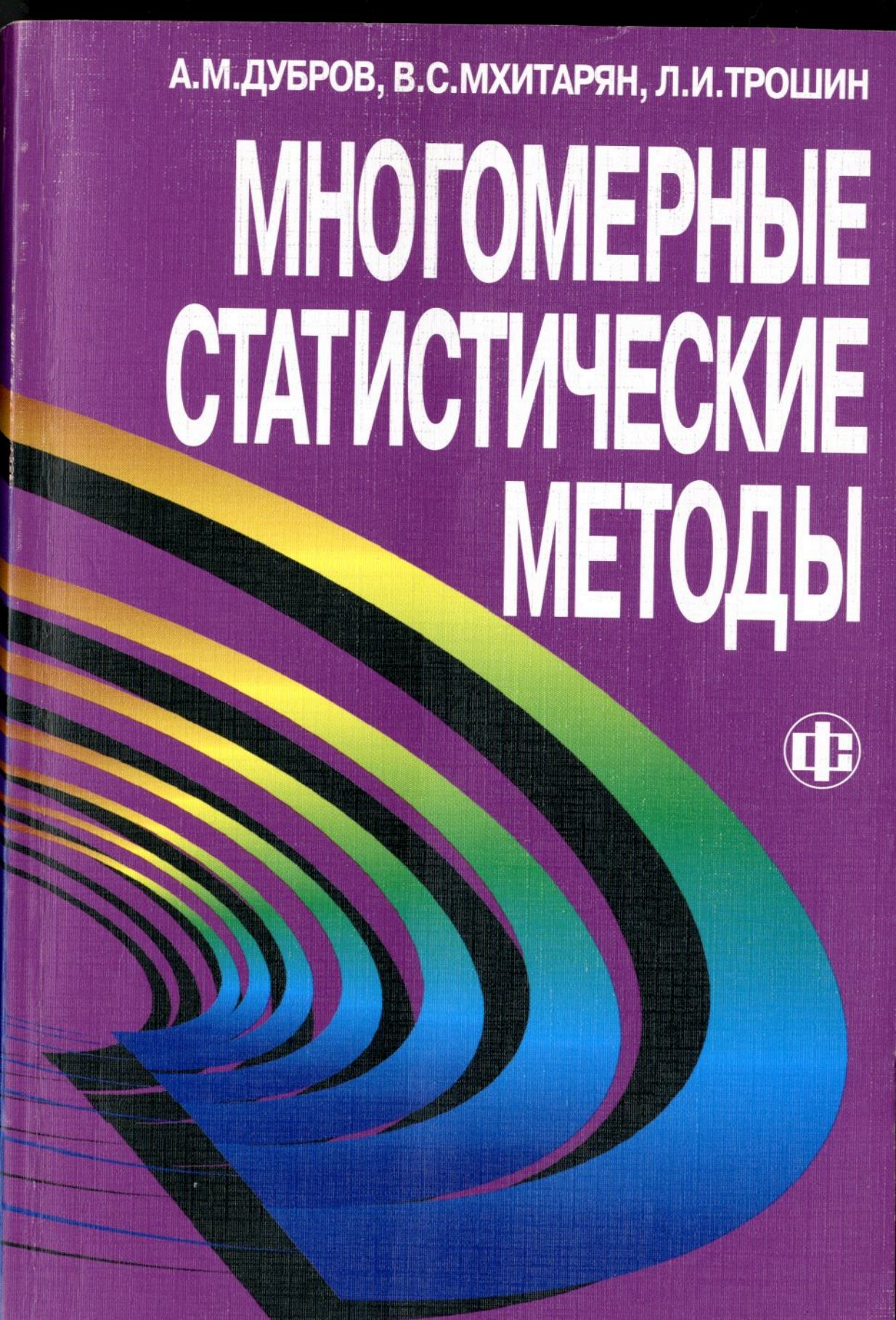
А.М.ДУБРОВ, В.С.МХИТАРЯН, Л.И.ТРОШИН

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ



А.М.ДУБРОВ, В.С.МХИТАРЯН, Л.И.ТРОШИН

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

ДЛЯ ЭКОНОМИСТОВ И МЕНЕДЖЕРОВ

2579

11703

Рекомендовано
Министерством образования
Российской Федерации
в качестве учебника
для студентов экономических специальностей
высших учебных заведений

Москва
"Финансы и статистика"
2003

УДК 519.237(075.8)
ББК 22.172я73
Д79

РЕЦЕНЗЕНТЫ:

И.И.Елисеева,
зав.кафедрой статистики
Санкт-Петербургского университета экономики и финансов,
д-р экон. наук, профессор, член-корр. РАН;

В.А.Колемаев,
зав. кафедрой прикладной математики
Государственной академии управления,
д-р экон. наук, профессор

Дубров А.М., Мхитарян В.С., Трошин Л.И.

Д79 Многомерные статистические методы: Учебник. – М.: Финансы и статистика, 2003. – 352 с.: ил.

ISBN 5-279-01945-3.

Рассматриваются многомерные генеральная и выборочная совокупности, корреляционный, регрессионный, компонентный, факторный анализ, канонические корреляции. Подробно изложены методы многомерной классификации, робастного оценивания. Учебник содержит математико-статистические таблицы, а также конкретные примеры и пояснения, необходимые для их выполнения, и упражнения с использованием ЭВМ, позволяющие закрепить изученный материал.

Для преподавателей, аспирантов, студентов бизнес-школ и экономических вузов, а также экономистов, менеджеров.

Д $\frac{1602000000-202}{010(01)-2003}$ 268 – 2003

УДК 519.237(075.8)
ББК 22.172я73

ISBN 5-279-019450-3

© А.М.Дубров, В.С.Мхитарян, Л.И.Трошин, 1998

ПРЕДИСЛОВИЕ

Учебник написан на основе двадцатилетнего опыта преподавания курса многомерного статистического анализа в Московском государственном университете экономики, статистики и информатики (МЭСИ).

Социально-экономические процессы и явления зависят от большого числа параметров, их характеризующих, что обусловливает трудности, связанные с выявлением структуры взаимосвязей этих параметров. В подобных ситуациях, т. е. когда решения принимаются на основании анализа стохастической, неполной информации, использование методов многомерного статистического анализа является не только оправданным, но и существенно необходимым.

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

К области приложения математической статистики могут быть отнесены задачи, связанные с исследованием поведения индивидуума, семьи или другой социально-экономической или производственной единицы, как представителя большой совокупности объектов.

Многомерный экономико-статистический анализ опирается на широкий спектр методов. Изложить все мыслимые подходы к исследованию в одном учебнике не представляется возможным, поэтому решено было ограничиться изложением тех методов, которые получили или в ближайшее время получат наибольшее распространение в практике работы.

Для установления связи пройденного курса теории вероятностей с математической статистикой введены главы "Многомерные генеральная и выборочная совокупности" и "Статистическое оценивание и сравнение многомерных генеральных совокупностей".

Проведение системного анализа до изучения взаимосвязей в многомерной совокупности требует иметь представление о связях между отдельной зависимой переменной и группой влияющих на нее показателей. Это может быть осуществлено при помощи множественного корреляционного и регрессионного анализа.

которая определяет вероятность того, что X примет значение, меньшее фиксированного действительного числа $x(x \in X)$.

В многомерном статистическом анализе изучаются генеральные совокупности с точки зрения нескольких признаков (обычно более двух). Рассматриваемое множество признаков обозначается вектором X, имеющим k компонент, каждая из которых характеризует соответствующий признак $x_j, j = 1,2,...,k$.

Таким образом, объектом исследования в многомерном анализе являются случайный вектор X (или случайная точка) в k-мерном евклидовом пространстве, система k случайных (одномерных) величин, k-мерная случайная величина $(x_1, x_2, ..., x_k)$.

Функцией распределения случайного вектора X называется **детерминированная неотрицательная величина**, определяемая по формуле:

$$F(x) = P(X_1 < x_1, X_2 < x_2, ... X_k < x_k) = P(X < x), \qquad (1.1)$$

где x - k-мерный вектор фиксированных действительных чисел

$$x = (x_1, x_2, ..., x_k)^T.$$

Детерминированная неотрицательная величина обладает следующими свойствами:

- $F(x) = 0$, если среди x_j имеется хотя бы одна компонента, равная $-\infty$.
- $F(x) = 1$, если все компоненты вектора X равны ∞.
- $F(x)$ удовлетворяет формуле вычисления вероятности попадания случайной точки в k-мерный параллелепипед с плоскостями, параллельными координатным.

Различают *непрерывные k-мерные случайные величины*, все компоненты которых - непрерывные (одномерные) случайные величины, *дискретные k-мерные случайные величины*, компоненты которых все дискретные, и *смешанные k-мерные случайные величины*, среди компонент которых есть как дискретные, так и непрерывные случайные величины.

Функция распределения $F(x)$ для непрерывной k-мерной случайной величины является **непрерывной** по определению.

Непрерывная k-мерная случайная величина имеет плотность распределения вероятностей

$$p(x) = p(x_1, x_2, ..., x_k) \geq 0, \qquad (1.2)$$

удовлетворяющую условию

$$F(x) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} ... \int_{-\infty}^{x_k} p(x_1, x_2, ..., x_k) dx_1 dx_2 ... dx_k. \qquad (1.3)$$

Плотность $p(x)$ обладает следующими свойствами:

$$\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} p(x_1, ..., x_k) dx_1 ... dx_k = 1, \qquad (1.4)$$

где k – общее число (кратность) интегралов.

Вероятность попадания точки $(x_1, x_2, ..., x_k)$ в какую-нибудь область G равна

$$\int_G^{(k)} ... \int p(x_1, ..., x_k) dx_1 ... dx_k. \qquad (1.5)$$

Из определения плотности следует, что

$$\begin{aligned}
&\int_{-\infty}^{+\infty} p(x_1, x_2) dx_1 = p(x_2); \\
&\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x_1, x_2, x_3) dx_2 dx_3 = p(x_1); \qquad (1.6) \\
&\int_{-\infty}^{+\infty} p(x_1, x_2, x_3) dx_1 = p(x_2, x_3).
\end{aligned}$$

Дискретная k-мерная случайная величина может быть задана указанием вероятностей ее попадания в любую точку счетного (или конечного) множества допустимых точек

$$P_{i_1, i_2, ... i_k} = P(X_1 = x_1^{(i_1)}, X_2 = x_2^{(i_2)}, ..., X_k = x_k^{(i_k)}),$$

где $x_j^{(i_j)}$ - значение случайной величины X_j из множества допустимых значений i_j,

$$i_1 = 1,2,...; i_2 = 1,2,..., i_k = 1,2,...$$

Функция распределения такой величины выражается формулой

$$F(x_1, x_2, ..., x_k) = \sum_{x_1^{(i_1)} < x_1} \sum_{x_2^{(i_2)} < x_2} ... \sum_{x_k^{(i_k)} < x_k} P_{i_1, i_2, ..., i_k} \qquad (1.7)$$

(суммирование ведется по всем номерам $i_j, j = 1,2,...,k$, для которых выполняются указанные неравенства).

Справедлива формула

$$\sum_{i_1}\sum_{i_2}\dots\sum_{i_k} P_{i_1,i_2,\dots,i_k} = 1.$$

Плотности, вероятности, функции распределения подсистем $L(1 \le L < k)$ случайных величин системы k случайных величин называют **частными**, или **маргинальными распределениями**.

Условными распределениями случайного вектора x называются **распределения подсистемы** $L(1 \le L < k)$ **его компонент** при условии, что остальные $k-L$ компоненты являются фиксированными. Эти компоненты от нефиксируемых будут отделяться **косой чертой**.

Справедливы, например, формулы, определяющие условное распределение двумерной случайной величины (x_1, x_2), являющейся подсистемой системы $(x_1, x_2, x_3, x_4, x_5)$ при условии, что в ней фиксированы три последние компоненты.

Для дискретной случайной величины

$$p_{i_1 i_2 / i_3 i_4 i_5} = \frac{p_{i_1 i_2 i_3 i_4 i_5}}{p_{i_3 i_4 i_5}} = \frac{p_{i_1 i_2 i_3 i_4 i_5}}{p_{**i_3 i_4 i_5}}, \qquad (1.8)$$

где $p_{**i_3 i_4 i_5} = \sum_{i_1}\sum_{i_2} p_{i_1 i_2 i_3 i_4 i_5}$.

Для непрерывной случайной величины

$$p(x_1, x_2 / x_3 x_4 x_5) = \frac{p(x_1, x_2, x_3, x_4, x_5)}{p(x_3, x_4, x_5)}, \qquad (1.9)$$

где $p(x_3, x_4, x_5) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x_1, x_2, x_3, x_4, x_5)\,dx_1 dx_2$.

Подсистема $L(1 \le L < k)$ компонент и дополнительная подсистема $k-L$ компонент вектора x называются **независимыми** (стохастически, вероятностно, статистически), если справедливо равенство

$$F(x_1, x_2, \dots, x_k) = F(x_{j1}, x_{j2}, \dots, x_{jl}) \cdot F(x_{j,l+1}, \dots, x_{jk}). \qquad (1.10)$$

В частности, компоненты x называются **независимыми**, если

$$F(x_1, x_2, \dots, x_k) = F(x_1)F(x_2)\dots F(x_k). \qquad (1.11)$$

В случае независимости справедливы аналогичные формулы для произведений плотностей или вероятностей маргинальных распределений и совпадение условных распределений с соответствующими маргинальными.

1.2. Некоторые характеристики генеральной совокупности

Закон распределения полностью характеризует генеральную совокупность, т.е. последняя может быть задана функцией распределения или плотностью. Однако такая исчерпывающая характеристика генеральной совокупности довольно сложна и не требуется для решения многих практических задач, в которых достаточно знать лишь числовые характеристики законов распределения. Например, для грубой характеристики одномерной случайной величины можно ограничиться ее средним значением и величиной разброса возможных значений.

Математическим ожиданием или генеральной средней дискретной случайной величины x называется сумма произведений возможных значений этой величины на соответствующие вероятности

$$Mx = \sum_{i=1}^{N} x_i p_i, \qquad (1.12)$$

где p_i - вероятность появлений i-го возможного значения x_i случайной величины $x (i=1,2,\dots,N)$.

Для непрерывной генеральной совокупности x с плотностью $p(x)$ математическое ожидание равно

$$Mx = \int_{-\infty}^{\infty} xp(x)dx. \qquad (1.13)$$

Аналогично можно определить математическое ожидание функции $\varphi(x)$ от случайной величины x:

$$\text{дискретной } M\varphi(x) = \sum_{i=1}^{N} \varphi(x_i)p_i$$

и

$$\text{непрерывной } M\varphi(x) = \int_{-\infty}^{\infty} \varphi(x_i)p(x)dx.$$

Начальным ν_k и центральным μ_k моментами k-го порядка скалярной случайной величины x называются **математические ожидания** соответственно **величин** x^k и $(x - Mx)^k$, т.е. $\nu_k = Mx^k$ и $\mu_k = M(x - Mx)^k$.

Отсюда следует, что начальный момент первого порядка есть математическое ожидание случайной величины $\nu_1 = Mx$.

Дисперсия Dx случайной величины x есть центральный момент второго порядка

$$Dx = M(x - Mx)^2. \qquad (1.14)$$

Средним квадратическим отклонением называется положительный квадратный корень из ее дисперсии $\sigma_x = \sqrt{Dx}$.

Одним из важнейших результатов применения многомерного анализа является статистический вывод о параметрах или характеристиках генеральной совокупности. Дадим определение основных генеральных характеристик.

Функция $Y_1 = g_1(x) = g_1(x_1, x_2, \ldots, x_k)$ является одномерной случайной величиной, если она вещественна, однозначно определена во всех точках $x \in R_k$ (R_k - k-мерное евклидово пространство), за исключением точек множества нулевой вероятности.

Функция распределения для нее вычисляется по формуле

$$F_{y_1}(y_1) = F(y_1) = P(Y_1 < y_1) = P(g_1(x) < y_1).$$

Нетрудно получить свойства математического ожидания исходя из определения математического ожидания функции

$$M(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = \beta_0 + \beta_1 Mx_1 + \ldots + \beta_k Mx_k, \qquad (1.15)$$

где $x_1, \ldots, x_k$ - случайные величины;

$\beta_0, \beta_1, \ldots, \beta_k$ - детерминированные величины;

$M(x_1, x_2, \ldots, x_k) = Mx_1 \cdot Mx_2 \cdot \ldots \cdot Mx_k$ - для взаимно независимых случайных величин $x_1, x_2, \ldots, x_k$.

Моментом l-го порядка $l = l_1 + l_2 + \ldots + l_k$ случайного вектора x относительно постоянного вектора c называется

$$M_{l_1, l_2, \ldots, l_k} = M[(x_1 - c_1)^{l_1} (x_2 - c_2)^{l_2} \ldots (x_k - c_k)^{l_k}]. \qquad (1.16)$$

Если $c_j = 0$, то моменты называются **начальными**, если $c_j = Mx_j$ - **центральными**.

На практике бывает достаточным ограничиваться моментами до второго порядка включительно. Приведем *названия, обозначения и формулы некоторых моментов*:

$Mx_j^l = M(x_j)^l$ - начальный момент l-го порядка, математическое ожидание l-й степени j-й компоненты вектора x;

$\underset{i \neq j}{M}(x_i x_j), i, j = 1 \div k$ - начальный смешанный момент второго порядка;

$$\sigma_{jj} = Mx_j^2 - (Mx_j)^2 = M(x_j - Mx_j)^2 = \sigma_j^2, \qquad (1.17)$$

где σ_{jj} - центральный момент второго порядка, дисперсия j-й компоненты вектора x, $j = 1 \div k$;

$$\sigma_{ij} = M\left[(x_i - Mx_i)(x_j - Mx_j)\right] = M(x_i x_j) - (Mx_i)(Mx_j), \qquad (1.18)$$

где σ_{ij} - центральный смешанный момент второго порядка, коэффициент ковариации i-й и j-й компонент вектора x, $i; j = 1 \div k$ при $i \neq j$.

Ковариационная матрица определяется как математическое ожидание произведения центрированного случайного вектора на этот же транспонированный вектор $\Sigma = M\left[(x - Mx)(x - Mx)^T\right]$.

Матрица

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1j} & \ldots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2j} & \ldots & \sigma_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \sigma_{i1} & \sigma_{i2} & \ldots & \sigma_{ij} & \ldots & \sigma_{ik} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_{kj} & \ldots & \sigma_{kk} \end{pmatrix} \qquad (1.19)$$

называется **ковариационной матрицей случайного вектора x**.

Нетрудно заметить, что она симметрическая. Покажем, что Σ неотрицательно определена.

Пусть $g_1(x) = c_1 x_1 + c_2 x_2 + \ldots + c_k x_k = c^T x$, где c - любой детерминированный вектор. Тогда дисперсия $Dg_1(x)$ (по определению математического ожидания и дисперсии) будет равна $Dg_1(x) = M[g_1(x) - Mg_1(x)]^2 =$
$= c^T M[(x - Mx)(x - Mx)^T]c = c^T \Sigma c \geq 0.$

Коэффициент ковариации нормированных случайных величин называется **коэффициентом корреляции, или коэффициентом парной корреляции**

$$\rho_{ij} = M\left(\frac{x_i - Mx_i}{\sigma_i} \cdot \frac{x_j - Mx_j}{\sigma_j}\right) = \frac{\sigma_{ij}}{\sigma_i \sigma_j}, \qquad (1.20)$$

где σ_i, σ_j - средние квадратические отклонения случайных величин x_i и x_j.

Квадрат коэффициента корреляции называют **коэффициентом детерминации**.

Так как для любых случайных величин x_i и x_j

$$M\left(\frac{x_i - Mx_i}{\sigma_i} \pm \frac{x_j - Mx_j}{\sigma_j}\right)^2 = 2 \pm 2\rho_{ij} \geq 0,$$

то $-1 \leq \rho_{ij} \leq 1$.

Причем при $\rho_{ij} = \pm 1$ x_i и x_j связаны линейной функциональной зависимостью. Если же x_i и x_j независимы, то $\rho_{ij} = 0$.

Матрица

$$R = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \dots & \dots & \dots & \dots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{pmatrix} \qquad (1.21)$$

называется **корреляционной**, она симметрическая и неотрицательно определена.

Имеет место очевидное свойство для любых компонент x_i и x_j

$$M(x_i x_j) = Mx_i Mx_j + \sigma_{ij} = Mx_i Mx_j + \rho_{ij}\sigma_i\sigma_j . \qquad (1.22)$$

Условным математическим ожиданием функции $y_1 = g(x) = g_1(x_1, x_2, \dots, x_p, x_{p+1}, \dots, x_k)$ при условии (фиксировании) $x_{p+1} = x_{p+1}, x_{p+2} = x_{p+2}, \dots, x_k = x_k$ называется функция компонент $x_{p+1}, x_{p+2}, \dots, x_k$, вычисляемая по формулам (для дискретной и непрерывной величин соответственно):.

$$My_1 / x_{p+1}, \dots, x_k = \begin{cases} \sum_{i_1} \dots \sum_{i_p} g_1(x_1^{(i_1)}, \dots, x_p^{(i_p)}, x_{p+1}, \dots, x_k) p_{i_1 \dots i_p / i_{p+1} \dots i_k} \\ \int_{-\infty}^{\infty} (p) \dots \int_{-\infty}^{\infty} g_1(x) p(x_1, x_2, \dots, x_p / x_{p+1}, \dots, x_k) dx_1 \dots dx_p \end{cases} \qquad (1.23)$$

1.3. Параметры связи между признаками в генеральной совокупности

В п.1.2 были даны некоторые параметры связи между двумя компонентами случайного вектора - коэффициенты ковариации и корреляции. Однако непосредственно измерителями связи эти параметры при размерности k, вектора x, большей двух,быть не могут. Они играют важную вспомогательную роль при вычислении действительных показателей связи между признаками генеральной совокупности и в образовании различных моделей многомерного статистического анализа.

Зависимость между компонентой x_1 и остальными компонентами $(x_2, \dots, x_k)$ вектора x можно изучать с помощью представления

$$x_1 = g_1(x_2, \dots, x_k) + \varepsilon_1(x_1, x_2, \dots, x_k), \qquad (1.24)$$

где $g_1(x_2, \dots, x_k)$ - некоторая функция компонент $x_2, \dots, x_k$;

$\varepsilon_1(x_1, x_2, \dots, x_k) \equiv \varepsilon_1(x)$ - остаточный (поправочный) член;

$\varepsilon_1(x) = x_1 - g_1(x_2, \dots, x_k)$.

Функция $g_1(x_2, \dots, x_k)$ называется **регрессией компоненты** x_1 на компоненты $(x_2, \dots, x_k)$.

Наиболее часто в качестве регрессии берут функцию g_1^*, минимизирующую математическое ожидание квадрата отклонения

$$Q(g_1) = M[x_1 - g_1(x_2, \dots, x_k)]^2 = M\varepsilon_1^2(x). \qquad (1.25)$$

Для отыскания такой функции рассмотрим тождество

$$x_1 - g_1 = (x_1 - Mx_1 / x_2, \dots, x_k) + (Mx_1 / x_2, \dots, x_k - g_1).$$

Возведем обе части тождества в квадрат и возьмем математическое ожидание правой и левой частей нового тождества. Так как

$$M[(x_1 - Mx_1 / x_2, \dots, x_k)(Mx_1 / x_2, \dots, x_k - g_1(x_2, \dots, x_k))] =$$
$$= [(Mx_1 / x_2, \dots, x_k - Mx_1 / x_2, \dots, x_k)(Mx_1 / x_2, \dots, x_k - g_1)] = 0,$$

то будем иметь

$$M(x_1 - g_1)^2 = M(x_1 - Mx_1 / x_2, \dots, x_k)^2 + M(Mx_1 / x_2, \dots, x_k - g_1)^2 . \qquad (1.26)$$

В правой части последнего тождества от неизвестной функции зависит только второе слагаемое, которое как неотрицательная величина достигает минимального значения, равного нулю, лишь при

$$g_1^* = g_1^*(x_2, \dots, x_k) \equiv Mx_1 / x_2, \dots, x_k .$$

Регрессия, являющаяся условным математическим ожиданием

$$g_1 = Mx_1 / x_2, \dots, x_k ,$$

называется **средней квадратической регрессией** x_1 на $x_2, \dots, x_k$.

В дальнейшем для краткости будем опускать слова "средняя квадратическая", но рассматривать следует именно указанную регрессию.

Показателем, характеризующим рассеяние случайной величины x_1 около фиксированной точки $g_1(x_2, \dots, x_k)$ поверхности регрессии, является условная дисперсия

$$Dx_1 / x_2, \dots, x_k = M(x_1 - Mx_1 / x_2, \dots, x_k)^2 . \qquad (1.27)$$

Математическое ожидание условной дисперсии является характеристикой рассеяния значений случайной величины x_1 около всей поверхности регрессии x_1 на $(x_2, \dots, x_k)$ и называется **остаточной дисперсией регрессии** x_1 на $(x_2, \dots, x_k)$:

$$D_{ОСТ x_1} = M(Dx_1 / x_2, \dots, x_k).$$

Нетрудно увидеть, что

$$D_{ОСТ X_1} = M(x_1 - Mx_1 / x_2,\ldots,x_k)^2 = Mh_1^2(x). \qquad (1.28)$$

Назовем дисперсией регрессии (факторной дисперсией) случайной величины x_1 на $(x_2,\ldots,x_k)$ детерминированную величину

$$D_{РЕГР X_1} = D_{ФАКТ X_1} = D(Mx_1 / x_2,\ldots,x_k). \qquad (1.29)$$

Справедлива следующая формула о разложении дисперсии величины x_1:

$$D_{ОБЩ X_1} = D_{РЕГР X_1} + D_{ОСТ X_1}, \qquad (1.30)$$

где $D_{ОБЩ X_1} = Dx_1$.

Мерой зависимости между величинами x_1 и $(x_2,\ldots,x_k)$ служит отношение

$$\eta^2_{x_1} = \eta^2_{x_1 / x_2,\ldots,x_k} = \frac{D_{РЕГР X_1}}{D_{ОБЩ X_1}}, \qquad (1.31)$$

показывающее долю рассеяния величины x_1, обусловленную функциональной зависимостью x_1 от остальных компонент.

Величина

$$1 - \eta^2_{x_1} = \frac{D_{ОСТ X_1}}{D_{ОБЩ X_1}} \qquad (1.32)$$

показывает долю рассеяния случайной величины x_1 за счет "случайного фактора" независимо от изменения компонент $x_2,\ldots,x_k$.

В качестве показателей зависимости иногда рассматривают арифметическое значение корня квадратного из приведенных отношений, такие показатели называют **корреляционными отношениями**.

Заметим, что при функциональной зависимости x_1 от $(x_2,\ldots,x_k)$ имеет место тождество

$$\varepsilon_1(x_1, x_2,\ldots,x_k) \equiv 0$$

и корреляционное отношение $\eta_{x_1} = 1$.

И наоборот, если $\eta_{x_1} = 1$, то $D_{ОСТ X_1} = 0$, следовательно, $h_1(x_1, x_2,\ldots, x_k) = C$. Однако константа на самом деле равна нулю, так как

$$Mh_1 / x_2,\ldots,x_k = 0$$

по определению регрессии. Поэтому x_1 зависит функционально от $(x_2,\ldots,x_k)$. Далее, если x_1 не зависит от $(x_2,\ldots,x_k)$, то $Mx_1 / x_2,\ldots,x_k = Mx_1$ и $\eta_{x_1} = 0$. Если же $\eta_{x_1} = 0$, то нельзя утверждать, что x_1 и $(x_2,\ldots,x_k)$ - независимые случайные величины.

Если имеет место равенство

$$Mx_1 / x_2,\ldots,x_k = g_1 = Mx_1 + \sum_{l=2}^{k} \beta_{1l}(x_l - Mx_l),$$

то говорят, что регрессия x_1 на $(x_2,\ldots,x_k)$ линейна, и β_{1l} называют **коэффициентами множественной регрессии** x_1 на x_l.

Очевидно коэффициент регрессии β_{1l} показывает, на сколько единиц (своего измерения) изменится в среднем результативный признак x_1, если факторный признак x_l изменится на единицу (своего измерения) при условии, что остальные факторные признаки (аргументы) не меняются.

Показателем меры линейной зависимости x_1 от $(x_2,\ldots,x_k)$ служит **множественный коэффициент корреляции**, или его **квадрат-коэффициент детерминации**, определяемый по формуле

$$R^2_{x_1} = R^2_{x_1 / x_2,\ldots,x_k} = 1 - \frac{|R|}{R_{11}}. \qquad (1.33)$$

Можно показать, что в случае линейной зависимости, т.е. когда

$$Mx_1 / x_2,\ldots,x_k = Mx_1 + \sum_{l=2}^{k} \beta_{1l}(x_l - Mx_l),$$

справедливо равенство

$$R^2_{x_1} = \eta^2_{x_1}. \qquad (1.34)$$

Частным коэффициентом корреляции компонент x_1 и x_2 (по отношению к остальным компонентам) называется величина

$$\rho_{12|34\ldots k} = -\frac{R_{12}}{\sqrt{R_{11}R_{22}}}, \qquad (1.35)$$

где R_{12} - алгебраическое дополнение 1-й строки и 2-го столбца корреляционной матрицы R.

Приведенный коэффициент измеряет связь между x_1 и x_2 после устранения влияния величин $(x_3, x_4,\ldots,x_k)$.

Легко показать, что, например,

$$\beta_{12} = -\frac{\sigma_1 R_{12}}{\sigma_2 R_{22}} \qquad (1.36)$$

и при k=3

$$\rho_{12/13} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}},$$

$$R_{1/23} = R_{x_1 / x_2 x_3} = \sqrt{\frac{\rho_{12}^2 + \rho_{13}^2 - 2\rho_{13}\rho_{12}\rho_{23}}{1-\rho_{23}^2}}. \qquad (1.37)$$

1.4. Многомерная нормально распределенная генеральная совокупность

При рассмотрении различных моделей статистического анализа часто предполагается нормальное распределение всех или некоторых признаков генеральной совокупности. Говорят, что непрерывная k-мерная случайная величина распределена нормально, если плотность распределения имеет вид

$$p(x) = [(2\pi)^k |\Sigma|]^{-1/2} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}, \qquad (1.38)$$

где $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_k \end{pmatrix}$ - k-мерный вектор математических ожиданий;

Σ^{-1} - матрица, обратная ковариационной матрице Σ размерности (k x k);

$|\Sigma|$ - определитель этой матрицы.

Напомним, что матрица Σ является симметрической и положительно определенной.

Таким образом, многомерный нормальный закон распределения определяется вектором математических ожиданий μ и ковариационной матрицей Σ, т.е. $[k + \frac{k(k+1)}{2}]$ параметрами генеральной совокупности.

Пример 1.1

Показать, что при k=1 имеет место одномерный нормальный закон распределения.

В самом деле, при k=1 $\Sigma = \sigma_{11} = \sigma^2$. Тогда $|\Sigma| = \sigma^2$, а обратная матрица $\Sigma^{-1} = \frac{1}{\sigma^2}$. Подставляя найденные значения в выражение (1.38), имеем

$$P(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \qquad (1.39)$$

Мы получили плотность распределения одномерного нормального закона, зависящего от двух параметров: математического ожидания μ и среднего квадратического отклонения σ.

Пример 1.2

Приняв k=2 в выражении (1.38), вывести плотность двумерного нормального закона распределения.

При k=2 имеем

$$(x-\mu) = \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}; \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

где $\sigma_{12} = \sigma_{21}$.

Тогда, приняв $\sigma_{11} = \sigma_1^2$ и $\sigma_{22} = \sigma_2^2$, получим

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}.$$

Откуда, согласно выражению (1.38)

$$Q(x_1 x_2) = (x-\mu)^T \Sigma^{-1}(x-\mu) = (x_1-\mu_1 x_2-\mu_2) \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \times$$

$$\times \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \times$$

$$\times((x_1-\mu_1)\sigma_1^2 - (x_2-\mu_2)\sigma_{12} - \sigma_{12}(x_1-\mu_1) + \sigma_1^2(x_2-\mu_2)) \times$$

$$\times \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \times$$

$$\times[(x_1-\mu_1)^2\sigma_1^2 - 2(x_1-\mu_1)(x_2-\mu_2)\sigma_{12} + (x_2-\mu_2)^2\sigma_2^2].$$

Разделив числитель и знаменатель на $\sigma_1^2\sigma_2^2$ и учитывая, что $\frac{\sigma_{12}}{\sigma_1\sigma_2} = \rho_{12} = \rho$, получим

$$Q(x_1, x_2) = \frac{1}{1-\rho^2}[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}]. \qquad (1.40)$$

Учитывая также, что

$$\frac{1}{\sqrt{(2\pi)^2|\Sigma|}} = \frac{1}{2\pi\sqrt{(\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}},$$

окончательно получим

$$P(x)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}Q(x_1,x_2)\right\}. \qquad (1.41)$$

Из выражений (1.40) и (1.41) следует, что плотность двумерного нормального закона распределения определяется пятью параметрами: математическими ожиданиями μ_1 и μ_2 случайных величин x_1 и x_2, их средними квадратическими отклонениями σ_1,σ_2 и коэффициентом корреляции ρ.

Определенная выше нормально распределенная генеральная совокупность является **невырожденной**.

В случае вырожденности совокупности компоненты случайного вектора x являются линейно зависимыми. Следовательно, часть их представляет собой линейные комбинации остальных компонент, образующих линейно независимую подсистему.

Можно показать, что всевозможные условные распределения нормально распределенной генеральной совокупности при фиксировании различных подмножеств компонент являются нормальными.

Также можно показать, что необходимым и достаточным условием взаимной независимости компонент (или подмножеств компонент) является равенство нулю соответствующих коэффициентов корреляции.

Всевозможные квадратические регрессии являются линейными. Следовательно, в качестве показателя связи можно эффективно использовать всевозможные коэффициенты корреляции (детерминации) и регрессии.

1.5. Выборка из генеральной совокупности

Выборкой из генеральной совокупности (X) называют результаты ограниченного ряда наблюдений $x_1,x_2,\ldots,x_n$, где n - объем выборки.

Выборку рассматривают как некий эмпирический аналог генеральной совокупности, с которым чаще всего на практике имеют дело, поскольку обследование всей генеральной совокупности бывает либо слишком трудоемко, либо принципиально невозможно.

Задачи математической статистики фактически сводятся к обоснованному суждению об объективных свойствах генеральной совокупности по результатам выборки.

Достоверность выводов, получаемых в результате статистической обработки данных, во многом зависит от успешного решения вопроса представительности выборки, т.е. полноты и адекватности представления свойств анализируемой генеральной совокупности.

Это достигается случайностью отбора, когда каждый элемент генеральной совокупности имеет одинаковую вероятность быть отобранным.

При оценке представительности выборки учитывается и то, как выборка получена, и то, насколько распределение показателей в выборке характерно для анализируемой генеральной совокупности в целом.

Необходимость выборочного обследования при решении практических задач может быть связана со следующими причинами:

- генеральная совокупность настолько многочисленна, что проведение обследования всех элементов совокупности (сплошное обследование) слишком трудоемко. С такой ситуацией приходится встречаться при контроле качества продукции крупносерийного и массового производства;
- при бесконечно большой генеральной совокупности, когда даже весьма большое множество наблюдений не исчерпывает всей совокупности. Например, при разработке статистически обоснованных временных нормативов на изготовление изделия. В этом случае выводы, полученные по результатам конечного числа наблюдений, распространяются на всю генеральную совокупность, охватывающую все детали, которые могут быть изготовлены на станках данного типа на протяжении ряда последующих лет;
- в процессе проведения испытания происходит разрушение отбираемых образцов (например, испытание службы изделия, предела прочности и т.д.);
- встречаются обстоятельства, когда мы располагаем результатами испытания всей совокупности, реально существующей на данный момент времени, но рассматриваем их как выборку из гипотетической генеральной совокупности. Так поступают в тех случаях, когда хотят выявить общую закономерность, по отношению к которой имеющаяся совокупность представляется лишь частным случаем.

Например, на протяжении ряда лет доля мальчиков среди новорожденных составляла 0,513 общего числа родившихся в стране. Это данные сплошного обследования, но если нас интересует общая закономерность и мы хотим распространить полученные результаты на последующие годы, то данные следует рассматривать как выборку из некоторой бесконечной гипотетической совокупности.

В математической статистике существуют два взгляда на выборку. С одной стороны (практической), $x_1,x_2,\ldots,x_n$ есть n наблюденных значений случайной величины X, т.е. конкретные числа или векторы. С другой стороны, это n независимых наблюдений, каждое из которых имеет один и тот же набор возможных значений X^* и их вероятностей $P(X^*)$. Отсюда следует, что последовательность наблюдений $(x_1,x_2,\ldots,x_n)$ можно представить как *n независимых случайных величин*

$x_1, x_2, \ldots, x_n$ *с одинаковой функцией* $p(x)$ плотности (если X - непрерывная величина) или вероятностей (если X - дискретна). Такая выборка называется **случайной** и представляет собой n-мерную случайную величину $(x_1, x_2, \ldots, x_n)$ с плотностью или функцией вероятностей (соответственно для непрерывной или дискретной величины X)

$$L(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2)\ldots P(x_n) = \prod_{i=1}^{n} P(x_i),$$

называемой **функцией правдоподобия**.

В этом случае говорят, что выборка взята из генеральной совокупности X с законом распределения $P(x)$.

Статистикой (выборочной характеристикой) называют функцию, зависящую только от результатов наблюдения $x_1, x_2, \ldots, x_n$

$$Q_n^* = y(x_1, x_2, \ldots, x_n).$$

Отсюда следует, что статистика представляет собой случайную величину с законом распределения, определяемым функцией правдоподобия, а следовательно, и законом распределения величины X.

Выборка объема n из многомерной генеральной совокупности X, имеющей функцию распределения

$$F(x) = F(x_1, x_2, \ldots, x_i, \ldots, x_k),$$

определенную в евклидовом пространстве R_k, называется nk-мерная случайная величина

$$F([x]) = \prod_{i=1}^{n} F(x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ik}) = \prod_{i=1}^{n} F(x^{(i)}),$$

определенная в R_{nk} - nk-мерном евклидовом пространстве.

Из определения следует:

1) каждая наблюдаемая k-мерная точка

$$x^{(i)} = (x_{i1}, x_{i2}, \ldots, x_{ik})$$

является случайной величиной, распределенной так же, как и генеральная совокупность;

2) наблюдаемые точки $x^{(i)}$ представляют собой независимые случайные величины.

ГЛАВА 2

СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ И СРАВНЕНИЕ МНОГОМЕРНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ

2.1. Точечные оценки параметров многомерной генеральной совокупности

Выборку объема n из k-мерной генеральной совокупности X можно представить в виде матрицы данных

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix},$$

строки которой рассматриваются как n независимых реализаций k-мерного случайного вектора. Таким образом, элементы x_{ij} матрицы X можно рассматривать либо как случайные (одномерные) величины (независимые по i), либо как конкретные наблюдаемые значения - координаты n точек в k-мерном евклидовом пространстве (или n точек в k-мерном пространстве).

Приведем точечные оценки моментов генеральной совокупности, которые получили наибольшее практическое применение.

Оценка начального момента m-го порядка l-й компоненты случайного вектора x вычисляется по формуле

$$\bar{x}_l^m = \frac{1}{n}\sum_{i=1}^{n} x_{il}^m, \quad l = \overline{1,k},$$

при m=1 имеем среднюю арифметическую.

Оценка ковариационной матрицы Σ случайного вектора x (матрица выборочных дисперсий и коэффициентов ковариации) определяется как

$$S = \begin{pmatrix} s_{11} & s_{12} & \ldots & s_{1k} \\ s_{21} & s_{22} & \ldots & s_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ s_{k1} & s_{k2} & \ldots & s_{kk} \end{pmatrix}; \quad s_{ij} = \frac{1}{n}\sum_{i=1}^{n}(x_{il} - \bar{x}_l)(x_{ij} - \bar{x}_j); l, j = \overline{1,k},$$

причем $s_{ll} = s_l^2$ - выборочная дисперсия l-й компоненты случайного вектора x, s_{lj} - выборочный коэффициент ковариации компонент l и j вектора x. Вместо S употребляют также несмещенную оценку Σ:

$$\hat{S} = \begin{pmatrix} \hat{s}_{11} & \hat{s}_{12} & \dots & \hat{s}_{1k} \\ \hat{s}_{21} & \hat{s}_{22} & \dots & \hat{s}_{2k} \\ \dots & \dots & \dots & \dots \\ \hat{s}_{k1} & \hat{s}_{k2} & \dots & \hat{s}_{kk} \end{pmatrix} = \frac{n}{n-1} S, \quad \hat{s}_{ij} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{il} - \bar{x}_l)(x_{ij} - \bar{x}_j);$$

$l, j = \overline{1,k}$.

При алгоритмизации задач многомерного анализа полезным может оказаться вычисление оценок параметров генеральной совокупности с использованием операций над матрицами.

Оценку корреляционной матрицы (ρ_{ij}) можно получить по формулам

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{bmatrix},$$

где $r_{lj} = \dfrac{S_{lj}}{S_l S_j}$ - оценка парного коэффициента корреляции между l-й и j-й компонентами $x, l, j = \overline{1,k}$.

Или

$$R = \frac{1}{n} Z^T Z, \qquad Z = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1k} \\ z_{21} & z_{22} & \dots & z_{2k} \\ \dots & \dots & \dots & \dots \\ z_{n1} & z_{n2} & \dots & z_{nk} \end{bmatrix}, \qquad z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}.$$

Для получения оценок параметров условных распределений требуются специально организованные выборки, т.е. выборки при фиксированных значениях части компонент генеральной совокупности. На практике такой подход можно осуществить с помощью группировки данных по закрепленным значениям части признаков, равных либо их естественным дискретным значениям, либо серединным, центральным значениям областей (интервалов, прямоугольников, брусов и т.д.) группирования.

Рассмотрим некоторые оценки параметров условных распределений на примере двумерной и трехмерной генеральных совокупностей.

Проанализируем выборку из генеральной совокупности (X,Y) объема n:

$$(x_1, y_1), (x_2, y_2), \dots, (x_j, y_j), \dots, (x_n, y_n).$$

Образуем двумерный группированный ряд в виде таблицы, называемой **корреляционной**:

Y \ X	...	x_k	...	m_y
...	...	...	...	...
y_l	...	m_{kl}	...	$m_{\bullet l}$
...	...	...	...	...
m_x	...	$m_{k\bullet}$	...	n

В строке x в возрастающем порядке расположены варианты x_k, а в столбце y варианты y_l. На пересечении столбца x_k и строки y_l находится частота m_{kl}, означающая число точек выборки, равных точке (x_k, y_l).

В столбце m_y помещены частоты одномерного группированного ряда y; а в строке m_x - частоты ряда x, полученные путем суммирования соответствующих частот m_{kl}. Наконец, n равно сумме частот любого из одномерных рядов x или y.

Если зафиксировать X на величине x_k, то получается одномерный группированный ряд со значениями (вариантами) $y_1, \dots, y_l, \dots, y_s$ с частотами $m_{k1}, \dots, m_{kl}, \dots, m_{ks}$ соответственно и с объемом $m_{k\bullet}$ группы x_k. Средние арифметические

$$\bar{y}_k = \frac{1}{m_{k\bullet}} \sum_{l=1}^{s} y_l m_{kl}, \qquad \bar{y}_k^2 = \frac{1}{m_{k\bullet}} \sum_{l=1}^{s} y_l^2 m_{kl},$$

и выборочная дисперсия $s_{yk}^2 = \bar{y}_k^2 - (\bar{y}_k)^2$ этого условного группированного ряда могут служить точечными оценками для условного математического ожидания $MY/x = x_k$ и дисперсии $DY/x = x_k$ генеральной совокупности (X, Y).

Последовательность пар

$$(x_1, \bar{y}_1), (x_2, \bar{y}_2), \dots, (x_k, \bar{y}_k), \dots, (x_r, \bar{y}_r)$$

может выступать в качестве оценки регрессии Y на X. Такая регрессия называется **эмпирической.**

Величина

$$s_y^2 = \frac{\sum_{k=1}^{r} s_{yk}^2 m_{*k}}{\sum_{k=1}^{r} m_{*k}}$$

служит оценкой остаточной дисперсии Y, а величина

$$s_{y\,регр}^2 = \frac{\sum_{l=1}^{s} (\bar{y}_k - \bar{y})^2 m_{k*}}{\sum_{k=1}^{s} m_{k*}}$$

называется **дисперсией эмпирической регрессии** y, которая служит оценкой генеральной дисперсии регрессии

$$Dy_{регр} = M\,[My/x - My]^2 .$$

Можно доказать тождество

$$s_y^2 = s_{y\,регр}^2 + \bar{s}_y^2 .$$

Рассмотрим выборку из генеральной совокупности (x, y, z) объема n:

$$(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n) .$$

Построим трехмерный группированный ряд в виде последовательности корреляционных таблиц, отвечающих различным значениям $z_j, j = \overline{1,t}$:

z_1

y \ x	..	x_k	..	m_y
...		...		...
y_l	..	m_{kl1}	..	m_{*l1}
...		...		...
m_x	..	m_{k*1}	..	n_{**1}

z_j

y \ x	..	x_k	..	m_y
...		...		...
y_l	..	m_{klj}	..	m_{*lj}
...		...		...
m_x	..	m_{k*j}	..	n_{**j}

z_t

y \ x	..	x_k	..	m_y
...		...		...
y_l	..	m_{klt}	..	m_{*lt}
...		...		...
m_x	..	m_{k*t}	..	n_{**t}

Корреляционная таблица z_j представляет собой двумерный группированный ряд, полученный по подвыборке объема $n_j = n_{**j}$, и дает оценку условного двумерного распределения случайной величины (x, y) при фиксированном $z = z_j$. Применяя уже известные формулы, можно получить точечные оценки параметров этого условного распределения. В частности, можно найти выборочный условный коэффициент корре-

ляции при условии $z = z_j$ по формуле

$$r(x, y/z_j) = \frac{\overline{[(x - \overline{x/z_j})(y - \overline{y/z_j})]}}{s_{x/z_j} s_{y/z_j}},$$

где $\overline{x/z_j} = \frac{\sum_{k=1}^{s} x_k m_{k*j}}{n_j}$, $\overline{y/z_j} = \frac{\sum_{l=1}^{r} y_l m_{*lj}}{n_j}$,

$$\overline{[(x - x/z_j)(y - y/z_j)]/z_j} = \frac{\sum_k \sum_l (x_k - \overline{x/z_j})(y_l - \overline{y/z_j}) m_{*lj}}{n_j},$$

$$s_{x/z_j} = \sqrt{\frac{1}{n_j} \sum_{k=1}^{s} (x_k - \overline{x/z_j})^2 m_{k*j}},$$

$$s_{y/z_j} = \sqrt{\frac{1}{n_j} \sum_{l=1}^{r} (y_l - \overline{y/z_j})^2 m_{*lj}} .$$

Эмпирическая регрессия на (x, y) может быть представлена $s \cdot r$ точками $(x_k, y_l, \overline{z_{kl}})$. Условные средние арифметические $\overline{z_{kl}}$ являются оценками условных математических ожиданий Mz/xy при фиксировании двух случайных величин x и y на уровнях x_k и y_l, $k = \overline{1,r}$; $l = \overline{1,s}$, и вычисляются по формулам

$$\overline{z_{kl}} = \frac{\sum_{j=1}^{t} z_j m_{klj}}{m_{kl*}}, \qquad m_{kl*} = \sum_{j=1}^{t} m_{klj} .$$

Другие точечные оценки параметров генеральной совокупности (в частности, при известной форме регрессии) рассматриваются в корреляционном и регрессионном анализе.

2.2. Доверительные области

При малых объемах выборок или подвыборок точечные оценки могут достаточно далеко отклоняться от оцениваемых параметров и поэтому вводится **понятие интервальной оценки параметра генеральной совокупности.**

Определения и понятия интервального оценивания можно перенести на случай векторного параметра $\theta=(\theta_1,\theta_2,...,\theta_m)^T$ с заменой доверительного интервала доверительной областью в соответствующем m-мерном пространстве.

Доверительной областью вектора параметров θ генеральной совокупности называется случайная область, полностью определяемая результатами наблюдений, которая с близкой к единице доверительной вероятностью (надежностью) γ содержит неизвестное значение вектора θ.

Очевидно, что существует бесконечное множество доверительных областей, соответствующих одному и тому же значению γ. Обычно стараются определить доверительные области, имеющие минимальные размеры при данной надежности γ. Часто этому условию удовлетворяют области, симметричные относительно вектора оценок $\hat{\theta}$ параметров θ.

Основную трудность в построении доверительной области представляет определение законов распределений подходящих статистик. В настоящее время эти вопросы достаточно хорошо разработаны только для нормального распределения наблюдаемых случайных величин.

2.2.1. Доверительная область для вектора математического ожидания

Пусть по результатам n наблюдений из генеральной совокупности X с k-мерным нормальным распределением $N_k(\mu,\Sigma)$ найдены вектор средних x и несмещенная оценка $\hat{S}$ ковариационной матрицы Σ.

Требуется найти с надежностью γ доверительную область для k-мерного вектора генеральных средних μ.

Предположим, что ковариационная матрица известна. Найдем такую подходящую статистику, чтобы ее распределение было известным и по которой однозначно можно определить доверительную область.

Напомним, что для одномерной нормально распределенной генеральной совокупности доверительный интервал для μ определяется из формул

$$|t|<\Phi^{-1}(\gamma), \qquad t=\frac{\bar{x}-\mu}{\sigma}\sqrt{n},$$

где статистика t подчиняется стандартному закону $N_1(0,1)$. Последнее равенство можно переписать в виде

$$t^2=n(x_1-\mu_1)(\sigma_{11})^{-1}(x_1-\mu_1),$$

полагая $x_1=x$, $\mu_1=\mu$ и $\sigma_{11}=\sigma^2$ при k=1.

Полученное равенство обобщается на случай k-мерной совокупности следующим образом:

$$t^2=n(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu).$$

Известно, что статистика t^2 при k=1 имеет хи-квадрат распределение с числом степеней свободы, равным $\nu=1$. Статистика t^2 для любого k также распределена по хи-квадрат с числом степеней свободы $\nu=k$.

Таким образом, с надежностью γ можно утверждать, что вектор μ накрывается доверительной областью, задаваемой неравенством

$$n(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)\le(\chi^2)^{-1}(1-\gamma).$$

Пусть теперь ковариационная матрица Σ не известна.

Чтобы при k=1 построить доверительный интервал для μ, используют статистику

$$t=\frac{\bar{x}-\mu}{\hat{s}}\sqrt{n},$$

которая имеет t-распределение с $\nu=n-1$ степенями свободы. Равенство можно переписать в эквивалентной форме

$$t^2=n(\bar{x}-\mu)(\hat{S}^2)^{-1}(\bar{x}-\mu).$$

По аналогии строится **статистика T^2 Хотеллинга**, которую используют при построении доверительной области для вектора средних μ:

$$T^2=n(\bar{x}-\mu)\hat{S}^{-1}(\bar{x}-\mu), \qquad (2.1)$$

где $\hat{S}^{-1}$ - матрица, обратная ковариационной матрице $\hat{S}$.

Учитывая, что F и T^2 распределения связаны соотношением

$$T^2_{\alpha,k,n-k}=\frac{k(n-1)}{n-k}F_{\alpha,k,n-k}, \qquad (2.2)$$

получим уравнение поверхности, ограничивающей доверительную область k генеральных средних с надежностью γ

$$(\bar{x}-\mu)^T\hat{S}^{-1}(\bar{x}-\mu)=\frac{k(n-1)}{n(n-k)}F_{\alpha,k,n-k},$$

где $F_{\alpha,k,n-k}$ - точка F-распределения, соответствующая уровню значимости α и числам степеней свободы k и n-k.

Уравнения (2.1) и (2.2) определяют k-мерный эллипсоид (эллипс при k=2) с центром $\bar{x}$, так как его левая часть представляет положительно определенную квадратичную форму относительно μ.

Пример 2.1

В таблице приводятся данные о фонде заработной платы работников централизованных бухгалтерий x_1 и товарообороте обслуживаемых аптек x_2.

Фонд заработной платы, тыс.руб.	x_1	20	36	28	51	70	45	30	56
Товарооборот, млн.руб.	x_2	3,5	5,4	2,7	9,8	10,1	6,2	2,4	9,5

Найти оценки математических ожиданий, дисперсии и коэффициента корреляции, доверительную область для вектора математических ожиданий с надежностью $\gamma = 0{,}95$.

Решение

Найдем средние арифметические $\bar{x}_1 = 42$ и $\bar{x}_2 = 6{,}2$ и перейдем к центрированным величинам $u_{ij} = x_{ij} - \bar{x}_j$.

$$U^{\tau} = \begin{pmatrix} -22 & -6 & -14 & 9 & 28 & 3 & -12 & 14 \\ -2{,}7 & -0{,}8 & -3{,}5 & 3{,}6 & 3{,}9 & 0 & -3{,}8 & 3{,}3 \end{pmatrix}.$$

Откуда

$$U^{\tau}U = \begin{pmatrix} -22 & -6 & \dots & 14 \\ -2{,}7 & -0{,}8 & \dots & 3{,}3 \end{pmatrix} \begin{pmatrix} -22 & -2{,}7 \\ -6 & -0{,}8 \\ \dots & \dots \\ 14 & 3{,}3 \end{pmatrix} = \begin{pmatrix} 1930 & 346{,}6 \\ 346{,}6 & 73{,}68 \end{pmatrix},$$

$$\hat{S} = \frac{1}{n-1} U^{\tau}U = \frac{1}{7} \begin{pmatrix} 1930 & 346{,}6 \\ 346{,}6 & 73{,}68 \end{pmatrix} = \begin{pmatrix} 275{,}714 & 49{,}514 \\ 49{,}514 & 10{,}526 \end{pmatrix}.$$

Итак, оценки дисперсий и средних квадратических отклонений следующие: $\hat{s}_1^2 = 275{,}714$; $\hat{s}_1 = 16{,}605$; $\hat{s}_2^2 = 10{,}526$; $\hat{s}_2 = 3{,}244$.

Найдем выборочный коэффициент корреляции

$$r = \frac{49{,}514}{16{,}605 \cdot 3{,}244} = 0{,}919.$$

Обратная матрица к $\hat{S}$ будет иметь вид

$$\hat{S}^{-1} = \begin{pmatrix} 0{,}023365 & -0{,}109909 \\ -0{,}109909 & 0{,}6120157 \end{pmatrix}.$$

Тогда согласно уравнению (2.2)

$$\begin{pmatrix} 42-\mu_1 & 6{,}2-\mu_2 \end{pmatrix} \begin{pmatrix} 0{,}023365 & -0{,}109909 \\ -0{,}109909 & 0{,}6120157 \end{pmatrix} \begin{pmatrix} 42-\mu_1 \\ 6{,}2-\mu_2 \end{pmatrix} = \frac{2 \cdot 7}{8 \cdot 6} 5{,}14,$$

где $F(0{,}05;2;6) = 5{,}14$ находим по таблице для $\alpha = 1 - \gamma = 0{,}05$ и чисел степеней свободы $\nu_1 = 2$ и $\nu_2 = 6$.

После преобразований получаем уравнение эллипса

$$0{,}0234(42-\mu_1)^2 + 0{,}0242(42-\mu_1)(6{,}2-\mu_2) + 0{,}612(6{,}2-\mu_2)^2 = 1{,}499,$$

которое определяет границы доверительной области для вектора $(\mu_1, \mu_2)^{\tau}$.

Пусть вектор x подчиняется k-мерному распределению $N_k(\mu, \Sigma)$, где Σ - ковариационная матрица, C при $(k \times l)$ - матрица $(l \le k)$ ранга l. Тогда $C^{\tau}X$ подчиняется l-мерному нормальному распределению $N_l(C^T\mu; C^T\Sigma C)$, так как линейные комбинации нормально-распределенных величин также распределены нормально.

В этом случае статистика T^2 имеет вид

$$T^2 = n(C^{\tau}\bar{x} - C^{\tau}\mu)(C^{\tau}SC)^{-1}(C^{\tau}\bar{x} - C^{\tau}\mu) \qquad (2.3)$$

со значением, отвечающим вероятности $\alpha = 1 - \gamma$

$$T^2_{\alpha,l,n-l} = \frac{l(n-1)}{n-l} F_{\alpha,l,n-l}.$$

Поскольку вектор $C^{\tau}\mu$ содержит l генеральных средних, то в отличие от уравнения (2.2) теперь число степеней свободы равно l и n–l. В частном случае, когда C - единичная матрица порядка k, уравнение (2.3) сводится к уравнению (2.2) и число степеней свободы становится равным, как и прежде, k и $(n$–$k)$.

Использование линейных комбинаций компонент вектора μ позволяет расширить область применения статистики T^2 Хотеллинга при интервальном оценивании в задачах сравнения.

С помощью линейных комбинаций можно, например, найти совместные доверительные интервалы или проверить гипотезу относительно первых $(l \le k)$ средних значений генеральной совокупности. Для этого достаточно принять

$$C^{\tau} = \left[\begin{matrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & & & \dots & \dots & & \dots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{matrix} \right\} l \Bigg], \quad \begin{matrix} \underbrace{}_{l} & \underbrace{}_{k-l} \end{matrix}$$

Поскольку C^τ имеет размерность $(l \times k)$ и ранг l, то вектор $C^\tau \mu$ имеет размерность $(l \times 1)$ и содержит l $(l \le k)$ генеральных средних.

Чтобы построить доверительный интервал для генерального среднего $\mu_j (j = 1,2,\ldots,k)$, достаточно принять, что C_j есть j-й столбец единичной матрицы размерности k. Тогда $C_j^\tau \mu = \mu_j$ и статистика Хотеллинга

$$T^2 = \frac{n[C_j^\tau (\bar{x} - \mu)]^2}{C_j^\tau SC_j}$$

имеют распределение, зависящее от чисел степеней свободы l и $n-1$.

Таким образом, согласно уравнению (2.3) доверительная область для $C^\tau \mu$ с надежностью γ будет ограничена поверхностью

$$(C^\tau \bar{x} - C^\tau \mu)^\tau (C^\tau SC)^{-1} (C^\tau \bar{x} - C^\tau \mu) = \frac{l(n-1)}{n(n-l)} F_{\alpha,l,n-l}.$$

В частности, с надежностью γ доверительные границы для линейной комбинации $C_j^\tau \mu$ (где C_j есть столбец матрицы С) определяются как

$$C^\tau \bar{x} \pm \sqrt{\frac{1}{n} C_j^\tau SC_j F_{\alpha,l,n-l}}.$$

Пример 2.2

По данным задачи 2.1 с помощью линейных комбинаций найти с надежностью $\gamma = 0,95$ интервальные оценки генеральных средних μ_1 и μ_2.

Решение

Для нашего примера $\alpha = 1 - \gamma = 0,05$, $\nu_1 = 1$, $\nu_2 = n - 1 = 7$. Согласно таблице F-распределения $F(0,05; 1; 7) = 5,59$.

Для построения интервальной оценки средней μ_1 примем $C_1 = (1,0)^\tau$, так что $C_1 x = 42$; $C_1^\tau SC_1 = \hat{S}_1^2 = 275,714$. Тогда границы доверительного интервала для μ_1 имеют вид

$$\mu_1 \in [42 \pm \sqrt{\frac{1}{8} 275,714 \cdot 5.59}] = [42 \pm 13,880].$$

Откуда

$$28,120 \le \mu_1 \le 55,880.$$

Для построения интервальной оценки генерального среднего μ_2 принимаем $C_2 = (0,1)^\tau$, откуда $C_2^\tau \bar{x} = 6,2$; $C_2^\tau SC_2 = 10,526$.

Тогда с надежностью $\gamma = 0,95$ границы доверительного интервала таковы:

$$6,2 \pm \sqrt{\frac{1}{8} \cdot 10,526 \cdot 5,59} = 6,2 \pm 2,712.$$

Откуда

$$3,488 \le \mu_2 \le 8,912.$$

2.2.2. Определение совместной доверительной области для математического ожидания и дисперсии

Один из простых подходов в построении многомерной доверительной области состоит в определении таких интервалов $I_1(\theta_1), I_2(\theta_2), \ldots, I_m(\theta_m)$ для координат $\theta_1, \theta_2, \ldots, \theta_m$ вектора параметров θ, для которых вероятность одновременного накрытия всех $\theta_1, \theta_2, \ldots, \theta_m$ соответствующими интервалами была бы не меньше заданного значения γ. Таким образом, речь идет о нахождении прямоугольной доверительной области для вектора θ соответствующей надежности, не меньше, чем γ.

Введем событие $A_i = \{\theta_i \in I_i(\theta_i)\} (i = 1,2,\ldots,m)$ и обозначим прямоугольную область в m-мерном пространстве, образованную интервалами $I_1(\theta_1), I_2(\theta_2), \ldots, I_m(\theta_m)$ через $I(\theta)$. На основании свойств вероятностей получим

$$P(\theta \in I(\theta)) = P(\bigcap_{i=1}^{m} A_i) = 1 - P(\bigcup_{i=1}^{m} \bar{A}_i) \ge 1 - \sum_{i=1}^{m} P(\bar{A}_i),$$

где $\bigcap_{i=1}^{m} A_i$ - пересечение событий A_i;

$\bigcup_{i=1}^{m} \bar{A}_i$ - объединение событий $\bar{A}_i$, противоположных событиям A_i.

Из этого следует, что для определения доверительной области достаточно найти доверительные интервалы $I_1(\theta_1), I_2(\theta_2), \ldots, I_m(\theta_m)$ для координат вектора θ, соответствующие надежности $[1 - \frac{1}{m}(1 - \gamma)]$. Тогда вероятность всех параметров $\theta_1, \theta_2, \ldots, \theta_m$ будет не меньше γ.

Пусть для нормальной генеральной совокупности x с неизвестными

параметрами μ и σ взята случайная выборка $x_1, x_2, \ldots, x_n$ с объемом n. Требуется с надежностью γ найти (совместную) доверительную область для μ и σ, т.е. для двумерного вектора параметров $\theta = (\mu, \sigma)^T$.

По данным выборки найдем среднюю арифметическую $\bar{x}$ и среднеквадратическое отклонение s. Тогда доверительные интервалы отдельно для математического ожидания μ и среднеквадратического отклонения σ, отвечающие надежности γ, имеют вид

$$\bar{x} - t_\alpha \frac{s}{\sqrt{n-1}} \le \mu \le \bar{x} + t_\alpha \frac{s}{\sqrt{n-1}} \qquad (2.4)$$

и

$$\sqrt{\frac{ns^2}{\chi^2(\alpha/2; n-1)}} \le \sigma \le \sqrt{\frac{ns^2}{\chi^2(1-\alpha/2; n-1)}}, \qquad (2.5)$$

где $\alpha = 1 - \gamma$, t_α и χ^2 находятся по таблицам распределения Стьюдента и хи-квадрат для числа степеней свободы $\nu = n - 1$ и вероятностей α и $(\frac{\alpha}{2}, 1 - \frac{\alpha}{2})$ соответственно.

Доверительную область для вектора θ можно определить формулой

$$\Delta_\alpha(\theta) = \left[(\mu, \sigma): \left\{ \begin{array}{l} \bar{x} - t\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + t\frac{\sigma}{\sqrt{n}} \\ \sqrt{\frac{ns^2}{\chi_2^2}} \le \sigma \le \sqrt{\frac{ns^2}{\chi_1^2}} \end{array} \right\} \right], \qquad (2.6)$$

где $\chi_2^2 = \chi^2(\frac{\alpha}{2}; n-1)$, $\chi_1^2 = \chi^2(1 - \frac{\alpha}{2}; n-1)$.

Эта область представляет собой трапецию.

Для построения такой области с заданной доверительной вероятностью γ можно руководствоваться следующим.

Так как при нормальном распределении генеральной совокупности x оценки $\bar{x}$ и s независимы, причем $\bar{x}$ имеет нормальное распределение $N(\mu, \frac{\sigma}{\sqrt{n}})$, а $\frac{ns^2}{\sigma^2}$ - распределение χ^2 с $\nu = n - 1$ степенями свободы, то

$$P(\theta \in \Delta_\gamma(\theta)) = P\{[|\bar{x} - \mu| \le t\sigma] \cdot [\sqrt{\frac{ns^2}{\chi_2^2}} \le \sigma \le \sqrt{\frac{ns^2}{\chi_1^2}}]\} = P[|\bar{x} - \mu| \le \frac{t\sigma}{\sqrt{n}}] \cdot$$

$$\cdot P[\sqrt{\frac{ns^2}{\chi_2^2}} \le \sigma \le \sqrt{\frac{ns^2}{\chi_1^2}}] = \Phi(t) P(\chi_1^2 \le \chi^2 \le \chi_2^2).$$

Определив t и χ_1^2, χ_2^2 из условий $\Phi(t) = C\sqrt{\gamma}$, $P(\chi_1^2 \le \chi^2 \le \chi_2^2) = \frac{\sqrt{\gamma}}{C}$, где C - произвольное число, принадлежащее интервалу $(\sqrt{\gamma}; \frac{1}{\gamma})$, найдем доверительную область вектора параметров $\theta = \left(\frac{\mu}{\sigma}\right)$, соответствующую надежности γ.

Пример 2.3

По результатам контроля n=14 изделий найдено, что средняя высота сальниковой камеры равна $\bar{x} = 88$ мм, а $s = 0{,}96$ мм. Требуется определить доверительную область для вектора параметров с надежностью $\gamma = 0{,}95$.

Решение

Для сравнения определим сначала с надежностью $\gamma = 0{,}95$ доверительные интервалы для математического ожидания μ и среднеквадратического отклонения σ.

По таблице t-распределения для числа степеней свободы $\nu = n - 1 = 13$ и уровня значимости $\alpha = 1 - \gamma = 0{,}05$ находим $t_\gamma = 2{,}160$. Согласно выражению (2.4) имеем

$$88 - 2{,}160\frac{0{,}96}{\sqrt{13}} \le \mu \le 88 + 2{,}160\frac{0{,}96}{\sqrt{13}},$$

откуда

$$87{,}425 \le \mu \le 88{,}575.$$

По таблице χ^2 - распределения для числа степеней свободы $\nu = 13$ и вероятности $\frac{1-\gamma}{2} = 0{,}025$ найдем верхнюю границу $\chi_2^2 = 24{,}736$ доверительного интервала для χ^2. Отсюда нижняя граница для σ равна (2.5):

$$\sqrt{\frac{ns^2}{\chi_2^2}} = \sqrt{\frac{14 \cdot 0{,}96^2}{24{,}736}} = 0{,}722.$$

Для числа степеней свободы $\nu = 13$ и вероятности $\frac{1+\gamma}{2} = 0{,}975$ найдем $\chi_1^2 = 5{,}008$. Отсюда верхняя граница σ имеет вид (2.5):

$$\sqrt{\frac{ns^2}{\chi_1^2}} = \sqrt{\frac{14 \cdot 0{,}96^2}{5{,}008}} = 1{,}605.$$

Таким образом,

$$0{,}722 \le \sigma \le 1{,}605.$$

Чтобы определить доверительную область для вектора (μ,σ) с надежностью $\gamma = 0{,}95$, найдем по таблице интегральной функции Лапласа t=2,24 из условия $\Phi(t) = \sqrt{\gamma} = 0{,}9747$.

По таблице χ^2-распределения находим

$$\chi_1^2 = \chi^2\left(\frac{1+\sqrt{\gamma}}{2}, n-1\right) = \chi^2(0{,}9873;13) = 4{,}285;$$

$$\chi_2^2 = \chi^2\left(\frac{1-\sqrt{\gamma}}{2}, n-1\right) = \chi^2(0{,}0127;13) = 27{,}090.$$

Теперь по формуле (2.6) найдем доверительную область:

$$88 - 0{,}599\sigma \le \mu \le 88 + 0{,}599\sigma; \quad 0{,}690 \le \sigma \le 1{,}735.$$

Для сравнения найдем прямоугольную доверительную область с коэффициентом доверия не меньше $\gamma = 0{,}95$. Для этого достаточно по формулам (2.4) и (2.5) найти доверительные интервалы для μ и σ, соответствующие надежности $\gamma_1 = 1 - \frac{1-\gamma}{2} = 0{,}975$.

В результате получим

$$\begin{cases} 87{,}357 \le \mu \le 88{,}643 \\ 0{,}690 \le \sigma \le 1{,}738 \end{cases}.$$

2.3. Гипотезы о параметрах многомерной нормально распределенной генеральной совокупности

2.3.1. Сравнение вектора генеральных средних со стандартом

Рассматривается k-мерная генеральная совокупность распределения $N_k(\mu,\Sigma), |\Sigma| \neq 0$. По выборке объема n из этой генеральной совокупности определены вектор средних арифметических и несмещенная оценка $\hat{s}$ ковариационной матрицы Σ.

Если ковариационная матрица известна, то для проверки гипотезы о равенстве вектора генеральных средних стандартному (заданному) значению, $H_0: \mu = \mu_0$, против альтернативы $H_1: \mu \neq \mu_0$ употребляется статистика

$$\chi^2 = n(\bar{x} - \mu_0)^T \Sigma^{-1} (\bar{x} - \mu_0),$$

имеющая распределение хи-квадрат с числом степеней свободы $\nu = k$ при справедливости гипотезы H_0.

Если же ковариационная матрица Σ не известна, то можно воспользоваться статистикой Хотеллинга

$$T^2 = n(\bar{x} - \mu_0)^T \hat{S}^{-1} (\bar{x} - \mu_0).$$

Как уже было сказано, при истинности H_0 имеет место равенство

$$T^2_{\alpha,k,n-k} = \frac{k(n-1)}{n-k} F_{\alpha,k,n-k}.$$

Поэтому для критической области

$$T^2 > T^2_{\text{кр}} = T^2_{\alpha,k,n-k}$$

можно вычислить $T^2_{\text{кр}}$ с помощью таблиц F-распределения.

Напоминаем, что применяемые статистики являются обобщениями соответствующих статистик для одномерной случайной величины. Это следует из того факта, что квадрат нормированной нормально распределенной (одномерной) случайной величины имеет распределение χ^2 с 1 степенью свободы, а квадрат случайной величины, имеющей t-распределение Стьюдента, распределен по F-распределению с числом степеней свободы 1 и n–1.

Пример 2.4

По данным годовых отчетов n=10 промышленных предприятий (табл. 2.1) с уровнем значимости $\alpha = 0{,}05$ проверить гипотезу о соответствии средних уровней экономических показателей работы группы предприятий (x_1 млн руб. - объем валовой продукции, x_2 тыс.руб./чел. – производительность труда, x_3 млн руб. – себестоимость товарной продукции) контрольным значениям при известных значениях указанных в таблице параметров генеральной совокупности, которая является нормально распределенной.

Таблица 2.1

Исходная информация для сравнения параметров

№ i	1	2	3	4	5	6	7	8	9	10	σ_j^2	Парная корреляция	Контрольные значения μ_{0j}
x_{i1}	3	4	5	5	5	5	6	7	10	10	4	$\rho_{12}=0{,}14\sqrt{5}$	7
x_{i2}	1,2	1,2	1,4	1,2	1,2	1,5	1,5	1,3	1,7	1,6	0,2	$\rho_{23}=0{,}12\sqrt{5}$	1,5
x_{i3}	2,1	2,8	3,2	4,5	4,8	4,9	5,5	6,5	8,5	8,2	4	$\rho_{31}=0{,}9$	5

Решение

Исходя из условия задачи требуется проверить гипотезу $H_0: \mu = \mu_0$ против $H_1: \mu \neq \mu_0$, причем $\mu_0 = (7;1{,}5;5)^T$. Здесь матрица Σ генеральных коэффициентов ковариации считается известной:

$$\sigma_1 = 2; \sigma_{12} = \sigma_{21} = \rho_{12}\sigma_1\sigma_2 = 0{,}14\sqrt{5}\cdot 2\sqrt{0{,}2} = 0{,}28;$$

$$\sigma_{13} = \sigma_{31} = \rho_{31}\sigma_3\sigma_1 = 0{,}9\cdot 2\cdot 2 = 3{,}6;$$

$$\sigma_2 = \sqrt{0{,}2}; \sigma_{23} = \sigma_{32} = \rho_{23}\sigma_2\sigma_3 = 0{,}12\sqrt{5}\cdot 2\sqrt{0{,}2} = 0{,}24; \sigma_3 = 2.$$

Таким образом

$$\Sigma = \begin{pmatrix} 4 & 0{,}28 & 3{,}6 \\ 0{,}28 & 0{,}2 & 0{,}24 \\ 3{,}6 & 0{,}24 & 4 \end{pmatrix}.$$

Найдем Σ^{-1}, воспользовавшись следующими обозначениями и известными формулами: $|\Sigma|$ - определитель матрицы Σ, M_{ml} - дополнительный к элементу σ_{ml} минор матрицы Σ, полученный вычеркиванием ее m-й строки и l-го столбца, σ_{ml}^{-1} - элемент обратной матрицы Σ^{-1}, $\sigma_{ml}^{-1} = (-1)^{m+l} M_{ml} : |\Sigma|$.

Будем иметь

$$|\Sigma| = 4^2\cdot 0{,}2 + 2\cdot 0{,}28\cdot 0{,}24\cdot 3{,}6 - 3{,}6^2\cdot 0{,}2 - 0{,}28^2\cdot 4 - 4\cdot 0{,}24^2 = 0{,}54784;$$

$$\sigma_{11}^{-1} = (-1)^2\begin{vmatrix} 0{,}2 & 0{,}24 \\ 0{,}24 & 4 \end{vmatrix} : 0{,}54784 = (4 - 0{,}24^2) : 0{,}54784 = 1{,}3551401;$$

$$\sigma_{12}^{-1} = (-1)^3\begin{vmatrix} 0{,}28 & 0{,}24 \\ 3{,}6 & 4 \end{vmatrix} : 0{,}54784 = -0{,}4672897 = \sigma_{21}^{-1};$$

$$\sigma_{13}^{-1} = (-1)^4\begin{vmatrix} 0{,}28 & 0{,}2 \\ 3{,}6 & 0{,}24 \end{vmatrix} : 0{,}54784 = -1{,}1915887 = \sigma_{31}^{-1};$$

$$\sigma_{22}^{-1} = (-1)^4\begin{vmatrix} 4 & 3{,}6 \\ 3{,}6 & 4 \end{vmatrix} : 0{,}54784 = 5{,}5490654;$$

$$\sigma_{23}^{-1} = (-1)^5\begin{vmatrix} 4 & 0{,}28 \\ 3{,}6 & 0{,}24 \end{vmatrix} : 0{,}54784 = 0{,}0876168 = \sigma_{32}^{-1};$$

$$\sigma_{33}^{-1} = (-1)^6\begin{vmatrix} 4 & 0{,}28 \\ 0{,}28 & 0{,}2 \end{vmatrix} : 0{,}54784 = 1{,}3171728.$$

Таким образом, обратная к ковариационной матрица имеет вид

$$\Sigma^{-1} = \begin{pmatrix} 1{,}3551401 & -0{,}4672897 & -1{,}1915887 \\ -0{,}4672897 & 5{,}5490654 & 0{,}0876168 \\ -1{,}1915887 & 0{,}0876168 & 1{,}3171728 \end{pmatrix}.$$

Критическая область задается следующим неравенством

$$n(\bar{x} - \mu_0)^{\tau}\Sigma^{-1}(\bar{x} - \mu_0) > \chi^2(\alpha, k).$$

Подставляя в него n=10, $\mu_0 = (7;1{,}5;5)^{\tau}$, Σ^{-1}, указанные выше, а также $\chi^2(0{,}05;\ 3) = 7{,}815$, получим неравенство относительно средних арифметических $\bar{x}_1, \bar{x}_2, \bar{x}_3$ - компонент вектора x.

Далее находим наблюдаемое значение вектора x:

$$\bar{x} = (6;1{,}4;5{,}1)^{\tau}.$$

Подставляя его в левую часть неравенства, получим

$$\chi^2_{набл} = 10\begin{pmatrix} -1 \\ -0{,}1 \\ 0{,}1 \end{pmatrix}^{\tau}\begin{pmatrix} 1{,}3551401 & -0{,}4672897 & -1{,}1915887 \\ -0{,}4672897 & 5{,}5490654 & 0{,}0876168 \\ -1{,}1915887 & 0{,}0876168 & 1{,}3171728 \end{pmatrix}\times\begin{pmatrix} -1 \\ -0{,}1 \\ 0{,}1 \end{pmatrix} =$$

$$= 10(-1{,}4275700;\ -0{,}0788551;\ 1{,}3145443)\times(-1;\ -0{,}1;\ 0{,}1)^{\tau} = 15{,}669099 = 15{,}669.$$

Так как $\chi^2_{набл} > \chi^2(0{,}05;3)$, т.е. $\bar{x}_{набл}$ принадлежит критической области, гипотеза H_0 отвергается с вероятностью ошибки 0,05. Таким образом, средние уровни экономических показателей работы группы предприятий не соответствуют контрольным цифрам. Анализ причин и последствий обнаруженного несоответствия выясняется планирующими и управляющими органами.

Пример 2.5

В условиях примера 2.4 решить задачу при неизвестных параметрах - генеральных дисперсиях и коэффициентах корреляции.

Решение

Требуется проверить гипотезу $H_0: \mu = \mu_0$ против $H_1: \mu \neq \mu_0$ при неизвестной матрице Σ генеральных коэффициентов ковариации. Прежде всего находим оценку ковариационной матрицы

$$\hat{S} = \frac{1}{9}\begin{pmatrix} 50 & 2{,}8 & 42{,}8 \\ 2{,}8 & 0{,}28 & 2{,}16 \\ 42{,}8 & 2{,}16 & 41{,}68 \end{pmatrix},$$

где, например,

$$\bar{x}_1 = 6; \bar{x}_2 = 1{,}4;$$

$$\hat{s}_{11} = \frac{1}{9}((3-6)^2 + (4-6)^2 + 4(5-6)^2 + (7-6)^2 + 2(10-6)^2) = \frac{1}{9}50;$$

$$\hat{s}_{12} = ((3-6)(1{,}4-1{,}4) + (4-6)(1{,}2-1{,}4) + (5-6)(1{,}4-1{,}4) + (5-6)(1{,}2-1{,}4)2 +$$

$$+ (5-6)\times(1{,}5-1{,}4) + (6-6)(1{,}5-1{,}4) + (7-6)(1{,}3-1{,}4) + (10-6)(1{,}7-1{,}4) +$$

$$+ (10-6)(1{,}6-1{,}4) = \frac{1}{9}2{,}8.$$

Обратная матрица к матрице несмещенных оценок ковариаций имеет вид

$$\hat{S}^{-1} = 9\begin{pmatrix} 7{,}0048 & -24{,}256 & -5{,}936 \\ -24{,}256 & 252{,}16 & 11{,}84 \\ -5{,}936 & 11{,}84 & 6{,}16 \end{pmatrix} : 28{,}2624 .$$

Критическая область задается следующим неравенством

$$n(\bar{x}-\mu_0)^{\tau}\hat{S}^{-1}(\bar{x}-\mu_0) > \frac{k(n-1)}{n-k}F_{\alpha,k,n-k} .$$

Подставляя в него $F_{0,05;3;7}$ = 4,35, полученное по таблицам F-распределения, и значения остальных символов, будем иметь после некоторых преобразований

$$10\begin{pmatrix} -10 \\ -0{,}1 \\ 0{,}1 \end{pmatrix}^{\tau}\begin{pmatrix} 7{,}0048 & -24{,}256 & -5{,}936 \\ -24{,}256 & 252{,}16 & 11{,}84 \\ -5{,}936 & 11{,}84 & 6{,}16 \end{pmatrix}\begin{pmatrix} -1 \\ -0{,}1 \\ 0{,}1 \end{pmatrix} > 28{,}2624\cdot\frac{3}{7}\cdot 4{,}35,$$

или 705,6872>368,82432. Так как 3988,104>368,82432 истинно, нулевая гипотеза отвергается с вероятностью ошибки 0,05.

2.3.2. Сравнение двух генеральных совокупностей

Назовем две генеральные совокупности однородными, если кроме одних и тех же признаков эти совокупности имеют одинаковые законы распределения вероятностей.

Рассмотрим две нормально распределенные совокупности X и Y. Их распределения полностью определяются заданием параметров μ_x, Σ_x и μ_y, Σ_y. Следовательно, для проверки однородности этих совокупностей достаточно сравнить их ковариационные матрицы Σ_x и Σ_y. Затем, в случае принятия гипотезы о равенстве этих ковариационных матриц, сравнить генеральные средние μ_x и μ_y совокупностей.

Для сравнения матриц генеральных коэффициентов ковариации проверяется гипотеза $H_0: \Sigma_x = \Sigma_y$ против $H_1: \Sigma_x \neq \Sigma_y$ с уровнем значимости α на основе выборок из совокупностей соответственно объемов n_x и n_y.

В качестве статистики критерия проверки берется случайная величина

$$W = ba,$$

где $b = 1 - \left(\frac{1}{n_x-1} + \frac{1}{n_y-1} - \frac{1}{n_x+n_y-2}\right)\frac{2k^2+3k-1}{6(k+1)}$;

$$a = (n_x+n_y-2)\ln\left|\hat{S}_{xy}\right| - ((n_x-1)\ln\left|\hat{S}_x\right| + (n_y-1)\ln\left|\hat{S}_y\right|);$$

$\hat{S}_x$ - несмещенная оценка ковариационной матрицы с элементами

$$\hat{s}_{xml} = \frac{1}{n_x-1}\sum_{i=1}^{n_x}(x_{im}-\bar{x}_m)(x_{il}-\bar{x}_l);\ m,l = 1 \div k;$$

$\hat{S}_y$ - несмещенная оценка ковариационной матрицы с элементами

$$\hat{s}_{yml} = \frac{1}{n_y-1}\sum_{i=1}^{n_y}(y_{im}-\bar{y}_m)(y_{il}-\bar{y}_l);\ m,l = 1 \div k;$$

$\hat{S}_{xy} = \frac{1}{n_x+n_y-2}[(n_x-1)\hat{S}_x + (n_y-1)\hat{S}_y]$ - несмещенная оценка одной и той же ковариационной матрицы $\Sigma_x = \Sigma_y$.

При справедливости гипотезы H_0, достаточно больших n_x и n_y и достаточно малой величине

$$C = \frac{k(k+1)}{48b^2}\{(k-1)(k+2)\left[\frac{1}{(n_x-1)^2} + \frac{1}{(n_y-1)^2} + \frac{1}{(n_x-1)^2+(n_y-1)^2}\right] - 6(1-b)^2\}$$

статистика W аппроксимируется распределением χ^2 с числом степеней свободы $\nu = \frac{k(k+1)}{2}$.

Таким образом, критическая область имеет вид

$$W > W_{кр} = \chi^2(\alpha; \frac{k(k+1)}{2}).$$

Если $W_{набл} = b\cdot a$ попадает в критическую область $(W_{набл} > W_{кр})$, то гипотеза $H_0: \Sigma_x = \Sigma_y$ отвергается с вероятностью ошибки α.

Тогда считается доказанным, что ковариационные матрицы Σ_x и Σ_y неодинаковы, и, следовательно, генеральные совокупности неоднородны. Задача проверки однородности совокупностей решена.

Если $W_{набл}$ не попало в критическую область $(W_{набл} \le W_{кр})$, то гипотеза $H_0: \Sigma_x = \Sigma_y$ не отвергается. Так как других выборок не производится, то гипотеза принимается, т.е. считается, что ковариационные матрицы Σ_x и Σ_y одинаковы. При таком условии можно сравнить генеральные средние, т.е. проверить гипотезу $H_0: \mu_x = \mu_y$ против $H_1: \mu_x \neq \mu_y$ с уровнем значимости α.

Для проверки применяется статистика Хотеллинга вида

$$T^2 = \frac{n_x n_y}{n_x+n_y}(\bar{x}-\bar{y})^{\tau}\hat{S}_{xy}^{-1}(\bar{x}-\bar{y}).$$

Если гипотеза $H_0: \mu_x = \mu_y$ справедлива, то статистики T^2 и F связаны формулой:

$$T^2_{\text{кр}} = \frac{(n_x+n_y-2)k}{n_x+n_y-k-1} F_{\alpha;k;n_x+n_y-k-1},$$

где $F_{\alpha;k;n_x+n_y-k-1}$ находится по таблицам F-распределения Фишера-Снедекора с числом степеней свободы в числителе $\nu_1 = k$ и в знаменателе $\nu_2 = n_x + n_y - k - 1$. Критическая область имеет вид $T^2 > T^2_{kp}$.

Если гипотеза $H_0: \mu_x = \mu_y$ отвергается с вероятностью ошибки α, то считается доказанной неоднородность генеральных совокупностей X и Y. Если же гипотеза $H_0: \mu_x = \mu_y$ не отвергается, то, принимая эту гипотезу, мы считаем, что генеральные совокупности однородны.

Пример 2.6

Проверить влияние способа организации производства на экономические показатели группы предприятий на основе следующих выборочных данных (табл.2.2).

Решение

Будем предполагать, что приведенные в табл.2.2 выборки взяты из двумерных нормально распределенных генеральных совокупностей $x = (x_1, x_2)'$ и $y = (y_1, y_2)'$ с неизвестными параметрами

$$\mu_x = \begin{pmatrix} \mu_{x1} \\ \mu_{x2} \end{pmatrix}; \ \Sigma_x = \begin{pmatrix} \sigma^2_{x1} & \sigma_{x1}\sigma_{x2}\cdot\rho_{x1x2} \\ \sigma_{x1}\sigma_{x2}\cdot\rho_{x1x2} & \sigma^2_{x2} \end{pmatrix}$$

и

$$\mu_y = \begin{pmatrix} \mu_{y1} \\ \mu_{y2} \end{pmatrix}; \ \Sigma_y = \begin{pmatrix} \sigma^2_{y1} & \sigma_{y1}\sigma_{y2}\cdot\rho_{y1y2} \\ \sigma_{y1}\sigma_{y2}\cdot\rho_{y1y2} & \sigma^2_{y2} \end{pmatrix}.$$

Таблица 2.2

Фондовооруженность и производительность труда для двух способов организации производства на предприятиях (в условных единицах)

1-й способ (X)		2-й способ (Y)	
Фондовооруженность (x_1)	Производительность (x_2)	Фондовооруженность (y_1)	Производительность (y_2)
56	52	58	74
51	64	56	75
52	63	62	71
72	68	69	83
67	59	62	78
51	62	65	72
52	64	59	80
50	68	71	70
59	70	38	60
60	70	47	65
-	-	74	60
-	-	59	76

Проверим на уровне значимости $\alpha = 0{,}05$ гипотезу о равенстве ковариационных матриц Σ_x и Σ_x. Вычислим оценки основных параметров двух генеральных совокупностей:

$$\bar{x} = \begin{pmatrix} 57 \\ 64 \end{pmatrix}; \quad \hat{S}_x = \frac{1}{9}\begin{pmatrix} 510 & 41 \\ 41 & 278 \end{pmatrix};$$

$$\bar{y} = \begin{pmatrix} 60 \\ 72 \end{pmatrix}; \quad \hat{S}_y = \frac{1}{11}\begin{pmatrix} 1106 & 246 \\ 246 & 592 \end{pmatrix};$$

$$\hat{S}_{xy} = \frac{1}{20}\begin{pmatrix} 1616 & 287 \\ 287 & 870 \end{pmatrix} = \begin{pmatrix} 80{,}8 & 14{,}35 \\ 14{,}35 & 43{,}5 \end{pmatrix}.$$

Для вычисления статистики критерия получим значения определителей матриц оценок ковариаций:

$$\left|\hat{S}_x\right| = \frac{1}{81}(510\cdot 278 - 41^2) = 1729{,}6172;$$

$$\left|\hat{S}_y\right| = \frac{1}{121}(1106\cdot 592 - 245^2) = 4911{,}0413;$$

$$\left|\hat{S}_{xy}\right| = \frac{1}{400}(1616\cdot 870 - 287^2) = 3308{,}8775.$$

Тогда

$$a = 20\cdot ln3308{,}8775 - 9\cdot ln1729{,}6172 - 11\cdot ln4911{,}0413 = 1{,}494748;$$

$$b = 1 - \left(\frac{1}{9}+\frac{1}{11}-\frac{1}{20}\right)\frac{2\cdot 4+3\cdot 2-1}{6\cdot 3} = 1 - 0{,}1520201\cdot 0{,}7222222 = 0{,}8903078.$$

Следовательно,

$$W_{набл} = 1{,}494748\cdot 0{,}8903078 = 1{,}331.$$

По таблицам хи-квадрат-распределения найдем при $\alpha = 0{,}05$ и $\nu = \frac{1}{2}2\cdot 3 = 3$ критическое значение $\chi^2(0{,}05;3) = 7{,}815$. Так как $W_{набл}$ не попало в критическую область $\left(\chi^2_{набл} < \chi^2(0{,}05;3)\right)$, то мы гипотезу H_0 не отвергаем и будем считать ковариационные матрицы генеральных совокупностей одинаковыми. Заметим, что в этом примере значение C, равное 0,0015316, подтверждает правильность аппроксимации распределения W распределением χ^2.

Теперь можно проверить гипотезу о равенстве генеральных средних, $H_0: \mu_x = \mu_y$ на уровне значимости $\alpha = 0{,}05$ против альтернативы $H_1: \mu_x \neq \mu_y$.

Находим обратную матрицу для $\hat{S}_{xy}$:

$$\hat{S}_{xy}^{-1} = \begin{pmatrix} 0{,}0131464 & -0{,}0043368 \\ -0{,}0043368 & 0{,}0244191 \end{pmatrix}.$$

Тогда наблюдаемое значение статистики T^2 есть

$$T^2_{набл} = \frac{10 \cdot 12}{22} \begin{pmatrix} -3 \\ -8 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 0{,}0131464 & -0{,}0043368 \\ -0{,}0043368 & 0{,}0244191 \end{pmatrix} \begin{pmatrix} -3 \\ -8 \end{pmatrix} =$$

$$= 5{,}4545454 \cdot 1{,}4729736 = 8{,}03,$$

а критическое значение T^2 есть

$$T^2_{кр} = \frac{(10+12-2)2}{10+12-2-1} \cdot F_{0{,}05;2;19} = \frac{40}{19} \cdot 3{,}52 = 7{,}41.$$

Так как $T^2_{набл} > T^2_{кр}$, гипотеза о равенстве векторов генеральных средних отвергается с вероятностью ошибки 0,05. Следовательно, можно считать доказанным, что генеральные совокупности не однородны. Таким образом, есть основания на указание существенности различия в двух способах организации производства по данным выборки.

ГЛАВА 3

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

3.1. Основные понятия

Корреляционный анализ, разработанный К.Пирсоном и Дж.Юлом, является одним из методов статистического анализа взаимозависимости нескольких признаков - компонент случайного вектора x.

Одним из основных показателей взаимозависимости двух случайных величин является **парный коэффициент корреляции**, *служащий мерой линейной статистической зависимости между этими величинами.* Следовательно, этот показатель соответствует своему прямому назначению, когда статистическая связь между соответствующими признаками в генеральной совокупности линейна. То же самое касается **частных и совокупных коэффициентов корреляции**. Одним из требований, определяющих корреляционный метод, является **требование линейности статистической связи**, т.е. линейности всевозможных уравнений (средней квадратической) регрессии.

Указанные условия выполняются, если генеральная совокупность распределена по многомерному нормальному закону.

В настоящее время корреляционный анализ (корреляционная модель) определяется как метод, применяемый тогда, когда данные наблюдений или эксперимента можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

Основная задача корреляционного анализа состоит в оценке $k(k+3)/2$ параметров, определяющих нормальный закон распределения k-мерного вектора x, *в частности, корреляционной матрицы генеральной совокупности X, по выборке.*

Для значимых парных коэффициентов корреляции имеет смысл указать более предпочтительные точечные или интервальные оценки.

Далее следует оценить и проверить значимость множественных коэффициентов корреляции или детерминации всевозможных подсистем системы x_j ($j=1,k$), содержащих три и более различных случайных величин x_j.

Для выяснения "чистых", истинных взаимозависимостей следует проанализировать выборочные частные коэффициенты корреляции.

Таким образом, основная задача позволяет определить расположение "облака" точек в пространстве k измерений, т.е. оценить природу взаимозависимости между наблюдаемыми переменными.

Дополнительная задача корреляционного анализа (являющаяся основной в регрессионном анализе) состоит в оценке уравнений регрессии, где в качестве результативного признака выступает признак, являющийся следствием других признаков (факторов) — причин. Причинно-следственная связь устанавливается из внестатистических соображений, например из аргументов, касающихся физической природы явлений.

Иногда имеет смысл оценить уравнение регрессии для измерения результативного признака по факторным, несмотря на то, что причинно-следственной связи на самом деле между ними не существует. Здесь причиной могут быть другие факторы, не рассматриваемые в модели, но действующие как на функцию, так и на аргументы уравнения регрессии. Так следует поступать в том случае, когда непосредственное измерение результативного признака затруднительно, но существует тесная корреляционная связь (коэффициент множественной корреляции достаточно близок к единице) между результативным признаком и факторными, измерять и наблюдать которые легче в последующих исследованиях.

Назовем параметр связи в генеральной совокупности значимо отличающимся от нуля (значимым), если гипотеза о равенстве нулю этого параметра отвергается с заданным уровнем значимости α. Если же эта гипотеза принимается, генеральный параметр связи называется незначимым. В корреляционной модели соответствующая связь между величинами считается недоказанной или отсутствующей.

3.2. Двумерная модель

3.2.1. Точечные оценки параметров

Рассмотрим генеральную совокупность с двумя признаками x и y, совместное распределение которых задано плотностью двумерного нормального закона

$$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\{Q_2(x,y)\}, \qquad (3.1)$$

где $Q_2(x,y)=\frac{1}{2\sqrt{1-\rho^2}}[(\frac{x-\mu_x}{\sigma_x})^2-2\rho\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}+(\frac{y-\mu_y}{\sigma_y})^2]$, определяемого пятью параметрами: $Mx=\mu_x$, $Dx=\sigma_x^2$, $My=\mu_y$, $Dy=\sigma_y^2$, $M[\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}]=\rho$,

$\rho^2\neq 1$.

Имея эти параметры, можно получить уравнения линий регрессии, показывающих изменение условных математических ожиданий в зависимости от изменения соответствующих значений случайных аргументов:

$My/x-My=\beta_{yx}(x-Mx)$ - прямая регрессии y на x;

$My/x-Mx=\beta_{xy}(y-My)$ - прямая регрессии x на y;

$\beta_{yx}=\rho\frac{\sigma_y}{\sigma_x}$ - коэффициент регрессии y на x;

$\beta_{xy}=\rho\frac{\sigma_x}{\sigma_y}$ - коэффициент регрессии x на y.

Полезно вспомнить, что квадрат коэффициента корреляции ρ^2, т.е. *коэффициент детерминации, в рассматриваемой модели указывает долю дисперсии одной случайной величины, обусловленную вариацией другой. Коэффициент регрессии β_{yx} показывает, на сколько единиц своего измерения увеличится ($\beta>0$) или уменьшится ($\beta<0$) в среднем $y(My/x)$, если x увеличить на единицу своего измерения.*

Задача двумерного корреляционного анализа состоит прежде всего в оценке пяти параметров, определяющих генеральную совокупность.

В качестве точечных оценок неизвестных начальных моментов первого и второго порядка генеральной совокупности берутся соответствующие выборочные моменты.

Точечные же оценки неизвестных других параметров получают с помощью формул, аналогичных формулам вычисления самих параметров через генеральные начальные моменты. Таким образом, будем иметь:

$\bar{x}$ - оценка для μ_x,

$\bar{y}$ - оценка для μ_y,

$\overline{x^2}$ - оценка для $M(x^2)$,

$\overline{y^2}$ - оценка для $M(y^2)$,

$\overline{xy}$ - оценка для $M(xy)$.

Откуда

$s_x^2=\overline{x^2}-(\bar{x})^2$ - оценка для σ_x^2,

$s_y^2=\overline{y^2}-(\bar{y})^2$ - оценка для σ_y^2,

$r=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{s_x s_y}$ - оценка для ρ.

Оценки генеральных коэффициентов регрессии β_{yx} и β_{xy} получаются соответственно по формулам:

$$b_{yx}=r\frac{s_y}{s_x}, \qquad b_{xy}=r\frac{s_x}{s_y},$$

откуда оценки уравнений регрессии имеют вид:

$$\overline{y/x} - \bar{y} = b_{yx}(x - \bar{x}), \qquad \overline{x/y} - \bar{x} = b_{xy}(y - \bar{y}).$$

При этом $\overline{y/x}$ и $\overline{x/y}$ - обозначения оценок для условных математических ожиданий My/x и Mx/y генеральной совокупности.

Следует отметить, что вышеприведенные точечные оценки являются состоятельными, а $\bar{x}$ и $\bar{y}$ несмещенными и эффективными. Кроме того, распределение выборочных средних $(\bar{x}, \bar{y})$ не зависит от распределения (s_x^2, s_y^2, r). Наконец, выборочный коэффициент корреляции r по абсолютной величине не превосходит единицы.

3.2.2. Приемы вычисления выборочных характеристик

Если объем выборки невелик, то наблюдаемые точки располагают в таблице в порядке их регистрации и обрабатывают по следующей схеме:

x	y	x^2	y^2	xy
x_j	y_j	x_j^2	y_j^2	$x_j y_j$
Σx_j	Σy_j	Σx_j^2	Σy_j^2	$\Sigma x_j y_j$

В схеме последовательно заполняют столбцы таблицы результатами операций, указанных сверху. В последней строке вычисляются соответствующие суммы элементов столбцов. Далее используют формулы:

$$\bar{x} = \frac{\Sigma x_i}{n}; \quad \bar{y} = \frac{\Sigma y_j}{n}; \quad s_x^2 = \frac{\Sigma x_j^2}{n} - (\bar{x})^2; \quad s_y^2 = \frac{\Sigma y_j^2}{n} - (\bar{y})^2;$$

$$r = \frac{\Sigma x_j y_j - (\Sigma x_j \Sigma y_j)/n}{\{[\Sigma x_j^2 - (\Sigma x_j)^2/n][\Sigma y_j^2 - (\Sigma y_j)^2/n]\}^{1/2}};$$

$$b_{yx} = \frac{\Sigma x_j y_j - (\Sigma x_j \Sigma y_j)/n}{\Sigma x_j^2 - (\Sigma x_j)^2/n}; \quad b_{xy} = \frac{\Sigma x_j y_j - (\Sigma x_j \Sigma y_j)/n}{\Sigma y_j^2 - (\Sigma y_j)^2/n}.$$

Если выборка многочисленна, то данные группируются путем построения двумерного интервального ряда, корреляционная таблица для которого имеет вид:

y \ x	...	$(a_k - b_k]$	...	m_y
$(c_l - d_l]$	...	m_{kl}	...	m_{*l}
m_x	...	m_{k*}	...	n

В таблице m_{kl} - частота прямоугольника, в основании которого лежит полуинтервал $(a_k - b_k]$, а по высоте - $(c_l - d_l]$, т.е. число точек выборки, попавших внутри или на часть границы прямоугольника, задаваемой полуинтервалами. При этом длины интервалов по x одинаковы и равны h_x; то же самое относится и к y (одинаковой длины h_y).

Для вычисления характеристик интервального вариационного ряда переходим к условному дискретному вариационному ряду с условными вариантами

$$x'_k = \frac{x_k - x_O}{h_x}; \quad y'_l = \frac{y_l - y_O}{h_y},$$

где x_o, y_o - рабочие средние - выбираются обычно равными центрам интервалов, лежащих в середине соответствующих одномерных рядов;

x_k - центры интервалов.

Таким образом, условные варианты - целые числа, наименее уклоняющиеся от нуля по абсолютной величине.

