

Так как на практике количество факторов значительно превосходит количество результативных показателей, то будем предполагать, что $p \leq g$.

Каноническая корреляция — это корреляция между новыми компонентами (каноническими переменными) U и V:

$$U = a_1x_1 + a_2x_2 + \ldots + a_gx_g,$$
$$V = b_1y_1 + b_2y_2 + \ldots + b_py_p. \qquad (11.1)$$

По аналогии с парной корреляцией теснота связи между каноническими переменными будет определяться каноническим коэффициентом корреляции r:

$$r = \frac{\mathrm{cov}(U,V)}{\sqrt{\mathrm{var}(U) * \mathrm{var}(V)}}.$$

В зависимости от того, какие значения принимают коэффициенты a_i и b_j ($i = 1, \ldots, g$, $j = 1, \ldots, p$), будут изменяться значения канонических переменных и канонический коэффициент корреляции. Одна из основных задач, решаемых в ходе анализа канонических корреляций, заключается в отыскании такой пары значений канонических переменных, которой соответствует максимальное значение канонического коэффициента корреляции.

11.2. Подготовка информации и вычисления канонических корреляций

Для вычисления канонических коэффициентов корреляции необходимо прежде всего определить матрицы ковариаций исходных переменных. Запишем расширенную матрицу ковариаций для обеих групп переменных:

$$S = \begin{pmatrix} \sigma_{x_1x_1} & \sigma_{x_1x_2} & \ldots & \sigma_{x_1x_g} & \sigma_{x_1y_1} & \ldots \sigma_{x_1y_p} \\ \sigma_{x_2x_1} & \sigma_{x_2x_2} & \ldots & \sigma_{x_2x_g} & \sigma_{x_2y_1} & \ldots \sigma_{x_2y_p} \\ \sigma_{x_3x_1} & \sigma_{x_3x_3} & \ldots & \sigma_{x_3x_g} & \sigma_{x_3y_1} & \ldots \sigma_{x_3y_p} \\ \sigma_{y_1x_1} & \sigma_{y_1x_2} & \ldots & \sigma_{y_1x_g} & \sigma_{y_1y_1} & \ldots \sigma_{y_1y_p} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \sigma_{y_px_1} & \sigma_{y_px_2} & \ldots & \sigma_{y_px_g} & \sigma_{y_py_1} & \ldots \sigma_{y_py_p} \end{pmatrix}.$$

Матрица S фактически разделена на четыре части, которые можно обозначить следующим образом:

$$S = \left(\begin{array}{c|c} S_{11} & S_{12} \\ \hline S_{21} & S_{22} \end{array}\right),$$

где S_{11} — ковариационная матрица исходных переменных X_1, X_2, ..., X_g размерности $g \times g$,

S_{22} — ковариационная матрица исходных переменных Y_1, Y_2, ..., Y_p размерности $p \times p$,

S_{12}, S_{21} — ковариационные матрицы исходных переменных X_1, X_2, ..., X_g и Y_1, Y_2, ..., Y_p размерности соответственно $g \times p$ и $p \times g$.

Матрица S_{21} представляет собой результат транспортирования матрицы S_{12}. В матрицах S_{11} и S_{22} элементы, расположенные на главной диагонали, являются дисперсиями соответствующих переменных. Например, величина $\sigma_{x_1x_1}$ — дисперсия переменной X_1:

$$\sigma_{x_1x_1} = \frac{\sum_{i=1}^{n}(x_{1i} - \overline{x_1})(x_{1i} - \overline{x_1})}{n} = \frac{\sum_{i=1}^{n} x_{1i} - \overline{x_1})^2}{n} = \mathrm{var}X_1.$$

Все остальные элементы матрицы S представляют собой величины ковариаций пар переменных. Например, $\sigma_{x_1y_1}$ — ковариация переменных X_1 и Y_1, определяемая по формуле:

$$\sigma_{x_1y_1} = \frac{\sum_{i=1}^{n}(x_{1i} - \overline{x_1})(y_{1i} - \overline{y_1})}{n} = \mathrm{cov}(X_1, Y_1).$$

Представим выражения (11.1) в матричном виде:

$$U = X \times A,$$
$$V = Y \times B, \qquad (11.2)$$

где U, V — векторы канонических переменных,

X, Y — матрицы исходных значений переменных,

A, B — векторы коэффициентов.

Если предположить, что средние значения канонических переменных U и V равны нулю, а их дисперсии равны единице, тогда формулу для расчета канонических коэффициентов корреляции можно записать следующим образом:

$$r = \frac{\mathrm{cov}(U,V)}{\sqrt{\mathrm{var}(U) \cdot \mathrm{var}(V)}} = \frac{\mathrm{cov}(XA,YB)}{\sqrt{\mathrm{var}(XA) \cdot \mathrm{var}(YB)}} = \frac{A'S_{12}B}{\sqrt{A'S_{11}AB'S_{22}B}}. \qquad (11.3)$$

Для того чтобы упростить процедуру нахождения оптимальных коэффициентов канонических переменных, предположим, что каждая из этих переменных имеет единичную дисперсию и нулевое математическое ожидание. Тогда можно записать:

$$\mathrm{var}(XA) = A'S_{11}A = 1 \text{ и } \mathrm{var}(YB) = B'S_{22}B = 1.$$

Выражение (11.3) примет вид: $r = A'S_{12}B$. Чтобы найти максимальный канонический коэффициент корреляции, воспользуемся способом множителей Лагранжа для нахождения условного экстремума. Продифференцируем функцию Лагранжа по

компонентам векторов A и B и приравняем их нулю. Получаем два выражения:

$$S_{12}B-\lambda S_{11}A=0, \quad (11.4)$$

$$S_{21}A-\lambda S_{22}B=0, \quad (11.5)$$

где λ — множитель Лагранжа.

Умножим выражение (11.4) на λ, а выражение (11.5) — на обратную матрицу:

$$S_{12}\lambda B=\lambda^2 S_{11}A,$$

$$S_{22}^{-1}S_{21}A=\lambda B,$$

откуда получаем:

$$S_{12}S_{22}^{-1}S_{21}A=\lambda^2 S_{11}.$$

Если умножить обе части равенства на S_{11}^{-1}, получаем выражение:

$$S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}A=\lambda^2 A \text{ или } S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}-\lambda^2 E=0. \quad (11.6)$$

Аналогичные преобразования приводят еще к одному выражению:

$$(S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}-\lambda^2 E)B=0. \quad (11.7)$$

Чтобы решить эти уравнения, необходимо найти характеристические корни и характеристические векторы. Из предположения, что $p \le g$, вытекает, что размерность вектора B меньше размерности вектора A. Следовательно, будем находить характеристические корни и векторы уравнения: $(S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}-\lambda^2 E)B$.

Из выражения (11.4) определим вектор A:

$$A=\frac{S_{11}^{-1}S_{12}B}{\lambda}. \quad (11.8)$$

Для того чтобы найти компоненты вектора A, необходимо определить векторы B и λ.

Значения λ^2 находим как собственные значения матрицы $C=S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$. Размерность этой матрицы равна $(p \times p)(p \times q)(q \times q)(q \times p)$. Значит, можно найти p собственных значений λ^2 и p соответствующих им собственных векторов B. Подставляя поочередно их значения в выражение (11.8), определяем вектор A.

Покажем, что $\lambda = r$. Для этого воспользуемся преобразованием выражений (11.4) и (11.5). Умножим выражение (11.4) на

транспонированный вектор A', а уравнение (11.5) — на транспонированный вектор B':

$$A'S_{12}B=\lambda A'S_{11}A,$$

$$B'S_{21}A=\lambda B'S_{22}B.$$

Так как $A'S_{11}A=B'S_{22}B=1$, получаем $A'S_{12}B=B'S_{21}A$, т.е. $r=B'S_{21}A=\lambda$. Следовательно, если ранжировать собственные значения λ^2 так, что $\lambda_1^2 \ge \lambda_2^2 \ge \ldots \lambda_p^2$, то λ_1^2 будет соответствовать максимальный канонический коэффициент корреляции. При определении вектора A и канонических коэффициентов корреляции знак при λ выбирается исходя из экономического содержания анализируемых множеств X и Y.

Канонические коэффициенты корреляции можно рассчитывать и исходя из выборочной корреляционной матрицы R:

$$R=\left(\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right).$$

При этом все рассуждения и выкладки аналогичны тем, которые приведены выше для ковариационной матрицы. Коэффициенты a_i (i = 1, ..., g) и b_j (j = 1, ..., p) для канонических переменных U и V будут иметь в этом случае различные значения. Если a_i и b_j были вычислены по ковариационной матрице S, то они относятся к исходным переменным — x_i и y_j. Если расчет коэффициентов a_i и b_j базировался на корреляционной матрице R, то они относятся к стандартизованным значениям исходных переменных x_i' и y_j':

$$x_i'=\frac{x_i-\bar{x}}{\sigma_{x_i}}; \quad y_j'=\frac{y_j-\bar{y}}{\sigma_{y_j}}.$$

Пример 11.1. Вычислить канонические коэффициенты корреляции для двух групп переменных (табл. 11.1):

Y_1 — производительность труда, млн. руб./чел.;

Y_2 — уровень рентабельности, %;

X_1 — трудоемкость единицы продукции, чел.-час.;

X_2 — оборачиваемость оборотных средств, дней;

X_3 — фонд оплаты труда промышленно-производственного персонала, млн. руб.

Таблица 11.1. **Показатели производственно-хозяйственной деятельности промышленных предприятий**

Номер предприятия	X_1	X_2	X_3	Y_1	Y_2
1	0,45	170	1860	10,1	23,0
2	0,21	185	1455	8,6	13,2
3	0,18	160	1290	9,5	11,0
4	0,38	175	1710	9,0	9,4
5	0,35	140	1850	7,6	9,2
6	0,50	105	1630	11,5	10,0
7	0,32	90	1935	12,0	19,5
8	0,54	134	1795	6,8	9,0
9	0,47	98	2800	8,5	12,0
10	0,38	100	1635	9,4	10,6

Матрица парных коэффициентов корреляции:

$$R = \begin{pmatrix} 1 & -0,409 & 0,520 & -0,108 & -0,060 \\ -0,409 & 1 & -0,488 & -0,351 & 0,047 \\ 0,520 & -0,488 & 1 & -0,117 & 0,143 \\ -0,108 & -0,351 & -0,117 & 1 & 0,530 \\ -0,060 & 0,047 & 0,143 & 0,530 & 1 \end{pmatrix}.$$

Вспомогательные матрицы R_{11}^{-1} и R_{22}^{-1} соответственно:

$$R_{11}^{-1} = \begin{pmatrix} 1,4375 & 0,2929 & -0,6045 \\ 0,2929 & 1,3766 & 0,5195 \\ -0,6045 & 0,5195 & 1,5712 \end{pmatrix}, \quad R_{22}^{-1} = \begin{pmatrix} 1,3890 & -0,7360 \\ -0,7360 & 1,3890 \end{pmatrix}.$$

Необходимо вычислить собственные значения (характеристические корни) и собственные векторы для матрицы $C = R_{22}^{-1} \cdot R_{21} \cdot R_{11}^{-1} \cdot R_{12}$. Ее размерность в нашем примере равна $(2 \times 2)(2 \times 3)(3 \times 3)(3 \times 2) = (2 \times 2)$, следовательно, она будет иметь два собственных значения и два собственных вектора:

$$C = R_{22}^{-1} \cdot R_{21} \cdot R_{12}^{-1} \cdot R_{12} = \begin{pmatrix} 0,4009 & -0,1299 \\ -0,2712 & 0,1212 \end{pmatrix}.$$

Для нахождения собственных значений нужно составить характеристический многочлен $\left|C - \lambda^2 \cdot E\right| = 0$ для матрицы C и найти его корни:

Корень	*Характеристический вектор*	
$\lambda_1^2 = 0,491$	−1,1470	0,8411
$\lambda_2^2 = 0,031$	0,2746	0,8269

Первый канонический коэффициент корреляции $r_1 = 0,701$, а второй — $r_2 = 0,176$. Чтобы найти вектор коэффициентов A, подставляем соответствующие значения в выражение (11.8). Первый вектор A_1, который соответствует первому каноническому коэффициенту, равен:

$$A_1 = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix} = \frac{1}{0,49} \begin{pmatrix} 1,4375 & 0,2929 & -0,6045 \\ 0,2929 & 1,3766 & 0,5195 \\ -0,6045 & 0,5195 & 1,5712 \end{pmatrix} \begin{pmatrix} -0,108 & -0,060 \\ -0,351 & 0,047 \\ -0,117 & 0,143 \end{pmatrix} \begin{pmatrix} -1,147 \\ 0,8411 \end{pmatrix},$$

$$A_1 = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix} = \begin{pmatrix} 0,1132 \\ 1,085 \\ 0,8345 \end{pmatrix}.$$

Аналогично находим второй вектор:

$$A_2 = \begin{pmatrix} a_{21} \\ a_{22} \\ a_{23} \end{pmatrix} = \begin{pmatrix} -1,0370 \\ -0,3291 \\ 0,8674 \end{pmatrix}.$$

Итак, максимальный коэффициент канонической корреляции равен 0,701. Ему соответствуют канонические переменные:

$$U_1 = 0,1132X_1' + 1,085X_2' + 0,8345X_3', \qquad (11.9)$$
$$V_1 = -1,147Y_1' + 0,8411Y_2'.$$

Второму коэффициенту канонической корреляции $r_2 = 0,176$ соответствует следующая пара канонических переменных:

$$U_1' = -1,037X_1' - 0,329X_2' + 0,8674X_3', \qquad (11.10)$$
$$V_1' = 0,2746Y_1' + 0,8269Y_2'.$$

Так как векторы коэффициентов канонических переменных вычислены на основании матрицы парных корреляций, то необ-

ходимо помнить, что они относятся к стандартизованным значениям (исходных) переменных.

11.3. Оценка значимости канонических корреляций

Прежде чем интерпретировать с экономической точки зрения полученные результаты, рассмотрим вопросы проверки значимости полученных коэффициентов канонической корреляции.

Если две группы исходных переменных (X_1, X_2, ..., X_g) и (Y_1, Y_2, ..., Y_p) независимы, то элементы корреляционной (ковариационной) матрицы будут равны нулю и все коэффициенты канонических корреляций будут тоже нулевыми.

Для проверки значимости коэффициентов канонической корреляции используется критерий Бартлетта. Если обозначить как W_0 произведение p множителей $(1 - \lambda_i^2)$, то проверить нулевую гипотезу о том, что множество переменных X_1, X_2, ..., X_g не коррелирует с множеством переменных Y_1, Y_2, ..., Y_p, можно при помощи χ^2-критерия:

$$\chi^2 = -\left[n-1-\frac{1}{2}(p+q+1)\right]\ln W_0 \qquad (11.11)$$

с числом степеней свободы, равным $p \times g$. Если расчетное χ^2 больше табличного значения критерия при выбранном уровне значимости, то можно утверждать, что по крайней мере $r_1 = \sqrt{\lambda_1^2}$ будет отличен от нуля. Чтобы проверить значимость второго коэффициента канонической корреляции, необходимо рассчитать величину $W_1 = \prod_{i=1}^{p}\left(1-\lambda_i^2\right)$ и определить χ^2 по формуле:

$$\chi^2 = -\left[n-2-\frac{1}{2}(p+q+1)\right]\ln W_1 \text{ для } (p-1)(g-1) \text{ степеней свободы.}$$

Если для первых m коэффициентов нулевая гипотеза подтверждается, т.е. они равны нулю, тогда значимость оставшихся $p-m$ коэффициентов проверяется аналогично при помощи статистики $W_k = \prod_{i=k+1}^{p}\left(1-\lambda_i^2\right)$, распределенной как величина:

$$\chi^2 = -\left[n-1-k-\frac{1}{2}(p+q+1)\right]\ln W_k \text{ для } (p-k)(q-k) \text{ степеней свободы.}$$

Воспользовавшись критерием Бартлетта, проверим значимость двух коэффициентов канонической корреляции в нашем примере. Нулевая гипотеза заключается в том, что $r_1 = r_2 = 0$.

Определим

$$W_0 = (1-\lambda_1)(1-\lambda_2) = (1-r_1^2)(1-r_2^2) = (1-0{,}491)(1-0{,}031) = 0{,}493221.$$

Тогда $\chi^2 = -\left[10-1-\frac{1}{2}(2+3+1)\right]\ln 0{,}493221 = 4{,}24$.

Находим табличное значение χ^2 для числа степеней свободы, равного $(p \times g) = 6$ и уровня значимости $\alpha = 0{,}05$:

$\chi^2_{\text{табл}} = 1{,}635 < \chi^2_{\text{рас}} = 4{,}24$,

следовательно, нулевая гипотеза отвергается, т.е. первый коэффициент отличен от нуля.

