Л.А. Сошникова, В.Н. Тамашевич, Л.А. Махнач

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

Практикум

Минск БГЭУ 2004

Л.А. Сошникова, В.Н. Тамашевич, Л.А. Махнач

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

Практикум

Допущено
Министерством образования Республики Беларусь
в качестве учебного пособия
для студентов специальности «Статистика»
учреждений образования, обеспечивающих получение
высшего образования

Минск БГЭУ 2004

УДК 657:381
ББК 65.052.242
С88

Рецензенты: заместитель директора по науке НИИ статистики, профессор *Л.П. Шахотько*; кафедра статистики и экономического анализа БГСХА (зав. кафедрой — доцент *Б.М. Шундалов*)

Сошникова Л.А.

С88 Многомерный статистический анализ: Практикум / Л.А. Сошникова, В.Н. Тамашевич, Л.А. Махнач. — Мн.: БГЭУ, 2004. — 162 с.

ISBN 985-426-973-6.

Практикум подготовлен как дополнение к учебному пособию для вузов «Многомерный статистический анализ в экономике» (под редакцией В.Н. Тамашевича). Содержит краткие методические рекомендации, решения типовых задач и контрольные задания по всем темам курса. Кроме того, по основным и наиболее трудоемким темам приведены методические рекомендации по решению задач на компьютере с использованием пакета Statistica.

Для студентов, аспирантов и преподавателей экономических вузов. Может быть полезным специалистам, занимающимся анализом статистических данных.

УДК 657:381
ББК 65.052.242

ISBN 985-426-973-6

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Курс многомерного статистического анализа (МСА) является неотъемлемой частью программы университетской подготовки современного экономиста. Этот курс значительно расширяет возможности социально-экономических исследований, включая математико-статистическое моделирование сложных явлений и процессов, происходящих в обществе.

Настоящий практикум представляет собой дополнение к учебному пособию «Многомерный статистический анализ в экономике», изданному в 1999 г. Содержание практикума определено типовой программой по дисциплине «Многомерные статистические методы» для специальности 1-25 01 05 «Статистика». Основная цель — помочь приобрести навыки применения на практике методов МСА и интерпретации аналитических результатов.

При определении структуры практикума принималось во внимание, что почти все многомерные статистические методы имеют объемный математический аппарат и сложные алгоритмы расчетов, требующие четкой постановки задач и использования вычислительной техники.

Последовательность расположения тем в практикуме обусловлена следующим: сначала рассматриваются традиционные статистические методы, широко используемые в аналитической практике, а затем приведены наиболее сложные и редко встречающиеся, например, многомерное шкалирование, кластерный и дискриминантный анализ, метод канонических корреляций.

Главы практикума разделены на параграфы, включающие краткое изложение теоретических основ и алгоритмы методов МСА, примеры решения типовых задач, наборы задач для практических занятий и самостоятельного решения. По отдельным темам приводятся также рекомендации по решению типовых задач на компьютере с использованием пакета STATISTICA.

Условия задач в пособии сформулированы таким образом, чтобы показать широкие возможности практического приложения того или иного аналитического метода.

В приложениях к практикуму приведены основные математико-статистические таблицы, необходимые при решении задач.

Авторы практикума:

Л.А. Сошникова — кандидат экономических наук, доцент (параграфы 3.1, 3.3, 3.4, 6.1, 6.3, 6.4, 7.1, 7.3, 7.4, 8.1—8.4); В.Н. Тамашевич — кандидат экономических наук, доцент (параграфы 1.1, 1.2, 2.1—2.4, 4.1—4.3, 5.1—5.3); Л.А. Махнач — ассистент (параграфы 3.2, 6.2, 7.2).

1. ПРОВЕРКА МНОГОМЕРНЫХ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

1.1. Методические рекомендации

Статистические гипотезы — это выдвигаемые теоретические предположения относительно параметров статистического распределения или закона распределения случайной величины. В соответствии с решаемой задачей различают параметрические и непараметрические гипотезы.

При проверке статистических гипотез используются понятия 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) принимается после того, как отвергнута основная. Ниже приведены примеры статистических гипотез, которые формулируются для проверки равенства параметров нормально распределенной одномерной и многомерной случайных величин (табл. 1.1).

Таблица 1.1

	Нулевые гипотезы	Альтернативные гипотезы
Одномерная случайная величина	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ $H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma \neq \sigma_0$	$\mu < \mu_0$, $\mu > \mu_0$ $\sigma^2 < \sigma_0^2$, $\sigma^2 > \sigma_0^2$
Многомерная случайная величина	$H_0: \mu_j = \mu_{0j}$ $H_0: \Sigma = \Sigma_0$	$H_1: \mu_j = \mu_{0j}$ $H_1: \Sigma = \Sigma_0$

В статистике рассматривают простые и сложные параметрические гипотезы. *Простая гипотеза* содержит только одно предположение относительно оцениваемого параметра, например, предположение о том, что среднее значение j-го признака $\overline{X}_j$ равно нулю: H_0: $\overline{X}_j = 0$, или H_1: $\overline{X}_j \neq 0$. *Сложная гипотеза* состоит из нескольких простых гипотез. Например, H_0: $\overline{X}_j > 0$ — это означает, что могут быть H_0: $\overline{X}_j = 1$; H_0: $\overline{X}_j = 2$ и т.д., т.е. здесь гипотеза состоит из набора простых гипотез.

Для того чтобы проверить гипотезу, используют статистические критерии, позволяющие выяснить, следует ли принять или отвергнуть нулевую гипотезу. Если расчетное значение критерия не превышает критического, то есть веские основания для принятия прямой (нулевой) гипотезы. В противоположном случае целесообразно предположить справедливость альтернативной гипотезы (H_1).

Проверка статистических гипотез всегда допускает определенную вероятность ошибки в выводах:

α — вероятность отвергнуть нулевую гипотезу, когда она справедлива;

$\beta = 1 - \alpha$ — вероятность принять нулевую гипотезу, когда она ложна.

В экономических исследованиях обычно используется α — вероятность ошибки первого рода. Наиболее распространенными в практике экономического анализа значениями α являются: 0,001; 0,005; 0,1.

В многомерном анализе для проверки статистических гипотез используются те же статистические критерии, что и в одномерном, но они изменяются с учетом природы многомерных случайных величин. Чаще всего это критерии для проверки параметрических гипотез: t-Стьюдента, F-Фишера, и проверки непараметрических гипотез-χ^2.

1.1.1. Проверка гипотезы о равенстве вектора средних значений заданному вектору

В многомерном статистическом анализе проверка гипотезы о равенстве вектора средних значений заданному вектору основывается на тех же подходах, что и для одномерных величин. Но в этом случае мы имеем дело уже с m-числом выборочных средних, т.е. с вектором средних значений: $\overline{X} = (\overline{X}_1, \overline{X}_2, ...\overline{X}_m)$. Вектор $\overline{X}$ сравнивается с постоянным вектором $\mu = (\mu_1, \mu_2, ..., \mu_m)$. Прямая гипотеза имеет вид

$$H_0: \overline{X} = \mu, \text{ при альтернативной } H_1: \overline{X} \neq \mu.$$

Для проверки многомерной гипотезы данного вида используется критерий, известный как критерий Хотеллинга

$$T_p^2 = n(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu), \tag{1.1}$$

где $\Sigma = \frac{1}{n-1}(\hat{X}^T \hat{X})$ — ковариационная матрица; X — матрица с центрированными значениями переменной: $\hat{x}_{ij} = x_{ij} - \overline{x}_{ij}$.

Расчетное значение (T_p^2) сравнивается с критическим значением, исчисляемым при заданном уровне вероятности (α) и числе степеней свободы $\nu_1 = m$ и $\nu_2 = n - m$

$$T_{\alpha,m,n-m}^2 = \frac{m(n-1)}{n-m} F_{\alpha,m,n-m}. \tag{1.2}$$

В формуле (1.2) $F_{\alpha,m,n-m}$ — табличное значение F-критерия Фишера для числа степеней свободы $\nu_1 = m$ и $\nu_2 = n - m$. Многомерная гипотеза о равенстве вектора средних величин заданному вектору подтверждается при $T_p^2 < T_{кр(\alpha,m,n-m)}^2$.

Пример 1. Для предприятий розничной торговли в административном районе установлены следующие нормативные показатели эффективности деятельности: средний уровень рентабельности хозяйственной деятельности — 20 % и средняя продолжительность оборота оборотных средств — 12 дней. Предположим, что более низкие значения уровня рентабельности и скорости оборота означают нарушение ритмичности товарно-денежных операций и снижение конкурентоспособности предприятий торговли.

С целью оперативного контроля за результатами коммерческой деятельности торговых предприятий района проведен анализ эффективности торговой деятельности и получены следующие данные (табл. 1.2).

Таблица 1.2

Рентабельность торговой деятельности и оборачиваемость оборотных средств предприятий

Номер объекта	Рентабельность, %	Продолжительность оборота, дней
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	19	20
11	12	22
12	14	23
Среднее значение $(\bar{X}_j)$	15,8	17,8

Оценим существенность различий между фактическими значениями рассматриваемых показателей и установленными нормативами. Уровень значимости α зададим равным 0,05.

Решение. Определим следующие параметры многомерной случайной величины:

вектор средних значений $\bar{X} = (15{,}8\ 17{,}8)$;

ковариационная матрица $\Sigma = \frac{1}{n-1}(\hat{X}^T \hat{X})$

$$\Sigma = \frac{1}{11} \times \begin{pmatrix} 14-15{,}8 & 12-15{,}8 & 16-15{,}8 & 14-15{,}8 & 15-15{,}8 & \dots & 14-15{,}3 \\ 19-17{,}8 & 15-17{,}8 & 19-17{,}8 & 17-17{,}8 & 24-17{,}8 & \dots & 23-17{,}8 \end{pmatrix} \times$$

$$\times \begin{pmatrix} 14-15{,}8 & 19-17{,}8 \\ 12-15{,}8 & 15-17{,}8 \\ 16-15{,}8 & 19-17{,}8 \\ 14-15{,}8 & 17-17{,}8 \\ 15-15{,}8 & 24-17{,}8 \\ \vdots & \vdots \\ 14-15{,}8 & 23-17{,}8 \end{pmatrix} = \frac{1}{11} \times \begin{pmatrix} 118{,}3 & -86{,}5 \\ -86{,}5 & 201{,}7 \end{pmatrix} = \begin{pmatrix} 10.75 & -7{,}86 \\ -7{,}86 & 18{,}34 \end{pmatrix}.$$

Обратная ковариационная матрица Σ^{-1} будет равна

$$\Sigma^{-1} = \frac{1}{|\Sigma|} a_{ij}\ \Sigma = \begin{pmatrix} 0{,}1355 & 0{,}0581 \\ 0{,}0581 & 0{,}1093 \end{pmatrix}.$$

Рассчитаем фактическое значение T^2-критерия Хотеллинга

$$T_н^2 = n(\bar{X}-\mu)^T \Sigma^{-1}(\bar{X}-\mu) = 12\ (15{,}8\ -\ 20\ \ 17{,}8\ -\ 12) \times \begin{pmatrix} 0{,}1355 & 0{,}0581 \\ 0{,}0581 & 0{,}1093 \end{pmatrix} \times$$

$$\times \begin{pmatrix} 15{,}8 & -20 \\ 17{,}8 & -12 \end{pmatrix} = 32{,}16.$$

Критическое значение для заданного уровня значимости $\alpha = 0{,}05$ составит

$$T_{кр}^2 = \frac{m(n-1)}{n-m} F_{0{,}05;2;10} = \frac{2(12-1)}{12-2} \times 4{,}459 = 9{,}8.$$

Как видим, расчетное значение T_p^2-критерия почти в три раза превосходит критическое (32,16 > 9,8), что свидетельствует о существенности расхождения между фактическими и нормативными значениями анализируемых показателей.

1.1.2. Проверка гипотезы о равенстве двух векторов средних значений

В многомерном статистическом анализе проверяется гипотеза о равенстве векторов средних значений многомерных величин

$$H_0 : (\bar{X}_{11}\bar{X}_{12}\bar{X}_{13}...\bar{X}_{1m}) = (\bar{X}_{21}\bar{X}_{22}\bar{X}_{23}...\bar{X}_{2m}),$$

$$H_1 : (\bar{X}_{11}\bar{X}_{12}\bar{X}_{13}...\bar{X}_{1m}) \neq (\bar{X}_{21}\bar{X}_{22}\bar{X}_{23}...\bar{X}_{2m}),$$

или в векторной форме

$$H_0 : \bar{X}_1 = \bar{X}_2; \quad \text{и} \quad H_1 : \bar{X}_1 \neq \bar{X}_2.$$

Для проверки данной гипотезы применяется многомерный T^2-критерий, исчисляемый по формуле

$$T_p^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^T \Sigma_*^{-1} (\bar{X}_1 - \bar{X}_2), \qquad (1.3)$$

где $\bar{X}_1$, $\bar{X}_2$ — векторы средних значений; Σ_*^{-1} — обратная матрица, рассчитанная для объединенной ковариационной матрицы

$$\Sigma_* = \frac{1}{n_1 + n_2 - 2} \left(\hat{X}_1^T \hat{X}_1 + \hat{X}_2^T \hat{X}_2 \right),$$

где $\hat{X}$ — матрица центрированных значений $\hat{x}_{ij} = x_{ij} - \bar{x}_j$.

Критические значения для T^2 находят по формуле

$$T_{кр\,(\alpha, m, n_1+n_2-m-2)}^2 = \frac{(n_1 + n_2 - 2)m}{n_1 + n_2 - m - 2} \times F_{\alpha, m, n_1+n_2-m-2}. \qquad (1.4)$$

При $T_p^2 < T_{кр(\alpha,m,n_1+n_2-m-2)}^2$ нулевая гипотеза $H_0: \overline{X}_1 = \overline{X}_2$ принимается с вероятностью $(1-\alpha)$. Если же $T_p^2 > T_{кр(\alpha,m,n_1+n_2-m-2)}^2$, то нулевая гипотеза о равенстве векторов средних значений отвергается.

Пример 2. С целью анализа различий показателей производственно-хозяйственной деятельности родственных предприятий, расположенных в рамках свободной экономической зоны (первая группа) и за ее пределами (вторая группа), было проведено выборочное обследование.

Из каждой группы предприятий были сформированы две выборки неравных объемов. Результаты выборочного наблюдения представлены в табл. 1.3.

Таблица 1.3

Первая группа			Вторая группа		
Номер объекта	X_1	X_2	Номер объекта	X_1	X_2
1	25	3,25	1	32	2,90
2	20	2,85	2	30	2,94
3	35	2,90	3	41	3,00
4	28	3,25	4	34	2,75
5	40	4,90	5	20	3,30
6	31	2,65	6	46	3,43
7	28	4,00	7	35	2,80
8	48	3,90			
9	50	5,24			

В таблице X_1 — валовая добавленная стоимость на одного работника, тыс. ден. ед.; X_2 — фондоотдача основных производственных фондов, ден. ед.

По приведенным выше данным следует оценить существенность различий двух групп предприятий по X_1 и X_2 при $\alpha = 0{,}01$.

Решение. Для того чтобы решить поставленную задачу, проверим гипотезу о равенстве векторов средних значений двух выборочных совокупностей.

1. Определим векторы средних значений и совместную ковариационную матрицу S_*, необходимые в последующем для расчета T^2-критерия

$$\overline{X}_1 = (34 \quad 3{,}66); \qquad \overline{X}_2 = (34 \quad 3{,}02),$$

$$S_* = \frac{1}{n_1+n_2-2}(\hat{X}_1^T\hat{X}_1 + \hat{X}_2^T\hat{X}_2) = \frac{1}{9+7-2}\left[\begin{pmatrix} 836 & 54{,}14 \\ 54{,}14 & 5{,}77 \end{pmatrix} + \begin{pmatrix} 410 & 1{,}20 \\ 1{,}2 & 0{,}39 \end{pmatrix}\right] =$$

$$= \frac{1}{14}\begin{pmatrix} 1246 & 55{,}34 \\ 55{,}34 & 6{,}16 \end{pmatrix} = \begin{pmatrix} 89 & 3{,}95 \\ 3{,}95 & 0{,}44 \end{pmatrix},$$

следовательно, обратная матрица равна $S_*^{-1} = \begin{pmatrix} 0{,}0187 & -0{,}1677 \\ -0{,}1677 & 3{,}7780 \end{pmatrix}$.

2. Теперь можно рассчитать наблюденное (расчетное) значение общего T^2-критерия Хотеллинга:

$$T_p^2 = \frac{n_1 n_2}{n_1+n_2}(\overline{X}_1 - \overline{X}_2)^T S_*^{-1}(\overline{X}_1 - \overline{X}_2) =$$

$$= \frac{63}{16}(0 \;\; 0{,}64) \times \begin{pmatrix} 0{,}0187 & -0{,}1677 \\ -0{,}1677 & 3{,}7780 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0{,}64 \end{pmatrix} = 6{,}093.$$

3. Найдем критическое значение T^2-критерия и произведем сравнение значений $T_{кр}^2$ и T_p^2

$$T_{кр}^2 = \frac{(n_1+n_2-2)m}{n_1+n_2-m-1} \times F_{0{,}1;2;17} = \frac{28}{13} \times 3{,}8 = 8{,}185.$$

Поскольку рассчитанное значение $T_p^2 = 6{,}093$ меньше $T_{кр}{}^2 = 8{,}185$, нулевую гипотезу о равенстве векторов средних значений следует принять и сделать вывод о несущественном влиянии различных условий функционирования предприятий.

Пример 3. Для того чтобы оценить уровень различия двух групп инженерно-технических работников на предприятиях легкой промышленности (мужчины и женщины) по двум признакам, было проведено выборочное обследование. В выборку попали 20 человек (10 мужчин и 10 женщин). Результаты наблюдения представлены в табл. 1.4.

Таблица 1.4

Мужчины		Женщины	
X_1	X_2	X_1	X_2
4	9,0	20	7,2
15	8,2	8	6,0
17	10,0	12	9,2
20	8,0	6	6,5
22	6,5	5	7,5
30	8,5	18	8,0
20	7,5	15	8,4
7	7,0	25	9,0
18	9,7	10	8,1
4	8,4	17	7,8

В таблице X_1 — стаж работы, лет; X_2 — средняя дневная заработная плата, ден. ед.

Проверьте гипотезу о сходстве двух групп работающих: а) по каждому признаку отдельно; б) по двум признакам вместе.

Решение. 1. Для каждой группы отдельно и для совокупности в целом рассчитаем средние значения и дисперсии анализируемых признаков.

I группа (мужчины)

$\overline{X}_1 = 15{,}7$ лет; $\sigma_1^2 = 63{,}810$; $\overline{X}_2 = 8{,}28$ ден. ед.; $\sigma_2^2 = 1{,}106$.

Ковариационная матрица S_1 равна

$$S_1 = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 63{,}810 & -0{,}576 \\ -0{,}576 & 1{,}106 \end{pmatrix}.$$

II группа (женщины)

$\overline{X}_1 = 13{,}6$лет; $\sigma_1^2 = 38{,}24$; $\overline{X}_2 = 7{,}77$ ден. ед.; $\sigma_2^2 = 0{,}926$.

Ковариационная матрица S_2 равна

$$S_2 = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 38{,}240 & 3{,}078 \\ 3{,}078 & 0{,}926 \end{pmatrix}.$$

2. Вычислим совместную ковариационную матрицу для двух групп (S_*)

$$S_* = \frac{1}{n_1 + n_2 - 2}(S_1 n_1 + S_2 n_2) = \frac{1}{10+10-2}\begin{pmatrix} 1020{,}5 & 25{,}0 \\ 25{,}0 & 20{,}3 \end{pmatrix} = \begin{pmatrix} 56{,}7 & 1{,}4 \\ 1{,}4 & 1{,}1 \end{pmatrix},$$

$$S^{-1} = \begin{pmatrix} 0{,}018 & -0{,}022 \\ -0{,}022 & 0{,}913 \end{pmatrix}.$$

Вычислим расчетное значение T^2-критерия Хотеллинга

$$T_p^2 = \frac{n_1 n_2}{n_1 + n_2 - 2}(\overline{X}_1 - \overline{X}_2)' S_*^{-1}(\overline{X}_1 - \overline{X}_2) =$$

$$= \frac{10 \cdot 10}{10+10-2}(2{,}1 \;\; 0{,}51) \times \begin{pmatrix} 0{,}018 & -0{,}022 \\ -0{,}022 & 0{,}913 \end{pmatrix} \times \begin{pmatrix} 2{,}10 \\ 0{,}51 \end{pmatrix} = 1{,}49.$$

Табличное значение критерия

$$T_{кр}^2 = \frac{(n_1 + n_2 - 2)m}{(n_1 + n_2 - m - 2)} F_{\alpha, m, (n_1 + n_2 - m - 2)} =$$

$$= \frac{(10+10-2)\cdot 2}{(10+10-2-2)} \cdot 3{,}592 = 8{,}082 \quad > \quad T_p^2 = 1{,}49.$$

Следовательно, нулевая гипотеза о равенстве векторов средних значений двух множеств принимается, т.е. расхождения между работающими мужчинами и женщинами по изучаемым признакам несущественные.

1.1.3. Проверка гипотезы о равенстве ковариационных матриц

Сравнение ковариационных матриц, отражающих взаимосвязи изучаемых признаков, открывает возможность дополнить и уточнить гипотетические предположения относительно самих признаков. Это имеет особое 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 банка, при группировке предприятий по уровню устойчивости финансового положения или при оценке эффективности производственной и коммерческой деятельности.

Решения многомерными методами статистики большинства задач изначально предполагают равенство ковариационных матриц различных выборочных совокупностей. Например, в дискриминантном анализе рассматриваются две генеральные совокупности, имеющие многомерный нормальный закон распределения и равные ковариационные матрицы.

На практике учет ковариаций изучаемого комплекса признаков и проверка равенства ковариационных матриц значительно снижает вероятность появления ошибки в выводах. Это происходит из-за весьма малой вероятности случайного совпадения одновременно большого числа сложных характеристик признаковых связей.

Для одномерных случайных величин проверка гипотезы о равенстве их дисперсий в разных выборочных совокупностях осуществляется при помощи критерия Бартлетта

$$\chi_p^2 = \frac{2{,}303}{c}(n-k)\lg S_*^2 - \sum_{j=1}^{k}(n_j - 1)\lg S_j^2), \qquad (1.5)$$

$$\text{при} \quad c = 1 + \frac{1}{3(k-1)}\left(\sum_{j=1}^{k}\frac{1}{n_j - 1} - \frac{1}{n_j - k}\right),$$

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, где $S_*^2 = \frac{1}{n-k}\sum_{j=1}^{k}(n_j - 1)S_j^2$.

Для χ^2-статистики критические значения находят по таблицам квантилей χ^2-распределения по заданному уровню значимости (α) и числу степеней свободы $\nu = k-1$. Нулевая гипотеза о равенстве дисперсий отклоняется, если $\chi_p^2 \geq \chi_{кр(\alpha,\nu)}^2$ и принимается, когда $\chi_p^2 \leq \chi_{кр(\alpha,\nu)}^2$.

В многомерном анализе сравниваются не дисперсии, а ковариационные матрицы двух m-мерных выборочных совокупностей. Критерий принимает следующий вид:

$$\omega_p = b(-2\ln v_1), \qquad (1.6)$$

где параметры b и $-2\ln v_1$ определяются по следующим формулам:

$$b = 1 - \left(\sum_{j=1}^{2} \frac{1}{n-1} - \frac{1}{\sum_{j=1}^{2}(n_j - 1)} \right) \times \left(\frac{2m^2 + 3m - 1}{6(m+1)} \right); \quad (1.7)$$

$$-2\ln v_1 = \left(\sum_{j=1}^{2}(n_j - 1) \right) \times |\Sigma_*| - \sum_{j=1}^{2}((n_j - 1)\ln|\Sigma_j|), \quad (1.8)$$

где m — число признаков, представляющих многомерную выборочную совокупность.

Величина многомерного ω-критерия сравнивается с $\chi^2_{\alpha,v}$ — табличными значениями и степенями свободы $v = n_1 + n_2 - 2$.

Пример 4. На основе данных из примера 2 произведем расчет ω-критерия по рассчитанным ранее ковариационным матрицам для каждой выборки

$$S_1 = \frac{1}{9}\begin{pmatrix} 836 & 54{,}14 \\ 54{,}14 & 5{,}77 \end{pmatrix}; \quad S_2 = \frac{1}{7}\begin{pmatrix} 410 & 1{,}20 \\ 1{,}20 & 0{,}39 \end{pmatrix}; \quad S_* = \begin{pmatrix} 89 & 3{,}95 \\ 3{,}95 & 0{,}44 \end{pmatrix}$$

и $n_1 = 9$; $n_2 = 7$; $|S_1| = 23{,}365$; $|S_2| = 3{,}226$; $|S_*| = 23{,}558$.

Предварительно рассчитаем значения величин b и $-2\ln v_1$:

$$b = 1 - \left(\frac{1}{8} + \frac{1}{6} - \frac{1}{14} \right) \left(\frac{2\times 4 + 3\times 2 - 1}{6(2+1)} \right) = 0{,}841;$$

$$-2\ln v_1 = 14\ln 23{,}558 - (8\ln 23{,}365 + 6\ln 3{,}226) = 11{,}996.$$

Расчетное значение ω-критерия равно $\omega_p = 0{,}841 \times 11{,}996 = 10{,}089$.

Критическое значение ω-критерия найдем по таблицам χ^2-распределения при $\alpha = 0{,}05$ и числе степеней свободы $m(m+1)/2$ или

$$v = (2\times 3)/2 = 3; \quad \chi^2_{0{,}05;3} = 7{,}815.$$

Так как $\omega > \chi^2_{0{,}05;3}$, следует отвергнуть нулевую гипотезу о равенстве ковариационных матриц S_1 и S_2. То есть при заданном уровне значимости $\alpha = 0{,}05$ их различие существенно.

Следовательно, вариация и ковариация признаков в рассматриваемых выборках существенно отличаются, при том, что гипотеза о равенстве векторов средних значений подтвердилась (см. предыдущий пример).

1.2. Контрольные задания

Задание 1. В табл. 1.5 приведены результаты статистического наблюдения, проводимого на предприятиях, входящих в состав концерна.

Таблица 1.5

Номер объекта	Вариант 1			Вариант 2		
	X_1	X_2	X_3	X_1	X_2	X_3
1	10,8	1,90	180	18,6	1,95	210
2	12,0	1,02	150	19,0	2,30	235
3	14,0	1,86	165	21,0	1,80	200
4	15,6	2,65	200	19,5	2,10	195
5	14,8	2,00	210	20,0	2,60	220
Среднее значение $(\overline{X}_j)$	13,5	1,70	185	19,8	2,20	210

Вариант 1 — функционирование предприятий в условиях старой системы управления концерном, вариант 2 — в условиях новой системы управления.

В таблице X_1 — производительность труда одного работающего, тыс. ден. ед./чел.; X_2 — фондоотдача активной части ОПФ, ден. ед.; X_3 — средняя месячная заработная плата, ден. ед.

Проверьте существенность различия значений показателей производственно-хозяйственной деятельности предприятий концерна в условиях старой и новой систем управления.

Задание 2. С целью оценки эффективности новой технологии содержания и кормления крупного рогатого скота производится сравнение среднего веса скота по реализации (X_1) и среднегодовых удоев молока на одну корову (X_2) по двум хозяйствам. В первом хозяйстве, где фермы работают по старой технологии, обследовано 300 коров и установлено, что $\overline{X}_1 = 400$ кг, $\overline{X}_2 = 27{,}0$ ц.

Во втором хозяйстве, внедрившем прогрессивную технологию, обследовано 250 животных и получены следующие показатели: $\overline{X}_1 = 600$ кг, $\overline{X}_2 = 42{,}0$ ц.

В результате анализа получена объединенная ковариационная матрица

$$S_* = \begin{pmatrix} 0{,}25 & 2{,}85 \\ 2{,}85 & 9{,}00 \end{pmatrix}.$$

Решите вопрос о существенности влияния новой технологии на продуктивность крупного рогатого скота в целом и ее влияние на увеличение среднего веса животных ($\overline{X}_1$) и средних надоев молока ($\overline{X}_2$) в отдельности, $\alpha = 0{,}01$.

Задание 3. Для оценки влияния химических производств на уровень заболеваемости населения было проведено медицинское обследование населения двух регионов:

I регион — высокий уровень концентрации промышленных предприятий и наличие нефтехимической промышленности;

II регион — предприятия нефтехимической промышленности на данной территории отсутствуют.

Результаты наблюдения приведены в табл 1.6.

Таблица 1.6

Первый регион					Второй регион				
Населенные пункты	Численность населения, тыс. чел.	X_1	X_2	X_3	Населенные пункты	Численность населения, тыс. чел.	X_1	X_2	X_3
1	160	1,00	56,0	1,80	1	50	0,91	29,0	1,6
2	800	2,00	40,0	2,20	2	150	0,90	28,0	1,8
3	200	1,80	34,0	2,80	3	350	0,53	26,0	1,5
4	340	2,10	29,0	2,40	4	200	0,46	20,0	2,1
5	500	0,86	30,0	2,50					
Итого по региону	2000	1,63	36,3	2,34	Итого по региону	750	610,9	25,0	1,73

В таблице X_1 — уровень заболеваемости эндокринной системы (число заболевших (тыс. чел.) на 100 тыс. чел. населения); X_2 — уровень заболеваемости органов дыхания; X_3 — уровень заболеваемости органов пищеварения.

1. Проверьте гипотезу о равенстве векторов средних значений.

2. Для каждого региона рассчитайте ковариационную матрицу и проверьте гипотезу о равенстве ковариационных матриц, полученных по каждому из регионов.

Задание 4. По десяти предприятиям района изучается влияние условий их работы на финансовые показатели. Результаты наблюдения представлены в табл. 1.7.

Таблица 1.7

Номер предприятия	X_1		X_2	
	I вариант	II вариант	I вариант	II вариант
1	0,63	0,83	0,29	0,21
2	0,51	0,62	0,19	0,15
3	0,70	0,73	0,20	0,19
4	0,68	0,70	0,10	0,08
5	0,55	0,60	0,30	0,15
6	0,72	0,73	0,11	0,09
7	0,60	0,71	0,10	0,08
8	0,62	0,69	0,19	0,20
9	0,64	0,69	0,20	0,21
10	0,69	0,72	0,18	0,16

I вариант — предприятия, работающие в условиях льготного налогообложения; II вариант — предприятия, работающие в условиях полного налогообложения.

В таблице X_1 — коэффициент автономии; X_2 — коэффициент заемных средств.

При помощи критериев Хотеллинга и Бартлетта проверьте гипотезы о равенстве векторов средних значений и равенстве ковариационных матриц. Поясните полученные результаты и сделайте выводы.

Примечание. Коэффициент автономии $= \dfrac{\text{Собственный капитал}}{\text{Общая сумма капитала}}$;

Коэффициент заемных средств $= \dfrac{\text{Сумма обязательств по платежам}}{\text{Собственный капитал}}$.

Задание 5. Для анализа влияния специализации магазина на показатели эффективности его работы было проведено выборочное наблюдение десяти предприятий розничной торговли и получены следующие результаты (табл. 1.8).

Таблица 1.8

Номер магазина	Универсальные магазины			Номер магазина	Специализированные магазины		
	X_1	X_2	X_3		X_1	X_2	X_3
1	200	1,85	0,030	1	60	16,25	0,067
2	180	2,75	0,045	2	80	7,69	0,100
3	150	3,00	0,050	3	10	5,55	0,150
4	120	4,54	0,025	4	75	6,00	0,085
$\overline{X}$	162,5	3,035	0,037	5	100	9,38	0,090
				6	110	7,64	0,110
				$\overline{X}$	72,5	8,75	0,1003

В таблице X_1 — число продавцов, чел.; X_2 — товарооборот на одного продавца, тыс. ден. ед.; X_3 — издержки обращения на один рубль товарооборота, ден. ед.

Проверьте при уровне значимости $\alpha = 0{,}01$ существенность различий двух групп магазинов по заданным переменным, а также равенство ковариационных матриц.

2. РОБАСТНОЕ СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

2.1. Методические рекомендации

2.1.1. Методы выявления грубых ошибок в статистической совокупности

Методы робастного оценивания — это статистические методы, которые позволяют получать достаточно надежные оценки статистической совокупности при неизвестном законе ее распределения и при наличии в ней данных, существенно отклоняющихся от основного массива. Единицы статистической совокупности, у которых значения анализируемого признака существенно отклоняются от основного массива, называются *аномальными наблюдениями*, *«грубыми ошибками»* или *выбросами*.

При робастном оценивании решаются задачи двух основных типов:

- при помощи специальных критериев в статистической совокупности выявляются аномальные наблюдения;
- при помощи одного из выбранных методов исчисляются устойчивые (робастные) оценки совокупности данных, в частности при нормальном законе распределения определяют среднее значение и дисперсию.

В некоторых случаях аномальное наблюдение в статистической совокупности можно обнаружить при помощи визуального анализа. Но чаще приходится использовать специальные статистические приемы и методы. Рассмотрим некоторые из них.

Выявление грубых ошибок на основании Т-критерия Граббса

$$T_p = \frac{x - \bar{x}}{s}, \qquad (2.1)$$

где $\bar{x}$ — выборочная средняя; s — выборочное среднеквадратическое отклонение случайной величины.

Расчетные значения T-критерия сравнивают с пороговыми значениями, заданными соответствующим распределением. Проверяемые значения переменной относят к классу выбросов, если $T_p > T_{kp}$ $(T_{kp} = T_{\alpha,h})$. Если $T_p < T_{kp}$, то считается, что эти значения несущественно отличаются от других значений и не являются аномальными для данной совокупности.

Выявление грубых ошибок на основании L- и E-критериев Титьена—Мура:

1. L-критерий используется для выявления грубых ошибок среди наименьших значений переменной

$$L = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x}_k)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad (2.2)$$

где x_i — значение анализируемого признака у i-го наблюдения; n — объем выборки; k — число предполагаемых аномальных наблюдений; $\bar{x}$ — среднее значение признака, рассчитанное по выборке; $\bar{x}_k$ — среднее значение признака, рассчитанное по «усеченной» совокупности данных, то есть по $(n-k)$-наблюдениям, остающимися после удаления из выборочной совокупности k грубых ошибок «сверху», т.е. значений, сильно отличающихся от средней в меньшую сторону

$$\bar{x}_k = \frac{\sum_{i=1}^{n-k} x_i}{n-k}.$$

2. L'-критерий используется для выявления грубых ошибок среди наибольших значений переменной

$$L' = \frac{\sum_{i=k+1}^{n} (x_i - \bar{x}_k)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad (2.3)$$

где $\bar{x}$ — среднее значение признака, рассчитанное по «усеченной» совокупности, то есть по $(n-k)$-наблюдениям, остающимся после удаления из выборочной совокупности 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 грубые ошибки расположены симметрично в верхней и нижней частях ранжированного ряда значений переменной

$$E = \frac{\sum_{i=k+1}^{n-k'} (x_i - \bar{x}_k)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad (2.4)$$

где $\bar{x}_k$ — средняя, рассчитанная по «усеченной совокупности», после удаления наименьших k и наибольших k' подозреваемых значений

$$\bar{x}_k = \frac{\sum_{i=k+1}^{n-k'} x_i}{n-(k+k')}.$$

Для всех названных критериев L, L' и E предельные значения при заданном уровне значимости α, известном объеме выборки n и предполагаемом числе ошибок k представлены в специальных таблицах (см. параграф 2.4). Если рассчитанные значения критериев оказываются меньше табличных $(C_{\alpha,k})$, то проверяемые значения переменной следует отнести к грубым ошибкам.

2.1.2. Методы устойчивого оценивания параметров статистической совокупности

После того как при помощи одного из рассмотренных критериев удалось выявить аномальные наблюдения, предстоит исчислить устойчивые (робастные) оценки среднего значения и дисперсии. При этом, как уже говорилось, используются два основных подхода: аномальные значения (грубые ошибки) либо удаляются из совокупности, либо модифицируются.

Наиболее простым способом устойчивую оценку средней можно получить по усеченной совокупности данных. Для этого из совокупности предварительно удаляются наблюдения, являющиеся грубыми ошибками. Американский статистик Пуанкаре предложил следующую формулу для расчета средней по усеченной совокупности

$$T(\alpha) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_i . \qquad (2.5)$$

В данной формуле k — это число грубых ошибок, $k \le \alpha \times n$ — есть целая часть от произведения $\alpha \times n$, где n — объем выборочной совокупности, а α — некоторая функция величины засорения выборки (ξ), значения которой находят по специальным таблицам (см. параграф 2.4).

Другой подход к исчислению устойчивой средней, предложенный Винзором, предполагает замену аномальных значений переменной модифицированными значениями. Средняя по Винзору определяется с известным заранее уровнем α $(0 < \alpha < 0{,}5)$ по формуле

$$W(\alpha) = \frac{1}{n}\left(\sum_{i=k+1}^{n-k} x_i + k\left(x_{k+1} + x_{n-k}\right)\right). \qquad (2.6)$$

По аналогии с оценками $T(\alpha)$ и $W(\alpha)$ по усеченной совокупности могут быть найдены не только средние величины, но и другие характеристики статистической совокупности, например, меры вариации, мода, медиана.

Широкое распространение при определении устойчивых оценок получил также подход Хубера, при котором используется некоторая исходная величина k, *определяемая с учетом степени* «засорения» статистической совокупности (ξ) (см. параграф 2.4). Расчеты в этом случае повторяются (имеют итеративный характер) и в итоге приводят к наилучшим оценкам.

Оценка средней величины по методу Хубера производится по формуле

$$\theta = \frac{1}{n}\left(\sum_{|x_i-\theta|<k} x_i + \sum_{|x_i-\theta|>k} x_i + (n_1 - n_2)k \right), \qquad (2.7)$$

где θ — устойчивая оценка среднего значения; k — допустимая величина отклонения от центра совокупности (определяется 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 формуле (2.7) в качестве начальной оценки (θ) может быть использована средняя арифметическая или медиана, рассчитанные по выборке. На каждой итерации производится разделение выборочной совокупности на три класса. В первый класс попадают значения признака, которые остаются без изменения $\left(|x_i - \theta|\right) < k$. Во второй и третий классы (для $x_i > \theta + k$ и $x < \theta - k$) — «грубые ошибки». Причем они не исключаются из рассмотрения, а модифицируются — заменяются соответственно на величины $(x - k)$ и $(x_i + k)$. По исходным и модифицированным значениям при каждой итерации определяется новая оценка средней (θ). Процесс продолжается до тех пор, пока все наблюдения не окажутся в интервале «истинных» значений $|x_i - \theta| < k$.

В многомерном случае «засорением» совокупности данных уже будут не отдельные значения, а вектор значений, характеризующий аномальное наблюдение.

Чтобы проверить, является ли многомерное наблюдение аномальным, обычно используют расстояние Махаланобиса

$$d^m = (X - \bar{X})^T \Sigma^{-1} (X - \bar{X}), \qquad (2.8)$$

где X — вектор значений признака у «подозреваемого» объекта; $\bar{X}$ — вектор средних значений для многомерной совокупности данных; Σ — матрица ковариаций.

В этом случае критерий (F) для проверки гипотезы о существенности отклонения случайного вектора X строится следующим образом:

$$F_p = \frac{(n-m)n}{2(n-1)m}(X - \bar{X})^T \Sigma^{-1} (X - \bar{X}). \qquad (2.9)$$

Если при заданном уровне значимости α и числе степеней свободы $\nu_1 = m$ и $\nu_2 = n - m - 1$ окажется, что $F_p > F_{\alpha,\nu_1,\nu_2}$, то проверяемое наблюдение действительно признается аномальным. В противном случае, когда $F_p \le F_{\alpha,\nu_1,\nu_2}$, *отклонение проверяемого вектора от вектора средних значений считается приемлемым*, а гипотеза о «засорении» совокупности отвергается.

К выявленным грубым ошибкам в многомерной совокупности можно применять уже известные для одномерного случая приемы обработки данных.

2.2. Примеры решения типовых задач

Пример 1. На основании приведенных ниже данных по предприятиям пищевой промышленности проверьте гипотезу о наличии грубых ошибок (ано-

мальных наблюдений) в начале и в конце ранжированного ряда по каждой переменной (табл. 2.1).

Таблица 2.1

Номер наблюдаемого объекта	Удельный вес сертифицированной продукции, %	Рентабельность производства, %
1	90,0	11,6
2	85,0	15,2
3	76,0	18,3
4	50,0	9,1
5	65,0	7,5
6	10,0	3,8
7	80,0	20,0
8	75,0	16,5
9	62,5	4,8
10	95,0	25,1
11	70,5	14,7
12	25,0	17,0

Рассчитайте устойчивые оценки средних значений переменных по Винзору и Пуанкаре.

Решение. 1. Прежде всего проверим статистическую гипотезу о наличии грубых ошибок для переменной X_1. Для этого следует ранжировать ее значения в порядке возрастания (убывания): 10,0 25,0 50,0 62,5 65,0 70,5 75,0 76,0 80,0 85,0 90,0 95,0.

Можно предположить, что значения 10,0 и 25,0 являются аномальными для данной совокупности. Проверим эти значения при помощи L'-критерия Титьена—Мура:

$$L' = \frac{\sum_{i=k+1}^{n}\left(x_i - \bar{x}_k\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}; \qquad \bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = 65{,}3\,\%;$$

$$\bar{x}_k = \frac{1}{10}\sum_{i=k+1}^{n} x_i = 74{,}9\,\%; \qquad \sum_{i=3}^{12}\left(x_i - \bar{x}_k\right)^2 = 1652{,}4;$$

$$\sum_{i=1}^{12}\left(x_i - \bar{x}\right)^2 = 7256{,}18; \qquad L' = \frac{1652{,}40}{7256{,}18} = 0{,}228.$$

Табличное значение критерия $L_{кр}$ равно 0,305. Так как $L' < L_{кр}$, гипотеза подтверждается, следовательно, проверяемые значения 10,0 и 25,0 являются грубыми ошибками.

2. Для расчета устойчивых средних применим формулы Винзора и Пуанкаре:

а) средняя по Винзору

$$W(\alpha) = \frac{1}{n}\left(\sum_{i=k+1}^{n-k} x_i + k\left(x_{k+1} + x_{n-k}\right)\right),$$

$$W(\alpha) = \frac{1}{12}\left((62{,}5 + 65 + 70{,}5 + 75{,}0 + 76{,}0 + 80{,}0 + 85 + 90{,}0 + 95) + 50 \cdot 3\right) = 70{,}75\,\%;$$

б) средняя по Пуанкаре

$$T(\alpha) = \frac{1}{n - 2k}\sum_{i=k+1}^{n-k} x_i,$$

$$n = 12,\ k \le \alpha \cdot n.$$

Параметр α определяем по специальной таблице (приложение) в зависимости от степени «засоренности» выборки ξ

$$\xi = \frac{2}{12} = 0{,}167; \quad \alpha \approx 0{,}164; \quad k \le 0{,}164 \cdot 12 = 2{,}004; \quad k = 2$$

$$T(\alpha) = \frac{1}{12 - 2}\left(50 + 62{,}5 + 65 + \ldots + 95\right) = 74{,}9\,\%.$$

Аналогичные расчеты выполните самостоятельно для переменной X_2 (рентабельность производства).

Пример 2. Имеются следующие данные о ежедневных объемах реализации продукции по 20 предприятиям розничной торговли за месяц (табл. 2.2).

Таблица 2.2

Номер предприятия	Реализовано, тыс. ден. ед.	Номер предприятия	Реализовано, тыс. ден. ед.
1	450	11	790
2	520	12	600
3	537	13	450
4	480	14	550
5	560	15	640
6	600	16	800
7	310	17	635
8	250	18	450
9	900	19	700
10	850	20	1200

1. При помощи критерия Граббса проверьте, являются ли значения 250 и 1200 грубыми ошибками.

2. Используя метод Хубера, проведите несколько итераций по расчету устойчивой оценки средней.

Решение. Прежде всего для рассматриваемой совокупности исчислим среднее значение признака и среднее квадратическое отклонение.