Вычисления удобно производить по схеме, последовательно заполняя строки, лежащие ниже таблицы двумерного ряда условных вариантов $(1 \div 4)$, и столбцы, лежащие справа от этой таблицы $(1 \div 2)$:

y' \ x'	...	x_k	...	m_y	$y' m_y^1$	$(y')^2 m_y^2$
y_l	...	m_{kl}	...	m_{xl}	$y'_l m_{*l}$	$(y'_l)^2 m_{*l}$
m_x	...	m_{k*}	...	n	$\Sigma y'_l m_{*l}$	$\Sigma (y'_l)^2 m_{*l}$
1. $x' m_x$	...	$x'_k m_{k*}$	...	$\Sigma x'_k m_{k*}$		
2. $(x')^2 m_x$	...	$(x'_k)^2 m_{k*}$	...	$\Sigma (x'_k)^2 m_{k*}$		
3. $\Sigma y' m_{xy}$	...	$\Sigma y'_l m_{kl}$	...	$\Sigma y'_l m_{*l}$		
4. $x' \Sigma y' m_{xy}$	...	$x' \Sigma y'_l m_{kl}$	...	$\Sigma\Sigma x'_k y'_l m_{kl}$		

Заметим, что для контроля вычислений можно использовать равенство $\Sigma\Sigma x'_k y'_l m_{kl} = \Sigma y'_l m_{*l}$, т.е. равенство чисел в конце строки 3 и столбца 1.

Далее используются формулы:

$$\bar{x}_{гр} \frac{\Sigma x'_k m_{k*}}{n} h_x + x_0; \quad \bar{y}_{гр} = \frac{\Sigma \bar{y}_l m_{*l}}{n} h_y + y_0;$$

$$S^2_{x_{гр}} = [\frac{\Sigma (x'_k)^2 m_{k*}}{n} - (\frac{\Sigma x'_k m_{k*}}{n})^2] h_x^2;$$

$$S^2_{y_{гр}} = [\frac{\Sigma (y'_l)^2 m_{l*}}{n} - (\frac{\Sigma y'_l m_{l*}}{n})^2] h_y^2;$$

$$r_{гр} = \frac{n\Sigma\Sigma x'_k y'_l m_{kl} - \Sigma x'_k m_{k*} \Sigma y'_l m_{*l}}{\{[n\Sigma (x'_k)^2 m_{k*} - (\Sigma x_k m_{k*})^2][n\Sigma (y'_l)^2 m_{*l} - (\Sigma y'_l m_{*l})^2]\}^{1/2}};$$

$$byx_{гр} = \frac{n\Sigma\Sigma x'_k y'_l m_{kl} - \Sigma x'_k m_{k*} \Sigma y'_l m_{*l}}{n\Sigma (y'_l) m_{*l} - (\Sigma y'_l m_{*l})^2} \cdot \frac{h_x}{h_y};$$

$$bxy_{гр} = \frac{n\Sigma\Sigma x'_k y'_l m_{kl} - \Sigma x'_k m_{k*} \Sigma y'_l m_{xl}}{n\Sigma (y'_l) m_{*l} - (\Sigma y'_l m_{*l})^2} \cdot \frac{h_x}{h_y}.$$

При группировке вычисленные характеристики могут сильно отличаться от выборочных. Оценки по группированным данным центральных моментов второго порядка s_x^2 и s_y^2 можно **улучшить поправками Шеппарда**:

$$s_x^2 \cong s^2_{x_{гр}} - \frac{1}{12} h_x^2; \quad s_y^2 \cong s^2_{y_{гр}} - \frac{1}{12} h_y^2. \qquad (3.2)$$

Эти поправки часто сглаживают ошибки, возникающие от группировки, если длина интервала (h) не превосходит восьмой части размаха соответствующего признака.

3.2.3. Проверка значимости параметров связи

В двумерной модели параметрами связи являются коэффициент корреляции ρ (или квадрат, называемый коэффициентом детерминации) и коэффициенты регрессии β_{yx} и β_{xy}.

Заметим, что в двумерной модели достаточно проверить значимость только коэффициента корреляции. Если коэффициент корреляции незначим, то признаки x и y считаются независимыми в генеральной совокупности.

Статистика r, вычисляемая для выборки из двумерной нормально распределенной совокупности с $\rho = 0$, связана со статистикой t, имеющей распределения Стьюдента с $\nu = n-2$ степенями свободы, формулой

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}. \qquad (3.3)$$

Зная границы для t, соответствующие обычным уровням значимости (α=10%, 5%, 2%, 1%), можно получить границы для r, воспользовавшись этой формулой. Границы для r табулированы. Таким образом, для проверки гипотезы H_0:$\rho = 0$ по данным α и $\nu = n-2$ находим $r_{табл}$. Если $|r_{набл}| > r_{табл}$, то гипотеза H_0 отвергается с вероятностью ошибки α, если же $|r_{набл}| \le r_{табл}$, то гипотеза не отвергается. При $\nu > 100$ для проверки H_0:$\rho = 0$ следует пользоваться нормированным нормальным законом распределения статистики

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{\nu}$$

или статистики

$$r\sqrt{n-1}.$$

Если наблюдаемая величина (t или $r\sqrt{n-1}$) расположена в доверительном интервале $[-t_{1-\alpha}, t_{1-\alpha}]$, то гипотеза H_0 не отвергается; в противном случае H_0 отвергается с уровнем значимости α.

3.2.4. Интервальные оценки параметров связи

Для значимых параметров связи имеет смысл найти интервальные оценки.

При нахождении доверительного интервала для коэффициента корреляции ρ используют статистику, введенную Фишером:

$$z_r = \frac{1}{2} \ln \frac{r+1}{1-r},$$

которая при $n > 10$ распределена приблизительно нормально с генеральным средним $MZ_r \approx \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$ и дисперсией $Dz_r \approx \frac{1}{n-3}$.

Тогда доверительный интервал, оценивающий MZ_r с надежностью $\gamma = 1-\alpha$, имеет вид

$$Z_r - t_\gamma \sqrt{\frac{1}{n-3}} \le MZ_r \le Z_r + t_\gamma \sqrt{\frac{1}{n-3}}, \qquad (3.4)$$

где t_γ - находится по таблицам интеграла Лапласа для данного γ (или $\gamma = 1-\alpha$).

Для перехода от Z к ρ имеется таблица, составленная Фишером и Иейтсом, после использования которой получаем интервальную оценку с надежностью γ вида

$$r_{min} \le \rho \le r_{max},$$

где r_{min} и r_{max} выбираются с учетом того, что Z_r - функция нечетная. При этом поправочным членом $\frac{\rho}{2(n-1)}$ у MZ_r пренебрегают.

Если коэффициент корреляции значим, то коэффициенты регрессии также значимо отличаются от нуля (с тем же уровнем α). Интервальные оценки для них получаются по формулам

$$|t| \le t(\gamma, \nu); \quad t = (b_{yx} - \beta_{yx}) \frac{s_x \sqrt{n-2}}{s_y \sqrt{1-r^2}}; \quad t = (b_{xy} - \beta_{xy}) \frac{s_y \sqrt{n-2}}{s_x \sqrt{1-r^2}},$$

где t имеет распределение Стьюдента с $\nu = n-2$ степенями свободы.

Переход от неравенства $|t| \le t(\gamma, \nu)$ к интервальным оценкам для коэффициента регрессии осуществляется с помощью тождественных алгебраических преобразований.

Для значимого коэффициента корреляции ρ некоторые авторы рекомендуют более предпочтительную оценку, чем r:

$$r(1 + \frac{1-r^2}{2(n-4)}).$$

Предпочтительной оценкой ρ^2 является выражение

$$\frac{(n-1)r^2 - 1}{n-2}.$$

Этими точечными оценками следует пользоваться при небольших объемах n выборки.

3.2.5. Задачи, решаемые при помощи статистики Фишера

Кроме нахождения интервальной оценки для ρ, с помощью преобразования

$$Z_r = \frac{1}{2} \ln \frac{1+r}{1-r}$$

можно решить следующие задачи.

1. Проверить, согласуется ли выборочный коэффициент корреляции r с предполагаемым значением генерального коэффициента корреляции ρ_o. Для этого, взяв уровень значимости α, проверяем, попадает ли абсолютная величина разности $|Z_r - Z_{\rho_o}|$ в интервал $[0, t_{1-\alpha} / \sqrt{n-3}]$. Если попадает, то гипотеза $H_o: \rho = \rho_o$ не отвергается. В противном случае отвергается с уровнем α.

2. Проверить гипотезу об однородности коэффициентов корреляции. Пусть $r_1, r_2, \ldots, r_k$ - коэффициенты корреляции, полученные из k нормально распределенных совокупностей по выборкам с объемами $n_1, n_2, \ldots, n_k$. Проверяется гипотеза

$$H_o: \rho_1 = \rho_2 = \ldots = \rho_k = \rho.$$

Статистика

$$\sum_{i=1}^{k} \frac{(z_{r_i} - z_\rho)^2}{1/(n_i - 3)}$$

имеет тогда распределение χ^2 с k степенями свободы. Если заменить z_ρ на среднее арифметическое

$$\bar{z}_r = \frac{\Sigma z_i n_i}{\Sigma n_i},$$

то получим, что

$$\sum_{i=1}^{k} \frac{(z_{r_i} - \bar{z}_r)^2}{1/(n_i - 3)}$$

распределена по закону χ^2 с $\nu = k-1$ степенями свободы.

Если теперь для заданных α и $\nu = k-1$

$$\chi^2_{набл} < \sum_{i=1}^{k} \frac{(z_{r_i} - \bar{z}_r)^2}{1/(n_i - 3)},$$

то гипотеза однородности отвергается с уровнем α. В противном случае гипотеза H_o не отвергается.

В случае принятия гипотезы однородности предпочтительной точечной оценкой ρ является значение r, полученное обратным преобразованием из z_r.

3.2.6. Корреляционный анализ затрат и оплаты труда

Пример 3.1

На основе следующих данных годового отчета 100 хозяйств об оплате труда (y) в хлопководстве, измеряемой в тыс.руб. за 1 чел.-ч, и затратах труда (x), измеряемых количеством человеко-часов на 1 га хлопчатника, провести корреляционный анализ. Рассмотрим табл. 3.1.

Таблица 3.1

Взаимосвязь между оплатой труда за единицу времени работы и затратами труда на единицу обрабатываемой площади

y \ x	500-650	650-800	800-950	950-1100	1100-1250	1250-1400	1400-1550	m_y
0,6-0,9					1	1	2	4
0,9-1,2			1	4	2	3	2	12
1,2-1,5		4	7	6	1	1		19
1,5-1,8	2	4	4	3	1			14
1,8-2,1	6	8	3	4				21
2,1-2,4	6	5	6	1				18
2,4-2,7	2	5						7
2,7-3,0	2	3						5
m_x	18	29	21	18	5	5	4	100

Решение

Будем считать, что приведенные в примере данные являются выборкой из двумерной нормально распределенной совокупности, на основании которой построен двумерный интервальный вариационный ряд (корреляционная таблица).

Используя формулы для условных вариантов $x_i' = (x_i - x_0):h_x$; $y_i' = (y_i - y_0):h_y$ при h_x=150, x_o=1025, h_y=0,3 и y_o=1,95, получим из вспомогательной таблицы корреляционного анализа значения величин, необходимых для вычисления точечных оценок основных параметров: n=100, $\Sigma x' m_x = -106$, $\Sigma x'^2 m_x = 360$, $\Sigma x'y' m_{xy} = -136$, $\Sigma y' m_y = -57$, $\Sigma y'^2 m_y = 353$. Тогда будем иметь следующие оценки этих параметров:

$$\bar{x} = \frac{1}{n}\Sigma x' m_x h_x + x_0 = -\frac{106}{100}150 + 1025 = 866 \text{ чел.-ч/га} = 8{,}66 \text{ чел.-ч/га};$$

$$\bar{y} = \frac{1}{n}\Sigma y' m_y h_y + y_0 = \frac{-57}{100}0{,}3 + 1{,}95 = 1{,}779 = 1{,}78 \text{ руб/чел.-ч.};$$

$$s_x^2 = [\frac{\Sigma x'^2 m_x}{n} - (\frac{\Sigma x' m_x}{n})^2] \cdot h_x^2 = [\frac{360}{100} - (\frac{-106}{100})^2] \cdot 150^2 = 55719;$$

$$s_y^2 = [\frac{\Sigma y'^2 m_y}{n} - (\frac{\Sigma y' m_y}{n})^2] \cdot h_y^2 = [\frac{353}{100} - (\frac{-57}{100})^2] \cdot 0{,}3^2 = 0{,}288459;$$

s_x=236,05 чел.-ч/га=2,36 чел.-ч/га; s_y=0,54 тыс.руб/чел.-ч;

$$r = \frac{n\Sigma x'y' m_{xy} - \Sigma x' m_x \Sigma y' m_y}{\sqrt{[n\Sigma x'^2 m_x - (\Sigma x' m_x)^2][n\Sigma y'^2 m_y - (\Sigma y^2 m_y)^2]}} =$$

$$= \frac{100(-136) - (-106)(-57)}{\{[100 \cdot 360 - (-106)^2 \cdot 100 \cdot 353 - (-57)^2]\}^{\frac{1}{2}}} =$$

$$= \frac{19642}{(24764 \cdot 32051)^{\frac{1}{2}}} = -\frac{19642}{28172{,}875} = -0{,}6972.$$

Проверим значимость коэффициента корреляции на уровне α=0,05. По таблицам распределения статистики r получаем $r_{кр}=r_{табл}(0{,}05;98)<r_{табл}(0{,}05;90)=0{,}205$.

Критическая область имеет вид $|r|>r_{кр}$. Так, $r_{табл}$=-0,6972 попадает в критическую область ($|-0{,}6972|>0{,}205>r_{кр}$). Гипотеза H_0: ρ=0 отвергается с вероятностью ошибки 0,05. Следовательно, коэффициент корреляции ρ значимо отличается от нуля. Можно считать доказанной связь между случайными величинами x и y. Корреляционный анализ можно углубить.

Найдем с надежностью γ=1-0,05=0,95 интервальную оценку генерального коэффициента корреляции ρ. Получим значение статистики Z по формуле

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\ln\frac{0{,}3028}{1{,}6972} = -0{,}86183.$$

Найдем точность интервальной оценки MZ по формуле

$$\Delta_Z = \Phi^{-1}(\gamma):\sqrt{n-3} = 1{,}96:\sqrt{97} = 0{,}19901.$$

Интервальная оценка для MZ имеет вид

$$Z - \Delta_Z \le MZ \le Z + \Delta_Z,$$

т.е.

$$-1{,}06084 \le MZ \le -0{,}66282.$$

С помощью обратной функции Z^{-1} получаем интервальную оценку коэффициента корреляции ρ:

$$Z^{-1}(Z - \Delta_Z) \le \rho \le Z^{-1}(Z + \Delta_Z)$$

или

$$-0{,}7860 \le \rho \le -0{,}5802.$$

Найдем точечные оценки коэффициентов регрессии с учетом измерения x в человеко-часах (в противном случае вычисления окажутся неточными).

$$b_{yx} = r\frac{s_y}{s_x} = -0{,}6972\sqrt{\frac{0{,}288459}{5{,}5719}} = -0{,}1586344;$$

$$b_{xy} = r\frac{s_y}{s_x} = -0{,}6972\sqrt{\frac{5{,}5719}{0{,}288459}} = -3{,}064199.$$

Точечные оценки уравнений регрессии имеют вид

$$\overline{y/x} - \bar{y} = b_{yx}(x - \bar{x}) \text{ или } \overline{y/x} - 1{,}78 = 0{,}15863(x - 8{,}66);$$

$$\overline{x/y} - \bar{x} = b_{xy}(y - \bar{y}) \text{ или } \overline{x/y} - 8{,}66 = -3{,}0642(y - 1{,}78).$$

Можно получить интервальные оценки коэффициентов регрессии:

$$b_{yx} - St^{-1}(\alpha)\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}} \le \beta_{yx} \le b_{yx} + St^{-1}(\alpha)\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}};$$

$$b_{xy} - St^{-1}(\alpha)\frac{s_x\sqrt{1-r^2}}{s_y\sqrt{n-2}} \le \beta_{xy} \le b_{xy} + St^{-1}(\alpha)\frac{s_x\sqrt{1-r^2}}{s_y\sqrt{n-2}}, \qquad (3.5)$$

где $St^{-1}(\alpha) = St^{-1}(0{,}05;\ 98) = 1{,}967$ находится по таблицам распределения Стьюдента с помощью линейной интерполяции.

Будем иметь

$$-0{,}19104 \le \beta_{yx} \le -0{,}12623; \qquad -3{,}6902 \le \beta_{yx} \le -2{,}4382.$$

Поясним содержательный смысл полученных результатов.

Доказана значимость, существенность связи между оплатой 1 чел.-ч затрат в хлопководстве и затратах труда, приходящихся на 1 га площади хлопчатника. Связь является отрицательной, т.е. увеличению одного из показателей соответствует уменьшение среднего значения другого показателя.

При этом увеличению оплаты труда на 1 тыс.руб. за человеко-час соответствует уменьшение затрат на 1 га на величину 3,0642 чел.-ч в среднем. С вероятностью 0,95 можно утверждать, что указанная величина может оказаться равной любому значению из интервала [2,4382; 3,6902]. Увеличению затрат на одну сотню человеко-часов на 1 га будет соответствовать в среднем уменьшение оплаты за 1 чел.-ч на величину 0,15863 тыс.руб., причем с вероятностью 0,95 можно утверждать, что эта величина может принимать любое значение из интервала 0,12623; 0,19104.

Так как коэффициент детерминации ρ^2 при доверительной вероятности 0,95 лежит в границах от $(-0{,}5802)^2 = 0{,}3366$ до $(-0{,}7860)^2 = 0{,}6178$, то можно сказать, что изменение одного из контролируемых в модели показателей соответствует изменению другого на величину, принимающую значение от 33,66 до 61,78%. Дополняющая до 100% величина соответствует изменению неконтролируемых (случайных) моделей показателей. В вычислительной табл.3.2 представлен корреляционный анализ двумерной модели.

Таблица 3.2

Вычислительная таблица корреляционного анализа двумерной модели

y' \ x'	−3	−2	−1	0	1	2	3	m_y	$y' m_y$	$y'^2 m_y$
−4					1	1	2	4	−16	64
−3			1	4	2	3	2	12	−36	108
−2		4	7	6	1	1		19	−38	76

Продолжение

y' \ x'	−3	−2	−1	0	1	2	3	m_y	$y' m_y$	$y'^2 m_y$
-1	2	4	4	3	1			14	-14	14
0	6	8	3	4				21	0	0
1	6	5	6	1				18	18	18
2	2	5						7	14	28
3	2	3						5	15	45
m_x	18	29	21	18	5	5	4	100	-57	353
$x' m_x$	-54	-58	-21	0	5	10	12	-106		
$x'^2 m_x$	162	116	21	0	5	20	36	360		
$\Sigma y' m_{xy}$	14	12	-15	-26	-13	-15	-14	-57		
$x' \Sigma y' m_{xy}$	-42	-24	15	0	-13	-30	-42	-136		

3.3. Трехмерная модель

3.3.1. Основные параметры модели

Для изучения основных задач и особенностей корреляционного анализа удобно рассматривать генеральную совокупность трех признаков x, y и z.

Трехмерная непрерывная случайная величина (x, y, z) называется **нормально распределенной**, если плотность совместного распределения одномерных случайных величин x, y и z задается в виде

$$\rho(x,y,z) = (2\pi)^{-\frac{3}{2}}\left[\sigma_x^2\sigma_y^2\sigma_z^2\left|\Re_3\right|\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}u^T\Re_3^{-1}u\right\},$$

где $\Re_3 = \begin{pmatrix} 1 & \rho_{xy} & \rho_{xz} \\ \rho_{yx} & 1 & \rho_{yz} \\ \rho_{zx} & \rho_{zy} & 1 \end{pmatrix}$ - симметрическая положительно определенная матрица парных коэффициентов корреляции, соответствующих частным двумерным распределениям случайных величин (x, y), (x, z) и (y, z);

$\left|\Re_3\right|$ - определитель матрицы $\Re_3$, обобщенная дисперсия случайной величины (x, y, z);

$$\left|\Re_3\right| = 1 + 2\rho_{xy}\rho_{xz}\rho_{yz} - \rho_{xz}^2 - \rho_{yz}^2 - \rho_{zy}^2 > 0;$$

$\Re_3^{-1} = \begin{pmatrix} \rho^{(11)} & \rho^{(12)} & \rho^{(13)} \\ \rho^{(21)} & \rho^{(22)} & \rho^{(23)} \\ \rho^{(31)} & \rho^{(32)} & \rho^{(33)} \end{pmatrix}$ - матрица, обратная $\Re_3$:

$\rho^{(ij)} = \dfrac{\Re_{ji}}{|\Re_3|}$; $\Re_{ji} = (-1)^{j+i} M_{ji}$, M_{ji} - минор матрицы $\Re_3$, дополнительный к элементу ρ_{ji}, j,i=1,2,3:

$$\rho^{(11)} = \frac{1}{|\Re_3|}(-1)^{1+1}\begin{pmatrix} 1 & \rho_{yz} \\ \rho_{yx} & 1 \end{pmatrix} = \frac{1-\rho_{yz}^2}{|\Re_3|};$$

$$\rho^{(12)} = \frac{1}{|\Re_3|}(-1)^{1+2}\begin{vmatrix} \rho_{xy} & \rho_{xz} \\ \rho_{yz} & 1 \end{vmatrix} = \frac{-\rho_{xy}+\rho_{xz}\rho_{yz}}{|\Re_3|};$$

$$\rho^{(13)} = \frac{1}{|\Re_3|}(-1)^{1+3}\begin{vmatrix} \rho_{xy} & \rho_{xz} \\ 1 & \rho_{yz} \end{vmatrix} = \frac{\rho_{xy}\rho_{yz}-\rho_{xz}}{|\Re_3|};$$

$$\rho^{(22)} = \frac{1}{|\Re_3|}(-1)^{2+2}\begin{vmatrix} 1 & \rho_{xz} \\ \rho_{yz} & 1 \end{vmatrix} = \frac{1-\rho_{xz}^2}{|\Re_3|};$$

$$\rho^{(23)} = \frac{1}{|\Re_3|}(-1)^{2+3}\begin{vmatrix} 1 & \rho_{xz} \\ \rho_{xy} & \rho_{yz} \end{vmatrix} = \frac{-\rho_{yz}+\rho_{xz}\rho_{xy}}{|\Re_3|};$$

$$\rho^{(33)} = \frac{1}{|\Re_3|}(-1)^{3+3}\begin{vmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{vmatrix} = \frac{1-\rho_{xy}^2}{|\Re_3|},$$

матрица $\Re_3^{-1}$ - симметрическая, положительно определенная;

$u = \begin{pmatrix} \dfrac{x-\mu_x}{\sigma_x} \\ \dfrac{y-\mu_y}{\sigma_y} \\ \dfrac{z-\mu_z}{\sigma_z} \end{pmatrix}$ - вектор значений нормированных случайных величин x, y и z;

u^T - транспонированный вектор u:

$$u^T\Re_3^{-1}u = \begin{pmatrix} \dfrac{x-\mu_x}{\sigma_x} & \dfrac{y-\mu_y}{\sigma_y} & \dfrac{z-\mu_z}{\sigma_z} \end{pmatrix}\begin{pmatrix} \rho^{(11)} & \rho^{(12)} & \rho^{(13)} \\ \rho^{(21)} & \rho^{(22)} & \rho^{(23)} \\ \rho^{(31)} & \rho^{(32)} & \rho^{(33)} \end{pmatrix} \times$$

$$\times\begin{pmatrix} \dfrac{x-\mu_x}{\sigma_x} \\ \dfrac{y-\mu_y}{\sigma_y} \\ \dfrac{z-\mu_z}{\sigma_z} \end{pmatrix} = \left[\frac{x-\mu_x}{\sigma_x}\rho^{(11)} + \frac{y-\mu_y}{\sigma_y}\rho^{(12)} + \frac{z-\mu_z}{\sigma_z}\rho^{(13)}\right]\frac{x-\mu_x}{\sigma_x} +$$

$$+\left[\frac{x-\mu_x}{\sigma_x}\rho^{(12)} + \frac{y-\mu_y}{\sigma_y}\rho^{(22)} + \frac{z-\mu_z}{\sigma_z}\rho^{(23)}\right]\frac{y-\mu_y}{\sigma_y} +$$

$$+\left[\frac{x-\mu_x}{\sigma_x}\rho^{(13)} + \frac{y-\mu_y}{\sigma_y}\rho^{(32)} + \frac{z-\mu_z}{\sigma_z}\rho^{(33)}\right]\frac{z-\mu_z}{\sigma_z}.$$

Таким образом, трехмерная нормально распределенная случайная величина определяется *девятью параметрами*:

тремя математическими ожиданиями:

$$Mx=\mu_x,\ My=\mu_y,\ Mz=\mu_z;$$

тремя дисперсиями (или средними квадратическими отклонениями):

$$Dx=\sigma_x^2,\quad Dy=\sigma_y^2,\quad Dz=\sigma_z^2\quad(\sigma_x,\sigma_y,\sigma_z);$$

тремя парными коэффициентами корреляции:

$$\rho_{xy} = M[\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}],$$

$$\rho_{xz} = M[\frac{x-\mu_x}{\sigma_x}\cdot\frac{z-\mu_z}{\sigma_z}],$$

$$\rho_{yz} = M[\frac{y-\mu_y}{\sigma_y}\cdot\frac{z-\mu_z}{\sigma_z}].$$

Следует отметить, что частные одномерные (x, y и z), двумерные ((x,y),(x,z) и (y,z)) распределения компонент, а также условные распределения при фиксировании одной ((x,y)/z, (x,z)/y, (y,z)/x) и двух (x/y,z; y/x,z; z/x,y) компонент являются нормальными. Поэтому поверхности и линии регрессии являются плоскостями и прямыми соответственно.

Для трехмерной (и других многомерных) корреляционной модели важную роль играют частные и множественные коэффициенты корреляции или детерминации (коэффициент детерминации равен квадрату соответствующего коэффициента корреляции).

Частным коэффициентом корреляции между x и y при фиксированных остальных компонентах (т.е. z) является выражение

$$\rho_{xy/z} = -\frac{\Re_{12}}{(\Re_{11}\cdot\Re_{12})^{1/2}} = \frac{\rho_{xy}-\rho_{xz}\rho_{yz}}{\sqrt{(1-\rho_{xz}^2)(1-\rho_{yz}^2)}}. \qquad (3.6)$$

Остальные частные коэффициенты корреляции $\rho_{xz/y}$ и $\rho_{yz/x}$ определяют путем замены соответствующих индексов в приведенных формулах.

Для нормального распределения частный коэффициент корреляции $\rho_{xy/z}$ совпадает с парным коэффициентом корреляции между величинами x и y при фиксированном z (в двумерном условном распределении ((x,y)/z). *Частный коэффициент корреляции обладает всеми свойствами парного коэффициента корреляции.*

Он служит показателем линейной связи между двумя переменными случайными величинами независимо от влияния остальных случайных

переменных. Если частный коэффициент детерминации меньше, чем соответствующий парный коэффициент детерминации, то взаимозависимость между двумя величинами обусловлена частично (или целиком при равенстве нулю частного коэффициента детерминации) воздействием на эту пару остальных, фиксируемых, случайных величин. Если же, наоборот, частный коэффициент детерминации больше соответствующего парного, то фиксируемые величины ослабляют, затушевывают связь.

Множественный коэффициент корреляции между одной величиной z и двумя другими величинами (x,y) определяется по формуле:

$$\rho_z = \rho_{z/xy} = \sqrt{1 - \frac{|\Re_3|}{\Re_{11}}} = \sqrt{\frac{\rho_{zx}^2 + \rho_{zy}^2 - 2\rho_{xy}\rho_{zx}\rho_{zy}}{1-\rho_{xy}^2}}. \qquad (3.7)$$

Для трехмерной нормально распределенной случайной величины (x,y,z) множественный коэффициент корреляции является мерой связи между одной случайной величиной и двумя остальными. Он заключен между нулем и единицей. При $\rho_z = 1$ связь между величинами z и (x,y) является функциональной, линейной: точки (x,y,z) расположены в плоскости регрессии z на (x,y). При $\rho_z = 0$ одномерная случайная величина z и двумерная случайная величина (x,y) являются независимыми (в силу нормальности распределения). *Множественный коэффициент детерминации ρ_z^2 показывает долю дисперсии случайной величины z, обусловленную изменением случайных величин (x,y).*

Из определяющей ρ_z формулы можно получить следующие неравенства:

$$\rho_z \geq |\rho_{zx}|, \quad \rho_z \geq |\rho_{zy}|, \quad \rho_z \geq |\rho_{zx/y}|, \quad \rho_z \geq |\rho_{zy/x}|.$$

Отсюда можно заметить, что коэффициент множественной корреляции может только увеличиться, если в модель включать дополнительные признаки - случайные величины, и не увеличиться, если из имеющихся признаков производить исключение.

Далее, если $\rho_z = 0$, то $\rho_{zx} = \rho_{zy} = \rho_{zx/y} = \rho_{zy/x} = 0$.

Если, например, $\rho_{xy}^2 \leq \rho_{xz}^2$ и $\rho_{xy}^2 \leq \rho_{yz}^2$, то $\rho_z^2 \geq \rho_x^2$, $\rho_z^2 \geq \rho_y^2$ и $\rho_{xy/z}^2 \leq \rho_{xz/y}^2$, $\rho_{xy/z}^2 \leq \rho_{yz/x}^2$. Последние неравенства можно получить исходя из формул:

$$\rho_{xy/z}^2 = \begin{cases} 1 - \dfrac{1-\rho_x^2}{1-\rho_{xz}^2}; \\ 1 - \dfrac{1-\rho_y^2}{1-\rho_{yz}^2}. \end{cases} \qquad \rho_{xz/y}^2 = \begin{cases} 1 - \dfrac{1-\rho_y^2}{1-\rho_{xy}^2}; \\ 1 - \dfrac{1-\rho_z^2}{1-\rho_{yz}^2}. \end{cases} \qquad \rho_{yz/x}^2 = \begin{cases} 1 - \dfrac{1-\rho_y^2}{1-\rho_{xy}^2} \\ 1 - \dfrac{1-\rho_z^2}{1-\rho_{xz}^2}. \end{cases}$$

Таким образом, наибольшему множественному коэффициенту детерминации соответствуют большие частные коэффициенты детерминации (например, ρ_x^2 соответствуют $\rho_{xz/y}^2$ и $\rho_{xy/z}^2$).

Приведем некоторые характеристики, подлежащие корреляционному анализу трехмерной случайной величины. При этом будем рассматривать лишь по одному условному распределению (двумерному и одномерному), так как остальные совпадают с рассматриваемыми с точностью до перестановки букв.

Условное распределение при заданном z

Так как это двумерное нормальное распределение $(x,y)/z$, то оно определяется пятью параметрами (двумя условными математическими ожиданиями $\mu_{x/z}$ и $\mu_{y/z}$, двумя условными дисперсиями $\sigma_{x/z}^2$ и $\sigma_{y/z}^2$; условным коэффициентом корреляции $\rho_{xy/z}$):

$$\mu_{x/z} = \mu_x + \rho_{zx}\frac{\sigma_x}{\sigma_z}(z-\mu_z); \qquad \mu_{y/z} = \mu_y + \rho_{zy}\frac{\sigma_y}{\sigma_z}(z-\mu_z);$$

$$\sigma_{x/z}^2 = \sigma_x^2(1-\rho_{zx}^2); \qquad \sigma_{y/z}^2 = \sigma_y^2(1-\rho_{zx}^2); \qquad \rho_{yx/z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\left[(1-\rho_{xz}^2)(1-\rho_{yz}^2)\right]^{1/2}}.$$

Форма зависимости выражается следующими линиями регрессии в плоскости $Z = \bar{z}$:

$$M(y/x)/z - \mu_{y/z} = \beta_{yx/z}(x - \mu_{x/z});$$
$$M(x/y)/z - \mu_{x/z} = \beta_{xy/z}(y - \mu_{y/z}).$$

Коэффициенты частной регрессии имеют вид:

$$\beta_{yx/z} = \rho_{xy/z}\frac{\sigma_{y/z}}{\sigma_{x/z}} = \frac{\beta_{yx} - \beta_{yz}\beta_{zx}}{1-\beta_{xz}\beta_{zx}}; \qquad (3.8)$$

$$\beta_{xy/z} = \rho_{xy/z}\frac{\sigma_{x/z}}{\sigma_{y/z}} = \frac{\beta_{xy} - \beta_{xz}\beta_{zy}}{1-\beta_{yz}\beta_{zy}},$$

причем

$$\rho_{xy/z}^2 = \beta_{xy/z}\beta_{yx/z}.$$

Условные средние квадратические отклонения (при двух условиях), характеризующие рассеяние относительно указанных линий регрессии и совпадающие с остаточными средними квадратическими отклонениями, определяются формулами:

$$\sigma_{y/zx} = \sigma_{y/z}\sqrt{1-\rho_{xy/z}^2} = \sigma_{y/x}\sqrt{1-\rho_{yz/x}^2};$$
$$\sigma_{x/yz} = \sigma_{x/z}\sqrt{1-\rho_{xy/z}^2} = \sigma_{x/y}\sqrt{1-\rho_{xz/y}^2}.$$

Центр условного двумерного распределения ($M(x/y)/z$, $M(y/x)/z$) при изменении Z описывает прямую в пространстве 0_{xyz}, в то время, как условные дисперсии $\sigma^2_{x/z}, \sigma^2_{y/z}$ и условный коэффициент корреляции $\rho_{xy/z}$ остаются постоянными.

Условное распределение при заданном (x, y)

Это распределение $z/(x,y)$ является одномерным и определяется своими математическим ожиданием и дисперсией (естественно условными):

$$Mz/(x,y) = M(z/x)/y = M(z/y)/x; \qquad Dz/(x,y) = \sigma^2_{z/xy}.$$

Если точку (x,y) менять, то будем иметь плоскость регрессии z на (x,y)

$$Mz/(x,y) - \mu_z = \beta_{zx/y}(x-\mu_x) + \beta_{zy/x}(y-\mu_y)$$

и остаточную дисперсию относительно плоскости регрессии (совпадающую с условной дисперсией)

$$\sigma^2_{z/xy} = \sigma^2_{z/y}(1-\rho^2_{zx/y}) = \sigma^2_{z/x}(1-\rho^2_{yz/x}).$$

Коэффициент множественной регрессии (совпадающий с соответствующим коэффициентом частной регрессии), например, $\beta_{zx/y}$ показывает, на сколько единиц своего измерения изменится признак z в среднем, если признак x изменится на единицу своего измерения, а остальные признаки не изменятся. Таким образом, коэффициент регрессии может выступать в качестве норматива.

Множественный коэффициент корреляции ρ_z можно вычислить в силу линейности регрессии и как корреляционное отношение z на (x,y):

$$\rho_z = \eta_z = \sqrt{1 - \frac{D_{\text{ост}Z}}{D_{\text{об щ}Z}}} = \sqrt{1 - \frac{\sigma^2_{z/xy}}{\sigma^2_z}}.$$

Если, например, $\rho_{zx/y} = 0$, то из последней формулы следует:

$$\rho^2_z = 1 - \frac{\sigma^2_{z/xy}}{\sigma^2_z} = 1 - \frac{\sigma^2_{z/y}}{\sigma^2_z} = \rho^2_{zy}.$$

3.3.2. Оценивание и проверка значимости параметров

Пусть дана выборка объемом из трехмерной нормально распределенной генеральной совокупности с признаками x, y и z:

$$(x_1,y_1,z_1),\ (x_2,y_2,z_2),\ \ldots,\ (x_n,y_n,z_n).$$

Обработку данных будем производить, руководствуясь таблицей:

x	y	z	x^2	y^2	z^2	xy	xz	yz
x_j	y_j	z_j	x_j^2	y_j^2	z_j^2	x_jy_j	x_jz_j	y_jz_j
$y\Sigma x$	Σy	Σz	Σx^2	Σy^2	Σz^2	Σxy	Σxz	Σyz

Точечные оценки девяти генеральных параметров $\mu_x, \mu_y, \mu_z, \sigma^2_x, \sigma^2_y, \sigma^2_z, \rho_{xy}, \rho_{xz}$ и ρ_{yz} можно вычислить по формулам

$$\bar{x} = \frac{1}{n}\sum x; \quad s_x^2 = \frac{\sum x^2}{n} - (\bar{x})^2; \quad r_{xy} = \frac{\frac{1}{n}\sum xy - \bar{x}\cdot\bar{y}}{s_x s_y};$$

$$\bar{y} = \frac{1}{n}\sum y; \quad s_y^2 = \frac{\sum y^2}{n} - (\bar{y})^2; \quad r_{xz} = \frac{\frac{1}{n}\sum xz - \bar{x}\cdot\bar{z}}{s_x s_z}; \qquad (3.9)$$

$$\bar{z} = \frac{1}{n}\sum z; \quad s_z^2 = \frac{\sum z^2}{n} - (\bar{z})^2; \quad r_{yz} = \frac{\frac{1}{n}\sum yz - \bar{y}\cdot\bar{z}}{s_y s_z}.$$

Затем вычисляются оценки условных средних квадратических отклонений при фиксировании одной компоненты, частных коэффициентов корреляции, условных средних квадратических отклонений при двух фиксированных компонентах и множественных коэффициентов корреляции, используя формулы, соответствующие формулам для вычисления параметров генеральной совокупности:

$$s_{x/y} = s_x\sqrt{1-r_{xy}^2}; \quad s_{x/z} = s_x\sqrt{1-r_{xz}^2}; \quad s_{y/z} = s_y\sqrt{1-r_{yz}^2};$$

$$s_{y/z} = s_y\sqrt{1-r_{xy}^2}; \quad s_{z/x} = s_z\sqrt{1-r_{xz}^2}; \quad s_{z/y} = s_z\sqrt{1-r_{yz}^2};$$

$$r_{xy/z} = \frac{r_{xy} - r_{xz}r_{yz}}{\left[(1-r_{xz}^2)(1-r_{yz}^2)\right]^{1/2}};$$

$$r_{xz/y} = \frac{r_{xz} - r_{xy}r_{yz}}{\left[(1-r_{xy}^2)(1-r_{yz}^2)\right]^{1/2}}; \qquad (3.10)$$

$$r_{yz/x} = \frac{r_{yz} - r_{xy}r_{xz}}{\left[(1-r_{xy}^2)(1-r_{xz}^2)\right]^{1/2}};$$

$$s_{x/yz} = s_{x/y}\sqrt{1-r^2_{xz/y}}; \quad s_{y/xz} = s_{y/z}\sqrt{1-r^2_{xy/z}}; \qquad (3.11)$$

$$s_{z/xy} = s_{z/x}\sqrt{1-r^2_{yz/x}};$$

$$r_x = \sqrt{1-\frac{s^2_{x/yz}}{s^2_x}}; \quad r_y = \sqrt{1-\frac{s^2_{y/xz}}{s^2_y}}; \quad r_z = \sqrt{1-\frac{s^2_{z/xy}}{s^2_z}}.$$

Проверка значимости множественного коэффициента детерминации ρ^2_M (следовательно, и ρ_M) осуществляется с помощью F-распределения.

Вычисляется

$$F_{набл} = \frac{r^2_M / 2}{(1-r^2_M)/(n-3)}. \qquad (3.12)$$

Затем с заданным уровнем значимости α и числами степеней свободы ν_1=2 (числителя) и ν_1=n-3 (знаменателя) находят $F_{табл}$. Если $F_{набл} > F_{табл}$, то гипотеза $H_0: \rho^2_M = 0$ отвергается с вероятностью ошибки α, т.е. ρ^2_M значимо отличается от нуля. Если коэффициент ρ_z незначим, связь между случайной величиной Z и случайной величиной (x,y) отсутствует.

Конечно, проверку значимости коэффициентов связи начинать с частных коэффициентов корреляции не обязательно. Можно в некоторых случаях сократить такую проверку, например, если ρ_z незначим, то коэффициенты $\rho_{zx/y}$ и $\rho_{zy/x}$ становятся незначимыми. Далее, если $\rho_{zx/y}$ незначим, то $\rho_z = |\rho_{zy}|$ (множественный коэффициент корреляции незначимо отличается от абсолютной величины парного коэффициента корреляции).

Для значимых множественных коэффициентов корреляции можно получить оценки уравнения регрессии.

Например, пусть ρ_z значимо отличается от нуля, тогда оценкой соответствующего уравнения регрессии служит

$$\overline{z/(x,y)} - \bar{z} = b_{zx/y}(x-\bar{x}) + b_{yz/x}(y-\bar{y}). \qquad (3.13)$$

При этом коэффициенты регрессии вычисляются по формулам:

$$b_{zx/y} = r_{zx/y}\frac{s_{z/y}}{s_{x/y}}; \quad b_{zy/x} = r_{zy/x}\frac{s_{z/x}}{s_{y/x}} \qquad (3.14)$$

и $\overline{z/(x,y)}$ является оценкой $Mz/(x,y)$.

Напомним, что если какой-либо частный коэффициент корреляции незначим, то соответствующий коэффициент плоскости регрессии также незначим. Поэтому, если позволяют условия практического анализа, с точки зрения надежности статистических выводов, предпочтительнее рассматривать модель взаимозависимости признаков такую, для которой множественный коэффициент детерминации - наибольший (и, конечно,

значимый): ему соответствует максимальное число значимых частных коэффициентов детерминации (корреляции).

Для значимых параметров связи представляет интерес найти интервальную оценку с надежностью $\gamma = 1-\alpha$.

Интервальная оценка для $\rho_{частн}$ находится с помощью статистики Фишера

$$Z_r = Z(r) = \frac{1}{2}\ln\frac{1+r}{1-r}.$$

По таблице указанного преобразования находят величину $Z/r_{частн}$. Затем вычисляют точность интервальной оценки для MZ воспользовавшись тем фактом, что статистика $Z(r)$ распределена приближенно нормально с параметрами $MZ \cong \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ и $DZ \cong \frac{1}{n-4}$:

$$\Delta Z = t_\gamma \frac{1}{\sqrt{n-4}},$$

где t_γ является решением уравнения $\Phi(t)=\gamma$ и находится по таблице интегральной функции Лапласа.

Затем вычисляются границы интервальной оценки для MZ по формуле

$$Z(r_{част}) \pm \Delta Z, \qquad (3.15)$$

и, наконец, доверительные границы для $\rho_{частн}$ получают по таблице обратного преобразования Фишера.

Для значимого множественного коэффициента корреляции интервальная оценка также находится с помощью Z-преобразования Фишера, с дисперсией, приблизительно равной $\frac{1}{n}$ для достаточно больших значений n. Имеются графики и таблицы (Эзекиела и Фокса; К.Крамера) для получения интервальных оценок ρ^2_M по значениям r^2_M.

Определение доверительных интервалов для коэффициентов плоскости регрессии производится исходя из статистик

$$t = \frac{(b_{zx/y} - \beta_{zx/y})s_{x/y}\sqrt{n-3}}{s_{z/y}\sqrt{1-r^2_{zx/y}}}; \qquad (3.16)$$

$$t = \frac{(b_{zy/x} - \beta_{zy/x})s_{y/x}\sqrt{n-3}}{s_{z/x}\sqrt{1-r^2_{zy/x}}},$$

которые имеют t-распределение Стьюдента с $\nu=n-3$ степенями свободы. Для этого достаточно решить относительно оцениваемого коэффициента

регрессии неравенство $|t| \le t(\alpha, n-3)$, где $t(\alpha, n-3)$ находится по таблице Стьюдента.

Для значимых частных и множественных коэффициентов детерминации можно указать более предпочтительные точечные оценки, чем выборочные коэффициенты, например,

$$\frac{(n-2)r^2_{xy/z}-1}{n-3}$$ - оценка для $\rho^2_{xy/z}$;

$$\frac{(n-1)r^2_{z}-2}{n-3}$$ - оценка для ρ_z.

3.3.3. Корреляционный анализ показателей эффективности работы предприятий

С целью анализа взаимосвязи показателей эффективности производства продукции: производительности труда, фондоотдачи и материалоемкости продукции была отобрана группа 25 однотипных машиностроительных предприятий.

На основании годовых отчетов предприятий были получены следующие данные (табл.3.3):

x - выработка валовой продукции в неизменных ценах на одного работающего средней списочной численности ППП (млн руб.);

y - выпуск валовой продукции на один рубль среднегодовой стоимости основных промышленно-производственных фондов (руб.);

z - материалоемкость в стоимостном выражении: стоимость материалов в валовой продукции в неизменных ценах (%).

Таблица 3.3

Исходные данные

№ предприятия	x	y	z	№ предприятия	x	y	z
1	6,0	2,0	25	14	5,7	2,2	25
2	4,9	0,8	30	15	5,1	1,3	30
3	7,0	2,7	20	16	5,2	1,5	14
4	6,7	3,0	21	17	7,3	2,7	20
5	5,8	1,0	28	18	6,1	2,4	27
6	6,1	2,3	26	19	6,2	2,2	28
7	5,0	0,9	30	20	5,9	2,0	26
8	6,9	2,6	22	21	6,0	2,0	26
9	6,8	3,0	20	22	4,8	0,9	31
10	5,9	1,1	29	23	7,3	3,2	19
11	5,0	0,8	27	24	7,2	3,3	20
12	5,6	2,2	25	25	7,0	3,0	20
13	6,0	2,4	24				

Предположим, что рассматриваемые признаки x, y и z в генеральной совокупности подчиняются нормальному закону распределения и указанные данные представляют выборку из этой совокупности.

Для получения точечных оценок генеральных средних, дисперсий средних квадратических отклонений и парных коэффициентов корреляции результаты промежуточных вычислений удобно поместить в расчетную таблицу (табл.3.4). С целью контроля вычислений данные разбиты на пятерки. Для каждой пятерки в итоговой (выделенной) строке приведены суммы элементов каждого столбца. В последнем контрольном столбце приводятся суммы элементов соответствующих строк. Сумма пяти элементов контрольного столбца должна совпадать с суммой элементов итоговой строки.

Таблица 3.4

Расчетная таблица

x	y	z	x^2	y^2	z^2	xy	xz	yz	Контроль
6,0	2,0	25	36,00	9,00	625	12,00	150,0	50,0	910,00
4,9	0,8	30	24,01	0,64	900	3,92	147,0	24,0	1135,27
7,0	2,7	20	49,00	7,29	400	18,90	140,0	54,0	698,89
6,7	3,0	21	44,89	9,00	441	20,10	140,7	63,0	749,39
5,8	1,0	28	33,64	1,00	784	5,80	162,4	28,0	1049,64
30,4	**9,5**	**124**	**187,54**	**21,93**	**3150**	**60,72**	**740,1**	**219,0**	**4543,19**
6,1	2,1	26	37,21	4,41	676	12,81	158,6	54,6	977,83
5,0	0,9	30	25,00	0,81	900	4,50	150,0	27,0	1143,21
6,9	2,6	22	47,61	6,76	484	17,94	151,8	57,2	796,81
6,8	3,0	20	46,21	9,00	400	20,40	136,0	60,0	701,44
5,9	1,1	29	34,81	1,21	841	6,49	171,1	31,9	1122,51
30,7	**9,7**	**127**	**190,87**	**22,19**	**3301**	**62,14**	**767,5**	**230,7**	**4741,80**
5,0	0,8	27	25,00	0,64	729	4,00	135,0	21,6	948,04
5,6	2,2	25	31,36	4,84	625	12,32	140,0	55,0	301,32
6,0	2,4	24	36,00	5,76	576	14,40	144,0	57,6	866,16
5,7	2,2	25	32,49	4,84	625	12,54	142,5	55,0	905,27
5,1	1,3	30	26,01	1,69	900	6,63	153,0	39,0	1162,73
27,4	**8,9**	**131**	**150,86**	**17,77**	**3455**	**49,89**	**714,5**	**228,2**	**4783,52**
5,2	1,5	24	27,04	2,25	576	7,80	124,8	36,0	804,59
7,3	2,7	20	53,29	7,29	400	19,71	146,0	54,0	710,29
6,1	2,4	27	37,21	5,76	729	14,64	164,7	64,8	1051,61
6,2	2,2	28	38,44	4,84	784	13,64	173,6	61,6	1112,52
5,9	2,0	26	34,81	4,00	676	11,80	153,4	52,0	965,91
30,7	**10,8**	**125**	**190,79**	**24,14**	**3165**	**67,59**	**762,5**	**268,4**	**4644,92**
6,0	2,0	26	36,00	4,00	676	12,00	156,0	52,0	970,00
4,8	0,9	31	23,04	0,81	961	4,32	148,8	27,9	1202,57
7,3	3,2	19	53,29	10,24	361	23,36	138,7	60,8	676,89
7,2	3,3	20	51,84	10,89	400	23,76	144,0	66,0	726,99
7,0	3,0	20	43,00	9,00	400	21,00	140,0	60,0	709,00
32,3	**12,4**	**116**	**213,17**	**34,94**	**2798**	**84,44**	**727,5**	**266,7**	**4285,45**

Например,

6,0+4,9+7,0+6,7+5,8=30,4;

6,0+2,0+25+36,0+9,0+625+12,0+150+50=910,0.

Контроль:

30,4+9,5+124+187,54+21,93+3150+60,72+740,1+219,0=910,00+

+11,3527+698,89+749,39+1049,64=4543,19.

Далее получаем таблицу итоговых строк (табл.3.5).

Таблица 3.5

Итоговая таблица

1	2	3	4	5	6	7	8	9	10
30,4	9,5	124	187,54	21,93	3150	60,72	740,1	219,0	4543,19
30,7	9,7	127	190,87	22,19	3301	62,14	767,5	230,7	4741,80
27,4	8,9	131	150,86	17,77	3455	49,89	714,5	228,2	4783,52
30,7	10,8	125	190,79	24,14	3165	67,59	762,5	268,4	4644,92
32,3	12,4	116	213,17	34,94	2798	84,44	727,5	266,7	4285,45
151,5	**51,3**	**623**	**933,23**	**120,97**	**15869**	**324,78**	**3712,1**	**1213,0**	**22798,88**

Из последней строки итоговой таблицы получаем

$$\Sigma x=151,5;\ \Sigma y=51,3;\ \Sigma z=623;$$
$$\Sigma x^2=933,23;\ \Sigma y^2=120,97;\ \Sigma z^2=15869;$$
$$\Sigma xy=324,78;\ \Sigma xz=3712,1;\ \Sigma yz=1213,0.$$

Следовательно,

$$\bar{x}=\frac{1}{25}\cdot 151,5=6,06;\quad \overline{x^2}=\frac{1}{25}\cdot 933,23=37,3292;$$
$$\bar{y}=\frac{1}{25}\cdot 51,3=2,052;\quad \overline{y^2}=\frac{1}{25}\cdot 120,97=4,8388;$$
$$\bar{z}=\frac{1}{25}\cdot 623=24,92;\quad \overline{z^2}=\frac{1}{25}\cdot 15869=634,76;$$
$$s_x^2=x^2-(\bar{x})^2=37,3292-6,06^2=0,6056;\ s_x=0,778203;$$
$$s_y^2=4,8388-2,052^2=0,628096;\ s_y=0,792525;$$
$$s_z^2=634,76-24,92^2=13,7536;\ s_z=3,7085846.$$

Для вычисления точечных оценок парных коэффициентов корреляции используем формулу

$$r_{xy}=\frac{\sum xy-(\sum x)(\sum y)/n}{\left\{\left[\sum x^2-(\sum x)^2/n\right]\left[\sum y^2-(\sum y)^2/n\right]\right\}^{1/2}}.$$

Получаем

$$r_{xy}=\frac{324,78-151,5\cdot 51,3/25}{\left[(933,23-151,5^2/25)(120,97-51,3^2/25)\right]^{1/2}}=$$
$$=\frac{13,902}{\sqrt{15,14\cdot 15,7024}}=0,9016362;$$

$$r_{xz}=\frac{3712,1-151,5\cdot 623/25}{\left[15,14(15869-623^2/25)\right]^{1/2}}=\frac{-63,28}{\sqrt{15,14\cdot 343,84}}=-0,8770519;$$

$$r_{yz}=\frac{1213,0-51,3\cdot 623/25}{\sqrt{15,7024\cdot 343,84}}=-0,8899999.$$

Вычисляем точечные оценки условных средних квадратических отклонений (при фиксировании одной переменной):

$$s_{x/y}=\sqrt{s_x^2(1-r_{xy}^2)}=\sqrt{0,6056(1-0,9016362^2)}=0,3365691;$$
$$s_{y/x}=\sqrt{s_y^2(1-r_{xy}^2)}=\sqrt{0,628096\cdot 0,1870522}=0,3427633;$$
$$s_{x/z}=0,3738452;\qquad s_{y/z}=0,3613602;$$
$$s_{z/x}=1,7815879;\qquad s_{z/y}=1,6909689.$$

Получаем точечные оценки частных коэффициентов корреляции:

$$r_{xy/z}=\frac{r_{xy}-r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}=\frac{0,9016362-0,8770519\cdot 0,8899999}{\sqrt{0,23078\cdot 0,2079002}}=$$
$$=0,5526811;$$
$$r_{xz/y}=\frac{-0,8770519+0,9016362\cdot 0,8899999}{\sqrt{0,1870552\cdot 0,2079002}}=-0,3782736;$$
$$r_{yz/x}=\frac{-0,8899999+0,9016362\cdot 0,8770519}{\sqrt{0,1870552\cdot 0,23078}}=-0,4775413.$$

Вычисляем точечные оценки остаточных дисперсий (при фиксировании двух переменных):

$$s_{x/yz}^2=s_{x/z}^2(1-r_{xy/z}^2)=0,3738452^2(1-0,5526811^2)=0,0970695;$$
$$s_{y/xz}^2=s_{y/z}^2(1-r_{xy/z}^2)=0,0906942;$$
$$s_{z/xy}^2=s_{z/x}^2(1-r_{yz/x}^2)=2,450226.$$

Получаем оценки множественных коэффициентов детерминации и корреляции:

$$r_x^2=1-\frac{s_{x/yz}^2}{s_x^2}=1-\frac{0,0970695}{0,6056}=0,8397136;\ r_x=0,9163588;$$
$$r_y^2=1-\frac{s_{y/xz}^2}{s_y^2}=1-\frac{0,906942}{0,628096}=0,8556046;\ r_y=0,9249889;$$

$$r_z^2 = 1 - \frac{s^2_{z/xy}}{s_z^2} = 1 - \frac{2{,}450226}{13{,}7536} = 0{,}8218484\,;\quad r_z = 0{,}9065585\,.$$

Для контроля вычислений полезно воспользоваться другими формулами точечных оценок множественных коэффициентов корреляции, например,

$$r_x = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - 2r_{xy}r_{xz}r_{yz}}{1 - r_{yz}^2}} = \sqrt{\frac{0{,}1745766}{0{,}2079002}} = \sqrt{0{,}8397134} = 0{,}9163587\,.$$

Ошибка в последнем разряде на единицу является допустимой.

Проверим с уровнем α=0,05 значимость множественных коэффициентов корреляции. Вычислим наблюдаемые значения F-критерия:

$$F_{набл}\ (r_x^2) = \frac{r_x^2\,/\,2}{(1 - r_x^2)\,/\,(n-3)} = \frac{0{,}8397136 \cdot 22}{2 \cdot 0{,}1602864} = 57{,}627157;$$

$$F_{набл}\ (r_y^2) = \frac{0{,}8556046 \cdot 22}{2 \cdot 0{,}1443954} = 65{,}179712;$$

$$F_{набл}\ (r_z^2) = \frac{0{,}8218484 \cdot 22}{2 \cdot 0{,}1781516} = 50{,}745165.$$

Находим, по таблице F-распределения критическое значение F-статистик для уровня значимости α=0,05, числа степеней свободы числителя ν_1=2 и знаменателя ν_2=n–3=22:

$$F_{кр}\ (0{,}05;\ 2;\ 22) = 3{,}44\,.$$

Так как все наблюдаемые значения F-статистик превосходят ее критическое значение, то гипотеза о равенстве нулю каждого множественного коэффициента корреляции генеральной совокупности отвергается с вероятностью ошибки, равной 0,05. Следовательно, эти коэффициенты значимо отличаются от нуля.

Проверим значимость частных коэффициентов корреляции с уровнем значимости α=0,05. По таблице распределения r-статистики для числа степеней свободы ν=n–3=25–3=22 и α=0,05 находим:

$$r_{кр}\ (0{,}05;\ 25) < r_{кр}\ (0{,}05;\ 22) < r_{кр}\ (0{,}05;\ 20)$$

или

$$0{,}381 < r_{кр}\ (0{,}05;\ 22) < 0{,}423\,.$$

Так как наблюдаемые значения $|r|$, $|r_{xy/z}|$ и $|r_{yz/x}|$ больше 0,423, то они будут превосходить и $r_{кр}$(0,05; 22). Следовательно, гипотеза о равенстве нулю генеральных частных коэффициентов корреляции $\rho_{xy/z}$ и $\rho_{yz/x}$ отвергается с вероятностью ошибки 0,05. Наблюдаемое значение r-статистики

для частного коэффициента корреляции между x и y меньше 0,381, т.е. и подавно меньше, чем $r_{кр}$(0,05; 22). Следовательно, гипотеза H_0 : $\rho_{xz/y}$=0 не отвергается. Примем эту гипотезу, т.е. будем считать, что генеральный частный коэффициент корреляции между x и z равен нулю, т.е. незначим.

Итак, значимыми оказались только два частных коэффициента корреляции $\rho_{xy/z}$ и $\rho_{yz/x}$.

Заметим, что в случае необходимости можно было бы вычислить $r_{кр}$(0,05; 22), например, с помощью линейной интерполяции:

$$r_{кр}\ (0{,}05;\ 22) = 0{,}381 + \frac{0{,}423 - 0{,}381}{25 - 20}(25 - 22) = 0{,}4062\,.$$

С надежностью γ =1–α =0,95 найдем интервальные оценки значимых частных коэффициентов корреляции.

Для $\rho_{xy/z}$ получаем по таблице функции $Z_r = \frac{1}{2}\ln\frac{1+r}{1-r}$ значение $Z_{0{,}55}$= =0,6184. Затем вычисляем точность интервальной оценки для MZ

$$\Delta Z = t_{0{,}95}\frac{1}{\sqrt{n-4}} = 1{,}96 \cdot \frac{1}{\sqrt{21}} = 0{,}4277$$

($t_{0{,}95}$ находится по таблице интеграла Лапласа). Интервальная оценка для MZ есть:

$$0{,}6184 - 0{,}4277 \le MZ \le 0{,}6184 + 0{,}4277$$

или

$$0{,}1907 \le MZ \le 1{,}0461\,.$$

Чтобы получить интервальную оценку для $\rho_{xy/z}$ по таблицам Z_r, находим числа, ближайшие к 0,1907 и к 1,0461. Это будут числа $Z_{r_{min}} = 0{,}1923$ и $Z_{r_{max}} = 1{,}0454$. Переходя к аргументу r, получим

$$0{,}19 \le \rho_{xy/z} \le 0{,}78.$$

Для $\rho_{yz/x}$ будем иметь $Z_{-0{,}48} = -Z_{0{,}48} = -0{,}5230$.

Тогда

$$-0{,}5230 - 0{,}4277 \le MZ \le -0{,}5230 + 0{,}4277$$

или

$$-0{,}9507 \le MZ \le 0{,}0953.$$

Ближайшими к числам 0,0953 и 0,9507 будут 0,1003 и 0,9505. Следовательно, интервальная оценка для $\rho_{yz/x}$ есть

$$-0{,}74 \le \rho_{yz/x} \le -0{,}10.$$

Для изучаемых трех признаков нет односторонней причинно-следст-

венной связи. Поэтому выберем в качестве модели взаимозависимости статистически наиболее надежную в условиях данной выборки. Значимым множественным коэффициентом детерминации, имеющим наибольшую оценку является ρ_y^2. Ему соответствуют оба значимых частных коэффициента детерминации $\rho_{xy/z}^2$ и $\rho_{yz/x}^2$ (имеющих и оценки больше, чем оценка $\rho_{xz/y}^2$).

Оценка соответствующего уравнения регрессии имеет вид

$$\overline{y/(x,z)} - \bar{y} = b_{yx/z}(x-\bar{x}) + b_{yz/x}(z-\bar{z}).$$

Вычислим коэффициенты множественной (частной) регрессии:

$$b_{yx/z} = r_{xy/z}\frac{s_{y/z}}{s_{x/z}} = 0{,}5526811\frac{0{,}3613601}{0{,}3738452} = 0{,}5342236;$$

$$b_{yz/x} = r_{yz/x}\frac{s_{y/x}}{s_{z/x}} = -0{,}4775413\frac{0{,}3427633}{1{,}7815879} = -0{,}0918751.$$

Таким образом, получаем

$$\overline{y/(x,z)} - 2{,}052 = 0{,}5342(x-6{,}06) - 0{,}091875(z-24{,}92).$$

Найдем с надежностью $\gamma = 1-\alpha = 0{,}95$ интервальные оценки для множественных коэффициентов регрессии. По таблице распределения Стьюдента находим $t(\alpha, n-3) = t(0{,}05;\ 22) = 2{,}074$ и решим относительно β неравенства

$$\left|\frac{(0{,}5342236-\beta_{yx/z})0{,}3738452\sqrt{22}}{0{,}3613602\sqrt{1-0{,}5526811^2}}\right| \le 2{,}074;$$

$$\left|\frac{(-0{,}0918751-\beta_{yz/x})1{,}7815879\sqrt{22}}{0{,}3427633\sqrt{1-0{,}4775413^2}}\right| \le 2{,}074.$$

Вычислим точности интервальных оценок:

$$\Delta b_{yx/z} = 2{,}074\frac{0{,}3613602\sqrt{1-0{,}5526811^2}}{0{,}3738452\sqrt{22}} = 0{,}3562008;$$

$$\Delta b_{yz/x} = 2{,}074\frac{0{,}3427633\sqrt{1-0{,}4775413^2}}{1{,}7815879\sqrt{22}} = 0{,}0747445.$$

Откуда

$$0{,}5342236-0{,}3562008 \le \beta_{yx/z} \le 0{,}5342236+0{,}3562008;$$
$$-0{,}0918751-0{,}0747445 \le \beta_{yz/x} \le -0{,}0918751+0{,}0747445$$

или

$$0{,}1780 \le \beta_{yx/z} \le 0{,}8904;$$

$$-0{,}0171306 \le \beta_{yz/x} \le -0{,}1666196.$$

Получим предпочтительные точечные оценки изучаемых коэффициентов детерминации:

$$\hat{\rho}_{xy/z}^2 = \frac{(n-2)r_{yx/z}^2 - 1}{n-3} = \frac{23\cdot 0{,}5526811^2 - 1}{22} = 0{,}2738861;$$

$$\hat{\rho}_{yz/x}^2 = \frac{23\cdot 0{,}4775413^2 - 1}{22} = 0{,}1929567;$$

$$\hat{\rho}_y^2 = \frac{(n-1)r_y^2 - 2}{n-3} = \frac{24\cdot 0{,}8556046 - 2}{22} = 0{,}8424777.$$

На основании полученных расчетов можно сделать следующие ***выводы***.

Доказана тесная взаимосвязь каждого из исследуемых показателей эффективности работы предприятия с другими (все множественные коэффициенты детерминации значимы и превышают 0,8).

Особенно тесная связь существует между фондоотдачей и двумя остальными показателями - производительностью труда и материалоемкостью. Изменение фондоотдачи в среднем на 84,25% объясняется изменением производительности труда и материалоемкости (изменение фондоотдачи в среднем на 15,75% объясняется влиянием неконтролируемых факторов, признаков). При этом при увеличении производительности труда на 1 млн руб. фондоотдача увеличивается в среднем на 0,55 руб. на рубль основных производственных фондов. При уменьшении материалоемкости на 1% фондоотдача увеличивается в среднем на 0,48%. Указанные нормативы относительно стабильны при условии, что изучаемые показатели отклоняются на небольшие величины от своих средних уровней (стабильность указывается доверительными интервалами и вероятностью 0,95). Взаимозависимость между материалоемкостью и производительностью труда (без учета фондоотдачи) не доказана (частный коэффициент корреляции $\rho_{xz/y}$ незначим) при данных условиях. Для более надежной проверки такой зависимости необходим большой объем выборки.

3.4. Некоторые вопросы корреляционного анализа многомерной модели

Напомним, что нормальный закон распределения генеральной совокупности с k изучаемыми признаками $x_1, x_2, \ldots, x_k$ определяется $2k + \frac{k(k-1)}{2}$ параметрами:

k - математическими ожиданиями $\mu_i = Mx_i$, $i = 1 \div k$;

k - дисперсиями $\sigma_i^2 = Dx_i$, $i=1\div k$;

$\frac{k(k-1)}{2}$ - парными коэффициентами корреляции

$$\rho_{ij} = M[\frac{x_i - \mu_i}{\sigma_i} \cdot \frac{x_j - \mu_j}{\sigma_j}];\ i<j;\ i,j=1\div k.$$

Корреляционная матрица

$$\Re = \Re_k = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2k} \\ \rho_{31} & \rho_{32} & 1 & \dots & \rho_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{k1} & \rho_{k2} & \rho_{k3} & \dots & 1 \end{pmatrix} \qquad (3.17)$$

симметрична и положительно определена.

Для анализа модели удобно внести понятие параметров связи l-го порядка.

Исходная система случайных величин состоит из всех рассматриваемых признаков $\{x_1,x_2,\dots,x_k\}$. Изучим корреляцию между какими-нибудь двумя признаками, когда среди оставшихся $(k-2)$ признаков фиксируются не все, а только l определенные признаки. В таком случае мы получаем частный коэффициент l-го порядка. Далее, если мы интересуемся множественным коэффициентом корреляции или коэффициентами регрессии одной величины на подмножество, содержащее l признаков среди оставшихся $(k-1)$ признаков, то мы имеем множественный коэффициент корреляции или коэффициенты регрессии l-го порядка. Будем обозначать коэффициенты l-го порядка, указывая справа от черты множество l фиксируемых признаков. Так, порядок $\rho_{12/45}$=2, порядок ρ_{12}=0, порядок $\beta_{12/346}$=3 (коэффициент регрессии x_1 на x_2 уравнения регрессии x_1 на множество $\{x_2,x_3,x_4,x_6\}$ коэффициент частной регрессии). $\rho_{3/12}$ - множественный коэффициент корреляции второго порядка; $\rho^2_{3/1} = \rho^2_{13}$ - множественный коэффициент детерминации первого порядка, равный парному коэффициенту детерминации; $\sigma^2_{1/3678}$ - остаточная дисперсия четвертого порядка. Индексы перед чертой называют **первичными**, после черты - **вторичными**.

Частный коэффициент корреляции, например, между x_1 и x_2 по отношению к величинам $x_3,x_4,\dots,x_k$ определяется формулой:

$$\rho_{12/34\dots k} = -\frac{\Re_{12}}{(\Re_{11}\Re_{22})^{1/2}}, \qquad (3.18)$$

где $\Re_{12}$ - алгебраическое дополнение к элементу ρ_{12} корреляционной матрицы $\Re_k$;

$\Re_{11}$ - алгебраическое дополнение к элементу $\rho_{11}=1$;

$\Re_{22}$ - алгебраическое дополнение к элементу $\rho_{22}=1$.

Другие частные коэффициенты корреляции определяются аналогично. Очевидно, здесь мы имеем частные коэффициенты корреляции порядка $k-2$.

Для определения частного коэффициента корреляции любого порядка l (от 0 до $k-2$) следует рассмотреть подматрицу $l+2$-го порядка $\Re_{l+2}$ матрицы $\Re_k$, составленную из строк и столбцов Q_k, отвечающих индексам вычисляемого коэффициента, и воспользоваться вышеприведенной определяющей формулой.

Получим, например, выражение для частного коэффициента второго порядка $\rho_{34/26}$. Рассмотрим подматрицу четвертого порядка, нумеруя ее элементы в обычном порядке:

$$\begin{pmatrix} 1 & \rho_{23} & \rho_{24} & \rho_{26} \\ \rho_{23} & 1 & \rho_{34} & \rho_{36} \\ \rho_{24} & \rho_{34} & 1 & \rho_{46} \\ \rho_{26} & \rho_{36} & \rho_{46} & 1 \end{pmatrix} = \begin{pmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{23} & q_{24} \\ q_{31} & q_{32} & q_{33} & q_{34} \\ q_{41} & q_{42} & q_{43} & q_{44} \end{pmatrix} = Q_4 .$$

Тогда

$$\rho_{34/26} = -\frac{Q_{23}}{(Q_{22}\cdot Q_{33})^{1/2}} =$$

$$= -\frac{(-1)^{2+3}\begin{vmatrix} q_{11} & q_{12} & q_{14} \\ q_{31} & q_{32} & q_{34} \\ q_{41} & q_{42} & q_{44} \end{vmatrix}}{\left((-1)^{2+2}\begin{vmatrix} q_{11} & q_{13} & q_{14} \\ q_{31} & q_{33} & q_{34} \\ q_{41} & q_{43} & q_{44} \end{vmatrix} (-1)^{3+3}\begin{vmatrix} q_{11} & q_{12} & q_{14} \\ q_{21} & q_{22} & q_{24} \\ q_{41} & q_{42} & q_{44} \end{vmatrix} \right)^{1/2}};$$

$$\rho_{34/26} = -\frac{\begin{vmatrix} 1 & \rho_{23} & \rho_{26} \\ \rho_{24} & \rho_{34} & \rho_{46} \\ \rho_{26} & \rho_{36} & 1 \end{vmatrix}}{\left(\begin{vmatrix} 1 & \rho_{24} & \rho_{26} \\ \rho_{24} & 1 & \rho_{46} \\ \rho_{26} & \rho_{46} & 1 \end{vmatrix} \cdot \begin{vmatrix} 1 & \rho_{28} & \rho_{26} \\ \rho_{23} & 1 & \rho_{36} \\ \rho_{26} & \rho_{36} & 1 \end{vmatrix} \right)^{1/2}}.$$

Формула для остаточной дисперсии, например, регрессии x_1 на x_2, x_3,..., x_k имеет вид

$$\sigma^2_{1/23\dots k} = \sigma^2_1 \frac{|\Re_k|}{\Re_{11}},$$

где $|\Re_k|$ - определитель матрицы $\Re_k$.

Формулы остаточных дисперсий $(k-1)$ порядка других регрессий имеют аналогичный вид. Для получения различных остаточных дисперсий

любого порядка следует рассматривать соответствующие корреляционные подматрицы, как было указано при вычислении частного коэффициента корреляции.

Множественный коэффициент корреляции ((*k*–1) порядка), например, между x_1 и $(x_2,\ldots, x_k)$ определяется выражением

$$\rho_1 = \rho_{1/23\ldots k} = \sqrt{1 - \frac{\sigma^2_{1/23\ldots k}}{\sigma_1^2}} = \sqrt{1 - \frac{|\mathfrak{R}_k|}{\mathfrak{R}_{11}}}. \qquad (3.19)$$

Множественные коэффициенты корреляции (*k*–1) порядка остальных случайных величин и других порядков определяются, очевидно, из соответствующих подматриц исходной матрицы $\mathfrak{R}_k$.

Уравнение регрессии, например, x_1 на $x_2,\ldots,x_k$ имеет вид:

$$Mx_1 / (x_2,\ldots,x_k) = \mu_1 + \sum_{i=2}^{k} \beta_{1i/23\ldots(i-1)(i+1)\ldots k}(x_i - \mu_i), \qquad (3.20)$$

где $\beta_{1i/23\ldots(i-1)(i+1)\ldots k}$ - коэффициенты частной регрессии (*k*–1) порядка, определяемые по формулам

$$\beta_{1i/23\ldots(i-1)(i+1)\ldots k} = -\frac{\sigma_1}{\sigma_i}\cdot\frac{\mathfrak{R}_{1i}}{\mathfrak{R}_{11}}, \qquad (3.21)$$

где i=1,2,...,k.

Другие регрессии определяются аналогично, исходя из соответствующих перемен индексов и выбранных подматриц.

Приведем полезное для анализа свойство некоторых показателей связи различных порядков.

Проверим при k=4 равенство

$$\sigma^2_{1/23\ldots k} = \sigma_1^2\frac{|\mathfrak{R}_k|}{\mathfrak{R}_{11}} = \sigma_1^2(1-\rho^2_{1k})(1-\rho^2_{1k-1/k})\cdot(1-\rho^2_{1k-2/k-1k})\ldots(1-\rho^2_{12/3\ldots k}).$$

Имеем

$$\sigma^2_{1/234} = \sigma_1^2\frac{|\mathfrak{R}_4|}{\mathfrak{R}_{11}} = \frac{\begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_{23} & \rho_{24} \\ \rho_{23} & 1 & \rho_{34} \\ \rho_{24} & \rho_{34} & 1 \end{vmatrix}}\cdot\sigma_1^2;$$

$$\sigma^2_{1/34} = \sigma_1^2\frac{|Q_3|}{Q_{11}} = \frac{\begin{vmatrix} 1 & \rho_{13} & \rho_{14} \\ \rho_{13} & 1 & \rho_{34} \\ \rho_{14} & \rho_{34} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_{34} \\ \rho_{34} & 1 \end{vmatrix}}\cdot\sigma_1^2;$$

тогда

$$\frac{\sigma^2_{1/234}}{\sigma^2_{1/34}} = \frac{\begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & \rho_1 \end{vmatrix}\cdot\begin{vmatrix} 1 & \rho_{34} \\ \rho_{31} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_{23} & \rho_{24} \\ \rho_{23} & 1 & \rho_{34} \\ \rho_{24} & \rho_{34} & 1 \end{vmatrix}\cdot\begin{vmatrix} 1 & \rho_{13} & \rho_{14} \\ \rho_{13} & 1 & \rho_{34} \\ \rho_{14} & \rho_{34} & 1 \end{vmatrix}}.$$

По отношению к элементам матрицы $\mathfrak{R}$ согласно **обобщенной теореме Якоби об определителях** получим

$$\frac{\sigma^2_{1/234}}{\sigma^2_{1/34}} = \frac{\begin{vmatrix} \mathfrak{R}_{11} & \mathfrak{R}_{12} \\ \mathfrak{R}_{12} & \mathfrak{R}_{22} \end{vmatrix}}{\mathfrak{R}_{11}\mathfrak{R}_{22}} = 1 - \frac{\mathfrak{R}^2_{12}}{\mathfrak{R}_{11}\mathfrak{R}_{22}},$$

откуда

$$\sigma^2_{1/234} = \sigma^2_{1/34}(1-\rho^2_{12/34}).$$

Согласно последней рекуррентной формуле будем иметь

$$\sigma^2_{1/34} = \sigma^2_{1/4}(1-\rho^2_{13/4});$$
$$\sigma^2_{1/4} = \sigma^2_1(1-\rho^2_{14}).$$

Таким образом,

$$\sigma^2_{1/234} = \sigma^2_1(1-\rho^2_{14})(1-\rho^2_{13/4})(1-\rho^2_{12/34}).$$

Из приведенного равенства следует, что

$$\sigma^2_{1/23\ldots k} \le \sigma_1 / \{l\},$$

где $\{l\}$ - любое подмножество индексов случайных величин из множества $\{2,3,\ldots,k\}$.

Но тогда для множественного коэффициента детерминации получаем цепочку неравенств:

$$\rho^2_{1/2}(=\rho^2_{12}) \le \rho^2_{1/23} \le \rho_{1/234} \le \ldots \le \rho^2_{1/23\ldots k}(=\rho^2_1),$$

указывающую на то, что коэффициент множественной детерминации не

уменьшается при добавлении признаков, относительно которых измеряется зависимость x_1.

Проверка значимости частных коэффициентов корреляции различных порядков осуществляется по тому же критерию, по которому проверяется значимость частного коэффициента нулевого порядка, но число степеней свободы теперь $\nu = n-l-2$ (l-порядок).

Проверка значимости множественного коэффициента корреляции (точнее, коэффициента детерминации) осуществляется с помощью F-распределения, где

$$F_{набл} = \frac{r_M^2 / l}{(1 - r_M^2)/(n - l - 1)}, \qquad (3.22)$$

l-порядок r_M, $\nu_1 = l$ и $\nu_2 = n-l-1$.

Для значимых частных коэффициентов корреляции интервальные оценки находятся с помощью логарифмического преобразования Фишера исходя из дисперсии $1/(n-l-3)$.

Для значимых коэффициентов частной регрессии интервальные оценки находятся с помощью распределения статистики Стьюдента, например,

$$t = (b_{12/\{l\}} - \beta_{12/\{l\}}) \left\{ \frac{s^2_{2/\{l+1\}(n-l-2)}}{s^2_{1/\{l+1\}} - b^2_{12/\{l\}} s^2_{2/\{l+1\}}} \right\}^{1/2}$$

с числом степеней свободы $\nu = n-l-2$, причем у $s^2_{2/\{l+1\}}$ множество $\{l+1\}$ образовано добавлением к множеству $\{l\}$ индекса 1, а у $s^2_{1/\{l+1\}}$ - индекса 2.

Для значимых частных и множественных коэффициентов детерминации можно получить более предпочтительные точечные оценки, чем выборочные коэффициенты:

$$\frac{(n-l-1)r^2_{12/\{l\}} - 1}{n-l-2} \quad \text{- оценка для } \rho^2_{12/\{l\}};$$

$$\frac{(r-1)r_M^2 - l}{n-l-1} \quad \text{- оценка для } \rho^2_M.$$

Оценку уравнения регрессии и его коэффициентов при достаточно больших k удобнее производить по методике регрессионного анализа. Следует отметить, что требования регрессионного анализа в корреляционном анализе выполнены. Можно также показать, что оценка линейного уравнения регрессии, проведенная по методике корреляционного анализа, совпадает с оценкой того же уравнения, полученной методом наименьших квадратов.

Анализ всевозможных коэффициентов корреляции (детерминации), как указывалось в начале главы, позволяет оценить природу зависимости между наблюдаемыми переменными. В результате можно отделить некоторые признаки, имеющие слабую взаимосвязь с остальными, или, наоборот, включить в систему дополнительные, ранее не учитываемые признаки. Вопрос о размерности пространства взаимозависимых величин изучается с разных точек зрения в регрессионном и факторном анализе. Вопрос о сильной взаимозависимости признаков-аргументов (мультиколлинеарности) рассматривается также в структурном анализе. Заметим, что оценка корреляционной матрицы может служить исходным материалом для перечисленных и других моделей многомерного статистического анализа.

3.5. Проверка гипотезы о нормальном распределении генеральной совокупности

Хотя общих методов проверки определяющего требования корреляционного анализа нет, *с помощью критериев согласия можно убедиться в том, что гипотеза о нормальном распределении генеральной совокупности не отвергается для частных распределений компонент $x_1, x_2, \ldots, x_k$.*

Для двумерных частных распределений следует построить поле корреляции (изображение точек выборки на координатной плоскости). Указанная гипотеза не отвергается, если "облако" точек имеет вытянутую форму, причем точки группируются около некоторой прямой.

Для проверки линейности взаимозависимости пары признаков полезно оценить разницу между квадратом корреляционного отношения и коэффициентом детерминации выборки исходя из того факта, что статистика

$$\frac{(\eta^2 - r^2)/(m-2)}{(1-\eta^2)/(n-m)}$$

имеет F-распределение Фишера с $\nu_1 = m-2$ и $\nu_2 = n-m$ степенями свободы, где η^2 - квадрат выборочного корреляционного отношения, m - число групп (интервалов) признака, по которому находится корреляционное отношение другого признака.

3.6. Задачи и упражнения

3.1. Доказать, что точечная оценка парного коэффициента корреляции ρ заключена в пределах от –1 до +1.

3.2. Доказать, что статистика при $\rho = 0$

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

имеет t-распределение с числом степеней свободы $\nu = n-2$.

3.3. Доказать, что статистика $t = \frac{b_{yx} - \beta_{yx}}{\sqrt{1-r^2}} \cdot \frac{s_x}{s_y}\sqrt{n-2}$ имеет t-распределение с $\nu = n-2$.

3.4. Найти вероятность попадания выборочного коэффициента корреляции r в заданный интервал $[-r_\alpha; r_\alpha]$ при условии $\rho=0$.

3.5. Доказать, что в корреляционном анализе при $\rho=0$ величины x и y независимы.

3.6. Доказать, что в двумерной корреляционной модели уравнения регрессии линейны и имеют вид

$$My/x = \mu_y + \beta_{yx}(x-\mu_x),$$

где $\beta_{yx} = \rho\frac{\sigma_y}{\sigma_x}$.

$$Mx/y = \mu_x + \beta_{xy}(y-\mu_y),$$

где $\beta_{xy} = \rho\frac{\sigma_x}{\sigma_y}$.

3.7. Доказать, что если в корреляционном анализе $\rho^2=1$, то x и y связаны линейной функциональной зависимостью.

3.8. Доказать, что в двумерной корреляционной модели условные дисперсии имеют вид

$$Dy/x = \sigma_y^2(1-\rho^2), \quad Dx/y = \sigma_x^2(1-\rho^2)$$

и характеризуют вариацию относительно линий регрессий.

3.9. Доказать, что формулы (3.10) для выборочных частных коэффициентов корреляции являются частным случаем (3.18).

3.10. Доказать, что формула для вычисления выборочного множественного коэффициента корреляции

$$r_1 = r_{1/23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}}$$

является частным случаем (3.19).

3.11. Доказать, что в трехмерной модели определитель корреляционной матрицы R больше или равен нулю, т.е. $|R| \geq 0$.

3.12. Доказать, что множественный коэффициент корреляции заключен между 0 и 1.

3.13. Доказать, что в корреляционном анализе, если $\rho_z = \rho_{z/xy} = 0$, то величины z и (x,y) независимы.

3.14. Доказать, что если $\rho_y = \rho_{y/xz} = 1$, то величины y и (x,z) связаны линейной функциональной зависимостью.

3.15. Доказать справедливость неравенств:

$$r_{z/xy}^2 \geq r_{zx}^2; \quad r_{z/xy}^2 \geq r_{zy}^2; \quad r_{z/xy}^2 \geq r_{zx/y}^2; \quad r_{z/xy}^2 \geq r_{zy/x}^2.$$

3.16. Доказать, что частный коэффициент корреляции меняется в пределах от -1 до $+1$.

3.17. По результатам выборки объемом $n=20$ единицам из двумерной нормальной генеральной совокупности получены выборочные характеристики: $b_{yx} = -0{,}50$; $s_x^2 = 31{,}16$; $s_y^2 = 9{,}21$. Проверить при $\alpha=0{,}05$ значимость генерального коэффициента корреляции.

3.18. Из двумерной нормальной генеральной совокупности взята выборка объемом $n=32$ единицы и найдены выборочные коэффициенты регрессии $b_{yx} = -0{,}50$ и $b_{xy} = -1{,}71$. Проверить значимость генерального коэффициента корреляции при $\alpha=0{,}01$.

3.19. На основании шестнадцати выборочных данных получены следующие характеристики: $s_x^2 = 10{,}5$ и $s_{x/y}^2 = 4{,}5$. Требуется при $\alpha=0{,}05$ проверить значимость генерального коэффициента корреляции.

3.20. По результатам трех выборок из совокупности (x,y) имеем: $n_1=20$; $r_1=0{,}80$; $n_2=30$; $r_2=0{,}85$ и $n_3=30$; $r_3=0{,}9$. Требуется при $\alpha=0{,}05$ проверить гипотезу однородности коэффициентов корреляции.

3.21. На основе ста выборочных данных об объеме и весе одного вида кондитерских изделий было выяснено, что при увеличении веса на 1г объем увеличивается в среднем на 0,3см³, увеличению объема на 1см³ соответствует увеличение веса в среднем на 2,7г. Необходимо:

а) при $\alpha=0{,}05$ проверить значимость генерального коэффициента корреляции;

б) с надежностью $\gamma=0{,}954$ найти доверительный интервал для генерального коэффициента корреляции между объемом и весом кондитерских изделий;

в) с надежностью $\gamma=0{,}95$ найти длину интервальной оценки генерального коэффициента регрессии объема на вес кондитерских изделий;

г) с надежностью $\gamma=0{,}9$ найти интервальную оценку коэффициента регрессии веса на объем кондитерских изделий.

3.22. На основании пятидесяти выборочных данных о выручке и массе реализованной сельскохозяйственной продукции было выяснено, что выборочная доля дисперсии выручки, вызванная случайной вариацией выручки без учета вариации массы, составляет 9%. Требуется:

а) при $\alpha=0{,}01$ проверить значимость генерального коэффициента корреляции;

б) с надежностью $\gamma=0{,}95$ найти интервальную оценку ρ;

в) при $\gamma=0{,}9$ определить интервальную оценку коэффициента регрессии выручки от массы реализованной продукции.

3.23. По данным ста выборочных наблюдений об объеме реализованной продукции некоторого вида и себестоимости единицы продукции было выяснено, что при увеличении объема на единицу его измерения себестоимость единицы продукции уменьшается в среднем на 0,4 единицы своего измерения, а при увеличении себестоимости на единицу объем реализованной продукции уменьшается в среднем на 1,6 единицы. С надежностью $\gamma=0{,}99$ найти:

а) интервальную оценку генерального коэффициента корреляции между объемом реализации и себестоимостью единицы продукции;

б) интервальную оценку коэффициента регрессии себестоимости единицы продукции от объема реализации.

3.24. На основании двадцати наблюдений было выяснено, что выборочная доля дисперсии случайной величины y, вызванная вариацией x, составляет 64%. Найти интервальную оценку коэффициента корреляции ρ с надежностью $\gamma=0{,}543$.

3.25. По данным обследования тридцати хозяйств для анализа зависимости между средней урожайностью (x) и средней себестоимостью (y) хлопка были получены следующие выборочные характеристики: $r=-0{,}56$; $S_x=5{,}3$ т/га; $s_y=2{,}5$ тыс.руб/т. С надежностью $\gamma=0{,}90$ найти нижнюю границу оценки генерального коэффициента регрессии себестоимости на урожайность.

3.26. По результатам шестнадцати выборочных наблюдений получены выборочные коэффициенты регрессии: $b_{yx} = -0{,}51$, $b_{xy} = -1{,}72$. С надежностью $\gamma=0{,}95$ найти доверительные интервалы для коэффициента корреляции ρ и коэффициента регрессии β_{yx}.

3.27. Выборочное среднее квадратическое отклонение и условное среднее квадратическое отклонение случайной величины y при фиксированной случайной величине x, полученные на основании выборки объемом 30 единиц, равны соответственно 6,25; 2,25. Проверить с $\alpha=0{,}05$ гипотезу о том, что генеральный коэффициент корреляции равен 0,95, если выборочный коэффициент корреляции положителен.

3.28. На основании пяти выборочных точек (2,5), (3,5), (1,3), (2,4), (3,4) из генеральной совокупности (x,y) проверить с α=0,02 гипотезу H_0:ρ=0,6.

3.29. По результатам пяти наблюдений (1,5), (2,4), (3,1), (4,1), (5,2) проверить с α=0,03 гипотезу H_0:ρ= –0,6.

3.30. Проверить с α=0,01 гипотезу H_0 :ρ= –0,7, если известны: b_{xy}= –0,3; s_x^2 =2,25; s_y^2 =6,25; n=40.

3.31. Для вычисления взаимозависимости между себестоимостью 1т песка (z), сменной добычей песка (y) и фондоотдачей (x) было обследовано восемь карьеров. В результате получены следующие данные:

X	30	20	40	35	45	25	50	30
Y, т	20	30	50	70	80	20	90	25
Z, тыс.руб.	20	25	20	15	10	30	10	20

Вычислить:

а) матрицу выборочных парных коэффициентов корреляции;

б) выборочный частный коэффициент корреляции между фондоотдачей и сменной добычей песка;

в) выборочный частный коэффициент корреляции между фондоотдачей и себестоимостью 1 т песка;

г) выборочный множественный коэффициент корреляции себестоимости 1 т песка на остальные показатели;

д) выборочный множественный коэффициент корреляции фондоотдачи на остальные показатели;

е) выборочный множественный коэффициент корреляции сменной добычи песка на остальные показатели.

3.32. На основании годовых отчетных данных пяти строительно-монтажных предприятий были получены значения следующих показателей:

№ предприятия	Объем выполненных работ, млрд.руб.	Численность рабочих, чел.	Фонд заработной платы, млрд.руб.
1	13	320	3,2
2	14	570	5,5
3	16	780	8,0
4	12	200	2,5
5	15	700	7,2

Найти:

а) матрицу выборочных парных коэффициентов корреляции;

б) выборочный частный коэффициент корреляции между объемом выполненных работ и численностью рабочих;

в) выборочный частный коэффициент корреляции между объемом выполненных работ и фондом заработной платы;

г) выборочный частный коэффициент корреляции между численностью рабочих и фондом заработной платы;

д) выборочный множественный коэффициент корреляции между объемом выполненных работ и остальными показателями;

е) выборочный множественный коэффициент корреляции между численностью рабочих и остальными показателями.

3.33. По результатам n=100 наблюдений найдены выборочные характеристики трехмерной генеральной совокупности:

$\bar{x}$ = 4; s_x=2; r_{xy}= –0,6;

$\bar{y}$ = 5; s_y=2; r_{xz}=0,8;

$\bar{z}$ = 7; s_z=3; r_{yz}= –0,6.

Требуется:

а) проверить значимость:

частных коэффициентов корреляции при $\rho_{xy/z}$, $\rho_{xz/y}$ и $\rho_{yz/x}$ при α=0,05;

множественного коэффициента корреляции $\rho_{x/yz}$ при α=0,01;

б) найти интервальную оценку для $\rho_{xy/z}$ с надежностью γ=0,925.

3.34. По результатам n=30 наблюдений найдено s_y^2 =4,1; r_{yz}=0,81; $s_{y/xz}^2$ =0,4. Требуется проверить при α=0,05 значимость частного коэффициента корреляции $\rho_{xy/z}$.

3.35. На основании n=20 наблюдений найдено s_z^2 =9,2; r_{yx}^2 =0,36 и $s_{z/xy}^2$ =1.

Требуется:

а) при α=0,05 проверить значимость частного коэффициента корреляции $\rho_{xy/z}$;

б) с надежностью γ=0,92 найти интервальную оценку $\rho_{xz/y}$.

3.36. По результатам n=100 наблюдений найдены частные коэффициенты регрессии $b_{xy/z}$ = –0,4 и $b_{yx/z}$= –0,9.

Требуется:

а) проверить α=0,05 значимость частного коэффициента корреляции $\rho_{xy/z}$;

б) с надежностью γ =0,95 найти интервальную оценку для $\rho_{yz/x}$.

3.37. На основании n=50 наблюдений найдены выборочные характеристики $b_{yz/x}$= –10, $b_{zy/x}$= –0,081. Проверить при α=0,01 значимость частного коэффициента корреляции $\rho_{yz/x}$ и с надежностью γ=0,95 найти интервальную оценку для $\rho_{yz/x}$.

3.38. По результатам n=50 наблюдений найдены выборочные характеристики s_y^2 =4,1, $r_{\hat{y}}^2$ =0,36, $r_{yz/x}^2$ =0,81. Определить значимость множественного коэффициента корреляции $\rho_{y/xz}$ при α=0,05.

3.39. На основании n=30 наблюдений найдено s_z^2 =4,1, r_{yz}^2 =0,36, $r_{xz/y}^2$ =0,49 . Проверить при α=0,01 значимость множественного коэффициента корреляции $\rho_{z/xy}$.

3.40. По данным задачи 3.33 найти интервальные оценки для:

а) частного коэффициента регрессии $\beta_{yx/z}$ с надежностью γ=0,95;

б) частного коэффициента регрессии $\beta_{xz/y}$ при γ=0,9.

Определить точечные оценки условного математического ожидания x при y=5 и z=7, а также условного математического ожидания y при x=4, z=7.

3.41. По отчетным данным n=14 машиностроительных предприятий (табл.3.6) провести корреляционный анализ взаимосвязи следующих показателей эффективности: x - производительность труда (млн руб/чел.); y - индекс снижения себестоимости продукции (%); z - рентабельность (%).

Таблица 3.6

№ предприятия	x	y	z	№ предприятия	x	y	z
1	9,4	62	11	8	6,6	77	19
2	9,9	53	9	9	5,5	32	7
3	9,1	56	23	10	9,4	200	14
4	5,6	30	10	11	5,7	91	8
5	6,7	18	9	12	5,2	82	18
6	4,3	14	5	13	10,0	76	17
7	7,4	90	10	14	6,7	37	13

3.42. По отчетным данным n=12 машиностроительных предприятий, представленных в табл.3.7, провести корреляционный анализ взаимосвязи следующих показателей эффективности: x - фондоотдача активной части на 1 руб. ОПФ (руб.); y - рентабельность (%); z - производительность труда (млн руб/чел.).

Таблица 3.7

№ предприятия	x	y	z	№ предприятия	x	y	z
1	2,0	13,2	9,4	7	0,7	6,6	5,5
2	1,2	12,9	6,7	8	1,3	19,1	6,6
3	2,6	17,2	10,0	9	1,1	9,9	7,4
4	1,8	17,5	5,2	10	0,6	5,4	4,3
5	1,4	8,0	5,7	11	0,9	9,1	6,6
6	2,3	14,2	9,4	12	1,0	9,7	5,5

3.43. По данным n=14 цементных заводов, представленных в табл.3.8, провести корреляционный анализ взаимосвязи следующих показателей эффективности: x - фондоотдачи, y - выработки натурального цемента на одного работающего (млн.т); и z - среднезаводской себестоимости 1т цемента.

Таблица 3.8

№ предприятия	x	y	z	№ предприятия	x	y	z
1	26,2	1,2	19,0	8	35,2	1,7	11,9
2	38,4	1,6	10,3	9	62,5	1,8	10,9
3	31,6	1,1	11,9	10	29,2	1,2	12,8
4	42,7	2,2	10,5	11	38,7	1,5	9,2
5	32,9	1,7	11,6	12	56,7	1,8	6,7
6	58,2	1,6	7,8	13	23,1	0,7	13,2
7	44,8	0,9	9,7	14	48,4	1,9	9,3

3.44. По данным годовой отчетности n=30 угольных шахт найдены следующие выборочные характеристики:

$\bar{x}_1 = 34,3;$ $s_1=12,9;$ $r_{12}=-0,427;$

$\bar{x}_2 = 0,02;$ $s_2=0,0053;$ $r_{13}=-0,750;$

$\bar{x}_3 = 568,8;$ $s_3=167,2;$ $r_{23}=0,102.$

Требуется провести корреляционный анализ, если x_1 - среднемесячная производительность труда рабочего по добыче угля (т); x_2 - фондоотдача; x_3 - трудоемкость работ по добыче (человеко-дни на 1000 т).

3.45. На основании данных годовых отчетов n=20 цементных заводов вычислены:

$\bar{x}_1 = 35,4;$ $s_1=12,1;$ $r_{12}=0,429;$

$\bar{x}_2 = 1417;$ $s_2=633;$ $r_{13}=-0,520;$

$\bar{x}_3 = 12,5;$ $s_3=3,2;$ $r_{23}=-0,706.$

Требуется провести корреляционный анализ, если x_1 - фондоотдача на 1 руб. ОПФ (кг); x_2 - выработка натурального цемента на одного работающего (т); x_3 - среднезаводская себестоимость 1 т цемента (тыс.руб.).

3.46. По данным годовых отчетов n=30 угольных шахт найдены следующие выборочные характеристики:

$\bar{x}_1 = 34,3;$ $s_1=12,9;$ $r_{12}=-0,229;$

$\bar{x}_2 = 59,5;$ $s_2=21,8;$ $r_{13}=-0,750;$

$\bar{x}_3 = 568,8;$ $s_3=167,2;$ $r_{23}=0,361.$

Провести корреляционный анализ, если x_1 - среднемесячная производительность труда рабочего по добыче угля (т); x_2 - фондоемкость (тыс.руб./т); x_3 - трудоемкость работ по добыче (человеко-дни на 1000 т).

3.7. Задания для самостоятельной работы на ЭВМ по корреляционному анализу

На основании данных приложений провести корреляционный анализ:

1. Определить оценки параметров шестимерного нормального закона распределения (векторы средних арифметических и среднеквадратического отклонения, матрица парных коэффициентов корреляции).
2. Получить оценку матрицы частных коэффициентов корреляции. Проверить значимость и найти интервальные оценки частных коэффициентов корреляции.
3. Найти оценки шести множественных коэффициентов корреляции (детерминации). Проверить их значимость, предварительно выбрав уровень α.
4. Построить уравнение регрессии, выбрав в качестве результативного показатель, которому соответствует наибольший множественный коэффициент корреляции.
5. Дать интерпретацию полученным результатам корреляционного анализа.

3.8. Ранговая корреляция

При изучении неколичественных признаков или количественных признаков с непрерывными и неизвестными законами распределения классические подходы корреляционного анализа либо не применены либо не эффективны. Для изучения тесноты связи в этих случаях применяются методы непараметрической статистики, среди которых наиболее распространены методы ранговой корреляции.

Рассмотрим вариационный ряд для признака x:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(j)}, \ldots, x_{(n)}.$$

Напомним, что рангом наблюдаемого значения x_j признака x называется номер этого наблюдения в вариационном ряду (т.е. $R(x_j)=j$) при условии, что неравенства - строгие. Если же в вариационном ряду встречаются одинаковые члены, то в качестве одинаковых (связных) рангов берется средняя арифметическая соответствующих номеров.

Например, суммы баллов, набранных студентами за выполнение двух контрольных заданий, были следующими:

5, 10, 8, 7, 9, 10, 5, 5.

Вариационный ряд имеет вид

5, 5, 5, 7, 8, 9, 10, 10.

Ранг трех студентов, попавших в начало ряда, равен (1+2+3)/3=2 или $R(5)=2$. Далее $R(7)=4$, $R(8)=5$, $R(9)=8$, $R(10)= (9+10)/2=9{,}5$. Очевидно, ранги могут быть не обязательно натуральными числами.

Для измерения связей между признаками, значения которых можно упорядочить по степени проявления ими анализируемых свойств, применяются коэффициенты ранговой корреляции.

Изучим сначала парную связь, т.е. связь между двумя признаками x и y.

Пусть имеется выборка объема n из непрерывно распределенной двумерной генеральной совокупности (x,y):

$$(x_1, y_1)\ldots(x_n, y_n).$$

Из непрерывности распределения следует, что множество выборок с некоторыми одинаковыми значениями по x или по y имеет меру нуль. Следовательно, можно считать, что в ранжировке значений x и y связные ранги отсутствуют. При изучении связи между x и y, можно предполагать, что выборка упорядочена по x, тогда такой выборке соответствует следующая матрица (подстановка)

$$\begin{pmatrix} 1 & 2 & \ldots & i & \ldots & n \\ R(1) & R(2) & \ldots & R(i) & \ldots & R(n) \end{pmatrix},$$

в которой первая строка состоит из рангов наблюдений x, а вторая - из рангов y.

Вместо изучения связи между x и y с помощью (количественных) значений x и y будем исследовать эту связь, используя соответствующие ранги. Очевидно, что жесткой (функциональной) положительной связи между x и y соответствует подстановка

$$\begin{pmatrix} 1 & 2 & \ldots & i & \ldots & n \\ 1 & 2 & \ldots & i & \ldots & n \end{pmatrix};$$

жесткой отрицательной связи - подстановка

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ n & n-2 & \ldots & 1 \end{pmatrix}.$$

Остальные $n-2$ подстановки получаются при той или иной степени связи, собственной стохастической (статистической).

Для измерения степени сходства между перестановками $(1,2,\ldots,n)$ и $(R(1),R(2),\ldots,R(n))$, степени связи между x и y назовем **инверсией (беспорядком) между элементами перестановки** $R(i)$ и $R(j)$ (второй строки подстановки), если $R(i)$ стоит левее $R(j)$ и больше $R(j)$. Если же при том же условии $R(i)$ меньше $R(j)$, то говорят, что элементы $R(i)$ и $R(j)$ **инверсии не образуют или образуют порядок.**

В качестве меры связи берут разность между суммами чисел порядков N и чисел беспорядков Q, образованных элементами второй строки подстановки.

