МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Московский государственный университет экономики, статистики и информатики
Московский международный институт эконометрики, информатики, финансов и права

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

Многомерные статистические методы и основы эконометрики

Москва 2002

УДК 519.2
ББК 22.172.6
Д 797

Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы и основы эконометрики. / Учебное пособие./ Московский государственный университет экономики, статистики и информатики. М.: МЭСИ, 2002г., 79 с.

Содержание

Введение

В условиях перехода страны к рыночной экономике возрастает интерес и потребность в познании статистических методов анализа и прогнозирования, к количественным оценкам социально-эконо-мических явлений, полученным с использованием многомерных статистических методов, реализованных на ПЭВМ.

В данном учебном пособии излагаются основные теоретические положения таких многомерных статистических методов, как корреляционный и регрессионный, компонентный и кластерный анализы, основы эконометрики.

Значительное внимание уделяется логическому анализу исходной информации и экономической интерпретации получаемых результатов. Пособие снабжено достаточным количеством экономических примеров и задач для самостоятельного решения.

Глава 1. Корреляционный анализ

1.1. Основы корреляционного анализа

Корреляционный анализ является одним из методов статистического анализа взаимозависимости нескольких признаков. Он применятся тогда, когда данные наблюдений можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону. Основная задача корреляционного анализа состоит в оценке корреляционной матрицы генеральной совокупности по выборке и определении на ее основе оценок частных и множественных коэффициентов корреляции и детерминации.

Парный (частный) коэффициент корреляции характеризует тесноту линейной зависимости между двумя переменными соответственно на фоне действия (при исключении влияния) всех остальных показателей, входящих в модель. Они изменяются в пределах от -1 до +1, причем чем, ближе коэффициент корреляции к ± 1, тем сильнее зависимость между переменными. Если коэффициент корреляции больше 0, то связь положительная, а если меньше нуля – отрицательная.

Множественный коэффициент корреляции характеризует тесноту линейной связи между одной переменной (результативной) и остальными, входящими в модель; изменяется в пределах от 0 до 1. Квадрат множественного коэффициент корреляции называется множественным коэффициентом детерминации. Он характеризует долю дисперсии одной переменной (результативной), обусловленной влиянием всех остальных (аргументов), входящих в модель.

Исходной для анализа является матрица:

$$X=\begin{pmatrix} x_{11} \cdots x_{1j} \cdots x_{1k} \\ \cdots\cdots\cdots\cdots\cdots \\ x_{i1} \cdots x_{ij} \cdots x_{ik} \\ \cdots\cdots\cdots\cdots\cdots \\ x_{n1} \cdots x_{nj} \cdots x_{nk} \end{pmatrix}$$

размерности (n x k), i-я строка которой характеризует i-е наблюдение (объект) по всем k-м показателям (j=1, 2, ..., k).

В корреляционном анализе матрицу X рассматривают как выборку объема n, из k-мерной генеральной совокупности, подчиняющейся k-мерному нормальному закону распределения.

По выборке определяют оценки параметров генеральной совокупности, а именно: вектор средних ($\bar{x}$), вектор средне-квадратических отклонений s и корреляционная матрица (R) порядка k:

$$\bar{x}=\begin{pmatrix}\overline{x_1}\\ \overline{x_2}\\ .\\ .\\ .\\ \overline{x_k}\end{pmatrix},\quad s=\begin{pmatrix}s_1\\ s_2\\ .\\ .\\ .\\ s_k\end{pmatrix},\quad R=\begin{pmatrix}1 & r_{12} & . & . & r_{1k}\\ r_{21} & 1 & . & . & r_{2k}\\ r_{k1} & r_{k2} & . & . & 1\end{pmatrix}.$$

Матрица R является симметричной ($r_{je} = r_{ej}$) и положительно определенной, где:

$$\bar{x}=\frac{1}{n}\sum x_{ij}\ ,\ s_j=\ \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}\ , \qquad (1.1)$$

$$r_{jl}=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)\left(x_{il}-\overline{x_l}\right)}{s_j s_l} \qquad (1.2)$$

x_{ij} – значение *i*-го наблюдения *j*-го фактора; r_{ie} – выборочный парный коэффициент корреляции, характеризует тесноту линейной связи между показателями x_j и x*e*. При этом r_{je} является оценкой генерального парного коэффициента корреляции.

Кроме того, находятся точечные оценки частных и множественных коэффициент корреляции любого порядка. Например, частный коэффициент корреляции (*k-2*)-го порядка между факторами X1 и X2 равен:

$$r_{12/3,4\ldots,k}=-\frac{R_{12}}{\sqrt{R_{11}R_{22}}} \qquad (1.3)$$

где R_{jl} – алгебраическое дополнение элемента rje корреляционной матрицы R. При этом, $R_{jl}=(-1)^{j+l}\times M_{jl}$, где M_{jl} – минор, определитель матрицы, получаемой из матрицы R, путем вычеркивания *j*-й строки и *l*-го столбца.

Множественный коэффициент корреляции (*k-1*)-го порядка фактора (результативного признака) X1 определяется по формуле:

$$r_{1/2,3,\ldots,k}=r_1=\sqrt{1-\frac{|R|}{R_{11}}}\ , \qquad (1.4)$$

где $|R|$ – определитель матрицы R.

Значимость частных и парных коэффициентов корреляции, т. е. гипотеза H0: ρ=0, проверяется по t-критерию Стъюдента. Наблюдаемое значение критерия находится по формуле:

$$t_{набл}=\frac{r}{\sqrt{1-r^2}}\sqrt{n-l-2} \qquad (1.5)$$

где r – соответственно оценка частного или парного, коэффициент корреляции; l – порядок частного коэффициент корреляции, т. е. число фиксируемых факторов. Для парного коэффициента корреляции l=0.

Напомним, что проверяемый коэффициент корреляции считается значимым, т. е. гипотеза H0: $\rho=0$ отвергается с вероятностью ошибки α, если tнабл по модулю будет больше, чем tкр, определяемое по таблицам t-распределение (см. приложения) для заданного α и $\nu = n - l - 2$.

Значимость коэффициентов корреляции можно также проверить с помощью таблиц Фишера-Иейтса (табл. 5 приложения).

При определении с надежностью γ доверительного интервала для значимого парного или частного коэффициентов корреляции ρ, используют Z-преобразование Фишера и предварительно устанавливают интервальную оценку для Z:

$$Z' - t_\gamma \sqrt{\frac{1}{n-l-3}} \le Z \le Z' + t_\gamma \sqrt{\frac{1}{n-l-3}} \qquad (1.6)$$

где t_γ вычисляют по таблице интегральной функции Лапласа (табл. 1 приложения) из условия:

$$\Phi(t_\gamma)=\gamma.$$

Значение Z' определяют по таблице Z-преобразования (табл. 6 приложения) по найденному значению r. Функция нечетная, т. е.:

$$Z'(-r) = -Z'(r).$$

Обратный переход от Z к ρ осуществляют также по таблице Z-преобразования, после использования которой получают интервальную оценку для ρ с надежностью γ:

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

Таким образом, с вероятностью γ гарантируется, что генеральный коэффициент корреляции ρ будет находиться в интервале (rmin, rmax).

Значимость множественного коэффициента корреляции (или его квадрата – коэффициента детерминации) проверяется по F-критерию.

Например, для множественного коэффициента корреляции проверка значимости сводится к проверке гипотезы, что генеральный множественный коэффициент корреляции равен нулю, т. е. H0: $\rho_{1/2,\dots,k}=0$, а наблюдаемое значение статистики находится по формуле:

$$F_{набл.} = \frac{\frac{1}{k-1} r^2_{1/2,\dots,k}}{\frac{1}{n-k}\left(1 - r^2_{1/2,\dots,k}\right)} \qquad (1.7)$$

Множественный коэффициент корреляции считается значимым, т. е. имеет место линейная статистическая зависимость, между X1 и остальными факторами X2,...,Xk, если: Fнабл. > Fкр. (α, *k - 1, n - k*) , где $F_{кр}$ определяется по таблице, F-распределения для заданных α, $\nu_1 = k - 1$, $\nu_2 = n - k$.

1.2. Тренировочный пример

Деятельность n = 8 карьеров характеризуется себестоимостью 1 т. песка (X1), сменной добычей песка (X2) и фондоотдачей (X3). Значения показателей представлены в таблице.

X_1 (тыс.руб)	30	20	40	35	45	25	50	30
X_2 (тыс.руб)	20	30	50	70	80	20	90	25
X_3	20	25	20	15	10	30	10	20

Требуется:

1. Оценить параметры генеральной совокупности, которая предполагается нормально распределенной;

2. При α = 0.05 проверить значимость частных коэффициентов корреляции $\rho_{12/3}$,$\rho_{13/2}$ и $\rho_{23/1}$ и при γ = 0.95, построить интервальную оценку для $\rho_{13/2}$.

3. Найти точечную оценку множественного коэффициента корреляции $\rho_{1/23}$ и при α = 0.05 проверить его значимость.

Решение:

1. Найдем значения средних арифметических (x_j) и средних квадратических отклонений (S_j) где *j*=1, 2, 3, а также парных коэффициентов корреляции r_{12}, r_{13} и r_{23} по формулам:

$$\bar{x}_1 = \frac{30+20+40+35+45+25+50+30}{8} = 34.375 \text{ тыс. руб.}$$

$$\bar{x}_2 = 48.125 \text{ т. руб.}$$

$$\bar{x}_3 = 18.75$$

S_1=9,49

$S_2 = 26{,}68$ т .руб

$S_3 = 6{,}48$

$$r_{12} = \frac{\overline{x_1 x_2}}{S_1 S_2} - \frac{\bar{x}_1 \bar{x}_2}{S_1 S_2} = \frac{1875 - 34.375 \times 48{,}125}{9.49 \times 26{,}68} = \frac{220.70}{9.49 \times 26.68} = 0.871$$

где $\overline{x_1 x_2} = \frac{1}{n}\sum_{i=1}^{n} x_{i1} x_{i2} = \frac{1}{8}(30\times 20 + 20\times 30 + 40\times 50 + \ldots + 30\times 25) = 1875$

В результате расчетов получим:

$$\overline{X} = \begin{pmatrix} 34.38 \\ 48.12 \\ 18.75 \end{pmatrix}; \quad S = \begin{pmatrix} 9.49 \\ 26.68 \\ 6.48 \end{pmatrix}; \quad R = \begin{pmatrix} 1 & 0.871 & -0.874 \\ 0.871 & 1 & -0.879 \\ -0.874 & -0.879 & 1 \end{pmatrix}$$

2. Предварительно найдем точечные оценки частных коэффициентов корреляции из выражения:

$r_{12/3} = -\frac{R_{12}}{\sqrt{R_{11}\times R_{22}}}$, где R12 – алгебраическое дополнение элемента r_{12} корреляционной матрицы R, а R11 и R22 алгебраические дополнения 1-го и 2-го диагонального элемента этой матрицы.

$$R_{12} = (-1)^3 \times \begin{vmatrix} 0.871 & -0.879 \\ -0.874 & 1 \end{vmatrix} = -0.103$$

$$R_{11} = (-1)^2 \times \begin{vmatrix} 1 & -0.879 \\ -0.879 & 1 \end{vmatrix} = 0.227$$

$$R_{22} = (-1)^4 \times \begin{vmatrix} 1 & -0.874 \\ -0.874 & 1 \end{vmatrix} = 0.236$$

$$r_{12/3} = \frac{0.103}{\sqrt{0.227\times 0.236}} = 0.445$$

Аналогично находим: $r_{13/2} = -0.462$ и $r_{23/1} = -0.494$

Для проверки значимости частных коэффициентов корреляции найдем rкр.(α = 0.05, $\nu = n - l - 2 = 5$)=0.754, где l – порядок коэффициента корреляции. В нашем примере l = 1.

Так как $|r| <$ rкр .= 0.754, то гипотезы H0: $\rho=0$ не отвергаются, т. е. предположение о равенстве его нулю не противоречит наблюдениям, но n = 8 мало.

Определим интервальную оценку для $\rho_{13/2}$ при γ =0.95. Для этого используем Z-преобразование Фишера и предварительно найдем интервальную оценку для Z из условия:

$$Z \in \left[Z' \pm t\sqrt{\frac{1}{n-l-3}} \right].$$

По таблице Z-преобразования Фишера для $r_{13/2} = -0.462$, учитывая, что Z'(–r)= – Z'(r), будем иметь Z'= – 0.497. По таблице нормального закона, из условия Ф(t) = 0.95, найдем t=1.96.

Тогда,

$$Z \in \left[-0.497 \pm 1.96\sqrt{\frac{1}{8-4}} \right],$$

откуда, $Z \in [-1.477, 0.483]$.

По таблице Z-преобразования для Zmin= – 1,477 и Zmax=0.483 найдем интервальную оценку для $\rho_{13/2}$:

$\rho_{13/2} \in [-0.9, 0.45]$.

Полученная интервальная оценка подтверждает вывод о незначимости частного коэффициента корреляции $\rho_{13/2}$, т. к. нуль находится внутри доверительного интервала.

3. Найдем точечную оценку множественного коэффициента корреляции $\rho_{13/2}$ и при α=0.05 проверим его значимость.

Точечная оценка определяется по формуле:

$r_{1/23} = \sqrt{1 - \frac{|R|}{R_{11}}}$, где $|R|$ – определитель корреляционной матрицы.

$|R|$= 1+0.871(-0.879)(-0.874)+0.871(-0.879)(-0.874) - $(0.874)^2$ - 0.871^2 - $(-0.879)^2$ - $(-0.879)^2$ =0.043

$$r_{1/23} = \sqrt{1 - \frac{0.043}{0.227}} = 0.90$$

Проверим гипотезу H0: $\rho_{1/23}$=0

$$\text{Fнабл.} = \frac{\frac{1}{2} r_{1/23}^2}{\frac{1}{n-l-1}\left(1 - r_{1/23}^2\right)} = \frac{\frac{1}{2} 0.81}{\frac{1}{5} 0.19} = 10.66,$$

где *l*=2. Критическое значение по таблице F-распределения,

Fкр.(α=0.05, ν1 =2, ν2 =5) = 5.79

Т. к. Fнабл.>Fкр., то гипотеза H0 отвергается, т. е. множественный коэффициент корреляции не равен нулю ($\rho_{1/23} \neq 0$).

Глава 2. Регрессионный анализ

2.1. Основы регрессионного анализа

Регрессионный анализ – это статистический метод исследования 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 $\tilde{Y} = \varphi(x_1,\ldots,x_k)$, являющимся функцией от аргументов xj, и с постоянной, не зависящей от аргументов дисперсией σ^2.

Для проведения регрессионного анализа из (*k+1*)-мерной генеральной совокупности (Y,X1,X2,...,X*j*,...,X*k*) берется выборка объемом *n* и каждое *i*-ое наблюдение (объект) характеризуется значениями переменных (y_i, x_{i1}, x_{i2},..., x_{ij},..., x_{ik}), где x_{ij} – значение *j*-ой переменной для i-го наблюдения (*i=1,2,...,n*), y_i – значение результативного признака для i-го наблюдения.

Наиболее часто используемая множественная линейная модель регрессионного анализа имеет вид:

$$y = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ij} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad (2.1)$$

где ε_i – случайные ошибки наблюдения, независимые между собой, имеют нулевую среднюю и дисперсию σ^2.

Отметим, что модель (2.1) справедлива для всех *i* = 1,2,.., *n* , линейна относительно неизвестных параметров β0, β1,..., βj,..., β*k* и аргументов.

Как следует из (2.1) *коэффициент регрессии* β_j *показывает, на какую величину в среднем изменится результативный признак Y, если переменную X j увеличить на единицу измерения, т. е. является нормативным коэффициентом.*

В матричной форме регрессионная модель имеет вид:

$$Y = X\beta + \varepsilon \qquad (2.2)$$

где Y – случайный вектор – столбец размерности (n x 1) наблюдаемых значений результативного признака (y*1*, y*2*,..., y*n*); X – матрица размерности [n x (k+1)] наблюдаемых значений аргументов. Элемент матрицы x_{ij} рассматривается как неслучайная величина (*i* =1,2,...,n; j=0,1,2,...k; x_{0i} =1); β-вектор – столбец размерности [(k+1)x1] неизвестных, подлежащих оценке параметров (коэффициентов регрессии) модели; ε-случайный вектор – столбец размерности (n x 1)

ошибок наблюдений (остатков). Компоненты вектора ε_i независимы между собой, имеют нормальный закон распределения с нулевым математическим ожиданием ($M\varepsilon_i = 0$) и неизвестной дисперсией σ^2 ($D\varepsilon_i = \sigma^2$).