Проверим существенность второго коэффициента:

$$W_1 = 1 - r_2^2 = 1 - 0{,}031 = 0{,}969,$$

$$\chi^2 = -\left[10-2-\frac{1}{2}(3+2+1)\right]\ln 0{,}969 = 0{,}157.$$

Так как $\chi^2_{\text{рас}} > \chi^2_{\text{табл}} = 0{,}103$ для числа степеней свободы, равного $(p-1)(q-1) = 2$, и уровня значимости $\alpha = 0{,}05$, следовательно, второй коэффициент канонической корреляции также отличен от нуля. В нашем примере коэффициенты канонических переменных вычислены по матрице парных коэффициентов корреляции R, следовательно, они относятся к стандартизованным переменным. Для того чтобы получить коэффициенты, относящиеся к исходным переменным, нужно выполнить следующие преобразования [17, с. 245]:

$a_1' = a_1/\sigma_{x1} = 0{,}1132/0{,}1184905 = 0{,}9553,$

$a_2' = a_2/\sigma_{x2} = 1{,}085/35{,}73063 = 0{,}0304,$

$a_3' = a_3/\sigma_{x3} = 0{,}8345/403{,}8069 = 0{,}0021,$

$b_1' = b_1/\sigma_{y1} = -1{,}147/1{,}604854 = -0{,}7147,$

$b_2' = b_2/\sigma_{y2} = 0{,}8411/4{,}766655 = 0{,}1764,$

где a_1, a_2, a_3 и b_1, b_2 — коэффициенты канонических переменных, относящихся к соответствующим значениям исходных переменных.

Тогда выражение (11.9) примет вид:

$$U_1 = 0{,}9553X_1 + 0{,}0304X_2 + 0{,}0021X_3,$$
$$V_1 = -0{,}7147Y_1 + 0{,}1764Y_2. \qquad (11.12)$$

При необходимости, используя выражение (11.12), можно рассчитать значения канонических переменных для каждого объекта наблюдения, для этого в выражение (10.9) следует подставить соответствующие значения исходных переменных.

11.4. Экономическая интерпретация результатов канонического анализа

Использование метода канонических корреляций в экономических исследованиях прежде всего предполагает возможность содержательной интерпретации полученных результатов. В противном случае теряет смысл применение этого метода в исследовании причинно-следственной связи (ее тесноты и формы) массовых экономических явлений. В начале главы уже сказано о возможности применения метода канонических корреляций при исследовании взаимосвязи большого числа факторных признаков и большого числа результативных показателей. Теперь попытаемся разобраться в тех результатах расчетов, которые получены при решении условного примера в § 11.2. Максимальный коэффициент канонической корреляции $r_1 = 0{,}701$ достигается в том случае, когда исходные стандартизованные переменные образуют следующую пару канонических переменных (11.9):

$$U_1 = 0{,}1132X_1' + 1{,}085X_2' + 0{,}8345X_3',$$
$$V_1 = -1{,}147Y_1' + 0{,}8411Y_2'.$$

Так как величина r_1 близка к единице, значит связь между полученными линейными комбинациями исходных переменных тесная, т. е. на уровни производительности труда и рентабельности существенное влияние оказывают выбранные факторы. Если судить по матрице R, то в обеих группах исходных переменных есть показатели, достаточно тесно связанные друг с другом. Например, Y_1 и Y_2 ($r = 0{,}53$), а также X_1 и X_3 ($r = 0{,}52$).

Второй по величине коэффициент канонической корреляции $r_2 = 0{,}176$ говорит о том, что любые другие линейные комбинации отобранных факторов и результативных показателей слабо связаны друг с другом. В то же время $r_1 = 0{,}701$ является

максимальным для данного набора исходных переменных. Следовательно, не исключена возможность подобрать другой по количеству и по составу набор факторов, который будет иметь очень тесную связь с показателями производительности труда и рентабельностью.

Коэффициенты в канонических переменных характеризуют силу влияния соответствующих признаков-факторов и результативных показателей на уровень связи между ними.

Рассматривая канонические переменные U_1 и U_2, можно заметить, что в обеих линейных комбинациях сохраняется наибольший коэффициент при X_3 (фонд оплаты труда промышленно-производственного персонала), коэффициенты при X_1 и X_2 не только меняют знак, но и существенно изменяются по величине. Из этого можно сделать вывод, что X_3 при любых линейных комбинациях несет примерно одинаковую информацию об уровнях результативных показателей.

С целью расширения исследования взаимосвязи экономических показателей рекомендуется повторить расчеты коэффициентов канонической корреляции и канонических переменных для других сочетаний факторных и результативных переменных. В таком случае удается проследить изменение величины первого коэффициента канонической корреляции, относительного и абсолютного влияния факторов на выходные показатели.

На каждом шаге отбрасывается одна переменная, которой соответствует наименьший коэффициент в канонической переменной. Для сокращенного набора опять рассчитываются коэффициенты канонических корреляций. Если максимальные коэффициенты для исходного и сокращенного наборов переменных различаются несущественно, то процесс сокращения продолжается. Оценка значимости расхождения максимальных коэффициентов корреляции осуществляется при помощи следующего критерия:

$$t_{\text{набл}} = (Z_k - Z_{k+1})(n-3)/2,$$

где Z_k — величина, исчисленная для исходного набора переменных,

Z_{k+1} — для сокращенного набора переменных;

n — число наблюдений,

$$Z_k = 0{,}5 \ln \frac{1+r}{1-r}.$$

Путем последовательного исключения или включения очередного фактора можно проследить, в каких случаях информативность факторного набора существенно меняется, т.е. существенно возрастает или снижается величина максимального коэффициента канонической корреляции.

Выводы

Метод канонических корреляций расширяет возможность исследования взаимосвязей отдельных сторон производственно-хозяйственной и финансовой деятельности предприятий или отдельных отраслей в результате вовлечения в процесс анализа сразу нескольких результативных показателей.

В отличие от множественного корреляционно-регрессионного анализа здесь нет необходимости добиваться независимости исходных переменных, участвующих в анализе, так как их линейные комбинации строятся таким образом, что они являются центрированными, нормированными и некоррелированными внутри множеств. Как и корреляционно-регрессионный анализ, метод канонических корреляций не дает возможности устанавливать причинно-следственный характер связи. Исследователь должен заранее определить, какие переменные являются факторами, а какие — результативными показателями.

По сравнению с другими многомерными методами (метод главных компонент, факторный анализ) результаты метода канонических корреляций легче интерпретируются.

Анализ структуры канонических переменных и величины канонических корреляций позволяет осуществлять отбор наиболее информативных переменных по характеристике тесноты связи между двумя множествами переменных и содержанию процесса.

Критерием оценки существенности или несущественности отбрасываемого признака на каждом шаге служит изменение величины канонической корреляции. Канонические корреляции по своей сути имеют много общего с парными коэффициентами корреляций Пирсона. В то же время знак канонического коэффициента корреляции не свидетельствует о направлении связи между каноническими переменными. В зависимости от экономического смысла изучаемых совокупностей можно выбрать положительный или отрицательный знак для канонической корреляции.

Вопросы и задачи

1. В чем состоит основное отличие метода канонических корреляций от многомерного корреляционно-регрессионного анализа?

2. Какие задачи решаются при помощи метода канонических корреляций?

3. Что такое канонические переменные?

4. Какие ограничения налагаются алгоритмом канонического анализа на количество факторных и результативных переменных?

5. Как проверить значимость коэффициентов канонической корреляции?

6. При помощи каких приемов канонического анализа можно сократить количество исследуемых факторов?

7. На основании приведенных ниже данных рассчитать матрицу парных коэффициентов корреляции для проведения канонического анализа:

Номер объекта	*Множество 1*		*Множество 2*		
	Y_1	Y_2	X_1	X_2	X_3
1	23,0	19,0	570	19,4	6,7
2	25,0	12,0	860	21,6	13,3
3	25,2	13,4	1150	28,8	12,2
4	21,0	9,2	610	20,5	8,4
5	24,5	14,0	502	23,3	6,8
6	18,5	11,5	670	17,8	9,2
7	18,8	10,9	540	20,0	7,0

X_1 — число работающих, тыс. чел.
X_2 — фондовооруженность труда одного работника, млн. руб.
X_3 — электровооруженность труда одного рабочего, тыс. кВт-час.
Y_1 — производительность труда одного работающего, млн. руб.
Y_2 — рентабельность производства, %

8. На основании приведенной матрицы парных коэффициентов корреляции определить степень зависимости между двумя группами признаков:

$$R = \begin{bmatrix} 1 & -0{,}3 & 0{,}5 & 0{,}7 & 0{,}2 \\ & 1 & 0{,}4 & -0{,}6 & 0{,}8 \\ & & 1 & 0{,}2 & 0{,}5 \\ & & & 1 & 0{,}3 \\ & & & & 1 \end{bmatrix}.$$

Примечание. В первую группу признаков входят переменные X_1, X_2, X_3, во вторую — переменные Y_1, Y_2.

9. Динамика цен по различным отраслям производства и товаров в отчетном периоде по сравнению с базисным характеризуется следующими индексами:

Месяц	Индексы производства, %			Индексы потребительских цен, %	Вклады населения, % (предыдущий месяц = 100)
	Промышленность	Сельское хозяйство	Транспорт		
	X_1	X_2	X_3	Y_1	Y_2
Январь	98,4	99,5	101,8	140,7	105,2
Февраль	96,0	98,0	105,0	118,7	119,2
Март	100,5	106,0	112,0	110,2	130,6
Апрель	104,0	113,0	124,0	128,6	129,6
Май	108,0	124,0	126,0	128,7	118,1
Июнь	110,0	108,6	124,0	119,5	118,5
Июль	101,0	115,0	120,5	126,6	173,5
Август	99,0	122,8	115,0	153,4	132,5
Сентябрь	109,0	135,5	110,3	125,5	104,4
Октябрь	115,0	130,0	107,0	125,7	146,4
Ноябрь	120,0	121,8	100,5	140,5	115,1
Декабрь	121,0	101,3	99,8	131,3	106,9

Рассчитать первый канонический коэффициент корреляции и проверить его значимость. Все необходимые вычисления выполнить на ПЭВМ.

10. На основании результатов расчетов, выполненных в п. **3**, определить, какая из линейных комбинаций индексов позволяет

наилучшим образом предсказать сводный индекс результативных показателей (изменение потребительских цен и вкладов населения). Попытайтесь это сделать.

11. На основании данных из годовых отчетов предприятий, приведенных в табл. 11.1 для двух множеств (X_1, X_2, X_3) и (Y_1, Y_2), определить канонические переменные, соответствующие максимальному коэффициенту канонической корреляции, и дать оценку существенности изменения максимального коэффициента корреляции после удаления наименее значимого факторного признака. Расчеты выполнить на ПЭВМ.

Номер предприятия	Y_1	Y_2	X_1	X_2	X_3
1	40,5	0,98	4,5	0,15	10,0
2	51,5	1,15	7,6	0,10	8,0
3	58,0	1,35	8,0	0,13	8,2
4	54,0	0,95	5,1	0,24	15,0
5	40,8	1,25	7,3	0,09	10,1
6	39,6	1,30	8,5	0,17	9,3
7	62,0	2,10	8,1	0,07	7,5
8	47,5	1,50	7,8	0,11	8,2

Y_1 — уровень производительности труда, млн. руб.;
Y_2 — фондоотдача активной части основных фондов, руб.;
X_1 — коэффициент обновления основных фондов, %;
X_2 — удельный вес потерь от брака, %
X_3 — коэффициент оборачиваемости оборотных средств, дней.

12. В результате канонического анализа получены следующие значения коэффициентов канонической корреляции: $r_1 = 0{,}971$, $r_2 = 0{,}663$.

Воспользовавшись критерием Бартлетта, проверить значимость двух коэффициентов и второго коэффициента при уровне значимости $\alpha = 0{,}05$. Известно, что число факторов, участвующих в анализе, равно 5, а число результативных показателей равно 3.

Глава 12

Компьютерный анализ многомерных статистических данных

12.1. Характеристика и особенности построения пакетов STATGRAPHICS и DSTAT

Процесс совершенствования системы экономического управления неизбежно приводит к постоянному расширению области применения экономико-математических и статистических методов анализа данных. Существенным сдерживающим фактором в этом направлении является большая вычислительная трудоемкость аналитических методов, в частности, всех методов многомерного статистического анализа, которые описаны в предыдущих главах. В связи с этим большое значение приобретают вопросы использования ЭВМ и разработки соответствующего программного обеспечения для проведения анализа экономической информации. Появление быстродействующих персональных компьютеров с большими объемами памяти послужило базой для создания сложных программ статистического анализа. В настоящее время существует несколько десятков наиболее известных статистических пакетов, разработанных ведущими компаниями мира. Эти пакеты широко применяются в органах государственной статистики, органах управления, производственных и научно-исследовательских организациях различных стран для сбора, обработки и анализа больших объемов данных.

К статистическим пакетам программ, или, как сейчас принято называть, программным продуктам относят как пакеты программ, специализированные в определенной области (например, обработка динамических рядов, реализация методов кластерного анализа и др.), так и универсального назначения, т.е. позволяющие осуществлять комплексный статистический анализ

(управление данными, расчет выборочных характеристик исходных переменных, анализ аномальных наблюдений, расчет устойчивых средних оценок, классификация, графическое представление структуры совокупности моделирования связей, прогнозирование и другие функции). Кроме того, отдельные процедуры статистического анализа включаются в современные интегрированные пакеты. Это прежде всего построение некоторых видов статистических графиков и таблиц, а также анализ данных (например, классификация объектов, регрессионный анализ и т.д.).

Тип программного обеспечения, применяемого в экономико-статистическом анализе, зависит от возможностей используемых ЭВМ, от предполагаемой области применения и от уровня профессиональной подготовленности пользователя. Поэтому не существует универсальной методики, по которой можно было бы дать сводную оценку качества статистического пакета. Однако с точки зрения пользователя можно выделить ряд наиболее важных характеристик статистических пакетов.

Прежде всего это наличие удобных средств для быстрого доступа пользователя к ресурсам программ, понятные названия процедур, возможность обработки неограниченного количества переменных и наблюдений, возможность прерывания и восстановления работы пакета, широкий набор процедур управления данными, а также совместимость с другими программными средствами.

Скорость восприятия пользователем выходной информации зависит от ее размещения на экране или в протоколе работы, выдаваемом на принтер, поэтому названия показателей, их условное обозначение, заголовки таблиц, графиков, должны быть общепринятые в статистике.

Несмотря на то, что пакеты программ статистического анализа относятся к наиболее важной и наиболее развитой категории программных продуктов, не все из них содержат качественные программы, реализующие методы многомерного статистического анализа. Как правило, из методов многомерного анализа в пакетах чаще всего встречаются корреляционный и регрессионный анализ, наиболее популярные алгоритмы иерархического кластерного анализа, дисперсионный анализ, а также некоторые методы факторного анализа. Гораздо реже встречаются методы многомерного шкалирования, метод канонических корреляций, а также итеративные методы кластерного анализа.

По оценке независимых экспертов, которые являются квалифицированными пользователями нескольких пакетов [85], можно сделать вывод, что в настоящее время наиболее популярны статистические пакеты SPSS, SAS, STATGRAPHICS.

Поскольку авторы не ставили цель дать исчерпывающий обзор всех современных статистических пакетов, в данной главе описываются возможности двух версий одного из лучших современных пакетов — STATGRAPHICS FOR DOS и STATGRAPHICS FOR WINDOWS, а также пакета DSTAT.

Статистический пакет STATGRAPHICS (Statistical Graphics System) — один из лучших пакетов прикладных программ для проведения статистического анализа на ПЭВМ. Существует несколько версий этого программного продукта; среди них STATGRAPHICS FOR DOS (версия 3.1., версия 7.0), STATGRAPHICS FOR WINDOWS (версия 1.0.), STATGRAPHICS PLUS FOR WINDOWS (версия 1.0.) и др.

Популярность данного пакета у пользователей объясняется несколькими факторами: дружественный пользовательский интерфейс, наличие современных инструментальных средств, широкие аналитические возможности, начиная с простой описательной статистики и кончая сложными статистическими методами анализа — регрессионным, факторным, дискриминантным. Кроме того, пакет STATGRAPHICS обеспечивает расширенную возможность графического отображения исходных данных и результатов анализа (итоговые графики).

Пакет STATGRAPHICS FOR DOS состоит из шести разделов, включающих 22 программы. Причем пакет для DOS обеспечивает более широкий набор функций статистического анализа по сравнению с версией для WINDOWS, но стоит дешевле. Пакет STATGRAPHICS обеспечивает выполнение базовых статистических функций, построение высококачественных графиков, реализацию широкого круга аналитических методов, но требует от пользователя хорошей подготовки в области статистического анализа и теории вероятностей.

Головное меню пакета STATGRAPHICS включает шесть разделов, структура которых показана на рис. 12.1.

