ФГБОУ ВПО «Воронежский государственный
технический университет»

Кафедра полупроводниковой электроники и наноэлектроники

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к лабораторным работам №1-4 по дисциплине
«Современные технологии обработки информации»
для бакалавров направления 152200 «Наноинженерия»
(профиль «Инженерные нанотехнологии
в приборостроении») очной формы обучения

Воронеж 2014

Составитель д-р техн. наук К. А. Разинкин

УДК 004.622

Методические указания к лабораторным работам № 1-4 по дисциплине «Современные технологии обработки информации» для бакалавров направления 152200 «Наноинженерия» (профиль «Инженерные нанотехнологии в приборостроении») очной формы обучения / ФГБОУ ВПО «Воронежский государственный технический университет»; сост. К. А. Разинкин. – Воронеж, 2014. 70 с.

Методические указания посвящены практическому изучению современных подходов, связанных с обработкой и построением зависимостей на основе статической и экспертной информации. В указаниях приведены основные положения изучаемого метода и пример решения типовой задачи по каждой из лабораторных работ.

Методические указания подготовлены в электронном виде в текстовом редакторе и содержатся в файле Разинкин_ЛР_СТОИ_№1-4.pdf.

Табл. 18. Ил. 37. Библиогр.: 7 назв.

Рецензент д-р техн. наук, проф. И.Я. Львович

Ответственный за выпуск зав. кафедрой д-р физ.-мат. наук, проф. С.И. Рембеза

Издается по решению редакционно-издательского совета Воронежского государственного технического университета

ВВЕДЕНИЕ

Целями изучения дисциплины «Современные технологии обработки информации» являются:

1) практическое освоение современных технологий обработки информации на основе статистических методов.

2) изучение дисциплины должно способствовать формированию у студентов основ научного мышления, в том числе: пониманию значения современных технологий обработки информации и осознания широкого спектра применимости указанных технологий в научной и инженерной деятельности.

Для достижения целей ставятся задачи:

- исследование средств современных технологий обработки информации в структуре современных подходов к организации и планированию научных исследований;

- формирование представлений об основных статистических методах, задачах прикладной статистики при обработке результатов эксперимента и принятии решений;

В методичке рассмотрены такие разделы современной технологии обработки информации как множественный регрессионный анализ, в том числе на основе реализации фунций пошаговой регрессии, кластерный анализ, и анализ временных рядов.

Лабораторная работа № 1
Регрессионный анализ. Линейная регрессия

Общие сведения

Анализ взаимосвязей, присущих изучаемым процессам и явлениям, — важнейшая задача многих исследований. В тех случаях, когда речь идет о явлениях и процессах, обладающих сложной структурой и многообразием свойственных им связей, такой анализ представляется сложным. Прежде всего, необходимо установить наличие взаимосвязей и их характер. Вслед за этим возникает вопрос о тесноте взаимосвязей и степени воздействия различных факторов (причин) на интересующий исследователя результат. Если черты и свойства изучаемых объектов могут быть измерены и выражены количественно, то анализ взаимосвязей может вестись с применением математических методов, что позволяет проверить гипотезу о наличии или отсутствии взаимосвязей между теми или иными признаками, выдвигаемую на основе содержательного анализа. Далее, лишь посредством математических методов можно установить тесноту и характер взаимосвязей или выявить силу (степень) воздействия различных факторов на результат. В таких исследованиях широко используются процедуры множественной регрессии. Регрессионный анализ тесно связан с другими статистическими методами — методами корреляционного и дисперсионного анализа. В отличие от корреляционного анализа, который изучает направление и силу статистической связи признаков, регрессионный анализ изучает вид зависимости признаков, т.е. параметры функции зависимости одного признака от одного или нескольких других признаков. В отличие от дисперсионного анализа, с помощью которого исследуется зависимость количественного признака от одного или нескольких качественных признаков, в регрессионном анализе обычно исследуется зависимость (количественного или качественного признака) от одного или нескольких количественных признаков [1].

Таким образом, в регрессионном анализе рассматривается односторонняя зависимость случайной зависимой пере-

менной от одной или нескольких независимых переменных. Независимые переменные называются факторами, или предикторами, а зависимая переменная — результативным признаком, или откликом.

Если число предикторов равно 1, регрессию называют простой, если число предикторов больше 1 — множественной. Множественная регрессия позволяет исследователю задать вопрос (и, вероятно, получить ответ) о том, «что является лучшим предиктором для...». Например, исследователь в области образования мог бы пожелать узнать, какие факторы являются лучшими предикторами успешной учебы в средней школе. А психолога мог быть заинтересовать вопрос, какие индивидуальные качества позволяют лучше предсказать степень социальной адаптации индивида.

Если в ходе количественного анализа выявлена и обоснована зависимость одного явления от других, то задача регрессионного анализа — измерение зависимости, в которой причинно-следственный механизм выступает в наглядной форме. Прогноз в этом случае лучше поддается содержательной интерпретации, становится более ясным воздействие отдельных факторов, и исследователь лучше понимает природу изучаемого явления. Кроме того, регрессии создают базу для расчетного экспериментирования с целью получения ответов на вопросы типа «Что будет, если...?». Регрессионный анализ предполагает решение двух задач.

Первая заключается в выборе независимых переменных, существенно влияющих на зависимую величину, и определении формы уравнения регрессии. Данная задача решается путем анализа изучаемой взаимосвязи.

Вторая задача — оценивание параметров — решается с помощью того или иного статистического метода обработки данных наблюдения.