Расчетное значение T-критерия Граббса для проверяемого значения признака 250 и уровня значимости $\alpha = 0{,}05$ равно

$$\bar{x} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{12\,272}{20} = 613{,}6 \text{ тыс. ден. ед.}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{20}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{900\,694{,}8}{20}} = 212{,}2 \text{ тыс. ден. ед.}$$

$$T_p = \frac{x - \bar{x}}{s} = \frac{250 - 613{,}6}{212{,}2} = 247{,}108,$$

табличное значение $T_{кр} = 2{,}779$. Поскольку $T_p > T_{кр}$, проверяемое значение признака 250 является грубой ошибкой, т.е. аномальным наблюдением.

Аналогичные расчеты выполним и для максимального значения 1200

$$T_p = \frac{x - \bar{x}}{s} = \frac{1200 - 613{,}6}{212{,}2} = 2{,}763.$$

В этом случае $T_p < T_{кр}$, следовательно, проверяемое значение 1200 не является грубой ошибкой для данной совокупности.

1. Используя метод Хубера, рассчитаем устойчивую оценку средней. Для первой итерации в качестве оценки величины θ возьмем среднюю $\theta = \bar{x} = 613{,}6$.

В нашем примере число подозреваемых грубых ошибок равно единице, значит $\xi = 1/20 = 0{,}05$. По таблице (см. параграф 2.4) находим $k(\xi) = 1{,}399$.

Разобьем всю совокупность на три класса (табл. 2.3).

Таблица 2.3

	I класс	II класс	III класс
Значения переменных	$\lvert x_i - \theta \rvert < \kappa$ $613{,}6 - 1{,}399 < x_i < 613{,}6 + 1{,}399$ $612{,}2 < x_i < 615$	$x_i > \theta + \kappa$ $612{,}2 < x_i$	$x_i < \theta - \kappa$ $x_i < 615$
Исходные		900 850 790 640 800 635 700 1200	450 520 537 480 560 600 310 250 600 450 550 450
Модифицированные		898,6 848,6 788,6 638,6 798,6 633,6 698,6 1198,6	451,4 521,4 538,4 481,4 561,4 601,4 311,4 251,4 601,4 451,4 551,4 451,4

Модифицированные значения переменных рассчитываются следующим образом: во втором классе от каждого значения отнимаем $k = 1{,}399$, а в третьем классе к каждому значению переменной прибавляем k.

Новое значение оценки величины θ определяется по модифицированным данным следующим образом:

$$\theta_2 = \frac{1}{n}\left(\sum_{|x_i - \theta| < k} x_i + (n_2 - n_1)k \right) = \frac{1}{20}(12277{,}6 + (8 - 12) \times 1{,}399) = 614{,}2.$$

Следующую итерацию начнем уже с данных, модифицированных на первом шаге (табл. 2.4). Тогда границы классов будут иметь вид:

I класс: $|x_i - \theta| < k$ $\quad \theta - k < x_i < \theta + k$ $\quad 614{,}2 - 1{,}399 < x_i < 614{,}2 + 1{,}399$,

II класс: $\theta + k < x_i$ $\quad 614{,}2 + 1{,}399 < x_i$ $\quad 615{,}6 < x_i$,

III класс: $x_i < \theta - k$ $\quad x_i < 614{,}2 - 1{,}399$ $\quad x_i < 612{,}8$.

Таблица 2.4

Значения переменных	I класс	II класс	III класс
	$614{,}2 - 1{,}399 < x_i < 614{,}2 + + 1{,}399$	$615{,}6 < x_i$	$x_i < 612{,}8$
Исходные		898,6 848,6 788,6 638,6 798,6 633,6 698,6 1198,6	451,4 521,4 538,4 481,4 561,4 601,4 311,4 251,4 601,4 451,4 551,4 451,4
Модифицированные		897,2 847,2 787,2 637,2 797,2 632,2 697,2 1197,2	452,8 522,8 539,8 482,8 562,8 602,8 312,8 252,8 602,8 452,8 552,8 452,8

Как видно из расчетов, после второй итерации ни одно значение не перешло в I класс. Пересчитаем новую оценку средней θ

$$\theta_2 = \frac{1}{20}\sum_{|x_i - \theta| < k} x_i + (n_2 - n_1)k = \frac{1}{20}(12283{,}2 + (8 - 12) \times 1{,}399) = 313{,}9.$$

Границы классов будут иметь следующий вид:

I класс: $|x_i - \theta| < k$, $\quad \theta - k < x_i < \theta + k$, $\quad 613{,}9 - 1{,}399 < x_i < 613{,}9 + 1{,}399$;

II класс: $\theta + k < x_i$, $\quad 613{,}9 + 1{,}399 < x_i$, $\quad 615{,}3 < x_i$;

III класс: $x_i < \theta - k$, $\quad x_i < 613{,}9 - 1{,}399$, $\quad x_i < 612{,}5$.

Поскольку значение оценки средней после очередной модификации изменились незначительно, можно предположить, что для достижения конечной цели (все значения попадают в I класс) потребуется очень большое число итераций.

2.3. Контрольные задания

Задание 1. В ходе текущего контроля содержания загрязняющих веществ в воздухе были получены следующие данные (табл. 2.5).

Таблица 2.5

Номер наблюдения	Содержание окиси углерода в воздухе, мг/м^3	Номер наблюдения	Содержание окиси углерода в воздухе, мг/м^3
1	2,6	11	2,3
2	2,1	12	2,9
3	3,0	13	1,0
4	1,5	14	1,8
5	1,2	15	1,3
6	1,8	16	2,4
7	2,0	17	2,0
8	1,7	18	2,4
9	2,5	19	2,0
10	3,5	20	3,4

Используя критерии Граббса и Титьена—Мура, проверьте наличие грубых ошибок (аномальных наблюдений) и рассчитайте устойчивые оценки средней по Пуанкаре и Винзору.

Задание 2. В результате ежедневного наблюдения за качеством производимой продукции в разные смены были получены следующие данные (табл. 2.6).

Таблица 2.6

День	Объем бракованной продукции, %		День	Объем бракованной продукции, %	
	I смена	II смена		I смена	II смена
1	1,5	1,9	11	4,0	6,0
2	1,7	2,4	12	3,5	3,8
3	2,0	1,9	13	2,1	4,0
4	1,2	1,7	14	1,9	2,4
5	1,8	2,0	15	1,7	1,9
6	2,5	2,1	16	2,4	7,0
7	3,7	3,0	17	1,3	6,1
8	2,3	5,0	18	4,2	3,8
9	1,8	1,5	19	4,0	8,0
10	5,0	2,8	20	2,6	2,5

Проведите комплексное исследование полученных данных.

1. Оцените существенность различия средних значений анализируемого признака, рассчитанных для I и II смен.

2. Используя критерии Титьена—Мура, проверьте наличие грубых ошибок отдельно для каждой группы наблюдений (для I и II смен).

3. Рассчитайте устойчивые средние на основе подходов Пуанкаре и Винзора. Сравните полученные значения и сделайте выводы.

Задание 3. При обследовании десяти промышленных предприятий концерна по двум признакам — уровню фондоотдачи основных производственных фондов, р. (X_1) и среднему размеру получаемой прибыли на одного работающего, тыс. ден. ед. (X_2) — были получены следующие данные (табл. 2.7).

Таблица 2.7

Номер предприятия	X_1	X_2	Номер предприятия	X_1	X_2
1	0,75	350	6	1,50	850
2	1,20	420	7	4,10	1000
3	0,90	50	8	0,55	100
4	1,75	235	9	2,65	670
5	2,80	1100	10	3,85	520

1. При помощи критериев Титьена—Мура проверьте наличие грубых ошибок по каждой переменной в начале и в конце ранжированного ряда данных.

2. Для каждого признака рассчитайте устойчивые средние оценки по Пуанкаре и Винзору.

2.4. Таблицы значений критериев, используемых при проверке многомерных статистических и в робастном оценивании

Таблица 2.8

Процентные точки критерия Смирнова—Граббса (7)

№ п/п	Доверительная вероятность $(1-\alpha)$			№ п/п	Доверительная вероятность $(1-\alpha)$		
	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

Окончание табл. 2.8

№ п/п	Доверительная вероятность $(1-\alpha)$			№ п/п	Доверительная вероятность $(1-\alpha)$		
	0,9	0,95	0,99		0,9	0,95	0,99
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 2.9

Значения C_α-оценки для L и L'-критериев Титьена — 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 2.10

Значения C_α-оценки для E-критерия Титьена—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 2.11

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 2.12

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		

3. МНОЖЕСТВЕННЫЙ КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

3.1. Методические рекомендации

Регрессионный анализ — это статистический метод исследования зависимости случайной величины Y-отклик от переменной (X) или переменных X_j $(j = 1, 2, ..., k)$ — предикторы; X_j рассматриваются в регрессионном анализе как неслучайные величины, независимо от истинного закона их распределения. В ходе регрессионного анализа при помощи выбранного метода [7, 16] строится математическая модель, описывающая форму связи переменных — *уравнение регрессии*. Как правило, регрессионному анализу предшествует *анализ корреляционной зависимости* переменных, который позволяет установить наличие связи между анализируемыми переменными, оценить ее тесноту и определить направление (прямая или обратная связь). Кроме того, в ходе корреляционного анализа происходит отбор существенных факторов, включаемых в уравнение регрессии. Наиболее простой формой корреляционного анализа является парная корреляция — анализируется связь между парой признаков — откликом Y и одним предиктором X. В этом случае уравнение регрессии принимает вид $y = f(x)$.

В ходе множественного корреляционного анализа рассчитываются следующие характеристики:

- *парные коэффициенты корреляции* r_{ij} — оценки тесноты линейной корреляционной связи между всеми парами анализируемых признаков с учетом их взаимного влияния и взаимодействия. Совокупность парных коэффициентов корреляции, относящихся ко всем исследуемым признакам, может быть представлена в виде корреляционной матрицы R, которая рассчитывается по формуле

$$R = \begin{pmatrix} 1 & r_{12} \ldots & r_{1m} \\ r_{21} & 1 \ldots & r_{2m} \\ \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} \ldots & 1 \end{pmatrix} = \frac{1}{n} Z^T Z, \qquad (3.1)$$

где Z — матрица стандартизованных значений исходных переменных. Ее элементы рассчитываются по формуле

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}.$$

На главной диагонали матрицы R стоят единицы, т.е. дисперсии стандартизованных переменных, а все другие элементы — парные коэффициенты корреляции r_{ij};

- *частные коэффициенты корреляции* $r_{ij/k,m}$, характеризующие тесноту линейной корреляционной связи между парой анализируемых признаков (x_i и

x_j) в условиях элиминирования влияния на эту пару других переменных (x_k, x_m и т.д.). Эти коэффициенты характеризуют так называемую чистую корреляцию. В матричном виде формулу для расчета частных коэффициентов корреляции можно записать следующим образом:

$$r_{ij/km} = \frac{-A_{ij}}{\sqrt{(A_{ii}A_{jj})}}, \qquad (3.2)$$

где A_{ii}, A_{jj}, A_{ij} — алгебраические дополнения соответствующих элементов матрицы парных корреляций R.

Знак частному коэффициенту корреляции присваивается такой же, как и у парного коэффициента корреляции;

- *множественный коэффициент корреляции* $R_{0/1,2,\ldots m}$ характеризует степень тесноты связи между результативным признаком (откликом) Y и всеми факторными признаками (предикторами — X_j);

- *множественный коэффициент детерминации* $R^2_{0/1,2,\ldots m}$ характеризует долю дисперсии результативной переменной, обусловленную влиянием факторных переменных, участвующих в анализе. На основе корреляционной матрицы R множественный коэффициент корреляции и множественный коэффициент детерминации могут быть исчислены следующим образом:

$$R_{0,1,2,\ldots m} = \sqrt{1 - \frac{|R|}{|R_{**}|}}; \qquad R^2_{0,1,2,\ldots m} = 1 - \frac{|R|}{|R_{**}|}, \qquad (3.3)$$

где $|R|$ — определитель матрицы парных корреляций, $|R_{**}|$ — определитель матрицы парных корреляций, полученной после вычеркивания строки и столбца, представляющих связи зависимой переменной (Y).

В множественном регрессионном анализе исследуется связь между несколькими независимыми переменными (предикторами) $X_1, X_2, \ldots X_m$ и результативным признаком (откликом) Y. Следовательно,

$$Y = f(X_1, X_2, \ldots X_m) + \varepsilon.$$

Обычно предполагается, что случайная величина (Y) имеет нормальный закон распределения с условным математическим ожиданием $\tilde{Y} = \varphi(X_1, X_2, \ldots, X_k)$ и постоянной, не зависящей от аргументов дисперсией σ_y^2.

В анализе чаще всего используются уравнения регрессии линейного вида

$$\tilde{Y} = a_0 + a_1X_1 + \ldots + a_jX_j + \ldots + a_mX_m.$$

Коэффициенты регрессии a_j показывают, на какую величину в среднем изменяется результативный признак Y, если независимая переменная X_j изменяется на единицу ее измерения.

В матричной форме регрессионная модель имеет вид

$$Y = XA + \varepsilon,$$

где Y — случайный вектор-столбец размерности $(n \times 1)$ наблюдаемых значений результативного признака $(y_1, y_2, \ldots, y_n)$; X — матрица размерности $(n \times (m+1))$ наблюдаемых значений аргументов. Элемент матрицы $\|x_{ij}\|$ рассматривается как неслучайная величина ($i = 1, 2, \ldots n$; $j = 0, 1, 2, \ldots m$; $x_{0j} = 1$); A — вектор-столбец размерности $(n \times (m+1))$ неизвестных параметров, подлежащих оценке в ходе регрессионного анализа (вектор коэффициентов регрессии); ε — случайный вектор-столбец размерности $(n \times 1)$ — вектор остатков, которые являются независимыми нормально распределенными случайными величинами с нулевым математическим ожиданием ($E(\varepsilon_i) = 0$) и неизвестной дисперсией $\sigma^2(D(\varepsilon_i) = \sigma^2)$. На практике рекомендуется, чтобы число наблюдений (n) превышало число анализируемых признаков (m) не менее, чем в пять-шесть раз.

Для расчета вектора оценок коэффициентов регрессии $A = (a_0, a_1, \ldots a_m)$ по методу наименьших квадратов используется формула

$$A = (X^T X)^{-1} X^T Y, \qquad (3.4)$$

где

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1m} \\ 1 & x_{21} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{nm} \end{pmatrix}; \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix};$$

X^T — транспонированная матрица X; $(X^T X)^{-1}$ — матрица, обратная матрице $X^T X$.

Для устранения влияния различия дисперсий и единиц измерения отдельных переменных на результаты регрессионного анализа в ряде случаев целесообразно вместо исходных значений переменных x_{ij} использовать стандартизованные значения $z_{ij} = (x_{ij} - x_j)/\sigma_j$. В этом случае уравнение множественной линейной регрессии будет иметь следующий вид:

$$\hat{Z}_0 = \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_m Z_m, \qquad (3.5)$$

где $\hat{Z}_0$ — стандартизованные значения отклика Y; Z_j — стандартизованные значения предикторов (независимых переменных — X_j); β_j — стандартизованные коэффициенты регрессии, которые могут быть вычислены исходя из следующей системы уравнений:

$$\begin{cases} r_{01} = \beta_1 + \beta_2 r_{21} + \ldots + \beta_m r_{m1}, \\ r_{02} = \beta_1 r_{12} + \beta_2 r_{22} + \ldots + \beta_m r_{m2}, \\ \ldots \quad \ldots \quad \ldots \quad \ldots \\ r_{0m} = \beta_1 r_{1m} + \beta_2 r_{2m} + \ldots + \beta_m. \end{cases}$$

Если решать данную систему по правилу Крамера, то β_j равно

$$\beta_j = |R_j|/|R|, \qquad (3.6)$$

где $|R|$ — определитель матрицы системы уравнений; $|R_j|$ — определитель матрицы системы линейных уравнений, в которой j-й столбец заменен столбцом свободных членов уравнений системы $(r_{01}, r_{02}, \ldots r_{0m})$.

Когда уравнение построено в стандартизованном масштабе, коэффициенты регрессии $\beta_1, \beta_2, \ldots \beta_m$ показывают, на сколько стандартных отклонений изменится Y при изменении каждой из X_j на одно стандартное отклонение. Между коэффициентами a_j и β_j существует следующая зависимость:

$$a_j = \beta_j \frac{\sigma_y}{\sigma_j}. \qquad (3.7)$$

Кроме того, при помощи коэффициентов β_j можно рассчитать частные (r^2_{ij}) и множественный $(R^2_{0/1,2,\ldots m})$ коэффициенты детерминации

$$R^2_{0/1,2,\ldots m} = \beta_1 r_{01} + \beta_2 r_{02} + \ldots + \beta_m r_{0m};$$

$$r^2_{01} = \beta_1 r_{01} \quad r^2_{02} = \beta_2 r_{02} \quad \ldots \quad r^2_{0m} = \beta_m r_{0m}, \text{ причем}$$

$$R^2_{0/1,2,\ldots m} = r^2_{01} + r^2_{02} + \ldots + r^2_{0m}.$$

После того как рассчитано само уравнение регрессии и перечисленные выше характеристики корреляционных связей, необходимо убедиться в адекватности полученных результатов.

Значимость уравнения регрессии в целом, т.е. нулевая гипотеза $H_0 : A = 0 \, (a_0 = a_1 = \ldots = a_m = 0)$, проверяется по F-критерию Фишера. Его наблюдаемое значение определяется по формуле

$$F_p = \frac{Q_R / m}{Q_{ост} / (n - m - 1)}, \qquad (3.8)$$

где $Q_R = (XA)^T (XA)$, $Q_{ост} = (Y - XA)^T (Y - XA)$.

По таблице распределения значений F-критерия Фишера, при заданных α, $\upsilon_1 = m$, $\upsilon_2 = n - m - 1$, находят $F_{кр}$. Гипотеза H_0 отклоняется с вероятностью α, если $F_p > F_{кр}$. Из этого следует, что уравнение является значимым, т.е. хотя бы один из коэффициентов регрессии существенно отличен от нуля.

Для проверки значимости отдельных коэффициентов регрессии, т.е. гипотез $H_0 : a_j = 0$, где $j = 1,2,\ldots,m$, используют t-критерий Стьюдента, фактическое значение которого вычисляют следующим образом:

$$t_p(\alpha_j) = a_j / \hat{S}_{a_j}; \quad \hat{S}^2_{a_j} = S^2_{ост} \times c_{ij}; \quad S^2_{ост} = (Y - XA)^T (Y - XA)/(n - m - 1), \qquad (3.9)$$

где $\hat{S}^2_{aj}$ — средняя ошибка коэффициента регрессии a_j; $\hat{S}^2_{ост}$ — оценка среднего квадрата ошибки; c_{jj} — соответствующие коэффициенту a_j диагональные элементы матрицы $(X^T X)^{-1}$.

По таблице значений t-критерия Стьюдента для заданного уровня значимости и числа степеней свободы $(n - m - 1)$ находят $t_{кр}$. Значимость проверяемого коэффициента a_j подтверждается, если $|t_p| > t_{кр}$. В противном случае коэффициент регрессии незначим, и соответствующая ему переменная не должна входить в модель.

Аналогичным образом осуществляется проверка значимости парных и частных коэффициентов корреляции. При этом табличное значение определяется для числа степеней свободы, равного $(n - m - 1)$, а расчетное значение критерия вычисляется по формуле

$$t_{кр} = \frac{r_{ij.}}{\left(1 - r^2_{ij}\right)^{1/2}} \sqrt{n - m - 1}. \qquad (3.10)$$

Значимость множественного коэффициента детерминации $(R^2_{0/1,2\ldots m})$ и соответственно множественного коэффициента корреляции $(R_{0/1,2\ldots m})$ оценивается по F-критерию Фишера. Расчетное значение этого критерия определяется по формуле

$$F_p = \frac{(n - m - 1) R^2_{0/1,2\ldots m}}{m(1 - R^2_{0/1,2\ldots m})}. \qquad (3.11)$$

Гипотеза о значимости множественного коэффициента детерминации принимается в том случае, если $F_p > F_{кр}$ для заданного уровня значимости α и числа степеней свободы $\upsilon_1 = m$ и $\upsilon_2 = n - m - 1$.

3.2. Примеры решения типовых задач

Пример 1. По результатам корреляционно-регрессионного анализа вычислены матрица парных коэффициентов корреляции и значения β-коэффициентов:

$$R = \begin{pmatrix} 1 & 0{,}75 & 0{,}55 \\ & 1 & 0{,}28 \\ & & 1 \end{pmatrix}; \quad \begin{matrix} \beta_2 = 0{,}6470 \\ \beta_3 = 0{,}3689 \end{matrix}.$$

Известно, что откликом (Y) является первая переменная (X_1), а число наблюдений (n) равно 20.

1. Рассчитаем частные и множественный коэффициенты детерминации и корреляции. Проанализируем полученные результаты и сделаем вывод.

2. Произведем оценку значимости коэффициентов. В анализе корреляции будем использовать t-критерий Стьюдента.

Решение. 1. Для оценки тесноты корреляционной связи рассчитаем следующие величины:

- определители матрицы парных корреляций (R) и матрицы R_*:

$$|R| = \begin{vmatrix} 1 & 0{,}75 & 0{,}55 \\ 0{,}75 & 1 & 0{,}28 \\ 0{,}55 & 0{,}28 & 1 \end{vmatrix} = 0{,}2876; \qquad |R_*| = \begin{vmatrix} 1 & 0{,}28 \\ 0{,}28 & 1 \end{vmatrix} = 0{,}9216;$$

- множественный коэффициент детерминации (R^2)

$$R^2 = 1 - \frac{|R|}{|R_*|} = 1 - \frac{0{,}2876}{0{,}9216} = 0{,}6879 \approx 0{,}688;$$

- множественный коэффициент корреляции (R)

$$R = \sqrt{R^2} = \sqrt{0{,}6879} = 0{,}829.$$

Следовательно, связь между факторными признаками и откликом тесная, поскольку значение множественного коэффициента корреляции близко к единице; вариация результативного показателя на 68,8 % определяется признаками-факторами;

- частные коэффициенты детерминации равны соответственно:

$$r_{12}^2 = \beta_2 \cdot r_{12} = 0{,}647 \cdot 0{,}75 = 0{,}485, \text{ или } 48{,}5\ \%;$$
$$r_{13}^2 = \beta_3 \cdot r_{13} = 0{,}3689 \cdot 0{,}55 = 0{,}203, \text{ или } 20{,}3\ \%.$$

На долю первого фактора приходится 48,5 % объясненной дисперсии результативного показателя, а на долю второго фактора — 20,3 %.

Выполним проверку правильности расчетов

$$R^2 = r_{12}^2 + r_{13}^2 = 0{,}485 + 0{,}203 = 0{,}688\ (68{,}8\ \%).$$

Таким образом, два фактора объясняют примерно 68,8 % вариации результативного показателя.

2. Оценим значимость парных коэффициентов корреляции по t-критерию Стьюдента.

Расчетное значение критерия определяем по формуле

$$t_r = \frac{r}{m_r},$$

где m_r — ошибка коэффициента корреляции. Ее находят по формуле

$$m_r = \frac{1 - r_{yx_j}^2}{\sqrt{n-2}}.$$

Определяем m_r для r_{yx_1} и для r_{yx_2}:

$$m_r = \frac{1 - 0{,}75}{\sqrt{20-2}} = 0{,}0589, \quad t_r = \frac{0{,}75}{0{,}0589} = 17{,}73,$$
$$m_r = \frac{1 - 0{,}55}{\sqrt{20-2}} = 0{,}1061, \quad t_r = \frac{0{,}55}{0{,}1061} = 5{,}184.$$

Критическое значение t-критерия для уровня значимости $\alpha = 0{,}05$ равно 1,725. Поскольку в обоих случаях $t_r > t_{кр}$, проверяемая гипотеза отвергается, т.е. парные коэффициенты корреляции являются значимыми.

Пример 2. По пяти промышленным предприятиям имеются следующие данные о фондовооруженности труда рабочих (X_1), уровне производительности труда (X_2), удельном весе потерь от брака (X_3) (табл. 3.1).

Таблица 3.1

Номер предприятия	Фондовооруженность труда рабочего, тыс. ден. ед.	Месячная производительность труда рабочего, тыс. ден. ед.	Удельный вес потерь от брака, %
	X_1	X_2	X_3
1	3,9	7,0	2,4
2	1,1	11,1	5,9
3	1,8	10,2	6,2
4	6,0	12,0	6,0
5	5,4	10,0	11,0
	$\bar{x}_1 = 3{,}64$	$\bar{x}_2 = 10{,}06$	$\bar{x}_3 = 6{,}3$
	$\sigma_1 = 1{,}93$	$\sigma_2 = 1{,}69$	$\sigma_3 = 2{,}74$

Определите:

1) матрицы парных и частных коэффициентов корреляции;

2) множественный коэффициент детерминации и множественный коэффициент корреляции при условии, что X_2 — зависимая переменная;

3) матрицу ковариаций.

Решение. 1. Парные коэффициенты корреляции рассчитываются следующим образом:

$$r_{12} = \frac{\frac{1}{5}\sum_{i=1}^{5}[(3{,}9 - 3{,}64)(7 - 10{,}06) + \ldots + (5{,}4 - 3{,}64)(10 - 10{,}06)]}{1{,}93 \times 1{,}69} = 0{,}048;$$

$$r_{13} = \frac{\frac{1}{5}[(3{,}9 - 3{,}64)(2{,}4 - 6{,}3) + \ldots + (5{,}4 - 3{,}64)(11 - 6{,}3)]}{1{,}93 \times 2{,}74} = 0{,}293;$$

$$r_{23} = \frac{\frac{1}{5}[(7 - 10{,}06)(2{,}4 - 6{,}3) + \ldots + (10 - 10{,}06)(11\text{-}6{,}3)]}{1{,}69 \times 2{,}74} = 0{,}460.$$

Матрица парных коэффициентов корреляции имеет вид

$$R = \begin{pmatrix} 1 & 0{,}048 & 0{,}293 \\ 0{,}048 & 1 & 0{,}460 \\ 0{,}293 & 0{,}460 & 1 \end{pmatrix}.$$

2. Частные коэффициенты корреляции рассчитываются по формуле

$$r_{1,2/3} \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}; \quad r_{1,3/2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}; \quad r_{2,3/1} = \frac{r_{23} - r_{12} \cdot r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}};$$

$$r_{2,3/1}=\frac{0,460-0,048\cdot 0,293}{\sqrt{(1-0,048^2)\cdot(1-0,293^2)}}=\frac{0,445936}{\sqrt{0,997696\times 0,914151}}=0,467;$$

$$r_{1,2/3}=\frac{0,048-0,293\cdot 0,460}{\sqrt{(1-0,293^2)\cdot(1-0,460^2)}}=\frac{-0,08678}{\sqrt{0,914151\times 0,7884}}=-0,102;$$

$$r_{1,3/2}=\frac{0,293-0,048\cdot 0,460}{\sqrt{(1-0,048^2)\cdot(1-0,460^2)}}=\frac{0,27092}{\sqrt{0,997696\times 0,7884}}=0,305.$$

Матрица частных коэффициентов корреляции (R') будет иметь вид

$$R'=\begin{pmatrix} & -0,102 & 0,305\\ -0,102 & & 0,467\\ 0,305 & 0,467 & \end{pmatrix}.$$

3. Множественный коэффициент корреляции определяется по формуле

$$R=\sqrt{1-\frac{|R|}{|R^*|}}=\sqrt{1-\frac{0,713}{0,914}}=0,469; \qquad R=0,220.$$

4. Элементы ковариационной матрицы $cov(x_i, x_j)$ определяются по формуле

$$cov(x_i,x_j)=\frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{n}.$$

В многомерном статистическом анализе ковариации принято иногда обозначать как σ_{ij} (по аналогии с дисперсиями).

Рассчитаем последовательно все элементы ковариационной матрицы:

$$\sigma_{11}=cov(x_1,x_1)=\frac{\sum_{k=1}^{5}(x_{1k}-\bar{x}_1)}{5}=\frac{(3,9-3,64)^2+\ldots+(5,4-3,64)^2}{5}=3,72;$$

$$\sigma_{12}=cov(x_1,x_2)=\frac{\sum_{k=1}^{5}(x_{1k}-\bar{x}_1)(x_{2k}-\bar{x}_2)}{5}=\frac{(3,9-3,64)(7-10,06)}{5}+$$

$$+\frac{(1,1-3,64)(11,1-10,06)+\ldots+(5,4-3,64)(10-10,06)}{5}=0,1556;$$

$$\sigma_{13}=cov(x_1,x_3)=-1,97; \qquad \sigma_{23}=cov(x_2,x_3)=2,128;$$

$$\sigma_{22}=cov(x_2,x_2)=2,85; \qquad \sigma_{33}=cov(x_3,x_3)=7,512.$$

Матрица ковариаций будет иметь следующий вид:

$$S=\begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{pmatrix}=\begin{pmatrix}3,72 & 0,16 & -1,97\\ 0,16 & 2,85 & 2,13\\ -1,97 & 2,13 & 7,51\end{pmatrix}.$$

На основании матрицы ковариаций можно сравнить вариацию признаков в исследуемой статистической совокупности. Для этого рассчитаем коэффициенты вариации по каждой переменной:

$$V_j=\frac{\sigma_j}{\bar{x}_j}\times 100; \qquad V_1=\frac{\sqrt{3,72}}{3,64}\times 100=53\,\%;$$

$$V_2=\frac{\sqrt{2,85}}{10,06}\times 100=16,8\,\%; \quad V_3=\frac{\sqrt{7,51}}{6,3}\times 100=43,5\,\%.$$

Как показывают расчеты, исследуемая совокупность наиболее однородна по второй переменной X_2 — месячная производительность труда, а наименее однородна по переменной X_1 — фондовооруженность труда рабочего.

Используя элементы ковариационной матрицы, можно также проверить правильность расчета парных линейных коэффициентов корреляции

$$r_{ij}=\frac{cov(x_i,x_j)}{\sigma_i\sigma_j}.$$

Например, коэффициент корреляции между переменными x_1 и x_2 будет равен

$$r_{12}=\frac{cov(x_1,x_2)}{\sigma_1\sigma_2}=\frac{0,16}{\sqrt{3,72\times 2,85}}=0,049,$$

а в корреляционной матрице он равен 0,048, т.е. имеется небольшое расхождение за счет округлений.

Пример 3. На основании приведенных данных табл. 3.2 по районам области постройте линейную регрессионную модель валового выпуска продукции сельского хозяйства в целом по области.

Таблица 3.2

Район	X_1	X_2	Y
1	28	12,22	121,0
2	31	8,96	43,0
3	25	11,69	69,0
4	32	5,38	21,0
5	22	8,66	58,0
6	30	9,35	29,0
7	15	8,92	66,0
8	18	7,61	54,0
9	14	11,32	86,0
10	23	9,53	81,0

Окончание табл. 3.2

Район	X_1	X_2	Y
11	30	6,75	52,0
12	27	7,00	35,0
13	36	6,58	27,0
14	20	6,79	74,0
15	18	9,12	83,0
16	21	4,79	57,0

Здесь: X_1 — нагрузка пашни на одного работника, га; X_2 — производительность труда одного работника, тыс. ден. ед.; Y — валовая продукция, млн ден. ед.

1. Рассчитайте уравнение множественной линейной регрессии.

2. Оцените тесноту связи между анализируемыми признаками с помощью коэффициентов корреляции и детерминации (парных и множественных).

3. Оцените значимость коэффициентов регрессии по t-критерию Стьюдента и качество модели по F-критерию Фишера. Поясните экономический смысл полученных результатов.

Решение. 1. Для оценки коэффициентов уравнения регрессии воспользуемся методом наименьших квадратов (МНК).

С этой целью строим систему нормальных уравнений для матрицы исходных значений переменных:

$$\begin{cases} \sum y = a_0 16 + a_1 \sum x_1 + a_2 \sum x_2, \\ \sum y x_1 = a_0 \sum x_1 + a_1 \sum x_1^2 + a_2 \sum x_1 \cdot x_2, \\ \sum y x_2 = a_0 \sum x_2 + a_1 \sum x_1 \cdot x_2 + a_2 \sum x_2^2, \end{cases}$$

$$\sum x_1 = 390; \quad \sum x_2 = 134{,}67; \quad \sum x_1^2 = 10142;$$

$$\sum x_2^2 = 1204{,}152; \quad \sum x_1 \cdot x_2 = 3232{,}5;$$

$$\sum y = 956;$$

$$\sum x_1 \cdot y = 21941; \quad \sum x_2 \cdot y = 8607{,}33.$$

Система уравнений будет иметь вид

$$\begin{cases} 956 = 16a_0 + 390a_1 + 134{,}67a_2, \\ 21941 = 390a_0 + 10142a_1 + 3232{,}5a_2, \\ 8607{,}33 = 134{,}67a_0 + 3232{,}5a_1 + 1204{,}152a_2, \end{cases}$$

$$a_0 = 41{,}7; \quad a_1 = -1{,}6; \quad a_2 = 6{,}8.$$

Уравнение регрессии можно записать следующим образом:

$$y = 41{,}7 - 1{,}6x_1 + 6{,}8x_2.$$

Рассмотрим экономический смысл полученных коэффициентов регрессии для нашего примера. Первый коэффициент $a_1 = -1{,}6$ показывает, что при увеличении нагрузки пашни на одного работника на 1 га объем выпуска продукции сельского хозяйства уменьшится на 1,6 млн ден. ед. Второй коэффициент регрессии $a_2 = 6{,}8$ показывает, что при увеличении производительности труда одного работника на 1 тыс. ден. ед., объем выпуска продукции увеличится на 6,8 млн ден. ед. при прочих равных условиях.

1. Для расчетов коэффициентов корреляции и детерминации (парных и множественных) проведем прежде всего стандартизацию исходных переменных и рассчитаем матрицу корреляций R

$$R = \frac{1}{16} Z^T Z,$$

где Z — матрица стандартизованных значений переменных

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}.$$

Матрица $Z^T Z$ будет иметь следующий вид:

$$Z^T Z = \begin{pmatrix} 16{,}018 & -3{,}768 & -8{,}542 \\ -3{,}768 & 16{,}021 & 10{,}570 \\ -8{,}542 & 10{,}570 & 16{,}056 \end{pmatrix},$$ а матрица корреляций будет равна

$$R = \frac{1}{16} \begin{pmatrix} 16{,}018 & -3{,}768 & -8{,}542 \\ -3{,}768 & 16{,}021 & 10{,}570 \\ -8{,}542 & 10{,}570 & 16{,}056 \end{pmatrix} = \begin{pmatrix} 1 & -0{,}236 & -0{,}534 \\ & 1 & 0{,}661 \\ & & 1 \end{pmatrix}.$$

Теперь рассчитаем множественный коэффициент детерминации (R^2) и множественный коэффициент корреляции $R_{y,x1,x2}$

$$R^2 = 1 - \frac{|R|}{|R_*|} = 1 - \frac{0{,}389}{0{,}944} = 0{,}588,$$

$$R_{y,x1,x2} = \sqrt{0{,}588} = 0{,}767.$$

Полученные результаты позволяют сделать следующий вывод: вариация объема выпуска продукции на 58,8 % зависит от исследуемых признаков-факторов; связь между результативным признаком (откликом) достаточно тесная, поскольку множественный коэффициент корреляции близок к единице $R_{y,x_1,x_2} = 0{,}767$.

Пример 4. На основе приведенных данных табл. 3.3 по десяти промышленным предприятиям проведите регрессионный анализ зависимости себестоимости произведенной продукции Y (млн ден. ед.) от объема произведенной

продукции X_1 (млн ден. ед.) и уровня производительности труда рабочих X_2 (тыс. ден. ед. на чел.).

Таблица 3.3

№ п/п	1	2	3	4	5	6	7	8	9	10	Итого
X_1	3,3	4,2	5,0	5,6	5,8	5,1	6,2	7,0	10,8	15,0	68
X_2	1,7	1,5	1,4	1,3	1,3	1,5	1,6	1,2	1,3	1,2	14
Y	2,5	2,7	3,7	4,0	4,3	4,6	5,0	6,0	7,2	10,0	50

Решение. 1. Определим вектор оценок коэффициентов регрессии A уравнения

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 .$$

Согласно методу наименьших квадратов, вектор A получается из выражения $A = \left(X^T X\right)^{-1} X^T Y$,

где $$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ . & .. & .. \\ 1 & x_{10,1} & x_{10,2} \end{pmatrix}, \; Y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_{10} \end{pmatrix}, \; A = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix};$$

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1} x_{i2} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i1} x_{i2} & \sum_{i=1}^{n} x_{i2}^2 \end{pmatrix}; \quad X^T Y = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \\ \sum_{i=1}^{n} x_{i2} y_i \end{pmatrix}$$

Для того чтобы рассчитать все необходимые элементы матрицы $X^T X$ и вектора $X^T Y$ заполним табл. 3.4.

Таблица 3.4

№ п/п	X_1	X_2	Y	X_1^2	X_2^2	YX_1	YX_2	$X_1 X_2$	$\hat{Y}$
1	3,3	1,7	2,5	10,89	2,89	8,25	4,25	5,61	2,59
2	4,2	1,5	2,7	17,64	2,25	11,34	4,05	6,37	3,25
3	5,0	1,4	3,7	25,00	1,96	18,50	5,18	7,0	3,79
4	5,6	1,3	4,0	31,36	1,69	22,40	5,20	7,28	4,21
5	5,8	1,3	4,3	33,64	1,69	24,94	5,59	7,54	4,32
6	5,1	1,5	4,6	26,01	2,25	23,46	6,90	7,65	3,78
7	6,2	1,6	5,0	38,44	2,56	31,00	8,00	9,92	4,35
8	7,0	1,2	6,0	49,00	1,44	42,00	7,20	8,40	5,10
9	10,8	1,3	7,2	116,64	1,69	77,76	9,36	14,04	7,26
10	15,0	1,2	10,0	225,00	1,44	150,0	12,0	18,00	9,79
Итого	68	14	50	573,62	19,86	409,65	67,73	91,74	48,5

$$X^T X = \begin{pmatrix} 10 & 68 & 14 \\ 68 & 573,62 & 91,74 \\ 14 & 91,74 & 19,86 \end{pmatrix} \quad X^T Y = \begin{pmatrix} 50 \\ 409,65 \\ 67,73 \end{pmatrix}.$$

Рассчитываем элементы обратной матрицы $(X^T X)^{-1}$ и вектор оценок коэффициентов регрессии A. Определитель матрицы $|X^T X| = 169,456$, а обратная матрица $(X^T X)^{-1}$ равна

$$(X^T X)^{-1} = \begin{pmatrix} 17,5576 & -0,3901 & -10,5749 \\ -0,3901 & 0,0153 & 0,2041 \\ -10,5749 & 0,2041 & 6,5619 \end{pmatrix}, \; A = \begin{pmatrix} 1,838 \\ 0,586 \\ -0,698 \end{pmatrix}$$

и оценку уравнения регрессии $\hat{y} = 1,838 + 0,586 x_1 - 0,698 x_2$. Получаем

$$Q_{ост} = \sum_{i=1}^{n} e_i^2 = 2,309324 \quad Q_R = \sum_{i=1}^{n} \hat{y}_i^2 = 275,948.$$

Тогда несмещенная оценка остаточной дисперсии равна

$$\hat{S}^2 = \frac{1}{n-3} \cdot Q_{ост} = \frac{1}{7} \cdot 2,309324 = 0,3299,$$

а оценка среднеквадратического отклонения составит $\hat{S} = \sqrt{\hat{S}^2} = 0,574$.

Проверяем на уровне значимости $\alpha = 0,05$ адекватность уравнения регрессии, т.е. гипотезу $H_0 : a = 0 \, (a_0 = a_1 = a_2 = 0)$. Для этого вычисляем наблюдаемое значение F-критерия

$$F_p = \frac{Q_R / (m)}{Q_{ост} / (n - m - 1)} = \frac{275,948/2}{2,309324/7} = 418,2.$$

По таблице F-распределения для заданного уровня значимости $\alpha = 0,05$ и числа степеней свободы $\nu_1 = 3$ и $\nu_2 = 7$ находим $F_{табл} = 4,35$.

Так как $F_p > F_{кр} (418,2 > 4,35)$, гипотеза отвергается с вероятностью ошибки 0,05. Таким образом, уравнение является значимым, т.е. хотя бы один из рассчитанных коэффициентов регрессии отличен от нуля.

Перед проверкой значимости отдельных коэффициентов регрессии найдем оценку ковариационной матрицы вектора коэффициентов регрессии (A):

$$\hat{S}(A) = \hat{S}^2 (X^T X)^{-1} = \begin{pmatrix} 5,79225 & -0,12869 & -3,48866 \\ -0,12869 & 0,00505 & 0,06733 \\ -3,48866 & 0,06733 & 2,16467 \end{pmatrix}.$$

Учитывая, что на главной диагонали ковариационной матрицы находятся дисперсии коэффициентов регрессии, получаем следующие несмещенные оценки этих дисперсий:

$$\hat{S}^2_{a_0} = 5{,}79225 \quad (\hat{S}_{a_0} = 2{,}40671);$$

$$\hat{S}^2_{a_1} = 0{,}00505 \quad (\hat{S}_{a_1} = 0{,}07106);$$

$$\hat{S}^2_{a_2} = 2{,}16467 \quad (\hat{S}_{a_2} = 1{,}47128);$$

и оценку корреляционной матрицы R_a с элементами, определяемыми по формуле

$$r_{j-1,k-1} = \frac{cov(a_{j-1}, a_{k-1})}{\hat{S}_{a_{j-1}} \hat{S}_{a_{k-1}}},$$

где $cov(a_{j-1}, a_{k-1})$ — элементы матрицы $\hat{S}(a)$, стоящие на пересечении j-строки и k-столбца, j, $k = 1,2,3$.

Находим оценку корреляционной матрицы R_a

$$\hat{R}_a = \begin{pmatrix} 1 & -0{,}752 & -0{,}985 \\ -0{,}752 & 1 & 0{,}644 \\ -0{,}985 & 0{,}644 & 1 \end{pmatrix}.$$

Для проверки значимости отдельных коэффициентов регрессии, т.е. гипотез $H_0 : a_m = 0$, где $m = 0,1,2$ находим по таблицам F-распределения для $\alpha = 0{,}05$, $\upsilon_1 = 1$, $\upsilon_2 = 7$ критическое значение $F_{кр} = 5{,}59$.

Вычисляем F_p для каждого из коэффициентов регрессии по формуле $F_p(a_m) = a_m^2 / \hat{S}^2_{a_m}$, $m = 1,2$. Подставляя данные, получаем

$$F_p(a_1) = \frac{0{,}343396}{0{,}00505} = 67{,}99; \quad F_p(a_2) = \frac{0{,}487204}{2{,}16467} = 0{,}225.$$

Так как $F_p(a_1) > F_{кр} = 5{,}59$, то коэффициент регрессии a_1 значимо отличается от нуля. Для коэффициента a_2 выполняется неравенство $F_p(a_2) < F_{кр} = 5{,}59$, поэтому данный коэффициент можно считать незначимым.

3.3. Расчет моделей линейной регрессии на компьютере

Задание 1. Для анализа зависимости уровня средней месячной заработной платы рабочих (X_1) от стажа (X_2) и средней производительности труда (X_3) на промышленном предприятии было проведено обследование 30 рабочих. Полученные результаты представлены в табл. 3.5.

Таблица 3.5

Номер предприятия	X_1	X_2	X_3	Номер предприятия	X_1	X_2	X_3
1	240	8	37	16	236	5	21
2	280	10	33	17	300	10	40
3	265	15	28	18	248	6	31

Окончание табл. 3.5

Номер предприятия	X_1	X_2	X_3	Номер предприятия	X_1	X_2	X_3
4	350	4	45	19	214	8	28
5	295	12	29	20	280	2	30
6	270	10	38	21	165	4	15
7	180	5	24	22	180	2	17
8	250	7	28	23	315	20	45
9	310	12	34	24	200	7	20
10	345	15	38	25	274	11	37
11	220	6	26	26	194	5	25
12	180	3	18	27	267	18	32
13	175	3	20	28	280	10	45
14	190	6	21	29	320	12	50
15	215	3	29	30	380	18	55

Проведите корреляционно-регрессионный анализ взаимосвязи приведенных признаков с использованием пакета STATISTICA. Распечатайте и поясните полученные результаты.