Реализация на ПЭВМ любого из предложенных в меню методов анализа начинается с подготовки исходных данных. Для этой цели в пакете STATGRAPHICS имеется собственный встроенный редактор для ввода и преобразования данных, а также система управления базами данных. Различают два уровня

представления данных: файлы (files) и переменные (variables). Файл может содержать произвольное число переменных, каждая из которых имеет пять характеристик: имя (*name*), тип (*type*), ранг (*rang*), длина (*lenght*) и комментарий (*comment*).

Для ввода в базу данных новой переменной необходимо в главном меню выбрать в разделе ***Управление данными и используемые системы*** программу ***Управление данными*** *(Data management)* и нажать клавишу *Enter* или *F6.*

После появления на экране подменю выбрать программу ***Операции с файлами*** *(File Operations)* и нажать клавишу *Enter.*

Для расчета всех демонстрационных примеров создадим базу данных на основе исходной информации, приведенной в табл. 11.1

Управление данными и используемые системы	**Анализ временных рядов**
A. Управление данными	L. Прогноз
B. Системное обеспечение	M. Контроль качества
C. Составление отчетов и вызов графиков	N. Сглаживание
D. Использование графопостроителя	O. Анализ временных рядов
Графики и описательные статистики	**Дополнительные процедуры**
E. Построение графиков функций	P. Анализ категоризованных данных
F. Описательные методы	Q. Многомерные методы
G. Оценивание и проверка гипотез	R. Непараметрические методы
H. Функции распределения вероятностей	S. Планирование выборочных обследований
I. Исследовательский анализ данных	
Дисперсионный и регрессионный анализ	**Математика и используемые процедуры**
J. Дисперсионный анализ	U. Математические функции
K. Регрессионный анализ	V. Дополнительные операции

Рис. 12.1. ***Головное меню пакета*** STATGRAPHICS

File Operations

STATGRAPHICS file name: ***RBASE***

Operations:	*A.* Copy	*D.* Erase	*G.* Recode	*J.* Update
	B. Create	*E.* Join	*H.* Rename	
	C. Edit	*F.* Print	*I.* Split	

Desired operation: *J*

Рис. 12.2. Меню файловых операций

Из меню файловых операций (рис. 12.2) выбираем функцию ***B. Создание*** (*Create*), указываем имя нового файла ***RBASE***, который будет содержать все вновь вводимые переменные, и нажимаем клавишу *F6*. Последовательно вводим значения каждой из пяти переменных и записываем их в созданный файл.

FILE RBASE

X_1	X_2	X_3	Y_1	Y_2
10.45	170	1860	10.1	23.0
20.21	185	1455	8.6	13.2
30.18	160	1290	9.5	11.0
40.38	175	1710	9.0	9.4
50.35	140	1850	7.6	9.2
60.50	105	1630	11.5	10.0
70.32	90	1935	12.0	19.5
80.54	134	1795	6.8	9.0
90.47	98	2800	8.5	12.0
100.38	100	1635	9.4	10.6

Рис. 12.3. База данных, используемая для анализа

После создания базы данных, содержащей исходные значения переменных, можно приступать к проведению статистического анализа. Начнем с расчета выборочных характеристик исходных переменных и построения некоторых видов графиков, затем рассмотрим реализацию более сложных статистических методов в STATGRAPHICS. Для вычисления выборочных характеристик, или, как их еще называют, одномерных статистик, необходимо в главном меню выбрать программу ***Одномерные статистики*** *(Summary Statistics)* в разделе ***Описательные методы*** *(F. «Descriptive Methods»)* и заполнить пустые окна:

Data vectors — указать имена переменных в столбик или в строку (с разделителем «&»), для которых будут вычисляться одномерные статистики;

Statistics — из всех предполагаемых величин можно выбрать по желанию определенный набор, а буквенные коды остальных стереть в данном *окне*. Например, если в окне оставить буквы *AMMVSSK*, то для соответствующей переменной будут вычислены: среднее значение, медиана, мода, дисперсия, среднее квадратическое отклонение, коэффициент асимметрии, коэффициент эксцесса.

На рис.12.4 приведены результаты расчета для трех переменных, записанных в базе данных в файле ***RBASE***

Variable:	X_1	X_2	X_3
Sample size	10	10	10
Average	0.378	135.7	1796
Median	0.38	137	1752.5
Mode	0.38	105	1635
Geometric mean	0.35834	131.361	1760.65
Variance	0.01404	1276.68	163060
Standard deviation	0.118491	35.7306	403.807
Standard error	0.03747	11.299	127.695
Minimum	0.18	90	1290
Maximum	0.54	185	2800
Range	0.36	95	1510
Lower quartile	0.32	100	1630
Upper quartile	0.47	170	1860
Interquartile range	0.15	70	230
Skewness	−0.459323	0.0360443	1.7646
Standardized skewness	−0.592984	0.046533	2.27808
Kurtosis	−0.625419	−1.80574	4.66597
Standardized kurtosis	−0.403706	−1.1656	3.01187

Рис. 12.4. Результаты расчета выборочных характеристик для трех переменных

В данном примере были вычислены одномерные статистики для несгруппированных данных. Если нужно провести аналогичные вычисления для сгруппированных данных, то следует выбирать программу ***Статистики для сгруппированных данных***

(Codebook Procedure). Вычисленные характеристики играют важную роль в анализе статистической совокупности. Они позволяют оценить ее структуру и степень однородности изучаемых признаков. В рассматриваемом наборе данных три переменные (X_1, X_2, X_3) являются факторами, а две (Y_1, Y_2) — результативными показателями.

Пакет прикладных статистических программ DSTAT предназначен для обработки данных специалистами различных областей знаний непосредственно на рабочих местах. Для его функционирования необходимо наличие интерпретатора языка Бейсик в составе операционной системы. Загрузочный файл пакета — DSTAT. BAT. Так как пользователи могут иметь различную математическую подготовку и в связи с тем, что особенности решаемых задач не требуют одинаково сложных методов, в составе пакета выделены две части: предварительный анализ данных (***Статистик новичок***) и основной анализ (***Статистик эксперт***). Каждая часть состоит из самостоятельных блоков, реализующих отдельные функции обработки данных. Полная характеристика пакета дана в работе [1].

Основные методы многомерного статистического анализа, а также дисперсионный анализ и анализ временных рядов представлены во второй части пакета ***Статистик эксперт***. Главное меню этой части пакета выглядит следующим образом:

1. Выход
2. Ввод исходных данных
3. Первичная обработка
4. Корреляционный анализ
5. Регрессионный анализ
6. Многомерный статистический анализ
7. Дисперсионный анализ
8. Анализ временных рядов
9. Непараметрическая статистика

Рис. 12.5. ***Пакет*** DSTAT.
Главное меню раздела Статистик эксперт

Из предложенного набора функций рассмотрим те, о которых уже шла речь при описании пакета STATGRAPHICS. Это

позволяет читателю и пользователю пакетов сравнить их возможности и в определенной мере оценить достоинства и недостатки каждого пакета.

Функция ***Ввод исходных данных*** позволяет вводить исходную информацию с клавиатуры или с диска, осуществлять просмотр и корректировку данных, выводить на экран дисплея, а также сохранять данные в файле на диске.

Если пользователь предварительно работал с табличным процессором и созданный файл (таблицу) записал на диск, то в разделе ***Статистик эксперт*** предусмотрена обработка таких файлов (процедура ***Ввод данных, подготовленных табличным процессором***). Для иллюстрации работы режимов и функций пакета DSTAT будем пользоваться исходными данными табл. 11.1 (как и при описании пакета STATGRAPHICS).

После ввода исходных данных анализ целесообразно начинать с расчета выборочных характеристик. Для этого выбираем в главном меню функцию ***Первичная обработка***. На экране появляется меню этой функции (рис. 12.6):

**** ПЕРВИЧНАЯ ОБРАБОТКА ****

1. Выход
2. Анализ аномальных наблюдений
3. Сглаживание скользящими средними
4. Расчет выборочных характеристик
5. Построение графиков факторов
6. Выбор формы связи
7. Кластерный анализ
8. Стандартные распределения
9. Построение актуарной кривой

ВЫБИРАЙТЕ ФУНКЦИЮ? 4

Рис. 12.6

Из предложенного меню рассмотрим функцию ***Расчет выборочных характеристик***. Для каждой переменной будут рассчитаны следующие характеристики (рис.12.7):

ФАКТОР #

Минимум = .18 Максимум = .54 Размах = .36

Число точек = 10 Медиана = .38

Среднее = .378 Дисперсия = .01404 Вариация = 31.3467%

Среднеквадратическое отклонение = .1184905

Среднее абсолютное отклонение = .0904

Асимметрия = - .3307126 Эксцесс = −1.298374

Винзоризованные оценки

Среднее	Дисперсия	порядок
.377	1.155667E-02	1
.393	4.356676E-03	2
.396	2.293322E-03	3
.38	−1.324548E-08	4
.38	−1.324548E-08	5

Рис. 12.7. ***Результаты выполнения функции Расчет выборочных характеристик для переменной*** X_1

После расчета выборочных характеристик для каждой переменной можно приступать к анализу взаимосвязей между отдельными признаками при помощи специальных методов, рассмотренных в следующем параграфе.

12.2. Реализация методов многомерного статистического анализа в пакетах STATGRAPHICS и DSTAT

12.2.1. Регрессионный анализ

Метод регрессионного анализа представлен в той или иной мере практически в любом статистическом пакете. Это объясняется несколькими причинами. Во-первых, регрессионный анализ — один из наиболее наглядных и доступных методов, позволяющий хотя бы приблизительно судить о характере связи между переменными, входящими в регрессионную модель. Во-вторых, аналитический способ выражения связи имеет ряд преимуществ по сравнению с графическим и табличным способами, так как позволяет использовать полученную математическую модель (уравнение регрессии) для интерполяции и экстраполяции.

В статистическом анализе задача исследования формы связи очень важна, поскольку без ее решения невозможно глубокое понимание изучаемого явления и прогнозирование хода дальнейшего его развития.

В самом регрессионном анализе существует множество методов и алгоритмов его реализации. Основные понятия и подходы к решению регрессионных задач, которые изложены в гл. 6, дают достаточно полное представление об этом методе и позволяют исследователю в каждом конкретном случае выбрать наиболее подходящий вариант решения.

В статистических пакетах регрессионный анализ, как правило, представлен традиционным набором программ, позволяющим рассчитать парную или множественную регрессию, сделать оценку параметров регрессионной модели. В рассматриваемых пакетах STATGRAPHICS и DSTAT имеется возможность выполнить регрессионный анализ несколькими способами.

Методы регрессионного анализа в пакете STATGRAPHICS реализуются в разделе ***Регрессионный анализ*** *(K. REGRESSION ANALYSIS)* головного меню и включают пять пунктов:

- Простая регрессия (*SIMPLE REGRESSION*);
- Интерактивное отбрасывание (*INTERACTIVE OUTLIER REJECTION*);
- Множественная регрессия (*MULTIPLE REGRESSION*);
- Пошаговая регрессия (*STEPWISE SELECTION*);
- Гребневая регрессия (*RIDGE REGRESSION*);
- Нелинейная регрессия (*NONLINEAR REGRESSION*).

Работа программы ***Простая регрессия*** предполагает выполнение следующих действий:

1) выбрать в меню ***REGRESSION ANALYSIS*** и нажать клавишу *Enter*;
2) выбрать программу ***SIMPLE REGRESSION*** и нажать клавишу *Enter*;
3) заполнить появившиеся на экране поля:

DEPENDENT VARIABLE — задать имя зависимой переменной;

INDEPENDENT VARIABLE — задать имя независимой переменной;

MODEL — выбрать одну из следующих моделей, предлагаемых программой:

LINEAR — линейная $y = a + bx$;
MULTIPLICATIVE — мультипликативная $y = ax^b$;
EXPONENTIAL — экспоненциальная $y = e^{ax + b}$;
RECIPROCAL — обратная $1/y = a + bx$.

Кроме того, необходимо задать в процентах коэффициент доверия (поле *Confidence limits*) для определения доверительного интервала для среднего значения Y при заданном значении X.

Поле *Point labels* (*метки наблюдений*) следует заполнять только в том случае, когда требуется разметить точки при графическом изображении результатов регрессионного анализа. После заполнения всех названных полей необходимо нажать клавишу *F6* и на экране появятся результаты регрессионного анализа.

Следующая программа, которая далее подробно рассмотрена на примере 12.1, — ***Множественная линейная регрессия*** *(MULTIPLE REGRESSION)*. Она позволяет построить уравнение линейной регрессии зависимой переменной Y от нескольких признаков-факторов X_1, X_2,..., X_m, провести анализ коэффициентов регрессии и анализ остатков, а также вывести на экран диаграмму рассеивания, линию регрессии и другие графики.

Для начала работы программы ***Множественная линейная регрессия*** необходимо в главном меню ***REGRESSION ANALYSIS*** выбрать пункт ***MULTIPLE REGRESSION*** и нажать клавишу *Enter*. В появившемся на экране меню заполнить следующие окна:

DEPENDENT VARIABLE — задать имя зависимой переменной;

INDEPENDENT VARIABLES — задать имена независимых переменных (X_1, X_2,..., X_m);

WEIGHTS — задать имя переменной, содержащей веса наблюдений;

CONSTANT — это поле служит для выбора одной из двух моделей уравнения регрессии: со свободным членом (*Yes*) или без него (*No*);

VERTICAL BARS — заполняется с целью отобразить на графике теоретических (расчетных) значений y вертикальные отрезки остатков;

CONFIDENCE LIMITS — указывается в процентах коэффициент доверия для построения доверительного интервала для коэффициентов регрессии.

После заполнения указанных полей необходимо нажать клавишу *F6* и на экране появятся результаты регрессионного анализа.

Пример 12.1. Матрица исходных данных для проведения регрессионного анализа содержит значения четырех переменных (исходные данные взяты из табл. 11.1):

Y — производительность труда, млн. руб./чел.;

X_1 — трудоемкость единицы продукции, чел.-дней;

X_2 — оборачиваемость оборотных средств, дней;

X_3 — фонд оплаты труда промышленно-производственного персонала.

Необходимо построить уравнение регрессии, характеризующее зависимость уровня производительности труда Y от трех факторов X_1, X_2, X_3.

Результаты расчетов по программе ***Multiple Regression*** для данного набора переменных приведены на рис. 12.8.

Model fitting results for: Y_1

Independent variable	*Coefficient*	*Std. Error*	*t-value*	*sig. Level*
CONSTANT	$a_0 = 15.902939$	4.848966	3.2797	0.0168
x_1	$a_1 = -2.514677$	5.703773	−0.4409	0.6747
CONSTANT	$a_0 = 15.902939$	4.848966	3.2797	0.0168
x_1	$a_1 = -2.514677$	5.703773	−0.4409	0.6747
x_2	$a_2 = -0.025805$	0.018513	−1.3939	0.2128
x_3	$a_3 = -0.001197$	0.00175	−0.6841	0.5194

R-SQ.(ADJ) = 0.0000 SE = 1.693967 MAE = 1.123219 DurbWat = 2.174
Previously: 0.0000 0.000000 0.000000 0.000

10 observations fitted, forecast (s) computed for 0 missing val. Of dep. Var.

Analysis of variance for the full regression

Source	Sum of Squares	DF	Mean Square	F-Ratio	F-value
Model	5.96285	3	1.98762	0.692663	0.5893
Error	17.2172	6	2.86953		
Total (Corr.)	23.1800	9			

R-squared = 0.257241 Stand. Error of est. = 1.69397
R-squared (Adj. For d. f.) = 0.5071889 Durbin-Watson statistic = 2.17378

Рис. 12.8. ***Результаты расчета множественной регрессии и ее характеристик***

Полученные результаты могут быть интерпретированы следующим образом.

Уравнение регрессии имеет вид:

$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$,

$y = 15{,}902939 - 2{,}514677x_1 - 0{,}025805x_2 - 0{,}001197x_3$.

Коэффициент Дарбина—Уотсона равен 2,174. Расчетные значения *t*-критерия для коэффициентов регрессии соответственно равны:

0,4409 для a_1,
1,3939 для a_2,
0,6841 для a_3.

Сравнив эти значения с табличным значением критерия, можно судить о значимости каждого из полученных коэффициентов.

Множественный коэффициент детерминации (*R-squared*) равен 0,257241, а множественный коэффициент корреляции равен 0,5071889, т.е. теснота связи между уровнем производительности труда *Y* и выбранными факторами средняя.

Вторая часть таблицы (*Analysis of Variance for Full Regression*) содержит результаты дисперсионного анализа регрессионной модели, которые в данной главе не рассматриваются. Для подробного изучения этих вопросов можно обратиться к работам [11, 85].

Кроме множественной линейной регрессии, в пакете STATGRAPHICS имеется возможность реализовать пошаговый регрессионный анализ при помощи программы ***Stepwise Variable Selection***.

