Министерство образования Российской Федерации
Московский государственный институт электронной техники
(Технический университет)

Э.А.Вуколов

Регрессионный анализ

Методические указания по курсу «Статистика»

Утверждено редакционно-издательским советом института

Москва 2000

УДК 519.246.8

Рецензенты: канд. техн. наук *С.А.Леонова*
канд. техн. наук *А.А.Мазаев*

Вуколов Э.А.
Регрессионный анализ. Методические указания по курсу «Статистика». - М.: МИЭТ, 2000. - 52 с.: ил.

Рассмотрены методы линейного регрессионного анализа: оценка параметров регрессионных моделей, анализ качества и проверка адекватности моделей результатам наблюдений, определение ошибки прогноза, мультиколлинеарность и ее следствия. Все методы представлены в матричном виде, что позволяет использовать их для любых линейных по параметрам моделей. Приведены задачи для самостоятельного решения.

Предназначены для студентов и аспирантов, изучающих методы регрессионного анализа.

Введение

Основная цель регрессионного анализа состоит в построении математических моделей объектов или явлений на основе данных, полученных в результате наблюдений или экспериментов. Составление модели начинается с содержательной постановки задачи, определения основных характеристик рассматриваемого явления - зависимых переменных $y_1, y_2, \ldots, y_k$ и факторов, обуславливающих изменение этих характеристик, - независимых переменных $x_1, x_2, \ldots, x_m$. Например, нас интересует зависимость спроса на товар от его цены и покупательной способности населения. В этом случае зависимой переменной является спрос, а независимыми переменными будут цена и покупательная способность населения. Очевидно, что спрос на товар определяется не только ценой и покупательной способностью населения, но и такими факторами, как расходы на рекламу товара, время года и т.п., число которых можно продолжать до бесконечности. Некоторые факторы могут иметь случайный характер, например, непредсказуемая человеческая реакция или влияние моды. Поэтому среди действующих факторов рассматривают наиболее важные.

Математическая модель, описывающая поведение зависимой переменной *Y*, представляет сумму детерминированной и случайной компонент

$$Y = f(x_1, x_2, \ldots, x_m) + \varepsilon,$$

где $f(x_1, x_2, \ldots, x_m)$ - функция независимых переменных; ε - случайная величина, отражающая суммарный эффект влияния всех остальных факторов и ошибок измерения зависимой переменной. Обычно предполагают, что ε имеет нормальное распределение. Основанием этому служит центральная предельная теорема, утверждающая, что сумма случайных величин с приблизительно одинаковыми дисперсиями имеет распределение, близкое к нормальному закону. Предполагается также, что математическое ожидание ε равно нулю, а дисперсия постоянна.

В настоящей работе рассматриваются методы анализа большого класса математических моделей - моделей линейной регрессии. Эти методы составляют предмет линейного регрессионного анализа. Практические расчеты, необходимые при построении математических моделей методами регрессионного анализа, могут быть полностью выполнены с помощью современных пакетов прикладных программ, например Statistica, Statgraphics, SPSS. Очень просто решаются задачи оценки параметров модели в пакете Matlab. Использование статистических пакетов позволяет не только произвести все необходимые расчеты при построении модели, но и подобрать наилучшую модель, описывающую данные (выбор наилучшей регрессии), использовать специальные методы в условиях мультиколлинеарности (например, метод гребневой регрессии) и рассчитать параметры нелинейных моделей.

1. Определение регрессионной модели

Пусть Y - зависимая переменная, а $x_1, x_2, \ldots x_m$ - независимые переменные (факторы), определяющие поведение зависимой переменной. При построении модели, описывающей зависимость Y от $x_1, x_2, \ldots x_m$, предполагается, во-первых, что у исследователя имеются результаты совокупных наблюдений зависимой переменной Y и независимых переменных $x_1, x_2, \ldots x_m$, во-вторых, что значения независимых переменных определяются точно (без ошибок), а значение зависимой переменной Y определяется с ошибками, имеющими случайный характер. Математическая модель, описывающая данные такого вида, выглядит следующим образом:

$$Y = f(x_1, x_2, \ldots x_m) + \varepsilon,$$

где ε – случайная ошибка наблюдений зависимой переменной. Математическое ожидание ε предполагается равным нулю: $M[\varepsilon] = 0$.

Регрессией (уравнением регрессии) называется условное математическое ожидание Y

$$M[Y / x_1, x_2, \ldots x_m] = f(x_1, x_2, \ldots x_m). \quad (1)$$

Таким образом, регрессия описывает поведение наблюдаемой зависимой переменной в среднем, представляя ее главную тенденцию. В связи с этим нахождение регрессии по результатам наблюдений называют *сглаживанием данных.*

Существуют различные регрессионные модели, определяемые выбором функции $f(x_1, x_2, \ldots x_m)$:

простая линейная регрессия

$$Y = \beta_0 + \beta_1 x + \varepsilon; \quad (2)$$

множественная регрессия

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1} + \varepsilon; \quad (3)$$

полиномиальная регрессия

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_{k-1} x^{k-1} + \varepsilon. \quad (4)$$

Зависимость, описываемая системой функций $\alpha_0(x)$, $\alpha_1(x), \ldots \alpha_{k-1}(x)$, имеет вид:

$$Y = \beta_0 \alpha_0(x) + \beta_1 \alpha_1(x) + \ldots + \beta_{k-1} \alpha_{k-1}(x) + \varepsilon. \quad (5)$$

Коэффициенты $\beta_0, \beta_1, \ldots, \beta_{k-1}$ называются *параметрами* регрессии. После выбора определенной модели параметры регрессии должны быть вычислены по результатам наблюдений зависимой переменной и факторов.

В приведенные регрессионные модели параметры $\beta_0, \beta_1, ..., \beta_{k-1}$ входят *линейно*. Такие модели называют *линейными* (по параметрам) моделями, а математические методы анализа этих моделей - *линейным регрессионным анализом*.

Модель $Y = \beta_0 e^{\beta_1 x_1} + \beta_1 e^{\beta_2 x_2}$ не линейна по параметрам. В некоторых случаях нелинейные модели с помощью специальных линеаризирующих преобразований могут быть преобразованы в линейные [1, 4]. Рассмотрим несколько примеров.

1. Функция $y = \beta_0 x^{\beta_1}$ с помощью логарифмирования и замены переменных преобразуется так: $\ln y = \ln \beta_0 + \beta_1 \ln x$. Заменив переменные $y' = \ln y;\ \beta_0' = \ln \beta_0;\ x' = \ln x,$ получим линейную по параметрам функцию

$$y' = \beta_0' + \beta_1 x'.$$

2. Функция $y = \dfrac{ax}{b + x}$ преобразуется так:

$$b + x = a\frac{x}{y}, \text{ или } \frac{x}{y} = \frac{b}{a} + \frac{1}{a}x.$$

После замены переменных $y' = \dfrac{x}{y},\ \beta_0 = \dfrac{b}{a},\ \beta_1 = \dfrac{b}{a}$ получим

$$y' = \beta_0 + \beta_1 x.$$

3. Логистическая функция $y = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ (рис.1) при помощи преобразования $y' = \ln\left(\dfrac{y}{1-y}\right)$ принимает вид:

$$y' = \beta_0 + \beta_1 x.$$

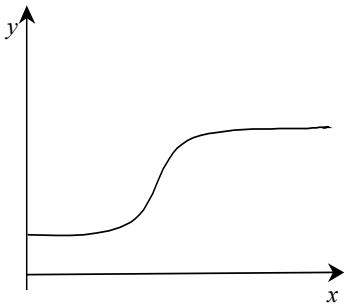

Рис.1. Логистическая функция

Далее будут анализироваться линейные регрессионные модели. В общем виде такую модель можно записать следующим образом:

$$Y = \beta_0 + \beta_1\varphi_1(x_1, x_2, \ldots x_m) + \ldots + \beta_{k-1}\varphi_{k-1}(x_1, x_2, \ldots x_m) + \varepsilon,$$

где $\varphi_i(x_1, x_2, \ldots, x_m)$, $i = 1, 2, \ldots, k-1$, - заданные функции факторов.

Для описания некоторых явлений необходимо использование "лаговых" переменных. Соответствующие модели называются *авторегрессионными* и записываются в виде

$$y_i = -\alpha_1 y_{i-1} - \alpha_2 y_{i-2} - \ldots - \alpha_k y_{i-k} + \gamma_1 x_{i-1} + \ldots + \gamma_k x_{i-k} + \varepsilon.$$

В этом случае предполагается, что наблюдения проводятся в дискретные моменты времени, отстоящие друг от друга на интервал Δt. Обозначим y_i значение зависимой переменной Y в i-й момент времени, т.е. $y_i = Y(i \cdot \Delta t)$. Тогда значение Y, наблюдаемое на k интервалов раньше, будет обозначаться как y_{i-k}. Аналогично x_{i-k} - значение фактора, наблюдаемое с запаздыванием на k интервалов по отношению к текущему моменту времени $t = i \cdot \Delta t$. Авторегрессионную модель можно записать в виде

$$y_i = \beta_1\varphi_1 + \beta_2\varphi_2 + \ldots + \beta_k\varphi_k + \beta_{k+1}\varphi_{k+1} + \ldots + \beta_{2k}\varphi_{2k} + \varepsilon_i,$$

если сделать следующую замену:

$$\beta_1 = -\alpha_1; \quad \beta_2 = -\alpha_2; \ldots; \beta_k = -\alpha_k;$$
$$\varphi_1 = y_{i-1}; \quad \varphi_2 = y_{i-2}; \ldots; \varphi_k = y_{i-k};$$

$\beta_{k+1} = \gamma_1; \ldots; \beta_{2k} = \gamma_k; \quad \varphi_{k+1} = x_{i-1}; \ldots \varphi_{2k} = x_{i-k}$.

Анализ авторегрессионных моделей имеет ряд особенностей, связанных с высокой коррелированностью лаговых переменных [4].