При помощи t-критерия Стьюдента и F-критерия Фишера оцените значимость показателей тесноты связи и адекватность уравнения регрессии. При помощи частных коэффициентов детерминации оцените информативность отдельных факторных признаков с точки зрения их влияния на результативную переменную.

Решение. После ввода исходных значений анализируемых переменных в стартовой панели анализа данных выбираем модуль «Линейная регрессия» (**Linear regression**). В открывшемся окне (рис. 3.1) указываем зависимую (**Dependent**) и независимые переменные (**Independent**). Нажимаем **ОК**.

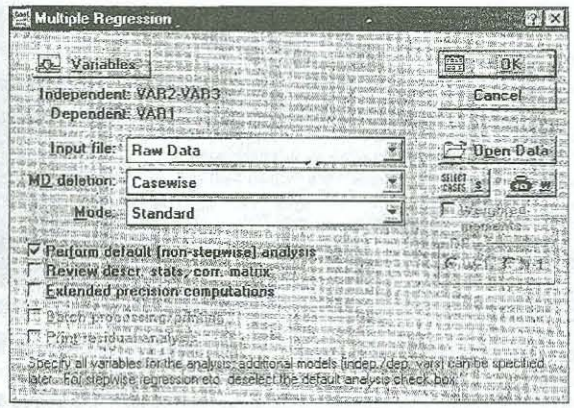


Рис. 3.1. Окно выбора переменных для многомерной регрессии

В качестве зависимой переменной в данном примере будет выступать величина средней заработной платы (VAR1), а две других переменных (VAR2, VAR3) выступают в качестве факторов. После выбора переменных, участвующих в анализе, на экране появляется окно с предварительными результатами и предлагаемым набором функций для дальнейшего анализа.

В диалоговом окне (рис. 3.2) содержатся начальные результаты работы модуля «Множественная регрессия» (**Multiple Regression**):

Dep. Var — зависимая переменная *Y*; **Multiple *R*** — множественный коэффициент корреляции, $R = 0{,}896$; **Standard error of estimate** — стандартная ошибка аппроксимации равна 26,7 %.

Кроме того, в окне указан свободный член уравнения и его стандартная ошибка: **Intercept** — свободный член уравнения равен 99,17; **Std. Error** — стандартная ошибка свободного члена равна 16,09 %;

F-критерий Фишера равен 55 при числе степеней свободы в числителе 2, а в знаменателе 27; расчетное значение *t*-критерия для оценки значимости свободного члена равно 6,16.

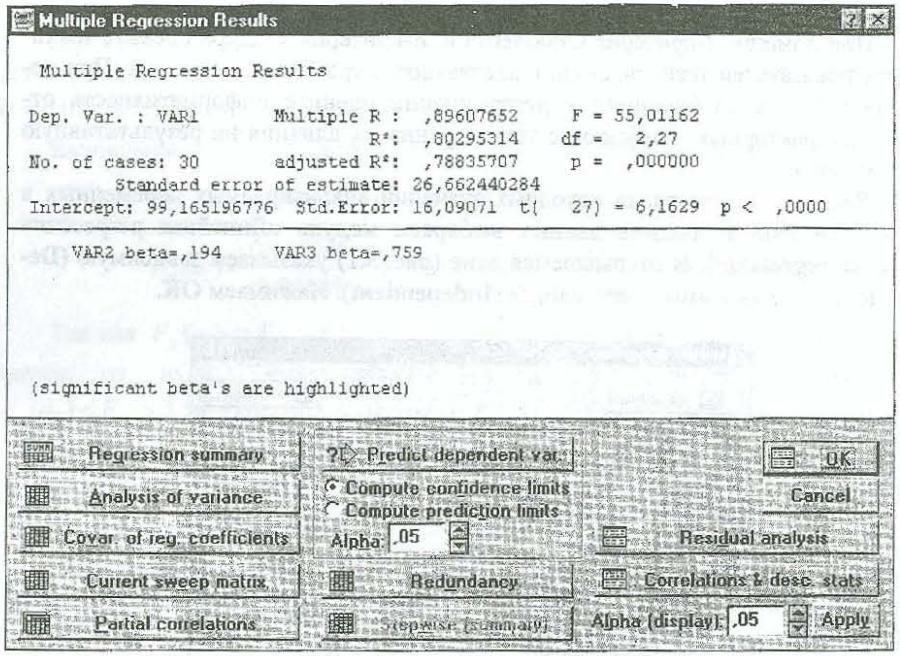


Рис. 3.2. Предварительные результаты регрессионного анализа

На рис. 3.3 даны также результаты дисперсионного анализа (факторная и остаточная дисперсия, а также расчетное значение *F*-критерия Фишера). Судя по критерию Фишера ($F_{расч.} = 55{,}0 > F_{табл.} = 3{,}369$), уравнение регрессии в целом следует признать адекватным.

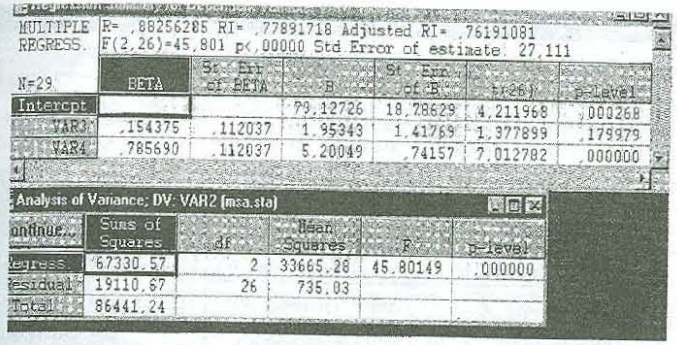
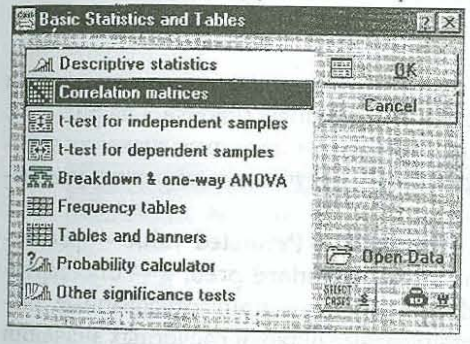


Рис. 3.3. Уравнение регрессии и оценка параметров регрессии

Множественный коэффициент детерминации $R^2 = 0{,}896$ можно разложить на частные коэффициенты детерминации. Для этого найдем попарные произведения β-коэффициентов на соответствующие парные коэффициенты корреляции факторов с результативной переменной. Рассчитать эти коэффициенты можно при помощи модуля **Basic Statistics and Tables** вторая функция **Correlation matrices**). Первый частный коэффициент детерминации равен 0,132 (0,194×0,68), а второй — 0,668 (0,759×0,88). Проверим правильность расчетов: 0,132 + 0,668 = 0,800, а множественный коэффициент детерминации равен $R^2 = 0{,}803$. Небольшое расхождение возникает за счет округления. Следовательно, на долю первого фактора приходится приблизительно 13,2 % объясненной дисперсии, а на долю второго фактора — 66,8 %. На основании этого, можно сделать вывод, что в данном случае вариация средней заработной платы в основном определяется вариацией производительности труда и слабо зависит от вариации стажа работы.



Для того чтобы продолжить регрессионный анализ, нажмем в левом верхнем углу активного окна кнопку **Continue,** вернемся в предыдущее окно (см. рис. 3.2) и выберем функцию **Residual analysis** — анализ остатков. На экране появится окно, содержащее шесть групп процедур для анализа остатков (рис. 3.4). Рассмотрим некоторые из них.

Первая группа процедур **STATISTICS** в активном окне (рис. 3.4) позволяет рассчитать основные характеристики остатков построенной регрессионной модели:

Statistics — статистики; **Casewise plot** — построчные графики; **Scatter plots** — диаграммы рассеяния; **Histograms** — гистограммы; **Probability plots** — вероятностные графики; **Bivariate Scatterplots** — диаграммы рассеяния, включающие зависимую переменную.

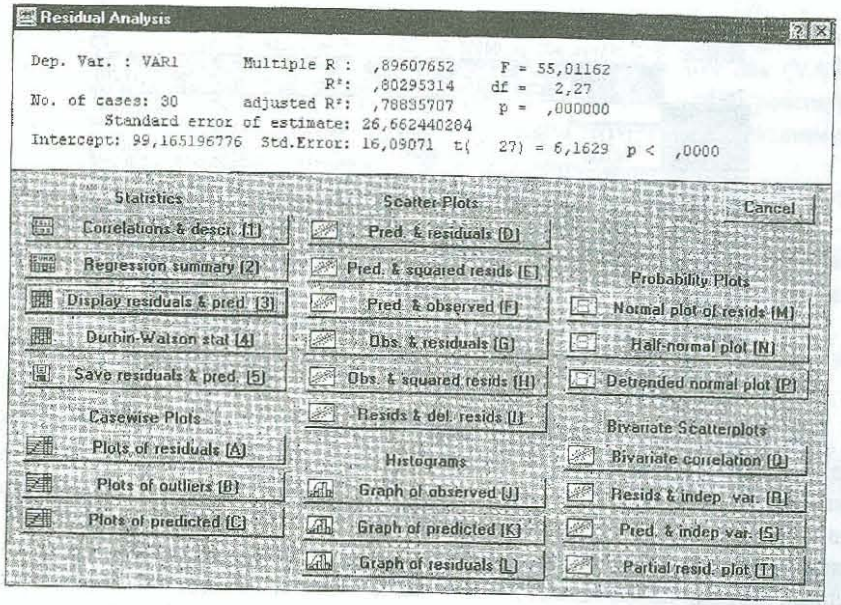


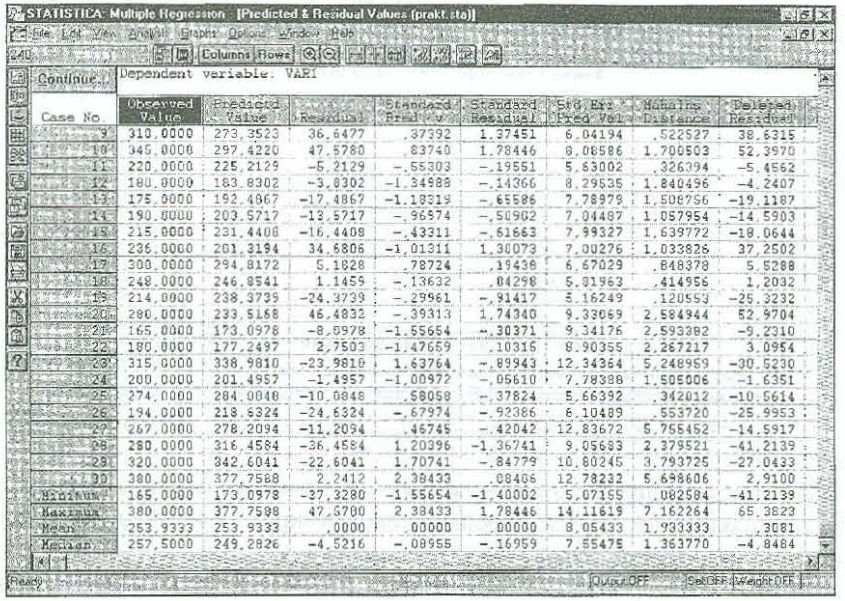
Рис. 3.4. Процедуры, предлагаемые для анализа остатков

Остальные процедуры относятся к графическому анализу остатков. Для примера рассмотрим процедуру **Prediceted and residuals** (кнопка 3 на рис. 3.4). В раскрывшемся окне появится таблица, содержащая исходные значения зависимой переменной и значения этой переменной, полученные на основе уравнения регрессии (рис. 3.5):

Observed value — наблюдаемые значения (Y); **Predicted value** — расчетные значения (Y_x); **Residual** — остатки ($Y—Y_x$); **Standard pred. v** — стандартизованные значения переменной; **Standard residual** — стандартизованные значения остатков; **Std. Err. Pred. val** — стандартные ошибки расчетных значений и другие характеристики.

В последних строках таблицы приведены обобщающие характеристики переменных (минимальное, максимальное, среднее значения и медиана). На рис. 3.6 представлен график расчетных значений переменной и остатков. По характеру распределения остатков можно судить о том, являются ли они случайными величинами или зависят от величины результативной переменной. В данном примере характер рассеяния остатков свидетельствует о том, что они являются случайными величинами. Следовательно, применение метода наименьших квадратов для нахождения оценок коэффициентов регрессии оправдано.

Модуль **«Анализ остатков» (Residual analysis)** позволяет также визуально оценить близость фактических и рассчитанных значений переменной. Для этого в меню модуля (см. рис. 3.4) нужно выбрать функцию **Pred. & residuals** (кнопка D). В развернувшемся окне будет представлен следующий график (рис. 3.7).



Case No.	Observed Value	Predicted Value	Residual	Standard Pred. v	Standard Residual	Std. Err. Pred. Val	Mahalano. Distance	Deleted Residual
9	310,0000	273,3523	36,6477	,37392	1,37451	6,04194	,522527	38,6315
10	345,0000	297,4220	47,5780	,83740	1,78446	8,08586	1,700503	52,3970
11	220,0000	225,2129	-5,2129	-,55303	-,19551	5,63002	,326394	-5,4562
12	180,0000	183,8302	-3,8302	-1,34988	-,14366	8,29535	1,840496	-4,2407
13	175,0000	192,4867	-17,4867	-1,18319	-,65586	7,78979	1,508956	-19,1187
14	190,0000	203,5717	-13,5717	-,96574	-,50902	7,04487	1,057954	-14,5903
15	215,0000	231,4408	-16,4408	-,43311	-,61663	7,99327	1,639772	-18,0644
16	236,0000	201,3194	34,6806	-1,01311	1,30079	7,00276	1,033826	37,2502
17	300,0000	294,8172	5,1828	,78724	,19436	6,67029	,848378	5,5288
18	248,0000	246,8541	1,1459	-,13632	,04298	5,91963	,414956	1,2032
19	214,0000	238,3739	-24,3739	-,29961	-,91417	5,16249	,120559	-25,3232
20	280,0000	233,5168	46,4832	-,39313	1,74340	9,33069	2,584944	52,9704
21	165,0000	173,0978	-8,0978	-1,55654	-,30371	9,34176	2,593382	-9,2310
22	180,0000	177,2497	2,7503	-1,47659	,10316	8,90355	2,267217	3,0954
23	315,0000	338,9810	-23,9810	1,63764	-,89943	12,34364	5,248959	-30,5230
24	200,0000	201,4957	-1,4957	-1,00972	-,05610	7,78388	1,506006	-1,6351
25	274,0000	284,0048	-10,0048	,58058	-,37824	5,66392	,342012	-10,5614
26	194,0000	218,6324	-24,6324	-,67974	-,92386	6,10489	,553720	-25,9953
27	267,0000	278,2094	-11,2094	,46745	-,42042	12,83672	5,755452	-14,5917
28	280,0000	316,4584	-36,4584	1,20396	-1,36741	9,05683	2,379521	-41,2139
29	320,0000	342,6041	-22,6041	1,70741	-,84779	10,80245	3,793725	-27,0433
30	380,0000	377,7588	2,2412	2,38433	,08406	12,78232	5,698606	2,9100
Minimum	165,0000	173,0978	-37,3280	-1,55654	-1,40002	5,07155	,082584	-41,2139
Maximum	380,0000	377,7588	47,5780	2,38433	1,78446	14,11619	7,162264	65,3823
Mean	253,9333	253,9333	,0000	,00000	,00000	8,05433	1,933333	,3081
Median	257,5000	249,2826	-4,5216	-,08955	-,16959	7,55475	1,363770	-4,8484

Рис. 3.5. Результаты расчетов на основании уравнения регрессии

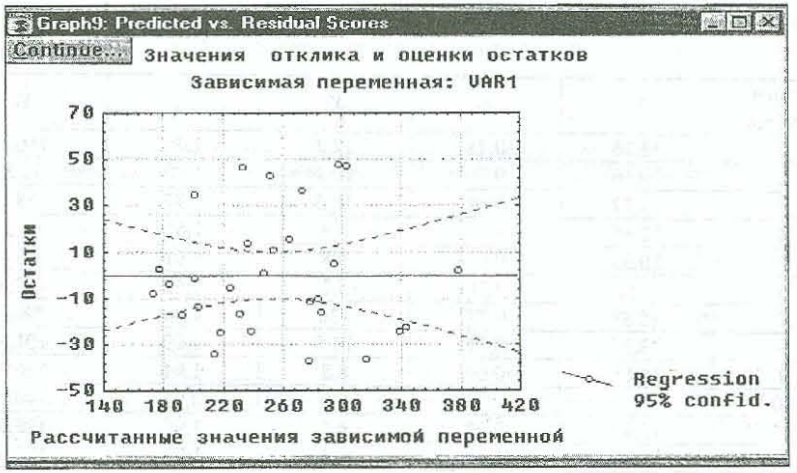


Рис. 3.6. График рассеяния остатков зависимой переменной

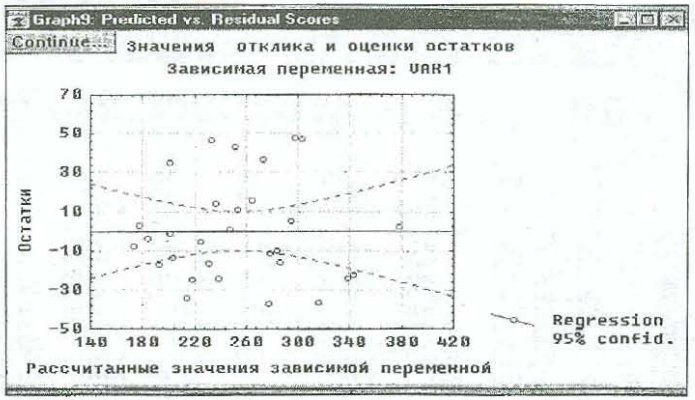


Рис. 3.7. График зависимости расчетных и наблюдаемых значений отклика

Судя по расположению точек на графике, можно утверждать, что построенная регрессионная модель хорошо аппроксимирует фактические значения зависимой переменной.

Задание 2. По приведенным ниже данным о работе 25 предприятий проведите корреляционный анализ взаимосвязи изучаемых признаков. Проверьте матрицу парных корреляций (R) на наличие мультиколлинеарности факторов.

Постройте две линейные регрессионные модели относительно отклика X_1, одну — по значениям всех факторных признаков, и вторую — без значений признака, порождающего мультиколлинеарность. Укажите «наилучшую» регрессионную модель и сделайте выводы (табл. 3.6).

Таблица 3.6

Номер объекта	X_1	X_2	X_3	X_4	X_5
1	13,26	0,78	12,3	1,45	166,3
2	10,16	0,75	10,4	1,30	92,8
3	13,72	0,68	18,0	1,37	158,0
4	12,85	0,70	4,3	1,65	93,9
5	10,63	0,62	8,8	1,91	173,9
6	9,12	0,76	5,7	1,68	162,3
7	25,83	0,73	170	1,94	88,6
8	23,39	0,71	17,8	1,89	101,2
9	14,68	0,69	8,8	1,94	166,3
10	10,05	0,73	6,0	2,06	140,8
11	13,99	0,68	8,2	1,96	128,5
12	9,68	0,74	8,4	1,02	177,8
13	10,03	0,66	6,7	1,35	114,5
14	9,13	0,72	10,4	0,88	93,2
15	5,37	0,68	6,6	0,62	126,7
16	9,86	0,77	8,6	1,09	91,8
17	12,62	0,78	7,9	1,60	69,1
18	5,02	0,78	3,4	1,53	66,1

Окончание табл. 3.6

Номер объекта	X_1	X_2	X_3	X_4	X_5
19	21,18	0,81	16,0	1,40	67,7
20	25,17	0,79	14,6	2,22	50,4
21	19,40	0,77	12,7	1,32	70,6
22	21,00	0,78	15,8	1,48	72,0
23	6,57	0,72	6,8	0,68	97,2
24	14,19	0,79	8,6	2,30	80,3
25	15,82	0,77	19,8	1,37	51,5

Здесь: X_1 — уровень рентабельности производства, %; X_2 — удельный вес рабочих в составе работающих; X_3 — размер премий и вознаграждений на одного работника, тыс. ден. ед.; X_4 — уровень фондоотдачи, ден. ед.; X_5 — оборачиваемость нормируемых оборотных средств, дней.

Решение. 1. Введем в таблицу исходные значения переменных и сохраним их в виде файла с именем *PRAKT2*. На главной панели инструментов щелкаем по кнопке **Analysis** и в раскрывшемся окне выбираем **Statup Panel** модуль **Basic Statistics** процедуру **Correlation matrices** и щелкаем по кнопке **OK**.

В раскрывшемся окне (рис. 3.8) указываем, для каких переменных должна быть построена матрица корреляций, выбираем и щелкаем на кнопке **Correlations.** На экране появится окно с матрицей парных корреляций, позволяющей судить о тесноте линейной корреляционной связи между признаками.

Как видно из этой матрицы, самая тесная прямая связь с результативной переменной у фактора X_3 ($r_{13} = 0{,}52$), а самая слабая связь у фактора X_2 ($r_{12} = 0{,}29$). Следует обратить внимание на тот факт, что между факторными переменными X_2 и X_5 существует достаточно тесная связь (r_{25}). Это может свидетельствовать о наличии мультиколлинеарности факторов и требует дальнейшей проверки данного предположения.

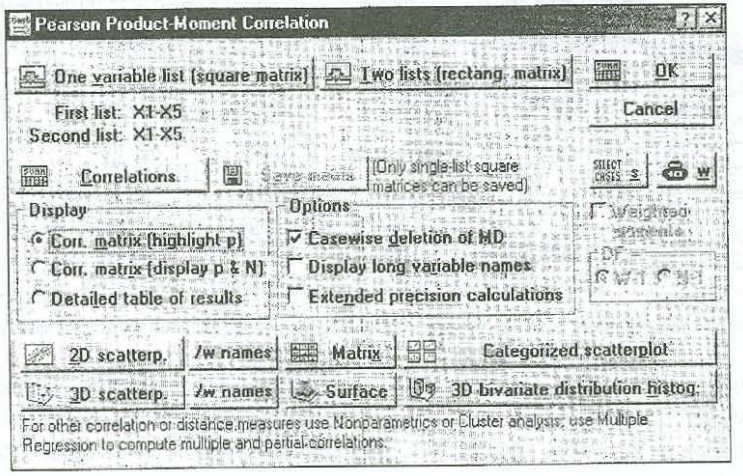


Рис. 3.8. Окно модуля **Correlation matrices**

После того как получены оценки парных коэффициентов корреляции, необходимо проверить их значимость. Для этого на рис. 3.8. отметим флажком процедуру **Corr. Matrix (display p & N)** и щелкнем по клавише **OK**.

В развернувшемся окне (рис. 3.9) появится корреляционная матрица, дополненная уровнями значимости рассчитанных коэффициентов корреляции. Из приведенных результатов видно, что в данном случае значимыми ($p < 0,05$) являются только коэффициенты корреляции между переменными

X_1 и X_3 ($r_{13} = 0,516$), X_1 и X_4 ($r_{14} = 0,508$), X_2 и X_5 ($r_{25} = -0,577$).

Correlations (prakt2.sta)

Continue... Marked correlations are significant at p < ,05000
N=25 (Casewise deletion of missing data)

Variable	X1	X2	X3	X4	X5
X1	1,00	,29	,52	,51	-,38
X2	,29	1,00	,01	,05	-,58
X3	,52	,01	1,00	,20	-,13
X4	,51	,05	,20	1,00	-,01
X5	-,38	-,58	-,13	-,01	1,00

Рис. 3.9. Матрица парных корреляций

Для остальных коэффициентов корреляции полученные значения и уровни их значимости не позволяют отвергнуть гипотезу о равенстве их нулю (рис. 3.10).

Correlations (prakt2.sta)

Continue... Marked correlations are significant at p < ,05000
N=25 (Casewise deletion of missing data)

Variable	X1	X2	X3	X4	X5
X1	1,0000	,2881	,5159	,5080	-,3790
	p= ---	p=,163	p=,008	p=,010	p=,062
X2	,2881	1,0000	,0088	,0547	-,5769
	p=,163	p= ---	p=,967	p=,795	p=,003
X3	,5159	,0088	1,0000	,2007	-,1343
	p=,008	p=,967	p= ---	p=,336	p=,522
X4	,5080	,0547	,2007	1,0000	-,0130
	p=,010	p=,795	p=,336	p= ---	p=,951
X5	-,3790	-,5769	-,1343	-,0130	1,0000
	p=,062	p=,003	p=,522	p=,951	p= ---

Рис. 3.10. Парные коэффициенты корреляции и уровни их значимости

Одним из простых приемов проверки наличия мультиколлинеарности факторов является сравнение с нулем определителя корреляционной матрицы. Проверяем гипотезу H_0: $|R| = 1$. В данном примере определитель матрицы R равен

$$|R| = \begin{vmatrix} 1 & 0,29 & 0,52 & 0,51 & -0,38 \\ & 1 & 0,01 & 0,05 & -0,58 \\ & & 1 & 0,2 & -0,13 \\ & & & 1 & -0,1 \\ & & & & 1 \end{vmatrix} = 0,28586.$$

Для проверки гипотезы воспользуемся критерием Уилкса

$$\chi^2_p = -\left(n - \frac{1}{6}(2m+5)\right)\ln|R| = -\left(25 - \frac{1}{6}(2\times 5+5)\right)\ln 0,28586 = 28,176,$$

а табличное (критическое) значение для уровня значимости $\alpha = 0,01$ и числа степеней свободы $\nu = 0,5m(m-1) = 0,5 \times 5(5-1) = 10$ равно $\chi^2_{кр} = 23,209$. Поскольку $\chi^2_{кр} < \chi^2_p$ нулевая гипотеза отклоняется, т.е. определитель $|R| \neq 1$, следовательно, наличие мультиколлинеарности факторных признаков считается доказанным.

2. Для продолжения анализа корреляционной связи между переменными необходимо исключить одну из двух факторных переменных, порождающих мультиколлинеарность (X_2 или X_5). Судя по значениям парных линейных коэффициентов корреляции ($r_{12} = 0,288$ $r_{15} = -0,379$), из рассмотрения следует удалить переменную X_2, так как у нее слабее связь с зависимой переменной X_1.

Рассчитаем два уравнения регрессии: первое, включающее все факторные переменные, а второе — без значений переменной X_2.

I вариант. Расчет уравнения регрессии по значениям всех факторных переменных и зависимой переменной X_1.

В окне переключения модулей (рис. 3.11) выбираем модуль **Multiple Regression.**

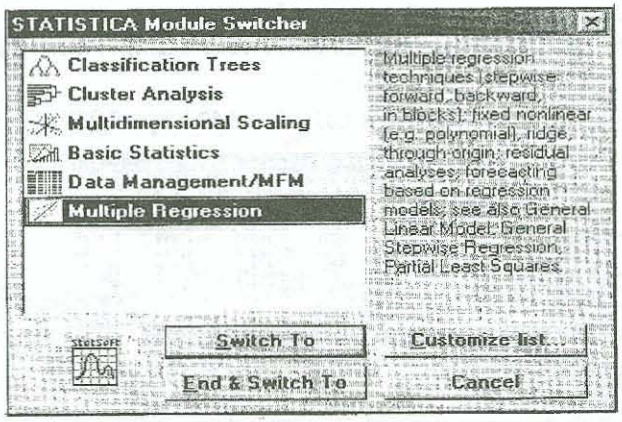


Рис. 3.11. Окно переключения модулей

В раскрывшемся окне указываем зависимую переменную X_1 и независимые переменные X_2, X_3, X_4, X_5. Щелкаем по кнопке **OK**, и на экране разворачивается окно с основными процедурами регрессионного анализа (рис. 3.12).

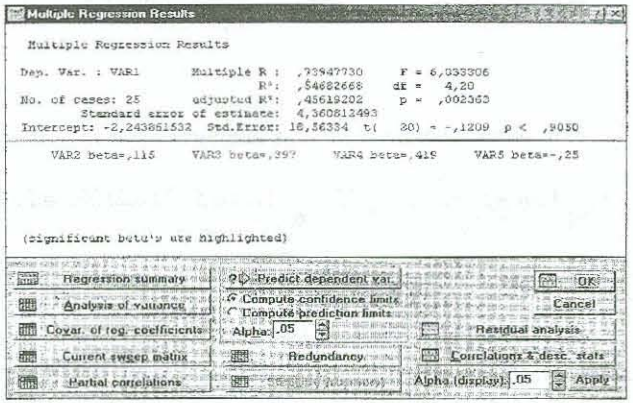


Рис. 3.12. Предварительные результаты множественной регрессии

Dep. Var — зависимая переменная X_1; ***Multiple R*** — множественный коэффициент корреляции $R = 0{,}739$; **Standard error of estimate** — стандартная ошибка аппроксимации $A = 4{,}36\ \%$; **Intercept** — свободный член уравнения $a_0 = -2{,}244$; **Std. Error** — стандартная ошибка свободного члена = 18,56 %.

Анализ полученных результатов показывает, что связь между зависимой переменной X_1 (уровень рентабельности производства) и факторными переменными (X_2, X_3, X_4, X_5) достаточно тесная ($R = 0{,}739$). Судя по величине множественного коэффициента детерминации ($R^2 = 0{,}547$), вариация X_1 на 54,7 % определяется рассматриваемыми факторами.

Чтобы получить на экране само уравнение регрессии, в активном окне (рис. 3.12) щелкнем по кнопке **Regression Summary.**

Результаты расчетов приведены на рис. 3.13. Поясним их смысл. Уравнение регрессии будет иметь вид

$$Y = -2{,}244 + 14{,}038X_2 + 0{,}073X_3 + 5{,}527X_4 - 0{,}037X_5.$$

MULTIPLE REGRESS. R= ,73947730 RI= ,54682668 Adjusted RI= ,45619202
F(4,20)=6,0333 p<,00236 Std.Error of estimate: 4,3608

N=25

	BETA	St. Err. of BETA	B	St. Err. of B	t(20)	p-level
Intercept			-2,24386	18,56334	-,12088	,904995
X2	,115298	,185496	14,03787	22,58479	,62156	,541250
X3	,396793	,155822	,07274	,02856	2,54645	,019223
X4	,418752	,154122	5,52697	2,03421	2,71701	,013274
X5	-,253771	,186932	-,03691	,02719	-1,35756	,189726

Analysis of Variance; DV: X1 (prakt2.sta)

Continue...	Sums of Squares	df	Mean Squares	F	p-level
Regress	458,9340	4	114,7335	6,033306	,002363
Residual	380,3337	20	19,0167		
Total	839,2677				

Рис. 3.13. Уравнение регрессии и его оценки для первого варианта

Для проверки значимости полученного уравнения регрессии в целом воспользуемся F-критерием Фишера. Расчетное значение $F_p = 6{,}033$, а табличное значение при числе степеней свободы в числителе 4, а в знаменателе 20, равно $F_{кр} = 2{,}87$. Так как $F_p > F_{кр.}$, уравнение регрессии в целом следует считать значимым.

На рис. 3.12 под чертой указаны значения β-коэффициентов для каждой переменной: $\beta_2 = 0{,}115$; $\beta_3 = 0{,}397$; $\beta_4 = 0{,}419$; $\beta_5 = -0{,}25$.

Для того чтобы определить влияние отдельных факторов на зависимую переменную, исчислим частные коэффициенты детерминации (r^2_j)

$$r_2^2 = r_{12}\beta_2 = 0{,}2881 \times 0{,}115 = 0{,}033; \qquad r_3^2 = r_{13}\beta_3 = 0{,}5159 \times 0{,}397 = 0{,}205;$$
$$r_4^2 = r_{14}\beta_4 = 0{,}508 \times 0{,}419 = 0{,}213; \qquad r_5^2 = r_{15}\beta_5 = -0{,}379 \times -0{,}25 = 0{,}095.$$

Проверим правильность расчетов

$$R^2 = r_2^2 + r_3^2 + r_4^2 + r_5^2 = 0{,}033 + 0{,}205 + 0{,}213 + 0{,}095 = 0{,}546\ (54{,}6\%).$$

Итак, на долю переменной X_2 приходится 3,3 % объясненной дисперсии зависимой переменной, на долю X_3 — 20,5 %, на долю X_4 — 21,3 % и на долю X_5 — 9,5 %. Следовательно, переменная X_2 является наименее значимой по своему влиянию на зависимую переменную X_1.

II вариант. Расчет уравнения регрессии после удаления значений переменной X_2, порождающей мультиколлинеарность (рис. 3.14).

Аналогично тому, как это было сделано в первом варианте, выполним все расчеты без переменной X_2.

Уравнение регрессии будет иметь вид

$$Y = 9{,}028 + 0{,}071X_3 + 5{,}624X_4 - 0{,}047X_5.$$

MULTIPLE REGRESS. R= ,73353440 RI= ,53807272 Adjusted RI= ,47208311
F(3,21)=8,1539 p<,00086 Std.Error of estimate: 4,2966

N=25

	BETA	St. Err. of BETA	B	St. Err. of B	t(21)	p-level
Intercept			9,028471	3,904331	2,31242	,030985
X3	,387238	,152779	,070987	,028007	2,53462	,019277
X4	,426091	,151407	5,623843	1,998376	2,81421	,010393
X5	-,321472	,149684	-,046762	,021773	-2,14767	,043567

Analysis of Variance; DV: X1 (prakt2.sta)

Continue...	Sums of Squares	df	Mean Squares	F	p-level
Regress	451,5870	3	150,5290	8,153900	,000865
Residual	387,6806	21	18,4610		
Total	839,2677				

Рис. 3.14. Уравнение регрессии и его оценки для второго варианта

Если сравнить его с уравнением, рассчитанным для первого варианта анализа, то мы увидим, что незначительно изменились коэффициенты регрессии для оставшихся переменных, но одновременно довольно сильно изменилось значение свободного члена уравнения.

Оценка нового уравнения по F-критерию Фишера показывает расчетное значение F-критерия — $F_p = 8{,}154$, а табличное значение при числе степеней свободы $\nu_1 = 3$ $\nu_2 = 21$ и уровне значимости $\alpha = 0{,}05$ равно $F_{кр} = 3{,}072$. Так как $F_p > F_{кр}$, то уравнение регрессии в целом следует считать значимым.

Повторим также анализ множественных коэффициентов корреляции и детерминации. Для этого нам понадобится новая матрица корреляций (рис. 3.15).

Continue... Marked correlations are significant at p < ,05000
N=25 (Casewise deletion of missing data)

Variable	X1	X3	X4	X5
X1	1,0000	,5159	,5080	-,3790
	p= ---	p=,008	p=,010	p=,062
X3	,5159	1,0000	,2007	-,1343
	p=,008	p= ---	p=,336	p=,522
X4	,5080	,2007	1,0000	-,0130
	p=,010	p=,336	p= ---	p=,951
X5	-,3790	-,1343	-,0130	1,0000
	p=,062	p=,522	p=,951	p= ---

Рис. 3.15. Парные коэффициенты корреляции и уровни их значимости для второго варианта анализа

Новые значения β-коэффициентов для переменных равны: $\beta_3 = 0{,}387$; $\beta_4 = 0{,}426$; $\beta_5 = -0{,}321$. Для того чтобы определить влияние отдельных факторов на зависимую переменную, исчислим частные коэффициенты детерминации (r^2_j):

$$r_3^2 = r_{13}\beta_3 = 0{,}516 \times 0{,}387 = 0{,}199;$$
$$r_4^2 = r_{14}\beta_4 = 0{,}508 \times 0{,}426 = 0{,}216;$$
$$r_5^2 = r_{15}\beta_5 = -0{,}379 \times -0{,}321 = 0{,}122.$$

Множественный коэффициент детерминации равен $R^2 = 0{,}538$, а множественный коэффициент корреляции $R = 0{,}734$. Проверим правильность расчетов. Множественный коэффициент детерминации равен сумме частных коэффициентов

$$R^2 = r_3^2 + r_4^2 + r_5^2 = 0{,}199 + 0{,}216 + 0{,}122 = 0{,}538 \ (53{,}8\ \%).$$

Итак, на долю переменной X_3 приходится около 20 % объясненной дисперсии зависимой переменной X_1, на долю X_4 — 21,6 %, на долю X_5 — 12,2 %. Следовательно, в данном варианте переменная X_5 является наименее значимой по своему влиянию на зависимую переменную X_1.

Что касается множественных коэффициентов детерминации и корреляции, то их значения уменьшились, но незначительно.

3.4. Контрольные задания

Задание 1. На основании данных, приведенных в табл. 3.7, проведите анализ зависимости индекса уровня человеческого развития от объемов государственного финансирования систем образования и здравоохранения.

Таблица 3.7

Страна	X_1	X_2	X_3
Германия	0,925	4,8	7,9
Бельгия	0,939	3,1	6,3
Дания	0,936	8,1	6,9
Франция	0,928	6,0	7,3
Греция	0,885	3,1	4,7
Великобритания	0,928	5,3	5,8
Ирландия	0,925	6,0	5,2
Италия	0,913	4,9	5,8
Люксембург	0,925	4,0	5,7
Нидерланды	0,935	5,1	6,0
Португалия	0,880	5,8	5,1
Испания	0,913	5,0	5,4
США	0,939	5,4	5,7

Здесь: X_1 — индекс человеческого развития за 2000 г.; X_2 — государственные расходы на образование в процентах к ВВП за 1995—1997 гг.; X_3 — государственные расходы на здравоохранение в процентах к ВВП за 1998 г.

Постройте линейную регрессионную модель зависимости переменной X_1 от переменных X_2 и X_3. В матричном виде рассчитайте парные и множественные коэффициенты корреляции и детерминации и оцените их значимость. Поясните полученные результаты и сделайте выводы.

Задание 2. По приведенным данным табл. 3.8 о результатах производственной и коммерческой деятельности в 14 хозяйствах области рассчитайте матрицу парных коэффициентов корреляции, проверьте ее на наличие мультиколлинеарности независимых переменных. Постройте модели линейной регрессии относительно зависимой переменной (X_1) в натуральном и стандартизованном масштабах.

Таблица 3.8

Номер хозяйства	X_1	X_2	X_3	X_4
1	460	1735	1141	649
2	937	3959	2110	448
3	1603	7441	2709	2549
4	1003	3650	1650	835
5	1037	5140	2753	589
6	677	3587	2164	330
7	558	3239	1507	684
8	2793	10610	4699	2327

Окончание табл. 3.8

Номер хозяйства	X_1	X_2	X_3	X_4
9	1925	11604	3865	2498
10	1036	4714	1795	795
11	772	4345	2245	632
12	1156	5777	2739	1155
13	1391	5834	2908	1383
14	677	3290	1721	723

Здесь: X_1 — прибыль от реализации продукции, тыс. ден. ед.; X_2 — объем реализованной продукции, тыс. ден. ед.; X_3 — запасы товарно-материальных ценностей на конец года, тыс. ден. ед.; X_4 — средства, используемые на развитие производства, тыс. ден. ед.

Рассчитайте значение F-критерия при $\alpha = 0{,}05$ и с его помощью оцените значимость уравнения регрессии и множественного коэффициента корреляции. Проанализируйте полученные результаты и сделайте выводы.

Задание 3. По данным табл. 3.9 об уровне цен импорта отдельных продуктов за ряд лет постройте матрицу парных корреляций (R). Оцените существующие взаимосвязи между уровнями цен на различные виды импортной продукции. Рассчитайте линейную регрессионную модель зависимости уровня цен одной тонны пшеницы от уровня цен на другие виды импортной продукции.

Оцените значимость регрессионной модели и матрицы парных корреляций при помощи статистических критериев.

Таблица 3.9

Средние цены импорта, USD за 1 т продукции			
пшеница Y	нефть X_1	бензин X_2	газ X_3
181,6	26,48	256,12	240,23
170,31	12,86	146,40	143,80
135,37	17,64	169,24	157,03
168,31	14,20	165,45	134,85
182,05	16,98	192,28	161,86
139,50	22,12	204,61	165,21
125,05	18,10	221,69	196,52

Все необходимые расчеты выполните на компьютере с использованием специального пакета прикладных программ STATISTICA. Протоколы работы соответствующих модулей распечатайте и поясните полученные результаты.

Задание 4. В ходе проведения множественного корреляционно-регрессионного анализа была получена система линейных уравнений относительно стандартизованных коэффициентов регрессии β_j

$$\begin{cases} 0{,}992 = \beta_2 + 0{,}938\beta_3 + 0{,}907\beta_4, \\ 0{,}955 = 0{,}938\beta_2 + \beta_3 + 0{,}930\beta_4, \\ 0{,}937 = 0{,}907\beta_2 + 0{,}930\beta_3 + \beta_4. \end{cases}$$

1. Рассчитайте значения стандартизированных коэффициентов β_j и постройте уравнение регрессии для зависимой переменной X_1 в стандартизованном масштабе.

2. В матричной форме рассчитайте множественные коэффициенты детерминации и корреляции, поясните полученные результаты.

Задание 5. На основании данных выборочной совокупности ($n = 30$) была рассчитана матрица парных корреляций

$$R = \begin{pmatrix} 1 & 0{,}680 & 0{,}880 \\ & 1 & 0{,}640 \\ & & 1 \end{pmatrix}.$$

1. Учитывая, что зависимой переменной (откликом) является переменная X_1, а независимыми переменными (предикторами) — X_2 и X_3, рассчитайте множественные коэффициенты детерминации и корреляции.

2. На основании критерия Снедекора оцените значимость множественного коэффициента детерминации при уровне значимости $\alpha = 0{,}01$.

3. Постройте уравнение множественной регрессии в стандартизированном масштабе и поясните смысл его параметров.

Задание 6. В результате проведения регрессионного анализа по 50 наблюдениям построена регрессионная модель

$$Y = 70{,}49 - 0{,}004X_1 + 0{,}182X_2 + 0{,}203X_3,$$

где Y — индекс физического объема ВВП, %; X_1 — индекс объема продукции промышленности, %; X_2 — индекс объема продукции сельского хозяйства, %; X_3 — индекс объема внешнеторгового оборота, %.

Множественный коэффициент детерминации $R^2 = 0{,}562$.

Наблюдаемые значения t-критерия Стьюдента для коэффициентов регрессии равны для: $a_1 = -0{,}004$; $t_p = -0{,}136$;
$a_2 = 0{,}182$; $t_p = 1{,}674$;
$a_3 = 0{,}203$; $t_p = 3{,}830$.

1. Проверьте достоверность полученной регрессионной модели в целом и отдельных ее коэффициентов.

2. Оцените степень тесноты связи между динамикой ВВП и рассматриваемыми факторами. Поясните полученные результаты.

4. ФАКТОРНЫЙ АНАЛИЗ

4.1. Методические рекомендации

Факторный анализ (ФА) представляет собой совокупность методов, которые на основе реально существующих связей анализируемых признаков, или связей самих наблюдаемых объектов, позволяют выявлять скрытые (неявные, латентные) обобщающие характеристики организационной структуры и механизма развития изучаемых явлений, процессов.

Методы факторного анализа в исследовательской практике применяются главным образом с целью сжатия информации, получения небольшого числа обобщающих признаков, объясняющих вариативность (дисперсию) элементарных признаков (*R*-техника факторного анализа) или вариативность наблюдаемых объектов (*Q*-техника факторного анализа).

Алгоритмы факторного анализа основываются на использовании редуцированной матрицы парных корреляций (ковариаций). Редуцированная матрица — это матрица, на главной диагонали которой расположены не единицы (оценки) полной корреляции или оценки полной дисперсии, а их редуцированные, несколько уменьшенные величины. При этом постулируется, что в результате анализа будет объяснена не вся дисперсия изучаемых признаков (объектов), а ее некоторая часть, обычно большая. Оставшаяся необъясненная часть дисперсии — это характерность, возникающая из-за специфичности наблюдаемых объектов, или ошибок, допускаемых при регистрации явлений, процессов, т.е. ненадежности вводных данных.