Для выполнения данной программы необходимо выполнить следующие действия:

1) выбрать в главном меню пакета функцию ***Regression Analysis*** и нажать клавишу *Enter*.
2) в появившемся меню выбрать программу ***Stepwise Analysis*** и нажать клавишу *Enter*.
3) заполнить необходимой информацией пустые окна:

Dependent variable — задать имя зависимой переменной *Y*;

Independent variables — задать имена независимых переменных X_1, X_2,..., X_m;

Weights — задать имя переменной, содержащей веса наблюдений. Это поле заполняется в том случае, когда каждому наблюдению присваивается определенный вес;

Constant — выбрать вид уравнения регрессии:
со свободным членом (*Yes*);
без свободного члена (*No*).

Vertical bars — задается один из вариантов вывода на экран графика теоретических значений зависимой переменной: с отмеченными вертикальными отрезками остатков (*Yes*) или без них (*No*);

Confidence level — выбирается коэффициент доверия для построения доверительных интервалов коэффициентов регрессии (90%, 95% и т. д.).

Method — в программе имеется возможность выбрать один из трех методов реализации пошагового регрессионного анализа:

1. *Forward* — анализ начинается с регрессионной модели, не включающей ни одной независимой переменной. Последующее включение или удаление ранее включенной переменной осуществляется на основании *F*-критерия. При помощи этого критерия проверяется значимость каждого частного коэффициента корреляции между включаемой новой переменной X_j и всеми переменными, которые уже вошли в регрессионную модель.

Проверяется гипотеза о том, что теоретическое значение зависимой переменной *Y* существенно не меняется при включении (или исключении) переменной X_j в уравнение.

Если речь идет о включении очередной переменной X_j в уже существующий набор, то значение *F*-критерия называется *F-включения* и вычисляется по формуле:

$$F_{yx_j \cdot c} = \frac{r^2_{yx_j \cdot c}(n-k-2)}{1-r^2_{yx_j \cdot c}},$$

где *c* — набор переменных, включенных в уравнение регрессии, не содержащий переменной X_j;

r^2 — частный коэффициент корреляции между переменной X_j и результативным показателем;

k — количество переменных, входящих в набор *c*.

Аналогично рассчитываются *F-удаления*, когда проверяется гипотеза о том, что после удаления переменной X_j из набора *c* значение *Y* существенно не изменяется:

$$F_{yx_j \cdot c'} = \frac{r^2_{yx_j \cdot c'}(n-k'-2)}{1-r^2_{yx_j \cdot c'}},$$

где *c'*- новый набор переменных, получающийся после удаления X_j из набора *c*;

k' — количество переменных в наборе *c'*.

Расчетные значения *F-удаления* и *F-включения* сравнивают с константами, заданными пользователем по запросу программы. По умолчанию в пакете STATGRAPHICS задается значение 4.0.

2. *Backward* — этот метод отличается от предыдущего метода тем, что алгоритм начинает работу с регрессионной модели, в которую на первом шаге включены все переменные. На каждом последующем шаге происходит удаление переменных на основании значений *F-удаления*.

3. *None* — стандартная процедура множественного регрессионного анализа, описанная в начале раздела.

Следующее поле *Max. steps* позволяет ограничить число шагов в пошаговом регрессионном анализе. Кроме того, при помощи поля *Control* можно задать вариант выдачи результатов анализа (только по последней модели или после каждого шага процедуры).

Когда все высвеченные на экране поля будут заполнены, необходимо нажать клавишу *F6*. Протокол работы процедуры ***Stepwise Variable Selection*** по своей сути мало отличается от результатов работы программы ***Multiple Regression***.

По желанию пользователя можно выдать на экран или на принтер набор статистик, оценивающих как отдельные компоненты уравнения регрессии, так и все уравнение в целом. Имеется возможность записать результаты анализа на диск. Для этого используется поле *Save results*.

Кроме описанных программ, в разделе ***Регрессионный анализ (Regression Analysis)*** имеется программа, позволяющая рассчитывать модели нелинейной регрессии, которые могут быть представлены в виде:

$$\tilde{y}_i = f(x_{1i}, x_{2i} \ldots x_{pi}; Q_1, Q_2 \ldots Q_m) + \varepsilon_i,$$

где f — нелинейная функция переменных X_1, X_2, ..., X_p и параметров Q_1, Q_2, ..., Q_m; $i = 1, 2, \ldots, n$;

ε_i — некоррелированные остатки.

Примером нелинейной регрессионной модели может служить следующая функция:

$$\tilde{y} = f(x; a_0, a_1, a_2) = a_0 + a_1 \cdot e^{a_2 \cdot x}.$$

В случае нелинейной зависимости оценки параметров (a_0, a_1, a_2...) можно получить при помощи *метода максимального правдоподобия*. Процесс вычисления их является итерационным и требует от пользователя задания начальных условий. Подробно все процедуры получения оценок параметров нелинейного уравнения регрессии приводятся в работе [11].

Рассмотрим работу пользователя с программой ***Нелинейный регрессионный анализ (Nonlinear Regression)***.

Для выполнения нелинейного регрессионного анализа необходимо в главном меню выбрать пункт ***Regression Analysis*** и нажать клавишу *Enter*, затем в появившемся меню выбрать программу ***Нелинейный регрессионный анализ (Nonlinear Regression)***.

Для начала работы программы следует заполнить появившиеся на экране поля:

Dependent variable — задать имя зависимой переменной;

Parameter vector — задать начальные значения для коэффициентов регрессии, т.е. предварительные оценки параметров модели;

Function — задать функцию регрессии, т.е. записать на специальном языке, как выглядит нелинейная регрессионная модель;

Maximum iterations — указать максимальное число итераций для процесса поиска оценок параметров модели;

Maximum function calls — задать максимальное число вызовов функций регрессии процедуры нелинейного регрессионного анализа.

Далее следует заполнить пять полей, которые служат критерием для завершения процедуры нелинейного регрессионного анализа. Подробное пояснение по этим параметрам приводится в описании пакета и в работе [11, с. 208].

После заполнения всех полей необходимо нажать клавишу *F6* и на экране появятся результаты нелинейного регрессионного анализа. Протокол работы этой программы очень похож на протоколы работы предыдущих программ регрессионного анализа. После нажатия клавиши *F10* можно выбрать один из режимов вывода результатов анализа на экран или на диск:

Analysis of variance (ANOVA) — вывод на экран таблицы дисперсионного анализа;

Plot fitted model — вывод на экран графика рассеивания остатков;

Summarise residuals — вычисление статистик для анализа остатков;

Save residuals — запись на диск остатков;

Save parameter estimates — запись на диск оценок параметров (a_0, a_1, a_2, ...);

Save covariance — запись на диск ковариационной матрицы оценок параметров.

Для возвращения в главное меню после завершения работы программы нужно нажать клавишу *F10*.

В пакете DSTAT для проведения множественного регрессионного анализа необходимо в разделе *Статистик эксперт* выбрать функцию 5 ***Регрессионный анализ*** и нажать клавишу *Enter*. На рис. 12.9 приведено меню этой функции.

```
*** Регрессионный анализ ***
1. Выход
2. Оценка параметров регрессии МНК
3. Оценка параметров регрессии методом средних
4. Оценка параметров нелинейной регрессии
5. Множественная пошаговая регрессия
6. Оценка параметров двухшаговым МНК
7. Прогнозные значения
8. Оценка параметров при наличии ограничений
Выбирайте функцию?
```

*Рис. 12.9. **Меню регрессионного анализа***

При выборе функции (2) расчет коэффициентов регрессии осуществляется методом наименьших квадратов, а само уравнение отражает зависимость случайной величины Y от факторов X_1, X_2, ..., X_m. Метод наименьших квадратов позволяет получить оценки коэффициентов регрессии, которые являются несмещенными, эффективными и состоятельными при выполнении следующих условий:

отклонения (остатки) некоррелированы;

отклонения имеют постоянную дисперсию;

наблюдения независимы;

исходные данные — случайные величины.

Вычисляемые значения t-критерия Стьюдента и F-критерия Фишера имеют смысл только в случае нормального распределения остатков, что можно проверить по результатам работы функции ***Расчет статистических характеристик и корреляция*** (раздел ***Статистик новичок***).

Пример 12.2. Воспользовавшись той же матрицей исходных данных, что и в примере ***12.1***, проведем регрессионный анализ с помощью метода наименьших квадратов в пакете DSTAT. В меню регрессионного анализа необходимо выбрать функцию (2) и нажать клавишу *Enter*. По запросу программы следует указать номер зависимой переменной. В качестве зависимой переменной возьмем переменную y_1 — производительность труда, млн. руб./чел. (см. рис. 12.3). Протокол работы программы ***Регрессионный анализ*** будет выглядеть следующим образом:

```
Укажите номер фактора, являющегося зависимой переменной? 4
Все факторы включаются в уравнение (да — 0, нет — 1)? 1
Укажите номер фактора, исключаемого из регрессии? 5

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ
Уравнение регрессии имеет вид:

y = 0.159E + 02 − 2.515E + 00*x1 − 2.581E − 02*x2 − 1.197E − 03*x3

                ДАННЫЕ                    ОТКЛОНЕНИЯ
        исходные      расчетные     абсолютные    относительные
1       10.10         8.16          1.94          19.24%
2       8.60          8.86          −0.26         −3.01%
3       9.50          9.78          −0.28         −2.91%
4       9.00          8.38          0.62          6.85%
5       7.60          9.19          −1.59         −20.98%
...     ...           ...           ...           ...

Анализ остатков
Критерий Дарбина-Уотсона          2.173782

Среднее      2.908707E − 06   Дисперсия    1.383119
Эксцесс      −1.187609        С.к.о.       .698741
Асимметрия   −9.654784E − 03   С.к.о.       .5793655

Относительная ошибка аппроксимации = 12.52%
t-статистика для коэффициентов:
b1   .4408717
b2   1.393876
b3   .6840923

Число степеней свободы — 6
ПРИМЕЧАНИЕ: индекс у  x  и  b — номер фактора в исходных данных.
Вычисленное F-значение                          1.443706
Число степеней свободы в числителе             6
Число степеней свободы в знаменателе           3
Множественный коэффициент корреляции (R)       .5071889
R-квадрат                                       .2572406
```

*Рис. 12.10. **Регрессионный анализ. Метод наименьших квадратов***

Судя по полученным результатам, факторы, включенные в уравнение регрессии, объясняют примерно 25,7% всей вариации результативного признака, так как множественный коэффициент детерминации (R-квадрат) равен 0,2572. А степень тесноты связи между откликом Y и факторами X_1, X_2, X_3 средняя — множественный коэффициент корреляции R равен 0,5072. О коррелированности остатков судят по величине вычисленного критерия Дарбина-Уотсона. В данном примере значение критерия равно 2,174, значит автокорреляция остатков отсутствует. Остальные величины, приведенные в протоколе, позволяют оценить качество регрессионной модели и проверить различные гипотезы, связанные с этой моделью.

12.2.2. Факторный анализ

Этот метод многомерного статистического анализа в пакетах STATGRAPHICS и DSTAT представлен в виде двух моделей: метод главных компонент и метод главных факторов. Рассмотрим работу соответствующих программ в каждом из пакетов.

В пакете STATGRAPHICS программа ***Principal components (Главные компоненты)*** входит в раздел ***Многомерные методы (Multivariate Methods)*** главного меню и предназначена для расчета главных компонент на основании матрицы значений исходных переменных. Для выполнения программы необходимо в главном меню пакета выбрать раздел ***Q. Multivariate methods*** и нажать клавишу *Enter*. В появившемся меню выбрать программу ***Principal components***, нажать клавишу *Enter* и заполнить появившиеся на экране поля:

Corr. or cov. matrix — необходимо указать имя файла, содержащего матрицу корреляций или матрицу ковариаций, если одна из них была предварительно рассчитана и сохранена;

Variable labels — указывается имя переменной, содержащей метки исходных переменных;

Data vect. or file name — в этом поле задается имя файла, содержащего значения исходных переменных, или последовательно указываются имена этих переменных (каждое в отдельной строке).

Третье окно заполняется в том случае, когда вычисления главных компонент начинаются с самого начала, т.е. с расчета матрицы парных коэффициентов корреляции или ковариационной матрицы.

Standardise — задается режим *Yes*, когда нужно рассчитать матрицу корреляций, или режим *No*, если нужно рассчитать ковариационную матрицу.

После заполнения всех названных полей нужно нажать клавишу *F6* и на экране появятся результаты работы программы.

Principal Components

Correlation or covariance matrix:
and
Variable labls:
or
Data vectors
or filename: x_1, x_2, x_3

Standardize: *Yes* Missing values: *Listwise* Point labels:

Рис. 12.11. Меню программы ***Главные компоненты***

На рис. 12.12 приведена таблица главных компонент с указанием процента объясняемой суммарной дисперсии исходных переменных, а также накопленных процентов. Чтобы выдать результаты расчетов на принтер или запомнить в файле, необходимо нажать клавишу *F5* и выбрать соответствующий режим вывода.

Principal Components Analysis

Component Number	Percent of Variance	Cumulative Percentage
1	64.87787	64.87787
2	19.78459	84.66246
3	15.33754	100.00000

Рис. 12.12. Результаты работы программы ***Главные компоненты***

Аналитические возможности данной программы можно расширить за счет графического представления результатов. Для этого нужно нажать клавишу *ESC* и в появившемся меню выбрать необходимую функцию:

Plot component weights — на основании матрицы факторных нагрузок каждая переменная будет изображена в виде точки в пространстве двух первых главных компонент (рис. 12.13).

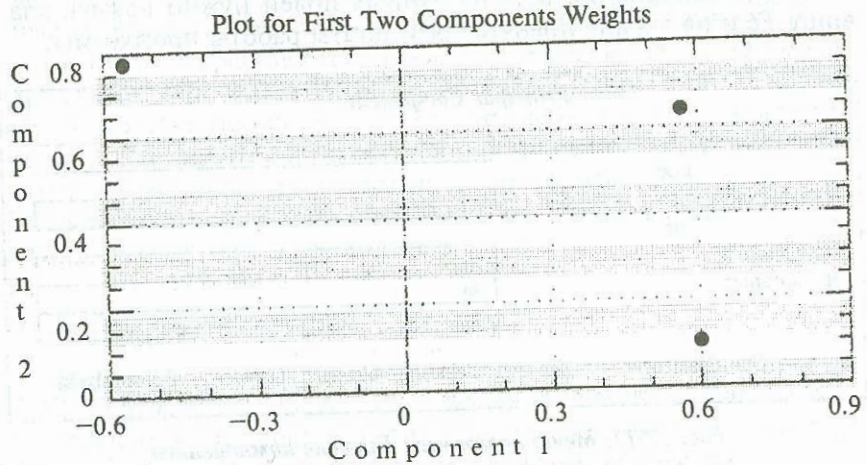


*Рис. 12.13. **Факторные нагрузки исходных переменных на две главные компоненты***

Scatterplot — эта функция предоставляет возможность построения поля рассеяния точек в пространстве первых двух главных компонент;

Biplot — печатается график значений главных компонент для каждого наблюдения (рис.12.14).

К недостаткам программы ***Главные компоненты*** в пакете STATGRAPHICS следует отнести то, что не выводится на экран матрица факторных нагрузок и матрица значений главных компонент.

Вторая модель факторного анализа — метод главных факторов в пакете STATGRAPHICS реализуется при помощи программы ***Факторный анализ*** *(Factor Analysis)* в разделе ***Q. Multivariate methods***. По сравнению с программой ***Главные компоненты*** производится оценка общностей для построения редуцированной корреляционной матрицы. В процессе работы программа запрашивает число факторов и тип вращения матрицы факторных нагрузок для получения простой структуры. На рис. 12.15 показаны результаты работы программы *Factor Analysis* в той последовательности, в которой они получены для нашего примера.