На практике рекомендуется, чтобы n превышало k не менее, чем в три раза.

$$X=\begin{pmatrix} 1 & x_{11} & . & . & x_{1k} \\ 1 & x_{i1} & . & . & x_{ik} \\ 1 & x_{n1} & . & . & x_{nk} \end{pmatrix}; \quad Y=\begin{pmatrix} y_1 \\ y_i \\ y_n \end{pmatrix}; \quad \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_j \\ \beta_k \end{pmatrix}.$$

В модели (2.2) единицы в первом столбце матрицы призваны обеспечить наличие свободного члена в модели (2.1). Здесь предполагается, что существует переменная x_0, которая во всех наблюдениях принимает значения равные 1.

Основная задача регрессионного анализа заключается в нахождении по выборке объемом n, оценки неизвестных коэффициентов регрессии $\beta 0, \beta 1, ..., \beta k$ модели (2.1) или вектора β в (2.2).

Так как, в регрессионном анализе x_j рассматриваются как неслучайные величины, а $M\varepsilon_i = 0$, то согласно (2.1) уравнение регрессии имеет вид:

$$\tilde{y}_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_j x_{ij} + ... + \beta_k x_{ik} \qquad (2.3)$$

для всех $i = 1,2,...,n$, или в матричной форме:

$$\tilde{Y} = X\beta \qquad (2.4)$$

где $\tilde{Y}$ – вектор-столбец с элементами $\tilde{y}_1, ..., \tilde{y}_i, ..., \tilde{y}_n$.

Для оценки вектора β наиболее часто используют метод наименьших квадратов (МНК), согласно которому в качестве оценки принимают вектор b , который минимизирует сумму квадратов отклонения наблюдаемых значений y_i от модельных значений $\tilde{y}_i$, т. е. квадратичную форму:

$$Q=(Y - X\beta)^T (Y - X\beta) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \Rightarrow \min_{\beta_0, \beta_1, ..., \beta_k}$$

Наблюдаемые и модельные значения показаны на рис. 2.1

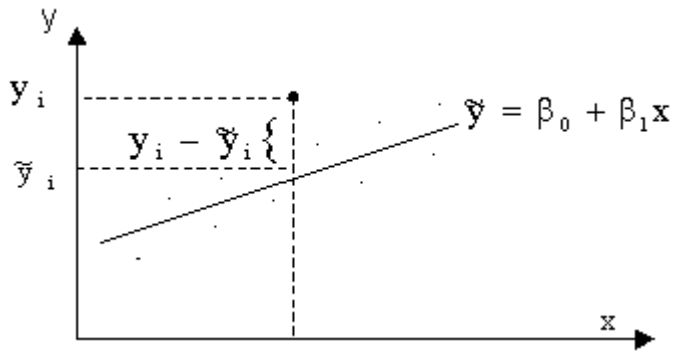


Рис. 2.1 *Наблюдаемые и модельные значения результативной величины y.*

Дифференцируя, с учетом (2.4) и (2.3) квадратичную форму Q по $\beta_0,\beta_1,...,\beta_к$ и приравнивая производные нулю, получим систему нормальных уравнений:

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_i} = 0 \\ \text{для всех } j = 0, k \end{cases}$$

решая которую и получаем вектор оценок b, где $b=(b_0\ b_1...b_k)^T$.

Согласно методу наименьших квадратов, вектор оценок коэффициентов регрессии получается по формуле:

$$b = (X^T X)^{-1} X^T Y \qquad (2.5)$$

$$b = \begin{pmatrix} b_0 \\ b_1 \\ M \\ b_j \\ M \\ b_k \end{pmatrix}$$

X^T – транспортированная матрица X;

$(X^TX)^{-1}$ – матрица, обратная матрице X^TX.

Зная вектор оценок коэффициентов регрессии b, найдем оценку $\hat{y}_i$ уравнения регрессии :

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} \qquad (2.6)$$

Или в матричном виде: $y = X\beta$

где $\hat{y} = (\hat{y}_1 \quad \hat{y}_2 ... \hat{y}_n)^T$.

Оценка ковариационной матрицы коэффициентов регрессии вектора b определяется из выражения:

$$S(b) = \hat{S}^2 (X^T X)^{-1}, \qquad (2.7)$$

где $$\hat{S}^2 = \frac{1}{n-k-1}(Y - Xb)^T (Y - Xb). \qquad (2.8)$$

Учитывая, что на главной диагонали ковариационной матрицы находятся дисперсии коэффициентов регрессии, имеем:

$$\hat{S}^2_{b_{(j-1)}} = \hat{S}^2 \left[(X^T X)^{-1} \right]_{jj} \text{ для } j=1,2,...,k,k+1 \qquad (2.9)$$

Значимость уравнения регрессии, т. е. гипотеза H_0: $\beta=0$ ($\beta_0=\beta_1=...=\beta_k=0$), проверяется по F-критерию, наблюдаемое значение которого определяется по формуле:

$$F_{набл} = \frac{Q_R / (k+1)}{Q_{ост} / (n-k-1)}, \qquad (2.10)$$

где, $$Q_R = (Xb)^T (Xb), \quad Q_{ост} = (Y - Xb)^T (Y - Xb) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \qquad (2.11)$$

По таблице F-распределения для заданных α, $\nu_1=\kappa+1$, $\nu_2=n-\kappa-1$ находят $F_{кр}$.

Гипотеза H_0 отклоняется с вероятностью α, если $F_{набл}>F_{кр}$. Из этого следует, что уравнение является значимым, т. е. хотя бы один из коэффициентов регрессии отличен от нуля.

Для проверки значимости отдельных коэффициентов регрессии, т. е. гипотез H_0: $\beta_j=0$, где j=1,2,...k, используют t-критерий и вычисляют: $t_{набл}(b_j) = b_j / \hat{S}_{b_j}$. По таблице t-распределения для заданного α и $\nu= n-k-1$, находят $t_{кр}$.

Гипотеза H_0 отвергается с вероятностью α, если $t_{набл}>t_{кр}$. Из этого следует, что соответствующий коэффициент регрессии β_j значим, т. е. $\beta_j \neq 0$. В противном случае коэффициент регрессии незначим и соответствующая переменная в модель не включается. Тогда реализуется алгоритм пошагового регрессионного анализа, состоящий в том, что исключается одна из незначимых переменных, которой соответствует минимальное по абсолютной величине значение $t_{набл}$. После этого вновь проводят регрессионный анализ с числом факторов, уменьшенным на единицу. Алгоритм заканчивается получением уравнения регрессии со значимым коэффициентами.

Существуют и другие алгоритмы пошагового регрессионного анализа, например с последовательным включением факторов.

Наряду с точечными оценками b_j генеральных коэффициентов регрессии β_j, регрессионный анализ позволяет получать и интервальные оценки последних с доверительной вероятностью γ.

Интервальная оценка с доверительной вероятностью γ для параметра β_j имеет вид:

$$b_j - t_\alpha \hat{S}_{b_j} \le \beta_j \le b_j + t_\alpha \hat{S}_{b_j}, \qquad (2.12)$$

где t_α находят по таблице t-распределения при вероятности $\alpha = 1-\gamma$ и числе степеней свободы $\nu = n-\kappa-1$.

Интервальная оценка для уравнения регрессий $\tilde{y}$ в точке, определяемой вектором начальных условий $X^0=(1,\ X_1^0,\ X_2^0,...,X_k^0)^T$, равна:

$$\tilde{y} \in \left[(X^0)^T b \pm t_\alpha \hat{S}\sqrt{(X^0)^T (X^T X)^{-1} X^0}\right]. \qquad (2.13)$$

Интервал оценки предсказания $\tilde{y}_{n+1}$ с доверительной вероятностью γ определяется как:

$$\tilde{y} \in \left[(X^0)^T b \pm t_\alpha \hat{S}\sqrt{(X^0)^T (X^T X)^{-1} X^0 + 1}\right], \qquad (2.14)$$

где t_α определяется по таблице t-распределения при $\alpha = 1-\gamma$ и $\nu = n-\kappa-1$.

По мере удаления вектора начальных условий x^0, от вектора средних $\bar{x}$, ширина доверительного интервала при заданном γ будет увеличиваться (рис. 2.2.), где $\bar{x} = (1, \bar{x}_1, ..., \bar{x}_k)$.

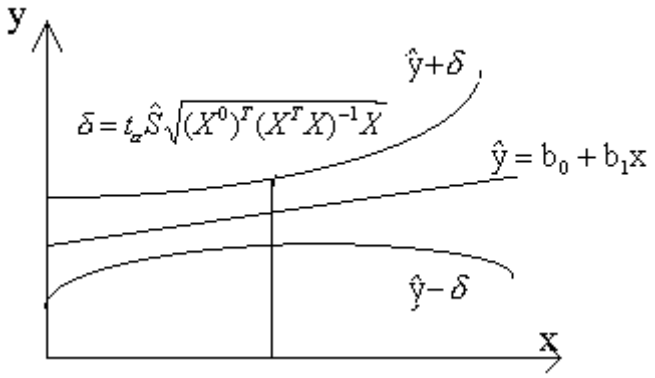


Рис. 2.2 *Точечная* $\hat{y}$ *и интервальная оценки* $\left[\hat{y} - \delta < \tilde{y} < \hat{y} + \delta\right]$ *уравнения регрессии* $\tilde{y} = \beta_0 + \beta_1 x$

Мультиколлинеарность

Одним из основных препятствий эффективного применения множественного регрессионного анализа, является мультиколлинеарность. Она связана с линейной зависимостью между аргументами $x_1, x_2, ..., x_k$. В результате мультиколлинеарности, матрица парных коэффициентов корреляции и матрица $(X^T X)$ становятся слабообусловленными, то есть их определители близки к нулю.

Это вызывает неустойчивость оценок коэффициентов регрессии (2.5), большие дисперсии $\hat{S}^2_{b_j}$ оценок этих коэффициентов (2.7), т. к. в их выражении входит обратная матрица $(X^T X)^{-1}$, получение которой связано с делением на определитель матрицы $\left|X^T X\right|$. Отсюда следуют

заниженные значения $t(b_j)$. Кроме того, мультиколлинеарность приводит к завышению значения множественного коэффициента корреляции.

На практике, о наличии мультиколлинеарности, обычно судят по матрице парных коэффициентов корреляции. Если один из элементов матрицы R больше 0.8, т. е. $|r_{je}| > 0{,}8$, то считают, что имеет место мультиколлинеарность и в уравнение регрессии следует включать только один из показателей x_j или x_e.

Чтобы избавиться от этого негативного явления, обычно используют алгоритм пошагового регрессионного анализа или строят уравнение регрессии на главных компонентах (раздел 1.3).

2.2. Пример построения регрессионного уравнения

По данным n=20 сельскохозяйственных районов требуется построить регрессионную модель урожайности на основе следующих показателей:

Y – урожайность зерновых культур (ц/га);

X_1 – число колесных тракторов (приведенной мощности) на 100 га;

X_2 – число зерноуборочных комбайнов на 100 га;

X_3 – число орудий поверхностной обработки почвы на 100 га;

X_4 – количество удобрений, расходуемых на гектар;

X_5 – количество химических средств оздоровления растений, расходуемых на гектар.

Исходные данные для анализа приведены в табл. 2.2

Таблица 2.2

Исходные данные для анализа

N п/п	Y	X_1	X_2	X_3	X_4	X_5
1	9.70	1.59	0.26	2.05	0.32	0.14
2	8.40	0.34	0.28	0.46	0.59	0.66
3	9.00	2.53	0.31	2.46	0.30	0.31
4	9.90	4.63	0.40	6.44	0.43	0.59
5	9.60	2.16	0.26	2.16	0.39	0.16
6	8.60	2.16	0.30	2.69	0.32	0.17
7	12.50	0.68	0.29	0.73	0.42	0.23
8	7.60	0.35	0.26	0.42	0.21	0.08
9	6.90	0.52	0.24	0.49	0.20	0.08
10	13.50	3.42	0.31	3.02	1.37	0.73
11	9.70	1.78	0.30	3.19	0.73	0.17
12	10.70	2.40	0.32	3.30	0.25	0.14
13	12.10	9.36	0.40	11.51	0.39	0.38
14	9.70	1.72	0.28	2.26	0.82	0.17
15	7.00	0.59	0.29	0.60	0.13	0.35
16	7.20	0.28	0.26	0.30	0.09	0.15
17	8.20	1.64	0.29	1.44	0.20	0.08
18	8.40	0.09	0.22	0.05	0.43	0.20
19	13.10	0.08	0.25	0.03	0.73	0.20
20	8.70	1.36	0.26	1.17	0.99	0.42

Решение: Предварительно, с целью анализа взаимосвязи показателей построена таблица парных коэффициентов корреляции R.

	Y	X_1	X_2	X_3	X_4	X_5
Y	1.00	0.43	0.37	0.40	0.58	0.33
X_1	0.43	1.00	0.85	0.98	0.11	0.34
X_2	0.37	0.85	1.00	0.88	0.03	0.46
X_3	0.40	0.98	0.88	1.00	0.03	0.28
X_4	0.58	0.11	0.03	0.03	1.00	0.57
X_5	0.33	0.34	0.46	0.28	0.57	1.00

Анализ матрицы парных коэффициентов корреляции показывает, что результативный показатель наиболее тесно связан с показателем X_4 – количеству удобрений, расходуемых на гектар (ryx_4=0.58).

В то же время связь между признаками-аргументами достаточно тесная. Так, существует практически функциональная связь между числом колесных тракторов (X_1) и числом орудий поверхностной обработки почвы (X_3) – rx_1x_3=0.98.

О наличии мультиколлинеарности свидетельствует также коэффициенты корреляции rx_1x_2=0.85 и rx_3x_2=0.88.

Чтобы продемонстрировать отрицательное влияние мультиколлинеарности, рассмотрим регрессионную модель урожайности, включив в нее все исходные показатели:

$$\hat{Y}=3.515 - 0.006X_1 + 15.542X_2 + 110X_3 + 4.475X_4 - 2.932X_5 \quad (2.15)$$
$$(-0.01) \quad (0.72) \quad (0.13) \quad (2.90) \quad (-0.95)$$

В скобках указаны $t_{набл}(b_j)$, расчетные значения t-критерия для проверки гипотезы о значимости коэффициента регрессии H_0: $\beta_j=0$, j=1, 2, 3, 4, 5. Критическое значение $t_{кр}$=1.76 найдено по таблице t-распределения при уровне значимости α=0.1 и числе степеней свободы ν=14. Из уравнения следует, что статистически значимым является коэффициент регрессии только при X_4, так как $|t_4|$=2.90>$t_{кр}$=1.76. Не поддаются экономической интерпретации отрицательные знаки коэффициентов регрессии при X_1 и X_5, из чего следует, что повышение насыщенности сельского хозяйства колесными тракторами (X_1) и средствами оздоровления растений (X_5) отрицательно сказывается на урожайности. Таким образом, полученное уравнение регрессии не приемлемо.

После реализации алгоритма пошагового регрессионного анализа, с исключением переменных и учетом того, что в уравнение должна войти только одна из трех тесно связанных переменных (X_1, X_2 или X_3) получаем окончательное уравнение регрессии:

$$\hat{Y} = 7.342 + 0.345X_1 + 3.294X_4 \qquad (2.16)$$
$$(11.12) \quad (2.09) \quad (3.02)$$

В уравнение (2.16) включен X_1, как определяющий из трех показателей.

Уравнение значимо при $\alpha=0.05$, т.к. $F_{набл}=266>F_{кр}=3.20$, найденного по таблице F-распределения при $\alpha=0.05$; $\nu_1=3$ и $\nu_2=17$. Значимы и все коэффициенты регрессии β_1 и β_4 в уравнении $|t_j|>t_{кр}$ ($\alpha=0.05$; $\nu=17$) = 2.11. Коэффициент регрессии β_1 следует признать значимым ($\beta_1\neq0$) из экономических соображений, при этом $t_1=2.09$ лишь незначительно меньше $t_{кр}=2.11$. При $\alpha=0.1$, $t_{кр}=1.74$ и β_1 статистически значим.

Из уравнения регрессии следует, что увеличение на 1 числа тракторов на 100 га пашни приводит к росту урожайности зерновых в среднем на 0.345 ц/га ($b_1=0.345$).

Коэффициенты эластичности $Э_1=0.068$ и $Э_4=0.161$ показывают, что при увеличении показателей X_1 и X_4 на 1% урожайность зерновых повышается соответственно на 0.068% и 0.161%, ($Э_j=b_j\frac{\overline{x_j}}{\overline{y}}$).