2. Оценка параметров регрессионной модели по результатам наблюдений

Рассмотрим регрессионную модель с одной независимой переменной *x*. Пусть проведено *n* независимых наблюдений случайной величины *Y* при значениях *x*, равных $x_1, x_2, \ldots x_n$, другими словами, имеется *n* пар наблюдений (x_i, y_i), $i = 1, 2, ..., n$. В случае простой линейной регрессии (2) для оценки двух параметров регрессии β_0 и β_1 имеем *n* уравнений:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1; \\ y_2 &= \beta_0 + \beta_1 x_2; \\ y_3 &= \beta_0 + \beta_1 x_3; \\ &\ldots\ldots\ldots\ldots \\ y_n &= \beta_0 + \beta_1 x_n, \end{aligned} \qquad (6)$$

или, в матричном виде

$$Y = A\beta, \qquad (7)$$

где $Y = (y_1, \ldots y_n)^T$; $\beta = (\beta_0, \beta_1)^T$; индекс “ T ” означает транспонирование; *A* - регрессионная матрица размера $(n \times 2)$, $A = \begin{pmatrix} 1 & x1 \\ 1 & x2 \\ . & . \\ 1 & xn \end{pmatrix}$.

Для моделей (4) и (5) регрессионные матрицы соответственно будут:

$$A = \begin{pmatrix} 1 & x_1 & . & x_1^{k-1} \\ 1 & x_2 & . & x_2^{k-1} \\ . & . & . & . \\ 1 & x_n & . & x_n^{k-1} \end{pmatrix}; \quad A = \begin{pmatrix} \alpha_0(x_1) & \alpha_1(x_1) & ... & \alpha_{k-1}(x_1) \\ \alpha_0(x_2) & \alpha_1(x_2) & ... & \alpha_{k-1}(x_2) \\ ... & ... & ... & ... \\ \alpha_0(x_n) & \alpha_1(x_n) & ... & \alpha_{k-1}(x_n) \end{pmatrix}.$$

В случае множественной линейной регрессии (3) регрессионная матрица имеет вид:

$$A = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{(k-1)1} \\ 1 & x_{12} & x_{22} & \dots & x_{(k-1)2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{(k-1)n} \end{pmatrix},$$

где x_{ij} - результат j-го измерения i-й переменной; $i = 1, 2,..., k-1$; $j = 1, 2,..., n$.

Число наблюдений n в регрессионных моделях (2) - (5) больше числа параметров k, которые надо оценить. Это приводит к необходимости решения несовместной системы линейных алгебраических уравнений

$$Y = A\beta,$$

где A - матрица размера $(n \times k)$, $n > k$.

Для решения таких систем применяется метод наименьших квадратов (МНК), состоящий в следующем.

Введем вектор невязки (вектор ошибок наблюдений):

$$\varepsilon = Y - A\beta . \qquad (8)$$

Предположим, что ранг матрицы A равен k и, следовательно, столбцы матрицы A - линейно-независимые векторы $A_1, A_2, \dots A_k$.

Вектор $A\beta$ является линейной комбинацией столбцов матрицы A и может быть записан в виде

$$A\beta = A_1\beta_0 + A_2\beta_1 + \dots + A_k\beta_{k-1}.$$

Это означает, что вектор $A\beta$ лежит в линейном подпространстве L_k размерности *k*, причем $L_k \subset L_n$, где L_n - линейное пространство векторов размерности *n*. Так как вектор наблюдений *Y* принадлежит L_n, то в качестве решения системы

$$Y = A\beta$$

можно взять вектор $\tilde{\beta}$, минимизирующий модуль вектора невязки

$$\|\varepsilon\| = \|Y - A\beta\|.$$

В геометрической интерпретации (рис.2) вектор $A\tilde{\beta}$ будет ортогональной проекцией вектора *Y* на подпространство L_k.

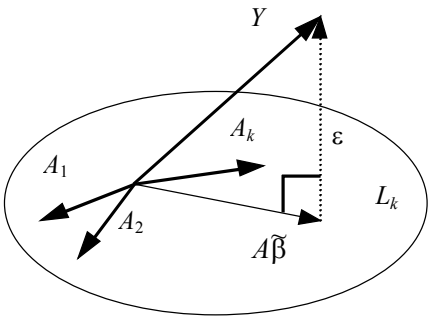

*Рис.2.*Геометрическая интерпретация метода наименьших квадратов

Из условия ортогональности вектора невязки и подпространства L_k следует, что скалярное произведение вектора ε и любого вектора $A\alpha \subset L_k$ равно нулю:

$$(A\alpha)^T (Y - A\beta) = 0$$

или

$$\alpha^T \left(A^T Y - A^T A\beta\right) = 0.$$

Решив уравнение $A^T Y - A^T A\beta = 0$, получим, что решением несовместной системы $\mathrm{Y} = \mathrm{A}\beta$, минимизирующим модуль вектора невязки ε, будет вектор $\tilde{\beta}$:

$$\tilde{\beta} = \left(A^T A\right)^{-1} A^T Y. \qquad (9)$$

Решение (9) определяет оценки параметров линейной регрессионной модели по результатам наблюдений. Эти оценки называются МНК-оценками.

Для простой линейной регрессии $Y = \beta_0 + \beta_1 x$, применив (9), получим МНК-оценки:

$$\widetilde{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \qquad (10)$$

$$\widetilde{\beta}_0 = \bar{y} - \widetilde{\beta}_1 \bar{x}, \qquad (11)$$

где $\bar{x} = \frac{1}{n}\sum x_i$, $\bar{y} = \frac{1}{n}\sum y_i$.

3. Статистический анализ МНК-оценок. Оценка качества аппроксимации данных с помощью линейной регрессионной модели

МНК-оценки параметров модели, вычисляемые по формуле

$$\widetilde{\beta} = (A^T A)^{-1} A^T Y,$$

представляют собой линейные функции случайного вектора Y и, значит, будут случайными величинами. Этот результат является следствием того, что оценки параметров модели вычисляются по выборке наблюдений переменных Y, x_1, x_2, ... и изменяются при вариациях наблюдений.

Задача статистического анализа состоит в исследовании свойств оценок, нахождении доверительных интервалов для параметров линейной модели и проверке гипотез о параметрах.

Статистический анализ проводится при следующих предположениях:

- вектор ошибок (невязки) имеет нулевое математическое ожидание: $M[\varepsilon]=0$ - это означает, что отсутствуют систематические ошибки наблюдений;

- дисперсии вектора ошибок постоянны: $D[\varepsilon_i]=\sigma^2$, $i = 1,2,...,n$, а его координаты некоррелированы: $\mathrm{Cov}(\varepsilon_i,\varepsilon_j)=0$, $i \neq j$, $i,j = 1,2, ...,n$ или $\mathrm{Cov}(\varepsilon)=\sigma^2 I$, где I - единичная ($n \times n$) матрица;

- вектор ошибок ε имеет n-мерное нормальное распределение $N(0, \sigma^2 I)$. При этом предположении из некоррелированности ошибок следует их независимость.

С учетом (8), (9) и данных предположений получим ряд следствий.

Следствие 1. В силу (8) вектор наблюдений в линейной регрессионной модели можно записать в виде

$$Y = A\beta + \varepsilon. \qquad (12)$$

Отсюда следует, что Y имеет n-мерное нормальное распределение с математическим ожиданием

$$M[Y] = M[A\beta + \varepsilon] = A\beta = m_Y$$

и ковариационной матрицей

$$\mathrm{Cov}(Y) = M\left[(Y - m_Y)(Y - m_Y)^T\right] = M[\varepsilon\varepsilon^T] = \mathrm{Cov}(\varepsilon) = \sigma^2 I.$$

Следствие 2. Пусть C - постоянная $(k \times n)$ матрица. Тогда легко проверяются следующие свойства:

$$M[CY] = CM[Y],$$
$$\mathrm{Cov}(CY) = C\mathrm{Cov}(Y)C^T.$$

Обозначим квадратную симметрическую матрицу $A^T A$ размера $(k \times k)$ как B (эта матрица называется информационной) и запишем вектор МНК-оценок в виде

$$\tilde{\beta} = B^{-1}A^T Y = B^{-1}A^T(A\beta + \varepsilon) = \beta + B^{-1}A^T\varepsilon.$$

Найдем математическое ожидание вектора $\tilde{\beta}$:

$$M[\tilde{\beta}] = \beta + B^{-1}A^T M[\varepsilon] = \beta.$$

Это означает, что МНК-оценки - несмещенные.

Вычислим ковариационную матрицу $\widetilde{\beta}$, используя формулу для ковариации произведения *CY*:

$$\mathrm{Cov}(\widetilde{\beta}) = \mathrm{Cov}(B^{-1}A^{T}Y) = B^{-1}A^{T}\mathrm{Cov}(Y)AB^{-1} = \sigma^{2}B^{-1}. \qquad (13)$$

Таким образом $\widetilde{\beta}$, как линейная функция ε, имеет *k*-мерное нормальное распределение с математическим ожиданием и ковариационной матрицей соответственно $M[\widetilde{\beta}] = \beta,\ \mathrm{Cov}(\widetilde{\beta}) = \sigma^{2}B^{-1}$.

Следствие 3. Если регрессионная матрица *A* имеет ортонормированные столбцы, то $B = A^{T}A = I$, где *I* - единичная $(k \times k)$ матрица. В этом случае ковариационная матрица МНК-оценок имеет вид: $\mathrm{Cov}(\widetilde{\beta}) = B^{-1}\sigma^{2} = \sigma^{2}I$, т.е. оценки параметров некоррелированы, а так как $\widetilde{\beta}$ имеет *k*-мерное нормальное распределение, то вектор $\widetilde{\beta}$ имеет независимые компоненты.

С вычислительной точки зрения ортогонализация матрицы *A* позволяет исключить операцию обращения матрицы $A^{T}A = B$. Пусть *A* представлена в виде *A* = *QR*, где *Q* - $(k \times k)$ матрица с ортонормированными столбцами, $Q^{T}Q = I;$ *R* - ($k \times k$) верхняя треугольная матрица (матрица перехода к ортогональному базису). В пакете Matlab эта операция выполняется с помощью оператора

$$[Q, R] = \mathrm{qr}(A).$$

Тогда, подставив в (9) *A* = *QR*, с учетом того, что $Q^{T}Q = I$, получим:

$$\widetilde{\beta} = R^{-1}\mathrm{Q}^{T}Y$$

или

$$\widetilde{\beta}' = R\widetilde{\beta} = \mathrm{Q}^{T}Y.$$

Последний результат показывает, что в ортонормированном базисе вектор оценок параметров регрессии $\widetilde{\beta}'$ получается перемножением Q^{T} на *Y*. Эта процедура используется для определения истинной размерности регрессионной модели.