Руководствуясь комбинаторными подсчетами, можно определить вероятности получения перестановок с заданной мерой связи. Так, например, для подстановок из четырех элементов рассмотрим следующую расчетную таблицу:

Число порядков N	Число инверсий Q	Мера сходства S_k	Подстановки	Вероятность
0	6	–6	4321	1/24
1	5	–4	3421, 4231, 4322	3/24
2	4	–2	3412, 4132, 4213, 2431, 3241	5/24
3	3	0	3214, 2413, 4123, 3142,1432, 2341	6/24
4	2	2	2143, 1423, 2314, 3124, 1342	5/24
5	1	4	2134, 1324, 1243	3/24
6	0	6	1234	1/24

Нетрудно заметить (и доказать для общего случая), что сумма числа порядков N и инверсий Q равна сумме номеров перестановки, т.е. $1+2+\ldots+n=n(n+1)/2$; распределение вероятностей симметрично относительно центра S_k, равного нулю; если сумма номеров четна, то все возможные значения меры S_k четны (если же $n(n+2)/2$ - нечетное число, то все S_k - нечетные числа). Из сказанного следует, что таблицы для решения задач проверки гипотез относительно меры сходства (или связи) можно давать для неотрицательных значений S_k (четных или нечетных).

Коэффициент ранговой корреляции Кендалла определяется путем нормирования случайной величины S_k, т.е. деления на $n(n-1)/2$:

$$r_k = \frac{2S_k}{n(n-1)} = 1 - \frac{4Q}{n(n-1)} = \frac{4N}{n(n-1)} - 1.$$

Он, очевидно, изменяется в пределах от –1 до +1, которые отвечают самым жестким значениям меры сходства - отрицательной и положительной связи между x и y. Нуль соответствует отсутствию связи между признаками в генеральной совокупности. Как обычный парный коэффициент корреляции $r_{xy}(\rho_{xy})$, коэффициент корреляции Кендалла не является абсолютным измерителем связи (если он равен нулю, то x и y могут оказаться зависимыми).

При больших объемах n выборки и независимости x и y можно использовать нормальный закон распределения r_k с параметрами $Mr_k=0$ и

$$Dr_k = \frac{2(2n+5)}{9n(n-1)}.$$

Другой **коэффициент ранговой корреляции, предложенный Спирменом,** использует меру сходства с учетом весов рангов:

$$r_C = 1 - \frac{6S_C}{n^3 - n}, \qquad S_C = \sum_{i=1}^{n} (R(i) - i)^2.$$

Этот коэффициент можно получить по формуле парного коэффициента корреляции r_{xy} для выборки $(i,R(i))$, $i=\overline{1,n}$.

При больших объемах n выборки и независимости x и y коэффициент ранговой корреляции Спирмена подчиняется нормальному закону распределения с параметрами $Mr_c=0$ и $Dr_c=\frac{1}{n-1}$.

В теории доказывается, что случайные величины r_c приблизительно в полтора раза больше r_k при больших n, если только их квадраты не слишком близки к единице. При изучении связей между числом признаков, измеряемых в порядковой шкале, число которых больше двух, применяют меру сходства (согласия) соответствующего числа ранжировок (перестановок). Мера является суммой квадратов отклонений сумм рангов наблюдений (объектов) от их общего среднего ранга

$$S_W=\sum_{i=1}^{n}(S_i-\bar{S})^2=\sum_{i=1}^{n}S_i^2-\frac{(\sum_{i=1}^{n}S_i)^2}{n},\quad \bar{S}=\frac{\sum_{i=1}^{n}S_i}{n},\quad S_i=\sum_{j=1}^{k}R_{ij},$$

где R_{ij} - ранг i-го наблюдения по j-му признаку, $i=\overline{1,n}, j=\overline{1,k}$.

В качестве показателя согласованности определяется **коэффициент конкордации Кендалла**, вычисляемый по формуле

$$W=\frac{12S_W}{k^2(n^3-n)}.$$

Этот коэффициент может изменяться от нуля (абсолютная несогласованность) до единицы (полное совпадение всех ранжировок), что легко проверяется.

Доказано, что при отсутствии связи между k признаками при больших значениях $n(n>7)$ случайные величины

$$k(n-1)W=\frac{12S_W}{kn(n+1)}$$

приближенно распределены как хи-квадрат с числом степеней свободы $\nu=n-1$.

При $k=2$ получается следующее соотношение между W и r_c:

$$W=\frac{1}{2}(r_c+1).$$

При малых значениях n и k существуют таблицы распределения S_k, S_c и S_W, пригодные для проверки гипотезы независимости признаков.

Приведем фрагменты этих таблиц:

при $n=10$, $\alpha_l=P(|S|\geq S_l)$, где l есть k или c

S_k	17	19	21	23	25	27	29	31
α_k	0,156	0,108	0,072	0,046	0,028	0,0166	0,0092	0,0046

S_c	248	258	268	278	288	298	308
α_c	0,144	0,096	0,060	0,034	0,0174	0,0072	0,0022

При $n=5$, $k=3$, $\alpha=P(|S|\geq S_w)$

S_w	50	56	60	62	66	74	78	82	86
α_w	0,163	0,096	0,063	0,056	0,038	0,015	0,0053	0,0028	0,0006

Пример

На основе следующих данных о стаже работы x и времени решения определенной контрольной задачи y десяти программистов вычислить коэффициенты ранговой корреляции Кендалла и Спирмена и проверить значимость этих коэффициентов на уровне значимости $\alpha=0,01$.

№ программистов	Стаж (в мес.) x	Время решения (в мин.) y
1	32	12
2	15	24
3	16	23
4	18	21
5	20	20
6	28	9
7	21	11
8	29	10
9	23	15
10	17	16

Решение

Расположив пары наблюдений (x, y) в порядке возрастания x, получаем следующую подстановку

$$\begin{pmatrix}1&2&3&4&5&6&7&8&9&10\\10&9&6&8&7&3&5&1&2&4\end{pmatrix}.$$

Подсчитаем меру сходства S_k, приписывая числу инверсий, образуемых элементом второй перестановки, знак минус:

–1	–1	–1	–1	–1	–1	–1	–1	–1	=	–9
	–1	–1	–1	–1	–1	–1	–1	–1	=	–8
		1	1	–1	–1	–1	–1	–1	=	–3
			–1	–1	–1	–1	–1	–1	=	–6
				–1	–1	–1	–1	–1	=	–5
					–1	–1	–1	+1	=	0
						–1	–1	–1	=	–3
							–1	+1	=	2
								1	=	1
								S_k	=	–31

Вычисляем коэффициент ранговой корреляции Кендалла:

$$r_k=\frac{2S_k}{n(n-1)}=\frac{2(-31)}{10\cdot 9}=-\frac{31}{45}=-0,6889.$$

Проверим гипотезу $H_0: \rho_k=0$ против альтернативы $H_1: \rho_k \neq 0$. При $n=10$ значению $|S_k|=31$ соответствует уровень значимости $\alpha=0{,}0046$. Этот уровень меньше, чем 0,01. Следовательно, гипотеза H_0 отвергается с вероятностью ошибки, меньшей 0,01, т.е. генеральный коэффициент ранговой корреляции ρ_k значимо отличается от нуля.

Вычислим коэффициент r_c по условиям предыдущего примера. Получим $S_c=(10-1)^2+(9-2)^2+(6-3)^2+(8-4)^2+(7-5)^2+(3-6)^2+(5-7)^2+(1-8)^2+$ $+(2-9)^2+(4-10)^2=306$,

$$r_c = 1-\frac{6\cdot 306}{3\cdot 330} = 1-\frac{306}{165} = 20{,}8545.$$

При $n=10$ и $S_c=308$ из фрагмента таблицы получаем, что коэффициент ранговой корреляции Спирмена значим на уровне значимости меньше, чем 0,0092.

Таким образом, связь между стажем программиста и временем, затраченным на решение задачи, можно считать доказанной (чем больше стаж, тем меньше в среднем затраты времени на решение задачи).

Пример

Трем судьям предстояло присвоить места пяти спортсменам. Каждый судья расставляет каждого спортсмена по местам от первого до пятого места в соответствии со степенью предпочтительности. Решения судей приведены в следующей таблице.

Спортсмен	Результат спортсмена	Место	Решение судьи
	1	2	3
А	2	1	3
Б	3	3	2
В	1	2	1
Г	5	4	4
Д	4	5	5

Требуется установить, имеется ли в приведенных результатах некоторый порядок предпочтения, общий для всех, и если имеется, то определить реальные места, занимаемые каждым спортсменом согласно действительной степени показанного мастерства.

Решение

Подсчитаем сумму квадратов отклонений S_w и коэффициент конкордации W.

Спортсмен	Сумма мест S_i	Квадраты
А	6	36
Б	8	64
В	4	16
Г	13	169
Д	14	196
Всего	45	481

$$S_w = \sum_{i=1}^{5} S_i^2 - \frac{(\sum S_i)^2}{n} = 481-(45^2)/5 = 76;$$

$$W = \frac{12S_w}{k^2(n^3-n)} = \frac{12\cdot 76}{3^2(5^3-5)} = 0{,}844.$$

Так как значению $S_w=76$ соответствует уровень значимости α, заключенный внутри интервала (0,0053; 0,015), гипотезу о несогласованности мнения судей следует отвергнуть с вероятностью ошибки, меньшей 0,015. Заключение об окончательной расстановке спортсменов по местам можно сделать, руководствуясь, например, суммами номеров S_i.

Так, в нашем примере спортсмен *А* должен занять второе место, спортсмена *Б* следует поставить на третье, спортсмена *В* - на первое место, затем идут спортсмены *Г* и *Д*.

При наличии неразличимых объектов по признакам (связных рангах) способы и формулы вычисления несколько меняются.

Для вычисления коэффициента ранговой корреляции Кендалла и других удобно рассматривать таблицу сопряженности признаков x и y, значения которых можно упорядочить (табл. 3.9).

Таблица 3.9

Таблица сопряженности

X \ Y	1	...	j	...	b	$n_{i\bullet}$
1	n_{11}	...	n_{1j}	...	n_{1b}	$n_{1\bullet}$
...	...	...	...	...	...	...
i	n_{i1}	...	n_{ij}	...	n_{ib}	$n_{i\bullet}$
...	...	...	...	...	...	...
a	n_{a1}	...	n_{aj}	...	n_{ab}	$n_{a\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	...	$n_{\bullet j}$	...	$n_{\bullet b}$	$n_{\bullet\bullet}$

В таблице n_{ij} - число (частота) объектов, наблюдений, имеющих i-й уровень (ранг) по признаку x и j-й уровень (ранг) по признаку y; $n_{i\bullet}=\sum_j n_{ij}$ - число объектов, имеющих i-й уровень по признаку x; $n_{\bullet j}=\sum_j n_{ij}$ - число объектов, имеющих j-й уровень по признаку y; $n_{\bullet\bullet}=\sum_i\sum_j n_{ij}=\sum_i n_{i\bullet}=$ $=\sum_j n_{\bullet j}=n$ - общее число объектов (объем выборки).

Вычисляются следующие величины

$$A=\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\Big(\sum\sum_{k>i,l>j} n_{kl}\Big), \qquad B=\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\Big(\sum\sum_{k>i,l<j} n_{kl}\Big),$$

$$T_1=\frac{1}{2}\sum_{j=1}^{a} n_{j\bullet}(n_{j\bullet}-1), \qquad A_{ij}=\sum_{k>i,l>j}\sum n_{kl}+\sum_{k<i,l<j}\sum n_{kl},$$

$$T_2=\frac{1}{2}\sum_{j=1}^{b} n_{\bullet j}(n_{\bullet j}-1), \qquad B_{ij}=\sum_{k>i,l<j}\sum n_{kl}+\sum_{k<i,l>j}\sum n_{kl}.$$

Коэффициент ранговой корреляции Кендалла вычисляется для квадратных таблиц сопряженности, т.е. при $a=b$ по следующей формуле:

$$r_k = (A-B)/\{[\frac{1}{2}n(n-1)-T_1][\frac{1}{2}n(n-1)-T_2]\}^{\frac{1}{2}}.$$

Коэффициент ранговой корреляции Спирмена вычисляется по формуле

$$r_c = \frac{12\sum_i\sum_j n_{ij}[\sum_{k<i} n_{k\bullet} + (n_{k\bullet}/2) - (n/2)][\sum_{l<j} n_{\bullet l} + (n_{\bullet l}/2) - (n/2)]}{\{[n^3 - n - \sum_i (n_{i\bullet}^3 - n_{i\bullet})][n^3 - n - \sum_j (n_{\bullet j}^3 - n_{\bullet j})]\}^{\frac{1}{2}}}.$$

При $n>10$ для проверки гипотезы $H_0: \rho_c=0$, против $H_1: \rho_c \neq 0$ на уровне значимости α можно использовать статистику r с числом степеней свободы $\nu=n-2$ (или статистику Стьюдента $t=\sqrt{\frac{1-r_c^2}{n-2}}$ с $\nu=n-2$).

Для прямоугольных таблиц сопряженности упорядоченных значений признаков при $a \neq b$ используется коэффициент связанности Стьюарта, определяемый по формуле

$$r_{cm} = 2min(a,b)(A-B)/(n^2(min(a,b)-1),$$

при этом для достаточно больших значений n

$$\sigma_{r_{ст}} = \frac{2\min(a,b)}{n^3(\min(a,b)-1)}[n^2\sum_i\sum_j n_{ij}(A_{ij}-B_{ij})^2 - 4n(A-B)^2]^{\frac{1}{2}}.$$

При наличии связанных рангов формула для вычисления коэффициента конкордации Кендалла модифицируется:

$$r_w = \frac{12S_w}{k^2(n^3-n) - k\sum_{j=1}^{k} T_j}; \qquad T_j = \sum_{l=1}^{m_j}(n_{lj}^3 - n_{lj}),$$

где m_j - число групп связных рангов у признака x_j, а n_{lj} - число объектов (рангов), образующих l-ю группу связных рангов (если неразличимые ранги отсутствуют, то $m_i=n$, $n_{lj}=1$ для $l=\overline{1,n}$ и, следовательно, $T_j=0$).

ГЛАВА 4

РЕГРЕССИОННЫЙ АНАЛИЗ

4.1. Основные понятия

После того как с помощью корреляционного анализа выявлено наличие статистически значимых связей между переменными и оценена степень их тесноты, обычно переходят к математическому описанию конкретного вида зависимостей с использованием регрессионного анализа. С этой целью подбирают класс функций, связывающий результативный показатель y и аргументы $x_1, x_2, ..., x_k$, отбирают наиболее информативные аргументы, вычисляют оценки неизвестных значений параметров уравнения связи и анализируют точность полученного уравнения.

Функция $f(x_1, x_2, ..., x_k)$, описывающая зависимость условного среднего значения результативного признака y от заданных значений аргументов, называется **функцией (уравнением) регрессии**.

Термин "регрессия" (лат. - "regression" - отступление, возврат к чему-либо) введен английским психологом и антропологом Ф.Гальтоном и связан только со спецификой одного из первых конкретных примеров, в котором это понятие было использовано.

Обрабатывая статистические данные в связи с вопросом о наследственности роста, Ф.Гальтон нашел, что если отцы отклоняются от среднего роста всех отцов на x дюймов, то их сыновья отклоняются от среднего роста всех сыновей меньше, чем на x дюймов. Выявленная тенденция была названа **"регрессией к среднему состоянию"**.

Термин "регрессия" широко используется в статистической литературе, хотя во многих случаях он недостаточно точно характеризует понятие статистической зависимости.

Для точного описания уравнения регрессии необходимо знать условный закон распределения результативного показателя y. В статистической практике такую информацию получить обычно не удается, поэтому ограничиваются поиском подходящих аппроксимаций для функции $f(x_1, x_2, ..., x_k)$, основанных на исходных статистических данных.

В рамках отдельных модельных допущений о типе распределения вектора показателей $(y, x_1, x_2, ..., x_k)$ может быть получен **общий вид уравне-**

ния регрессии $f(x)=M(y/x)$, где $x=(x_1,x_2,...,x_k)^T$. Например, в предложении, что исследуемая совокупность показателей подчиняется $(k+1)$ - мерному нормальному закону распределения с вектором математических ожиданий

$$M=\begin{pmatrix} My \\ Mx \end{pmatrix},$$

где $Mx=\begin{pmatrix} Mx_1 \\ . \\ . \\ . \\ Mx_k \end{pmatrix}$, $\mu_y = MY$

и ковариационной матрицей $\Sigma=\begin{pmatrix} \sigma_{yy} & \Sigma_{yx}^T \\ \Sigma_{yx} & \Sigma_{xx} \end{pmatrix}$,

где $\sigma_{yy}=\sigma_y^2=M(y-My)^2$;

$$\Sigma_{yx}=\begin{pmatrix} \sigma_{y1} \\ \sigma_{y2} \\ . \\ . \\ . \\ \sigma_{yk} \end{pmatrix}; \quad \Sigma_{xx}=\begin{pmatrix} \sigma_{11} & \sigma_2 & . & . & . & \sigma_{1k} \\ \sigma_{12} & \sigma_{22} & . & . & . & \sigma_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_{k1} & \sigma_{k2} & . & . & . & \sigma_{kk} \end{pmatrix};$$

$$\sigma_{ij}=M(x_i-Mx_i);(x_j-Mx_j); \quad \sigma_{jj}=\sigma_j^2=M(x_j-Mx_j)^2.$$

Из (1.34) следует, что уравнение регрессии (условное математическое ожидание) имеет вид:

$$M(y/x)=\mu_y+\Sigma_{yx}^T\cdot\Sigma_{xx}^{-1}(x-Mx).$$

Таким образом, если многомерная случайная величина $(y, x_1,x_2...,x_k)$ подчиняется $(\kappa+1)$-мерному нормальному закону распределения, то уравнение регрессии результативного показателя y по объясняющим переменным $x_1,x_2...,x_k$ имеет линейный по x вид.

Однако в статистической практике обычно приходится ограничиваться поиском подходящих аппроксимаций для неизвестной истинной функции регрессии $f(x)$, так как исследователь не располагает точным знанием условного закона распределения вероятностей анализируемого результатирующего показателя y при заданных значениях аргументов $x=x$.

Рассмотрим взаимоотношение между истиной $f(x)=M(y/x)$, модельной $\tilde{y}$ и оценкой $\hat{y}$ регрессии.

Пусть результативный показатель y связан с аргументом x соотношением:

$$y=2x^{1,5}+\varepsilon,$$

где ε - случайная величина, имеющая нормальный закон распределения, причем $M\varepsilon=0$ и $D\varepsilon=\sigma^2$.

Истинная функция регрессии в этом случае имеет вид:

$$f(x)=M(y/x)=2x^{1,5}.$$

Предположим, что точный вид истинного уравнения регрессии нам не известен, но мы располагаем девятью наблюдениями над двумерной случайной величиной, связанной соотношением $y_i=2x_i^{1,5}+\varepsilon_i$ и представленной на рис. 4.1.

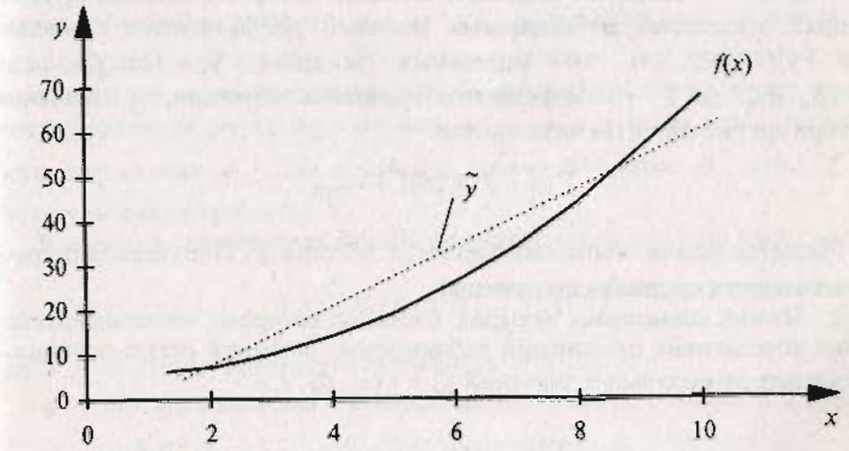


Рис. 4.1. Взаимное расположение истинной $f(x)$ и теоретической $\tilde{y}$ модели регрессии.

Расположение точек на рис. 4.1 позволяет ограничиться классом линейных зависимостей вида: $\tilde{y}=\beta_0+\beta_1 x$.

С помощью метода наименьших квадратов найдем оценку уравнения регрессии $\hat{y}=b_0+b_1x$.

Для сравнения на рис.4.1 приводятся графики истинной функции регрессии $f(x)=2x^{1,5}$, теоретической аппроксимирующей функции регрессии $\hat{y}=\beta_0+\beta_1x$. К последней сходится по вероятности оценка уравнения регрессии $\hat{y}$ при неограниченном увеличении объема выборки $(n\to\infty)$.

Поскольку мы ошиблись в выборе класса функции регрессии, что, к сожалению, достаточно часто встречается в практике статистических исследований, то наши статистические выводы и оценки не будут обладать свойством состоятельности, т.е., как бы мы ни увеличивали объем наблюдений, наша выборочная оценка $\hat{y}$ не будет сходиться к истинной функции регрессии $f(x)$.

Если бы мы правильно выбрали класс функций регрессии, то неточность в описании $f(x)$ с помощью $\hat{y}$ объяснялась бы только ограниченностью выборки и, следовательно, она могла бы быть сделана сколько угодно малой при $n \to \infty$.

С целью наилучшего восстановления по исходным статистическим данным условного значения результатирующего показателя $y(x)$ и неизвестной функции регрессии $f(x)=M(y/x)$ наиболее часто используют следующие **критерии адекватности** (функции потерь).

1. *Метод наименьших квадратов*, согласно которому минимизируется квадрат отклонения наблюдаемых значений результативного показателя $y_i (i=1,2,\ldots,n)$ от модельных значений $\tilde{y}_i = f(x_i,\beta)$, где $\beta=(\beta_0,\beta_1,\ldots,\beta_k)^t$ - коэффициенты уравнения регрессии, x_i - значения вектора аргументов в i-м наблюдении:

$$\sum_{i=1}^{n}(y_i - f(x_i,\beta))^2 \to \min_{\beta}.$$

Решается задача отыскания оценки $\hat{\beta}$ вектора β. Получаемая регрессия называется **среднеквадратической**.

2. *Метод наименьших модулей*, согласно которому минимизируется сумма абсолютных отклонений наблюдаемых значений результативного показателя от модульных значений $\tilde{y}_i = f(x_i,\beta)$, т. е.

$$\sum_{i=1}^{n}\left|y_i - f(x_i,\beta)\right| \to \min_{\beta}.$$

Получаемая регрессия называется **среднеабсолютной (медианной)**.

3. *Метод минимакса* сводится к минимизации максимума модуля отклонения наблюдаемого значения результативного показателя y_i от модельного значения $f(x_i,\beta)$, т. е.

$$\max_{1\le i\le n}\left|y_i - f(x_i,\beta)\right| \to \min_{\beta}.$$

Получаемая при этом регрессия называется **минимаксной**.

В практических приложениях часто встречаются задачи, в которых изучается случайная величина y, зависящая от некоторого множества переменных $x_1,x_2,\ldots,x_k$ и неизвестных параметров β_j (j=0,1,2,...,k). Будем рассматривать $(y, x_1,x_2,\ldots,x_k)$ как (k+1)-мерную генеральную совокупность, из которой взята случайная выборка объемов n, где $(y_i, x_{i1},x_{i2},\ldots,x_{ik})$ результат i-го наблюдения i=1,2,...,n. Требуется по результатам наблюдений оценить неизвестные параметры β_j (j=0,1,2,...,k). Описанная выше задача относится к задачам регрессионного анализа.

Регрессионным анализом называется метод статистического анализа зависимости случайной величины y от переменных x_j (j= 1,2,...,k), рассматриваемых в регрессионном анализе как неслучайные величины, независимо от истинного закона распределения x_j.

Обычно предполагается, что случайная величина y имеет нормальный закон распределения с условным математическим ожиданием $\tilde{y}$, являющимся функцией от аргументов x_j (j= 1,2,...,k) и постоянной, не зависящей от аргументов дисперсий σ^2, т.е. следует помнить, что требование нормальности закона распределения y необходимо лишь для проверки значимости уравнения регрессии и его параметров β_j, а также для интервального оценивания β_j. Для получения точечных оценок β_j (j=0,1,2,...,k) этого условия не требуется.

В общем виде **линейная модель** регрессионного анализа имеет вид:

$$y = \sum_{j=0}^{k}\beta_j\varphi_j(x_1,x_2,\ldots,x_k)+\varepsilon,$$

где φ_j - некоторая функция его переменных $x_1,x_2,\ldots,x_k$;

ε - случайная величина с нулевым математическим ожиданием и дисперсией σ^2.

Примечание.

В регрессионном анализе под **линейной моделью** подразумевают модель, линейно зависящую от неизвестных параметров β_j.

Собственно линейной будем называть модель, линейно зависящую как от параметров β_j, так и от переменных x_j.

В регрессионном методе вид уравнения регрессии выбирают исходя из анализа физической сущности изучаемого явления и результатов наблюдения.

Наиболее часто встречаются следующие **виды уравнений регрессии**:

- собственно линейное многомерное

$$\tilde{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots \beta_kx_k;$$

• полиномиальное

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \ldots \beta_k x_k^k;$$

• гиперболическое

$$\tilde{y} = \beta_0 + \beta_1 \frac{1}{x};$$

• степенное

$$\tilde{y} = \beta_0' \cdot x_1^{\beta_1} x_2^{\beta_2} \ldots x_k^{\beta_k}.$$

Путем логарифмирования степенные уравнения регрессии могут быть преобразованы в линейные уравнения относительно параметров β_j.

Логарифмируя, получим:

$$\lg \tilde{y} = \lg \beta_0' + \beta_1 \lg x_1 + \beta_2 \lg x_2 + \ldots + \beta_k \lg x_k.$$

Пусть $\lg x_j = u_j$ для $j=1,2,\ldots,\kappa$; $\lg \tilde{y} = \tilde{z}$ и $\lg \beta_0' = \beta_0$, тогда после подстановки будем иметь собственно линейные уравнения регрессии:

$$\tilde{z} = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \ldots + \beta_k u_k.$$

Путем подстановок $\frac{1}{x} = u$ и $x^j = u_j$ гиперболическое и полиномиальное уравнения могут быть преобразованы в собственно линейные, теория которых разработана наиболее полно.

Оценки неизвестных параметров уравнения регрессии находят обычно методом наименьших квадратов. Ниже мы остановимся более подробно на методе наименьших квадратов и свойствах оценок, найденных этим методом.

4.2. Простейшее линейное уравнение регрессии

4.2.1. Нахождение оценки уравнения

Пусть на основании анализа исследуемого явления предполагается, что в "среднем" y есть линейная функция от x, т.е. имеет место уравнение регрессии

$$\tilde{y} = M(y/x) = \beta_0 + \beta_1 x, \qquad (4.1)$$

где $M(y/x)$ - условное математическое ожидание случайной величины y при заданном x;

β_0 и β_1 - неизвестные параметры генеральной совокупности, которые подлежит оценить по результатам выборочных наблюдений.

Предположим, что для оценки параметров β_0 и β_1 из двухмерной генеральной совокупности (x, y) взята выборка объемом n, где (x_i, y_i) результат

i-го наблюдения (i–1,2,...,n). В этом случае модель регрессионного анализа имеет вид

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_1, \qquad (4.2)$$

где ε_i - независимые нормально распределенные случайные величины с нулевым математическим ожиданием и дисперсией σ^2, т.е. $M\varepsilon_i = 0$;

$D\varepsilon_i = \sigma^2$ для всех $i=1,2,\ldots,n$ и $M\varepsilon_i \varepsilon_j = 0$ при $i \neq j$.

Согласно методу наименьших квадратов в качестве оценок неизвестных параметров β_0 и β_1 следует брать такие значения выборочных характеристик b_0 и b_1, которые минимизируют сумму квадратов отклонений значений результативного признака y_i от условного математического ожидания $\tilde{y}_i$, т.е.

$$Q = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_i x_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2. \qquad (4.3)$$

Так как Q дифференцируема по β_0 и β_1, то для отыскания минимума функции (4.3) найдем частные производные по β_0 и β_1:

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum\limits_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1) \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum\limits_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \end{cases} \qquad (4.4)$$

Приравняв производные нулю и подставив в (4.4) вместо β_0 и β_1 их оценки b_0 и b_1, получим:

$$\begin{cases} \sum\limits_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \\ \sum\limits_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0 \end{cases} \quad \text{или} \quad \begin{cases} b_0 n + b_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i \end{cases} \qquad (4.5)$$

Данная система уравнений называется **системой нормальных уравнений.**

Решая систему (4.5) относительно b_0 и b_1, получим

$$b_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i - \frac{1}{n}\sum\limits_{i=1}^{n} x_1 \sum\limits_{i=1}^{n} y_i}{\sum\limits_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum\limits_{i=1}^{n} x_i)^2}; \quad b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - b_1 \bar{x}. \qquad (4.6)$$

Таким образом, имеем оценку уравнения регрессии:

$$\hat{y} = b_0 + b_1 x. \qquad (4.7)$$

Докажем, что в случае нормального закона распределения случайной величины y оценки метода наименьших квадратов и наибольшего правдоподобия совпадают.

Пусть из двумерной генеральной совокупности (x,y) взята независимая выборка (x_i, y_i), где i=1,2,...,n, объемом n.

Будем рассматривать y_i как независимые нормальные случайные величины с математическим ожиданием $\tilde{y} = M(y_i / x_i)$, являющимся функцией от x_i согласно (4.1), и постоянной дисперсией σ^2.

Тогда $p(y_i) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(y_i-\tilde{y}_i)^2}{2\sigma^2}}$, где $\tilde{y}_i = \beta_0 + \beta_1 x_i$, и функция правдоподобия примет вид

$$L = P(y_1, x_1; y_2, x_2; \ldots; y_n, x_n; \beta_0, \beta_1; \sigma^2) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(y_i-\tilde{y}_i)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} e^{-\frac{\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}{2\sigma^2}}.$$

Согласно **методу наибольшего правдоподобия** в качестве оценок параметров β_0, β_1 и σ^2 возьмем значения b_0, b_1 и s^2, максимизирующие L. При заданных $x_1, x_2, \ldots, x_n$ и постоянном σ^2 функция правдоподобия L достигнет максимума, когда показатель степени при e будет минимальным, т.е. при условии минимума функции $Q = \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 =$

$= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, что совпадает с условием (4.3) нахождения оценок b_0, b_1 по методу наименьших квадратов. Таким образом, оценки b_0, b_1 обладают свойствами оценок наибольшего правдоподобия.

Однако функция правдоподобия L зависит также и от параметра σ. Из условия $\frac{\partial L}{\partial \sigma} = 0$ найдем оценку s^2 наибольшего правдоподобия параметра σ^2:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2. \qquad (4.8)$$

Несмещенная оценка параметра σ^2 равна:

$$\hat{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2. \qquad (4.9)$$

Исследуем свойства b_0 и b_1.

4.2.2. Определение интервальной оценки для β_0

Будем рассматривать модель регрессионного анализа

$$y_i = \beta_0 + \beta_1 x_i' + \varepsilon_i \quad \text{или} \quad y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i \qquad (4.10)$$

где $x_i' = (x_i - \bar{x})$ - центрированные величины, удовлетворяющие условию

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Тогда оценки b_0 и b_1 метода наименьших квадратов согласно (4.6) равны:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i. \qquad (4.11)$$

Учитывая (4.10), получим

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}[\beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i] = n\beta_0 + \beta_1\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}\varepsilon_i.$$

Откуда $b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$.

Учитывая (4.11), получим

$$b_0 = \beta_0 + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i. \qquad (4.12)$$

Величина b_0 есть линейная функция нормальных случайных величин ε_i. Следовательно, она также имеет нормальный закон распределения с математическим ожиданием:

$$Mb_0 = M(\beta_0 + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i) = \beta_0 + \frac{1}{n}\sum_{i=1}^{n} M\varepsilon_i = \beta_0, \qquad (4.13)$$

так как по условию $M\varepsilon_i = 0$, и дисперсией:

$$Db_0 = D(\beta_0 + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i) = D(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i) = \frac{1}{n^2}\sum_{i=1}^{n} D\varepsilon_i = \frac{\sigma^2}{n}. \qquad (4.14)$$

Здесь учитывалось, что ε_i независимые случайные величины с дисперсией $D\varepsilon_i = \sigma^2$ для всех i=1,2,...,n.

Подставляя вместо σ^2 несмещенную оценку $\hat{s}^2$, получим оценку дисперсии $\hat{s}^2_{b_0}$, оценки b_0, $\hat{s}^2_{b_0} = \frac{\hat{s}^2}{n}$.

Таким образом, b_0 есть случайная величина, имеющая нормальный закон распределения $b_0 \in N(\beta_0; \frac{\sigma}{\sqrt{n}})$.

Отсюда следует, что величина

$$z = \frac{b_0 - \beta_0}{\sigma}\sqrt{n} \in N(0,1) \qquad (4.15)$$

имеет нормированный нормальный закон распределения.

С другой стороны, статистика

$$u^2 = \frac{ns^2}{\sigma^2} \in \chi^2(\nu = n-2) \qquad (4.16)$$

имеет χ^2-распределение с $\nu = n-2$ степенями свободы, так как уравнение регрессии определяется двумя параметрами b_0 и b_1, которые подлежат оцениванию.

Отсюда следует, что статистика

$$t = \frac{z}{u}\sqrt{\nu} = \frac{b_0 - \beta_0}{\sigma}\sqrt{n} \cdot \frac{\sigma}{\sqrt{n}\cdot s}\sqrt{n-2} = \frac{b_0 - \beta_0}{s}\sqrt{n-2} \in St(\nu = n-2)$$

имеет t-распределение с $\nu = n-2$ степенями свободы.

С помощью статистики t построим с доверительной вероятностью γ интервальную оценку для β_0 из условия:

$$P\left\{-t_\gamma \le \frac{b_0 - \beta_0}{s}\sqrt{n-2} \le t_\gamma\right\} = \gamma .$$

Откуда получим

$$\beta_0 \in \left[b_0 \pm t_\gamma \frac{s}{\sqrt{n-2}}\right] \qquad (4.17)$$

или, учитывая, что $\hat{s}_{b_0} = \frac{\hat{s}}{\sqrt{n}} = \frac{s}{\sqrt{n-2}}$, будем иметь: $\beta_0 \in \left[b_0 \pm t_\gamma \hat{s}_{b_0}\right]$, где t_γ определяется по таблице распределения Стьюдента (t-распределение) для уровней значимости $\alpha = 1-\gamma$ и числа степеней свободы $\nu = n-2$. Выражение (4.17) показывает, что β_0 принадлежит интервалу, границы которого заданы в квадратных скобках.

4.2.3. Определение интервальной оценки и проверка значимости β_1

С учетом (4.10) рассмотрим выражение

$$\sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n}(x_i - \bar{x})[\beta_0 + \beta_1(x_0 - \bar{x}) + \varepsilon_i] =$$

$$= \beta_0\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i =$$

$$= \beta_1\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i .$$

Учитывалось, что $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Решив уравнение относительно β_1, получим:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

а с учетом (4.11) будем иметь:

$$b_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (4.18)$$

Из (4.18) следует, что β_1 есть линейная функция независимых нормально распределенных случайных величин $\varepsilon_i \in N(0,\sigma)$, где i=1,2,...,n. Следовательно, она также имеет нормальный закон распределения.

Определим математическое ожидание и дисперсию b_1.

Учитывая, что математическое ожидание суммы равно сумме математических ожиданий; неслучайный множитель x_1 можно выносить за знак математического ожидания и $M\varepsilon_1 = 0$, получим:

$$Mb_1 = M\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})M\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 . \qquad (4.19)$$

Так как ε_1 есть независимые между собой случайные величины с дисперсией $D\varepsilon_1 = \sigma^2$, а дисперсия постоянной величины равна нулю, т.е. $D\beta_1 = 0$, то

$$Db_1 = D\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 D\varepsilon_i}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2},$$

откуда получим

$$Db_1 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (4.20)$$

Мы доказали, что в b_1 есть случайная величина, имеющая нормальный закон распределения:

$$b_1 \in N\left(\beta_1; \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2}}\right).$$

Отсюда следует, что

$$z = \frac{b_1 - \beta_1}{\sigma}\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2} \in N(0,1). \qquad (4.21)$$

Учитывая независимость случайных величин (4.16) и (4.21), получим статистику, имеющую t-распределение с $\nu = n-2$ степенями свободы:

$$t = \frac{z}{u}\sqrt{n-2} = \frac{(b_1 - \beta_1)\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2}}{s}\sqrt{\frac{n-2}{n}}; \qquad (4.22)$$

$$t = \frac{(b_1 - \beta_1)}{s}\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2} \in St(\nu = n-2),$$

где $\hat{s} = s\sqrt{\frac{n}{n-2}}$.

Интервальную оценку для β_1 с надежностью γ найдем из условия: $P(-t_\gamma \le t \le t_\gamma) = \gamma$.

После преобразования с учетом (4.22) получим

$$\beta_1 \in \left[b_1 \pm t_\gamma \frac{s}{\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2}}\right]$$

или

$$\beta_1 \in \left[b_1 \pm t_\gamma \hat{s}_{b_1}\right], \qquad (4.23)$$

где t_γ - находят по таблице t-распределения при $\alpha = 1-\gamma$ и $\nu = n-2$;

$\hat{s}^2_{b_1} = \frac{\hat{s}^2}{\sum_{i=1}^{n}(x_1 - \bar{x})^2}$ - несмещенная оценка дисперсии D_{b_1};

$\hat{s}_{b_1}$ - оценка среднего квадратического отклонения величины b_1.

Интервальную оценку с надежностью γ для σ^2 найдем с помощью статистики (4.16):

$$\sigma^2 \in \left[\frac{ns^2}{\chi_1^2}; \frac{ns^2}{\chi_2^2}\right], \qquad (4.24)$$

где χ_1^2, χ_2^2 находят по таблице χ^2-распределения для числа степеней свободы $\nu = n-2$ и вероятностей соответственно $\frac{1+\gamma}{2}$ и $\frac{1-\gamma}{2}$.

Установление значимости простейшего линейного уравнения регрессии $\tilde{y} = \beta_0 + \beta_1 x$ сводится к проверке при заданном α нулевой гипотезы о значимости коэффициента регрессии β_1, т.е. гипотезы $H_0: \beta_1 = 0$ при альтернативной гипотезе $H_1: \beta_1 \neq 0$.

С этой целью используется t-критерий и значение статистики критерия

$$t_1 = \frac{b_1}{\hat{s}_{b_1}} \qquad (4.25)$$

сравнивают с критическим значением $t_{кр}(\alpha; \nu = n-2)$, найденным по таблице t-распределения при заданном α и $\nu = n-2$.

Гипотеза $H_0: \beta_1 = 0$ отвергается с вероятностью ошибки α при выполнении неравенства $|t_1| > t_{кр}(\alpha, \nu = n-2)$ и уравнение регрессии считается значимым. В противном случае, т.е. если $|t_1| \le t_{кр}$, гипотеза $H_0: \beta_1 = 0$ не отвергается и уравнение регрессии считают незначимым и на этом регрессионный анализ заканчивается.

Для значимого уравнения регрессии представляет интерес построение интервальных оценок для коэффициента регрессии β_1, свободного члена β_0 и самого уравнения $\tilde{y}$.

4.2.4. Определение интервальной оценки для условного математического ожидания

Пусть имеем уравнение регрессии:

$$\tilde{y} = \beta_0 + \beta_1 (x - \bar{x}) \qquad (4.26)$$

и его оценку:

$$\hat{y} = b_0 + b_1 (x - \bar{x}),$$

где b_0, b_1 - оценки метода наименьших квадратов параметров уравнения β_0, β_1.

Величина $\hat{y}$ есть линейная функция двух случайных величин b_0 и b_1, имеющих нормальный закон распределения. Следовательно, $\hat{y}$ также имеет нормальный закон распределения. Определим параметры этого закона. Учитывая выражения (4.12) и (4.19), получим:

$$M\hat{y} = M\left[b_0 + b_1 (x - \bar{x})\right] = Mb_0 + (x - \bar{x})\, Mb_1 .$$

Откуда $M\hat{y} = \beta_0 + \beta_1 (x - \bar{x}) = \tilde{y}$.

Для определения дисперсии $D\hat{y}$ предварительно докажем независимость величин b_0 и b_1.

Так как величины b_0 и b_1 имеют нормальный закон распределения, то независимость этих величин следует из их некоррелированности. Следовательно, нам достаточно доказать, что

$$M(b_0 - \beta_1)(b_1 - \beta_1) = 0 .$$

Учитывая выражения (4.12) и (4.18) и, что x_i есть неслучайная величина, получим:

$$M(b_0 - \beta_0)(b_1 - \beta_1) = M\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right)\left(\frac{\sum_{j=1}^{n}(x_j - \bar{x})\varepsilon_j}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) =$$

$$= \frac{1}{n\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]} \sum_{i=1}^{n} M\varepsilon_i \sum_{j=1}^{n}(x_j - \bar{x})\varepsilon_j .$$

Так как ε_i (i= 1, 2, ...,n) по условию есть независимые случайные величины с $M\varepsilon_i = 0$, то $M\varepsilon_i \varepsilon_j = 0$ при $i \neq j$, где i, j=1,2,...,n. Следовательно,

$$M\varepsilon_i \sum_{j=1}^{n}(x_i - \bar{x})\varepsilon_j = (x_i - \bar{x})\, M\varepsilon_i^2 = (x_i - \bar{x})\sigma^2 ,$$

где $M\varepsilon_i^2 = \sigma^2$. Учитывая, что $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, после подстановки окончательно получим:

$$M(b_0 - \beta_0)(b_1 - \beta_1) = \frac{1}{n\sum_{i=1}^{n}(x_1 - \bar{x})^2}\sigma^2 \sum_{i=1}^{n}(x_i - \bar{x}) = 0 .$$

Этот результат получен для центрированных величин $(x_i - \bar{x})$, для которых выполняется условие $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. В этом случае b_0 и b_1 независимые случайные величины. Тогда согласно выражению (4.26) дисперсия величины $\hat{y}$ равна сумме дисперсий слагаемых, т.е.:

$$D\hat{y} = Db_0 + (x - \bar{x})^2\, Db_1 .$$

Подставив (4.14) и (4.20) в последнее выражение, получим:

$$D\hat{y} = \frac{\sigma^2}{n}\left(1 + n\frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Таким образом,

$$\hat{y} \in N\left(\tilde{y}; \frac{\sigma}{\sqrt{n}}\sqrt{1 + n\frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right). \qquad (4.27)$$

Тогда нормированный нормальный закон распределения имеет величину:

$$z = \frac{\hat{y} - \tilde{y}}{\frac{\sigma}{\sqrt{n}}\sqrt{1 + n\frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} b_0 \in N(0,1) . \qquad (4.28)$$

Учитывая (4.28) и (4.16), получим выборочную характеристику:

$$t = \frac{\hat{y} - \tilde{y}}{s\sqrt{1 + n\frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}\sqrt{n-2} \in St(\nu = n - 2), \qquad (4.29)$$

которая имеет распределение Стьюдента (t-распределение) с $\nu = n - 2$ степенями свободы.

Тогда с надежностью γ доверительный интервал для $\tilde{y}$ при заданном $x = x_0$ равен:

$$\tilde{y} \in \left[(b_0 + b_1 x) \pm t_\gamma \hat{s}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right], \qquad (4.30)$$

где t_γ определяется по таблице распределения Стьюдента для уровня значимости $\alpha = 1 - \gamma$ и числа степеней свободы $\nu = n - 2$.

Из (4.30) следует, что по мере удаления x_0 от среднего значения ($\bar{x}$) ширина доверительного интервала увеличивается, а точность оценки $\tilde{y}$ снижается. Доверительный интервал имеет наименьшую величину, когда $x_0 = \bar{x}$, т.е. наблюдаемое значение признака равно $\bar{x}$. Расположение доверительного интервала для $\tilde{y}$, найденного с надежностью γ, иллюстрирует рис. 4.2.

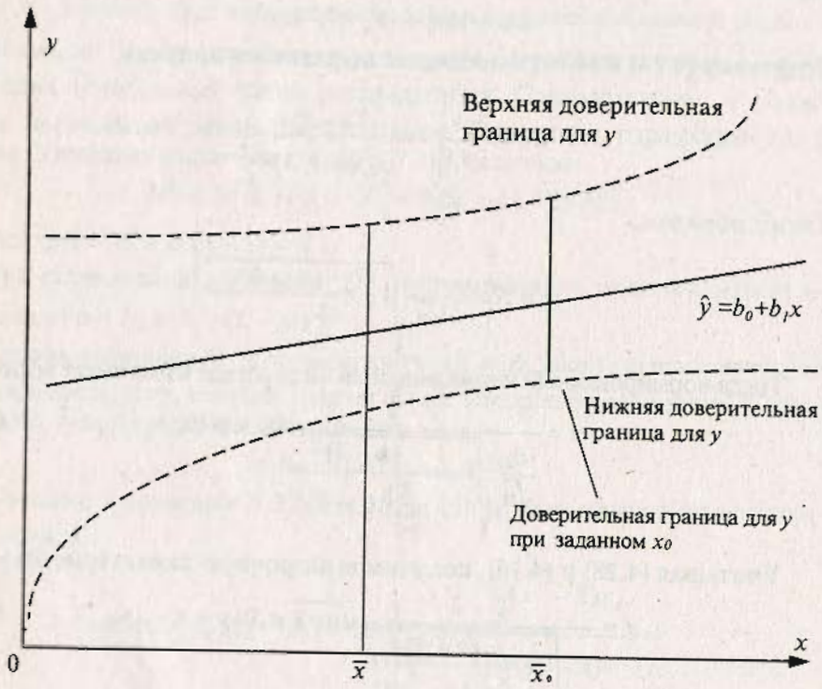


Рис. 4.2. Расположение доверительных границ в случае линейной регрессии

4.3. Множественное линейное уравнение регрессии

4.3.1. Нахождение и анализ свойств оценки уравнения регрессии

Рассмотрим общий случай линейной регрессии, когда $\tilde{y}$ есть функция k переменных $x_1, x_2, \ldots, x_k$:

$$\tilde{y} = M(Y / X_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k .$$

Пусть для оценки неизвестных параметров β_i (i=0, 1, 2, ..., k) уравнения регрессии взята случайная выборка объемом n из (k+1)-мерной случайной величины ($y, x_1, x_2, \ldots, x_k$) и $y_i, x_{i1}, \ldots, x_{ik}$ - результат i-го наблюдения, где i=1, 2, ..., n.

Модель множественной собственно линейной регрессии можно представить для i=1,2,...,n в виде

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i , \qquad (4.31)$$

где ε_i - взаимно некоррелированные случайные величины с нулевым математическим ожиданием и дисперсией σ^2, т.е. $M\varepsilon_i = 0$; $D\varepsilon_i = \sigma^2$.

В матричной форме линейная модель имеет вид:

$$Y = X\beta + \varepsilon , \qquad (4.32)$$

где $Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$ - вектор-столбец наблюдений размерности n;

$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$ - матрица размерности $n(k+1)$ известных коэффициентов ($n>k$);

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{pmatrix}$$ - вектор-столбец (размерности (κ+1)) неизвестных параметров, которые подлежат оцениванию;

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$ - вектор-столбец (размерности n) случайных "ошибок".

Причем

$$M\varepsilon = 0, \tag{4.33}$$

а ковариационная матрица $\sum(\varepsilon) = M\varepsilon\varepsilon^T$, где

$$M(\varepsilon\varepsilon^T) = \left[\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix} (\varepsilon_1 \varepsilon_2 \ldots \varepsilon_n) \right] = M \begin{pmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \ldots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \ldots & \varepsilon_2\varepsilon_n \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \ldots & \varepsilon_n^2 \end{pmatrix}.$$

Из условия следует, что для i=1,2,...,n $M\varepsilon_i^2 = \sigma^2$ и $M\varepsilon_{i_1}\varepsilon_{i_2} = 0$ при $i_1 \neq i_2$, тогда

$$\sum(\varepsilon) = M(\varepsilon\varepsilon^T) = \sigma^2 \mathrm{E}_n, \tag{4.34}$$

где $\mathrm{E}_n = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}$ - единичная матрица размерности (n x n).

Нам необходимо найти методом наименьших квадратов оценки неизвестных параметров β. Они определяются из условия минимизации скалярной суммы квадратов Q по компонентам вектора β.

$$Q = (Y - X\beta)^T (Y - X\beta). \tag{4.35}$$

В самом деле:

$$(Y - X\beta) = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} - \begin{pmatrix} \beta_0 + \sum_{j=1}^{k} x_{1j}\beta_j \\ ..... \\ ..... \\ ..... \\ \beta_0 + \sum_{j=1}^{k} x_{nj}\beta_j \end{pmatrix} = \begin{pmatrix} y - \beta_0 - \sum_{j=1}^{k} x_{2j}\beta_j \\ ..... \\ ..... \\ ..... \\ y_n - \beta_n - \sum_{j=1}^{k} x_{nj}\beta_j \end{pmatrix},$$

подставив найденное выражение в (4.35), получим скалярную сумму квадратов: $Q = \sum_{j=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2$.

Условием обращения Q в минимум является система уравнений $\frac{\partial Q}{\partial \beta_j} = 0$, где j=0,1,2,...,κ. Дифференцируя, получим

$$-2X^T(Y - X\beta) = 0,$$

где X^T - транспонированная матрица X.

Заменяя вектор b на оценку метода наименьших квадратов, получим

$$X^T Y = X^T X b,$$

где b - вектор-столбец (размерности (κ+1)).

Умножив обе части уравнения слева на матрицу $(X^T X)^{-1}$, обратную матрице $(X^T X)$, получим:

$$(X^T X)^{-1}(X^T Y) = \underbrace{(X^T X)^{-1}(X^T X)}_{\mathrm{E}_{(k+1)}} b,$$

откуда

$$b = \hat{\beta} = (X^T X)^{-1} X^T Y. \tag{4.36}$$

Докажем, что оценки наименьших квадратов являются несмещенными оценками.

Учитывая, что операции над матрицами ассоциативны и дистрибутивны, на основании (4.36) и (4.32) получим:

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) =$$
$$= (X^T X)^{-1}(X^T X)\beta + (X^T X)^{-1} X^T \varepsilon,$$

откуда

$$b = \beta + (X^T X)^{-1} X^T \varepsilon. \tag{4.37}$$

Из (4.37) следует, что вектор b подчиняется (κ+1)-мерному нормальному закону распределения как линейная функция нормально распределенного вектора ε. Закон распределения b зависит от вектора математических ожиданий Mb и ковариационной матрицы $\Sigma(b)$. Так как матрица X постоянна (по условию элементы матрицы X - неслучайные величины), то с учетом (4.33) получим:

$$Mb = \beta + (X^T X)^{-1} X^T M\varepsilon = \beta. \qquad (4.38)$$

Мы доказали несмещенность оценок метода наименьших квадратов.

В случае **линейной модели** b является несмещенной оценкой с минимальной дисперсией вектора β.

Ковариационная матрица вектора b равна:

$$\Sigma(b) = M\left[(b-\beta)(b-\beta)^T\right].$$

После подстановки в нее (4.37) получим:

$$\Sigma(b) = M\left\{\left[(X^T X)^{-1} X^T \varepsilon\right]\left[(X^T X)^{-1} X^T \varepsilon\right]^T\right\}.$$

Учитывая свойства матриц, будем иметь:

$$\left[(X^T X)^{-1} X^T \varepsilon\right]^T = \varepsilon^T X (X^T X)^{-1},$$

откуда

$$\Sigma(b) = (X^T X)^{-1} X^T M\varepsilon\varepsilon^T X (X^T X)^{-1}. \qquad (4.39)$$

Учитывая (4.34) и что $XE=EX$, получим:

$$\Sigma(b) = \sigma^2 (X^T X)^{-1} E (X^T X)(X^T X)^{-1}$$

и окончательно ковариационная матрица примет вид:

$$\Sigma(b) = \sigma^2 (X^T X)^{-1}. \qquad (4.40)$$

Рассмотрим статистический смысл элементов этой матрицы. Элементы главной диагонали матрицы (4.40) представляют собой дисперсии вектора оценок b. Вне главной диагонали ковариационной матрицы расположены значения коэффициентов ковариации. Например, на пересечении i-й строки и j-го столбца матрицы расположен коэффициент ковариации:

$$\operatorname{cov}(b_i b_j) = M(b_i - \beta_i)(b_j - \beta_j),$$

где $i, j = 0,1,2,...,\kappa$.

Таким образом, оценка b_j коэффициента линейной регрессии β_j (j=0,1, 2,...,κ) есть линейная функция (4.36) от y. Она имеет нормальный закон распределения с математическим ожиданием β_j и дисперсией согласно (4.40).

$$Db_{j-1} = \sigma^2 \left[(X^T X)^{-1}\right]_{jj}, \qquad (4.41)$$

где $\left[(X^T X)^{-1}\right]_{jj}$ - диагональный элемент обратной матрицы $(X^T X)^{-1}$, соответствующий j-й строке и j-му столбцу.

Найдем несмещенную оценку $\hat{s}^2$ для генеральной дисперсии σ^2, рассмотрим вектор остатков $e = Y - \hat{Y} = Y - Xb$.

Учитывая, что $b = (X^T X)^{-1} X^T Y$ и $\beta + \varepsilon$, получим:

$$e = X\beta + \varepsilon - X\left[(X^T X)^{-1} X^T (X\beta + \varepsilon)\right] =$$
$$= X\beta + \varepsilon - X(X^T X)^{-1}(X^T X)\beta - X(X^T X)^{-1} X^T \varepsilon =$$
$$= X\beta + \varepsilon - X\beta - X(X^T X)^{-1} X^T \varepsilon,$$

откуда

$$e = \varepsilon - X(X^T X)^{-1} X^T \varepsilon \qquad (4.42)$$

и

$$e^T = \varepsilon^T - \varepsilon^T X (X^T X)^{-1} X^T.$$

Тогда

$$M(Y - Xb)^T (Y - Xb) = Me^T e = M[\varepsilon^T - \varepsilon^T X (X^T X)^{-1} X^T] \times$$
$$\times [\varepsilon - X(X^T X)^{-1} X^T \varepsilon] = M\varepsilon^T \varepsilon - M\varepsilon^T X (X^T X)^{-1} X^T \varepsilon -$$
$$- M\varepsilon^T X (X^T X)^{-1} X^T \varepsilon + M\varepsilon^T X (X^T X)^{-1} (X^T X)(X^T X)^{-1} X^T \varepsilon.$$

Исходя из того, что произведение $(X^T X)$ на обратную $(X^T X)^{-1}$ дает единичную матрицу, последнее слагаемое преобразуется к виду $Me^T X (X^T X)^{-1} X^T$.

Таким образом,

$$Me^T e = M\varepsilon^T \varepsilon - M\varepsilon^T X (X^T X)^{-1} X^T \varepsilon. \qquad (4.43)$$

Учитывая, что скалярное произведение векторов $\varepsilon^T \varepsilon = \sum_{i=1}^{n} \varepsilon_i^2$ и $M\varepsilon_i^2 = \sigma^2$ для всех i=1,2,...,n, будем иметь

$$M\varepsilon^T \varepsilon = \sum_{i=1}^{n} M\varepsilon_i^2 = n\sigma^2. \qquad (4.44)$$

Легко убедиться, что матрица $B = X(X^T X)^{-1} X^T$ является симметричной, т.е. $B^T = B$. Для этого достаточно учесть, что $[(X^T X)^{-1}]^T = [(X^T X)^T]^{-1} = (X^T X)^{-1}$.

Из симметричности матрицы B следует:

$$M(\varepsilon^T B\varepsilon) = M\sum_{i\neq j}^{n} b_{ij}\varepsilon_i\varepsilon_j,$$

где $i,j=1,2,...,n$.

Принимая во внимание, что математическое ожидание суммы равно сумме математических ожиданий и что элементы матрицы B являются неслучайными величинами, получим:

$$M(\varepsilon^T B\varepsilon) = \sum_{i=1}^{n} b_{ii} M\varepsilon_i^2 + \sum_{i\neq j} b_{ij} M\varepsilon_i\varepsilon_j = \sigma^2 \sum_{i=1}^{n} b_{ii} = \sigma^2 trB.$$

Так как $M\varepsilon_i^2 = \sigma^2$ по условию, $M\varepsilon_i\varepsilon_j = 0$ из независимости величины ε_i и $\sum_{i=1}^{n} b_{ii} = trB$ - след матрицы B равняется сумме диагональных элементов квадратной матрицы.

Подставив вместо B его выражение, получим:

$$\begin{aligned} M(\varepsilon^T B\varepsilon) &= M\varepsilon^T X(X^T X)^{-1} X^T \varepsilon = \sigma^2 tr[X(X^T X)^{-1} X^T] = \\ &= \sigma^2 tr[(X^T X)(X^T X)^{-1}] = \sigma^2 tr\mathrm{E}_{k+1} = \sigma^2 (k+1). \end{aligned} \qquad (4.45)$$

Здесь учитывается свойство следа матрицы $tr(AC)=tr(CA)$; E_{k+1} - единичная матрица размерности $(k+1)$x$(k+1)$. Сумма ее диагональных элементов равна $(k+1)$, т.е. $tr\,\mathrm{E}_{k+1} = k+1$.

Подставив в (4.43) найденные выражения (4.44) и (4.45), получим:

$$M(Y - Xb)^T (Y - Xb) = (n - k - 1)\sigma^2 \qquad (4.46)$$

и несмещенная оценка остаточной дисперсии σ^2 равна:

$$\hat{s}^2 = \frac{1}{n-k-1}(Y - Xb)^T (Y - Xb). \qquad (4.47)$$

4.3.2. Проверка значимости уравнения регрессии

Для проверки значимости уравнения регрессии воспользуемся критерием дисперсионного анализа (F-критерием).

Предполагается, что вектор ε имеет нормальный закон распределения

$$\varepsilon \in N_n(\mu, \sigma^2 \cdot \mathrm{E}_n).$$

Предварительно докажем, что

$$Y^T Y = (Y - Xb)^T (Y - Xb) + (Xb)^T (Xb). \qquad (4.48)$$

Преобразуем правую часть и поставим в нее выражение (4.37):

$$(Y^T - b^T X^T)(Y - Xb) + (b^T X^T) Xb = Y^T Y - Y^T Xb - b^T X^T Y +$$

$$+ 2b^T X^T Xb = Y^T Y - Y^T Xb - (Y^T Xb)^T + 2b^T (X^T X)(X^T X)^{-1} X^T Y = Y^T Y.$$

Так как величина $Y^T Xb = b_0 \sum_{i=1}^{n} y_i + \sum_{j=1}^{k} b_j \sum_{i=1}^{n} y_i x_{ij}$ есть скалярная величина, то $(Y^T Xb) = (Y^T Xb)^T$.

Мы доказали справедливость тождества (4.48). Проанализируем теперь смысл слагаемых этого тождества.

$Q_{общ} = Y^T Y = \sum_{i=1}^{n} y_i^2$ есть сумма квадратов отклонений y_i от нуля.

Первое слагаемое (4.48)

$$Q_{ост} = (Y - Xb)^T (Y - Xb) = e^T e = \sum_{i=1}^{n} e_i^2$$

есть сумма квадратов отклонений результатов наблюдения от регрессии $\hat{Y} = Xb$.

Второе слагаемое

$$Q_R = (Xb)^T (Xb) = \sum_{i=1}^{n} \hat{y}^2$$

есть сумма квадратов отклонений от нуля, обусловленных регрессией.

Таким образом, $Q_{общ} = Q_R + Q_{ост}$.

Мы имеем разложение общей вариации на составляющие.

Так как ранг n квадратической формы $Q_{общ}$ равен сумме рангов $(k+1)$ и $(n-k-1)$ квадратических форм Q_R и $Q_{ост}$, стоящих в правой части (4.48), то согласно теореме Кохрана слагаемые правой части Q_R и $Q_{ост}$ независимы.

Согласно выражению (4.47) $M\frac{Q_{ост}}{n-k-1} = \sigma^2$.

Определим теперь математическое ожидание второго слагаемого Q_R:

$$MQ_R = M[(Xb)^T (Xb)] = M(b^T X^T Xb).$$

Подставив выражения (4.32) и (4.36) и учитывая, что $b^T = Y^T X(X^T X)^{-1}$, будем иметь:

$$\begin{aligned} MQ_R &= M[(\varepsilon^T + \beta^T X^T) X(X^T X)^{-1} X^T (X_\beta + \varepsilon)] = \\ &= M[\varepsilon^T X(X^T X)^{-1} X^T \varepsilon] + \beta^T X^T X\beta = (k+1)\sigma^2 + \beta^T X^T X\beta. \end{aligned}$$

Если предположить, что $\beta = 0$, где 0 - нулевой вектор, то $MQ_R = (k+1)\sigma^2$, откуда $M\frac{Q_R}{k+1} = \sigma^2$.

Мы показали, что при выполнении условия $\beta=0$, т.е. когда $\beta_0 = \beta_1 = \ldots = \beta_k = 0$, $\frac{Q_{ост}}{n-k-1}$ и $\frac{Q_R}{k+1}$ являются независимыми оценками одной и той же дисперсии σ^2.

В этой связи для проверки гипотезы $H_0: \beta = 0$ используется статистика:

$$F_{набл} = \frac{\frac{1}{k+1} Q_R}{\frac{1}{n-k-1} Q_{ост}}, \qquad (4.49)$$

которая при выполнении гипотезы H_0 имеет F-распределение с $(k+1)$ и $(n-k-1)$ степенями свободы.

Если уравнение регрессии незначимо, т.е. все коэффициенты уравнения регрессии для генеральной совокупности равны нулю, то на этом анализ уравнения регрессии заканчивается.

Если же нулевая гипотеза $H_0: \beta = 0$ отвергается, то представляет интерес проверка значимости отдельных коэффициентов регрессии и построение интервальных оценок для значимых коэффициентов.

Значимость коэффициентов регрессии можно проверить с помощью t-критерия, основанного на статистике:

$$t_j = \frac{b_j}{\hat{s}[(X^T X)^{-1}]_{jj}^{1/2}}, \qquad (4.50)$$

которая при выполнении гипотезы $H_0: \beta_j = 0$ имеет t-распределение с числом степеней свободы $n-k-1$.

4.3.3. Доверительные интервалы для параметров линейной модели

Пусть вектор ошибок ε имеет нормальное распределение. В этом случае вектор Y наблюдений (4.32) также имеет нормальное распределение и из некоррелированности следует их независимость.

Согласно выражению (4.36) оценка b_j (j=0,1,2,...,k) имеет нормальный закон распределения с математическим ожиданием β_j и дисперсией (4.41).

$$D(b_j) = \sigma^2[(X^T X)^{-1}]_{jj}.$$

Откуда $z = \frac{b_j - \beta_j}{\sigma[(X^T X)^{-1}]_{jj}^{1/2}}$ имеет нормальный закон распределения.

Согласно выражению (4.47) статистика

$$u^2 = \frac{ns^2}{\sigma^2} \qquad (4.51)$$

имеет распределение χ^2 с $n-k-1$ степенями свободы.

Напомним, что случайная величина

$$t = \frac{z}{u}\sqrt{\nu} \qquad (4.52)$$

имеет распределение Стьюдента (t-распределение) с ν степенями свободы, если z имеет нормированное нормальное распределение $z \in N(0,1)$, u^2 имеет распределение χ^2 с ν степенями свободы $u^2 \in \chi^2(\nu)$, причем z и u независимыми.

Подставив z и u в (4.52), получим выборочную характеристику:

$$t_j = \frac{b_j - \beta_j}{\hat{s}[(X^T X)^{-1}]_{jj}^{1/2}}, \qquad (4.53)$$

которая имеет распределение Стьюдента с $(n-k-1)$ степенями свободы. Используя (4.50), построим с надежностью γ интервальную оценку для

$$\beta_j \in \left\{ b_j \pm t_\gamma \hat{s}[(X^T X)^{-1}]_{jj} \right\}^{1/2}. \qquad (4.54)$$

Теперь определим интервальную оценку $\tilde{y}$ в точке, определяемой вектором X^0 начальных условий, размерности $(k+1)$:

$$x^o = (1, x_1^o, x_2^o, \ldots, x_k^o)^T.$$

Тогда несмещенная оценка с минимальной дисперсией значения признака $\tilde{y}$ в точке, определяемой условиями x^o, равна $\hat{y} = (x^o)^T b$.

В самом деле,

$$M\hat{y} = (x^o)^T M b = (x^o)^T \beta = \tilde{y},$$
$$D\hat{Y} = (x^o)^T \Sigma(b) x^o = \sigma^2 (x^o)^T (X^T X)^{-1} x^o,$$

где $\Sigma(b)$ - ковариационная матрица вектора оценок b.

Найдем доверительный интервал для $\tilde{y}$, учитывая (4.52) и используя статистику:

$$t = \frac{\hat{y} - \tilde{y}}{\hat{s}\sqrt{(x^o)^T (X^T X)^{-1} x^o}}, \qquad (4.55)$$

которая имеет распределение Стьюдента с $(n-k-1)$ степенями свободы. Из условия $P(|T| < t_\gamma) = \gamma$, раскрыв неравенство, стоящее в скобках, найдем

интервальную оценку для $\tilde{y}$ в точке, определяемой условиями x^o с надежностью γ:

$$\tilde{y} \in [\hat{y} \pm t_\gamma \hat{s} \sqrt{(x^o)^T (X^T X)^{-1} x^o}], \qquad (4.56)$$

где t_γ определяется по таблице t-распределения Стьюдента для уровня значимости $\alpha = 1 - \gamma$ и числа степеней свободы $(n-k-1)$.

Доверительная оценка для интервала предсказания $\hat{y}_{n+1}$ с надежностью γ определяется как:

$$\tilde{y}_{n+1} \in [(X^o)^T b \pm t_\gamma \hat{s} \sqrt{(x^o)^T (X^T X)^{-1} x^o + 1}]. \qquad (4.57)$$

4.3.4. Регрессионный анализ фондоотдачи

Пример.

На основе линейной регрессионной модели исследовать зависимость фондоотдачи в процентах на единицу ОПФ (y) от среднечасовой производительности вращающихся печей (x_1) и удельного веса активной части ОПФ (x_2).

В табл. 4.1. приводятся данные n=15 цементных заводов страны.

Таблица 4.1

Исходные данные

№ п/п	Фондоотдача, y	Среднечасовая производительность печей, x_1	Удельный вес активной части ОПФ (%), x_2
1	26	37	39
2	33	33	40
3	24	15	35
4	29	36	48
5	42	26	53
6	24	24	42
7	52	15	54
8	56	33	54
9	26	44	50
10	45	34	53
11	27	63	46
12	54	8	50
13	34	44	43
14	48	43	55
15	45	31	51

Решение

Для определения вектора оценок $b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$ согласно (4.6) найдем предварительную симметричную матрицу $X^T X$, которая имеет вид:

$$X^T X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} \\ \sum_i x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_i x_{i1} x_{i2} \\ \sum_i x_{i2} & \sum_i x_{i1} x_{i2} & \sum_i x_{i2}^2 \end{pmatrix}$$

и равна для нашего примера:

$$X^T X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 37 & 33 & \dots & 31 \\ 39 & 40 & \dots & 51 \end{pmatrix} \begin{pmatrix} 1 & 37 & 39 \\ 1 & 33 & 40 \\ \vdots & \vdots & \vdots \\ 1 & 31 & 51 \end{pmatrix} = \begin{pmatrix} 15 & 486 & 713 \\ 486 & 18416 & 23132 \\ 713 & 23132 & 34455 \end{pmatrix}.$$

Вектор $X^T Y$ имеет вид:

$$X^T Y = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \\ \sum_{i=1}^{n} x_{i2} y_i \end{pmatrix}$$

и для нашего примера равен:

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ 32 & 33 & 15 & \dots & 43 & 31 \\ 39 & 40 & 35 & \dots & 55 & 51 \end{pmatrix} \begin{pmatrix} 26 \\ 33 \\ 24 \\ \vdots \\ 48 \\ 45 \end{pmatrix} = \begin{pmatrix} 565 \\ 17513 \\ 27656 \end{pmatrix}.$$

Для получения обратной матрицы $(X^T X)^{-1}$ воспользуемся методом полного исключения **переменных Жордана-Гаусса**. С этой целью припишем справа к матрице $(X^T X)$ единичную матрицу и контрольный столбец, элементы которого равны сумме соответствующих строк матрицы $(X^T X)$ и единичной матрицы. Для рассматриваемого примера имеем:

$$\left(\begin{array}{ccc|ccc|c} 15 & 486 & 713 & 1 & 0 & 0 & 1215 \\ 486 & 18416 & 23132 & 0 & 1 & 0 & 42035 \\ 713 & 23132 & 34455 & 0 & 0 & 1 & 58301 \end{array} \right).$$

Применим элементарные преобразования к строкам матрицы $(X^T X)$ таким образом, чтобы на ее месте получить единичную матрицу. Тогда на месте единичной матрицы в (4.57) получим обратную матрицу $(X^T X)^{-1}$. Первоначально умножим первую строку выражения (4.57), включая и контрольный столбец, на $\left(-\frac{486}{15}\right)$ и прибавим полученные значения к соответ-

ствующим элементам второй строки. Затем умножим вновь первую строку (4.57) на $\left(-\frac{713}{15}\right)$ и прибавим полученные значения к элементам третьей строки. В результате этих действий получим нуль в качестве первого элемента второй и третьей строк. В самом деле:

$$15\left(-\frac{486}{15}\right)+486=0 \quad \text{и} \quad 15\left(-\frac{713}{15}\right)+713=0.$$

Второй и третий элементы второй строки соответственно равны:

$$\left(-\frac{486}{15}\right)\cdot 486+18416=2669{,}6; \quad \left(-\frac{486}{15}\right)\cdot 713+23132=30{,}8.$$

Первый элемент второй строки на месте единичной матрицы равен:

$$\left(-\frac{486}{15}\right)\cdot 1+0=-32{,}4.$$

Второй элемент контрольного столбца имеет вид

$$\left(-\frac{486}{15}\right)\cdot 1215+42035=2669{,}0.$$

Таким образом, окончательно будем иметь:

$$\begin{pmatrix} 15 & 486 & 713 \\ 0 & 2669{,}6 & 30{,}8 \\ 0 & 30{,}801 & 563{,}734 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ -32{,}4 & 1 & 0 \\ -47{,}533333 & 0 & 1 \end{pmatrix}=\begin{vmatrix} 1215 \\ 2669{,}0 \\ 548{,}00167 \end{vmatrix}. \qquad (4.58)$$

Если вычисления выполнены правильно, то сумма элементов строки двух матриц будет равна соответствующему элементу контрольного столбца. В самом деле:

$$15+486+713+1+0+0=1215;$$
$$0+2669{,}6+30{,}8-32{,}4+1+0=2669{,}0;$$
$$0+30{,}801+563{,}734-47{,}533333+0+1=548{,}00167.$$

Умножим теперь все элементы второй строки (4.58) на $\left(-\frac{486}{2669{,}6}\right)$ и прибавим к соответствующим элементам первой строки. Затем умножим элементы второй же строки на $\left(-\frac{30{,}801}{2669{,}6}\right)$ и прибавим к элементам третьей строки. После выполнения этих действий получим:

$$\begin{pmatrix} 15 & 0 & 707{,}39287 \\ 0 & 2669{,}6 & 30{,}8 \\ 0 & 0 & 563{,}37865 \end{pmatrix}\begin{pmatrix} 6{,}8984102 & -0{,}1820497 & 0 \\ -32{,}4 & 1 & 0 \\ -47{,}159515 & -0{,}0115376 & 1 \end{pmatrix}=\begin{vmatrix} 729{,}1094 \\ 2669{,}0 \\ 517{,}20782 \end{vmatrix}. \qquad (4.59)$$

Вновь проверим правильность вычислений, для чего просуммируем элементы строк матриц и сравним полученный результат с элементом контрольного столбца.

$$15+707{,}39287+6{,}8984102-0{,}1820497=729{,}10924\approx 729{,}10940;$$
$$2669{,}6+30{,}8-32{,}4+1=2669{,}0;$$
$$563{,}37865-47{,}159515-0{,}0115376+1=517{,}20761\approx 517{,}20782.$$

Невязки на уровне двух последних значащих цифр связаны конечной точностью наших вычислений.

Умножим теперь все элементы третьей строки (4.59) на $\left(-\frac{707{,}39287}{563{,}37865}\right)$ и прибавим к соответствующим элементам первой строки, а затем умножим на $\left(-\frac{30{,}8}{563{,}37865}\right)$ и прибавим к элементам второй строки. После выполнения этих действий получим:

$$\begin{pmatrix} 15 & 0 & 0 \\ 0 & 2669{,}6 & 0 \\ 0 & 0 & 563{,}378 \end{pmatrix}\begin{pmatrix} 66{,}113118 & -0{,}1675628 & -1{,}2556259 \\ -29{,}821785 & 1{,}0006307 & -0{,}0546701 \\ -47{,}159515 & -0{,}0115376 & 1 \end{pmatrix}=\begin{vmatrix} 79{,}68987 \\ 2640{,}7242 \\ 517{,}20782 \end{vmatrix}. \qquad (4.60)$$

Проверка показала правильность наших вычислений. Разделим в выражении (4.30) элементы первой, второй и третьей строк соответственно на 15, 2669,6 и 563,37865, тогда будем иметь:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 4{,}4075412 & -0{,}0111708 & -0{,}0837088 \\ -0{,}0111708 & 0{,}0003748 & -0{,}0000204 \\ -0{,}0837088 & -0{,}0000204 & 0{,}001775 \end{pmatrix}=\begin{vmatrix} 5{,}3126580 \\ 0{,}9891834 \\ 0{,}9180465 \end{vmatrix}.$$

Проверим правильность вычислений:

$$1+4{,}4075412-0{,}0111708-0{,}0837088=5{,}3126616;$$
$$1-0{,}0111708+0{,}0003748-0{,}0000204=0{,}9891836;$$
$$1-0{,}0837088-0{,}0000204+0{,}001775=0{,}9180458.$$

Полученные значения соответствуют контрольному столбцу в пределах точности проводимых вычислений.

Таким образом, мы получили обратную матрицу:

$$(X^TX)^{-1}=\begin{pmatrix} 4{,}4075412 & -0{,}0111708 & -0{,}0837088 \\ -0{,}0111708 & 0{,}0003748 & -0{,}0000204 \\ -0{,}0837088 & -0{,}0000204 & 0{,}001775 \end{pmatrix}.$$

Подставив найденный вектор X^TY и матрицу $(X^TX)^{-1}$ в выражение (4.36), найдем вектор оценок:

$$b=(X^TX)^{-1}X^TY=\begin{pmatrix} 4{,}4075412 & -0{,}0111708 & -0{,}0837088 \\ -0{,}0111708 & 0{,}0003748 & -0{,}0000204 \\ -0{,}0837088 & -0{,}0000204 & 0{,}001775 \end{pmatrix}\begin{pmatrix} 565 \\ 17513 \\ 27656 \end{pmatrix}=$$

$$=\begin{pmatrix} 4{,}4075412\cdot 565-0{,}0111708\cdot 17513-0{,}0837088\cdot 27656 \\ -0{,}0111708\cdot 565+0{,}0003748\cdot 17513-0{,}0000204\cdot 27656 \\ -0{,}0837088\cdot 565-0{,}0000204\cdot 17513+0{,}001775\cdot 27656 \end{pmatrix}=\begin{pmatrix} -20{,}4102 \\ -0{,}3118120 \\ 1{,}505803 \end{pmatrix}.$$

Таким образом,

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} -20{,}4102 \\ -0{,}3118120 \\ 1{,}505803 \end{pmatrix}$$

и оценка уравнения регрессии имеет вид:

$$\hat{Y} = -20{,}4102 - 0{,}3118120\,X_1 + 1{,}505803\,X_2 .$$

Для проверки значимости уравнения регрессии согласно выражению (4.49) нужно найти $Q_{ост} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ и $Q_R = \sum_{i=1}^{n}\hat{y}_i^{\,2}$.

Составим вспомогательную таблицу (табл.4.2).

Таблица 4.2

Вспомогательная таблица

№ п/п (i)	y_i	x_{1i}	x_{2i}	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	26	37	39	6,779073	0,6069547
2	33	33	40	29,532124	12,026163
3	24	15	35	27,615725	13,073467
4	29	36	48	40,643112	135,56205
5	42	26	53	51,290247	86,308689
6	24	24	42	35,350038	128,82336
7	52	15	54	56,225982	17,858923
8	56	33	54	50,613366	29,015825
9	26	44	50	41,160222	229,83233
10	45	34	53	48,795751	14,407725
11	27	63	46	29,212582	4,8955191
12	54	8	50	52,385454	2,6067587
13	34	44	43	30,619601	11,427097
14	48	43	55	49,001049	1,0020991
15	45	31	51	46,719581	2,9569588
Итого	565	486	713		690,40385

Из табл. 4.2 следует, что $Q_{ост}$ =690,40385. Тогда несмещенная оценка остаточной дисперсии σ^2 равна:

$$\hat{s}^2 = \frac{1}{n-3}Q_{ост} = \frac{1}{12}690{,}40385 = 57{,}533654 \quad \text{и} \quad \hat{s} = 7{,}5850941 .$$

По данным табл. 4.2 найдем: $Q_R = \sum_{i=1}^{n}\hat{y}_i^{\,2} = 26801{,}319$.

Проверим на уровне значимости α=0,05 значимость уравнения регрессии, т.е. гипотезу $H_0{:}\beta = 0$. Согласно выражению (4.49)

$$F_{набл} = \frac{\frac{1}{k+1}Q_R}{\frac{1}{n-k-1}Q_{ост}} = \frac{\frac{1}{3}26801{,}319}{\frac{1}{12}690{,}40385} = 155{,}27907 .$$

По таблице F-распределения для α=0,05 и чисел степеней свободы $\nu_1 = 3$ и $\nu_2 = 12$ найдем критическое значение $F_{кр}$ (0,05;3;12). Так как $F_{набл} > F_{кр}$, гипотеза $H_0{:}\beta = 0$ отвергается, т.е. хотя бы один элемент вектора $\beta = (\beta_0, \beta_1, \beta_2)^T$ не равен нулю.

Перед проверкой значимости отдельных коэффициентов регрессии найдем оценку ковариационной матрицы (4.40) вектора b. Для этого достаточно элементы обратной матрицы $(X^T X)^{-1}$ умножить на $\hat{s}^2 = 58{,}533654$. Тогда будем иметь:

$$\hat{V}(b) = \hat{s}^2 (X^T X)^{-1} = \begin{pmatrix} 253{,}582 & -0{,}643 & -4{,}816 \\ 0{,}643 & 0{,}0216 & -0{,}0012 \\ -4{,}816 & -0{,}0012 & 0{,}1021 \end{pmatrix}.$$

Из статистического смысла ковариационной матрицы следует, что оценки дисперсии коэффициентов уравнения регрессии b_0, b_1, b_2 соответственно равны:

$$\hat{s}^2_{b_0} = 253{,}582 , \quad \hat{s}^2_{b_1} = 0{,}0216 , \quad \hat{s}^2_{b_2} = 0{,}1021 .$$

Проверим значимость коэффициента β_1, т.е. гипотезу $H_0{:}\beta_1 = 0$.

Согласно (4.50)

$$t_{b_1} = \frac{b_1}{\hat{s}_{b_1}} = \frac{(-0{,}311812)}{(0{,}0216)^{1/2}} = -2{,}122 .$$

По таблице t-распределения для значений $\alpha = 0{,}05$ и $\nu_2 = 12$; $t_{кр}$ =2,179. Так как $|t_{b_1}| < t_{кр}$, гипотеза о том, что $\beta_1 = 0$, не отвергается, т.е. β_1 незначим.

Проверим теперь гипотезу $H_0{:}\beta_2 = 0$.

$$t_{b_2} = \frac{b_2}{\hat{s}_{b_2}} = \frac{1{,}505803}{(0{,}1021)^{1/2}} = 4{,}713 .$$

Так как $t_{b_2} > t_{кр} = 2{,}18$, гипотеза $H_0{:}\beta_2 = 0$ отвергается, т.е. β_2 не равен нулю ($\beta_2 \neq 0$).

Перейдем к алгоритму пошагового регрессионного анализа и исключим из рассмотрения переменную x_1, имеющую незначимый коэффициент β_1 уравнения регрессии. Уравнение регрессии будем искать в виде $\tilde{y} = \beta_0 + \beta_1 x_2$. Исходные данные для оценки коэффициентов β_0 и β_1 представлены в табл. 4.3.

Таблица 4.3

Исходные данные

№ п/п (i)	y_i	x_{2i}	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	26	39	25,46791	0,2831197
2	33	40	26,8840	37,405456
3	24	35	19,80355	17,610192
4	29	48	38,21272	84,874209

Продолжение

№ п/п (i)	y_i	x_{2i}	$\hat{y}_i$	$(y_i-\hat{y}_i)^2$
5	42	53	45,29317	10,844968
6	24	42	29,71618	32,674713
7	52	54	46,70926	27,991929
8	56	54	46,70926	86,317849
9	26	50	41,0449	226,34901
10	45	53	45,29317	0,0859486
11	27	46	35,38054	70,23345
12	54	50	41,0449	167,83461
13	34	43	31,13227	8,2238758
14	48	55	43,12535	0,157126
15	45	51	42,46099	6,4465717
Итого	565	713		777,19156

Тогда матрица X будет иметь вид:

$$X^T=\begin{pmatrix}1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\39&40&35&48&53&42&54&54&50&53&46&50&48&55&51\end{pmatrix}$$

и

$$(X^TX)=\begin{pmatrix}1&1&\dots&1\\39&40&\dots&51\end{pmatrix}\begin{pmatrix}1&39\\1&40\\\vdots&\vdots\\1&51\end{pmatrix}=\begin{pmatrix}15&713\\713&34455\end{pmatrix}.$$

Обратную матрицу $(X^TX)^{-1}$ вычислим по формуле:

$$(X^TX)^{-1}=\frac{1}{|X^TX|}\begin{pmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{pmatrix},$$

где A_{ij} - алгебраическое дополнение к элементу a_{ij} матрицы (X^TX).

Определитель

$$|X^TX|=15\cdot34455-713\cdot713=8456{,}0,$$

а обратная матрица:

$$(X^TX)^{-1}=\frac{1}{8456}\begin{pmatrix}34455&-713\\-713&15\end{pmatrix}=\begin{pmatrix}4{,}0746515&-0{,}0843188\\-0{,}0843188&0{,}0017738\end{pmatrix}.$$

Найдем вектор $X^TY=\begin{pmatrix}1&1&\dots&1\\39&40&\dots&51\end{pmatrix}\cdot\begin{pmatrix}26\\33\\\vdots\\45\end{pmatrix}=\begin{pmatrix}565\\27656\end{pmatrix}$.

Тогда вектор оценок (4.36):

$$b=(X^TX)^{-1}X^TX=\begin{pmatrix}4{,}0746512&-0{,}0843188\\-0{,}0843188&0{,}0017738\end{pmatrix}\begin{pmatrix}565\\27656\end{pmatrix}=\begin{pmatrix}-29{,}7596\\1{,}416090\end{pmatrix}.$$

Оценка уравнения регрессии имеет вид: $\hat{Y}=-29{,}7596+1{,}416090\,x_2$.

По данным табл. 4.3 найдем несмещенную оценку остаточной дисперсии:

$$\hat{s}^2=\frac{1}{n-k-1}Q_{ост}=\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{1}{13}\cdot777{,}19156=59{,}784,$$

где $k=1$ и $\hat{s}=7{,}732$.

Найдем оценку ковариационной матрицы вектора b.

$$\hat{S}(b)=\hat{s}^2(X^TX)^{-1}=\begin{pmatrix}243{,}597&-5{,}0409\\-5{,}0409&0{,}1060\end{pmatrix}.$$

Отсюда $\hat{s}^2_{b_1}=0{,}106$.

Для проверки значимости коэффициента β_1, т.е. гипотезы $H_0:\beta_1=0$, найдем

$$t_{набл}=\frac{b_1}{\hat{s}_{b_1}}=\frac{(1{,}41609)}{(0{,}106)^{1/2}}=4{,}349.$$

Определим критическое значение для $\alpha=0{,}05$; $\nu=n-2=13$ по таблице $t_{кр}=2{,}16$. Так как $t_{набл}>t_{кр}$, нулевая гипотеза отвергается ($\beta_1\neq0$). Таким образом, окончательно оценка регрессии со значимыми коэффициентами имеет вид $\hat{y}=-29{,}7596+1{,}41609x_2$.

Коэффициент регрессии при x_2 показывает, что при росте удельного веса активной части ОПФ на единицу (%) фондоотдача в среднем увеличивается на 1,41609 единиц.

Найдем теперь с доверительной вероятностью $\gamma=0{,}96$ интервальную оценку для коэффициента регрессии β_1. Согласно (4.54)

$$\beta_1\in\left[1{,}41609\pm2{,}16\cdot\sqrt{0{,}106}\right],$$

где $t_\gamma=2{,}16$ находим по таблице t-распределения при $\alpha=1-\gamma=0{,}05$ и $\nu=n-k-1=14-2=13$.

Отсюда

$$1{,}41609-0{,}70324\le\beta_1\le1{,}41609+0{,}70324;$$
$$0{,}71285\le\beta_1\le2{,}11933.$$

Доверительную границу для условного математического ожидания y найдем в точке, определяемой вектором начальных условий: $x^0=\begin{pmatrix}1\\x_2^0\end{pmatrix}=$

$=\begin{pmatrix}1\\50\end{pmatrix}$ с надежностью $\gamma=0{,}95$.

Предварительно найдем матричное произведение

$$(x^0)^T(X^TX)^{-1}x^0=\begin{pmatrix}1&50\end{pmatrix}\begin{pmatrix}4{,}0746515&-0{,}0843188\\-0{,}0843188&0{,}0017738\end{pmatrix}\begin{pmatrix}1\\50\end{pmatrix}=$$

$$=\begin{pmatrix}-0{,}1413185&0{,}0043712\end{pmatrix}\begin{pmatrix}1\\50\end{pmatrix}=0{,}0772415.$$

Согласно (4.56) интервальная оценка для $\tilde{y}$ равна:

$$\tilde{y} \in \left[(-29{,}7596 + 1{,}41609) \pm 2{,}16 \cdot 7{,}732\sqrt{0{,}0772415}\right].$$

Отсюда

$$\tilde{y} \in \left[41{,}0449 \pm 4{,}6416\right]$$

и окончательно

$$36{,}4033 \le \tilde{y} \le 45{,}6865.$$

Таким образом, с доверительной вероятностью $\gamma = 0{,}95$ мы можем гарантировать, что при значении $x_2^0 = 50$ фондоотдача завода будет находиться в интервале от 36,4033 до 45,6865.

4.3.5. Взвешенный метод наименьших квадратов

В рассмотренной выше линейной регрессионной модели, которую часто называют **классической**, предполагается, что случайные составляющие ε_i не коррелируют друг с другом и ковариационная матрица случайного вектора ε имеет вид $\Sigma(\varepsilon) = \sigma^2 E_n$. Однако на практике данное ограничение выполняется далеко не всегда. В экономических исследованиях часто приходится иметь дело с зависимыми и неравноточными наблюдениями.

Рассмотрим регрессионную модель вида

$$Y = X\beta + \xi,$$

где

$$\Sigma(\xi) = \sigma^2 \Omega, \tag{4.61}$$

а случайный вектор ξ имеет n-мерный нормальный закон распределения $\xi \in N_n(0; \sigma^2\Omega)$, Ω - известная положительно определенная матрица.

Всегда существует такая невырожденная матрица A, что $AA^T = \Omega$.

Исследуем закон распределения случайного вектора

$$\varepsilon = A^{-1}\xi. \tag{4.62}$$

Как линейная функция вектора ξ вектор ε также подчиняется нормальному закону распределения. Легко показать, что $M\varepsilon = 0$, а ковариационная матрица имеет вид:

$$\Sigma_\varepsilon = M(A^{-1}\xi)(A^{-1}\xi)^T = A^{-1} M\xi\xi^T (A^{-1})^T = A^{-1}\sigma^2\Omega(A^{-1})^T.$$

Здесь учитывается, что $M\xi\xi^T = \Sigma(\xi) = \sigma^2\Omega$.

Подставляя вместо Ω его выражение и учитывая, что умножение матрицы на обратную дает единичную матрицу, будем иметь:

$$\Sigma\varepsilon = \sigma^2 A^{-1} AA^T (A^{-1})^T = \sigma^2 E_n, \text{ так как } (A^{-1})^T = (A^T)^{-1}.$$

Таким образом, мы доказали, что

$$\varepsilon \in N_n(0; \sigma^2 E_n). \tag{4.63}$$

Умножим все элементы уравнения (4.61) слева на матрицу A^{-1}.

Получим $A^{-1}Y = A^{-1}X\beta + A^{-1}\xi$.

Принимая, что $A^{-1}Y = U$; $A^{-1}X = Z$ и $A^{-1}\xi = \varepsilon$, будем иметь:

$$U = Z\beta + \varepsilon, \tag{4.64}$$

где случайный вектор ε удовлетворяет условию (4.63).

Таким образом, мы получили классическую линейную регрессионную модель. Тогда согласно (4.36) оценки b параметров β уравнения регрессии, полученные методом наименьших квадратов, имеют вид:

$$b = (Z^T Z)^{-1} Z^T U. \tag{4.65}$$

Подставив вместо z и u их выражения, получим

$$b = \left[(A^{-1}X)^T (A^{-1}X)\right]^{-1} (A^{-1}X)^T (A^{-1}Y) = \left[X^T (AA^T)^{-1} X\right]^{-1} \cdot X^T (AA^T)^{-1} Y.$$

Так как $AA^T = \Omega$, то

$$b = (X^T\Omega^{-1}X)^{-1} X^T\Omega^{-1}Y. \tag{4.66}$$

Покажем, что вектор b есть несмещенная оценка вектора β, т.е., что $Mb = \beta$.

Учитывая (4.61) и что $M\xi = 0$, будем иметь

$$Mb = (X^T\Omega^{-1}X)^{-1} X^T\Omega^{-1} M(X\beta + \varepsilon) = (X^T\Omega^{-1}X)^{-1}(X^T\Omega^{-1}X)\beta = \beta.$$

Ковариационная матрица вектора оценок коэффициентов регрессии b согласно (4.40) имеет вид:

$$\Sigma(b) = \sigma^2 (Z^T Z)^{-1} = \sigma^2 (X^T\Omega^{-1}X)^{-1}. \tag{4.67}$$

Как линейная функция вектора Y вектор b имеет $(k+1)$-мерный закон распределения. Таким образом,

$$b \in N_{(k+1)}(\beta; \sigma^2 (X^T\Omega X)^{-1}). \tag{4.68}$$

По аналогии с (4.47) можно записать несмещенную оценку остаточной дисперсии σ^2:

$$\hat{s}^2 = \frac{1}{n-k-1}(U - Zb)^T (U - Zb).$$

Подставляя вместо z и u их выражения, окончательно получим:

$$\hat{s}^2 = \frac{1}{n-k-1}(Y - xb)^T \Omega^{-1} (Y - xb). \tag{4.69}$$

Как и в случае классической линейной регрессии (п. 4.3), проверка значимости уравнения сводится к проверке на уровне значимости α нулевой

гипотезы $H_0:\beta=0$, для чего используется F-критерий, основанный на статистике (4.49):

$$F_{набл}=\frac{\frac{1}{k+1}Q_R}{\frac{1}{n-k-1}Q_{ост}}.$$

В нашем случае

$$Q_R=(Ub)^T(Ub)=(Xb)^T\Omega^{-1}(Xb); \qquad (4.70)$$

$$Q_{ост}=(Y-Xb)^T\Omega^{-1}(Y-Xb). \qquad (4.71)$$

Значение $F_{набл}$ сравнивают с критическим значением $F_{кр}$, определяемым по таблице F-распределения при заданном уровне значимости α и числах степеней свободы числителя $\nu_1=k+1$ и знаменателя $\nu_2=n-k-1$. Уравнение регрессии значимо, т.е. гипотеза $H_0:\beta=0$ отвергается, если $F_{набл}>F_{кр}(\alpha;\nu_1;\nu_2)$. В противном случае, когда $F_{набл}\le F_{кр}$, уравнение незначимо и на этом регрессионный анализ заканчивается.

Для значимого уравнения регрессии представляет интерес проверка значимости отдельных коэффициентов регрессии, т.е. нулевой гипотезы $H_0:\beta_j=0$.

Гипотеза H_0 отвергается, если

$$t_{набл}=\frac{b_j}{\hat{s}_{b_j}},$$

где

$$\hat{s}^2_{b_j}=\hat{s}^2\left[(X^T\Omega^{-1}X)^{-1}\right]_{jj} \qquad (4.72)$$

есть j-й диагональный элемент оценки ковариационной матрицы

$$\hat{\Sigma}(b)=\hat{s}^2(X^T\Omega^{-1}X)^{-1}, \qquad (4.73)$$

окажется больше критического значения $t_{кр}$ $(\alpha;\nu=n-k-1)$, найденного по таблице t-распределения при заданных значениях α и $\nu=n-k-1$.

Интервальную оценку для коэффициента регрессии с доверительной вероятностью γ можно определить как

$$\beta_j\in\left[b_j\pm t_\gamma\hat{s}_{b_j}\right],$$

где t_γ - определяют по таблице t-распределения по заданным значениям γ и $\nu_2=n-k-1$.

Во взвешенном методе наименьших квадратов предполагают, что матрица Ω известна, так как на основании наблюдений Y ее определить невозможно. Это связано с необходимостью оценить $n(n+1)/2$ элементов этой матрицы по результатам n наблюдений.

Рассмотрим частный случай использования взвешенного метода наименьших квадратов, когда матрица Ω является диагональной с элементами $\Omega_{ii}=\sigma_i^2$:

$$\Omega=\begin{pmatrix}\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & \sigma_n^2\end{pmatrix}.$$

Здесь принимается, что отклонения ε_i некоррелированы между собой и имеют разные, но известные дисперсии. Однако и в этом случае использованию взвешенного метода наименьших квадратов должны предшествовать определенные априорные предположения относительно дисперсий σ_i^2 случайных ошибок ε_i.

Например, если предположить, что относительная ошибка измерения y_i постоянна и равна σ_0, то среднеквадратическое отклонение σ_i будет пропорционально математическому ожиданию $\tilde{y}_i=My_i/x$ этой величины, т.е. $\sigma_i=\sigma_0\tilde{y}_i$. В этом случае для решения задачи можно предположить двушаговую процедуру, согласно которой на первом шаге обычным методом наименьших квадратов находят значения оценки уравнения регрессии $(\hat{y}_i)$ и среднюю относительную ошибку аппроксимации $\bar{\delta}=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i-\hat{y}_i}{\hat{y}_i}=$ $=\hat{\sigma}_0$, а на втором, полагая, что оценка среднеквадратического отклонения величины y_i равна $\hat{s}_i=\hat{\sigma}_0\hat{y}_i$, используют взвешенный метод наименьших квадратов с матрицей

$$\Omega^{\frac{1}{2}}=\hat{\sigma}_0\begin{pmatrix}\hat{y}_1 & 0 & . & 0\\ 0 & \hat{y}_2 & . & 0\\ . & . & . & .\\ 0 & 0 & . & \hat{y}_n\end{pmatrix}.$$

При планировании экспериментов встречаются случаи, когда в каждой точке факторного пространства ($x_{i1},x_{i2},\ldots,x_{ik}$, где $i=1,2,\ldots,n$) наблюдается различное число n_i значений результативного показателя $y_{i1},y_{i2},\ldots,y_{in_i}$.

В этом случае при построении уравнения регрессии за значение результативного показателя принимают его среднее значение $\bar{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}$ в i-й

точке факторного пространства. В предположении, что измерения y_{ij} равноточны с дисперсией $Dy_{ij} = \sigma^2$ для всех i=1,2,...,m и j=1,2,...,n, имеем $D\bar{y}_i = \frac{\sigma^2}{n_i}$. Оценку уравнения регрессии можно найти с помощью взвешенного метода наименьших квадратов, когда матрица Ω имеет вид

$$\Omega = \begin{pmatrix} n_1 & 0 & . & 0 \\ 0 & n_2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & n_m \end{pmatrix}.$$

4.4. Задачи и упражнения

4.1. Из генеральной совокупности (x, y), где y имеет нормальный закон распределения с условным математическим ожиданием $\tilde{y} = M(y/x) = \beta_0 + \beta_1 x$ и дисперсией σ^2, взята случайная выборка объемом n. Пусть (x_i, y_i) результат i-го наблюдения (i=1,2,...,n).

Определить:

а) закон распределения оценки метода наименьших квадратов b_0 параметра β_0;

б) закон распределения оценки b_1 параметра β_1;

в) закон распределения оценки $\hat{y} = b_0 + b_1 x$.

4.2. По условию задачи 4.1 определить математическое ожидание суммы квадратов отклонений, обусловленных регрессией, т.е. MQ_R, где $Q_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$.

4.3. По условию задачи 4.1 определить математическое ожидание суммы квадратов отклонений, обусловленных остаточной вариацией относительно линии регрессии, т.е. $MQ_{ост}$, где $Q_{ост} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

4.4. Доказать, что вектор b, полученный методом наименьших квадратов, является несмещенной оценкой вектора генеральных коэффициентов регрессии β.

4.5. Доказать, что выражение $b = (X^T X)^{-1} X^T Y$ - есть оценка метода наименьших квадратов вектора генеральных коэффициентов регрессии β.

4.6. Доказать, что статистика $\hat{s}^2 = \frac{1}{n-k-1}(Y - Xb)^T (Y - Xb)$ есть несмещенная оценка остаточной дисперсии σ^2.

4.7. Доказать справедливость следующего равенства относительно ковариационной матрицы вектора коэффициентов b:

$$\Sigma_b = M\left[(b-\beta)\cdot(b-\beta)^T\right] = \sigma^2 (X^T X)^{-1}.$$

4.8. Доказать, что в линейном анализе, когда остатки ε_i независимы и $\varepsilon_i \in N(0,\sigma)$ для всех i=1,2,...,n, статистика $t_j = \frac{b_j - \beta_j}{\hat{s}_{b_j}}$ имеет t-распределение Стьюдента с числом степеней свободы $\nu = n - k - 1$.

4.9. Доказать, что в точке, задаваемой вектором начальных условий x^0, оценка уравнения регрессии $\hat{y}_{x^0}$ имеет нормальный закон распределения $\hat{y}_{x^0} \in N(\tilde{y}_{x^0}, \hat{s} \cdot \sqrt{(x^0)^T (X^T X)^{-1} x^0})$, где $\tilde{y}_{x^0}$-значение генерального уравнения регрессии в точке x^0.

4.10. Доказать справедливость следующего равенства $Y^T Y = (Xb)^T (Xb) + (Y - Xb)^T (Y - Xb)$, характеризующего разложение общей вариации результативного показателя y на составляющие, обусловленные регрессией Q_R и влиянием неучтенных факторов $Q_{ост}$.

4.11. Доказать, что математическое ожидание квадратичной формы $Q_R = (Xb)^T (Xb)$, характеризующей вариацию y, обусловленную регрессией, равно следующему выражению:

$$MQ_R = (X\beta)^T (X\beta) + (k+1)\sigma^2.$$

4.12. Доказать, что математическое ожидание квадратичной формы $Q_{ост} = (Y - Xb)^T (Y - Xb)$, характеризующей вариацию y, обусловленную влиянием неучтенных в модели факторов, равно выражению: $MQ_{ост} = (n - k - 1)\sigma^2$.

4.13. Доказать, что при β=0 статистика $F = \frac{\frac{1}{k+1} Q_R}{\frac{1}{n-k-1} Q_{ост}}$ имеет F-распределение с числом степеней свободы $\nu_1 = k+1$ и $\nu_2 = n-k-1$.