Множественный коэффициент детерминации $r_y^2=0.469$ свидетельствует о том, что только 46.9% вариации урожайности объясняется вошедшими в модель показателями (X_1 и X_4), то есть насыщенностью растениеводства тракторами и удобрениями. Остальная часть вариации обусловлена действием неучтенных факторов (X_2, X_3, X_5, погодных условий и др.). Средняя относительная ошибка аппроксимации $\overline{\delta}=10.5\%$ характеризует адекватность модели, также как и величина остаточной дисперсии $S^2=1.97$.

2.3. Тренировочный пример

По данным годовых отчетов, десяти (n=10) машиностроительных предприятий, провести регрессионный анализ зависимости производительности труда y (млн. руб. на чел.), от объема производства x (млрд. руб.). Предполагается линейная модель, т.е. $\tilde{y} = \beta_0 + \beta_1 x$.

Таблица 2.1

Исходная информация для анализа и результаты расчетов

N п/п (i)	y_i	x_i	$\hat{y}_i$	$e_i=y_i-\hat{y}_i$
1	2,1	3	2,77	-0,67
2	2,8	4	3,52	-0,72
3	3,2	5	4,27	-1,07
4	4,5	5	4,27	0,23
5	4,8	5	4,27	0,53
6	4,9	5	4,27	0,63
7	5,5	6	5,02	0,48
8	6,5	7	5,77	0,73
9	12,1	15	11,75	0,35
10	15,1	20	15,50	-0,4

Решение: Определим вектор оценок b коэффициентов регрессии. Согласно методу наименьших квадратов, вектор b получается из выражения:

$$b = (x^T x)^{-1} x^T y .$$

Воспользовавшись правилами умножения матриц будем иметь:

$$x^T x = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 4 & 5 & 5 & 5 & 5 & 6 & 7 & 15 & 20 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 15 \\ 1 & 20 \end{pmatrix} = \begin{pmatrix} 10 & 75 \\ 75 & 835 \end{pmatrix}$$

В матрице $(x^T x)$ число 10, лежащее на пересечении 1-й строки и 1-го столбца, получено как сумма произведений элементов 1-й строки матрицы x^T и 1-го столбца матрицы x, а число 75, лежащее на пересечении 1-й строки и 2-го столбца, как сумма произведений элементов 1-й строки матрицы x^T и 2-го столбца матрицы x и т.д.

$$x^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 4 & 5 & 5 & 5 & 5 & 6 & 7 & 15 & 20 \end{pmatrix} \begin{pmatrix} 2{,}1 \\ 2{,}8 \\ 3{,}2 \\ 4{,}5 \\ 4{,}8 \\ 4{,}9 \\ 5{,}5 \\ 6{,}5 \\ 12{,}1 \\ 15{,}1 \end{pmatrix} = \begin{pmatrix} 61{,}4 \\ 664{,}5 \end{pmatrix}$$

Найдем обратную матрицу:

$$\left(x^T x\right)^{-1} = \frac{1}{10 \cdot 835 - (75)^2} \begin{pmatrix} 835 & -75 \\ -75 & 10 \end{pmatrix} = \begin{pmatrix} 0{,}306422 & -0{,}0275229 \\ -0{,}0275227 & 0{,}0036697 \end{pmatrix},$$

Тогда вектор оценок коэффициентов регрессии равен:

$$b = \begin{pmatrix} 0{,}306422 & -0{,}0275229 \\ -0{,}0275299 & 0{,}0036697 \end{pmatrix} \cdot \begin{pmatrix} 61{,}4 \\ 664{,}5 \end{pmatrix} = \begin{pmatrix} 0{,}5253430 \\ 0{,}7486096 \end{pmatrix},$$

а оценка уравнения регрессии будет иметь вид: $\hat{y} = 0{,}52534 + 0{,}74861x$.

Перейдем к статистическому анализу полученного уравнения регрессии: проверке значимости уравнения и его коэффициентов, исследованию абсолютных $e_i = y_i - \hat{y}$ и относительных $\delta_i = \frac{y_i - \hat{y}_i}{y_i} 100\%$ ошибок аппроксимации.

Предварительно определим вектор модельных значений результативного показателя $\hat{y}$:

$$\hat{y} = xb = \begin{pmatrix} 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 15 \\ 1 & 20 \end{pmatrix} \cdot \begin{pmatrix} 0{,}5253430 \\ 0{,}7486096 \end{pmatrix} = \begin{pmatrix} 2{,}77 \\ 3{,}52 \\ 4{,}27 \\ 4{,}27 \\ 4{,}27 \\ 4{,}27 \\ 5{,}02 \\ 5{,}77 \\ 11{,}75 \\ 15{,}50 \end{pmatrix}.$$

Тогда,

$$Q_{\text{ост}} = (y - \hat{y})^T (y - \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 3{,}9847314.$$

Откуда согласно (2.8.) несмещенная оценка остаточной дисперсии равна:

$$\hat{S}^2 = \frac{1}{8} \cdot 3{,}9847314 = 0{,}49809176,$$

а оценка среднего квадратического отклонения:

$$\hat{S} = \sqrt{\hat{S}^2} = 0{,}70575616.$$

Проверим на уровне значимости α=0,05 значимость уравнения регрессии, т.е. гипотезу H_0:β=0 (β_0=β_1=0). Для этого вычисляем согласно (2.10.) величину

$$F_{\text{набл}} = \frac{\frac{1}{2} Q_R^{|}}{\frac{1}{8} Q_{\text{ост}}^{|}} = \frac{264{,}84958}{0{,}49809176} = 531{,}72849$$

По таблице F-распределения для α=0,05, $\nu_1^{|} = 2$ и $\nu_2^{|} = 8$ находим $F_{кр}$=4,46. Так как $F_{набл}$>$F_{кр}$, то уравнение является значимым.

Найдем оценку ковариационной матрицы вектора b:

$$\hat{S}(b) = \hat{s}^2 (X^T X)^{-1} = \begin{pmatrix} 0{,}306422 & -0{,}0275299 \\ -0{,}0275229 & 0{,}0036697 \end{pmatrix} \cdot 0{,}49809176 =$$

$$= \begin{pmatrix} 0{,}15262627 & -0{,}013712416 \\ -0{,}013712416 & 0{,}0018278473 \end{pmatrix}$$

Отсюда получаем несмещенные оценки дисперсий и среднеквадратических отклонений коэффициентов регрессии:

$$\hat{s}_{b_0}^2 = 0{,}15262627 \qquad \hat{s}_{b_0} = 0{,}3906741$$

$$\hat{s}_{b_1}^2 = 0{,}0018278473 \qquad \hat{s}_{b_1} = 0{,}0427527.$$

Для проверки значимости коэффициента регрессии, т.е. гипотезы H_0:β_1=0, находим по таблице t-распределения при α=0,05, ν=8 значение $t_{кр}$=2,31:

$$t_{\text{набл}}(b_1) = \frac{b_1}{\hat{S}_{b_1}} = \frac{0{,}74861}{0{,}0427527} = 17{,}5102.$$

Так как $t_{\text{набл}}(b_1^{|}) = 17{,}51$, больше $t_{кр}$=2,31, то коэффициент регрессии β_1 значимо отличается от нуля. Таким образом, окончательное уравнение регрессии имеет вид: $\hat{y} = 0{,}52534 + 0{,}74861x$.

Определим интервальные оценки коэффициентов уравнения с доверительной вероятностью γ=0,95.

Из (2.12.) следует:

$\beta_0 \in [0{,}525 \pm 2{,}31 \times 0{,}391]$, откуда $-0{,}378 \le \beta_0 \le 1{,}428$ и

$\beta_1 \in [0{,}74861 \pm 2{,}31 \times 0{,}0428]$, откуда $0{,}650 \le \beta_1 \le 0{,}847$.

Глава 3. Компонентный анализ

3.1. Основы компонентного анализа

Компонентный анализ предназначен для преобразования системы k исходных признаков, в систему k новых показателей (главных компонент). Главные компоненты не коррелированны между собой и упорядочены по величине их дисперсий, причем, первая главная компонента, имеет наибольшую дисперсию, а последняя, *k-я*, наименьшую. При этом выявляются неявные, непосредственно не измеряемые, но объективно существующие закономерности, обусловленные действием как внутренних, так и внешних причин.

Компонентный анализ является одним из основных методов факторного анализа. В задачах снижения размерности и классификации обычно используются m первых компонент ($m << k$).

При наличии результативного показателя Y может быть построено уравнение регрессии на главных компонентах.

На основании матрицы исходных данных:

$$X=\begin{pmatrix} x_{11} \dots x_{1j} \dots x_{1k} \\ x_{i1} \dots x_{ij} \dots x_{ik} \\ \dots\dots\dots\dots\dots \\ x_{n1} \dots x_{nj} \dots x_{nk} \end{pmatrix}$$

размерности (n×k) , где x_{ij} – значение *j-го* показателя у *i-го* наблюдения (i=1,2,...,n; j=1,2,...,k) вычисляют средние значения показателей $\overline{x}_1, \overline{x}_2, \dots \overline{x}_k$, а также s_1, ..., s_k и матрицу нормированных значений:

$$Z=\begin{pmatrix} z_{11} \cdots z_{1j} \cdots z_{1k} \\ \dots\dots\dots\dots\dots \\ z_{i1} \cdots z_{ij} \cdots z_{ik} \\ \dots\dots\dots\dots\dots \\ z_{n1} \cdots z_{nj} \cdots z_{nk} \end{pmatrix}$$

с элементами:

$$z_{ij}=\frac{x_{ij}-\overline{x_j}}{s_j}.$$

Рассчитывается матрица парных коэффициентов корреляции:

$$R=\frac{1}{n}Z^{T}Z \qquad (3.1)$$

с элементами:

$$r_{jl}=\frac{1}{n}\sum_{i=1}^{n} z_{ij}z_{il}=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)\left(x_{il}-\overline{x_l}\right)}{Sj\times S_e} \qquad (3.2)$$

где, j1 = 1,2, ..., k.

На главной диагонали матрицы R, т.е. при j=l,

$$r_{jj}=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)^2}{s_j^2}=1.$$

Модель компонентного анализа имеет вид:

$$Z_{ij}=\sum_{\nu=1}^{k} a_{j\nu}f_{i\nu} \qquad (3.3)$$

где:

$a_{i\nu}$ – "вес", факторная нагрузка, ν-ой главной компоненты на j-ой переменной;

$f_{i\nu}$ – значение ν-й главной компоненты для i-го наблюдения (объекта), где ν=1,2, ... ,k.

В матричной форме модель (3.3) имеет вид:

$$Z = F\ A^T \qquad (3.4)$$

где:

$$F=\begin{pmatrix} f_{11}\cdots f_{1\nu}\cdots f_{1k} \\ \cdots\cdots\cdots\cdots \\ f_{i1}\cdots f_{i\nu}\cdots f_{ik} \\ \cdots\cdots\cdots\cdots \\ f_{n1}\cdots f_{n\nu}\cdots f_{nk} \end{pmatrix}$$ – матрица значений главных компонент размерности k)

$$A=\begin{pmatrix} a_{11}\cdots a_{1\nu}\cdots a_{1k} \\ \cdots\cdots\cdots\cdots \\ a_{i1}\cdots a_{i\nu}\cdots a_{ik} \\ \cdots\cdots\cdots\cdots \\ a_{k1}\cdots a_{k\nu}\cdots a_{kk} \end{pmatrix}$$ – матрица факторных нагрузок размерности (k×k).

A^T – транспонированная матрица A;

$f_{i\nu}$ – значение ν-й главной компоненты у i-го наблюдения (объекта);

$a_{j\nu}$ – значение факторной нагрузки ν-й главной компоненты на j-й переменной.

Матрица F описывает n наблюдений в пространстве k главных компонент. При этом элементы матрицы F нормированы, то есть: $\overline{f_\nu} = \frac{1}{n}\sum_{i=1}^{n} f_{i\nu} = 0$, $S_{f_\nu}^2 = \frac{1}{n}\sum_{i=1}^{n} f_{j\nu}^2 = 1$, а главные компоненты не коррелированы между собой. Из этого следует, что,

$$(1/n)\, F^T F = E \qquad (3.5)$$

где,

$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ – единичная матрица размерности (k×k).

Выражение (3.5) может быть также представлено в виде:

$$\frac{1}{n}\sum_{i=1}^{n} f_{i\nu} f_{i\nu'} = \begin{cases} 1 & \text{при} \quad \nu = \nu' \\ 0 & \text{при} \quad \nu \neq \nu' \end{cases} \qquad (3.6)$$

$\nu, \nu' = 1,2,\ldots,k.$

С целью интерпретации элементов матрицы A, рассмотрим выражение для парного коэффициента корреляции, между Z_j-переменной и, например, f_1-й главной компонентой. Так как, z_j и f_1 нормированы, будем иметь с учетом (3.3.):

$$r_{z_j f_1} = \frac{1}{n}\sum_{i=1}^{n} z_{ij} f_{i1} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\nu=1}^{k} a_{j\nu} f_{i\nu}\right) f_{i1} = a_{j1}\frac{1}{n}\sum_{i=1}^{n} f_{i1}^2 + \sum_{\nu=2}^{k} a_{j\nu}\left(\frac{1}{n}\sum_{i=1}^{n} f_{i1} f_{i\nu}\right).$$

Принимая во внимание (3.6), окончательно получим:

$$r_{z_j f_\nu} = a_{j1}.$$

Рассуждая аналогично, можно записать в общем виде:

$$r_{z_j f_\nu} = a_{j\nu} \qquad (3.7)$$

для всех j=1,2, ... , k и ν=1,2, ... , k.

Таким образом, элемент $a_{j\nu}$ матрицы факторных нагрузок A, характеризует тесноту линейной связи между z_j-исходной переменной и f_ν-й главной компонентой, то есть $-1 \le a_{j\nu} \le +1$.

Рассмотрим теперь выражение для дисперсии z_j-й нормированной переменной. С учетом (3.3) будем иметь:

$$S_j^2 = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\nu=1}^{k} a_{j\nu} f_{i\nu}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\nu=1}^{k} a_{j\nu}^2 f_{i\nu}^2 + 2\sum_{\nu\neq\nu'} a_{j\nu} a_{j\nu'} \times f_{i\nu} f_{i\nu'}\right] =$$

$$= \sum_{\nu=1}^{k} a_{j\nu}^2\left(\frac{1}{n}\sum_{i=1}^{n} f_{i\nu}^2\right) + 2\sum_{\nu\neq\nu'} a_{j\nu} a_{j\nu'}\left(\frac{1}{n}\sum_{i=1}^{n} f_{i\nu} f_{i\nu'}\right),$$

где ν, ν'=1,2, ... , k.

Учитывая (3.6), окончательно получим:

$$S_j^2 = \sum_{\nu=1}^{k} a_{j\nu}^2 = 1. \qquad (3.8)$$

По условию переменные z_j нормированы и $s_j^2=1$. Таким образом, дисперсия z_j-й переменной согласно (3.8), представлена своими составляющими, определяющими долю вклада в нее всех k главных компонент.

Полный вклад ν-й главной компоненты в дисперсию всех k исходных признаков вычисляется по формуле:

$$\lambda_k = \sum_{j=1}^{k} a_{j\nu}^2. \qquad (3.9)$$

Одно из основополагающих условий метода главных компонент, связано с представлением корреляционной матрицы R, через матрицу факторных нагрузок A. Подставив для этого (3.4) в (3.1), будем иметь:

$$R=(1/n)Z^TZ=(1/n)(FA^T)^TFA^T=A((1/n)F^TF)A^T.$$

Учитывая (3.5), окончательно получим:

$$R = AA^T \qquad (3.10)$$

Перейдем теперь непосредственно к отысканию собственных значений и собственных векторов корреляционной матрицы R.

Из линейной алгебры известно, что для любой симметрической матрицы R, всегда существует такая ортогональная матрица U, что выполняется условие:

$$U^TRU=\Lambda\ , \qquad (3.11)$$

где,

$\Lambda = \begin{pmatrix} \lambda_1 & 0 \cdots 0 \\ 0 & \lambda_2 \cdots 0 \\ \cdots\cdots & \\ 0 & 0 \cdots \lambda_k \end{pmatrix}$ – диагональная матрица собственных значений размерности (k×k);

$U = \begin{pmatrix} u_{11} \cdots u_{1\nu} \cdots u_{1\kappa} \\ \cdots\cdots\cdots\cdots\cdots \\ u_{j1} \cdots u_{j\nu} \cdots u_{jk} \\ \cdots\cdots\cdots\cdots\cdots \\ u_{k1} \cdots u_{k\nu} \cdots u_{kk} \end{pmatrix}$ – ортогональная матрица собственных векторов размерности (k×k).