Следствие 4. Подставим МНК-оценки $\widetilde{\beta}$ в уравнение невязки (8):

$$Y - A\widetilde{\beta} = e. \tag{14}$$

Вектор e (вектор остатков) определяет разницу между результатами наблюдений и значениями, предсказываемыми моделью. Квадрат модуля вектора остатков, равный $e^T e$, называется *остаточной суммой квадратов* Q_e, которая характеризует качество аппроксимации данных регрессионной моделью.

Линейная регрессионная модель называется *адекватной результатам наблюдений*, если предсказанные по ней значения зависимой переменной Y согласуются с результатами наблюдений. Если модель адекватна, то остатки e_i являются реализациями случайных ошибок наблюдений ε_i, i = 1,2, ..., n, и, следовательно, в силу вышеприведенных предположений независимы и имеют нормальное распределение $N\left(0, \sigma^2\right)$.

Проверка выполнения предположений различными статистическими методами лежит в основе оценки адекватности модели по остаткам. Если для каждого или для некоторых значений независимой переменной имеются несколько значений зависимой переменной Y (повторные наблюдения), то проверка адекватности проводится на основе дисперсионного анализа (см. ниже).

В дальнейшем предполагается, что регрессионная модель (12) адекватна результатам наблюдений, тогда можно показать [2], что отношение $\frac{\mathrm{Q}_e}{\sigma^2}$ имеет распределение χ^2 с (n – k) степенями свободы, где σ^2 - дисперсия ошибок наблюдений; n - число наблюдений; k - число параметров модели. Отсюда следует, что статистика

$$S^2 = \frac{\mathrm{Q}_e}{n-k} \tag{15}$$

будет несмещенной оценкой дисперсии ошибок наблюдений, которая, как правило, неизвестна. Доверительный интервал для σ^2 имеет вид:

$$\frac{Q_e}{\chi^2_{1-\frac{\alpha}{2}}(n-k)} < \sigma^2 < \frac{Q_e}{\chi^2_{\frac{\alpha}{2}}(n-k)}, \qquad (16)$$

где $\chi^2_{1-\frac{\alpha}{2}}(n-k)$, $\chi^2_{\frac{\alpha}{2}}(n-k)$ - квантили распределения $\chi^2(n-k)$ соответственно порядков $1-\alpha/2$ и $\alpha/2$, а α - заданный уровень значимости.

Следствие 5. Рассмотрим координаты вектора МНК-оценок. Из следствия 2 получим:

$$M[\tilde{\beta}_i] = \beta_i; \quad D[\tilde{\beta}_i] = \sigma^2 (B^{-1})_{ii},$$

где $\left(B^{-1}\right)_{ii}$ - диагональный элемент матрицы B^{-1}, находящийся на пересечении i-й строки и i-го столбца, $i = 0,1, ..., k-1$,

$$\tilde{\beta}_i \infty N(\beta_i, D[\tilde{\beta}_i]).$$

Из следствия 4 имеем:

$$\frac{Q_e}{\sigma^2} = \chi^2(n-k).$$

Для оценки неизвестной дисперсии ошибок наблюдений σ^2 воспользуемся статистикой $S^2 = \frac{Q_e}{n-k}$. Из двух последних соотношений следует:

$$\frac{S^2}{\sigma^2} = \frac{\chi^2(n-k)}{n-k}.$$

Рассмотрим статистику:

$$\frac{\frac{\tilde{\beta}_i - \beta_i}{\sqrt{D[\tilde{\beta}_i]}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\frac{\tilde{\beta}_i - \beta_i}{\sigma\sqrt{(B^{-1})_{ii}}}}{\frac{S}{\sigma}} = \frac{\tilde{\beta}_i - \beta_i}{S\sqrt{(B^{-1})_{ii}}} = T(n-k).$$

Эта статистика имеет распределение Стьюдента с $(n-k)$ степенями свободы. С учетом этого результата получим доверительные интервалы для МНК-оценок параметров β_i:

$$\widetilde{\beta}_i - t_{1-\frac{\alpha}{2}}(n-k)\ \cdot S\sqrt{(B^{-1})_{ii}} < \beta_i < \widetilde{\beta}_i + t_{1-\frac{\alpha}{2}}(n-k)\ \ S\sqrt{(B^{-1})_{ii}}, \quad (17)$$
$$i = 1,2,...,k-1$$

где $t_{1-\frac{\alpha}{2}}(n-k)$ - квантиль распределения Стьюдента порядка $1-\frac{\alpha}{2}$ с ($n - k$) степенями свободы; α - заданный уровень значимости.

Следствие 6. Рассмотрим доверительный интервал для предсказанного значения.

Обозначим строку регрессионной матрицы A через $a^T(x)$. Для простой линейной регрессии $Y = \beta_0 + \beta_1 x$ это вектор-строка $a^T(x) = (1, x)$; для множественной регрессии $Y = \beta_0 + \beta_1 x_1 + ... + \beta_{k-1} x_{k-1}$ это вектор $a^T(x) = (1, x_1, x_2, ..., x_{k-1})$; для полиномиальной регрессии $Y = \beta_0 + \beta_1 x + ... + \beta_{k-1} x^{k-1}$ это вектор $a^T(x) = \left(1,\ x,\ x^2, \ldots, x^{k-1}\right)$.

Значение зависимой переменной Y, предсказанное регрессионной моделью в точке x_0, можно вычислить так: $\hat{Y}(x_0) = \widetilde{\beta}^T \cdot a(x_0)$, где $\widetilde{\beta}$ - вектор-столбец МНК-оценок.

Так как $M\left[\hat{Y}\right] = \beta^T a(x_0)$, то дисперсия для $\hat{Y}$ равна:

$$D\left[\hat{Y}\right] = M\left[\left(\hat{Y} - M\left[\hat{Y}\right]\right)^2\right] = M\left[\left(\hat{Y} - M\left[\hat{Y}\right]\right)^T \left(\hat{Y} - M\left[\hat{Y}\right]\right)\right] =$$
$$= M\left[a^T(x_0) \cdot \left(\widetilde{\beta}^T - \beta^T\right)^T \cdot \left(\widetilde{\beta}^T - \beta^T\right) a(x_0)\right] =$$
$$= a^T(x_0) \cdot M\left[\left(\widetilde{\beta} - \beta\right) \cdot \left(\widetilde{\beta} - \beta\right)^T\right] \cdot a(x_0) =$$
$$= a^T(x_0)\mathrm{Cov}(\widetilde{\beta}) \cdot a(x_0) = \sigma^2 a^T(x_0) \cdot B^{-1} a(x_0).$$

Оценка дисперсии для $\hat{Y}$ по результатам наблюдений равна:

$$S^2_{\hat{Y}} = S^2 a^T(x_0) B^{-1} a(x_0),$$

где $S^2 = \frac{Q_e}{n-k}$ - несмещенная оценка дисперсии ошибок наблюдений.

Предсказанное значение $\hat{Y}$ имеет нормальное распределение, поэтому статистика

$$\frac{\hat{Y} - M\left[\hat{Y}\right]}{S_{\hat{Y}}} = T(n-k)$$

имеет распределение Стьюдента с (n – k) степенями свободы. Границы доверительного интервала для предсказанного значения при доверительной вероятности (1 – α) имеют вид:

$$\hat{Y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-k)S = \hat{Y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-k)S\sqrt{a^T(x_0)\cdot B^{-1}a(x_0)},$$

где $t_{1-\frac{\alpha}{2}}(n-k)$ - квантиль распределения Стьюдента с (n – k) степенями свободы.

Границы доверительного интервала зависят от значения независимых переменных в точке x_0.

В качестве примера найдем доверительный интервал для значения, предсказанного простой регрессией

$$Y = \beta_0 + \beta_1 x.$$

Предположим, что проведено пять наблюдений при значениях x = –2, –1, 0, 1, 2 и что оценка дисперсии ошибок наблюдений $S^2 = 4.$

Для простой линейной регрессии вектор $a^T(x) = (1, x),$ а матрица A равна:

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

Далее вычисляем:

$$B = A^T A = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix};\ B^{-1} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{pmatrix}.$$

При $\alpha = 0{,}05$, квантиль распределения Стьюдента $t_{1-\frac{\alpha}{2}}(n-k) = t_{0{,}975}(3) = 3{,}182$ (см. табл.П6 в [3]).

Таким образом, доверительный интервал при доверительной вероятности 0,95 для предсказанного значения в точке $x = x_0$, $-2 \le x_0 \le 2$, будет:

$$\hat{Y}(x_0) - 3{,}182 \cdot 2 \cdot \sqrt{\frac{1}{5} + \frac{1}{10}x_0^2} \le Y(x_0) \le \hat{Y}(x_0) + 3{,}182 \cdot 2 \cdot \sqrt{\frac{1}{5} + \frac{1}{10}x_0^2}.$$

4. Дисперсионный анализ и проверка гипотез о параметрах линейной регрессии

Преобразуем остаточную сумму квадратов Q_e, применив (14):

$$\mathrm{Q}_e = e^T e = (Y - A\tilde{\beta})^T (Y - A\tilde{\beta}) = Y^T Y - 2\tilde{\beta}^T A^T Y + \tilde{\beta}^T A^T A\tilde{\beta}. \quad (18)$$

МНК-оценки $\tilde{\beta}$ были получены из условия минимума Q_e и, следовательно, $\left.\frac{\partial \mathrm{Q}_e}{\partial \beta}\right|_{\beta=\tilde{\beta}} = 0.$

Вычислив производную $\left.\frac{\partial \mathrm{Q}_e}{\partial \beta}\right|_{\beta=\tilde{\beta}}$ и приравняв ее к нулю, получим:

$$\left.\frac{\partial \mathrm{Q}_e}{\partial \beta}\right|_{\beta=\tilde{\beta}} = -2A^T Y + 2A^T A\tilde{\beta} = 0.$$