4.14. На основании данных о зависимости усушки формового хлеба от продолжительности хранения (табл. 4.4) требуется:

Таблица 4.4

Продолжительность хранения (ч)	1	3	6	8	10
Усушка (% к весу горячего хлеба)	1,6	2,4	2,8	3,2	3,3

а) найти оценки b_0, b_1 и остаточной дисперсии s^2 в предположении, что генеральное уравнение регрессии имеет вид $\tilde{y} = \beta_0 + \beta_1 x$;

б) проверить при α =0,05 значимости уравнения регрессии, т.е. гипотезу $H_0: \beta_1 = 0$;

в) с надежностью γ =0,9 определить интервальные оценки параметров β_0, β_1;

г) с надежностью γ =0,95 определить интервальную оценку условного математического ожидания $\tilde{y}$ при x^0=6;

д) определить при γ =0,95 доверительный интервал предсказания y_{n+1} в точке x^0=12.

4.15. На основании данных о динамике процента хронических больных на 1000 жителей, приведенных в табл. 4.5.

Таблица 4.5

Годы, x	0	1	2	3	4
%, y	10	8	5	3	4

и предположения, что генеральное уравнение регрессии имеет вид $\tilde{y} = \beta_0 + \beta_1 x$:

а) определить оценки b_0 и b_1 параметров уравнения регрессии и остаточной дисперсии s^2;

б) проверить при α =0,01 значимость уравнения регрессии, т.е. гипотезу $H_0: \beta_1 = 0$;

в) с надежностью γ =0,95 определить интервальные оценки параметров β_0 и β_1;

г) с надежностью γ =0,9 определить интервальную оценку условного математического ожидания $\tilde{y}$ при x^0=4;

д) определить при γ =0,9 доверительный интервал предсказания $\tilde{y}_{n+1}$ в точке x=5.

4.16. Результаты исследования динамики веса поросят приведены в табл. 4.6.

Таблица 4.6

Возраст (недели) (x)	0	1	2	3	4	5	6
Вес (кг) (y)	1,2	2,5	3,9	5,2	6,4	7,7	9,2

Предполагая, что генеральное уравнение регрессии - линейное:

а) определить оценки b_0 и b_1 параметров уравнения регрессии и остаточной дисперсии s^2;

б) проверить при уровне значимости α =0,05 значимость уравнения регрессии, т.е. гипотезу $H_0: \beta_1 = 0$;

в) с надежностью γ =0,8 определить интервальную оценку параметров β_0, β_1;

г) с надежностью γ =0,98 определить интервальную оценку условного математического ожидания $\tilde{y}$ при x^0=3 и x^0=6;

д) определить при γ =0,98 доверительный интервал предсказания y_{n+1} в точке x=8.

4.17. При анализе зависимости удельного объема валовой продукции совхозов области (y) от средней мощности тракторов (x), приходящихся на одного работника, получены следующие результаты (табл.4.7):

Таблица 4.7

Средняя мощность тракторов, л.с.	Число совхозов	Валовая продукция на одного работника, тыс. руб.
4,1	10	1,4
6,1	27	2,0
7,8	34	2,5
9,9	22	2,6
12,4	7	3,2

Требуется:

а) определить оценки b_0 и b_1 параметров линейного уравнения регрессии и остаточной дисперсии s^2;

б) проверить при α =0,01 значимость уравнения регрессии, т.е. гипотезу $H_0: \beta_1 = 0$;

в) с надежностью γ =0,9 определить интервальную оценку параметров β_0, β_1 и условного математического ожидания $\tilde{y}$ при x=10.

4.18. Себестоимость y (тыс.руб.) одного экземпляра книги в зависимости от тиража x (тыс. экз.) характеризуется данными, собранными издательством (табл.4.8). Определить оценки b_0 и b_1 параметров уравнения регрессии гиперболического вида $\tilde{y} = \beta_0 + \beta_1 \frac{1}{x}$. С надежностью γ =0,9 построить доверительные интервалы для параметров β_0 и β_1, а также условного математического ожидания y при x=10.

Таблица 4.8

x (тыс. экз.)	1	2	3	5	10	20	30	50
y (тыс.руб.)	9,10	5,30	4,11	2,83	2,11	1,62	1,41	1,30

4.19. Данные о расходе электроэнергии (кВт.ч) на изготовление 1 т цемента (y) в зависимости от объема выпуска (x) продукции (тыс. т) цементными заводами приводятся в табл. 4.9.

Таблица 4.9

Выпуск продукции (тыс. т) x	5	10	15	20	25	30
Расход электроэнергии (кВт.ч) y	10,0	8,2	7,3	6,3	6,4	5,2

Определить оценки b_0 и b_1 параметров уравнения регрессии вида $\tilde{y} = \beta_0 + \beta_1 \frac{1}{x}$, проверив при α =0,05 гипотезу $H_0: \beta_1 = 0$, и построить с надежностью γ =0,9 доверительные интервалы для параметров β_0 и β_1 и условного математического ожидания y при x=20.

4.20. Измерение некоторой величины y через равные интервалы x дали результаты, приведенные в табл. 4.10.

Таблица 4.10

x	−3	−2	−1	0	1	2	3
y	−1,70	−1,01	−0,21	0,52	0,73	1,30	1,50

Считая, что генеральное уравнение регрессии имеет вид $\tilde{y} = \beta_0 + \beta_1 x + \beta_2 x^2$, определить:

а) оценки b_0, b_1 и b_2 параметров уравнения;

б) оценку остаточной дисперсии s^2;

в) оценки дисперсий $\sigma^2_{b_j}$ выборочных характеристик b_j, где j=0,1,2.

Проверить:

а) значимость при α =0,05 уравнения регрессии, т.е. гипотезу $H_0: \beta_0 = \beta_1 = \beta_2 = 0$;

б) при α =0,05 значимость отдельных коэффициентов регрессии.

4.21. Осредненные данные по группе хозяйств, характеризующие связь между выходом продукции на 1 га сельскохозяйственных угодий (y) и основными средствами совхозов, также приходящимися на 1 га угодий (x), приведены в табл. 4.11.

Таблица 4.11

(млн руб.)

x	11,3	12,9	13,6	16,8	18,8	22,0	22,2	23,7
y	13,2	15,6	17,2	18,8	20,2	23,3	22,4	23,0

Требуется:

а) найти оценки параметров линейного уравнения регрессии, т.е. $\tilde{y} = \beta_0 + \beta_1 x$;

б) вычислить несмещенную оценку s^2 остаточной дисперсии;

б) при α =0,05 проверить значимость линейного уравнения регрессии;

в) построить с надежностью γ =0,95 доверительные интервалы параметров β_0, β_1 и условного математического ожидания $\tilde{y}$ при x=22.

4.22. По данным задачи 4.21:

а) определить оценки b_0, b_1 и b_2 параметров уравнения регрессии вида $\tilde{y} = \beta_0 + \beta_1 x + \beta_2 x^2$;

б) вычислить несмещенную оценку s^2 остаточной дисперсии и сравнить ее с соответствующей оценкой в задаче 4.21;

в) проверить при α =0,05 значимость уравнения регрессии, т.е. гипотезы $H_0: \beta_0 = \beta_1 = \beta_2 = 0$;

г) проверить при α =0,05 значимость коэффициентов регрессии, т.е. гипотезу $H_0: \beta_j$, где j=0,1,2;

д) построить с надежностью γ =0,9 доверительные интервалы параметров $\beta_0, \beta_1, \beta_2$ и условного математического ожидания $\tilde{y}$ при x=22.

4.23. Данные анализа зависимости средней себестоимости хлопка (y) от урожайности приведены в табл. 4.12.

Таблица 4.12

Средняя урожайность (ц/га) x	13	16	22	28	33
Средняя себестоимость 1 ц (тыс.руб.) y	26	22	18	15	14

Считая, что генеральное уравнение регрессии имеет вид $\tilde{y} = \beta_0 + \beta_1 \frac{1}{x}$, определить:

а) оценки b_0, b_1 параметров уравнения регрессии;

б) оценки остаточной дисперсии s^2;

в) точечные оценки дисперсий выборочных характеристик b_0, b_1 и условной средней $\tilde{y} = \beta_0 + \beta_1 \frac{1}{x}$ при X=16.

4.24. Результаты лабораторных испытаний прочности стальных проволок различных диаметров приведены в табл. 4.13.

Таблица 4.13

Диаметр проволоки (мм) x	1	1,2	1,4	1,5	1,6
Разрывное усилие (т) y	0,15	0,20	0,3	0,4	0,6

Приняв, что генеральное уравнение регрессии имеет вид $\tilde{y} = \beta_0 + \beta_1 x^2$, определить:

а) оценки b_0, b_1 параметров β_0, β_1;

б) несмещенную оценку $\hat{s}^2$ остаточной дисперсии;

в) интервальные оценки неизвестных параметров β_0, β_1 и условного математического ожидания $\tilde{y}$ при x^0=1, приняв γ =0,9.

4.25. По приведенным в табл. 4.14 данным n=14 машиностроительных предприятий провести регрессионный анализ зависимости индекса снижения себестоимости продукции (y) от трудоемкости единицы продукции (x) и удельного веса покупных изделий.

Таблица 4.14

№ п/п	y	x_1	x_2	№ п/п	y	x_1	x_2
1	204	0,23	0,40	8	56	0,26	0,44
2	209	0,24	0,26	9	52	0,49	0,17
3	222	0,19	0,40	10	46	0,36	0,39
4	236	0,17	0,50	11	53	0,37	0,33
5	62	0,23	0,40	12	31	0,43	0,25
6	53	0,43	0,19	13	146	0,35	0,32
7	172	0,31	0,25	14	18	0,38	0,02

4.26. По данным n=12 угольных шахт (табл. 4.15) провести регрессионный анализ зависимости полной себестоимости добычи 1 т угля y (тыс.руб.) от среднесуточной добычи угля на шахте - x_1 (т) и удельного веса комбайновой проходки выработки x_2 (%).

Таблица 4.15

№ п/п	y	x_1	x_2	№ п/п	y	x_1	x_2
1	12,2	4795	69	7	12,7	4888	55
2	7,6	6962	82	8	10,5	6237	81
3	10,4	6571	87	9	15,1	2997	65
4	9,9	4249	92	10	10,6	2990	98
5	15,7	9540	23	11	15,2	1748	100
6	14,0	3488	31	12	17,2	2128	69

4.27. На основании данных n=16 цементных заводов (табл. 4.16) провести регрессионный анализ зависимости выработки натурального цемента на одного работающего y (т/чел.) от среднечасовой производительности вращающихся печей x_1 (т/ч) и среднечасовой производительности цементных мельниц x_2 (т/ч).

Таблица 4.16

№ п/п	y	x_1	x_2	№ п/п	y	x_1	x_2
1	996	37	46	9	1620	44	38
2	1362	23	44	10	3018	34	32
3	759	15	26	11	1831	63	50
4	1216	36	34	12	1167	8	23
5	1350	26	26	13	1424	44	38
6	1026	24	31	14	2387	43	35
7	1099	15	20	15	1438	30	29
8	1726	33	32	16	1199	19	38

4.5. Задания для самостоятельной работы на ЭВМ по регрессионному анализу

По данным Приложения 1 провести регрессионный анализ.

1. Найти оценку уравнения регрессии вида:

а) линейное - $\tilde{y} = \beta_0 + \beta_1 x_1 + ... + \beta_5 x_5$;

б) степенное - $\hat{y} = \beta_0 x_1^{\beta_1} x_2^{\beta_2} ... x_5^{\beta_5}$.

2. Проверить значимость уравнения и коэффициентов регрессии $\beta_0, \beta_1, ..., \beta_5$.

3. С помощью алгоритма пошагового регрессионного анализа получить уравнение регрессии с максимальным числом значимых коэффициентов регрессии.

4. Для полученного уравнения регрессии:

а) дать экономическую интерпретацию коэффициентам регрессии b_j и эластичности $Э_j = b_j \frac{\bar{x}_j}{\bar{y}}$, а также множественному коэффициенту детерминации;

б) оценить адекватность полученной модели по величине абсолютных $e_j = y_i - \hat{y}_i$ и относительных отклонений $\delta_i = \frac{e_i}{y_i}$, где i=1,2,...,n, а также по величине $\hat{s}$ и средней относительной ошибки $\bar{\delta} = \frac{1}{n} \sum_{i=1}^{n} |\delta_i|$;

в) построить интервальные оценки для коэффициентов регрессии.

ГЛАВА 5

КОМПОНЕНТНЫЙ АНАЛИЗ

Компонентный анализ относится к многомерным методам снижения размерности. Он содержит один метод - метод главных компонент. В этом методе линейные комбинации случайных величин определяются характеристическими векторами ковариационной матрицы. **Главные компоненты** представляют собой ортогональную систему координат, в которой дисперсии компонент характеризуют их статистические свойства.

5.1. Статистический подход в методе главных компонент

В зависимости от конкретных задач, решаемых в экономике, используется один из методов факторного анализа, или метод главных компонент.

Метод главных компонент считается статистическим методом. Однако есть другой подход, приводящий к методу главных компонент, но не являющийся статистическим. Этот подход связан с получением наилучшей проекции точек наблюдения в пространстве меньшей размерности. Для решения подобной задачи необходимо знать матрицу вторых моментов.

В статистическом, основном для нас, подходе, которому посвящен данный учебник, задача будет заключаться в выделении линейных комбинаций случайных величин, имеющих максимально возможную дисперсию. Он опирается на ковариационную или корреляционную матрицу этих случайных величин. У этих двух разных подходов есть общий аспект: *использование матрицы вторых моментов как исходной для начала анализа.*

Из сказанного следует, что для овладения методом главных компонент необходимо пользоваться методами теории вероятностей и математической статистики на основе моделей линейной алгебры. Рассмотрим основные положения этих математических дисциплин, на которые опирается метод главных компонент.

Учитывая, что объекты исследования в экономике (фирма, завод, министерство, отрасль народного хозяйства, экономика страны) характеризуются большим, но конечным количеством признаков (характеристик), влияние которых подвергается воздействию большого количества случайных причин, в качестве моделей в статистическом плане возьмем многомерные распределения, а в алгебраическом - многомерное пространство признаков.

5.1.1. Многомерное нормальное распределение как модель

Математической моделью, на которой основываются методы многомерного статистического анализа (в том числе и методы факторного и компонентного анализа), является многомерное нормальное распределение. Это распределение является зачастую наиболее хорошим приближением к действительному распределению на практике в экономических задачах.

Из центральной предельной теоремы следует, что предельным распределением одномерных независимых случайных величин является одномерный нормальный закон.

Из обобщенной центральной предельной теоремы получаем, что предельным распределением в случае нескольких измерений является многомерное нормальное распределение.

В настоящее время многомерные методы, основанные на нормальном распределении, нашли широкое распространение при изучении различных процессов в экономике.

Среди математических методов многомерного анализа выделяют:

1) корреляцию;

2) аналоги одномерных статистических методов в многомерном анализе;

3) проблемы системы координат;

4) проблемы классификации;

5) зависимость наблюдений.

1. При изучении корреляции нас интересуют *различные коэффициенты корреляции.*

Выборочные коэффициенты корреляции используются для оценки соответствующих параметров распределения.

Частный коэффициент корреляции измеряет зависимость между случайными величинами, когда действие других коррелированных случайных величин исключено.

При помощи **множественного коэффициента корреляции** распространяется понятие коэффициента корреляции на измерение зависимости между одной случайной величиной и множеством случайных величин.

2. *Аналоги одномерных статистических методов.*

Следует отметить, что многие проблемы, решаемые в многомерном статистическом анализе, когда изучаются многомерные совокупности, имеют свои аналоги при изучении одномерных совокупностей. Представим эти проблемы в виде таблицы (табл.5.1).

Таблица 5.1

Аналоги одномерных статистических методов

Одномерное случайное распределение	Многомерное случайное распределение
Проверка гипотезы о математическом ожидании $M[x]=\mu$	Проверка гипотезы о векторе математических ожиданий $M[x_1, x_2, \ldots, x_n]^T=\mu$

Продолжение

Одномерное случайное распределение	Многомерное случайное распределение
t-критерий Стьюдента	Обобщенный T^2 критерий для многомерного распределения
Метод наименьших квадратов	Обобщение метода наименьших квадратов на многомерный случай
Дисперсионный анализ	Обобщение дисперсионного анализа на многомерное распределение

Для этих проблем выбор системы координат связан с линейным преобразованием переменных.

3. *Проблемы системы координат.*

В ряде случаев удачный выбор новой системы координат может наиболее экономным способом выявить некоторые важные для исследователя свойства многомерной случайной совокупности.

Примером может служить выявление главных компонент, т.е. отыскание такой нормализованной линейной комбинации случайных величин, чтобы ее дисперсия была максимальной или минимальной. Это равноценно повороту осей, который приводит ковариационную матрицу к диагональной форме. Другой пример - нахождение канонических корреляций. Для решения подобных задач требуется определение характеристических векторов и характеристических корней различных систем линейных алгебраических уравнений.

4. *Проблемы классификации.*

Это разбиение множества случайных величин на подмножества. Возникает важный вопрос проверки гипотезы о независимости подмножеств. Факторный анализ, метод главных компонент и кластерный анализ обычно используют в задачах многомерной классификации.

5. *Зависимость наблюдений.*

Если в экономических исследованиях мы занимаемся анализом временных рядов, то сталкиваемся с наблюдениями над рядами случайных величин, последовательными во времени. Наблюдения в данный момент времени могут зависеть от ранее произведенных наблюдений. Это требует, например, изучения внутрирядной корреляции.

Поскольку в качестве основной статистической модели выступает многомерное нормальное распределение, остановимся более подробно на этом распределении, которое полностью определяется своей квадратичной формой, а последняя зависит от вектора математических ожиданий и ковариационной матрицы. Эта зависимость четко определяется следующей теоремой.

Теорема 5.1. *Если даны вектор μ и положительно определенная матрица Σ, то существует такая многомерная нормальная плотность распределения вероятностей:*

$$Nn(x/\mu,\Sigma) = (2\pi)^{-\frac{1}{2}n} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \qquad (5.1)$$

что математическое ожидание случайного вектора x с этой плотностью распределения есть μ и ковариационная матрица есть Σ.

Обычно плотность распределения вероятностей обозначают так, как записано слева в равенстве (5.1), а многомерный нормальный закон распределения обозначают $N(\mu,\Sigma)$. В данном распределении нас должна заинтересовать структура ковариационной матрицы и ее связь с корреляционной матрицей. Это можно сделать в общем виде для случайного вектора n-го порядка. Однако удобней обратиться к простейшему многомерному распределению - двумерному.

При рассмотрении двумерного нормального распределения мы легко убеждаемся в том, что коэффициенты корреляции и дисперсии случайных величин являются основными числовыми характеристиками наряду с математическими ожиданиями. Если конечное число случайных величин превосходит n=2, то роль дисперсий выполняет ковариационная (корреляционная) матрица. Элементы этой матрицы получаются из экспериментальных или статистических данных и являются статистическими величинами, требующими своей оценки. В методе главных компонент в дальнейшем потребуется также оценка и весовых коэффициентов модели.

5.1.2. Геометрическая интерпретация плотности вероятности двумерного нормального распределения

Плотность распределения вероятностей n-мерного случайного вектора в n-мерном евклидовом пространстве постоянна на эллипсоидах, определяемых положительно определенными квадратичными формами

$$(x-\mu)^T \Sigma^{-1} (x-\mu) = P_1$$

для каждого положительного значения P_1. Центром каждого эллипсоида является точка μ. Форма и положение эллипсоида определяются значением ковариационной матрицы.

Для заданной ковариационной матрицы размеры эллипсоида определяются значением P_1.

Для двумерного случая преобразуем координаты посредством нормирования отклонения от математического ожидания

$$y_i = \frac{(x_{ij} - \mu_j)}{\sigma_j} \qquad (j=1,2)$$

таким образом, чтобы центры линий, на которых плотность распределения постоянна, находились в начале координат.

Уравнения этих линий запишутся для двумерного случая так:

$$\frac{1}{1-\rho^2}(y_1^2 - 2\rho y_1 y_2 + y_2^2) = P_1.$$

При $\rho > 0$ большая ось эллипса наклонена под углом 45⁰ к оси x. Длина большой полуоси равна $\left[P_1\left(1+\rho\right)\right]^{1/2}$, длина малой полуоси равна $\left[P_1\left(1-\rho\right)\right]^{1/2}$. Если же $\rho < 0$, то большая ось эллипса наклонена под углом 135⁰ к оси x. Длина большой полуоси равна $\left[P_1\left(1-\rho\right)\right]^{1/2}$, а малой равна $\left[P_1\left(1+\rho\right)\right]^{1/2}$. Вероятностная поверхность представляет собой "нормальный холм" над плоскостью. Контуры равных плотностей аналогичны контурам равных высот на топографической карте. Если от эллипсоида в координатах y_j перейти обратно к координатам x_j

$$x_j = \sigma_j y_j + \mu_j,$$

то это означает, что каждый контур растягивается в σ_j раз в направлении j-й оси, а центр переносится в точку (μ_1, μ_2).

Для двумерных случайных величин имеются таблицы. Рассмотрим некоторые примеры квадратичных функций и их представление в матричной форме.

Пример 5.1

Задана квадратичная функция без линейной составляющей

$$f(x) = x_1^2 - 4x_1x_2 + x_1x_3 - 2x_2x_3 - x_3^2.$$

Определить матрицу A.

Решение

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

Пример 5.2

Используя функцию примера 1, представить квадратичную функцию в матричном виде, зная симметрическую матрицу А.

Решение

$$f(x) = x^T Ax,$$

т.е.

$$f(x) = \begin{pmatrix} x_1 x_2 x_3 \end{pmatrix} \begin{pmatrix} 1 & -2 & 0{,}5 \\ -2 & 0 & -1 \\ 0{,}5 & -1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Пример 5.3

Используя функцию, имеющую линейную часть, представить ее в матричном виде:

$$f(x) = 4x_1 + 5x_2 + 2x_1x_2 - 2x_1^2 - 2x_2^2.$$

Решение

$$f(x) = a^T x + x^T Ax,$$

где

$$a^T = (4 \quad 5);$$

$$f(x) = (4 \quad 5)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (x_1 \quad x_2)\begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Таким образом, при наличии многомерного нормального распределения о характере и тесноте стохастической связи между признаками можно судить по квадратичной форме, которая определяется ковариационной матрицей. Всестороннее изучение ковариационной или корреляционной матрицы и ее преобразование к новой системе координат позволят вскрыть важные объективно существующие закономерности методами факторного анализа и главных компонент.

5.2. Линейная модель метода главных компонент

5.2.1. Общие положения

Рассмотрим модель метода главных компонент:

$$y'_j = \sum_{i=1}^{n} a_{jr} f_r, \qquad (5.2)$$

где f_r - r-я главная компонента;

a_{jr} - вес r-й компоненты на j-й переменной;

y'_j - центрированное (нормированное) значение j-го признака.

По структуре она напоминает (6.4), но имеется существенная разница. ***Во-первых***, отсутствуют специфические (индивидуальные) факторы. ***Во-вторых***, число главных компонент $r=n$ здесь соответствует числу признаков n. Значит, в полной модели главных компонент исчерпывается вся дисперсия исследуемого процесса.

Как будет показано ниже, главные компоненты являются характеристическими векторами ковариационной матрицы.

Множество главных компонент представляет собой удобную систему координат, а соответствующие **дисперсии главных компонент** характеризуют их статистические свойства. Из общего числа главных компонент для исследования, как правило, оставляют m ($m<n$) наиболее весомых, т.е. вносящих максимальный вклад в объясняемую часть общей дисперсии. Опыт показал, что $m \approx (0{,}1 \div 0{,}25)n$. Для экономической интерпретации полученных результатов самыми наглядными являются случаи, когда m=1, 2 или 3.

Таким образом, несмотря на то, что в методе главных компонент для точного воспроизведения корреляций и дисперсий между переменными необходимо найти все компоненты, большая доля дисперсии объясняется небольшим числом главных компонент. Кроме того, можно по признакам описать главные компоненты, а по главным компонентам описать признаки. Для центроидного метода факторного анализа это принципиально невозможно; можно лишь добиваться, чтобы дисперсия остатков была минимальной. С.Р.Рао показал, что метод главных компонент одинаково хорошо приближает дисперсии и ковариации. Наконец, следует отметить еще одно существенное свойство метода - это его линейность и аддитивность. Центроидный метод, например, несет в себе только гипотезу линейности. Если она не верна, то результаты могут быть использованы только для первого приближения. В настоящее время часто используется центроидный метод для получения приближенных оценок, которые затем уточняются методом максимума правдоподобия.

5.2.2. Получение матрицы весовых коэффициентов по алгоритму метода главных компонент

Остановимся на получении матрицы весовых коэффициентов в компонентном анализе. Обратимся к теореме, не приводя доказательства.

Теорема 5.2. *Для любой симметрической положительно определенной матрицы R существует ортогональная матрица U, такая, что*

$$U^T RU = \Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix},$$

а все элементы матрицы Λ *положительны* ($\lambda_r > 0$), *определитель матрицы R положителен* $|R| > 0$, *все главные миноры матрицы положительны. При этом:*

1. λ_r *- собственное значение, или характеристический корень матрицы R.*

2. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$.

Не доказывая теоремы, воспользуемся U и Λ для получения матрицы нагрузок A. Составим характеристическое уравнение

$$|R - \lambda E| = 0. \qquad (5.3)$$

При его преобразовании с помощью матрицы U получим:

$$0 = |U^T (R - \lambda E) U| = |U^T (R - \lambda E) U| = |U^T RU - \lambda E| =$$

$$= |\Lambda - \lambda E| = \begin{vmatrix} \lambda_1 - \lambda & 0 & \dots & 0 \\ 0 & \lambda_2 - \lambda & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n - \lambda \end{vmatrix} = \prod_{j=1}^{n} (\lambda_j - \lambda). \qquad (5.4)$$

Значит, характеристические корни матрицы R совпадают с диагональными элементами преобразованной матрицы.

Воспользуемся основным уравнением метода главных компонент для ненормированных признаков x_j:

$$x_i = Af_i, \qquad (5.5)$$

$$x_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + \dots + a_{jn} f_{ni}, \qquad (5.6)$$

где x_{ji} - центрированное значение изучаемого признака;

f_{ri} - r-я главная компонента у i-го объекта исследования;

a_{jr} - весовой коэффициент r-й главной компоненты;

N - число объектов;

n - число признаков.

В матричном виде выражение (5.5) запишется

$$X = A_1 F_1, \qquad (5.7)$$

если X является матрицей центрированных значений переменных.

Если возьмем нормированные значения переменных - матрицу Y, то получим выражение

$$Y = AF. \qquad (5.8)$$

В выражениях (5.7) и (5.8) размерности матриц будут следующие:

X и $Y \rightarrow (n \times N)$, A и $A_1 \rightarrow (n \times n)$, F и $F_1 \rightarrow (n \times N)$.

При извлечении всех главных компонент матрица весовых коэффициентов будет квадратной порядка n.

Признаки, составляющие матрицу X или Y, подчинены нормальному закону распределения, как и главные компоненты. Следовательно, взаимосвязи между признаками определяются ковариационной матрицей Σ. Она является симметрической матрицей, на главной диагонали которой стоят дисперсии признаков, а внедиагональные элементы представляют собой ковариации.

Введем ортогональную матрицу U и перейдем вместо переменных x к новым переменным z, причем

$$\left.\begin{aligned} z &= U^T x \\ x &= Uz \end{aligned}\right\}, \qquad (5.9)$$

где

$$z = \begin{pmatrix} z_n \\ z_2 \\ \dots \\ z_n \end{pmatrix}; \qquad \overline{j = 1,n};$$

x - вектор центрированных значений параметров (5.9) иллюстрирует связь между переменными x и z;

вектор u_r - r-й столбец ортогональной матрицы U;

u_r^T - r-й вектор-строка в матрице U^T.

Вектор u_1 выбирается таким образом, чтобы дисперсия z_1 была максимальной. После получения вектора u_1 выбирается вектор u_2 таким образом, чтобы дисперсия z_2 была максимальной при условии, что z_2 не коррелирует с z_1. Короче говоря, процедура проводится для любого z_r так, чтобы дисперсия была максимальной при условии отсутствия корреляции со всеми переменными $z_1, z_2, \dots, z_{r-1}$. Обозначим невырожденную ковариационную матрицу через Σ, а дисперсию z_r через λ_r. Из (5.9) следует, что элемент z_r вектора z равен: $z_r = u_r^T x$, а λ_r (из теоремы 5.2) является собственным значением, которому соответствует собственный вектор u_r. Поэтому можно записать характеристическое уравнение

$$(\Sigma - \lambda_r E)\, u_r = 0$$

или

$$\Sigma u_r = \lambda_r u_r. \tag{5.10}$$

Слева умножим обе части равенства на вектор u_r^T:

$$u_r^T \Sigma u_r = u_r^T \lambda_r u_r.$$

Так как у ортогональных матриц $u_r^T u_r = 1$, то

$$\lambda_r = u_r^T \Sigma u_r,$$

а так как z_r не коррелированы, то при $r \neq s$ $u_r^T \Sigma u_s = 0$.

Это значит, что

$$\Lambda = U^T \Sigma U \tag{5.11}$$

является диагональной матрицей с элементами

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n.$$

Из (5.10) видно, что λ_r является собственным значением собственного вектора u_r.

Проведя нормирование z_r, получим r-ю главную компоненту f_r. При нормировании z_r дисперсии должны равняться единицам для всех индексов $r = \overline{1,n}$ при переменных z_r. Для этого надо разделить z_r на среднеквадратическое отклонение $\sqrt{\lambda}_r$. Получим

$$f_r = \lambda_r^{-1/2} z_r = \lambda_r^{-1/2} u_r^T x;$$

$$f = \Lambda^{-1/2} z = \Lambda^{-1/2} U^T x.$$

Отсюда

$$U\Lambda^{1/2} f = U\Lambda^{1/2} \Lambda^{-1/2} U^T x = UEU^T x = EUU^T x = EEx = x,$$

поэтому $x = U\Lambda^{1/2} f$. Итак

$$U\Lambda^{1/2} = A. \tag{5.12}$$

Запишем r-й столбец матрицы весовых коэффициентов

$$a_r = \lambda_r^{1/2} u_r. \tag{5.13}$$

Определим норму вектора весовых коэффициентов

$$a_r^T a_r = \lambda_r^{1/2} u_r^T u_r \lambda_r^{1/2} = \lambda_r.$$

Таким образом, нормой вектора a_r будет $\lambda_r^{1/2}$.

Из (5.12) вытекает, что

$$A^T A = \Lambda^{1/2} U^T U \Lambda^{1/2} = \Lambda E = \Lambda,$$

а также

$$AA^T = U\Lambda^{1/2} \Lambda^{1/2} U^T = U\Lambda U^T = UU^T \Sigma UU^T = E\Sigma E = \Sigma.$$

Следовательно:

$$\left.\begin{aligned} a_r^T a_r &= \lambda_r, \\ A^T A &= \Lambda, \\ AA^T &= \Sigma. \end{aligned}\right\}. \tag{5.14}$$

Если исходные данные нормированы, т.е. вместо матрицы X решение проводится на основе матрицы Y, тогда $AA^T = R$. При помощи матрицы весовых коэффициентов восстанавливается (редуцируется) корреляционная матрица.

Следы матриц Σ и Λ равны:

$$tr\Sigma = tr\Lambda. \tag{5.15}$$

Суммарная дисперсия всех n ненормированных исходных показателей x_j равна суммарной дисперсии ненормированных переменных z_r.

В корреляционной матрице R на главной диагонали стоят дисперсии нормированных исходных показателей, равные единицам. Следовательно,

$$trR = n, \qquad (5.16)$$

где n - число исходных показателей.

Представим схему алгоритма метода главных компонент для получения матрицы A.

$$X \to Y \to R \begin{matrix} \nearrow \Lambda \searrow \\ \searrow U \nearrow \end{matrix} A, \qquad (5.17)$$

где X - матрица исходных данных (нецентрированных и ненормированных);
Y - матрица нормированных исходных данных;
R - матрица парных коэффициентов корреляции (с единицами на главной диагонали);
Λ - диагональная матрица собственных значений;
U - ортогональная матрица собственных векторов;
A - матрица весовых коэффициентов.

В методе главных компонент, где матрица факторного отображения и матрица структуры совпадают, элемент матрицы A a_{jr} является коэффициентом корреляции между переменной y_j и главной компонентой f_r.

Определение 5.1. *Дисперсия одномерного нормального распределения имеет два аналога: ковариационную матрицу и определитель этой матрицы (скаляр), называемый* ***обобщенной дисперсией многомерного распределения.***

Возникает естественный вопрос, не изменяются ли сумма дисперсий исходных показателей (дисперсия процесса) и обобщенная дисперсия в результате ортогональных преобразований.

Теорема 5.3. *Ортогональное преобразование $x = Cf$ случайного центрированного вектора f оставляет инвариантной обобщенную дисперсию и сумму дисперсий компонент.*

Доказательство.

По условию

$$\begin{cases} Mf = 0 \\ Mff^T = \Sigma. \end{cases} \qquad (5.18)$$

Следовательно,

$$Mx = 0$$
$$Mxx^T = MCff^T C^T = C(Mff^T)C^T = C\Sigma C^T.$$

Обобщенная дисперсия x равна $|C\Sigma C^T| = |C||\Sigma||C^T| = |\Sigma|$. Она совпадает с обобщенной дисперсией f. Сумма дисперсий компонент x равна:

$$\sum_{j=1}^{n} Mx_j^2 = tr(C\Sigma C^T) = tr(\Sigma C^T C) = tr(\Sigma E) = tr\Sigma = \sum_{j=1}^{n} Mf_j^2. \qquad (5.19)$$

Следствие из теоремы 5.3. *Обобщенная дисперсия вектора главных компонент равна обобщенной дисперсии исходного вектора, а сумма дисперсий главных компонент равна сумме дисперсий исходных величин.*

Таким образом, рассмотрены кардинальные вопросы получения матрицы A. Если известна выборочная ковариационная матрица (или корреляционная), то можно получить матрицу A, элементы которой учитывают тесноту стохастической связи между признаками и главными компонентами. Зная весовые коэффициенты, можно редуцировать ковариационную (корреляционную) матрицу, а также диагональную матрицу Λ.

На основе вышеизложенного более подробно рассмотрим ***схему алгоритма метода главных компонент*** (5.17).

Введем пояснения к схеме:

I. *Исходная информация*:

а) N - число объектов исследования $(i = \overline{1, N})$;

б) n - число исходных показателей, по каждому из которых имеется значение x_{ij} j-го показателя на i-м объекте исследования;

в) δ - значение доли дисперсии, которая должна быть пояснена извлекаемыми первыми m главными компонентами;

г) данные, связанные с ограничениями используемого математического обеспечения;

д) данные о возможностях и режиме работы вычислительного центра.

II. *Решение.*

Ввод исходной информации, в том числе и матрицы X размерности (n x N):

1-й шаг - $\sum_{i=1}^{N} x_{ji}$;

2-й шаг - $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ji}$;

3-й шаг - выдача на печать $\bar{x}_j$, $(j = \overline{1,n})$;

4-й шаг - $x_{ji}^0 = x_{ji} - \bar{x}_j$ - центрирование;

5-й шаг - $(x_{ji}^0)^2 = (x_{ji} - \bar{x}_j)^2$;

6-й шаг - $\sum_{i=1}^{N}(x_{ji} - \bar{x}_j)^2$;

7-й шаг - $\frac{1}{N-1}\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2$;

8-й шаг - $S_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{ji} - \bar{x}_j)^2}$;

9-й шаг - выдача на печать дисперсий и средних квадратических отклонений признаков;

10-й шаг - расчет и формирование матрицы парных коэффициентов корреляции;

10-2821

11-й шаг - выдача на печать матрицы парных коэффициентов корреляции;

12-й шаг - получение матрицы собственных значений;

13-й шаг - получение матрицы собственных векторов;

14-й шаг - суммирование текущих собственных значений и остановка извлечения главных компонент как только $\sum \lambda_j$ первый раз превзойдет заданное значение объясняемой дисперсии первыми наиболее весомыми главными компонентами;

15-й шаг - получение матрицы индивидуальных значений главных компонент;

16-й шаг - выдача на печать матрицы собственных значений Λ;

17-й шаг - выдача на печать матрицы собственных векторов U;

18-й шаг - формирование матрицы $A = U\Lambda^{1/2}$;

19-й шаг - выдача на печать матрицы весовых коэффициентов.

Таким образом, мы рассмотрели основной алгоритм метода главных компонент. В описанном алгоритме большинство шагов очевидно. Пояснения требуются только для двух шагов.

Во-первых, это относится к получению матрицы индивидуальных значений главных компонент F. Для получения матрицы F надо извлечь все n главных компонент и получить квадратную матрицу A порядка n. Тогда в выражении $Y=AF$ окажутся известными две матрицы из трех. Это матрица нормированных значений исходных признаков Y и квадратная матрица A. Умножая слева на A^{-1}, получим

$$A^{-1}Y = A^{-1}AF = F.$$

Во-вторых, при получении по стандартным программам диагональной матрицы собственных значений Λ и ортогональной матрицы U собственных векторов необходимо остановиться на элементах теории линейных операторов, характеристических матрицах и характеристических многочленах.

5.2.3. Линейный оператор и отвечающая ему матрица

А. Дано числовое поле К. Рассмотрим линейное преобразование:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_m = a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \end{cases} . \qquad (5.20)$$

Коэффициенты данного преобразования принадлежат числовому полю K. Рассмотрим также два векторных пространства над этим полем: n-мерное R и m-мерное S. В R выберем некоторый базис $\bar{l}_1, \bar{l}_2, \ldots, \bar{l}_n$. В S - базис $\bar{g}_1, \bar{g}_2, \ldots, \bar{g}_m$.

Итак, дано

1) K - поле;

2) R и S - векторные пространства с базисами соответственно

$$\bar{l}_1, \bar{l}_2, \ldots, \bar{l}_n \quad \text{и} \quad \bar{g}_1, \bar{g}_2, \ldots, \bar{g}_m.$$

Тогда преобразование (5.20) относит каждому вектору из R

$$x = \sum_{i=1}^{n} x_i \bar{l}_i , \qquad (5.21)$$

а вектор y из S

$$y = \sum_{k=1}^{m} y_k \bar{g}_k . \qquad (5.22)$$

Значит (5.20) определяет некоторый оператор A_0, относящий вектору x вектор y. Откуда $y = A_0 x$. Следовательно, преобразование (5.20) при заданных базисах в R и S определяет некоторый линейный оператор, отображающий R в S.

Покажем обратное.

Для произвольного линейного оператора A_0, отображающего R в S, и произвольных базисов $\bar{l}_1, \bar{l}_2, \ldots, \bar{l}_n$ в R и $\bar{g}_1, \bar{g}_2, \ldots, \bar{g}_m$ в S существует такая прямоугольная матрица с элементами из поля K

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}, \qquad (5.23)$$

что составленное при помощи этой матрицы линейное преобразование (5.20) выражает координаты преобразованного вектора

$$y = A_0 x \qquad (5.24)$$

через координаты исходного вектора x.

Доказательство.

1. Применим оператор A_0 к базисному вектору $\bar{l}_k$ и координаты полученного вектора $A_0\bar{l}_k$ в базисе $\bar{g}_1, \bar{g}_2, \ldots, \bar{g}_m$ обозначим $a_{1k}, a_{2k}, \ldots, a_{mk}$, т.е. при $k = \overline{1,n}$

$$A_0\bar{l}_k = \sum_{i=1}^{m} a_{ik}\bar{g}_i . \qquad (5.25)$$

Надо показать $y_i = \sum_{k=1}^{n} a_{ik}x_k \quad (i = \overline{1,m})$.

2. Умножим на x_k (5.25) обе части равенства и просуммируем от 1 до n:

$$\sum_{k=1}^{n} x_k A_0 \bar{l}_k = \sum_{i=1}^{m} (\sum_{k=1}^{n} a_{ik} x_k) \bar{g}_i .$$

3. Откуда

$$y = A_0 x = A_0 (\sum_{k=1}^{n} x_k \bar{l}_k) = \sum_{k=1}^{n} x_k A_0 \bar{l}_k = \sum_{i=1}^{m} (\sum_{k=1}^{n} a_{ik} x_k) \bar{g}_i = \sum_{i=1}^{m} y_i \bar{g}_i ,$$

где $y_i = \sum_{k=1}^{n} a_{ik} x_k \quad (i = \overline{1,m})$, что и требовалось доказать.

Вывод. При заданных базисах R и S каждому линейному оператору A_0, отображающему R и S, отвечает некоторая прямоугольная матрица (5.23) с размерами m x n и, наоборот, каждой такой матрице отвечает некоторый линейный оператор, отображающий R в S.

При этом в матрице A, отвечающей оператору A_0, k-й столбец состоит из последовательных координат вектора $A_0 \bar{l}_k \ (k = \overline{1,n})$. Обозначим столбцы координат векторов x и y

$$x = (x_1 x_2 \ldots x_n)^T ;$$

$$y = (y_1 y_2 \ldots y_n)^T .$$

Тогда векторному равенству

$$y = A_0 x \tag{5.26}$$

соответствует матричное равенство

$$y = Ax , \tag{5.27}$$

которое является матричной записью преобразования (5.20).

Б. Характеристические числа и собственные векторы линейного оператора.

Дано :

1) K - поле;

2) над полем пространство R;

3) вектор $x \neq 0$;

4) $\lambda \in K$.

Исследуем структуру линейного оператора A_0 в пространстве R. При этом большую роль играют векторы x, для которых

$$A_0 x = \lambda x \qquad (\lambda \in K ; \ x \neq 0), \tag{5.28}$$

где x - собственный вектор линейного оператора A_0;

λ - собственное число линейного оператора A_0.

Необходимо определить характеристические числа и собственные векторы оператора A_0.

1. Выберем произвольно базис $\bar{l}_1, \bar{l}_2, \ldots, \bar{l}_n$ в R.

2. Пусть $x = \sum_{i=1}^{n} \bar{l}_i$, $A = (a_{ik})$,

где A - квадратная матрица порядка n, отвечающая оператору A_0, в базисе $\bar{l}_1, \bar{l}_2, \ldots, \bar{l}_n$.

3. Приравниваем между собой соответствующие координаты векторов, стоящих в левых и правых частях (5.28), получим систему скалярных уравнений

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = \lambda x_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = \lambda x_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = \lambda x_n \end{cases} . \tag{5.29}$$

Перепишем систему (5.29).

$$\begin{cases} (a_{11} - \lambda)x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = 0 \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + \ldots + a_{2n}x_n = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + (a_{nn} - \lambda)x_n = 0 \end{cases} . \tag{5.30}$$

По условию искомый вектор не должен быть равен нулю, поэтому среди его координат $x_1, x_2, \ldots, x_n$ по крайней мере одна координата должна быть отлична от нуля.

Чтобы система линейных однородных уравнений (5.30) имела ненулевое решение, необходимо и достаточно, чтобы определитель этой системы был равен нулю.

$$\begin{vmatrix} (a_{11} - \lambda) & a_{12} & \ldots & a_{1n} \\ a_{21} & (a_{22} - \lambda) & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n1} & \ldots & (a_{nn} - \lambda) \end{vmatrix} = 0 . \tag{5.31}$$

Уравнение (5.31) представляет собой алгебраическое уравнение n-й степени относительно λ. Коэффициент этого уравнения принадлежит тому же числовому полю, что и элементы матрицы $A = (a_{ik})$, т.е. полю K.

Уравнение (5.31) часто применяется в геометрии, механике, астрономии, физике. Его называют как:

а) вековое уравнение матрицы A;

б) характеристическое уравнение матрицы A.

Кроме того, левая часть уравнения (5.31) называется характеристический многочлен:

$$|A-\lambda E|=(-\lambda)^n+S_1(-\lambda)^{n-1}+S_2(-\lambda)^{n-2}+\ldots+S_{n-1}(-\lambda)+S_n=0, \tag{5.32}$$

где $S_1=\sum_{i=1}^{n}a_{ii}$ матрицы A;

S_r - сумма миноров r-го порядка $(r=\overline{1,n})$ $S_n=|A|$.

Выводы.

1. Каждое характеристическое число линейного оператора A_0 является корнем характеристического уравнения (5.31).

2. Наоборот, если некоторое число λ является корнем уравнения (5.31), то при этом значении λ система (5.30) и, следовательно, (5.29) имеет ненулевое решение $x_1,x_2,\ldots,x_n$, т.е. числу λ отвечает собственный вектор x оператора A_0:

$$x=\sum_{i=1}^{n}x_i\bar{l}_i.$$

Итак, любой линейный оператор A_0 в R имеет не более n различных характеристических чисел. Если K - есть произвольное алгебраически замкнутое поле, т.е. такое поле, которому принадлежат корни всех алгебраических уравнений с коэффициентами из данного поля, то любой линейный оператор в R всегда имеет по крайней мере один собственный вектор в R и соответствующее этому собственному вектору характеристическое число λ.

5.2.4. Характеристический многочлен подобных матриц

Рассмотрим два базиса в n-мерном пространстве:

$$\bar{l}_1,\bar{l}_2,\ldots,\bar{l}_n;\ (A=(a_{ik});\ y=Ax), \tag{5.33}$$

$$\tilde{l}_1,\tilde{l}_2,\ldots,\tilde{l}_n;\ (A'=(a'_{ik});\ y'=A'x'). \tag{5.34}$$

Запишем в матричной форме формулы преобразования координат:

$$x=Tx'. \tag{5.35}$$

Тогда из (5.33) и (5.35) находим

$$y=Ty'. \tag{5.36}$$

Отсюда

$$y'=T^{-1}y=T^{-1}Ax=T^{-1}ATx'. \tag{5.37}$$

Сопоставляем (5.37) и (5.34). Получим

$$A'=T^{-1}AT.$$

Обозначим

$$B=T^{-1}AT. \tag{5.38}$$

Определение. *Две матрицы A и B, связанные соотношением*

$$B=T^{-1}AT, \tag{5.39}$$

где T - неособенная матрица, называются ***подобными.***

Таким образом, две матрицы, соответствующие одному и тому же линейному оператору в R при различных базисах, подобны между собой. При этом матрица T, связывающая эти матрицы, совпадает с матрицей преобразования координат при переходе от первого базиса l_i ко второму $\tilde{l}_i$; $(i=\overline{1,n})$. Линейному оператору A_0 в R отвечает целый класс подобных между собой матриц. Эти матрицы представляют данный оператор в различных базисах. Отметим, что две подобные матрицы имеют равные определители:

$$|B|=|T|^{-1}|A||T|=|A|. \tag{5.40}$$

Вернемся к характеристическому многочлену подобных матриц.

Отсюда: из $A'=T^{-1}AT$ следует $(A'-\lambda E)=T^{-1}(A-\lambda E)T$.

Следовательно,

$$|A'-\lambda E|=|A-\lambda E|.$$

Значит, подобные матрицы имеют один и тот же характеристический многочлен. Этот многочлен называют характеристическим многочленом оператора A_0 и обозначают:

$$|A_0-\lambda E|. \tag{5.41}$$

5.2.5. Собственные векторы при одном и том же характеристическом числе

Если $x,y,z,\ldots$ - собственные векторы оператора, соответствующие одному и тому же характеристическому числу λ, а $\alpha,\beta,\gamma,\ldots$ - произвольные числа на K, то вектор $\alpha x+\beta y+\gamma z+\ldots$ либо равен нулю, либо также является собственным вектором оператора A_0 при том же числе λ. Действительно, из

$$A_0x=\lambda x,$$
$$A_0y=\lambda y,$$
$$A_0z=\lambda z,$$
$$\ldots\ldots\ldots$$

следует

$$A_0(\alpha x+\beta y+\gamma z+\ldots)=\lambda(\alpha x+\beta y+\gamma z), \tag{5.42}$$

поэтому "линейно независимые" собственные векторы, отвечающие одному и тому же характеристическому числу λ, образуют базис некоторого "собственного" подпространства, каждый вектор которого есть собственный вектор при том же λ. В частности, каждый собственный вектор порождает одномерное собственное подпространство, т.е. образует собственное направление.

Лемма. *Собственные векторы, соответствующие попарно различным характеристическим числам, всегда линейно независимы.*

Из леммы вытекает следующий **вывод:** *если характеристическое уравнение оператора имеет n различных корней и эти корни принадлежат полю K, то на основании леммы собственные векторы, соответствующие этим корням, линейно независимы.*

Для пояснения методов получения собственных значений и собственных векторов необходимо остановиться на характеристических матрицах и характеристических многочленах.

5.2.6. Характеристическая матрица и характеристический многочлен

Определение. *Если задана квадратная матрица* $A=(a_{ik})$ *порядка n, то характеристической матрицей для A называется матрица* $(\lambda E - A)$. *Определитель характеристической матрицы*

$$\Delta(\lambda)=|\lambda E - A|=|\lambda\delta_{ik} - a_{ik}|_1^n \qquad (5.43)$$

является скалярным многочленом относительно λ *и называется* ***характеристическим многочленом матрицы A.***

Примечания: 1. δ_{ik} - символ Кронекера, при $i=k$;

$\delta_{ik}=1$, а при $i\neq k$ $\delta_{ik}=0$.

2. $\Delta(\lambda)$ отличается от $|A-\lambda E|$ на $(-1)^n$.

Рассмотрим матрицу $B(\lambda)=(b_{ik})_1^n$, где $b_{ik}(\lambda)$ - элемент этой матрицы является алгебраическим дополнением элемента $\lambda\delta_{ik}-a_{ki}$ в определителе $\Delta(\lambda)$. Матрица $B(\lambda)$ называется *присоединенной матрицей для матрицы A.*

Пример 5.4

Дана матрица

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

Найти характеристическую матрицу для матрицы A, определитель характеристической матрицы и присоединенную матрицу.

Решение

1) $\lambda E - A=\begin{pmatrix} \lambda - a_{11} & -a_{12} \\ -a_{21} & \lambda - a_{22} \end{pmatrix}$;

2) $\Delta(\lambda)=|\lambda E - A|=\lambda^2 - \lambda(a_{11}+a_{22})+a_{11}a_{22}-a_{12}a_{21}$;

3) $K=1$, $i=1$; $b_{ki}(\lambda)=b_{11}=\lambda - a_{22}$;

$b_{12}=a_{21}$; $b_{21}=a_{12}$; $b_{22}=\lambda - a_{11}$.

$$(b_{ki}(\lambda))=\begin{pmatrix} \lambda - a_{22} & a_{21} \\ a_{12} & \lambda - a_{11} \end{pmatrix}.$$

$$B(\lambda)=(b_{ik}(\lambda))=\begin{pmatrix} \lambda - a_{22} & a_{12} \\ a_{21} & \lambda - a_{11} \end{pmatrix}.$$

Пример 5.5

Дана матрица

$$A=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Определить $(\lambda E - A)$, $\Delta(\lambda)$ и $B(\lambda)$.

Решение

1) $(\lambda E - A)=\begin{pmatrix} (\lambda - a_{11}) & -a_{12} & -a_{13} \\ -a_{21} & (\lambda - a_{22}) & -a_{23} \\ -a_{31} & -a_{32} & (\lambda - a_{33}) \end{pmatrix}$.

2) $\Delta(\lambda)=|\lambda E - A|=\lambda^3 - (a_{11}+a_{22}+a_{33})\lambda^2 - \ldots$.

3) Для получения первого столбца матрицы $B(\lambda)$ найдем алгебраические дополнения элементов первой строки матрицы $(\lambda E - A)$.

а) $b_{11}=\begin{vmatrix} \lambda - a_{22} & -a_{23} \\ -a_{32} & \lambda - a_{33} \end{vmatrix}=\lambda^2 - (a_{33}+a_{22})\lambda + a_{22}a_{33} - a_{23}a_{32}$;

б) $b_{21}=\begin{vmatrix} -a_{21} & -a_{23} \\ -a_{31} & \lambda - a_{33} \end{vmatrix}=a_{21}\lambda - a_{21}a_{33} + a_{23}a_{31}$;

в) $b_{31}=\begin{vmatrix} -a_{21} & \lambda - a_{22} \\ -a_{31} & -a_{32} \end{vmatrix}=a_{31}\lambda + a_{21}a_{32} - a_{22}a_{31}$.

Второй и третий столбцы матрицы предоставим читателю найти самос-

тоятельно. Тогда

$$B(\lambda)=\begin{pmatrix} \lambda^2-(a_{22}+a_{33})\lambda+a_{22}a_{33}-a_{23}a_{32} & * & * \\ a_{21}\lambda+a_{23}a_{31}-a_{21}a_{33} & * & * \\ a_{31}\lambda+a_{21}a_{32}-a_{22}a_{31} & * & * \end{pmatrix},$$

где $\begin{matrix} * & * \\ * & * \end{matrix}$ - элементы второго и третьего столбцов матрицы $B(\lambda)$.

Обратимся к теореме Гамильтона-Кэли.

Теорема Гамильтона-Кэли. *Всякая квадратная матрица удовлетворяет своему характеристическому уравнению, т.е.* $\Delta(A)=0$.

Доказательство. Из определений характеристической, присоединенной и обратной матриц можно записать:

$$(\lambda E-A)^{-1}=\frac{B(\lambda)}{\Delta(\lambda)}; \tag{5.44}$$

$$E=\frac{B(\lambda)}{\Delta(\lambda)}(\lambda E-A)=(\lambda E-A)\frac{B(\lambda)}{\Delta(\lambda)}. \tag{5.45}$$

Следовательно,

$$(\lambda E-A)B(\lambda)=\Delta(\lambda)E, \tag{5.46}$$

$$B(\lambda)(\lambda E-A)=\Delta(\lambda)E. \tag{5.47}$$

Правые части последних равенств можно рассматривать как многочлены с матричными коэффициентами. Каждый из коэффициентов равен произведению скаляра на единичную матрицу E. Присоединенную матрицу $B(\lambda)$ можно представить в виде многочлена, расположенного по степеням λ. Равенства (5.46) и (5.47) показывают, что $\Delta(\lambda)E$ делится слева и справа на $(\lambda E-A)$ без остатка. Согласно обобщенной **теореме Безу** это возможно только тогда, когда остаток

$$\Delta(A)E=\Delta(A)=0. \tag{5.48}$$

Теорема доказана.

Пример 5.6

Дана квадратная матрица

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

Показать, что: 1) она удовлетворяет своему характеристическому уравнению; 2) $\Delta(A)=0$.

Решение

1) $\Delta(\lambda)=\begin{vmatrix} \lambda-2 & -1 \\ 1 & \lambda-3 \end{vmatrix}=\lambda^2-5\lambda+7$.

2) $\Delta(A)=A^2-5A+7E=\begin{pmatrix} 3 & 5 \\ -5 & 8 \end{pmatrix}-5\begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix}+7\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}=\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}=0$.

5.2.7. Выражение присоединенной матрицы B(λ) через характеристический многочлен Δ(λ).

Представим два характеристических многочлена $\Delta(\lambda)$ и $\Delta(\mu)$:

$$\Delta(\lambda)=\lambda^n-p_1\lambda^{n-1}-p_2\lambda^{n-2}-\ldots-p_n; \tag{5.49}$$

$$\Delta(\mu)=\mu^n-p_1\mu^{n-1}-p_2\mu^{n-2}-\ldots-p_n. \tag{5.50}$$

Разность $\Delta(\lambda)-\Delta(\mu)$ делится без остатка на $\lambda-\mu$.

Вспомним формулу сокращенного деления многочлена:

$$\frac{x^n-y^n}{x-y}=x^{n-1}+x^{n-2}y+x^{n-3}y^2+\ldots+xy^{n-2}+y^{n-1}. \tag{5.51}$$

Обозначим $\delta(\lambda,\mu)$ разность характеристических многочленов, деленную на $\lambda-\mu$:

$$\delta(\lambda,\mu)=\frac{\Delta(\lambda)-\Delta(\mu)}{\lambda-\mu}=\lambda^{n-1}+(\mu-p_1)\lambda^{n-2}+ \tag{5.52}$$

$$+(\mu^2-p_1\mu-p_2)\lambda^{n-3}+\ldots$$

Значит, $\delta(\lambda,\mu)$ - многочлен относительно λ и μ.

Получим тождество

$$\Delta(\lambda)-\Delta(\mu)=\delta(\lambda,\mu)(\lambda-\mu). \tag{5.53}$$

Это тождество не нарушится, если вместо λ и μ подставить перестановочные между собой матрицы: вместо λ матрицу λE и вместо μ матрицу A.

$$\Delta(\lambda)E-\Delta(A)E=\delta(\lambda E,A)(\lambda E-A). \tag{5.54}$$

По теореме Гамильтона-Кэли $\Delta(A)=0$, так как деление у нас будет без остатка. В результате

$$\Delta(\lambda)E=\delta(\lambda E,A)(\lambda E-A). \tag{5.55}$$

Сопоставим (5.55) с (5.47) и из однозначности частного получаем

$$B(\lambda)=\delta(\lambda E,A). \tag{5.56}$$

Отсюда согласно (5.52)

$$B(\lambda) = \lambda^{n-1} + B_1\lambda^{n-2} + B_2\lambda^{n-3} + \ldots + B_{n-1}, \qquad (5.57)$$

где

$$\begin{cases} B_1 = A - p_1 E, \\ B_2 = A^2 - p_1 A - p_2 E, \\ \ldots \\ B_k = A^k - p_1 A^{k-1} - p_2 A^{k-2} - \ldots - p_k E \end{cases}$$

$$(k = 1, 2, \ldots (n-1))$$

Матрицы B_k можно вычислять из рекуррентного соотношения:

$$B_k = AB_{k-1} - p_k E \qquad (5.58)$$

$$(k = \overline{1,(n-1)}; B_0 = E).$$

При этом

$$AB_{n-1} - P_n E = 0. \qquad (5.59)$$

Если A - неособенная матрица, то

$$P_n = (-1)^{n-1}|A| \neq 0$$

и из (5.59) следует $A^{-1}AB_{n-1} = A^{-1}P_n E$;

$$A^{-1} = \frac{B_{n-1}}{P_n}. \qquad (5.60)$$

Пусть λ_0 - характеристическое число матрицы A, т.е. $\Delta(\lambda_0) = 0$. Подставив в (5.46) вместо λ значения λ_0, получим:

$$(\lambda_0 E - A)B(\lambda_0) = 0.$$

Пусть $B(\lambda_0 \neq 0)$ и b - любой ненулевой столбец матрицы $B(\lambda_0)$. Тогда

$$(\lambda_0 E - A)b = 0$$

или

$$Ab = \lambda_0 b. \qquad (5.61)$$

Значит, любой ненулевой столбец матрицы $B(\lambda_0)$ определяет собственный вектор, соответствующий характеристическому числу λ_0.

Итак:

1) если коэффициенты характеристического многочлена известны, то присоединенная матрица может быть найдена по формуле (5.56);

2) если A - неособенная матрица, то по формуле (5.60) определяется обратная матрица A^{-1};

3) если λ_0 - характеристическое число матрицы A, то ненулевые столбцы матрицы $B(\lambda_0)$ являются собственными векторами матрицы A для $\lambda = \lambda_0$.

Пример 5.7

Дана матрица

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

Определить: $\Delta(\lambda)$, $\delta(\lambda,\mu)$, $B(\lambda)$ и A^{-1}.

1) $\Delta(\lambda) = \lambda^3 - 4\lambda^2 + 5\lambda - 2$;

2) $\delta(\lambda,\mu) = \dfrac{\Delta(\lambda) - \Delta(\mu)}{\lambda - \mu} = \lambda^2 + \lambda(\mu - 4) + \mu^2 - 4\mu + 5$;

3) $B(\lambda) = \delta(\lambda E, A) = \lambda^2 E + \lambda(A - 4E) + A^2 - 4A + 5E$;

$$B_1 = A - 4E = \begin{pmatrix} -2 & -1 & 1 \\ 0 & -3 & 1 \\ -1 & 1 & -3 \end{pmatrix};$$

$$B_2 = AB_1 + 5E = \begin{pmatrix} 0 & 2 & -2 \\ -1 & 3 & -2 \\ 1 & -1 & 2 \end{pmatrix};$$

$$B(\lambda) = \begin{pmatrix} \lambda^2 - 2\lambda & -\lambda + 2 & \lambda - 2 \\ -1 & \lambda^2 - 3\lambda + 3 & \lambda - 2 \\ -\lambda + 1 & \lambda - 1 & \lambda^2 - 3\lambda + 2 \end{pmatrix};$$

$$A^{-1} = \frac{1}{2}B_2 = \begin{pmatrix} 0 & 1 & -1 \\ -\frac{1}{2} & \frac{3}{2} & -1 \\ \frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix};$$

$|A| = 2$.

5.2.8. Метод Фаддеева - одновременное вычисление коэффициентов характеристического многочлена и присоединенной матрицы

При помощи метода Фаддеева одновременно определяются:

а) $P_1, P_2, \ldots, P_n$ - скалярные коэффициенты характеристического многочлена

$$\Delta(\lambda) = \lambda^n - P_1\lambda^{n-1} - \ldots - P_n; \qquad (5.62)$$

б) $B_1, B_2, \ldots, B_{n-1}$ - матричные коэффициенты присоединенной матрицы.

При помощи $tr\,A$ следа матрицы получаем

$$trA = \sum_{i=1}^{n} a_{ii} = P_1 = \sum_{i=1}^{n} \lambda_i,$$

если $\lambda_1, \lambda_2, \ldots, \lambda_n$ - характеристические числа матрицы A, т.е. $\Delta(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)\ldots(\lambda - \lambda_n)$.

Теорема. *Если $\lambda_1, \lambda_2, \ldots, \lambda_n$ - все характеристические числа (с учетом кратностей) матрицы A, а $g(\lambda)$ - некоторый скалярный многочлен, то $g(\lambda_1), g(\lambda_2), \ldots, g(\lambda_n)$ - являются характеристическими числами матрицы $g(A)$.*

Частный случай. Дана матрица A; $\lambda_1, \lambda_2, \ldots, \lambda_n$ - ее характеристические числа. Определить характеристические числа матрицы A^k.

В соответствии с теоремой $g(A) = A^k$.

Поэтому $g(\lambda_1) = \lambda_1^k$, $g(\lambda_2) = \lambda_2^k, \ldots, g(\lambda_n) = \lambda_n^k$. $(k = 0,1,2,\ldots)$.

Отсюда следует, что $trA^k = S_k = \sum_{i=1}^{n} \lambda_i^k$, ($k$=0,1,2, ...).

Суммы S_k (k=1,2, ... ,n) степеней корней многочлена (5.62) связаны с коэффициентами этого уравнения формулами Ньютона.

$$KP_k = S_k - P_1S_{k-1} - \ldots - P_{k-1}S_1, \ (k = \overline{1,n}). \qquad (5.63)$$

Метод Леверрье. Определение коэффициентов характеристического многочлена по следам степеней матрицы заключается в следующем:

1) определяются $S_1, S_2, \ldots, S_n$ - следы матрицы $A, A^2, \ldots, A^n$.

2) по (5.63) последовательно определяются $P_1, P_2, \ldots, P_n$.

Метод Фаддеева

Фаддеев предложил вместо следов степеней матриц $A, A^2, \ldots, A^n$ вычислять последовательно следы других матриц $A_1, A_2, \ldots, A_n$ и с их помощью определять $P_1, P_2, \ldots, P_n$ и $B_1, B_2, \ldots, B_n$.

$$\begin{array}{lll} A_1 = A; & P_1 = tr(A_1); & B_1 = A_1 - P_1E; \\ A_2 = AB_1; & P_2 = tr(A_2); & B_2 = A_2 - P_2E; \\ \ldots & \ldots & \ldots \\ A_{n-1} = AB_{n-2}; & P_{n-1} = \dfrac{1}{n-1}tr(A_{n-1}); & B_{n-1} = A_{n-1} - P_{n-1}E; \\ A_n = AB_{n-1}; & P_n = \dfrac{1}{n}tr(A_n); & B_n = A_n - P_nE = 0. \end{array} \qquad (5.64)$$

Для контроля вычислений можно воспользоваться последней формулой (B_n=0). Убедимся, что по системе (5.64) $P_1, P_2, \ldots, P_n$; $B_1, B_2, \ldots, B_{n-1}$ последовательно определяемые, являются коэффициентами $\Delta(\lambda)$ и $B(\lambda)$.

Используя систему (5.64) для A_k и B_k, ($k = \overline{1,n}$) получим:

$$A_k = A^k - P_1A^{k-1} - \ldots - P_{k-1}A; \qquad (5.65)$$

$$B_k = A^k - P_1A^{k-1} - \ldots - P_{k-1}A - P_kE. \qquad (5.66)$$

Приравняем следы левой и правой частей (5.65)

$$KP_k = S_k - P_1S_{k-1} - \ldots - P_{k-1}S_1. \qquad (5.67)$$

Выражения (5.67) и (5.63) совпадают с формулами Ньютона, по которым последовательно определяются коэффициенты характеристического многочлена $\Delta(\lambda)$. Значит, числа $P_1, P_2, \ldots, P_n$ системы (5.64) являются коэффициентами $\Delta(\lambda)$.

По формуле (5.66) определяют матричные коэффициенты $B_1, B_2, \ldots, B_{n-1}$ присоединенной матрицы $B(\lambda)$.

Значит, система (5.64) определяет коэффициенты $B_1, B_2, \ldots, B_n$ матричного многочлена $B(\lambda)$.

Пример 5.8

Вычислить собственные значения и собственные векторы матрицы

$$A = \begin{pmatrix} 7 & -2 & 0 \\ -2 & 6 & -2 \\ 0 & -2 & 5 \end{pmatrix}.$$

Решение методом Фаддеева

1. Составим характеристическое уравнение

$$\Delta(\lambda) = |\lambda E - A| = \begin{vmatrix} \lambda - 7 & 2 & 0 \\ 2 & \lambda - 6 & 2 \\ 0 & 2 & \lambda - 5 \end{vmatrix}$$

2. Запишем его в виде многочлена 3-й степени относительно

$$\Delta(\lambda) = \lambda^3 - P_1\lambda^2 - P_2\lambda - P_3.$$

3. $P_1 = tr(A) = 7 + 6 + 5 = 18$.

4. $P_2 = \dfrac{1}{2}tr(A_2)$;

$A_2 = AB_1$;

$B_1 = A - P_1E$;

$$B_1 = \begin{pmatrix} 7-18 & -2 & 0 \\ 2 & 6-18 & -2 \\ 0 & -2 & 5-18 \end{pmatrix} = \begin{pmatrix} -11 & -2 & 0 \\ -2 & -12 & -2 \\ 0 & -2 & -13 \end{pmatrix};$$

$$A_2 = AB_1 = \begin{pmatrix} 7 & -2 & 0 \\ -2 & 6 & -2 \\ 0 & -2 & 5 \end{pmatrix}\begin{pmatrix} -11 & -2 & 0 \\ -2 & -12 & -2 \\ 0 & -2 & -13 \end{pmatrix} =$$

$$= \begin{pmatrix} -77+4 & -14+24 & 4 \\ 22-12 & 4-72+4 & 0-12+26 \\ 4 & 24-10 & 4-65 \end{pmatrix} = \begin{pmatrix} -73 & 10 & 4 \\ 10 & -64 & 14 \\ 4 & 14 & -61 \end{pmatrix}.$$

5. $P_3 = |A|$;

$$|A| = -(-2)\begin{vmatrix} 7 & 0 \\ -2 & -2 \end{vmatrix} + 5\begin{vmatrix} 7 & -2 \\ -2 & 6 \end{vmatrix} = -28 + 190 = 162.$$

6. Получены все члены характеристического уравнения

$$\lambda^3 - 18\lambda^2 + 99\lambda - 162 = 0.$$

7. Определим корни характеристического уравнения

$$\lambda_3 = 3; \quad \lambda_2 = 6; \quad \lambda_1 = 9; \qquad \Lambda = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 9 \end{pmatrix}.$$

8. Определим собственный вектор, соответствующий $\lambda_1 = 9$; подставим в систему уравнений $\lambda = 9$.

$$\begin{cases} (9-7)x_1 + 2x_2 = 0 \\ 2x_1 + (9-6)x_2 + 2x_3 = 0. \\ 2x_2 + (9-5)x_3 = 0 \end{cases}$$

Система однородная, все b_i, т.е. определители, равны 0. Система неполная (уравнения зависимы) и имеет бесконечное множество решений. Одно решение может быть выбрано произвольно. В этом случае можно определить отношение корней:

$$x_1 : x_2 : \ldots : x_n = A_{i1} : A_{i2} : \ldots : A_{in},$$

где A_{ij} - алгебраические дополнения элементов любой строки.

Решение этой системы уравнений позволяет определить следующие соотношения:

$$x_1 : x_2 : x_3 = 2 : (-2) : 1.$$

Значит, собственный вектор

$$x_1 = C\begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}.$$

9. Определим собственный вектор, соответствующий $\lambda_2 = 6$.

$$\begin{cases} (6-7)x_1 + 2x_2 = 0 \\ 2x_1 + (6-6)x_2 + 2x_3 = 0; \\ 2x_2 + (6-5)x_3 = 0 \end{cases}$$

$$x_1 : x_2 : x_3 = (-2) : (-1) : 2;$$

$$x_2 = C\begin{pmatrix} -2 \\ -1 \\ 2 \end{pmatrix}.$$

10. Определим собственный вектор, соответствующий $\lambda_3 = 3$.

$$\begin{cases} (3-7)x_1 + 2x_2 = 0 \\ 2x_1 + (3-6)x_2 + 2x_3 = 0; \\ 2x_2 + (3-5)x_3 = 0 \end{cases}$$

$$x_1 : x_2 : x_3 = 1 : 2 : 2;$$

$$x_3 = C\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}.$$

Решение при помощи присоединенной матрицы

$$B(\lambda) = \begin{pmatrix} \lambda^2 - 11\lambda + 26 & 10 - 2\lambda & 4 \\ 10 - 2\lambda & \lambda^2 - 12\lambda + 35 & 14 - 2\lambda \\ 4 & 14 - 2\lambda & \lambda^2 - 13\lambda + 38 \end{pmatrix}.$$

Найти x_1.

1. Возьмем первый столбец

$$B_1(9) = C\begin{pmatrix} 8 \\ -8 \\ 4 \end{pmatrix} = 4\begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}; \quad x_1 = C\begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}.$$

Вектор $\begin{pmatrix} 8 \\ -8 \\ 4 \end{pmatrix}$ получен из $B(\lambda)$.

Это тоже собственный вектор, так как у вектора собственного $x = C\begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}$ есть целое семейство векторов, элементы которых пропорциональны соответственным элементам вектора $x = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}$, поэтому везде введены коэффициенты C. Обычно, когда говорят о собственном векторе, то речь идет о векторе, у которого элементы сокращены на их наибольший общий делитель.

2. Найдем x_2 при $\lambda_2 = 6$. Подставим в первый столбец

$$B_1(6) = \begin{pmatrix} -4 \\ -2 \\ 4 \end{pmatrix} = 2\begin{pmatrix} -2 \\ -1 \\ 2 \end{pmatrix}; \qquad x_2 = C\begin{pmatrix} -2 \\ -1 \\ 2 \end{pmatrix}.$$

3. Определим x_3 при $\lambda_3 = 3$.

$$B_1(3) = \begin{pmatrix} 2 \\ 4 \\ 4 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}; \qquad x_3 = C\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}.$$

5.3. Квадратичные формы и главные компоненты

Для того чтобы представить в геометрическом плане главные компоненты, рассмотрим простейшие случаи: плоскости и пространства трех измерений.

Пусть дано уравнение линии второго порядка:

$$Ax^2 + 2Bxy + Cy^2 = H. \qquad (5.68)$$

Левая часть уравнения (5.68) не меняется при замене x, y на $-x$, $-y$. Значит, *во-первых*, точки линии (5.68) расположены парами симметрично относительно начала координат. *Во-вторых*, линия второго порядка, заданная (5.68), обладает центром симметрии и, *в-третьих*, начало координат помещено в центр. Левая часть (5.68) представляет собой однородный многочлен второй степени. Такой многочлен называют **квадратичной формой от двух переменных.**

$$Ax^2 + 2Bxy + Cy^2. \qquad (5.69)$$

Приведем данную квадратичную форму (5.69) к каноническому виду. Для этого надо будет повернуть так координатные оси x и y, чтобы в новых

координатах исчез член с произведением новых текущих координат. Переход к новым координатам производится по известным формулам:

$$\begin{cases} x' = l_1 x + m_1 y \\ y' = l_2 x + m_2 y \end{cases}. \qquad (5.70)$$

Старые координаты связаны с новыми по формулам:

$$\begin{cases} x = l_1 x' + l_2 y' \\ y = m_1 x' + m_2 y' \end{cases}, \qquad (5.71)$$

где x' и y' - новые координаты.

Характеристика коэффициентов со старыми координатами представлена на рис.5.1.

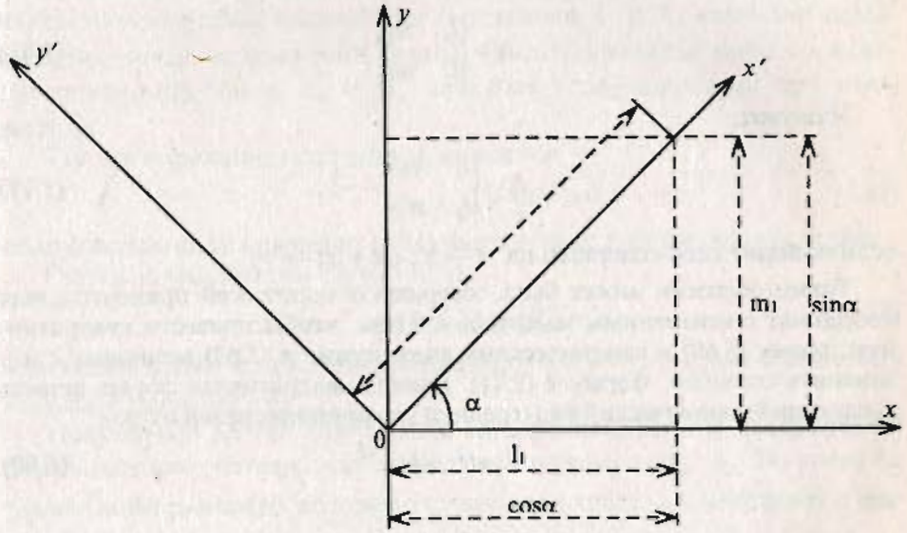


Рис. 5.1. Единичный вектор и его компоненты

На рис. 5.1 на новой оси абсцисс отложен отрезок OX_1 единичной длины, тогда его проекции на старые координатные оси составят:

$$\begin{cases} l_1 = \cos\alpha \\ m_1 = \sin\alpha \end{cases}, \qquad (5.72)$$

где α - угол поворота осей x и y.

Значит, вектор с компонентами l_1 и m_1 является единичным вектором, определяющим направление новой оси абсцисс x':

$$\begin{pmatrix} l_1 \\ m_1 \end{pmatrix} = \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix}. \tag{5.73}$$

Аналогично единичный вектор, определяющий направление новой оси y' ординат, имеет вид:

$$\begin{pmatrix} l_2 \\ m_2 \end{pmatrix} = \begin{pmatrix} -\sin\alpha \\ \cos\alpha \end{pmatrix}. \tag{5.74}$$

Рассматриваемые коэффициенты обладают следующими свойствами:

$$l_1^2 + m_1^2 = 1; \tag{5.75}$$

$$l_2^2 + m_2^2 = 1; \tag{5.76}$$

$$l_1 l_2 + m_1 m_2 = 0; \tag{5.77}$$

$$\Delta = \begin{vmatrix} l_1 & m_1 \\ l_2 & m_2 \end{vmatrix} = 1. \tag{5.78}$$

Наконец,

$$\Delta = \begin{vmatrix} l_1 & m_1 \\ l_2 & m_2 \end{vmatrix} = -1, \tag{5.79}$$

если поворот осей совершен на $\alpha \to x$; $(\alpha + \pi) \to y$.

Таким образом, может быть совершен поворот осей прямоугольных координат с неизменным масштабом. Итак, чтобы привести квадратичную форму (5.69) к каноническому виду, нужно в (5.69) величины x и y заменить согласно формуле (5.71). Данная квадратичная форма примет следующий канонический вид (средний коэффициент равен нулю):

$$\lambda_1 x'^2 + \lambda_2 y'^2, \tag{5.80}$$

т.е.

$$Ax^2 + 2Bxy + Cy^2 = \lambda_1 x'^2 + \lambda_2 y'^2. \tag{5.81}$$

Для решения (5.81) достаточно подобрать так коэффициенты (5.71) и числа λ_1, λ_2, чтобы

$$\begin{cases} Al_1 + Bm_1 = \lambda_1 l_1 \\ Bl_1 + Cm_1 = \lambda_1 m_1 \end{cases}; \qquad \begin{cases} Al_2 + Bm_2 = \lambda_2 l_2 \\ Bl_2 + Cm_2 = \lambda_2 m_2 \end{cases}.$$

Значит, надо решить систему уравнений

$$\begin{cases} Al + Bm = \lambda l \\ Bl + Cm = \lambda m \end{cases}. \tag{5.82}$$

В системе (5.82) перенесем правые части влево и получим

$$\begin{cases} (A - \lambda)l + Bm = 0 \\ Bl + (C - \lambda)m = 0 \end{cases}. \tag{5.83}$$

Определитель данной системы

$$\begin{vmatrix} A - \lambda & B \\ B & C - \lambda \end{vmatrix} = 0 \tag{5.84}$$

можно представить в виде

$$\lambda^2 - (A + C)\lambda + (AC - B^2) = 0. \tag{5.85}$$

Откуда

$$\lambda_{1,2} = \frac{1}{2}[(A + C) \pm \sqrt{(A + C)^2 - 4(AC - B^2)}]. \tag{5.86}$$

Уравнение (5.84) представляет собой характеристическое уравнение квадратичной формы, а корни этого уравнения λ_1 и λ_2 являются характеристическими числами этой формы. После приведения формы к каноническому виду числа λ_1 и λ_2 являются коэффициентами при неизвестных.

Так как выражение под радикалом, равное

$$(A - C)^2 + 4B^2 \geq 0, \tag{5.87}$$

неотрицательно, то уравнение (5.84) имеет только действительные корни.

Отдельно рассмотрим случай, когда

$$(A - C)^2 + 4B^2 > 0. \tag{5.88}$$

При этом условии $\lambda_1 \neq \lambda_2$. Подставим в (5.83) $\lambda = \lambda_1$. Система будет иметь ненулевое решение l и m.

Полученный вектор будет иметь главное направление квадратичной формы, которое соответствует характеристическому числу λ_1. По этому же главному направлению, которое соответствует числу λ_1, направлен и вектор $\begin{pmatrix} l_1 \\ m_1 \end{pmatrix}$, т.е.

$$\begin{cases} l_1 = \mu l \\ m_1 = \mu m \end{cases}, \tag{5.89}$$

где $\mu \neq 0$.

Если примем, что $\sqrt{l_1^2 + m_1^2} = \pm 1$, то по системе (5.89)

$$l_1^2 + m_1^2 = 1.$$

Вектор $\begin{pmatrix} l_1 \\ m_1 \end{pmatrix}$ является единичным вектором главного направления.

Естественно, что вектор $\begin{pmatrix} l_2 \\ m_2 \end{pmatrix}$ определяет другое главное направление квадратичной формы.

Согласно выражению (5.77), если $\lambda_1 \neq \lambda_2$, векторы главных направлений взаимно перпендикулярны.

Другой случай соответствует

$$(A-C)^2 + 4B^2 = 0. \tag{5.90}$$

В данном случае

$$\begin{cases} \lambda_1 = \lambda_2 = \lambda \\ A = C \\ B = 0 \end{cases}. \tag{5.91}$$

Из выражения (5.86) $\lambda = A = C$.

Подставим в выражение (5.85) полученное значение λ и убедимся в том, что все коэффициенты системы обращаются в нуль. Таким образом, система (5.83) будет состоять из тождеств. Ей подходят любые числа l и m.

В результате можно заключить, что если $\lambda = \lambda_1$, то для квадратичной формы любое направление является главным. При повороте осей на любой угол форма сохранит свой канонический вид

$$Ax^2 + Ay^2.$$

При любом преобразовании квадратичной формы к любым прямоугольным координатам не меняются ее инварианты

$$\begin{cases} A + C = A' + C' \\ AC - B^2 = A'C' - B'^2 \end{cases}. \tag{5.92}$$

Согласно ***теореме Виета***

$$AC - B^2 = \lambda_1 \lambda_2. \tag{5.93}$$

1. Если $\lambda_1 \neq 0$; $\lambda_2 \neq 0$ имеют одинаковые знаки, то квадратичная форма называется **эллиптической:**

$$AC - B^2 > 0. \tag{5.94}$$

2. Если $\lambda_1 \neq 0$; $\lambda_2 \neq 0$, но знаки у них разные, то форма называется **гиперболической:**

$$AC - B^2 < 0. \tag{5.95}$$

3. Если одно из чисел λ_1, λ_2 равно нулю, т.е. $AC - B^2 = 0$, то форма называется **параболической.**

В методе главных компонент характеристические числа по своему физическому смыслу не могут равняться нулю и быть отрицательными. Значит, $\lambda_1 > 0$ и $\lambda_2 > 0$. В этом случае квадратичная форма будет называться **положительно определенной эллиптической формой.**

На рис.5.2 показаны переход от произвольной системы координат к системе с точкой нуль в центре эллипса и поворот осей, осуществленный для приведения квадратичной формы к каноническому виду. После приведения к каноническому виду ось абсцисс, соответствующая λ_1, направлена по одной главной оси эллипса (главному направлению), а ось координат, соответствующая другому главному направлению, направлена перпендикулярно к ней вдоль другой главной оси эллипса. Вдоль главной оси эллипса *oy* направлена первая главная компонента, а вдоль оси *ox* направлена вторая главная компонента.

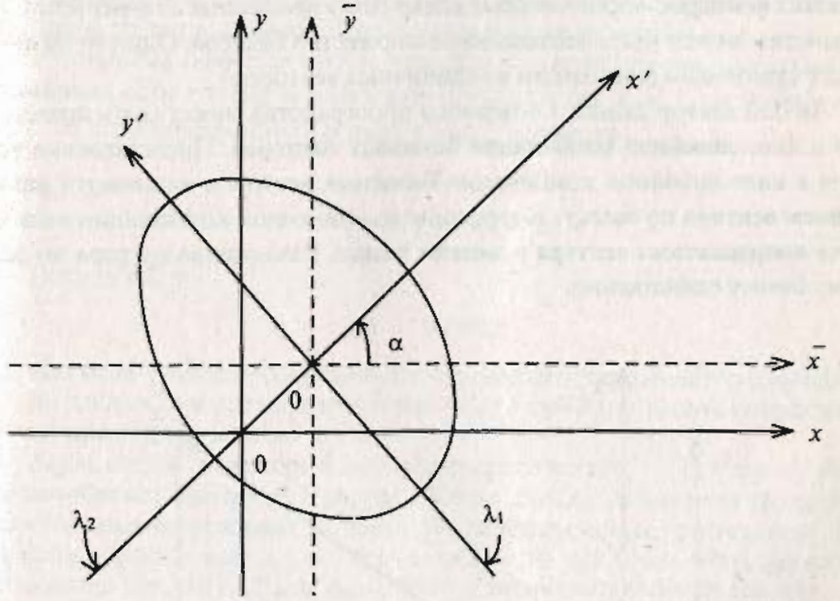


Рис. 5.2. Перенос системы координат $(x,0,y)$ в центр эллипса $(\bar{x}, 0, \bar{y})$ и поворот на угол α

На рис.5.2 первое главное направление (y') определяется λ_1, а второе главное направление (x') определяется характеристическим числом λ_2.

5.3.1. Главные компоненты трехмерного и конечно-мерного пространства

Введем некоторые известные определения.

Множество всех n-мерных векторов, для которых определены операции сложения и умножения вектора на число, называется ***n*-мерным векторным пространством.**

Трехмерное векторное пространство можно представить как частный случай n-мерного векторного пространства (при n=3). Определения и свойства векторов в трехмерном и n-мерном пространстве совпадают.

В n-мерном пространстве любая система, содержащая более чем n векторов, линейно зависима.

Наибольшее число линейно независимых векторов системы называется ее рангом. Ранг n-мерного пространства совпадает с его размерностью. *Базисом множества векторов называется любой набор линейно независимых векторов, число которых равно рангу множеств.* В n-мерном пространстве может быть бесчисленное множество базисов. Один из базисов будет единичным (состоящим из единичных векторов).

Любой вектор данного n-мерного пространства может быть представлен в виде **линейной комбинации базисных векторов.** Представление вектора в виде линейной комбинации базисных векторов называется **разложением вектора по базису.** Коэффициенты линейной комбинации называются **координатами вектора в данном базисе.** Разложение вектора по данному базису единственно.

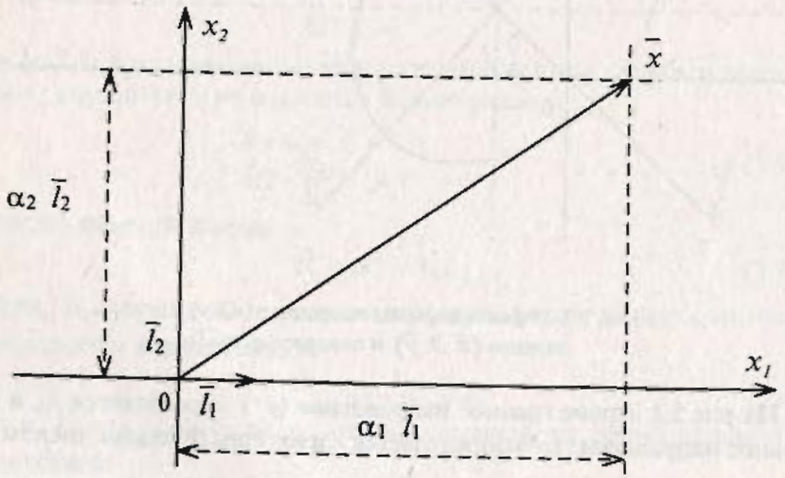


Рис. 5.3. Разложение вектора X по векторам ортогонального двумерного базиса

Рассмотрим случай, чаще всего используемый в методе главных компонент (рис.5.3).

Вектор x разложим по векторам ортогонального двумерного базиса:

$$x = \alpha_1 \bar{l}_1 + \alpha_2 \bar{l}_2 , \qquad (5.96)$$

где α_1 и α_2 - произвольные числа;

$\bar{l}_1$ и $\bar{l}_2$ - единичные векторы;

$\alpha_1 l_1$ и $\alpha_2 l_2$ - координаты вектора x.

Ознакомившись с основными определениями конечно-мерного пространства, остановимся на линейных преобразованиях.

Преобразование линейного пространства R представляет собой закон, ставящий каждому вектору из R в соответствие снова некоторый вполне определенный вектор из R. Если при данном преобразовании произведение числа на вектор переводится в произведение того же числа на собственный вектор, а сумма векторов переводится в сумму соответственных векторов, то данное преобразование называется линейным.

В принятых символах можно записать, что преобразование называется **линейным**, если для любых векторов пространства $\bar{x}_1$ и $\bar{x}_2$ и любого числа α из поля коэффициентов имеют место равенства:

$$(\alpha x)A = \alpha(xA); \qquad (5.97)$$

$$(x + y)A = xA + yA . \qquad (5.98)$$

Если α =0, то

$$0A = 0, \qquad (5.99)$$

т.е. всякое линейное преобразование переводит нулевой вектор в нулевой.

Возвратимся к интересующей нас задаче преобразования координат в трехмерном пространстве.

Задан вектор x, который надо перевести в вектор x'. Вектор x' называется **образом вектора x.** Преобразование $x' = Ax$ называется линейным, если соблюдены условия (5.97) и (5.98). Запишем общее уравнение второй степени относительно x, y, z, применяя вместо заглавных букв для коэффициентов при неизвестных одну букву с двумя числовыми индексами:

$$a_{11}x^2 + a_{22}y^2 + a_{33}z^2 + 2a_{12}xy + 2a_{13}xz + 2a_{23}yz + 2a_{14}x + \\ +2a_{24}y + 2a_{34}z + a_{44} = 0. \qquad (5.100)$$

Из уравнения (5.100) рассмотрим неполное уравнение второй степени: квадратичную форму от трех переменных:

$$a_{11}x^2 + a_{22}y^2 + a_{33}z^2 + 2a_{12}xy + 2a_{13}xz + 2a_{23}yz = H . \qquad (5.101)$$

В нем нет членов первой степени, значит, левая часть не меняется при замене x, y, z на $-x$, $-y$, $-z$. Точки поверхности расположены парами симметрично относительно начала координат. Таким образом, поверхность второго порядка, описываемая равенством (5.101), находится в центре начала координат и обладает центром симметрии. Приведем формулу к каноническому виду, т.е. произведем такое ортогональное преобразование координат, чтобы исчезли все члены с произведением новых текущих координат. Найдем

$$\begin{cases} x = l_1 x' + l_2 y' + l_3 z' \\ y = m_1 x' + m_2 y' + m_3 z' \\ z = n_1 x' + n_2 y' + n_3 z' \end{cases} . \quad (5.102)$$

Из данных формул следует тождество:

$$a_{11}x^2 + a_{22}y^2 + a_{33}z^2 + 2a_{12}xy + 2a_{13}xz + 2a_{23}yz = \\ = \lambda_1 x'^2 + \lambda_2 y'^2 + \lambda_3 z'^2 . \quad (5.103)$$

Решение по схеме для случая двух переменных приведет нас к характеристическому определителю:

$$\begin{vmatrix} (a_{11}-\lambda) & a_{12} & a_{13} \\ a_{21} & (a_{22}-\lambda) & a_{23} \\ a_{31} & a_{32} & (a_{33}-\lambda) \end{vmatrix} = 0 .$$

Отметим, что характеристический многочлен матрицы линейного преобразования A не зависит от выбора базиса. После развертывания определителя получим характеристический многочлен, где наивысшая степень λ будет равна трем. Решив его, найдем характеристические корни данного уравнения.

Нас интересует случай, когда $\lambda_1 > 0$, $\lambda_2 > 0$, $\lambda_3 > 0$. Значит, квадратичная форма данного уравнения будет эллиптической. При этом H>0, и в трехмерном пространстве получим поверхность второго порядка - эллипсоид. У данного эллипсоида три главных направления, соответствующие трем значениям собственных чисел: λ_1, λ_2 и λ_3, которые совпадают с его главными осями.

Если будем осуществлять линейные преобразования в n-мерном пространстве, то получим некоторые гиперповерхности, главные оси которых будут совпадать с главными направлениями после приведения их к каноническому виду. Все главные направления считаем взаимно перпендикулярными.

Всякому собственному числу λ_j соответствует свой собственный вектор, совпадающий с j-м главным направлением.

Таким образом, при решении задач методом главных компонент мы переходим к новой системе взаимно перпендикулярных осей, при которой взаимно перпендикулярные главные направления совпадают соответственно с этой новой системой координат.

5.3.2. Дисперсия исследуемых признаков в методе главных компонент

Главные компоненты ортогональны между собой, поэтому полная дисперсия j-го признака равна:

$$\sigma_j^2 = \sum_{r=1}^{n} a_{jr}^2 = 1 , \quad (5.104)$$

где a_{jr}^2 - доля полной дисперсии j-го признака, вносимая r-й главной компонентой.

Если при помощи n нормированных признаков изучается процесс, то его дисперсия будет равна n (сумме единиц).

На практике не всегда извлекают все главные компоненты. Как правило, извлекают только их часть, равную $m<n$. Оставляют наиболее весомую часть главных компонент; вкладом остальных $(n-m)$ компонент в этом случае сознательно пренебрегают, так как каждая из $(n-m)$ главных компонент вносит малую долю в общую дисперсию изучаемого процесса.

В данном случае вместо

$$y_j = \sum_{r=1}^{n} a_{jr} f_r \quad \text{имеем} \quad y_j = \sum_{r=1}^{m} a_{jr} f_r . \quad (5.105)$$

Полный вклад всех m извлеченных главных компонент в общую дисперсию всех признаков равен

$$V_0 = \sum_{r=1}^{m} \lambda_r , \quad (5.106)$$

а относительный вклад

$$K = \frac{1}{n} V_0 , \quad (5.107)$$

где K - показатель факторизации в методе главных компонент.

5.4. Задачи и упражнения

5.1. Доказать, что $\frac{1}{N} Y^T Y$ является матрицей выборочных коэффициентов корреляции с элементами r_{jl}, где $j, l = 1,2,...,n$, размерность матрицы $Y(N \times n)$.

5.2. Доказать, что след матрицы Λ равен числу исходных показателей n, т.е. $tr\Lambda = \sum_{r=1}^{n} \lambda_r = n$.

5.3. Доказать, что определитель корреляционной матрицы R равен произведению собственных чисел (дисперсий) главных компонент $|R| = \prod_{r=1}^{n} \lambda_r$.

5.4. Показать, что матрица выборочных коэффициентов корреляции R главных компонент определяется матрицей $\frac{1}{N} FF^T$, если размерность $F(n \times N)$.

5.5. Доказать, что матрица факторных нагрузок A обладает следующими свойствами:

а) сумма квадратов элементов n-го столбца матрицы A равна дисперсии r-й главной компоненты, т.е.

$$\sum_{j=1}^{n} a_{jr}^2 = \lambda_r ;$$

б) элемент, стоящий на пересечении j-й строки и r-го столбца матрицы A, есть выборочный коэффициент корреляции между j-м признаком и r-й главной компонентой, т.е. $a_{jr} = r_{x_j f_r}$.

5.6. Доказать, что для матрицы факторных нагрузок A, размерности $(n \times n)$ справедливо равенство $(A^{-1})^T = (A^T)^{-1}$.

5.7. Пусть i-й объект характеризуется n значениями главных компонент $f_{i1}, f_{i2}, \ldots, f_{ir}, \ldots, f_{in}$ и известна матрица факторных нагрузок A с элементами a_{jr}, где $j, r = 1,2,\ldots,n$. Найти выражение для нормированного значения j-го показателя у i-го объекта (y_{ij}).

5.8. Хозяйственная деятельность предприятий приборостроения (N=10) характеризуется следующими показателями (n=2) (табл.5.2).

Таблица 5.2

№ п/п	Общие затраты на 1 руб. продукции, x_1	Фондоотдача, x_2
1	0,92	0,51
2	0,72	0,59
3	0,83	1,03
4	0,81	1,21
5	0,82	0,63
6	0,93	0,68
7	0,84	0,57
8	0,89	1,52
9	0,89	1,04
10	0,95	0,99

Требуется:

а) ранжировать предприятия по первой главной компоненте;

б) графически изобразить предприятия в пространстве двух главных компонент. Дать экономическую интерпретацию полученным результатам.

5.9. Деятельность предприятий приборостроения (N=10) характеризуется следующими показателями (n=2) (табл. 5.3).

Таблица 5.3

№ п/п	Общие затраты на единицу товарной продукции, x_1	Среднегодовая заработная плата на единицу ППП, x_2
1	0,92	2,33
2	0,72	2,31
3	0,83	2,39
4	0,81	2,34
5	0,82	2,38
6	0,93	2,43
7	0,84	2,20
8	0,79	2,33
9	0,89	2,23
10	0,95	2,13

Приняв за результативный показатель y производительность труда работника, построить уравнение регрессии на главной компоненте, наиболее тесно связанной с y, где

y^T=(3,5; 4,4; 3,9; 3,8; 3,6; 3,5; 3,0; 4,1; 4,0; 3,1).

Дать интерпретацию полученным результатам.

5.10. Значения экономических показателей, отражающих работу предприятий приборостроения (N=10), приводятся в табл. 5.4.

Таблица 5.4

№ п/п	Фондоотдача, x_1	Среднегодовая заработная плата на единицу ППП, x_2
1	0,98	2,24
2	0,73	2,29
3	0,73	2,40
4	1,47	2,46
5	0,75	2,58
6	0,81	2,18
7	0,83	2,12
8	0,92	2,10
9	0,46	2,29
10	1,11	2,36

Требуется:

а) определить собственные значения главных компонент и их вклад в суммарную дисперсию исходных показателей;

б) определить ортогональную матрицу собственных векторов корреляционной матрицы R;

в) дать экономическую интерпретацию главных компонент.

5.11. Найдены значения главных компонент для i-го предприятия $f_{i1} = 0,661$; $f_{i2} = -2,151$ и матрица факторных нагрузок

$$A = \begin{pmatrix} -0,756 & 0,654 \\ 0,756 & 0,654 \end{pmatrix}.$$

Требуется найти значения исходных показателей для этого предприятия (x_{i1}, x_{i2}), если выборка характеризуется векторами средних $\bar{x} = \begin{pmatrix} 0,850 \\ 0,877 \end{pmatrix}$ и средних квадратических отклонений $S = \begin{pmatrix} 0,072 \\ 0,333 \end{pmatrix}$.

5.12. В матрице значений главных компонент F оказалось пропущенным последнее десятое наблюдение (i=10)

$$F = \begin{pmatrix} -0,484 & 1,053 \\ 1,227 & -1,572 \\ 0,773 & 0,527 \\ 0,606 & -0,175 \\ 0,799 & 0,314 \\ 0,136 & 2,097 \\ -0,671 & -1,109 \\ 0,720 & -0,507 \\ -0,921 & -0,231 \end{pmatrix}.$$

Матрица факторных нагрузок

$$A=\begin{pmatrix}-0,791 & 0,611\\ 0,791 & 0,611\end{pmatrix}.$$

Определить:

а) значения главных компонент f_{i1}, f_{i2} для десятого наблюдения (i=10);

б) нормированные значения исходных показателей для десятого наблюдения (y_{i1}, y_{i2}), где i=10;

в) значения исходных показателей для i=10 наблюдения (x_{i1}, x_{i2}), если выборка характеризуется векторами средних $\bar{x}=\begin{pmatrix}0,850\\ 2,307\end{pmatrix}$; $s=\begin{pmatrix}0,072\\ 0,093\end{pmatrix}$;

г) значения исходных показателей для шестого наблюдения.

5.13. Хозяйственная деятельность машиностроительных предприятий (N=10) характеризуется следующими показателями (n=3) (табл. 5.5).

Таблица 5.5

№ п/п	Трудоемкость единицы продукции, x_1	Удельный вес рабочих в составе ППП, x_2	Коэффициент сменности оборудования, x_3
1	0,38	0,72	1,40
2	0,24	0,70	1,20
3	0,31	0,66	1,15
4	0,42	0,69	1,09
5	0,51	0,71	5,00
6	0,31	0,73	1,36
7	0,37	0,65	1,15
8	0,16	0,82	1,87
9	0,18	0,80	2,17
10	0,43	0,83	1,61

Требуется:

а) ранжировать предприятия по первой главной компоненте;

б) графически изобразить предприятия в пространстве двух первых компонент;

в) сопоставить и дать интерпретацию полученным результатам.

5.14. Экономическая деятельность предприятий машиностроения (N=10) характеризуется показателями (n=3) (табл. 5.6).

Таблица 5.6

№ п/п	Удельный вес покупных изделий, x_1	Коэффициент сменности оборудования, x_2	Премии и вознаграждения на одного работника, x_3	Y
1	0,30	1,40	0,67	9,8
2	0,56	1,20	0,98	13,2
3	0,42	1,15	1,16	17,3
4	0,26	1,09	0,54	7,1
5	0,16	1,26	1,23	11,5
6	0,45	1,36	0,78	12,1
7	0,31	1,15	1,16	15,2
8	0,08	1,87	2,44	31,3
9	0,63	2,17	1,06	11,6
10	0,03	1,61	2,13	30,1

Приняв за результативный показатель y - рентабельность, построить уравнение регрессии на главной компоненте, наиболее тесно связанной с y. Дать экономическую интерпретацию полученным результатам.

5.15. Значения экономических показателей предприятий машиностроения (N=10) характеризуются следующими показателями табл. 5.7:

Таблица 5.7

№ п/п	Удельный вес рабочих в составе ППП, x_1	Удельный вес покупных изделий, x_2	Премии и вознаграждения на одного работника, x_3
1	0,68	0,33	0,82
2	0,74	0,25	0,84
3	0,66	0,32	0,67
4	0,72	0,02	1,04
5	0,68	0,06	0,66
6	0,77	0,15	0,86
7	0,78	0,08	0,79
8	0,78	0,20	0,34
9	0,81	0,20	1,60
10	0,79	0,30	1,46

Требуется:

а) определить собственные значения главных компонент и их вклад в суммарную дисперсию исходных признаков;

б) определить ортогональную матрицу собственных векторов V;

в) дать экономическую интерпретацию главных компонент.