Несколько особняком от метода факторного анализа стоит метод главных компонент (МГК). Этот метод не относят к ФА, хотя он имеет схожий алгоритм и решает схожие аналитические задачи. Его главное отличие заключается в том, что обработке подлежит не редуцированная, а обычная матрица парных корреляций, ковариаций, на главной диагонали которой расположены единицы (в матрице ковариаций — оценки полной дисперсии). Иными словами, здесь предполагается объяснение всей дисперсии анализируемых признаков (необходимых объектов), а явление «характерности» во внимание не принимается.

При классификации методов ФА можно выделить следующие группы:

1. *Метод главных компонент* (Г. Хотеллинг).
2. *Упрощенные методы ФА*, обычно это методы, которые появились раньше, в первой половине двадцатого столетия, во время появления и формирования базисных теоретических разработок ФА. Эти методы отличаются, с одной стороны, сравнительно простыми вычислительными процедурами, а с другой стороны, ограниченными возможностями в выделении латентных факторов и аппроксимации факторных решений. В данную группу входят методы:

- *однофакторная модель Ч. Спирмена*, позволяет выделять только один латентный фактор;

- *бифакторная модель Г. Хользингера*, ориентирована на выделение двух латентных факторов;

- *центроидный метод Л. Тэрстоуна* — множество корреляций между переменными рассматривается как пучок векторов, латентный фактор в этом пучке появляется как некий уравновешивающий вектор, проходящий через его центр.

3. *Современные аппроксимирующие методы ФА* — методы, имеющие, по сравнению с предыдущей группой, более гибкую модель выделения латентных факторов (искусственно не ограничивающую их число), а также позволяющую оптимизировать полученные решения. В этой группе наиболее представительными являются:

- *метод главных факторов Г. Томсона* используется на практике особенно часто, наиболее близок методу главных компонентов;

- *групповой метод Л. Гуттмана и П. Хорста* основывается на исследовании не простого набора данных, а на предварительно отобранных каким-либо образом группах анализируемых признаков (наблюдаемых объектов).

4. *Методы с повышенными аппроксимирующими свойствами* — современные методы, позволяющие получать и последовательно улучшать аналитические результаты. Эти методы отличаются сложностью алгоритмов и высокой трудоемкостью вычислительных процедур, практически нереализуемы без технических средств. К этой группе относятся методы:

- максимального правдоподобия Д. Лоули и Д. Максвелла;
- минимальных остатков Г. Хармана;
- двухфакторного анализа Г. Кайзера и И. Кэффри;
- канонического факторного анализа К. Рао.

Несмотря на различия, многочисленные методы факторного анализа имеют общую алгоритмическую схему реализации (рис. 4.1).

В факторном анализе предполагается объяснение не всей дисперсии варьирующих элементарных признаков, а только некоторой ее части. Таким образом, элементы полученной матрицы факторного отображения (A) представляют только объясненную часть дисперсии — общность (h_j^2). Кроме того, остается необъясненная часть дисперсии, или характерность (d_j^2). Полное разложение дисперсии в факторном анализе, в зависимости от возможностей применяемых методов, можно представить в виде следующей схемы связей дисперсионных показателей.

Общность — доля дисперсии, объясненная действием общих факторов $h_j^2 = a_{j1}^2 + a_{j2}^2 + ... + a_{jm}^2$ и $h_j^2 = 1 - d_j^2$.

Характерность — доля дисперсии, необъясненной действием общих факторов $d_j^2 = 1 - h_j^2 = b_j^2 + l_j^2$.

Специфичность — доля дисперсии, обусловленной специфичной вариабельностью анализируемого признака (X_j) $b_j^2 = d_j^2 - l_j^2$.

Ненадежность — доля дисперсии, обусловленной несовершенством измерений (ошибками измерений) $l_j^2 = 1 - h_j^2 - b_j^2$.

Надежность — доля дисперсии характерного фактора без измерений ошибки $c_j^2 = h_j^2 + b_j^2 = 1 - l_j^2$.

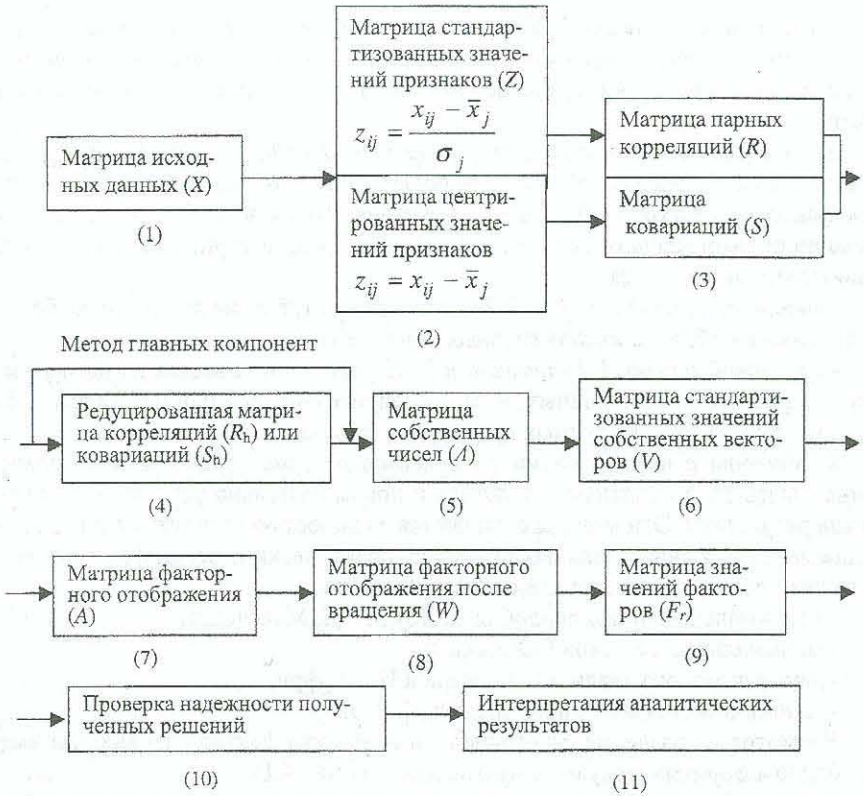


Рис. 4.1. Схема алгоритмов методов факторного анализа

Полная дисперсия — дисперсия варьирующих значений стандартизированного признака (Z_j) $h_j^2 + d_j^2 = h_j^2 + b_j^2 + l_j^2 = 1$.

Общие факторы выделяют последовательно: первый фактор, объясняющий наибольшую долю дисперсии исходных переменных, затем второй, объясняющий следующую по величине долю дисперсии, и т.д. Процесс выделения общих факторов может быть прерван, если объяснена достаточная доля дисперсии элементарных признаков.

В ходе анализа важное значение приобретает оценка достаточности числа выделенных общих (латентных) факторов; она находится с использованием критериев Бартлетта и Лоули. Для метода главных компонент применяется критерий Бартлетта

$$\chi_p^2 = -\left(n - \frac{1}{6}(2m+5) - \frac{2}{3}r\right)\ln R_{m-r}$$

где r — число оставленных в анализе главных компонент;

$$R_{m-r} = \frac{|R|}{\lambda_1 \times \lambda_2 \times \ldots \times \lambda_3 \left(\frac{m - \lambda_1 - \lambda_2 - \ldots - \lambda_r}{m-r}\right)^{m-r}}.$$

При $\chi_n^2 < \chi_{\alpha,\nu}^2$ $(\nu = 1/2((m-r) - m - r - 1))$ принимается гипотеза о достаточности числа выделенных, оставленных в анализе главных компонент, в противном случае число главных компонент в анализе должно быть увеличено, дисперсия элементарных признаков объясняется в недостаточной мере.

Для методов факторного анализа используется критерий Лоули

$$\chi_p^2 = (n-1)\ln\frac{|R^+|}{|R|},$$

где $|R^+|$ и $|R|$ — определители воспроизведенной ($R^+ = AA^T$) и исходной матриц парных корреляций.

Достаточность числа общих (латентных) факторов подтверждается, когда $\chi^2{}_p < \chi^2{}_{\alpha,\nu}$, со степенью свободы $\nu = 1/2((m - r^2) - m \cdot r)$.

Корректность решений, полученных при помощи методов факторного анализа, обеспечивается, прежде всего, значимостью матрицы корреляций (ковариаций). В случае, когда элементы корреляционной (ковариационной) матрицы незначимы, ее собственные числа будут близки к единице и найденные факторы (латентные переменные) почти не отличаются от исходных элементарных признаков, т.е. проведение факторного анализа теряет смысл.

Проверка значимости матрицы R производится с помощью критерия Уилкса-χ^2

$$\chi_p^2 = -\left(n - \frac{1}{6}(2m+5)\right)ln|R|,$$

где R — матрица парных корреляций, определитель $|R| = \lambda_1\lambda_2\ldots\lambda_m$; n, m —число наблюдаемых объектов и число элементарных признаков, участвующих в анализе, соответственно.

Значимость корреляционной матрицы (R) подтверждается, если $\chi_p^2 > \chi_{\alpha,\nu}^2$, при заданном уровне значимости α и числе степеней свободы $\nu = 1/2m(m-1)$.

В ходе построения матрицы факторного отображения (A), при необходимости, когда ее столбцы плохо структурированы и слабо поддаются интерпретации, производится вращение общих факторов. Вращение может быть ортогональным (при сохранении линейной независимости общих факторов) или косоугольным (в ходе вращения появляются линейнозависимые общие факторы). Посредством вращения решается задача упрощения структуры общих факторов. Другими словами, в процессе вращения факторные нагрузки одних элементарных признаков (наиболее значимых) возрастают, а других (менее значимых) — снижаются. В результате получают упрощенную структуру факторов, которая легче поддается объяснению.

Проблема вращения общих факторов, вследствие объемности материала и разнообразия методических подходов к ее решению, в данном пособии не рассматривается [16, 22].

На основе матрицы факторного отображения (A) (без вращения или после вращения) определяют значения главных факторов по каждому из наблюдаемых объектов. При условии, что в анализе остаются все главные факторы и их число равно числу элементарных признаков, матрица значений главных факторов определяется как $F = A^{-1}Z^T$. В другом случае, который является более естественным для ФА и встречается чаще, число главных факторов значительно меньше числа элементарных признаков ($r < m$), и матрица A — не квадратная, тогда матрица значений главных факторов определяется как $F = (A^TA)^{-1}A^TZ^T$.

Результатом решения задачи методом главных факторов, также как и методом главных компонент, будут системы линейных уравнений вида $Z = AF^T$; $F = \Lambda^{-1}A^TZ$ и матрица значений главных факторов (F).

4.2. Примеры решения типовых задач

Рассмотрим приведенный выше алгоритм факторного анализа (см. рис. 4.1) на основе решения прикладной задачи с применением метода главных компонент и метода главных факторов, имеющих достаточно простую и наглядную конструкцию.

Пример 1. Пусть имеется матрица исходных данных, представляющая совокупность четырех промышленных предприятий, оцененных по трем признакам: X_1 — уровень выработки на одного среднегодового работника; X_2 — уровень рентабельности продукции, %; X_3 — уровень фондоотдачи основных фондов, ден. ед. Обратим внимание, что в факторном анализе исходная матрица (X) строится как транспонированная

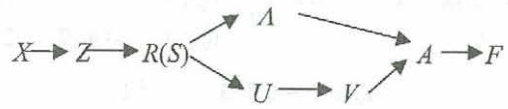
$$X^T = \begin{pmatrix} 9{,}26 & 9{,}38 & 12{,}11 & 10{,}81 \\ 13{,}26 & 10{,}16 & 13{,}72 & 12{,}85 \\ 1{,}45 & 1{,}30 & 1{,}37 & 1{,}65 \end{pmatrix}.$$

Используем сначала метод главных компонент, его реализация предусматривает решение следующих формальных уравнений:

$$R = \frac{1}{n}ZZ^T; \quad |R - \lambda E| = 0; \quad (R - \lambda E)U = 0; \quad A = V\Lambda^{1/2}, \text{ с } V = \|V_j\| \text{ и } V_j = \frac{U_j}{|U_j|};$$

$F = A^{-1}Z^T$. В приведенных формулах $Z = \|z_{ij}\|$, $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$ и $X = X^T$ — транспонированная матрица исходных данных, размерности $(n \times m)$.

Схематично алгоритм метода главных компонент имеет вид



1. Стандартизуем значения изучаемых признаков

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \text{ получим } Z' = \begin{pmatrix} -0{,}971 & -0{,}868 & 1{,}478 & 0{,}361 \\ 0{,}549 & -1{,}684 & 0{,}882 & 0{,}253 \\ 0{,}076 & -1{,}069 & -0{,}534 & 1{,}527 \end{pmatrix}.$$

2. Найдем матрицу парных корреляций $R = \dfrac{1}{n}Z^TZ$

$$R = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}.$$

3. Поиск матрицы собственных чисел в конечном счете означает решение характеристического уравнения $(R - \lambda E)V = 0$, откуда $|R - \lambda E| = 0$.

Существует множество различных подходов для вычисления элементов матрицы λ, остановимся на одном из них, широко применяемом на практике и использующем рекуррентные соотношения Фаддеева: пусть A — некоторая симметрическая матрица, размерностью $m \times m$, тогда ее определитель находим по следу матриц, производных от A:

$A_1 = A$	$P_1 = trA_1$	$B_1 = A_1 - P_1E$
$A_2 = AB_1$	$P_2 = 1/2trA_2$	$B_2 = A_2 - P_2E$
…	…	…
$A_{m-1} = AB_{m-2}$	$P_{m-1} = 1/(m-1)trA_{m-1}$	$B_{m-1} = A_{m-1} - P_{m-1}E$
$A_m = AB_{m-1}$	$P_m = 1/m \cdot trA_m$	$B_m = A_m - P_mE;\ B_m = 0$

На заключительном этапе расчетов $P_m = |A|$ можем записать характеристический многочлен

$$P_m(\lambda) = \lambda^m - P_1\lambda^{m-1} - P_2\lambda^{m-2} - \ldots - P_m.$$

Приравнивая характеристический многочлен нулю, можем получить его корни, т.е. множество значений λ.

В нашем случае: $R = A$, $A = A_1$, $P_1 = trA_1 = 1 + 1 + 1 = 3$,

$$B_1 = A_1 - P_1E = \begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & 2 & 0{,}439 \\ 0{,}154 & 0{,}439 & 2 \end{pmatrix};$$

$$A_2 = AB_1 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix} \begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & -2 & 0{,}439 \\ 0{,}154 & 0{,}439 & -2 \end{pmatrix} =$$

$$= \begin{pmatrix} -1{,}638 & -0{,}513 & 0{,}101 \\ -0{,}513 & -1{,}469 & -0{,}350 \\ 0{,}101 & -0{,}350 & -1{,}783 \end{pmatrix};$$

$$P_2 = 1/2 tr A_2 = 1/2[(-1{,}638) + (-1{,}469) + (-1{,}783)] = -2{,}445;$$

$$B_2 = A_2 - P_2 E = \begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix};$$

$$A_3 = AB_2 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix} \begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}524 & 0 & 0 \\ 0 & 0{,}524 & 0 \\ 0 & 0 & 0{,}524 \end{pmatrix};$$

$$P_3 = 1/3(3 \times 0{,}524) = 0{,}524; \; B_3 = A_3 - P_3 E = 0.$$

Таким образом, $|R| = 0{,}524$ и характеристический многочлен будет $\lambda^3 - 3\lambda^2 + 2{,}445\lambda - 0{,}524 = 0$, откуда $\lambda_1 = 1{,}798$, $\lambda_2 = 0{,}875$, $\lambda_3 = 0{,}327$.

$$\lambda = \begin{pmatrix} 1{,}798 & 0 & 0 \\ 0 & 0{,}875 & 0 \\ 0 & 0 & 0{,}327 \end{pmatrix}.$$

4. Матрицу собственных векторов определим, решая системы линейных уравнений, каждому из собственных чисел (λ_j) соответственно. Так как подобные системы уравнений имеют бесконечное множество решений, каждый раз одному из неизвестных признаков будем задавать произвольное значение, например единицу:

для $\lambda_1 = 1{,}798$ имеем

$$\begin{array}{l|l} (1 - 1{,}798)u_{11} + 0{,}581u_{22} + 0{,}154u_{33} = 0 & u_{11} = 1{,}262 \\ 0{,}581u_{11} + (1 - 1{,}798)u_{21} + 0{,}439u_{31} = 0 & u_{21} = 1{,}469 \\ 0{,}154u_{11} + 0{,}439u_{21} + (1 - 1{,}798)u_{31} = 0 & u_{31} = 1{,}000 \end{array}$$

для λ_2=0,875

$$\begin{array}{l|l} (1 - 0{,}875)u_{12} + 0{,}581u_{22} + 0{,}154u_{32} = 0 & u_{31} = 1{,}307 \\ 0{,}581u_{12} + (1 - 0{,}875)u_{22} + 0{,}439u_{32} = 0 & u_{22} = 1{,}779 \\ 0{,}154u_{12} + 0{,}439u_{22} + (1 - 0{,}875)u_{32} = 0 & u_{32} = 1{,}000 \end{array}$$

для $\lambda_3 = 0{,}327$

$$\begin{array}{l|l} (1 - 0{,}327)u_{13} + 0{,}581u_{23} + 0{,}154u_{33} = 0 & u_{13} = 1{,}307 \\ 0{,}581u_{13} + (1 - 0{,}327)u_{23} + 0{,}439u_{33} = 0 & u_{23} = 1{,}779 \\ 0{,}154u_{13} + 0{,}39u_{23} + (1 - 0{,}327)u_{33} = 0 & u_{33} = 1{,}000 \end{array}$$

Матрица собственных векторов принимает вид

$$U = \begin{pmatrix} 1{,}262 & -0{,}144 & 1{,}307 \\ 1{,}469 & -0{,}234 & -1{,}779 \\ 1{,}000 & 1{,}000 & 1{,}000 \end{pmatrix}.$$

5. Поскольку в ходе расчетов собственные векторы приобретают различные шкалы измерения, следует привести их к нормируемому виду $V_j = Uj / |Uj|$

$$V = \begin{pmatrix} 0{,}579 & -0{,}139 & 0{,}539 \\ 0{,}674 & -0{,}225 & -0{,}734 \\ 0{,}459 & 0{,}964 & 0{,}413 \end{pmatrix}.$$

6. Теперь можем построить матрицу факторного отображения (A), $A = V\lambda^{1/2}$

$$A = \begin{pmatrix} 0{,}579 & -0{,}139 & 0{,}539 \\ 0{,}674 & -0{,}225 & -0{,}734 \\ 0{,}459 & 0{,}964 & 0{,}413 \end{pmatrix} \begin{pmatrix} \sqrt{1{,}798} & 0 & 0 \\ 0 & \sqrt{0{,}875} & 0 \\ 0 & 0 & \sqrt{0{,}327} \end{pmatrix} =$$

	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

Элементы матрицы A — это коэффициенты частной корреляции, характеризующие тесноту связей X_j — элементарных признаков с F_r — латентными факторами (главными компонентами).

Для матрицы A выполняется условие $A^T A = \lambda$, т.е. $\sum_j a_{jr}^2 = \lambda_r$,

или $\sum_j a_{j1}^2 = 0{,}776^2 + 0{,}904^2 + 0{,}616^2 = 1{,}798$; $\sum_j a_{j2}^2 = 0{,}875$; $\sum_j a_{j3}^2 = 0{,}327$.

Матрица A, представляющая признаковую структуру каждой из главных компонент, позволяет определять названия главных компонент:

F_1 — объясняет (1,798/3 = 0,599) более 60 % общей дисперсии элементарных признаков — может быть названа эффективностью производства;

F_2 — объясняет (0,875/3 = 0,292) около 30 % общей дисперсии признаков — назовем ее «эффективность использования основных средств»;

F_3 — объясняет (0,327/3 = 0,109) около11 % общей дисперсии. Ввиду низкой значимости этой главной компоненты в дальнейшем анализе она может не рассматриваться.

С целью более точной оценки структуры каждой из главных компонент может применяться специальный коэффициент уровня информативности составляющих ее элементарных признаков

$$K_u = \frac{\sum_j a_{jr}^2}{\sum_{j=1}^{m} a_{jr}^2},$$

где $\sum_{j=1}^{m} a_{jr}^2$ — сумма квадратов всех значений нагрузок элементарных признаков для главной компоненты F_r; $\sum_j a_{jr}^2$ — сумма квадратов нагрузок тех элементарных признаков, которые наиболее значимы и в основном формируют название главной компоненты (F_r).

Исчисленная матрица A позволяет записать уравнения связи элементарных признаков с главными компонентами:

$$Z_1 = 0{,}776F_1 - 0{,}130F_2 + 0{,}308F_3;$$
$$Z_2 = 0{,}904F_1 - 0{,}210F_2 - 0{,}420F_3;$$
$$Z_3 = 0{,}616F_1 + 0{,}902F_2 + 0{,}236F_3,$$

и наоборот — главных компонент с элементарными признаками:

$$F_1 = 1/1{,}798(0{,}776Z_1 + 0{,}904Z_2 + 0{,}616Z_3);$$
$$F_2 = 1/0{,}875(-0{,}130Z_1 - 0{,}210Z_2 + 0{,}236Z_3).$$

7. Расчет матрицы F позволяет получить значения главных компонент по каждому наблюдаемому объекту $F = A^{-1}Z^T$:

$$F = \left(\begin{array}{ccc|cccc} 0{,}542 & 0{,}507 & 0{,}196 & -0{,}971 & -0{,}868 & 1{,}478 & 0{,}361 \\ -0{,}776 & -0{,}010 & 0{,}994 & 0{,}549 & -1{,}684 & 0{,}882 & 0{,}253 \\ 1{,}554 & -1{,}283 & -0{,}075 & 0{,}076 & -1{,}069 & -0{,}534 & 1{,}527 \end{array}\right) =$$

Проверка: $\sum f_{ri} = 0$

	n_1	n_2	n_3	n_4
F_1	−0,233	−1,533	1,143	0,623
F_2	0,823	−0,372	−1,687	1,236
F_3	−2,218	0,892	1,205	0,121

Теперь для обработки тех же исходных данных используем метод главных факторов. Построим редуцированную матрицу корреляций. С этой целью можно применить один из следующих подходов.

1. *Метод наибольшей корреляции.* На главной диагонали с положительным знаком записывается наибольший по величине коэффициент корреляции столбца.

2. *Метод Барта.* По каждому столбцу матрицы R вначале находят среднее значение коэффициентов корреляции (r_{ij}), затем, если оно относительно велико, за общность принимается значение несколько выше наибольшего в столбце коэффициента корреляции, а если сравнительно мало, несколько меньше наибольшего в столбце коэффициента корреляции.

3. *Метод триад.* Для каждого столбца общность вычисляют по формуле

$$h_j^2 = \frac{r_{kj} r_{ij}}{r_{kl}}$$

где r_{kj}, r_{ij} — коэффициенты корреляции наибольшие в столбце.

4. *Метод малого центроида.* Для каждой j-й переменной строится корреляционная матрица, размерностью 4×4. Включая саму переменную, в эту матрицу записывают оценки корреляции трех других переменных, наиболее тесно связанных с первой. По данным малой матрицы корреляций (малого центроида) рассчитываются общности

$$h_j^2 = \frac{\left(\sum r_{i1}\right)^2}{\sum_{ij} r_{ij}},$$

где $\sum r_{i1}$ — сумма элементов первого столбца; $\sum_{ij} r_{ij}$ — сумма всех элементов малого центроида.

Используя простой метод Барта, построим для нашего примера редуцированную матрицу корреляций (R_h)

$$R_h = \begin{pmatrix} 0{,}670 & 0{,}581 & 0{,}154 \\ 0{,}581 & 0{,}630 & 0{,}439 \\ 0{,}154 & 0{,}439 & 0{,}420 \end{pmatrix}.$$

В дальнейшем определение собственных чисел и собственных векторов может осуществляться выполнением уже продемонстрированных ранее шагов алгоритма метода главных компонент или применением подхода Хотеллинга, предусматривающего процедуру многократного возведения в квадрат исходной матрицы (R_h). Выберем подход Хотеллинга.

Для имеющейся у нас редуцированной матрицы выполним операцию возведения в степень $R^2 = R^T R$. Процедура должна повторяться до тех пор, пока некоторые величины (α-оценки матрицы R) до и после возведения в степень не перестанут существенно различаться, т.е. $d = (a_{(i)} - a_{(i-1)})$ должны быть минимальны, меньше некоторого заранее заданного порогового уровня. Оценки a — это приближения факторного отображения, $a = p/p_{\max}$, где p — скаляр ($p_i^{i+1} = R_h^1 S^1$), соответствующий величине суммы коэффициентов корреляции по каждой строке ($S = \sum_j r_{ij}$), значения величин p и S взаимно контролируются.

Матрицу R_h будем возводить в степень, одновременно вычисляя оценки a, S и p, результаты расчетов сведем в табл. 4.1— 4.4.

После первого возведения в квадрат матрицы R_h значения α-характеристики еще достаточно велики. Продолжим операцию умножения матриц. Так как мы всякий раз имеем дело с симметрической матрицей, расчеты можно значительно сократить, вычисляя только элементы по главной диагона-

ли и над главной диагональю. Кроме того, несколько шагов возведения в квадрат можно пропустить, используя матрицы исходных данных во второй, четвертой и т.д. степени (см. табл. 4.3).

Таблица 4.1

Исходная редуцированная корреляционная матрица

Признак	R_h				
	X_1	X_2	X_3	$S_i^{(1)} = \sum_j r_{ij}$	$a^{(1)} = \frac{S_i^{(1)}}{S_{\max}}$
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 4.2

Первый цикл итерации — возведение в квадрат корреляционной матрицы

Признак	$R_h^2 = R_h' R_h$			$S^{(2)} = \sum_j r_{ij}$	$p_i^{(2)} = R_h S^{(1)}$	$a_i^{(2)} = p_i / p_{\max}$	$d = \lvert a^{(2)} - a^{(1)} \rvert$
	X_1	X_2	X_3				
X_1	0,811	0,823	0423	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 4.3

Второй цикл итерации — корреляционная матрица в четвертой степени

Признак	$R_h^4 = R_h^2 R_h^2$			$S_i^{(3)}$	$p_i^{(3)} = R_h^2 S^{(2)}$	$a^{(3)} = p_i / p_{\max}$	$d = \lvert a^{(3)} - a^{(2)} \rvert$
	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	2,637	0,585	0,008

После возведения корреляционной матрицы в четвертую степень α-разности резко уменьшились Вычислим R_h^8 и завершим итерацию.

Таблица 4.4

Третий цикл итерации — корреляционная матрица в восьмой степени

Признак	$R_h^8 = R_h^4 R_h^4$			$S_i^{(4)}$	$p_i^{(4)} = R_h^4 S^{(3)}$	$a^{(4)} = p_i / p_{\max}$	$d = \lvert a^{(4)} - a^{(3)} \rvert$
	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 итерации не превы-

сило пяти тысячных. Таким образом, оценки компонент первого собственного вектора можно считать достоверными. Собственный вектор — это ненормированные значения $a^{(4)}$, т.е. $a_1^{(4)} = U_1$.

Перейдем к определению нагрузок первого главного фактора (табл. 4.5).

Таблица 4.5

Вычисление нагрузок первого главного фактора

Признак	$a_1^{(4)} = U_1^*$	$\beta_1 = R_h a_1^{(4)}$	$A = \frac{U_1 \sqrt{\lambda_1}}{\left(\sum U_{1i}^2\right)^{1/2}}$
X_1	0,906	1,278	0,732
X_2	1,000	1,412	0,808
X_3	0,583	0,823	0,471

* Данные представлены в табл. 4.4

В табл. 4.5 собственное число λ_1 — это наибольшая величина вектора β_1. Компоненты a_{j1} легко рассчитываются по известному λ_1=1,412

$$\frac{U_1 \sqrt{\lambda_1}}{\left(\sum U_{1i}^2\right)^{1/2}} = \frac{\sqrt{1,412}}{(0,821 + 1,000 + 0,340)^{1/2}} = 0,808.$$

Итогами первой итерации будут первое собственное число λ_1=1,412 и вектор факторных нагрузок A'_1=(0,732, 0,808, 0,471). Проверим выполнение требования

$$\sum a_{j1}^2 = \lambda_1, \text{ или } 0,732^2 + 0,808^2 + 0,471^2 = 1,412.$$

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 только по одному, первому, вектору факторного отображения будет $R^+_h = AA^T$

$$R_h^+ = \begin{pmatrix} 0,732 \\ 0,808 \\ 0,471 \end{pmatrix} \times (0,732 \quad 0,808 \quad 0,471) = \begin{pmatrix} 0,536 & 0,591 & 0,345 \\ 0,591 & 0,653 & 0,381 \\ 0,345 & 0,381 & 0,222 \end{pmatrix}.$$

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. 4.6 получены более грубые оценки элементов собственного вектора, чем в первом случае (см. табл. 4.5). Тем не менее, итерация прервана с учетом того, 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 4.6

Матрица первых остаточных коэффициентов корреляции

Признак	R_1			$S^{(1)}$	$a^{(1)}$
	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

По данным табл. 4.7 определим нагрузки второго главного фактора (табл. 4.8)

Таблица 4.7

Первый и второй циклы итерации для матрицы первых остаточных коэффициентов корреляции

Признак	R_1			$S^{(3)}$	$p^{(3)}$	$a^{(3)}$	$d = \lvert a^{(3)} - a^{(2)} \rvert$
	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	0,00061	0,00061	0,170	0,027
X_3	–0,00870	0,00179	0,01051	0,00360	0,00358	1,000	0,000

Таблица 4.8

Вычисление нагрузок второго главного фактора

Признак	$a_2^{(3)} = U_2$	$\beta_2 = R_1 a_2^{(3)}$	$A_2 = \dfrac{U_2\sqrt{\lambda_{22}}}{\left(\sum U_{2i}^2\right)^{1/2}}$
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 $R_1 - A_2 \times A_2^T$.

Элементы матрицы R_2 имеют малые значения, и выделение третьего главного фактора не целесообразно. Такой же вывод следует и по данным собственных чисел λ, а именно: при помощи первого и второго главных факторов полностью воспроизведена общность ($trR_h = 1{,}720$) и почти на 60 % удалось объяснить вариацию элементарных признаков X_1, X_2, X_3 $\left(\dfrac{1{,}412 + 0{,}366}{3}\right)$

$$R_2 = \begin{pmatrix} 0{,}134 & -0{,}010 & -0{,}191 \\ -0{,}010 & -0{,}023 & 0{,}058 \\ -0{,}191 & 0{,}058 & 0{,}198 \end{pmatrix} - \begin{pmatrix} -0{,}383 \\ 0{,}079 \\ 0{,}462 \end{pmatrix} \cdot (-0{,}383 \quad 0{,}079 \quad 0{,}462) =$$

$$= \begin{pmatrix} -0{,}013 & 0{,}020 & -0{,}014 \\ 0{,}020 & -0{,}029 & 0{,}022 \\ 0{,}014 & 0{,}022 & -0{,}015 \end{pmatrix}.$$

В заключение обобщим итоги решения задачи методом главных факторов, выделяя общности и характерности для каждого из элементарных X_j признаков (табл. 4.9).

Таблица 4.9

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. 4.9 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

$$\sum h_j^2 + \sum d_j^2 = 3.$$

В дальнейшем может быть найдена матрица значений главных факторов

$$F = (A^T A)^{-1} A^T Z'.$$

4.3. Контрольные задания

Задание 1. По двенадцати промышленным предприятиям имеются данные об уровне заработной платы — X_1 (дол. США) и размеру получаемой за месяц прибыли в расчете на одного работника — X_2 (дол. США) (табл. 4.10).

Таблица 4.10

Предприятие	X_1	X_2	Предприятие	X_1	X_2
1	210	250	7	90	130
2	240	310	8	240	510
3	115	265	9	160	280
4	220	405	10	120	150
5	125	270	11	75	120
6	130	225	12	185	315

Представьте результаты выборочного обследования предприятий в двумерной системе координат, покажите предположительно векторы, которые обобщали бы распределение значений признаков (X_1, X_2).

Задание 2. По данным экспертных опросов покупателей натуральных соков построена матрица корреляционных связей характеристик товарной продукции[1]

$$R = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline X_1 & 1 & 0{,}712 & 0{,}932 & 0{,}913 \\ X_2 & 0{,}712 & 1 & 0{,}749 & 0{,}787, \\ X_3 & 0{,}932 & 0{,}749 & 1 & 0{,}851 \\ X_4 & 0{,}913 & 0{,}787 & 0{,}851 & 1 \end{array}$$

где X_1 — стоимость; X_2 — дизайн упаковки; X_3 — калорийность; X_4 — сохраняемость продукта.

С помощью рекуррентных соотношений Фаддеева постройте характеристический многочлен.

При условии, что известны значения λ_j ($\lambda_1 = 3{,}477$, $\lambda_2 = 0{,}336$, $\lambda_3 = 0{,}145$, $\lambda_4 = 0{,}042$), произведите расчет матрицы собственных векторов (U, V) и матрицы факторных нагрузок (A). Используйте алгоритм метода главных компонент.

Задание 3. На основе нижеприведенной матрицы факторных нагрузок (A), построенной с помощью метода главных компонент, определите вклад и долю вариации каждой главной компоненты в общую дисперсию элементарных признаков.

Отберите оптимальное число компонент для анализа и попытайтесь дать название первым двум главным компонентам (F_1 и F_2) при условии, что изучаемые элементарные признаки следующие: X_1 — трудоемкость производства продукции, X_2 — рентабельность производства, X_3 — уровень фондоотдачи основных средств, X_4 — оборачиваемость оборотных средств, X_5 — удельный вес брака в выпускаемой продукции.

[1] Оценки экспертов строились по 10-балльной системе (1÷10), максимальный уровень оценки – 10 баллов.

$$A = \begin{pmatrix} -0{,}874 & -0{,}223 & 0{,}432 & 0{,}007 & -0{,}014 \\ 0{,}955 & 0{,}134 & 0{,}181 & -0{,}125 & 0{,}148 \\ 0{,}961 & -0{,}108 & 0{,}094 & 0{,}234 & -0{,}030 \\ 0{,}927 & 0{,}294 & 0{,}154 & -0{,}051 & -0{,}168 \\ -0{,}664 & 0{,}740 & 0{,}051 & 0{,}079 & -0{,}040 \end{pmatrix}.$$

Задание 4. По пяти странам СНГ за 1999 г. имеются следующие данные об объеме выполненных НИОКР (в % к ВВП) — X_1 и численности работников, осуществляющих проведение полученных исследований и разработок (тыс. чел.) — X_2 (табл. 4.11).

Таблица 4.11

Страна	(X_1)	(X_2)
Беларусь	1,1	21,3
Молдова	0,6	9,0
Россия	1,2	493,0
Казахстан	0,2	10,8
Туркменистан	0,1	2,2
Украина	1,2	126,0

По приведенным данным произведите расчеты, реализующие полный алгоритм метода главных компонент и поясните полученные результаты.

Задание 5. В результате анализа данных о доходах и потребительском поведении населения, проведенного с помощью метода главных компонент, получена следующая матрица факторных нагрузок (A) (табл. 4.12).

Таблица 4.12

Факторные нагрузки переменных X_j	F_1	F_2
Свободное время, уделяемое семье, детям, хозяйству	0,097	–0,480
Денежные доходы населения	0,670	0,248
Расходы на отдых, туризм	0,222	–0,359
Потребление социальных (дешевых) товаров первой необходимости	–0,309	0,721
Расходы на обучение, повышение квалификации	0,463	0,103
Затраты времени на работу (бизнес)	0,594	–0,095
Потребление платных услуг	0,644	0,467
Состав семьи (включая число детей)	–0,058	0,002
Расходы на питание в ресторанах, кафе	0,030	–0,097
Удельный вес расходов населения на продукты питания	–0,702	0,151

Оцените информативность выделенных главных компонент, а также их структуру с точки зрения поиска названий.

В ходе анализа структуры компонент рассчитайте коэффициенты информативности для подмножеств значимых признаков (X_j). Определите названия главных компонент.

Задание 6. Осуществите переход всеми известными способами от обычной (R) к редуцированной (R_h) матрице парных корреляций, если известно, что

$$R = \begin{pmatrix} 1 & 0{,}287 & 0{,}512 & 0{,}025 & 0{,}009 & 0{,}207 \\ 0{,}287 & 1 & 0{,}214 & 0{,}220 & 0{,}094 & 0{,}541 \\ 0{,}512 & 0{,}214 & 1 & 0{,}079 & -0{,}036 & 0{,}192 \\ 0{,}025 & 0{,}220 & 0{,}079 & 1 & 0{,}297 & 0{,}173 \\ 0{,}009 & 0{,}094 & -0{,}036 & 0{,}297 & 1 & 0{,}091 \\ 0{,}207 & 0{,}541 & 0{,}192 & 0{,}173 & 0{,}0911 & 1 \end{pmatrix}.$$

Задание 7. Имеется матрица парных корреляций

$$R = \begin{pmatrix} 1 & 0{,}910 & 0{,}906 \\ 0{,}910 & 1 & 0{,}878 \\ 0{,}906 & 0{,}878 & 1 \end{pmatrix}.$$

По приведенным данным постройте матрицу стандартизированных значений собственных векторов (V), используйте два способа реализации метода главных факторов:

а) рекурентные соотношения Фаддеева;

б) процедуру многократного возведения (до 8-й степени) в квадрат исходной матрицы парных корреляций (подход Хоттелинга).

Редактированную корреляционную матрицу постройте простым способом «наибольшей корреляции» (с учетом поправок Барта).

Задание 8. По известной матрице факторных нагрузок (A) постройте воспроизведенную матрицу парных корреляций (R^+) и определите, насколько хорошо она описывает исходную матрицу R, т.е. найдите матрицу корреляционных остатков ($R-R^+$)

$$A = \begin{pmatrix} 1 & 0{,}910 & 0{,}906 \\ 0{,}910 & 1 & 0{,}878 \\ 0{,}906 & 0{,}878 & 1 \end{pmatrix} \quad R = \begin{pmatrix} 1 & 0{,}910 & 0{,}906 \\ 0{,}910 & 1 & 0{,}878 \\ 0{,}906 & 0{,}878 & 1 \end{pmatrix}.$$

Задание 9. Используя метод главных факторов по приведенным в табл. 4.13 динамическим данным, обобщите значения набора признаков, количественно отражающих экономический рост Республики Беларусь, в одном латентном признаке F_2 — «экономический рост». Покажите тренд значений латентного признака на графике, сделайте выводы.

Таблица 4.13

	1995 г.	1996 г.	1997 г.	1998 г.	1999 г.	2000 г.	2001 г.
Валовой внутренний продукт	89,6	102,8	111,4	108,4	103,4	105,8	104,1
Продукция промышленности	88,3	103,5	118,8	112,4	110,3	107,8	105,4
Продукция сельского хозяйства	95,3	102,4	95,1	99,3	91,7	109,3	101,8

	1995 г.	1996 г.	1997 г.	1998 г.	1999 г.	2000 г.	2001 г.
Реальные денежные доходы на душу населения	74	118	106	120	99	120	125
Розничный товарооборот	77,2	130,5	117,9	126,1	110,7	111,8	121,2
Транспортные перевозки грузов	74,7	80,6	105,8	105,6	95,6	89,5	95,2

При решении задачи используйте пакет прикладных программ.

Задание 10. Заполните табл. 4.14 недостающими данными.

Таблица 4.14

Компоненты дисперсии	Условное обозначение	Вариант задачи					
		I	II	III	IV	V	VI
Общность	h^2	0,65	–	–	0,80	–	–
Характерность	d^2	–	0,35	–	–	–	0,15
Специфичность	b^2	–	–	0,35	–	0,25	–
Дисперсия ошибки (ненадежность)	l^2	0,20	–	0,15	–	–	0,10
Надежность	c^2	–	0,70	–	0,90	0,80	–

Задание 11. Известна матрица факторных нагрузок (A), полученная в ходе реализации метода главных факторов

$$A = \begin{pmatrix} F_1 & F_2 \\ 0{,}92330 & -0{,}29100 \\ 0{,}65430 & -0{,}14500 \\ 0{,}91224 & -0{,}26720 \\ 0{,}39726 & 0{,}88415 \\ 0{,}28930 & 0{,}92760 \end{pmatrix}.$$

Определите:

1) собственные числа λ_1 и λ_2; уровень информативности главных факторов;

2) дисперсионные показатели общности и характерности для каждого из элементарных признаков (X_j);

3) воспроизведенную матрицу парных корреляций (R^+).

Задание 12. По пяти странам СНГ за 1997 г. известны данные индекса человеческого развития (X_1) и интегрального показателя результативности экономических реформ (X_2). Реализуя алгоритм метода главных факторов, включающий процедуру квадрирования матрицы парных корреляций, выделите первый главный фактор (F_1), обобщающий элементарные признаки X_1 и X_2. Определите структуру и название F_1, а также уровень его информативности и характеристики общности.

Страна	X_1	X_2
Беларусь	0,763	0,370
Казахстан	0,740	0,510
Молдова	0,702	0,660
Россия	0,747	0,720
Украина	0,721	0,520

Замечание: так как элементарные признаки X_1 и X_2 имеют единый масштаб измерения, то стандартизацию их значений производить не обязательно, а расчеты можно строить, используя ковариационную матрицу (S).

Задание 13. Для улучшения структуры главных факторов имеются две матрицы — факторных нагрузок (A) и вращения (T):

$$A = \begin{pmatrix} 0{,}854 & -0{,}256 \\ 0{,}596 & -0{,}310 \\ 0{,}340 & 0{,}490 \\ -0{,}420 & -0{,}278 \\ 0{,}375 & 0{,}018 \end{pmatrix}; \quad T = \begin{pmatrix} 0{,}574 & 0{,}819 \\ -0{,}819 & 0{,}574 \end{pmatrix}.$$

1. Ответьте на вопрос: T-матрица косоугольного, или ортогонального вращения. Какое вращение она задает: по часовой стрелке или против?
2. Выполните 1-й и 2-й шаги вращения.
3. Сделайте выводы о качественных изменениях структуры главных факторов после каждого шага вращения.

Задание 14. Имеется матрица (F) значений двух главных компонент по десяти предприятиям сельского хозяйства.

	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}
F_1	−2,901	0,701	0,321	1,521	−0,202	0,640	−0,056	0,105	−3,140	0,879
F_2	−1,560	0,280	3,610	1,104	0,756	−2,235	1,670	−4,201	−0,015	0,654

При условии, что F_1 — выход продукции растениеводства на единицу сельскохозяйственных угодий, а F_2 — продуктивность отрасли животноводства, покажите графически распределение предприятий в пространстве двух главных компонент, выделите однородные группы предприятий, используя иерархический кластерный анализ (один алгоритмический шаг), сделайте обобщающие выводы.

Задание 15. В результате анализа деятельности шести продовольственных магазинов, проведенного методом главных факторов, получена матрица значений главных факторов (F)

	n_1	n_2	n_3	n_4	n_5	n_6
F_1	−1,214	−0,483	0,571	1,664	−0,635	0,097
F_2	−1,250	−0,270	0,180	−0,884	0,947	1,277

,

где F_1 — оснащенность магазина основными средствами; F_2 — уровень квалификации работников.

Постройте линейную регрессионную модель, отражающую влияние F_1, F_2 на уровень доходности торговой деятельности (Y) — поступление прибыли в расчете на одного работника за месяц, дол. США, если

$$Y' = (0{,}120 \;\; 0{,}150 \;\; 0{,}258 \;\; 0{,}364 \;\; 0{,}420 \;\; 1{,}500).$$

Примечание. Регрессионная модель вида $\hat{y} = a_1F_1 + a_2F_2$, поиски параметров регрессионной модели осуществляются решением известного матричного уравнения $A = (F^TF)^{-1}F^TY$.