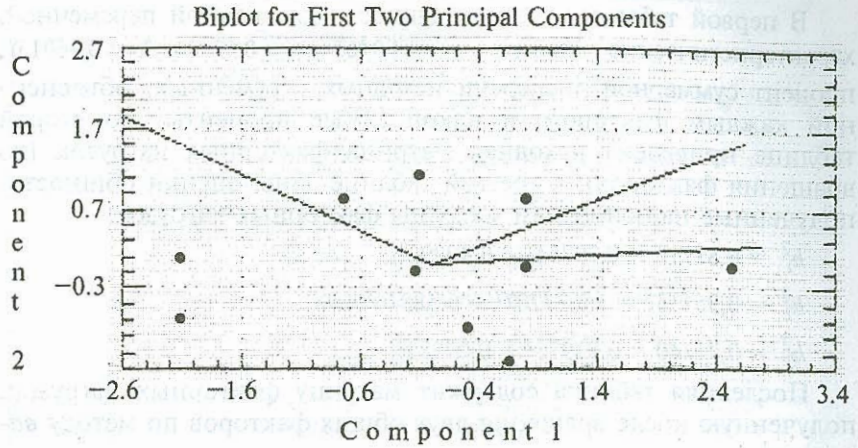


*Рис. 12.14. **График значений главных компонент для десяти наблюдений***

Variable	Communality	Factor	Eigenvalue Percent	Var	Cum Percent
X_1	0.30197	1	1.94634	64.9	64.9
X_2	0.27134	2	.59354	19.8	84.7
X_3	0.36181	3	.46013	15.3	100.0

Factor Matrix (Матрица факторных нагрузок)		
Variable/Factor	1	2
X_1	0.79754	0.49096
X_2	−0.77787	0.58846
X_3	0.83975	0.07882

Variable	Est Communality (Оцененные общности)
X_1	0.87710
X_2	0.95137
X_3	0.71140

VARIMAX ROTATED FACTOR MATRIX		
Variable / Factor	1	2
X_1	0.92601	−0.14001
X_2	−0.21408	0.95160
X_3	0.69216	−0.48199

*Рис. 12.15. **Результаты факторного анализа***

В первой таблице указаны общности для каждой переменной, характеристические корни ($\lambda_1 = 1{,}94634, \lambda_2 = 0{,}59354, \lambda_3 = 0{,}46013$), процент суммарной дисперсии исходных переменных, объясненной каждым фактором, и накопленные проценты. Во второй таблице приведена исходная матрица факторных нагрузок (до вращения факторов), в третьей таблице даны оценки общности, полученные на основании матрицы факторных нагрузок:

$$h_1^2 = 0{,}87710 = 0{,}79754^2 + 0{,}49096^2,$$

$$h_2^2 = 0{,}951137 = -0{,}77787^2 + 0{,}58846^2,$$

$$h_3^2 = 0{,}71140 = 0{,}83975^2 + 0{,}07882^2.$$

Последняя таблица содержит матрицу факторных нагрузок, полученную после вращения двух общих факторов по методу *варимакс*.

В пакете ***DSTAT*** методы факторного анализа реализуются двумя программами раздела ***Статистик эксперт***:

- Компонентный анализ.
- Факторный анализ.

Рассмотрим работу этих программ на примере тех же значений исходных переменных, которые использовались для демонстрации других методов (рис. 12.3 файл ***RBASE***).

Так как в расчете главных компонент участвуют только признаки-факторы, необходимо предварительно файл ***RBASE*** скорректировать — удалить из него зависимые переменные Y_1 и Y_2.

Для проведения анализа методом главных компонент нужно в разделе ***Многомерный статистический анализ*** выбрать функцию 2 (*компонентный анализ*) и нажать клавишу *Enter*. В процессе работы программа запросит точность вычислений собственных значений, а также предельное число итераций. По умолчанию вычисления будут выполнены исходя из значений этих параметров, заложенных в программе.

Так как в нашем примере три независимые переменные (X_1, X_2, X_3), то будут определены три главные компоненты; для каждой из них будет определена доля суммарной дисперсии исходных переменных, объясненной этой компонентой, и накопленный процент дисперсий. Кроме того, рассчитываются матрица факторных нагрузок и значения всех главных компонент для каждого наблюдения. Ниже приведен протокол работы программы ***Компонентный анализ*** (рис. 12.16).

Дисперсии главных компонент

1.9463 0.5935 0.4601

Доли дисперсий главных компонент (Y_i)

64.88% 19.78% 15.34%

Накопленный процент дисперсии

64.88% 84.66% 100.00%

Главные компоненты

	F_1	F_2	F_3
x_1	+ 0.7975	+ 0.4916	− 0.3498
x_2	− 0.7779	+ 0.5880	+ 0.2214
x_3	+ 0.8398	+ 0.0778	+ 0.5373

Индивидуальные значения

	F_1	F_2	F_3
1	− 0.07	1.55	0.20
2	− 1.58	0.09	0.80
3	− 1.58	− 0.92	0.14
4	− 0.55	1.13	0.28
5	− 0.09	− 0.06	0.42
6	0.62	− 0.06	− 1.77
7	0.48	− 1.72	0.17
8	0.61	1.14	− 1.12
9	1.91	− 0.08	1.90
10	0.25	− 1.08	− 1.01

*Рис. 12.16. **Протокол работы программы компонентный анализ***

Аналогично выполняется и программа ***Факторный анализ***. Протокол работы этой программы приведены на рис. 12.17.

По желанию пользователя индивидуальные значения главных компонент или общих факторов могут быть сохранены в файле и в дальнейшем использованы для проведения регрессионного анализа.

В отличие от аналогичных программ, рассмотренных в пакете STATGRAPHICS, программы ***Компонентный анализ*** и ***Факторный анализ*** не позволяют графически изобразить результаты анализа, что с нашей точки зрения является их существенным недостатком.

— Укажите режим определения общностей (автомат. — 0, ручной — 1)?

— Задайте точность вычисления собственных чисел (по умолчанию 0,01)?

— Задайте предельное число итераций?

Для преждевременного завершения нажмите *ВВОД*

— Задайте число определяемых факторов? 2

Итерация # 6 Точность = 2.547972E − 04

— Задайте число определяемых факторов? 2

Факторные нагрузки

	F_1	F_2
x_1	+ 0.6952	+ 0.2058
x_2	−0.6623	+ 0.2302
x_3	+ 0.7313	+ 0.0129

ОЦЕНКИ

	общностей	*специфичностей*
x_1	0.526	0.005
x_2	0.492	0.003
x_3	0.535	0.015

Индивидуальные значения

	F_1	F_2
1	− 0.15	3.82
2	− 1.37	0.29
3	− 1.17	− 2.07
4	− 0.52	2.83
5	− 0.18	− 0.23
6	0.93	0.16
7	0.37	− 4.36
8	0.71	2.98
9	0.90	− 0.92
10	0.49	− 2.50

Рис. 12.17. ***Протокол работы программы Факторный анализ***

12.2.3. Кластерный анализ

В состав процедур первичной обработки данных можно включить и кластерный анализ, который позволяет провести классификацию наблюдений с целью изучения структуры статистической совокупности.

В пакете STATGRAPHICS для проведения кластерного анализа предназначена программа ***Cluster analysis*** из раздела ***Многомерные методы*** *(Multivariate methods)* главного меню. После загрузки программы необходимо для ее выполнения заполнить следующие поля:

Date — указывается имя файла с переменными, матрицей корреляций или матрицей расстояний. Выбор нужного варианта осуществляется в зависимости от того, какая мера сходства используется при объединении объектов (или переменных) в кластеры, а также с какого этапа начинается процесс классификации. Если предварительно уже была рассчитана матрица расстояний между объектами и записана в виде файла, то указывается имя этого файла. В случае, когда определяется сходство между переменными, предварительно должна быть рассчитана матрица парных коэффициентов корреляции и записана в файле;

Label — задается имя переменной, содержащей метки;

Method — выбирается один из методов проведения кластерного анализа.

В пакете STATGRAPHICS реализованы следующие методы:

Seeded — неиерархический метод (аналог метода *k*-средних) (см. § 9.4.);

Average — метод средней связи;

Centroid — центроидный метод;

Furthest — метод «дальнего соседа»;

Nearest — метод «ближайшего соседа»;

Median — метод медианной связи.

Distance — в этом поле указывается тип входных данных, используемых для анализа: матрица корреляций или матрица расстояний (*Matrix*) или матрица значений исходных переменных, по которой рассчитываются евклидовы расстояния между наблюдениями (*Euclidean*);

Standardise — это поле позволяет выбрать режим предварительной обработки данных: *Yes* — стандартизовать значения исходных переменных; *No* — без предварительной стандартизации.

Следующие три поля задаются для графического изображения результатов кластерного анализа в виде трехмерного графика:

X-axis — имя переменной, значения которой откладываются на оси абсцисс;

Y-axis — имя переменной, значения которой откладываются на оси ординат;

Z-axis — имя переменной, значения которой откладываются на оси аппликат.

Circle — это поле позволяет выбрать режим изображения полученных кластеров на графике: *Yes* — оконтурование кластеров; *No* — без оконтурования.

Если кластеризация проводится по методу *Seeded*, необходимо указать номера наблюдений, которые будут точками притяжения будущих кластеров. После заполнения всех перечисленных полей необходимо нажать клавишу *F6* и на экране появятся результаты расчетов. На рис. 12.18 показаны заполненные поля для проведения кластерного анализа десяти наблюдений по пяти переменным (рис.12.3, база данных ***RBASE***) по методу *Seeded*. Предполагается выделить три кластера, используя евклидово расстояние.

Cluster Analysis

Data: X_1 X_2 X_3 Y_1 Y_2

Labels:

Method: *Seeded* Clusters: 3 Distance: *Euclidean* Standardize: *Yes*

X-axis: 1 *Y*-axis: 2 *Z*-axis: 0 Circle: *Yes*

Codes: *ABCEDFGHIJKLMNOPQRSTUVWXYZ* Colors: *12312312312312312312*

*Рис. 12.18. **Меню кластерного анализа***

Значения исходных переменных предварительно будут стандартизованы; контуры полученных кластеров при построении трехмерного графика очерчены; объекты первого кластера обозначены буквой *A*, второго — *B*, третьего — *C*.

Результаты кластерного анализа представлены на рис. 12.19.

Results of Clustering by Seeded Method

Observation	Cluster	Cluster	Frequency	Percentage	Seed
1. Obs. 1	1	1	2	20.0000	1
2. Obs. 2	2	2	7	70.0000	5
3. Obs. 3	2	3	1	10.0000	9
4. Obs. 4	2				
5. Obs. 5	2				
6. Obs. 6	2				
7. Obs. 7	1				
8. Obs. 8	2				
9. Obs. 9	3				
10. Obs. 10	2				

*Рис. 12.19. Результат выполнения программы **Кластерный анализ** (метод Seeded)*

Видно, что десять наблюдений оказались разбиты на три кластера. В первый кластер попало два наблюдения, во второй — семь, в третий — одно наблюдение. Из протокола работы программы видно, какие наблюдения попали в тот или иной кластер.

Если необходимо получить графическое представление, следует нажать клавишу *ESC* и выбрать режим *Plot clusters*. Кроме того, есть возможность выдать результаты анализа на принтер (клавиша *F5*) или заполнить их в файле, а также сохранить матрицу расстояний (*Save distance matrix*) и переменную с номерами кластеров для каждого наблюдения (*Save cluster numbers*). Для возвращения в главное меню пакета необходимо нажать клавишу *F10*.

В пакете DSTAT кластерный анализ представлен только одним методом — иерархическим агломеративным кластерным анализом. В качестве меры сходства между единицами наблюдения в программе используется евклидова метрика:

$$d_{ke} = \sqrt{\sum_{j=1}^{m}(x_{jk} - x_{je})^2}.$$

Расстояние между объединяемыми кластерами определяется по алгоритму «ближайшего соседа». Перед началом процесса классификации на экран выводится запрос о нормировании значений исходных переменных. В ходе работы алгоритма пользователю выдается информация о числе существующих класте-

ров, их составе, а также запрос о продолжении процедуры классификации. Процесс объединения кластеров завершается образованием одного кластера, в состав которого входят все наблюдаемые единицы. Если число наблюдений не больше 20, то на экран выводится дендрограмма. Ниже приведен фрагмент протокола работы программы ***Кластерный анализ*** (рис. 12.20), а на рис. 12.21 — окончательный результат работы иерархического агломеративного алгоритма в виде дендрограммы:

Число кластеров = 10

Минимальное расстояние между 2-м и 3-м кластерами равно 1.177631

Кластер # 2 содержит наблюдения: 2
Кластер # 3 содержит наблюдения: 3

Объединить эти кластеры (да — 1, нет — 0)? 1

Число кластеров = 9

Минимальное расстояние между 3-м и 4-м кластерами равно 1.455485

Кластер # 3 содержит наблюдения: 4
Кластер # 4 содержит наблюдения: 5

Объединять эти кластеры (да - 1, нет - 0)? 1 и т.д.

. . . . . .

Число кластеров = 2

Минимальное расстояние между 1-м и 2-м кластерами равно 3.869156

Кластер # 1 содержит наблюдения: 1 2 3 6 8 10
Кластер # 2 содержит наблюдения: 4 5 7 9

Объединять эти кластеры (да — 1, нет — 0)? 1

Число кластеров = 1

Кластер # 1 содержит наблюдения 1 2 3 4 5 6 7 8 9 10

Рис. 12.20. Протокол работы программы ***Кластерный анализ***

На рис. 12.21 видно, что объединение началось на расстоянии 1,178 — объединены кластеры 2 и 3, затем на расстоянии 1,455 объединены кластеры 4 и 5, и так далее до тех пор, пока все наблюдения не оказались в одном кластере. Наглядное изображение в виде дендрограммы позволяет проследить последовательность объединения и по величине расстояния судить о степени сходства объединяемых кластеров. Если число классифицируемых наблюдений больше 20, то дендрограмма не строится.

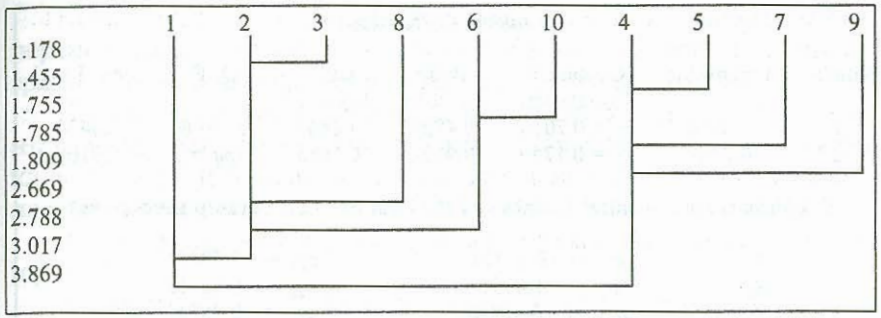


Рис. 12.21. Дендрограмма иерархического агломеративного анализа для десяти наблюдений

12.2.4. Канонические корреляции

В пакете STATGRAPHICS этот метод реализуется при помощи программы ***Canonical correlation*** (*канонические корреляции*) раздела ***Multivariate methods*** (*многомерные методы*) главного меню, которая позволяет для двух множеств переменных рассчитать значения канонических коэффициентов корреляции. Одно множество переменных — это признаки-факторы X_1, X_2,..., X_p, а второе — результативные признаки Y_1, Y_2,..., Y_q. Все переменные должны быть записаны в одном файле, который должен быть открыт к началу работы данной программы.

Для выполнения программы необходимо в главном меню пакета выбрать строку ***Canonical correlation*** и нажать клавишу *Enter*. На экране появятся два поля, которые нужно заполнить:

First set of variables — имена переменных первого множества (X_1, X_2,..., X_p);

Second set of variables — имена переменных второго множества (Y_1, Y_2,..., Y_q).

После заполнения указанных полей нажать клавишу *F6* для выполнения расчетов. На экране появится таблица канонических корреляций (рис. 12.22).

Для $r_1 = 0.701$:

$U_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = -0{,}11321x_1 - 1{,}08509x_2 - 0{,}83446x_3$;

$V_1 = b_{11}y_1 + b_{12}y_2 = 1{,}1471y_1 - 0{,}8411y_2$.

Для $r_2 = 0.1759$:

$U_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = -1{,}03674x_1 - 0{,}32912x_2 + 0{,}86742x_3$;

$V_2 = b_{21}y_1 + b_{22}y_2 = 0{,}27459y_1 + 0{,}82691y_2$.

Canonical Correlations

Number	Eigenvalue	Canonical Correlation	Wilks Lambda	Chi - Squre χ^2	D. F.	Sign. Level
1	0.4914	$r_1 = 0.7010$	0.4929	4.2449	$pq = 6$	0.6436
2	0.0309	$r_2 = 0.1759$	0.9691	0.1885	$1/pq = 2$	0.9100

Coefficients for Canonical Variables of the First Set (для первого множества)

X_1	$a_{11} = -0.11321$	$a_{21} = -1.03674$
X_2	$a_{12} = -1.08509$	$a_{22} = -0.32912$
X_3	$a_{13} = -0.83446$	$a_{23} = -0.86742$

Coefficients for Canonical Variables of the Second Set (для второго множества)

Y_1	$b_{11} = 0.14710$	$b_{21} = 0.27459$
Y_2	$b_{12} = -0.84110$	$b_{22} = 0.82691$

*Рис. 12.22. **Результаты расчетов по программе Канонические корреляции***

На рис. 12.22 приведены результаты работы программы ***Канонические корреляции***. Для двух заданных множеств исходных переменных определены два канонических коэффициента корреляции ($r_1 = 0{,}701$, $r_2 = 0{,}1759$), значения λ-критерия и χ^2-критерия (критерия Бартлетта), число степеней свободы для оценки значимости каждого из коэффициентов $p \cdot q = 6$, $(p-1)(q-1) = 2$. Расчетное значение χ^2-критерия определяется по формуле:

$$\chi^2 = -[n - m - 1/2(p + q + 1)] \ln U,$$

где $U = (1 - r_1^2)(1 - r_2^2)$ для r_1 и $U = (1 - r_2^2)$ для r_2.