5.16. Деятельность предприятий (N=10) характеризуется следующими показателями (n=3) (табл. 5.8).

Таблица 5.8

№ п/п	Трудоемкость единицы продукции, x_1	Удельный вес покупных изделий, x_2	Коэффициент сменности оборудования, x_3	Y
1	0,51	0,20	1,47	21,9
2	0,36	0,64	1,27	48,4
3	0,23	0,42	1,51	173,5
4	0,26	0,27	1,46	74,1
5	0,27	0,37	1,27	68,6
6	0,29	0,38	1,43	60,8
7	0,01	0,35	1,50	355,6
8	0,02	0,42	1,35	264,8
9	0,18	0,32	1,41	526,6
10	0,25	0,33	1,47	118,6

Приняв за результативный показатель Y - индекс снижения себестоимости продукции, построить уравнение регрессии на главные компоненты, наиболее тесно связанные с y. Дать экономическую интерпретацию полученным результатам.

5.17. Заданы значения главных компонент для i-го объекта f_{i1} = –0,252; f_{i2} = = 0,335; f_{i3} = –0,542 и матрица факторных нагрузок

$$A=\begin{pmatrix} -0,718 & 0,691 & -0,081 \\ 0,879 & 0,436 & 0,191 \\ 0,964 & 0,116 & -0,235 \end{pmatrix}.$$

Найти значения исходных показателей для этого объекта (x_{i1} , x_{i2} , x_{i3}), если выборка характеризуется векторами средних

$$\bar{x}=\begin{pmatrix} 0,331 \\ 0,731 \\ 1,426 \end{pmatrix}$$

и средних квадратических отклонений

$$s=\begin{pmatrix} 0,107 \\ 0,051 \\ 0,337 \end{pmatrix}.$$

5.18. В матрице значений главных компонент F оказалось пропущенным восьмое наблюдение (i=8) (табл.5.9):

Таблица 5.9

1	–0,25	0,34	–0,54
2	–0,22	–1,30	1,31
3	–0,75	–1,08	–0,14
4	–0,97	0,24	0,39
5	–0,89	1,39	–0,87
6	–0,03	–0,24	0,60
7	–1,00	–0,62	–0,92
8			
9	1,86	–0,33	–1,91
10	0,58	2,07	1,09

Матрица факторных нагрузок равна:

$$A=\begin{pmatrix} -0,72 & 0,69 & -0,08 \\ 0,88 & 0,44 & 0,19 \\ 0,96 & 0,12 & -0,24 \end{pmatrix}.$$

Определить:

а) значения главных компонент f_{i1} , f_{i2} , f_{i3} для i=8;

б) нормированные значения исходных показателей восьмого наблюдения (y_{81} , y_{82} , y_{83});

в) значения исходных показателей для восьмого наблюдения (x_{81} , x_{82} , x_{83}), если выборка характеризуется векторами средних $\bar{x}=\begin{pmatrix} 0,33 \\ 0,73 \\ 1,43 \end{pmatrix}$ и средних квадратических отклонений $s=\begin{pmatrix} 0,11 \\ 0,06 \\ 0,34 \end{pmatrix}$;

г) значения исходных показателей для третьего наблюдения.

5.5. Задания для самостоятельной работы на ЭВМ

По данным Приложения 2 провести компонентный анализ.

1. Для факторных признаков x найти оценку матрицы R парных коэффициентов корреляции.

2. На основании матрицы собственных значений Λ определить вклад компонент в суммарную дисперсию. Отобрать и указать вклад (m<k) первых главных компонент.

3. Используя матрицу факторных нагрузок A, дать экономическую интерпретацию полученным главным компонентам.

4. По матрице значений главных компонент F провести классификацию объектов по двум первым главным компонентам. Дать интерпретацию полученным результатам.

5. Используя вектор значений результативного признака y и матрицу F, построить уравнение регрессии на главные компоненты с максимальным числом значимых коэффициентов регрессии.

ГЛАВА 6

ФАКТОРНЫЙ АНАЛИЗ

6.1. Основные понятия факторного анализа

6.1.1. Основные проблемы факторного анализа

При использовании факторного анализа исследователь сталкивается со значительным числом проблем. Наиболее часто они возникают в процессе содержательной (экономической) интерпретации. Многие из них носят частный характер, присущий определенному классу задач, например, наличие слабо обусловленных матриц парных коэффициентов корреляций, присущее классу экономико-статистических задач. Однако это не относится непосредственно к факторному анализу.

Исходя из логической последовательности решения задач факторного анализа, первой следует считать проблему робастности (гл.9), второй - общности, третьей - факторов, четвертой - вращения, пятой - оценки значений факторов, шестой - проблему динамических моделей.

Введем следующие пояснения к общей схеме факторного анализа, представленной на рис. 6.1.

На рис.6.1 представлена матрица исходных признаков X размерности $n \times N$, где число признаков равно n, а количество исследуемых объектов - N.

Элементы матрицы x_{ij} представляют собой значения j-х признаков на i-м объекте исследования. Каждый признак имеет свое наименование. Например, x_1 - производительность труда (выработка в год на одного производственного рабочего), x_2 - фондоотдача и т.д.

В классическом факторном анализе следующей матрицей является матрица нормированных значений исходных данных, которая получалась путем нормирования элементов X. Однако современные исследования и опыт решения практических задач показали, что отсутствие учета возможного засорения исходных признаков грубыми ошибками при многомерном анализе может привести к ошибкам интерпретации. В настоящее время эти две матрицы объединяет **первая проблема - робастности**. Технически на ЭВМ эта задача решается методом устойчивой (робастной) оценки среднего значения ($\bar{x}$) и оценки среднего квадратического отклонения (s) (гл. 9).

При этом используется тест на засорение. Если засорение отсутствует, то $\bar{x}$ и s определяются так же, как и в классическом случае. Если же засорение обнаружено, то вводится тест на симметричность распределения.

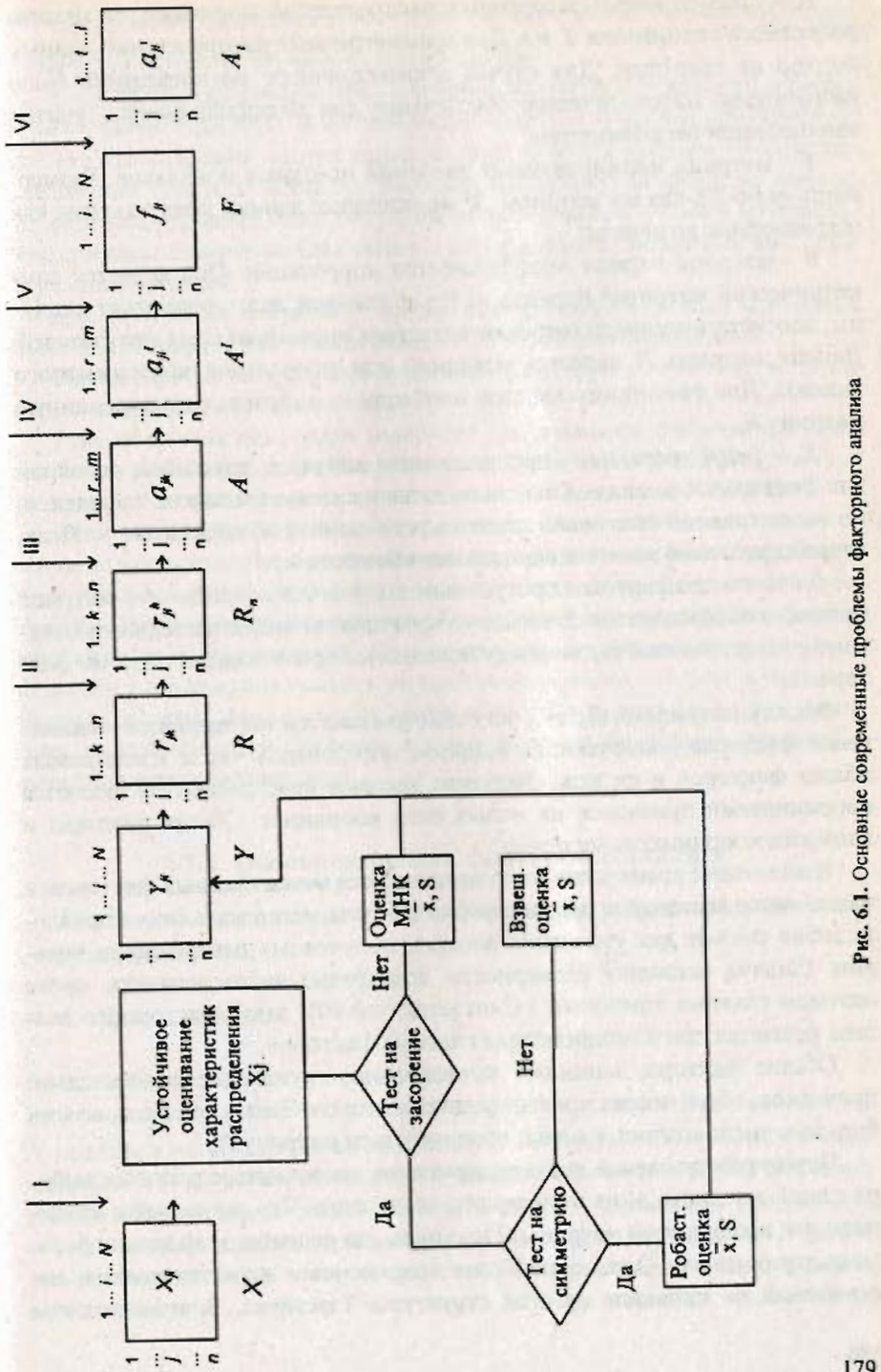


Рис. 6.1. Основные современные проблемы факторного анализа

Для симметричных засоренных распределений применяются методы робастного оценивания $\bar{x}$ и s. Для асимметричных распределений данный подход не пригоден. Для случая асимметричного распределения было разработано математическое обеспечение для джекнайф-оценки, учитывающей наличие асимметрии.

Y - матрица нормированных значений исходных признаков. Размерность ее ($n \times N$), как и у матрицы X, но исходные данные представлены как безразмерные величины.

R - матрица парных коэффициентов корреляции. Она является симметрической матрицей порядка n. На ее главной диагонали стоят единицы, соответствующие дисперсиям исходных нормированных показателей. Данная матрица R является исходной для проведения компонентного анализа. Для факторного анализа необходимо получить редуцированную матрицу R_h.

R_h - редуцированная корреляционная матрица, служащая основной для факторного анализа. Она также является симметрической порядка n, но на ее главной диагонали стоят вместо единиц общности (h_j^2). Итак, **второй проблемой** является **определение общности** h_j^2.

Далее по стандартным программам получается матрица A - матрица весовых коэффициентов. Весовые коэффициенты являются характеристиками стохастической связи между исходными признаками и общими факторами.

Между матрицами R_h и A на схеме указана **третья проблема - нахождение факторов**, включающая вопросы определения числа извлекаемых общих факторов и их вида. Значения весовых коэффициентов являются координатами признаков на новых осях координат. Общие факторы и являются координатными осями.

В настоящее время чаще всего используется **метод главных факторов**, а также **метод максимума правдоподобия**. Иногда метод максимума правдоподобия служит для уточнения данных, полученных **центроидным методом**. (Задача снижения размерности достаточно часто решается также методом главных компонент.) Считается, что 90% задач факторного анализа решается при помощи метода главных факторов.

Общие факторы занимают произвольное положение относительно признаков, образующих конфигурацию векторов. Таким образом, весьма большое число матриц A может воспроизвести матрицу R_h.

Четвертой проблемой является **вращение**, включающее вопросы выбора одной матрицы A^1 из возможного множества. Что же является критерием для выбора этой матрицы? Подходов для решения этой задачи было выдвинуто много. Однако наиболее современным является **подход, построенный на принципе простой структуры Тэрстоуна.** В этом подходе

$A^1 = (a_{jr}^1)$ является матрицей размерности $n \times m$, элементы (весовые коэффициенты) которой получены после вращения.

F - матрица индивидуальных значений факторов для каждого объекта исследования. Размерность данной матрицы ($m \times N$). Матрица примечательна тем, что содержит оценки значений факторов для каждого объекта исследования. Вначале на основе значений исходных признаков на каждом объекте исследования были получены общие факторы - новые интегральные единицы измерения. Они явились результатом исследования массовых случайных явлений.

Для получения матрицы F новые интегральные единицы измерения были использованы для оценки каждого из объектов исследования.

Пятая проблема - оценка значений факторов - включает вопросы перехода от матрицы A^1 к матрице F.

Решением этих пяти задач (проблем) заканчивается статический вариант факторного анализа. На основе полученных данных (например, в конце года) имеем результаты измерения признаков на каждом из N объектов.

Если в течение нескольких лет проводить факторный анализ, затем сопоставить его результаты за несколько лет, то в ряде случаев можно получить интересные результаты исследования. Такой подход определяет **шестую проблему** факторного анализа. Она раскрывает **принципы построения динамических моделей.** При помощи динамических моделей представляется возможным выявить те признаки, влияние которых в будущем будет снижаться или, наоборот, возрастать. Динамические модели родились из потребностей практики в сопоставлении факторных решений в течение длительного периода.

6.1.2. Основная модель факторного анализа

Пусть задана система переменных $x_1, x_2, \ldots, x_n$. Дадим им некоторый содержательный смысл:

x_1 - производительность труда;

x_2 - фондоотдача;

...

x_n - себестоимость.

Значения переменных или признаков $x_1, x_2, \ldots, x_n$ известны для каждого из N предприятий. Представим полученную исходную информацию в виде матрицы $X=(x_{ji})$ размерности ($n \times N$). Каждая строка состоит из значений одного показателя для каждого из N объектов исследования. Предполагаем, что каждый элемент этой матрицы (x_{ji}) является результатом воздействия некоторого числа m гипотетических общих факторов и одного характерного фактора.

$$x_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + \ldots + a_{jm} f_{mi} + d_j v_{ji}; \qquad (6.1)$$

$$(j = \overline{1,n}; r = \overline{1,m}; i = \overline{1,N}),$$

где x_{ji} - центрированное значение j-го показателя (переменной) у i-го объекта исследования;

f_r - r-й общий фактор;

v_j - j-й характерный (индивидуальный) фактор, присущий только данной j-й переменной;

f_{ri} - значение r-го общего фактора на i-м объекте исследования;

v_{ji} - значение j-го характерного фактора на i-м объекте исследования;

a_{jr} - весовой коэффициент j-й переменной на r-м общем факторе или нагрузка j-й переменной на r-м общем факторе;

d_j - нагрузка или весовой коэффициент j-й переменной на j-м характерном факторе.

Обычно предполагают, что характерные факторы некоррелированны между собой и с общими факторами.

Факторы, связанные значимыми коэффициентами веса более чем с одной переменной, называются **общими**. Общий фактор, связанный значимыми коэффициентами веса со всеми признаками, называется **генеральным**.

В выражении (6.1) x_{ji} имеют свою размерность. Для того чтобы перейти к безразмерным переменным, удобно провести нормирование исходных показателей. Проведем центрирование:

$$x_{ji} = x'_{ji} - \bar{x}_j, \qquad (6.2)$$

где $\bar{x}_j$ - среднее значение j-й переменной;

x_{ji} - исходное значение j-й переменной на i-м объекте исследования.

Размерность при этом, естественно, не изменилась.

Проведем нормирование:

$$y_{ji} = \frac{x_{ji}}{s_j} = \frac{x'_{ji} - \bar{x}_j}{s_j}, \qquad (6.3)$$

где y_{ji} - нормированное значение j-й переменной на i-м объекте (безразмерная величина);

s_j - среднее квадратическое отклонение j-го признака (исправленное).

$$s_j = \frac{1}{\sqrt{n-1}} \sqrt{\sum_{i=1}^{N} x_{ji}^2}.$$

После этих преобразований получим $s_j^2 = 1$.

$$y_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + \ldots + a_{jm} f_{mi} + d_j v_{ji}, \qquad (6.4)$$

$$(j = \overline{1,n}; r = \overline{1,m}; i = \overline{1,N}).$$

Средние значения переменных Y=0, а дисперсии $s_j^2 = 1$. Формулы в виде (6.1) или (6.4) напоминают выражение для регрессионного анализа. Действительно, признак (зависимая переменная) описывается при помощи m других показателей плюс остаток $d_j v_{ji}$. Однако здесь есть большая разница. В регрессионном анализе система переменных предполагается измеряемой непосредственно. Например, взятые нами переменные могут быть получены из отчетной информации предприятий (объектов исследования). И действительно по ним можно было бы провести регрессионный анализ. Однако выражения (6.1) и (6.4) предназначены для решения другой задачи. В факторном анализе общие факторы f_r и индивидуальные (характерные) являются гипотетическими. Мы должны их оценить при помощи методов математической статистики и линейной алгебры.

В дальнейшем будем пользоваться в основном выражением (6.4). Обращение к выражению (6.1) будет каждый раз оговариваться.

Из выражения (6.4) следует, что наблюдаемые значения переменных являются линейными комбинациями ненаблюдаемых, гипотетических, непосредственно не измеряемых факторов. Следовательно любой метод факторного анализа имеет одну главную задачу: представить элемент матрицы y в виде линейной комбинации некоторого числа общих факторов (m) и одного характерного фактора.

Представим теперь выражение (6.4) в матричной форме. В матричной форме из выражения (6.4) следует, что A и D относятся к переменным разных типов. A - относится к m общим факторам, каждый из которых связан со всеми признаками, значимыми или незначимыми коэффициентами веса, следовательно, является прямоугольной матрицей. D связана только с индивидуальными факторами, каждый из которых связан только с одной переменной.

Следовательно, D - диагональная матрица. Матрица A не содержит весовых коэффициентов характерных факторов, а D - коэффициентов при общих факторах. Эти две матрицы надо сложить. При описанных условиях известная нам матрица A должна быть дополнена n нулевыми столбцами порядка n справа, а D - m нулевыми столбцами порядка n - слева.

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} & \overbrace{0 & 0 & \ldots & 0}^{n} \\ a_{21} & a_{22} & \ldots & a_{2m} & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nm} & 0 & 0 & \ldots & 0 \end{pmatrix}; \qquad (6.5)$$

$$D=\begin{pmatrix} 0 & 0 & \dots & 0 & d_1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \underbrace{0 \quad 0 \quad \dots \quad 0}_{m} & & & & 0 & 0 & \dots & d_n \end{pmatrix}. \qquad (6.6)$$

Сумма матриц (6.5) и (6.6) дает

$$M=\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} & d_1 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2m} & 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} & 0 & 0 & \dots & d_n \end{pmatrix}. \qquad (6.7)$$

Представим матрицу индивидуальных значений общих факторов F:

$$F=\begin{pmatrix} f_{11} & f_{12} & \dots & f_{1N} \\ f_{21} & f_{22} & \dots & f_{2N} \\ \dots & \dots & \dots & \dots \\ f_{m1} & f_{m2} & \dots & f_{mN} \end{pmatrix}; \qquad (6.8)$$

$$V=\begin{pmatrix} v_{11} & v_{12} & \dots & v_{1N} \\ v_{21} & v_{22} & \dots & v_{2N} \\ \dots & \dots & \dots & \dots \\ v_{n1} & v_{n2} & \dots & v_{nN} \end{pmatrix}. \qquad (6.9)$$

Иногда для удобства их представляют в одной таблице

$$F^*=\begin{pmatrix} f_{11} & f_{12} & \dots & f_{1N} \\ f_{21} & f_{22} & \dots & f_{2N} \\ \dots & \dots & \dots & \dots \\ f_{m1} & f_{m2} & \dots & f_{mN} \\ v_{11} & v_{12} & \dots & v_{1N} \\ v_{21} & v_{22} & \dots & v_{2N} \\ \dots & \dots & \dots & \dots \\ v_{n1} & v_{n2} & \dots & v_{nN} \end{pmatrix}. \qquad (6.10)$$

После этих уточнений модель (6.4) можно записать в матричной форме

$$Y=AF, \qquad (6.11)$$

при этом учитывается, что нагрузки характерных факторов вычисляются из общности.

Отметим, что матрица A имеет размерность $(n\times m)$, матрица F^* - размерность $[(m+n)\times N]$, а матрица Y - размерность $(n\times N)$. M - матрица факторного отображения, включающая нагрузки характерных факторов.

Ее называют **полной факторной матрицей**. F^* - матрица значений факторов (общих и характерных) для всех объектов исследования.

Из трех матриц выражения (6.11) не известны две A и F.

6.1.3. Компоненты дисперсии в факторном анализе

Из теории измерений известно, что измеренная величина содержит, по крайней мере, две компоненты: истинное значение и ошибку измерения

$$x_{ji}=x_{mi}+x_{li},$$

где x_{ji} - измеренное (зарегистрированное) значение j-го признака у i-го объекта;
x_{mi} - истинное значение измеряемого признака;
x_{li} - ошибка измерения при определении j-го признака у i-го объекта.

Когда же измерения ведутся в биологии, психологии, экономике, медицине и т.д., то добавляется третья компонента, зависящая от вариабельности изучаемого признака на данных объектах (субъектах) исследования. Следовательно, измеренное значение признака становится суммой трех составляющих

$$x_{ji}=x_{mi}+x_{si}+x_{li}, \qquad (6.12)$$

где x_{si} - вариативное значение измеряемого признака у i-го объекта исследования.

Необходимо выяснить, что представляет собой истинное значение измеряемого признака и какова составляющая x_{mi}. Речь может идти о математическом ожидании, поэтому для случайных величин x_{mi} - математическое ожидание изучаемого признака, а две другие компоненты характеризуют отклонение от математического ожидания.

Ошибка измерения обычно оказывается значительно меньше вариативной компоненты, поэтому их часто объединяют. Однако поскольку вариативная составляющая и ошибки измерения возникают независимо друг от друга, то их легко разъединить и вычислить отдельно.

Если первая компонента x_{mi} является общей статистической характеристикой совокупности объектов исследования, то вторая и третья компоненты характеризуют отклонение отдельного показателя на данном объекте от среднего. Они являются носителем особенностей, присущих данному объекту и методу измерения. В процессе управления важнейшим моментом являются знание и умение учитывать индивидуальные возможности и способности отдельных объектов (субъектов) исследования.

Простейшей характеристикой этих особенностей является разность

$$x_{ji}-x_{mi}=x_{si}+x_{li}. \qquad (6.13)$$

Однако при множестве объектов исследования пользуются более обобщенными характеристиками. Самой распространенной характеристикой изменчивости объектов исследования является **дисперсия** σ_j^2 или σ_j - **среднее квадратическое отклонение**. Рассмотрим компоненты

дисперсии в факторном анализе. Для этого обратимся к выражению (6.4). Так как y_j является нормированной случайной величиной, то его дисперсия равна единице и представляет собой сумму квадратов значений показателей по всем объектам исследования, деленную на число объектов N или N-1 (с учетом смещения характеристики)

$$\hat{s}_j^2 = 1 = \frac{1}{N}\sum_{i=1}^{N} y_{ji}^2 = \frac{1}{N}[a_{j1}^2 \sum_{i=1}^{N} f_{1i}^2 + a_{j2}^2 \sum_{i=1}^{N} f_{2i}^2 + \ldots + \tag{6.14}$$

$$+ a_{jm}^2 \sum_{i=1}^{N} f_{mi}^2 + d_j^2 \sum_{i=1}^{N} v_{ji}^2 + 2(a_{j1}a_{j2}\sum_{i=1}^{N} f_{1i}f_{2i} + a_{j1}a_{j3}\sum_{i=1}^{N} f_{1i}f_{3i} +$$

$$+ \ldots + a_{j(m-1)}a_{jm}\sum f_{(m-1)i}f_{mi} + a_{jm}d_j \sum_{i=1}^{N} f_{mi}v_{ji})],$$

где $j = \overline{1,n}$; $r = \overline{1,m}$; $i = \overline{1,N}$.

Рассмотрим слагаемые, содержащие сомножитель $\frac{1}{N}\sum_{i=1}^{N} f_{ri}^2$.

В рассматриваемом случае величина

$$\frac{1}{N}\sum_{i=1}^{N} f_{ri}^2 = \hat{s}_{f_r}^2 = 1$$

является дисперсией нормированного общего фактора f_r и равняется единице.

Значит,

$$f_{ri} = \frac{f'_{ri} - \bar{f}_r}{s_{f_r}},$$

где f'_{ri} - ненормированное значение r-го общего фактора;

$\bar{f}_r$ - среднее значение ненормированного фактора;

s_{f_r} - среднее квадратическое отклонение ненормированного общего фактора.

Рассмотрим в (6.14) слагаемые, содержащие сомножитель $\frac{1}{N}\sum_{i=1}^{N} f_{ri}f_{li}$. Это коэффициент корреляции между двумя общими факторами, т.е.

$$\frac{1}{N}\sum_{i=1}^{N} f_{ri}f_{li} = r_{f_r f_l},$$

где $r = \overline{1,m}$; $l = \overline{1,m}$; $r \neq l$.

Очевидно, что

$$\frac{1}{N}\sum_{i=1}^{N} f_{mi}v_{ji} = r_{f_m v_j}.$$

Следовательно, выражение (6.14) можно представить в виде

$$\hat{s}_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2 + d_j^2 + 2(a_{j1}a_{j2}r_{f_1 f_2} + a_{j1}a_{j3}r_{f_1 f_3} + \ldots$$
$$\ldots + a_{jm}d_j r_{f_m v_j}).$$

Откуда следует, что

$$\hat{s}_j^2 = 1 = \sum_{r=1}^{m} a_{jr}^2 + d_j^2 + 2\sum_{l>r=1}^{m} a_{jr}a_{jl}r_{f_r f_l} + 2d_j \sum_{r=1}^{m} a_{jr}r_{f_r v_j}. \tag{6.15}$$

Так как характерный фактор присущ только данной j-й переменной и всегда не коррелирован с общими факторами, следовательно, $r_{f_r v_j} = 0$ и выражение (6.15) можно упростить:

$$s_j^2 = \sum_{r=1}^{m} a_{jr}^2 + d_j^2 + 2\sum_{l>r=1}^{m} a_{jr}a_{jl}r_{f_r f_l}. \tag{6.16}$$

Дальнейшее упрощение может быть получено для некоррелированных общих факторов, когда и $r_{f_r f_l} = 0$, тогда

$$\hat{s}_j^2 = d_j^2 + \sum_{r=1}^{m} a_{jr}^2. \tag{6.17}$$

В этом случае дисперсия признака y_j равна сумме относительных вкладов в дисперсию этого признака каждого из m общих и одного характерного фактора.

Выражение

$$\sum_{r=1}^{m} a_{jr}^2 = h_j^2 \tag{6.18}$$

называется **общностью показателя** y_j, т.е. суммой относительных вкладов всех m общих факторов в дисперсию признака y_j. Вклад в дисперсию признака y_j характерного фактора v_j или характерность представляет собой d_j^2.

Дисперсия характерного фактора состоит из двух компонент: компоненты, связанной со спецификой параметра S_j, и компоненты, связанной с ошибками измерений E_j.

Если факторы специфичности S_j и ошибки E_j не коррелированны между собой, то модель факторного анализа примет вид

$$y_j = a_{j1}f_1 + a_{j2}f_2 + \ldots + a_{jm}f_m + b_j S_j + c_j E_j. \tag{6.19}$$

Вклад же характерного фактора в дисперсию признака может быть представлен следующим образом:

$$d_j^2 = b_j^2 + c_j^2. \tag{6.20}$$

Если выделить из дисперсии признака составляющую ошибки, то получим характеристику, называемую **надежностью**:

$$r_j^2 = h_j^2 + b_j^2. \tag{6.21}$$

Составим сводную таблицу формул для определения вкладов факторов в дисперсию признака (табл.6.1).

Если исследователя интересует вклад фактора f_r в суммарную дисперсию всех признаков, то это выразится так:

$$V_r = \sum_{j=1}^{n} a_{jr}^2. \tag{6.22}$$

Таблица 6.1

Формулы долей дисперсии

Характеристика	Обозначение	Расчетные соотношения
Полная дисперсия	s_j^2	$h_j^2 + b_j^2 + c_j^2 = h_j^2 + d_j^2 = 1$
Надежность	r_j^2	$h_j^2 + b_j^2 = 1 - c_j^2$
Общность	h_j^2	$1 - d_j^2$
Характерность	d_j^2	$b_j^2 + c_j^2 = 1 - h_j^2$
Специфичность	b_j^2	$d_j^2 - c_j^2$
Дисперсия ошибки	c_j^2	$1 - r_j^2$

Вклад всех общих факторов в суммарную дисперсию признаков рассчитывается следующим образом:

$$V_o = \sum_{r=1}^{m} v_r \,. \qquad (6.23)$$

Полнотой факторизации называют отношение

$$k = \frac{V_o}{n}. \qquad (6.24)$$

При проведении анализа полученных результатов бывает полезным построение диаграмм, в которых отражаются доли дисперсии каждой переменной и вклады исходных признаков в дисперсии общих факторов.

Пример 6.1

В результате решения задачи, имеющей семь признаков, получено два общих фактора. Необходимо определить:

1) вклады общих и характерного факторов в дисперсию признаков, %;

2) вклад всех семи признаков в каждый общий фактор, %;

3) вклад каждого общего фактора в суммарную дисперсию, построить график вкладов признаков в каждый из общих факторов, %;

4) составить таблицу относительного вклада факторов в суммарную дисперсию.

Матрица A весовых коэффициентов общих факторов имеет вид:

$$A = \begin{pmatrix} 0,90 & -0,30 \\ 0,80 & -0,30 \\ 0,60 & 0,30 \\ 0,50 & 0,20 \\ 0,50 & 0,50 \\ -0,30 & 0,60 \\ 0,20 & 0,80 \end{pmatrix}. \qquad (6.25)$$

Признаки y_j матрицы A отображены на рис.6.2.

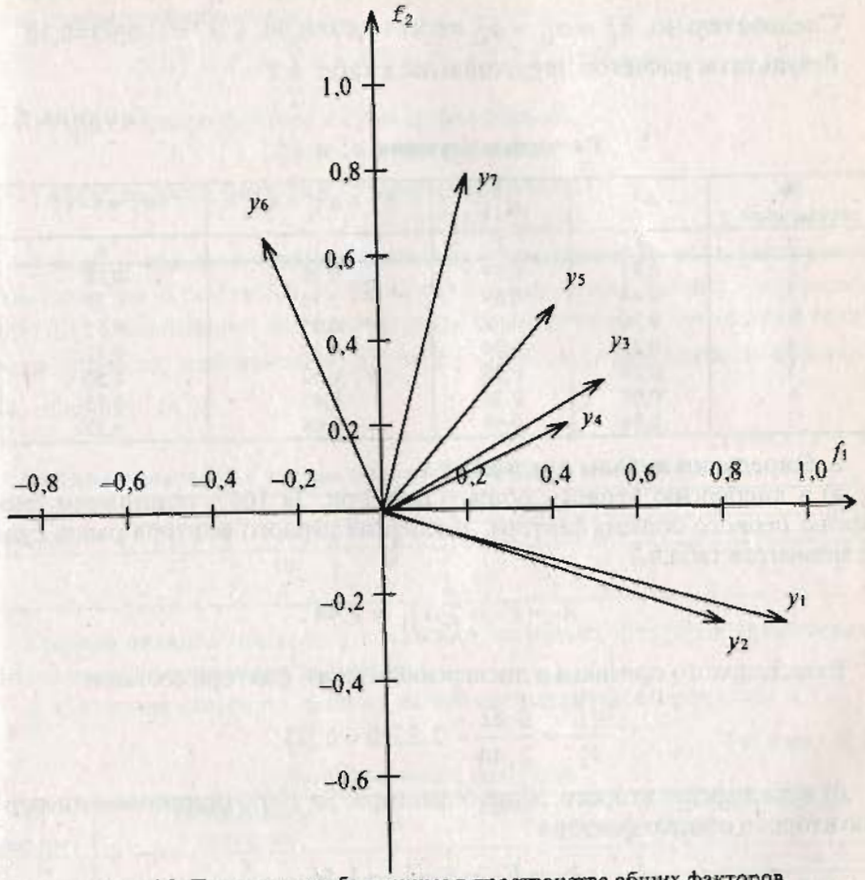


Рис. 6.2. Признаки, отображенные в пространстве общих факторов

Пояснения. Первый столбец данной матрицы a_1 представляет собой вектор весовых коэффициентов первого общего фактора. Второй столбец матрицы A - вектор a_2 весовых коэффициентов второго общего фактора. Так, a_{31} является весовым коэффициентом связи между признаком y_3 и первым общим фактором, а $a_{31}^2 = 0,36$ - вклад третьей переменной в дис-персию первого общего фактора. Вклад первого признака в дисперсию второго общего фактора составляет $a_{12}^2 = 0,09$.

Решение

1. Определим вклады общих и характерного факторов в дисперсию признаков.

а) вклад первого признака в дисперсию первого фактора составит $a_{11}^2 = 0,90^2 = 0,81$, а его вклад во второй фактор - $a_{12}^2 = 0,30^2 = 0,09$.

Следовательно, $h_1^2 = a_{11}^2 + a_{12}^2 = 0,81+0,09=0,90$; а $d_1^2 = 1-0,90=0,10$.

Результаты расчетов представлены в табл. 6.2.

Таблица 6.2

Расчетные значения h_j^2 и d_j^2

№ переменной j	a_{j1}^2	a_{j2}^2	$h_j^2 = a_{j1}^2 + a_{j2}^2$	$d_j^2 = 1 - h_j^2$
A	1	2	3	4
1	0,81	0,09	0,90	0,10
2	0,64	0,09	0,73	0,27
3	0,36	0,09	0,45	0,55
4	0,25	0,04	0,29	0,71
5	0,25	0,25	0,50	0,50
6	0,09	0,36	0,45	0,55
7	0,04	0,64	0,68	0,32

2. Определим вклады признаков:

а) в дисперсию первого общего фактора. За 100% принимаем дисперсию первого общего фактора. Дисперсия первого фактора равна сумме элементов табл.6.2

$$\lambda_1 = V_1 = \sum_{j=1}^{7} a_{j1}^2 = 2,44.$$

Вклад первого признака в дисперсию первого фактора составит

$$\frac{a_{11}^2}{V_1} = \frac{0,81}{2,44} = 0,3320 \approx 0,33;$$

б) в дисперсию второго общего фактора. За 100% принимаем дисперсию второго общего фактора

$$\lambda_2 = V_2 = \sum_{j=1}^{7} a_{j2}^2 = 1,56.$$

Вклад первой переменной в дисперсию второго фактора

$$\frac{a_{12}^2}{V_2} = \frac{0,09}{1,56} = 0,0577 \approx 0,06.$$

в) составим таблицу вкладов переменных в дисперсию общих факторов (табл. 6.3).

Таблица 6.3

Вклады признаков в дисперсии общих факторов

№ фактора r	Вклады признаков, %						
	y_1	y_2	y_3	y_4	y_5	y_6	y_7
1	33	26	15	10	10	4	2
2	6	6	6	2.5	16	23	41

3. Рассчитаем вклады общих факторов в суммарную общность и определим:

а) суммарную общность

$$\sum_{o=1}^{7} h_j^2 = V_o = \sum_{r=1}^{2} V_r = V_1 + V_2 = 2,44 + 1,56 = 4,00;$$

б) вклад первого фактора в суммарную общность:

$$V_1 : V_o = 2,44 : 4,0 = 0,61;$$

в) вклад второго фактора в суммарную общность:

$$1 - V_1 : V_o = 1 - 0,61 = 0,39;$$

г) вклады каждого признака в общность первого и второго факторов с точностью до 1% (табл.6.4). Для этого надо вклад каждого признака (табл.6.3) умножить на вес соответствующего фактора в суммарной общности процесса, или значения a_{j1}^2 и a_{j2}^2 (табл.6.2) разделить на суммарную общность (4,0).

Таблица 6.4

Вклады признаков с учетом вкладов факторов в суммарную общность

№ фактора r	Вклады признаков, %						
	y_1	y_2	y_3	y_4	y_5	y_6	y_7
1	20	16	9	6	6	2	1
2	2	2	2	1	6	9	16

График вкладов признаков в каждый из общих факторов представим читателю построить самостоятельно.

4. Составим итоговую таблицу долей дисперсий факторов (табл.6.5).

Таблица 6.5

Доли дисперсий факторов

№ фактора r	Наименование дисперсии	Формула	Значение вклада	Вклад, %
A	1	2	3	4
1	Дисперсия процесса (полная дисперсия)	S_Σ^2	7,00	100
2	Дисперсия первого фактора	$V_1 = \sum_{j=1}^{7} a_{j1}^2$	2,44	34,86
3	Дисперсия второго фактора	$V_2 = \sum_{j=1}^{7} a_{j2}^2$	1.56	22,29
4	Общность процесса (суммарная общность)	$V_o = \sum_{j=1}^{7} h_j^2 = V_1 + V_2$	4,0	57,14
5	Суммарная характерная дисперсия	$V_x = \sum_{j=1}^{7} d_j^2$	3,0	42,86

Следует отметить, что дисперсия процесса равна 7 и совпадает с числом признаков. Дисперсия каждого нормированного признака равна 1, поэтому полная дисперсия при семи показателях равна 7.

Естественно, что $V_0 + V_x = 4{,}0 + 3{,}0 = 7{,}0$. Необходимо также заметить, что весь анализ дисперсий был проведен только на основе заданной матрицы весовых коэффициентов общих факторов. Значит, значение a_{jr} общих факторов однозначно определяет значения весовых коэффициентов характерных факторов.

6.1.4. Получение матрицы коэффициентов парной корреляции и ее преобразование в факторном анализе

Исходные данные матрицы Y позволяют получить матрицу R. Эта матрица коэффициентов парной корреляции или корреляционная матрица. Из матриц (6.5), (6.6) и (6.7) видно, что $M=A+D$. Значит, для воспроизведения всех связей переменных в корреляционной матрице надо воспользоваться матрицей M.

Тогда

$$R = MM^T = (A+D)(A+D)^T = (A+D)(A^T + D^T) = \\ = AA^T + AD^T + DA^T + DD^T. \qquad (6.26)$$

Из матриц (6.5) и (6.6) следует, что

$$AD^T = DA^T = 0. \qquad (6.27)$$

Следовательно,

$$R = AA^T + DD^T; \qquad (6.28)$$

$AA^T = R_h$, а $DD^T = D^2$, так как D - диагональная матрица. Возведение ее в квадрат приводит к диагональной матрице, у которой на главной диагонали стоят элементы матрицы D в квадрате.

Таким образом, матрица коэффициентов парной корреляции, полученная из исходных показателей, может быть воспроизведена при помощи матрицы M.

$$R = R_h + D^2 \qquad (6.29)$$

или

$$R = AA^T + D^2. \qquad (6.30)$$

Матрица R - корреляционная матрица с единицами на главной диагонали, а матрица R_h - это корреляционная матрица с общностями на главной диагонали. Матрица R является симметрической, как и ковариационная матрица Σ. Элементы ее вещественные. Элементы ее главной диагонали являются дисперсиями соответствующих случайных величин. Так, r_{11} - дисперсия случайной величины-показателя y_1, а r_{kk} - дисперсия k-го признака y_k. Все дисперсии равны единицам. Следовательно, суммарная дисперсия всех изучаемых признаков будет равна следу матрицы R - сумме ее

диагональных элементов, т.е. сумме дисперсий признаков. Представим элементы матрицы R в развернутом виде

$$\begin{cases} r_{11} = 1 \\ r_{12} = \dfrac{1}{N}(y_{11}y_{21} + y_{12}y_{22} + \ldots + y_{1N}y_{2N}) \\ \ldots \\ r_{jk} = \dfrac{1}{N}(y_{j1}y_{k1} + y_{j2}y_{k2} + \ldots + y_{jN}y_{kN}) \\ \ldots \\ r_{nn} = 1 \end{cases}, \qquad (6.31)$$

где $r_{jj}=r_{kk}=1$; $r_{jk}=r_{kj}$.

В матричной форме на основе (6.31) можно записать

$$R = \frac{1}{N} YY^T. \qquad (6.32)$$

Воспользуемся формулой (6.11) и преобразуем матрицу R в **редуцированную матрицу**

$$R_h = \frac{1}{N} AF(AF)^T = \frac{1}{N} AFF^T A^T = A\frac{1}{N} FF^T A^T.$$

Выражение, стоящее между A и A^T по аналогии с (6.32), является **корреляционной матрицей стохастических связей между общими факторами.**

Обозначим ее

$$\frac{1}{N} FF^T = C,$$

тогда

$$R_h = ACA^T. \qquad (6.33)$$

Если общие факторы не коррелированны между собой, то C будет единичной матрицей E_n, тогда

$$R_h = AA^T. \qquad (6.34)$$

Выражения (6.33) и (6.34) называются **фундаментальной теоремой факторного анализа.**

Пример 6.2

Иллюстрация воспроизведения матрицы R_h при известной матрице весовых коэффициентов A.

Пользуясь матрицей A (пример 6.1), получить:

1) редуцированную матрицу R_h;

2) матрицу рассеяния R^I;

3) матрицу остатков $R_1 = R_h - R^I$.

Провести анализ вкладов в дисперсию общих и характерных факторов.

Решение

1. Получим матрицу R_h.

Для этого произведем умножение матрицы A на A^T и получим редуцированную корреляционную матрицу R_h, т.е. восстановленную из модели факторного анализа при условии, что факторы не коррелированны.

$$R_h = AA^T = \begin{pmatrix} 0{,}90 & -0{,}30 \\ 0{,}80 & -0{,}30 \\ 0{,}60 & 0{,}30 \\ 0{,}50 & 0{,}20 \\ 0{,}50 & 0{,}50 \\ -0{,}30 & 0{,}60 \\ 0{,}20 & 0{,}80 \end{pmatrix} \begin{pmatrix} 0{,}90 & 0{,}80 & 0{,}60 & 0{,}50 & 0{,}50 & -0{,}30 & 0{,}20 \\ -0{,}30 & -0{,}30 & 0{,}30 & 0{,}20 & 0{,}50 & 0{,}60 & 0{,}80 \end{pmatrix};$$

$$R_h = \begin{pmatrix} 0{,}90 & 0{,}81 & 0{,}45 & 0{,}39 & 0{,}30 & -0{,}45 & -0{,}06 \\ 0{,}81 & 0{,}73 & 0{,}39 & 0{,}34 & 0{,}25 & -0{,}42 & -0{,}08 \\ 0{,}45 & 0{,}39 & 0{,}45 & 0{,}36 & 0{,}45 & 0{,}00 & 0{,}36 \\ 0{,}39 & 0{,}34 & 0{,}36 & 0{,}29 & 0{,}35 & -0{,}03 & 0{,}26 \\ 0{,}30 & 0{,}25 & 0{,}45 & 0{,}35 & 0{,}50 & 0{,}15 & 0{,}50 \\ -0{,}45 & -0{,}42 & 0{,}00 & -0{,}03 & 0{,}15 & 0{,}45 & 0{,}42 \\ -0{,}06 & -0{,}08 & 0{,}36 & 0{,}26 & 0{,}50 & 0{,}42 & 0{,}68 \end{pmatrix}.$$

2. Получим матрицу рассеяния R^I.

Ответим на следующие вопросы: что было бы, если бы мы пренебрегли вторым общим фактором и провели интерпретацию на основании только первого общего фактора? Какая матрица рассеяния R^I была бы воспроизведена?

Для этого надо получить внешнее произведение

$$R^I = a_1 a_1^T = \begin{pmatrix} 0{,}90 \\ 0{,}80 \\ 0{,}60 \\ 0{,}50 \\ 0{,}50 \\ 0{,}50 \\ -0{,}30 \\ 0{,}20 \end{pmatrix} \begin{pmatrix} 0{,}90 & 0{,}80 & 0{,}60 & 0{,}50 & 0{,}50 & -0{,}30 & 0{,}20 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}81 & 0{,}72 & 0{,}54 & 0{,}45 & 0{,}45 & -0{,}27 & 0{,}18 \\ 0{,}72 & 0{,}64 & 0{,}48 & 0{,}40 & 0{,}40 & -0{,}24 & 0{,}16 \\ 0{,}54 & 0{,}48 & 0{,}36 & 0{,}30 & 0{,}30 & -0{,}18 & 0{,}12 \\ 0{,}45 & 0{,}40 & 0{,}30 & 0{,}25 & 0{,}25 & -0{,}15 & 0{,}10 \\ 0{,}45 & 0{,}40 & 0{,}30 & 0{,}25 & 0{,}25 & -0{,}15 & 0{,}10 \\ -0{,}27 & -0{,}24 & -0{,}18 & -0{,}15 & -0{,}15 & 0{,}09 & -0{,}06 \\ 0{,}18 & 0{,}16 & 0{,}12 & 0{,}10 & 0{,}10 & -0{,}06 & 0{,}04 \end{pmatrix}.$$

Воспроизведенная или редуцированная по первому общему фактору матрица восстанавливает связи, объясняемые первым вектором - вектором матрицы A. Как первая, так и вторая воспроизведенные матрицы не отражают всей информации процесса. При этом вторая матрица R^I отражает меньше информации, чем первая R^I. Это объясняется тем, что R^I воспроизводит связи, соответствующие V_1=2,44, а R_1 - воспроизводит связи при V_2=1,56 . Однако и более полная матрица R_h не производит связей, определяемых характерными факторами, так как матрица R_h объединяет весовые коэффициенты только общих факторов.

3. Определим матрицу остатков R_1 .

$$R_1 = R_h - R^I = \begin{pmatrix} 0{,}09 & 0{,}09 & -0{,}09 & -0{,}06 & -0{,}15 & -0{,}18 & -0{,}24 \\ 0{,}09 & 0{,}09 & -0{,}09 & 0{,}06 & -0{,}15 & -0{,}18 & -0{,}24 \\ -0{,}09 & -0{,}09 & 0{,}09 & 0{,}06 & 0{,}15 & 0{,}18 & 0{,}24 \\ -0{,}08 & -0{,}06 & 0{,}06 & 0{,}04 & 0{,}10 & 0{,}12 & 0{,}16 \\ -0{,}15 & -0{,}15 & 0{,}15 & 0{,}10 & 0{,}25 & 0{,}30 & 0{,}40 \\ -0{,}18 & -0{,}18 & 0{,}18 & 0{,}12 & 0{,}30 & 0{,}36 & 0{,}48 \\ -0{,}24 & -0{,}24 & 0{,}24 & 0{,}16 & 0{,}40 & 0{,}48 & 0{,}64 \end{pmatrix}.$$

Матрица R_1 содержит необъясненную часть информации матрицей R^I по сравнению с информацией, объясняемой матрицей R_h. Где же содержится эта часть информации? Матрица R_h объясняет всю информацию, представленную матрицей общих факторов A. Следовательно, необъясненная часть информации должна содержаться во втором векторе матрицы A, т.е. в a_2 . Проверим! Если внешнее произведение вектора a_2 на самого себя даст матрицу рассеяния, совпадающую с R_1, то R_1 отражает информацию, содержащуюся в векторе a_2 . (Представим читателю самостоятельно получить произведение вектора a_2 на самого себя и убедиться в том, что $R_1 = a_2 a_2^T$.)

Необъясненная же часть информации матрицей R_h и A приходится на характерные факторы.

4. Анализ вкладов в дисперсию общих и характерных факторов проведем на основе табл.6.2, где даны расчеты вкладов признаков в общие и характерные факторы.

Выводы

1. Вклад первого общего фактора $V_1=2,44$, вклад второго общего фактора $V_2=1,56$ (табл.6.5, гр.3). Суммарный вклад в дисперсию процесса общих факторов составляет V_0. В данном примере суммарный вклад характерных факторов дополняет V_0 до значения $n=7$.

2. Элементы главной диагонали матрицы остатков R_1 совпадают с соответствующими элементами табл.6.2 (гр.3.). В этом столбце заключены вклады признаков в дисперсию второго фактора a_{j2}^2.

3. В матрице R^1 на главной диагонали расположены вклады в дисперсию первого фактора соответствующих переменных a_{j2}^2 (табл.6.2, гр.2).

4. В матрице R_h на главной диагонали находятся дисперсии, которые представляют собой общности, а также суммарный вклад в переменные имеющихся двух общих факторов (табл.6.2, гр.4).

5. Матрица R^1 характеризует связи между переменными, объясняемые только первым общим фактором, а матрица R_1 характеризует связи между переменными, объясняемые вторым общим фактором.

Связи между переменными, объясняемые всеми общими факторами, характеризует матрица R_h, являющаяся суммой $(R^1 + R_1)$.

Матрица R_h данного примера, так же как и R^1, называется **редуцированной матрицей.**

Как правило на практике для получения R_h вместо единиц по главной диагонали матрицы R ставят оценки общностей h_j^2.

Согласно выводам примера 6.2 фундаментальную теорему можно представить в виде

$$R = (R_h + D^2) = M\begin{pmatrix} C & 0 \\ 0 & E \end{pmatrix}M' . \qquad (6.35)$$

Матрица R_h отличается от матрицы R на матрицу D^2. Следовательно, необходимо рассмотреть методы перехода от матрицы R с единицами на главной диагонали к матрице R_h. Существует большое число необходимых методов, однако мы остановимся только на тех, которые наиболее часто используются на практике.

6.1.5. Корреляционная матрица R_h с общностями на главной диагонали

Дисперсия нормированного признака равна единице. Следовательно, каждый показатель должен иметь общность h_j^2, меньшую единицы. Это

видно из табл.6.1. Значит, в факторном анализе на главной диагонали корреляционной матрицы не могут стоять единицы, а должны быть общности. Из табл.6.1 ясно видно, что элементы матрицы D определяются однозначно при известном значении h_j^2. А поскольку в начале исследования значения h_j^2 не известны, то приходится находить оценки элементов главной диагонали редуцированной матрицы R_h.

Существует два различных подхода. **В первом подходе** определяют общности, а затем число факторов. Такой подход используется в методе главных факторов, центроидном методе, методе треугольной декомпозиции по матрице R_h. **Во втором подходе** предварительно устанавливается количество факторов m, которое должно быть выделено, а затем подбирают значения общностей так, чтобы ранг матрицы R_h приближался к m. Предварительная оценка числа общих факторов используется в методах: максимального правдоподобия и максимальных остатков. Однако в настоящее время из рассматриваемых многомерных методов статистического анализа в основном используются **метод главных компонент и метод главных факторов.** Последний метод представляет факторный анализ, при котором вначале определяется общность, а потом число факторов. Остановимся на методе оценки общностей, необходимых для образования исходной матрицы R_h.

Известно, что нижней границей оценки общности является квадрат множественного коэффициента корреляции j-й переменной с остальными $(n\text{-}1)$ признаками, а верхней границей оценки общности - квадрат коэффициента надежности:

$$R_{j.12...j...n} \le h_j^2 \le r_j^2 , \qquad (6.36)$$

где $j = \overline{1,n}$.

Лоули и Рао дали определение общностей как величин, которые при статистически значимых факторах позволяют наилучшим образом воспроизвести корреляционную матрицу. При решении практических задач исследователь отыскивает такие значения h_j^*, которые определяют общую дисперсию каждого признака и находятся в пределах (6.36).

В настоящее время исследователь зависит от метода, заложенного в доступной ему программе. Однако в ряде случаев, когда число переменных невелико, приходится вникать и в возможные расчеты h_j^*. Если же число переменных более 10-20, то методы оценки общностей не могут оказать существенного влияния на конечный результат, подлежащий содержательной интерпретации.

Рассмотрим методы определения общности h_j^*.

Определение h_j^2 при помощи квадрата коэффициента множественной корреляции.

По данным Иберла известно, что с ростом числа переменных при постоянном числе факторов нижняя граница оценки общности (6.36) сходится к истинному значению общности. Следовательно, метод квадрата коэффициента множественной корреляции является не только практическим, но и наиболее обоснованным теоретически. Привлекательной стороной этого метода является возможность вычисления его при помощи обратной корреляционной матрицы R^{-1}.

$$R_{j.12...j...n} = 1 - \frac{1}{r^{jj}}, \qquad (6.37)$$

где r^{jj} - диагональный элемент обратной матрицы R^{-1}.

Определение h_j^2 при помощи наибольшего коэффициента корреля-ции по строке (столбцу).

При большом числе переменных часто используется наибольший коэффициент корреляции данной переменной с остальными переменными. Теоретически этот подход не обоснован, но имеет весьма широкое распространение благодаря своей простоте. Наибольший коэффициент корреляции не связан непосредственно с общностью. Как отмечалось выше, он является случайной величиной. Разумеется, что данный метод дает значение общности, которое может оказаться выше или ниже ее истинного значения. Однако если число признаков окажется выше 20, то конечные результаты незначительно отличаются от результатов, полученных при более точных методах.

Суть метода заключается в том, что в строке матрицы R, соответствующей данному признаку, выбирается элемент с наибольшим абсолютным значением. Это наибольшее значение коэффициента корреляции записывается на главной диагонали (со знаком плюс, разумеется).

Метод оценки h_j^2 при помощи среднего коэффициента корреляции по строке (столбцу).

Полученное значение должно оказаться меньше наибольшего коэффициента корреляции.

$$h_j^2 = \frac{1}{n}\sum_{k=1}^{n} r_{jk}, \qquad (6.38)$$

где $j \neq k$.

Метод триад для оценки h_j^2.

При данном методе в j-й строке (столбце) матрицы R отыскиваются два наибольших значения коэффициентов корреляции r_{jk} и r_{jl} и составляется триада

$$h_j^2 = \frac{r_{jk} r_{jl}}{r_{kl}}, \qquad (6.39)$$

где r_{jk} - коэффициент корреляции в строке j, имеющий наибольшее значение стохастической связи между признаком y_j и переменной y_k;

r_{jl} - коэффициент корреляции, имеющий наибольшее значение, не превосходящее r_{jk}, характеризующий связь признака y_j с y_l.

Таким образом, индексы k и l относятся к k-му и l-му параметрам, коэффициенты корреляции которых с признаком превосходят все остальные коэффициенты корреляции в j-й строке матрицы R; r_{kl} - коэффициент корреляции между признаками y_k и y_l.

При помощи этого метода усиливается влияние наибольших коэффициентов корреляции.

Использование метода первого центроидного фактора для определения h_j^2.

На главной диагонали матрицы R ставятся наибольшие коэффициенты корреляции каждой строки (столбца). По новой матрице вычисляется отношение квадрата суммы элементов соответствующей строки (столбца) к сумме всех элементов матрицы. При этом не нужно менять направление признаков как при построении первого центроида.

$$\hat{h}_j^2 = \frac{\left(\sum_{k=1}^{n} r_{jk}\right)^2}{\sum_{k=1}^{n}\sum_{l=1}^{n} r_{kl}}. \qquad (6.40)$$

Данной формулой дается оценка общности снизу.

Пример 6.3

С целью изучения помощи факторного анализа взаимосвязей между восемью показателями по данным N=305 наблюдений была построена матрица парных коэффициентов корреляции R. Требуется определить $\hat{h}_j^2$, используя следующие методы:

1) наибольшего элемента по строке (столбцу);
2) среднего коэффициента корреляции;
3) триад;
4) первого центроидного фактора.

Составить таблицу и построить график по полученным результатам. Сравнить полученные результаты с данными Хармана.[1] Матрица парных коэффициентов корреляции R представлена в виде треугольной, хотя она является симметрической.

[1] См.: *Харман Г.* Современный факторный анализ / Пер. с англ. - М.: Статистика, 1972.

x_j \ x_k	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	1							
x_2	0,846	1						
x_3	0,805	0,881	1					
x_4	0,859	0,826	0,801	1				
x_5	0,473	0,376	0,380	0,436	1			
x_6	0,398	0,326	0,319	0,329	0,762	1		
x_7	0,301	0,277	0,237	0,327	0,730	0,583	1	
x_8	0,382	0,415	0,345	0,365	0,629	0,577	0,539	1

Решение

1. Запишем матрицу R_h, используя метод наибольшего элемента по строке (столбцу). Для этого переписывается матрица R без ее диагональных элементов. В качестве диагональных элементов проставляются наибольшие элементы по строке:

а) для начала возьмем первый столбец, который совпадает с первой строкой, так как матрица R является симметрической. В этом столбце поставим вместо h_j^2 его оценку $\hat{h}_j^2 = r_{41} = r_{14} = 0,859$, которую для отличия от внедиагональных элементов поместим в скобки;

б) определим теперь $\hat{h}_4^2$. Для этого просматриваем элементы четвертой строки до 1 и элемент четвертого столбца после 1, так как эти элементы совпадают с отсутствующими элементами четвертой строки. Наибольшим является $r_{41}=r_{14}=0,859$. Повторяя эту операцию все восемь раз, получим следующую матрицу R_h.

$$R_h = \begin{pmatrix} (0,856) & & & & & & & \\ 0,846 & (0,881) & & & & & & \\ 0,805 & 0,881 & (0,881) & & & & & \\ 0,859 & 0,826 & 0,801 & (0,859) & & & & \\ 0,473 & 0,376 & 0,380 & 0,436 & (0,762) & & & \\ 0,398 & 0,326 & 0,319 & 0,329 & 0,762 & (0,762) & & \\ 0,301 & 0,277 & 0,237 & 0,327 & 0,730 & 0,583 & (0,730) & \\ 0,382 & 0,415 & 0,345 & 0,365 & 0,629 & 0,577 & 0,539 & (0,629) \end{pmatrix}$$

2. Определим $\hat{h}_j^2$ методом среднего коэффициента корреляции (без диагонального элемента матрицы R):

а) найдем среднюю элементов первой строки без диагонального элемента:

$$\frac{1}{7}(0,846+0,805+0,859+0,473+0,398+0,301+0,382)=\frac{1}{7}\cdot 4,064 \approx 0,581;$$

б) определим суммы коэффициентов корреляции по остальным семи строкам:

$$7\hat{h}_2^2=3,947; \quad 7\hat{h}_5^2=3,786; \quad 7\hat{h}_8^2=3,252;$$
$$7\hat{h}_3^2=3,768; \quad 7\hat{h}_6^2=3,294;$$
$$7\hat{h}_4^2=3,943; \quad 7\hat{h}_7^2=2,994;$$

в) сведем значения $\hat{h}_j^2$ в таблицу (табл.6.6).

Таблица 6.6

Значения $\hat{h}_j^2$, полученные методом среднего коэффициента корреляции

	$\hat{h}_1^2$	$\hat{h}_2^2$	$\hat{h}_3^2$	$\hat{h}_4^2$	$\hat{h}_5^2$	$\hat{h}_6^2$	$\hat{h}_7^2$	$\hat{h}_8^2$
Значение $\hat{h}_j^2$	0,581	0,564	0,538	0,563	0,541	0,471	0,428	0,465

3. Определим метод триад $\hat{h}_j^2$:

а) выберем в первом столбце матрицы два наибольших коэффициента корреляции. Это $r_{14}=0,859=r_{jk}$ и $r_{12}=0,846=r_{ji}$;

б) найдем $r_{kl}=r_{42}=0,826$;

в) определим $\hat{h}_1^2 = \dfrac{0,859\cdot 0,846}{0,826} = 0,880$ при условии, если при расчете получается $\hat{h}_j^2 \geq 1$, то общность $\hat{h}_j^2 = 1$.

Составим таблицу значений r_{jk}, r_{ji}, r_{kl}, $\hat{h}_j^2$ (табл.6.7).

Таблица 6.7

Значения r_{jk}, r_{ji}, r_{kl}, $\hat{h}_j^2$

Показатели \ $\hat{h}_j^2$	$\hat{h}_1^2$	$\hat{h}_2^2$	$\hat{h}_3^2$	$\hat{h}_4^2$	$\hat{h}_5^2$	$\hat{h}_6^2$	$\hat{h}_7^2$	$\hat{h}_8^2$
r_{jk}	0,859	0,881	0,881	0,859	0,762	0,762	0,730	0,629
r_{ji}	0,846	0,846	0,805	0,826	0,730	0,583	0,583	0,577
r_{kl}	0,826	0,805	0,846	0,846	0,583	0,730	0,762	0,762
$\hat{h}_j^2$	0,880	0,926	0,838	0,839	0,954	0,609	0,559	0,476

4. Используя метод первого центроидного фактора найдем $\hat{h}_j^2$:

а) методом выбора наибольшего элемента составим матрицу (1-й пункт данного примера);

б) определим сумму корреляционных коэффициентов матрицы (из пункта 1 данного примера)

$$\sum_{j,k=1}^{n} r_{jk} = \sum_{j=1}^{n} r_{jj} + \sum_{j,k=1}^{n} r_{kj} = 0,859+0,881+0,881+0,859+0,762+0,762+$$
$$+0,730+0,623+2(0,846+0,805+...+0,577+0,539)=6,363+2(4,064+$$
$$3,101+2,082+1,457+2,121+1,16+0,539)=6,363+29,048=35,411;$$

в) квадраты сумм коэффициентов столбцов для $\hat{h}_j^2$

$$(\sum_{k=1}^{8} r_{jk})^2 = (4{,}064+0{,}859)^2 = (4{,}923)^2 = 24{,}239.$$

Отсюда

$$\hat{h}_1^2 = \frac{(\sum_{k=1}^{8} r_{jk})^2}{\sum_{j,k=1}^{8} r_{jk}} = \frac{24{,}236}{35{,}411} = 0{,}684.$$

Составим таблицу значений элементов столбцов, их квадратов и $\hat{h}_j^2$ (табл.6.8).

Таблица 6.8

Значения квадратов сумм и оценок общностей

Показатели \ $\hat{h}_j^2$	$\hat{h}_1^2$	$\hat{h}_2^2$	$\hat{h}_3^2$	$\hat{h}_4^2$	$\hat{h}_5^2$	$\hat{h}_6^2$	$\hat{h}_7^2$	$\hat{h}_8^2$
$\sum_{k=1}^{n} r_{jk}$	4,923	4,828	4,649	4,802	4,548	4,056	3,734	3,881
$(\sum_{k=1}^{n} r_{jk})^2$	24,236	23,310	21,613	23,059	20,684	16,451	13,943	15,062
$\hat{h}_j^2$	0,684	0,658	0,610	0,651	0,584	0,465	0,394	0,425

Для сравнения полученных автором расчетов общности со средними оценками Хармана составим таблицу и построим графики (табл.6.9, рис.6.3).

Таблица 6.9

Результаты расчетов $\hat{h}_j^2$ разными методами

№ п/п	Метод	Результаты							
		$\hat{h}_1^2$	$\hat{h}_2^2$	$\hat{h}_3^2$	$\hat{h}_4^2$	$\hat{h}_5^2$	$\hat{h}_6^2$	$\hat{h}_7^2$	$\hat{h}_8^2$
1	Наибольшего элемента	0,859	0,881	0,881	0,859	0,762	0,762	0,730	0,629
2	Среднего коэффициента корреляции	0,581	0,564	0,538	0,563	0,541	0,471	0,428	0,465
3	Триад	0,880	0,926	0,838	0,839	0,954	0,609	0,559	0,576
4	Первого центроидного фактора	0,684	0,658	0,610	0,651	0,584	0,465	0,394	0,425
5	Средние оценки Хармана	0,842	0,881	0,817	0,815	0,872	0,647	0,584	0,502

Методы среднего коэффициента корреляции и первого центроидного фактора дали заниженные результаты во всех точках по сравнению с методом наибольшего элемента по строке. Методом триад получены три точки, расположенные ниже, чем на графике для первого метода.

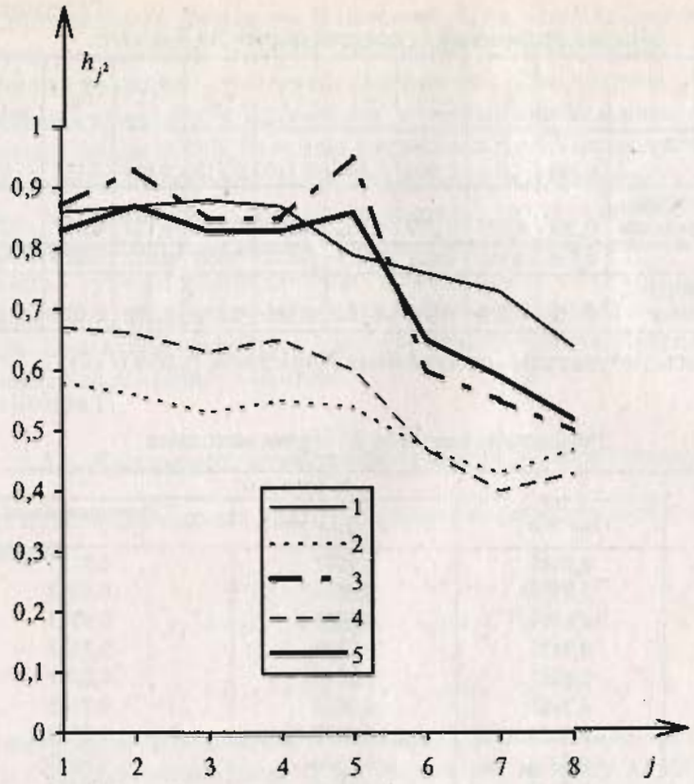


Рис 6.3. Сравнение расчетов общности со средними оценками Хармана

Интересно отметить, что в данном примере наиболее близкими к средним оценкам Хармана оказались оценки, полученные по триадам. Кроме того только одна оценка $\hat{h}_5^2$ у Хармана оказалась выше результата, полученного при помощи максимального элемента по строке. Все остальные точки на рис.6.3 лежат ниже графика.

Если для характеристики взять модули отклонений в восьми точках, то получим следующий результат (табл.6.10).

Из табл. 6.10 видно, что в данном примере по сумме модулей отклонений ближе всего к средним значениям Хармана находится метод триад, за ним следует метод максимальных элементов по строке (столбцу). Существенно заниженные значения дают методы среднего значения по столбцу и первого центроидного фактора (1,809 и 1,486 соответственно). Можно привести ряд примеров использования других методов оценки общностей, которые дают близкие результаты с методом максимального коэффициента корреляции в строке.

Таблица 6.10

Модули отклонений от средних оценок по Харману

Метод	$\hat{h}_1^2$	$\hat{h}_2^2$	$\hat{h}_3^2$	$\hat{h}_4^2$	$\hat{h}_5^2$	$\hat{h}_6^2$	$\hat{h}_7^2$	$\hat{h}_8^2$	$\sum_{j=1}^{8}\lvert\Delta_j\rvert$
Наибольшего элемента	0,017	0	0,064	0,044	0,110	0,115	0,146	0,127	0,623
Среднего коэффициента корреляции	0,261	0,317	0,279	0,252	0,331	0,176	0,156	0,037	1,809
Триад	0,038	0,145	0,021	0,024	0,082	0,038	0,025	0,026	0,299
Первого центроидного фактора	0,158	0,223	0,207	0,164	0,288	0,179	0,190	0,077	1,486

Приведем результаты, полученные Харманом (табл.6.11).

Таблица 6.11

Результаты расчетов $\hat{h}_j^2$ тремя методами

Признак y_j	Методы		
	$r_{max} = \hat{h}_j^2$	$R_{j.12\ldots)\,j(\ldots n} = \hat{h}_j^2$	Итерационный
y_1	0,9082	0,9077	0,9135
y_2	0,8388	0,8332	0,8393
y_3	0,8919	0,8909	0,9023
y_4	0,7422	0,7308	0,7427
y_5	0,8897	0,8849	0,8907
y_6	0,7165	0,7052	0,7185
y_7	0,9154	0,9138	0,9242
y_8	0,8373	0,8344	0,8439
y_9	0,7179	0,6902	0,6933
y_{10}	0,8784	0,8875	0,8961
y_{11}	0,9306	0,9310	0,9394
y_{12}	0,8354	0,8326	0,8392

Следует отметить, что в итерационном методе в качестве первого значения для определения $\hat{h}_j^2$ использовалось значение множественного коэффициента корреляции.

Таким образом, анализ табл.6.11 и рис.6.3 показывает возможности использования ряда практических методов получения оценок общностей. Однако до сих пор не разработано теоретическое обоснование для выбора наиболее приемлемого метода. В то же время широкая практика использования факторного анализа показала, что при числе переменных, большем $10 \div 20$, ошибки в определении $\hat{h}_j^2$ незначительно влияют на результаты содержательной интерпретации. В экономических исследованиях число признаков обычно превосходит $10 \div 20$. При числе признаков, меньшем 10, реже обращаются к факторному анализу.

Итак, существует ряд методов преобразования корреляционной матрицы R к корреляционной матрице R_h, идентичной матрице R, за исключением диагональных элементов. В матрице R_h на главной диагонали должны стоять не единицы, а только доли единицы, приходящиеся на дисперсию общих факторов в дисперсии переменной. Эта матрица R_h служит основой для проведения дальнейших преобразований одним из методов факторного анализа. При изучении возможных путей перехода от матрицы R к матрице R_h нами была рассмотрена проблема общностей в факторном анализе. Мы убедились в том, что определить точное значение общностей не представляется возможным. Существует возможность найти оценку общности - сумму дисперсий общих факторов, которая оказывается меньше единицы. В общем случае общность является частью дисперсии переменной. Она может равняться единице только при отсутствии у данной переменной характерного фактора.

6.1.6. Факторное отображение и факторная структура

Из модели факторного анализа вытекает следующая система линейных уравнений.

$$\begin{cases} y_1 = a_{11} f_1 + a_{12} f_2 + \ldots + a_{1m} f_m + d_1 v_1 \\ y_2 = a_{21} f_1 + a_{22} f_2 + \ldots + a_{2m} f_m + d_2 v_2 \\ \ldots \\ y_n = a_{n1} f_1 + a_{n2} f_2 + \ldots + a_{nm} f_m + d_n v_n \end{cases} \qquad (6.41)$$

Факторным отображением называется система линейных уравнений (6.41). Отображением можно было бы назвать матрицу M (6.11), играющую роль линейного оператора преобразования, отображающего полное факторное пространство в пространство признаков. Но d_j однозначно определяется общностью признака h_j^2 (табл.6.1), поэтому факторным отображением называется и ее подматрица A, состоящая только из весовых коэффициентов при общих факторах.

В факторном анализе общие факторы могут быть коррелированными и некоррелированными. Характерные факторы предполагаются некоррелированными между собой и общими факторами.

Из фундаментальной теоремы факторного анализа (6.33) видно, что корреляционная матрица R может быть воспроизведена при помощи факторного отображения (матрицы A) и матрицы C, содержащей коэффициенты корреляции между факторами. Следовательно, если общие факторы не ортогональны между собой (не коррелированны), и $C \neq E$, то коэффициенты корреляции между признаками (исходными переменными) и общими факторами будут воспроизводиться при помощи произведения двух матриц: A и C.

Обычно это произведение обозначается

$$S = V_{fs} = AC. \qquad (6.42)$$

Матрица коэффициентов корреляции S между параметрами и факторами называется **факторной структурной**.

Раскроем связи между факторной структурной и факторным отображением.

Для этого обратимся к модели факторного анализа

$$y_j = a_{j1} f_1 + a_{j2} f_2 + \ldots + a_{jm} f_m + d_j v_j .$$

Используя данное выражение для нормированных случайных величин, получим коэффициент корреляции между j-м признаком и первым общим фактором. Для этого сумму произведений двух нормированных переменных y_j и f_i по всем N объектам следует разделить на N левую и правую часть.

$$r_{y_j f_1} = \frac{1}{N}\sum_{i=1}^{N} y_{ji} f_{1i} = \frac{1}{N}\left(a_{j1}\sum_{i=1}^{N} f_{1i} f_{1i} + a_{j2}\sum_{i=1}^{N} f_{2i} f_{1i} + \ldots + d_j \sum_{i=1}^{N} v_{ij} f_{1i}\right).$$

Первый член в скобках, деленный на N,

$$a_{j1}\frac{1}{N}\sum_{i=1}^{N} f_{1i}^2 = a_{j1} .$$

Второй член

$$a_{j2}\frac{1}{N}\sum_{i=1}^{N} f_{2i} f_{1i} = a_{j2} r_{f_2 f_1} .$$

Последний член

$$d_j r_{f_1 v_j} = d_j 0 = 0 .$$

По аналогии выпишем систему уравнений

$$\begin{cases} r_{y_j f_1} = a_{j1} + a_{j2} r_{f_1 f_2} + \ldots + a_{jr} r_{f_1 f_r} + \ldots + a_{jm} r_{f_1 f_m} \\ r_{y_j f_2} = a_{j1} r_{f_2 f_1} + a_{j2} + \ldots + a_{jr} r_{f_2 f_r} + \ldots + a_{jm} r_{f_2 f_m} \\ \ldots \\ r_{y_j f_r} = a_{j1} r_{f_r f_1} + a_{j2} r_{f_r f_2} + \ldots + a_{jr} + \ldots + a_{jm} r_{f_r f_m} \\ \ldots \\ r_{y_j f_m} = a_{j1} r_{f_m f_1} + a_{j2} r_{f_m f_2} + \ldots + a_{jr} r_{f_m f_r} + \ldots + a_{jm} \\ r_{y_j v_j} = d_j \end{cases} \qquad (6.43)$$

Последнее равенство (6.43) вытекает из

$$d_j = \frac{1}{N}\sum_{i=1}^{N} V_{ji} V_{ji} = d_j .$$

Система уравнений (6.43) есть факторная структура. Следовательно, и матрицу

$$S_0 = \begin{pmatrix} S_{11} & S_{12} & \ldots & S_{1m} & d_1 & 0 & 0 & \ldots & 0 \\ S_{21} & S_{22} & \ldots & S_{2m} & 0 & d_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ S_{n1} & S_{n2} & \ldots & S_{nm} & 0 & 0 & 0 & \ldots & d_n \end{pmatrix}, \qquad (6.44)$$

где $S_{jr} = S_{y_j f_r}$, также называют факторной структурой.

Так как у характерных факторов коэффициент корреляции между ними и признаком всегда равен коэффициенту при характерном факторе в соответствующем уравнении факторного отображения, то матрицей (6.44) не пользуются. Факторной структурой называют также **матрицу коэффициентов корреляции признаков с общими факторами**, т.е. **матрицу, у которой отсутствует диагональная подматрица *D*.**

$$\hat{S} = \begin{pmatrix} S_{11} & S_{12} & \ldots & S_{1m} \\ S_{21} & S_{22} & \ldots & S_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ S_{n1} & S_{n2} & \ldots & S_{nm} \end{pmatrix}. \qquad (6.45)$$

Если же общие факторы не коррелированны между собой, то согласно системе уравнений (6.43) в правой части все коэффициенты корреляции между факторами будут равны нулю. Тогда

$$S_{jr} = r_{y_j f_r} = a_{jr} . \qquad (6.46)$$

Таким образом, если характерные факторы отсутствуют и общие факторы не коррелированны между собой, то справедливо (6.46) и факторная структура совпадает с факторным отображением

$$S=A, \qquad (6.47)$$

так как матрица $C=E$.

В этом случае полное решение может быть получено непосредственно из факторного отображения.

Во всех остальных случаях для проведения факторного анализа необходимо знать факторное отображение, выражающее линейные взаимосвязи признаков и факторов, а также факторную структуру, чтобы выявить существующие корреляционные связи между параметрами и факторами. Например, факторное отображение может быть использовано для воспроизведения корреляций и проверки значимости решения. Идеальным был бы случай, если бы выборочная матрица корреляции полностью совпала с матрицей корреляции, полученной из модели после решения факторной задачи. Но на практике такого совпадения не обнаруживается. Это в определенной степени объясняется преимуществами, которыми обладает теоретическая модель. При помощи модели изучаются не все закономерности, а только основные, наиболее существенные, так как при ее использовании

могут быть ошибки из-за неточности эксперимента, непредставительности выборки и т.д. Поэтому коэффициенты корреляции, полученные из исходных признаков, могут отличаться от тех, которые можно получить из модели r'_{jk}. Для получения коэффициента корреляции, воспроизведенного из модели, достаточно перемножить любые два признака y_j и y_k, затем просуммировать произведения по всем N объектам и разделить полученные результаты на N. Тогда, взяв два уравнения из системы (6.41), получим справа выражение для воспроизведенного коэффициента корреляции r_{jk} после приведения подобных членов

$$r'_{jk} = a_{j1}a_{k1} + a_{j2}a_{k2} + \ldots + a_{jm}a_{km} + (a_{j1}a_{k2} + a_{k1}a_{j2})r_{f_1f_2} + \ldots$$
$$\ldots + (a_{j1}a_{km} + a_{k1}a_{jm})r_{f_1f_m} + \ldots + a_{j1}d_k r_{f_1v_k} + a_{k1}d_j r_{f_1v_j} + \ldots \quad (6.48)$$
$$\ldots + a_{jm}d_k r_{f_mv_k} + a_{km}d_j r_{f_mv_j} + d_j d_k r_{v_jv_k}.$$

Если характерные факторы не коррелированны между собой и с общими факторами, то

$$r_{f_rv_j} = r_{f_rv_k} = r_{v_jv_k} = 0.$$

Из (6.48) станут нулевыми все члены, начиная с $a_{j1}d_k r_{f_1v_k}$. Если при этом и общие факторы не коррелированны, то $r_{f_rf_l} = 0$ $(r \neq l;\ l,r = \overline{1,m})$.

Из выражения (6.48) получим

$$r'_{jk} = a_{j1}a_{k1} + a_{j2}a_{k2} + \ldots + a_{jm}a_{km}. \quad (6.49)$$

где $j \neq k;\ j,k = \overline{1,n}$.

Данный коэффициент может отличаться от выборочного коэффициента корреляции r_{jk}, полученного из эксперимента. Это отличие может быть оценено разностью

$$r_{jko} = r_{jk} - r'_{jk}, \quad (6.50)$$

где r_{jko} называют **остаточным коэффициентом корреляции.**

По матрице остатков можно судить о близости редуцированных коэффициентов корреляции к полученным коэффициентам из эксперимента. Значения остаточных коэффициентов корреляции должны быть близки к нулю. Следовательно, можно полагать, что остатки будут распределены как выборка равного объема с нулевой корреляцией. Если факторы некоррелированны, то среднее квадратическое отклонение выборки с нулевой корреляцией определится

$$s_{ro} = \frac{1}{\sqrt{N-1}}. \quad (6.51)$$

Подобный критерий является грубым приближением. В данном выражении учитывается только объем выборки N, но не учитывается количество признаков. Несмотря на это, благодаря простоте, метод факторного отображения используется на практике.

Если получено

$$s_{ro} >> \frac{1}{\sqrt{N-1}}, \quad (6.52)$$

то это означает, что какие-то связи между признаками остались недоучтенными.

Если получено

$$s_{ro} << \frac{1}{\sqrt{N-1}}, \quad (6.53)$$

то надо полагать, что какие-то связи между признаками оказались несущественными. Надо увеличить число признаков.

Если же s_{ro} незначительно меньше критерия для выборки с нулевой корреляцией, то факторное решение можно считать допустимым по данному критерию.

6.1.7. Пространство общих факторов и полное факторное пространство

Пространство общих факторов

Проанализируем r_{jk} в пространстве общих факторов. В факторном анализе рассматривается пространство общих факторов и полное факторное пространство. Рассмотрим пространство общих факторов, пространство наименьшей размерности, в котором можно представить n переменных в виде векторов.

Пространство с меньшей размерностью не включило бы все переменные. Какова же минимально необходимая размерность для представления всех переменных в пространстве? Эта размерность соответствует рангу корреляционной матрицы или рангу матрицы исходных данных (X или Y).

При графическом изображении координатными осями пространства общих факторов являются столбцы матрицы весовых коэффициентов A. Значения весовых коэффициентов могут изменяться от –1 до 1.

Переменные можно отразить в пространстве общих факторов. Определим длину вектора переменной:

$$d_j^{\Phi} = \sqrt{\sum_{r=1}^{m} a_{jr}^2} = h_j. \quad (6.54)$$

Длина вектора переменной равна корню квадратному из общности. Квадрат длины вектора определяет суммарную дисперсию общих факторов в единичной дисперсии нормированного признака.

В пространстве общих факторов косинус угла между двумя векторами-признаками определяет коэффициент корреляции между ними.

$$r_{jk}^{\Phi} = \cos(\varphi_{jk}^{\Phi}) = \frac{1}{h_j h_k} \sum_{r=1}^{m} a_{jr} a_{kr} . \qquad (6.55)$$

Пример 6.4

Использовав данные примера 6.1 построить график, на котором отобразить признаки в пространстве общих факторов.

Решение

В примере 6.1 матрица состоит из двух общих факторов, в пространстве которых графически представим исходные признаки (рис.6.3).

Координаты первого признака представляют собой соответствующие весовые коэффициенты с первым и вторым общими факторами - y_1(0,90; –0,30) - и являются первой строкой матрицы A. Седьмой признак - y_7(0,20; 0,80) - последняя строка матрицы A.

Пример 6.5

Используя данные примера 6.1 определить в пространстве общих факторов коэффициенты корреляции между признаками: а) пятым и шестым; б) первым и седьмым.

Решение

1. Найдем h_5 и h_6.

$$h_5^2 = 0{,}5^2 + 0{,}5^2 = 0{,}5; \quad h_5 = \sqrt{0{,}5} = 0{,}707;$$

$$h_6^2 = 0{,}3^2 + 0{,}6^2 = 0{,}45; \quad h_6 = \sqrt{0{,}45} \approx 0{,}671.$$

2. Определим r_{56}^{Φ}.

$$r_{56}^{\Phi} = \frac{1}{h_5 h_6} \sum_{r=1}^{2} a_{jr} a_{kr} ;$$

$$r_{56}^{\Phi} = \frac{1}{0{,}707 \cdot 0{,}671}(a_{51}a_{61} + a_{52}a_{62}) = 2{,}11[0{,}50(-0{,}30) + 0{,}50(0{,}60)] =$$

$$= 2{,}11(-0{,}15 + 0{,}30) \approx 0{,}317.$$

3. $h_1 = \sqrt{0{,}9^2 + 0{,}3^2} = \sqrt{0{,}90} \approx 0{,}949$;

$h_2 = \sqrt{0{,}2^2 + 0{,}8^2} = \sqrt{0{,}68} \approx 0{,}825$;

$$r_{17}^{\Phi} = \frac{1}{0{,}949 \cdot 0{,}825}[0{,}90 \cdot 0{,}20 + (-0{,}30)0{,}80] \approx 1{,}28(-0{,}06) \approx -0{,}077 .$$

Перейдем к представлению d_j и r_{jk} в полном факторном пространстве.

Полное факторное пространство

Данное пространство имеет размерность $(m+n)$. Осями координат будут m общих факторов и n характерных факторов. Таким образом, оно включает все факторы: общие и индивидуальные. Следовательно, пространство

общих факторов является подпространством полного факторного пространства.

С точки зрения удобства содержательной интерпретации пространство общих факторов обладает существенными преимуществами.

С теоретической точки зрения важно обратить внимание на представление вектора-признака в полном факторном пространстве, где общие факторы взаимно не коррелированны, а характерные, - не коррелированны ко всем факторам.

$$S_j = \sqrt{a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2 + d_j^2} = 1.$$

Итак, в отличие от пространства общих факторов, в полном факторном пространстве концы всех векторов-признаков лежат на поверхности $(n+m)$-мерного гипершара с радиусом, равным единице.

В полном факторном пространстве косинус угла между двумя векторами-переменными, т.е. коэффициент корреляции, равен скалярному произведению этих векторов. Коэффициент корреляции может быть выражен формулой

$$r_{jk}^{\Pi} = \cos(\varphi_{jk}^{\Pi}) = \sum_{r=1}^{m} a_{jr} a_{kr} ,$$

что соответствует скалярному произведению векторов a_j и a_k.

Пример 6.6

На основе матрицы A (пример 6.1) представить матрицу M, заданную в (6.7), и определить коэффициенты корреляции между признаками: а) пятым и шестым; б) первым и седьмым.

Решение

1. Представим матрицу M.

Для этого к матрице A надо приписать подматрицу D, элементами которой будут d_j.

$$M = \begin{pmatrix} 0{,}90 & -0{,}30 & 0{,}316 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0{,}80 & -0{,}30 & 0 & 0{,}520 & 0 & 0 & 0 & 0 & 0 \\ 0{,}60 & 0{,}30 & 0 & 0 & 0{,}742 & 0 & 0 & 0 & 0 \\ 0{,}50 & 0{,}20 & 0 & 0 & 0 & 0{,}843 & 0 & 0 & 0 \\ 0{,}50 & 0{,}50 & 0 & 0 & 0 & 0 & 0{,}707 & 0 & 0 \\ -0{,}30 & 0{,}60 & 0 & 0 & 0 & 0 & 0 & 0{,}742 & 0 \\ 0{,}20 & 0{,}80 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}566 \end{pmatrix}.$$

Значения d_j получены извлечением квадратного корня из элементов столбца d_j^2 (табл.6.2).

2. Определим $r_{jk}^{\text{п}} = \sum_{r=1}^{m} a_{jr} a_{kr}$ при j=5, k=6.

$$a_{51}\, a_{61} + a_{52}\, a_{62} = 0{,}50(-0{,}30)+0{,}50(0{,}60)=0{,}15.$$

3. Найдем $r_{jk}^{\text{п}}$ при j=1, k=7.

$$r_{17}^{\text{п}} = a_{11}\, a_{71} + a_{12}\, a_{72}=0{,}90(0{,}20)+(-0{,}30)0{,}80=0{,}18-0{,}24=-0{,}06.$$

Таким образом, содержательную интерпретацию удобней и проще проводить в пространстве общих факторов, в котором еще можно представить n переменных в виде векторов.

6.1.8. Связь факторных решений, полученных разными методами

Как известно, факторный анализ в современном математическом обеспечении в основном представлен методом главных факторов, центроидным методом и методом максимума правдоподобия. Исследователь, использующий несколько методов для решения задачи снижения размерности, всегда будет интересоваться эквивалентностью решений.

Исходной матрицей вне зависимости от метода решений является матрица R_h с общностями на главной диагонали. Эту матрицу при вращении можно представить в бесконечно большом числе систем координат.

Пусть мы получили A-факторное решение одного метода и B-факторное решение другого метода. В этом случае всегда существует матрица преобразования T, при которой выполняется равенство

$$AT=B, \qquad (6.56)$$

если AA^T и BB^T не вырожденные.

Определим эту матрицу преобразования. Для этого умножим слева обе части равенства (6.56) на такое выражение, чтобы в левой стороне на T умножалась единичная матрица A^{-1} (если A - квадратная матрица). В факторном анализе $n>m$, поэтому матрица A является прямоугольной и не имеет обратной A^{-1}, а значит используется другое соотношение. Хорошо известно соотношение для любой матрицы

$$(A^T A)^{-1}(A^T A) = E\,. \qquad (6.57)$$

Следовательно, для получения T надо в (6.56) умножить слева обе части равенства на $(A^T A)^{-1} A^T$ и получить

$$T = (A^T A)^{-1} A^T B\,. \qquad (6.58)$$

Мы условились, что факторные решения ортогональны, поэтому матрица $(A^T A)$ является диагональной. У диагональной матрицы всегда существует и без труда определяется обратная матрица.

Таким образом, мы проиллюстрировали и обосновали возможность перевода одного результата в другой.

Рассмотрим пример, в котором сопоставимы результаты решения задачи центроидным методом и методом главных факторов.

Пример 6.7

Имеется два решения Хармана, представленные матрицами A и B, полученными соответственно методами главных факторов и центроидным методом по одной и той же матрице R_h.

$$A = \begin{pmatrix} 0{,}8619 & -0{,}2703 \\ 0{,}7448 & -0{,}3051 \\ 0{,}6815 & -0{,}1782 \\ 0{,}3477 & 0{,}7246 \\ 0{,}3785 & 0{,}5971 \\ 0{,}2339 & 0{,}4435 \end{pmatrix}; \qquad B = \begin{pmatrix} 0{,}7392 & -0{,}5190 \\ 0{,}6232 & -0{,}5076 \\ 0{,}5983 & -0{,}3735 \\ 0{,}5460 & 0{,}5875 \\ 0{,}5393 & 0{,}4588 \\ 0{,}3569 & 0{,}3537 \end{pmatrix}$$

Определим матрицу преобразования T и оценим расхождение между весовыми коэффициентами заданной матрицы B и вычисленной матрицы AT.

Решение

1. Получим произведение $(A^T A)$.

$$A^T A = \begin{pmatrix} 0{,}8619 & 0{,}7448 & 0{,}6815 & 0{,}3477 & 0{,}3785 & 0{,}2339 \\ -0{,}2703 & -0{,}3051 & -0{,}1782 & 0{,}7246 & 0{,}5971 & 0{,}4435 \end{pmatrix} \times$$

$$\times \begin{pmatrix} 0{,}8619 & -0{,}2703 \\ 0{,}7448 & -0{,}3051 \\ 0{,}6815 & -0{,}1782 \\ 0{,}3477 & 0{,}7246 \\ 0{,}3785 & 0{,}5971 \\ 0{,}2339 & 0{,}4435 \end{pmatrix} = \begin{pmatrix} 2{,}0809 & 0 \\ 0 & 1{,}2762 \end{pmatrix}.$$

Вычисленная диагональная матрица подтверждает, что матрица A, полученная методом главных факторов, является ортогональной.

На главной диагонали матрицы $(A^T A)$ стоят суммы квадратов компонент соответствующих векторов матрицы A. Норма вектора a_1 равна $\sqrt{2{,}0809} \approx 1{,}4425$. Норма вектора a_2 равна $\sqrt{1{,}2762} \approx 1{,}1297$. Скалярное произведение векторов $a_1^T a_2$ равно нулю.

2. Получим обратную матрицу $(A^T A)^{-1}$. Так как матрица $(A^T A)$ диагональная, то ее обратная матрица также является диагональной. Элементы

обратной матрицы равны единице, деленной на элементы исходной матрицы. Следовательно,

$$(A^T A)^{-1} = \begin{bmatrix} 0,4806 & 0 \\ 0 & 0,7836 \end{bmatrix},$$

так как $\frac{1}{2,0809} \approx 0,4806$; $\frac{1}{1,2762} \approx 0,7836$.

3. Вычислим матрицу $(A^T A)^{-1} A^T$.

$(A^T A)^{-1} A^T$ равна A^T, у которой элементы верхней строки умножены на 0,4806, а элементы нижней строки умножены на 0,7836, поэтому

$$(A^T A)^{-1} A^T = \begin{pmatrix} 0,4142 & 0,3579 & 0,3275 & 0,1671 & 0,1819 & 0,1124 \\ -0,2118 & -0,2391 & -0,1396 & 0,5678 & 0,4679 & 0,3475 \end{pmatrix}.$$

4. Определим матрицу T.

$$T = (A^T A)^{-1} A^T B \approx \begin{pmatrix} 0,9546 & -0,2976 \\ 0,2976 & 0,9546 \end{pmatrix}.$$

5. Получим матрицу AT.

$$AT = \begin{pmatrix} 0,8619 & -0,2703 \\ 0,7448 & -0,3051 \\ 0,6815 & -0,1782 \\ 0,3477 & 0,7246 \\ 0,3785 & 0,5971 \\ 0,2339 & 0,4435 \end{pmatrix} \times \begin{pmatrix} 0,9546 & -0,2976 \\ 0,2976 & 0,9546 \end{pmatrix} = \begin{pmatrix} 0,7423 & -0,5145 \\ 0,6202 & -0,5129 \\ 0,5975 & -0,3729 \\ 0,5476 & 0,5882 \\ 0,5390 & 0,4574 \\ 0,3554 & 0,3538 \end{pmatrix}.$$

6. Вычислим отклонения элементов матрицы AT от B. В качестве меры k принято отношение бóльшего элемента одной матрицы к соответственному мéньшему элементу другой матрицы: $k = \frac{b_{jr\,\max}}{b_{jr\,\min}}$. Для удобства в табл.6.12 проставлена величина $\Delta b_{jr} = k - 1$.

Так, при вычислении Δb_{jr} для b_{11}: $b_{11\,\max} = 0,7423$ стоит в матрице AT, а $b_{11\,\min} = 0,7392$ - в матрице B.

$$\Delta b_{11} = \frac{0,7423}{0,7392} - 1 = 1,004194 - 1 \approx 0,0042.$$

Для вычисления Δb_{jr} для коэффициента b_{21}: $b_{21\,\max}$ находится в матрице B, поэтому $\Delta b_{21} = \frac{0,6232}{0,6202} - 1 = 1,004837 - 1 \approx 0,0048$. Результаты расчетов сведем в табл. 6.12, в которой b_{jr} - весовой коэффициент j-й переменной на r-ом общем факторе; $\overline{\Delta}_r$ - среднее значение отклонений элементов r-го общего фактора; $\overline{B}$ - среднее значение отклонений элементов матриц AT и B; $\overline{B}$ - среднее значение отклонений элементов матрицы, в %.

Таблица 6.12

Отклонения значений соответствующих элементов матриц AT и B

Δb_{11} 0,0054	Δb_{21} 0,0048	Δb_{31} 0,0013	Δb_{41} 0,0029	Δb_{51} 0,0006	Δb_{61} 0,0042	Δb_{12} 0,0087	Δb_{22} 0,0104
Δb_{32} 0,0016	Δb_{42} 0,0012	Δb_{52} 0,0031	Δb_{62} 0,0003	$\overline{\Delta}_1$ 0,0032	$\overline{\Delta}_2$ 0,0042	$(\overline{B})$ 0,0037	$(\overline{B})$,% 0,37

Из табл.6.12 видно, что размах ошибок колеблется от 0,03 до 1,04%. Выше 0,54% наблюдаются отклонения у элементов b_{12} и b_{22}. Следовательно, можно допустить, что основной причиной наблюдаемых отклонений являются **ошибки округления.**

Харман выяснил, насколько могут различаться фактические значения общностей для набора из n показателей, при которых можно получать эквивалентные решения, если применяются различные методы факторного анализа. Для одного и того же набора данных из восьми признаков решение проводилось методами главных факторов, максимального правдоподобия и минимальных остатков.

Все факторные нагрузки совпали с точностью до третьего десятичного знака. Вклады факторов и значения общностей также различались в третьем десятичном знаке. Решения также оказались эквивалентными, несмотря на различия общностей. Условия эксперимента были более сложными, чем в примере, рассмотренном нами более подробно.

Поскольку разные факторные решения приводятся к эквивалентному виду, исследователь использует имеющийся метод в доступном ему пакете прикладных программ. При решении задач факторного анализа чаще используется метод главных факторов.

6.2. Метод главных факторов и его алгоритм

Метод главных факторов имеется в пакетах прикладных программ для ЭВМ.

1. Алгоритм метода главных факторов начинается с получения матрицы парных коэффициентов корреляции с единицами на главной диагонали.

2. Определяются общности и получают матрицу R_h с общностями **на главной диагонали.**

3. Определяют первый общий фактор при условии, чтобы его вклад в дисперсию процесса v_1 был максимальным.

Нам необходимо максимум вклада первого общего фактора в суммарную общность

$$v_1 = \overline{a}_1 \overline{a}_1 = \sum_{j=1}^{n} a_{j1}^2 = \max. \qquad (6.59)$$

Условия, при которых должен быть обеспечен *max* (6.59) выражается формулой

$$r_{jk} = \sum_{r=1}^{m} a_{jr} a_{kr}, (j,k = \overline{1,n}), \qquad (6.60)$$

где $r_{jj} = h_j^2$.

Функция (6.59) имеет n переменных a_{ji}; условия (6.60), ограничивающие коэффициенты a_{ji}, будут $\frac{n(n+1)}{2}$. **Воспользуемся методом множителей Лагранжа**. Обозначим множители Лагранжа $\mu_{jk} = \mu_{kj}$.

Построим функцию

$$2T = V_1 - \sum_{j,k=0}^{n} \mu_{jk} r_{jk}. \qquad (6.61)$$

Преобразуем ее, используя (6.60),

$$2T = V_1 - \sum_{j,k}^{n} \sum_{r=1}^{m} \mu_{jk} a_{jr} a_{kr}. \qquad (6.62)$$

Возьмем частные производные функции T по переменным a_{ji} и приравняем их нулю

$$\frac{\partial T}{\partial a_{j1}} = a_{j1} - \sum_{k=1}^{n} \mu_{jk} a_{k1} = 0. \qquad (6.63)$$

Получим частные производные по остальным переменным и также приравняем их к нулю

$$\frac{\partial T}{\partial a_{jr}} = -\sum_{k=1}^{n} \mu_{jk} a_{kr} = 0. \qquad (6.64)$$

Таким образом, получены две системы уравнений, которые могут быть объединены при помощи **символа Кронекера**: $\delta_{jr} = 1$, если $r=1$ и $\delta_{jr} = 0$, если $r \neq 1$.

При объединении получим

$$\frac{\partial T}{\partial a_{jr}} = \delta_{jr} a_{j1} - \sum_{k=1}^{n} \mu_{jr} a_{kr} = 0, \qquad (6.65)$$

где $r = \overline{1,m}$.

Умножим (6.65) на a_{j1} и просуммируем по j:

$$\delta_{jr} \sum_{j=1}^{n} a_{j1} - \sum_{j=1}^{n} \sum_{k=1}^{n} \mu_{jk} a_{j1} a_{kr} = 0. \qquad (6.66)$$

Используя (6.63), можно заключить, что $\sum_{j=1}^{n} \mu_{jk} a_{j1} = a_{k1}$ и, обозначив

$$\lambda_1 = \sum_{j=1}^{n} a_{j1}^2, \qquad (6.67)$$

перепишем (6.66) более коротко

$$\delta_{jr} \lambda_1 - \sum_{k=1}^{n} a_{k1} a_{kr} = 0. \qquad (6.68)$$

Полученное выражение теперь умножим на a_{jr} и просуммируем по r (при $r=1$, $a_{jr}=a_{j1}$)

$$a_{j1} \lambda_1 - \sum_{k=1}^{n} a_{k1} \left(\sum_{r=1}^{m} a_{jr} a_{kr} \right) = 0. \qquad (6.69)$$

Из (6.60), подставив в (6.69) r_{jk} получим

$$\sum_{k=1}^{n} r_{jk} a_{k1} - \lambda_1 a_{j1} = 0. \qquad (6.70)$$

Выражение (6.70) является **системой n однородных алгебраических уравнений относительно n неизвестных a_{ji}.** Необходимым и достаточным условием существования нетривиального решения является равенство нулю определителя матрицы коэффициентов этих уравнений.

Развернем систему (6.70), учитывая, что $r_{jj} = h_j^2$.

$$\begin{cases} (h_1^2 - \lambda) a_{11} + r_{12} a_{21} + \ldots + r_{1n} a_{n1} = 0 \\ r_{21} a_{11} + (h_2^2 - \lambda) a_{21} + \ldots + r_{2n} a_{n1} = 0 \\ \ldots \\ r_{n1} a_{11} + r_{n2} a_{21} + \ldots + (h_n^2 - \lambda) a_{n1} = 0 \end{cases} \qquad (6.71)$$

Определитель матрицы коэффициентов этой системы уравнений должен быть равен нулю

$$\begin{vmatrix} (h_1^2 - \lambda) & r_{12} & \ldots & r_{1n} \\ r_{22} & (h_2^2 - \lambda) & \ldots & r_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} & \ldots & r_{nn} \end{vmatrix} = 0. \qquad (6.72)$$

Уравнение (6.72) называется **характеристическим уравнением** или **вековым уравнением.** Левая часть уравнения (6.72) называется характеристическим многочленом. **Характеристический многочлен** представля-

ет собой полином n-й степени относительно λ. Все корни этого уравнения действительны.

Если вместо λ подставить простой корень, то ранг матрицы, для которой представлен определитель (6.72), уменьшится на единицу и будет равен *(n–1)*. Если же вместо λ подставить l-кратный корень, то ранг матрицы уменьшится на l и будет равен *(n–l)*.

Система уравнений и определитель (6.72) получены в результате оптимизации выражения (6.59). Системе (6.71) соответствует множество решений, пропорциональных одному частному решению. Частное решение должно удовлетворять (6.67), которое соответствует максимизируемой функции V_1 (6.59). Значит, V_1 равно наибольшему корню λ, который в (6.71) и (6.72) был записан без индекса 1.