Так как матрица R положительно определена, т.е. ее главные миноры положительны, то все собственные значения положительны – $\lambda_\nu > 0$ для всех $\nu = 1,2, ..., K$.

В компонентном анализе элементы матрицы Λ ранжированы $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_\nu \geq ... \geq \lambda_\kappa > 0$. Как будет показано ниже, собственное значение $\lambda\nu$ характеризует вклад ν-й главной компоненты в суммарную дисперсию исходного признакового пространства.

Таким образом, первая главная компонента вносит наибольший вклад в суммарную дисперсию, а последняя k-я – наименьший.

В ортогональной матрице U собственных векторов, ν-й столбец является собственным вектором, соответствующим λ_ν-му значению.

Собственные значения $\lambda_1 \geq ... \geq \lambda_\nu \geq ... \geq \lambda_\kappa$ находятся как корни характеристического уравнения:

$$/\Lambda E - R/ = 0. \qquad (3.12)$$

Собственный вектор V_ν, соответствующий собственному значению λ_ν корреляционной матрицы R, определяется как отличное от нуля решение уравнения, которое следует из (3.11):

$$(\lambda_\nu E - R)V_\nu = 0. \qquad (3.13)$$

Нормированный собственный вектор $U\nu$ равен:

$$U_\nu = \frac{V_\nu}{\sqrt{V_\nu^T V_\nu}}.$$

Из условия ортогональности матрицы U следует, что $U^{-1}=U^{T}$, но тогда по определению матрицы R и Λ подобны, так как они согласно (3.11) удовлетворяют условию:

$$U^{-1}RU=\Lambda.$$

Так как следы, т.е. суммы диагональных элементов у подобных матриц равны, то:

$$tr\Lambda=tr(U^{-1}RU)=tr[R(UU^{-1})]=trR.$$

Напомним из линейной алгебры, что умножение матрицы U на обратную матрицу U^{-1}, дает единичную матрицу E. Следы матричных произведений $(U^{-1})\times(RU)$ и $R\times(UU^{-1})$ также равны.

Учитывая, что сумма диагональных элементов матрицы R равна k, будем иметь:

$$tr\Lambda=trR=k.$$

Таким образом,

$$\sum_{\nu=1}^{k}\lambda_{\nu}=k. \qquad (3.14)$$

Представим матрицу факторных нагрузок A в виде:

$$A=UA^{1/2}, \qquad (3.15)$$

а ν-й столбец матрицы A:

$$A\nu=U\nu\cdot\lambda_{\nu}^{1/2},$$

где U_{ν} – собственный вектор матрицы R, соответствующий собственному значению λ_{ν}.

Найдем норму вектора A_{ν}:

$$/A_{\nu}/^{2}=A_{\nu}^{T}A_{\nu}=\lambda_{\nu}^{1/2}U_{\nu}^{T}U_{\nu}\lambda_{\nu}^{1/2}=\lambda_{\nu}. \qquad (3.16)$$

Здесь учитывалось, что вектор U_{ν} нормированный и $U_{\nu}^{T}U_{\nu}=1$. Таким образом,

$$\lambda_\nu = \sum_{j=1}^{k} a_{j\nu}^2.$$

Сравнив полученный результат с (3.9), можно сделать вывод, что собственное значение λ_ν характеризует вклад ν-й главной компоненты в суммарную дисперсию всех исходных признаков. Из (3.15) следует:

$$A^T A = \Lambda \qquad (3.17)$$

Согласно (3.14) общий вклад всех главных компонент в суммарную дисперсию равен k. Тогда удельный вклад ν-й главной компоненты определяется по формуле:

$$\frac{\lambda_\nu}{k} 100\%.$$

Суммарный вклад m первых главных компонент определяется из выражения:

$$\frac{\sum_{\nu=1}^{m} \lambda_\nu}{k} 100\%.$$

Обычно для анализа используют m первых главных компонент, суммарный вклад которых превышает 60–70%.

Матрица факторных нагрузок A используется для экономической интерпретации главных компонент, которые представляют линейные функции исходных признаков. Для экономической интерпретации f_ν используются лишь те x_j, для которых, $/a_{j\nu}/>0{,}5$.

Значения главных компонент для каждого i-го объекта (i=1,2,...,n) задаются матрицей F.

Матрицу значений главных компонент можно получить из формулы:

$$Z = FA^T,$$

откуда,

$$F = Z(A^T)^{-1} = ZV\Lambda^{-1/2},$$

где,

$$F=\begin{pmatrix} f_{11}\cdots f_{1v}\cdots f_{1k} \\ \cdots\cdots\cdots\cdots\cdots \\ f_{i1}\cdots f_{iv}\cdots f_{ik} \\ \cdots\cdots\cdots\cdots\cdots \\ f_{n1}\cdots f_{nv}\cdots f_{nk} \end{pmatrix}$$

Z – матрица нормированных значений исходных показателей.

Уравнение регрессии на главных компонентах строится по алгоритму пошагового регрессионного анализа, где в качестве аргументов используются главные компоненты, а не исходные показатели. К достоинству последней модели следует отнести тот факт, что главные компоненты не коррелированы. При построении уравнений регрессии следует учитывать все главные компоненты.

3.2. Тренировочный пример

По данным о численности (x_1) и фонде зарплаты (x_2) пяти(n=5) строительных организаций провести компонентный анализ.

$$X=\begin{pmatrix} 3 & 4 \\ 6 & 5 \\ 8 & 9 \\ 2 & 3 \\ 7 & 6 \end{pmatrix}$$

Решение: Рассчитаем выборочные характеристики переменных x_1 и x_2:

$\bar{x}_1=5,2$ $S_1=2,315$
$\bar{x}_2=5,4$ $S_2=2,059$

Выборочный коэффициент корреляции равен:

$$r=\frac{\overline{x_1x_2}-\bar{x}_1\cdot\bar{x}_2}{S_1\cdot S_2}=\frac{32,4-5,2\cdot 5,4}{2,315\cdot 2,059}=0,906,$$

преобразуем матрицу X в матрицу нормированных значений Z, с элементами: $z_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j}$, где i=1,2,3,4,5; j=1,2.

$$Z=\begin{pmatrix} -0,950 & -0,680 \\ 0,346 & -0,194 \\ 1,210 & 1,748 \\ -1,382 & -1,166 \\ 0,778 & 0,291 \end{pmatrix}$$

Матрица парных коэффициентов корреляции имеет вид:

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

Для определения собственных значений матрицы R, рассмотрим характеристическое уравнение (3.12).

$$\left|\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}-\begin{pmatrix} 1-\lambda & 0,906 \\ 0,906 & 1-\lambda \end{pmatrix}\right|=\begin{vmatrix} 1-\lambda & 0,906 \\ 0,906 & 1-\lambda \end{vmatrix}=0$$

Отсюда следует,

$$(1-\lambda)^2-(0,906)^2=0 \text{ или } (1-\lambda)=\pm 0,906,$$

Т.к. по условию компонентного анализа $\lambda_1>\lambda_2$, то $\lambda_1=1,9062$,

$\lambda_2=0,0938$,

где λ_1, λ_2 соответственно дисперсии и вклад 1-й и 2-й главных компонент в суммарную дисперсию, равную $\lambda_1+\lambda_2=k=2$.

Относительный вклад компонент в суммарную дисперсию равен :

$$\frac{\lambda_1}{k}100\%=\frac{1,906}{2}100\%=95,3\%$$
$$\frac{\lambda_{21}}{k}100\%=\frac{1,094}{2}100\%=4,7\%$$

Таким образом,

$$\Lambda=\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}=\begin{pmatrix} 1,9062 & 0 \\ 0 & 0,09838 \end{pmatrix}.$$

Определим матрицу собственных векторов из уравнения $(R-\lambda E)V=0$.

Откуда собственный вектор V_1 находим из условия:

$$\begin{pmatrix} 1-\lambda_1 & r \\ r & 1-\lambda_1 \end{pmatrix}\begin{pmatrix} V_{11} \\ V_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

где,

$$V_1 = \begin{pmatrix} v_{11} \\ v_{21} \end{pmatrix}.$$

Подставляя полученные значения получим:

$$\begin{pmatrix} 0{,}9062 & 0{,}9062 \\ 0{,}9062 & 0{,}9062 \end{pmatrix}\begin{pmatrix} v_{12} \\ v_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Откуда, $-0{,}9062v_{11} + 0{,}9062v_{21}=0$ или $v_{11}=v_{21}=1$, т.е. $v_1=\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Нормированный собственный вектор, соответствующий λ_1, равен:

$$U_1=\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

Собственный вектор v_2 найдем решая уравнение:

$$\begin{pmatrix} 0.9062 & 0.9062 \\ 0.9062 & 0.9062 \end{pmatrix}\begin{pmatrix} v_{12} \\ v_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Откуда, $0{,}9062V_{12}+0{,}9062V_{22}=0$ или $-V_{12}=V_{22}$, $V_2=\begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

Нормированный собственный вектор, соответствующий λ_2 равен:

$$U_2=\begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix},$$

тогда нормированная матрица собственных векторов имеет вид:

$$U=\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 0{,}707 & -0{,}707 \\ 0{,}707 & 0{,}707 \end{pmatrix}.$$

Матрицу факторных нагрузок найдем по формуле:

$$A=U\Lambda^{1/2}, \text{ где } \Lambda^{1/2}=\begin{pmatrix}\sqrt{\lambda_1} & 0\\ 0 & \sqrt{\lambda_2}\end{pmatrix}.$$

Подставив полученные значения, получим:

$$A=\begin{pmatrix}0{,}707 & -0{,}707\\ 0{,}707 & 0{,}707\end{pmatrix}\times\begin{pmatrix}1{,}3807 & 0\\ 0 & 0{,}3063\end{pmatrix}=\begin{pmatrix}0{,}9763 & -0{,}2166\\ 0{,}9763 & 0{,}2166\end{pmatrix}.$$

Матрицу факторных нагрузок используют для интерпретации главных компонент, т.к. элементы матрицы $a_{jv}=r_{jv}$ характеризуют тесноту связи между x_j-м признаком и f_v главной компонентой. В нашем примере первая главная компонента тесно связана с показателями X_1 и X_2, f_1 – характеризует размер предприятия.

Матрицу значений главных компонент F можно получить по формуле:

$$F = Z(A^T)^{-1}$$

Предварительно найдем обратную матрицу $(A^T)^{-1}$

Так как,

$$A^T=\begin{pmatrix}0{,}9763 & 0{,}9763\\ 0{,}2166 & -0{,}2166\end{pmatrix},$$

то,

$$\left(A^T\right)^{-1}=\frac{-1}{0{,}4229}\begin{pmatrix}-0{,}2166 & -0{,}9763\\ -0{,}2166 & 0{,}9763\end{pmatrix}=\begin{pmatrix}0{,}5121 & -2{,}3084\\ 0{,}5121 & 2{,}3084\end{pmatrix}$$

Тогда,

$$F=\begin{pmatrix}-0{,}9503 & -0{,}6799\\ 0{,}3456 & -0{,}1943\\ 1{,}2095 & 1{,}7484\\ -1{,}3823 & -1{,}1656\\ 0{,}7775 & 0{,}2914\end{pmatrix}\times\begin{pmatrix}0.5121 & -2.3084\\ 0.5121 & 2.3084\end{pmatrix}=\begin{pmatrix}-0{,}835 & 0{,}624\\ 0{,}077 & -1{,}246\\ 1{,}515 & 1{,}244\\ -1{,}305 & 0{,}500\\ 0{,}547 & -1{,}122\end{pmatrix}.$$

Как уже отмечалось, матрица F, которую мы получили, характеризует пять строительных организаций в пространстве главных компонент. Ее можно использовать в задачах классификации и регрессионного анализа. Например, классификация организаций по

первой главной компоненте f_1, характеризующих размер предприятия, позволяет их ранжировать в порядке возрастания следующим образом: 1; 4; 2; 5; 3, что согласуется с матрицей X.

3.3. Тренировочный пример

По данным примера 1.2.3 провести компонентный анализ и построить уравнение регрессии урожайности Y на главные компоненты.

Решение: В примере 1.2.2. пошаговая процедура регрессионного анализа позволила исключить отрицательное влияние мультиколлинеарности на качество регрессионной модели, за счет значительной потери информации. Из 5 исходных показателей-аргументов в нашу, окончательную модель, вошли только два (x_1 и x_4). Более рациональным в условиях мультиколлинеарности, можно считать построение уравнения регрессии на главных компонентах, которые являются линейными функциями от всех исходных показателей и не коррелированы между собой.

Воспользовавшись методом главных компонент, найдем собственные значения, и на их основе вклад главных компонент в суммарную дисперсию исходных показателей X_1, X_2, X_3, X_4, X_5 (табл.3.1).

Таблица 3.1

Собственные значения главных компонент

Главные компоненты f_ν	f_1	f_2	f_3	f_4	f_5
Собственные значения λ_ν	3.04	1.41	0.43	0.10	0.02
Относительный вклад ν-й главной компоненты (%) в суммарную дисперсию	60.8	28.2	8.6	2.0	0.4
Относительный вклад первых главных компонент (%)	60.8	89.0	97.6	99.6	100.0

Ограничимся экономической интерпретацией двух первых главных компонент, общий вклад которых, в суммарную дисперсию составляет 89.0%. В матрице факторных нагрузок A:

$$A = \begin{bmatrix} 0{,}95^{*} & -0{,}19 & -0{,}18 & -0{,}15 & -0{,}08 \\ 0{,}94^{*} & -0{,}17 & 0{,}18 & 0{,}26 & -0{,}02 \\ 0{,}94^{*} & -0{,}28 & -0{,}15 & -0{,}06 & 0{,}09 \\ 0{,}24 & 0{,}88^{*} & -0{,}39 & 0{,}08 & 0{,}00 \\ 0{,}56 & 0{,}67^{*} & 0{,}43 & -0{,}11 & 0{,}01 \end{bmatrix},$$

звездочкой (*) указаны элементы $a_{jv}=r_{xjfv}$, учитывающиеся при интерпретации главных компонент f_v, где j, v=1, 2, ... 5.

Из матрицы факторной нагрузки A следует, что первая главная компонента наиболее тесно связана с показателями: X_1 – число колесных тракторов ($a_{11}=r_{x1f1}=0{,}95$); X_2 – число зерноуборочных комбайнов ($r_{x2f1}=0{,}97$); X_3 – число орудий поверхностной обработки почвы на 100 га ($r_{x3f1}=0{,}94$). В этой связи, первая главная компонента f_1, интерпретирована как уровень механизации работ.

Вторая главная компонента f_2, тесно связана с количеством удобрения (X_4) и средств защиты растений (X_5), расходуемых на гектар и f_2 была интерпретирована как уровень химизации растениеводства.

Уравнение регрессии на главных компонентах строится по данным вектора значений результативного показателя Y и матрицы значений главных компонент F, представленных в таблице 3.2

Таблица 3.2

Значения главных компонент

№ п\п	Y	f_1	f_2	f_3	f_4	f_5
1	9.70	–0.42	–0.52	–0.46	–0.95	0.76
2	8.40	–0.09	1.52	2.18	–0.88	1.42
3	9.00	0.27	–0.35	0.72	0.02	–1.49
4	9.90	1.98	–0.03	1.70	1.35	0.88
5	9.60	–0.29	–0.38	–0.69	–1.28	–0.33
6	8.60	0.04	–0.64	–0.13	0.47	–0.20
7	12.50	0.40	–0.01	0.37	1.24	–0.51
8	7.60	–0.89	–0.70	0.02	0.24	–0.11
9	6.90	–1.00	–0.68	–0.18	–0.95	0.18
10	13.50	1.15	2.79	–0.44	–0.31	–0.81
11	9.70	0.14	0.17	–1.33	1.59	2.01
12	10.70	0.24	–0.97	–0.03	1.22	–0.06
13	12.10	3.08	–1.35	–0.96	–1.74	0.02
14	9.70	–0.09	0.48	–1.64	1.01	0.59
15	7.00	–0.38	–0.26	1.90	–0.01	–0.27
16	7.20	–0.87	–0.74	0.78	–0.35	–0.01
17	8.20	–0.37	–0.96	–0.07	0.74	–1.82
18	8.40	–1.08	0.21	–0.25	–1.71	1.24
19	13.10	–0.80	0.73	–0.86	0.59	0.46
20	8.70	–0.24	1.70	–0.63	–0.27	–1.93

Некоррелированность главных компонент между собой и тесноту их связи с результативным показателем Y, показывает матрица парных коэффициентов корреляции (табл.3.3).