Для r_1:

$$\chi^2_{\text{расч}} = -[10 - 1 - 1/2(3 + 2 + 1)] \ln (1 - 0{,}701^2) = 4{,}245,$$

Для r_2:

$$\chi^2_{\text{расч}} = -[10 - 1 - 1/2(3 + 2 + 1)] \ln (1 - 0{,}1759^2) = 0{,}1885.$$

Табличное значение χ^2 для числа степеней свободы $p \times q = 3 \times 2 = 6$ и уровня значимости $\alpha = 0{,}05$ равно 1,635. Так как $\chi^2_{\text{расч}} = 4{,}245 > \chi^2_{\text{табл}} = 1{,}635$, подтверждается гипотеза о значимости первого коэффициента канонической корреляции. Аналогично проверяется значимость второго коэффициента.

Если необходимо результаты работы запомнить в файле или распечатать, нужно нажать клавишу *F5*.

Кроме того, программа рассчитывает значения канонических переменных (два набора) для каждого объекта наблюдения и строит графики канонических переменных (*Plot canonical variables*).

По желанию пользователя значения канонических переменных могут быть сохранены в файле. Для этого следует нажать клавишу *ESC* и выбрать в появившемся меню функцию ***Save canonical variables***.

Для возвращения в главное меню необходимо нажать клавишу *F10*.

В пакете DSTAT канонический корреляционный анализ можно осуществить, войдя в раздел ***Статистик эксперт***. В меню ***Многомерного статистического анализа*** нужно выбрать программу ***Канонические корреляции*** и нажать клавишу *Enter*. На экране появятся следующее сообщение и запросы:

В каноническом корреляционном анализе все исходные факторы разбиваются на два множества: $\{Y_1, \ldots, Y_q\}$ и $\{X_1, \ldots, X_p\}, g \le p$.

На этих множествах определяются линейные комбинации, которые:

- центрированы;
- нормированы;
- не коррелированы внутри множеств;
- корреляция между соответствующими линейными комбинациями (1-я с 1-й, ..., q-я с q-й) суть канонические корреляции, а для несоответствующих равна нулю.

Разбиение исходных факторов на два множества задается пользователем.

В множество *Y* включается фактор # 4

В множество *Y* включается фактор # 5

Все остальные факторы включаются в множество *X*

Точность вычислений — максимальный внедиагональный элемент

Задайте точность вычисления собственных чисел (по умолчанию .01)?

Задайте предельное число итераций (по умолчанию 18)? 50

После ответа на запросы программы ,нужно нажать клавишу *Enter* и на экране появятся результаты канонического корреляционного анализа (рис. 2.23)

Итерация # 2 Точность = 1.490116E - 08

Каноническая корреляция = 0.701

хи-квадрат = 4.24 Число степеней свободы = 6

Канонические величины:

$-1.147E + 00*y_4 + 8.411E - 01*y_5$

$+ 1.132E - 01*x_1 + 1.085E + 00*x_2 + 8.345E - 01*x_3$

Каноническая корреляция = 0.176

хи-квадрат = 0.22 Число степеней свободы = 2

Канонические величины:

$+ 2.746E - 01* y_4 + 8.269E - 01* y_5$

$-1.037E + 00* x_1 - 3.291E - 01* x_2 + 8.674E - 01* x_3$

Рис. 12.23. ***Результаты расчетов по программе Канонические корреляции в пакете*** DSTAT

Из протокола работы видно, что за две итерации была достигнута заданная точность вычислений собственных значений и получены две пары канонических переменных.

Выводы

Рассмотренные в данной главе специальные статистические пакеты и опыт их использования позволяют сделать следующие выводы.

Оценивать качество статистического программного продукта необходимо с точки зрения конкретного пользователя, а не по каким-то общим критериям.

Если пользователь имеет высокую квалификацию в статистическом анализе и ему предстоит обрабатывать большие объемы информации, то предпочтение следует отдать таким специальным пакетам, как SPSS, STATGRAPHICS, SAS и т. п.

В учебном процессе удобно пользоваться простыми статистическими пакетами, которые не требуют больших объемов памяти. Такие пакеты легко эксплуатировать в компьютерных

классах при проведении практических занятий. Примером может служить пакет DSTAT, который содержит достаточно большой набор процедур статистического анализа, очень удобную форму представления результатов и при этом занимает небольшой объем памяти.

Для начинающих пользователей многомерных методов наиболее удобной формой представления результатов работы компьютерной программы являются подробные русифицированные протоколы, например, как в пакете DSTAT. Это позволяет в определенной мере проследить порядок проведения анализа, а не только получить готовый ответ в виде таблицы или графика.

Если существует несколько алгоритмов реализации выбранного метода анализа в компьютерной программе, то это представляет большой интерес для пользователя только при наличии понятной документации по статистическому пакету. Иначе даже специалисту в области прикладной статистики бывает трудно правильно ответить на запросы программы.

При выборе пакета статистических программ следует учитывать, что более поздние версии не всегда содержат расширенный набор методов и функций многомерного анализа по сравнению с предыдущими. Необходимо внимательно изучить имеющуюся документацию по пакету.

Учитывая популярность, которую все больше завоевывают статистические графики в аналитической практике, предпочтительнее пользоваться статистическими пакетами, которые предоставляют широкий выбор графического изображения исходных данных и результатов анализа.

Вопросы и задачи

1. Назовите важнейшие с точки зрения пользователя характеристики пакетов статистических программ.

2. Какие методы MCA реализованы в пакетах DSTAT и STATGRAPHICS?

3. Перечислите алгоритмы регрессионного анализа в пакетах DSTAT и STATGRAPHICS. Укажите, в чем заключаются

достоинства и недостатки реализации этого метода в названных пакетах.

4. Какие алгоритмы кластерного анализа можно использовать при работе с пакетом STATGRAPHICS?

5. Проведите иерархический агломеративный кластерный анализ при помощи пакетов DSTAT и STATGRAPHICS. Распечатайте протоколы работы программ и дайте их сравнительную характеристику.

6. На основании исходных данных, приведенных в «Вопросах и задачах» гл. 11, выполните расчеты по методу канонических корреляций при помощи пакетов DSTAT и STATGRAPHICS. Распечатайте протоколы работы программ и поясните их содержание.

Библиографический список

1. *Автоматизированное* рабочее место для статистической обработки данных. Шураков В.В., Дайитбегов Д.М., Мизрохи С.В., Ясеновский С.В. — М.: Финансы и статистика, 1990. — 190 с.
2. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983. — 472 с.
3. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. — М.: Статистика, 1974. — 240 с.
4. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.
5. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Исследование зависимостей — М.: Финансы и статистика, 1985. — 487 с.
6. *Александров А.Д., Нецветаев Н.Ю.* Геометрия. — М.: Наука, 1990. — С. 11 — 393, 655 — 660.
7. *Александров В.В., Алексеев А.И., Горский Н.Д.* Анализ данных на ЭВМ (на примере системы СИТО). — М.: Финансы и статистика, 1990. — 192 с.
8. *Андерсон Т.* Введение в многомерный статистический анализ. — М.: Физматгиз, 1963. — 500 с.
9. *Андерсон Т.* Статистический анализ временных рядов. — М.: Мир, 1976. — 755 с.
10. *Апатенок Р.Ф., Маркина А.М., Попова Н.В., Хейнман В.Б.* Элементы линейной алгебры и аналитической геометрии. — Мн.: Вышэйшая школа, 1986. — 272 с.
11. *Афифи А., Эйзен С.* Статистический анализ: подход с использованием ЭВМ. — М.: Мир, 1982. — 488 с.

12. *Бард Й.* Нелинейное оценивание параметров. — М.: Статистика, 1979. — 349 с.
13. *Беллман Р.* Введение в теорию матриц. Пер. с англ. — М.: Наука, 1969. — 367 с.
14. *Бендат Дж., Пирсол А.* Прикладной анализ случайных данных. — М.: Мир, 1989. — 540 с.
15. *Благуш П.* Факторный анализ с обобщениями. — М.: Финансы и статистика, 1988. — 248 с.
16. *Бокс Дж., Дженкинс Г.* Анализ временных рядов: Прогноз и управление. Вып. 1. — М.: Мир, 1974. — 408 с.
17. *Болч Б., Хуань К. Дж.* Многомерные статистические методы для экономики. — М.: Финансы и статистика, 1979. — 317 с.
18. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. — М.: Наука, 1983. — 416 с.
19. *Браверман Э.М., Мучник И.Б.* Структурные методы обработки эмпирических данных. — М.: Наука, 1983. — 464 с.
20. *Бро Г.Г., Шнайдман Л.М.* Математические методы экономического анализа на предприятии. — М.: Экономика, 1976. — 183 с.
21. *Булдык Г.М.* Теория вероятностей и математическая статистика. — Мн.: Вышэйшая школа, 1989. — 285 с.
22. *Бухштабер В.М., Маслов В.И., Зеленюк Е.А.* Методы анализа и построения алгоритмов автоматической классификации на основе математических моделей // Прикладная статистика. — М.: Наука, 1983. — С. 126 — 144.
23. *Вучков И., Бояджиева Л., Солаков Е.* Прикладной линейный регрессионный анализ. — М.: Финансы и статистика, 1987. — 239 с.
24. *Гантмахер Ф.Р.* Теория матриц. — М.: Наука, 1988. — 548 с.
25. *Герасимович А.И., Матвеева Я.И.* Математическая статистика. — Мн.: Вышэйшая школа, 1978. — 200 с.
26. *Дадаян А.А., Дударенко В.А.* Алгебра и геометрия. — Мн.: Вышэйшая школа, 1989. — 288 с.
27. *Джонсон Дж.* Эконометрические методы. — М.: Статистика, 1980. — 444 с.
28. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Кн. 1, 2. — М.: Финансы и статистика, 1986. — 702 с.
29. *Дружинин Н.К.* Логика оценки статистических гипотез. — М.: Статистика, 1973. — 211 с.
30. *Дубров А.М.* Обработка статистических данных методом главных компонент. — М.: Статистика, 1978. — 135 с.
31. *Дубровский С.А.* Прикладной многомерный статистический анализ. — М.: Финансы и статистика, 1982. — 216 с.

32. *Дэйвисон М.* Многомерное шкалирование. — М.: Финансы и статистика, 1988. — 254 с.
33. *Дюк В.А.* Компьютерная психодиагностика. — Санкт-Петербург: «Братство», 1994. — 364 с.
34. *Дюран Б., Оделл П.* Кластерный анализ. — М.: Статистика, 1977. — 128 с.
35. *Елисеева И.И., Рукавишников В.О.* Логика прикладного статистического анализа. М.: Финансы и статистика, 1982. — С. 35 — 71.
36. *Елисеева И.И., Рукавишников В.О.* Группировка, корреляция, распознавание образов. — М.: Статистика, 1977. — 144 с.
37. *Елисеева И.И., Семенова Е.В.* Основные процедуры многомерного статистического анализа. — Л.: УЭФ, 1993. — 78 с.
38. *Елкина В.Н., Загоруйко Н.Г.* Количественные критерии качества таксономии и их использование в процессе принятия решений // Вычислительные системы. — Новосибирск: Наука, 1969. — С. 29 — 47.
39. *Енюков И.С.* Методы, алгоритмы, программы многомерного статистического анализа: Пакет ППСА. — М.: Финансы и статистика, 1986. — 232 с.
40. *Жамбю М.* Иерархический кластер-анализ и соответствия. — М.: Финансы и статистика, 1988. — 342 с.
41. *Загоруйко Н.Г.* Методы распознавания и их применение. — М.: Советское радио, 1972. — 288 с.
42. *Закс Л.* Статистическое оценивание. — М.: Статистика, 1976. — 598 с.
43. *Закс Л.* Теория статистических выводов. — М.: Мир, 1975. — 776 с.
44. *Иберла К.* Факторный анализ. М.: Статистика, 1980. — 398 с.
45. *Кади Дж.* Количественные методы в экономике. — М.: Прогресс, 1977 — С. 85 — 239.
46. *Кейн Э.* Экономическая статистика и эконометрия. Вып.2. — М.: Статистика, 1977. — 230 с.
47. *Кендалл М. Дж., Стьюарт А.* Многомерный статистический анализ и временные ряды. — М.: Наука, 1976. — 736 с.
48. *Кендалл М., Стьюарт А.* Статистические выводы и связи. — М.: Наука, 1973. — 900 с.
49. *Кильдишев Г.С., Аболенцев Ю.И.* Многомерные группировки. — М.: Статистика, 1978. — 160 с.
50. *Ким Дж.-О., Мьюллер Ч.У. и др.* Факторный, дискриминантный и кластерный анализ. Пер. с англ. — М.: Финансы и статистика, 1989. — 215 с.

51. *Клейн Ф.* Элементарная математика с точки зрения высшей. Том 2. – М.: Наука, 1987. — С. 10 — 327.
52. *Клигер С.А., Косолапов М.С., Толстова Ю.Н.* Шкалирование при сборе и анализе социологической информации. — М.: Наука, 1978. — 112 с.
53. *Колемаев В.А., Староверов О. В, Турундаевский В.Б.* Теория вероятностей и математическая статистика. — М.: Высшая школа, 1991. – 400 с.
54. *Колмогоров А.Н.* Теория вероятностей и математическая статистика. — М.: Наука, 1986. — 536 с.
55. *Колмогоров А.Н., Фомин С.В.* Элементы теории функций и функционального анализа. — М.: Наука, 1989. — 623 с.
56. *Кулаичев А.П.* Методы и средства анализа данных в среде Windows Stadia 6.0 – М.: Информатика и компьютеры, 1996. – 257 с.
57. *Курош А.Г.* Курс высшей алгебры. — М.: Физматиз, 1962. — 431 с.
58. *Ллойд Э., Ледерман У.* Справочник по прикладной статистике, Т.1 и 2 т. — М.: Финансы и статистика, 1990. — 1036 с.
59. *Лоули Д., Максвелл А.* Факторный анализ как статистический метод. — М.: Мир, 1967. — 144 с.
60. *Львовский Е.Н.* Статистические методы построения эмпирических формул. — М.: Высшая школа, 1988. — 239 с.
61. *Мандель И.Д.* Кластерный анализ. — М.: Финансы и статистика, 1988. – 176 с.
62. *Митропольский А.К.* Техника статистических вычислений. — М.: Наука, 1971. — 576 с.
63. *Мозоров И.Д.* Матричные расчеты в статистике. — М.: Финансы и статистика, 1983. — 216 с.
64. *Мостлер Ф., Тьюки Дж.* Анализ данных и регрессия. Вып. 1, 2. — М.: Финансы и статистика, 1982. — 520 с.
65. *Мэйндональд Дж.* Вычислительные алгоритмы в прикладной статистике. – М.: Финансы и статистика, 1988. — С. 79—120.
66. *Окунь Я.* Факторный анализ. — М.: Статистика, 1974. — 200 с.
67. *Орлов А.И.* Устойчивость в социально-экономических моделях. — М.: Наука, 1979. – 296 с.
68. *Петрович М.Л., Давидович М.В.* Статистическое оценивание и проверка гипотез на ЭВМ. — М.: Финансы и статистика, 1989. — 190 с.
69. *Плошко Б.Г., Елисеева И.И.* История статистики. — М.: Финансы и статистика, 1990. — С. 171 — 176.

70. *Плюта В.* Сравнительный многомерный анализ в экономическом моделировании. Пер. с польского. — М.: Финансы и статистика, 1989. — 175 с.
71. *Райков Д.А.* Многомерный математический анализ. — М.: Высшая школа, 1989. — 271 с.
72. *Райс Дж.* Матричные вычисления и математическое обеспечение. Пер. с англ. — М.: Мир, 1984. — 264 с.
73. *Рао С.Р.* Линейные статистические методы и их применения. — М.: Наука, 1968. — 548 с.
74. *Севастьянов Б.А.* Вероятностные модели. — М.: Наука, 1992. — 176 с.
75. *Сербер Дж.* Линейный регрессионный анализ. — М.: Финансы и статистика, 1980. — 456 с.
76. *Сирл С., Госхман У.* Матричная алгебра в экономике. Пер. с англ. — М.: Статистика, 1974. — 374 с.
77. *Смирнов Е.С.* Таксономический анализ. — М.: Изд-во МГУ, 1969. — 187 с.
78. *Смоляк С.А., Титаренко Б.П.* Устойчивые методы оценивания. М.: Статистика, 1980. — 208 с.
79. *Статистические* методы для ЭВМ / Под ред. Эйнслейна, Э. Релстоуна, Г.С. Уолфа. — М.: Наука, 1986. — 459 с.
80. *Статистический* словарь/Под ред. Ю.А. Юркова. — М.: Финстатинформ, 1996. — 479 с.
81. *Стренг Г.* Линейная алгебра и ее применения. Пер. с англ. — М.: Мир, 1980. — 454 с.
82. *Терехина А.Ю.* Анализ данных методами многомерного шкалирования. — М.: Наука, 1986. — 168 с.
83. *Терехина А.Ю.* Методы многомерного шкалирования и визуализации данных//Автоматика и телемеханика. 1973. — № 7. — С. 80 — 94.
84. *Торгерсон У.С.* Многомерное шкалирование. Теория и метод// Статистическое измерение качественных характеристик. — М.: Статистика, 1972. — С. 95 — 118.
85. *Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютере / Под ред. В.Э. Фигурнова. — М.: ИНФРА — М.: Финансы и статистика, 1997. – 528 с.
86. *Уилкс С.* Математическая статистика. Пер. с англ. — М.: Наука, 1967. — 632с.
87. *Ферестер Э., Ренц Б.* Методы корреляционного и регрессионного анализа. – М.: Финансы и статистика, 1988. — 302 с.
88. *Фигурнов В.Э.* IBM для пользователя. 5-е издание. — М.: Финансы и статистика, 1994. — 368 с.