Задание 16. По следующим данным оцените достаточность для значимых выводов числа главных компонент, выделенных в ходе анализа, если

$$R = \begin{pmatrix} 1 & 0{,}012 & 0{,}197 & -0{,}110 & 0{,}436 \\ 0{,}012 & 1 & 0{,}488 & 0{,}246 & 0{,}223 \\ 0{,}197 & 0{,}488 & 1 & 0{,}812 & 0{,}803 \\ -0{,}110 & 0{,}246 & 0{,}812 & 1 & 0{,}785 \\ 0{,}436 & 0{,}223 & 0{,}803 & 0{,}785 & 1 \end{pmatrix};$$

число наблюдений n = 46; собственные числа: $\lambda_1 = 2{,}822$; $\lambda_2 = 1{,}136$; $\lambda_3 = 0{,}863$; $\lambda_4 = 0{,}131$; $\lambda_5 = 0{,}048$; выделено две главных компоненты (F_1, F_2).

Задание 17. Методами факторного анализа предусматривается обработка массива исходных данных, размерностью 6×50, имеющих матрицу парных корреляций (R), для которой: $\lambda_1 = 1{,}477$; $\lambda_2 = 1{,}085$; $\lambda_3 = 1{,}033$; $\lambda_4 = 0{,}943$; $\lambda_5 = 0{,}838$; $\lambda_6 = 0{,}655$.

По приведенным данным, при $\alpha = 0{,}05$, оцените значимость матрицы R и справедливость принятого решения о применении методов ФА.

Задание 18. В результате приложения в анализе метода главных факторов получены следующие результаты:

- значения собственных чисел: $\lambda_1 = 4{,}294$; $\lambda_2 = 1{,}027$; $\lambda_3 = 0{,}435$; $\lambda_4 = 0{,}163$; $\lambda_5 = 0{,}060$; $\lambda_6 = 0{,}021$;
- факторные нагрузки двух первых выделенных главных факторов

$A =$

	n_1	n_2	n_3	n_4	n_5	n_6
F_1	−0,816	−0,837	−0,600	0,885	0,935	0,954
F_2	−0,268	0,340	0,671	0,260	0,320	0,125

Справочно: число наблюдений $n = 36$.

Используя критерий Лоули, оцените достаточность числа оставленных в анализе главных факторов, при $\alpha = 0{,}025$.

5. МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ

5.1. Методические рекомендации

5.1.1. Сущность методов и алгоритм многомерного шкалирования

Термин «многомерное шкалирование» (МШ) является синонимом понятия «ординация», т.е. упорядочения. Имеется в виду упорядочение наблюдаемых объектов в m-мерном признаковом (стимульном) пространстве. Так же, как и в факторном анализе, в МШ появляются латентные признаки, однако они определяются не на основе интегрирования элементарных признаков, а исходя из различий наблюдаемых объектов и их распределения в некотором пространстве.

Методы многомерного шкалирования в статистических исследованиях используются для решения следующих задач:

- сжатие признакового пространства;
- визуализация расположения наблюдаемых объектов относительно друг друга в теоретическом пространстве;
- выявление латентных факторов, предопределяющих пространственное расположение, различие наблюдаемых объектов.

С помощью МШ открывается возможность моделирования сложных явлений, процессов и построения прогнозов их развития.

Всю совокупность методов МШ подразделяют на два больших класса: *метрические* и *неметрические*.

Метрические методы используются для обработки количественных данных.

Неметрические методы применяют, когда исходными являются неколичественные данные (порядковые, ранговые и т.п.).

В зависимости от объекта исследования многомерное шкалирование различают по направлениям:

- *анализ стимулов* — изучение некоторой совокупности объектов и моделирование их пространственного расположения в соответствии с определенными признаковыми различиями;
- *анализ индивидуальных различий* — изучение субъективного восприятия наблюдаемых объектов и их различий;
- *анализ предпочтений* — изучение совокупности объектов с учетом существования некоторых идеальных объектов, другими словами, существование представлений об идеальных объектах;
- *анализ идеальных точек* — поиск, формальное описание, пространственное отображение идеального положения изучаемых объектов (стимулов).

5.1.2. Представление и первичная обработка данных

В основе построения алгоритмов МШ лежат два типа формальных моделей:

- *дистанционные*, в большинстве случаев базируются на евклидовой метрике, при этом различия самих наблюдаемых объектов описываются расстояниями (d_{ij}) в теоретическом шкальном пространстве

$$\hat{d}_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2};$$

- *векторные*, с их помощью различия объемов аппроксимируются скалярными произведениями векторов, соединяющих начало координат с точками пространственного расположения стимулов

$$\hat{b}_{ij} = \sum_k x_{ik} x_{jk}.$$

При записи моделей МШ используются следующие обозначения:

x_{ik}, x_{jk} — значения k-го признака, наблюдаемые у i-го и j-го объектов (стимулов); $\hat{d}_{ij}$, $\hat{b}_{ij}$ — статистически оцененные меры различия i-го и j-го стимулов.

В практике исследований чаще используются дистанционные модели. Общий алгоритм, базирующийся на подходе Торгерсона, в основе которого лежит дистанционная модель методов МШ, включает шаги:

1) систематизация данных статистического наблюдения, экспертных оценок и представление результатов в виде матрицы различий симметрического вида (Δ). На главной диагонали этой матрицы расположены нули (меры различия одних и тех же стимулов), отражающие полное сходство;

2) переход от матрицы различий к матрице с двойным центрированием (Δ^*), которая в последующем позволяет выявлять латентные признаки;

3) определение латентных признаков (X_r) с помощью метода главных компонент или какого-либо из методов факторного анализа;

4) интерпретация аналитических результатов, при необходимости их визуальное (графическое) представление.

Построение матрицы различий (*первый шаг алгоритма*) предполагает формирование матрицы исходных данных. Эту матрицу получают, используя оценки экспертов или регистрируя непосредственно признаковые значения исследуемых явлений и процессов в ходе статистического наблюдения.

Экспертные оценки обычно систематизируют в матрице условных вероятностей (матрице идентификаций) или матрице совместных вероятностей.

Матрица условных вероятностей представляет относительные данные по узнаванию стимулов. Например, экспертам для узнавания предъявляется стимул A; семьдесят процентов его узнают, двадцать пять процентов — принимают за стимул В, пять процентов за стимул C и т.д. (рис. 5.1, *а*).

Обычно матрица условных вероятностей несимметрическая, поэтому предусматривается использование определенных методов по приведению ее к симметрическому виду. Простейший метод — когда на главной диагонали проставляются нули, а элементы, равноудаленные от главной диагонали (снизу и сверху от нее), находят как полусумму исходных значений (рис. 5.1, *б*).

Матрица совместных вероятностей отражает взаимодействие стимулов (i, j). Она содержит согласованные данные и всегда симметрическая.

Значения признаков исследуемых явлений, процессов первоначально представляются в виде матрице, которую называют матрицей мер различия профилей.

а)

	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

$\frac{P_{ij}+P_{ji}}{2}$;

б)

	А	В	С	Д
А	0	0,28	0,08	0,01
В	0,28	0	0,28	0,04
С	0,08	0,28	0	0,15
Д	0,01	0,04	0,15	0

Рис. 5.1. Матрица условных вероятностей и полученная из нее матрица различий

Переход от нее к матрице различий предусматривает предварительное нормирование данных и исчисление мер признаковых различий стимулов, с использованием какой-либо из метрических формул.

Нормирование исходных значений признаков обычно выполняется с помощью одного из следующих приемов:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \qquad z_{ij} = x_{ij} / \bar{x}_j, \qquad z_{ij} = \log x_{ij}.$$

Меры различия находят по метрическим формулам. Среди них наиболее распространены следующие:

$$\delta_{ij}^{E} = \left(\sum_k \left(x_{ik} - x_{jk} \right)^2 \right)^{1/2}$$ евклидова метрика;

$$\delta_{ij}^{E} = \sum_k \left(x_{ik} - x_{jk} \right)^2$$ квадрат евклидова расстояния;

$$\delta_{ij}^{M} = \left(\sum_k \left(x_{ik} - x_{jk} \right)^p \right)^{1/p}$$ метрика Минковского;

$$\delta_{ij}^{L} = \sum_k \left| x_{ij} - x_{jk} \right|$$ метрика city-block.

Меры различия обобщаются в матрице различий (Δ) симметрического вида.

Второй шаг алгоритма — переход от матрицы различий (Δ) к матрице с двойным центрированием (Δ^*) — осуществляется по формуле

$$\delta_{ij}^{*} = \frac{1}{2}\left(\delta_{ij}^2 - \delta_{i}^2 - \delta_{j}^2 + \delta_{..}^2 \right),$$

где δ_i^2 — средняя для характеристик различий в j-столбцах i-й строки, возведенных в квадрат $\delta_i^2 = \frac{1}{J}\sum_j \delta_{ij}^2$; δ_j^2 — средняя для характеристик различий в i-строках j-го столбца, возведенных в квадрат $\delta_j^2 = \frac{1}{I}\sum_i \delta_{ij}^2$; $\delta_{..}^2$ — средняя величина для квадратов характеристик различий всей матрицы различий (Δ) $\delta_{..}^2 = \frac{1}{IJ}\sum_i \sum_j \delta_{ij}^2$.

Правильность построения матрицы с двойным центрированием легко проверяется: суммы ее элементов, полученные по любой строке или столбцу, должны быть равны нулю.

По Торгерсону для матрицы с двойным центрированием существует равенство

$$\Delta^* = X^T X, \qquad (5.1)$$

где X — матрица значений обобщенных (латентных) признаков[1]. Важно учитывать, что их определенность обусловливается не признаковым составом (x_j) как в факторном анализе, а составом стимулов, обычно — наблюдаемых объектов.

На третьем шаге алгоритма, исходя из равенства (5.1), находят сами латентные признаки (X). С этой целью используют методы главных компонент или факторного анализа (главных факторов, центроидный, максимального правдоподобия и т.д., см. гл. 4).

На завершающем, *четвертом шаге алгоритма* МШ, производится интерпретация полученных аналитических результатов и их визуальное представление. При объяснении выходных данных МШ исходят из того, что название латентных признаков формируется структурой наблюдаемых объектов (стимулов h_i), а не признаков (X_j), как в факторном анализе.

Графическое изображение стимульного пространства, с погружением в него стимулов строится на основе значений одного — трех латентных признаков (X_j), как правило, первых, имеющих наибольшую информативную нагрузку.

В отличие от обработки количественных данных методами МШ, алгоритмы обработки неколичественных данных имеют дополнительные шаги, они сводятся к выполнению следующих операций:

- оцифровка неколичественных данных;
- получение стартовой конфигурации стимулов;
- стандартизация текущих координатных оценок;
- вычисление различий стимулов по теоретическим данным;
- поиск улучшенных оценок координат (с использованием формулы Лингоса — Роскама);
- оценка степени улучшения значений координат стимулов. Если улучшение мало — алгоритм завершается, если улучшение существенно — алгоритм возобновляется, начиная с шага «стандартизировать оценки координат стимулов».

Кроме анализа стимулов в МШ нередко решаются задачи анализа индивидуальных различий, предпочтений и идеальных точек.

В анализе индивидуальных различий одновременно обрабатывается несколько матриц различий (Δ_s), по числу экспертов, оценивающих стимулы. В ходе анализа устанавливается не только пространственное расположение стимулов, но и весовые значения экспертов, которые придаются каждой из k-шкал стимульного пространства. Расчеты при этом производятся с использованием

[1] В многомерном шкалировании латентные признаки принято обозначать так же, как и элементарные признаки в традиционной статистике (X).

модифицированных дистанционных моделей. В большинстве случаев это взвешенная евклидова модель или трехмодельная модель Такера.

Анализ предпочтений и анализ идеальных точек в определенной мере связаны. В первом случае оценивается удаленность от представляемого «идеала», во втором — производится оценка координат идеального стимула.

В перечисленных трех направлениях МШ исследовательские цели достигаются применением дистанционной или векторной модели. В ходе анализа определяются координаты стимулов (в т.ч. и идеальных точек), а так же предпочтения экспертов, их суждения об «идеале».

Более подробно о неметрическом шкалировании, анализе индивидуальных различий, предпочтений и идеальных точек [12].

5.2. Примеры решения типовых задач

В табл. 5.1 приведены статистические показатели по трем отраслям экономики Республики Беларусь за 2001 г.

Таблица 5.1

Исходные данные

Отрасль	Объем продукции, млрд р.	Численность работающих, тыс. чел.	Рентабельность продукции, %	Среднемесячный уровень заработной платы, тыс. р.	Кредиторская задолженность на конец года, в расчет на один рубль валовой продукции, р.
Промышленность	14650	1223	10,8	148,4	0,264
Сельское хозяйство	4364,3	606	0,2	76,5	0,234
Капитальное строительство	1400	306	8,2	157,3	0,274
В среднем по трем отраслям	6804,8	711,7	6,4	127,4	0,256

Требуется определить признаки, отражающие различие трех отраслей и показать их пространственное расположение.

В ходе анализа будем использовать метод нормирования исходных данных $z_{ij} = x/\bar{x}$ и евклидову метрику для оценки различий отраслей.

Решение. 1. Нормируем исходные данные $z_{11} = 2{,}15$; $z_{12} = 1{,}72$; $z_{13} = 1{,}69$ и т.д.

Матрица нормированных значений стимулов принимает следующий вид:

$$Z = \begin{pmatrix} 2{,}15 & 1{,}72 & 1{,}69 & 1{,}16 & 1{,}03 \\ 0{,}64 & 0{,}85 & 0{,}03 & 0{,}60 & 0{,}91 \\ 0{,}20 & 0{,}43 & 1{,}28 & 1{,}23 & 1{,}07 \end{pmatrix}.$$

Используя евклидову метрику, определим различия отраслей:

$$\delta_{1,2} = \left(\sum_k (x_{ik} - x_{jk})^2\right)^{\frac{1}{2}} = \left((2{,}15 - 0{,}64)^2 + \ldots + (1{,}03 - 0{,}91)^2\right)^{\frac{1}{2}} = 2{,}74;$$

$$\delta_{1,3} = \left((2{,}15 - 0{,}20)^2 + (1{,}72 - 0{,}43)^2 + \ldots + (1{,}03 - 1{,}07)^2\right)^{\frac{1}{2}} = 2{,}38;$$

$$\delta_{2,3} = \left((0{,}64 - 0{,}20)^2 + (0{,}85 - 0{,}43)^2 + \ldots + (0{,}91 - 1{,}07)^2\right)^{\frac{1}{2}} = 1{,}54.$$

Обобщим метрические данные в исходной матрице различий (Δ)

$$\Delta = \begin{pmatrix} 0 & 2{,}47 & 2{,}38 \\ 2{,}47 & 0 & 1{,}54 \\ 2{,}38 & 1{,}54 & 0 \end{pmatrix}.$$

Для получения координат стимулов в теоретическом шкальном пространстве необходимо от матрицы различий (Δ) перейти к матрице с двойным центрированием (Δ^*). Элементы матрицы Δ^* исчислим по известной формуле

$$\delta_{ij}^* = -\frac{1}{2}\left(\delta_{ij}^2 - \delta_{i\cdot}^2 - \delta_{j}^2 + \delta_{\cdot\cdot}^2\right)$$

при этом

$$\delta_{1,1}^* = -\frac{1}{2}(0 - 3{,}922 - 3{,}922 + 3{,}142) = 2{,}351;$$

$$\delta_{1,2}^* = -\frac{1}{2}(6{,}101 - 3{,}922 - 2{,}824 + 3{,}142) = -1{,}248 \text{ и т.д.}$$

С учетом симметричности матрицы с двойным центрированием (Δ^*) общий объем вычислений можно сократить, в результате получаем

$$\Delta^* = \begin{pmatrix} 2{,}351 & -1{,}248 & -1{,}103 \\ -1{,}248 & 1{,}253 & -0{,}005 \\ -1{,}103 & -0{,}005 & 1{,}108 \end{pmatrix}.$$

Правильность построения матрицы Δ^* подтверждается при суммировании ее элементов по каждой строке и каждому столбцу, все суммы равны нулю.

2. Определим координаты стимулов (отраслей экономики). Эта задача решается, исходя из равенства $\Delta^* = X^T X$, методом главных компонент или каким-либо из методов факторного анализа (см. параграф 4.1).

Методом главных компонент получаем матрицу

$$X = \begin{pmatrix} -0{,}934 & -0{,}119 & -0{,}354 \\ 0{,}811 & 0{,}557 & 0{,}193 \\ 0{,}806 & -0{,}418 & 0{,}437 \end{pmatrix}.$$

Каждый элемент этой матрицы x_{ij} — значение j-й главной компоненты для i-й отрасли.

В приведенной матрице (X) значения главных компонент по отраслям нарушают логику распределения «плохих» — «хороших» отраслей. Так, промышленность, которая по своим экономическим параметрам выгодно отличается от двух других отраслей, оказывается в координатной зоне с отрицательными значениями обобщенных признаков. В то же время сельское хозяйство и капитальное строительство получили положительные значения обобщенных признаков. Восстановить логику распределения наблюдаемых объектов можно вращением теоретического пространства шкал. Одним из вариантов такого вращения является простое изменение знаков значений главных компонент на противоположный (поворот каждой из координатных осей на 180°). После вращения получим новую матрицу координат стимулов

$$X=\begin{pmatrix} 0{,}934 & 0{,}119 & 0{,}354 \\ -0{,}811 & -0{,}557 & -0{,}193 \\ -0{,}806 & 0{,}418 & -0{,}437 \end{pmatrix}.$$

Значимость первых двух главных компонент объясняет более 85 % колеблемости анализируемых признаков. Оставляя эти компоненты в анализе, покажем размещение стимулов в двумерном шкальном пространстве ($R^{X1,X2}$): $n_1 = (0{,}934;\ 0{,}119)$; $n_2 = (-0{,}811;\ -0{,}557)$; $n_3 = (-0{,}806;\ 0{,}418)$.

Распределение стимулов, показанное на рис. 5.1, позволяет следующим образом идентифицировать шкалы: X_1 — состояние производственного потенциала (масштабы производства, эффективность ресурсопотребления); X_2 — трудоемкость производства продукции.

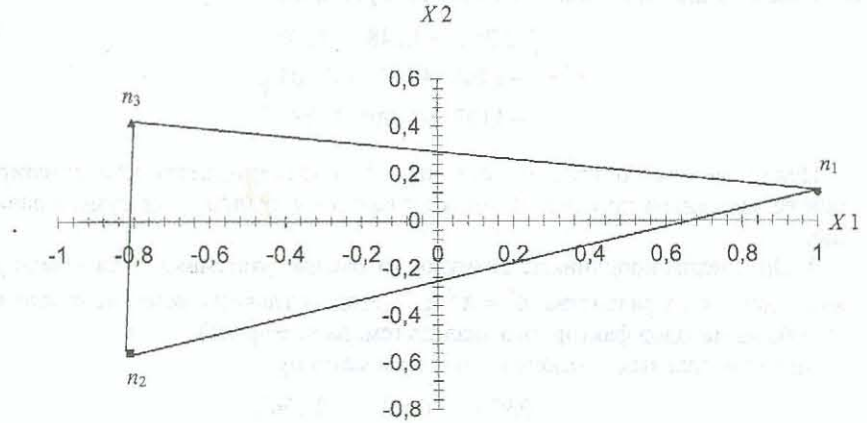


Рис. 5.1. Конфигурация отраслей в пространстве главных компонент

5.3. Контрольные задания

Задание 1. Экспертами оценены различия трех марок автомобилей: «Волга» (n_1), «Фольксваген» (n_2) и «Тойота» (n_3) по двум обобщенным признакам: экономичность (X_1) и надежность (X_2). По 10-балльной шкале оценок получены следующие усредненные данные:

	X_1	X_2
n_1	0,35	0,20
n_2	0,72	0,65
n_3	0,80	0,75

Исходя из оценок экспертов, используя евклидову метрику, определите расстояние между различными марками автомобилей.

Покажите пространственное расположение автомобилей трех марок и сделайте выводы об их различии.

Задание 2. Предположим, что в оценке автомобилей (задание 1) участвовали два эксперта — Александр и Николай. По предоставленной ими информации построены матрицы с данными различий автомобилей.

Александр

	X_1	X_2
n_{11}	0,20	0,10
n_{21}	0,60	0,80
n_{31}	0,82	0,75

Николай

	X_1	X_2
n_{12}	0,40	0,45
n_{22}	0,70	0,75
n_{32}	0,75	0,60

Основываясь на экспертных оценках, покажите распределение автомобилей в двухшкальном пространстве анализируемых признаков (X_1, X_2), с идентификацией каждого из экспертов.

Используя метрику city-block, определите различия мнений экспертов относительно автомобилей сначала по каждому из признаков (X_1, X_2), а затем одновременно по значениям двух признаков.

Задание 3. Группе экспертов предложено ранжировать по степени действенности направления экономической политики, обусловливающие выход из кризиса. Для оценки предъявлены четыре стимула:

- сдерживание инфляции, S;
- повышение инвестиционной активности, K;
- наращивание объемов производства, Q;
- повышение доходов и покупательной способности населения, D.

При предъявлении первого стимула (S) голоса экспертов распределились следующим образом: S — 35 %, K — 25 %, Q — 25 % и D — 15 %.

Для второго стимула (K) получены результаты: K — 50 %, S — 30 %, Q — 15 %, D — 5 %.

Для третьего стимула (Q) имеем: Q —25 %, S — 30 %, K — 25 %, D — 20 %.

Наконец, для четвертого стимула (D) получено: D — 30 %, S — 30 %, K — 25 %, Q — 15 %.

На основе данных опроса экспертов (общее число голосов 100 %), постройте матрицу условных вероятностей и осуществите от нее переход к матрице различий симметрического вида (совместных вероятностей).

Сделайте общие выводы о характере распределения мнений экспертов относительно действенности рычагов экономической политики.

Задание 4. В ходе маркетинговых исследований сбыта оборудования (приборов) для дома и огорода предложено оценить различия качества товара, поступающего от отечественного производителя (От. пр.), из стран СНГ и стран Западной Европы (Зап. Евр.). Выступая в роли экспертов, попробуйте оценить эти различия и заполнить матрицу совместных вероятностей следующего вида:

	От. пр.	СНГ	Зап. Евр.
От. пр.	—		
СНГ		—	
Зап. Евр.			—

Сделайте обобщающие выводы о различиях качества товара, поступающего из различных стран.

Задание 5. Группа стран СНГ характеризуется динамикой (индексными значениями) основных макроэкономических показателей (данные за 2001 г. в %) (табл. 5.2).

Таблица 5.2

	ВВП	Продукция промышленности	Инвестиции в основной капитал	Реальная заработная плата	Потребительские цены
Беларусь	104,1	105,4	94	130,4	161
Россия	105,4	104,9	109	123,5	119
Молдова	106,1	114,2	98	116,4	110
Казахстан	113,2	113,5	121	111,1	108
Украина	109,0	114,2	117	120,5	112

По приведенным данным определите различия стран, используя метрики city-block и евклидово расстояние.

Постройте матрицы различий и объясните несовпадения оценок различий, полученных с применением двух метрик.

Задание 6. По приведенным данным табл. 5.3 определите различия показателей производства основных видов сельскохозяйственной продукции за четыре года.

Таблица 5.3

Год	Производство, тыс. т					
	скот и птица на убой (в живом весе)	молоко	яйца	зерно и зернобобовые	картофель	льноволокно
1990	1758,1	7457,3	3657,0	7035	8590	52
1994	1137,6	5510,0	3400,0	6095	8241	49
1998	980,5	5232,4	3480,8	4831	7574	36
2001	896,7	4818,6	3142,8	5153	7768	32

В расчетах используйте метрику евклидово расстояние. Сделайте выводы. Какую, по вашему мнению, размерность будет иметь теоретическое пространство стимулов (сколько шкал будет в нем)?

Задание 7. На рис. 5.2 координатные оси представляют данные за 2000 г. о числе построенных за год квартир в расчете на 1000 чел. (X_1) и уровне естественного прироста населения, % (X_2). В координатном поле, соответственно значениям анализируемых признаков (X_1, X_2), показано распределение четырех стран: Беларусь, США, Франция, Великобритания.

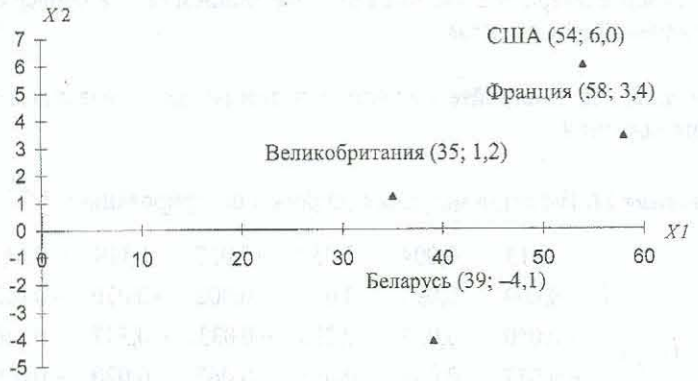


Рис. 5.2

Используя метод нормирования данных $z_{ij} = x_{ij} / \bar{x}_j$ и евклидову метрику, определите расстояния между странами.

При условии, что координата «число построенных квартир на 1000 чел.» (X_1) игнорируется, какая из стран окажется ближе (более похожей) на США и Францию? И, тот же вопрос, при условии, что игнорируется ось абсцисс — «естественный прирост населения» (X_2).

Задание 8. Известна корреляционная матрица, отражающая взаимосвязи трех признаков: X_1 — уровень образования работников, X_2 — стаж работы на предприятии, X_3 — процент выполнения норм выработки

$$R=\begin{pmatrix}1{,}000 & 0{,}670 & 0{,}760\\ 0{,}670 & 1{,}000 & 0{,}820\\ 0{,}760 & 0{,}820 & 1{,}000\end{pmatrix}.$$

Каким образом от оценок сходства анализируемых признаков можно перейти к оценкам различия? Постройте матрицу различий (Δ) на основе приведенных выше значений парных коэффициентов корреляции. Распределение каких стимулов можно получить в теоретическом шкальном пространстве?

Задание 9. Пусть известна исходная матрица различий (Δ)

$$\Delta=\begin{pmatrix}0 & 0{,}340 & 0{,}634 & 0{,}394\\ 0{,}340 & 0 & 0{,}578 & 0{,}394\\ 0{,}634 & 0{,}578 & 0 & 0{,}650\\ 0{,}394 & 0{,}399 & 0{,}650 & 0\end{pmatrix}.$$

Постройте матрицу с двойным центрированием (Δ^*) и проверьте правильность значений ее элементов.

Задание 10. Постройте матрицу с двойным центрированием по данным решения задания 4.

Задание 11. Известна матрица с двойным центрированием (Δ^*)

$$(\Delta^*)=\begin{pmatrix}0{,}013 & -0{,}004 & -0{,}050 & -0{,}037 & 0{,}119 & -0{,}041\\ -0{,}004 & 0{,}095 & 0{,}025 & 0{,}002 & -0{,}016 & -0{,}102\\ -0{,}050 & 0{,}025 & 0{,}289 & -0{,}033 & -0{,}347 & 0{,}116\\ -0{,}037 & 0{,}002 & -0{,}033 & 0{,}067 & 0{,}020 & -0{,}019\\ 0{,}119 & -0{,}016 & -0{,}347 & 0{,}020 & 0{,}632 & -0{,}408\\ -0{,}041 & -0{,}102 & 0{,}116 & -0{,}019 & -0{,}408 & 0{,}454\end{pmatrix}.$$

. Используя соответствующий пакет прикладных программ и реализуя последовательно два различных метода факторного анализа: метод главных факторов и метод максимального правдоподобия, определите координаты стимулов в теоретическом пространстве шкал.

Покажите на рисунках каждую из полученных двух конфигураций стимулов. В чем причина не совпадения конфигураций, найденных с помощью различных методов факторного анализа?

Задание 12. Известны данные об уровне антропогенной нагрузки в областях Республики Беларусь за 2000 г. (табл. 5.4).

Таблица 5.4

Область	Показатели антропогенной нагрузки			
	плотность населения, чел./км^2	использование свежей воды, млн м^3	выбросы вредных веществ в атмосферу, тыс. т [1]	удельный вес территории, загрязненной цезием 137, %
Брестская	45	227	33,6	12
Витебская	34	233	99,3	0,05
Гомельская	38	294	76,3	67
Гродненская	47	174	29,7	6
Минская	81	296	50,0	4
Могилевская	41	196	38,8	33
В среднем по Республике Беларусь	48	237	54,6	21

По приведенным данным:

1) используя метод нормирования данных $z_{ij}=x_{ij}/\bar{x}_j$ и евклидову метрику, постройте матрицу различий областей (Δ);

2) осуществите переход от матрицы различий (Δ) к матрице с двойным центрированием (Δ^*), проверьте правильность построения матрицы Δ^*;

3) используя пакет прикладных программ STATISTICA, реализуйте метод главных факторов (факторный анализ) с целью определения координат стимулов;

4) покажите на рисунке распределение стимулов в двумерном шкальном пространстве (R^{F_1,F_2}), т.е. пространстве двух первых главных факторов;

5) определите названия X_1 и X_2;

Сделайте обобщающие выводы относительно распределения областей в теоретическом шкальном пространстве.

Задание 13. В результате опроса студентов относительно предпочтительности сферы профессиональной занятости получены ранговые оценки различий четырех отраслей экономики (10 — максимальная ранговая оценка различий)

	Промышленность	Сельское хозяйство	Наука	Банковская деятельность
Промышленность	–	6	4	5
Сельское хозяйство	6	–	8	10
Наука	4	8	–	3
Банковская деятельность	5	10	3	–

На основе приведенной матрицы различий определите стартовую конфигурацию отраслей экономики и представьте ее визуально на рисунке в двухмерном шкальном пространстве.

[1] Выбросы от стационарных источников.

Задание 14. Белорусскими экспертами произведено ранжирование рынков стран СНГ с учетом их удаленности, емкости и платежеспособности покупателей. По результатам экспертного оценивания были рассчитаны исходные координаты стимулов (табл. 5.5).[1]

Таблица 5.5

Страна СНГ	Ранг предпочтения	Координаты стимулов	
		X_1	X_2
Азербайджан	7	0,41	0,16
Армения	9	0,32	0,12
Грузия	8	0,26	0,29
Казахстан	4	0,68	0,40
Кыргызстан	5	0,54	0,36
Молдова	3	0,95	0,77
Россия	1	1,15	1,02
Таджикистан	10	0,25	–0,10
Туркменистан	6	0,58	0,46
Узбекистан	6	0,45	0,18
Украина	4	0,75	0,27

Используя векторную модель линейного типа, постройте регрессионное уравнение предпочтений и приведите оценки его надежности (коэффициенты множественной детерминации и корреляции, F-критерий Фишера).

Покажите расположение стран СНГ в трехмерном теоретическом пространстве шкал X_1, X_2, $\hat{\delta}_i$.

Задание 15. Двумя экспертами оценивается качество продукции (масло сливочное), поступающей от восьми различных производителей. В результате получены ранжированные ряды данных предпочтения (табл. 5.6).

Таблица 5.6

Город (производитель)	Ранг предпочтения, max = 1, min = 8	
	эксперт А	эксперт Б
Минск	2	1
Несвиж	4	3
Дзержинск	1	2
Витебск	5	4
Крупки	8	6
Столбцы	3	5
Борисов	7	7
Гомель	6	8

На основе приведенных данных, используя евклидову метрику, постройте индивидуальные матрицы различий и среднюю матрицу различий.

[1] Ранги предпочтения: 1 — максимальный, 10 — минимальный.

По данным усредненной матрицы различий определите стартовую конфигурацию стимулов и покажите ее на рисунке.

Задание 16. Имеются сведения для каждого земельного участка под строительство дома о ценовой ставке, которую согласны оплачивать два потенциальных покупателя, а также значения координат стимулов (табл. 5.7).

Таблица 5.7

Номер земельного участка	Цена, которую согласны оплачивать покупатели, тыс. дол. США		Координаты стимулов, x_{ik}		Оценки предпочтений, δ_{i3}		Процентное соотношение предпочтений Б к А, %
	покупатель А	покупатель Б	x_{i1}	x_{i2}	покупатель А	покупатель Б	
1	20	16	3,50	2,70			
2	5	10	1,99	0,85			
3	12	8	2,20	3,10			
4	40	20	5,60	3,75			
5	25	35	4,30	5,85			

На основании приведенных данных, используя простую евклидову модель, найдите оценки предпочтений для покупателей А и Б.

Определите процентный уровень соответствия предпочтений у двух покупателей. Заполните графы 5, 6, 7 таблицы.

Покажите пространственное расположение предпочтений покупателей в трехмерном пространстве X_1, X_2, δ_{i3}.

Задание 17. Используя метод метрического многомерного шкалирования, проанализируйте и обобщите нижеприведенные данные за 1998 г. о состоянии сельского хозяйства в шести странах (табл. 5.8).

Таблица 5.8

Страна	Производство продукции на душу населения, кг				
	зерновых и зернобобовых	картофеля	сахарной свеклы	мяса в убойном весе	молока
Беларусь	483	757	143	67	299
Австрия	597	82	363	112	383
Болгария	655	56	5	53	145
Великобритания	394	110	165	63	235
Венгрия	1349	100	316	109	206
Германия	552	143	328	71	347
США	1299	80	109	129	264

Необходимые расчеты выполните на компьютере с использованием программы STATISTICA (модуль **Multidimentional Scaling** (многомерное шкалирование).

6. КЛАСТЕРНЫЙ АНАЛИЗ

6.1. Методические рекомендации

6.1.1. Иерархический кластерный анализ

Кластерный анализ — это совокупность методов многомерной классификации, целью которой является образование групп (кластеров) схожих между собой объектов. В отличие от традиционных группировок, рассматриваемых в общей теории статистики, кластерный анализ приводит к разбиению на группы с учетом всех группировочных признаков одновременно. Например, если наблюдаемый объект характеризуется двумя признаками X_1 и X_2, то при использовании методов кластерного анализа оба эти признака учитываются одновременно при отнесении наблюдения в ту или иную группу. Методы кластерного анализа позволяют решать следующие задачи:

- проведение классификации объектов с учетом признаков, отражающих сущность, саму природу объектов;
- проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов, т.е. поиск существующей структуры;
- построение новых классификаций для слабо изученных явлений, когда необходимо установить наличие связей внутри совокупности и попытаться привнести в нее структуру.

Все методы кластерного анализа можно разделить на две группы: иерархические методы (агломеративные и дивизимные) и итеративные (метод k-средних, метод поиска сгущений и т.д.). Достаточно подробный обзор и систематизация различных методов кластерного анализа приводятся в работах [20, 22].

Для записи формализованных алгоритмов кластерного анализа введем следующие условные обозначения:

$X_1, X_2 \ldots, X_n$ — совокупность объектов наблюдения; $X_i = (X_{i1}, X_{i2}, \ldots, X_{im})$ — i-е многомерное наблюдение в m-мерном пространстве признаков $(i = 1,2 \ldots n)$; d_{kl} — расстояние между k-м и l-м объектами; z_{ij} — нормированные значения исходных переменных; D — матрица расстояний между объектами.

Для реализации любого метода кластерного анализа необходимо ввести понятие «сходство объектов». Причем в процессе классификации в каждый кластер должны попадать объекты, имеющие наибольшее сходство друг с другом с точки зрения наблюдаемых переменных.

В кластерном анализе для количественной оценки сходства вводится понятие метрики. Каждый объект описывается m-признаками и представлен как точка в m-мерном пространстве. Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. В кластерном анализе используются различные меры расстояния между объектами:

- евклидово расстояние $$d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2};$$

- взвешенное евклидово расстояние $$d_{ij} = \sqrt{\sum_{k=1}^{m}\omega_k (x_{ik} - x_{jk})^2};$$

- расстояние city-block $$d_{ij} = \sum_{k=1}^{m}\left|x_{ik} - x_{jk}\right|;$$

- расстояние Минковского $$d_{ij} = \left(\sum_{k=1}^{m}\left|x_{ik} - x_{jk}\right|^p\right)^{1/p};$$

- расстояние Махаланобиса $$d_{ij} = (X_i - X_j)^T S_*^{-1}(X_i - X_j),$$

где d_{ij} — расстояние между i-ым и j-ым объектами; x_{il}, x_{jl} — значения l-й переменной и 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; f_l — вес, приписываемый l-й переменной.

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 объектами.

Из всех методов кластерного анализа наиболее 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 расстояний первоначально имеет размерность $(m \times m)$, то полностью процесс объединения завершается за $(m - 1)$ шагов. В итоге все объекты будут объединены в один кластер. Последовательность объединения может быть представлена в виде следующей дендрограммы (рис. 6.1).

Дендрограмма, изображеная на рис. 6.1, показывает, что в данном случае на первом шаге были объединены в один кластер второй и третий объекты при расстоянии между ними 0,15. На втором шаге к ним присоединился первый объект. Расстояние от первого объекта до кластера, содержащего второй и третий объекты, было 0,3 и т.д.

Множество методов иерархического кластерного анализа отличаются алгоритмами классификации, из которых наиболее распространенными являются: метод одиночной связи, метод полных связей, метод средней связи, метод Уорда.

Метод одиночной связи — 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, т.е. тот объект, который имеет наибольшее сходство хотя бы с одним из объектов кластера. Например, имеется матрица расстояний между объектами.

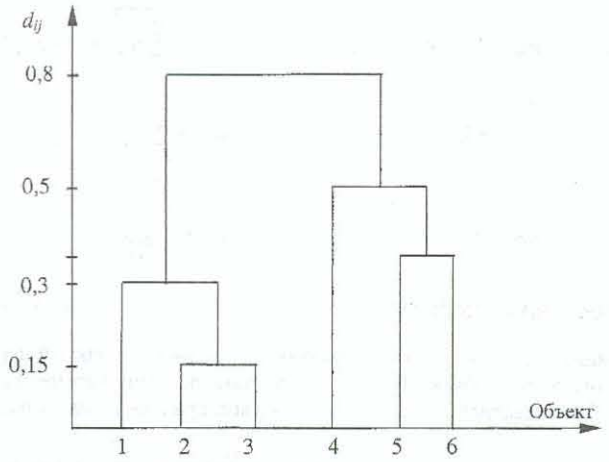


Рис. 6.1. Дендрограмма иерархического кластерного анализа

$$D = \begin{pmatrix} 0 & 1,55 & 2,06 & 1,80 \\ & 0 & 1,72 & 23,70 \\ & & 0 & 3,61 \\ & & & 0 \end{pmatrix}.$$

В первый кластер будут включены первый и второй объекты, так как расстояние между ними минимальное ($d_{12} = 1,55$).

На следующем шаге к этому кластеру будет подключен третий объект, так как расстояние $d_{23} = 1,72 — \min\{d_{13}, d_{23}, d_{14}, d_{24}\}$. На последнем шаге в кластер будет включен четвертый объект. Графически это будет выглядеть следующим образом (рис. 6.2):

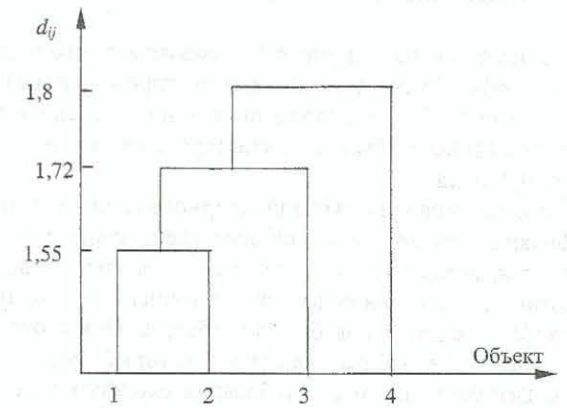


Рис. 6.2. Последовательность объединения четырех предприятий

Метод полных связей — включение нового объекта в кластер происходит только в том случае, если сходство между всеми объектами не меньше некоторого заданного уровня сходства (рис. 6.3).

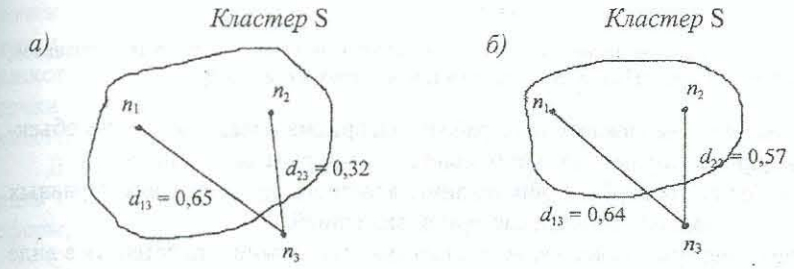


Рис. 6.3. Определение нового состава кластера при различных уровнях сходства наблюдаемых объектов:

если задан уровень сходства 0,25, тогда третий объект будет включен в кластер S, так как $d_{13} > 0,25$ и $d_{23} > 0,25$.

если задан уровень сходства 0,58, тогда третий объект не будет включен в кластер S, так как $d_{13} > 0,58$, а $d_{23} < 0,58$.

Метод средней связи — при включении нового объекта в уже существующий кластер вычисляется среднее значение меры сходства, которое затем сравнивается с заданным пороговым уровнем. Если речь идет об объединении двух кластеров, то вычисляют меру сходства между их центрами и сравнивают ее с заданным пороговым значением. Рассмотрим геометрический пример с двумя кластерами (рис. 6.4).

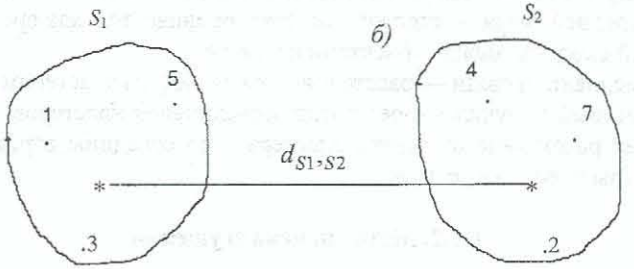


Рис. 6.4. Объединение двух кластеров по методу средней связи:

а) первый кластер б) второй кластер

Если мера сходства между центрами кластеров (d_{S_1,S_2}) будет не меньше заданного уровня, то кластеры S_1 и S_2 будут объединены в один.

Метод Уорда — на первом шаге каждый кластер состоит из одного объекта. Первоначально объединяются два ближайших кластера. Для них определя-

отклонений σ_k

$$\sigma_k = \sum_{i=1}^{nk} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{jk})^2, \qquad (6.1)$$

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 σ_k.

Метод Уорда приводит к образованию кластеров приблизительно равных размеров с минимальной внутрикластерной вариацией.

Алгоритм иерархического кластерного анализа можно представить в виде последовательности процедур:

- нормирование исходных значений переменных;
- расчет матрицы расстояний или матрицы мер сходства;
- определение пары самых близких объектов (кластеров) и их объединение по выбранному алгоритму;
- повторение первых трех процедур до тех пор, пока все объекты не будут объединены в один кластер.

Мера сходства для объединения двух кластеров определяется следующими методами:

- метод «ближайшего соседа» — степень сходства между кластерами оценивается по степени сходства между наиболее схожими (ближайшими) объектами этих кластеров;
- метод «дальнего соседа» — степень сходства оценивается по степени сходства между наиболее отдаленными (несхожими) объектами кластеров;
- метод средней связи — степень сходства оценивается как средняя величина степеней сходства между объектами кластеров;
- 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.

6.1.2. Метод поиска сгущений

Одним из итеративных методов классификации является метод поиска сгущений. Подробно этот метод описан в работах [20, 22]. Суть итеративного алгоритма данного метода заключается в применении гиперсферы заданного радиуса, которая перемещается в пространстве классификационных признаков с целью поиска локальных сгущений объектов.

Метод поиска сгущений требует прежде всего вычисления матрицы расстояний (или матрицы мер сходства) между объектами и выбора первоначального центра сферы. Обычно на первом шаге центром сферы служит объект (точка), в ближайшей окрестности которого расположено наибольшее число соседей. На основе заданного радиуса сферы (R) определяется совокупность то-

(вектор средних значений признаков).

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. Перечисленные процедуры повторяются для всех оставшихся точек. Работа алгоритма завершается за конечное число шагов, и все точки оказываются распределенными по кластерам. Число образовавшихся кластеров заранее неизвестно и сильно зависит от 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.

Существуют 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.