Таблица 3.3

Матрица парных коэффициентов корреляции

№ п/п	Y	f_1	f_2	f^3	f4	f_5
Y	1.00	0.48	0.34	–0.37	0.18	0.01
f_1	0.48	1.00	0.00	0.00	0.00	–0.00
f_2	0.34	0.00	1.00	–0.00	0.00	–0.00
f_3	–0.37	0.00	–0.00	1.00	0.00	–0.00
f_4	0.18	0.00	0.00	0.00	1.00	–0.00
f_5	0.01	–0.00	–0.00	–0.00	–0.00	1.00

Из матрицы парных коэффициентов корреляции следует, что Y наиболее тесно связан с первой (r_{yf2}=0.48), третьей (r_{yf3}=-0.37) и второй (r_{yf2}=0.34) главными компонентами. Можно предположить, что только эти главные компоненты войдут в регрессионную модель Y.

Первоначально в модель Y включили все главные компоненты:

$$\hat{Y} = 9.52 + 0.93f_1 + 0.66f_2 - 0.71f_3 + 0.34f_4 + 0.01f_5 \qquad (3.19)$$

(26.6) (2.59) (1.85) (-1.99) (0.95) (0.03)

В скобках указаны расчетные значения t-критерия.

Качество модели характеризует множественный коэффициент детерминации r_y^2=0,517, средняя относительная ошибка аппроксимации $\bar{\delta}$=10.4%, остаточная дисперсия S^2=1.79 и $F_{набл}$=121.

В виду того, что $F_{набл}>F_{кр}$ (α=0.05; ν_1=6; ν_2=14)=2.85, то уравнение регрессии значимо и хотя бы один из коэффициентов регрессии β_1, β_2, β_3, β_4 не равен нулю.

Если значимость уравнения регрессии (H_0: β_1= β_2= β_3= β_4=0) проверялась при α=0.05, то значимость коэффициентов регрессии, т.е. гипотезы H_0; β_j=0 (j=1, 2, 3, 4) следует проверять при уровне значимости большим 0.05. Например, при α=0.1. Тогда, $t_{кр}$ (α=0.1; ν=14)=1.76 и значимыми, как следует из уравнения (3.19) являются коэффициенты регрессии β_1, β_2, β_3.

Учитывая, что главные компоненты не коррелированы между собой, можно сразу исключить из уравнения все не значимые коэффициенты и уравнение примет вид:

$$\hat{Y} = 9.52 + 0.93f_1 + 0.66f_2 - 0.71f_3 \qquad (3.20)$$

(27.6) (2.68) (1.92) (-2.06)

Сравнив уравнения (3.19) и (3.20) видим, что исключение не значимых главных компонент f_4 и f_5 не отразилось на значениях коэффициентов уравнения $b_0=9.52$; $b_1=0.93$ и $b_2=0.66$ и соответствующих t_j (j=0, 1, 2, 3).

Это обусловлено некоррелированностью главных компонент. Здесь интересна параллель уравнений регрессии по исходным показателям (2.15), (2.16) и главным компонентам (3.19), (3.20).

Уравнение (3.20) значимо $F_{набл}=194>F_{кр}(\alpha=0.05;\ \nu_1=4;\ \nu_2=16)=3.01$. Значимы и коэффициенты уравнения, $/t_j/>t_{кр}$ ($\alpha=0.01$; $\nu=16$)=1.746 для j=0, 1, 2, 3. Коэффициент детерминации $r^2_{(y)}=0.486$ свидетельствует, что 48.6% вариации обусловлено влиянием трех первых главных компонент.

Уравнение характеризуется средней относительной ошибкой аппроксимации $\delta=9.99\%$ и остаточной дисперсией $S^2=1.91$.

Уравнение регрессии на главных компонентах (3.20) обладает несколько лучшими аппроксимирующими свойствами по сравнению с регрессионной моделью (2.16) по исходным показателям: $r^2_{y(f)}=0.486>r^2_{y(x)}=0.469$; $\bar{\delta}_{(f)}=9.99\%<\bar{\delta}_{(x)}=10.5\%$ и $S^2_{(f)}=1.91<S^2_{(x)}=1.97$. Кроме того, в уравнении (3.20) главные компоненты являются линейными функциями всех исходных показателей, в то время как в уравнение (2.16) входят только две переменные (x_1 и x_4). В ряде случаев приходится учитывать, что модель (3.20) трудно интерпретируема, т.к. в нее входит третья главная компонента f_3, которая нами не интерпретирована и вклад которой в суммарную дисперсию исходных показателей (X_1 ... X_5) всего 8.6%. Однако, исключение f_3 из уравнения (3.20), значительно ухудшает аппроксимирующие свойства модели: $r^2_{y(f)}=0.349$; $S_{(f)}=12.4\%$ и $S^2_{(f)}=2.41$. Тогда, в качестве регрессионной модели урожайности, целесообразно выбрать уравнение (2.16).

Глава 4 Кластерный анализ

4.1 Основы кластерного анализа

4.1.1 Основные понятия

В статистических исследованиях группировка первичных данных является основным приемом решения задачи классификации, а поэтому 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 не представляется возможным упорядочить классификационные признаки, 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 или компонентного анализа.

При наличии нескольких признаков (исходных или обобщенных), задача классификации может быть решена методами *кластерного анализа,* которые отличаются от других методов многомерной классификации отсутствием обучающих выборок, т.е. априорной информации о распределении генеральной совокупности, которая представляет собой вектор X.

Различия между схемами решения задачи по классификации во многом определяются тем, что понимают под понятием “сходство” и “степень сходства”.

После того как сформулирована цель работы, естественно 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, цель классифицировать оборудование 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[♣] (таксонами[♣♣], образами), методы их нахождения – кластер-анализом (соответственно численной таксономией или распознаванием образов с самообучением).

При этом, необходимо с самого начала, 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 находились друг от друга по возможности на небольшом расстоянии.

Решение другой задачи, 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.

Хотя, многие методы кластерного анализа довольно элементарны, основная часть работ, в которых они были предложены, относится к последнему десятилетию. Это объясняется тем, что эффективное решение задач поиска кластеров требует большого числа арифметических и логических операций, и поэтому стало возможным только с возникновением и развитием вычислительной техники.

Обычной формой представления исходных данных в задачах кластерного анализа служит прямоугольная таблица:

$$X=\begin{pmatrix} x_{11}\cdots & x_{1j}\cdots & x_{1k} \\ \vdots & \vdots & \vdots \\ x_{i1}\cdots & x_{ij}\cdots & x_{ik} \\ \vdots & \vdots & \vdots \\ x_{n1}\cdots & x_{nj}\cdots & x_{nk} \end{pmatrix},$$

[♣] Claster (англ.) – группа элементов, характеризуемых каким-либо общим свойством.

[♣♣] Taxon (англ.) – систематизированная группа любой категории.

каждая строка которой, представляет результат измерений k, рассматриваемых признаков на одном из обследованных объектов. В конкретных ситуациях, может представлять интерес как группировка объектов, так и группировка признаков. В тех случаях, когда разница между двумя этими задачами не существенна, например, при описании некоторых алгоритмов, мы будем пользоваться только термином “объект”, включая в это понятие и “признак”.

Матрица X не является единственным способом представления данных в задачах кластерного анализа. Иногда, исходная информация задана в виде квадратной матрицы:

$$R=(r_{ij}),\ i,j=1,\ 2,\ ...,\ k,$$

элемент r_{ij} , который определяет степень близости i-го объекта к j-му.

Большинство алгоритмов кластерного анализа 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.

Относительно проще решается вопрос об определении близости 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. Мерами близости в этом случае служат различные статистические коэффициенты связи.

4.1.2. Расстояние между объектами (кластерами) и мера близости

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов.

В общем случае, понятие однородности объектов задается либо введение правила вычисления расстояний $\rho(x_i,x_j)$ 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 к одному классу. Очевидно, что необходимо при этом сопоставлять $\rho(x_i,x_j)$ с некоторыми пороговыми значениями, определяемыми в каждом конкретном случае по-своему.

Аналогично используется и мера близости $r(x_i,x_j)$, при задании которой мы должны помнить о необходимости выполнения следующих условий: симметрии $r(x_i,x_j)= r(x_j,x_i)$; максимального сходства объекта с самим собой $r(x_i,x_i)= \max\limits_{i\,j} r(x_i,x_j)$, при $1\le i,j\le n$, и монотонного убывания

$r(x_i,x_j)$ по мере увеличения $\rho(x_i,x_j)$, т.е. из $\rho(x_k,x_l)\geq\rho(x_i,x_j)$ должно следовать неравенство $r(x_k,x_l)\leq r(x_i,x_j)$.

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 широко используемые в задачах кластерного анализа расстояния и меры близости.

Обычное Евклидово расстояние

$$\rho_E(x_i, x_j) = \sqrt{\sum_{e=1}^{k}(x_{ie} - x_{je})^2} \qquad (4.1)$$

где $x_{ie,}$, x_{je} – величина e-ой компоненты у i-го (j-го) объекта (e=1,2,...,к, i,j=1,2,...,n)

Использование этого расстояния оправдано в следующих случаях:

а) наблюдения берутся из генеральной совокупности, имеющей многомерное нормальное распределение с ковариационной матрицей вида $\sigma^2 E_к$, т.е. компоненты X взаимно независимы и имеют одну и ту же дисперсию, где $E_к$ – единичная матрица;

б) компоненты вектора наблюдений X однородны по физическому смыслу и одинаково важны для классификации;

в) признаковое пространство совпадает с геометрическим пространством.

Естественное, с геометрической точки зрения, евклидово пространство может оказаться бессмысленным (с точки зрения содержательной интерпретации), если признаки измерены в разных единицах. Чтобы исправить положение, прибегают к нормированию каждого признака путем деления центрированной величины на среднее квадратическое отклонение и переходят от матрицы X, к нормированной матрице с элементами:

$$t_{ie} = \frac{x_{ie} - \overline{x}_e}{S_e},$$

где, $\overline{X_{ie}}$ – значение e-го признака у i-го объекта;

$\overline{x_e}$ - среднее значение e-го признака;

$S_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ie} - \overline{x}_e)^2}$ – среднее квадратическое отклонение e-го признака.

Однако, эта операция может привести к нежелательным последствиям. Если, кластеры хорошо разделены по одному признаку, и не разделены по другому, то после нормирования дискриминирующие возможности первого признака, будут уменьшены в связи с увеличением "шумового" эффекта второго.

"Взвешенное" Евклидово пространство

$$\rho_{BE}(x_i, x_e) = \sqrt{\sum_{e=1}^{k} \omega_e (x_{ie} - x_{je})^2} \qquad (4.2)$$

применяется в тех случаях, когда каждой компоненте x_l вектора наблюдений X, удается приписать некоторый "вес" ω_l, пропорционально степени важности признака в задаче классификации. Обычно принимают $0 \le \omega_e \le 1$, где e=1,2,...k.

Определение "весов", как правило, связано с дополнительными исследованиями, например, организацией опроса экспертов и обработкой их мнений. Определение весов ω_l, только по данным выборки, может привести к ложным выводам.

Хеммингово расстояние

Используется как мера различия объектов, задаваемых дихотомическими признаками. Это расстояние определяется по формуле:

$$\rho_H(x_i, x_j) = \sum_{e=1}^{k} \left| x_{ie} - x_{je} \right| \qquad (4.3)$$

и равно числу несовпадений значений соответствующих признаков, в рассматриваемых i-м и j-м объектах.

В некоторых задачах классификации объектов, в качестве меры близости объектов, можно использовать некоторые физические содержательные параметры, так или иначе характеризующие взаимоотношения между объектами. Например, задачу классификации отраслей народного хозяйства, с целью агрегирования, решают на основе матрицы межотраслевого баланса [1].

В данной задаче, объектом классификации является отрасль народного хозяйства, а матрица межотраслевого баланса представлена элементами s_{ij}, характеризующими сумму годовых поставок i-ой 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-ой отрасли в j-ю заменяют долей этих поставок по отношению ко всем поставкам i-ой отрасли. Симметризацию, нормированной матрицы межотраслевого баланса можно проводить, выразив близость между i-й и j-й отраслями через среднее значение из взаимных поставок, 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, ..., $x_к$, наблюдаемых векторов X, сравнительно небольшое число наиболее существенно информативных, т.е. уменьшить размерность наблюдаемого пространства.

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” –

$$\rho_{min}(S_e, S_m) = \min_{x_i \in S_e, x_j \in S_m} \rho(x_i, x_j); \qquad (4.4)$$

– расстояние, измеряемого по принципу “дальнего соседа” –

$$\rho_{max}(S_e, S_m) = \max_{x_i \in S_e, x_j \in S_m} \rho(x_i, x_j); \qquad (4.5)$$

– расстояние, измеряемое по “центрам тяжести” групп –

$$\rho_{ц.т.}(S_e, S_m) = \rho(\bar{x}_e, \bar{x}_m); \qquad (4.6)$$

– расстояние, измеряемое по принципу “средней связи”, определяется как среднее арифметическое всех попарных расстояний между представителями рассматриваемых групп –

$$\rho_{ср}(S_e, S_m) = \frac{1}{n_e n_m} \sum_{x_i \in S_e} \sum_{x_j \in S_m} \rho(x_i, x_j). \qquad (4.7)$$

Академиком А.Н. Колмогоровым было предложено “обобщенное расстояние” между классами, которое включает в себя, в качестве частных случаев, все рассмотренные выше виды расстояний.

Расстояния между группами элементов особенно важно, в так называемых, агломеративных иерархических кластер-процедурах, так как принцип работы таких алгоритмов состоит в последовательном объединении элементов, а затем и целых групп, сначала самых близких, а затем все более и более отдаленных друг от друга.

При этом расстояние между классами s_l и $s_{(m,q)}$, являющиеся объединением двух других классов s_m и s_q, можно определить по формуле:

$$\rho_{e,(m,q)} = \rho(S_e, S_{(m,q)}) = \alpha\rho_{em} + \beta\rho_{eq} + \gamma\rho_{mq} + \delta(\rho_{em} - \rho_{eq}), \qquad (4.8)$$

где, $\rho_{em} = \rho(S_e, S_m)$; $\rho_{eq} = \rho(S_e, S_q)$ *И* $\rho_{mq} = \rho(S_m, S_q)$

– расстояния между классами s_l, s_m и s_q;

– α, β, δ и γ – числовые коэффициенты, значения которых определяют специфику процедуры, ее алгоритм.

Например, при $\alpha = \beta = -\delta = 1/2$ и $\gamma = 0$ приходим к расстоянию, построенному по принципу “ближайшего соседа”. При $\alpha = \beta = \delta = 1/2$ и $\gamma = 0$ – расстояние между классами определяется по принципу “дальнего соседа”, то есть 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 (4.11) приводит к расстоянию $\rho_{ср}$ между классами, вычисленному как среднее из расстояний между всеми парами элементов, один из которых берется из одного класса, а другой из другого.

4.1.3. Функционалы качества разбиения

Существует большое количество различных способов разбиения заданной совокупности элементов на классы. Поэтому представляет интерес, задача сравнительного анализа качества этих способов разбиения Q(S), определенного на множестве всех возможных разбиений.

Тогда, под наилучшим разбиением S^*, понимаем такое разбиение, при котором достигается экстремум выбранного функционала качества. Следует отметить, что выбор того или иного функционала качества, как правило, опирается на эмпирические соображения.

Рассмотрим некоторые наиболее распространенные функционалы качества разбиения. Пусть исследованием выбрана метрика ρ, в пространстве X и пусть $S=(s_1,s_2,...,s_p)$ – некоторое фиксированное разбиение наблюдений $x_1, ...,x_n$ на заданное число p классов $s_1,s_2,...,s_p$.

За функционал качества берут сумму (“взвешенную”) внутриклассовых дисперсий:

$$Q_1(S)=\sum_{e=1}^{p}\sum_{x_i\in S_e}\rho^2(x_i,x_j). \qquad (4.9)$$

4.1.4.Иерархические кластер-процедуры

Иерархические (древообразные) процедуры, являются наиболее распространенными (в смысле реализации на ЭВМ), алгоритмами кластерного анализа. 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 вычисления на каждом шаге матрицы 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.

В качестве примера рассмотрим агломеративный иерархический алгоритм. На первом шаге алгоритма 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, предусматривает графическое представление результатов классификации в виде дендрограммы.

4.2. Тестовый пример

Провести классификацию n=6 объектов, каждый из которых характеризуется двумя признаками.

Номер 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 этих точек на плоскости показано на рис. 4.1.