89. *Харин Ю.С., Степанова М.Д.* Практикум на ЭВМ по математической статистике. — Мн.: Университетское, 1987. — 304 с.
90. *Харман Г.* Современный факторный анализ. — М.: Статистика, 1972. — 486 с.
91. *Хемминг Р.В.* Численные методы. Пер. с англ. — М.: Наука, 1972. — 400 с.
92. *Хорн Р., Джонсон Ч.* Матричный анализ. Пер. с англ. — М.: Мир: 1989. — 655 с.
93. *Шураков В.В., Дайитбегов Д.М. и др.* Автоматизированное рабочее место для статистической обработки данных. — М.: Финансы и статистика, 1990. — 190 с.
94. *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. Пер. с англ. — М.: Финансы и статистика. 1988. — 263 с.
95. *Backhaus K., Erichson B., Plinke W., Weiber R.* Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 7. Auflage, Berlin, 1994. — P. 433 — 497, 594.
96. *Bleymuler J., Gehlert G., Gulicher H.* Statistik für Wirtschaftswissenschaftler. 8. Auflage. — München, 1992.
97. *Borg I.* Anwendungsorientierte Multidimensinale Skalierung. Berlin, 1981.
98. *Braun, Martin* (1975): Differential Equations and Their Applications. Springer Verlag. New York, Heidelberg, Berlin.
99. *Dhrymes P.J.* Introductory Econometrics. Springer, New York, 1978.
100. *Green P.E., Carmone F., Smith S.M.* Multidimensional Scaling: Concepts and Applikations. Bosten, London u. a., 1989.
101. *Greene W.H.* Econometric Analysis. Macmillan, New York, 1991.
102. Greub, Werner (1975): Linear Algebra. (Fourth Edition). Springer Verlag. New York, Heidelberg, Berlin.
103. *Hartung J., Elpelt B.* Multivariate Statistik. Oldenbourg Verlag, München-Wien, 1986.
104. *Hartung J., Werner H.J.* Hypothesenprüfung im restringierten linearen Modell: Theorie und Anwendungen. Vandenhoeck & Ruprecht, Gottingen, 1984.
105. *Intriligator M.* Econometric Models. Techniques and Applications. North Holland, 1978, P. 159 — 165.
106. *Jänich, Klaus* (1979): Lineare Algebra. Springer Verlag. New York, Heidelberg, Berlin.
107. *Johnston J.* Econometric Methods. Third edition. McGraw-Hill International Editions, Auckland etc., 1984.
108. *Kemper F.J.* Multidimensionale Skalierung. Bremen, 1984.

109. *Kendall M.* Multivariate Analysis. — 2 hd ed, hondos, 1980.
110. *Morrison D.F.* Multivariate Statistical Methods. — 2 hd ed, New York, 1976.
111. *Oksanen E.H.* A Simple Approach to Teaching Generalized Least Squares Theory. The American Statistician 45, No 3, 1991. — P. 229 — 233.
112. *Rasch D.* Einführung in die mathematische Statistik I und II. Deutscher Verlag der Wissenschaften, Berlin, 1978.
113. Searle S.R. Matrix Algebra for the Biological Sciences, Wiley, New York, 1966, — P. 166 — 167, 203.
114. *Überla K.* Faktorenanalyse, Springer, Berlin, 1977.
115. *Uebe G., Schäfer M.* Einführung in die Statistik für Wirtschaftswissenschaftler. Oldenbourg, München—Wien, 1991.
116. *Zellner A.* Econometrics. An Introduction to Bayesian Inference in Econometrics. Wiley, New York, 1971.

Приложения

Приложение 1

Плотность вероятностей нормированного нормального распределения: $f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$.

z	0	1	2	3	4	5	6	7	8	9
0.0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0.2	3910	3902	3894	3885	3876	3867	3957	3847	3836	3825
0.3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0.4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0.5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0.6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0.7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0.8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0.9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1.0	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1.2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1.3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1.4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1.6	1109	1092	1047	1057	1040	1023	1006	0989	0973	0957
1.7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1.8	0790	0775	0761	0748	0734	0721	0707	0694	1681	0669
1.9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2.0	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2.2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2.3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2.4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2.5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2.6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2.7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2.8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2.9	0060	0058	0056	0053	0051	0051	0050	0048	0047	0046
3.0	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3.2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3.3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3.4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3.5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3.6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3.7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3.8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3.9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Приложение 2

Значения функции $\varphi(z)=\frac{2}{\sqrt{2\pi}}\int_0^z e^{\frac{-z^2}{2}}dt=\frac{1}{\sqrt{2\pi}}\int_{-z}^{+z}e^{\frac{-z^2}{2}}dz$.

Целые и десятые доли z_i	Сотые доли z_i									
	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0080	0.0160	0.0239	0.0319	0.0399	0.0478	0.0558	0.0638	0.0717
0.1	0.0797	0.0876	0.055	0.1034	0.1113	0.1192	0.1271	0.1350	0.1428	0.1507
0.2	0.1585	0.1663	0.1741	0.1819	0.1897	0.1974	0.2051	0.2128	0.2205	0.2282
0.3	0.2358	0.2434	0.2510	0.2586	0.2661	0.2737	0.2812	0.2886	0.2960	0.3035
0.4	0.3108	0.3128	0.3255	0.3328	0.3401	0.3473	0.3545	0.3616	0.3688	0.3759
0.5	0.3829	0.3899	0.3969	0.4039	0.4108	0.4177	0.4245	0.4313	0.4381	0.4448
0.6	0.4515	0.4581	0.4647	0.4713	0.4778	0.4843	0.4907	0.4971	0.5035	0.5098
0.7	0.5161	0.5223	0.5285	0.5346	0.5407	0.5467	0.5527	0.5587	0.5646	0.5705
0.8	0.5763	0.5821	0.5878	0.5935	0.5991	0.6047	0.6102	0.6157	0.6211	0.6265
0.9	0.6319	0.6372	0.6424	0.6476	0.6528	0.6579	0.6629	0.6679	0.6729	0.6778
1.0	0.6827	0.6875	0.6923	0.6970	0.7017	0.7063	0.7109	0.7154	0.7199	0.7243
1.1	0.7287	0.7330	0.7373	0.7415	0.7457	0.7499	0.7540	0.7580	0.7620	0.7660
1.2	0.7699	0.7737	0.7775	0.7813	0.7850	0.7887	0.7923	0.7959	0.7994	0.8029
1.3	0.8064	0.8098	0.8132	0.8165	0.8198	0.8282	0.8262	0.8293	0.8324	0.8355
1.4	0.8385	0.8415	0.8444	0.8473	0.8501	0.8529	0.8557	0.8584	0.8611	0.8638
1.5	0.8664	0.8690	0.8715	0.8740	0.8764	0.8789	0.8812	0.8836	0.8859	0.8882
1.6	0.8904	0.8926	0.8948	0.8969	0.8990	0.9011	0.9031	0.9051	0.9070	0.9090
1.7	0.9109	0.9127	0.9146	0.9164	0.9181	0.9199	0.9216	0.9233	0.9249	0.9265
1.8	0.9281	0.9297	0.9312	0.9327	0.9342	0.9357	0.9371	0.9385	0.9399	0.9412
1.9	0.9426	0.9439	0.9451	0.9446	0.9476	0.9488	0.9500	0.9512	0.9523	0.9534
2.0	0.9545	0.9556	0.9566	0.9576	0.9586	0.9596	0.9606	0.9616	0.9625	0.9634
2.1	0.9643	0.9651	0.9660	0.9668	0.9666	0.9684	0.9692	0.9700	0.9707	0.9715
2.2	0.9722	0.9729	0.9736	0.9743	0.9749	0.9756	0.9762	0.9768	0.9774	0.9780
2.3	0.9786	0.9791	0.9797	0.9802	0.9801	0.9812	0.9817	0.9822	0.9827	0.9832
2.4	0.9836	0.9841	0.9845	0.9849	0.9853	0.9857	0.9861	0.9865	0.9869	0.9872
2.5	0.9876	0.9879	0.9883	0.9886	0.9889	0.9892	0.9895	0.9898	0.9901	0.9904
2.6	0.9907	0.9910	0.9912	0.9915	0.9917	0.9920	0.9922	0.9924	0.9926	0.9928
2.7	0.9931	0.9933	0.9935	0.9937	0.9939	0.9940	0.9942	0.9944	0.9046	0.9947
2.8	0.9949	0.9951	0.9952	0.9953	0.9955	0.9956	0.9958	0.9959	0.9960	0.9961
2.9	0.9963	0.9964	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972
3.0	0.9973	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980
3.1	0.9981	0.9982	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3.2	0.9986	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.3	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.4	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	0.9995
3.5	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	0.9997
3.6	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	0.9998	0.9998
3.7	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
4.0	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
4.1	0.9999									

Приложение 3

Распределение Стьюдента (t-распределение)

Значения $t_{\alpha,\nu}$ удовлетворяют условию: $P(t \geq t_{\alpha,\nu}) = \int_{t_{\alpha,\nu}}^{\infty} S(t,\nu)\,dt = \alpha$

ν \ α	0.10	0.050	0.010	ν \ α	0.10	0.050	0.010
1	3.078	6.314	31.82	20	1.325	1.725	2.528
2	1.886	2.920	6.965	21	1.323	1.721	2.518
3	1.638	2.353	4.541	22	1.321	1.717	2.508
4	1.533	2.132	3.747	23	1.319	1.714	2.500
5	1.476	2.015	3.365	24	1.318	1.711	2.492
6	1.440	1.943	3.143	25	1.316	1.708	2.485
7	1.415	1.895	2.998	26	1.315	1.706	2.479
8	1.397	1.860	2.896	27	1.314	1.703	2.473
9	1.838	1.833	2.821	28	1.313	1.701	2.477
10	1.372	1.812	2.764	29	1.311	1.699	2.462
11	1.363	1.796	2.718	30	1.310	1.697	2.457
12	1.356	1.782	2.681	40	1.303	1.684	2.423
13	1.350	1.771	2.650	50	1.298	1.676	2.403
14	1.345	1.761	2.624	60	1.296	1.671	2.390
15	1.341	1.753	2.602	80	1.292	1.664	2.374
16	1.337	1.746	2.583	100	1.290	1.660	2.365
17	1.333	1.740	2.567	200	1.286	1.653	2.345
18	1.330	1.734	2.552	500	1.283	1.648	2.334
19	1.328	1.729	2.539		1.282	1.645	2.326

Приложение 4

Распределение Пирсона (χ^2-распределение)

ν \ α	0.10	0.005	0.002	0.01
1	2.706	3.841	5.412	6.635
2	4.605	5.991	7.824	9.210
3	6.251	7.815	9.837	11.345
4	7.779	9.488	11.668	13.277
5	9.236	11.070	13.388	15.086
6	10.645	12.592	15.033	16.812
7	12.017	14.067	16.622	18.475
8	13.362	15.507	18.168	20.090
9	14.684	16.919	19.679	21.666
10	15.987	18.307	21.161	23.209
11	17.275	19.675	22.618	24.725
12	18.549	21.026	24.054	26.217
13	19.812	22.362	25.472	27.688
14	21.064	23.685	26.783	29.141
15	22.307	24.996	28.259	30.578
16	23.542	26.296	29.296	32.000
17	24.769	27.587	30.995	33.409
18	25.989	28.869	32.346	34.805
19	27.204	30.144	33.687	36.191
20	28.412	31.410	35.020	37.566
21	29.615	32.671	36.343	38.932
22	30.813	33.924	37.659	48.268
23	32.007	35.172	38.968	41.638
24	33.196	36.415	40.270	42.980
25	34.382	37.652	41.566	44.314
26	35.563	38.885	42.856	45.642
27	36.741	40.113	44.140	46.963
28	37.916	41.337	45.419	48.278
29	39.087	42.557	46.693	49.558
30	40.256	43.773	47.962	50.892

Распределение Фишера—Снедокора
(*F*-распределение)

$$P(F \geq F_{\alpha;\nu_1;\nu_2}) = 0.05$$

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	161.45	199.50	215.57	224.57	230.17	233.97	238.89	243.91	249.04	254.32
2	18.215	18.999	19.163	19.248	19.329	19.329	19.371	19.414	19.453	19.496
3	10.129	9.552	9.276	9.118	9.014	8.941	8.844	8.744	8.638	8.527
4	7.710	6.945	6.591	6.388	6.257	6.164	6.041	5.912	5.774	5.628
5	6.607	5.786	5.410	5.192	5.050	4.950	4.818	4.678	4.527	4.365
6	5.987	5.143	4.756	4.534	4.388	4.284	4.147	4.000	3.841	3.669
7	5.591	4.737	4.347	4.121	3.972	3.866	3.725	3.574	3.410	3.230
8	5.317	4.459	4.067	3.838	3.688	3.580	3.438	3.284	3.116	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.230	3.073	2.900	2.707
10	4.965	4.103	3.408	3.478	3.326	3.217	3.072	2.913	2.737	2.538
11	4.844	3.982	3.587	3.357	3.204	3.094	2.948	2.788	2.609	2.405
12	4.747	3.885	3.490	3.259	3.106	2.999	2.848	2.686	2.505	2.296
13	4.667	3.805	3.410	3.179	3.025	2.915	2.767	2.604	2.420	2.207
14	4.600	3.739	3.344	3.112	2.958	2.848	2.699	2.534	2.349	2.131
15	4.543	3.683	2.287	3.056	2.901	2.790	2.641	2.475	2.288	2.066
16	4.494	3.634	3.239	3.007	2.853	2.741	2.591	2.424	2.235	2.010
17	4.451	3.592	3.197	2.965	2.810	2.699	2.548	2.381	2.190	1.961

Продолжение приложения 5

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
18	4.414	3.565	3.160	2.928	2.773	2.661	2.510	2.342	2.150	1.917
19	4.381	3.522	3.127	2.895	2.740	2.629	2.477	2.308	2.114	1.878
20	4.351	3.493	3.098	2.866	2.711	2.599	2.447	2.278	2.083	1.843
21	4.325	3.467	3.072	2.840	2.685	2.573	2.421	2.250	2.054	1.812
22	4.301	3.443	3.049	2.817	2.661	2.549	2.397	2.226	2.028	1.783
23	4.279	3.422	3.028	2.795	2.640	2.528	2.375	2.203	2.005	1.757
24	4.260	3.403	3.009	2.777	2.621	2.503	2.355	2.183	1.984	1.733
25	4.242	3.385	2.991	2.759	2.603	2.490	2.337	2.165	1.965	1.711
26	4.225	3.369	2.975	2.743	2.587	2.474	2.421	2.148	1.947	1.691
27	4.210	3.354	2.961	2.728	2.572	2.459	2.305	2.132	1.930	1.672
28	4.196	3.340	2.947	2.714	2.558	2.445	2.292	2.118	1.915	1.654
29	4.183	3.328	2.934	2.702	2.545	2.432	2.278	2.104	1.901	1.638
30	4.171	3.316	2.922	2.690	2.534	2.421	2.266	2.092	1.887	1.622
40	4.085	3.232	2.839	2.606	2.449	2.336	2.180	2.004	1.793	1.509
60	4.001	3.151	2.758	2.525	2.368	2.254	2.097	1.918	1.700	1.389
120	3.920	3.072	2.680	2.447	2.290	2.175	2.016	1.834	1.608	1.254
∞	3.841	2.996	2.605	2.372	2.214	2.098	1.938	1.762	1.517	1.000