Применив это соотношение в (18), найдем:

$$\mathrm{Q}_e = e^T e = Y^T Y - \tilde{\beta}^T A^T Y. \quad (19)$$

Обозначим сумму квадратов отклонений y_i от $\bar{y}$ как Q_y:

$$\mathrm{Q}_y = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = Y^T Y - n\bar{y}^2,$$

где $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Тогда

$$\mathrm{Q}_e = Y^T Y - \tilde{\beta}^T A^T Y = \mathrm{Q}_y - (\tilde{\beta}^T A^T Y - n\bar{y}^2) = \mathrm{Q}_y - \mathrm{Q}_R, \qquad (20)$$

где

$$\mathrm{Q}_R = \tilde{\beta}^T A^T Y - n\bar{y}^2 = \tilde{\beta}^T A^T Y - \frac{1}{n}(\sum y_i)^2. \qquad (21)$$

Q_R называется *суммой квадратов*, *обусловленной регрессией.* В практических расчетах остаточную сумму квадратов Q_e вычисляют по формуле (19). Переписав (20), получим основное тождество дисперсионного анализа для линейной регрессии:

$$Q_y = Q_R + Q_e. \qquad (22)$$

Можно показать [1, 2, 4], что если модель адекватна данным, то статистики, входящие в (22), независимы и связаны с распределением χ^2:

$$\frac{\mathrm{Q}_y}{\sigma^2} = \chi^2(n-1); \qquad \frac{\mathrm{Q}_R}{\sigma^2} = \chi^2(k-1); \qquad \frac{\mathrm{Q}_e}{\sigma^2} = \chi^2(n-k).$$

Этот результат используется для проверки гипотезы $H_0: \beta_1 = \beta_2 = \ldots \beta_{k-1} = 0$, утверждающей, что независимые переменные $x_1, x_2, \ldots, x_{k-1}$ не улучшают предсказание Y по сравнению с $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Статистикой критерия для проверки гипотезы H_0 является отношение

$$F = \frac{\mathrm{Q}_R/(k-1)}{\mathrm{Q}_e/(n-k)}. \qquad (23)$$

Если гипотеза H_0 верна, то F имеет распределение Фишера с $k-1$ и $n-k$ степенями свободы [3]. Гипотеза H_0 принимается на уровне значимости α, если вычисленное значение F меньше квантили распределения Фишера порядка $1-\alpha$ $F_{1-\alpha}$ $(k-1, n-k)$. В этом случае говорят, что *регрессионная модель незначима*. Если гипотеза H_0 отклоняется, то *регрессионная модель* называется *значимой*.

Полезной характеристикой значимой регрессионной модели является *коэффициент детерминации* R^2, вычисляемый по формуле

$$R^2 = \frac{\mathrm{Q}_R}{\mathrm{Q}_y} = 1 - \frac{\mathrm{Q}_e}{\mathrm{Q}_y}. \qquad (24)$$

Коэффициент детерминации равен той доле дисперсии ошибок наблюдений, которая объясняется регрессионной моделью. Арифметический корень из R^2 равен коэффициенту корреляции между наблюдаемыми y_i и предсказываемыми моделью $\widetilde{y}_\mathrm{i}$ значениями зависимой переменной:

$$r_{y\widetilde{y}} = +\sqrt{R^2} = R.$$

R называют также *множественным коэффициентом корреляции*, так как он является мерой линейной зависимости между Y и независимыми переменными $x_1, x_2, \ldots, x_{k-1}$. В случае простой линейной регрессии R равен модулю коэффициента корреляции между Y и x:

$$r_{Yx} = (\text{Знак } \beta_1)R.$$

Приведенный коэффициент детерминации $\overline{R}^2$ вычисляется по формуле

$$\overline{R}^2 = 1 - \frac{\mathrm{Q}_e/(n-k)}{\mathrm{Q}_y/(n-1)}.$$

Относительно параметров регрессии также можно проверить гипотезу $H_0: \beta_i = 0$, утверждающую, что переменная x_i не улучшает предсказание *Y* по сравнению с предсказанием, полученным с помощью регрессии *Y* по остальным $(k-2)$ независимым переменным. Гипотеза $H_0: \beta_i = 0$, как и гипотеза $H_0: \beta_i = \beta_i^{(0)}$, где $\beta_i^{(0)}$ - заданная константа, проверяется с помощью доверительного интервала (17) для β_i: если доверительный интервал для β_i накрывает число $\beta_i^{(0)}$, то гипотеза $H_0: \beta_i = \beta_i^{(0)}$ принимается на уровне значимости α.

Рассмотрим проверку еще одной гипотезы: о равенстве нулю некоторого подмножества из $(k-1)$ параметров регрессии. Предположим, что это подмножество состоит из первых *m* параметров. Тогда гипотеза $H_0: \beta_1 = \beta_2 = \ldots = \beta_m = 0$ утверждает, что введение в регрессионную модель *m* переменных $x_1, x_2, \ldots, x_m$ не улучшает предсказание *Y* по сравнению с предсказанием, полученным с помощью регрессии *Y* по переменным $x_{m+1}, x_{m+2}, \ldots, x_{k-1}$. Для проверки H_0 сначала вычисляются параметры регрессии *Y* по $x_{m+1}, x_{m+2}, \ldots, x_{k-1}$ и находится остаточная сумма квадратов Q'_e для "урезанной" модели, а затем определяются параметры регрессии *Y* по $x_1, x_2, \ldots, x_{k-1}$ и вычисляется остаточная сумма квадратов Q_e для полной модели. Статистика критерия равна отношению

$$F = \frac{(\mathrm{Q}'_e - \mathrm{Q}_e)/m}{\mathrm{Q}_e/(n-k)}. \qquad (25)$$

Если гипотеза H_0 верна, то *F* имеет распределение Фишера с *m* и $n-k$ степенями свободы. H_0 принимается на уровне значимости α, если $F < F_{1-\alpha}(m,\ n-k)$, где $F_{1-\alpha}(m,\ n-k)$ - квантиль распределения Фишера порядка $1-\alpha$.

5. Проверка адекватности модели

Для проверки адекватности полученной модели данным надо найти остатки, т.е. разности между наблюдаемыми и предсказанными моделью значениями переменной *Y*. Вектор остатков равен:

$$e = Y - A\tilde{\beta}. \tag{26}$$

Далее вычисляются статистики, на основе которых можно проверять выполнение основных предположений регрессионного анализа: оценка математического ожидания $M[e]$, статистика Дарбина-Уотсона

$$d = \sum_{i=2}^{n} \frac{(e - e_{i-1})^2}{\mathrm{Q}_e}, \tag{27}$$

а также проверяется гипотеза о нормальном распределении остатков по критерию χ^2.

Критерий Дарбина-Уотсона позволяет проверить гипотезу H_0: остатки некоррелированы $(\rho = 0)$. Процедура проверки состоит в следующем. В зависимости от числа наблюдений *n*, числа оцениваемых параметров в модели k и уровня значимости α по таблице (см. приложение 2) находят два числа d_1 и d_2. В зависимости от формулировки альтернативной гипотезы H_1 решение принимается по одному из следующих правил:

1) $H_1: \rho > 0$:

H_0 принимается, если $d > d_2$;

H_0 отклоняется, если $d < d_1$;

при $d_1 \le d \le d_2$ решение не принимается;

2) $H_1: \rho < 0$:

H_0 принимается, если $4 - d > d_2$;

H_0 отклоняется, если $4-d<d_1$;

при $d_1 \le 4-d \le d_2$ решение не принимается;

3) H_1: $\rho \neq 0$:

H_0 принимается на уровне значимости 2α, если $d>d_2$ или $4-d>d_2$;

H_0 отклоняется на уровне значимости 2α, если $d<d_1$ или $4-d<d_1$.

Если гипотеза H_0 о некоррелированности остатков отклоняется, то либо в модели не учтен один или несколько существенных факторов, влияющих на зависимую переменную, либо неправильно выбрана форма связи между переменными.

Для адекватной модели, кроме равенства нулю математического ожидания остатков, их некоррелированности и нормального распределения, должно выполняться условие гомоскедаксичности, т.е. постоянства дисперсии ошибок наблюдений для всех наблюдений. Оценка выполнимости этого условия проводится по графику остатков в зависимости от номера наблюдений: если все остатки укладываются в симметричную относительно нулевой линии полосу, то можно считать, что дисперсия ошибок наблюдений постоянна.

Более тщательная проверка адекватности регрессионной модели может быть проведена, если для зависимой переменной *Y* проведены повторные наблюдения. В этом случае для проверки адекватности модели используется следующая процедура дисперсионного анализа.

Пусть при *i*-м наборе независимых переменных проведено n_i повторных наблюдений переменной *Y*, $i=1,2,\ldots,m$. Объем всей выборки $n=\sum_{i=1}^{m} n_i$. Обозначим y_{ij}, $j=1,2,\ldots,n_i$, результаты повторных наблюдений *Y* при *i*-м наборе независимых переменных. Если модель адекватна данным, то средние $\overline{Y_i}=\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $i=1,2,\ldots,m$, должны быть близки к значениям $\tilde{y}_i$, предсказанным регрессионной моделью:

$$\widetilde{Y} = A\widetilde{\beta}.$$

Мерой неадекватности модели будет сумма квадратов:

$$\mathrm{Q}_n = \sum_{i=1}^{m} n_i (\bar{y}_i - \widetilde{y}_i)^2 . \qquad (28)$$

Чем меньше будет Q_n, тем лучше результаты наблюдений согласуются с моделью. Возведя обе части тождества $y_{ij} - \widetilde{y}_{ij} = (\bar{y}_i - \widetilde{y}_i) + (y_{ij} - \bar{y}_i)$ в квадрат и просуммировав их по i и по j, получим, что остаточная сумма квадратов может быть разбита на две суммы:

$$\mathrm{Q}_e = \sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - \widetilde{y}_j)^2 = \sum_{i=1}^{m} n_i (\bar{y}_i - \widetilde{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2, \quad (29)$$

$$\text{или } Q_e = Q_n + Q_p ,$$

где второе слагаемое Q_p называется суммой квадратов чистой ошибки.