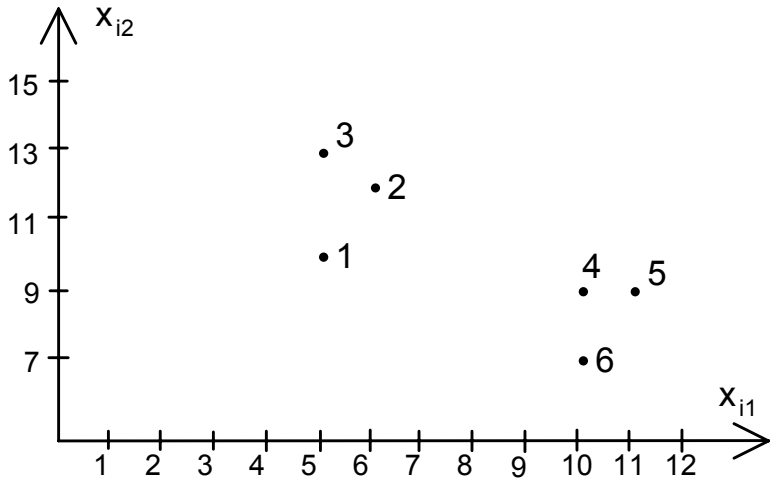


Рис. 4.1

Воспользуемся агломеративным иерархическим алгоритмом классификации. В качестве расстояния между объектами примем обычное евклидово расстояние. Тогда, согласно (4.1), расстояние между объектами 1 и 2 равно:

$$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = 2{,}24,$$

а между объектами 1 и 3 –

$$\rho_{13} = \sqrt{(5-5)^2 + (10-13)^2} = 3,$$

очевидно, что,

$$\rho_{11} = 0.$$

Аналогично находим расстояния между всеми шестью объектами и строим матрицу расстояний:

$$R_1 = \{\rho(x_i, x_j)\} = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 6{,}08 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 5{,}83 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}21 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 1 & 2 \\ 6{,}08 & 5{,}83 & 7{,}21 & 1 & 0 & 2{,}24 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2 & 2{,}24 & 0 \end{pmatrix}.$$

Из матрицы расстояний следует, что объекты 4 и 5 наиболее близки $\rho_{4,5}$=1,00 и поэтому объединяются в один кластер.

После объединения имеем пять кластеров.

Номер кластера	1	2	3	4	5
Состав кластера	(1)	(2)	(3)	(4,5)	(6)

Расстояние между кластерами будем находить по принципу “ближайшего соседа”, воспользовавшись формулой пересчета (4.11). Так, расстояние между объектом s_1 и кластером $s_{(4,5)}$, равно:

$$\rho_{1,(4,5)} = \rho(S_1, S_{(4,5)}) = \frac{1}{2}\rho_{14} + \frac{1}{2}\rho_{15} - \frac{1}{2}|\rho_{14} - \rho_{15}| = \frac{1}{2}(5{,}10 + 6{,}08) -$$
$$- \frac{1}{2}(|5{,}10 - 6{,}08|) = 5{,}10.$$

Мы видим, что 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 равна:

$$R_2 = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 2 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2 & 0 \end{pmatrix}.$$

Объединим объекты 2 и 3, имеющие наименьшее расстояние $\rho_{2,3}$=1,41. После объединения имеем четыре кластера:

$$s_{(1)}, s_{(2,3)}, \quad s_{(4,5)}, \quad s_{(6)}.$$

Вновь найдем матрицу расстояний. Для этого необходимо рассчитать расстояние до кластера $s_{(2,3)}$. Для этого воспользуемся матрицей расстояний R_2.

Например, расстояние между кластерами

$s_{(4,5)}$ и $s_{(2,3)}$ равно:

$$\rho_{(4,5),(2,3)} = \frac{1}{2}\rho_{(4,5),2} + \frac{1}{2}\rho_{(4,5),3} - \frac{1}{2}\left|\rho_{(4,5),2} - \rho_{(4,5),3}\right| = \frac{5}{2} + \frac{6{,}40}{2} - \frac{1{,}40}{2} = 5 .$$

Проведя аналогичные расчеты, получим:

$$R_3 = \begin{pmatrix} 0 & 2{,}24 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 5 & 6{,}40 \\ 5{,}10 & 5 & 0 & 2 \\ 5{,}83 & 6{,}40 & 2 & 0 \end{pmatrix}.$$

Объединенные кластеры $s_{(4,5)}$ и $s_{(6)}$, расстояние между которыми, согласно матрице R_3, наименьшее: $\rho_{(4,5),6}=2$.

В результате этого получим три кластера: s_1, $s_{(2,3)}$ и $s_{(4,5,6)}$.

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 равно $\rho_{1,(2,3)} = 2{,}24$. В результате получим два кластера: $s_{(1,2,3)}$ и $s_{(4,5,6)}$, расстояние между которыми, найденное по принципу “ближайшего соседа”, равно:

$$\rho_{(1,2,3);(4,5,6)} = 5 .$$

Результаты иерархической классификации объектов представлены на рис. 4.2 в виде дендрограммы.

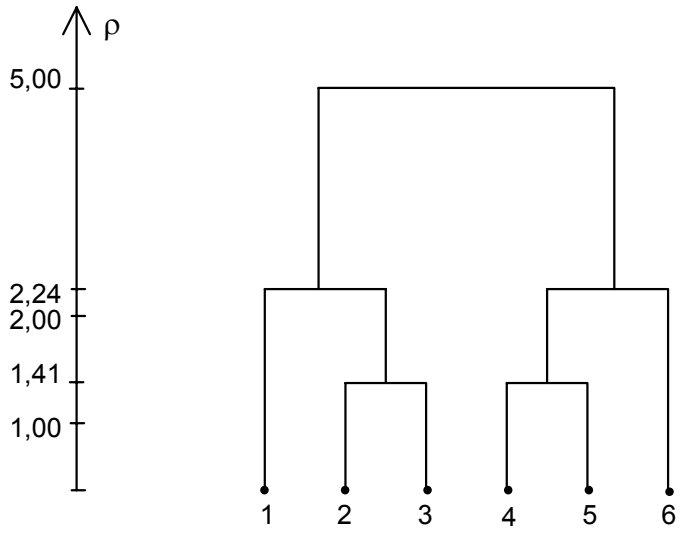


Рис. 4.2. *Дендрограмма*

Слева на рисунке приводится расстояние между объединяемыми на данном этапе кластерами (объектами).

В задаче предпочтение следует отдать предпоследнему этапу классификации, когда все объекты объединены в два кластера:

$$S_{(1,2,3)} \quad \text{и} \quad S_{(4,5,6)},$$

что наглядно видно на рис. 4.1.

Глава 5. Основы эконометрики

5.1. Основные понятия эконометрики

Эконометрика – это дисциплина, объединяющая совокупность теоретических результатов, методов и приемов, позволяющих на базе экономической теории, экономической статистики и математико-статистического инструментария получать количественное выражение качественным закономерностям. Курс эконометрики призван научить различным способам выражения связей и закономерностей, через эконометрические модели и методы проверки их адекватности, основанные на данных наблюдений. От математико-статистического, эконометрический подход, отличается тем вниманием, которое уделяется в нем вопросу соответствия выбранной модели изучаемому объекту, рассмотрению причин, приводящих к необходимости пересмотра модели на основе более точной системы представлений. Эконометрика занимается, по существу, статистическими выводами, т. е. использованием выборочной информации для получения некоторого представления о свойствах генеральной совокупности. Наиболее распространенными эконометрическими моделями, являются производственные функции и модели, описываемые системой одновременных уравнений. Кратко остановимся на них.

5.1.1. Производственные функции

Производственная функция представляет собой математическую модель, характеризующую зависимость объема выпускаемой продукции, от объема трудовых и материальных затрат. При этом, модель может быть построена как, для отдельной фирмы и отрасли, так и всей национальной экономики. Рассмотрим производственную функцию, включающую два фактора производства: затраты капитала (K) и трудовые затраты (L), определяющих объем выпуска Q. Тогда можно записать:

$$Q=f(K,L).$$

Определенного уровня выпуска можно достигнуть с помощью различного сочетания капитальных и трудовых затрат, а кривые, описываемые условиями f(K,L)=const, обычно называют изоквантами. Предполагается, что по мере роста значений одной из независимых переменных, предельная норма замещения данного фактора производства уменьшается. Поэтому, при сохранении постоянного объема производства, экономия одного вида затрат, связанная с увеличением затрат другого фактора, постепенно уменьшается. На примере производственной функции Кобба-Дугласа, рассмотрим основные выводы, которые можно получить, исходя из предположений о

том или ином виде производственной функции. Производственная функция Кобба-Дугласа, включающая два фактора производства, имеет вид:

$$Q = A \times K^{\alpha} \times L^{\beta} \qquad (5.1),$$

где A, α, β – параметры модели. Величина A зависит от единиц измерения Q, K и L, а также от эффективности производственного процесса.

При фиксированных значениях K и L функции, характеризующейся большей величиной параметра A, соответствует большее значение Q, следовательно, и производственный процесс, описываемый такой функцией, более эффективен. Описываемая функция однозначна и непрерывна (при положительных K и L). Параметры α и β называют коэффициентами эластичности. Они показывают, на какую величину в среднем изменится Q, если α или β увеличить соответственно на один процент. Рассмотрим поведение функции при изменении масштабов производства. Предположим для этого, что затраты каждого фактора производства увеличились в C раз. Тогда, новое значение будет определяться следующим образом:

$$Q_1 = A \times (C \times K)^{\alpha} \times (C \times L)^{\beta} = C^{\alpha+\beta} \times Q \qquad (5.2)$$

При этом, если $\alpha + \beta = 1$, то уровень эффективности не зависит от масштабов производства. Если $\alpha + \beta < 1$, то средние издержки, рассчитанные на единицу продукции, растут, а при $\alpha + \beta > 1$ – убывают по мере расширения масштабов производства. Следует отметить, что эти свойства не зависят от численных значений K, L и сохраняют силу в любой точке производственной функции. Для определения параметров и вида производственной функции, необходимо провести дополнительные наблюдения. Как правило, пользуются двумя видами данных – динамическими рядами и данными одновременных наблюдений (пространственной информацией). Динамические ряды данных характеризуют поведение одной и той же фирмы во времени, тогда как, данные второго вида, обычно относятся к одному и тому же моменту, но к различным фирмам. В случаях, когда исследователь располагает временным рядом, например, годовыми данными, характеризующими деятельность одной и той же фирмы, возникают трудности, с которыми не пришлось бы столкнуться при работе с пространственными данными. Так, относительные цены со временем становятся иными, а следовательно, меняется и оптимальное сочетание затрат отдельных факторов производства. Кроме того, с течением времени меняется и уровень административного управления. Однако, основные проблемы при использовании временных рядов, порождают последствия технического процесса, в результате которого меняются нормы затрат производственных факторов, соотношения, в которых они могут

замещать друг друга, и параметры эффективности. Отсюда, с течением времени могут меняться не только параметры, но и формы производственной функции. Технический прогресс может быть учтен в форме некоторого временного тренда, включаемого в состав производственной функции. Тогда,

$$Q_t = \varphi(K_t, L_t, t).$$

Производственная функция Кобба-Дугласа с учетом технического прогресса имеет вид:

$$Q_t = A \times e^{\theta \cdot t} \times K_t^{\alpha} \times L_t^{\beta} \qquad (5.3)$$

В этом выражении параметр θ, с помощью которого характеризуется технический прогресс, показывает, что объем выпускаемой продукции ежегодно увеличивается на θ процентов, независимо от изменений в затратах производственных факторов, и, в частности, от размера новых инвестиций. Такая форма технического прогресса, не связанная с какими-либо затратами труда или капитала, называется “нематеризованным техническим прогрессом”. Однако подобный подход не вполне реалистичен, т. к. новые открытия не могут повлиять на функционирование старых машин, а расширение объема производства возможно только посредством новых инвестиций. При другом подходе к учету технического прогресса, для каждой возрастной группы капитала, строят свою производственную функцию. В этом случае функция Кобба-Дугласа будет иметь вид:

$$Q_t(\nu) = Ae^{\theta \times \nu} \times K_t^{\alpha}(\nu) \times L_t^{\beta}(\nu), \qquad (5.4)$$

где $Q_t(\nu)$ – объем продукции, произведенной в период t на оборудовании, введенном в строй в период ν; $L_t(\nu)$ – труд, занятый в период t обслуживанием оборудования, введенного в строй в период ν, и $K_t(\nu)$ – основной капитал, введенный в строй в период ν и использованный в период t. Параметр ν, в такой производственной функции, отражает состояние технического прогресса. Затем, для периода t строится агрегированная производственная функция, представляющая собой зависимость совокупного объема выпускаемой продукции Q_t, от общих затрат труда L_t и капитала K_t, на момент t. При использовании для построения производственной функции пространственной информации, т. е. данных нескольких фирм, относящихся к одному и тому же времени, возникают проблемы другого рода. Так как наблюдения относятся к разным фирмам, то при их использовании предполагается, что поведение всех фирм может быть описано с помощью одной и той же функции. Для успешной экономической интерпретации полученной модели желательно, чтобы все эти фирмы принадлежали одной и той же отрасли. Кроме того, предполагается, что они располагают примерно одинаковыми производственными возможностями и уровнями административного управления. Рассмотренные выше производственные функции, носили

детерминированный характер и не учитывали влияние случайных возмущений, присущих каждому экономическому явлению. Поэтому, в каждое уравнение, параметры которого предстоит оценить, необходимо ввести еще случайную переменную ε , которая будет отражать воздействие на процесс производства всех тех факторов, которые не вошли в состав производственной функции в явном виде. Таким образом, в общем виде производственную функцию Кобба-Дугласа можно представить как:

$$Q = A \times K^{\alpha} \times L^{\beta} \times e^{\varepsilon} \qquad (5.5)$$

Мы получили степенную регрессионную модель, оценки параметров которой A, α и β можно найти с помощью метода наименьших квадратов, лишь прибегнув предварительно к логарифмическому преобразованию. Тогда для i-го наблюдения имеем:

$$\ln Q_i = \ln A + \alpha \times \ln K_i + \beta \times \ln L_i + \varepsilon_i \ , \qquad (5.6)$$

где, Q_i, K_i и L_i – соответственно объемы выпуска, капитальных и трудовых затрат для i-го наблюдения (i=1,2,...,n), а n – объем выборки,

число наблюдений, используемых для получения оценок $\ln \hat{A}$, $\hat{\alpha}$ и $\hat{\beta}$ параметров производственной функции. Относительно ε_i обычно предполагается, что они взаимно независимы между собой и $\varepsilon_i \in N(0,\sigma)$. Исходя из априорных соображений, значения α и β должны удовлетворять условиям: $0<\alpha<1$ и $0<\beta<1$. Если предположить, что с изменением масштабов производства уровень эффективности остается постоянным, то, приняв $\beta = 1-\alpha$, имеем:

$$Q = A \times K^{\alpha} \times L^{1-\alpha} \times e^{\varepsilon} = A \times (\tfrac{K}{L})^{\alpha} \times L \times e^{\varepsilon}, \qquad (5.7)$$

или

$$\tfrac{Q}{L} = A \times (\tfrac{K}{L})^{\alpha} \times e^{\varepsilon}$$

и

$$\ln(\tfrac{Q}{L}) = \ln A + \alpha \times \ln(\tfrac{K}{L}) + \varepsilon \qquad (5.8)$$

Прибегнув к такой форме выражения производственной функции, можно устранить влияние мультиколлинеарности между lnK и lnL. В качестве примера, приведем полученную на основе данных о 180 предприятий, выпускающих верхнюю одежду, модель Кобба-Дугласа:

$$\ln(\tfrac{Q}{L}) = \underset{}{1{,}43} + \underset{(4{,}67)}{0{,}14} \ln L + \underset{(3{,}80)}{0{,}19} \ln(\tfrac{K}{L})$$

В скобках указаны значения t-критерия для коэффициентов регрессии уравнения. При этом множественный коэффициент детерминации и расчетное значение статистики F-критерия, соответственно равны: $r^2 = 0{,}16$ и F=12,7. Расчетное значение F указывает на то, что полученное значение не носит случайный характер. Оценки параметров α и β функции Кобба-Дугласа соответственно равны: $\hat{\alpha} = 0{,}19$ и $\hat{\beta} = 0{,}95$ (1–0,19+0,14). Так как $\hat{\alpha} + \hat{\beta} = 1{,}14 > 1$, то можно

предположить некоторое повышение эффективности по мере расширения масштабов производства. Параметры модели показывают также, что при увеличении капитала K на 1%, объем выпуска увеличивается в среднем на 0,19%, а при увеличении трудовых затрат L на 1%, объем выпуска в среднем увеличится на 0,95%.