для $\alpha = 0.01$

$P(F \geq F_{\alpha;\nu_1;\nu_2}) = 0.01$

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	4052.1	4999.0	5403.5	5625.1	5764.1	5859.4	5981.4	6105.8	6234.2	6366.5
2	98.495	99.008	99.167	99.247	99.305	99.325	99.365	99.425	99.464	99.504
3	34.117	30.815	29.459	28.709	28.236	27.910	27.489	27.053	26.597	26.122
4	21.200	18.001	16.693	15.978	15.521	15.208	14.800	14.374	13.930	13.464
5	16.258	13.274	12.059	11.391	10.966	10.672	10.266	9.888	9.467	9.019
6	13.744	10.924	9.779	9.149	8.746	8.465	8.101	7.718	7.313	6.880
7	12.246	9.564	9.452	7.846	6.460	7.191	6.840	6.469	6.074	5.650
8	11.259	8.649	7.591	7.006	6.631	6.371	6.029	5.667	5.279	4.650
9	10.561	8.022	6.992	6.423	6.057	5.802	5.467	6.111	4.730	4.311
10	10.044	7.560	6.552	5.994	5.636	5.386	5.057	4.706	4.327	3.909
11	9.647	7.205	6.217	5.668	5.317	5.069	4.745	4.397	4.021	3.602
12	9.330	6.927	5.953	5.412	5.064	4.820	4.500	4.156	3.780	3.361
13	9.074	6.701	5.740	5.205	4.862	4.620	4.302	3.961	3.586	3.165
14	8.862	6.514	5.563	5.035	4.695	4.456	4.140	3.800	3.427	3.005
15	5.683	6.359	5.417	4.893	4.556	4.318	4.004	3.668	3.294	2.869

Продолжение приложения 5

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
16	8.532	6.227	5.292	4.772	4.437	4.201	3.889	3.553	3.181	2.753
17	8.400	6.112	5.185	4.669	4.336	4.102	3.791	3.455	3.083	2.653
18	8.285	6.013	5.092	4.579	4.248	4.015	3.706	3.370	2.999	2.566
19	8.184	5.926	5.010	4.501	4.170	3.939	3.631	3.296	2.925	2.489
20	8.096	5.849	4.938	4.431	4.103	3.871	3.565	3.231	2.859	2.421
21	8.017	5.780	4.875	4.368	4.042	3.811	3.506	3.173	2.801	2.360
22	7.944	5.719	4.816	4.314	3.988	3.759	3.453	3.121	2.749	2.305
23	7.881	5.663	4.765	4.264	3.939	3.710	3.406	3.074	2.702	2.256
24	7.823	5.614	4.718	4.218	3.895	3.666	3.363	3.031	2.659	2.210
25	7.770	5.568	4.676	4.177	3.855	3.627	3.324	2.993	2.620	2.169
26	7.722	5.527	4.637	4.140	3.818	3.591	3.288	2.958	2.585	2.132
27	7.677	5.488	4.601	4.106	3.785	3.558	3.256	3.925	2.551	2.096
28	7.636	5.453	4.568	4.074	3.754	3.528	3.226	2.896	2.522	2.064
29	7.597	5.421	4.538	4.045	3.726	3.499	3.198	2.869	2.494	2.034
30	7.563	5.390	4.510	4.018	3.699	3.474	3.173	2.843	2.469	2.006
40	7.314	5.179	4.312	3.828	3.512	3.291	2.993	2.665	2.287	1.805
60	7.077	4.978	4.126	3.649	3.339	3.119	2.823	2.496	2.115	1.601
120	6.851	4.786	3.949	3.479	3.173	2.956	2.663	2.336	1.950	1.380
∞	6.635	4.605	3.782	3.320	3.017	2.802	2.511	2.182	1.791	1.000

Предметный указатель

Алгоритм
- взаимного поглощения 496
- классификации 474
- типа «Форель» 494

Анализ
- внешний 447
- внутренний 447
- дискриминантный 21, 507
- кластерный 21, 468
- корреляционный 21, 215
- регрессионный 21, 216
- факторный 21, 333

Базис 40
Биллинеарность 82

Вектор 32
- собственный 70

Величина
- случайная 89

Винзорированные данные 201
Вращение факторного пространства 386—390
- косоугольное 344, 387
- ортогональное 344, 387

Выборка
- обучающая 513
- случайная 172

Гиперсфера 494
- радиус гиперсферы 494
- центр гиперсферы 494

Гипотеза статистическая 172
- альтернативная 172
- иерархия гипотез 282
- нуль-гипотеза 173
- о равенстве векторов 185
- о равенстве ковариационных матриц 190
- простая 173
- прямая 172
- семейство гипотез 236
- сложная 173

Дендрограмма 474
Дискриминантный анализ 21, 507
Дисперсия 117, 118
- многомерной случайной величины 347
- признаковых значений 117
- разложение дисперсии в факторном анализе 369, 370
- разложение дисперсии в регрессионном анализе 234, 235

Доверительная область 173, 175
- совместная 180, 181

Закон
- ассоциативный 33
- дистрибутивный 33
- коммутативный 33
- распределения случайной величины 90, 96
- слабый больших чисел 228

Значимость канонических корреляций 534

Идеальная точка 445, 456
Инструментальное оценивание 246
Интервал
- величина интервала 236
- доверительный 177, 180, 236
- оценивание интервала 236

Квадратичная форма 82
Классификация
- двумерная (линейной модели *ANOVA)* 293
- многомерная 468
- одномерная 468

Кластер 468
- вес кластера 487

Кластерный анализ
- агломеративный 474
- дивизимный 484
- иерархический 474
- итеративный 486

Ковариация 126, 137
Константа
- аддитивная 223
- дискриминации 513

Корреляция
- каноническая 527
- множественная 216
- парная 215
- частная 215

Коэффициент
- аппроксимации (*MAPE*) 221
- вариации 121, 122
- весовой 433
- дискриминантной функции 511
- информативности 358
- корреляции 529
- ковариации 128
- конгруэнтности 395
- множественной детерминации 216
- множественной корреляции 216
- неопределенности 216
- определенности 130, 255
- парной корреляции 215
- первичных затрат 62, 63
- прямых затрат 62, 63
- регрессии 216
- структурный 520
- частной корреляции 215

Критерий
- Бартлетта 190
- варимакс 390
- Граббса 198
- квартимакс 390
- квартимин 391
- критерии Титьена и Мура (L, L', E) 199, 200
- линейной независимости 39
- облимакс 391
- облимин 391
- односторонний 237, 238
- регрессионного параметра 221
- симметрический236—238
- сравнения двух средних величин 183
- средней величины 177
- статистический 173
- функционал 498
- χ^2-Бартлетта 394
- χ^2-Лоули 394
- χ^2-Уилкса 393
- F-Фишера 221
- *MSE* 157
- *STRESS* 430
- t-Стьюдента 177, 183, 222
- T^2-Хотеллинга 179, 186

Лаг
- Альмона 269

- оценивания 268
- репараметризации 268
- треугольный 269

Линейная зависимость векторов 39
Линейная независимость векторов 39
Линейная оболочка 38
Линейные комбинации 37

Математическое ожидание 115, 116
Матрица
- Вандермонда 306
- весовых коэффициентов 435, 437
- воспроизведенная корреляционная 341
- Гессе 144, 225
- диагональная 351
- значений факторов 346
- идемпотентная 266
- квадратная 43, 57
- конечного спроса 61
- коэффициентов первичных затрат 62, 63
- коэффициентов прямых затрат 62
- межотраслевых потоков 59, 61
- нормированных векторов 352
- обратная57, 58
- остатков корреляций 373, 377
- парных корреляций 215, 350
- различий 412, 414
- расстояний 470
- расширенная 51, 52
- расширенная ковариаций 271
- симметрическая 82, 411
- скалярных произведений (с двойным центрированием) 417
- след матрицы 351
- собственных векторов 352
- собственных чисел (диагональная) 351
- совместных вероятностей 411
- транспонированная 56
- условных вероятностей 407
- факторного отображения 344, 353
- частных корреляций 215, 216

Мера
- различия профилей 412
- сходства 474

Метод
- α-факторного анализа 338
- *k*-средних 486
- Барта 344
- ближайшего соседа 480
- главных компонент 349, 350
- главных факторов 372
- групповой 338
- дальнего соседа 480
- дискриминантного анализа (методы) 507
- исходной (якорной) точки 423
- канонических корреляций 526
- кластерного анализа (методы) 468
- максимального правдоподобия (в факторном анализе) 140
- максимального правдоподобия 223
- малого центроида 344
- матричных проекций 245, 248
- минимальных остатков 339
- моментов 153, 229
- наибольшей корреляции 343
- наименьших квадратов 228, 243
- одиночной связи 475
- поиска сгущений 493
- полных связей 476
- последовательной рандомизации 423

- программирования 245
- рейтинговой оценки 423
- средней связи 477, 480
- триад 344
- Уорда 477
- центроидный 337

Метрика 471
- доминирования 422
- евклидово расстояние 471
- Махаланобиса 471
- Минковского 471
- city-block (l_1 - норма) 471

Многомерное шкалирование
- метрическое 404, 415
- неметрическое 404, 421

Множество весовых коэффициентов 433, 473
Модель
- авторегрессионная 243, 301
- аналитическая модель затрат, выпуска 63, 64
- *ANOVA* 290
- без константны 227
- билинейная 309
- бифакторная Г. Хользингера 337
- векторная 406, 447
- взвешенная евклидова 435, 447
- гетероскедастичная 242, 300
- гомоскедастичная 241
- дистанционная 406
- Койка 271, 302
- линейная 223
- линии роста 307
- на значениях латентных (общих) факторов 364
- на стандартизированных значениях признаков 218
- на центрированных значениях признака 225, 231
- нелинейная 304, 305, 306
- обобщенная евклидова 447
- однофакторная Ч. Спирмена 337
- регрессионная 216
- с эквивалентной корреляцией остатков 242, 301
- Торгерсона 415—417
- трехмодальная 434, 443
- факторного анализа 340, 341

Мультиколлинеарность 284, 285

Надежность 370, 371
Ненадежность 370, 371
Нелинейность 304
Неравенство 229
- Коши—Шварца 77
- Маркова 117
- Шварца 286

Несмещенность
- асимптотическая 155

Норма вектора 75
Нормирование статистических данных 472

Область взаимного поглощения 496, 497
Область доверительная 174—176
Общность 370
Ограничения
- в линейной регрессионной модели 274
- в нелинейной регрессионной модели 283
- гомогенные 275
- негомогенные 276
- нуль-ограничения 275, 282
- ранговые 286, 288

Определитель матрицы (детерминант) 64, 65
Ошибка грубая 197
Оценивание
- интервальное 236
- точечное 139
- линейное 232, 265
- методом наименьших квадратов 239, 228

- методом моментов 153, 229
- методом максимального правдоподобия 140, 226
- по Айткину 264
- ридж 289

Оценки координат
- стандартизованные 422
- нестандартизованные 422, 426

Пакеты программ многомерного статистического анализа 542
Параметр
- статистического распределения 120—123
- регрессионной модели 217, 225
- начального разбиения 486, 494
- оценки сходства 401, 474

Параметрический вектор 217
Переменная
- дискриминантная 508
- зависимая (отклик) 214
- каноническая 526
- латентная 333
- независимая (предиктор) 214
- экзогенная 269
- эндогенная 273, 315

Плотность вероятностей 92, 93
Подмножество весовых коэффициентов 358, 359
Подпространство 36
Поиск индивидуальных различий 406, 432
Подход Гаусса—Маркова 232, 235
Правило Крамера 69
Предпочтение 405
Преобразования элементарные 46, 52
Признак
- латентный 334
- элементарный 333, 334

Произведение матриц скалярное 73
Пространство
- векторное 33, 34
- главных компонент 350, 362
- двумерное 28
- многомерное 29, 30
- нуль 30, 34
- общих факторов 334, 345
- одномерное 28
- ортонормированное 76, 350
- трехмерное 29
- шкальное 405

Принцип
- лезвия Оккама 230
- максимального правдоподобия Фишера 140

Процесс
- *VAR /p/* 272

Равенство Тэрстоуна 340
Ранг матрицы 46
Распределение
- Бернули 97
- Бета 112
- биномиальное 97, 98
- Вейбулла 107
- Вейбулла-гамма 114
- гамма 113
- геометрическое 99
- гипергеометрическое 100
- гипер-экспоненциальное 108
- дискретное 90, 91
- двумерное экспоненциальное 105
- двумерное нормальное 133
- двустороннее экспоненциальное 108, 109
- Коши 111
- логнормальное 110, 111
- логистическое 313
- нормальное (Гаусса) 109, 110
- Паскаля 100
- Парето 104, 105
- Пуассона 103
- плотности вероятностей 92, 93

- равномерное дискретное 96
- совместное 131
- Стьюдента (*t*-критерия)
- треугольное 104
- табулированных критериальных значений 138
- условное 132
- экспоненциальное 105, 106
- Эрланга 107
- *F*-критерия Фишера 112, 113
- χ-квадрат 113, 114

Регрессия
- линейная 223
- множественная 216, 250
- нелинейная 304
- нормальная с одним параметром 223
- нормальная с числом параметров $m \geq 1$
- обратная 230, 255
- полиномиальная 220, 305
- смешанная 277

Рекуррентные соотношения Фаддеева 351, 354
Репараметризация
- данных (шкалирование) 264, 268
- параметрического вектора 268

Робастное (устойчивое) оценивание 196
Робастные статистические оценки
- Винзора 201
- Пауанкаре 200
- Хубера 201, 205

Свойства
- определителя матрицы 64, 66
- статистических оценок 154, 261

Система
- уравнений 49, 52
- координат 28—30

Случайная величина 89
- двумерная 123, 131
- дискретная 90
- многомерная 136
- независимость (зависимость) случайных величин 126, 137
- непрерывная 103
- нормированная 472
- одномерная 89
- распределение случайной величины 90
- стандартизированная 122
- центрированная 265, 349

Смещенность статистической оценки 155, 234
Совокупность
- выборочная 172
- генеральная 172

Согласованность статистических оценок 159
Состоятельность статистической оценки 157, 158
Среднее квадратическое отклонение 177
Средняя квадратическая ошибка (стандартное отклонение) 157, 219
Специфичность 370, 371
Стартовая конфигурация 422, 425
Статистические оценки
- авторегрессионной модели 243, 301
- *BLU* 261
- *GLS* 265, 302
- итеративные 309—311
- линейные 245
- оптимальные 233

Стимул 404

Таблица затраты-выпуск 59
Трансформация
- базиса 41, 42
- данных 305, 306
- линейная 137, 265

- логит 313
- монотонная 141
- по Айткину 264
- случайной величины 118

Фактор
- общий 368, 371
- характерный 371

Факторные нагрузки 341
Формула
- Гуттмана 430
- Краскала 430
- Лингоса—Роскама 429
- Стресс 405, 430
- Юнга 430

Функция
- авторегрессионная 243
- дискриминантная 509
- дистанционная 406
- Кобба—Дугласа 283, 308
- максимального правдоподобия 140, 141
- монотонная 91, 421
- непрерывная 34
- предпочтений линейная 421
- распределения плотности вероятностей 93
- тренда 308
- циклических колебаний 317

Форма
- билинейная симметрическая 81
- квадратичная 81, 82

Характеристическое уравнение 70
Характерность 370

Число собственное 70
Числовое поле 30

Шкалирование данных 402

Эластичность
- шкалы признаковых значений 284

Эффективность
- асимптотическая157
- статистической оценки 156, 262

Учебное пособие

Сошникова Людмила Антоновна,
Тамашевич Виктор Николаевич,
Уебе Готц,
Шефер Мартин

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ В ЭКОНОМИКЕ

Редактор *Л.Н. Вылегжанина*
Корректор *К.В. Федорова*
Оформление художника *А.В. Лебедева*

Оригинал-макет выполнен в
ИЗДАТЕЛЬСТВЕ ЮНИТИ-ДАНА

Лицензия № 071252 от 04.01.96
Подписано в печать 23.07.99. Формат 60x88 1/16
Усл. печ. л. 37,5. Уч.-изд. л. 24,3
Тираж 10000 экз. (1-й завод - 5000). Заказ 1476

ООО "ИЗДАТЕЛЬСТВО ЮНИТИ-ДАНА"
Генеральный директор *В.Н. Закаидзе*

123298, Москва, Тепличный пер., 6
Тел.: (095) 194-00-15. Тел./Факс: (095) 194-00-14
E.mail: unity@tech.ru

Отпечатано в ГУП ИПК "Ульяновский дом печати"
432601, г. Ульяновск, ул. Гончарова, 14

ISBN 5-238-00099-5

ИНТЕРНЕТ

- ПОДКЛЮЧЕНИЕ
- ПОЧТОВЫЙ ЯЩИК (E-mail)
- ЛИЧНАЯ СТРАНИЧКА
- ВАША РЕКЛАМА В ИНТЕРНЕТЕ
- ЦИФРОВАЯ СВЯЗЬ
- ЛЕГКИЙ ДОЗВОН

а также:

- WEB-ХОСТИНГ
- WEB-ДИЗАЙН
- РЕГИСТРАЦИЯ ДОМЕНОВ

30 $

ООО "Топ Технолоджи"

Тел./Факс: 723 90 01

E-mail: info@tech.ru
WWW: http://www.tech.ru

МЫ ИСПОЛЬЗУЕМ НОВЕЙШИЕ ТЕХНОЛОГИИ