6.1.3. Оценка качества многомерной классификации

Использование в ходе проведения кластерного анализа различных методов и алгоритмов каждый раз приводит к образованию кластеров с различными характеристиками. После завершения многомерной классификации необходимо оценить полученные результаты. Для этой цели используются специальные характеристики — *функционалы качества* [22]. Наилучшим разбиением считается такое, при котором достигается экстремальное (минимальное или максимальное) значение выбранного функционала качества.

Рассмотрим некоторые функционалы качества.

1. Сумма квадратов расстояний до центров кластеров (F_1)

$$F_1 = \sum_{l=1}^{k} \sum_{i \in S_l} d^2\left(X_i, \overline{X}_l\right), \qquad (6.2)$$

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; $d\left(X_l, \overline{X}_l\right)$ — расстояние между i-м объектом и центром l-го кластера.

При использовании этого критерия наилучшим является такое разбиение совокупности объектов, при котором значение F_1 было бы минимальным.

2. Сумма внутрикластерных расстояний между объектами (F_2)

$$F_2 = \sum_{l=1}^{k} \sum_{i,j \in S_l} d_{ij}^2 . \qquad (6.3)$$

В этом случае наилучшим следует считать такое разбиение, при котором достигается минимальное значение F_2, т.е. получены кластеры большой «плотности».

3. Сумма внутрикластерных дисперсий (F_3)

$$F_3 = \sum_{l=1}^{k} \sum_{j=1}^{p} \sigma_{lj}^2, \qquad (6.4)$$

где σ_{lj}^2 — дисперсия j-й переменной в l-м кластере.

В данном случае оптимальным следует считать разбиение, при котором сумма внутрикластерных (внутригрупповых) дисперсий будет минимальной.

На принципе минимизации внутрикластерной дисперсии основаны алгоритмы метода k-средних и метода Уорда [20].

Судить о качестве разбиения позволяют и некоторые простейшие приемы. Например, можно сравнивать средние значения признаков в отдельных кластерах (группах) со средними значениями в целом по всей совокупности объектов. Если групповые средние существенно отличаются от общего среднего значения, то это может являться признаком хорошего разбиения. Оценка существенности, различий может быть выполнена с помощью t-критерия Стьюдента.

6.2. Примеры решения типовых задач

Пример 1. На основании приведенных данных табл. 6.1 необходимо провести классификацию пяти предприятий при помощи иерархического агломеративного кластерного анализа.

Таблица 6.1

Номер предприятия	X_1	X_2	X_3
1	220,0	94,0	264,0
2	185;0	75,0	192,0
3	245,0	80,0	220,0
4	178,0	75,2	96,0
5	170,0	73,1	105,0
Среднее значение $(\bar{x}_j)$	199,6	79,5	175,4
Среднее квадратическое отклонение (σ)	28,4	7,6	65,4

Здесь: X_1 — среднегодовая стоимость основных производственных фондов, млрд р.; X_2 — материальные затраты на один рубль произведенной продукции, коп.; X_3 — объем произведенной продукции, млрд р.

Решение. 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 иметь вид

$$Z = \begin{pmatrix} 0,718 & 1,908 & 1,355 \\ -0,514 & -0,592 & 0,254 \\ 1,596 & 0,066 & 0,682 \\ -0,761 & -0,566 & -1,214 \\ -1,042 & -0,842 & 1,076 \end{pmatrix}.$$

Классификацию проведем при помощи иерархического агломеративного метода. Для построения матрицы расстояний воспользуемся евклидовым расстоянием. Тогда, например, расстояние между первым и вторым объектами будет

$$d_{12} = \left((0,718-(-0,514))^2 + (1,908-(-0,592))^2 + (1,355-0,254)^2\right)^{1/2} = 2,99.$$

Матрица расстояний D_0 характеризует расстояния между объектами, каждый из которых, на первом шаге представляет собой отдельный кластер

$$D_0 = \begin{pmatrix} 0 & 2,990 & 2,149 & 3,861 & 3,277 \\ & 0 & 2,251 & 1,489 & 1,008 \\ & & 0 & 3,090 & 2,818 \\ & & & 0 & 2,324 \\ & & & & 0 \end{pmatrix}.$$

Как видно из матрицы D_0, наиболее близкими являются объекты n_2 и n_5 $d_{45} = 1,008$. Объединим их в один кластер и присвоим ему номер S_2. Пересчитаем расстояния всех оставшихся объектов (кластеров) до кластера S_2, получим новую матрицу расстояний D_1

$$D_1 = \begin{pmatrix} 0 & 2,990 & 3,277 & 3,861 \\ & 0 & 2,818 & 2,324 \\ & & 0 & 3,090 \\ & & & 0 \end{pmatrix}.$$

В матрице D_1 расстояния между кластерами определены по алгоритму «дальнего соседа». Тогда расстояние между объектом n_1 и кластером S_2 равно

$$d_{S_1,S_4} = \max\{d_{12}, d_{15}\} = \max\{2,990; 3,277\} = 3,277 \text{ и т.д.}$$

В матрице D_1 опять находим самые близкие кластеры. Это будут S_1 и S_3 $d_{13} = 2,149$. Следовательно, на этом шаге объединяем S_1 и S_3 кластеры; получим новый кластер, содержащий объекты n_1, n_3. Присвоим ему номер S_1. Теперь имеем три кластера S_1 {1,3}, S_2 {2,5}, S_3 {4}.

$$D_2 = \begin{pmatrix} 0 & 3,277 & 3,861 \\ & 0 & 2,324 \\ & & 0 \end{pmatrix}.$$

Судя по матрице D_2, на следующем шаге объединяем кластеры S_2 и S_3 $d_{23} = 2{,}324$ в один кластер и присвоим ему номер S_2. Теперь имеем только два кластера:

$$\left.\begin{array}{l} S_1 \text{ кластер (объекты } n_1, n_3) \\ S_2 \text{ кластер (объекты } n_2, n_4, n_5) \end{array}\right\} \quad d_{12} = \max\{3{,}277;\ 3{,}861\} = 3{,}861.$$

$$D_3 = \begin{pmatrix} 0 & 3{,}861 \\ & 0 \end{pmatrix}.$$

И, наконец, на последнем шаге объединим кластеры S_1 и S_2 на расстоянии 3,861.

Представим результаты классификации в виде дендрограммы (рис. 6.5). 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.

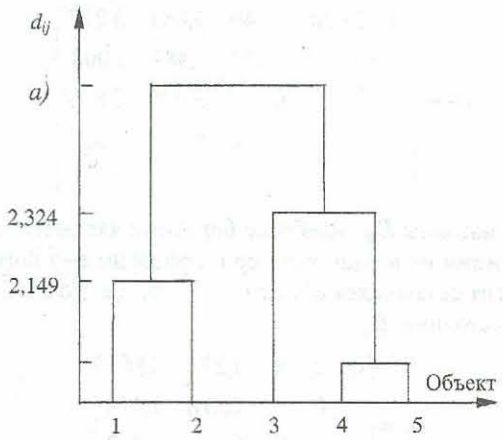


Рис. 6.5. Дендрограмма кластеризации пяти объектов

Пример 2. На основании данных, приведенных ниже, проведите классификацию магазинов по трем признакам: X_1 — площадь торгового зала, м^2, X_2 — товарооборот на одного продавца, ден. ед., X_3 — уровень рентабельности, %.

Номер магазина	X_1	X_2	X_3	Номер магазина	X_1	X_2	X_3
1	100	160	25	6	85	200	35
2	130	200	30	7	60	170	28
3	80	180	20	8	110	150	18
4	40	100	22	9	55	110	15
5	150	90	15	10	110	100	12

Для классификации магазинов используйте метод поиска сгущений (необходимо выделить первый кластер).

Решение. 1. Рассчитаем расстояния между объектами по евклидовой метрике

$$d_{ij} = \sqrt{\sum_{i=1}^{m} (z_{ik} - z_{jk})^2},$$

где z_{ik}, z_{jk} — стандартизированные значения исходных переменных соответственно у i-го и j-го объектов; m — число признаков.

$$z_{ik} = \frac{x_{ik} - \overline{x}_k}{\sigma_k}$$

$$Z = \begin{pmatrix} 0{,}243 & 0{,}345 & 0{,}426 \\ 1{,}156 & 1{,}338 & 1{,}136 \\ -0{,}365 & 0{,}838 & -0{,}284 \\ -1{,}582 & -1{,}134 & 0{,}000 \\ 1{,}764 & -1{,}381 & -0{,}994 \\ -0{,}273 & 1{,}332 & 1{,}847 \\ -0{,}973 & 0{,}592 & 0{,}852 \\ 0{,}547 & 0{,}099 & -0{,}568 \\ -1{,}125 & -0{,}888 & -0{,}994 \\ 0{,}547 & -1{,}134 & -1{,}420 \end{pmatrix}.$$

2. На основе матрицы Z рассчитаем квадратную симметричную матрицу расстояний между объектами (D).

$$D = \begin{bmatrix} 0 & 1{,}524 & 1{,}052 & 2{,}609 & 2{,}324 & 1{,}805 & 1{,}312 & 1{,}069 & 2{,}325 & 2{,}385 \\ & 0 & 2{,}140 & 3{,}860 & 3{,}507 & 1{,}596 & 2{,}274 & 2{,}193 & 3{,}833 & 3{,}608 \\ & & 0 & 2{,}335 & 3{,}156 & 2{,}189 & 1{,}312 & 1{,}208 & 2{,}015 & 2{,}452 \\ & & & 0 & 3{,}499 & 3{,}348 & 2{,}019 & 2{,}525 & 1{,}121 & 2{,}559 \\ & & & & 0 & 4{,}425 & 3{,}846 & 1{,}963 & 2{,}931 & 1{,}313 \\ & & & & & 0 & 1{,}424 & 2{,}833 & 3{,}705 & 4{,}175 \\ & & & & & & 0 & 2{,}138 & 2{,}371 & 3{,}233 \\ & & & & & & & 0 & 1{,}988 & 1{,}499 \\ & & & & & & & & 0 & 1{,}743 \\ & & & & & & & & & 0 \end{bmatrix}.$$

Анализ матрицы расстояний D помогает определить положение первоначального центра сферы и выбрать радиус сферы.

В данном примере большинство «маленьких» расстояний находятся в первой строке, т.е. у первого объекта достаточно много «близких» соседей. Следовательно, первый объект можно взять в качестве центра сферы.

3. Зададим радиус сферы $R = 2$. В этом случае в сферу попадают объекты, расстояние которых до первого объекта меньше 2.

$$d_{12} = 1{,}524, \quad d_{13} = 1{,}052, \quad d_{16} = 1{,}805, \quad d_{17} = 1{,}312, \quad d_{18} = 1{,}069.$$

Для шести точек (объекты 1, 2, 3, 6, 7, 8) определяем координаты центра тяжести: $\bar{x}_* = (0{,}056;\ 0{,}757;\ 0{,}568)$.

4. На следующем шаге алгоритма помещаем центр сферы в точку $\bar{x}_*$ и определяем расстояние каждого объекта до нового центра:

$$d_{1*} = 0{,}474 \quad d_{2*} = 1{,}367 \quad d_{3*} = 0{,}954 \quad d_{4*} = 2{,}565 \quad d_{5*} = 3{,}151$$

$$d_{6*} = 1{,}440 \quad d_{7*} = 1{,}080 \quad d_{8*} = 1{,}401 \quad d_{9*} = 2{,}557 \quad d_{10*} = 2{,}787.$$

Следовательно, в сферу опять попали объекты 1, 2, 3, 6, 7, 8, расстояния которых до центра меньше радиуса сферы. Поскольку в этом случае центр сферы не изменит своих координат, выделение первого кластера закончено, в его состав вошли шесть объектов (1, 2, 3, 6, 7, 8).

5. Чтобы начать формирование второго кластера, нужно поместить центр сферы в одну из точек, не вошедших в первый кластер (объекты 4, 5, 9, 10).

Судя по матрице расстояний D, целесообразно в качестве центра сферы выбрать объекты 9 или 10. Если взять объект 9 в качестве центра сферы, то в сферу попадают четыре точки (объекты 4, 8, 9, 10). Рассчитаем для них координаты нового центра тяжести $\bar{x}_* = (-0{,}403; -0{,}764; -0{,}746)$.

6. Определим расстояние каждого из десяти объектов до точки $\bar{x}_*$:

$$d_{1*} = 1{,}738 \quad d_{2*} = 2{,}646 \quad d_{3*} = 1{,}668 \quad d_{4*} = 1{,}443 \quad d_{5*} = 3{,}151$$

$$d_{6*} = 1{,}888 \quad d_{7*} = 2{,}172 \quad d_{8*} = 1{,}296 \quad d_{9*} = 1{,}888 \quad d_{10*} = 1{,}222.$$

В сферу попадают объекты, которые имеют расстояние до центра $\bar{x}_*$ меньше двух (объекты 1, 3, 4, 8, 9, 10).

На основании матрицы Z по евклидовой метрике определяем новые координаты центра для этих точек $\bar{x}^* = (-0{,}289; -0{,}312; -0{,}473)$.

Для нового центра повторяем п. 6 данного алгоритма:

$$d_{1*} = 1{,}234 \quad d_{2*} = 2{,}720 \quad d_{3*} = 1{,}379 \quad d_{4*} = 1{,}603 \quad d_{5*} = 2{,}373$$

$$d_{6*} = 2{,}843 \quad d_{7*} = 1{,}744 \quad d_{8*} = 0{,}936 \quad d_{9*} = 1{,}141 \quad d_{10*} = 1{,}507.$$

После выполнения этого шага видно, что в сферу с радиусом $R = 2$ попадают объекты 1, 3, 4, 7, 8, 9, 10, т.е. состав второго кластера опять изменился. Следовательно, повторяются процедуры п. 6 и п. 7:

$$\bar{x}_* = (-0{,}387;\ -0{,}183;\ -0{,}284)$$

$$d_{1*} = 1{,}086 \quad d_{2*} = 2{,}590 \quad d_{3*} = 1{,}021 \quad d_{4*} = 1{,}553 \quad d_{5*} = 2{,}562$$

$$d_{6*} = 2{,}617 \quad d_{7*} = 1{,}495 \quad d_{8*} = 1{,}016 \quad d_{9*} = 1{,}243 \quad d_{10*} = 1{,}751.$$

Как видно из полученных расстояний каждого из десяти объектов до центра второго кластера, состав кластера не изменился. На этом выделение второго кластера завершается. В его состав вошли семь объектов 1, 3, 4, 7, 8, 9, 10.

Результаты выделения первых двух кластеров представлены на рис. 6.6.

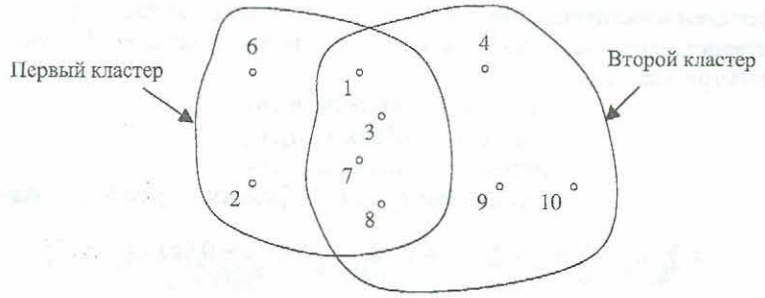


Рис. 6.6. Два выделенных пересекающихся кластера

Так как полученные кластеры являются пересекающимися и в области пересечения находятся четыре объекта из десяти, есть все основания считать результаты классификации неудовлетворительными. По-видимому, следует повторить классификацию, предварительно изменив радиус сферы или выбрать другой метод.

Пример 3. Рассмотрим на примере расчет функционала F_3. Предположим, что шесть объектов наблюдения, приведенные ниже в таблице, 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.

Номер объекта	X_1	X_2	X_3
1	0,10	10	4,9
2	0,80	13	1,9
3	0,40	11	2,8
4	0,18	10	3,7
5	0,25	12	3,2
6	0,67	14	2,3

Рассчитаем дисперсии для каждой переменной в каждом кластере (σ_{ij}^2):

$$\sigma_{11}^2 = 0; \qquad \sigma_{21}^2 = 0{,}004225; \qquad \sigma_{31}^2 = 0{,}008422;$$

$$\sigma_{12}^2 = 0; \qquad \sigma_{22}^2 = 0{,}25; \qquad \sigma_{32}^2 = 0{,}666(6);$$

$$\sigma_{13}^2 = 0; \qquad \sigma_{23}^2 = 0{,}20; \qquad \sigma_{33}^2 = 0{,}1866(6).$$

$$\sum_{j=1}^{3} \sigma_{1j}^2 = 0; \qquad \sum_{j=1}^{3} \sigma_{2j}^2 = 0{,}454; \qquad \sum_{j=1}^{3} \sigma_{3j}^2 = 0{,}862.$$

Тогда суммарная дисперсия всех переменных по трем кластерам будет равна

$$F_3 = \sum_{j=1}^{3} \sigma_{1j}^2 + \sum_{j=1}^{3} \sigma_{2j}^2 + \sum_{j=1}^{3} \sigma_{3j}^2 = 1{,}316.$$

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. Критерии F_1 и F_2 аналогичны по смыслу.

Если оценивать качество разбиения по степени удаленности кластеров друг от друга, то можно использовать функционал F_4 — средние межклассовые расстояния $F_4 = \sum_{\substack{i\in S_q \\ j\in S_l}} d_{ij} \Big/ \sum_{l<q} n_l n_q - \max$.

6.3. Реализация методов кластерного анализа на компьютере

В системе STATISTICA кластерный анализ представлен тремя методами: иерархический кластерный анализ; метод k-средних; метод двухкластерного разбиения. На практике явное предпочтение отдается иерархическим агломеративным процедурам, которые позволяют исследователю проследить по шагам весь процесс образования кластеров и представить его наглядно в графическом виде. Продемонстрируем на примере методику проведения иерархического кластерного анализа, на компьютере с использованием пакета прикладных программ STATISTICA.

Пример. В результате наблюдения выборочной совокупности домашних хозяйств получены следующие данные об уровне среднедушевого потребления нескольких видов продуктов питания за месяц (табл. 6.2).

Таблица 6.2

Номер семьи	Потребление, кг			Номер семьи	Потребление, кг		
	картофель	мясо	фрукты		картофель	мясо	фрукты
1	13	3	2	11	7	6	5
2	8	5	4	12	8	5	7
3	9	6	5	13	9	8	6
4	5	7	4	14	11	7	3
5	9	4	5	15	12	3	6
6	7	5	3	16	8	4	5
7	11	7	1	17	11	3	2
8	14	2	2	18	10	7	1
9	12	5	2	19	5	6	8
10	10	3	4	20	6	7	5

Проведите классификацию семей по уровню душевого потребления, используя метод иерархического агломеративного кластерного анализа. Все расчеты выполните на компьютере с использованием системы STATISTICA (модуль **Cluster Analysis).**

Решение. После ввода исходных данных в электронную таблицу, при помощи кнопки **ANALYS** в основном меню модулей выбираем модуль **Cluster Analysis** (рис. 6.7) и нажимаем кнопку **RAPLAYS**, чтобы развернуть окно выбора метода классификации.

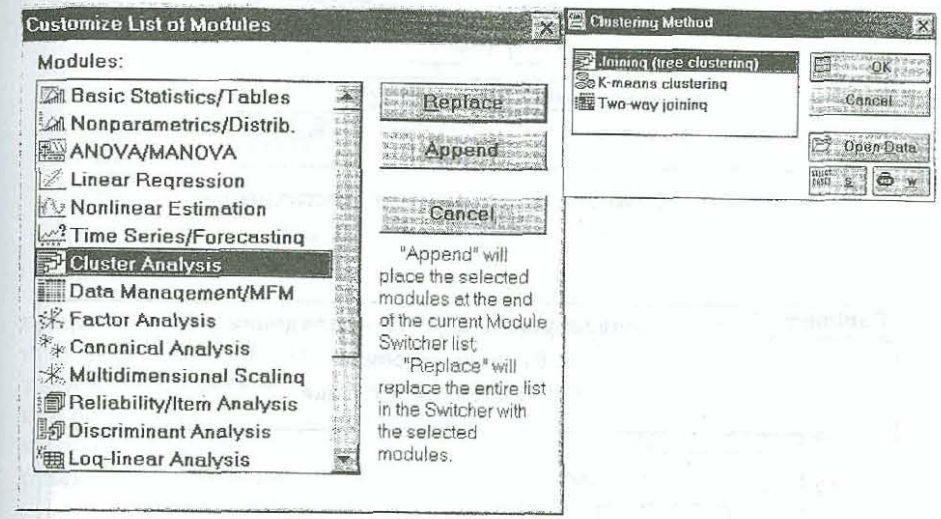


Рис. 6.7. Окно выбора методов кластерного анализа

Из трех предложенных методов выберем **Joining (tree clustering)** — иерархический кластерный анализ. Следующим шагом должен быть выбор переменных, участвующих в классификации. Нажмем кнопку **OK**.

Для двадцати домашних хозяйств в качестве классификационных переменных будем использовать: X_1 — потребление картофеля, кг; X_2 — потребление мяса, кг; X_3 — потребление фруктов, кг.

В развернувшемся окне (рис. 6.8) необходимо выбрать переменные для проведения классификации (кнопка **Variables**), указать, что будем классифицировать наблюдения, а не переменные (в окне **Cluster** задан вариант **Cases**), выбрать метрику сходства объектов (в окне **Distance measure** задать вариант **Euclidean distances**) и нажать кнопку **OK**.

Результаты классификации могут быть представлены в виде вертикальной дендрограммы (рис. 6.9).

Чтобы изобразить матрицу расстояний, на основе которой происходило объединение в кластеры, нужно в окне процедур щелкнуть по кнопке **Distance matrix.** На рис. 6.10 изображен фрагмент этой матрицы.

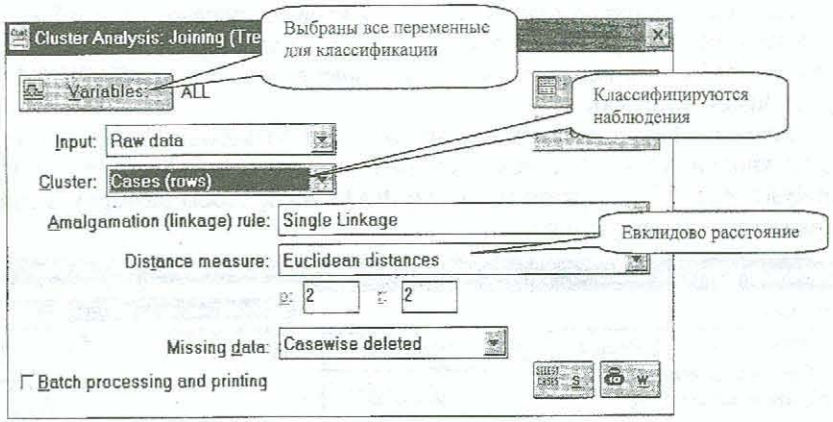


Рис. 6.8. Окно задания параметров классификации

Рис. 6.9. Дендрограмма иерархического кластерного анализа

Continue...	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10
1	,0	5,74	5,83	9,2	5,10	6,40	4,58	1,4	2,24	3,61
2	5,7	,00	1,73	3,6	1,73	1,41	4,69	7,0	4,47	2,83
3	5,8	1,73	,00	4,2	2,00	3,00	4,58	7,1	4,36	3,32
4	9,2	3,61	4,24	,0	5,10	3,00	6,71	10,5	7,55	6,40
5	5,1	1,73	2,00	5,1	,00	3,00	5,39	6,2	4,36	1,73
6	6,4	1,41	3,00	3,0	3,00	,00	4,90	7,7	5,10	3,74
7	4,6	4,69	4,58	6,7	5,39	4,90	,00	5,9	2,45	5,10
8	1,4	7,00	7,07	10,5	6,16	7,68	5,92	,0	3,61	4,58
9	2,2	4,47	4,36	7,5	4,36	5,10	2,45	3,6	,00	3,46
10	3,6	2,83	3,32	6,4	1,73	3,74	5,10	4,6	3,46	,00
11	7,3	1,73	2,00	2,4	2,83	2,24	5,74	8,6	5,92	4,36
12	7,3	3,00	2,45	4,7	2,45	4,12	7,00	8,4	6,40	4,12
13	7,5	3,74	2,24	4,6	4,12	4,69	5,48	8,8	5,83	5,48
14	4,6	3,74	3,00	6,1	4,12	4,47	2,00	5,9	2,45	4,24
15	4,1	4,90	4,36	8,3	3,32	6,16	6,48	4,6	4,47	2,83
16	5,9	1,41	2,24	4,4	1,00	2,45	5,83	7,0	5,10	2,45
17	2,0	4,12	4,69	7,5	3,74	4,58	4,12	3,2	2,24	2,24
18	5,1	4,12	4,24	5,8	5,10	4,12	1,00	6,5	3,00	5,00
19	10,4	5,10	5,00	4,1	5,39	5,48	9,27	11,5	9,27	7,07
20	8,6	3,00	3,16	1,4	4,24	3,00	5,40	9,9	7,00	5,74

Рис. 6.10. Матрица евклидовых расстояний между двадцатью домохозяйствами

Чтобы проследить последовательность объединения кластеров, нужно в окне процедур щелкнуть по кнопке **Amalgamation Schedule** (спецификация объединения). В раскрывшемся окне (рис. 6.11) представлен фрагмент схемы последовательности объединения объектов. В крайнем левом поле указаны значения расстояний при очередном объединении кластеров. Как видно у схемы, самыми близкими на основе евклидовой метрики оказались 5-й и 16-й объекты, а также 7-й и 18-й объекты. Их объединение произошло на первом шаге на расстоянии $d_{ij} = 1{,}000$.

CLUSTER ANALYSIS	Single Linkage Euclidean distances						
linkage distance	Obj No 1	Obj No 2	Obj No 3	Obj No 4	Obj No 5	Obj No 6	C
1,000000	C_5	C_16					
1,000000	C_7	C_18					
1,414214	C_1	C_8					
1,414214	C_2	C_5	C_16				
1,414214	C_2	C_5	C_16	C_6			
1,414214	C_4	C_20					
1,414214	C_4	C_20	C_11				
1,732051	C_2	C_5	C_16	C_6	C_3		
1,732051	C_2	C_5	C_16	C_6	C_3	C_4	
1,732051	C_2	C_5	C_16	C_6	C_3	C_4	
2,000000	C_1	C_8	C_17				
2,000000	C_7	C_18	C_14				
2,236068	C_1	C_8	C_17	C_2	C_5	C_16	
2,236068	C_1	C_8	C_17	C_2	C_5	C_16	
2,236068	C_1	C_8	C_17	C_2	C_5	C_16	
2,236068	C_1	C_8	C_17	C_2	C_5	C_16	
2,449490	C_1	C_8	C_17	C_2	C_5	C_16	
2,828427	C_1	C_8	C_17	C_2	C_5	C_16	
3,316625	C_1	C_8	C_17	C_2	C_5	C_16	

Рис. 6.11. Схема последовательного объединения кластеров

При необходимости полученные результаты классификации можно сохранить при помощи процедуры **Save distance matrix**.

6.4. Контрольные задания

Задание 1. Имеются следующие данные о потребительских расходах населения по двенадцати районам (табл. 6.3).

Таблица 6.3

Район	X_1	X_2	X_3
1	1,32	0,55	0,08
2	1,29	0,59	0,09
3	1,28	0,58	0,08
4	1,32	0,61	0,09
5	1,39	0,63	0,01
6	1,45	0,67	0,02
7	1,46	0,73	0,01
8	1,49	0,76	0,08
9	1,53	0,77	0,07
10	1,53	0,82	0,06
11	1,55	0,84	0,03
12	1,61	0,88	0,02

Здесь X_1 — расходы на питание; X_2 — расходы на одежду; X_3 — расходы на лекарства, ден. ед.

Проведите группировку районов при помощи методов кластерного анализа, используя алгоритмы «ближайшего соседа» и «дальнего соседа» . Сравните полученные результаты.

Задание 2. В результате проведенной многомерной классификации десяти регионов по трем признакам получено следующее разбиение на три группы (табл. 6.4).

Таблица 6.4

Номер группы и номер объекта	Значения группировочных признаков			
I группа (3, 7)	$X_{31} = 65$	$X_{71} = 64$		
	$X_{32} = 15,4$	$X_{72} = 15,2$		
	$X_{33} = 9,4$	$X_{73} = 9,9$		
II группа (4, 5, 6, 8)	$X_{41} = 60$	$X_{51} = 61$	$X_{61} = 59$	$X_{81} = 62$
	$X_{42} = 16$	$X_{52} = 16$	$X_{62} = 16,5$	$X_{82} = 16,7$
	$X_{43} = 11,1$	$X_{53} = 11$	$X_{63} = 11,7$	$X_{83} = 12,0$
III группа (1, 2, 9, 10)	$X_{11} = 63$	$X_{21} = 64$	$X_{91} = 61$	$X_{10\,1} = 65$
	$X_{12} = 17$	$X_{22} = 16,6$	$X_{92} = 16,4$	$X_{10\,2} = 17$
	$X_{13} = 12$	$X_{23} = 12,1$	$X_{93} = 12,3$	$X_{10\,3} = 12,4$

Здесь: X_1 — плата за жилье, ден. ед.; X_2 — плата за бензин, ден. ед.; X_3 — плата за лекарства, ден. ед.

Проведите классификацию этих же регионов методом поиска сгущений и сравните качество полученных разбиений при помощи функционала F_2.

Задание 3. По результатам статистического наблюдения получены следующие данные (табл. 6.5).

Таблица 6.5

Отрасль (подотрасль)	X_1	X_2	X_3
Электроэнергетика	18	3,3	103
Топливная промышленность	19	1,5	102
Черная металлургия	14	1,3	101
Химическая и нефтехимическая промышленность	24	5,9	105
Машиностроение и металлообработка	23	10,9	102
Лесная, деревообрабатывающая и целлюлозно-бумажная промышленность	30	1,1	101
Промышленность строительных материалов	15	2,0	110
Текстильная промышленность	22	1,4	94
Швейная промышленность	28	0,5	103
Кожевенная, меховая и обувная промышленность	29	0,3	110
Пищевкусовая промышленность	11	0,9	102
Мясная и молочная промышленность	10	0,9	105
Рыбная промышленность	34	0,1	101

Здесь: X_1 — средний уровень рентабельности предприятий по отраслям промышленности, %; X_2 — среднегодовая стоимость промышленно-производственных основных фондов по отраслям промышленности, млн ден. ед.; X_3 — индекс физического объема производства, %.

Произведите многомерную классификацию отраслей, используя алгоритмы «ближайшего соседа» и «дальнего соседа». Постройте дендрограмму для каждого алгоритма и сравните полученные результаты разбиения.

Задание 4. По двенадцати регионам страны известны помесячные значения демографических показателей (табл. 6.6).

Таблица 6.6

Регион	Браки на 1000 человек	Родившиеся на 1000 человек
1	5,7	8,1
2	7,0	9,2
3	3,2	9,7
4	8,6	9,0
5	2,8	9,1
6	6,4	10,1
7	8,7	9,9
8	10,7	9,4
9	8,5	9,4
10	9,0	9,0
11	10,1	8,5
12	5,6	8,4

Используя метод иерархического кластерного анализа, проведите классификацию регионов по двум переменным, используя алгоритм «средней связи».

Задание 5. Дана матрица расстояний между объектами (D)

$$D = \begin{pmatrix} 0 & 5{,}78 & 1{,}14 & 3{,}16 & 4{,}05 \\ & 0 & 4{,}23 & 2{,}03 & 5{,}01 \\ & & 0 & 3{,}68 & 2{,}15 \\ & & & 0 & 1{,}12 \\ & & & & 0 \end{pmatrix}.$$

Проведите классификацию наблюдаемых объектов, используя дивизимный алгоритм кластерного анализа.

Процесс классификации отразите в виде дендрограммы и поясните полученные результаты.

Задание 6. Известны следующие данные по пяти сельскохозяйственным предприятиям области (табл. 6.7).

Таблица 6.7

Номер хозяйства	X_1	X_2	X_3
1	24	0,9	0,2
2	23	0,5	0,4
3	24	1,2	0,4
4	24	1,1	0,5
5	26	1,3	2,0

Здесь: X_1 — урожайность зерновых, ц/га; X_2 — внесено минеральных удобрений на 1 га ц; X_3 — доля площади полей с защитными лесными полосами во всей площади.

Рассчитайте матрицы расстояний, используя евклидову метрику и метрику city-block, предварительно нормировав данные по способу $z_{ij} = \frac{x_{ij}}{x_j}$.

Сравните и проанализируйте полученные результаты.

Задание 7. Используя данные табл. 6.8, проведите иерархический кластерный анализ по алгоритмам «средней связи» и «медианной связи».

Таблица 6.8

Номер продукта	Уровень качества, баллов	Выпуск продукта в натуральном выражении, тыс. шт.	Цена единицы продукта, тыс. ден. ед.
1	6	80	40
2	8	150	20
3	12	100	70

Окончание табл. 6.8

Номер продукта	Уровень качества, баллов	Выпуск продукта в натуральном выражении, тыс. шт.	Цена единицы продукта, тыс. ден.ед.
4	15	115	30
5	14	135	80
6	16	120	60
7	17	140	75
8	13	110	50
9	19	125	85
10	7	145	90

Рассчитайте для каждого разбиения сумму квадратов расстояний до центров классов (функционал F_1) и определите, какой из двух алгоритмов приводит к лучшим результатам.

Задание 8. По приведенным данным табл. 6.9 проведите иерархический кластерный анализ по алгоритмам «ближайшего соседа» и «дальнего соседа». Для определения расстояний воспользуйтесь метрикой city-block.

Таблица 6.9

Регион	Численность населения, тыс. чел.	Число магазинов, единиц	Число предприятий общественного питания, единиц
1	600	1560	600
2	180	522	252
3	320	704	288
4	450	1390	675
5	750	1575	685
6	120	336	168
7	900	1890	1260
8	640	1600	768
9	1220	3660	1342
10	250	675	275

Рассчитайте для каждого разбиения сумму внутриклассовых расстояний между объектами (функционал F_2). Исходные данные предварительно нормируйте по способу $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$.

Задание 9. Используя иерархический агломеративный кластерный анализ и метрику «расстояние Махаланобиса», определите по данным табл. 6.10

группы стран, однородных по силе колебаний уровней макроэкономических показателей[1].

Таблица 6.10

Страна	Индексы (1999 г. в процентах к 1998 г.)	
	потребительских цен (X_1)	физического объема ВВП (X_2)
Беларусь	351,2	103,4
Россия	136,5	103,2
Украина	119,2	99,6
Австрия	100,6	102,2
Бельгия	101,1	102,5
Болгария	101,8	102,6
Великобритания	101,6	102,1
Венгрия	110,0	104,5
Германия	100,6	101,5
Польша	107,3	104,1
Словакия	110,6	101,9
США	102,2	104,2
Франция	100,5	102,9
Чешская Республика	102,1	99,8
Швеция	100,5	103,8

Постройте также типологическую группировку, выделив страны с незначительными и существенными колебаниями уровня потребительских цен. Сравните состав полученных групп с результатами иерархической агломеративной классификации.

[1] Беларусь и страны мира // Стат. сб. Мн., 2000.

7. ДИСКРИМИНАНТНЫЙ АНАЛИЗ

7.1. Методические рекомендации

Дискриминантный анализ — это совокупность методов, позволяющих решать задачи идентификации объектов по заданному набору характерных признаков.

Весь процесс проведения дискриминантного анализа разбивается на два этапа и каждый из них можно рассматривать как совершенно самостоятельный метод.

Первый этап — выявление и формальное описание различий между существующими множествами (группами) наблюдаемых объектов.

Второй этап — непосредственная классификация новых объектов, т.е. отнесение каждого объекта к одному из существующих множеств.

Пусть имеется множество единиц наблюдения, каждая из которых характеризуется несколькими признаками (переменными): x_{ij} — значения j-й переменной у i-го объекта $i=\overline{1,n}$; $j=\overline{1,p}$.

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) $k=\overline{1,q}$.

Признаки, которые используются для того чтобы отличать одно подмножество от другого, называются *дискриминантными переменными*.

Число дискриминантных переменных не ограничено, но на практике выбор должен осуществляться на основании логического анализа исходной информации. Число объектов наблюдения должно превышать число дискриминантных переменных, т.е. $p < n$. Предполагается, что дискриминантные переменные — линейно независимые нормально распределенные многомерные величины.

Рассмотрим случай для двух дискриминантных переменных. Функция $f(X)$ называется *канонической дискриминантной функцией*, а величины X_1 и X_2 — дискриминантными переменными

$$f(X) = a_1X_1 + a_2X_2. \qquad (7.1)$$

Дискриминантная функция может быть как линейной, так и нелинейной. Выбор вида этой функции зависит от геометрического расположения разделяемых классов в пространстве дискриминантных переменных.

Коэффициенты дискриминантной функции (a_i) определяются таким образом, чтобы $\overline{f}_1(X)$ и $\overline{f}_2(X)$ как можно больше отличались между собой.

Вектор коэффициентов дискриминантной функции (A) определяется по формуле

$$A = S_x^{-1}(\overline{X}_1 - \overline{X}_2). \qquad (7.2)$$

Полученные значения коэффициентов подставляют в формулу (7.1) и для каждого объекта в обоих множествах вычисляют дискриминантные функции $f(X)$, затем находят среднее значение для каждой группы $(\overline{f}_k)$. Таким образом, каждому i-му наблюдению, которое первоначально описывалось m-переменными,

будет соответствовать одно значение дискриминантной функции, и размерность признакового пространства снижается.

Классификация при наличии двух обучающих выборок. Перед тем как приступить непосредственно к процедуре классификации, нужно определить границу, разделяющую два множества. Такой величиной может быть значение функции, равноудаленное от $\overline{f}_1$ и $\overline{f}_2$, т.е.

$$c = \frac{1}{2}(\overline{f}_1 + \overline{f}_2). \qquad (7.3)$$

Величина c называется *константой дискриминации*.

Объекты, расположенные над разделяющей поверхностью $f(x) = a_1x_1 + a_2x_2 + \cdots + a_px_p = c$ находятся ближе к центру множества M_1, следовательно, могут быть отнесены к первой группе, а объекты, расположенные ниже этой поверхности, ближе к центру второго множества, т.е. относятся ко второй группе. Если граница между группами будет выбрана как сказано выше, то в этом случае суммарная вероятность ошибочной классификации будет минимальной.

Классификация при наличии *k*-обучающих выборок. Рассмотрим особенности классификации объектов, возникающие при наличии k-обучающих выборок ($k > 2$). Как и в случае с двумя обучающими выборками, предполагается, что каждое множество является нормально распределенным с различными векторами средних значений. Оценка совместной ковариационной матрицы S_* рассчитывается по следующей формуле:

$$S_* = \frac{\sum_{i=1}^{k}(n_i - 1)\cdot S_i}{\sum_{i=1}^{k} n_i}, \qquad (7.4)$$

где k — количество обучающих выборок; S_i — матрица ковариации для i-й выборки; n_i — численность i-й выборки.

В этом случае каждому множеству ставится в соответствие своя дискриминантная функция вида

$$f_i = a_{0i} + a_{1i}x_1 + a_{2i}x_2 + \ldots + a_{mi}x_m.$$

Вектор коэффициентов этой функции a_{ij} $(j = \overline{1,m})$ рассчитывается по формуле $a_i = \overline{X}_i \cdot S_*^{-1}$, а свободный член $a_{0i} = -\frac{1}{2}\overline{X}_i S_*^{-1}\overline{X}_i$.

Новый классифицируемый объект с переменными $Y_1, Y_2, \ldots, Y_m$ будет отнесен к тому множеству M_i, для которого величина $f_i = c_i + a_{1i}Y_1 + a_{2i}Y_2 + \ldots + a_{mi}Y_m$ будет максимальной.

В заключение необходимо отметить, что в данном параграфе рассмотрен только один из методов проведения дискриминантного анализа. Более подробно другие методы и алгоритмы дискриминантного анализа описаны в специальной литературе [4, 22, 27].

7.2. Примеры решения типовых задач

Пример 1. Имеются следующие данные по двум группам промышленных предприятий (табл. 7.1).

Таблица 7.1

Первая группа (k_1)			Вторая группа (k_2)		
Номер предприятия	Удельный вес потерь от брака, % (X_1)	Фондоотдача активной части основных фондов, ден. ед. (X_2)	Номер предприятия	Удельный вес потерь от брака, % (Y_1)	Фондоотдача активной части основных фондов, ден. ед. (Y_2)
1	0,15	1,91	1	0,48	0,88
2	0,34	1,68	2	0,41	0,62
3	0,09	1,89	3	0,62	1,09
4	0,21	2,30	4	0,50	1,32
			5	1,20	0,68
–	$\overline{X}_1 = 0{,}198$	$\overline{X}_2 = 1{,}945$	–	$\overline{Y}_1 = 0{,}642$	$\overline{Y}_2 = 0{,}918$

1. На основании приведенных данных следует найти оценки векторов средних $\overline{X}$, $\overline{Y}$ и ковариационных матриц (S_x, S_y), а также оценку суммарной ковариационной матрицы (S_*) и обратной к ней (S_*^{-1}).

2. Рассчитайте вектор оценок коэффициентов дискриминантной функции (A) и определите ее средние значения для каждого множества. Определите константу дискриминации (c).

Вычислите оценки значений дискриминантной функции для предприятия, у которого переменные принимают значения: удельный вес потерь от брака (Z_1) равен 0,2 %; фондоотдача активной части основных фондов (Z_2) равна 0,75 ден. ед.

Определите, к какой из двух групп следует отнести данное предприятие.

Решение. 1. Для каждой группы предприятий рассчитаем ковариационные матрицы.