Если модель адекватна, то статистики $\frac{Q_n}{\sigma^2}$ и $\frac{Q_p}{\sigma^2}$ независимы и имеют распределение χ^2 соответственно с $(m-k)$ и $(n-m)$ степенями свободы.

В этом случае статистика

$$F = \frac{\mathrm{Q}_n/(m-k)}{\mathrm{Q}_P/(n-m)} \qquad (30)$$

имеет распределение Фишера с $(m-k)$ и $(n-m)$ степенями свободы.

Вычисленное значение статистики (30) сравнивается с квантилью распределения Фишера $F_{1-\alpha}(m-k, n-m)$. Если $F < F_{1-\alpha}(m-k, n-m)$, то гипотеза об адекватности модели принимается на уровне значимости α.

6. Пример множественной регрессии

Рассмотрим следующий пример.

Руководство авиакомпании по результатам анализа деятельности 15 своих представительств получило следующие данные за март месяц:

Y	$X1$	$X2$	$X3$
79,3	2,5	10,0	3,0
200,1	5,5	8,0	6,0
163,2	6,0	12,0	9,0
200,1	7,9	7,0	16,0
146,0	5,2	8,0	15,0
177,7	7,6	12,0	9,0
30,9	2,0	12,0	8,0
291,9	9,0	5,0	10,0
160,0	4,0	8,0	4,0
339,4	9,6	5,0	16,0
159,6	5,5	11,0	7,0
88,3	3,0	12,0	8,0
237,5	6,0	6,0	10,0
107,2	5,0	10,0	4,0
155,0	3,5	10,0	4,0

где Y - общий доход от проданных билетов, млн руб.; $X1$ - средства на развитие компании в регионе, млн руб.; $X2$ - число конкурирующих компаний; $X3$ - процент пассажиров, летавших бесплатно по разным причинам.

Найти уравнение множественной регрессии по этим данным. Проверить адекватность модели по остаткам. Существенно ли влияет на доход число пассажиров, летавших бесплатно? Можно ли считать, что увеличение затрат на развитие компании в регионе на 1 млн руб. увеличит доход в 40 раз? Принять уровень значимости $\alpha = 0{,}05$.

Решение.

1. Для решения примера используем пакет Matlab.

Введем векторы-строки:

```
Y  = [79,3   200,1  ...  155];
X0 = ones(1, 15);
X1 = [2,5   5,5     ...  3,5];
X2 = [10   8        ...  10];
```

$X3$ = [3 6 ... 4].

Регрессионная матрица A = [$X0'$; $X1'$; $X2'$; $X3'$].

По формуле (9) найдем МНК-оценки:

b = inv (A'* A)*A'*Y'

b = 170,7600
25,4233
–13,0035
–2,7059.

Уравнение множественной регрессии имеет вид:

Y= 170,76 + 25,42$X1$ – 13$X2$ – 2,7$X3$.

Найдем остаточную сумму квадратов по формуле (19):

QE = Y*Y' – b'*A'*Y'

QE = 8499,9.

Оценка дисперсии ошибок наблюдений равна:

$S2$ = QE/(15 – 4)

$S2$ = 772,7161.

Проверим гипотезу о значимости регрессионной модели. Среднее арифметическое для Y равно:

$y1$ = mean(Y)

$y1$ = 169,0800.

Далее вычислим QY и QR по формуле (21):

QY = Y * Y' – 15 * ($y1$)^2

QY = 89220

QR = b' * A' * Y' – 15 * ($y1$)^2

QR = 80720.

Для проверки вычислений подставим результаты в тождество:

Qy = QR + Qe.

Вычислим статистику критерия по формуле (23):

F = (QR/(4 – 1))/(QE/(15 – 4))

F = 34,8211.

Квантиль распределения Фишера по табл.П7 [3] равна:

$F_{0,95}(3, 11) = 3,59$,

что меньше вычисленного значения статистики F, следовательно, гипотеза H_0 о незначимости модели отклоняется при $\alpha = 0,05$.

Коэффициент детерминации модели, определенный по формуле (24), равен:

$R2$ = QR/QY

$R2$= 0,9047.

Это означает, что 90,47 % дисперсии объясняется моделью.

Коэффициент множественной корреляции равен:

$$r = \text{sgrt}(R2),$$
$$r = 0{,}9512.$$

Найдем доверительные интервалы для параметров модели. Вычислим матрицу

$$B^{-1} = B1$$

$$B1 = \text{inv}\,(A'*A)$$

$$B1 = \begin{matrix} 3{,}5110 & -0{,}1992 & -0{,}2328 & -0{,}0279 \\ -0{,}1992 & 0{,}0306 & 0{,}0108 & -0{,}0078 \\ -0{,}2328 & 0{,}0108 & 0{,}0180 & 0{,}0012 \\ -0{,}0279 & -0{,}0078 & 0{,}0012 & 0{,}0069 \end{matrix}$$

Диагональные элементы матрицы $B^{-1} = B1$ равны:

$d = \text{diag}(B1)$

$$d = \begin{matrix} 3{,}5110 \\ 0{,}0306 \\ 0{,}0180 \\ 0{,}0069 \end{matrix}$$

Вычислим оценку среднего квадратического отклонения ошибки наблюдений:

$S = \text{sgrt}(S2)$

$S = 27{,}7978$

Значение квантили распределения Стьюдента найдем по табл.П6 [3]:

$$t_{0{,}975}(11) = T_{\text{tab}} = 2{,}201.$$

Нижние и верхние границы доверительных интервалов $\beta_1, \beta_2, \beta_3$ определим по формуле (17):

$b1 = b - T\text{tab}*S*\text{sgrt}(d)$

$b2 = b + T\text{tab}*S*\text{sgrt}(d)$

	b1	b2
β_1 :	14,7144	36,1321
β_2 :	–21,2104	–4,79660
β_3 :	–7,80273	2,39092.

Доверительный интервал для β_3 накрывает нуль, это означает, что на уровне значимости $\alpha = 0{,}05$ принимается гипотеза $H_0 : \beta_3 = 0$. Таким образом, пассажиры, летавшие бесплатно (переменная *X*3), не оказывают существенного влияния на доход компании.

Проверим гипотезу $H_0 : \beta_3 = 0$ другим способом, применив *F*-критерий (25). Вычислим параметры регрессии *Y* по переменным *X*1 и *X*2.

Регрессионная матрица *A*1 равна:

*A*1 = [*X*0'; *X*1'; *X*2'].

Найдем МНК-оценки по формуле (9)

*b*3 = inv(*A*1'**A*1)**A*1'**Y*'
*b*3 = 159,8629
22,3882
–12,5316

и остаточную сумму квадратов по формуле (19)

*QE*1= *Y*'**Y*' – *b*3'**A*1'**Y*'
*QE*1 = 9,5555

Вычислим статистику (25):

*F*1 = (*QE*1 – *QE*)/(*QE* / (15 – 4))
*F*1 = 1,3661.

Так как квантиль распределения Фишера [3] $F_{0,95}(1,11) = 4{,}84$, что больше выборочного значения статистики, то гипотеза H_0 принимается.

МНК-оценка коэффициента β_1 равна 25,4333, а 95%-ный доверительный интервал для β_1 составляет [14,7144; 36,1321]. Это означает, что при увеличении затрат на развитие компании в регионе на 1 млн руб. (переменная *X*1) доход возрастает на $1 \times 25{,}4333 =$ = 25,4333 млн руб., причем с вероятностью 0,95 эта величина может изменяться в пределах от 14,71 до 36,13 млн руб.

2. Рассмотрим решение примера в пакете Statgraphics (процедура K.3 Multiple Regression).

Результаты анализа оформим в виде таблицы:

Independent variable	Coefficient	Std.error	t-value	Sig.level
CONSTANT	170,76	52,09	3,28	0,0074
var2	25,42	4,86	5,22	0,0003
var3	–13,00	3,73	–3,49	0,0051
var4	–2,71	2,32	–1,17	0,2672

R-SQ. (ADJ) = 0,8787 SE = 27,797771 MAE = 20,094079 Durbwat = **2,211**

Во втором столбце таблицы приведены МНК-оценки параметров $\widetilde{\beta}_i$, в третьем - стандартные ошибки $\widetilde{\beta}_i$: $\sqrt{D[\widetilde{\beta}_i]} = S\sqrt{B_{ii}^{-1}}$; в четвертом - *t*-значения для проверки гипотезы $H_0 : \beta_i = 0$:

$$t_i = \frac{\widetilde{\beta}_i}{\sqrt{D[\widetilde{\beta}_i]}};$$

в пятом - уровни значимости $SL = P\left[T(n \times k) > |t_i|\right]$; если $SL < \alpha$, где α - заданный уровень значимости, то гипотеза $H_0 : \beta_i = 0$ отклоняется.

В данном случае гипотеза $H_0 : \beta_3 = 0$ принимается, так как соответствующий уровень значимости *SL* = 0,2672 превышает заданный уровень значимости $\alpha = 0{,}05$.

Далее в таблице описаны: R-SQ(ADJ) - приведенный коэффициент детерминации $\overline{R}^2$, средняя квадратичная ошибка наблюдений $SE = \sqrt{S^2}$, средняя абсолютная ошибка прогноза MAE и статистика Дарбина-Уотсона *d*, вычисляемая по формуле (27).

Если нажать функциональную клавишу F10, то на экране появится меню, содержащее опции для дальнейшего анализа:

1. Analysis of variance	(дисперсионный анализ)
2. Conditional sums of sguares	(условные суммы квадратов)
3. Plot residuals	(график остатков)
4. Summarize residuals	(статистики остатков)
5. Plot predicted values	(график предсказанных значений)
6. Probability plot	(проверка нормальности распределения остатков по их графику на вероятностной бумаге)
7. Component effects plot	(графики отдельных переменных)
8. Influence measures	(меры влияния)
9. Correlation matrix	(корреляционная матрица оценок коэффициентов регрессии)
10. Generate reports	(вывод на экран различных результатов)
11. Confidence intervals	(доверительные интервалы для коэффициентов регрессии)

12. Interval plots (доверительные интервалы в виде графиков)
13. Save results (сохранение результатов на диске)

Выбрав опцию 1, получим на экране таблицу дисперсионного анализа:

Analysis of Variance for the Full Regression

Source	Sum of Sguares	DF	Mean Sguare	F-Ratio	P-value
Model	80720,4	3	26906,8	34,8211	0,0000
Error	8499,88	11	772,716		
Total (Corr.)	89220,3	14			

R-squared = 0,904732 Std. error of est. = 27,7978
R-squared (Adj. for d.f.) = 0,878749 Durbin-Watson statistic = 2,21139

Во втором столбце таблицы приведены суммы квадратов Q_R, Q_e, Q_y, в пятом - статистика *F*, вычисляемая по формуле (23), а в шестом – *P*-значение, равное *SL*.