5.1.2. Система одновременных эконометрических уравнений

Систему взаимосвязанных тождеств и регрессионных уравнений, в которой переменные могут одновременно выступать, как результирующие в одних уравнениях, и как объясняющие в других, принято называть системой одновременных (эконометрических) уравнений. При этом, в соотношения могут входить переменные, относящиеся не только к моменту времени t , но и к предшествующим моментам. Такие переменные называются лаговыми (запаздывающими). Тождества относятся к функциональной связи переменных и вытекают из содержательного смысла этих переменных. Техника оценивания параметров системы эконометрических уравнений имеет свои особенности. Это связано с тем, что в регрессионных уравнениях системы независимых переменных и случайных погрешностей оказываются коррелированы между собой. Достаточно хорошо изучены статистические свойства и вопросы оценивания систем линейных уравнений. Будем рассматривать линейную модель вида:

$$\beta_{i1}y_{1t} + \beta_{i2}y_{2t} + \ldots + \beta_{iG}y_{Gt} + \gamma_{i1}x_{1t} + \ldots + \gamma_{ik}x_{kt} = u_{it} \ , \qquad (5.9)$$

где, t=1,2,...,n ; i=1,2,...,G;

y_{it} – значение эндогенной (результирующей) переменной в момент времени t ;

x_{jt} – значение предопределенной переменной, т. е. экзогенной (объясняю-щей) переменной в момент t, или лаговой эндогенной переменной;

u_{it} – случайные возмущения, имеющие нулевые средние.

Совокупность равенств (5.9) называется системой одновременных уравнений в структурной форме. Наличие априорных ограничений, связанных, например, с тем, что часть коэффициентов считаются равными нулю, обеспечивает возможность статистического оценивания оставшихся. В матричном виде систему уравнений можно представить как:

$$By_t + \Gamma x_t = \varepsilon_t \ , \qquad (5.10)$$

где, B – матрица порядка $G \times G$, состоящая из коэффициентов при текущих значениях эндогенных переменных;

Γ – матрица порядка $G \times K$, состоящая из коэффициентов экзогенных переменных.

$y_t = (y_{1t},\ldots,y_{Gt})^T$; $x_t = (x_{1t},\ldots,x_{Kt})^T$; $\varepsilon_t = (\varepsilon_{1t},\ldots,\varepsilon_{Gt})^T$ – векторы-столбцы значений соответственно эндогенных и экзогенных переменных, случайных ошибок. При этом, $M\varepsilon_t = 0$; $\Sigma_{(\varepsilon)} = M\varepsilon_t\varepsilon_t^T = \sigma_t^2 E_n$, где E_n –

единичная матрица. Таким образом, если $M\varepsilon_{t_1}\varepsilon_{t_2}=0$ при $t_1 \neq t_2$ и $t_1, t_2 = 1,2,\ldots,n$, то случайные ошибки независимы между собой. Если дисперсия ошибки постоянна, $M\varepsilon_t^2 = \sigma_t^2 = \sigma^2$ и не зависит от t и x_t , то это свидетельствует о гомоскедастичности остатков.

Условием гетероскедастичности является зависимость значений $M\varepsilon_t^2 = \sigma_t^2$ от t и x_t. Умножив все элементы уравнения (5.10) слева на обратную матрицу B^{-1}, получим приведенную форму системы одновременных уравнений:

$$y_t = B^{-1}\Gamma x_t + B^{-1}\varepsilon_t \qquad (5.11)$$

Среди систем одновременных уравнений наиболее простыми являются рекурсивные системы, для оценивания коэффициентов которых, можно использовать метод наименьших квадратов. Систему (5.10) одновременных уравнений называют рекурсивной, если выполняются следующие условия:

1) Матрица значений эндогенных переменных:

$$B=\begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ \beta_{21} & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \beta_{i1} & \beta_{i2} & \cdots & \beta_{ij} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \beta_{G1} & \beta_{G2} & \cdots & \beta_{Gj} & \cdots & 1 \end{pmatrix},$$

является нижней треугольной матрицей, т. е. $\beta_{ij}=0$ при j>i и $\beta_{ii}=1$;

2) случайные ошибки независимы между собой, т. е. $\sigma_{ii}>0, \sigma_{ij}=0$, при $i \neq j$, где i,j=1,2,...,G. Отсюда следует, что ковариационная матрица ошибок $M\varepsilon_t\varepsilon_t^T = \Sigma_{(\varepsilon)}$ диагональна;

3) каждое ограничение на структурные коэффициенты относится к отдельному уравнению. Процедура оценивания коэффициентов рекурсивной системы, с помощью метода наименьших квадратов, примененного к отдельному уравнению, приводит к состоятельным оценкам. В качестве примера рассмотрим ситуацию, которая приводит к рекурсивной системе уравнений. Предположим, что цены на рынке P_t в день t, зависят от объема продаж в предыдущий день q_{t-1} , а объем покупок q_t в день t, зависит от цены товара в день t . Математически систему уравнений можно представить в виде:

$$P_t = \alpha_0 + \alpha_1 q_{t-1} + \varepsilon_t \ ,$$
$$q_t = \beta_0 + \beta_1 P_t + \xi_t \quad .$$

Случайные возмущения ε_t и ξ_t можно считать независимыми. Мы получили рекурсивную систему двух уравнений, причем в правую

часть первого уравнения входит предопределенная переменная q_{t-1}, а второго – эндогенная переменная P_t .

Применение метода наименьших квадратов, для получения оценок системы одновременных уравнений, приводит к смещенным и несостоятельным оценкам, поэтому область его применения ограничена рекурсивными системами. Для оценивания систем одновременных уравнений, в настоящее время, наиболее часто используют двухшаговый метод наименьших квадратов, применяемый к каждому уравнению системы в отдельности, и трехшаговый метод наименьших квадратов, предназначенный для оценивания всей системы в целом. Двухшаговый метод наименьших квадратов (2 МНК) применяют для оценки отдельного уравнения системы одновременных уравнений. Сущность этого метода состоит в том, что для оценивания параметров структурного уравнения, метод наименьших квадратов применяют в два этапа. Он дает состоятельные, но в общем случае смещенные оценки коэффициентов уравнения, является достаточно простым с теоретической точки зрения и удобным для вычисления. Запишем исходное i -е структурное уравнение системы в виде:

$$y_i = Y_i\beta_i + X_i\gamma_i + \varepsilon_i$$

где y_i – вектор n наблюдений над i -й эндогенной переменной;

Y_i – матрица порядка $(n \times q_i)$ значений эндогенных переменных, входящих в i-е уравнение (кроме y_i -й);

β_i – вектор размерности $(q_i \times 1)$ значений структурных коэффициентов эндогенных переменных из матрицы Y_i ;

X_i – матрица порядка $(n \times k_i)$ значений экзогенных переменных, входящих в уравнение;

γ_i – вектор размерности $(k_i \times 1)$ коэффициентов, относящихся к переменным X_i;

ε_i – вектор случайных возмущений, имеющий размерность $(n \times 1)$, причем $M\varepsilon_i = 0$; $\Sigma_{(\varepsilon)} = \sigma_i^2 E_n$.

Непосредственно применить в данном случае метод наименьших квадратов нельзя, так как эндогенные переменные, содержащиеся в матрице Y_i коррелированы со случайными составляющими ε_i .

В этой связи представим эндогенные переменные Y_i , входящие в уравнение, как функцию всех содержащихся в модели экзогенных переменных (X). Найдем оценку $\hat{Y}_i$ матрицы Y_i , которая согласно методу наименьших квадратов определяется из выражения:

$$\hat{Y}_i = X_i(X_i^T X_i)^{-1} X_i^T Y_i \quad .$$

Тогда,

$Y_i = \hat{Y}_i + \hat{U}$, где $\hat{U}$ – матрица оценок остаточных величин преобразованной системы. Исходное структурное уравнение может быть преобразовано к виду:

$$y_i = \hat{Y}_i\beta_i + X_i\gamma + \nu_i \quad ,$$

где,

$$\nu_i = \varepsilon_i + \hat{U}\beta_i \quad .$$

Применяя метод наименьших квадратов, для нахождения оценок параметров вновь полученного уравнения регрессии, будем иметь:

$$d = \begin{pmatrix} \hat{\beta}_i \\ \hat{\gamma}_i \end{pmatrix} = \begin{pmatrix} \hat{Y}_i^T\hat{Y}_i & \hat{Y}_i^T X_i \\ X_i^T\hat{Y}_i & X_i^T X_i \end{pmatrix}^{-1} \begin{pmatrix} \hat{Y}_i^T y_i \\ X_i^T y_i \end{pmatrix},$$

где, d – вектор оценок коэффициентов размерности $((q_i + k_i)\times 1)$. Перейдя к исходным переменным, получим:

$$d = \begin{pmatrix} \hat{\beta}_i \\ \hat{\gamma}_i \end{pmatrix} = \begin{pmatrix} Y_i^T X_i (X_i^T X_i)^{-1} X_i^T Y_i & Y_i^T X_i \\ X_i^T Y_i & X_i^T X_i \end{pmatrix} \begin{pmatrix} Y_i^T X_i (X_i^T X)^{-1} X_i^T y_i \\ X_i^T y_i \end{pmatrix} \quad .$$

Полученная оценка и носит название оценки двухшагового метода наименьших квадратов параметров β и γ .

Таким образом, двухшаговый метод наименьших квадратов, состоит в замене матрицы Y_i расчетной матрицей $\hat{Y}_i$, после чего оцениваются коэффициенты обыкновенного уравнения регрессии y_i на $\hat{Y}_i$ и X_i. Согласно алгоритму трехшагового метода наименьших квадратов, первоначально с целью оценки коэффициентов каждого структурного уравнения, применяют двухшаговый метод наименьших квадратов, а затем определяют оценку для ковариационной матрицы случайных возмущений. После этого, с целью оценивания коэффициентов всей системы, применяется обобщенный метод наименьших квадратов. Рассмотрим систему одновременных уравнений, содержащую G эндогенных и K экзогенных переменных, принимаемых как неслучайные. Преобразуем i-е уравнение (5.11) к виду:

$$y_i = Z_i\delta_i + \varepsilon_i \quad ,$$

где,

$$Z_i = (Y_i X_i) \ , \quad \delta_i = \begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix} \ .$$

Умножив левую и правую части уравнения слева, на транспонированную матрицу X^T, значений всех экзогенных переменных модели, получим:

$$X^T y_i = X^T Z_i\delta_i + X^T\varepsilon_i \quad .$$

Записав таким образом все уравнения системы, получим:

$$\begin{pmatrix} X^T y_1 \\ \vdots \\ X^T y_i \\ \vdots \\ X^T y_G \end{pmatrix} = \begin{pmatrix} X^T Z_1 & 0 & \cdots & \cdots & 0 \\ 0 & X^T Z_2 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & X^T Z_G \end{pmatrix} \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_i \\ \vdots \\ \delta_G \end{pmatrix} + \begin{pmatrix} X^T \varepsilon_1 \\ \vdots \\ X^T \varepsilon_i \\ \vdots \\ X^T \varepsilon_G \end{pmatrix} \quad .$$

Для применения обобщенного метода наименьших квадратов, построим ковариационную матрицу вектора возмущений:

$$\Sigma_{(U)} = \begin{pmatrix} \sigma_{11} X^T X & \cdots & \sigma_{1i} X^T X & \cdots & \sigma_{1G} X^T X \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{i1} X^T X & \cdots & \sigma_{ii} X^T X & \cdots & \sigma_{2G} X^T X \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{G1} X^T X & \cdots & \sigma_{Gi} X^T X & \cdots & \sigma_{GG} X^T X \end{pmatrix} = \Sigma \otimes X^T X \quad .$$

Заменив матрицу $\Sigma = (\sigma_{ij})$ ее оценкой $S = (s_{ij})$, получим оценку ковариационной матрицы вектора возмущений –

$$S_{(U)} = S \otimes X^T X$$

и соответствующую обратную матрицу –

$$S^{-1}_{(U)} = S^{-1} \otimes (X^T X)^{-1} \quad .$$

Тогда, искомая оценка трехшагового метода наименьших квадратов, имеет вид:

$$\hat{\delta} = (A^T S^{-1}_{(U)} A)^{-1} A^T S^{-1}_{(U)} Z \quad ,$$

где,

$$A = \begin{pmatrix} X^T Z_1 & 0 & \cdots & 0 \\ 0 & X^T Z_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & X^T Z_G \end{pmatrix} \quad ; \quad Z = \begin{pmatrix} X^T y_1 \\ \vdots \\ X^T y_i \\ \vdots \\ X^T y_G \end{pmatrix} \quad .$$

В случае, когда матрица Σ не является диагональной, т. е. когда возмущения, входящие в различные структурные уравнения, зависимы, трехшаговая процедура имеет лучшую асимптотическую эффективность по сравнению с двухшаговой.

5.2. Тренировочный пример

Построение эконометрической модели мирового рынка нефти.

Очевидно, что модель должна отражать взаимосвязь между тремя основными элементами рыночного механизма – спросом, ценой и предложением (эндогенными переменными). В свою очередь, состояние указанных элементов, в каждый момент времени, можно охарактеризовать с помощью системы объясняющих, экзогенных переменных.

Система включает общехозяйственные и товарно-рыночные показатели. Общехозяйственные показатели отражают экономические процессы, происходящие в мире и отдельных странах, и дают представление о фоне, на котором происходит развитие рынка. Вторая группа показателей отражает явления, которые характерны для рынка нефти. Особый интерес представляют показатели, обладающие опережающим эффектом (временным лагом), по отношению к динамике эндогенных переменных конъюнктуры рынка нефти.

При выборе экзогенных переменных учитывалось, что состояние рынка нефти в любой момент времени определяется не только его внутренними факторами, но и состоянием внешней среды, т.е. общехозяйственной конъюнктуры всего мирового хозяйства, и, в первую очередь, динамикой воспроизводственного цикла, состоянием деловой активности в отраслях-потребителях, положением в кредитно-денежной и валютно-финансовой сферах экономики.

Завершающим этапом разработки модели исследуемого рынка является ее реализация. На данном этапе математическая модель формируется в общем виде, оцениваются ее параметры, проводится содержательная экономическая интерпретация, выясняются статистические и прогностические свойства модели.

При построении модели использовалась система показателей, основанная на ежеквартальных динамических рядах за последние 15 лет, которая характеризует основные стороны рынка нефти в экономическом, временном и географическом аспектах.

Использование корреляционного анализа, на этапе предварительной обработки данных, позволило ограничить круг используемых показателей (первоначально их было более 100), выбрать для дальнейшего анализа такие, которые отражают воздействие основных факторов на рынок нефти и наиболее тесно связаны с динамикой показателей конъюнктуры. При этом, решалась также, задача исключения влияния мультиколлинеарности.

Модель строилась исходя из предпосылки, что величина спроса играет более активную роль, чем факторы предложения и цены. Рекурсивная модель включает линейные регрессионные уравнения для следующих эндогенных переменных в момент времени t:

$y_{1,t}$ – экспорт нефти из стран ОПЕК;

$y_{2,t}$ – добыча нефти в странах ОПЕК;

$y_{3,t}$ – цена на нефть легкую аравийскую.

В модель вошли предопределенные переменные:

$y_{3,t-1}$ –- цена на нефть легкую аравийскую с лагом в 1 квартал;

$x_{6,t}$ – поставки нефти на переработку в Японию;

$x_{7,t-1}$ – поставки нефти на переработку в США в момент t-1;

$x_{9,t}$ – коммерческие запасы нефти в странах Западной Европы;

$x_{10,t-1}$ – коммерческие запасы нефти в США, с лагом в 1 квартал;

$x_{12,t}$ – экспорт нефти из бывшего СССР в развитые страны;

$x_{20,t-2}$ – индекс экспортных цен ООН на топливо с лагом в 2 квартала, а $x_{20,t-3}$ – в 3 квартала;

$x_{23,t-1}$ - загрузка производственных мощностей обрабатывающей промышленности США;

$\frac{y_{1,t}}{y_{2,t}}$ – показатель, учитывающий дисбаланс на рынке нефти в момент времени t.