Первая группа (k_1):

$$X^T X = \begin{pmatrix} -0{,}0475 & 0{,}1425 & -0{,}1075 & 0{,}0125 \\ -0{,}035 & -0{,}265 & -0{,}055 & 0{,}355 \end{pmatrix} \times \begin{pmatrix} -0{,}0475 & -0{,}035 \\ 0{,}1425 & -0{,}265 \\ -0{,}1075 & -0{,}055 \\ 0{,}0125 & 0{,}355 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}03428 & -0{,}02575 \\ -0{,}02575 & 0{,}20050 \end{pmatrix},$$

$$S_1 = \frac{1}{4}\begin{pmatrix} 0{,}034275 & -0{,}02575 \\ -0{,}02575 & 0{,}20050 \end{pmatrix} = \begin{pmatrix} 0{,}00856875 & -0{,}0064375 \\ -0{,}0064375 & 0{,}050125 \end{pmatrix}.$$

Вторая группа (k_2):

$$Y^T Y = \begin{pmatrix} -0{,}162 & -0{,}232 & -0{,}022 & -0{,}142 & 0{,}558 \\ -0{,}038 & -0{,}298 & 0{,}172 & 0{,}402 & -0{,}238 \end{pmatrix} \times \begin{pmatrix} -0{,}162 & -0{,}038 \\ -0{,}232 & -0{,}298 \\ -0{,}022 & 0{,}175 \\ -0{,}142 & 0{,}402 \\ 0{,}558 & -0{,}238 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}41208 & -0{,}11838 \\ -0{,}11838 & 0{,}33808 \end{pmatrix},$$

$$S_2 = \frac{1}{5}\begin{pmatrix} 0{,}41208 & -0{,}11838 \\ -0{,}11838 & 0{,}33808 \end{pmatrix} = \begin{pmatrix} 0{,}082416 & -0{,}023676 \\ -0{,}023676 & 0{,}067616 \end{pmatrix}.$$

Совместная ковариационная матрица S_* и обратная матрица S_*^{-1} будут иметь следующий вид:

$$S_* = \frac{1}{n_1 + n_2 - 2}(n_1 S_1 + n_2 S_2) = \frac{1}{7}\begin{pmatrix} 0{,}4464 & -0{,}1441 \\ -0{,}1441 & 0{,}5386 \end{pmatrix} = \begin{pmatrix} 0{,}0638 & -0{,}0206 \\ -0{,}0206 & 0{,}0769 \end{pmatrix};$$

$$S^{-1} = \begin{pmatrix} 17{,}1661 & 4{,}5938 \\ 4{,}5938 & 14{,}2266 \end{pmatrix}.$$

2. Вектор коэффициентов дискриминантной функции (A) равен

$$A = S^{-1}(\overline{X} - \overline{Y}) = \begin{pmatrix} 17{,}1661 & 4{,}5938 \\ 4{,}5938 & 14{,}2266 \end{pmatrix} \times \begin{pmatrix} -0{,}445 \\ 1{,}027 \end{pmatrix} = \begin{pmatrix} -2{,}912 \\ 12{,}569 \end{pmatrix}.$$

Вектор значений дискриминантной функции для первого подмножества

$$f_1 = X \cdot A = \begin{pmatrix} 0{,}15 & 1{,}91 \\ 0{,}34 & 1{,}68 \\ 0{,}09 & 1{,}89 \\ 0{,}21 & 2{,}30 \end{pmatrix} \times \begin{pmatrix} -2{,}913 \\ 12{,}569 \end{pmatrix} = \begin{pmatrix} 23{,}569 \\ 20{,}125 \\ 23{,}493 \\ 28{,}297 \end{pmatrix}.$$

Вектор значений дискриминантной функции для второго подмножества

$$f_2 = Y \cdot A = \begin{pmatrix} 0{,}48 & 0{,}88 \\ 0{,}41 & 0{,}62 \\ 0{,}62 & 1{,}09 \\ 0{,}50 & 1{,}32 \\ 1{,}20 & 0{,}68 \end{pmatrix} \times \begin{pmatrix} -2{,}913 \\ 12{,}569 \end{pmatrix} = \begin{pmatrix} 9{,}663 \\ 6{,}599 \\ 11{,}894 \\ 15{,}135 \\ 5{,}052 \end{pmatrix}.$$

Средние значения дискриминантной функции и константа дискриминации

$$\left.\begin{aligned} \bar{f}_1 &= 23{,}871 \\ \bar{f}_2 &= 9{,}668 \end{aligned}\right\} \qquad c = \frac{1}{2}(23{,}871 + 9{,}668) = 16{,}770.$$

3. Рассчитаем значение дискриминантной функции для предприятия со значениями переменных равными $Z_1 = 0{,}2$; $Z_2 = 0{,}75$

$$f_z = (0{,}2 \quad 0{,}75) \times \begin{pmatrix} -2{,}913 \\ 12{,}569 \end{pmatrix} = 8{,}844.$$

Так как полученное значение дискриминантной функции для рассматриваемого предприятия $f_z = 8{,}844 < c = 16{,}770$, его следует отнести ко второму подмножеству.

Пример 2. Рассмотрим пример проведения дискриминантного анализа при наличии трех обучающих выборок.

Для того чтобы проще было изобразить на рисунке объекты каждого из подмножества, предположим, что каждый объект (предприятие) характеризуется только двумя переменными:

Первое подмножество

Номер объекта	X_1	X_2
1	9,4	1,9
2	9,9	1,7
3	9,1	2,3
4	10,0	2,6
5	9,4	2,0
6	9,0	1,9
	$\overline{X}_1 = 9{,}47$	$\overline{X}_2 = 2{,}07$

Второе подмножество

Номер объекта	X_1	X_2
1	7,4	1,09
2	6,7	1,23
3	6,6	1,33
4	7,0	1,25
5	7,5	1,15
	$\overline{X}_1 = 7{,}04$	$\overline{X}_2 = 1{,}21$

Третье подмножество

Номер объекта	X_1	X_2
1	5,5	0,9
2	5,1	0,88
3	5,4	1,20
4	5,8	1,25
	$\overline{X}_1 = 5{,}45$	$\overline{X}_2 = 1{,}06$

Здесь: X_1 — выработка продукции на одного работающего, тыс. ден. ед.; X_2 — фондоотдача основных производственных фондов, ден. ед.

Для каждого подмножества объектов определяем ковариационную матрицу:

- для первого подмножества

$$S_1 = \begin{pmatrix} 0{,}1389 & 0{,}0222 \\ 0{,}0222 & 0{,}0889 \end{pmatrix};$$

- для второго подмножества

$$S_2 = \begin{pmatrix} 0{,}1304 & -0{,}0264 \\ -0{,}0264 & 0{,}0069 \end{pmatrix};$$

- для третьего подмножества

$$S_3 = \begin{pmatrix} 0{,}0625 & 0{,}0286 \\ 0{,}0286 & 0{,}0284 \end{pmatrix}.$$

Для каждой из существующих пар подмножеств рассчитаем совместные ковариационные матрицы и матрицы, обратные к ним.

Совместная ковариационная матрица и обратная к ней матрица для первого и второго подмножеств

$$S_{*(1,2)} = \frac{1}{6+5-2}\left[6S_1 + 5S_2\right] =$$

$$= \frac{1}{9}\left[\begin{pmatrix} 0{,}8334 & 0{,}1334 \\ 0{,}1334 & 0{,}5334 \end{pmatrix} + \begin{pmatrix} 0{,}652 & -0{,}132 \\ -0{,}132 & 0{,}0344 \end{pmatrix}\right] = \begin{pmatrix} 0{,}1650 & -0{,}00016 \\ -0{,}00016 & 0{,}0631 \end{pmatrix}$$

$$S^{-1}_{*(1,2)} = \begin{pmatrix} 6{,}0673 & 0{,}0154 \\ 0{,}0154 & 15{,}8654 \end{pmatrix}.$$

Определяем разность векторов средних значений $\overline{X}_1 - \overline{X}_2 = \begin{pmatrix} 2{,}43 \\ 0{,}86 \end{pmatrix}$.

Совместная ковариационная матрица и обратная к ней матрица для первого и третьего подмножеств

$$S_{*(1,3)} = \frac{1}{6+4-2}\left[6S_1 + 4S_3\right] =$$

$$= \frac{1}{8}\left[\begin{pmatrix} 0{,}8334 & 0{,}1334 \\ 0{,}1334 & 0{,}5334 \end{pmatrix} + \begin{pmatrix} 0{,}2500 & 0{,}1145 \\ 0{,}1145 & 0{,}1137 \end{pmatrix}\right] = \begin{pmatrix} 0{,}1354 & 0{,}031 \\ 0{,}031 & 0{,}0809 \end{pmatrix}$$

$$S^{-1}_{*(1,3)} = \begin{pmatrix} 8{,}098 & -0{,}3103 \\ -0{,}3103 & 13{,}5535 \end{pmatrix}.$$

Разность векторов средних значений

$$\overline{X}_1 - \overline{X}_3 = \begin{pmatrix} 4{,}02 \\ 1{,}01 \end{pmatrix}.$$

Наконец, совместная ковариационная матрица и обратная к ней матрица для второго и третьего подмножеств

$$S_{*(2,3)} = \frac{1}{5+4-2}\left[\begin{pmatrix} 0{,}652 & -0{,}132 \\ -0{,}132 & 0{,}0344 \end{pmatrix} + \begin{pmatrix} 0{,}2500 & 0{,}1145 \\ 0{,}1145 & 0{,}1137 \end{pmatrix}\right] =$$

$$= \begin{pmatrix} 0{,}1288 & -0{,}0025 \\ -0{,}0025 & 0{,}0212 \end{pmatrix};$$

$$S^{-1}_{*(2,3)} = \begin{pmatrix} 7{,}794 & 0{,}919 \\ 0{,}919 & 47{,}353 \end{pmatrix}.$$

Разность векторов средних значений

$$\overline{X}_1 - \overline{X}_2 = \begin{pmatrix} 1{,}59 \\ 0{,}15 \end{pmatrix}.$$

На основании обратных матриц $S^{-1}_{*(1,2)}$, $S^{-1}_{*(1,3)}$, $S^{-1}_{*(2,3)}$ и разностей векторов средних значений определим для каждой пары подмножеств векторы коэффициентов дискриминантных функций:

$$\overline{A_1} = S^{-1}_{1,2} \cdot \left(\overline{X}_1 - \overline{X}_2\right) = \begin{pmatrix} 6{,}0673 & 0{,}0154 \\ 0{,}0154 & 15{,}8654 \end{pmatrix} \times \begin{pmatrix} 2{,}43 \\ 0{,}86 \end{pmatrix} = \begin{pmatrix} 14{,}76 \\ 13{,}68 \end{pmatrix};$$

$$A_2 = S^{-1}_{1,3} \cdot \left(\overline{X}_1 - \overline{X}_3\right) = \begin{pmatrix} 8{,}098 & -0{,}3103 \\ -0{,}3103 & 13{,}5535 \end{pmatrix} \times \begin{pmatrix} 4{,}02 \\ 1{,}01 \end{pmatrix} = \begin{pmatrix} 32{,}24 \\ 12{,}44 \end{pmatrix};$$

$$A_3 = S^{-1}_{2,3} \cdot \left(\overline{X}_2 - \overline{X}_3\right) = \begin{pmatrix} 7{,}794 & 0{,}919 \\ 0{,}913 & 47{,}353 \end{pmatrix} \times \begin{pmatrix} 1{,}59 \\ 0{,}15 \end{pmatrix} = \begin{pmatrix} 12{,}53 \\ 8{,}55 \end{pmatrix}.$$

Определяем значения дискриминантных функций по матрице значений исходных переменных в каждом подмножестве.

Для разграничения первого и второго подмножеств имеем

$$f_{11} = X_1 \cdot A_1 = \begin{pmatrix} 9{,}4 & 1{,}9 \\ 9{,}9 & 1{,}7 \\ 9{,}1 & 2{,}3 \\ 10{,}0 & 2{,}6 \\ 9{,}4 & 2{,}0 \\ 9{,}0 & 1{,}9 \end{pmatrix} \times \begin{pmatrix} 14{,}76 \\ 13{,}68 \end{pmatrix} = \begin{pmatrix} 164{,}7 \\ 169{,}4 \\ 165{,}8 \\ 183{,}2 \\ 166{,}1 \\ 158{,}8 \end{pmatrix};$$

- среднее значение дискриминантной функции f_1 для первого подмножества $\bar{f}_{11} = 168{,}0$.

$$f_{12} = X_2 \cdot A_1 = \begin{pmatrix} 7,4 & 1,09 \\ 6,7 & 1,23 \\ 6,6 & 1,33 \\ 7,0 & 1,25 \\ 7,5 & 1,15 \end{pmatrix} \times \begin{pmatrix} 14,76 \\ 13,68 \end{pmatrix} = \begin{pmatrix} 124,1 \\ 115,7 \\ 115,6 \\ 120,4 \\ 126,4 \end{pmatrix};$$

- среднее значение дискриминантной функции f_1 для второго подмножества $\bar{f}_{12} = 120,44$.

Для первого и второго подмножеств константа дискриминации c_1 равна

$$c_1 = \frac{168 + 120,44}{2} = 144,22$$

Для разграничения первого и третьего подмножеств имеем

$$f_{13} = X_1 \cdot A_2 = \begin{pmatrix} 9,4 & 1,9 \\ 9,9 & 1,7 \\ 9,1 & 2,3 \\ 10,0 & 2,6 \\ 9,4 & 2,0 \\ 9,0 & 1,9 \end{pmatrix} \times \begin{pmatrix} 32,24 \\ 12,44 \end{pmatrix} = \begin{pmatrix} 326,7 \\ 340,3 \\ 321,9 \\ 354,7 \\ 327,9 \\ 313,8 \end{pmatrix};$$

- среднее значение дискриминантной функции f_3 для первого подмножества $\bar{f}_{13} = 330,5$;

$$f_{33} = X_3 \cdot A_2 = \begin{pmatrix} 5,5 & 0,90 \\ 5,1 & 0,88 \\ 5,4 & 1,20 \\ 5,8 & 1,25 \end{pmatrix} \times \begin{pmatrix} 32,24 \\ 12,44 \end{pmatrix} = \begin{pmatrix} 188,5 \\ 175,4 \\ 189,0 \\ 202,5 \end{pmatrix};$$

- среднее значение дискриминантной функции f_3 для третьего подмножества $\bar{f}_{33} = 188,9$.

Для разграничения первого и третьего подмножеств константа дискриминации c_2 равна

$$c_2 = \frac{330,5 + 188,9}{2} = 259,7.$$

Для разграничения второго и третьего подмножеств имеем

$$f_{22} = X_2 \cdot A_3 = \begin{pmatrix} 7,4 & 1,09 \\ 6,7 & 1,23 \\ 6,6 & 1,33 \\ 7,0 & 1,25 \\ 7,5 & 1,15 \end{pmatrix} \times \begin{pmatrix} 12,53 \\ 8,55 \end{pmatrix} = \begin{pmatrix} 102,0 \\ 94,5 \\ 94,1 \\ 98,4 \\ 103,8 \end{pmatrix};$$

- среднее значение дискриминантной функции f_2 для второго подмножества $\bar{f}_{22} = 98,6$

$$f_{23} = X_3 \cdot A_3 = \begin{pmatrix} 5,5 & 0,90 \\ 5,1 & 0,88 \\ 5,4 & 1,20 \\ 5,8 & 1,25 \end{pmatrix} \times \begin{pmatrix} 12,53 \\ 8,55 \end{pmatrix} = \begin{pmatrix} 76,6 \\ 71,4 \\ 77,9 \\ 83,4 \end{pmatrix};$$

- среднее значение дискриминантной функции f_2 для третьего подмножества $\bar{f}_{23} = 77,3$.

Для разграничения второго и третьего подмножеств константа дискриминации c_3 равна:

$$c_3 = \frac{98,6 + 77,3}{2} = 87,9.$$

В общем виде дискриминантные функции для трех подмножеств имеют вид:

$$\begin{cases} f_1 = 14,76X_1 + 13,68X_2 - 144,2, \\ f_2 = 32,24X_1 + 12,44X_2 - 259,9, \\ f_3 = 12,53X_1 + 8,5X_2 - 87,9. \end{cases}$$

Следовательно, границы трех классов будут определяться системой уравнений

$$\begin{cases} f_1 - f_2 = -17,48X_1 + 1,24X_2 + 115,7, \\ f_1 - f_3 = 2,23X_1 + 5,13X_2 - 56,3, \\ f_2 - f_3 = 19,71X_1 + 3,89X_2 - 172,0. \end{cases} \quad (7.5)$$

Чтобы начать классификацию новых объектов, подставим значения исходных переменных в выражение (7.5) и сравним полученные результаты с нулем,

если: $f_1 - f_2 > 0$ и $f_1 - f_3 > 0$ объект принадлежит первому множеству;

$f_1 - f_2 < 0$ и $f_2 - f_3 > 0$ объект принадлежит второму множеству;

$f_1 - f_3 < 0$ и $f_2 - f_3 < 0$ объект принадлежит третьему множеству.

Пример 3. По результатам анализа финансовой устойчивости предприятий экспертным путем были выделены три группы предприятий, характеризующиеся двумя переменными: X_1 — рентабельность основной деятельности предприятия; X_2 — коэффициент абсолютной ликвидности (табл. 7.2).

Таблица 7.2

I группа		II группа		III группа	
X_1	X_2	X_1	X_2	X_1	X_2
10	10	20	10	32	50
9	8	15	12	40	62
8,7	9	30	20	43	57
6	7,5	22	15	45	70
7	9,8	31	27	35	60

Окончание табл. 7.2

I группа		II группа		III группа	
X_1	X_2	X_1	X_2	X_1	X_2
		35	30	31	54
		40	18	41	58
				42	61
				38	67

Примечание.

$$X_1 = \frac{\text{Прибыль от реализации}}{\text{Затраты на поизводство продукции}} \times 100;$$

$$X_2 = \frac{\text{Денежные средства + Краткосрочные вложения}}{\text{Краткосрочные обязательства}} \times 100.$$

Решение.

1. Для каждой из групп рассчитаем вектор средних значений и матрицу ковариаций:

- для I группы $\overline{X} = (8{,}14 \quad 8{,}86)$;

$$S_1 = \begin{pmatrix} 2{,}0784 & -0{,}322 \\ -0{,}322 & 0{,}96 \end{pmatrix} \quad S_1^{-1} = \begin{pmatrix} 0{,}5075 & 0{,}1702 \\ 0{,}1702 & 1{,}0988 \end{pmatrix};$$

- для II группы $\overline{X} = (27{,}57 \quad 18{,}86)$;

$$S_2 = \begin{pmatrix} 67{,}67 & 30{,}41 \\ 30{,}41 & 47{,}55 \end{pmatrix} \quad S_2^{-1} = \begin{pmatrix} 0{,}0207 & -0{,}0133 \\ -0{,}0133 & 0{,}0295 \end{pmatrix};$$

- для III группы $\overline{X} = (39{,}0 \quad 59{,}9)$;

$$S_3 = \begin{pmatrix} 21{,}778 & 17{,}733 \\ 17{,}733 & 33{,}654 \end{pmatrix} \quad S_3^{-1} = \begin{pmatrix} 0{,}0804 & -0{,}0424 \\ -0{,}0424 & 0{,}0520 \end{pmatrix}.$$

2. Рассчитаем три дискриминантные функции:

I группа

Вектор коэффициентов дискриминантной функции:

$$A_1 = (8{,}14 \quad 8{,}86) \times \begin{pmatrix} 0{,}508 & 0{,}170 \\ 0{,}170 & 1{,}099 \end{pmatrix} = (5{,}64 \quad 11{,}12);$$

свободный член

$$a_0 = -\frac{1}{2}(8{,}14 \quad 8{,}86) \times \begin{pmatrix} 0{,}508 & 0{,}170 \\ 0{,}170 & 1{,}099 \end{pmatrix} \times \begin{pmatrix} 8{,}14 \\ 8{,}86 \end{pmatrix} = -72{,}216.$$

Дискриминантная функция будет иметь следующий вид:

$$f_1 = -72{,}216 + 5{,}64X_1 + 11{,}12X_2$$

II группа

Вектор коэффициентов дискриминантной функции

$$A_2 = (27{,}57 \quad 18{,}86) \times \begin{pmatrix} 0{,}021 & -0{,}013 \\ -0{,}013 & 0{,}029 \end{pmatrix} = (0{,}3199 \quad 0{,}1897);$$

свободный член

$$a_0 = -\frac{1}{2}(27{,}57 \quad 18{,}86) \times \begin{pmatrix} 0{,}021 & -0{,}013 \\ -0{,}013 & 0{,}029 \end{pmatrix} \times \begin{pmatrix} 27{,}57 \\ 18{,}86 \end{pmatrix} = -6{,}199.$$

Дискриминантная функция будет иметь следующий вид:

$$f_2 = -6{,}199 + 27{,}57X_1 + 18{,}86X_2.$$

III группа

Вектор коэффициентов дискриминантной функции

$$A_3 = (39 \quad 59{,}9) \times \begin{pmatrix} 0{,}0804 & -0{,}0424 \\ -0{,}0424 & 0{,}0520 \end{pmatrix} = (0{,}596 \quad 1{,}461);$$

свободный член

$$a_0 = -\frac{1}{2}(39 \quad 59{,}9) \times \begin{bmatrix} 0{,}0804 & -0{,}0424 \\ -0{,}0424 & 0{,}0520 \end{bmatrix} \times \begin{bmatrix} 39 \\ 59{,}9 \end{bmatrix} = -55{,}379.$$

Дискриминантная функция будет иметь следующий вид:

$$f_3 = -55{,}379 + 0{,}596X_1 + 1{,}461X_2.$$

Итак, получены три дискриминантные функции:

$$f_1 = -72{,}216 + 5{,}64X_1 + 11{,}12X_2;$$

$$f_2 = -6{,}199 + 27{,}57X_1 + 18{,}86X_2;$$

$$f_3 = -55{,}379 + 0{,}596X_1 + 1{,}461X_2.$$

Теперь можно начинать процедуру классификации новых объектов. Предположим, что рассматривается предприятие, которое имеет следующие значения дискриминантных переменных: $X_1 = 0{,}30$, $X_2 = 48$.

Подставляем поочередно эти значения в каждую из дискриминантных функций:

$$f_1 = -72{,}216 + 5{,}64 \cdot 0{,}30 + 11{,}12 \cdot 48 = 463{,}236;$$

$$f_2 = -6{,}199 + 27{,}57 \cdot 0{,}30 + 18{,}86 \cdot 48 = 907{,}352;$$

$$f_3 = -55{,}379 + 0{,}596 \cdot 0{,}30 + 1{,}461 \cdot 48 = 14{,}928.$$

Так как $f_2 > f_1$ и $f_2 > f_3$, новый объект следует отнести ко второй группе предприятий.

7.3. Проведение дискриминантного анализа на компьютере

По двадцати четырем крупнейшим банкам Москвы имеются следующие данные о результатах деятельности на конец 2000 г. (X_1 — достаточность капитала, %; X_2 — рентабельность активов, %; X_3 — коэффициент ликвидности, %) (табл. 7.3).[1]

Таблица 7.3

Банк	X_1	X_2	X_3
Альфа-банк	64	0	68
Внешторгбанк	93	3	83
Газпромбанк	44	1	37
Международный промышленный банк	65	1	40
Сбербанк России	17	4	38
БИН	37	5	32
Гута-банк	30	1	19
Доверительный и инвестиционный банк	17	3	39
«Еврофинанс»	35	5	52
Конверсбанк	41	1	42
МДМ-банк	35	1	47
НОМОС-банк	40	1	36
Росбанк	24	8	28
Собинбанк	54	1	25
Автобанк	20	1	29
«Зенит»	27	5	17
Банк Москвы	9	1	23
«Глобэкс»	82	0	90
Московский индустриальный банк	36	3	27
Национальный резервный банк	44	1	39
«Петрокоммерц»	30	6	42
Транскредитбанк	41	9	40
«Возрождение»	19	0	20
Международный московский банк	10	0	10

1. Используя пакет STATISTICA, метод иерархического агломеративного кластерного анализа, проведем классификацию банков. Определим три группы банков, которые могут быть использованы в качестве обучающих выборок в дискриминантном анализе.

2. Рассчитаем значения дискриминантных функций для каждой группы банков.

3. Определим, к какой из выделенных групп следует отнести банк, у которого переменные принимают значения: $X_1 = 35$ %; $X_2 = 4$ %; $X_3 = 70$ %.

Решение. 1. Сначала при помощи пакета STATISTICA проведем классификацию двадцати четырех банков по методу иерархического агломеративного кластерного анализа (см. параграф 6.3). Для наглядности представим результаты классификации в виде дендрограммы (рис. 7.1).

[1] Эксперт. 2001. № 11.

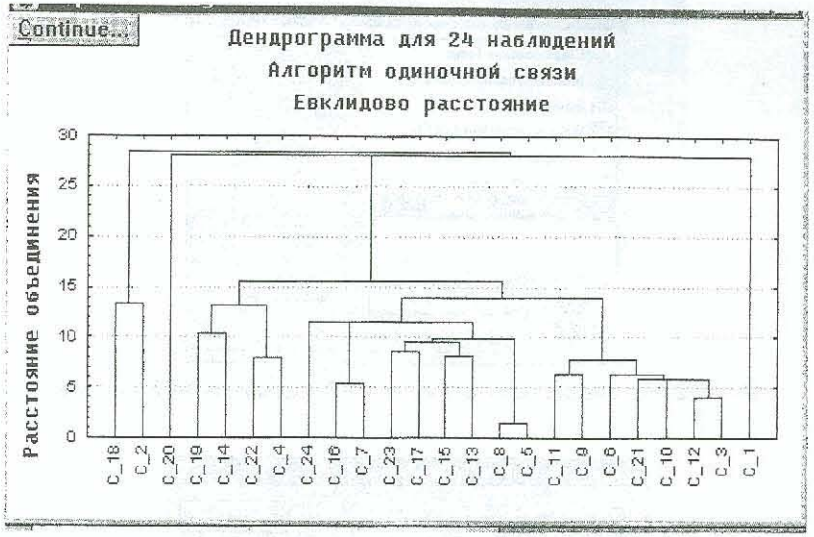


Рис. 7.1. Дендрограмма многомерной классификации двадцати четырех банков Москвы

По результатам классификации в совокупности можно выделить три достаточно плотных кластера (максимальное расстояние при объединении немногим более 15), которые в дальнейшем послужат нам в качестве обучающих выборок:

I кластер — объекты 11, 9, 6, 21, 10, 12, 3;

II кластер — объекты 24, 16, 7, 23, 17, 15, 13, 8, 5;

III кластер — объекты 19, 14, 22, 4.

Остальные объекты (18, 2, 20,1) относятся к перечисленным кластерам уже на достаточно большом расстоянии (более 28), поэтому их можно рассматривать как самостоятельные кластеры и впоследствии не включать в состав обучающих выборок.

Для продолжения анализа введем в электронную таблицу еще одну переменную (group), в которой укажем номер группы (кластера) для каждого наблюдения. Выберем модуль **Discriminant Analysis** и щелкнем на кнопке **Switch To** (рис. 7.2).

В раскрывшемся окне (рис. 7.3) для проведения пошагового анализа дискриминантных функций необходимо задать следующие параметры:

Variables — переменные, участвующие в анализе, в том числе: **Grouping** — классификационные признаки; **Independent** — независимые (дискриминантные) переменные; **Codes for grouping variables** — значения классификационных признаков, которые будут участвовать в образовании групп.

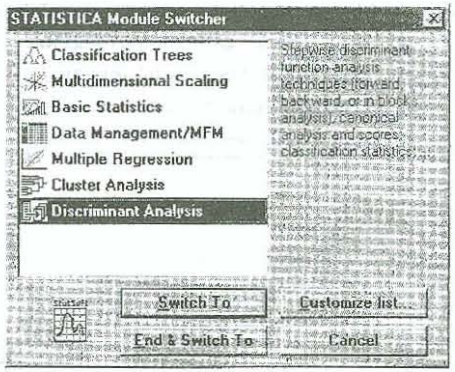


Рис. 7.2. Включение рабочего модуля «Дискриминантный анализ»

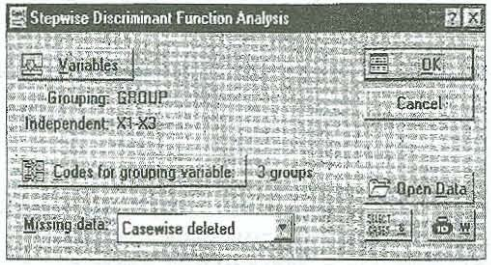


Рис. 7.3. Окно выбора параметров дискриминантного анализа

В нашем примере это будут значения 1, 2, 3, соответствующие трем выделенным кластерам.

После задания всех необходимых параметров, щелкнем на кнопке **OK**. Программа выдаст окно для задания вида дискриминантной функции. В этом окне (рис. 7.4) необходимо еще раз уточнить набор дискриминантных переменных (в нашем примере это X_1, X_2, X_3) и выбрать метод отбора дискриминантных переменных.

В системе STATISTICA реализованы следующие методы дискриминации:

Standard — стандартный метод, который предполагает использование всех дискриминантных переменных, первоначально указанных пользователем (см. рис. 7.3), независимо от уровня их информативности.

Forward stepwise — прямая процедура пошагового отбора переменных, начиная с переменной, обеспечивающей наилучшее различение множеств. На каждом шаге этого алгоритма отбирается очередная переменная, которая в сочетании с ранее отобранными, дает наилучшее различение групп.

Backward stepwise — обратная процедура пошагового отбора дискриминантных переменных, когда на первом шаге алгоритма все переменные включаются в дискриминантную функцию. На каждом последующем шаге происходит исключение из набора той переменной, которая вносит наименьший вклад в различение множеств.

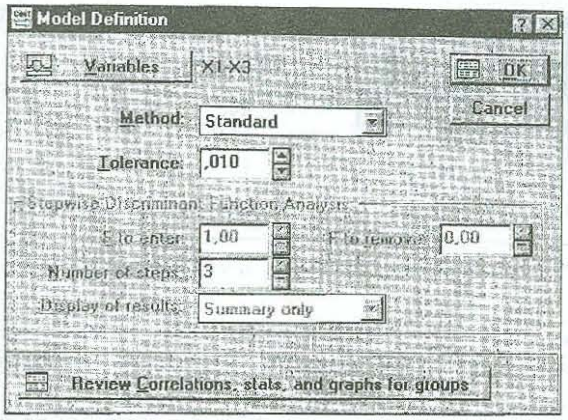


Рис. 7.4. Окно для определения методов построения модели

В нашем примере был выбран стандартный метод, т.е. все переменные (X_1, X_2, X_3) включены в состав дискриминантной функции.

На следующем шаге проведения дискриминантного анализа нам необходимо оценить коэффициенты дискриминантных функций. Для этого в окне на рис. 7.5 щелкнем на кнопку **Classification Functions** (классификационные, т.е. дискриминантные функции).

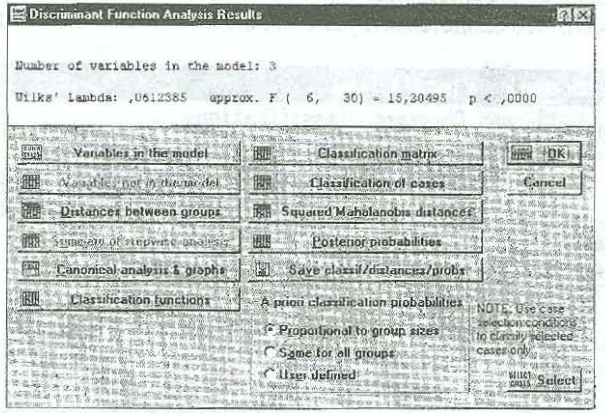


Рис. 7.5. Окно для проведения анализа дискриминантных функций

Полученные результаты (рис. 7.6) позволяют нам записать выражения всех трех дискриминантных функций:

$$f_1 = -53{,}0377 + 1{,}5136\,X_1 - 1{,}0414\,X_2 + 1{,}1489\,X_3,$$

$$f_2 = -17{,}6576 + 0{,}8489\,X_1 - 0{,}4749\,X_2 + 0{,}6561\,X_3,$$

$$f_3 = -83{,}6792 + 2{,}0651\,X_1 - 1{,}1156\,X_2 + 1{,}2123\,X_3.$$

Classification Functions; grouping: GROUP (мса(73~1.sta)

Continue...	G_1:1 p=,35000	G_2:2 p=,45000	G_3:3 p=,20000
X1	1,5136	,8489	2,0651
X2	-1,0414	-,4749	-1,1156
X3	1,1489	,6561	1,2123
Constant	-53,0377	-17,6576	-83,6792

Рис. 7.6. Расчетные значения коэффициенты дискриминантных функций

Это так называемые линейные дискриминантные функции Фишера, с которыми подробно можно познакомиться в специальной литературе [12, 25].

На основе рассчитанных классификационных функций по определенному правилу производится повторная классификация объектов всех трех подмножеств [12]. Чтобы увидеть результаты этой процедуры, нужно в окне (см. рис. 7.5) щелкнуть по кнопке **Classification Matrix**.

На рис. 7.7 мы видим, что произошли изменения в первоначальном составе подмножеств. Например, с учетом применения классификационных функций в первую группу отнесены три объекта, хотя первоначально в ней находились четыре объекта. Следовательно, процент корректной классификации составляет 75 % (3/4×100). Аналогично трактуются и другие результаты. Средний по всем группам процент корректной классификации составил около 84,2 % (последняя строка таблицы). Это свидетельствует о хорошем качестве классификации.

Classification Matrix (мса(73~1.sta)

Continue... Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	G_1:1 p=,21053	G_2:2 p=,47368	G_3:3 p=,10526	G_4:4 p=,21053
G_1:1	75,0000	3	1	0	0
G_2:2	88,8889	1	8	0	0
G_3:3	100,0000	0	0	2	0
G_4:4	75,0000	1	0	0	3
Total	84,2105	5	9	2	3

Рис. 7.7. Результаты применения классификационных функций

Чтобы получить полную картину классификации, т.е. узнать какие объекты и на каком основании были отнесены к соответствующему множеству, нужно в окне (см. рис. 7.5) выбрать процедуру **Classification of cases.** В раскрывшемся окне (рис. 7.8) мы видим детальную картину классификации. Поясним некоторые ее фрагменты. В первой графе таблицы указаны номера объектов анализируемой совокупности; во второй графе приведены номера тех групп, к которым мы первоначально причислили изучаемые объекты (прочерками отмечены четыре объекта, которые по результатам кластерного анализа мы не отнесли ни к одной из выделенных групп). Далее рассмотрим ситуацию с третьим объектом. На первом шаге мы причислили его к первому подмножеству, а по результатам вычислений на базе классификационных функций самой высокой оказалась вероятность его принадлежности к третьему подмножеству ($p = 0{,}45$). Именно на основании этой (максимальной) вероятности он и был причислен к нему. Для всех остальных случаев рассуждения аналогичны.

Continue... Incorrect classifications are marked with *

Case	Observed Classif.	1 p=,35000	2 p=,45000	3 p=,20000
1	----	G_3:3	G_1:1	G_2:2
2	----	G_3:3	G_1:1	G_2:2
3	G_1:1	G_1:1	G_3:3	G_2:2
4	G_3:3	G_3:3	G_1:1	G_2:2
5	G_2:2	G_2:2	G_1:1	G_3:3
6	G_1:1	G_1:1	G_2:2	G_3:3
7	G_2:2	G_2:2	G_1:1	G_3:3
8	G_2:2	G_2:2	G_1:1	G_3:3
9	G_1:1	G_1:1	G_3:3	G_2:2
10	G_1:1	G_1:1	G_3:3	G_2:2
11	G_1:1	G_1:1	G_3:3	G_2:2
12	G_1:1	G_1:1	G_3:3	G_2:2
13	G_2:2	G_2:2	G_1:1	G_3:3
14	G_3:3	G_3:3	G_1:1	G_2:2
15	G_2:2	G_2:2	G_1:1	G_3:3
16	G_2:2	G_2:2	G_1:1	G_3:3
17	G_2:2	G_2:2	G_1:1	G_3:3
18	----	G_3:3	G_1:1	G_2:2
19	G_3:3	G_3:3	G_1:1	G_2:2
20	----	G_3:3	G_1:1	G_2:2
21	G_1:1	G_1:1	G_3:3	G_2:2
22	G_3:3	G_3:3	G_1:1	G_2:2
23	G_2:2	G_2:2	G_1:1	G_3:3
24	G_2:2	G_2:2	G_1:1	G_3:3

Рис. 7.8. Результаты классификации объектов трех подмножеств банков на основании классификационных функций (f)

В заключение отметим, что пользователь при желании может получить геометрическую интерпретацию результатов классификации, т.е. на графике увидеть расположение анализируемых объектов в пространстве дискриминантных функций. Для этого нужно в окне (см. рис. 7.5) выбрать процедуру **Canonical analysis & graphs** (канонический анализ и графики) — **Scatter plot of canonical scores** (диаграмма рассеяния значений канонических оценок). На рис. 7.9 показано расположение каждого из объектов трех подмножеств в новом признаковом пространстве — первых двух дискриминантных функций (f_1, f_2). Четыре объекта $(n_1, n_2, n_{18}, n_{20})$, которые после проведения кластерного анализа были исключены нами из дальнейшего исследования, образуют самостоятельную группу, достаточно удаленную от объектов трех выделенных подмножеств.

На диаграмме также видно, что первое подмножество образовано семью схожими между собой объектами. В него были отнесены следующие банки: Газпромбанк, БИН, «Еврофинанс», Конверсбанк, МДМ-банк, НОМОС-банк и «Петрокоммерц». Рассеивание точек этого подмножества по значениям координат наименьшее.

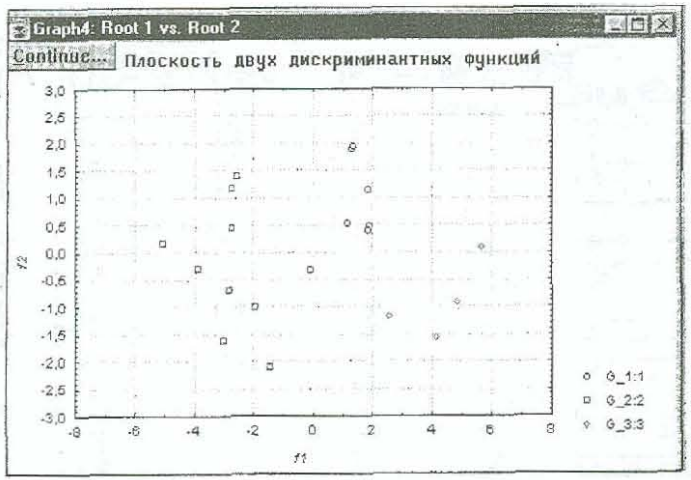


Рис. 7.9. Диаграмма рассеивания двадцати наблюдаемых объектов в координатной системе двух дискриминантных функций.

7.4. Контрольные задания

Задание 1. В ходе проведенного наблюдения были получены следующие данные по двум группам факультетов университета (табл. 7.4).

Таблица 7.4

Группа факультета	Факультет	X_1	X_2	X_3
Первая группа	1	0,30	0,55	0,12
	2	0,29	0,48	0,20
	3	0,33	0,52	0,10
Вторая группа	4	0,40	0,58	0,02
	5	0,41	0,55	0,03
	6	0,42	0,56	0,01
	7	0,41	0,55	0,02

Здесь: X_1 — доля студентов, успевающих на «отлично»; X_2 — доля студентов, успевающих на «хорошо» и «отлично»; X_3 — доля студентов, успевающих на «хорошо» и «удовлетворительно»:

Проведите распределение следующих факультетов по имеющимся группам:

8-й факультет $X_1 = 0,32$, $X_2 = 0,49$, $X_3 = 0,15$;

9-й факультет $X_1 = 0,41$, $X_2 = 0,57$, $X_3 = 0,01$;

10-й факультет $X_1 = 0,31$, $X_2 = 0,54$, $X_3 = 0,14$.

Все необходимые расчеты выполните на компьютере и постройте график распределения объектов в пространстве дискриминантных функций.

Задание 2. По группе сельскохозяйственных предприятий известны следующие результаты производственно-хозяйственной деятельности (табл. 7.5).

Таблица 7.5

Номер хозяйства	X_1	X_2	X_3	X_4	X_5
1	1669	357	649	324	1095
2	2501	671	448	1032	2344
3	3959	980	2549	1197	2147
4	2505	618	835	891	695
5	2117	745	584	620	2211
6	6387	759	330	1105	2642
7	7344	360	684	2135	3263
8	2067	1100	2327	2162	3085
9	3599	1430	2498	965	2394
10	2554	806	795	787	2253
11	2395	760	632	957	2276
12	2990	812	1155	1391	794
13	3170	765	1383	743	2724
14	3730	511	723	777	1448

Здесь: X_1 — сумма производственных затрат, тыс. ден. ед.; X_2 — сумма затрат живого труда, тыс. чел.-ч; X_3 — затраты средств из фонда на развитие производства, тыс. ден. ед.; X_4 — использование долгосрочных кредитов; X_5 — размер пашни, га.

При помощи кластерного анализа из первых десяти предприятий сформируйте две обучающие выборки, рассчитайте дискриминантную функцию и константу дискриминации.

Проведите дискриминацию оставшихся четырех предприятий. Поясните полученные результаты.

Задание 3. Используя пакет прикладных программ STATISTICA — модуль **Discriminant Analysis** (дискриминантный анализ), классифицируйте пять промышленных предприятий, характеризующихся переменными: X_1 — среднегодовая численность рабочих, тыс. чел.; X_2 — валовая продукция, млн ден. ед.; X_3 — среднегодовая стоимость основных производственных фондов, млн ден. ед.; X_4 — рентабельность продукции, % (табл. 7.6).

Таблица 7.6

Номер объекта	X_1	X_2	X_3	X_4
1	9,2	14,0	17,5	15,9
2	8,3	13,5	20,0	12,8
3	2,5	5,8	11,6	9,1
4	7,1	11,2	19,4	16,2
5	3,2	6,4	12,0	7,6

Классификацию проведите при помощи дискриминантного анализа, используя в качестве обучающих выборок следующие группы объектов.

Первая группа

Номер объекта	X_1	X_2	X_3	X_4
1	1,8	5,1	10,7	6,3
2	2,0	0,3	11,4	7,0
3	3,5	4,8	10,3	6,9
4	2,7	3,2	13,8	8,7
5	3,0	5,4	12,1	7,4

Вторая группа

Номер объекта	X_1	X_2	X_3	X_4
1	6,9	10,2	17,8	15,0
2	10,1	11,1	18,0	14,1
3	7,7	12,3	19,2	15,2
4	8,6	12,7	16,7	13,3
5	9,4	13,1	16,9	12,7

Распечатайте протоколы работы модуля Discriminant Analysis и поясните полученные результаты классификации.

Задание 4. По четырем отраслям экономики за два периода (первый период — четыре благоприятных года; второй период — пять неблагоприятных лет) известны темпы роста инвестиций в основной капитал, % (табл. 7.7).

Таблица 7.7

Год		Промышленность X_1	Сельское хозяйство X_2	Транспорт X_3	Связь X_4
Первый период	1	110	113	99	89
	2	115	119	104	90
	3	112	110	116	84
	4	119	120	112	87
Второй период	1	84	68	81	56
	2	91	70	65	60
	3	83	71	77	61
	4	87	64	80	59
	5	94	76	81	55

При помощи методов дискриминантного анализа определите, к какому периоду можно отнести годы, в которые рассматриваемые переменные X_j равны:

первый год: $X_1 = 89$; $X_2 = 73$; $X_3 = 68$; $X_4 = 58$;
второй год: $X_1 = 111$; $X_2 = 115$; $X_3 = 110$; $X_4 = 91$.

Задание 5. В соответствии с тремя аналитическими признаками по данным за 1990 г. десять европейских стран распределены на две группы (табл. 7.8).

Таблица 7.8

Первая группа			
	X_1	X_2	X_3
Великобритания	75	9	2,8
Франция	76	7	1,9
Бельгия	74	9	2,0
Германия	75	8	2,1
Швейцария	77	7	1,9
Вторая группа			
	X_1	X_2	X_3
Чешская Республика	72	11	2,5
Польша	71	15	1,3
Венгрия	70	16	2,3
Югославия	71	24	1,0
Румыния	69	27	1,6

Здесь: X_1 — ожидаемая продолжительность жизни, лет; X_2 — уровень младенческой смертности, %; X_3 — коэффициент разводимости, %.

Используя дискриминантный анализ, вычислите дискриминантную функцию (f) и определите принадлежность каждой из названных ниже шести стран к одной из двух выделенных групп.

	X_1	X_2	X_3
Беларусь	71	12	3,7
Литва	71	10	3,4
Латвия	69	14	4,0
Россия	69	17	1,4
Узбекистан	69	35	1,5
Туркмения	66	45	1,4

Расчеты выполните на компьютере, продемонстрируйте их на графике и поясните полученные результаты.

Задание 6. Известны следующие данные по предприятиям одной из отраслей промышленности (табл. 7.9).

Таблица 7.9

Основные производственные фонды, млн ден. ед. (X_1)	Среднесписочное число рабочих, тыс. чел. (X_2)	Произведено продукции за год, млн ден. ед. (X_3)
4,01	4,27	9,38
3,01	4,25	7,03
1,95	2,85	3,90
4,10	4,75	9,87
3,11	3,55	6,33

Окончание табл. 7.9

Основные производственные фонды, млн ден. ед. (X_1)	Среднесписочное число рабочих, тыс. чел. (X_2)	Произведено продукции за год, млн ден. ед. (X_3)
3,27	3,90	6,65
2,10	2,85	3,38
2,99	3,65	5,72
2,16	2,75	3,68
3,10	3,20	4,84
2,04	2,15	3,05
2,54	3,15	5,28
3,40	3,75	6,60
2,50	3,15	5,02
3,15	3,55	5,82

Используя любой алгоритм кластерного анализа, сформируйте из первых десяти наблюдений две обучающие выборки. На основании полученных выборок с помощью дискриминантного анализа проведите классификацию пяти предприятий, не вошедших в обучающие выборки.

Дайте экономическую интерпретацию полученных результатов.

Задание 7. Найдите оценку дискриминантной функции и константу дискриминации, если известны следующие матрицы исходных значений переменных по двум выборкам.