В данном случае гипотеза H_0 : модель незначима отклоняется для любого заданного уровня значимости α, так как *P*-значение равно нулю.

В опциях 12 и 11 меню вычисляются доверительные интервалы для

95 percent confidence intervals for coefficient estimates

Estimate	**Strandard error**	**Lower limit**	**Upper limit**
CONSTANT	170,760	52,0862	56,0889	285,431
.var2	5,4233	4,86420	14,7144	36,1321
.var3	–13,0035	3,72776	–21,2104	–4,7966
.var4	–2.70591	2,31510	–7,80273	2,3909

параметров модели и строятся их графики:

Correlation matrix for coefficient estimates

	CONST.	.var2	.var3	.var4
CONSTANT	**1,00**		**–,61**	**–,93**	**–,179**
var2	–,61	1,00	,46	–,53
var3	–,93	4612	1,00	1083
var4	–,18	–,53	1083	1,0

В опции 9 вычисляется корреляционная матрица оценок параметров:

График предсказанных значений зависимой переменной *Y* выводится в опции 5, а графики, показывающие влияния отдельных компонент, - в опции 7. Анализ остатков приводится в опциях 3, 4 и 6:

Приведенные статистики остатков показывают, что в данном примере модель адекватна данным. Графики остатков (опция 3) и нормальный вероятностный график остатков (опция 6) показывают, что распределение остатков хорошо аппроксимируется нормальным распределением $N(0,\sigma^2)$. Значение статистики Дарбина-Уотсона $d = 2{,}21$ и $4 - d = 1{,}79$ больше табличного критического значения

Residual Summary

Number of observations = 15 (2 missing values excluded)
Residual average = 6,01593E-14
Residual variance = 772,716
Residual standard error = 27,7978

Coeff. of skewness = –0,44422 standardized value = –0,70
Coeff. of kurtosis = –0,311981 standardized value = –0,25

$d_2 = 1{,}75$, что означает, что гипотеза H_0 о некоррелированности остатков при $H_1: \rho > 0$ и $H_1: \rho < 0$ принимается при $\alpha = 0{,}05$.

7. Вычислительные проблемы регрессионного анализа: мультиколлинеарность и плохая обусловленность информационной матрицы

Вычисление МНК-оценок параметров линейной регрессионной модели по формуле

$$\tilde{\beta} = \left(A^T A\right)^{-1} A^T Y$$

предполагает, что регрессионная матрица *A* имеет ранг *k*, где *k* - число параметров модели, а информационная матрица

$$B = A^T A$$

является невырожденной, т.е. определитель матрицы B не равен нулю: $|B| \neq 0$.

Если $|B| = 0$, то ранг матрицы A будет меньше k. Это условие является следствием того, что между столбцами матрицы A существует линейная зависимость, т.е. хотя бы один из них является линейной комбинацией других столбцов. Если столбцы матрицы A рассматривать как векторы в n-мерном линейном пространстве, то некоторые из них будут коллинеарны. Это явление называется *строгой мультиколлинеарностью*. Случай, когда линейная зависимость между столбцами матрицы выполняется лишь приблизительно, т.е. когда $|B| \approx 0$, называется *мультиколлинеарностью*. При этом одно или несколько собственных чисел матрицы $B = A^T A$ и определитель матрицы будут близки к нулю.

Основные следствия мультиколлинеарности таковы:

1) падает точность оценивания: ошибки оценок некоторых параметров становятся очень большими, резко возрастают дисперсии оценок;

2) коэффициенты регрессионной модели коррелированы между собой, что затрудняет их интерпретацию;

3) некоторые переменные становятся незначимыми и должны исключаться из модели, хотя истинная причина состоит не в том, что эти переменные не влияют на зависимую переменную, а в том, что выборочные данные не позволяют отобразить это влияние;

4) оценки параметров становятся неустойчивыми: добавление нескольких наблюдений приводит к большим изменениям в оценках параметров.

Мультиколлинеарность имеет место при больших значениях элементов обратной матрицы $B^{-1} = \left(A^T A\right)^{-1}$ и, следовательно, оценок параметров регрессии. В этом случае и предсказание, и остаточная сумма квадратов будут неточными.

Рассмотрим модель множественной регрессии

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \varepsilon. \qquad (31)$$

Преобразуем исходные данные: зависимую переменную Y и независимые переменные x_1, x_2,...,x_{k-1}, используя процедуру стандартизации

$$y'_i = \frac{y_i - \overline{y}}{\gamma_y},\ x'_{ji} = \frac{x_{ji} - \overline{x_j}}{\gamma_{x_j}},$$

где $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\gamma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$, $\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ji}$,

$\gamma_{x_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ji} - \overline{x_j}\right)^2}$, $j = 1,2,...,k-1$, $i = 1,2,...,n$.

Регрессионная модель в этом случае имеет вид:

$$y'_i = \beta'_1 x'_{1i} + \beta'_2 x'_{2i} + ... + \beta'_{k-1} x'_{k-1\,i} + \varepsilon_i . \qquad (32)$$

В этой модели свободный член (аналогичный β_0) отсутствует, так как если $x_{0i} = 1$, $i = 1,2,...,n$, то $\overline{x}_0 = 1$ и $x'_{0i} = 0$, $i = 1,2,...,n$.

Записав регрессионную матрицу A' и вектор Y' в виде

$$A' = \begin{pmatrix} x'_{11} & x'_{21} & \dots & x'_{k-11} \\ x'_{12} & x'_{22} & \dots & x'_{k-12} \\ \dots & \dots & \dots & \dots \\ x'_{1n} & x'_{2n} & \dots & x'_{k-1n} \end{pmatrix},\ Y' = \begin{pmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_n \end{pmatrix},$$

представим регрессионную модель так:

$$Y' = A' \cdot \beta' + \varepsilon .$$

Элементы информационной матрицы $(A')^T \cdot A' = B'$ равны:

$$B'_{fg} = \sum_{i=1}^{n} x'_{fi} \cdot x'_{gi} = \frac{\sum_{i=1}^{n}\left(x_{fi} - \overline{x}_f\right)\cdot\left(x_{gi} - \overline{x}_g\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{fi} - \overline{x}_f\right)^2\left(x_{gi} - \overline{x}_g\right)^2}},$$

где $f, g = 1,2,..., k-1$.

Таким образом, элементами матрицы B' являются коэффициенты корреляции между независимыми переменными $x_1, x_2,...,x_{k-1}$:

$$B' = R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1(k-1)} \\ r_{21} & 1 & \cdots & r_{2(k-1)} \\ \cdots & \cdots & \ddots & \cdots \\ r_{(k-1)1} & r_{(k-1)2} & \cdots & 1 \end{pmatrix}.$$

Между параметрами моделей (31) и (32) существуют простые соотношения:

$$\beta_j = \beta'_j \cdot \frac{\gamma_y}{\gamma_j}, \quad j = 1,2, \ldots, k-1,$$

$$\beta_0 = \overline{y} - \sum_{j=1}^{k-1} \left(\beta'_j \cdot \frac{\gamma_y}{\gamma_j} \right) \overline{x}_j = \overline{y} - \sum_{j=1}^{k-1} \left(\beta_j \cdot \overline{x}_j \right).$$

Если какой-либо внедиагональный элемент матрицы $R\left|r_{fg}\right| \geq 0{,}9$, то между соответствующими переменными x_f и x_g существует сильная линейная зависимость и возможны эффекты мультиколлинеарности.

Наличие мультиколлинеарности определяется и по определителю корреляционной матрицы: в этом случае $|R| << 1$, при строгой мультиколлинеарности $|R| = 0$.

Более точен следующий метод: переменную x_i относят к мультиколлинеарным, если коэффициент детерминации R_i^2, определяющий линейную зависимость переменной x_i от всех других аргументов, больше коэффициента детерминации R^2 между зависимой переменной y и всеми аргументами.

Для каждой переменной x_i вычисляют F-статистику

$$F_i = \frac{R_i^2 / (k-2)}{(1-R_i^2) / (n-k+1)}, \quad i = 1, 2, \ldots, k,$$

и проверяют значимость каждой статистики F_i, сравнивая ее с квантилью распределения Фишера $F_{1-\alpha}$ ($k - 2$, $n - k + 1$). Значения F_i показывают, какие из переменных могут быть в большей степени подвержены мультиколлинеарности.

В линейной алгебре для определения мультиколлинеарности используют числа обусловленности. Число обусловленности определяется как отношение максимального собственного числа $\lambda_{\max}$ матрицы $B = (A^T A)$ к ее минимальному собственному числу $\lambda_{\min}$.

Можно считать, что при $\frac{\lambda_{\max}}{\lambda_{\min}} \geq 10^5 \div 10^6$ имеет место сильная мультиколлинеарность. Плохая обусловленность матрицы B часто проявляется при использовании полиномиальных моделей, особенно если измерения независимой переменной x выполняются через равные интервалы, а при высоких степенях (выше 4) проявляется почти всегда.

Для устранения или уменьшения мультиколлинеарности используются следующие меры:

1) привлечение дополнительной информации;

2) преобразование множества независимых переменных в несколько ортогональных множеств. При этом применяются методы многомерного статистического анализа: факторный анализ и метод главных компонент, а также специальные процедуры регрессионного анализа: регрессия на главные компоненты, гребневая регрессия;

3) стандартизация и центрирование исходных данных;

4) исключение из рассмотрения одного или нескольких линейно связанных аргументов, это можно делать только после детального анализа данных.

Литература

1. ***Дрейпер Н., Смит Г.*** Прикладной регрессионный анализ: В 2-х ч. - М.: Финансы и статистика, 1986.