Решив (6.71), наибольший корень λ_1 подставим в (6.72) и получим одно из возможных решений $\alpha_{11},\alpha_{21},\ldots,\alpha_{n1}$. Эти значения являются **элементами первого собственного вектора**, которому соответствует первое собственное число λ_1. Данные значения необходимо умножить на весовой коэффициент, определяющий отношение среднего квадратического отклонения первого общего фактора $\sqrt{\lambda_1}$ к корню из полученной (суммарной) общности

$$\sqrt{\alpha_{11}^2+\alpha_{21}^2+\ldots+\alpha_{n1}^2};$$

$$a_{j1}=\frac{\alpha_{j1}\sqrt{\lambda_1}}{\sqrt{\alpha_{11}^2+\alpha_{21}^2+\ldots+\alpha_{n1}^2}}. \qquad (6.73)$$

Таким образом, получен первый общий фактор, весовые коэффициенты которого обеспечили ему максимальный вклад в суммарную общность.

Следующим шагом является поиск второго общего фактора, который обеспечил бы максимум вклада в суммарную общность процесса.

Из примера 6.2 знаем, что матрица остатков R_1 равна

$$R_1=R_h-R^1=R_h-a_1a_1^T. \qquad (6.74)$$

Далее рассматривается естественный алгоритм оптимизации вклада второго общего фактора в суммарную общность матрицы R_1. Только вместо ограничений (6.60) мы получим другие ограничения. Они должны соответствовать коэффициентам корреляции, полученным из остаточной матрицы после исключения влияния первого фактора

$${}_1r_{jk}=r_{jk}-a_{j1}a_{k1}=a_{j2}a_{k2}+a_{j3}a_{k3}+\ldots+a_{jm}a_{km}. \qquad (6.75)$$

При данных ограничениях можно отыскивать максимум функции

$$V_2=a_{12}^2+a_{22}^2+\ldots+a_{n2}^2=\max. \qquad (6.76)$$

Однако оказывается, что решение задачи оптимизации для обеспечения максимального вклада первого фактора в суммарную дисперсию является вынужденной процедурой. В дальнейшем более простой будет процедура использования алгоритма метода главных компонент для получения второго и остальных *(m–2)* общих факторов из матрицы R_1.

Итак, получен первый общий фактор, у которого дисперсия λ_1 будет максимальной. Получили матрицу остатков R_1, из которой было исключено влияние первого общего фактора.

При переходе к процедуре метода главных компонент надо убедиться в том, что собственные значения и собственные векторы корреляционной матрицы R_h и собственные значения и собственные векторы матрицы R_1 будут совпадать. Обозначим нормированные собственные векторы матрицы R_h через $a_1,a_2,\ldots,a_m$. Необходимо проверить, не являются ли они собственными векторами матрицы остатков R_1.

Умножим левую и правую части (6.74) на a_r справа

$$\begin{cases} R_1a_r=(R_h-a_1a_1^T)a_r \\ R_1a_r=R_ha_r-a_1a_1^Ta_r \end{cases}. \qquad (6.77)$$

Из определения собственного вектора следует, что $R_ha_r=\lambda_ra_r$, поэтому можно переписать (6.77) так:

$$R_1a_r=\lambda_ra_r-a_1a_1^Ta_r. \qquad (6.78)$$

Рассмотрим случай, когда $r=1$. При этом из (6.68) вытекает, что $\delta_{j1}=1$, $a_{kr}=a_{k1}$, поэтому $\lambda_1=\sum_{k=1}^{n}a_{k1}a_{k1}=a_1^Ta_1$.

Следовательно,

$$R_1a_1=\lambda_1a_1-\lambda_1a_1=0. \qquad (6.79)$$

Итак, собственный вектор матрицы R_h, обеспечивающий максимум дисперсии первому общему фактору, является собственным вектором и матрицей остатков R_1, а его собственное значение равно нулю.

Обратимся к другому случаю, когда $r\neq 1$. Согласно выражению (6.68) $\delta_{jr}=0$ и произведение $a_1^Ta_r=0$. Используя (6.78) получим выражение при $r\neq 1$

$$R_1a_r=\lambda_ra_r-a_1\cdot 0=\lambda_ra_r. \qquad (6.80)$$

Таким образом, собственные значения и собственные векторы матрицы остатков R_1 равны соответствующим собственным значениям и собственным векторам матрицы R_h. Исключение составляет только вектор a_1. Его собственное значение в R_1 равно нулю, а в матрице R_h равно $\max_j\sum_{j=1}^{n}a_{j1}^2=\lambda_1$.

Итак, собственное значение λ_2 матрицы R_h является максимальным собственным значением матрицы R_1. По значению λ_2 можно получить коэффициенты веса при втором общем факторе. Для третьего общего фактора надо воспользоваться значением λ_3 и т.д. Разумеется, если R_h положительно полуопределена, то число выделенных общих факторов будет равно $m<n$. Все собственные значения будут действительные и неотрицательные. Процесс извлечения общих факторов прекращается при равенстве суммы собственных значений следу матрицы R_h. След же матрицы R_h становится известным сразу после замены единиц на главной диагонали выборочной корреляционной матрицы R оценками общностей $\hat{h}_j^2$. Харман отмечает, что при замене матрицы R на R_h, у которой след меньше, чем у матрицы R, в практических задачах могут появиться как положительные, так и отрицательные собственные значения. Это связано с тем, что R_h потеряла свойство положительной полуопределенности. Разумеется, в данных случаях, где собственные значения являются дисперсиями общих факторов, отрицательные значения не имеют смысла. Сумма же положительных и отрицательных собственных значений равна следу матрицы R_h. Следовательно, сумма положительных собственных значений превосходит сумму оценок общностей. В общем случае выбор в качестве критерия окончания процедуры извлечения общих факторов суммы общностей матрицы R_h является оправданным, так как при сравнении данного критерия с текущей суммой собственных значений он ориентируется только на положительные λ_r.

При необходимости проверки значимости модели факторного анализа, можно воспользоваться **критерием Бартлетта**. Проверяется **нулевая гипотеза,** которая заключается в том, что m общих факторов достаточно для объяснения выборочных коэффициентов корреляции. Для статистической проверки гипотез находят расчетное значение

$$\chi^2 = N' \ln \frac{|AA^T|}{|R|}, \qquad (6.81)$$

где $N' = N - \frac{1}{6}(2n+5) - \frac{2}{3}m$;

$|AA^T|$ - определитель воспроизведенной матрицы корреляций;

$|R|$ - определитель исходной корреляционной матрицы;

n - число переменных;

m - число выделенных общих факторов;

N - число объектов исследования.

Если вычисленное значение $(\chi^2 > \chi_T^2)$ больше табличного при выбранном уровне значимости α и число степеней свободы для (6.81):

$$\nu = \frac{1}{2}[(n-m)^2 - n - m], \qquad (6.82)$$

то нулевая гипотеза отклоняется. Это значит, что m общих факторов недостаточно. Необходимо выделить факторов больше, чем m, хотя бы $(m+1)$ факторов. После этого процедура проверки повторяется.

На этом закончим рассмотрение проблем, связанных с получением общих факторов и перейдем к проблемам содержательной интерпретации полученных результатов.

6.3. Проблема вращения

6.3.1. Понятие ортогонального и косоугольного вращения

Представление всех переменных в пространстве общих факторов или главных компонент в виде совокупности векторов называется **конфигурацией.** На рис.6.4, 6.5 и 6.6 представлены соответственно случайная, простая-ортогональная и косоугольная структуры расположения наблюдений в пространстве общих факторов f_1 и f_2.

Будем вращать систему координат вокруг ее начала. При таком вращении остаются неизменными расстояния от точек до начала координат (длина векторов) и углы между векторами.

Цель вращения - получение простой структуры, чтобы большинство наблюдений находились вблизи координатных осей. При случайной конфигурации наблюдений (рис.6.4) невозможно получить простую структуру.

На рис.6.5 наблюдения группируются вокруг ортогональных общих факторов f_1' и f_2'. На рис.6.6 общие факторы f_1' и f_2', вокруг которых группируются наблюдения, не ортогональны.

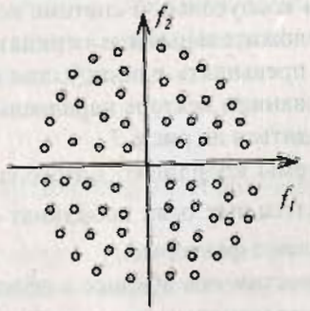


Рис. 6.4. Случайная конфигурация векторов

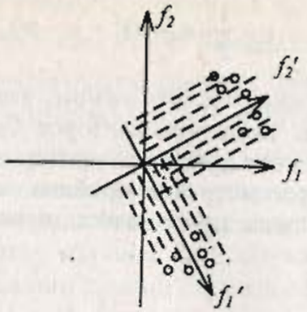


Рис. 6.5. Ортогональная простая структура

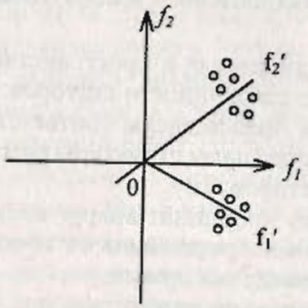


Рис. 6.6. Косоугольная простая структура

При таком косоугольном расположении общих факторов можно найти коэффициенты корреляции между факторами. При этом надо различать факторное отображение и факторную структуру (6.42). Они совпадают в методе главных компонент и в факторном анализе при некоррелированных общих факторах. При косоугольной системе координат факторные нагрузки a_{jr} могут быть положительными и отрицательными. По абсолютному значению они могут превышать единицу, так как факторные нагрузки a_{jr} получаются проецированием вектора параллельно другой неортогональной оси. В этом легко убедиться на рис.6.7.

Возьмем две системы координат: ортогональную $f_1 0 f_2$ и косоугольную $f_1' 0 f_2'$. На косоугольных осях координат отложим единичные векторы, которые представляют факторы.

Для удобства совместим ось абсцисс в обеих системах. В системе f_1 и f_2 - общие факторы в ортогональной системе координат, а f_1' и f_2' - общие факторы в косоугольной системе координат. Весовые коэффициенты между признаком y_j и факторами f_1 и f_2 равны соответственно $h_j \sin\theta$ и a_2.

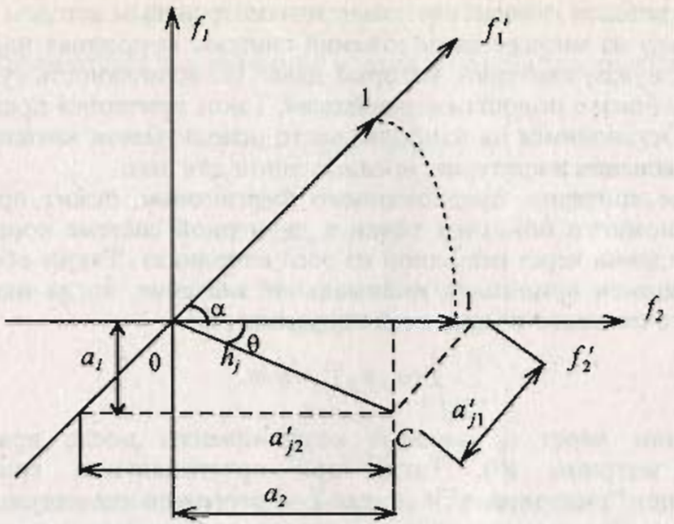


Рис. 6.7. Косоугольная система координат

Они меньше единицы, так как являются сторонами прямоугольного треугольника, в котором гипотенуза $h_j \le 1$. Зато в косоугольной системе координат весовой коэффициент между признаком y_j и фактором $f_2' : a'_{j2} > 1$, а фактором $f_1' : a'_{j1} < 1$.

Из определения коэффициента корреляции между двумя векторами можно записать, что в косоугольной системе координат коэффициенты корреляции всегда меньше единицы. Они могут быть как положительными, так и отрицательными. (Точка C имеет координаты $C(a'_{j1}, a'_{j2})$ в неортогональной системе координат.)

6.3.2. Вращение при помощи ортогональных матриц

Остановимся на ортогональном вращении, которое широко применяется в факторном анализе при решении практических задач. Конечной целью факторного анализа является получение содержательно интерпретируемых факторов, которые воспроизводили бы выборочную корреляционную матрицу между переменными. Например, в методе главных факторов это дос-

тигается путем вращения. Вращению можно подвергнуть и результаты, получаемые методом главных компонент или центроидным методом.

Поскольку из множества положений системы координат надо выбрать одну, нужен критерий, который давал бы возможность судить о том, что мы близко подошли к своей цели. Таких критериев предложено много. Остановимся на наиболее часто используемом **методе варимаксного вращения и критерии**, предложенном для него.

В основе критерия, предложенного **Фергюсоном**, лежит принцип самого экономного описания точки в двумерной системе координат при прохождении через нее одной из осей координат. Таким образом, критерий должен принимать минимальное значение, когда наибольшее число точек лежит вблизи осей координат, т.е.

$$\sum_{1<r=1}^{m}\sum_{j=1}^{n}(a_{jr}a_{jl})^2 = \min.$$

Обозначим через a_{jr} весовые коэффициенты после вращения (элементы матрицы A^1). Тогда при ортогональном вращении ("варимаксном") матрица $AT = A^1$, где T - ортогональная матрица преобразования. При ортогональном вращении общности переменных останутся без изменений.

$$\sum_{r=1}^{m}(a'_{jr})^2 = \sum_{r=1}^{m}a^2_{jr} = h_j^2, \quad (j=1,2,\ldots,n). \tag{6.83}$$

Используя основную идею Фергюсона, **Кайзер предложил варимакс-критерий** следующего вида

$$n\sum_{r=1}^{m}\sum_{j=1}^{n}(a'_{jr}/h_j)^4 - \sum_{r=1}^{m}(\sum_{j=1}^{n}a'^2_{jr}/h_j^2)^2 = \max. \tag{6.84}$$

Максимум (6.84) удовлетворяет требованиям ортогональной простой структуры.

При вращении по часовой стрелке на плоскости можно воспользоваться матрицей

$$T = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}. \tag{6.85}$$

С методической точки зрения удобно расчленить матрицу T на ряд матриц поворота на плоскости. Если комбинаций пар факторов будет p, то матрица

$$T = T_1 T_2 \ldots T_p; \tag{6.86}$$

$$p = \frac{1}{2}m(m-1),$$

где m - число общих факторов.

Полная матрица преобразования T в (6.86) также ортогональна. На практике одновременно производится поворот всех осей координат по стандартным программам при помощи матрицы T.

Так, в случае трех факторов можно получить три последовательных поворота

$$T = T_{12}T_{13}T_{23}, \tag{6.87}$$

где в общем случае при вращении против часовой стрелки на разные углы

$$T_{12} = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix};$$

$$T_{13} = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix};$$

$$T_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & -\sin\gamma \\ 0 & \sin\gamma & \cos\gamma \end{pmatrix}.$$

В заключение следует отметить, что для матрицы T (6.86) остается в силе условие $T^T T = TT^T = E$.

6.3.3. Обобщенные факторы

Если вращение является одной из обязательных проблем факторного анализа, то для метода главных компонент к вращению прибегали не часто. Однако применение варимаксного вращения для ортогональных главных компонент может оказать существенную пользу для содержательной интерпретации полученных результатов. Главные компоненты в общем случае представляют собой линии, идущие вдоль главных осей эллиптической гиперповерхности (в частности, в двумерном случае - вдоль осей эллипса, а в трехмерном - вдоль осей эллипсоида). При применении ортогонального вращения новые повернутые "главные компоненты" не являются главными компонентами и называются обобщенными факторами. На рис. 6.4 показаны оси f_1 и f_2 для главных компонент, а оси f_1' и f_2' после поворота системы координат по часовой стрелке на угол θ служат для обобщенных факторов.

В 1935 г. Тэрстоун разработал три условия, характеризующие принцип простой структуры, используемый при вращении. Продолжая работать в этом направлении, в 1947 г. он сформулировал уже **пять условий, которым должна удовлетворять простая структура:**

1) каждая строка матрицы факторной структуры должна содержать хотя бы один нулевой элемент;

2) в каждом столбце матрицы факторной структуры должно быть не менее m нулей (так как m - число общих факторов, то каждый фактор в своей гиперплоскости координат должен определяться не менее чем m признаками);

3) для каждой пары столбцов матрицы факторной структуры найдется несколько признаков, соответствующие элементы которых в матрице равны нулю в одном столбце и не равны нулю в другом;

4) если число факторов равно или превышает четыре, то достаточно велика доля признаков, имеющих в любой паре столбцов одновременно нулевые элементы;

5) для любой пары столбцов найдется мало параметров, соответствующие элементы которых в обоих столбцах отличны от нуля.

Главным условием является первое. Оно требует, чтобы каждая переменная (признак) лежала хотя бы в одной из гиперплоскостей.

Второе и третье условия были введены для того, чтобы исключить совпадение координатных гиперплоскостей и избежать неопределенности.

Все пять условий применимы для некоррелированных общих факторов.

Если рассматривается косоугольная система общих факторов, то, как известно, структура не совпадает с факторным отображением. Однако, принцип простой структуры применим не только к ортогональным общим факторам, но и к косоугольным. Тэрстоун утверждает, что можно подобрать систему координат таким образом, чтобы каждый из представляющих векторов лежал в одной или более гиперплоскостей. Полученная при этом система называется **простой структурой.**

Пример 6.8

Дана матрица A первых двух главных компонент, объясняющих 70% дисперсии всех семи признаков. Требуется преобразовать данную матрицу в матрицу, состоящую из обобщенных факторов, путем поворота на угол 30° по часовой стрелке (рис.6.8). Результаты представить аналитически. Проверить суммарную дисперсию первых двух главных компонент и полученных обобщенных факторов.

$$A = \begin{pmatrix} 0{,}90 & -0{,}30 \\ 0{,}80 & -0{,}30 \\ 0{,}60 & 0{,}30 \\ 0{,}50 & 0{,}20 \\ 0{,}80 & 0{,}60 \\ -0{,}80 & 0{,}50 \\ 0{,}20 & 0{,}80 \end{pmatrix}.$$

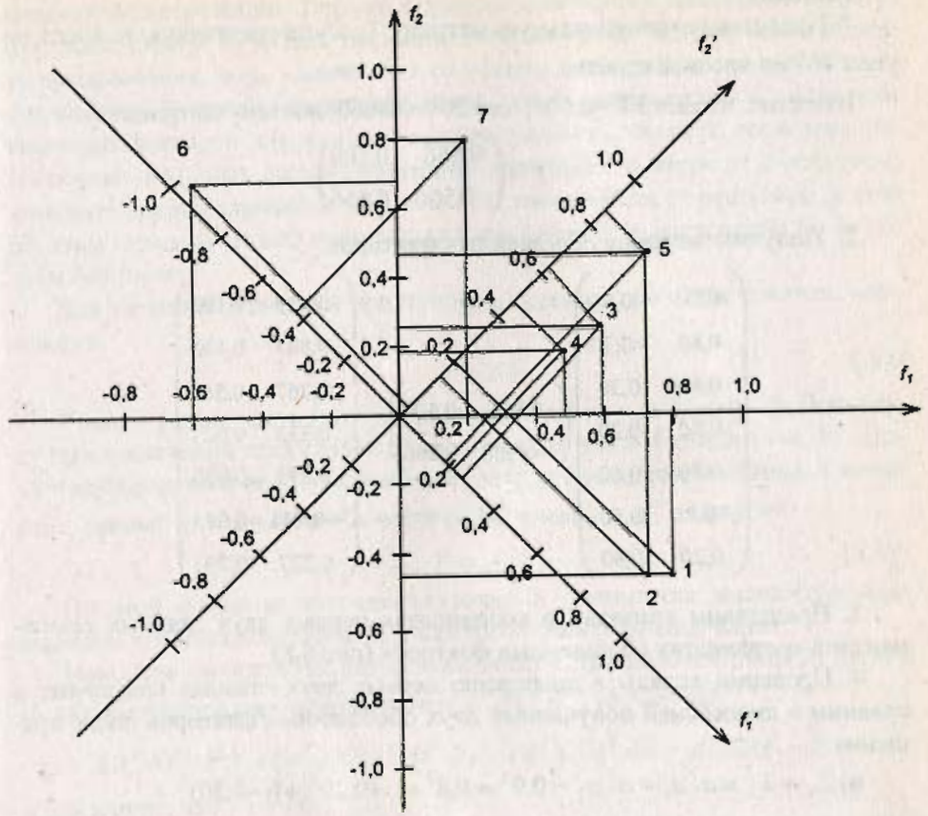


Рис. 6.8. Преобразование главных компонент в обобщенные факторы вращением осей на 30°

Решение

1. Представим ортогональную матрицу T, осуществляющую поворот на угол 30° по часовой стрелке.

Известно, что $sin\,30° = 0{,}500$; $cos\,30° \approx 0{,}8660$, поэтому матрица

$$T = \begin{pmatrix} 0{,}866 & 0{,}500 \\ -0{,}500 & 0{,}866 \end{pmatrix}.$$

2. Получим матрицу обобщенных факторов.

$$\begin{pmatrix} 0{,}90 & -0{,}30 \\ 0{,}80 & -0{,}30 \\ 0{,}60 & 0{,}30 \\ 0{,}50 & 0{,}20 \\ 0{,}80 & 0{,}60 \\ -0{,}80 & 0{,}50 \\ 0{,}20 & 0{,}80 \end{pmatrix} \begin{pmatrix} 0{,}866 & 0{,}500 \\ -0{,}500 & 0{,}866 \end{pmatrix} = \begin{pmatrix} 0{,}929 & 0{,}190 \\ 0{,}843 & 0{,}140 \\ 0{,}367 & 0{,}560 \\ 0{,}333 & 0{,}423 \\ 0{,}393 & 0{,}920 \\ -0{,}943 & 0{,}033 \\ 0{,}227 & 0{,}793 \end{pmatrix}.$$

3. Представим признаки в координатах первых двух главных компонент и в координатах обобщенных факторов (рис.6.8).

4. Проверим вклады в дисперсию первых двух главных компонент и сравним с дисперсией полученных двух обобщенных факторов после вращения:

а) $\lambda_1 + \lambda_2 = a_1^T a_1 + a_2^T a_2 = 0{,}9^2 + 0{,}8^2 + \ldots + 0{,}20^2 + [(-0{,}30)^2 + (-0{,}30)^2 + \ldots + 0{,}80^2] = 3{,}38 + 1{,}56 = 4{,}94;$

б) $\lambda_1' + \lambda_2' = (a_1')^T a_1' + (a_2')^T a_2' = 0{,}929^2 + 0{,}843^2 + \ldots + (-0{,}943^2) + (-0{,}227)^2 + 0{,}190^2 + 0{,}140^2 + \ldots + 0{,}033^2 + 0{,}793^2 = 2{,}915 + 2{,}025 = 4{,}940.$

6.4. Проблема оценки факторов и задачи классификации

6.4.1. Измерение факторов

Для лучшей содержательной интерпретации факторов следует воспользоваться вращением первично полученной матрицы. Основной задачей факторного анализа является получение **содержательно интерпретируемой матрицы,** которая лучшим образом воспроизводит матрицу коэффициентов корреляции. Первые формулы факторного анализа (6.1), (6.4) предусматривали не только значение признака y_{ji}, полученное на i-м объекте исследования, но f_{ri} - значение r-го общего фактора для i-го объекта исследования. В данном параграфе наша задача заключается в получении значений факторов для каждого индивидуального объекта исследования. На основе исходных данных в матрице значений Y и матрицы A возможно получить оценки элементов матрицы F. В зависимости от решаемой задачи по этим оценкам можно судить о каждом объекте исследования по m общим факторам.

Для уяснения методики приступим к оценке F в методе главных компонент.

$$Y = AF. \qquad (6.88)$$

Y имеет размерность $(n \times N)$; порядок A равен n, а F - $(n \times N)$. Поскольку при извлечении всех главных компонент матрица A квадратная, то задача получения матрицы F не вызывает затруднений, если матрица R имеет ранг, равный n. Умножим обе части (6.88) слева на A^{-1}, получим

$$F = A^{-1}Y. \qquad (6.89)$$

По этой формуле получаются точно и однозначно индивидуальные значения главных компонент для каждого объекта исследования.

Наиболее сложной процедурой является обращение матрицы A. Если в (6.88) обе части слева умножить на $(A^T A)^{-1} A^T$, то получим

$$(A^T A)^{-1} A^T Y = (A^T A)^{-1} A^T AF = A^{-1}(A^T)^{-1} A^T AF = A^{-1} EAF = F,$$

а это значит, что

$$F = \Lambda^{-1} A^T Y. \qquad (6.90)$$

В (6.90) нам не надо будет обращать матрицу A, если извлечены все главные компоненты. Чаще всего мы извлекаем не все главные компоненты, а только (m<n), поэтому матрица A не имеет обратной матрицы. Обратная от диагональной матрицы представляет собой диагональную матрицу Λ^{-1}, у которой все элементы являются обратными величинами к соответственным элементам матрицы Λ. Если же A не квадратная матрица, то $A^T A = \Lambda$ будет иметь порядок m.

В случае, если в методе главных компонент было использовано вращение при помощи ортогональной матрицы T, то матрица F может быть получена с учетом проведенной процедуры

$$F = T^T \Lambda^{-1} A^T Y. \qquad (6.91)$$

Отметим, что матрица A квадратная, когда извлечены все главные компоненты и матрица R имеет ранг n. Тогда формулы (6.89), (6.90) и (6.91) дают точные и однозначные значения главных компонент для каждого объекта исследования.

Таким образом, в процессе получения матрицы F самым интересным познавательным элементом является получение матриц AA^T, $(A^TA)^{-1}$, $(A^TA)^{-1}A^T$.

В факторном анализе имеются свои особенности получения матрицы F. Это обусловлено тем, что

- кроме общих существуют и характерные факторы. В этой связи матрица A заменяется матрицей M,
- первичная матрица общих факторов подвергается вращению,
- факторы могут быть и не ортогональны.

Дать точное определение индивидуальных значений факторов, как в случае (6.89), не возможно. (Кстати, в матрице M содержится $(m+n)$ факторов, при наличии n признаков.)

Поскольку задача не решается однозначно, то можно методом наименьших квадратов получить оценки индивидуальных значений общих факторов. Удобно обратиться к методу регрессионного анализа, когда имеется одна зависимая нормированная переменная и n независимых переменных, которые связаны между собой линейно.

$$y_{0i} = \beta_1 y_{1i} + \beta_2 y_{2i} + ... + \beta_n y_{ni} + e_i. \qquad (6.92)$$

Обозначим оценку

$$\hat{y}_{0i} = \hat{\beta}_1 y_{1i} + \hat{\beta}_2 y_{2i} + ... + \hat{\beta}_n y_{ni}. \qquad (6.93)$$

где y_{0i} - нормированное значение зависимой переменной на i-м объекте исследования;

$\hat{y}_{0i}$ - оценка нормированного значения зависимой переменной на i-м объекте.

Коэффициенты β_j выбираются таким образом, чтобы сумма квадратов ошибок оценок e_i^2 была минимальной

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_{0i} - \hat{y}_{0i})^2 = \min. \qquad (6.94)$$

Решение подобной задачи изложено выше. Известно, что произведение корреляционной матрицы R на вектор-столбец коэффициентов регрессии (β) равен вектору-столбцу коэффициентов корреляции между оценками нормированных значений зависимой переменной и всеми исходными признаками V, т.е.

$$R\beta = V, \qquad (6.95)$$

отсюда

$$\beta = R^{-1}V. \qquad (6.96)$$

Нас интересует β^T-строка

$$\beta^T = (R^{-1}V)^T = V^T(R^{-1})^T = V^T R^{-1}. \qquad (6.97)$$

Таких строк должно быть n в матрице регрессионных коэффициентов B.

Матрица оценок индивидуальных значений факторов может быть определена по матрице

$$\hat{F} = B^T Y. \qquad (6.98)$$

После проведения факторного анализа нам известны матрица V_{fs} и, естественно, матрица R.

Из (6.97) вместо векторов возьмем матрицы, тогда получим

$$B^T = V_{fs} R^{-1}. \qquad (6.99)$$

Элементами матрицы V_{fs} являются коэффициенты корреляции между переменными и факторами; B^T - матрица коэффициентов регрессии факторов по признакам; R^{-1} - матрица обратная матрице коэффициентов корреляции между переменными R.

Итак, матрица оценок индивидуальных значений факторов имеет вид:

$$\hat{F} = V_{fs} R^{-1} Y. \qquad (6.100)$$

Как правило, при проведении факторного анализа используется метод оценки индивидуальных значений факторов для каждого объекта исследования.

Оценки факторов для каждого объекта (субъекта) исследования могут быть проведены и при помощи обобщенных факторов.

Пример 6.9

Оценка деятельности N=20 молодых специалистов фирмы проводилась по показателям, характеризующим их успеваемость в вузе и результатам первого года работы на фирме.

Кроме признаков, характеризующих успеваемость в вузе, были использованы еще четыре признака, содержащие оценки деятельности выпускников. Последние данные были получены на основе анкетного опроса. Это теоретическая подготовка, работа в коллективе, организаторские способности.

По матрице A была получена интерпретация двух обобщенных факторов.

Первый обобщенный фактор (f_1) оказался характеристикой теоретической подготовки и организаторских способностей специалиста, а второй (f_2) характеристикой трудовой деятельности и вузовской подготовки.

В табл.6.13 приведена матрица F, индивидуальные значения двух обобщенных факторов для 20 специалистов. Специалисты, обладающие более высокими положительными значениями индивидуальных характеристик имели более высокие оценки по исходным признакам, определяющим название обобщенного фактора. Специалисты, имеющие наименьшие значения обобщенных факторов (отрицательные) f_n, имели более низкие значения оценок в исходных признаках, определяющих название обобщенных факторов.

Таблица 6.13

Индивидуальные значения первого и второго обобщенных факторов

№№	f_1	f_2	№№	f_1	f_2
1	1,5839	-0,0444	11	-1,8209	-1,0015
2	0,8712	2,0970	12	0,4630	0,8741
3	0,0135	-1,1269	13	-1,1591	0,5454
4	2,2035	0,8592	14	-0,9704	-0,4913
5	0,5843	0,0478	15	-0,8277	1,2095
6	-0,8614	-0,2446	16	-0,2517	-0,4999
7	-0,0748	0,9606	17	-0,7745	-1,0104
8	-1,3056	0,3229	18	-0,4680	-1,1695
9	-0,1432	-0,7014	19	-0,1890	0,1160
10	-0,6779	-0,1520	20	1,9109	-0,2201

Так как оба фактора представляют собой важные интегральные показатели, то по ним можно отобрать одного специалиста с лучшими организаторскими способностями и второго специалиста с наилучшей вузовской подготовкой по специальности.

Решение

1. Данную задачу легко решить по табл.6.13, выбрав f_{41}=2,2035 и f_{22}=2,0970, что соответствует специалистам под № 4 и № 2. В более сложных случаях подобные задачи решаются графически или методом ранжирования.

Данный пример был облегчен еще и тем, что вклады f_1 и f_2 в дисперсию процесса близки.

6.4.2. Факторный анализ и методы классификации многомерных наблюдений

При помощи факторного и компонентного анализа возможно сложные задачи большой размерности (по исходным показателям) приводить к новым интегральным показателям - общим факторам, число которых значительно меньше. Следовательно, факторный анализ позволяет проводить классификацию не по исходным признакам, а по интегральным показателям, построенным как линейная комбинация исходных.

Так, если первые два общих или обобщенных фактора объясняют большую долю дисперсии, то можно провести классификацию объектов исследования, например, по этим двум факторам, даже в том случае, если исходных показателей было несколько десятков. Но при этом факторный анализ не может подменить анализ кластерный. На практике используется как первый, так и второй подход. При помощи кластерного анализа можно провести классификацию при любых соотношениях между числом объектов и числом показателей. В факторном анализе число объектов, по крайней мере, в 2 раза должно превосходить количество исходных показателей.

Пример 6.10

По данным 143 студентов, обучающихся по специальности статистика, исследовалась взаимосвязь успеваемости в школе (по трем профилирующим предметам) с результатами вступительных экзаменов и экзаменами по математике и информатике в вузе.

С помощью компонентного анализа были получены две главные компоненты ($V_\Sigma = 0,78$), которые интерпретированы следующим образом: f_1 - характеристика общих способностей к информатике и математическим дисциплинам и f_2 - характеристика школьной успеваемости и вступительных экзаменов. Матрица факторных нагрузок A приведена в табл.6.14.

Таблица 6.14

Весовые коэффициенты матрицы A

Признаки	f_1	f_2
Алгебра в школе	0,80	0,56
Геометрия в школе	0,75	0,58
Информатика в школе	0,87	0,47
Дискретная математика в вузе	0,86	–0,20
Высшая математика в вузе	0,76	–0,12
Информатика в вузе	0,58	–0,18
Математика - 1 (вступительный экзамен)	0,86	–0,31
Математика -2 (вступительный экзамен)	0,78	–0,48
Информатика (вступительный экзамен)	0,69	–0,44
λ_r	5,4371	1,4778
$\frac{1}{9}\sum_{r=1}^{2}\lambda_r$		0,768

Как видно из табл.6.14, весовые коэффициенты у второй главной компоненты имеют разные знаки. Требуется:

1. Перейти к обобщенным факторам при помощи ортогонального вращения для того, чтобы полученные данные разместились в одном квадранте.

2. Результаты поворота осей f_1 и f_2 представить графически; на графике выделить классы признаков.

3. Составить таблицу, характеризующую вклады каждого из классов признаков в обобщенные факторы.

Решение

1. Повернем оси координат f_1 и f_2 на 45° при помощи матрицы

$$T = \begin{pmatrix} \cos 45^\circ & \sin 45^\circ \\ -\sin 45^\circ & \cos 45^\circ \end{pmatrix}.$$

$$AT = \begin{pmatrix} 0{,}80 & 0{,}56 \\ 0{,}75 & 0{,}58 \\ 0{,}87 & 0{,}47 \\ 0{,}86 & -0{,}20 \\ 0{,}76 & -0{,}12 \\ 0{,}58 & -0{,}18 \\ 0{,}86 & -0{,}31 \\ 0{,}78 & -0{,}48 \\ 0{,}69 & -0{,}44 \end{pmatrix} \times \begin{pmatrix} 0{,}7071 & 0{,}7071 \\ -0{,}7071 & 0{,}7071 \end{pmatrix} = \begin{pmatrix} 0{,}170 & 0{,}962 \\ 0{,}120 & 0{,}940 \\ 0{,}283 & 0{,}948 \\ 0{,}750 & 0{,}467 \\ 0{,}622 & 0{,}453 \\ 0{,}537 & 0{,}283 \\ 0{,}827 & 0{,}389 \\ 0{,}891 & 0{,}217 \\ 0{,}799 & 0{,}177 \end{pmatrix}.$$

2. Построим график (рис.6.9).

На графике выделены три класса признаков. Первый класс объединяет признаки школьной успеваемости (S_1). Второй класс выявляет признаки вступительных экзаменов (S_2). Третий класс характеризует успеваемость в вузе (S_3).

3. Вычислим вклады классов признаков в обобщенные факторы (табл. 6.15).

Таблица 6.15

Характеристики успеваемости и результаты экзаменов

Классы признаков	Успеваемость в школе		Вступительные экзамены в вуз		Успеваемость в вузе	
Вклады дисперсии в обобщенные факторы f_1 и f_2	S_1		S_2		S_3	
Оценка вкладов	f_1'	f_2'	f_1'	f_2'	f_1'	f_1'
V_{S_i} ,$(i=\overline{1,3})$	0,123	2,710	2,116	0,230	1,238	0,503
V_{S_i} / λ_r ,$(r=1,2)$	0,035	0,788	0,609	0,067	0,356	0,146

Примечания: 1) $\lambda_{f_1'} = 3{,}477353$; $\lambda_{f_2'} = 3{,}438729$;

2) $\lambda_{f_1'} + \lambda_{f_2'} = 6{,}916259 \approx 6{,}92$.

3) Сопоставим данные с дисперсией главных компонент до вращения $\lambda_{f_1} + \lambda_{f_2} = 5{,}4371 + 1{,}4778 = 6{,}9149 \approx 6{,}91$. С учетом округлений промежуточных расчетов можно считать, что поворот осей проведен правильно.

Выводы

1. Из табл. 6.14, 6.15 и рис. 6.9 видно, что успехи в школе (S_1) в основном определяются вторым обобщенным фактором.

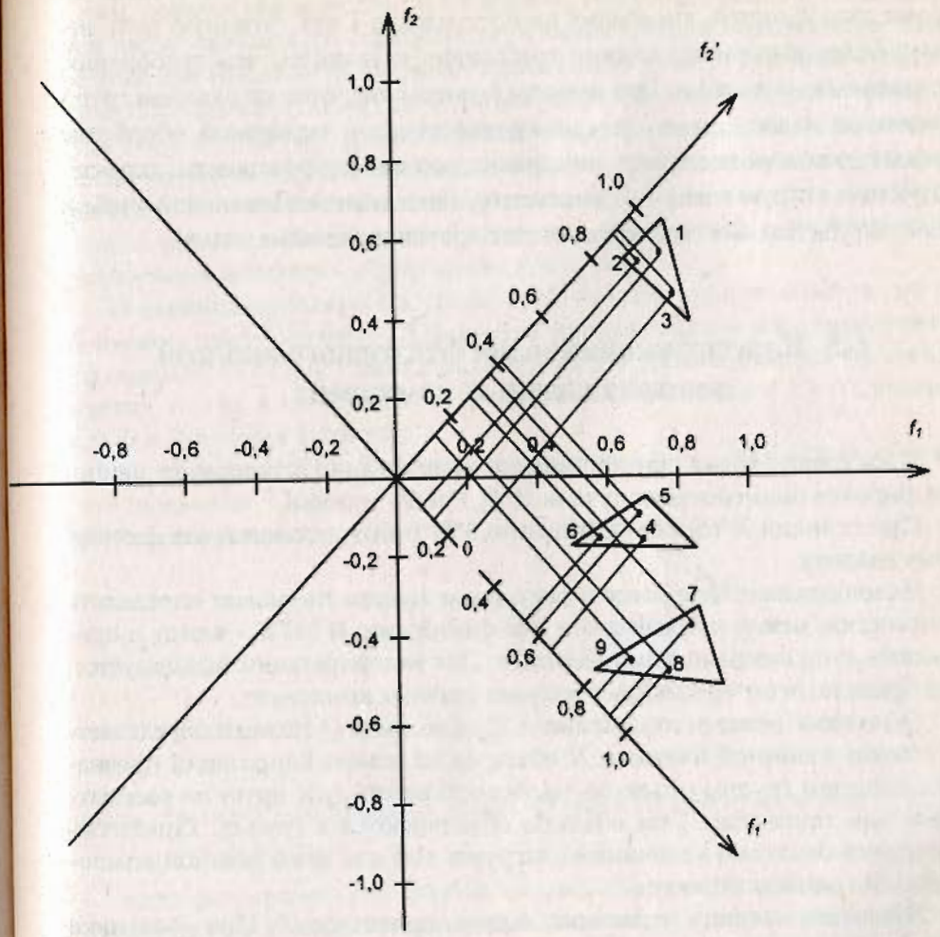


Рис. 6.9. Классификация по обобщенным факторам

2. Успехи в вузе в основном определяются первым обобщенным фактором.

3. Первый обобщенный фактор характеризует успехи при сдаче вступительных экзаменов в вуз.

Следовательно, из полученного эксперимента можно судить о том, что требования, предъявляемые к абитуриентам при сдаче вступительных экзаменов в вуз, существенно отличаются от требований к учащимся средних школ. Они ближе к требованиям вуза.

Близость классов S_2 и S_3 подтверждает предположение о том, что кроме способностей, влияющих на поступление в вуз, большую роль играют более высокие вузовские требования к знаниям, чем требования, предъявляемые в школе. Эти выводы оценены численно по вкладам групп признаков в дисперсию. Это не противоречит и первичной обработке данных до поворота главных компонент: весовые коэффициенты, характеризующие вторую главную компоненту, относящиеся к школьной учебе и сдаче вступительных экзаменов, имеют противоположные знаки.

6.5. Классификация задач факторного анализа и метода главных компонент

В настоящее время при проведении исследований в экономике наиболее широкое применение получили *R, Q, P* и *O* - техники.

При помощи *R*-техники выполнено 95% работ, посвященных факторному анализу.

Использование ***R*-техники** в факторном анализе позволяет определить взаимосвязь между n признаками и m факторами. В МГК - между n признаками и n главными компонентами. Для интерпретации используется, как правило, m ($m<n$) наиболее весомых главных компонент.

***Q*-техника** менее распространена. При помощи *Q*-техники определяется степень взаимной близости N объектов на основе корреляций признаков. Выделим группы объектов, наиболее близких друг другу по рассматриваемым признакам. Эти объекты объединяются в группу, характеризующуюся близкими величинами нагрузки той или иной главной компоненты для данных объектов.

Приведем матрицу первичных оценок признаков X. При *R*-технике изучаются корреляции между признаками (строками) матрицы. При *Q*-технике исследуются корреляции между объектами (столбцами) матрицы:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nN} \end{pmatrix} \begin{matrix} \\ \\ \\ \end{matrix} \quad \begin{matrix} Q \\ \downarrow \\ \\ \\ R \end{matrix}$$

***P*-техника** применяется не к N объектам исследования, а к одному объекту. В этом случае одно предприятие оценивается по небольшому числу

признаков в разные интервалы времени. Можно производить оценку через сутки, неделю, месяц, год и т.д. Обычно изучается корреляция между каждой парой признаков и определяется, в какой степени периодически замеряемые два признака подвержены одинаковым изменениям своих показаний. Исследования могут проводиться через разные интервалы времени. При этом, сталкиваясь с колебаниями условий процесса, имеется возможность получать контролируемые измерения признаков при меняющихся условиях. В отличие от *R*-техники *P*-техника позволяет изучать индивидуальные различия объектов, в то время как при *R*-технике анализируются различные комбинации общих свойств объектов.

***O*-техника** используется также при изучении одного объекта, но по большому числу признаков. Оценка проводится также в некоторые моменты времени по всем признакам. Корреляция рассматривается не между парами тестов, а между парами дней (недель, месяцев и т.д.). Условиями для *O* и *P*-техники являются:

N - число замеров в N промежутков времени $t_1, t_2, \dots, t_N$ у одного объекта по n признакам.

Рассмотрим матрицу значений исходных признаков

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nN} \end{pmatrix} \begin{matrix} \\ \\ \\ \downarrow \end{matrix} \quad \begin{matrix} \\ \\ \\ P \end{matrix}$$

$$O$$

,

где x_{lk} - значение k-го признака у одного объекта в l-й момент времени t_l ($k=1; \overline{l=1, N}$).

Если рассмотреть два столбца матрицы X, то корреляция между ними изучается при помощи *O*-техники. Если рассмотреть две строки матрицы X, то корреляции между ними будут исследоваться при помощи *P*-техники.

Примером применения *O*-техники на практике является реализация n видов изделий в течение N дней одним магазином.

Любая из рассмотренных техник факторного или компонентного анализа может быть использована для проведения классификации признаков или объектов.

В методах снижения размерности многомерная классификация может быть проведена по наиболее существенной информации, заключенной в первых наиболее весомых главных компонентах или общих факторах. Исследователю представляется также возможность выбора общих факторов или главных компонент в зависимости от того, какие внутренние, непосредственно не измеряемые, но объективно существующие закономерности они вскрывают.

6.6. Задачи и упражнения

Доли дисперсии и коэффициенты корреляции.

6.1. Составить уравнения для факторного анализа при n=5, m=2.

$y_1=a_{11}f_1+a_{12}f_2+d_1v_1$; $y_4=a_{41}f_1+a_{42}f_2+d_4v_4$;
$y_2=a_{21}f_1+a_{22}f_2+d_2v_2$; $y_5=a_{51}f_1+a_{52}f_2+d_5v_5$.
$y_3=a_{31}f_1+a_{32}f_2+d_3v_3$;

6.2. Чему равны вклады общих факторов и характерного фактора в дисперсию y_1:

$$y_1=0{,}5f_1+0{,}8f_2+0{,}33v_1;$$

$$\sigma^2_{f_1}=0{,}5^2=0{,}25;\quad \sigma^2_{f_2}=0{,}64;\quad \sigma^2_{v_1}=0{,}11;$$

$$\sigma^2_{y_1}=0{,}25+0{,}64+0{,}1089=0{,}9989\approx 1.$$

6.3. Найти общность в задаче 6.2.

$$h_1^2=a_{11}^2+a_{12}^2=0{,}5^2+0{,}8^2=0{,}25+0{,}64=0{,}89.$$

6.4. Найти общность и характерность параметра y_1, если его надежность r_1^2=0,95, а специфичность равна 0,15.

Решение

$$\sigma^2_{y1}=1=h_1^2+d_1^2=h_1^2+b_1^2+c_1^2;$$

$$r_1^2=h_1^2+b_1^2=0{,}95;$$

$$c_1^2=1-r_1^2=1-0{,}95=0{,}05;$$

$$h_1^2=r_1^2-b_1^2=0{,}95-0{,}15=0{,}80;$$

$$d_1^2=b_1^2+c_1^2=0{,}15+0{,}05=0{,}20$$

или $$d_1^2=1-h_1^2=1-0{,}80=0{,}20.$$

6.5. По таблице, содержащей две составляющие дисперсии, найти три другие составляющие.

Компоненты дисперсии признака	№ заданий							
	1	2	3	4	5	6	7	8
Специфичность - b_j^2	0,10	0,15	0,20	0,25				
Дисперсия ошибок - c_j^2	0,20				0,05	0,10		
Характерность - d_j^n			0,35			0,45		0,30
Общность - h_j^2		0,75			0,60		0,50	
Надежность - r_j^2				0,85			0,75	0,90

Решение

Компоненты дисперсии признака	Формула	№ заданий							
		1	2	3	4	5	6	7	8
Специфичность	$b_j^2=d_j^2-c_j^2$					0,35	0,35	0,25	0,20
Дисперсия ошибок	$c_j^2=1-r_j^2=d_j^2-b_j^2$		0,10	0,15	0,15			0,25	0,10
Характерность	$d_j^2=1-h_j^2=b_j^2+c_j^2$	0,30	0,25		0,40	0,40		0,50	
Общность	$h_j^2=1-d_j^2$	0,70		0,65	0,60		0,55		0,70
Надежность	$r_j^2=1-c_j^2$	0,80	0,90	0,85		0,95	0,90		

6.6. Заданы два общих фактора: a_{11}=0,6; a_{12}=0,5, а также надежность r_1^2=0,84. Составить полное уравнение для параметра y_1.

Решение

а). Определим b_j:

$b_j^2=r_j^2-h_j^2$;
$h_j^2=a_{11}^2+a_{12}^2=0{,}6^2+0{,}5^2=0{,}36+0{,}25=0{,}61$;
$b_j^2=0{,}84-0{,}61=0{,}23$;
$b_j=\sqrt{0{,}23}=0{,}48$.

б). Рассчитаем дисперсию ошибок c_j^2 и c_j:

$$c_j^2=1-r_j^2=1-0{,}84=0{,}16;\ c_j=\sqrt{0{,}16}=0{,}4.$$

в). Запишем полное уравнение признака y_1:

$$y_1=0{,}6f_1+0{,}5f_2+0{,}48S_1+0{,}4c_1.$$

6.7. Найти долю суммарной общности, вносимую двумя общими факторами и каждым в отдельности (в %), по заданной матрице A (знаки весовых коэффициентов не приводятся):

$$A=\begin{pmatrix}0{,}7 & 0{,}3\\ 0{,}8 & 0{,}0\\ 0{,}7 & 0{,}0\\ 0{,}8 & 0{,}6\\ 0{,}6 & 0{,}5\\ 0{,}5 & 0{,}0\\ 0{,}6 & 0{,}4\\ 0{,}7 & 0{,}6\end{pmatrix}.$$

Решение

а). Найдем общность двух общих факторов

$$\sum_{j=1}^{8}(h_{j1}^2+h_{j2}^2)=(0{,}7^2+0{,}8^2+0{,}7^2+0{,}8^2+0{,}6^2+0{,}5^2+0{,}6^2+0{,}7^2)+$$
$$+(0{,}3^2+0{,}6^2+0{,}5^2+0{,}4^2+0{,}6^2)=3{,}72+1{,}22=4{,}94=\lambda_1+\lambda_2.$$

б). Определим вклад первого общего фактора в суммарную общность, %:

$$\frac{\sigma^2_{f1}}{\sigma^2_{f1}+\sigma^2_{f2}}\cdot 100=\frac{3{,}72}{4{,}94}=75{,}3;$$

$$\frac{\sigma^2_{f2}}{\sigma^2_{f1}+\sigma^2_{f2}}\cdot 100=24{,}7.$$

6.8. Вычислить долю суммарной дисперсии, объясняемую каждым из общих факторов в задаче 6.7.

Решение

а). Вся суммарная дисперсия процесса равна числу признаков n=8.

б). Определим вклад первого фактора в суммарную дисперсию процесса, %.

$$\lambda_1=3{,}72;\ \frac{\lambda_1}{n}\cdot 100=\frac{3{,}72\times 100}{8}=46{,}50.$$

в). Вклад второго общего фактора определяется аналогично, %.

$$\frac{\lambda_2}{n}\cdot 100=\frac{1{,}22\times 100}{8}=15{,}25.$$

6.9. Вычислить коэффициенты корреляции, редуцированные на основе примера 6.7: r'_{12}, r'_{14}, r'_{25}, r'_{26}.

Решение

а). $r'_{12} = \sum_{r=1}^{2} a_{jr} a_{kr}$, при j=1, k=2.

$$r'_{12} = a_{11}a_{21} + a_{12}a_{22} = 0{,}7\times0{,}8 + 0{,}3\times0 = 0{,}56.$$

б). $r'_{14} = a_{11}a_{41} + a_{12}a_{42} = 0{,}7\times0{,}8 + 0{,}3\times0{,}6 = 0{,}56 + 0{,}18 = 0{,}74.$

в). $r'_{25} = a_{21}a_{51} + a_{22}a_{52} = 0{,}8\times0{,}6 + 0{,}0\times0{,}5 = 0{,}48.$

г). $r'_{26} = a_{21}a_{61} + a_{22}a_{62} = 0{,}8\times0{,}0 + 0{,}5\times0 = 0{,}0.$

6.10. По данным примера 6.7 вычислить наибольшее и наименьшее значения характерности.

Решение

а). Максимальное значение характерности

$$\max d_j^2 = 1 - h_{j\min}^2 = 1 - h_5^2 = 1 - 0{,}25 = 0{,}75 = d_6^2.$$

б). Наименьшее значение характерности

$$\min d_j^2 = 1 - h_{j\max}^2 = 1 - h_4^2 = 1 - 0{,}64 - 0{,}36 = 0 = d_4^2.$$

6.7. Задания для самостоятельной работы на ЭВМ

По данным приложения 2 провести факторный анализ.

1. Получить матрицу парных коэффициентов корреляции R.
2. Используя метод наибольшего элемента в строке, преобразовать матрицу R в редуцированную матрицу R_h.
3. Получить первые три общих фактора и дать им экономическую интерпретацию по матрице факторных нагрузок A.
4. На основании матрицы индивидуальных значений F провести графически классификацию предприятий по первым двум общим факторам.
5. Провести графически классификацию признаков по первым двум общим факторам.

ГЛАВА 7

МЕТОДЫ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ

7.1. Классификация без обучения. Кластерный анализ

7.1.1. Основные понятия

В статистических исследованиях группировка первичных данных является основным приемом решения задачи классификации, а значит и основой всей дальнейшей работы с собранной информацией.

Традиционно эта задача решается следующим образом. Из множества признаков, описывающих объект, отбирается один, наиболее информативный с точки зрения исследователя, и производится группировка в соответствии со значениями данного признака. Если требуется провести классификацию по нескольким признакам, ранжированным между собой по степени важности, то сначала производится классификация по первому признаку, затем каждый из полученных классов разбивается на подклассы по второму признаку, и т.д. Подобным образом строится большинство комбинационных статистических группировок.

В тех случаях, когда упорядочить классификационные признаки не представляется возможным, применяется наиболее простой метод многомерной группировки - **создание интегрального показателя (индекса),** функционально зависящего от исходных признаков, с последующей классификацией по этому показателю.

Развитием этого подхода является вариант классификации по нескольким обобщающим показателям (главным компонентам), полученным с помощью **методов факторного анализа.**

При наличии нескольких признаков (исходных или обобщенных) задача классификации может быть решена **методами кластерного анализа,** которые от других методов многомерной классификации отличаются отсутствием обучающих выборок, т.е. априорной информации о распределении генеральной совокупности, которая представляет собой вектор X.

Различия между схемами решения задач классификации во многом определяются тем, что понимают под понятиями "сходство" и "степень сходства".

После того, как сформулирована цель работы, необходимо попытаться определить **критерии качества, целевую функцию,** значения которой позволят сопоставить различные схемы классификации.

В экономических исследованиях целевая функция, как правило, должна минимизировать некоторый **параметр,** определенный на множестве объектов (например, целью классификации оборудования может явиться группировка, минимизирующая совокупность затрат времени и средств на ремонтные работы) .

В случаях, когда формализовать цель задачи не удается, критерием качества классификации может служить **возможность содержательной интерпретации найденных групп.**

Рассмотрим следующую задачу. Пусть исследуется совокупность n объектов, каждый из которых характеризуется по k замеренным на нем признакам X. Требуется разбить эту совокупность на однородные в некотором смысле группы (классы). При этом практически отсутствует априорная информация о характере распределения измерений X внутри классов.

Полученные в результате разбиения группы обычно называются **кластерами** (от англ. cluster - группа элементов, характеризуемых каким-либо общим свойством), а также таксонами (от англ. taxon - систематизированная группа любой категории) или образами. Методы нахождения кластеров называются **кластер-анализом** (соответственно численной таксономией или распознаванием образов с самообучением).

При этом с самого начала необходимо четко представить, какая из двух задач классификации подлежит решению. Если решается **обычная задача типизации**, то совокупность наблюдений разбивают на сравнительно небольшое число областей группирования (например, интервальный вариационный ряд в случае одномерных наблюдений) так, чтобы элементы одной такой области по возможности находились друг от друга на небольшом расстоянии.

Решение **другой задачи типизации** заключается в определении естественного расслоения исходных наблюдений на четко выраженные кластеры, лежащие друг от друга на некотором расстоянии.

Если первая задача типизации всегда имеет решение, то при второй постановке может оказаться, что множество исходных наблюдений не обнаруживает естественного расслоения на кластеры, т.е. образует один кластер.

Несмотря на то что многие методы кластерного анализа довольно элементарны, применение методов кластерного анализа стало возможным только в 80-е годы с возникновением и развитием вычислительной техники. Это объясняется тем, что эффективное решение задачи поиска кластеров требует большего числа арифметических и логических операций. Рас-

смотрим три различных подхода к проблеме кластерного анализа: эвристический, экстремальный и статистический.

Эвристический подход характеризуется отсутствием формальной модели изучаемого явления и критерия для сравнения различных решений. Его основой является алгоритм, построенный исходя из интуитивных соображений.

При **экстремальном подходе** также не формулируется исходная модель, а задается критерий, определяющий качество разбиения на кластеры. Такой подход особенно полезен, если цель исследования четко определена. В этом случае качество разбиения может измеряться эффективностью выполнения цели.

Основой **статистического подхода** решения задачи кластерного анализа является вероятностная модель исследуемого процесса. Статистический подход особенно удобен для теоретического исследования проблем, связанных с кластерным анализом. Кроме того, он дает возможность ставить задачи, связанные с воспроизводимостью результатов кластерного анализа.

Рассмотрим формы представления исходных данных и определение мер близости.

В задачах кластерного анализа обычной формой представления исходных данных служит **прямоугольная таблица,** каждая строка которой представляет результат измерения k рассматриваемых признаков на одном из обследованных объектов:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}.$$

В конкретных ситуациях может представлять интерес как группировка объектов, так и группировка признаков. В случаях, когда разница между этими двумя задачами несущественна, например при описании некоторых алгоритмов, мы будем пользоваться только термином "объект", подразумевая в этом понятии и "признак".

Числовые значения, входящие в матрицу X, могут соответствовать трем типам переменных: количественным, ранговым и качественным. **Количественные** переменные обладают свойством упорядоченности и над ними можно производить арифметические операции. Значения **ранговых** переменных тоже упорядочены, и их можно пронумеровать натуральными числами. Однако использование этих чисел в арифметических операциях будет некорректным. **Качественными** называются переменные, принимающие два (дихотомные) или более значений. Этим значениям также можно поставить в соответствие некоторые числа, которые, однако, не будут от-

ражать какой-либо упорядоченности значений качественной переменной. Исключением являются дихотомные переменные, два значения которых (как правило, они обозначаются числами 0 и 1) можно считать упорядоченными.

Желательно, чтобы таблица исходных данных соответствовала одному типу переменных. В противном случае разные типы переменных стараются свести к какому-то одному типу переменных. Например, все переменные можно свести к дихотомным, используя следующую процедуру. Количественные переменные переводят в ранговые, разбивая области значений количественной переменной на интервалы, которые затем нумеруются числами натурального ряда. Ранговые переменные автоматически становятся качественными, если не учитывать упорядоченности их значений. Что касается качественных переменных, то каждому из возможных ее значений приходится сопоставлять дихотомную переменную, которая будет равна 1, если качественная переменная приняла данное значение, и 0 - в противном случае.

Отметим, что форма записи исходных данных, их сведение к одному типу, возможность использования только части данных и т.п., играют определенную роль при оценке практической эффективности вычислительного комплекса, предназначенного для решения задач классификации.

Матрица X не является единственным способом представления исходных данных в задачах кластерного анализа. Иногда исходная информация задана в виде **квадратной матрицы**

$$R=(r_{ij}),\ i,j=1,2,...,k,$$

элемент r_{ij} которой определяет степень близости i-го объекта к j-му.

Большинство алгоритмов кластерного анализа либо полностью исходит из матрицы расстояний (или близостей), либо требует вычисления отдельных ее элементов, поэтому, если данные представлены в форме X, то первым этапом решения задачи поиска кластеров будет выбор способа вычисления **расстояний или близости между объектами или признаками** (в этом отношении различие между объектами и признаками является существенным).

Относительно просто определяется близость между признаками. Как правило, кластерный анализ признаков преследует те же цели, что и факторный анализ - выделение групп связанных между собой признаков, отражающих определенную сторону изучаемых объектов. В этом случае мерами близости служат различные статистические коэффициенты связи.

Если признаки количественные, то можно использовать оценки обычных парных выборочных коэффициентов корреляции r_{ij}, $i,j=1,2,...,k$. Однако коэффициент корреляции измеряет только линейную связь, поэтому если связь нелинейна, то следует использовать корреляционное отношение, либо произвести подходящее преобразование шкалы признаков.

Существуют также различные коэффициенты связи, определенные для ранговых, качественных и дихотомных переменных

7.1.2. Расстояние между обьектами и мера близости

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов.

В общем случае **понятие однородности объектов** задается либо введением правила вычислений расстояния $\rho(X_i,X_j)$ между любой парой исследуемых объектов $(X_1,X_2,...,X_n)$, либо заданием некоторой функции $r(X_i,X_j)$, характеризующей степень близости i-го и j-го объектов. Если задана функция $\rho(X_i,X_j)$, то близкие с точки зрения этой метрики объекты считаются однородными, принадлежащими одному классу. При этом необходимо сопоставлять $\rho(X_i,X_j)$ с некоторым пороговым значением, определяемым в каждом конкретном случае по-своему.

Аналогично используется и **мера близости** $r(X_i,X_j)$, при задании которой надо помнить о необходимости выполнения условий *симметрии* $r(X_i,X_j)=r(X_j,X_i)$; *максимального сходства объекта с самим собой* $r(X_i,X_i)=\max_j r(X_i,X_j)$, при $1\le j\le n$, и *монотонного убывания* $r(X_i,X_j)$ по $\rho(X_i,X_j)$, т.е. из $\rho(X_k,X_l)\ge\rho(X_i,X_j)$ должно следовать неравенство $r(X_k,X_l)\le r(X_i,X_j)$.

Выбор метрики или меры близости является узловым моментом исследования, от которого в основном зависит окончательный вариант разбиения объектов на классы при данном алгоритме разбиения. В каждом конкретном случае этот выбор должен производиться по-своему в зависимости от целей исследования, физической и статистической природы вектора наблюдений X, априорных сведений о характере вероятностного распределения X.

Рассмотрим наиболее часто используемые расстояния и меры близости в задачах кластерного анализа.

РАССТОЯНИЕ МАХАЛАНОБИСА (ОБЩИЙ ВИД)

В случае зависимых компонент $x_1,x_2,...,x_k$ вектора наблюдений X и их различной значимости в решении вопроса классификации обычно используют обобщенное (взвешенное) расстояние Махаланобиса, задаваемое формулой

$$\rho_0(X_i,X_j)=\sqrt{(X_i-X_j)^T\Lambda^T\Sigma^{-1}\Lambda(X_i-X_j)},\qquad(7.1)$$

где Σ - ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения;

Λ - некоторая симметрическая неотрицательно-определенная матрица "весовых" коэффициентов, которая чаще всего выбирается диагональной.

Следующие три вида расстояний являются частными случаями метрики ρ_0.

ОБЫЧНОЕ ЕВКЛИДОВО РАССТОЯНИЕ

$$\rho_E(X_i, X_j) = \sqrt{\sum_{l=1}^{k}(x_{il} - x_{jl})^2}, \qquad (7.2)$$

где x_{il}, x_{jl} - величина l-й компоненты у i-го (j-го) объекта (l=1,2,...,k; i,j=1,2,...,n).

Использование этого расстояния оправдано в случаях, если:

а) наблюдения берутся из генеральных совокупностей, имеющих многомерное нормальное распределение с ковариационной матрицей вида $\sigma^2 E_k$, т.е. компоненты X взаимно независимы и имеют одну и ту же дисперсию;

б) компоненты вектора наблюдений X однородны по физическому смыслу и одинаково важны для классификации;

в) признаковое пространство совпадает с геометрическим пространством.

Естественно с геометрической точки зрения и содержательной интерпретации евклидово расстояние может оказаться бессмысленным, если его признаки имеют разные единицы измерения. Для приведения признаков к одинаковым единицам прибегают к нормировке каждого признака путем деления центрированной величины на среднее квадратическое отклонение и переходят от матрицы X к нормированной матрице с элементами

$$x_{il}^{H} = \frac{x_{il} - \bar{x}_l}{s_l},$$

где x_{il} - значение l-го признака у i-го объекта;

$\bar{x}_l$ - среднее арифметическое значение l-го признака;

$s_l = \sqrt{\frac{1}{n}\sum_i (x_{il} - \bar{x}_l)^2}$ - среднее квадратическое отклонение l-го признака.

Однако эта операция может привести к нежелательным последствиям. Если кластеры хорошо разделены по одному признаку и не разделены по другому, то после нормировки дискриминирующие возможности первого признака будут уменьшены в связи с увеличением "шумового" эффекта второго.

"ВЗВЕШЕННОЕ" ЕВКЛИДОВО РАССТОЯНИЕ

$$\rho_{BE}(X_i, X_j) = \sqrt{\sum_{l=1}^{k} w_l (x_{il} - x_{jl})^2} \qquad (7.3)$$

применяется в случаях, когда каждой компоненте x_l вектора наблюдений X удается приписать некоторый "вес" w_l, пропорциональный степени важности признака в задаче классификации. Обычно принимают $0 \le w_l \le 1$, где l=1, 2,..., k.

Определение "весов", как правило, связано с дополнительными исследованиями, например, организацией опроса экспертов и обработкой их мнений. Определение весов w_l только по данным выборки может привести к ложным выводам.

ХЕММИНГОВО РАССТОЯНИЕ

$$\rho_H(X_i, X_j) = \sum_{l=1}^{k} |x_{il} - x_{jl}| \qquad (7.4)$$

используется как мера различия объектов, задаваемых дихотомическими признаками. Хеммингово расстояние равно числу несовпадений значений соответствующих признаков в рассматриваемых i-м и j-м объектах.

В некоторых задачах классификации в качестве меры близости объектов можно использовать некоторые физически содержательные параметры, так или иначе характеризующие взаимоотношение между объектами. Например, задачу классификации отраслей народного хозяйства с целью агрегирования решают на основе матрицы межотраслевого баланса.

В данной задаче объектом классификации является отрасль народного хозяйства, а матрица межотраслевого баланса представлена элементами S_{ij}, характеризующими сумму годовых поставок i-й отрасли в j-ю в денежном выражении. В качестве меры близости $\{r_{ij}\}$ принимают **симметризованную нормированную матрицу межотраслевого баланса.** С целью нормирования денежное выражение поставок i-й отрасли в j-ю заменяют долей этих поставок по отношению ко всем поставкам i-й отрасли. Симметризацию нормированной матрицы межотраслевого баланса можно проводить выразив через среднее значение близость взаимных поставок между i-й и j-й отраслью так, что в этом случае $r_{ij} = r_{ji}$.

Как правило, решение задач классификации многомерных данных предусматривает в качестве предварительного этапа исследования реализацию методов, позволяющих выбрать из компонент $x_1, x_2, \ldots, x_k$ наблюдаемых векторов x сравнительно небольшое число наиболее существенных информативных признаков, т.е. уменьшить размерность наблюдаемого пространства. С этой целью каждую из компонент $x_1, x_2, \ldots, x_k$ рассматривают как объект, подлежащий классификации. После разбиения на небольшое число однородных в некотором смысле групп для дальнейшего исследования оставляют по одному представителю от каждой группы. При этом предполагается, что признаки, попавшие в одну группу, в определенном смысле связаны друг с другом и несут информацию о каком-то одном свойстве объекта.

В качестве близости между отдельными признаками обычно используют различные характеристики степени их коррелированности, в первую очередь коэффициенты корреляции. В ряде задач применяются и другие

расстояния (метрики). Выбор метрики определяется структурой признакового пространства и целью классификации. Формализовать этот этап задачи классификации пока не представляется возможным.

7.1.3. Расстояние между кластерами

В ряде процедур классификации (кластер-процедур) используют понятия **расстояния между группами объектов и меры близости двух групп объектов.**

Пусть S_i - i-я группа (класс, кластер), состоящая из n_i объектов;

$\bar{x}_i$ - среднее арифметическое векторных наблюдений S_i группы, т.е. "центр тяжести" i-й группы;

$\rho(S_l, S_m)$ - расстояние между группами S_l и S_m.

Наиболее употребительными расстояниями и мерами близости между классами объектов являются:

- расстояние, измеряемое по принципу "ближайшего соседа"

$$\rho_{\min}(S_l, S_m) = \min_{\substack{x_l \in S_l \\ x_j \in S_m}} \rho(x_l, x_j); \qquad (7.5)$$

- расстояние, измеряемое по принципу "дальнего соседа"

$$\rho_{\max}(S_l, S_m) = \max_{\substack{x_i \in S_l \\ x_j \in S_m}} \rho(x_i, x_j); \qquad (7.6)$$

- расстояние, измеряемое по "центрам тяжести" групп

$$\rho(S_l, S_m) = \rho(\bar{x}_l, \bar{x}_m); \qquad (7.7)$$

- расстояние, измеряемое по принципу "средней связи". Это расстояние определяется как среднее арифметическое всех попарных расстояний между представителями рассматриваемых групп

$$\rho_{cp}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} \rho(x_i, x_j). \qquad (7.8)$$

Академиком А.Н.Колмогоровым было предложено "обобщенное расстояние" между классами, которое в качестве частных случаев включает в себя все рассмотренные выше виды расстояний.

Обобщенное расстояние основано на понятии так называемого "обобщенного среднего", а точнее - степенного среднего и определяется формулой:

$$\rho_{об}(S_l, S_m) = \left[\frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} \rho^r(x_i, x_j) \right]^{1/r}. \qquad (7.9)$$

Можно показать, что при $r \to \infty$

$$\rho_{об}(S_l, S_m) = \rho_{\max}(S_l, S_m),$$

при $r \to -\infty$

$$\rho_{об}(S_l, S_m) = \rho_{\min}(S_l, S_m),$$

при $r = 1$

$$\rho_{об}(S_l, S_m) = \rho(S_l, S_m).$$

Из формулы (7.9) следует, что если $S_{(m,q)} = S_m \cup S_q$ - группа элементов, полученная путем объединения кластеров S_m и S_q, то обобщенное расстояние между кластерами S_l и $S_{(m,q)}$ определяется по формуле

$$\rho_{об}(S_l, S_{(m,q)}) = \left\{ \frac{n_m \left[\rho_{об}(S_l, S_m)\right]^r + n_q \left[\rho_{об}(S_l, S_q)\right]^r}{n_m + n_q} \right\}^{1/r}. \qquad (7.10)$$

Расстояние между группами элементов особенно важно в так называемых **агломеративных иерархических кластер-процедурах**, так как принцип работы таких алгоритмов состоит в последовательном объединении сначала самых близких элементов, а затем и целых групп все более и более отдаленных друг от друга элементов.

При этом расстояние между классами S_l и $S_{(m,q)}$, являющимися объединением двух других классов S_m и S_q, можно определить по формуле:

$$\rho_{l,(m,q)} = \rho(S_l, S_{(m,q)}) = \alpha\rho_{lm} + \beta\rho_{lq} + \gamma\rho_{mq} + \delta\left|\rho_{lm} - \rho_{lq}\right|, \qquad (7.11)$$

где $\rho_{lm} = \rho(S_l, S_m)$; $\rho_{lq} = \rho(S_l, S_q)$; $\rho_{mq} = \rho(S_m, S_q)$ - расстояния между классами S_l, S_m и S_q;

α, β, γ и δ - числовые коэффициенты, значение которых определяет специфику процедуры, ее алгоритм.

Например, при $\alpha = \beta = -\delta = \frac{1}{2}$ и $\gamma = 0$ приходим к расстоянию, построенному по принципу "ближайшего соседа". При $\alpha = \beta = \delta = \frac{1}{2}$ и $\gamma = 0$ расстояние между классами определяется по принципу "дальнего соседа", как расстояние между двумя самыми дальними элементами этих классов. И наконец, при

$$\alpha = \frac{n_m}{n_m + n_q}, \quad \beta = \frac{n_q}{n_m + n_q}, \quad \gamma = \delta = 0$$

соотношение (7.11) приводит к расстоянию ρ между классами, вычисленному как среднее из расстояний между всеми парами элементов, один из которых берется из одного класса, а другой - из другого класса.

7.1.4. Функционалы качества разбиения

Существует большое количество различных способов разбиения на классы заданной совокупности элементов. Поэтому представляет интерес задача сравнительного анализа качества этих способов разбиения. С этой целью вводится понятие функционала качества разбиения Q (S), определенного на множестве всех возможных разбиений.

Наилучшее разбиение S^* представляет собой такое разбиение, при котором достигается экстремум выбранного **функционала качества.** Следует отметить, что выбор того или иного функционала качества разбиения, как правило, опирается на эмпирические соображения.

Рассмотрим некоторые наиболее распространенные функционалы качества разбиения. Пусть исследованием выбрана метрика ρ в пространстве X и $S = (S_1, S_2, ..., S_p)$ некоторое фиксированное разбиение наблюдений $X_1, X_2, ..., X_n$ на заданное число p классов $S_1, S_2, ..., S_p$.

Существуют следующие характеристики функционала качества:

• сумма внутриклассовых дисперсий

$$Q_1(S) = \sum_{l=1}^{p} \sum_{x_i \in S_l} \rho^2(x_i, \bar{x}_l); \qquad (7.12)$$

• сумма попарных внутриклассовых расстояний между элементами

$$Q_2(S) = \sum_{l=1}^{p} \sum_{x_i, x_j \in S_l} \rho^2(x_i, x_j); \qquad (7.13)$$

или

$$Q_2'(S) = \sum_{l=1}^{p} \frac{1}{n_l} \sum_{x_i, x_j \in S_l} \rho^2(x_i, x_j).$$

$Q_1(S)$ и $Q_2(S)$ широко используются в задачах кластерного анализа для сравнения качества процедур разбиения;

• обобщенная внутриклассовая дисперсия

$$Q_3(S) = \det\left(\sum_{l=1}^{p} n_l W_l\right), \qquad (7.14)$$

где $\det A$ - определитель матрицы A;

W_l - выборочная ковариационная матрица класса S_l, элементы которой определяются по формуле

$$w_{qm}(l) = \frac{1}{n_l} \sum_{x_i \in S_l} (x_{iq} - \bar{x}_q)(x_{im} - \bar{x}_m), \quad q, m = 1, 2, ..., k,$$

где x_{iq} - q-я компонента многомерного наблюдения x_i;

$\bar{x}_q$ - среднее значение q-й компоненты, вычисленное по наблюдениям l-го класса.

Качество разбиения характеризуют и другим видом обобщенной дисперсии, в которой операция суммирования W_l заменена операцией умножения

$$Q_4(S) = \prod_{l=1}^{p} (\det W_l)^{n_l}.$$

Отметим, что функционалы $Q_3(S)$ и $Q_4(S)$ обычно используют при решении вопроса: не сосредоточены ли наблюдения, разбитые на классы, в пространстве размерности, меньшей, чем k.

7.1.5. Иерархические кластер-процедуры

Иерархические (деревообразные) процедуры являются наиболее распространенными алгоритмами кластерного анализа по их реализации на ЭВМ. Они бывают двух типов: агломеративные и дивизимные. В **агломеративных** процедурах начальным является разбиение, состоящее из n одноэлементных классов, а конечным - из одного класса; в **дивизимных** наоборот.

Принцип работы иерархических агломеративных (дивизимных) процедур состоит в последовательном объединении (разделении) групп элементов сначала самых близких (далеких), а затем все более отдаленных (близких) друг от друга. Большинство этих алгоритмов исходит из матрицы расстояний (сходства).

К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации. Алгоритмы требуют на каждом шаге матрицы вычисления расстояний, а следовательно, емкой машинной памяти и большого количества времени. В этой связи реализация таких алгоритмов при числе наблюдений, большем нескольких сотен, нецелесообразна, а в ряде случаев и невозможна.

Приведем пример агломеративного иерархического алгоритма. На первом шаге каждое наблюдение X_i (i=1,2,...,n) рассматривается как отдельный кластер. В дальнейшем на каждом шаге работы алгоритма происходит объединение двух самых близких кластеров, и, с учетом принятого расстояния, по формуле пересчитывается матрица расстояний, размерность которой, очевидно, снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс. Большинство программ, реализующих алгоритм иерархической классификации, предусматривают графическое представление классификации в виде дендрограммы.

Пример 7.1

Провести классификацию n=6 объектов, каждый их которых характеризуется двумя признаками:

№ объекта i	1	2	3	4	5	6
x_{i1}	5	6	5	10	11	10
x_{i2}	10	12	13	9	9	7

Расположение объектов в виде точек на плоскости показано на рис.7.1.

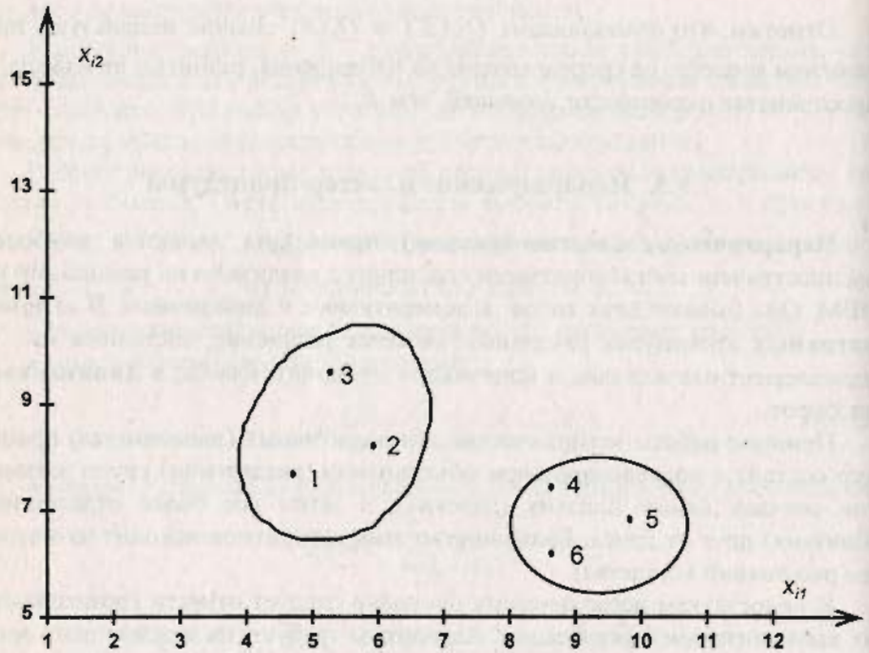


Рис.7.1. Классификация объектов

Решение

Воспользуемся агломеративным иерархическим алгоритмом классификации. В качестве расстояния между объектами возьмем обычное евклидово расстояние. Тогда согласно формуле (7.2) расстояние между первым и вторым объектами

$$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = 2{,}24,$$

а между первым и третьим объектами

$$\rho_{13} = \sqrt{(5-5)^2 + (10-13)^2} = 3.$$

Очевидно, что

$$\rho_{11} = 0.$$

Аналогично находим расстояния между шестью объектами и строим матрицу расстояний

$$R_1 = \left\{\rho(x_i, x_j)\right\} = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 6{,}08 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 5{,}83 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}21 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 1 & 2 \\ 6{,}08 & 5{,}83 & 7{,}21 & 1 & 0 & 2{,}24 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2 & 2{,}24 & 0 \end{pmatrix}.$$

Из матрицы расстояний следует, что четвертый и пятый объекты наиболее близки $\rho_{4,5} = 1{,}00$ и поэтому объединяются в один кластер.

После объединения объектов имеем пять кластеров:

Номер кластера	1	2	3	4	5
Состав кластера	(1)	(2)	(3)	(4,5)	(6)

Расстояние между кластерами определим по принципу "ближайшего соседа", воспользовавшись формулой пересчета (7.11). Так расстояние между объектом S_1 и кластером $S_{(4,5)}$

$$\rho_{1,(4,5)} = \rho(S_1, S_{(4,5)}) = \frac{1}{2}\rho_{14} + \frac{1}{2}\rho_{15} - \frac{1}{2}\left|\rho_{14} - \rho_{15}\right| =$$
$$= \frac{1}{2}\left[5{,}10 + 6{,}08\right] - \frac{1}{2}\left[\left|5{,}10 - 6{,}08\right|\right] = 5{,}10.$$

Таким образом, расстояние $\rho_{1,(4,5)}$ равно расстоянию от объекта 1 до ближайшего к нему объекта, входящего в кластер $S_{(4,5)}$, т.е. $\rho_{1,(4,5)} = \rho_{1,4} = 5{,}10$. Тогда матрица расстояний

$$R_2 = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 2 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2 & 0 \end{pmatrix}.$$

Объединим второй и третий объекты, имеющие наименьшее расстояние $\rho_{2,3} = 1{,}41$. После объединения объектов имеем четыре кластера:

$$S_{(1)}, \quad S_{(2,3)}, \quad S_{(4,5)}, \quad S_{(6)}.$$

Вновь найдем матрицу расстояний. Для того чтобы рассчитать расстояние до кластера $S_{(2,3)}$ воспользуемся матрицей расстояний R_2. Например, расстояние между кластерами $S_{(4,5)}$ и $S_{(2,3)}$ равно

$$\rho_{(4,5),(2,3)} = \frac{1}{2}\rho_{(4,5),2} + \frac{1}{2}\rho_{(4,5),3} - \frac{1}{2}\left|\rho_{(4,5),2} - \rho_{(4,5),3}\right| = \frac{5}{2} + \frac{6{,}40}{2} - \frac{1{,}40}{2} = 5.$$

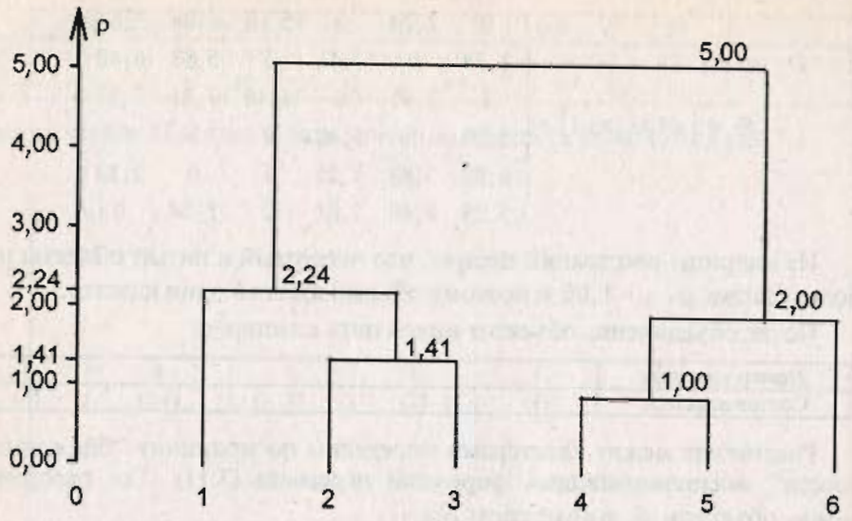


Рис. 7.2. Дендрограмма

Проведя аналогичные расчеты, получим

$$R_3 = \begin{pmatrix} 0 & 2,24 & 5,10 & 5,83 \\ 2,24 & 0 & 5 & 6,40 \\ 5,10 & 5 & 0 & 2 \\ 5,83 & 6,40 & 2 & 0 \end{pmatrix}.$$

Объединим кластеры $S_{(4,5)}$ и $S_{(6)}$, расстояние между которыми, согласно матрице R_3, наименьшее $\rho_{(4,5),6} = 2$. В результате получим три кластера

$$S_{(1)}, \quad S_{(2,3)} \quad \text{и} \quad S_{(4,5,6)}.$$

Матрица расстояний будет иметь вид:

$$R_4 = \begin{pmatrix} 0 & 2,24 & 5,10 \\ 2,24 & 0 & 5 \\ 5,10 & 5 & 0 \end{pmatrix}.$$

Объединим теперь кластеры S_1 и $S_{(2,3)}$, расстояние между которыми $\rho_{1,(2,3)} = 2,24$. В результате получим два кластера: $S_{(1,2,3)}$ и $S_{(4,5,6)}$, расстояние между которыми, найденное по принципу "ближайшего соседа", $\rho_{(1,2,3),\,(4,5,6)} = 5$.

Результаты иерархической классификации объектов представлены на рис.7.2 в виде дендрограммы.

На рис.7.2 приводятся расстояния между объединяемыми на данном этапе кластерами (объектами). В нашем примере предпочтение следует отдать предпоследнему этапу классификации, когда все объекты объединены в два кластера (рис.7.2)

$$S_{(1,2,3)} \quad \text{и} \quad S_{(4,5,6)}.$$

7.2. Дискриминантный анализ

7.2.1. Методы классификации с обучением

Дискриминантный анализ, как раздел **многомерного статистического анализа,** включает в себя статистические методы классификации многомерных наблюдений в ситуации, когда исследователь обладает так называемыми обучающими выборками ("классификация с учителем").

В общем случае задача различения (дискриминации) формулируется следующим образом. Пусть результатом наблюдения над объектом является реализация k-мерного случайного вектора $x=(x_1,x_2,...,x_k)^T$. Требуется установить правило, согласно которому по наблюденному значению вектора x объект относят к одной из возможных совокупностей π_i, $i=1,2,...,l$. Для построения правила дискриминации все выборочное пространство R значений вектора x разбивается на области R_i, $i=1,2,...,l$, так, что при попадании x в R_i объект относят к совокупности π_i.

Правило дискриминации выбирается в соответствии с определенным принципом оптимальности на основе априорной информации о совокупностях p_i извлечения объекта из π_i. При этом следует учитывать размер убытка от неправильной дискриминации. Априорная информация может быть представлена как в виде некоторых сведений о функции k-мерного распределения признаков в каждой совокупности, так и в виде выборок из этих совокупностей. Априорные вероятности p_i могут быть либо заданы, либо нет. Очевидно, что рекомендации будут тем точнее, чем полнее исходная информация.

С точки зрения применения дискриминантного анализа наиболее важной является ситуация, когда исходная информация о распределении представлена выборками из них. В этом случае задача дискриминации ставится следующим образом.

Пусть $x_1^{(i)},...,x_j^{(i)},...,x_{n_i}^{(i)}$ - выборка из совокупности π_i, $i=1,2,...,l$; $j=1,2,...,n_i$; причем каждый j-й объект выборки представлен k-мерным вектором параметров $x_j^{(i)} = (x_{j1}^{(i)} ... x_{iq}^{(i)} ... x_{ik}^{(i)})^T$. Произведено дополнительное

наблюдение $x=(x_1,...,x_k)^T$ над объектом, принадлежащим одной из совокупностей π_i. Требуется построить правило отнесения наблюдения x к одной из этих совокупностей.

Обычно в задаче различения переходят от вектора признаков, характеризующих объект, к линейной функции от них, дискриминантной функции - гиперплоскости, наилучшим образом разделяющей совокупность выборочных точек.

Наиболее изучен случай, когда известно, что распределение векторов признаков в каждой совокупности нормально, но нет информации о параметрах этих распределений. Здесь естественно заменить неизвестные параметры распределения в дискриминантной функции их наилучшими оценками. Правило дискриминации можно основывать на отношении правдоподобия.

Непараметрические методы дискриминации не требуют знаний о точном функциональном виде распределений и позволяют решать задачи дискриминации на основе незначительной априорной информации о совокупностях, что особенно ценно для практических применений.

В **параметрических методах** эти точки используются для оценки параметров статистических функций распределения. В параметрических методах построения функции как правило используется нормальное распределение.

7.2.2. Линейный дискриминантный анализ

Предположения:

1) имеются разные классы объектов;

2) каждый класс имеет нормальную функцию плотности от k переменных

$$x=\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_{ni} \end{pmatrix};$$

$$f_i(x)=(2\pi)^{ni/2}\left|\Sigma_i\right|^{-1/2}\exp\left(-\frac{1}{2}(x-\mu^{(i)})^T\Sigma_i^{-1}(x-\mu^{(i)})\right), \quad (7.15)$$

где $\mu^{(i)}$ - вектор математических ожиданий переменных размерности k;

Σ_i - ковариационная матрица при $n=n_i$;

Σ_i^{-1} - обратная матрица.

Матрица Σ_i - положительно определена.

В случае если параметры известны дискриминацию можно провести следующим образом.

Имеются функции плотности $f_1(x), f_2(x),..., f_l(x)$ нормально распределенных классов. Задана точка x в пространстве k измерений. Предполагая, что $f_i(x)$ имеет наибольшую плотность, отнесем точку x к i-му классу. Существует доказательство, что если априорные вероятности для определяемых точек каждого класса одинаковы и потери при неправильной классификации i-й группы в качестве j-й не зависят от i и j, то решающая процедура минимизирует ожидаемые потери при неправильной классификации.

Приведем пример оценки параметра многомерного нормального распределения μ и Σ.

μ и Σ могут быть оценены по выборочным данным: $\hat{\mu}$ и $\hat{\Sigma}$ для классов. Задано l выборок $(x_1^{(i)}, x_2^{(i)}, ..., x_n^{(i)}) \equiv x_i ; (i=\overline{1,l})$ из некоторых классов. Математические ожидания $\mu_1, \mu_2, ..., \mu_k$ могут быть оценены средними значениями

$$\hat{\mu}_q^{(i)}=\frac{1}{n_i}\sum_{j=1}^{n_i}x_{jq}^{(i)}; \; q=\overline{1,k}. \quad (7.16)$$

Несмещенные оценки элементов ковариационной матрицы Σ есть

$$(\hat{\Sigma}_i)_{rs}=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{jr}^{(i)}-\hat{\mu}_r^{(i)})(x_{jr}^{(i)}-\hat{\mu}_r^{(i)}); r,s=\overline{1,k}. \quad (7.17)$$

Следовательно, можно определить $\hat{\mu}^{(i)}$ и $\hat{\Sigma}_i$ по l выборкам в каждом классе при помощи (7.16), (7.17). Получив оценки, точку x отнесем к классу, для которой функция $f(x)$ максимальна.

Введем предположение, что все классы, среди которых должна проводиться дискриминация, имеют нормальное распределение с одной и той же ковариационной матрицей Σ.

В результате существенно упрощается выражение для дискриминантной функции.

Класс, к которому должна принадлежать точка x, можно определить на основе неравенства

$$f_i(x) > f_j(x). \quad (7.18)$$

Воспользуемся формулой (7.15) для случая, когда их ковариационные матрицы равны: $\Sigma_i=\Sigma_j=\Sigma$, а $\mu^{(i)}$ есть вектор математических ожиданий класса i. Тогда (7.18) можно представить неравенством их квадратичных форм

$$-[(x-\mu^{(i)})^T\Sigma^{-1}(x-\mu^{(i)})] > -[(x-\mu^{(j)})^T\Sigma^{-1}(x-\mu^{(j)})].$$

Раскроем скобки

$$-(x^T\Sigma^{-1}x-x^T\Sigma^{-1}\mu^{(i)}-\mu^{(i)T}\Sigma^{-1}x+\mu^{(i)T}\Sigma^{-1}\mu^{(i)}) > \\ > -(x^T\Sigma^{-1}x-x^T\Sigma^{-1}\mu^{(j)}-\mu^{(j)T}\Sigma^{-1}x+\mu^{(j)T}\Sigma^{-1}\mu^{(j)}). \quad (7.19)$$

Вспомним, если имеем два вектора Z и W, то скалярное произведение можно записать $Z^T W = W^T Z = (Z,W)$. В выражении (7.19) исключим $x^T\Sigma^{-1}x$ справа и слева, поменяем у всех членов суммы знаки. Теперь преобразуем

$$x^T\Sigma^{-1}\mu^{(j)} + \mu^{(j)^T}\Sigma^{-1}x = 2(x^T\Sigma^{-1}\mu^{(j)}) = 2(x,\Sigma^{-1}\mu^{(j)});$$

$$\mu^{(j)^T}\Sigma^{-1}\mu^{(j)} = (\mu^{(j)},\Sigma^{-1}\mu^{(j)}).$$

Аналогично провести преобразования по индексу i представим читателю самостоятельно.

Сократим правую и левую часть неравенства (7.19) на 2 и, используя запись квадратичных форм, получим

$$(x,\Sigma^{-1}\mu^{(i)}) - \frac{1}{2}(\mu^{(i)},\Sigma^{-1}\mu^{(i)}) > (x,\Sigma^{-1}\mu^{(j)}) - \frac{1}{2}(\mu^{(j)},\Sigma^{-1}\mu^{(j)}). \quad (7.20)$$

Введем обозначения в выражение (7.20):

$$\nu^{(i)} = \Sigma^{-1}\mu^{(i)} \;;\; i = \overline{1,m};$$

$$\lambda_i = \frac{1}{2}(\mu^{(i)},\Sigma^{-1}\mu^{(i)});\; i = \overline{1,m}.$$

Тогда выражение (7.20) примет вид

$$(x,\nu^{(i)}) - \lambda_i > (x,\nu^{(j)}) - \lambda_j. \quad (7.21)$$

Следствие: проверяемая точка x относится к классу i, для которого линейная функция

$$h_i(x) = (x,\nu^{(i)}) - \lambda_i = \max. \quad (7.22)$$

Преимущество метода линейной дискриминации Фишера заключается в линейности дискриминантной функции (7.22) и надежности оценок ковариационных матриц классов.

Пример

Имеются два класса с параметрами $(\mu^{(1)},\Sigma_1)$ и $(\mu^{(2)},\Sigma_2)$. По выборкам из этих совокупностей объемом n_1 и n_2 получены оценки $\hat{\Sigma}_1$ и $\hat{\Sigma}_2$. Первоначально проверяется гипотеза о том, что ковариационные матрицы Σ_1 и Σ_2 равны. В случае если оценки $\hat{\Sigma}_1$ и $\hat{\Sigma}_2$ статистически неразличимы, то принимается, что $\Sigma_1=\Sigma_2=\Sigma$ и строится общая оценка $\hat{\Sigma}$, основанная на суммарной выборке объемом n_1+n_2, после чего строится линейная дискриминантная функция Фишера (7.22).

Существуют и другие методы. Так, в математическом обеспечении пакета "Олимп" используется пошаговый дискриминантный анализ.

7.2.3. Дискриминантный анализ при нормальном законе распределения показателей

Имеются две генеральные совокупности X и Y, имеющие трехмерный нормальный закон распределения с неизвестными, но равными ковариационными матрицами. Из них взяты обучающие выборки с объемами n_1 у X и n_2 у Y.

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \dots & \dots & \dots \\ x_{n_1 1} & x_{n_1 2} & x_{n_1 3} \end{pmatrix}, \quad (7.23)$$

$$Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \dots & \dots & \dots \\ y_{n_2 1} & y_{n_2 2} & y_{n_2 3} \end{pmatrix}. \quad (7.24)$$

Целью дискриминантного анализа является отнесение нового наблюдения (строки матрицы Z) либо к X, либо к Y.

$$Z = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ \dots & \dots & \dots \\ z_{l1} & z_{l2} & z_{l3} \end{pmatrix}. \quad (7.25)$$

Для решения задачи по обучающим выборкам определим векторы средних

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} \text{ и } \bar{Y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix}.$$

1. Определим оценки ковариационных матриц

$$S_x = \{s_{ki}\}_x \text{ и } S_y = \{s_{ki}\}_y; \quad \bar{x}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}.$$

Найдем элемент матрицы S_x:

$$s_{kj}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = \overline{x_j x_k} - \bar{x}_j\bar{x}_k \;;\; j,k = 1,2,3,$$

где $\bar{x}_j$ и $\bar{x}_k$ - средние значения.

2. Рассчитаем несмещенную оценку суммарной ковариационной матрицы

$$\hat{S} = \frac{1}{n_1 + n_2 - 2}(n_1 S_x + n_2 S_y).$$

3. Определим матрицу $\hat{S}^{-1}$, обратную к $\hat{S}$.

4. Вычислим вектор оценок коэффициентов дискриминантной функции $a = \hat{S}^{-1}(\overline{X} - \overline{Y})$.

5. Рассчитаем оценки векторов значений дискриминантной функции для матриц исходных данных $\hat{U}_x = Xa$, $\hat{U}_y = Ya$.

6. Вычислим средние значения оценок дискриминантной функции $\overline{\hat{u}}_x = \frac{1}{n_1}\sum_{i=1}^{n_1}\hat{u}_{xi}$, $\overline{\hat{u}}_y = \frac{1}{n_2}\sum_{i=1}^{n_2}\hat{u}_{yi}$.

7. Определим константу $\hat{C} = \frac{1}{2}(\overline{\hat{u}}_x + \overline{\hat{u}}_y)$.

Дискриминантную функцию для ν-ого наблюдения, подлежащего дискриминации, получим решив уравнение

$$\hat{u}_\nu = z_{\nu 1}a_1 + z_{\nu 2}a_2 + z_{\nu 3}a_3.$$

Если $\hat{u}_\nu \geq \hat{C}$, то ν-е наблюдение надо отнести к совокупности x, если же $\hat{u}_\nu < \hat{C}$, то ν-е наблюдение следует отнести к совокупности y.

Пример дискриминантного анализа

Деятельность каждого производственного объединения отрасли оценивалась по следующим трем показателям:

- среднегодовой стоимости основных производственных фондов (ОПФ);
- среднесписочной численности промышленно-производственного персонала (ППП);
- балансовой прибыли.

В отрасли выделены две группы: передовая, состоящая из четырех объединений, и остальная, включающая пять объединений.

Данные представлены в табл.7.1.

Отрасли передано объединение Z, у которого по принятым трем показателям получены следующие результаты: стоимость ОПФ - 55,451; численность ППП - 9,592 тыс.человек; балансовая прибыль - 12,840.

Определить, можно ли отнести новое объединение к передовой группе предприятий отрасли.

Решение

1. Запишем исходные данные в виде матриц X и Y согласно (7.23) и (7.24).

Таблица 7.1

Исходные данные

Группа объединений \ Показатели	Стоимость ОПФ	Численность ППП	Балансовая прибыль
Передовая	224,228	17,115	22,981
	151,827	14,904	21,481
	147,313	13,627	28,669
	152,253	10,545	10,199
Остальная	46,757	4,428	11,124
	29,033	5,510	6,091
	52,134	4,214	11,842
	37,050	5,527	11,873
	63,979	4,211	12,860

$$X = \begin{pmatrix} 224,228 & 17,115 & 22,981 \\ 151,827 & 14,904 & 21,481 \\ 147,313 & 13,627 & 28,669 \\ 152,253 & 10,545 & 10,199 \end{pmatrix};$$

$$Y = \begin{pmatrix} 46,751 & 4,428 & 11,124 \\ 29,033 & 5,510 & 6,091 \\ 52,134 & 4,214 & 11,842 \\ 37,050 & 5,527 & 11,873 \\ 63,979 & 4,211 & 12,860 \end{pmatrix},$$

где $n_1 = n_x = 4$; $n_2 = n_y = 5$;

строка матрицы Z: $Z^T = \begin{pmatrix} 55,451 & 9,592 & 12,840 \end{pmatrix}$.

2. Получим векторы средних

$$\overline{X} = \begin{pmatrix} 168,92025 \\ 14,04775 \\ 20,8325 \end{pmatrix}; \quad \overline{Y} = \begin{pmatrix} 45,7926 \\ 4,778 \\ 10,758 \end{pmatrix}.$$

3. Определим оценку ковариационных матриц

$$S_x = \begin{pmatrix} 1025,61 & 55,66575 & 28,94475 \\ & 5,6468625 & 10,27365 \\ & & 44,879675 \end{pmatrix};$$

$$S_y = \begin{pmatrix} 145,8666 & -6,60952 & 22,78694 \\ & 0,371782 & -0,902484 \\ & & 5,750302 \end{pmatrix}.$$

4. Получим несмещенную оценку суммарной ковариационной матрицы

$$\hat{S} = \frac{1}{4+5-2}(4S_x + 5S_y);$$

$$\hat{S} = \begin{pmatrix} 690,25328 & 27,087914 & 32,816242 \\ & 3,4923371 & 5,2260257 \\ & & 29,752887 \end{pmatrix}.$$

5. Определим обратную матрицу к $\hat{S}$

$$\hat{S}^{-1} = \begin{pmatrix} 0,0020945371 & -0,017349116 & 0,00073714 \\ & 0,53214303 & -0,07433441 \\ & & 0,04565381 \end{pmatrix}.$$

6. Найдем вектор оценок коэффициентов дискриминации

$$a = \hat{S}^{-1}(\overline{X} - \overline{Y}) = \hat{S}^{-1}\begin{pmatrix} 123,12765 \\ 9,26975 \\ 10,0745 \end{pmatrix} - \begin{pmatrix} 0,10449979 \\ 2,0475006 \\ -0,13634981 \end{pmatrix}.$$

7. Вычислим оценки дискриминантной функции

$$\hat{U}_x = Xa = \begin{pmatrix} 55,346433 \\ 43,457381 \\ 39,3990544 \\ 36,113833 \end{pmatrix}; \quad \hat{U}_y = \begin{pmatrix} 12,437003 \\ 13,486817 \\ 12,46277 \\ 13,571031 \\ 13,555623 \end{pmatrix}.$$

8. Определим средние значения оценок дискриминантной функции

$$\overline{\hat{u}}_x = 43,577047; \qquad \overline{\hat{u}}_y = 13,102648.$$

9. Получим константу

$$\hat{C} = \frac{1}{2}(43,577047 + 13,102648) = 28,339847.$$

10. Определим возможность включения объединения Z в группу передовых. Так как матрица Z представлена одной строкой, то $\hat{U}_y$ обозначим $\hat{U}_z$.

$$\hat{u}_z = a_1 z_1 + a_2 z_2 + a_3 z_3,$$

$$\hat{u}_z = 0,10449979 \cdot 55,451 + 2,0478006 \cdot 9,952 - 0,13634981 \cdot 12,840 \approx 23,69.$$

Среднее значение дискриминантной функции $\hat{u}_z$ меньше чем константа $\hat{C}$

$$\hat{u}_z < \hat{C};$$

$$23,69 < 28,34.$$

Вывод. Объединение z с характеристиками Z^T не может быть отнесено к группе передовых предприятий отрасли.