Эконометрическая модель конъюнктуры рынка нефти имеет вид:

$$\begin{cases} \hat{y}_{1,t} = \underset{(8,5)}{4,2}x_{6,t} + \underset{(9,7)}{0,8}x_{7,t-1} + \underset{(9,7)}{1,5}x_{9,t} - \underset{(-9,0)}{0,6}x_{10,t} + \underset{(9,0)}{2,1}x_{12,t} - \underset{(-9,4)}{0,4}x_{20,t-2} - \underset{(-2,5)}{169,2} \\ \hat{y}_{2,t} = \underset{(12,0)}{0,9}y_{1,t} + \underset{(2,4)}{0,8}x_{7,t-1} + \underset{(1,8)}{0,3}x_{20,t-3} - \underset{(-1,1)}{64,0} \\ \hat{y}_{3,t} = \underset{(5,1)}{0,5}y_{3,t-1} + \underset{(1,4)}{16,2}\frac{y_{1,t}}{y_{2,t}} + \underset{(4,1)}{0,2}x_{20,t-3} + \underset{(4,1)}{0,3}x_{23,t-1} - \underset{(-2,0)}{32,6} \end{cases}.$$

Анализ статистических характеристик модели показал, что в целом она адекватно описывает рынок нефти – все уравнения значимы, объясняют от 67% до 92% дисперсии эндогенных переменных и характеризуются незначительными отклонениями расчетных значений эндогенных переменных от фактических. Значимость коэффициентов модели проверялась по t-критерию, расчетные значения которых указаны в скобках под соответствующими коэффициентами.

Построенная модель позволяет анализировать различные ситуации развития рынка нефти.

Выводы

В учебном пособии рассматриваются основные теоретические положения наиболее часто встречаемых в практике экономического анализа многомерных статистических методов исследования зависимости (корреляционный и регрессионный анализы), снижение размерностей (компонентный анализ) и классификации (кластерный анализ), а также основы эконометрики.

Значительное внимание уделяется логическому анализу исходной информации и экономической интерпретации получаемых результатов. Пособие снабжено достаточным числом подробно разработанных типовых примеров и сквозных задач, взятых из экономической практики и решенных с использованием ЭВМ.

Усвоению теоретического материала курса призваны способствовать тесты, завершающие каждый раздел учебного пособия, а также итоговый тест.

Сквозные примеры иллюстрируют необходимость комплексного применения многомерных статистических методов или решения социально-экономических задач. При этом, корреляционный анализ с одной стороны, используется на этапе предварительного анализа для выявления мультиколлинеарности, а с другой – при оценке адекватности регрессионной модели, компонентный анализ используется в задачах снижения размерности, а также при построении уравнения регрессии на главных компонентах и в задачах классификации. При окончательном выборе модели, в работе рекомендуется использовать как экономические, так и статистические критерии. Наряду с точечными, в учебном пособии рассматриваются методы построения интервальных оценок коэффициентов и уравнения регрессии.

В разделе “Основы эконометрики” рассматриваются производственные функции и системы одновременных эконометрических уравнений, двухшаговый и трехшаговый методы наименьших квадратов.

Учебное пособие предназначено для студентов, изучающих многомерные статистические методы, и специалистов, желающих повысить свою квалификацию по современным эконометрическим методам.

Литература

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. М., Финансы и статистика, 1983, 471 с.;

2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. М., Финансы и статистика, 1985, 487 с.;

3. Айвазян С.А., Бухштабер В. М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерностей. М., Финансы и статисика, 1989, 607 с.;

4. Мхитарян В.С., Трошин Л.И. Исследование зависимостей методами корреляции и регрессии. М., МЭСИ, 1995, 120 с.;

5. Мхитарян В.С., Дубров А.М., Трошин Л.И. Многомерный статистический анализ в экономике. М., МЭСИ, 1995, 149 с.;

6. Дубров А.М., Мхитарян В.С., Трошин Л.И. Математическая статистика для бизнесменов и менеджеров. М., МЭСИ, 1996, 140 с.;

7. Джонстон Дж. Эконометрические методы, М.: Статистика, 1980, 446 с.

Приложения

Приложение 1

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

Методические указания к использованию некоторых таблиц

В таблице 1 протабулирована функция:

$$\Phi(t) = \frac{2}{\sqrt{2\pi}} \int_0^t e^{-\frac{x^2}{2}} dx ,$$

f(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:

$$P(t_1 < T < t_2) = \frac{1}{2}[\Phi(t_2) - \Phi(t_1)]$$

Φ(t) обладает следующими свойствами:

$$\Phi(-t) = -\Phi(t); \quad \Phi(\infty) = 1; \quad \Phi(3) = 0{,}9973.$$

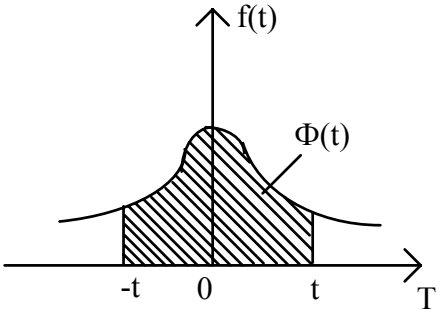
Пример:

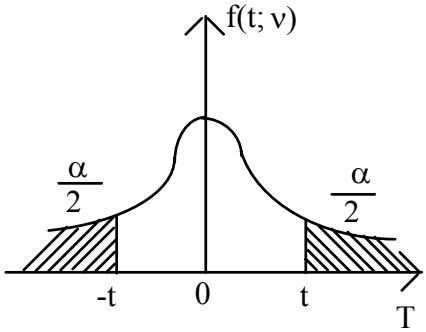
$$P(-1{,}36 < T < 2{,}15) = \frac{1}{2}[\Phi(2{,}15) - \Phi(-1{,}36)] = \frac{1}{2}[0{,}9684 + 0{,}8262] = 0{,}8973$$

В таблице 2 протабулирована вероятность выхода за пределы интервала от –t до +t случайной величины, имеющей распределение Стьюдента (t – распределение) с числом степеней свободы ν .

$$\alpha = St(t;\nu) = P(|T| > t)$$

f (t; ν) – 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);

St (∞) = 0; St (-∞) = 2; St (0) = 1.

Пример: при ν = 10 определить:

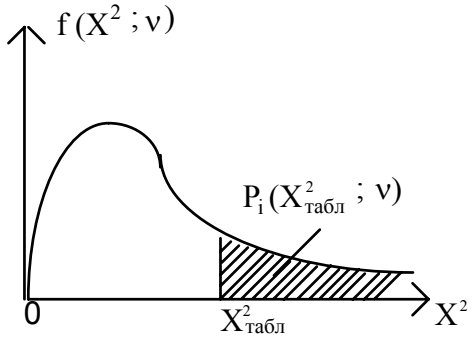
$$P(-1{,}36 < T < 2{,}15) = \frac{1}{2}[St(-1{,}36) - St(2{,}15)] = \frac{1}{2}[2 - St(1{,}36) -$$

$$- St(2{,}15)] \cong \frac{1}{2}[2 - St(1{,}372) - St(2{,}228)] = \frac{1}{2}[2 - 0{,}2 - 0{,}05] = \quad 75$$

Чтобы не прибегать к интерполяции, в строке, соответствующей ν =10, мы взяли ближайшие к заданным значениям 1,36 и 2,15.

Каждая строка таблицы отвечает t-распределению, с соответствующим числом степеней свободы ν .

В таблице 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 $\chi^2_{табл}$.



На рис. 3 представлен график функции $f(X^2_{табл})$ – плотности χ^2 – распределения с числом степеней свободы ν.

f ($\chi^2_{табл}$; у) – плотность χ^2 – распределения с числом степеней свободы ν .

Вероятность попадания случайной величины χ^2 в интервал от χ^2_1 до χ^2_2 вычисляется по формуле:

$$P(\chi_1^2 < \chi^2 < \chi_2^2) = P(\chi^2 > \chi_1^2) - P(\chi^2 > \chi_2^2) = P_i(\chi_1^2) - P_i(\chi_2^2)$$

.

Функция $P_i(\chi^2_{табл})$ обладает следующими свойствами:

P_i (0) = 1; P_i (∞) = 0.

Пример: при ν =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 таблицы, 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 таблице 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):

$$\alpha = P(F > F_{табл}) = 0{,}05; \quad 0{,}01 \quad и \quad 0{,}001.$$

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 2

Нормальный закон распределения

З н а ч е н и е ф у н к ц и и Ф(t) = P (|T| ≤ $t_{табл}$)

Таблица 1

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

Окончание табл. 1

Целые и десятичные доли, t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
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	-	-	-	-	-	-	-	-	-

Приложение 3

Таблица 2

Р а с п р е д е л е н и е С т ь ю д е н т а (t - распределение)

	Вероятность $\alpha = St(t) = P(\|T\| > t_{табл})$												
ν	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	1,943	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	0,327	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	1,101	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

Окончание табл. 2

	Вероятность $\alpha = St(t) = P(\|T\| > t_{табл})$												
ν	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
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	0,868	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
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

Приложение 4

Таблица 3

Р а с п р е д е л е н и е П и р с о н а (χ^2-распределение)

Значения $\chi^2_{табл}$ для вероятностей Р ($\chi^2 > \chi^2_{табл}$)

	В е р о я т н о с т ь										
ν	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
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

Окончание табл. 3

	Вероятность									
ν	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
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

Приложение 5

Р а с п р е д е л е н и е Ф и ш е р а - С н е д е к о р а (F-распределение) 4

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	199,5	215,7	224,6	230,2	234,0	238,9	243,9	249,0	253,3	12,71
	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366	63,66
	406523	500016	536700	562527	576449	585953	598149	610598	623432	636535	636,2
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50	4,30
	98,49	99,01	00,17	99,25	99,30	99,33	99,36	99,42	99,46	99,50	9,92
	998,46	999,00	999,20	999,20	999,20	999,20	999,40	999,60	999,40	999,40	31,00
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53	3,18
	34,12	30,81	29,46	28,71	28,24	27,91	27,49	27,05	26,60	26,12	5,84
	67,47	148,51	141,10	137,10	134,60	132,90	130,60	128,30	125,90	123,50	12,94
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63	2,78
	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,37	13,93	13,46	4,60
	74,13	61,24	56,18	53,43	51,71	50,52	49,00	47,41	45,77	44,05	8,61
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36	2,57
	16,26	13,27	12,06	11,39	10,97	10,67	10,27	9,89	9,47	9,02	4,03
	47,04	36,61	33,20	31,09	20,75	28,83	27,64	26,42	25,14	23,78	6,86
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67	2,45
	13,74	10,92	9,78	9,15	8,75	8,47	8,10	7,72	7,31	6,88	3,71
	35,51	26,99	23,70	21,90	20,81	20,03	19,03	17,99	16,89	15,75	5,96
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23	2,36
	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,47	6,07	5,65	3,50
	29,22	21,69	18,77	17,19	16,21	15,52	14,63	13,71	12,73	11,70	5,40

Продолжение табл. 4

ν_1 \ ν_2	1	2	3	4	5	6	8	12	24	∞	t
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,99	2,31
	11,26	8,65	7,59	7,10	6,63	6,37	6,03	5,67	5,28	4,86	3,36
	25,42	18,49	15,83	14,39	13,49	12,86	12,04	11,19	10,30	9,35	5,04
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71	2,26
	10,56	8,02	6,99	6,42	6,06	5,80	5,47	5,11	4,73	4,31	3,25
	22,86	16,39	13,90	12,56	11,71	11,13	10,37	9,57	8,72	7,81	4,78
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54	2,23
	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,71	4,33	3,91	3,17
	21,04	14,91	12,55	11,28	10,48	9,92	9,20	8,45	7,64	6,77	4,59
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40	2,20
	9,65	7,20	6,22	5,67	5,32	5,07	4,74	4,40	4,02	3,60	3,11
	19,69	13,81	11,56	10,35	9,58	9,05	8,35	7,62	6,85	6,00	4,49
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30	2,18
	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,16	3,78	3,36	3,06
	18,64	12,98	10,81	9,63	8,89	8,38	7,71	7,00	6,25	5,42	4,32
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21	2,16
	9,07	6,70	5,74	5,20	4,86	4,62	4,30	3,96	3,59	3,16	3,01
	17,81	12,31	10,21	9,07	8,35	7,86	7,21	6,52	5,78	4,97	4,12
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13	2,14
	8,86	6,51	5,56	5,03	4,69	4,46	4,14	3,80	3,43	3,00	2,98
	17,14	11,78	9,73	8,62	7,92	7,44	6,80	6,13	5,41	4,60	4,14
15	4,45	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07	2,13
	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,67	3,29	2,87	2,95
	16,59	11,34	9,34	8,25	7,57	7,09	6,47	5,81	5,10	4,31	4,07
16	4,41	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01	2,12
	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,55	3,18	2,75	2,92
	16,12	10,97	9,01	7,94	7,27	6,80	6,20	5,55	4,85	4,06	4,02

Продолжение табл. 4

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞	t
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96	2,11
	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,45	3,08	2,65	2,90
	15,72	10,66	8,73	7,68	7,02	6,56	5,96	5,32	4,63	3,85	3,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92	2,10
	8,28	6,01	5,09	4,58	4,25	4,01	3,71	3,37	3,01	2,57	2,88
	15,38	10,39	8,49	7,46	6,81	6,35	5,76	5,13	4,45	3,67	3,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88	2,09
	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,30	2,92	2,49	2,86
	15,08	10,16	8,28	7,26	6,61	6,18	5,59	4,97	4,29	3,52	3,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84	2,09
	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,23	2,86	2,42	2,84
	14,82	9,95	8,10	7,10	6,46	6,02	5,44	4,82	4,15	3,38	3,85
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,82	2,08
	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,17	2,80	2,36	2,83
	14,62	9,77	7,94	6,95	6,32	5,88	5,31	4,70	4,03	3,26	3,82
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78	2,07
	7,94	5,72	4,82	4,31	3,99	3,75	3,45	3,12	2,75	2,30	2,82
	14,38	9,61	7,80	6,81	6,19	5,76	5,19	4,58	3,92	3,15	3,79
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76	2,07
	7,88	5,66	4,76	4,26	3,94	3,71	3,41	3,07	2,70	2,26	2,81
	14,19	9,46	7,67	6,70	6,08	5,56	5,09	4,48	3,82	3,05	3,77
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73	2,06
	7,82	5,61	4,72	4,22	3,90	3,67	3,36	3,03	2,66	2,21	2,80
	14,03	9,34	7,55	6,59	5,98	5,55	4,99	4,39	3,74	2,97	3,75

Окончание табл. 4

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞	t
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71	2,06
	7,77	5,57	4,68	4,18	3,86	3,63	3,32	2,99	2,62	2,17	2,79
	13,88	9,22	7,45	6,49	5,89	5,46	4,91	4,31	3,66	2,89	3,72
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69	2,06
	7,72	5,53	4,64	4,14	3,82	3,59	3,29	2,96	2,58	2,13	2,78
	13,74	9,12	7,36	6,41	5,80	5,38	4,83	4,24	3,59	2,82	3,71
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67	2,05
	7,68	5,49	4,60	4,11	3,78	3,56	3,26	2,93	2,55	2,10	2,77
	13,61	9,02	7,27	6,33	5,73	5,31	4,76	4,17	3,52	2,76	3,69
28	4,19	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65	2,05
	7,64	5,54	4,57	4,07	3,75	3,53	3,23	2,90	2,52	2,06	2,76
	13,50	8,93	7,18	6,25	5,66	5,24	4,69	4,11	3,46	2,70	3,67
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64	2,05
	7,60	5,42	4,54	4,04	3,73	3,50	3,20	2,87	2,49	2,03	2,76
	13,39	8,85	7,12	6,19	5,59	5,18	4,65	4,05	3,41	2,64	3,66
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62	2,04
	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,84	2,47	2,01	2,75
	13,29	8,77	7,05	6,12	5,53	5,12	4,58	4,00	3,36	2,59	3,64
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39	2,00
	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,50	2,12	1,60	2,66
	11,97	7,76	6,17	5,31	4,76	4,37	3,87	3,31	2,76	1,90	3,36
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,03	1,96
	6,64	4,60	3,78	3,32	3,02	2,80	2,51	2,18	1,79	1,04	2,58
	10,83	6,91	5,42	4,62	4,10	3,74	3,27	2,74	2,13	1,05	3,29

Приложение 6

Таблица 5

Т а б л и ц а Ф и ш е р а - И е й т с а

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			

Приложение 7

Таблица 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	996	2,7587	2,8257	2,9031	2,9945	3,1063	3,2504	3,4534	3,8002

Таблица 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
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551

Приложение 9

Таблица 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

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

Приложение 10

Таблица 9

G - распределение

Пяти- и однопроцентное пределы для отношения G наибольшей выборочной дисперсии к сумме L выборочных дисперсий, полученных из L независимых выборок объемом n.
Первое значение соответствует уровню значимости $\alpha = 0,05$, а второе – $\alpha = 0,01$

L \ n-1	1	2	3	4	5	6	7	8	9	10	16	36	144	∞
	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
2	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
	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
3	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
	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
4	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
	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
5	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
	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
6	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
	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
7	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
	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
8	0,795		0,521	0,463	0,423	0,393	0,370	0,352	0,337	0,325	0,278	0,221	0,170	0,125
	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
9	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