2. **Себер Дж.** Линейный регрессионный анализ. - М.: Мир, 1980.

3. Сборник задач по математике для ВТУЗов. Теория вероятностей и математическая статистика /Под ред. ***А.В.Ефимова.*** - 2-е изд. - М.: Наука, 1990. - Т.3.

4. ***Джонстон Дж.*** Эконометрические методы. - М.: Статистика, 1980.

5. ***Levin R., Rubin D.*** Statics for Management. - Prentice-Hall, Inc., 1991.

Приложение 1. Задачи для самостоятельного решения*

1. Консультанта корпорации интересует, насколько аккуратно сделаны некоторые измерения. Один из способов проверки - исследовать отношение между индексом выполнения работ и окладами служащих. Были выбраны 8 служащих, собрана информация об их окладах и индексы выполнения работ (от 1 до 10):

Индекс выполнения работ X	9	7	8	4	7	5	5	6
Оклад, тыс.руб.	36	25	33	15	28	19	20	22

*Часть задач взята из [5].

1. Найдите уравнение линейной регрессии.

2. Рассчитайте стандартную ошибку *S* для этих данных.

3. Рассчитайте выборочный коэффициент детерминации.

2. Общество изучения аистов в противовес распространенному мнению надеется доказать с помощью статистики, что аисты все-таки *приносят* детей. Были собраны данные о количестве детей и аистов в нескольких больших городах Центральной Европы:

Аисты *X*, тыс. птиц	27	38	13	24	6	19	15
Дети *Y*, тыс. чел.	35	46	19	32	15	31	20

1. Подсчитайте коэффициенты детерминации и корреляции.

2. Опровергает ли статистика распространенное мнение об аистах?

3. Заполните пробелы:

"Регрессионный и корреляционный анализ изучает ... между переменными. Регрессионный анализ с помощью ... уравнений позволяет нам ... неизвестную переменную с помощью множества известных. Неизвестная переменная называется ... переменной. Известные переменные называются ... переменными. Корреляция между двумя переменными показывает ... линейной зависимости и таким образом дает понятие о том, насколько ... в регрессии выражается зависимость между переменными".

4. Президента компании интересует зависимость между приростом годового дохода и качеством работы коммерческих агентов в будущем году. Он выбрал 12 агентов и определил размеры дохода, приносимого компании каждым из них (в процентах от окладов), а также количество продаж, проведенных каждым агентом в течение года:

Размер дохода *X*, %	7,8	6,9	6,7	6,0	6,9	5,2	6,3	8,4	7,2	10,1	10,8	7,7
Количество продаж *Y*	64	73	42	49	71	46	32	88	53	84	85	93

1. Найдите уравнение линейной регрессии.

2. Рассчитайте стандартную ошибку *S* для этих данных.

3. Постройте 90%-ный доверительный интервал для числа продаж, проведенных агентом с уровнем дохода свыше 9,6 %.

5. Менеджер анализирует зависимость стоимости продукции *C* от количества исходного сырья *R*. Если угловой коэффициент линейной регрессии для приведенных ниже данных значительно больше 0,5, то в производственном процессе есть недостатки и

конвейер необходимо отрегулировать. Нужно ли регулировать конвейер при уровне значимости 0,05? Сформулируйте соответствующие гипотезы и сделайте выводы.

C, *тыс. руб.*	10	7	5	6	7	6
R, *сотни кг*	25	20	16	17	19	18

6. Экстраполяция для прогнозирования значений вне множества данных, используемых для построения прямой регрессии, ненадежна. Укажите причину этого (выберите одну):

1) зависимость между переменными различна для разных значений переменных;

2) независимая переменная не имеет причинной связи с зависимой переменной;

3) значения переменных изменяются с течением времени;

4) не существует прямой связи для объяснения зависимости.

7. Экономисты часто заинтересованы в оценке потребительских функций, что достигается посредством анализа зависимости между потреблением и доходами. В этом случае экономисты называют коэффициент регрессии предельной склонностью к потреблению. На примере 25 семей были рассчитаны коэффициент регрессии (0,87) и стандартная ошибка коэффициента регрессии (0,35). Определите, упала ли предельная склонность к потреблению ниже стандартной (0,94). Уровень значимости равен 0,05. Сформулируйте соответствующие гипотезы и сделайте выводы.

8. В отличие от коэффициента детерминации коэффициент корреляции (выберите одно):

1) показывает, какой (положительный или отрицательный) наклон имеет прямая регрессии;

2) измеряет степень соотношения между двумя переменными более точно;

3) никогда не может иметь абсолютную величину больше 1;

4) измеряет в процентах долю дисперсии, объясняемую регрессией.

9. Чтобы ответить на вопрос: важно ли иметь хорошие успехи в колледже для получения хорошего оклада в будущем, студент, изучающий экономическую статистику, произвольно выбрал несколько начальных окладов своих друзей, недавно окончивших колледж, и уровней их оценок в колледже:

Начальный оклад *x*, тыс. руб.	36	30	30	24	27	33	21	27
Уровень оценок *y*, баллы	4	3	3,5	2	3	3,5	2,5	2,5

1. Постройте диаграмму рассеяния.
2. Найдите уравнение линейной регрессии.
3. Начертите график этого уравнения на диаграмме рассеяния.

10. Арендодатель интересуется, типичны ли его расценки на квартиры. Он случайным образом взял 11 расценок и количество спален в 11 квартирах нескольких жилых комплексов:

Расценки *x*, долл.	230	190	450	310	218	185	340	245	125	350	280
Количество спален *y*	2	1	3	2	2	2	2	1	1	2	2

1. Найдите уравнение линейной регрессии.
2. Рассчитайте коэффициент детерминации и определите расценки на квартиры с двумя спальнями.

11. Установлено, что годовой доход фирмы сильно колеблется в течение последних нескольких лет. На это влияют многочисленные причины, так что трудно определить, какая из них оказывает наибольшее влияние на доход от продаж. Отдел изучения рынка рассмотрел множество соотношений и уверен, что наибольшее влияние оказывают ежемесячные расходы на рекламу. В течение семи месяцев были получены следующие данные:

Ежемесячные расходы на рекламу *X*, сотни тыс. руб.	25	16	42	34	10	21	19
Ежемесячный доход от продаж *Y*, млн руб.	34	14	48	32	26	29	20

1. Найдите уравнение линейной регрессии.
2. Рассчитайте стандартную ошибку для этих данных.
3. Найдите 90%-ный доверительный интервал для ожидаемого объема продаж, если ежемесячные расходы на рекламу составляют 28 сотен тыс. руб.

12. В лаборатории по изучению воздушного транспорта при анализе деятельности 18 компаний было установлено, что коэффициент регрессии между количеством пилотов и количеством задействованных самолетов составляет 4,3. Согласно предварительным исследованиям

этот коэффициент равен 4,0. Если стандартная ошибка регрессионного коэффициента была рассчитана как 0,17, есть ли причина утверждать на уровне значимости 0,05, что реальное значение коэффициента регрессии изменилось?

13. Используя приведенные ниже данные, установите, есть ли значимая зависимость между объемом инвестиций и ценой за акцию:

Объем инвестиций *X*, млн руб.	108	4,4	3,5	3,6	39	68,4	7,5	5,5	375	12	51
Цена за акцию *Y*, руб.	12	4	5	6	13	19	8,5	5	15	6	12

14. Проводятся испытания новой модели - радарного детектора на батарейках. В лаборатории собраны следующие данные:

Время ежедневной работы детектора, ч	Продолжительность безотказной работы детектора, мес.	
	на Li-батарейках	на щелочных батарейках
2	3,1	1,3
1,5	4,2	1,6
1	5,1	1,8
0,5	6,3	2,2

1. Составьте два линейных уравнения регрессии: одно - прогнозирующее продолжительность безотказной работы радара при ежедневном использовании с Li-батарейкой, другое - со щелочной батарейкой.

2. Рассчитайте 90%-ный доверительный интервал продолжительности работы детектора при ежедневном использовании в течение 1,25 часа для каждого типа батареек.

3. Может ли лаборатория на основании этих цифр установить, с какими батарейками детектор будет служить дольше?

15. Были проведены исследования, чтобы установить соотношение между весом новорожденных мальчиков и их взрослым ростом. Используя нижеприведенные данные, рассчитайте уравнение прямой регрессии. Какой процент дисперсии наблюдений описывается этой прямой регрессии?

Вес новорожденных, г	2494	3175	2830	3395	3680	3054
Взрослый рост, см	175	182	167	179	185	177

16. Чтобы правильно оценить потенциал студентов 3-го курса колледжа, декан проводит сравнительный анализ средних оценок за первые два и последние два года обучения:

Оценки 1-го и 2-го курсов *x*, баллы	3,7	4,5	4,3	4,6	4,0	3,8	4,4	3,9	4,0	4,7
Оценки 3-го и 4-го курсов *y*, баллы	4,4	4,7	4,0	4,5	4,9	4,0	4,5	3,8	4,7	4,8

1. Постройте уравнение линейной регрессии, которое декан должен использовать для прогноза средних оценок на старших курсах колледжа.

2. Если нижняя граница 90%-ного доверительного интервала (для оценок на старших курсах) не будет выше 4,1, декан не переведет студента на 3-й курс. Переведет ли он студента со средним уровнем оценок на первых курсах 4,4?

17. Используя приведенные ниже данные, с помощью пакета программ найдите уравнение множественной регрессии и ответьте на следующие вопросы:

1) каковы оценки коэффициентов регрессии и стандартные ошибки этих оценок?

2) каков коэффициент детерминации?

3) каково ожидаемое или прогнозируемое значение для *Y* при *x*1 = 5,8, *x*2 = 4,2, *x*3 = 5,1?

Y	*x*1	*x*2	*x*3
64,7	3,5	5,3	8,5
80,9	7,4	1,6	2,6
24,6	2,5	6,3	4,5
43,9	3,7	9,4	8,8
77,7	5,5	1,4	3,6
20,6	8,3	9,2	2,5
66,9	6,7	2,5	2,7
34,3	1,2	2,2	1,3

18. Используя приведенные ниже данные, с помощью пакета программ найдите уравнение множественной регрессии и ответьте на следующие вопросы:

1) каковы оценки коэффициентов регрессии и стандартные ошибки этих оценок?