7.3. Задачи и упражнения

Дискриминантный анализ

7.1. В табл.7.2 представлены группы машиностроительных предприятий с высоким и низким уровнями организации управления производством. Характеризуя деятельность предприятий показателями рентабельности и производительности труда с помощью дискриминантного анализа требуется классифицировать три последних предприятия.

Таблица 7.2

№ предприятия	Показатель / Группа предприятий	Рентабельность, %	Производительность труда, млн руб./чел.
1	Высокий уровень,	23,4	9,1
2	*X*	19,1	6,6
3		17,5	5,2
4		17,2	10,1
5	Низкий уровень,	5,4	4,3
6	*Y*	6,6	5,5
7		8,0	5,7
8		9,7	5,5
9		9,1	6,6
10	Подлежат	9,9	7,4
11	дискриминации,	14,2	9,4
12	*Z*	12,9	6,7

7.2. При оценке эффективности деятельности предприятий приборостроения были получены два класса предприятий *A* и *B* (табл.7.3). Используя их как обучающие выборки провести дискриминацию трех последних предприятий по показателям фондоотдачи и материалоемкости.

Таблица 7.3

№ предприятия	Показатель / Предприятия	Фондоотдача	Материалоемкость
1	Класс *A*,	1,47	3,90
2	*X*	1,52	6,18
3		1,38	13,80
4		1,27	8,18
5	Класс *B*,	0,59	9,08
6	*Y*	0,57	2,10
7		0,46	6,5
8		0,53	13,3
9	Подлежат	0,92	13,80
10	дискриминации,	1,21	4,80
11	*Z*	0,63	12,33

7.3. Анализ эффективности использования земельных угодий в сельскохозяйственных районах области позволил выделить регионы с низким A и высоким B уровнями использования земли (табл.7.4). С помощью дискриминантного анализа провести классификацию трех последних районов по показателям объема реализованной продукции растениеводства и животноводства с 1га посевной площади.

Таблица 7.4

№ района	Показатель / Уровень использования земли	Объем реализованной продукции	
		Растениеводства	Животноводства
1	Низкий A,	0,25	0,41
2	X	0,51	0,51
3		0,27	0,42
4		0,33	0,56
5	Высокий B,	1,17	0,28
6	Y	4,99	0,67
7		5,18	0,45
8		2,49	0,38
9		2,73	0,33
10	Подлежат	0,32	0,45
11	дискриминации,	0,67	0,32
12	Z	4,6	0,56

7.4. По эффективности работы выделены две группы, состоящие из n_x=4 и n_y=5 ферм. Для этих групп по показателям производительности труда и объема реализованной продукции растениеводства с 1га получены оценки векторов средних

$$\overline{X}=\begin{pmatrix}6,72\\0,340\end{pmatrix};\quad \overline{Y}=\begin{pmatrix}4,05\\2,51\end{pmatrix}$$

и ковариационных матриц

$$S_x=\begin{pmatrix}1,07 & 0,23\\0,23 & 0,08\end{pmatrix};\quad S_y=\begin{pmatrix}0,30 & 0,16\\0,16 & 0,24\end{pmatrix}.$$

Вычислить значения дискриминантной функции для ферм, показатели которых представлены в матрице

$$Z=\begin{pmatrix}4,92 & 0,32\\5,22 & 0,67\\5,62 & 4,60\end{pmatrix}.$$

7.5. При анализе уровня медицинского обслуживания населения выделены две группы регионов. При этом первая группа включает n_x=4, а вторая - n_y=5 регионов.

Для этих групп по двум показателям - число врачей на 10 тыс. жителей; число больничных коек на 10 тыс. человек - получены оценки векторов средних

$$\overline{X}=\begin{pmatrix}34,47\\124,75\end{pmatrix};\quad \overline{Y}=\begin{pmatrix}31,08\\112,00\end{pmatrix}$$

и ковариационных матриц

$$S_x=\begin{pmatrix}1,10 & 2,79\\2,79 & 7,58\end{pmatrix};\quad S_y=\begin{pmatrix}2,41 & 3,45\\3,45 & 5,02\end{pmatrix}.$$

Вычислить оценки значений дискриминантной функции для регионов, показатели которых представлены в матрице

$$Z=\begin{pmatrix}31,3 & 123\\31,8 & 110\\34,6 & 117\end{pmatrix}.$$

7.6. По уровню механизации добычи выделены две группы, включающие n_x=5 и n_y=4 угольных шахт. Для этих групп по двум показателям - удельный вес добычи угля из комплексно-механизированных забоев по отношению ко всей добыче (%); удельный вес комбайновой проходки выработок (%) - получены оценки векторов средних

$$\overline{X}=\begin{pmatrix}43,26\\28,54\end{pmatrix};\quad \overline{Y}=\begin{pmatrix}14,98\\38,95\end{pmatrix}$$

и ковариационных матриц

$$S_x=\begin{pmatrix}84,61 & 29,54\\29,54 & 57,91\end{pmatrix};\quad S_y=\begin{pmatrix}79,88 & 26,98\\26,98 & 54,77\end{pmatrix}.$$

Определить оценки значений дискриминантной функции для угольных шахт, показатели которых представлены в матрице

$$Z=\begin{pmatrix}34,2 & 24,8\\17,0 & 60,0\\38,0 & 9,7\end{pmatrix}.$$

7.7. Эффективность деятельности фермерских хозяйств оценивалась по трем показателям: производительность труда (млн руб./чел.); объем реализованной продукции растениеводства и животноводства с 1га посевной площади (млн руб./га). В результате были выделены хозяйства с высоким A и низким B уровнями использования земли (табл. 7.5). Используя данные таблицы, с помощью дискриминантного анализа произвести классификацию трех последних хозяйств.

Таблица 7.5

№ района	Показатель / Уровень использования земли	Производительность труда	Объем реализованной продукции	
			Растениеводства	Животноводства
1	Высокий A,	8,22	0,25	0,41
2	X	6,43	0,51	0,51
3		6,39	0,27	0,42
4		5,84	0,33	0,56
5	Низкий B,	3,14	1,17	0,28
6	Y	4,02	4,99	0,67
7		4,19	5,18	0,45
8		4,33	2,49	0,38
9		4,58	2,73	0,33
10	Подлежат	4,92	0,32	0,45
11	дискриминации,	5,22	0,67	0,32
12	Z	5,62	4,60	0,56

7.8. Имеется 12 предприятий, характеризуемых тремя экономическими показателями: производительность труда; удельный вес потерь от брака (%) и фондоотдача активной части основных производственных фондов. Из этих предприятий выделены две обучающие вы-

борки, первая из которых включает n_x=4 предприятий группы A, а вторая - n_y=5 предприятий группы B. Используя данные табл.7.6 требуется:

- найти оценки векторов средних $\overline{X}$, $\overline{Y}$ и ковариационных матриц S_x и S_y;
- определить несмещенную оценку суммарной ковариационной матрицы $\hat{S}$ и обратной матрицы $\hat{S}^{-1}$;
- получить вектор оценок коэффициентов дискриминантной функции;
- найти оценки значений дискриминантной функции $\hat{U}_x$ и $\hat{U}_y$ для матриц исходных данных X и Y (табл.7.6);
- определить оценку константы $\hat{C}$;
- вычислить оценки значений дискриминантной функции для группы оставшихся предприятий и провести их классификацию.

Таблица 7.6

№ предприятия	Показатель / Предприятия	Производительность труда	Удельный вес потерь от брака	Фондоотдача
1	Группы A,	9,4	0,15	1,91
2	X	9,9	0,34	1,68
3		9,1	0,09	1,89
4		9,4	0,21	2,30
5	Группы B,	6,6	0,48	0,88
6	Y	4,3	0,41	0,62
7		7,4	0,62	1,09
8		6,6	0,50	1,32
9		5,5	1,20	0,68
10	Подлежат	5,5	0,05	1,02
11	дискриминации,	5,7	0,66	1,43
12	Z	10,0	0,32	2,62

Кластерный анализ

7.9. По иерархическому агломеративному алгоритму провести классификацию n=6 предприятий машиностроения, деятельность которых характеризуется показателями: x_1 - рентабельности (%); и x_2 - производительности труда.

№ предприятия	1	2	3	4	5	6
x_1	23,4	17,5	9,7	18,2	6,6	8,0
x_2	9,1	5,2	5,5	9,4	7,5	5,7

В качестве расстояния между объектами принять:

а) обычное евклидово расстояние;

б) взвешенное евклидово расстояние с весами $\omega_1 = 0,3$ и $\omega_2 = 0,7$.

Сравнить разбиение на два кластера по критерию минимума суммы внутриклассовых дисперсий.

Расстояние между кластерами определить по принципу "ближайшего соседа". Данные не нормализовывать.

7.10. Решить задачу 7.9 предварительно нормировав исходные данные.

7.11. Решить задачу 7.9, измеряя расстояние между кластерами по принципу "дальнего соседа", нормируя предварительно исходные данные.

7.12. Решить задачу 7.9, измеряя расстояние между кластерами по "центрам тяжести" групп, не нормируя предварительно исходные данные.

7.13. По агломеративному алгоритму провести классификацию n=5 фермерских хозяйств, работа которых характеризуется показателями объема реализованной продукции: x_1 - растениеводства и x_2 - животноводства с 1га пашни.

№ колхоза	1	2	3	4	5
x_1	2,49	1,51	1,17	1,67	2,73
x_2	0,38	0,51	0,28	0,29	0,34

В качестве расстояния между объектами принять обычное евклидово расстояние, а расстояния между кластерами измерять по принципу:

а) "ближайшего соседа";

б) "дальнего соседа".

Сравнить разбиение на два кластера по критерию минимума суммы внутриклассовых дисперсий. Исходные данные не нормализовывать.

7.14. Решить задачу 7.13 предварительно нормировав исходные данные.

7.15. Решить задачу 7.13 при условии, что расстояния между кластерами измеряются по принципу "центра тяжести" групп, нормируя предварительные исходные данные.

7.16. Решить задачу 7.13 при условии, что расстояние между кластерами измеряется по принципу:

а) "средней связи";

б) "центра тяжести" групп, не нормируя предварительно исходные данные.

7.17. По агломеративному алгоритму провести классификацию n=6 регионов по уровню медицинского обслуживания населения, который характеризуется показателями: x_1 - число врачей на 10 тыс. жителей и x_2 - число больничных коек на 10 тыс. жителей.

№ региона	1	2	3	4	5	6
x_1	34,8	31,2	32,1	35,7	30,2	34,2
x_2	126	112	123	128	115	123

За расстояние между объектами принять взвешенное евклидово расстояние с $\omega_1 = 0,4$ и $\omega_2 = 0,6$, а расстояние между кластерами измерять по принципу:

а) "ближайшего соседа";

б) "дальнего соседа".

Сравнить разбиение на кластеры по критерию минимума суммы внутриклассовых дисперсий. Исходные данные не нормализовывать.

7.18. Решить задачу 7.17, предварительно нормировав исходные данные.

7.19. Решить задачу 7.17 при условии, что расстояние между кластерами измеряются по принципу "центра тяжести" групп.

7.20. Решить задачу 7.17 при условии, что расстояние между кластерами измеряются по принципу "средней связи".

7.21. По агломеративному алгоритму провести классификацию n=5 районов, работа которых оценивалась по показателям: x_1 - производительности труда, объема реализованной продукции: x_2 - растениеводства и x_3 - животноводства с 1га посевной площади.

В качестве расстояния между объектами принять:

а) обычное евклидово расстояние;

б) взвешенное евклидово расстояние с весами: $\omega_1 = 0,2$; $\omega_2 = 0,4$; $\omega_3 = 0,4$.

Сравнить разбиение на два кластера по критерию минимума суммы внутриклассовых дисперсий.

Расстояние между кластерами определять по принципу "ближайшего соседа". Исходные данные не нормировать.

№ колхоза	1	2	3	4	5
x_1	8,22	4,33	6,43	6,39	4,92
x_2	0,25	0,49	0,51	0,27	0,32
x_3	0,41	0,38	0,51	0,42	0,55

7.22. Решить задачу 7.21, предварительно нормировав исходные данные.

7.23. Решить задачу 7.21, измеряя расстояние между кластерами по принципу "дальнего соседа". Исходные данные не нормировать.

7.24. Решить задачу 7.21, измеряя расстояние между кластерами по принципу "центра тяжести", не нормируя исходные данные.

7.25. По агломеративному алгоритму провести классификацию n=6 машиностроительных предприятий, работа которых оценивалась по показателям: x_1 - производительности труда, x_2 - удельного веса потерь от брака (%) и x_3 - фондоотдачи активной части производственных фондов.

№ колхозов	1	2	3	4	5	6
x_1	9,4	6,6	7,4	10,0	6,6	9,1
x_2	0,15	0,48	0,62	0,32	0,50	0,90
x_3	1,91	0,88	1,09	2,62	1,32	1,89

В качестве расстояния между объектами принять обычное евклидово расстояние, а расстояние между кластерами измерить по принципу:

а) "ближайшего соседа";

б) "дальнего соседа".

Сравнить разбиение на два кластера по критерию минимума суммы внутриклассовых дисперсий. Исходные данные не нормировать.

7.26. Решить задачу 7.25, предварительно пронормировав исходные данные.

7.27. Решить задачу 7.25 при условии, что расстояние между кластерами измеряется по принципу "центра тяжести" групп, не нормируя предварительно исходные данные.

7.28. Решить задачу 7.25 при условии, что расстояние между кластерами измеряется по принципу "средней связи".

7.4. Задания для самостоятельной работы на ЭВМ по дискриминантному и кластерному анализу

Дискриминантный анализ

Предприятия характеризуются пятью экономическими показателями:

y_1 - производительность труда;

x_5 - удельный вес рабочих в составе промышленно-производственного персонала;

x_7 - коэффициент сменности оборудования (смен);

x_9 - удельный вес потерь от брака (%);

x_{10} - фондоотдача активной части основных производственных фондов.

Значения этих показателей у предприятий представлены в Приложении 2.

В каждом варианте (табл.7.7) даны две обучающие выборки, первая из которых включает 9 предприятий группы *A*, а вторая - 8 предприятий группы *B*.

Требуется вычислить оценки значений дискриминантной функции для оставшихся предприятий и провести их классификацию. Дать экономическую интерпретацию результатов дискриминации.

Таблица 7.7

Варианты самостоятельных работ на ЭВМ по дискриминантному анализу

№ варианта	Номер предприятия	
	Группа *A*	Группа *B*
1	1,2,3,4,5,6,8,24,29	12,14,15,16,21,23,27,28
2	1,2,3,4,5,6,8,24,43	12,14,15,16,21,23,27,41
3	1,2,3,4,5,6,8,29,43	12,14,15,16,21,23,28,41
4	1,2,3,4,5,6,24,29,43	12,14,15,16,21,27,28,41
5	1,2,3,4,5,8,24,29,43	12,14,15,16,23,27,28,41
6	1,2,3,4,6,8,24,29,43	12,14,15,21,23,27,28,41
7	1,2,3,5,6,8,24,29,43	12,14,16,21,23,27,28,41
8	1,2,4,5,6,8,24,29,43	12,15,16,21,23,27,28,41
9	1,3,4,5,6,8,24,29,43	14,15,16,21,23,27,28,41
10	2,3,4,5,6,8,24,29,43	12,14,15,16,21,23,27,28
11	1,2,3,4,5,6,8,24,29	12,14,15,16,21,23,27,41
12	1,2,3,4,5,6,8,24,43	12,14,15,16,21,23,28,41
13	1,2,3,4,5,6,8,29,43	12,14,15,16,21,27,28,41
14	1,2,3,4,5,6,24,29,43	12,14,15,16,23,27,28,41
15	1,2,3,4,5,8,24,29,43	12,14,15,21,23,27,28,41
16	1,2,3,4,6,8,24,29,43	12,14,16,21,23,27,28,41
17	1,2,3,5,6,8,24,29,43	12,15,16,21,23,27,28,41
18	1,2,4,5,6,8,24,29,43	14,15,16,21,23,27,28,41
19	1,3,4,5,6,8,24,29,43	12,14,15,16,21,23,27,28
20	2,3,4,5,6,8,24,29,43	12,14,15,16,21,23,27,41
21	1,2,3,4,5,6,8,24,29	12,14,15,16,21,23,28,41
22	1,2,3,4,5,6,8,24,43	12,14,15,16,21,27,28,41
23	1,2,3,4,5,6,8,29,43	12,14,15,16,23,27,28,41
24	1,2,3,4,5,6,24,29,43	12,14,15,21,23,27,28,41
25	1,2,3,4,5,8,24,29,43	12,14,16,21,23,27,28,41

Кластерный анализ

По данным Приложения 2 по алгоритму кластерного анализа провести классификацию предприятий и построить дендрограмму.

По дендрограмме выбрать окончательный вариант разбиения предприятий на кластеры и дать интерпретацию полученным результатам, используя статистические характеристики кластеров (векторы средних и средних квадратических отклонений).

ГЛАВА 8

КАНОНИЧЕСКИЕ КОРРЕЛЯЦИИ

8.1. Основные понятия

Метод канонических корреляций позволяет находить максимальные корреляционные связи между двумя группами случайных величин. Эта зависимость определяется при помощи новых аргументов - канонических величин, вычисленных как линейные комбинации исходных признаков. Новые канонические величины выбираются таким образом, чтобы новые координаты непосредственно указывали значение корреляции. В каждой группе отыскиваются линейные комбинации исходных величин, имеющие максимальную корреляцию, они и являются первыми координатами новых систем. После этого в каждой группе рассматриваются следующие линейные комбинации, у которых корреляция больше, чем между любыми другими линейными комбинациями, не коррелированными с первыми линейными комбинациями. Построение продолжается до тех пор, пока не будут полностью получены две новые координатные системы.

Метод канонических корреляций используется для научного обоснования системы показателей при проведении многомерного статистического анализа, а также как основной инструментарий в каноническом **факторном анализе.**

Пример

Для пояснения многомерного метода канонических корреляций возьмем многомерный случайный вектор x порядка p с дисперсиями компонент, равными единице, математическим ожиданием, равным нулю, и положительно определенной корреляционной матрицей R.

Вектор x разбивается на два подвектора размерности p_1 и p_2, где $p_1+p_2=p$; $p_1 \le p_2$.

$$x = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix}.$$

Пусть подвектор $x^{(1)}$ содержит p_1 признаков, относящихся к характеристикам эффективности производства: производительность труда, фондоотдача, себестоимость и т.д., а $x^{(2)}$ содержит признаки-аргументы, характери-

зующие факторы производства: составляющие материальных и трудовых затрат,организационные, технологические и др.

Задача заключается в выявлении максимальных связей между этими двумя группами показателей. Для этого вводятся новые переменные U и V, которые представляют собой линейные комбинации результативных показателей-компонент подвектора $x^{(1)}$ и аргументов-признаков $x^{(2)}$:

$$U = \sum_{j=1}^{p_1} l_j x_j^{(1)}, \qquad V = \sum_{i=1}^{p_2} m_i x_i^{(2)}, \tag{8.1}$$

где U и V - канонические величины.

Общая корреляционная матрица вектора x будет состоять из подматриц, которые позволяют решить задачу.

$$XX^T = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix} \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix}^T = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix},$$

где R_{11} - корреляционная матрица группы показателей эффективности, размерности $p_1 \times p_1$;

R_{22} - корреляционная матрица показателей-аргументов размерности $p_2 \times p_2$;

R_{12} и R_{21} - матрицы взаимных корреляций первой и второй групп признаков;

$R_{21} = R_{12}^T$, где R_{12}^T - транспонированная матрица.

Решение

На основе определения собственных значений и собственных векторов от некоторой функции корреляционной матрицы исходных признаков размерности $p \times p$

$$B = R_{11}^{-1} \cdot R_{12} \cdot R_{22}^{-1} \cdot R_{21}, \tag{8.2}$$

где R_{11}^{-1} и R_{22}^{-1} - соответствующие обратные матрицы.

Собственные значения матрицы B, ранжированные по убыванию, равняются квадратам канонических корреляций.

Метод канонических корреляций является обобщением множественной корреляции.

Обратимся к **множественной корреляции**, которая является мерой связи между одной случайной величиной и множеством других случайных величин. Но, как известно, множественная корреляция является максимальной корреляцией между одной случайной величиной и линейной функцией других случайных величин. Эта концепция была обобщена Хотеллингом на случай связи между множествами случайных величин.

Рассмотренный статистический подход очень полезен, если у исследователя имеется два больших множества величин и он хочет определить взаимосвязи между ними. При этом можно ограничиться рассмотрением небольшого числа **наиболее коррелированных линейных комбинаций из каждого подмножества.**

Взаимосвязь между этими подмножествами почти полностью описывается **коэффициентом корреляции** между несколькими первыми каноническими случайными величинами.

8.2. Канонические корреляции и канонические величины генеральной совокупности

Пусть дан случайный вектор x размерности p центрированных величин $x^{(1)},x^{(2)},...,x^{(p)}$, имеющий нулевое математическое ожидание Mx=0 и ковариационную матрицу Σ. Вектор x разбивается на два подвектора размерности p_1 и p_2 т.е. $x=\begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$, причем $p_1 \le p_2$, а ковариационная матрица на подматрицы

$$\Sigma=\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

где Σ_{11} - ковариационная матрица, характеризующая взаимосвязь результативных показателей, элементами которой являются коэффициенты ковариации

$$\sigma_{ij}^{(1)} = M(x_i^{(1)} - Mx_i^{(1)})(x_j^{(1)} - Mx_j^{(1)}), \quad i,j=1,2,...,p_1;$$

Σ_{22} - ковариационная матрица, характеризующая взаимосвязь определяющих показателей (аргументов);

Σ_{12} - ковариационная матрица размерности $(p_1 \times p_2)$, характеризующая взаимосвязь показателей первой и второй групп ($\Sigma_{21}=\Sigma_{12}^T$).

Рассмотрим линейные комбинации вида:

$$U=\alpha^T X^{(1)} \text{ компонент вектора } X^{(1)}$$

и

$$V=\beta^T X^{(2)} \text{ компонент вектора } X^{(2)},$$

где α и β - векторы размерности соответственно p_1 и p_2.

Математическое ожидание U и V равно нулю:

$$\begin{aligned} MU &= M\alpha^T X^{(1)} = \alpha^T MX^{(1)} = 0; \\ MV &= M\beta^T X^{(2)} = \beta^T MX^{(2)} = 0. \end{aligned} \tag{8.3}$$

Выберем векторы α и β так, чтобы дисперсии величин U и V были равны единице:

$$\begin{cases} 1 = MU^2 = M\alpha^T X^{(1)} X^{(1)T}\alpha = \alpha^T\Sigma_{11}\alpha; \\ 1 = MV^2 = M\beta^T X^{(2)} X^{(2)T}\beta = \beta^T\Sigma_{22}\beta. \end{cases} \tag{8.4}$$

Таким образом, U и V - нормированные случайные величины.

Коэффициент корреляции между U и V

$$MUV^T = M\alpha^T X^{(1)} X^{(2)T}\beta = \alpha^T\Sigma_{12}\beta. \tag{8.5}$$

Задача состоит в определении векторов α и β, при которых выражение (8.5) достигает максимума при условиях (8.3) и (8.4).

Применим **метод множителей Лагранжа.** Пусть λ, μ - множители Лагранжа, тогда функция Лагранжа определяется выражением

$$\psi = \alpha^T\Sigma_{12}\beta - \frac{1}{2}\lambda(\alpha^T\Sigma_{11}\alpha - 1) - \frac{1}{2}\mu(\beta^T\Sigma_{22}\beta - 1). \tag{8.6}$$

Продифференцируем функцию ψ по компонентам векторов α и β. Приравняв векторы производных нулю, получим:

$$\begin{cases} \dfrac{\partial\psi}{\partial\alpha} = \Sigma_{12}\beta - \lambda\Sigma_{11}\alpha = 0; \\ \dfrac{\partial\psi}{\partial\beta} = \Sigma_{12}^T\alpha - \mu\Sigma_{22}\beta = 0. \end{cases} \tag{8.7}$$

Умножив выражение (8.7) слева соответственно на α^T и β^T, получим:

$$\begin{cases} \alpha^T\Sigma_{12}\beta - \lambda\alpha^T\Sigma_{11}\alpha = 0; \\ \beta^T\Sigma_{12}\alpha - \mu\beta^T\Sigma_{22}\beta = 0. \end{cases} \tag{8.8}$$

Из (8.3) и (8.4) следует, что $\alpha^T\Sigma_{11}\alpha = 1$; $\beta^T\Sigma_{22}\beta = 1$. Тогда $\lambda = \mu = \alpha^T\Sigma_{12}\beta$.

Перепишем выражение (8.7)

$$\begin{cases} -\lambda\Sigma_{11}\alpha + \Sigma_{12}\beta = 0; \\ \Sigma_{21}\alpha - \lambda\Sigma_{22}\beta = 0. \end{cases} \tag{8.9}$$

С учетом $\Sigma_{12}^T = \Sigma_{21}$ выражение (8.9) запишем в матричном виде

$$\begin{pmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0. \tag{8.10}$$

Для того чтобы решение удовлетворяло требованиям (8.3) и (8.4), т.е. существовало нетривиальное решение, матрица слева должна быть вырожденной, а ее определитель - равен нулю:

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} = 0. \tag{8.11}$$

Определитель в (8.11) представляет собой многочлен степени p_1 относительно λ. Он имеет p_1 корней $\lambda_1 \ge \lambda_2 \ge ... \lambda_{p_1}$.

Из выражения (8.8) видно, что $\lambda = \alpha^T\Sigma_{12}\beta$. Тогда λ равна коэффи-

циенту корреляции между $U = \alpha^T X^{(1)}$ и $V = \beta^T X^{(2)}$, где α и β удовлетворяют уравнению (8.10) при некотором значении λ.

Нам же требуется определить максимальный коэффициент корреляции. Пусть это будет значение $\lambda = \lambda_1$.

Если максимальное значение коэффициента корреляции λ_1, то решение уравнения (8.10) при $\lambda = \lambda_1$ представляет собой $\alpha^{(1)}$ и $\beta^{(1)}$. Тогда можно записать, что

$$\begin{aligned} U_1 &= \alpha^{(1)T} X^{(1)}; \\ V_1 &= \beta^{(1)T} X^{(2)}. \end{aligned} \tag{8.12}$$

Следовательно, U_1 и V_1 являются нормированными линейными комбинациями компонент векторов $X^{(1)}$ и $X^{(2)}$ соответственно, а коэффициент корреляции между ними максимален.

Определим вторую линейную комбинацию векторов $X^{(1)}$ и $X^{(2)}$ и r-ю линейную комбинацию.

Пусть это будет

$$\begin{aligned} U_2 &= \alpha^{(2)T} X^{(1)}; \\ V_2 &= \beta^{(2)T} X^{(2)}; \\ &\cdots\cdots\cdots \\ U_r &= \alpha^{(r)T} X^{(1)}; \\ V_r &= \beta^{(r)T} X^{(2)}. \end{aligned} \tag{8.13}$$

Получено r линейных комбинаций для U и V, для которых имеются r коэффициентов корреляции, являющихся корнями уравнения (8.11): $\lambda^{(1)} = \lambda_1$, $\lambda^{(2)} = \lambda_2, \ldots, \lambda^{(r)} = \lambda_r$.

Линейные комбинации U_r и V_r не коррелированны с $U_1, V_1, \ldots, U_{r-1} V_{r-1}$ и их коэффициент корреляции больше коэффициента корреляции между двумя линейными комбинациями, некоррелированными с $U_1, V_1, \ldots, U_{r-1} V_{r-1}$, U_r, V_r.

Запишем условие некоррелированности для U и U_i.

$$O = MUU_i = M\alpha^T X^{(1)} X^{(1)T} \alpha^{(i)} = \alpha^T \Sigma_{11} \alpha^{(i)}. \tag{8.14}$$

Определим условия некоррелированности для V и V_i, а также для V и U_i.

Если $\lambda^{(i)} \neq 0$, то $\Sigma_{11} \alpha^{(i)} = \dfrac{1}{\lambda^{(i)}} \Sigma_{12} \beta^{(i)}$,

следовательно, $$O = \alpha^1 \Sigma_{12} \beta^{(i)} = MUV_i. \tag{8.15}$$

Если $\lambda^{(i)} = 0$, то $\Sigma_{12} \beta^{(i)} = 0$ и выполняется (8.15).

Условие некоррелированности V и V_i имеет вид

$$0 = MVV_i = \beta^T \Sigma_{22} \beta^{(i)}. \tag{8.16}$$

Условие некоррелированности V и U_i запишется

$$0 = MVU_i = \beta^T \Sigma_{21} \alpha^{(i)}. \tag{8.17}$$

Отметим, что корни уравнения (8.11) при $\lambda_i \neq 0$ и $\lambda_i \neq \lambda_k$ составляют матрицу

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & . & 0 \\ 0 & \lambda_2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & \lambda_{p1} \end{pmatrix}, \tag{8.18}$$

так как число линейно независимых комбинаций U_i не может превзойти $\min(p_1, p_2)$, а по условию $p_1 \leq p_2$.

8.3. Канонические корреляции и их интерпретация

Пусть $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$, где вектор $X^{(1)}$ p_1-мерный; вектор $X^{(2)}$ p_2-мерный; $p_2 = p - p_1 \geq p_1$. Тогда r-й парой канонических величин является пара линейных комбинаций $U_r = \alpha^{(r)T} X^{(1)}$ и $V_r = \beta^{(r)T} X^{(2)}$, каждая из которых имеет единичную дисперсию и не коррелированна с $(r-1)$ парами канонических величин.

Коэффициент корреляции между элементами этой пары максимален. Этот коэффициент корреляции называется **r-й канонической корреляцией.**

Коэффициенты в линейных комбинациях $\alpha^{(r)T} X^{(1)}$ и $\beta^{(r)T} X^{(2)}$, определяющие r-ю пару канонических величин, удовлетворяют уравнению (8.10) при условии $\lambda = \lambda_r$ и выполнения (8.3) и (8.4).

Рассмотрим интерпретацию канонических корреляций.

Даны две случайные величины U и V со средними значениями, равными нулю $MU=0$; $MV=0$; дисперсиями σ_U^2 и σ_V^2 и коэффициентом корреляции ρ.

При помощи bV, где b - множитель, найдем приближение величины U.

Определим среднюю квадратическую ошибку приближения

$$\begin{aligned} M(U - bV)^2 &= \sigma_U^2 - 2b\sigma_U \sigma_V \rho + b^2 \sigma_V^2 = \\ &= \sigma_U^2 (1 - \rho^2) + (b\sigma_V - \sigma_U)^2. \end{aligned} \tag{8.19}$$

Она имеет минимум при $b = \frac{\sigma_U \rho}{\sigma_V}$.

Величина bV представляет собой линейную оценку величины U по значениям величины V, тогда $\sigma_U^2(1-\rho^2)$ равна дисперсии оценки.

Отношение дисперсии ошибки предсказания к дисперсии величины U равно

$$\frac{\sigma_U^2(1-\rho^2)}{\sigma_U^2} = 1-\rho^2. \qquad (8.20)$$

Величина (8.20) является **мерой относительного влияния** V на U или мерой относительной эффективности величины V при предсказании U.

Следовательно, чем больше ρ^2 или $|\rho|$, тем более точно можно по величине V оценить величину U.

Рассмотрим теперь вектор

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}.$$

Пусть линейная комбинация $V = \beta^T X^{(2)}$ служит для предсказания линейной комбинации $U = \alpha^T X^{(1)}$.

Величина V будет наилучшим образом предсказывать величину U, если коэффициент корреляции между U и V максимален. Тогда можем утверждать, что $\alpha^{(1)T} X^{(1)}$ представляет собой линейную комбинацию компонент $X^{(1)}$, которую можно наилучшим образом предсказывать. Такое наилучшее предсказание осуществляет комбинация $\beta^{(1)T} X^{(2)}$.

За меру точности определения величины U по V можно выбрать

$$M(bV)^2 = \rho^2 \frac{\sigma_U^2}{\sigma_V^2} MV^2 = \rho^2 \sigma_U^2. \qquad (8.21)$$

Мерой относительного эффекта может служить выражение

$$\frac{M(bV)^2}{MU^2} = \rho^2. \qquad (8.22)$$

Максимальный эффект, который может оказывать линейная комбинация компонент вектора $X^{(2)}$ на линейную комбинацию компонент вектора $X^{(1)}$, оказывает $\beta^{(1)T} X^{(2)}$ на $\alpha^{(1)T} X^{(1)}$.

Перейдем к частному случаю, когда p_1=1. Тогда единственная каноническая корреляция представляет собой вектор множественной корреляции между $X^{(1)} = X_1$ и $X^{(2)}$.

8.4. Оценка канонических корреляций и канонических величин

Пусть $x_1, x_2, \ldots, x_n$ – n наблюдений над совокупностью $X \in N(\mu, \Sigma)$.

Пусть вектор X разбит на два подвектора размерностей p_1 и p_2 соответственно. Тогда

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \qquad (8.23)$$

и

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \qquad (8.24)$$

Оценками наибольшего правдоподобия для величин $\lambda_1, \lambda_2, \ldots, \lambda_p$ являются корни уравнения

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} = 0. \qquad (8.25)$$

Векторы $\hat{\alpha}^{(j)}$ и $\hat{\beta}^{(j)}$ удовлетворяют уравнениям

$$\begin{pmatrix} -\hat{\lambda}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\hat{\lambda}\Sigma_{22} \end{pmatrix} \begin{pmatrix} \hat{\alpha}^{(j)} \\ \hat{\beta}^{(j)} \end{pmatrix} = 0; \qquad (8.26)$$

$$\hat{\alpha}^{(j)T} \Sigma_{11} \hat{\alpha}^{(j)} = 1; \qquad (8.27)$$

$$\hat{\beta}^{(j)T} \Sigma_{22} \hat{\beta}^{(j)} = 1 \text{ и т.д.} \qquad (8.28)$$

Таким образом, рассмотрены подходы к изучению связей между двумя множествами случайных величин.

Канонический анализ, как и компонентный анализ, реализуется в форме задачи нахождения собственных значений и собственных векторов от некоторой функции корреляционной матрицы исходных признаков $B = R_{11}^{-1} R_{12} R_{22}^{-1} R_{21}$, где R_{11} и R_{22} - корреляционные матрицы групп результативных признаков $X^{(1)}$ и показателей-аргументов $X^{(2)}$ (их размерности соответственно равны $p_1 \times p_1$ и $p_2 \times p_2$); R_{12} - матрица взаимных корреляций первой и второй групп ($R_{21} = R_{12}^T$).

Собственные значения R, ранжированные по убыванию, равняются квадратам канонических корреляций ρ_k^2. Левые и правые собственные векторы - соответствующим каноническим переменным групп исходных признаков $X^{(1)}$ и показателей $X^{(2)}$.

С точки зрения канонического анализа обе группы равноценны.

Для разрешимости задачи требуется, чтобы корреляционные матрицы R_{11} и R_{22} были положительно определены, т.е. в соответствующей группе не должны существовать линейно зависимые признаки (показатели). В противном случае следует один или несколько признаков-показателей исключить из рассмотрения.

Канонические переменные обладают следующими свойствами.

Они являются линейными комбинациями исходных показателей соответствующих групп.

Канонические переменные одной группы взаимно не коррелированны.

Канонические переменные выбраны таким образом, чтобы соответствующие канонические корреляции были максимальны.

Канонические переменные упорядочены по мере убывания соответствующих канонических корреляций;

Число используемых канонических корреляций обычно значительно меньше числа исследуемых показателей p_2.

Канонические корреляции всегда неотрицательны, причем их основные свойства совпадают со свойствами обыкновенных множественных коэффициентов корреляций. Чем больше канонические корреляции, тем сильнее связаны рассматриваемые группы признаков $X^{(1)}$ и показателей $X^{(2)}$.

Значимость канонических переменных (или, что то же, отличие от нуля канонических корреляций) проверяется при помощи критерия χ^2. Если вычислено p_1 канонических корреляций, то для каждого m ($m=\overline{1,p_1}$) следует проверить гипотезы:

$$H_0^m: \rho_m = \rho_{m+1} = \ldots = \rho_{p_1} = 0,$$

т.е. все канонические корреляции, начиная с ρ_m, равняются нулю.

$H_1^m: \rho_m \neq 0$ (по крайней мере ρ_m отличается от нуля).

При этом учитывается, что $\rho_m > \rho_{m+1} > \ldots \rho_{p1}$.

Вычисление χ^2-статистики для каждой из канонических корреляций предусмотрено программой канонического анализа и определяется по формуле:

$$\chi^2_{набл} = -\left\{ N - m - \frac{1}{2}(p_1 + p_2 + 1) + \sum_{k=1}^{m-1} r_k^2 \right\} \ln \prod_{k=m}^{p_1} (1 - r_k^2), \quad (8.29)$$

где N - объем выборки,

r_k^2 - оценка канонического коэффициента детерминации.

Если значение $\chi^2_{набл}$ больше критического при выбранном уровне значимости α (например, 0,05) и числе степеней свободы $\nu = \left[(p_2 - m + 1)(p_1 - m + 1)\right]$, то принимается гипотеза H_1^m. Процедура повторяется для следующей (m+1) канонической корреляции. Если вычисленное значение χ^2-статистики меньше соответствующего табличного значения, то нулевая гипотеза H_1^{m+1} не отвергается, т.е. зависимость между группами исследуемых признаков показателей уже описана каноническими переменными с индексами 1,2,...,m.

Если при некотором значении m_0 нулевая гипотеза не отвергается, т.е. каноническая корреляция ρ_{m_0} равна нулю, то равны нулю и все последующие ρ_m при $m = m_0+1, m_0+2, \ldots, p_1$.

Следует интерпретировать только такие канонические переменные, которые соответствуют значимым (отличным от нуля) каноническим корреляциям. При интерпретации результатов можно использовать ту же схему, что и в компонентном анализе.

В процессе канонического анализа исходные данные $X^{(2)}$ и $X^{(1)}$ приводятся к **стандартизированному виду,** поэтому коэффициенты в выражениях для канонических переменных характеризуют силу влияния соответствующих исходных признаков и показателей, что позволяет получить их **ранжированные последовательности.** Отсев несущественных переменных может осуществляться на основе **многошаговой процедуры,** при которой на каждом шагу отбрасывается только одна переменная, наименее существенная в исходной последовательности. Для сравнения канонических корреляций исходного ρ_k и ρ_{k+1} наборов факторов используется **z-преобразование Фишера и критерий**

$$t_{набл} = (z_k - z_{k+1}) \sqrt{\frac{N-3}{2}}, \quad (8.30)$$

где $z_l = \frac{1}{2} \ln \frac{1+r_l}{1-r_l}$;

$t_{набл}$ - имеет нормированный нормальный закон распределения;

l-индекс, относящийся к признакам вектора $X^{(1)}$.

Если канонические корреляции ρ_k и ρ_{k+1} отличаются незначимо, процесс сокращения продолжается.

При каноническом анализе **процедура отсева** учитывает всю сложность структуры связей как внутри групп признаков и показателей, так и между этими группами. Признак, значимо влияющий хотя бы на один показатель и являющийся незначимым для других, уже не может быть отброшен. Процедура отсева основывается на принципе дополнительности: признаки $X^{(2)}$ исключаются с учетом того, какие показатели $X^{(1)}$ исключаются.

Достаточно компактная, максимально информативная система исходных признаков и показателей, полученная при помощи метода канонического анализа, может служить основой для дальнейших исследований при помощи методов регрессионного, компонентного и факторного анализа.

8.5. Использование канонических корреляций в практике экономического анализа

Пример 8.1

Для статистического анализа эффективности подземной угледобычи была собрана информация по 30 угольным шахтам. В качестве исходных параметров были отобраны 8 показателей эффективности угледобычи и 19 признаков, ее определяющих.

Показатели эффективности угледобычи:

y_1 - выработка валовой продукции на одного работающего, т ;

y_2 - среднемесячная производительность труда рабочего по добыче угля, т ;

y_3 - среднемесячная производительность труда промышленно-производственного персонала по добыче угля, т ;

y_4 - среднемесячная производительность труда одного рабочего на очистных работах, т.;

y_5 - полная себестоимость добычи 1т угля;

y_6 - фондоотдача;

y_7 - фондоемкость;

y_8 - трудоемкость работ по добыче, чел.-дни на 1000 т.

Показатели, определяющие эффективность угледобычи:

x_1 - среднесуточная добыча угля (нагрузка на шахту), т;

x_2 - среднединамическая мощность пласта;

x_3 - максимальная глубина разработки, м;

x_4 - приток воды в шахту, м³/час ;

x_5 - количество метана на 1т среднемесячной добычи, м³;

x_6 - среднесуточная нагрузка на забой, т;

x_7 - среднедействующая длина линии очистных забоев, м;

x_8 - среднемесячное продвижение линии действующих очистных забоев, м;

x_9 - удельный вес добычи угля из комплексно-механизированных забоев по отношению ко всей добыче из очистных забоев, м;

x_{10} - удельный вес комбайновой проходки выработок, %;

x_{11} - удельный объем проведения подготовительных выработок на 1000 т добычи, м³;

x_{12} - удельный объем выработок, закрепленных металлической и железобетонной крепью, %;

x_{13} - протяженность выработок на 1000 т добычи, м;

x_{14} - среднемесячная производительность работающего электровоза, тыс. км;

x_{15} - протяженность действующих рельсовых путей, км;

x_{16} - удельный вес забалластированных рельсовых путей, %;

x_{17} - удельная емкость вагонеточного парка на 1000 т среднесуточной выработки, т;

x_{18} - объем породы на 1000 т добычи, т;

x_{19} - зольность угля, %.

Вычислить первую пару канонических величин и каноническую корреляцию по показателям ($x_1 \div x_8$, $x_{11} \div x_{19}$) и показателям эффективности (y_2, y_5, y_6).

Решение

В качестве исходных групп были взяты признаки ($x_1 \div x_8$, $x_{11} \div x_{19}$) и показатели y_2, y_5, y_6. Было получено три пары канонических переменных. Первая пара канонических переменных имеет вид:

$$\begin{aligned} &0{,}352x_1+0{,}290x_2-0{,}334x_3+0{,}084x_4-0{,}408x_5+0{,}134x_6-0{,}023x_7- \\ &-0{,}407x_8+0{,}212x_{11}+0{,}110x_{12}-0{,}013x_{13}+0{,}337x_{14}-0{,}503x_{15}+ \\ &+0{,}060x_{16}-0{,}143x_{17}+0{,}146x_{18}-0{,}019x_{19}; \\ &0{,}769y_2-0{,}611y_5+0{,}187\,y_6. \end{aligned} \qquad (8.31)$$

Для удобства приняли $x^{(1)} = y$, $x^{(2)} = x$.

Величина соответствующей канонической корреляции составила 0,93523; вычисленное значение $\chi^2_{набл}$ =78,99, которое превосходит $\chi^2_{0,05}$ = =68,67. Таким образом, первая каноническая корреляция значима. Для второй пары канонических переменных величина канонической корреляции равна 0,81420, вычисленное значение χ^2 =38,50 меньше $\chi^2_{0,05}$ =46,19, поэтому вторая (а следовательно, и третья) каноническая корреляция незначимая.

Пример 8.2

Используя данные примера 8.1, провести сокращение числа показателей, определяющих эффективность угледобычи при помощи канонического анализа.

Решение

Для сокращения исходного набора признаков на основе (8.31) была составлена их ранжированная по силе влияния в первой канонической компоненте последовательность:

$$\begin{aligned} &x_{15} > x_5 > x_8 > x_1 > x_{14} > x_3 > x_2 > x_{11} > x_{18} > x_{17} > x_6 > x_{12} > x_4 > \\ &> x_{16} > x_7 > x_{19} > x_{13}. \end{aligned} \qquad (8.32)$$

Наименее значимыми в данной последовательности оказались признаки x_7, x_{19}, x_{13}. Поскольку порядок признаков, расположенных в "хвосте" последовательности (8.30), случаен, то исходя из эффективности производства на первом шаге было решено оставить признак x_{13} - протяженность выборок на 1000 т добычи, а признаки x_7 и x_{19} - исключить.

Таким образом, на втором шаге рассматривалась зависимость между набором из 15 признаков и тремя показателями эффективности угледобычи. Значимой оказалась только первая каноническая корреляция, равная 0,93518. Для сравнения канонических корреляций исходного и сокращенного наборов использовалось Z-преобразование Фишера и критерий (8.30). В нашем случае n=30; $\left(\sqrt{\frac{n-3}{2}}\right)^{-1} = 0,272$, поэтому для α=0,05

$$\Delta = Z(\rho_1) - Z(\rho_2) < 0,533\% . \qquad (8.33)$$

Согласно критерию (8.33) различие между каноническими корреляциями исходного и сокращенного наборов признаков незначимо. Анализ канонической переменной, полученной на втором шаге, показал, что наименьшее влияние на показатели эффективности оказывает признак x_4. Поэтому признак x_4 был исключен из дальнейшего рассмотрения.

Данные, характеризующие процедуру сокращения исходного набора признаков при помощи канонического анализа, приведены в табл.8.1.

Таблица 8.1

Сокращение исходного набора признаков при помощи метода канонического анализа

№ шага	Число переменных	Каноническая корреляция	$\Delta = Z_1 - Z_i$, $i = \overline{2,7}$	Удаление переменных
1	17	0,93523	-	x_{19}, x_7
2	15	0,93518	0,004	x_4
3	14	0,93466	0,0145	x_{12}
4	13	0,93286	0,0187	x_6
5	12	0,93213	0,0241	x_{16}
6	11	0,93210	0,0244	x_{17}
7	10	0,93178	0,0268	-

Как видно из табл.8.1, сокращение исходного набора признаков с 17 до 10 не привело к значимому различию соответствующих канонических корреляций.

Канонические переменные, полученные на последнем шаге, имеют вид:

$$0,387x_1+0,285x_2-0,331x_3-0,376x_5+0,463x_8+0,095x_{11}- \\ -0,142x_{13}+0,394x_{14}-0,476x_{15}+0,157x_{18}; \qquad (8.34) \\ 0,836\,y_2-0,524\,y_5+0,157\,y_6.$$

Если интерпретировать вторую каноническую переменную, как показатель уровня эффективности производства, то направление влияния всех признаков первой канонической переменной соответствует увеличению общего уровня эффективности подземной угледобычи.

Ранжированная последовательность признаков, влияющих на показатели эффективности производственной деятельности шахт, имеет вид:

$$x_{15} > x_8 > x_{14} > x_1 > x_5 > x_3 > x_2 > x_{18} > x_{13} > x_{71}. \qquad (8.35)$$

Сравнение выражений (8.32) и (8.35) показывает, что группа факторов, оказывающих существенное влияние на показатели эффективности, после сокращения оставалась без изменений, хотя их порядок в ранжированной последовательности несколько изменился. В эту группу входят признаки, характеризующие горно-геологические условия производства - x_2, x_3, x_5; работу внутришахтного транспорта - x_{14}, x_{15}; очистные работы x_1, x_8.

8.6. Задание для самостоятельной работы

Провести исследование информации по совокупности шахт, приведенной в табл.8.2 и 8.3, методом канонических корреляций для заданных наборов показателей эффективности подземной угледобычи y и признаков x, ее определяющих. Названия признаков даны в примере 8.1.

Таблица 8.2

Варианты для самостоятельной работы

№ варианта	№ переменной 1 группы, y	№ переменной 2 группы, x
1	1,2	x_1-x_5
2	1,3	x_1-x_5
3	2,3	x_1-x_5
4	1,2,3	x_1-x_6
5	1,2,3	x_1-x_7
6	1,2	x_3-x_8
7	1,3	x_3-x_8
8	2,3	x_3-x_8
9	1,2	x_2-x_7
10	1,3	x_2-x_7
11	2,3	x_2-x_7
12	1,2	x_4-x_8
13	1,3	x_4-x_8
14	2,3	x_1-x_8
15	1,2	x_1-x_8
16	1,3	x_1-x_8
17	2,3	x_1-x_8
18	1,2	x_1-x_3 и x_5-x_8
19	1,3	x_1-x_3 и x_5-x_8
20	2,3	x_1-x_3 и x_5-x_8

Таблица 8.3

№ шахты	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y_1	y_2	y_3
1	879	2,80	311	250	146	198	114	34,7	3090	23,0	19,0
2	1533	3,24	250	420	57	212	212	30,9	3388	23,9	20,0
3	2526	3,66	220	400	18,7	260	103	41,4	7525	56,9	46,5
4	1646	1,55	307	200	62	243	582	33,8	3795	27,4	23,4
5	1997	3,26	360	400	123	115	2978	3,60	0,40	22,9	25,3
6	6005	2,60	275	420	21,0	166	940	20,2	0,00	18,7	50,9
7	1824	2,96	343	120	19,68	160	858	10,4	0,00	44,3	36,5
8	2693	2,53	280	420	9,92	128	206	37,8	5360	40,2	33,1
9	3416	3,0	190	800	14,0	168	509	34,2	6514	47,5	40,2
10	1386	2,44	258	439	27,0	93	194	24,7	4995	37,8	30,7
11	2547	3,00	330	335	39,5	222	164	67,0	4889	35,9	30,1
12	4217	2,75	250	520	16,0	190	439	55,8	7361	54,7	45,4
13	3506	3,17	247	320	15,0	195	409	40,3	4878	36,1	30,0
14	2531	3,63	280	380	29,5	179	222	37,1	5344	39,4	32,8
15	2229	3,09	260	430	10,0	28	177	45,8	5938	43,4	36,2
16	2161	2,99	240	90,0	27,0	251	58	68,7	5268	39,0	32,4
17	3169	2,97	126	400	4,50	313	325	43,8	8299	62,3	51,1
18	3552	2,10	88,0	475	29,0	178	93	99,8	7041	52,0	43,1
19	1332	2,72	310	92,0	12,0	208	117	57,2	5155	40,6	31,8
20	3419	3,18	145	250	5,30	445	241	56,6	9011	66,8	55,5
21	1194	3,03	290	165	58,0	184	150	31,9	4697	35,8	28,7
22	2828	2,86	330	260	29,0	257	219	48,0	7008	51,5	42,7
23	2748	2,17	502	180	17,0	329	613	36,4	5467	36,3	31,5
24	1196	3,21	263	485	7,00	193	127	39,6	3535	25,7	21,7
25	2265	3,69	100	128	1,00	177	11	105,1	10644	76,7	65,7
26	1241	1,80	720	320	43,0	294	432	23,6	2768	19,2	16,7
27	3217	1,77	420	400	7,00	465	920	36,3	5225	37,0	32,1
28	1997	1,82	360	220	8,00	311	643	31,0	5167	38,1	31,5
29	1983	1,66	230	474	1,00	414	742	31,9	8393	45,8	39,1
30	842	3,47	289	250	21,0	165	175	23,6	3810	28,4	23,2

ГЛАВА 9

РОБАСТНЫЕ МЕТОДЫ ОЦЕНИВАНИЯ

Методы оценивания, чувствительные к "грубым ошибкам" (резко выделяющимся наблюдениям) называются **неустойчивыми.**

Методы оценивания, учитывающие наличие "грубых ошибок" и позволяющие при этом достаточно точно определять оценки параметров, называются **робастными** или устойчивыми. Например, методы Хубера, Винзора, Пуанкаре для симметричных распределений, метод джеккнайф-оценок для асимметричных распределений и ряд других методов.

9.1. Грубые ошибки

При обработке экономической информации, как правило, считают, что экономические показатели подчиняются нормальному распределению. Однако практика обработки такой информации показывает, что экономические показатели не так часто подчиняются теоретическому нормальному распределению. Наблюдаются отклонения как односторонние, так и двух-сторонние, когда "хвосты" дифференциального закона оказываются более тяжелыми, чем можно было предположить, исходя из данных таблиц нормального распределения. Иногда статистическая информация по данному показателю, подлежащая обработке, представляет собой смесь нескольких законов распределения с разными дисперсиями. Встречаются смеси основного нормального распределения с распределениями других видов. Наблюдаются случаи, когда из-за малого объема выборки не представляется возможным достаточно точно определить вид закона, засоряющего распределения. Кроме того, хорошо известно, что при применении метода наименьших квадратов небольшое число грубых ошибок может существенно исказить значение характеристик распределения. Следовательно, необходимо остановиться на таких методах обработки экономической информации, которые были бы менее чувствительными к виду закона распределения и влиянию небольшого числа больших случайных отклонений.

Хорошо известен метод выявления аномальных наблюдений и их удаления из совокупности. Но в этом случае у исследователя всегда остается чувство неудовлетворенности при применении этого метода. При обработке многомерной статистической информации такой подход может привести к отбрасыванию слишком большого количества точек наблюдения. А если объектом наблюдения является предприятие отрасли, то такое отбрасывание может привести к анализу узкой группы предприятий вместо изучения

экономических закономерностей во всей отрасли. По этим причинам понятен интерес многих статистиков к разработке устойчивых (робастных) методов оценки экономической информации. Основы робастных методов оценки экономической информации были разработаны академиками А.Н.Колмогоровым, Н.В.Смирновым и Б.С.Ястремским. Дальнейшее развитие робастные методы получили в работах американского статистика Д.Тьюки и швейцарского математика Хубера. В настоящее время создан целый раздел прикладной математической статистики, включающий методы устойчивого оценивания показателей, подчиняющихся различным законам распределения. В современной науке используется несколько робастных методов оценки экономической информации.

При определении структуры неоднородных совокупностей здесь возникают две задачи. Первая задача заключается в разбиении общей неоднородной совокупности на некоторое число однородных совокупностей, а вторая - в оценке параметров совокупностей, которые содержат грубые ошибки.

При решении первой задачи необходимо:

- классифицировать элементы по однородным совокупностям;
- оценить параметры распределения однородных составляющих, входящих в общую неоднородную совокупность.

При решении второй задачи чаще используются методы непосредственного выявления грубых ошибок и методы, которые сводят к минимуму искажения, создаваемого грубыми ошибками, а также комбинированные методы, которые выделяют грубые ошибки и дают наиболее правдоподобную оценку параметров распределения, не искаженную грубыми ошибками.

Грубые ошибки могут появиться при сборе исходной информации, а также в результате искажения информации в каналах ее передачи.

Грубые ошибки появляются при построении регрессионных моделей при недостаточно строгом предварительном экономическом анализе исходных показателей объектов. В результате в совокупность могут быть включены "нетипичные" предприятия. Если же в выборку входит малое количество предприятий, то даже одно нетипичное предприятие может существенно исказить результаты анализа. При проведении многомерного анализа методы устойчивого оценивания параметров распределения приобретают особую актуальность.

9.2. Устойчивые методы оценки. Основные понятия

Пусть совокупность вместе с "обычными" значениями элементов содержит "грубые ошибки". Таким образом, основная масса элементов является реализациями случайной величины, закон распределения которой известен с точностью до некоторого параметра. Вероятность появления этих элементов в совокупности равна $1-\varepsilon$, где ε - вероятность

появления другой случайной величины - η, определяющей **грубые ошибки.** Однако это название является условным.

Известно, что средняя арифметическая оценка является несмещенной, состоятельной и эффективной оценкой математического ожидания. Однако, эффективность ее падает с утяжелением "хвостов" распределения, т.е. наличием достаточно большого числа наблюдений, значительно удаленных от среднего значения.

В прошлом веке Лаплас и Гаусс выявили преимущества и недостатки **средней абсолютной ошибки**

$$d_N = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \bar{x}\right|, \tag{9.1}$$

а также **средней квадратической ошибки**

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{9.2}$$

где $i = \overline{1, N}$;

x_i - i-е значение случайной величины;

$\bar{x}$ - среднее значение случайной величины.

При наличии серии наблюдений $x_1, x_2, \ldots, x_N$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i . \tag{9.3}$$

Средняя абсолютная (9.1) и средняя квадратическая (9.2) ошибки определяют разные характеристики распределения ошибок. Отношение их предельных значений для нормального распределения ошибок

$$\frac{d_\infty}{s_\infty} = \sqrt{\frac{2}{\pi}} . \tag{9.4}$$

Еще в начале XIX в. Ф.Бессаль отметил, что в большинстве случаев реальные распределения имеют утяжеленные хвосты, по сравнению с табличным нормальным распределением. В XX в. утяжеление хвоста реальных распределений отмечено во многих наборах статистической информации. Д.Тьюки предложил свою модель для оценки характеристик распределения с утяжеленными относительно нормальной совокупности хвостами. В ней предусматривается наличие нормальной совокупности с математическим ожиданием μ, дисперсией σ^2, которая засоряется другой нормальной совокупностью с этим же математическим ожиданием и с дисперсией $(3\sigma)^2 = 9\sigma^2$. **Распределение Тьюки** имеет вид:

$$F(x) = (1-\varepsilon)\Phi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon\Phi\left(\frac{x-\mu}{3\sigma}\right), \tag{9.5}$$

где

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt . \tag{9.6}$$

Часто для сравнения средней абсолютной ошибки со средней квадратической ошибкой используется **асимптотическая характеристика**

$$e(\varepsilon) = \lim_{N\to\infty} \frac{D\left(\frac{s_N}{s_\infty}\right)}{D\left(\frac{d_N}{d_\infty}\right)}, \qquad (9.7)$$

где $e(\varepsilon)$ - относительная асимптотическая эффективность d_N по отношению к s_N.

Если засорения нет, то для определения s_N требуется на 12% меньше наблюдений, чем для определения d_N. Однако уже при малом засорении это преимущество быстро падает. Наоборот, с ростом засорения относительная эффективность d_N быстро растет. При засорении, равном 0,18%, $e(0,18)=1$. Следовательно, если на 1000 наблюдений приходятся две грубые ошибки (в выборке $N=1000$), то преимущества s_n перед d_n исчезнут (табл.9.1).

Таблица 9.1

Относительная эффективность оценок с учетом степени засорения совокупности

ε	0,000	0,001	0,002	0,005	0,010	0,020	0,050	0,100	0,150	0,250	0,500
$e(\varepsilon)$	0,876	0,948	1,016	1,198	1,439	1,752	2,035	1,903	1,689	1,371	1,017

Как следует из табл.9.1 при засорении совокупности выше 0,2% эффективность d_N выше эффективности среднего квадратического отклонения s_N.

Итак, легко убедиться в том, что процедуры, предусмотренные теорией нормальных ошибок, не устойчивы к "грубым" ошибкам. Более устойчивыми оказываются процедуры, связанные с определением средней абсолютной ошибки d_N. Известен целый ряд методов исключения резко выделяющихся наблюдений.

Наиболее доступным и распространенным является **анализ измерений с точки зрения экономической сущности полученных наблюдений.** Для выявления резко выделяющихся наблюдений имеется ряд критериев, которые являются несмещенными, инвариантными по отношению к преобразованиям совокупности и требуют добавления константы или умножения каждого члена совокупности на положительное число.

Для **критериев Граббса и Смирнова-Граббса** получены точные распределения и их процентные точки. Они служат для выявления отдельных выделяющихся наблюдений, например, если из совокупности наблюдений $x_1, x_2, \ldots, x_N$ надо выделить одно наблюдение, которое можем предположить как ошибочное. Для обнаружения нескольких экстремальных наблюдений используют **критерий Титьена-Мура,** который выявляет "маскирующий эффект", имеющий место, когда в совокупности находятся два и более близких экстремальных значений случайной величины.

В экономических исследованиях очень важно выявить грубые ошибки при построении регрессионных моделей, проведении многомерного стохастического анализа.

9.3. Устойчивые оценки Хубера

Рассматриваются оценки параметров генеральной совокупности, когда основное и засоряющее распределение симметричны.

Примером может служить нормальное распределение

$$P(x,\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad (9.8)$$

распределение Лапласа

$$P(x,\mu) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}}, \qquad (9.9)$$

где μ - центр, относительно которого распределение обладает симметрией.

Для подобных распределений в общем виде можно записать

$$P(x,\mu) = P(x-\mu). \qquad (9.10)$$

Из (9.10) легко понять почему μ называется **параметром сдвига.** Плотность распределения с параметром μ отличается от плотности, соответствующей нулевому значению ($\mu=0$), сдвигом на величину μ.

Оценка μ по выборочным данным ($\hat{\mu}$) случайной величины ξ может быть получена из решения уравнения

$$\sum_{i=1}^{N} f(x_i - \hat{\mu}) = 0. \qquad (9.11)$$

Хубер предположил состоятельные несмещенные оценки, весьма удобные при проведении экономических исследований.

Обозначим теоретическое распределение при отсутствии засорения через $P(x)$, засоряющее распределение через $H(x)$, а их плотности распределения соответственно через $p(x)$ и $h(x)$. Функцию $f(x)$ следует выбрать таким образом, чтобы получить оценку с наименьшим средним квадратом смещения при наихудшем засорении. Для этого запишем выражение

$$\sup_H M\left[\sqrt{N}(\hat{\mu}-\mu)\right]^2 = \min, \qquad (9.12)$$

где N - число наблюдений или реализаций случайной величины;

M - символ математического ожидания;

$\hat{\mu}$ - оценка математического ожидания; $(\hat{\mu}-\mu)$ - смещение оценки;

$f(x)$ - функция, подобранная для решения (9.12).

Смещение оценки приближенно определяется по формуле:

$$\hat{\mu}-\mu = \frac{\frac{1}{N}\sum_{i=1}^{N} f(x_i-\mu)}{\frac{1}{N}\sum_{i=1}^{N} f'(x_i-\mu)}, \qquad (9.13)$$

где $f'(x_i-\mu)$ - производная по μ функции $f(x_i-\mu)$.

Асимптотическая дисперсия оценки $\hat{\mu}$, полученная из (9.11), обозначается

$$M(f,R) = \frac{M_R f^2(x-\mu)}{\left[M_R f'(x-\mu)\right]^2}. \qquad (9.14)$$

При этом распределение "грубых ошибок" в модели имеет вид:

$$R = (1-\varepsilon)p + \varepsilon H. \qquad (9.15)$$

Распределение (9.15) является произвольным и имеет математическое ожидание функции f, равное нулю:

$$M_R f = 0.$$

Хубер доказал, что асимптотическая дисперсия оценки $\hat{\mu}$ (9.14) имеет седловую точку (f_0, R_0), где f_0 определяет наилучшую процедуру оценивания, а R_0 - наименее благоприятное распределение, т.е. распределение, полученное по наихудшему засорению. Остановимся на теореме Хубера.

Теорема Хубера. Если распределение $P(x)$ имеет дважды непрерывно дифференцируемую плотность $p(x)$, такую, что $\log p(x)$ является выпуклой функцией, а $\hat{\mu}$ является оценкой, получаемой из

$$\sum_{i=1}^{N} f(x_i-\hat{\mu}) = 0, \qquad (9.16)$$

то

1. Асимптотическая дисперсия $W(f,R)$ оценки $\hat{\mu}$, полученной из (9.16) с использованием некоторой функции f, имеет седловую точку, т.е. существуют такие

$$f_0 R_0 = (1-\varepsilon)P + \varepsilon H_0, \qquad (9.17)$$

что

$$\sup_R W(f_0,R) = W(f_0,R_0) = \inf_f W(f_1,R_0), \qquad (9.18)$$

где R - произвольное распределение в модели "грубых ошибок" для которого математическое ожидание функции f равно нулю $(M_R f=0)$.

2. Пусть $x_0 < x_1$ - концы интервала (один конец интервала или оба могут обратиться в бесконечность), где $\left|\frac{p'(x)}{p(x)}\right| \le k$; k и ε связаны уравнением

$$\frac{1}{1-\varepsilon} = \int_{x_0}^{x_1} p(x)dx + \frac{p(x_0)+p(x_1)}{k}, \qquad (9.19)$$

тогда плотность $r_0(x)$, отвечающая "наихудшему" распределению $R_0(x)$, имеет вид

$$r_0(x) = \begin{cases} (1-\varepsilon)P(x_0)e^{k(x-x_0)} & x \le x_0 \\ (1-\varepsilon)p(x) & x_0 < x < x_1 \\ (1-\varepsilon)p(x_1)e^{-k(x-x_1)} & x \ge x_1 \end{cases}. \qquad (9.20)$$

3. Функция $f_0(x) = -\frac{r'_0(x)}{r_0(x)}$ монотонна и ограничена. Рассматриваемый метод оценивания является методом максимального правдоподобия для распределения $R_0(x)$

$$f_0(x) = \begin{cases} -k & \text{для } x \le x_0 \\ -\frac{p'(x)}{p(x)} & \text{для } x_0 < x < x_1 \\ k & \text{для } x \ge x_1 \end{cases}. \qquad (9.21)$$

4. Наихудшее засорение $H_0(x)$ имеет плотность

$$h_0(x) = \begin{cases} \frac{1-\varepsilon}{\varepsilon}\left[P(x_0)e^{k(x-x_0)} - p(x)\right] & \text{для } x \le x_0 \\ 0 & \text{для } x_0 < x < x_1 \\ \frac{1-\varepsilon}{\varepsilon}\left[P(x_1)e^{-k(x-x_1)} - p(x)\right] & \text{для } x \ge x_1 \end{cases}.$$

5. Асимптотическая дисперсия оценки

$$W(f_0,R_0) = \frac{(1-\varepsilon)\int_{-\infty}^{\infty} f_0^2(x)p(x)dx + \varepsilon k^2}{\left[(1-\varepsilon)\int_{-\infty}^{\infty} f_0'(x)p(x)dx\right]^2}. \qquad (9.22)$$

Доказательство теоремы Хубера не приводится.

Пример

$P(x)$ - нормальное распределение с плотностью

$$p(x:\mu;1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}}.$$

Имеется выборка в модели "грубых ошибок" $x_1, x_2,...,x_N$.

Каждый элемент этой выборки с вероятностью $(1-\varepsilon)$ подчиняется нормальному распределению со средней μ и единичной дисперсией и с вероятностью ε является "грубой ошибкой" с произвольным законом распределения, но с той же средней μ. Так как рассматриваемая модель относится к симметричным распределениям относительно μ, то надо полагать, что "грубые ошибки" с равной вероятностью появляются как в области малых, так и в области больших значений.

Воспользуемся теоремой Хубера и определим функцию $f_0(x)$:

$$f_0(x)=\begin{cases}-k & \text{для } x\le -k\\ x & \text{для } -k<x<k\\ k & \text{для } x\ge k\end{cases}. \qquad (9.23)$$

Следовательно, наихудшее засорение сосредоточено вне интервала $(-k, k)$. Длина этого интервала зависит от степени засорения $\varepsilon k=k(\varepsilon)$.

Из теоремы Хубера имеем

$$\frac{1}{1-\varepsilon}=\int_{x_0}^{x}p(x)dx+\frac{p(x_0)+p(x_1)}{k},$$

т.е.

$$\frac{1}{1-\varepsilon}=\frac{1}{\sqrt{2\pi}}\int_{-k}^{k}e^{-\frac{x^2}{2}}dx+\frac{2}{k\sqrt{2\pi}}e^{-\frac{k^2}{2}}.$$

Перепишем выражение, умножив правую и левую части на εk. Тогда для дальнейших рассуждений данное выражение более удобно представить в виде

$$\frac{\varepsilon k}{1-\varepsilon}=\frac{2\varepsilon}{\sqrt{2\pi}}\left(e^{-\frac{k^2}{2}}-k\int_{k}^{\infty}e^{-\frac{x^2}{2}}dx\right). \qquad (9.24)$$

Пусть $\varepsilon\to 0$, когда $k\to\infty$, а $f_0(x)\to x$. Значит $f_0(x)\equiv x$ на всей числовой прямой в предельном случае, когда $\varepsilon=0$, $k=\infty$. Следовательно, в этом случае оценка для μ будет определяться по методу наименьших квадратов

$$\sum_{i=1}^{N}(x_i-\hat{\mu})^2=\min_{\hat{\mu}}.$$

Согласно теоремы Хубера (9.16), когда $f(x)\equiv x$, данное выражение сводится к решению $\sum_{i=1}^{N}(x_i-\hat{\mu})=0$.

Решение данного уравнения является средним арифметическим

$$\hat{\mu}=\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i.$$

Если же $\varepsilon\to 0$, $k\to\infty$, а $f_0(x)\to sign(x)$, то при $\varepsilon=1$, $k=0$ оценка для μ определится по методу наименьших модулей

$$\sum_{i=1}^{N}|x-\hat{\mu}|=\min_{\hat{\mu}}.$$

Согласно (9.16) теоремы Хубера, когда $f(x)=sign(x)$, решение сводится к $\sum_{i=1}^{N}sign(x-\hat{\mu})=0$. Решение данного уравнения является **выборочной медианой.**

Таким образом, выборочная средняя арифметическая и выборочная медиана являются предельными положениями оценки на основе (9.23).

Для общего случая, когда $0<\varepsilon<1$, оценка получается при решении задачи на определение минимума

$$\sum_{i=1}^{N}p(x_i-\hat{\mu})=\min_{\hat{\mu}}. \qquad (9.25)$$

В этом случае

$$p(x)=\begin{cases}\frac{1}{2}x^2 & \text{для } |x|<k\\ k|x|-\frac{1}{2}k^2 & \text{для } |x|\ge k\end{cases}. \qquad (9.26)$$

Согласно выражениям (9.25) и (9.26) в зависимости от величины наблюдаемых значений процедура Хубера разделяет наблюдения на основные и на засоряющие.

Если величины наблюдаемых значений меньше k по абсолютной величине ($|x|<k$), то наблюдения относятся к **основным**, и их значения учитываются в квадрате.

Если же значения наблюдаемых величин равны или превосходят по модулю k, то наблюдения относят к **засоряющим**. В этом случае учитывается модуль их значений.

Таким образом, оценка Хубера включает в себя свойства выборочной средней арифметической и выборочной медианы. Она обладает высокой эффективностью и достаточной устойчивостью (робастностью). Это означает, что оценка Хубера мало меняется при наличии резко выделяющихся наблюдений. Можно заметить, что оценка Хубера получается за счет усовершенствования метода наименьших квадратов. Для получения оценки Хубера, зависящей от k и ε, существует таблица значений функции $k=k(\varepsilon)$ (табл.9.2).

Таблица 9.2

Значения ε и k для оценок Хубера

ε	k	ε	k	ε	k	ε	k
0	∞	0,01	1,945	0,15	0,980	0,4	0,550
0,001	2,630	0,02	1,717	0,20	0,862	0,5	0,436
0,002	2,435	0,05	1,399	0,25	0,766	0,65	0,291
0,005	2,160	0,10	1,140	0,3	0,685	0,80	0,162

В методе Хубера, как мы уже отмечали, критерием качества оценки $\hat{\mu}$ является

$$\sup_R W(f, R). \qquad (9.27)$$

Соотношение (9.27) определяет границу гарантированной дисперсии оценки при наихудшем засорении. Хубером определено, что оценка $\hat{\mu}$ недостаточно чувствительна к изменениям критерия k в интервале от 1 до 2. По этой причине рекомендуется в случае неизвестного значения доли засорения $\varepsilon \geq 0{,}2$ в качестве k брать величину в интервале [1; 2].

Для иллюстрации потери эффективности средней арифметической оценки по сравнению с оценкой Хубера полезно рассмотреть табл.9.3.

Таблица 9.3

Эффективность оценок Хубера ($\hat{\mu}$) и $\bar{x}$ при данном ε

Засоренность, ε	0	0,01	0,05	0,10	0,20	0,50
$\bar{x}$	1,00	1,10	1,55	2,23	4,08	20,10
$\hat{\mu}$	1,00	1,06	1,26	1,49	2,05	5,93

9.4. Оценки, построенные при помощи порядковых статистик

Оценки параметра сдвига симметричных распределений появились раньше оценок Хубера. Их формализация проведена в 50-е гг. XX века американским ученым Д.Тьюки и его последователями. Оценки, предложенные Пуанкаре и Винзором, используют основные идеи, известные нам по оценкам Хубера.

Пусть $x_1, x_2,\ldots,x_N$ - выборка. Упорядочим ее, построив вариационный ряд $x_{(1)}, x_{(2)},\ldots,x_{(N)}$, где $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}$.

Пуанкаре рассматривал выборку, в которой основные элементы подчинены нормальному распределению, а некоторая доля α является грубой ошибкой. Он предложил вместо средней арифметической выборки определять **α-урезанную среднюю.**

Для выборочной совокупности $x_1, x_2,\ldots,x_N$ α-урезанная средняя $(0<\alpha<\frac{1}{2})$

определяется формулой

$$T(\alpha) = \frac{1}{N - 2[\alpha N]} \sum_{i=[\alpha N]+1}^{N-[\alpha N]} x_{(i)}, \qquad (9.28)$$

где $[\alpha N]$ - целая часть от αN, т.е. наибольшее целое число, не превосходящее αN.

Из формулы (9.28) видно, что для получения из вариационного ряда удаляются 100α % минимальных элементов и 100α % максимальных элементов выборки. По оставшимся элементам определяется среднее арифметическое значение.

Новое распределение $P(z)$ симметрично, а полученная оценка асимптотически нормальна. Оценка имеет асимптотическую дисперсию (9.31).

В экономических задачах более интересной является **α-винзоризованная оценка,** предложенная Винзором.

По совокупности $x_1, x_2,\ldots,x_N$ среднее значение для уровня α, где $0<\alpha<\frac{1}{2}$, определяется формулой:

$$W(\alpha) = \frac{1}{N}\left[\sum_{i=[\alpha N]+2}^{N-[\alpha N]-1} x_i + [\alpha N](x_{([\alpha N]+1)} + x_{(N-[\alpha N])})\right]. \qquad (9.29)$$

Данная процедура отличается от средней по Пуанкаре, так как $[\alpha N]$ значения не исключаются ни с левого, ни с правого конца вариационного ряда $x_{(1)}, x_{(2)},\ldots,x_{(N)}$. Эти значения проецируются в ближайшую точку оставшейся части, упорядоченной в вариационный ряд выборки. Таким образом, при определении средней арифметической участвуют все N наблюдений.

При наложении некоторых условий на распределение $P(z)$ полученная оценка асимптотически нормальна и ее асимптотическая дисперсия $s_W^2(\alpha)$ имеет следующий вид:

$$s_W^2(\alpha) = \int_{z_\alpha}^{z_{1-\alpha}} z^2 p(z)dz + 2\alpha\left(z_{1-\alpha} + \frac{\alpha}{p(z_\alpha)}\right)^2. \qquad (9.29a)$$

Как $T(\alpha)$, так и $W(\alpha)$ дают хорошее приближение, если засоряющее распределение тоже симметрично.

Известно, что асимптотическая дисперсия оценки Хубера совпадает с асимптотической дисперсией оценки $T(\alpha)$ при уровне урезания $\alpha=P(x_0)$, зависящим от доли грубых ошибок в выборке $\alpha=\alpha(\varepsilon)$. Если распределение $P(z)$ нормально, то:

$$\alpha_{опт} = P(-k); \qquad (9.30)$$

$$z_\alpha = P^{-1}(\alpha), \text{ а } z_{1-\alpha} = P^{-1}(1-\alpha).$$

При известных значениях ε для определения $\alpha_{опт} = \alpha(\varepsilon)$ для оценки $T(\alpha)$ можно воспользоваться табл.9.4.

Приведем следующие рекомендации.

Если ε неизвестно, то для ε<0,2, значение α выбирается в интервале 0,05÷0,15.

Таблица 9.4

Значения $\alpha_{опт}$ при известном уровне засорения

ε	α	ε	α	ε	α
0	0	0,05	0,081	0,40	0,291
0,001	0,004	0,10	0,127	0,50	0,332
0,002	0,008	0,15	0,164	0,65	0,386
0,005	0,015	0,20	0,194	0,80	0,436
0,01	0,026	0,25	0,222	1,0	0,500
0,02	0,043	0,30	0,247		

Величину ε можно также определить по выборке. Более того, можно найти оценки нижней и верхней доверительных границ. Имеются работы, где значения ε_j определялись по каждому исходному показателю y_j отдельно.

Пример 9.1

На основе данных о деятельности предприятий требуется с учетом и без учета робастности провести факторный анализ, получить уравнения регрессии по главным факторам, выделить признаки, определяющие название факторов.

Анализ проводится по следующим 13 технико-экономическим показателям:

x_1 - фактический уровень использования мощностей, %;

x_2 - численность промышленно-производственного персонала (ППП), чел.;

x_3 - удельный вес продукции машиностроения в товарной продукции, %;

x_4 - коэффициент сменности оборудования, смены;

x_5 -вооруженность рабочих оборудованием с учетом коэффициента сменности, млн руб./чел.;

x_6 - среднегодовая рабочая площадь на одного рабочего, м²/чел.;

x_7 - среднегодовая стоимость основных производственных фондов (ОПФ), млн руб.;

x_8 - удельный вес покупных изделий, полуфабрикатов и услуг кооперируемых предприятий с учетом их сырья и материалов, %;

x_9 - удельный вес оборудования в среднегодовых ОПФ, %;

x_{10} - удельный вес оборудования возрастом свыше 10 лет по отношению ко всему оборудованию, %;

x_{11} - оборачиваемость нормируемых оборотных средств, дни;

x_{12} - процент брака в себестоимости;

x_{13} - выпуск валовой продукции, млн руб.

Решение

Предварительно по алгоритму факторного анализа главных факторов была рассчитана матрица факторных нагрузок A с учетом (a^y_{jr}) и без учета робастности (a_{jr}). Результаты расчетов представлены в табл. 9.5.

Таблица 9.5

Весовые коэффициенты с учетом робастности f_r^y

и без учета f_r (a^y_{jr} и a_{jr})

Факторы / Переменные	f_1^y	f_1	f_2^y	f_2	f_3^y	f_3	f_4^y	f_4
x_1	0,29	0,18	0,11	0,06	0,27	0,05	0,44	0,33
x_2	0,94	0,94	-0,22	0,05	-0,02	-0,20	-0,03	-0,02
x_3	-0,11	0,01	0,01	0,02	-0,01	-0,01	0,92	0,85
x_4	0,49	0,37	-0,29	-0,12	0,66	-0,21	0,14	-0,11
x_5	0,03	-0,02	0,90	-0,17	0,26	0,74	0,05	0,40
x_6	-0,16	-0,12	0,63	-0,56	-0,31	0,66	-0,05	0,02
x_7	0,94	0,93	0,19	-0,11	-0,06	0,19	-0,08	-0,01
x_8	0,17	0,32	-0,13	-0,39	-0,67	-0,19	-0,03	0,13
x_9	0,03	-0,03	-0,09	0,85	-0,01	-0,15	0,26	0,38
x_{10}	0,14	0,12	-0,10	-0,16	-0,63	-0,08	0,04	-0,04
x_{11}	-0,09	-0,09	0,09	-0,74	-0,27	0,02	0,12	0,26
x_{12}	0,06	0,03	0,06	0,07	0,26	0,70	0,11	-0,44
x_{13}	0,93	0,95	-0,01	0,12	-0,13	-0,04	0,01	0,03

Выделим признаки, определяющие название факторов при использовании устойчивого (метод Винзора) и классического методов оценивания, если $a_{кр1} = 0,20$ и $a_{кр2} = 0,50$. С этой целью:

1. Составим табл.9.6, в которую будут включены признаки x_j, при условии, что $a_{jr} > a_{кр2} = 0,50$.

Таблица 9.6

Признаки, определяющие названия факторов

Факторы / Вид оценки	f_1	f_2	f_3	f_4
Устойчивая	x_2, x_7, x_{13}	x_5, x_6	x_4, x_8, x_{10}	x_3
Классическая	x_2, x_7, x_{13}	x_6, x_9, x_{11}	x_5, x_6, x_{12}	x_3

Как видно из табл.9.6. названия первого и четвертого факторов не зависят от методов оценивания, так как признаки их определяющие совпадают. Для второго фактора совпадает только один признак. Остальной набор признаков различен. Для третьего фактора наборы признаков вооб-

ще не совпадают. Это свидетельствует о целесообразности использования робастных оценок при проведении многомерного статистического анализа. В этом нас убеждают и уравнения регрессии.

Получим следующие уравнения регрессии для рентабельности по ортогональным факторам с использованием классического ($\hat{y}$) и робастного ($\hat{y}^y$) методов оценивания:

$$\hat{y} = 44{,}75 + 5{,}90 f_1 + 10{,}43 f_2 + 3{,}92 f_5;$$

$$\hat{y}^y = 44{,}78 + 6{,}72 f_1^y - 9{,}81 f_2^y + 0{,}33 f_4^y + 3{,}16 f_5^y + 1{,}14 f_6^y.$$

Так, при классическом методе оценивания мы считаем существенно влияющими на рентабельность факторы f_1, f_2 и f_5, а при робастном оценивании убеждаемся, что кроме этих факторов влияют также f_4 и f_6 (значения a_{jr} для f_5 и f_6 не приведены в табл.9.6).

2. Выделим признаки, влияние которых незначимо, при условии, что $a_{кр1} \le 0{,}20$ (табл.9.7).

Таблица 9.7

Признаки, влияние которых незначимо

Факторы / Вид оценки	f_1	f_2	f_3	f_4
Устойчивая	$x_3, x_5, x_6, x_8, x_9, x_{10}, x_{11}, x_{12}$	$x_1, x_3, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}$	$x_2, x_3, x_7, x_9, x_{13}$	$x_2, x_4, x_5, x_6, x_7, x_8, x_{10}, x_{11}, x_{12}, x_{13}$
Классическая	$x_1, x_3, x_5, x_6, x_9, x_{11}, x_{12}$	$x_1, x_2, x_3, x_4, x_5, x_7, x_{10}, x_{11}, x_{12}, x_{13}$	$x_1, x_2, x_3, x_7, x_8, x_9, x_{10}, x_{11}, x_{13}$	$x_2, x_4, x_6, x_7, x_8, x_{10}, x_{13}$

Анализ табл.9.7 показывает, что при робастном и классическом оценивании наблюдается переход весовых коэффициентов из подмножества значимых в подмножество незначимых. Например, при устойчивом оценивании второго фактора признак x_5 - вооруженность рабочих оборудованием с учетом коэффициента сменности - имеет $a_{52}^y = 0{,}90$, а при классической оценке $a_{52} = -0{,}17$. В первом случае x_5 - признак, существенно влияющий на фактор и участвующий в формировании его названия, а во втором - x_5 приходится относить к незначимым признакам. При устойчивом оценивании третьего фактора - $a_{103}^y = -0{,}63$, при классической оценке $a_{103} = -0{,}08$, или $a_{83}^y = -0{,}67$, при классической оценке $a_{83} = -0{,}19$ и т.д. В данном примере убеждаемся в том, что переход из подмножества значимых в подмножество незначимых и наоборот наблюдается у многих признаков. В первом факторе - у двух признаков (x_1, x_8), во втором факторе - у пяти признаков (x_2, x_4, x_5, x_8, x_9) и т.д. Если в первом случае при $a_{jr} > 0{,}50$ изменение названия не

влияет на состав и структуру фактора, то во втором случае при $a_{jr} \le 0{,}20$ переход признака из подмножества значимых в подмножество незначимых меняет вес и структуру фактора.

9.5. Применение устойчивого оценивания при наличии асимметрии распределения

В экономических исследованиях построение уравнений линейной регрессии, проведение факторного или компонентного анализа, как правило, связывают с нормальным распределением. Если исходные статистические данные не подчиняются нормальному закону распределения, то расчеты существенно усложняются. При проведении анализа данных технических измерений в теории точности часто сталкиваются с хорошей сходимостью к нормальному закону распределения. В экономике же сходимость к нормальному распределению является скорее исключением, чем правилом.

Однако гипотеза нормальности часто принимается как некоторое свойство экономических данных. Подобная гипотеза приводит к неадекватным уравнениям линейной регрессии и, как следствие, к неудовлетворительным прогнозам. Когда же исходные данные оказываются мультиколлинеарными, то получение уравнения линейной регрессии становится неэффективным.

В экономических исследованиях однородные группы объектов часто бывают относительно немногочисленны, а число используемых одновременно показателей является соизмеримым с числом объектов исследования. Кроме того, нарушение требований, предъявляемых к нормальности распределения, в ряде случаев кажущееся незначительным и трудно обнаруживаемым, может привести к существенным искажениям результатов обработки статистических данных.

Следовательно, в экономических задачах необходимо применять устойчивые правила, которые были бы малочувствительными к отступлениям от основных требований модели и к наличию грубых ошибок.

В первую очередь к таким устойчивым оценкам следует отнести минимаксные оценки Хубера или оценки Пуанкаре (α-урезанные), Винзора (α-винзоризованные). Эти оценки по своим конечным результатам близки. Однако при построении уравнений регрессии по исходным показателям наиболее хорошо разработанными следует считать оценки Хубера.

При построении уравнения регрессии по главным компонентам, проведении факторного или компонентного анализа более предпочтительной является оценка Винзора, для которой средняя арифметическая рассчитывается с учетом всех объектов исследования.

Существуют следующие ограничения методов робастного оценивания:

- распределение должно быть симметричным;
- уровень засорения ε должен быть известен.

При известном уровне засорения по таблице Хубера можно определить $W(\alpha)$. Однако на практике ε в большинстве экономических задач бывает неизвестным.

Существует и другой подход, который определяет $\alpha_{опт}$ при оценках, предложенных Пуанкаре. В научной литературе он известен как **оценка Джакела**, Л.Джакел также предлагает проводить оценку α по выборке. При этом следует считать, что искомой $\alpha_{опт}$ будет значение, при котором выборочная дисперсия минимальна

$$S_T^2(\alpha) = \frac{1}{(1-2\alpha)^2}\left(\int_{z_\alpha}^{z_{1-\alpha}} z^2 p(z)dz + 2z_\alpha^2\right), \qquad (9.31)$$

где z_α и $z_{1-\alpha}$ определяются как: $z_\alpha = P^{-1}(\alpha)$; $z_{1-\alpha} = P^{-1}(1-\alpha)$;
$P(z)$ - плотность распределения.

Если же $P(z)$ является не просто симметричным законом, а нормальным распределением, то для оценок Винзора и Пуанкаре $\alpha_{опт}=P(-k)$. Значение k определяется из соотношения (9.24). При ε, стремящимся к нулю, k стремится к бесконечности.

Хубером приведена таблица зависимостей между k и ε. Если по выборке получено ε, то k может быть найдено из упомянутой таблицы.

Однако все рассуждения справедливы до тех пор, пока исходные экономические показатели подчиняются симметричному распределению или, в частном случае нормальному закону распределения. При асимметричности распределения оценки Хубера, Пуанкаре и Винзора теряют свою эффективность, становятся несостоятельными и смещенными. Поэтому необходимы подходы, учитывающие асимметричность.

При несимметричном распределении наиболее удобным можно считать **джеккнайф-оценку Тьюки и Квенсулла.** Суть ее заключается в том, что выборочные данные разбиваются на группы. Эффект каждой группы оценивается по результату, полученному при исключении данной группы из рассмотрения. Такая оценка позволяет уменьшить смещение параметра положения при асимметричных распределениях. При построении уравнения регрессии данный метод не оправдывает надежды исследователей, так как он не доведен до получения оценок с минимальным средним смещением.

Более эффективной оказывается взвешенная **джеккнайф-оценка,** которая была предложена **Хинкли.**

Весовые коэффициенты выбираются как расстояния, отражающие недостаток симметрии, и определяют вклад каждого наблюдения в дисперсию показателя. При оценке параметров вклад наблюдений с большими весами уменьшается.

Джеккнайф-оценки и оценки взвешенного джеккнайфа могут быть применены как к методу наименьших квадратов, так и к устойчивому (робастному) оцениванию. Мы вынуждены обращаться к оценкам Хубера, Пуанкаре и Винзора, так как при засорении метод наименьших квадратов теряет свои оптимальные свойства. Созданы программы для взвешенного джеккнайф-оценивания, согласно которым в первую очередь выявляется наличие засорения. При его отсутствии используется метод наименьших квадратов. Если засорение обнаружено, то проводится устойчивое оценивание. При необходимости устойчивого оценивания включается тест на асимметрию. При отсутствии асимметрии применяются методы Хубера, Пуанкаре и Винзора. При наличии асимметрии используется взвешенная джеккнайф-оценка. Данный подход применяется как в одномерном, так и в многомерном статистическом анализе.

9.6. Методы обнаружения засорения (грубые ошибки)

9.6.1. Метод выявления грубых ошибок Смирнова-Граббса

Проверка максимального наблюдения

Пусть $x_1, x_2, \ldots, x_N$ результаты наблюдения. По полученным наблюдениям построим вариационный ряд.

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(N)}. \qquad (9.32)$$

Если известны среднее квадратическое отклонение σ и математическое ожидание μ, то можно воспользоваться критерием

$$t_{(N)} = \frac{x_{(N)} - \mu}{\sigma} \qquad (9.33)$$

и сравнить его с данными таблицы процентных точек критерия Смирнова-Граббса

$$t_{\max} = \frac{\max_i \left| x_{(i)} - \mu \right|}{\sigma}. \qquad (9.34)$$

Если же σ и μ неизвестны, то необходимо воспользоваться табл.9.8. Для этого надо определить

$$T_{(N)} = \frac{x_{(N)} - \bar{x}}{s}, \qquad (9.35)$$

где $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$; $s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$.

Затем сравнить значение $T_{(N)}$ с C_α из табл.9.8.

Таблица 9.8

Таблица Граббса. Критические значения C_α

№ наблюдения	α=0,10	α=0,05	α=0,025	№ наблюдения	α=0,10	α=0,05	α=0,025
1	-	-	-	14	2,297	2,461	2,602
2	-	-	-	15	2,326	2,493	2,638
3	1,406	1,412	1,414	16	2,354	2,523	2,670
4	1,645	1,689	1,710	17	2,380	2,551	2,701
5	1,791	1,869	1,917	18	2,404	2,577	2,728
6	1,894	1,996	2,067	19	2,426	2,600	2,754
7	1,974	2,093	2,182	20	2,447	2,623	2,778
8	2,041	2,172	2,273	21	2,467	2,644	2,801
9	2,097	2,237	2,349	22	2,486	2,664	2,823
10	2,146	2,294	2,414	23	2,504	2,683	2,843
11	2,190	2,343	2,470	24	2,520	2,701	2,862
12	2,229	2,387	2,519	25	2,537	2,717	2,880
13	2,264	2,426	2,562	26	2,553	2,734	2,897

Если $T_{(N)} < C_\alpha$, то верна гипотеза H_0 о том, что $x_{(N)}$ не является грубой ошибкой. При $T_{(N)} > C_\alpha$, $x_{(N)}$ значимо отклоняется от $\bar{x}$, следовательно, является грубой ошибкой. Встречаются таблицы, где используется вместо α понятие доверительной вероятности γ, где $\gamma=1-\alpha$ и, тогда критическое значение $C_\alpha=C_{1-\gamma}$; C_α соответствует величине $T_{(N)}$ при гипотезе H_0, отвечающей доверительной вероятности $P(T_{(N)}\le C_\alpha)=\gamma$.

Пример 9.2

На основе данных средней производительности труда (выработка на одного работающего) 10 предприятий подотрасли (табл.9.9). Проверить наличие грубых ошибок.

Таблица 9.9

Выработка рабочих

№ предприятия	1	2	3	4	5	6	7	8	9	10
Выработка	12	11	13	12	14	12	18	15	14	13

Решение

1. Составим вариационный ряд:

11, 12, 12, 12, 13, 13, 14, 14, 15, 18.

Наибольшее значение равно $x_{(10)}$=18.

2. Вычислим среднее значение: $\bar{x}$=13,4.

3. Определим оценку среднего квадратического отклонения:

s=1,91.

4. Наибольшее значение

$$T_{(10)} = \frac{18-13{,}4}{1{,}91} = 2{,}41.$$

5. По табл.9.8 для N=10 определим $C_\alpha = C_{0,05}$=2,294.

6. Проверим гипотезу H_0:

$C_{0,05}<T_{(10)}$, т.е. 2,294<2,41.

Следовательно, гипотеза об однородности ряда отвергается. Значение выработки на одного работающего на предприятии № 7 является нетипичным. Это значение можно считать грубой ошибкой при уровне значимости α=0,05.

Проверка минимального наблюдения

Проведем проверку гипотезы H_0 с альтернативной H_1 при помощи критерия

$$T_{(1)} = \frac{\bar{x} - x_{(1)}}{s}; \qquad x_1 = \min_i x_{(i)}. \tag{9.36}$$

Если H_0 верна, то $T_{(1)}$ имеет такое же распределение, как и $T_{(N)}$, поэтому для проверки гипотезы можно воспользоваться табл.9.8.

9.6.2. Критерий Граббса для обнаружения одного экстремального наблюдения

Пусть дан ряд наблюдений $x_1, x_2, \ldots, x_N$, по которому построен вариационный ряд $x_{(1)}, x_{(2)}, \ldots, x_{(N)}$ и получены оценки

$$\bar{x}_1 = \frac{1}{N-1}\sum_{i=1}^{N-1} x_{(i)}; \qquad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_{(i)}.$$

Для проверки максимального значения на наличие грубой ошибки используется критерий

$$G_N = \frac{\sum_{i=1}^{N-1}(x_{(i)} - \bar{x}_1)^2}{\sum_{i=1}^{N}(x_{(i)} - \bar{x})^2}. \tag{9.37}$$

Если надо проверить минимальное значение, то применяется критерий

$$G_1 = \frac{\sum_{i=2}^{N}(x_{(i)} - \underline{x}_1)^2}{\sum_{i=1}^{N}(x_{(i)} - \bar{x})^2}, \tag{9.38}$$

где $\underline{x}_1 = \frac{1}{N-1}\sum_{i=2}^{N} x_{(i)}$.

Критические значения C_α для критериев G_1 и $G_{(N)}$ можно определить по табл.9.10.

Наблюдения $x_{(1)}$ или $x_{(N)}$ относят к грубым ошибкам, если наблюдаемые значения статистик G_1 или $G_{(N)}$ окажутся больше критического G_α.

Таблица 9.10

Критические значения C_α при использовании критериев $G_{(1)}$ и $G_{(N)}$

№ наблюдения	α=0,10	α=0,05	α=0,025	№ наблюдения	α=0,10	α=0,05	α=0,025
1	-	-	-	14	0,5942	0,5340	0,4792
2	-	-	-	15	0,6134	0,5559	0,5030
3	0,0109	0,0027	0,0007	16	0,6306	0,5755	0,5246
4	0,0975	0,0494	0,0248	17	0,6461	0,5933	0,5442
5	0,1984	0,1270	0,0808	18	0,6601	0,6095	0,5621
6	0,2826	0,2032	0,1453	19	0,6730	0,6243	0,5785
7	0,3503	0,2696	0,2066	20	0,6848	0,6379	0,5937
8	0,4050	0,3261	0,2616	21	0,6958	0,6504	0,6076
9	0,4502	0,3742	0,3101	22	0,7058	0,6621	0,6206
10	0,4881	0,4154	0,3526	23	0,7151	0,6728	0,6327
11	0,5204	0,4511	0,3901	24	0,7238	0,6829	0,6439
12	0,5483	0,4822	0,4232	25	0,7319	0,6923	0,6544
13	0,5727	0,5097	0,4528				

Пример 9.3

Выработка на одного работающего в строительно-монтажных организациях описывается логарифмически нормальным распределением. Из чего следует, что логарифм величины выработки имеет нормальное распределение

$$P(\ln x; m, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\ln x - m)^2}{2\sigma^2}}, \text{ где } x>0,$$

с дисперсией σ^2 и средней m. В табл.9.11 представлены центрированные значения логарифма выработки.

Таблица 9.11

Центрированные значения логарифма выработки в строительно-монтажных организациях

№ организации	$\ln x_i - m$	№ организации	$\ln x_i - m$	№ организации	$\ln x_i - m$	№ организации	$\ln x_i - m$	№ организации	$\ln x_i - m$
1	–0,13	4	0,03	7	–0,60	10	–0,02	13	0,04
2	0,08	5	0,21	8	–0,06	11	–0,19	14	0,17
3	0,43	6	–0,09	9	–0,09	12	–0,10	15	0,27

Построить вариационный ряд и выявить не являются ли крайние правое или левое значения грубыми ошибками.

Решение

1. Строим вариационный ряд

(–0,60); (–0,19); (–0,13); (–0,10); (–0,09); (–0,06); (–0,02);
0,03; 0,04; 0,08; 0,09; 0,17; 0,21; 0,27; 0,43.

2. Проверяем при помощи критериев $G_{(1)}$ и $T_{(1)}$ не является ли $x_{(1)}$= =(–0,60), относящийся к организации №7, грубой ошибкой.

а) определяем $\bar{x} = \frac{1}{15}\sum_{i=1}^{15} x_{(i)} = 0{,}008$;

б) находим среднее квадратическое отклонение S=0,228;

в) вычисляем $\underline{x}_1 = \frac{1}{14}(x_{(2)} + x_{(3)} + \ldots + x_{(15)})$; $\underline{x}_1 = 0{,}051$;

г) определяем $G_{(1)} = \frac{\sum_{i=2}^{N}(x_{(i)} - \underline{x}_1)^2}{\sum_{i=1}^{N}(x_{(i)} - \bar{x})^2} = \frac{0{,}385}{0{,}780} \approx 0{,}494$;

д) находим $C_{0,05}$ по табл.9.10 при N=15: $C_{0,05}$=0,5559.

Величина $C_{0,05} > G_{(1)}$: 0,5559>0,494.

При исследовании минимального наблюдения вариационного ряда при критерии $G_{(1)}$ гипотеза об однородности отвергается, гипотеза H_0 неверна. Значение -0,60 является грубой ошибкой.

Проверим решение при помощи критерия $T_{(1)}$ и табл.9.8.

а) получим $T_{(1)} = \frac{\bar{x} - x_{(1)}}{S} = \frac{0{,}008 + 0{,}60}{0{,}228} = 2{,}710$;

б) по табл.9.8 при N=15 найдем $C_{0,05}$, т.е. $C_{0,05}$=2,493.

Величина $C_{0,05} < T_{(1)}$, т.е. 2,493<2,710.

Следовательно, гипотеза H_0 об однородности совокупности отвергается. На предприятии №7 значение отклонения от средней величины m логарифма величины выработки является грубой ошибкой.

Примечание: статистики G_N и T_N связаны соотношением\

$$G_{(N)} = 1 - \frac{1}{N-1}\left(\frac{x_{(N)} - \bar{x}}{S}\right)^2 = 1 - \frac{1}{N-1}T_N^2.$$

Итак, определено, что $x_{(1)}$ - грубая ошибка. Осталось выявить в оставшихся 14 наблюдениях не является ли 0,43 грубой ошибкой (организация №3).

3. Проверим максимальное наблюдение у оставшихся 14 наблюдений.

а) рассчитаем $\bar{x}_1 = \frac{1}{13}(x_{(2)} + x_{(3)} + \ldots + x_{(14)})$ (не меняя индексы в исходном вариационном ряде). Итак, $\bar{x}_1 = 0{,}022$.

б) $G_{(15)} = \frac{\sum_{i=2}^{14}(x_{(i)} - \bar{x}_1)^2}{\sum_{i=2}^{15}(x_{(i)} - \bar{x})^2} = \frac{0{,}532}{0{,}898} = 0{,}592$;

в) определим $C_{0,05}$ по табл.9.10 для N=14;

$C_{0,05}$=0,5340 $G_{(15)} > C_{0,05}$; $(G_N > C_\alpha)$.

Следовательно, нет основания считать $x_{(15)}$=0,43 грубой ошибкой. $G_{(15)}$ своим значением не противоречит гипотезе однородности совокупности.

Проверим данный вывод при помощи критерия $T_{(15)}$.

$$T_{(15)} = \frac{x_{(15)} - \bar{x}}{S} = \frac{0{,}43 - 0{,}051}{0{,}166} = 2{,}283;$$

Определим $C_{0,05}$ по табл.9.8, т.е. $C_{0,05}$=2,461 ($T_{(15)}<C_\alpha$).

Следовательно, совокупность однородна и значение $x_{(15)}$=0,43 не является грубой ошибкой.

9.6.3. Критерий исключения нескольких грубых ошибок

Подход Титьена-Мура - обобщение критерия Граббса на несколько экстремальных наблюдений.