Первая выборка			Вторая выборка		
Номер наблюдения	X_1	X_2	Номер наблюдения	X_1	X_2
1	0,23	2,1	1	0,40	1,5
2	0,25	2,3	2	0,35	1,3
3	0,18	4,0	3	0,55	2,0
4	0,21	3,8	4	0,65	1,7

Здесь: X_1 — средняя трудоемкость единицы продукции, чел.-ч; X_2 — удельный вес бракованной продукции, %.

Покажите на графике дискриминантную функцию $f = a_1 z_1 + a_2 z_2$, а также расположение объектов обеих выборок в двумерном пространстве нормированных значений исходных переменных.

Примечание. Нормирование исходных значений переменных проведите по формуле $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$.

Задание 8. По результатам дискриминантного анализа на основе четырех обучающих выборок были получены следующие оценки параметров дискриминантных функций:

$$f_1 = 12{,}5X_1 + 17X_2 - 48{,}2,$$
$$f_2 = 20{,}5X_1 + 15{,}2X_2 - 65{,}8,$$
$$f_3 = 26{,}0X_1 + 19{,}5X_2 - 102{,}0,$$
$$f_4 = 15{,}8X_1 + 6{,}0X_2 - 32{,}2.$$

Воспользовавшись полученными оценками дискриминантных функций, проведите классификацию новых объектов, у которых анализируемые переменные принимают следующие значения:

Номер объекта	X_1	X_2
1	2,1	5,5
2	5,0	2,8
3	4,3	3,2
4	2,8	2,5

Решение задачи проиллюстрируйте на графике, поясните полученные результаты классификации.

Задание 9. По результатам экспертного оценивания эффективности производства (X_1 — прибыль на одного работника, млн ден. ед.; X_2 — рентабельность производства, %; X_3 — производительность труда, млн ден. ед.) в наблюдаемой статистической совокупности были выделены четыре группы промышленных предприятий:

I группа — высокая эффективность производства;

II группа — средняя эффективность производства;

III группа — низкая эффективность производства;

IV группа — очень низкая эффективность производства.

В ходе проведения дискриминантного анализа для каждой группы получены следующие оценки параметров дискриминантных функций:

Параметр	Номер группы			
	I	II	III	IV
a_0	–295,6	–456,7	–68,7	–40,9
a_1	–2,7	–3,4	–0,84	–0,85
a_2	2,9	3,3	1,4	1,1

Изобразите на графике разделяющие прямые и области, соответствующие каждой из выделенных групп предприятий.

Задание 10. Для двух множеств наблюдаемых объектов (две обучающие выборки) дискриминантная функция имеет вид $f = -30{,}5X_1 + 6{,}04X_2$; векторы средних значений соответственно равны: для первого подмножества $\bar{X} = (0{,}15;\ 2{,}2)$; для второго подмножества $\bar{X} = (0{,}35;\ 3{,}7)$.

Определить, к какому из двух подмножеств следует отнести объекты № 1, 2, 3, 4, если имеются следующие данные:

Номер объекта	1	2	3	4
X_1	0,18	0,11	0,39	0,52
X_2	3,0	2,7	2,9	3,5

Поясните полученные результаты и изобразите на графике расположение их объектов в пространстве исходных переменных.

Задание 11. На основе приведенных ниже данных выполните классификацию пяти промышленных предприятий.

Номер предприятия	X_1	X_2	X_3	X_4
1	8,4	0,62	17,5	81,5
2	9,1	0,78	10,0	94,0
3	5,5	0,73	16,1	74,0
4	4,3	0,65	15,0	70,8
5	9,7	0,70	14,0	92,5

Здесь: X_1 — производительность труда одного работника, млн ден. ед.; X_2 — доля рабочих в общей численности работающих; X_3 — рентабельность продукции, %; X_4 — коэффициент использования сырья и материалов, %.

Классификацию проведите при помощи дискриминантного анализа, используя в качестве обучающих выборок следующие группы предприятий (табл. 7.10).

Таблица 7.10

Номер предприятия	X_1	X_2	X_3	X_4
Первая группа				
1	4,0	0,63	15,0	80,0
2	4,9	0,60	16,0	78,6
3	6,1	0,61	17,0	75,9
4	5,3	0,62	17,5	74,0
5	5,8	0,60	20,0	81,5
Вторая группа				
1	8,7	0,70	19,0	90,7
2	10,3	0,78	20,5	94,6
3	11,6	0,75	22,0	94,0
4	10,8	0,77	21,0	92,5

Поясните полученные результаты дискриминантного анализа, дайте экономическую интерпретацию выявленных различий между группами предприятий на основании расчета выборочных характеристик (средние значения переменных, сумма внутригрупповых дисперсий).

8. МЕТОД КАНОНИЧЕСКИХ КОРРЕЛЯЦИЙ

8.1. Методические рекомендации

Метод канонических корреляций — статистический метод анализа связей между массовыми общественными явлениями и процессами, применяемыми в том случае, когда рассматриваются несколько независимых переменных $X_j \ (j=\overline{1,q})$ и несколько результативных показателей $Y_k \ (k=\overline{1,p})$, т.е. канонический корреляционный анализ можно рассматривать как вариант распространения парной корреляции на случай двух многомерных величин.

Важнейшим достоинством метода канонических корреляций является то, что при его применении не требуется подтверждение отсутствия корреляции как в группе зависимых переменных (Y_k), так и в группе независимых переменных (X_j).

Цель применения метода — поиск максимальных корреляционных связей между факторными и результативными переменными.

Исходные данные в анализе канонических корреляций представляются в следующем виде.

Наблюдаемый объект	X_1	X_2	....	X_q	Y_1	Y_2 ...	Y_p
1	x_{11}	x_{12}	....	x_{1q}	y_{11}	y_{12}...	y_{1p}
2	x_{21}	x_{22}	....	x_{2q}	y_{21}	y_{22}...	y_{2p}
3	x_{31}	x_{32}	....	x_{3q}	y_{31}	y_{32}...	y_{3p}
...	...	...	...	...	...	...	...
n	x_{n1}	x_{n2}	....	x_{nq}	y_{n1}	y_{n2}...	y_{np}

Здесь: X_1 , X_2 ... X_q — независимые переменные (факторные признаки); Y_1, Y_2 ... Y_p — зависимые переменные (результативные признаки).

В ходе канонического корреляционного анализа оценивается теснота связи между новыми каноническими переменными U и V, вычисляемыми по формулам:

$$U = a_1X_1 + a_2X_2 + ... + a_qX_q;$$
$$V = b_1Y_1 + b_2Y_2 + ... + b_pY_p. \qquad (8.1)$$

По аналогии с парной корреляцией теснота связи между каноническими переменными оценивается при помощи канонического коэффициента корреляции r

$$r = \frac{cov(U,V)}{\sqrt{var(U) \times var(V)}}, \qquad (8.2)$$

где $cov(U, V)$ — ковариация канонических переменных U и V; $var(U) = = \sigma_{UU}$, $var(V) = \sigma_{VV}$ — вариации (дисперсии) канонических переменных. Один из возможных вариантов расчета

$$\sigma_{UU} = \frac{\sum_{i=1}^{n}(u_{1i} - \overline{u_1})(u_{1i} - \overline{u_1})}{n} = \frac{\sum_{i=1}^{n}(u_{1i} - \overline{u_1})^2}{n} = var(U); \qquad (8.3)$$

$$\sigma_{UV} = \frac{\sum_{i=1}^{n}(u_{1i} - \overline{u_1})(v_{1i} - \overline{v_1})}{n} = cov(U,V). \qquad (8.4)$$

В зависимости от того, какие значения принимают коэффициенты a_i и b_j $\left(j = \overline{1, p};\ \ k = \overline{1, q}\right)$, в выражении (8.1) будут изменяться значения канонических переменных и канонический коэффициент корреляции.

Одна из основных задач анализа канонических корреляций заключается в том, чтобы найти такую пару значений канонических переменных, которой будет соответствовать максимальный канонический коэффициент корреляции.

Для вычисления канонических коэффициентов корреляции необходимо прежде всего определить матрицы ковариаций исходных переменных. Запишем расширенную матрицу ковариаций для обеих групп переменных

$$S = \begin{pmatrix} \sigma_{x1x1} & \sigma_{x1x2} & \dots & \sigma_{x1xq} & \sigma_{x1y1} & \dots \sigma_{x1yp} \\ \sigma_{x2x1} & \sigma_{x2x2} & \dots & \sigma_{x2xq} & \sigma_{x2y1} & \dots \sigma_{x2yp} \\ \sigma_{x3x1} & \sigma_{x3x3} & \dots & \sigma_{x3xq} & \sigma_{x3y1} & \dots \sigma_{x3yp} \\ \sigma_{y1x1} & \sigma_{y1x2} & \dots & \sigma_{y1xq} & \sigma_{y1y1} & \dots \sigma_{y1yp} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{ypx1} & \sigma_{ypx2} & \dots & \sigma_{ypxq} & \sigma_{ypy1} & \dots \sigma_{ypyp} \end{pmatrix}.$$

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 факторных переменных $X_{12}, \dots X_q$ размерности $(q \times q)$; S_{22} — ковариационная матрица исходных результативных переменных $(Y_1, Y_2, \ Y_p)$ размерности $(p \times p)$; S_{12}, S_{21} — ковариационные матрицы исходных переменных $(X_1, X_2 \dots X_q)$ и $(Y_1, Y_2 \dots Y_p)$ соответственно, размерности $(q \times p)$ и $(p \times q)$.

В матрицах S_{11} и S_{22} элементы, расположенные на главной диагонали, являются дисперсиями соответствующих переменных. Все остальные элементы матрицы S представляют собой значения ковариаций пар переменных. Представим выражение (8.1) в матричном виде

$$U = XA, \quad V = YB,$$

где U, V — векторы значений канонических переменных; X, Y — матрицы исходных значений переменных; A, B — векторы коэффициентов канонических переменных.

С учетом такой записи векторов значений канонических переменных формулу для расчета канонических коэффициентов корреляции можно записать следующим образом:

$$r = \frac{cov(U,V)}{\sqrt{var(U) \times var(V)}} = \frac{A'S_{12}B}{\sqrt{A'S_{11}AB'S_{22}B}}. \qquad (8.5)$$

Для упрощения процедуры вычисления оптимальных коэффициентов канонических переменных предположим, что каждая из этих переменных имеет единичную дисперсию и нулевое математическое ожидание. Тогда можем записать

$$var(XA) = A'S_{11}A = 1 \quad \text{и} \quad var(YB) = B'S_{22}B = 1,$$

откуда $r = A'S_{12}B$.

Максимальный канонический коэффициент корреляции рассчитывается по формуле

$$A = \frac{S_{11}^{-1}S_{12}B}{\lambda}. \qquad (8.6)$$

Для того чтобы найти компоненты вектора A, необходимо определить векторы B и λ. Значения λ^2 находим как собственные числа матрицы $C = S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$. Размерность этой матрицы равна $(p \times p)(p \times q)(q \times q)(q \times p)$, следовательно, можно вычислить собственные числа (λ_j^2) и соответствующие им собственные векторы B_j. Подставляя поочередно полученные значения λ_j и B_j в формулу (8.6), вычислим вектор A_j и соответствующий канонический коэффициент корреляции r_j.

Канонические коэффициенты корреляции можно рассчитывать и исходя из выборочной матрицы корреляционной (R)

$$R = \left(\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right). \qquad (8.7)$$

При этом все рассуждения и выкладки аналогичны тем, которые приведены выше для ковариационной матрицы. Но значения коэффициентов канонических переменных a_i $(i = \overline{1, q}$ и b_j $j = \overline{1, p})$ будут отличаться.

Если коэффициенты канонических переменных были вычислены на основании ковариационной матрицы S, то они относятся к исходным переменным — X_i и Y_j. Если расчет этих коэффициентов осуществлялся на основе корреляционной матрицы R, то они относятся к стандартизованным значениям исходных переменных: $x'_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_j}$; $y'_{ij} = \frac{y_{ij} - \overline{y}_j}{\sigma_j}$.

Для проверки значимости коэффициентов канонической корреляции используется критерий Бартлетта.

Проверить нулевую гипотезу о том, что множество переменных $X_1, X_2, \ldots X_q$ не коррелирует с множеством переменных $Y_1, Y_2, \ldots, Y_p$, можно при помощи χ^2-критерия

$$\chi_p^2 = -\left[n-1-\frac{1}{2}(p+q+1)\right]\ln W_0, \qquad (8.8)$$

где $W_0 = \prod_{j=1}^{p}\left(1-\lambda_j^2\right)$.

Если χ^2 расчетное больше табличного значения критерия при выбранном уровне значимости и с числом степеней свободы равным $(p \times q)$, то можно утверждать, что, по крайней мере, первый канонический коэффициент корреляции $r_1 = \sqrt{\lambda_1^2}$ будет отличен от нуля.

Чтобы проверить значимость второго коэффициента канонической корреляции, необходимо рассчитать χ^2 по формуле

$$\chi_p^2 = -\left[n-2-\frac{1}{2}(p+q+1)\right]\ln W_1,$$

где $W_1 = \prod_{j=2}^{p}\left(1-\lambda_j^2\right)$.

В этом случае, если χ^2 расчетное больше табличного значения критерия при выбранном уровне значимости и с числом степеней свободы равным $(p-1)(q-1)$, то можно утверждать, что второй канонический коэффициент корреляции $r_2 = \sqrt{\lambda_2^2}$, также будет отличен от нуля. Значимость оставшихся коэффициентов проверяется аналогично.

8.2. Пример решения типовой задачи

Пример. На основании приведенных данных табл. 8.1 вычислим канонические коэффициенты корреляции для двух групп переменных.

Таблица 8.1

Номер предприятия	X_1	X_2	X_3	Y_1	Y_2
1	0,25	175	2560	10,1	45
2	0,23	105	2100	8,6	50
3	0,31	120	1865	9,5	68
4	0,28	112	1640	9,0	90
5	0,34	95	1950	7,6	70
6	0,17	75	2100	11,5	65
7	0,11	110	2350	12,0	85
8	0,26	130	1645	6,8	94
9	0,35	85	2125	8,5	76
10	0,28	103	1480	9,4	50

Здесь: Y_1 — производительность труда, млн ден. ед./чел.; Y_2 — процент сертифицированной продукции; X_1 — трудоемкость единицы продукции, чел.-ч; X_2 — оборачиваемость оборотных средств, дн.; X_3 — фонд оплаты труда работающих, млн ден. ед.

Решение. Для определения матрицы парных корреляций рассчитаем сначала средние значения исходных переменных

$$\overline{X}_1 = 0{,}258; \quad \overline{X}_2 = 111; \quad \overline{X}_3 = 1981{,}5; \quad \overline{Y}_1 = 7{,}2; \quad \overline{Y}_2 = 69{,}3.$$

Матрица парных коэффициентов корреляции для обеих групп равна

$$R = \left(\begin{array}{ccc|cc} 1 & -0{,}021 & -0{,}393 & -0{,}437 & -0{,}046 \\ -0{,}021 & 1 & 0{,}296 & -0{,}037 & -0{,}199 \\ -0{,}393 & 0{,}296 & 1 & 0{,}296 & -0{,}300 \\ \hline -0{,}437 & -0{,}037 & 0{,}296 & 1 & -0{,}117 \\ -0{,}046 & -0{,}199 & -0{,}300 & -0{,}117 & 1 \end{array}\right).$$

Вспомогательные матрицы (R_{11}^{-1} и R_{22}^{-1}) соответственно равны

$$R_{11}^{-1} = \begin{pmatrix} 1{,}197 & -0{,}125 & 0{,}507 \\ -0{,}125 & 1{,}109 & -0{,}377 \\ 0{,}507 & -0{,}377 & 1{,}311 \end{pmatrix}, \quad R_{22}^{-1} = \begin{pmatrix} 1{,}014 & 0{,}119 \\ 0{,}119 & 1{,}014 \end{pmatrix}.$$

На следующем шаге алгоритма вычислим собственные числа и собственные векторы матрицы $C = R_{22}^{-1} \times R_{21} \times R_{11}^{-1} \times R_{12}$. Ее размерность в нашем примере будет (2×2), следовательно, она имеет два собственных числа и два собственных вектора

$$C = R_{22}^{-1} \times R_{21} \times R_{12}^{-1} \times R_{12} = \begin{pmatrix} 0{,}2189 & -0{,}0023 \\ 0{,}0080 & 0{,}1308 \end{pmatrix}.$$

Для нахождения собственных чисел составим характеристический многочлен $\left|C-\lambda^2 E\right| = 0$ и вычислим его корни

$$\lambda^4 - 0{,}3498\lambda^2 + 0{,}0286 = 0.$$

Решая данное уравнение, получим два корня

$$\lambda_1^2 = 0{,}2195 \text{ и } \lambda_2^2 = 0{,}1303.$$

Далее для каждого λ_j^2 рассчитаем собственный вектор B_j. Для этого воспользуемся выражением $\left(C-\lambda^2 E\right)B = 0$.

Вычислим первый собственный вектор (B_1) для $\lambda_1^2 = 0{,}2195$

$$\begin{cases} \left(0{,}2189-\lambda_1^2\right)b_1 - 0{,}0023b_2 = 0, \\ 0{,}0080b_1 + \left(0{,}1308-\lambda_1^2\right)b_2 = 0, \end{cases} \qquad \begin{cases} -0{,}00052b_1 - 0{,}0023b_2 = 0, \\ 0{,}00801b_1 - 0{,}08867b_2 = 0. \end{cases}$$

Для получения нетривиальных решений системы уравнений зададим $b_2 = 1$, тогда $b_1 = 11{,}07$.

Вычислим второй собственный вектор (B_2) для $\lambda_2^2 = 0{,}1303$

$$\begin{cases} (0{,}2189 - \lambda_2^2) b_1 - 0{,}0023 b_2 = 0, \\ 0{,}0080 b_1 + (0{,}1308 - \lambda_2^2) b_2 = 0, \end{cases} \qquad \begin{cases} 0{,}0887 b_1 - 0{,}0023 b_2 = 0, \\ 0{,}00801 b_1 - 0{,}00053 b_2 = 0. \end{cases}$$

Аналогично зададим $b_2 = 1{,}000$ тогда $b_1 = 0{,}026$.

Занесем рассчитанные собственные числа и собственные (характеристические) векторы матрицы C в следующую таблицу:

λ_j^2	Характеристический вектор (B_j)	
$\lambda_1^2 = 0{,}2195$	11,07	1,0
$\lambda_2^2 = 0{,}1303$	0,026	1,0

Канонические коэффициенты корреляции равны:

$$r_1 = \sqrt{\lambda_1^2} = \sqrt{0{,}2195} = 0{,}469; \qquad r_2 = \sqrt{\lambda_2^2} = \sqrt{0{,}1303} = 0{,}361.$$

Чтобы найти векторы коэффициентов A_j, подставляем соответствующие значения λ_j и B_j в выражение (8.6). Первый вектор (A_1), соответствующий первому каноническому коэффициенту (r_1), равен

$$A_1 = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix} = \frac{1}{0{,}2195} \begin{pmatrix} 1{,}197 & -0{,}125 & 0{,}507 \\ -0{,}125 & 1{,}109 & -0{,}377 \\ 0{,}507 & -0{,}377 & 1{,}311 \end{pmatrix} \times \begin{pmatrix} -0{,}437 & -0{,}046 \\ -0{,}037 & -0{,}199 \\ 0{,}296 & -0{,}300 \end{pmatrix} \times \begin{pmatrix} 11{,}07 \\ 1{,}00 \end{pmatrix},$$

$$A_1 = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix} = \begin{pmatrix} -4{,}259 \\ -1{,}187 \\ 1{,}656 \end{pmatrix}.$$

Аналогичным образом вычислим второй вектор A_2

$$A_2 = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix} = \frac{1}{0{,}1303} \begin{pmatrix} 1{,}197 & -0{,}125 & 0{,}507 \\ -0{,}125 & 1{,}109 & -0{,}377 \\ 0{,}507 & -0{,}377 & 1{,}311 \end{pmatrix} \times \begin{pmatrix} -0{,}437 & -0{,}046 \\ -0{,}037 & -0{,}199 \\ 0{,}296 & -0{,}300 \end{pmatrix} \times \begin{pmatrix} 0{,}026 \\ 1{,}000 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} a_{21} \\ a_{22} \\ a_{23} \end{pmatrix} = \begin{pmatrix} 0{,}014 \\ -0{,}209 \\ -0{,}674 \end{pmatrix}.$$

Итак, максимальный коэффициент канонической корреляции равен 0,469. Ему соответствует пара канонических переменных:

$$U_1 = -4{,}259 X_1' - 1{,}187 X_2' + 1{,}656 X_3',$$

$$V_1 = 11{,}07 Y_1' + Y_2'.$$

Второму коэффициенту канонической корреляции $r_2 = 0{,}361$ соответствует пара канонических переменных:

$$U_1' = 0{,}014 X_1' - 0{,}209 X_2' - 0{,}674 X_3',$$

$$V_1' = 0{,}026 Y_1' + Y_2'.$$

Так как векторы коэффициентов канонических переменных вычислены на основании матрицы парных корреляций (R), они будут относиться к стандартизованным значениям исходных переменных X_j' и Y_j'.

8.3. Проведение канонического анализа на компьютере

Динамика макроэкономических показателей в Республике Беларусь по сравнению с базисным характеризуется следующими индексами (табл. 8.2).

Таблица 8.2

Месяц	Индекс, %				
	Производство			Потребительские цены, %	Вклады населения, %
	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

1. Рассчитаем первый канонический коэффициент корреляции и проверим его значимость с помощью χ^2-критерия Бартлетта.

2. Оценим коэффициенты канонических переменных для стандартизованных значений исходных переменных X_1, X_2, X_3, и Y_1, Y_2.

3. На основании результатов расчетов, выполненных в пп. 1—2, определим, какая из линейных комбинаций индексов позволяет наилучшим образом предсказать сводный индекс результативных показателей (изменение потребительских цен и вкладов населения). Все необходимые расчеты выполним на компьютере с использованием пакета программ STATISTICA.

Решение. В системе STATISTICA метод канонических корреляций реализуется при помощи модуля **Canonocal Analis**. Введем значения всех исходных переменных в электронную таблицу и сохраним эти данные в файле с именем Primer. Откроем окно выбора модулей и выберем модуль **Canonical Analysis** (рис. 8.1).

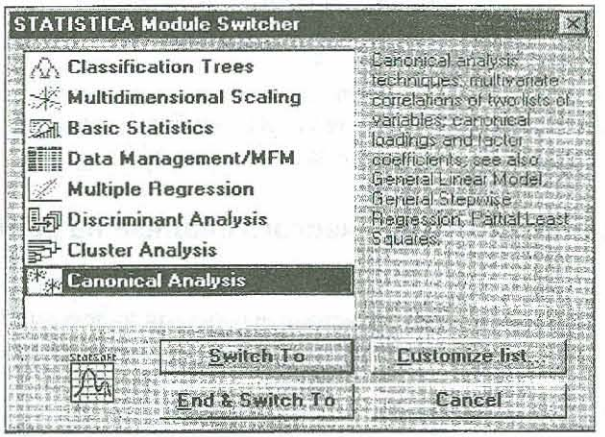


Рис. 8.1. Окно выбора аналитического модуля

После выбора модуля на экране появится окно (рис. 8.2), в котором необходимо указать имена переменных, принадлежащих первому (First List) и второму (Second List) множеству анализируемых признаков.

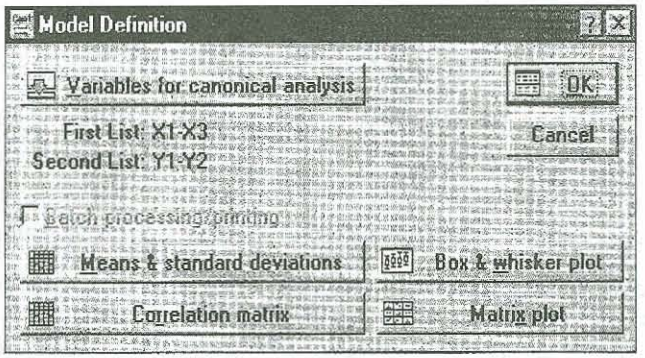


Рис. 8.2. Окно диалога для ввода переменных двух множеств

В нашем примере к первому множеству (независимые переменные) относятся переменные X_1, X_2, X_3, а ко второму множеству (зависимые переменные) — Y_1 и Y_2. Нажимаем кнопку **OK**, чтобы продолжить анализ или предварительно просмотрим матрицу парных корреляций для исходных переменных (рис. 8.3).

Correlations (primer.sta)

Continue...	X1	X2	X3	Y1	Y2
X1	1,00	,32	-,29	,07	-,24
X2	,32	1,00	,23	,22	,17
X3	-,29	,23	1,00	-,21	,39
Y1	,07	,22	-,21	1,00	-,12
Y2	-,24	,17	,39	-,12	1,00

Рис. 8.3. Матрица парных корреляций, *R*

В окне основных процедур канонического анализа (рис. 8.4) видны первые результаты: максимальный канонический коэффициент корреляции ($R = 0{,}445$) и его оценка по χ^2-критерию ($\chi^2 = 2{,}829$).

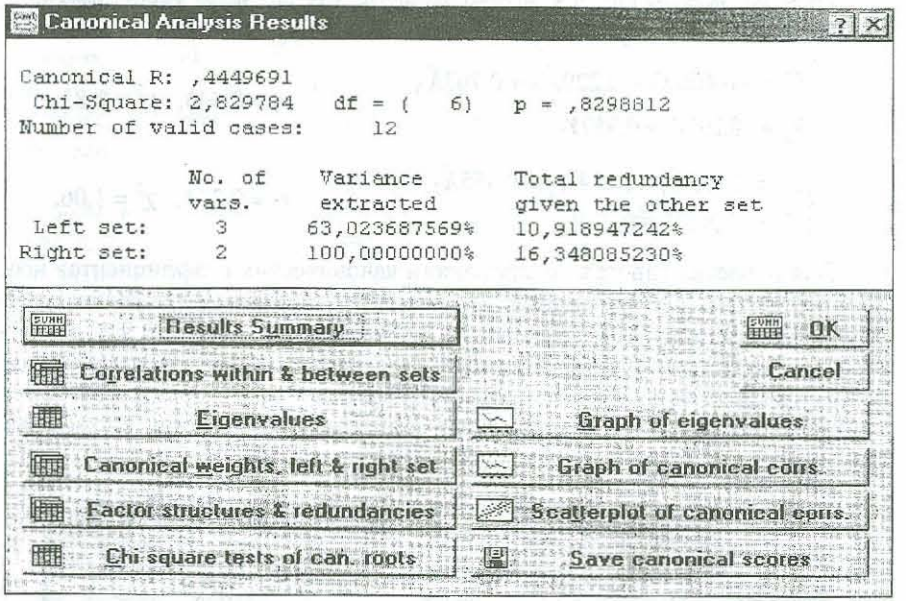
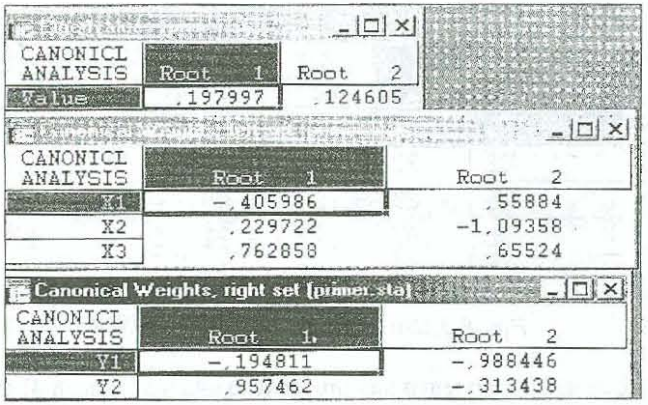


Рис. 8.4. Окно основных процедур канонического анализа

Для того чтобы продолжить анализ, последовательно выберем, например, процедуры **Eigenvalues** (собственные значения), **Canonical weights, left & right set** (канонические веса для левого и правого множеств).

В развернувшемся окне (рис. 8.5) мы видим два собственных числа матрицы C ($C = R_{22}^{-1} \times R_{21} \times R_{11}^{-1} \times R_{12}$): $\lambda_1^2 = 0{,}198$ и $\lambda_2^2 = 0{,}125$. Кроме того, показаны два варианта (Root 1 и Root 2) канонических весовых коэффициентов для переменных обоих множеств.



CANONICL ANALYSIS	Root 1	Root 2
Value	,197997	,124605

CANONICL ANALYSIS	Root 1	Root 2
X1	-,405986	,55884
X2	,229722	-1,09358
X3	,762858	,65524

Canonical Weights, right set (primer.sta)

CANONICL ANALYSIS	Root 1	Root 2
Y1	-,194811	-,988446
Y2	,957462	-,313438

Рис. 8.5. Собственные числа и коэффициенты канонических переменных

Судя по данным рис. 8.5, можем записать, как выглядят канонические переменные и соответствующие им канонические коэффициенты корреляции

$$\begin{cases} U_1 = -0{,}406\tilde{X}_1 + 0{,}229\tilde{X}_2 + 0{,}763\tilde{X}_3, \\ V_1 = -0{,}195\tilde{Y}_1 + 0{,}957\tilde{Y}_2; \end{cases} \qquad r_1 = 0{,}445, \quad \chi^2 = 2{,}83;$$

$$\begin{cases} U_2 = 0{,}559\tilde{X}_1 - 1{,}094\tilde{X}_2 + 0{,}655\tilde{X}_3, \\ V_1 = -0{,}988\tilde{Y}_1 - 0{,}313\tilde{Y}_2; \end{cases} \qquad r_2 = 0{,}354, \quad \chi^2 = 1{,}06.$$

Для проверки гипотезы о значимости канонических коэффициентов необходимо сравнить расчетные значения (рис. 8.6) с табличным для уровня значимости $\alpha = 0{,}01$ $\chi^2_{кр} = 16{,}80$ для шести степеней свободы и $\chi^2_{кр} = 9{,}2$ для двух степеней свободы, т.е. проверяемая гипотеза о равенстве канонического коэффициента корреляции нулю принимается. Следовательно, первый и второй канонические коэффициенты корреляции незначимы. Связь между множествами переменных (Y_1, Y_2) и (X_1, X_2, X_3, X_4) средняя, так как $r_1 = 0{,}445$.

Chi-Square Tests with Successive Roots Removed (primer.sta)

Continue...	Canonicl R	Canonicl R-sqr	Chi-sqr	df	p	Lambda Prime
0	,444969	,197997	2,829784	6	,829881	,702069
1	,352994	,124605	1,064636	2	,587247	,875395

Рис. 8.6. Канонические коэффициенты корреляции и их оценки

Судя по коэффициентам канонических функций, самую большую информационную нагрузку в определении тесноты связи имеют переменные X_3 и Y_2.

В сложившейся ситуации можно рекомендовать исследователю либо изменить набор исходных переменных, дополнив их более информативными, либо увеличить число наблюдений (n).

8.4. Контрольные задания

Задание 1. На основании приведенных данных табл. 8.3 по группе предприятий рассчитайте матрицу парных коэффициентов корреляции для последующего проведения канонического анализа.

Таблица 8.3

Номер объекта	Первое множество переменных		Второе множество переменных		
	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

Здесь: X_1 — число работающих, тыс. чел.; X_2 — фондовооруженность труда одного работника, млн ден. ед.; X_3 — электровооруженность труда одного рабочего, тыс. кВт.-ч; Y_1 — производительность труда одного работающего, млн ден. ед.; Y_2 — рентабельность производства, %.

Поясните смысл полученных элементов матрицы и сделайте выводы о степени тесноты связи между переменными внутри каждого из множеств.

Задание 2. По данным приведенной ниже матрицы парных коэффициентов корреляции, используя канонический корреляционный анализ, определите степень зависимости между двумя группами признаков.

$$R = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & Y_1 & Y_2 \end{matrix} \\ \begin{pmatrix} 1 & -0{,}25 & 0{,}48 & 0{,}72 & 0{,}31 \\ & 1 & 0{,}43 & -0{,}62 & 0{,}81 \\ & & 1 & 0{,}21 & 0{,}53 \\ & & & 1 & 0{,}32 \\ & & & & 1 \end{pmatrix} \end{array}.$$

Прежде чем приступить к расчетам, ответьте на вопрос: Сколько коэффициентов канонической корреляции можно рассчитать в данном примере?

Примечание. В первую группу признаков входят переменные X_1, X_2, X_3, во вторую группу — переменные Y_1, Y_2.

На основании полученных результатов сделайте выводы.

Задание 3. На основании данных, приведенных в табл. 8.4 для двух множеств (X_1, X_2, X_3) и (Y_1, Y_2), определите канонические переменные, соответствующие максимальному коэффициенту канонической корреляции.

Дайте оценку существенности изменения максимального коэффициента корреляции после удаления наименее значимой переменной (X_1, X_2 или X_3).

Таблица 8.4

Номер предприятия	Y_1	Y_2	X_1	X_2	X_3
1	50,4	1,2	3,8	0,18	40,0
2	51,5	2,1	6,9	0,10	28,0
3	49,6	1,5	7,8	0,18	18,2
4	54,0	1,9	5,1	0,24	15,0
5	40,8	2,5	7,3	0,10	20,1
6	35,8	1,3	8,0	0,13	29,3
7	59,7	3,5	8,1	0,08	27,5
8	47,5	2,7	7,3	0,11	182

Здесь: Y_1 — уровень производительности труда работающего, млн ден. ед.; Y_2 — фондоотдача основных производственных фондов, ден. ед.; X_1 — коэффициент обновления основных фондов, %; X_2 — удельный вес потерь от брака, %; X_3 — коэффициент оборачиваемости оборотных средств, дн.

Расчеты выполните на компьютере в пакете STATISTICA и поясните полученные результаты.

Задание 4. В ходе проведения канонического анализа были получены следующие значения коэффициентов канонической корреляции: $r_1 = 0{,}851$, $r_2 = 0{,}568$.

Воспользовавшись критерием Бартлетта, проверьте значимость каждого из двух коэффициентов при уровне значимости $\alpha = 0{,}05$.

Известно, что число независимых переменных, участвующих в анализе, равно пяти, число зависимых — трем, а объем выборки составил 30 наблюдений ($n = 30$).

Задание 5. На основании приведенных данных табл. 8.5 по группе предприятий легкой промышленности был проведен канонический анализ и получены следующие результаты:

- максимальный коэффициент корреляции $r = 0{,}583$;
- соответствующие ему канонические переменные:

$$\begin{cases} U = -0{,}014X_1 + 0{,}718X_2 - 0{,}422X_3, \\ V = -0{,}941Y_1 - 0{,}093Y_2. \end{cases}$$

Таблица 8.5

Номер предприятия	X_1	X_2	X_3	Y_1	Y_2
1	–10	120	105,0	25,5	25,0
2	–2	145	110,0	14,8	34,0
3	12	110	104,2	45,3	110,0
4	8	110	100,0	74,5	215,0
5	14	135	98,6	60,0	87,0
6	–7	75	130,0	80,0	135,0
7	–20	80	115,0	65,3	145,0

Окончание табл. 8.5

Номер предприятия	X_1	X_2	X_3	Y_1	Y_2
8	–3	95	101,0	64,1	180,0
9	15	125	95,0	34,5	94,5
10	10	140	94,2	46,3	120,0
11	2	90	94,8	33,4	84,3
12	–17	81	120,0	85,5	180,0
13	5	90	105,0	75,3	65,2
14	3	124	104,0	40,0	170,1
15	–6	80	113,5	80,2	210,0
16	6	65	108,6	35,0	65,0
17	–13	45	118,0	64,0	190,0
18	–25	60	108,0	75,0	285,0
19	4	95	101,0	45,5	210,0
20	8	115	115,0	54,8	250,4

Здесь: X_1 — средний процент снижения себестоимости продукции, %; X_2 — оборачиваемость оборотных средств, дн.; X_3 — индекс производительности труда рабочего, %; Y_1 — удельный вес экспорта в объеме реализованной продукции, %; Y_2 — прибыль на одного работника, тыс. ден. ед.

На основе анализа полученных коэффициентов канонических переменных выберите наименее информативную из числа независимых переменных. Удалив выбранную переменную, повторите расчеты канонических переменных и максимального канонического коэффициента корреляции. Оцените существенность его изменения для усеченного набора переменных.

Необходимые расчеты выполните на компьютере с использованием пакета STATISTICA.

Задание 6. Матрица парных коэффициентов корреляции, рассчитанная для двух групп переменных, равна

$$R = \left(\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right) = \left(\begin{array}{ccc|cc} 1 & -0{,}021 & -0{,}393 & -0{,}437 & -0{,}046 \\ -0{,}021 & 1 & 0{,}296 & -0{,}037 & -0{,}199 \\ -0{,}393 & 0{,}296 & 1 & 0{,}296 & -0{,}300 \\ \hline -0{,}437 & -0{,}037 & 0{,}296 & 1 & -0{,}117 \\ -0{,}046 & -0{,}199 & -0{,}300 & -0{,}117 & 1 \end{array}\right).$$

Вычислите собственные значения матрицы C: $C = R_{22}^{-1} \times R_{21} \times R_{12}^{-1} \times R_{12}$ и поясните полученные результаты.

Задание 7. На основании собственных значений и собственных векторов матрицы C (см. задание 6) рассчитайте оценки вектора коэффициентов канонической переменной $U = a_1X_1 + a_2X_2 + a_3X_3$ и максимальный коэффициент канонической корреляции для двух групп переменных.

Задание 8. По результатам выполнения задания 5 рассчитайте значения канонических переменных U и V для каждого наблюдаемого объекта.

Покажите на графике расположение этих объектов в пространстве канонических переменных. Расчеты выполните на компьютере с помощью пакета EXCEL.

Задание 9. Используя исходные данные, приведенные в задании 5, а также рассчитанные на их основе значения канонических переменных, проведите методом иерархического кластерного анализа классификацию двадцати объектов по назначениям:

- исходных переменных;
- канонических переменных.

Сравните полученные результаты двух классификаций. Оцените каждый из вариантов классификации с помощью одного из функционалов качества [23].

Задание 10. Воспользовавшись приведенными ниже ковариационной и корреляционной матрицами, вычисленными для одной и той же выборки, определите коэффициенты канонической корреляции для двух множеств переменных (X_1, X_2 — первое множество; Y_1, Y_2 — второе множество).

$$S = \left(\begin{array}{c|c} S_{11} & S_{12} \\ \hline S_{21} & S_{22} \end{array}\right) = \left(\begin{array}{cc|cc} 0{,}0126 & -1{,}558 & -0{,}0184 & -0{,}031 \\ & 1149 & -18{,}14 & 7{,}167 \\ \hline & & 2{,}318 & 3{,}651 \\ & & & 20{,}450 \end{array}\right),$$

$$R = \left(\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right) = \left(\begin{array}{cc|cc} 1 & -0{,}351 & 0{,}047 & -0{,}409 \\ & 1 & 0{,}530 & -0{,}108 \\ \hline & & 1 & -0{,}060 \\ & & & 1 \end{array}\right).$$

Сравните полученные значения максимальных коэффициентов канонической корреляции, поясните их смысл и оцените оба коэффициента по χ^2-критерию Бартлетта.

Задание 11. В ходе проведения канонического анализа были получены собственные числа и собственные векторы матрицы C:

$$C = R_{22}^{-1} \times R_{21} \times R_{11}^{-1} \times R_{12}:$$

$$\lambda_1^2 = 0{,}339; \qquad \lambda_2^2 = 0{,}038;$$

$$B_1 = (-0{,}941;\ -0{,}093); \quad B_2 = (-0{,}823;\ 1{,}247).$$

Вычислите значения соответствующих канонических коэффициентов корреляции и оцените их значимость по χ^2-критерию Бартлетта.

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

Значения функции $\varphi(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{\frac{-z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{-z}^{+z} e^{\frac{-z^2}{2}} dz$

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									

Распределение Стьюдента (*t*-распределение)

Значения $t_{\alpha,\nu}$ удовлетворяют условию $P(t \geq t_{\alpha,\nu}) = \int_{\alpha,\nu}^{\infty} S(t,\nu)dt = \alpha$

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

Приложение 5

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
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

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
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
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

Окончание прил. 5

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
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

ЛИТЕРАТУРА

1. Автоматизированное рабочее место для статистической обработки данных / В.В. Шураков, Д.М. Дайитбегов, С.В. Мизрохи и др. М.: Финансы и статистика, 1990.

2. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. М.: Статистика, 1974.

3. Прикладная статистика. Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков и др. М.: Финансы и статистика, 1989.

4. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.

5. *Андерсон Т.* Введение в многомерный статистический анализ. М.: Физматгиз, 1963.

6. *Благуш П.* Факторный анализ с обобщениями. М.: Финансы и статистика, 1988.

7. *Болч Б., Хуань К.Д.* Многомерные статистические методы для экономики: Пер. с англ. М.: Статистика, 1979.

8. *Браверман Э.М., Мучник И.Б.* Структурные методы обработки эмпирических данных. М.: Наука, 1983.

9. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. М.: Финансы и статистика, 1986. Кн. 2.

10. *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы: Учеб. М.: Финансы и статистика, 1998.

11. *Дубров А.М.* Обработка статистических данных методом главных компонент. М.: Статистика, 1978.

12. *Дэйвисон М.* Многомерное шкалирование. М.: Финансы и статистика, 1988.

13. *Дюк В.* Обработка данных на ПК в примерах. СПб.: Питер, 1997.

14. *Енюков И.С.* Методы, алгоритмы, программы многомерного статистического анализа (пакет ППСА). М.: Финансы и статистика, 1986.

15. *Жамбю М.* Иерархический кластерный анализ. М.: Финансы и статистика, 1988.

16. *Иберла К.* Факторный анализ. М.: Статистика, 1980.

17. *Кендалл М. Дж., Стьюарт А.* Многомерный статистический анализ и временные ряды. М.: Наука, 1976.

18. *Клигер С.А., Косолапов М.С., Толстова Ю.И.* Шкалирование при сборе и анализе социологической информации. М.: Наука, 1987.

19. *Ллойд Э., Ледерман У.* Справочник по прикладной статистике: Пер. с англ. М.: Финансы и статистика, 1990. Т. 2.

20. *Мандель И.Д.* Кластерный анализ. М.: Финансы и статистика, 1988.

21. *Миркин Б.Г.* Группировки в социально-экономических исследованиях. М.: Финансы и статистика, 1985.

22. Многомерный статистический анализ / Под ред. В.Н. Тамашевича. М.: ЮНИТИ, 1999.

23. *Окунь Я.* Факторный анализ. М.: Статистика, 1974.

24. *Плюта В.* Сравнительный многомерный анализ в экономическом моделировании: Пер. с пол. М.: Финансы и статистика, 1989.

25. *Терехина А.Ю.* Анализ данных методами многомерного шкалирования. М.: Наука, 1986.

26. *Торгерсон У.С.* Многомерное шкалирование. Теория и метод // Статистическое измерение количественных характеристик. М.: Статистика, 1972.

27. Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др. М.: Финансы и статистика, 1989.

28. *Харман Г.* Современный факторный анализ. М.: Статистика, 1972.

29. Эконометрика: Учеб. / Под ред. И.И. Елисеевой. М.: Финансы и статистика, 2001.

Учебное издание

Сошникова Людмила Антоновна
Тамашевич Виктор Николаевич
Махнач Любовь Александровна

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

Практикум

Ответственный за выпуск Л.А. Чеснокова

Редактор *А.К. Лапуста*
Корректор *Е.Г. Сазончик*
Технический редактор *О.В. Амбарцумова*
Компьютерный дизайн *Т.В. Бесчетнова*

Подписано в печать 12.02.2004. Формат 60×84/16. *Офсетная печать.* Гарнитура Times New Roman. Усл. печ. л. 9,5. Уч.-изд. л. 5,3. Тираж 200 экз. Заказ 31.

УО «Белорусский государственный экономический университет».
Лицензия ЛВ № 170 от 21.01.2003.
220070, Минск, просп. Партизанский, 26.

Отпечатано в УО «Белорусский государственный экономический университет».
Лицензия ЛП № 336 от 16.03.1999.
220070, Минск, просп. Партизанский, 26.