2) каков коэффициент детерминации?

3) каков 95%-ный доверительный интервал для значения *Y* при *x*1, *x*2, *x*3 и *x*4, равных 52,4; 41,6; 35,8; 3 соответственно?

*x*1	*x*2	*x*3	*x*4	**Y**
21,4	62,9	21,9	–2	22,8
51,7	40,7	42,9	5	93,7
41,8	81,8	69,8	2	64,9
11,8	41,0	90,9	–4	19,2
71,6	22,6	12,9	8	55,8
91,9	61,5	30,9	1	23,1

19. Владелец бухгалтерской фирмы считает, что целесообразно прогнозировать заранее количество налоговых деклараций, приходящихся на период с 1 марта по 15 апреля, так как в этом случае он сможет лучше спланировать работу на этот период. Он предполагает, что многие факторы могут быть использованы при таком прогнозе. Данные об этих факторах и количестве налоговых деклараций приведены ниже:

Экономический индекс $x1$	Население в радиусе 1 км от фирмы $x2$, тыс. чел.	Средний доход в районе $x3$, тыс. руб.	Количество деклараций на период с 1 марта по 15 апреля Y, тыс. шт.
99	10,188	21,465	2,306
106	8,566	22,228	1,266
100	10,557	27,665	1,422
129	10,219	25,200	1,721
179	9,662	26,300	2,544

1. Используя любой пакет программ, определите уравнение множественной регрессии для этих данных.

2. Какой процент дисперсии данных описывается этим уравнением?

3. В 1996 году экономический индекс был 169, население в радиусе 1 км - 10212 чел., средний доход - 26925 руб. Какое количество деклараций должен ожидать владелец фирмы с 1 марта по 15 апреля?

20. Мы пытаемся предсказать годовой спрос на продукцию, используя следующие независимые переменные:

- цена за одну единицу продукции, руб.;
- доход потребителя, руб.;
- замена (цена на заменитель этого товара), руб.

Данные были собраны с 1975 по 1988 год.

Год	Спрос, шт.	Цена	Доход	Замена
1975	40	9	400	10
1976	45	8	500	14
1977	55	8	700	13
1978	60	7	800	11
1979	70	6	900	15
1980	65	6	1000	16

1981	65	8	1100	17
Год	Спрос, шт.	Цена	Доход	Замена
1982	75	5	1200	22
1983	75	5	1300	19
1984	80	5	1400	20
1985	100	3	1500	23
1986	90	4	1600	18
1987	95	3	1700	24
1988	85	4	1800	21

1. Можно ли точно определить знак (“+” или “–”) регрессионных коэффициентов для независимых переменных? Дайте краткое объяснение. (Заметьте, это не статистический вопрос. Вы должны просто подумать о смысле регрессионного коэффициента.)

2. Используя любой пакет программ, найдите уравнение множественной регрессии.

3. Найдите коэффициент множественной детерминации для данного примера и объясните его смысл.

4. Найдите стандартную ошибку оценки для данного примера и объясните ее смысл.

5. Используя уравнение, оцените, какой спрос можно ожидать, если цена - 6 руб., доход - 1200 руб. и цена на заменитель - 17 руб.

21. Профессор статистики высшей экономической школы заинтересован в том, чтобы определить, какие факторы влияют на успехи студентов на экзаменах. Промежуточный экзамен в последнем семестре дал широкий разброс оценок, но профессор догадывается, каковы факторы, объясняющие этот разброс:

- он позволял студентам обучаться по любому нужному им количеству книг;
- коэффициент интеллекта IQ студентов различен;
- студенты разного возраста;
- они тратят разное время на подготовку к экзамену.

Чтобы составить формулу прогноза оценок на экзамене, профессор попросил каждого студента ответить после экзамена, сколько времени он потратил на подготовку и сколько книг использовал. Профессор также располагает данными о коэффициенте интеллекта и о возрасте студентов. Сопоставив эти данные для всей группы и определив уравнение множественной регрессии с помощью пакета StatGraphics, он получил выходные данные этой программы:

ROOT MSE = 11,66КОЭФ. ДЕТЕРМ. = 0,77

ПЕРЕМЕННЫЕ	DF	ПАРАМЕТР	СТАНДАР. ОШИБКА	*T*-СТАТИСТИКА ДЛЯ ПРОВЕРКИ H_0: ПАРАМЕТР=0	УРОВЕНЬ ЗНАЧИМ.
СВ. ЧЛЕН	1	–49,95	41,55	–1,202	0,2684
ЧАСЫ	1	1,07	0,98	1,089	0,3121
IQ	1	1,36	0,38	3,627	0,0084
КНИГИ	1	2,04	1,51	1,353	0,2182
ВОЗРАСТ	1	–1,78	0,67	–2,672	0,0319

1. Напишите уравнение множественной регрессии.

2. Какой процент дисперсии оценок описывается этим уравнением?

3. Какой оценки может ожидать студент в возрасте 21 года, с коэффициентом интеллекта 113, который занимался по 5 часов и использовал 3 книги?

22. Компания планирует расширить сеть своих магазинов. Чтобы выбрать место расположения новых магазинов, она собрала данные о еженедельных продажах каждого из 23 магазинов. Были собраны также данные о переменных, которые, по мнению компании, влияют на продажи. Переменные таковы:

- продажи - средний еженедельный объем продаж для каждого магазина, тыс. руб.;
- транспорт - средний еженедельный объем уличного движения, тыс. автомобилей;
- вход - возможность посещения магазина, измеренная от 1 до 100;
- годовой доход - среднегодовой доход домашних хозяйств в данном районе, млн руб.;
- расстояние. Расстояние от магазина до ближайшего супермаркета, км.

Данные были проанализированы с помощью пакета StatGraphics. Выходные данные приведены ниже:

ROOT MSE = 85.5865 КОЭФ.ДЕТЕРМ. = 0,9579

ПЕРЕМЕННЫЕ	DF	ПАРАМЕТР	СТАНДАРТ. ОШИБКА	*T*-СТАТИСТИКА ДЛЯ ПРОВЕРКИ H_0: ПАРАМЕТР=0		УРОВЕНЬ ЗНАЧИМ.
ПРОДАЖИ	1	175,37	92,62	1,90		0,07
ТРАНСПОРТ	1	–0,03	0,31	–0,09	0,93	
ВХОД	1	3,78	1,27	2,97	0,01	
ГОДОВОЙ ДОХОД	1	2,00	4,51	0,44	0,66	
РАССТОЯНИЕ	1	212,41	28,09	7,56	0,00	

1. Каково регрессионное уравнение, рассчитанное с помощью StatGraphics?

2. Какова стандартная ошибка оценки для этого уравнения?

3. Какой процент дисперсии данных описывается этой регрессией?

4. Какой объем продаж можно предсказать для расположенного по соседству магазина, если его среднегодовой доход 20 млн руб., расстояние до ближайшего супермаркета - 2 км, объем движения - 100000 машин, а возможность посещения - 50?

23. Леня Голубков собирается продать дом. Чтобы решить, какую цену запросить, он собрал данные о 12 недавних продажах. Он принимал во внимание цену, площадь дома, количество этажей, количество ванных и возраст дома.

Цена, тыс. долл.	Площадь, сотни кв. футов	Этажность	Количество ванных	Возраст дома, года
49,65	8,9	1	1,0	2
67,95	9,5	1	1,0	6
81,15	12,6	2	1,5	11
81,60	12,9	2	1,5	8
91,50	19,0	2	1,0	22
95,25	17,6	1	1,0	17
100,35	20,0	2	1,5	12
104,25	20,6	2	1,5	11
112,65	20,5	1	2,0	9
149,70	25,1	2	2,0	8
160,65	22,7	2	2,0	18
232,50	40,8	3	4,0	12

1. Используя любой пакет программ, найдите уравнение множественной регрессии.

2. Каков коэффициент детерминации для этого уравнения и что он определяет?

3. Если дом имеет площадь 1800 кв. футов, один этаж, полторы ванны и возраст 6 лет, по какой цене Леня сможет продать дом?

24. Сталелитейная корпорация рассматривает факторы, влияющие на объем ежегодно продаваемой стали. Руководство предполагает, что важнейшими являются следующие факторы: годовой национальный уровень инфляции, средняя цена за тонну импортируемой стали,

сбивающая цены корпорации, количество автомобилей, которое планируют выпустить производители машин в данном году. Были собраны следующие данные за 7 лет:

	Продано, млн т	Темп инфляции, %	Средняя цена за тонну импортируемой стали, тыс. долл.	Количество машин, млн шт.
1985	4,2	3,1	3,10	6,2
1984	3,1	3,9	5,00	5,1
1983	4,0	7,5	2,20	5,7
1982	4,7	10,7	4,50	7,1
1981	4,3	15,5	4,35	6,5
1980	3,7	13,0	2,60	6,1
1979	3,5	11,0	3,05	5,9

1. Используя любой пакет программ, найдите уравнение множественной регрессии для этих данных.

2. Какой процент дисперсии данных описывается этим уравнением?

3. Сколько тонн стали сможет продать корпорация, если в данном году темп инфляции будет 7,1 %, автомобилестроители планируют выпустить 6 млн машин, а средняя цена импортируемой стали - 3,5 тыс. долл. за тонну?

Приложение 2.Таблица критических точек критерия Дарбина-Уотсона

Критические точки d_1 и d_2 для уровня 5 % $(\alpha = 0{,}05)$

n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	0,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	0,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77

65	1,57	1,62	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,75	1,78

Содержание

Вуколов *Эдуард Александрович*

Регрессионный анализ

Методические указания по курсу «Статистика»

Редактор *Е.Г.Кузнецова*. Выпускающий редактор *С.В.Козинцева*. Технический редактор *Е.Н.Романова*. Корректор *Л.Г.Лосякова.*

ЛР № 020516 от 12.05.97. Подписано в печать с оригинала-макета 25.12.2000. Формат 60×84 1/16. Печать офсетная. Усл. печ. л. 3,02. Уч.-изд. л. 2,6. Тираж 200 экз. Заказ 64.

Отпечатано в типографии МИЭТ.
103498, Москва, МИЭТ.