Пусть по выборке объемом N построен вариационный ряд

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(N)}. \qquad (9.39)$$

Тогда решающее правило для отношения k наибольших наблюдений к грубым ошибкам основано на статистике

$$L_{(k)} = \frac{\sum_{i=1}^{N-k} (x_{(i)} - \bar{x}_k)^2}{\sum_{i=1}^{N} (x_{(i)} - \bar{x})^2}; \qquad (9.40)$$

где

$$\bar{x}_k = \frac{\sum_{i=1}^{N-k} x_{(i)}}{N-k}, \qquad (9.41)$$

где $\bar{x}_k$ - средняя арифметическая (N-k) наблюдений после отбрасывания k наибольших наблюдений;

$\bar{x}$ - выборочная средняя арифметическая всей выборки.

Решающее правило для отнесения k наименьших наблюдений к грубым ошибкам основано на статистике

$$\tilde{L}_{(k)} = \frac{\sum_{i=k+1}^{N} (x_{(i)} - \underline{x}_k)^2}{\sum_{i=1}^{N} (x_{(i)} - \bar{x})^2}, \qquad (9.42)$$

где $\underline{x}_k = \frac{\sum_{i=k+1}^{N} x_{(i)}}{N-k}$; (9.43)

(N-k) - число наибольших наблюдений в выборке;

$\underline{x}_k$ - средняя арифметическая (N-k) наблюдений, оставшихся после удаления из выборки k наименьших наблюдений;

$L_{(1)}$ - эквивалентна $G_{(N)}$, а $\tilde{L}_{(1)}$ - величине $G_{(1)}$.

Пример 9.4

Стоимость (млрд руб.) строительно-монтажных работ на объектах составила:

0,916; 0,944; 1,428; 1,452; 1,524; 1,604; 1,632.

Требуется проверить методом Граббса (табл.9.10) значение $x_{(1)}$=0,916, а по критерию $\tilde{L}_{(2)}$ - табл.9.12 - оба наименьших значения на содержание грубой ошибки.

Таблица 9.12

Критические значения C_α при α=0,05 для $L_{(k)}$ и $\tilde{L}_{(k)}$

k \ n	3	4	5	6	7	8	9	10	11
1	0,003	0,05	0,125	0,203	0,273	0,326	0,372	0,418	0,454
2	-	0,001	0,018	0,055	0,106	0,146	0,194	0,233	0,270
3	-	-	-	0,010	0,032	0,064	0,099	0,129	0,162

k \ n	12	13	14	15	16	17	18	19	20
1	0,489	0,517	0,540	0,556	0,575	0,594	0,608	0,624	0,639
2	0,305	0,337	0,363	0,387	0,410	0,427	0,447	0,462	0,484
3	0,196	0,224	0,250	0,276	0,300	0,322	0,337	0,354	0,377

Решение

Определим $G_{(1)} = \frac{0{,}333}{0{,}547} = 0{,}6088.$

При α=0,05 $C_{0,05}$=0,3261. Так как $G_{(1)}>C_{0,05}$, то значение $x_{(1)}$=0,916 нельзя признать грубой ошибкой.

При k=2 по формуле (9.42) найдем

$$\tilde{L}_{(2)} = \frac{0{,}079}{0{,}547} = 0{,}144.$$

Определим C_α при α=0,05 по табл.9.12. При N=8 и α=0,05 находим $C_{0,05}$=0,146. Так как $\tilde{L}_{(k)} < C_{0,05}$, то выборка неоднородна и наблюдения $x_{(1)}$=0,916 и $x_{(2)}$=0,944 могут быть исключены (как нетипичные) при определении средней стоимости строительно-монтажных работ.

9.6.4. Критерий обнаружения экстремальных наблюдений (наибольших и наименьших) одновременно

При помощи алгоритма метода по выборке вычисляется среднее арифметическое значение $\bar{x}$ и проводится расчет абсолютных отклонений от среднего:

$$r_1 = |x_1 - \bar{x}|, \quad r_2 = |x_2 - \bar{x}|, \quad \ldots, \quad r_N = |x_N - \bar{x}|.$$

Строится возрастающий ряд с переиндексацией: z_i - наблюдение r_i - i-е по величине. Самое близкое к $\bar{x}$ значение обозначим z_1, а через z_N - самое удаленное от $\bar{x}$ наблюдение.

Для проверки гипотезы о том, что k наибольших по модулю наблюдений являются грубыми ошибками, используется величина:

$$E_{(k)} = \frac{\sum_{i=1}^{N-k} (z_i - \bar{z}_{(k)})^2}{\sum_{i=1}^{N} (z_i - \bar{z})^2}, \qquad (9.44)$$

где $\bar{z}_{(k)} = \frac{\sum_{i=1}^{N-k} z_i}{N-k}$; ($\bar{z}_{(k)}$ - средняя арифметическая из N-k наблюдений, оставшихся после исключения из выборки k подозрительных элементов);

$\bar{z}$ - средняя арифметическая всей выборки.

Критические значения для $E_{(k)}$ даны в табл.9.13.

Таблица 9.13

Значения C_α для $E_{(k)}$ (α=0,05)

k \ n	3	4	5	6	7	8	9	10	11
1	0,001	0,125	0,081	0,146	0,208	0,265	0,314	0,356	0,386
2	-	0,001	0,010	0,034	0,065	0,099	0,137	0,172	0,204
3	-	-	-	0,004	0,016	0,034	0,057	0,083	0,107

k \ n	12	13	14	15	16	17	18	19	20
1	0,424	0,455	0,484	0,509	0,526	0,544	0,562	0,581	0,597
2	0,234	0,262	0,293	0,317	0,340	0,362	0,382	0,397	0,416
3	0,133	0,165	0,179	0,206	0,227	0,248	0,267	0,287	0,302

Значение $E_{(k)}$ сравнивается с критическим значением C_α. Если $E_{(k)} < C_\alpha$, то k рассматриваемых наблюдений являются грубыми ошибками.

Пример 9.5

Обратимся к примеру 9.3 исследования выборки 15 предприятий по выработке, которая описывалась логарифмически нормальным распределением. Наблюдение –0,60 по критерию Граббса исключили как грубую ошибку. Подозрительное максимальное наблюдение из 14 оставшихся наблюдений отнесли к основной совокупности при α=0,05. Определим теперь при k=2 (одно минимальное наблюдение –0,60 и одно максимальное наблюдение $x_{(15)}$=0,43). Относятся ли крайние наблюдения выборки к грубым ошибкам?

Решение

Определим $E_{(2)} = \frac{0{,}228}{0{,}781} = 0{,}292$.

По табл.9.13 для k=2 и N=15 найдем $C_{0,05}$=0,317.

Так как $E_2 < C_{0,05}$, то по данному критерию оба наблюдения –0,60 и 0,43 следует считать грубыми ошибками.

ПРИЛОЖЕНИЕ 1

ДИНАМИЧЕСКИЕ (ВРЕМЕННЫЕ) ФАКТОРНЫЕ МОДЕЛИ

Идея создания факторной модели связана с известным фактом, что эффективность применения методов факторного и компонентного анализа зависит от качества проведения экономической интерпретации полученных результатов.

Содержательная интерпретация является наиболее трудным этапом анализа. При содержательной интерпретации за ряд лет (промежутков времени) большое значение имеет изучение выявленных закономерностей в течение всего анализируемого периода.

Взаимосвязи между исходными признаками и факторами, выявленные в ходе решения задачи в статике, могут изменяться с течением времени. Поэтому существенно повысить адекватность факторной модели реальному процессу, а следовательно, и объективность проводимой интерпретации, можно путем многомерного статистического анализа системы исходных показателей за ряд лет. Для этого необходимо использовать временную факторную модель, которая дает возможность проводить анализ не только в статике, но и в динамике.

Временная факторная модель позволяет исследователю идентифицировать те исходные признаки, которые слабо проявили себя в прошлом, но существенно влияли в течение последних лет. С помощью этой модели становится возможным получение дополнительной информации об изучаемом экономическом процессе путем выявления признаков, ослабивших или усиливших свое влияние за последние годы, а также признаков, существенно влияющих на фактор в течение всего изучаемого периода времени. Идентификацией этих признаков можно гибко улавливать произошедшие изменения в исследуемом процессе.

Обозначим весовой коэффициент временной модели для t_i-го года через $a^{t_i}_{jr}$. Полученные значения r-го фактора за l лет представим в виде матрицы A, в которой каждый столбец характеризует фактор r за t_i-й год ($i=\overline{1,l}$). Каждая строка определяет вес j-го признака в r-м факторе за l лет.

Наряду с l вертикальными подмножествами W_β рассмотрим n горизонтальных подмножеств $K_{xj}(j=\overline{1,l})$ весовых коэффициентов $a^{t_i}_{jr}$, связывающих r-й фактор (главную компоненту) с каждым из n признаков в течение l лет.

Представим вертикальные нечеткие множества за l лет в виде:

$$W_1 = \{0 \le a_{jr}^{t_i} \le a_{kp_1} \mid \mu_{W_1}(a_{jr}^{t_i})\}; \quad (1.1)$$

$$W_2 = \{a_{kp_1} < a_{jr}^{t_i} \le 1 \mid \mu_{W_2}(a_{jr}^{t_i})\}; \quad (1.2)$$

$$W_3 = \{a_{kp_2} < a_{jr}^{t_i} \le 1 \mid \mu_{W_3}(a_{jr}^{t_i})\}; \quad (1.3)$$

$$W_4 = \{a_{kp_1} < a_{jr}^{t_i} \le a_{kp_2} \mid \mu_{W_4}(a_{jr}^{t_i})\}. \quad (1.4)$$

Горизонтальные подмножества для фиксированного значения $j=\overline{1,n}$ и $i=\overline{1,l}$ запишутся

$$K_{x_j} = \{a_{jr}^{t_i} \mid \mu_{K_{x_j}}(a_{jr}^{t_i}) = 1\}. \quad (1.5)$$

В каждом из рассматриваемых столбцов матрицы A эти весовые коэффициенты могли принадлежать подмножествам W_1, W_2, W_3, W_4.

При фиксированных j и r, но разных значениях i ($i=\overline{1,l}$), элемент $a_{jr}^{t_i}$ мог принадлежать разным подмножествам. Это значит, что в течение l лет в матрице A могли наблюдаться некоторые флуктуации, в результате которых весовой коэффициент, связывающий r-й фактор и j-й признак, мог переходить из одного подмножества $W_\beta(\beta=\overline{1,4})$ в другое. Оценка этих колебаний может дать полезную информацию об исходной переменной за заданный период времени.

Для исследования этих колебаний необходимо рассмотреть пересечение подмножеств K_{x_j} с каждым из подмножеств W_β. Обозначим через $W_\beta \cap K_{x_j}$ большее нечеткое подмножество, содержащееся одновременно в $W_\beta(\beta=\overline{1,4})$ и $K_{x_j}(j=\overline{1,n})$. В соответствии с теорией нечетких множеств можно записать:

$$W_\beta \cap K_{x_j} = \{a_{jr}^{t_i} \mid \mu_{W_\beta \cap K_{x_j}}(a_{jr}^{t_i}) = \min(\mu_{W_\beta}(a_{jr}^{t_i});\ \mu_{K_{x_j}}(a_{jr}^{t_i}))\}. \quad (1.6)$$

Для перехода от нечетких подмножеств весовых коэффициентов $a_{jr}^{t_i}$ к подмножествам признаков обозначим $x=\{x_j\}$; $(j=\overline{1,n})$ и приведем нечеткое подмножество исходных признаков

$$z_\beta = \{(x_j \mid \mu_{z_\beta}(x_j))\}. \quad (1.7)$$

По аналогии с подмножествами W_β для r-го фактора z_1 является нечетким подмножеством исходных признаков, имеющих незначимые коэффициенты веса. Для этого же фактора z_2 является нечетким подмножеством исходных признаков, имеющих значимые коэффициенты веса. Фактор

z_3 является нечетким подмножеством исходных признаков, имеющих значимые коэффициенты веса, участвующие в формировании названия фактора или главной компоненты, а z_4 - нечеткое подмножество исходных признаков, имеющих значимые коэффициенты веса с фактором, но не участвующих в формировании факторов.

Значение функции $F = F(t,\alpha) = t^\alpha$ для t=1,2,3,...,25, α=1,0;1,1;...;2,0 представлено в табл. 1.1.

Таблица 1.1

Значение функции $F=F(t,\alpha)=t^\alpha$ для t=1,2,3,...,25, α=1,0; 1,1; ...; 2,0.

t \ α	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
1	1	1	1	1	1	1	1	1	1	1
2	2,144	2,297	2,246	2,639	2,828	3,031	3,249	3,482	3,732	4, 000
3	3,348	3,737	4,171	4,656	5,196	5,800	6,473	7,225	8,064	9, 000
4	4,595	5,278	6,063	6,964	8,000	9,190	10,556	12,126	13,929	16, 000
5	5,873	6,899	8,103	9,518	11,180	13,133	15,426	18,119	21,283	25,000
6	7,177	8,568	10,271	12,286	14,697	17,581	21,031	25,158	30,095	36,000
7	8,504	10,330	12,550	15,245	18,520	22,499	27,332	33,203	40,335	49,000
8	9,849	12,126	14,929	18,379	22,627	27,858	34,297	42,224	51,984	64,000
9	11,212	13,967	17,399	21,674	27,000	33,635	41,900	52,196	65,022	81,000
10	12,589	15,849	19,935	25,119	31,623	39,811	50,119	63,096	79,433	100,000
11	13,981	17,769	22,584	28,704	36,483	46,369	58,934	74,904	95,202	121,000
12	15,385	19,725	25,289	32,423	41,569	53,295	68,329	87,604	112,317	144,000
13	16,801	21,714	28,062	36,268	46,872	60,577	78,290	101,181	130,765	169,000
14	18,228	23,733	30,901	40,233	52,383	68,203	88,801	115,619	150,537	196,000
15	19,665	25,782	33,800	44,313	58,095	76,163	99,852	130,907	171,622	225,000
16	21,112	27,858	36,758	48,503	64,000	84,448	111,430	147,033	194,012	265,000
17	22,568	29,960	39,773	52,799	70,093	93,050	123,527	163,986	217,697	289,000
18	24,033	32,087	42,840	57,198	76,368	101,961	136,133	181,757	242,671	324,000
19	25,505	34,238	45,960	61,196	82,819	111,175	149,239	200,335	268,925	361,000
20	26,968	36,411	49,129	66,289	89,443	120,684	162,836	219,712	296,454	400,000
21	28,474	38,607	52,346	70,975	96,234	130,482	176,918	239,880	325,249	441,000
22	29,968	40,823	55,610	75,752	103,189	140,565	191,478	260,832	355,306	484,000
23	31,470	43,060	58,917	80,616	110,304	150,926	206,508	282,560	386,619	529,000
24	32,979	45,316	62,269	85,565	117,576	161,562	222,003	305,056	419,181	576,000
25	34,493	47,.591	65,663	90,597	132,574	172,466	237,957	328,316	452,987	625,000

Для идентификации признаков большое значение имеет выбор функции принадлежности. Для фиксированного значения β величина $\mu_{z_\beta}(x_j)$ является функцией, отображающей x во множество принадлежностей M. Под M будем подразумевать интервал [0,1]. В этом интервале нуль означает низшую, а единица - высшую степень принадлежности. Функция принадлежности должна быть чувствительна к изменению исследуемых показателей (признаков) в динамике, учитывая роль коэффициентов веса по годам. Только в этом случае с помощью функции принадлежности можно будет выявить имеющиеся в системе исходных признаков флуктуации. Исследования показали, что достаточно чувствительной является функция принадлежности к нечеткому подмножеству W_β весового коэффициента $a_{jr}^{t_i}$, связывающего j-й признак и r-й фактор в году $t_m(1 \le m \le l)$ следующего вида:

$$\mu_{W_\beta}(a_{jr}^{t_i}) = \frac{t_m^\alpha}{\sum_{i=1}^{l} t_i^\alpha}, \qquad (1.8)$$

где $\beta = \overline{1,4}$; m является индексом года, в котором $a_{jr}^{t_i} \in W_\beta$; $i = 1,2,\ldots,m,\ldots,l$;

$\alpha = (1,0;1,1;\ldots;2,0)$. При необходимости значения α могут быть продолжены. (Для экономических рядов достаточно иметь эти значения, приведенные в табл.1.1).

Функция принадлежности признака x_j подмножеству $z_\beta(\beta = \overline{1,4})$ за ряд лет определяется следующим образом:

$$\mu_{z_\beta}(x_j) = \sum_{l} \mu_{W_\beta} \mathrm{I}\ K_{x_j}(a_{jr}^{t_i}). \qquad (1.9)$$

Сопоставив выражения (1.5), (1.6) и (1.8) можно (1.9) представить в виде

$$\mu_{z_\beta}(x_j) = \frac{\sum_{m=1}^{q} t_m^\alpha}{\sum_{i=1}^{l} t_i^\alpha}; \qquad (1.10)$$

где q - число лет, в которых $x_j \in z_\beta$, или $a_{jr}^{t_i} \in W_\beta$; $q \le l$.

Функцию принадлежности при помощи параметра α можно "настраивать" для работы с короткими или длинными экономическими временными рядами. При анализе структуры матрицы весовых коэффициентов r-го фактора за короткий интервал времени рекомендуется выбирать значения α, близкие к 2,0. Для длинных временных рядов предпочтительней задание значений α, близких к 1. Если в анализируемом периоде отдельные интервалы времени представляют для исследования не одинаковый интерес, то можно задавать свои значения параметра α, для каждого интервала

$$\mu_{z_\beta}(x_j) = \frac{\sum_{m=1}^{m_k} t_m^{\alpha 1} + \sum_{m=m_k+1}^{m_p} t_m^{\alpha 2} + \ldots + \sum_{m=m_q+1}^{q} t_m^{\alpha f}}{\sum_{i=1}^{i_1} t_i^{\alpha 1} + \sum_{i=i_1}^{i_2} t_i^{\alpha 2} + \ldots + \sum_{i=i_n+1}^{l} t_i^{\alpha f 1}}, \qquad (1.11)$$

где $m=1,2,\ldots,m_k,\ldots,m_p,\ldots,m_q,\ldots,q$;
$i=1,2,\ldots,i_1,\ldots,i_2,\ldots,i_n,\ldots,l$;
$\alpha=\{\alpha_1,\alpha_2,\ldots,\alpha_f\}$;
$j=1,2,\ldots,n$.

Выбор и расчет функции принадлежности в виде выражения (1.10) могут быть проведены при помощи данных табл.1.1.

Пример

Дана временная факторная модель функционирования швейных объединений службы быта. На основе содержательного экономического анали-

за с использованием методов кластерного анализа были отобраны следующие исходные признаки (показатели):

x_1 - объем реализации бытовых услуг населению;

x_2 - среднемесячная численность работников ППП;

x_3 - удельный вес рабочих в среднесписочной численности работников ППП;

x_4 - среднегодовая стоимость ОПФ;

x_5 - удельный вес машин и оборудования в среднегодовой стоимости ОПФ;

x_6 - фондовооруженность рабочих;

x_7 - удельный вес заработной платы производственных рабочих в себестоимости реализованных бытовых услуг;

x_8 - удельный вес накладных расходов в себестоимости реализованных бытовых услуг;

x_9 - удельный вес сырья в себестоимости продукции;

x_{10} - среднемесячный размер премии одного работника ППП;

x_{11} - среднемесячный размер заработной платы ППП;

x_{12} - среднемесячный размер заработной платы рабочего.

По перечисленным 12 показателям был проведен факторный анализ за 6 лет. В табл.1.2 приведены обобщенные факторы за один год, позволяющие оценить характер соответствующих весовых коэффициентов и характеристики дисперсий первых шести факторов, вклады которых больше 8%.

Решение

Проведем содержательную интерпретацию полученных результатов. Потребуем, чтобы коэффициент интерпретации был не менее 0,80 для всех оставленных обобщенных факторов, дающих суммарный вклад в общую дисперсию процесса 91%. Судя по имеющимся весовым коэффициентам (для обеспечения требования, чтобы коэффициент был не менее 80%), a_{kp2} должно быть не более 0,458.

Первый обобщенный фактор f_1 имеет два весовых коэффициента, превышающих 0,458. Значит, в формировании названия должны участвовать x_{11} и x_{12}. По этой причине первый фактор был назван "Характеристика оплаты по труду". Он объясняет 18% дисперсии.

Второй обобщенный фактор f_2 объясняет 23% дисперсии. Он связан тремя весовыми коэффициентами a_{12}, a_{22}, a_{42} превышающими 0,91, с признаками, характеризующими размер предприятия (x_1, x_2, x_4), поэтому он был назван "Размер предприятия".

Третий обобщенный фактор f_3 связан с показателями x_5 и x_6. Данные показатели являются характеристикой технической оснащенности предприятий. Однако нагрузки на эти показатели имеют разные знаки. Следовательно, обобщенный фактор отражает не техническую оснащенность, а соотношение между важнейшим элементом активной части ОФП - машинами и оборудованием - и остальными составляющими производственных

фондов. Поэтому третий фактор f_3 был назван "Характеристика структуры ОПФ". Данный обобщенный фактор объясняет 16% дисперсии процесса.

Таблица 1.2

Матрица нагрузок

Переменные	Объясняемая групповая дисперсия	Обобщенные факторы					
		f_1	f_2	f_3	f_4	f_5	f_6
x_1	0,965	–0,031	0,945	–0,197	0,177	–0,036	–0,015
x_2	0,989	–0,036	0,965	–0,076	0,105	0,007	0,010
x_3	0,657	–0,458	–0,229	–0,229	–0,532	0,202	–0,135
x_4	0,962	–0,031	0,912	0,290	0,095	0,135	–0,136
x_5	0,888	–0,083	–0,016	–0,893	0,051	–0,209	0,194
x_6	0,900	0,263	–0,183	0,878	–0,127	–0,024	–0,098
x_7	0,923	0,219	0,109	–0,046	0,824	0,253	–0,343
x_8	0,871	–0,014	–0,155	0,203	–0,895	–0,024	–0,064
x_9	0,940	–0,236	–0,101	–0,274	–0,077	–0,025	0,890
x_{10}	0,996	0,174	0,067	0,146	0,103	0,932	–0,027
x_{11}	0,965	0,942	–0,075	0,135	0,098	0,133	–0,165
x_{12}	0,983	0,949	–0,062	0,153	0,131	0,130	–0,141
Вклады факторов		2,211	2,800	1,930	1,877	1,073	1,046
Накопленные доли вкладов		0,184	0,417	0,578	0,734	0,823	0,910
Доли вкладов факторов		0,184	0,233	0,161	0,156	0,089	0,087
Коэффициент интерполяции, %		81	95	81	94	81	85

Четвертый обобщенный фактор f_4 связан с коэффициентами веса, превышающими по модулю $a_{кр2}$=0,458 с тремя показателями x_3, x_7 и x_8 (a_{34}=-0,532, a_{74}=0,824, a_{84}=-0,895). Известно, что для предприятий пошива и ремонта одежды заработная плата производственных рабочих и накладные расходы являются основными составными частями себестоимости бытовых услуг. Это связано со спецификой учета в себестоимости сырья и материалов для швейной подотрасли. Четвертый обобщенный фактор f_4, названный "Обобщенный показатель структуры себестоимости", объясняет 15,6% дисперсии процесса.

Названия пятого f_5 и шестого f_6 обобщенных факторов формировались по другому. Они объясняют около 9% дисперсии процесса. В каждом из них имеется только по одному весовому коэффициенту, значение которого превышает 0,458 (a_{105}=0,932; a_{96}=0,890).

Для формирования названия были использованы показатели x_{10} и x_9 соответственно. Так, f_5 был назван "Характеристика материального поощрения", а f_6 - "Удельный вес сырья в себестоимости".

Таблицы, аналогичные табл.1.1, были получены по данным за 6 лет. По каждому обобщенному фактору построены матрицы A, у которых столбцы

представляют r-й обобщенный фактор за i-й год ($i=\overline{1,l}$). Строки матрицы A представляют значения весовых коэффициентов связи признака a_j с фактором f_r за l лет ($j=\overline{1,n}$).

По табл.1.1 при α=1,7 были определены по формуле (1.10) функции принадлежности.

По результатам примера 1 построена табл.1.3 для первых четырех обобщенных факторов f_1, f_2, f_3 и f_4. На пересечении столбца фактора $\mu_{z_\beta}(x_j)$ со строкой признака x_j указана функция принадлежности к нечеткому подмножеству z_β данного признака x_j.

Определим во все ли годы показатели x_{11} и x_{12} определяли фактор f_1, как в табл.1.2. Согласно табл.1.3 функции принадлежности $\mu_{z_3}(x_{11})=1$ и $\mu_{z_3}(x_{12})=1$. Это подсказывает, что все шесть лет $a^{l_i}_{11,1}$ и $a^{l_i}_{12,1}$ входили в подмножество значимых коэффициентов, участвующих в формировании факторов. В результате флуктуации в отдельные годы в это подмножество входили показатели x_3 и x_7 (табл.1.3). Так, $\mu_{z_3}(x_3)=0{,}38$ (это значение получено за счет его весовых коэффициентов, превышающих $a_{кр2}$=0,458 в первом и шестом годах). Значение $\mu_{z_3}(x_7)=0{,}37$ связано с превышением $a^{l_i}_{jr}$ величины $a_{кр2}$=0,458 с первого по четвертый годы. Фактору f_2 - размер предприятия - к подмножеству значимых, участвующих в формировании названия обобщенного фактора f_2, принадлежали только показатели x_1, x_2 и x_4. При этом $\mu_{z_3}(x_j)=1$ у всех трех показателей. Для этого же фактора показатели x_7, x_9 и x_{10} имели функцию принадлежности к подмножеству незначимых коэффициентов веса ($\mu_{z_3}(x_j)$=0,82) и по 0,18 для подмножества значимых z_2 и z_4.

Анализ показывает, что только во втором и третьем годах все три показателя оказались в подмножестве значимых по своему влиянию показателей z_4, хотя все остальные четыре года они относились к подмножеству z_1. У показателя x_6 в этом же факторе f_2 функция принадлежности к подмножеству незначимых z_1 равна 0,38, а к подмножеству значимых, но не участвующих в формировании названия фактора, функция равна 0,62. В первый и шестой годы показатель x_6 относился к z_1, а во все остальные годы x_6 принадлежал к z_2 и его подмножеству z_4. Следовательно, в первый и шестой годы роль фондовооруженности снижалась в f_2 и нагрузка на f_2 оказывалась в подмножестве W_1.

На f_3 - характеристику структуры ОПФ в течение всего шестилетнего периода - существенно влияли показатели x_5 и x_6. Их $\mu_{z_3}(x_j)$=1. Однако $\mu_{z_3}(x_4)$=0,32. Это значение получено за счет того, что во втором и пятом

Таблица 1.3

Значения функции принадлежности признаков ($j=\overline{1,12}$) нечетким подмножествам $z_\beta(\beta=1,2,3,4)$ при $\alpha=1,7$; $l=6$; $a_{kpl}=0,21$

№ признака	Обобщенные факторы																	
	Характеристика оплаты по труду, f_1				Размер предприятия, f_2				Характеристика структуры ОПФ, f_3				Обобщенный показатель структуры себестоимости, f_4					
x_j	μ_{z1} (x_j)	μ_{z2} (x_j)	μ_{z3} (x_j)	μ_{z4} (x_j)	μ_{z1} (x_j)	μ_{z2} (x_j)	μ_{z3} (x_j)	μ_{z4} (x_j)	μ_{z1} (x_j)	μ_{z2} (x_j)	μ_{z3} (x_j)	μ_{z4} (x_j)	μ_{z1} (x_j)	μ_{z2} (x_j)	μ_{z3} (x_j)	μ_{z4} (x_j)		
1	1	0	0	0	0	1	1	0	0,74	0,26	0	0,26	1	0	0	0		
2	1	0	0	0	0	1	1	0	0,56	0,44	0	0,44	1	0	0	0		
3	0,56	0,44	0,38	0,06	0,64	0,36	0	0,36	0,63	0,37	0	0,37	0,56	0,44	0,44	0		
4	1	0	0	0	0	1	1	1	0	1	0,32	0,68	1	0	0	0		
5	1	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0		
6	0,62	0,38	0	0,38	0,38	0,62	0	0,62	0	1	1	0	0,89	0,11	0	0,11		
7	0	1	0,37	0,63	0,82	0,18	0	0,18	0,74	0,26	0	0,26	0	1	1	0		
8	0,81	0,19	0	0,19	1	0	0	0	0,68	0,32	0	0,32	0	1	1	0		
9	0,19	0,81	0	0,81	0,82	0,18	0	0,18	0,45	0,55	0	0,55	0,82	0,18	0	0,18		
10	0,74	0,26	0	0,26	0,82	0,18	0	0,18	0,74	0,26	0	0,26	1	0	0	0		
11	0	1	1	0	1	0	0	0	1	0	0	0	1	0	0	0		
12	0	1	1	0	1	0	0	0	0,98	0,02	0	0,02	1	0	0	0		

годах $a^{l_3}_{43} \in W_3$, а $x_4 \in z_3$. Следует отметить, что все шесть лет для f_3 показатель x_4 постоянно принадлежал только подмножеству значимых показателей z_2.

Этот вывод позволяет определить, что среднегодовая стоимость ОПФ всегда в большей или меньшей степени влияла на фактор, учитывающий структуру ОПФ. Следовательно, его роль в динамической модели должна постоянно учитываться. Поэтому исходя из экономической политики подотрасли, были приняты меры не просто по увеличению ОПФ, а по увеличению их активной части - машин и оборудования, что позволило повысить производительность труда в швейной подотрасли и объем реализации услуг.

Таблица 1.4

Нечеткие подмножества исходных признаков z_β (β=2,3)

Обобщенные факторы	Нечеткие подмножества исходных признаков
Характеристика оплаты по труду, f_1	$z_2 = \{(x_3/0,44),(x_6/0,38),(x_7/1),(x_8/0,19),(x_9/0,81),(x_{10}/0,26),(x_{11}/1),(x_{12}/1)\}$ $z_3 = \{(x_3/0,38),(x_7/0,37),(x_{11}/1),(x_{12}/1)\}$
Размер предприятия, f_2	$z_2 = \{(x_1/0,26),(x_2/0,44),(x_3/0,37),(x_4/1),(x_5/1),(x_6/1),(x_7/0,26),(x_8/0,32)\}$ $z_3 = \{(x_4/0,32),(x_5/1),(x_{61}/1)\}$
Характеристика структуры ОПФ, f_3	$z_2 = \{(x_1/0,26),(x_2/0,44),(x_3/0,37),(x_4/1),(x_5/1),(x_6/1),(x_7/0,26),(x_{12}/0,02)\}$ $z_3 = \{(x_4/0,32),(x_5/1),(x_{61}/1)\}$
Обобщенный показатель структуры себестоимости, f_4	$z_2 = \{(x_3/0,44),(x_6/0,11),(x_7/1),(x_8/1),(x_9/0,18)\}$ $z_3 = \{(x_3/0,44),(x_7/1),(x_{81}/1)\}$

Поскольку в первую очередь нас интересуют нечеткие подмножества значимых признаков z_2 и z_3, представим эти признаки в отдельной таблице (табл.1.4) элементами $(x_i|\mu_{z_\beta}(x_j))$. Табл.1.4 является удобной при изучении значимых элементов нечетких множеств. Например, в течение рассматриваемого периода в третьем факторе в те или иные годы все показатели, кроме $(x_{11}|0)$, оказывались в подмножестве значимых признаков z_2. Однако нижестоящая строка подмножества z_3 показывает, что среди этих показателей , наиболее существенно и постоянно влияющими являются только x_5 и x_6. Показатель x_4 оказывал влияние только в отдельные годы $a^{l_3}_{j_3} > 0,458$ (втором и пятом годах).

Рассмотрим f_4. В течение всего периода времени в z_3 с высокими весовыми коэффициентами постоянно действуют показатели x_7 и x_8. Показатель x_3 имел весовые коэффициенты во множестве W_3 в течение первого, второго и шестого годов. В шестом году (последнем из исследованных)

весовой коэффициент $a_{34}^{t_6} = |-0{,}532| > 0{,}458$, хотя в третьем, четвертом и пятом годах $a_{34}^{t_i} \notin W_2$. Таким образом, для окончательного решения вопроса о роли показателя x_3 в швейной подотрасли необходимо будет провести дополнительные исследования методами содержательного экономического анализа.

Временная факторная модель позволяет более объективно оценивать факторы и признаки за определенный период времени и иметь дополнительную информацию, которую не представляется возможным получить из факторной модели за один период времени. Построим сводную таблицу факторов и признаков, существенно влияющих на них, с учетом, когда они оказывались существенно влияющими (в z_3) в течение всего периода времени $\mu_{z_3}(x_j)=1$ и эпизодически $\mu_{z_3}(x_j)<1$ (табл.1.5).

В табл.1.5 две строки: в верхней строке указаны признаки постоянно действующие, а в нижней - признаки, действующие периодически в течение всего шестилетнего периода. Для этих элементов подмножества z_3 под чертой записаны цифрами годы, когда признаки x_j принадлежали z_3.

Вторая строка табл.1.5 обращает внимание на показатель x_3, который имел $b_{f_1} a_{31}^{t_1}$ и $a_{31}^{t_6} \geq |0{,}458|$.

Таблица 1.5

Анализ признаков в нечетком подмножестве z_3 $a_{кр1}$=0,458

Функция принадлежности \ Факторы	Признаки			
	f_1	f_2	f_3	f_4
$\mu_{z3}(x_j)=1$	x_{11}, x_{12}	x_1, x_2, x_4	x_5, x_6	x_7, x_8
$\mu_{z3}(x_j)<1$	$\frac{(x_3\,\|0{,}38)}{1{,}6}$ $\frac{(x_7\,\|0{,}37)}{1,2,3,4}$		$\frac{(x_4\,\|0{,}32)}{2{,}5}$	$\frac{(x_3\,\|0{,}44)}{1,2,6}$

Из табл.1.5 видно, что в течение шести рассматриваемых лет в первых четырех наиболее весомых обобщенных факторах только 9 из 12 признаков постоянно играли существенную роль $(\mu_{z_3}(x_j)=1)$.

Далее рассмотрим показатель x_4 в f_3, роль которого была существенна во втором году и возросла в пятом году. Зато признак x_7 в первом факторе существенно влиял во все предшествующие годы, но в последние годы его значение ослабло. Остановимся на табл.1.6, учитывающей подмножество z_1 незначимых признаков в соответствующем факторе.

Табл.1.6 дополняет табл.1.4 в соответствующих элементах $\mu_{z_\beta}(x_j)$ до единицы. Так, в табл.1.6 $\mu_{z_1}(x_3)=0{,}56$, а в табл.1.4 $\mu_{z_2}(x_3)=0{,}44$, и т.д. В табл.1.6 можно увидеть, в каком году $a_{j2}^{t_i} \in W_1$, где W_1 подмножество незначимых коэффициентов стохастической связи между признаками и обобщенными факторами.

Таблица 1.6

Анализ признаков в нечетком подмножестве z_1, при $a_{кр1}$ = 0,21

Функция принадлежности \ Факторы	Признаки			
	f_1	f_2	f_3	f_4
$\mu_{z1}(x_j)=1$	x_1, x_2, x_4, x_5	x_5, x_8, x_{11}, x_{12}	x_{11}	$x_1, x_2, x_4, x_5, x_{10}, x_{11}, x_{12}$
$\mu_{z1}(x_j)<1$	$\frac{(x_3\,\|0{,}56)}{3,4,5}$ $\frac{(x_6\,\|0{,}62)}{2,3,4,5}$ $\frac{(x_8\,\|0{,}81)}{4,5,6}$ $\frac{(x_9\,\|0{,}19)}{1,2,3}$ $\frac{(x_{10}\,\|0{,}74)}{3,5,6}$	$\frac{(x_3\,\|0{,}64)}{1,2,3,4,5}$ $\frac{(x_6\,\|0{,}38)}{1,6}$ $\frac{(x_7\,\|0{,}82)}{1,4,5,6}$ $\frac{(x_9\,\|0{,}82)}{1,4,5,6}$ $\frac{(x_{10}\,\|0{,}82)}{1,4,5,6}$	$\frac{(x_1\,\|0{,}74)}{3,5,6}$ $\frac{(x_2\,\|0{,}56)}{3,4,5}$ $\frac{(x_3\,\|0{,}63)}{1,2,3,4,5}$ $\frac{(x_7\,\|0{,}74)}{3,5,6}$ $\frac{(x_8\,\|0{,}68)}{1,3,4,6}$ $\frac{(x_9\,\|0{,}45)}{4,5}$ $\frac{(x_{10}\,\|0{,}74)}{3,5,6}$ $\frac{(x_{12}\,\|0{,}98)}{2,3,4,5,6}$	$\frac{(x_3\,\|0{,}56)}{3,4,5}$ $\frac{(x_6\,\|0{,}89)}{1,2,4,5,6}$ $\frac{(x_9\,\|0{,}82)}{1,2,3,5,6}$

Из табл.1.6 следует, что показатель x_8 в первом факторе в последние три года находился в подмножестве z_1 незначимых признаков. В подмножестве z_1, во втором обобщенном факторе, этот же показатель находился все шесть лет, а в третьем факторе - в первом, третьем, четвертом и шестом годах. В четвертом факторе показатель x_8 в подмножестве незначимых не был ни разу.

Показатель x_5 в трех рассматриваемых факторах f_1, f_2 и f_4 все годы принадлежал к подмножеству незначимых z_1. Только в факторе f_3 показатель x_5 всегда принадлежал к подмножеству z_2.

В примере 1 представлено исследование, охватывающее шестилетний период. Если же период оказывается более длительным, а также в случаях, когда значение α в функции принадлежности постоянное число, или меняющееся от периода к периоду, как указано в формуле (1.11), то представляется возможным воспользоваться табл.1.1, рассчитанной на период до 25 лет и далее.

ПРИЛОЖЕНИЕ 2

ВАРИАНТЫ ЗАДАНИЙ И ИСХОДНЫЕ ДАННЫЕ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ НА ЭВМ

Варианты заданий 1-25 по корреляционному, регрессионному, компонентному и кластерному анализу даны в табл.2.1, а значения показателей производственно-хозяйственной деятельности предприятий машиностроения приведены в табл.2.2.

Рассматриваются следующие показатели:

Y_1 - производительность труда;

Y_2 - индекс снижения себестоимости продукции;

Y_3 - рентабельность;

X_4 - трудоемкость единицы продукции;

X_5 - удельный вес рабочих в составе ППП;

X_6 - удельный вес покупных изделий;

X_7 - коэффициент сменности оборудования;

X_8 - премии и вознаграждения на одного работника;

X_9 - удельный вес потерь от брака;

X_{10} - фондоотдача;

X_{11} - среднегодовая численность ППП;

X_{12} - среднегодовая стоимость ОПФ;

X_{13} - среднегодовой фонд заработной платы ППП;

X_{14} - фондовооруженность труда;

X_{15} - оборачиваемость нормируемых оборотных средств;

X_{16} - оборачиваемость ненормируемых оборотных средств;

X_{17} - непроизводственные расходы.

Таблица 2.1

Варианты заданий 1-25 по корреляционному, регрессионному, компонентному и кластерному анализу

№ варианта	Результативный признак, Y	Номера факторных признаков, X
1	1	6, 8, 11, 12, 17
2	1	6, 8, 11, 13, 17
3	1	8, 11, 12, 13, 17
4	1	6, 8, 13, 14, 17
5	1	8, 11, 13, 14, 17
6	1	6, 8, 12, 13, 17
7	1	7, 11, 12, 13, 17
8	1	7, 9, 12, 13, 17
9	1	8, 11, 12, 13, 17

Продолжение

№ варианта	Результативный признак, Y	Номера факторных признаков, X
10	1	8, 9, 13, 14, 17
11	1	5, 6, 7, 9, 17
12	1	5, 7, 9, 11, 17
13	1	5, 6, 12, 13, 17
14	1	5, 7, 10, 14, 17
15	1	5, 6, 10, 14, 17
16	3	8, 10, 15, 16, 17
17	3	5, 6, 10, 15, 17
18	3	5, 6, 7, 11, 12
19	3	8, 9, 10, 11, 17
20	3	8, 9, 10, 12, 17
21	2	4, 5, 6, 8, 9
22	2	4, 5, 6, 7, 9
23	2	4, 5, 6, 8, 9
24	2	4, 5, 8, 9, 17
25	2	4, 5, 7, 9, 17

Таблица 2.2

Таблица исходных данных

№ предприятия	Y_1	Y_2	Y_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	9,26	204,2	13,26	0,23	0,78	0,40	1,37	1,23	0,23	1,45
2	9,38	209,6	10,16	0,24	0,75	0,26	1,49	1,04	0,39	1,30
3	12,11	222,6	13,72	0,19	0,68	0,40	1,44	1,80	0,43	1,37
4	10,81	236,7	12,85	0,17	0,70	0,50	1,42	0,43	0,18	1,65
5	9,35	62,0	10,63	0,23	0,62	0,40	1,35	0,88	0,15	1,91
6	9,87	53,1	9,12	0,43	0,76	0,19	1,39	0,57	0,34	1,68
7	8,17	172,1	25,83	0,31	0,73	0,25	1,16	1,72	0,38	1,94
8	9,12	56,5	23,39	0,26	0,71	0,44	1,27	1,70	0,09	1,89
9	5,88	52,6	14,68	0,49	0,69	0,17	1,16	0,84	0,14	1,94
10	6,30	46,6	10,05	0,36	0,73	0,39	1,25	0,60	0,21	2,06
11	6,22	53,2	13,99	0,37	0,68	0,33	1,13	0,82	0,42	1,96
12	5,49	30,1	9,68	0,43	0,74	0,25	1,10	0,84	0,05	1,02
13	6,50	146,4	10,03	0,35	0,66	0,32	1,15	0,67	0,29	1,85
14	6,61	18,1	9,13	0,38	0,72	0,02	1,23	1,04	0,48	0,88
15	4,32	13,6	5,37	0,42	0,68	0,06	1,39	0,66	0,41	0,62
16	7,37	89,8	9,86	0,30	0,77	0,15	1,38	0,86	0,62	1,09
17	7,02	62,5	12,62	0,32	0,78	0,08	1,35	0,79	0,56	1,60
18	8,25	46,3	5,02	0,25	0,78	0,20	1,42	0,34	1,76	1,53
19	8,15	103,5	21,18	0,31	0,81	0,20	1,37	1,60	1,31	1,40
20	8,72	73,3	25,17	0,26	0,79	0,30	1,41	1,46	0,45	2,22
21	6,64	76,6	19,40	0,37	0,77	0,24	1,35	1,27	0,50	1,32
22	8,10	73,01	21,0	0,29	0,78	0,10	1,48	1,58	0,77	1,48
23	5,52	32,3	6,57	0,34	0,72	0,11	1,24	0,68	1,20	0,68
24	9,37	199,6	14,19	0,23	0,79	0,47	1,40	0,86	0,21	2,30
25	13,17	598,1	15,81	0,17	0,77	0,53	1,45	1,98	0,25	1,37
26	6,67	71,2	5,23	0,29	0,80	0,34	1,40	0,33	0,15	1,51
27	5,68	90,8	7,99	0,41	0,71	0,20	1,28	0,45	0,66	1,43
28	5,22	82,1	17,50	0,41	0,79	0,24	1,33	0,74	0,74	1,82
29	10,02	76,2	17,16	0,22	0,76	0,54	1,22	0,03	0,32	2,62

Продолжение

№ предприятия	Y_1	Y_2	Y_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
30	8,16	119,5	14,54	0,29	0,78	0,40	1,28	0,99	0,89	1,75
31	3,78	21,9	6,24	0,51	0,62	0,20	1,47	0,24	0,23	1,54
32	6,48	48,4	12,08	0,36	0,75	0,64	1,27	0,57	0,32	2,25
33	10,44	173,5	9,49	0,23	0,71	0,42	1,51	1,22	0,54	1,07
34	7,65	74,1	9,28	0,26	0,74	0,27	1,46	0,68	0,75	1,44
35	8,77	68,6	11,42	0,27	0,65	0,37	1,27	1,00	0,16	1,40
36	7,00	60,8	10,31	0,29	0,66	0,38	1,43	0,81	0,24	1,31
37	11,06	355,6	8,65	0,01	0,84	0,35	1,50	1,27	0,59	1,12
38	9,02	264,8	10,94	0,02	0,74	0,42	1,35	1,14	0,56	1,16
39	13,28	526,6	9,87	0,18	0,75	0,32	1,41	1,89	0,63	0,88
40	9,27	118,6	6,14	0,25	0,75	0,33	1,47	0,67	1,10	1,07
41	6,70	37,1	12,93	0,31	0,79	0,29	1,35	0,96	0,39	1,24
42	6,69	57,7	9,78	0,38	0,72	0,30	1,40	0,67	0,73	1,49
43	9,42	51,6	13,22	0,24	0,70	0,56	1,20	0,98	0,28	2,03
44	7,24	64,7	17,29	0,31	0,66	0,42	1,15	1,16	0,10	1,84
45	5,39	48,3	7,11	0,42	0,69	0,26	1,09	0,54	0,68	1,22
46	5,61	15,0	22,49	0,51	0,71	0,16	1,26	1,23	0,87	1,72
47	5,59	87,5	12,14	0,31	0,73	0,45	1,36	0,78	0,49	1,75
48	6,57	108,4	15,25	0,37	0,65	0,31	1,15	1,16	0,16	1,46
49	6,54	267,3	31,34	0,16	0,82	0,08	1,87	4,44	0,85	1,60
50	4,23	34,2	11,56	0,18	0,80	0,68	1,17	1,06	0,13	1,47
51	5,22	26,8	30,14	0,43	0,83	0,03	1,61	2,13	0,49	1,38
52	18,00	43,6	19,71	0,40	0,70	0,02	1,34	1,21	0,09	1,41
53	11,03	72,0	23,56	0,31	0,74	0,22	1,22	2,20	0,79	1,39

Продолжение таблицы 2.2

№ предприятия	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}
1	26006	167,69	47750	6,40	166,32	10,08	17,72
2	23935	186,10	50391	7,80	92,88	14,76	18,39
3	22589	220,45	43149	9,76	158,04	6,48	26,46
4	21220	169,30	41089	7,90	93,96	21,96	22,37
5	7394	39,53	14257	5,35	173,88	11,88	28,13
6	11586	40,41	22661	9,90	162,30	12,60	17,55
7	26609	102,96	52509	4,50	88,56	11,52	21,92
8	7801	37,02	14903	4,88	101,16	8,28	19,52
9	11587	45,74	25587	3,46	166,32	11,52	23,99
10	9475	40,07	16821	3,60	140,76	32,40	21,76
11	10811	45,44	19459	3,56	128,52	11,52	25,68
12	6371	41,08	12973	5,65	177,84	17,28	18,13
13	26761	136,14	50907	4,28	114,48	16,20	25,74
14	4210	42,39	6920	8,85	93,24	13,32	21,21
15	3557	37,39	5736	8,52	126,72	17,28	22,97
16	14148	101,78	26705	7,19	91,80	9,72	16,38
17	9872	47,55	20068	4,82	69,12	16,20	13,21
18	5975	32,61	11487	5,46	66,24	24,84	14,48
19	16662	103,25	32029	6,20	67,68	14,76	13,38
20	9166	38,95	18946	4,25	50,40	7,56	13,69
21	15118	81,32	28025	5,38	70,56	8,64	16,66

Продолжение

№ предприятия	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}
22	11429	67,26	20968	5,88	72,00	8,64	15,06
23	6462	59,92	11049	9,27	97,20	9,00	20,09
24	24628	107,34	45893	4,36	80,28	14,76	15,98
25	49727	512,60	99400	10,31	51,48	10,08	18,27
26	11470	53,81	20719	4,69	105,12	14,76	14,42
27	19448	80,83	36813	4,16	128,52	10,44	22,76
28	18963	59,42	33956	3,13	94,68	14,76	15,41
29	9185	36,96	17016	4,02	85,32	20,52	19,35
30	17478	91,43	34873	5,23	76,32	14,40	16,83
31	6265	17,16	11237	2,74	153,00	24,84	30,53
32	8810	27,29	17306	3,10	107,64	11,16	17,98
33	17659	184,33	39250	10,44	90,72	6,48	22,09
34	10342	58,42	19074	5,65	82,44	9,72	18,29
35	8901	59,40	18452	6,67	79,92	3,24	26,05
36	8402	49,63	17500	5,91	120,96	6,48	26,20
37	32625	391,27	7888	11,99	84,60	5,40	17,26
38	31160	258,62	58947	8,30	85,32	6,12	18,83
39	46461	75,66	94697	1,63	101,52	8,64	19,70
40	13833	123,68	29626	8,94	107,64	11,88	16,87
41	6391	37,21	11688	5,82	85,32	7,92	14,63
42	11115	53,37	21955	4,80	131,76	10,08	22,17
43	6555	32,87	12243	5,01	116,64	18,72	22,62
44	11085	45,63	20193	4,12	138,24	13,68	26,44
45	9484	48,41	20122	5,10	156,96	16,56	22,26
46	3967	13,58	7612	3,49	137,52	14,76	19,13
47	15283	63,99	27404	4,19	135,72	7,92	18,28
48	20874	104,55	39648	5,01	155,52	18,36	28,23
49	19418	222,11	43799	11,44	48,60	8,28	12,39
50	3351	25,76	6235	7,67	42,84	14,04	11,64
51	6338	29,52	11524	4,66	142,20	16,92	8,62
52	9756	41,99	17309	4,30	145,80	11,16	20,10
53	11795	78,11	22225	6,62	120,52	14,76	19,41

ПРИЛОЖЕНИЕ 3

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

Таблица 3.1

Нормальный закон распределения. Значение функции $\Phi(t)=P(|T|\le t_{табл})$

Целые и десятичные доли, t	Сотые доли, t									
	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	0797	0876	0955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2960	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	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	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7994	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9476	9488	9500	9512	9523	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	9643	9651	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9841	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9910	9912	9915	9917	9920	9922	9924	9926	9928
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947
2,8	9949	9951	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	0,9973	0,9974	0,9975	0,9976	0,9976	0,9977	0,9978	0,9979	0,9979	0,9980
3,1	9981	9981	9982	9983	9983	9984	9984	9985	9985	9986
3,5	9995	9996	9996	9996	9996	9996	9996	9996	9997	9997
3,6	9997	9997	9997	9997	9997	9997	9997	9998	9998	9998
3,7	9998	9998	9998	9998	9998	9998	9998	9998	9998	9998
3,8	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
3,9	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
4,0	0,999936	9999	9999	9999	9999	9999	9999	9999	9999	9999
4,5	0,999994	-	-	-	-	-	-	-	-	-
5,0	0,99999994	-	-	-	-	-	-	-	-	-

Таблица 3.2

Распределение Стьюдента (t-распределение)

ν	Вероятность $\alpha=S(t)=P(\|T\|>t_{табл})$												
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,043	6,859
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,953	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,583
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,833
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,402	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725

Продолжение

ν	Вероятность $\alpha=St(t)=P(\|T\|>t_{табл})$											
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Таблица 3.3

Распределение Пирсона (χ^2-распределение)
Значения $\chi^2_{табл}$ для вероятностей P ($\chi^2 > \chi^2_{табл}$)

ν	Вероятность										
	0,999	0,995	0,99	0,98	0,975	0,95	0,90	0,80	0,75	0,70	0,50
1	0,05157	0,04393	0,03157	0,03628	0,03982	0,00393	0,0158	0,0642	0,102	0,148	0,455
2	0,00200	0,0100	0,0201	0,0404	0,0506	0,103	0,211	0,446	0,575	0,713	1,386
3	0,0243	0,0717	0,115	0,185	0,216	0,352	0,584	1,005	1,213	1,424	2,366
4	0,0908	0,207	0,297	0,429	0,484	0,711	1,064	1,649	1,923	2,195	3,357
5	0,210	0,412	0,554	0,752	0,831	1,145	1,610	2,343	2,675	3,000	4,351
6	0,381	0,676	0,872	1,134	1,237	1,635	2,204	3,070	3,455	3,828	5,348
7	0,598	0,989	1,239	1,564	1,690	2,167	2,833	3,822	4,255	4,671	6,346
8	0,857	1,344	1,646	2,032	2,180	2,733	3,490	4,594	5,071	5,527	7,344
9	1,152	1,735	2,088	2,532	2,700	3,325	4,168	5,380	5,899	6,393	8,343
10	1,479	2,156	2,558	3,059	3,247	3,240	4,865	6,179	6,737	7,267	9,342
11	1,834	2,603	3,053	3,609	3,816	4,575	5,578	6,989	7,584	8,148	10,341
12	2,214	3,074	3,571	4,178	4,404	5,226	6,304	7,807	8,438	9,034	11,340
13	2,617	3,565	4,107	4,765	5,009	5,892	7,042	8,634	9,299	9,926	12,340
14	3,041	4,075	4,660	5,368	5,629	6,571	7,790	9,467	10,165	10,821	13,339
15	3,483	4,601	5,229	5,985	6,262	7,261	8,547	10,307	11,036	11,721	14,339
16	3,942	5,142	5,812	6,614	6,908	7,962	9,312	11,152	11,912	12,624	15,338
17	4,416	5,697	6,408	7,255	7,564	8,672	10,085	12,002	12,892	13,531	16,338
18	4,905	6,265	7,015	7,906	8,231	9,390	10,865	12,857	13,675	14,440	17,338
19	5,407	6,844	7,633	8,567	8,907	10,117	11,651	13,716	14,562	15,352	18,338
20	5,921	7,434	8,260	9,237	9,591	10,871	12,443	14,578	15,452	16,266	19,337
21	6,447	8,034	8,897	9,915	10,283	11,591	13,240	15,445	16,344	17,182	20,337
22	6,983	8,643	9,542	10,600	10,982	12,338	14,041	16,314	17,240	18,101	21,337
23	7,529	9,260	10,196	11,293	11,688	13,091	14,848	17,187	18,137	19,021	22,337
24	8,035	9,886	10,856	11,992	12,401	13,848	15,659	18,062	19,037	19,943	23,337

Продолжение

ν	Вероятность										
	0,999	0,995	0,99	0,98	0,975	0,95	0,90	0,80	0,75	0,70	0,50
25	8,649	10,520	11,524	12,697	13,120	14,611	16,173	18,940	19,939	20,887	24,337
26	9,222	11,160	12,198	13,409	13,844	15,379	17,292	19,820	20,843	21,792	25,336
27	9,803	11,808	12,879	14,125	14,573	16,151	18,114	20,703	21,749	22,719	26,136
28	10,391	12,461	13,565	14,847	15,308	16,928	18,937	21,588	22,657	23,617	27,336
29	10,986	13,121	14,256	15,574	16,047	17,708	19,768	22,475	23,567	24,577	28,336
30	11,588	13,787	14,953	16,306	16,791	18,493	20,599	23,364	24,478	25,508	29,336

Продолжение таблицы 3.3

ν	Вероятность									
	0,30	0,25	0,20	0,10	0,05	0,025	0,02	0,01	0,005	0,001
1	1,074	1,323	1,642	2,706	3,841	5,024	5,412	6,635	7,879	10,827
2	2,408	2,773	3,219	4,605	5,991	7,378	7,824	9,210	10,597	13,815
3	3,665	4,108	4,642	6,251	7,815	9,348	9,837	11,345	12,838	16,268
4	4,878	5,385	5,989	7,779	9,488	11,143	11,668	13,277	14,860	18,465
5	6,064	6,626	7,289	9,236	11,070	12,839	13,388	15,086	16,750	20,517
6	7,231	7,841	8,558	10,645	12,592	14,449	15,033	16,812	18,548	22,457
7	8,383	9,037	9,803	12,017	14,067	16,013	16,622	18,475	20,278	24,322
8	9,524	10,219	11,030	13,362	15,507	17,535	18,168	20,090	21,955	26,125
9	10,656	11,389	12,242	14,684	16,919	19,023	19,679	21,666	23,589	27,877
10	11,781	12,549	13,412	15,987	18,307	20,483	21,161	23,209	25,188	29,588
11	12,899	13,701	14,631	17,275	19,675	21,920	22,618	24,725	26,757	31,264
12	14,011	14,845	15,812	18,549	21,026	23,337	24,054	26,217	28,300	32,909
13	15,119	15,984	16,985	19,812	22,362	24,736	25,472	27,688	29,819	34,528
14	16,222	17,117	18,151	21,064	23,685	26,119	26,873	29,141	31,319	36,123
15	17,322	18,245	19,311	22,307	24,996	27,488	28,259	30,578	32,801	37,697
16	18,418	19,369	20,465	23,542	26,296	28,845	29,633	32,000	34,267	39,252

Продолжение

ν	Вероятность									
	0,30	0,25	0,20	0,10	0,05	0,025	0,02	0,01	0,005	0,001
17	19,511	20,489	21,615	24,769	27,587	30,191	30,995	33,409	35,718	40,790
18	20,601	21,605	22,760	25,989	28,869	31,526	32,346	34,805	37,156	42,312
19	21,689	22,718	23,900	27,204	30,144	32,852	33,687	36,191	38,582	43,820
20	22,775	23,828	25,038	28,412	31,410	34,170	35,020	37,566	39,997	45,315
21	23,858	24,935	26,171	29,615	32,671	35,479	36,343	38,932	41,401	46,797
22	24,939	26,039	27,301	30,813	33,924	36,781	37,659	40,289	42,796	48,268
23	26,018	27,141	28,429	32,007	35,172	38,076	38,968	41,638	44,181	49,728
24	27,096	28,241	29,553	33,196	36,415	39,364	40,270	42,980	45,558	51,170
25	28,172	29,339	30,675	34,382	37,652	40,046	41,566	44,314	46,928	52,620
26	29,246	30,434	31,795	35,563	38,885	41,923	42,856	45,642	48,290	54,052
27	30,319	31,528	32,912	36,741	40,113	43,194	44,140	46,963	49,645	55,476
28	31,391	32,620	34,027	37,916	41,337	44,461	45,419	48,278	50,993	56,893
29	32,461	33,711	35,139	39,087	42,557	45,722	46,693	49,588	52,336	58,302
30	33,530	34,800	36,250	40,256	43,773	46,979	47,962	50,892	53,672	59,703

Таблица 3.4.

Распределение Фишера-Снедекора (F-распределение)

Значения $F_{табл}$, удовлетворяющие условию $P(F>F_{табл})$. Первое значение соответствует вероятности 0,05; второе - вероятности 0,01 и третье - вероятности 0,001; ν_1 - число степеней свободы числителя; ν_2 - знаменателя

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞	t
1	161,4 4052 406523	199,5 4999 500016	215,7 5403 536700	224,6 5625 562527	230,2 5764 576449	234,0 5859 585953	238,9 5981 598149	243,9 6106 610598	249,0 6234 623432	253,3 6366 636535	12,71 63,66 636,2
2	18,51 98,49 998,46	19,00 99,01 999,00	19,16 00,17 999,20	19,25 99,25 999,20	19,30 99,30 999,20	19,33 99,33 999,20	19,37 99,36 999,40	19,41 99,42 999,60	19,45 99,46 999,40	19,50 99,50 999,40	4,30 9,92 31,00
3	10,13 34,12 67,47	9,55 30,81 148,51	9,28 29,46 141,10	9,12 28,71 137,10	9,01 28,24 134,60	8,94 27,91 132,90	8,84 27,49 130,60	8,74 27,05 128,30	8,64 26,60 125,90	8,53 26,12 123,50	3,18 5,84 12,94
4	7,71 21,20 74,13	6,94 18,00 61,24	6,59 16,69 56,18	6,39 15,98 53,43	6,26 15,52 51,71	6,16 15,21 50,52	6,04 14,80 49,00	5,91 14,37 47,41	5,77 13,93 45,77	5,63 13,46 44,05	2,78 4,60 8,61
5	6,61 16,26 47,04	5,79 13,27 36,61	5,41 12,06 33,20	5,19 11,39 31,09	5,05 10,97 20,75	4,95 10,67 28,83	4,82 10,27 27,64	4,68 9,89 26,42	4,53 9,47 25,14	4,36 9,02 23,78	2,57 4,03 6,86
6	5,99 13,74 35,51	5,14 10,92 26,99	4,76 9,78 23,70	4,53 9,15 21,90	4,39 8,75 20,81	4,28 8,47 20,03	4,15 8,10 19,03	4,00 7,72 17,99	3,84 7,31 16,89	3,67 6,88 15,75	2,45 3,71 5,96
7	5,59 12,25 29,22	4,74 9,55 21,69	4,35 8,45 18,77	4,12 7,85 17,19	3,97 7,46 16,21	3,87 7,19 15,52	3,73 6,84 14,63	3,57 6,47 13,71	3,41 6,07 12,73	3,23 5,65 11,70	2,36 3,50 5,40

Продолжение

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞	t
8	5,32 11,26 25,42	4,46 8,65 18,49	4,07 7,59 15,83	3,84 7,10 14,39	3,69 6,63 13,49	3,58 6,37 12,86	3,44 6,03 12,04	3,28 5,67 11,19	3,12 5,28 10,30	2,99 4,86 9,35	2,31 3,36 5,04
9	5,12 10,56 22,86	4,26 8,02 16,39	3,86 6,99 13,90	3,63 6,42 12,56	3,48 6,06 11,71	3,37 5,80 11,13	3,23 5,47 10,37	3,07 5,11 9,57	2,90 4,73 8,72	2,71 4,31 7,81	2,26 3,25 4,78
10	4,96 10,04 21,04	4,10 7,56 14,91	3,71 6,55 12,55	3,48 5,99 11,28	3,33 5,64 10,48	3,22 5,39 9,92	3,07 5,06 9,20	2,91 4,71 8,45	2,74 4,33 7,64	2,54 3,91 6,77	2,23 3,17 4,59
11	4,84 9,65 19,69	3,98 7,20 13,81	3,59 6,22 11,56	3,36 5,67 10,35	3,20 5,32 9,58	3,09 5,07 9,05	2,95 4,74 8,35	2,79 4,40 7,62	2,61 4,02 6,85	2,40 3,60 6,00	2,20 3,11 4,49
12	4,75 9,33 18,64	3,88 6,93 12,98	3,49 5,95 10,81	3,26 5,41 9,63	3,11 5,06 8,89	3,00 4,82 8,38	2,85 4,50 7,71	2,69 4,16 7,00	2,50 3,78 6,25	2,30 3,36 5,42	2,18 3,06 4,32
13	4,67 9,07 17,81	3,80 6,70 12,31	3,41 5,74 10,21	3,18 5,20 9,07	3,02 4,86 8,35	2,92 4,62 7,86	2,77 4,30 7,21	2,60 3,96 6,52	2,42 3,59 5,78	2,21 3,16 4,97	2,16 3,01 4,12
14	4,60 8,86 17,14	3,74 6,51 11,78	3,34 5,56 9,73	3,11 5,03 8,62	2,96 4,69 7,92	2,85 4,46 7,44	2,70 4,14 6,80	2,53 3,80 6,13	2,35 3,43 5,41	2,13 3,00 4,60	2,14 2,98 4,14
15	4,45 8,68 16,59	3,68 6,36 11,34	3,29 5,42 9,34	3,06 4,89 8,25	2,90 4,56 7,57	2,79 4,32 7,09	2,64 4,00 6,47	2,48 3,67 5,81	2,29 3,29 5,10	2,07 2,87 4,31	2,13 2,95 4,07

Продолжение

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞	t
16	4,41 8,53 16,12	3,63 6,23 10,97	3,24 5,29 9,01	3,01 4,77 7,94	2,85 4,44 7,27	2,74 4,20 6,80	2,59 3,89 6,20	2,42 3,55 5,55	2,24 3,18 4,85	2,01 2,75 4,06	2,12 2,92 4,02
17	4,45 8,40 15,72	3,59 6,11 10,66	3,20 5,18 8,73	2,96 4,67 7,68	2,81 4,34 7,02	2,70 4,10 6,56	2,55 3,79 5,96	2,38 3,45 5,32	2,19 3,08 4,63	1,96 2,65 3,85	2,11 2,90 3,96
18	4,41 8,28 15,38	3,55 6,01 10,39	3,16 5,09 8,49	2,93 4,58 7,46	2,77 4,25 6,81	2,66 4,01 6,35	2,51 3,71 5,76	2,34 3,37 5,13	2,15 3,01 4,45	1,92 2,57 3,67	2,10 2,88 3,92
19	4,38 8,18 15,08	3,52 5,93 10,16	3,13 5,01 8,28	2,90 4,50 7,26	2,74 4,17 6,61	2,63 3,94 6,18	2,48 3,63 5,59	2,31 3,30 4,97	2,11 2,92 4,29	1,88 2,49 3,52	2,09 2,86 3,88
20	4,35 8,10 14,82	3,49 5,85 9,95	3,10 4,94 8,10	2,87 4,43 7,10	2,71 4,10 6,46	2,60 3,87 6,02	2,45 3,56 5,44	2,28 3,23 4,82	2,08 2,86 4,15	1,84 2,42 3,38	2,09 2,84 3,85
21	4,32 8,02 14,62	3,47 5,78 9,77	3,07 4,87 7,94	2,84 4,37 6,95	2,68 4,04 6,32	2,57 3,81 5,88	2,42 3,51 5,31	2,25 3,17 4,70	2,05 2,80 4,03	1,82 2,36 3,26	2,08 2,83 3,82
22	4,30 7,94 14,38	3,44 5,72 9,61	3,05 4,82 7,80	2,82 4,31 6,81	2,66 3,99 6,19	2,55 3,75 5,76	2,40 3,45 5,19	2,23 3,12 4,58	2,03 2,75 3,92	1,78 2,30 3,15	2,07 2,82 3,79
23	4,28 7,88 14,19	3,42 5,66 9,46	3,03 4,76 7,67	2,80 4,26 6,70	2,64 3,94 6,08	2,53 3,71 5,56	2,38 3,41 5,09	2,20 3,07 4,48	2,00 2,70 3,82	1,76 2,26 3,05	2,07 2,81 3,77

Продолжение

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞	t
24	4,26 7,82 14,03	3,40 5,61 9,34	3,01 4,72 7,55	2,78 4,22 6,59	2,62 3,90 5,98	2,51 3,67 5,55	2,36 3,36 4,99	2,18 3,03 4,39	1,98 2,66 3,74	1,73 2,21 2,97	2,06 2,80 3,75
25	4,24 7,77 13,88	3,38 5,57 9,22	2,99 4,68 7,45	2,76 4,18 6,49	2,60 3,86 5,89	2,49 3,63 5,46	2,34 3,32 4,91	2,16 2,99 4,31	1,96 2,62 3,66	1,71 2,17 2,89	2,06 2,79 3,72
26	4,22 7,72 13,74	3,37 5,53 9,12	2,98 4,64 7,36	2,74 4,14 6,41	2,59 3,82 5,80	2,47 3,59 5,38	2,32 3,29 4,83	2,15 2,96 4,24	1,95 2,58 3,59	1,69 2,13 2,82	2,06 2,78 3,71
27	4,21 7,68 13,61	3,35 5,49 9,02	2,96 4,60 7,27	2,73 4,11 6,33	2,57 3,78 5,73	2,46 3,56 5,31	2,30 3,26 4,76	2,13 2,93 4,17	1,93 2,55 3,52	1,67 2,10 2,76	2,05 2,77 3,69
28	4,19 7,64 13,50	3,34 5,54 8,93	2,95 4,57 7,18	2,71 4,07 6,25	2,56 3,75 5,66	2,44 3,53 5,24	2,29 3,23 4,69	2,12 2,90 4,11	1,91 2,52 3,46	1,65 2,06 2,70	2,05 2,76 3,67
29	4,18 7,60 13,39	3,33 5,42 8,85	2,93 4,54 7,12	2,70 4,04 6,19	2,54 3,73 5,59	2,43 3,50 5,18	2,28 3,20 4,65	2,10 2,87 4,05	1,90 2,49 3,41	1,64 2,03 2,64	2,05 2,76 3,66
30	4,17 7,56 13,29	3,32 5,39 8,77	2,92 4,51 7,05	2,69 4,02 6,12	2,53 3,70 5,53	2,42 3,47 5,12	2,27 3,17 4,58	2,09 2,84 4,00	1,89 2,47 3,36	1,62 2,01 2,59	2,04 2,75 3,64
60	4,00 7,08 11,97	3,15 4,98 7,76	2,76 4,13 6,17	2,52 3,65 5,31	2,37 3,34 4,76	2,25 3,12 4,37	2,10 2,82 3,87	1,92 2,50 3,31	1,70 2,12 2,76	1,39 1,60 1,90	2,00 2,66 3,36
∞	3,84 6,64 10,83	2,99 4,60 6,91	2,60 3,78 5,42	2,37 3,32 4,62	2,21 3,02 4,10	2,09 2,80 3,74	1,94 2,51 3,27	1,75 2,18 2,74	1,52 1,79 2,13	1,03 1,04 1,05	1,96 2,58 3,29

Продолжение

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞	t
∞	3,84 6,64 10,83	2,99 4,60 6,91	2,60 3,78 5,42	2,37 3,32 4,62	2,21 3,02 4,10	2,09 2,80 3,74	1,94 2,51 3,27	1,75 2,18 2,74	1,52 1,79 2,13	1,03 1,04 1,05	1,96 2,58 3,29

Таблица 3.5

Таблица Фишера-Иейтса

Зачения $r_{кр}$, найденные для уровня значимости α и чисел степеней свободы $\nu=n-2$ в случае парной корреляции и $\nu=n-l-2$, где l - число исключенных величин в случае частной корреляции

	Двусторонние границы					Двусторонние границы			
ν	0,05	0,02	0,01	0,001	ν	0,05	0,02	0,01	0,001
1	0,997	1,000	1,000	1,000	16	0,468	0,543	0,590	0,708
2	0,950	0,980	0,990	0,999	17	0,456	0,529	0,575	0,693
3	0,878	0,934	0,959	0,991	18	0,444	0,516	0,561	0,679
4	0,811	0,882	0,917	0,974	19	0,433	0,503	0,549	0,665
5	0,754	0,833	0,875	0,951	20	0,423	0,492	0,537	0,652
6	0,707	0,789	0,834	0,925	25	0,381	0,445	0,487	0,597
7	0,666	0,750	0,798	0,898	30	0,349	0,409	0,449	0,554
8	0,632	0,715	0,765	0,872	35	0,325	0,381	0,418	0,519
9	0,602	0,685	0,735	0,847	40	0,304	0,358	0,393	0,490
10	0,576	0,658	0,708	0,823	45	0,288	0,338	0,372	0,465
11	0,553	0,634	0,684	0,801	50	0,273	0,322	0,354	0,443
12	0,532	0,612	0,661	0,780	60	0,250	0,295	0,325	0,408
13	0,514	0,592	0,641	0,760	70	0,232	0,274	0,302	0,380
14	0,497	0,574	0,623	0,742	80	0,217	0,257	0,283	0,338
15	0,482	0,558	0,606	0,725	90	0,205	0,242	0,267	0,338
					100	0,195	0,230	0,254	0,321
ν	0,025	0,01	0,005	0,0005	ν	0,025	0,01	0,005	0,0005
	Односторонние границы					Односторонние границы			

Таблица 3.6

Таблица Z-преобразования Фишера

$$Z=\frac{1}{2}\{ln(1+r)-ln(1-r)\}$$

r	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0101	0,0200	0,0300	0,0400	0,0501	0,0601	0,0701	0,0802	0,0902
1	0,1003	0,1104	0,1206	0,1308	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
3	0,3095	0,3205	0,3316	0,3428	0,3541	0,3654	0,3767	0,3884	0,4001	0,4118
4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
5	0,5493	0,5627	0,5764	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
6	0,6932	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
7	0,8673	0,8872	0,9077	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
9	1,4722	1,5275	1,5890	1,6584	1,7381	1,8318	1,9459	2,0923	2,2976	2,6467
0,99	2,6466	2,6996	2,7587	2,8257	2,9031	2,9945	3,1063	3,2504	3,4534	3,8002

Таблица 3.7

Значение плотности $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ **значение для нормированного нормального закона распределения, f(–t) = f(t)**

Целые и десятые доли t	Сотые доли t									
	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	3857	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	3525	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	3-56	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2631	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	3012	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	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0762	0748	0734	0721	0707	0694	0681	0669

Продолжение

Целые и десятые доли t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
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	0,478	0,0568	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	0055	0053	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
4,0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

Таблица 3.8

Значение функции Пуассона $P(X = m) = \frac{\lambda^m}{m!}e^{-\lambda}$

m \ λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0,2223	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1216	0,1438	0,1547	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0000	0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0007	0,0012	0,0020	0,0031
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

Продолжение

m \ λ	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0
0	0,1353	0,0498	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001	0,0001
1	0,2707	0,1494	0,0733	0,0337	0,0149	0,0064	0,0027	0,0011	0,0005
2	0.2707	0,2240	0,1465	0,0842	0,0446	0,0223	0,0107	0,0050	0,0023
3	0,1805	0,2240	0,1954	0,1404	0,892	0,0521	-,0286	0,0150	0,0076
4	0,0902	0,1681	0,1954	0,1755	0,1339	0,0912	0,0572	0,0337	0,0189
5	0,0361	0,1008	0,1563	0,1755	0,1606	0,1277	0,0916	0,0607	0,0378
6	0,0120	0,0504	0,1042	0,1462	0,1606	0,1490	0,1221	0,0911	0,0631
7	0,0034	0,0216	0,0595	0,1045	0,1377	0,1490	0,1396	0,1171	0,0901
8	0,0009	0,0081	0,0298	0,0653	0,1033	0,1304	0,1396	0,1318	0,1126
9	0,0002	0,0027	0,0132	0,363	0,0689	0,1014	0,1241	0,1318	0,1251
10	0,0000	0,0008	0,0053	0,0181	0,0413	0,0710	0,0993	0,1186	0,1251
11	0,0000	0,0002	0,0019	0,0082	0,0225	0,0452	0,0722	0,970	0,1137
12	0,0000	0,0001	0,0006	0,0034	0,0113	0,0264	0,0481	0,728	0,0948
13	0,0000	0,0000	0,0002	0,0013	0,0052	0,0142	0,0296	0,0504	0,0729
14	0,0000	0,0000	0,0001	0,0005	0,0022	0,0071	0,0169	0,0324	0,0521
15	0,0000	0,0000	0,0000	0,0002	0,0009	0,0033	0,0090	0,0194	0,0347
16	0,0000	0,0000	0,0000	0,0000	0,0003	0,0015	0,0045	0,0109	0,0217
17	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0021	0,0058	0,0128
18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0029	0,0071
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0037

Таблица 3.9

G - распределение

5 и 1%-ные пределы для отношения G наибольшей выборочной дисперсии к сумме L выборочных дисперсий, полученных из L независимых выборок объемом n.

Первое значение соответствует уровню значимости α = 0,05, а второе - α = 0,01

L \ n-1	1	2	3	4	5	6	7	8	9	10	16	36	144	∞
2	0,998	0,975	0,939	0,906	0,877	0,853	0,838	0,816	0,801	0,788	0,734	0,660	0,518	0,500
	0,999	0,995	0,979	0,959	0,937	0,917	0,809	0,882	0,867	0,854	0,795	0,700	0,606	0,500
3	0,967	0,871	0,798	0,746	0,707	0,677	0,653	0,633	0,617	0,603	0,547	0,475	0,403	0,333
	0,993	0,942	0,883	0,834	0,903	0,761	0,734	0,711	0,691	0,674	0,606	0,515	0,423	0,333
4	0,906	0,768	0,684	0,629	0,590	0,560	0,537	0,518	0,502	0,488	0,437	0,372	0,309	0,250
	0,968	0,864	0,781	0,721	0,676	0,641	0,613	0,590	0,570	0,554	0,488	0,406	0,325	0,250
5	0,841	0,684	0,598	0,544	0,507	0,478	0,456	0,439	0,424	0,412	0,365	0,307	0,251	0,200
	0,928	0,789	0,696	0,633	0,588	0,553	0,526	0,504	0,485	0,470	0,409	0,335	0,254	0,200
6	0,781	0,616	0,532	0,480	0,445	0,418	0,398	0,382	0,368	0,357	0,314	0,261	0,212	0,167
	0,883	0,722	0,626	0,564	0,520	0,487	0,461	0,440	0,423	0,408	0,353	0,286	0,223	0,167
7	0,727	0,561	0,480	0,431	0,397	0,373	0,354	0,338	0,326	0,315	0,276	0,228	0,183	0,143
	0,838	0,664	0,569	0,508	0,466	0,435	0,411	0,391	0,375	0,362	0,311	0,249	0,193	0,143
8	0,680	0,516	0,438	0,391	0,360	0,336	0,319	0,304	0,293	0,283	0,246	0,202	0,162	0,15
	0,795	0,615	0,521	0,463	0,423	0,393	0,370	0,352	0,337	0,325	0,278	0,221	0,170	0,125
9	0,639	0,478	0,403	0,358	0,329	0,307	0,290	0,277	0,266	0,257	0,223	0,182	0,145	0,111
	0,754	0,573	0,481	0,425	0,387	0,359	0,338	0,321	0,307	0,295	0,251	0,199	0,152	0,111

Методические указания к использованию некоторых таблиц

В табл.3.1 протабулирована функция

$$\Phi(t)=\frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{x^2}{2}}dx,$$

где $\Phi(t)$ - плотность нормированной нормально распределенной случайной величины $T\in N(0,1)$.

Вероятность попадания случайной величины T в интервал от t_1 до t_2 вычисляется по формуле

$$P(t_1<T<t_2)=\frac{1}{2}[\Phi(t_2)-\Phi(t_1)].$$

$\Phi(t)$ обладает следующими свойствами:

$$\Phi(-t)=-\Phi(t);\quad \Phi(\infty)=1;\quad \Phi(3)=0{,}9973.$$

Пример

$$P(-1{,}36<T<2{,}15)=\frac{1}{2}[\Phi(2{,}15)-\Phi(-1{,}36)]=$$

$$=\frac{1}{2}[0{,}9684+0{,}8262]=0{,}8973.$$

В табл.3.2 протабулирована вероятность выхода за пределы интервала от $-t$ до $+t$ случайной величины, имеющей распределение Стьюдента (t-распределение) с числом степеней свободы ν.

$$\alpha=St(t;\nu)=P(|T|>t),$$

где $f(t;\ \nu)$ - плотность распределения Стьюдента с числом степеней свободы ν.

Вероятность попадания случайной величины T в интервал от t_1 до t_2 вычисляется по формуле:

$$P(t_1<T<t_2)=\frac{1}{2}[St(t_1)-St(t_2)].$$

Функция $St(t)$ обладает следующими свойствами:

$$St(-t)=2-St(t);\qquad St(\infty)=0;\quad St(-\infty)=2;\quad St(0)=1.$$

Пример

При $\nu=10$ определить

$$P(-1{,}36<T<2{,}15)=\frac{1}{2}[St(-1{,}36)-St(2{,}15)]=\frac{1}{2}[2-St(1{,}36)-$$

$$-St(2{,}15)]\cong\frac{1}{2}[2-St(1{,}372)-St(2{,}228)]=\frac{1}{2}[2-0{,}2-0{,}05]=0{,}875.$$

Чтобы не прибегать к интерполяции, в строке, соответствующей $\nu=10$, мы взяли ближайшие к заданным значениям 1,36 и 2,15. Каждая строка табл.3.2 отвечает t-распределению, с соответствующим числом степеней свободы ν. В табл.3.3 протабулирована вероятность того, что наблюдаемое значение случайной величины χ^2, имеющей распределение Пирсона (хи-квадрат распределение) с числом степеней свободы ν, превысит табличное значение $\chi^2_{табл}$.

Вероятность попадания случайной величины χ^2 в интервал от χ^2_1 до χ^2_2 вычисляется по формуле

$$P(\chi^2_1<\chi^2<\chi^2_2)=P(\chi^2>\chi^2_1)-P(\chi^2>\chi^2_2)=P_i(\chi^2_1)-P_i(\chi^2_2).$$

Функция $P_i(\chi^2_{табл})$ обладает следующими свойствами:

$$P_i(0)=1;\qquad P_i(\infty)=0.$$

Пример

При $\nu=10$ определить

$$P(2{,}5<\chi^2<19{,}0)=P_i(2{,}5)-P_i(19{,}0)\cong P_i(2{,}558)-P_i(18{,}307)=$$

$$=0{,}99-0{,}05=0{,}94.$$

Чтобы не прибегать к интерполяции в строке табл.3.3, соответствующей $\nu=10$, мы взяли ближайшие к заданным значениям 2,5 и 19,0.

Каждая строка таблицы отвечает χ^2-распределению с соответствующим числом степеней свободы ν.

В табл.3.4 для случайной величины F, имеющей закон распределения Фишера-Снедекора (F-распределение) с числами степеней свободы числителя ν_1 и знаменателя ν_2, протабулированы три табличных значения, соответствующие трем вероятностям (уровням значимости):

$$\alpha=P(F>F_{табл})=0{,}05;\ 0{,}01\ \text{и}\ 0{,}001.$$

Пример

Уровню значимости $\alpha=0{,}01$ и числам степеней свободы числителя $\nu_1=5$ и знаменателя $\nu_2=7$ соответствует $F_{табл}=7{,}46$.

Статистика F строится таким образом, чтобы наблюдаемое значение было не меньше единицы.

ПРИЛОЖЕНИЕ 4

БУТСТРЕП-ОЦЕНКИ

Сущность метода сводится к дополнению данными численного моделирования фактических наблюдений. При этом моделирование производится только на основе фактических данных.

Входными параметрами метода являются:

$\{x_1, x_2, \ldots, x_n\}$ - исходная выборка;

k - количество моделируемых выборок (k>50);

p - вероятностный уровень оценки математического ожидания (рекомендуемые значения 0,7-0,9).

Оценку математического ожидания для малой выборки найдем по следующему алгоритму.

1. Моделирование выборок с использованием датчика натуральных чисел, равномерно распределенных в интервале от 1 до n,

$$V_j = \left\{ x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)} \right\}.$$

2. Для каждой выборки V_j определим оценку математического ожидания:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_i^{(j)}, \quad j = \overline{1,k}.$$

3. Для вариационного ряда средних значений, найденных по выборкам, построим интервальный ряд.

4. С хвостов построенного интервального ряда отсечем интервалы таким образом, чтобы суммарная часть отброшенных интервалов не превосходила (1-p). По оставшимся интервалам определим интервальную оценку математического ожидания.

Пример

В результате шести испытаний сложного прибора были получены следующие значения времени наработки на отказ, час.:

2,1; 3,4; 6,5; 8,2; 7,3; 5,6.

Необходимо определить интервальную оценку математического ожидания времени наработки на отказ указанного прибора.

Решение

Задача решается с помощью программы "ОЛИМП: СтатЭксперт" в среде электронной таблицы Excel 7.0. Исходные данные вводятся в табличном виде. Выбирается пункт меню "Дескриптивные статистики", параметр "Число итераций бутстреп-метода" устанавливается в 1000 (рис.4.1).

Наработка на отказ: 2,1; 3,4; 6,5; 8,2; 7,3; 5,6.

Фрагмент выходной таблицы и графика, касающиеся бутстреп-оценок среднего значения, приведены ниже.

Интервальный ряд бутстреп-средних					
№ интервала	**Начало**	**Середина**	**Конец**	**Частота**	**Частость**
1	2,062	2,317	2,571	2	0,002
2	2,571	2,826	3,080	12	0,012
3	3,080	3,335	3,589	27	0,027
4	3,589	3,844	4,098	70	0,070
5	4,098	4,353	4,608	138	0,138
6	4,608	4,862	5,117	195	0,195
7	5,117	5,371	5,626	222	0,222
8	5,626	5,880	6,135	182	0,182
9	6,135	6,389	6,644	100	0,100
10	6,644	6,898	7,153	41	0,041
11	7,153	7,408	7,662	9	0,009
12	7,662	7,917	8,171	2	0,002
Число итераций метода 1000					

Рис.4.1. Гистограмма бутстреп

Следующий этап решения задачи связан с отбрасыванием интервалов с малыми значениями частости. Например, если мы отбросим крайние ин-

тервалы, имеющие значения частости по 0,002 (сумма частостей отброшенных интервалов равна 0,004), то результаты бутстреп-моделирования можно интерпретировать следующим образом: с вероятностью 0,996 среднее значение случайной величины (в данном случае время наработки прибора на отказ) находится в диапазоне 2,571 - 7,662.

Если необходимо уточнить полученные результаты, можно отбросить еще два крайних интервала (сумма частостей отброшенных интервалов равна 0,004+0,012+0,009=0,025).

В этом случае результаты бутстреп-моделирования можно интерпретировать следующим образом: с вероятностью 0,975 среднее значение случайной величины (в данном случае, время наработки прибора на отказ) находится в диапазоне 3,08 - 7,153.

Как следует из рассмотренных примеров, интерпретация результатов бутстреп-моделирования зависит от заранее заданного вероятностного уровня, с которым необходимо определять интервальную оценку среднего значения. С другой стороны, процесс отбрасывания интервалов существенно зависит от конкретного распределения частостей, ширины интервалов и плохо поддается формализации.

ЛИТЕРАТУРА

1. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. – М.: Статистика, 1974. – 240 с.

2. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1983. – 472 с.

3. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Исследование зависимостей. – М.: Финансы и статистика, 1985. – 488 с.

4. *Айвазян С.А., Бухштабер В.М., Енюков И.С. и др.* Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989.

5. *Андерсон Т.* Введение в многомерный статистический анализ / Пер. с англ. – М.: ГИФМЛ, 1963. – 500 с.

6. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики: Учебник. – М.: ЮНИТИ, 1998.

7. *Бикел П., Доксам К.* Математическая статистика: Вып. 2. – М.: Финансы и статистика, 1983.

8. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. – М.: Наука, 1965. – 464 с.

9. *Болч Б., Хуань К.* Многомерные статистические методы экономики / Пер. с англ. – М.: Статистика, 1979. – 317 с.

10. *Боровков А.А.* Теория вероятностей. – М.: Наука, 1976.

11. *Ван-дер-Варден Б.Л.* Математическая статистика /Пер. с нем. – М.: Иностранная литература, 1960. – 434 с.

12. *Гнеденко Б.В.* Курс теории вероятностей. – М.: Наука, 1969.

13. *Демиденко Е.З.* Линейная и нелинейная регрессия. – М.: Финансы и статистика, 1981. – 302 с.

14. *Дубров А.М.* Обработка статистических данных методом главных компонент. – М.: Статистика, 1978. – 136 с.

15. *Дубров А.М.* Последовательный анализ в статистической обработке информации. – М.: Статистика, 1976. – 160 с.

16. *Заде Л.* Понятие лингвистической переменной и его применение к принятию приближенных решений. – М.: Мир, 1975. – 228 с.

17. *Захаров В.Н., Севастьянов Б.А., Чистяков В.П.* Теория вероятностей. – М.: Наука, 1983.

18. *Иберла К.* Факторный анализ /Пер. с нем. – М.: Статистика, 1980. – 400 с.

19. *Иванова В.М., Калинина В.Н. и др.* Математическая статистика. – М.: Высшая школа, 1981.

20. Информатика в статистике: словарь-справочник. – М.: Финансы и статистика, 1994.

21. *Кендалл М., Стьюарт А.* Статистические выводы и связи - М.: Наука, 1973.
22. *Кендалл М., Стьюарт А.* Многомерный статистический анализ и временные ряды. - М.: Наука, 1976.
23. *Колмогоров А.Н.* Основные понятия теории вероятностей. - М.: Наука, 1974. - 120 с.
24. *Кофман А.* Введение в теорию нечетких множеств. - М.: Радио и связь, 1982. - 432 с.
25. *Колемаев В.А., Староверов О.В., Турундаевский В.Б.* Теория вероятностей и математическая статистика. - М.: Высшая школа, 1990.
26. *Маленво Э.* Статистические методы эконометрии /Пер. с фр.: Вып. 1. - М.: Статистика, 1975. - 423 с.
27. *Мхитарян В.С.* Статистические методы в управления качеством продукции. - М.: Финансы и статистика, 1982.
28. ОТЭКС. Пакет прикладных программ для обработки таблиц экспериментальных данных (версия 3.0.). - Новосибирск, 1981. - 27 с.
29. Пакет программ по прикладному статистическому анализу (ППСА). - М.: ЦЭМИ АН СССР, 1983. - 187 с.
30. *Пугачев В.С.* Теория вероятностей и математическая статистика. - М.: Наука, 1979.
31. *Рао С.Р.* Линейные статистические методы и их применение /Пер. с англ. - М.: Наука, 1968. - 548 с.
32. *Смоляк С.А., Титаренко Б.П.* Устойчивые методы оценивания. - М.: Статистика, 1980. - 208 с.
33. *Тутубалин В.Н.* Теория вероятностей. - М.: МГУ, 1972. - 68 с.
34. *Уилкс С.* Математическая статистика. - М.: Наука, 1967.
35. *Феллер В.* Введение в теорию вероятностей и ее приложение: В 2-х т. - М.: Мир, 1967. - Т.1. - 499 с.; Т.2. - 752 с.
36. *Хальд А.* Математическая статистика с технологическими приложениями. - М.: Иностранная литература, 1956.
37. *Харман Г.* Современный факторный анализ /Пер. с англ. - М.: Статистика, 1972. - 486 с.
38. *Хьюбер Дж.* Робастность в статистике. - М.: Мир, 1984. - 304 с.

ОГЛАВЛЕНИЕ

Предисловие 3

Глава 1. **МНОГОМЕРНЫЕ ГЕНЕРАЛЬНАЯ И ВЫБОРОЧНАЯ СОВОКУПНОСТИ** 5
1.1. Распределение генеральной совокупности 5
1.2. Некоторые характеристики генеральной совокупности 9
1.3. Параметры связи между признаками в генеральной совокупности 12
1.4. Многомерная нормально распределенная генеральная совокупность 16
1.5. Выборка из генеральной совокупности 18

Глава 2. **СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ И СРАВНЕНИЕ МНОГОМЕРНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ** 21
2.1. Точечные оценки параметров многомерной генеральной совокупности 21
2.2. Доверительные области 25
2.2.1. Доверительная область для вектора математического ожидания 26
2.2.2. Определение совместной доверительной области для математического ожидания и дисперсии 31
2.3. Гипотезы о параметрах многомерной нормально распределенной генеральной совокупности 34
2.3.1. Сравнение вектора генеральных средних со стандартом 34
2.3.2. Сравнение двух генеральных совокупностей 38

Глава 3. **КОРРЕЛЯЦИОННЫЙ АНАЛИЗ** 43
3.1. Основные понятия 43
3.2. Двумерная модель 44
3.2.1. Точечные оценки параметров 44
3.2.2. Приемы вычисления выборочных характеристик 46
3.2.3. Проверка значимости параметров связи 48
3.2.4. Интервальные оценки параметров связи 49
3.2.5. Задачи, решаемые при помощи статистики Фишера 50
3.2.6. Корреляционный анализ затрат и оплаты труда 51
3.3. Трехмерная модель 55
3.3.1. Основные параметры модели 55
3.3.2. Оценивание и проверка значимости параметров 60
3.3.3. Корреляционный анализ показателей эффективности работы предприятий 64

Глава 3. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ
3.4. Некоторые вопросы корреляционного анализа многомерной модели ... 71
3.5. Проверка гипотезы о нормальном распределении генеральной совокупности ... 77
3.6. Задачи и упражнения ... 77
3.7. Задания для самостоятельной работы на ЭВМ по корреляционному анализу ... 83
3.8. Ранговая корреляция ... 83

Глава 4. **РЕГРЕССИОННЫЙ АНАЛИЗ** ... 91
4.1. Основные понятия ... 91
4.2. Простейшее линейное уравнение регрессии ... 96
4.2.1. Нахождение оценки уравнения ... 96
4.2.2. Определение интервальной оценки для β_0 ... 99
4.2.3. Определение интервальной оценки и проверка значимости β_1 ... 101
4.2.4. Определение интервальной оценки для условного математического ожидания ... 104
4.3. Множественное линейное уравнение регрессии ... 107
4.3.1. Нахождение и анализ свойств оценки уравнения регрессии ... 107
4.3.2. Проверка значимости уравнения регрессии ... 112
4.3.3. Доверительные интервалы для параметров линейной модели ... 114
4.3.4. Регрессионный анализ фондоотдачи ... 116
4.3.5. Взвешенный метод наименьших квадратов ... 124
4.4. Задачи и упражнения ... 128
4.5. Задания для самостоятельной работы на ЭВМ по регрессионному анализу ... 133

Глава 5. **КОМПОНЕНТНЫЙ АНАЛИЗ** ... 134
5.1. Статистический подход в методе главных компонент ... 134
5.1.1. Многомерное нормальное распределение как модель ... 135
5.1.2. Геометрическая интерпретация плотности вероятности двумерного нормального распределения ... 137
5.2. Линейная модель метода главных компонент ... 139
5.2.1. Общие положения ... 139
5.2.2. Получение матрицы весовых коэффициентов по алгоритму метода главных компонент ... 140
5.2.3. Линейный оператор и отвечающая ему матрица ... 146
5.2.4. Характеристический многочлен подобных матриц ... 150
5.2.5. Собственные векторы при одном и том же характеристическом числе ... 151
5.2.6. Характеристическая матрица и характеристический многочлен ... 152

Глава 5. КОМПОНЕНТНЫЙ АНАЛИЗ
5.2.7. Выражение присоединенной матрицы $B(\lambda)$ через характеристический многочлен $\Delta(\lambda)$ ... 155
5.2.8. Метод Фаддеева - одновременное вычисление коэффициентов характеристического многочлена и присоединенной матрицы ... 157
5.3. Квадратичные формы и главные компоненты ... 162
5.3.1. Главные компоненты трехмерного и конечно-мерного пространства ... 167
5.3.2. Дисперсия исследуемых признаков в методе главных компонент ... 170
5.4. Задачи и упражнения ... 171
5.5. Задания для самостоятельной работы на ЭВМ ... 177

Глава 6. **ФАКТОРНЫЙ АНАЛИЗ** ... 178
6.1. Основные понятия факторного анализа ... 178
6.1.1. Основные проблемы факторного анализа ... 178
6.1.2. Основная модель факторного анализа ... 181
6.1.3. Компоненты дисперсии в факторном анализе ... 185
6.1.4. Получение матрицы коэффициентов парной корреляции и ее преобразование в факторном анализе ... 192
6.1.5. Корреляционная матрица R_h с общностями на главной диагонали ... 196
6.1.6. Факторное отображение и факторная структура ... 205
6.1.7. Пространство общих факторов и полное факторное пространство ... 209
6.1.8. Связь факторных решений, полученных разными методами ... 212
6.2. Метод главных факторов и его алгоритм ... 215
6.3. Проблема вращения ... 221
6.3.1. Понятие ортогонального и косоугольного вращения ... 221
6.3.2. Вращение при помощи ортогональных матриц ... 223
6.3.3. Обобщенные факторы ... 225
6.4. Проблема оценки факторов и задачи классификации ... 228
6.4.1. Измерение факторов ... 228
6.4.2. Факторный анализ и методы классификации многомерных наблюдений ... 232
6.5. Классификация задач факторного анализа и метода главных компонент ... 236
6.6. Задачи и упражнения ... 238
6.7. Задания для самостоятельной работы на ЭВМ ... 240

Глава 7. **МЕТОДЫ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ** ... 241
7.1. Классификация без обучения. Кластерный анализ ... 241
7.1.1. Основные понятия ... 241
7.1.2. Расстояние между объектами и мера близости ... 245

Глава 7. МЕТОДЫ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ
7.1.3. Расстояние между кластерами ... 248
7.1.4. Функционалы качества разбиения ... 250
7.1.5. Иерархические кластер-процедуры ... 251
7.2. Дискриминантный анализ ... 255
7.2.1. Методы классификации с обучением ... 255
7.2.2. Линейный дискриминантный анализ ... 256
7.2.3. Дискриминантный анализ при нормальном законе распределения показателей ... 259
7.3. Задачи и упражнения ... 263
7.4. Задания для самостоятельной работы на ЭВМ по дискриминантному и кластерному анализу ... 268

Глава 8. **КАНОНИЧЕСКИЕ КОРРЕЛЯЦИИ** ... 270
8.1. Основные понятия ... 270
8.2. Канонические корреляции и канонические величины генеральной совокупности ... 272
8.3. Канонические корреляции и их интерпретация ... 275
8.4. Оценка канонических корреляций и канонических величин ... 277
8.5. Использование канонических корреляций в практике экономического анализа ... 280
8.6. Задание для самостоятельной работы ... 283

Глава 9. **РОБАСТНЫЕ МЕТОДЫ ОЦЕНИВАНИЯ** ... 285
9.1. Грубые ошибки ... 285
9.2. Устойчивые методы оценки. Основные понятия ... 286
9.3. Устойчивые оценки Хубера ... 289
9.4. Оценки, построенные при помощи порядковых статистик ... 294
9.5. Применение устойчивого оценивания при наличии асимметрии распределения ... 299
9.6. Методы обнаружения засорения (грубые ошибки) ... 301
9.6.1. Метод выявления грубых ошибок Смирнова-Граббса ... 301
9.6.2. Критерий Граббса для обнаружения одного экстремального наблюдения ... 303
9.6.3. Критерий исключения нескольких грубых ошибок ... 306
9.6.4. Критерий обнаружения экстремальных наблюдений (наибольших и наименьших) одновременно ... 307

Приложение 1 **ДИНАМИЧЕСКИЕ (ВРЕМЕННЫЕ) ФАКТОРНЫЕ МОДЕЛИ** ... 309
Приложение 2 **ВАРИАНТЫ ЗАДАНИЙ И ИСХОДНЫЕ ДАННЫЕ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ НА ЭВМ** ... 320
Приложение 3 **МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ** ... 324
Приложение 4 **БУТСТРЕП-ОЦЕНКИ** ... 342
Литература ... 345

Учебное издание

Дубров Абрам Моисеевич
Мхитарян Владимир Сергеевич
Трошин Лев Иванович

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Заведующая редакцией *Л.А. Табакова*
Редактор *Е.В. Стадниченко*
Художественный редактор *Ю.И. Артюхов*
Технический редактор *Е.В. Кузьмина*
Корректоры *Т.М. Колпакова, Н.Б. Вторушина*
Обложка художника *О.В. Толмачева*
Компьютерный оригинал-макет выполнен *О.А. Микембо*

ИБ № 3842

Подписано в печать 27.08.2003. Формат 70×100/16. Гарнитура «Таймс»
Печать офсетная. Усл. п. л. 28,6. Уч.-изд. л. 23,48
Тираж 3000 экз. Заказ 2821. «С» 202

Издательство «Финансы и статистика»
101000, Москва, ул. Покровка, 7
Телефон (095) 925-35-02, факс (095) 925-09-57
E-mail: mail@finstat.ru http://www.finstat.ru

ГУП «Великолукская городская типография»
Комитета по средствам массовой информации Псковской области.
182100, Великие Луки, ул. Полиграфистов, 78/12
Тел./факс: (811-53) 3-62-95
E-mail: VTL@MART.RU

Издательство
“ФИНАНСЫ И СТАТИСТИКА”
предлагает учебник

Теория статистики

Под редакцией Р.А. Шмойловой

**Четвертое издание,
дополненное и переработанное
656 с.: ил.**

Излагаются общие вопросы теории статистики. Рассматриваются метод группировок, расчет относительных и средних величин, показатели вариации, корреляционный и регрессионный анализ, анализ временных рядов. Четвертое издание (3-е изд. – 1999 г.) дополнено и переработано: заново написаны главы, посвященные выборочному методу и анализу частотных распределений, экономическим индексам. Учебник снабжен приложением, содержащим математико-статистические таблицы, основные формулы и понятия теории статистики, а также предметным указателем.

Для аспирантов и преподавателей вузов, магистров, учащихся спецфакультетов второго высшего образования и программ MBA, специалистов аналитических служб и менеджеров.

**Книгу можно приобрести в киоске издательства
или заказать по почте**

**Адрес: 101000, Москва, ул.Покровка, 7
(метро “Китай-город”, выход на ул.Маросейка)
Варшавское шоссе, 9
(метро "Тульская", “Нагатинская”)
Тел.: (095)925-35-02, 923-18-68
Факс (095)925-09-57
E-mail: mail@finstat.ru http://www.finstat.ru**