Функция *F(X),* описывающая зависимость условного среднего значения результативного признака *Y* от заданных значений фактора, называется функцией (уравнением) регрессии [9]. Для точного описания уравнения регрессии необ-

ходимо знать условный закон распределения результативного признака *Y*. В статистической практике такую информацию получить обычно не удается, поэтому ограничиваются поиском подходящих аппроксимаций для функции *F(X),* основанных на исходных статистических данных. Значения переменной *X* в *i*-м опыте будем обозначать через x_i, соответствующие им значения величины *Y* — через $y_i, i = 1,..., n.$

Рассмотрим самую простую регрессионную модель — линейную. Для линейной модели предполагается, что наблюдаемые величины связаны между собой зависимостью вида

$$y_i = b_0 + b_1 x_i + c_i$$

где b_0, b_1 — неизвестные параметры (коэффициенты уравнения), c_i — независимые нормально распределенные случайные величины с нулевым математическим ожиданием и дисперсией σ^2. Иногда c_i называют ошибками наблюдения. Общая задача регрессионного анализа состоит в том, чтобы по наблюдениям $x_i y_i$ оценить параметры модели b_1, b_0 «наилучшим образом»; построить доверительные интервалы для b_1, b_0 проверить гипотезу о значимости уравнения и коэффициентов регрессии; оценить степень адекватности, полученной зависимости и т.д.

Если под «наилучшим образом» понимать минимальную сумму квадратов расстояний до прямой от наблюдаемых точек, вычисленных вдоль оси ординат, то такой метод построения уравнения регрессии называется методом наименьших квадратов. В качестве меры «наилучшим образом» можно использовать минимум суммы квадратов расстояний от точек до прямой, вычисленных вдоль оси абсцисс; минимум суммы квадратов расстояний длин перпендикуляров, опущенных из точек на прямую и т.д.

Линейная модель с несколькими предикторами называется линейной множественной регрессионной моделью, а именно:

$$y_i = b_1 x_{1i} + b_2 x_{2i} + ... + b_p x_{pi} + b_0 + c_i$$

где $b_0, b_1, b_2, ,..., b_p$ - неизвестные параметры модели, которые вычисляются при помощи систем нормальных уравнений. Например, система нормальных уравнений для регрессии с двумя предикторами имеет следующий вид [2]:

$$\begin{cases} nb_0 + b_1 \sum_i x_{1i} + b_2 \sum_i x_{2i} = \sum_i y_i \\ b_0 \sum_i x_{1i} + b_1 \sum_i x_{1i}^2 + b_2 \sum_i x_{1i} x_{2i} = \sum_i x_{1i} y_i \\ b_0 \sum_i x_{2i} + b_1 \sum_i x_{1i} x_{2i} + b_2 \sum_i x_{2i}^2 = \sum_i x_{2i} y_i \end{cases}$$

Описание модуля Multiple Regression (пакет STATISTICA)

Кратко рассмотрим основные обозначения и понятия, используемые в модуле **Multiple Regression** (множественная регрессия).

Predictable values (предсказанные значения) — значения Y, вычисленные по уравнению регрессии. Обозначим их $\Pr Y_i$.

Residuals (остатки) — разность между наблюдаемыми значениями и предсказанными: $\mathrm{Re}\, s = Y_i - \Pr Y_i$

SS (сумма квадратов Y_i, скорректированная на среднее):

$SS = \sum_i (Y_i - Y)^2$, где $Y = \sum_i Y_i / n.$

SSPr (сумма квадратов $\Pr Y_i$, скорректированная на среднее): $SS\,\Pr = \sum_i (\Pr Y_i - Y)^2$, *SSRes* (сумма квадратов остатков): $SS\,\mathrm{Re}\, s = \sum_i (\Pr Y_i - Y_i)^2$

$R^2 = 1 - SS\,\mathrm{Re}\, s / SS$ (коэффициент детерминации). Чем меньше разброс значений остатков около линии регрессии по отношению к общему разбросу значений, тем, очевидно, лучше прогноз. Например, если связь между предиктором *X* и откликом *Y* отсутствует, то отношение остаточной изменчивости переменной У к исходной дисперсии равно 1. Если X *и* Y связаны функциональной зависимостью, то остаточная изменчивость отсутствует, и отношение дисперсий будет равно 0. В

общем случае отношение будет лежать между этими экстремальными значениями, т.е. между 0 и 1. Коэффициент детерминации *R* интерпретируется следующим образом. Если, например, R^2 = 0,4, то изменчивость значений переменной *Y* около линии регрессии составляет 1-0,4 от исходной дисперсии; другими словами, 40% от исходной изменчивости могут быть объяснены, а 60% остаточной изменчивости остаются необъясненными. В идеале желательно иметь объяснение если не *для* всей, то хотя бы для большей части исходной изменчивости. Значение R^2 является индикатором степени подгонки модели к данным (значение R^2 близкое к 1, показывает, что модель объясняет почти всю изменчивость соответствующих переменных).

$R = \sqrt{R^2}$ — коэффициент множественной корреляции. Характеризует тесноту связи между предикторами и откликом, а также является оценкой качества предсказания. Изменяется в пределах от 0 до 1.

Adjusted $R^2 = 1 - (1 - R^2)(n/(n - k))$ — скорректированное R^2, где *k* — число параметров в регрессионном уравнении.

Задание и тестовый пример построения простой линейной регрессии

По выборке из заданий к лабораторной работе (стр.19) выполнить следующие расчеты и задания[4]:

1. Построить диаграмму рассеяния выборки (построение сделать точно на бумаге в клеточку или миллиметровке).

2. Вычислить оценки параметров линейной регрессии *Y* на *x:* $y = \beta_0 + \beta_1 x$ и *X* на $y : x = \beta_0' + \beta_1' y$, используя суммы квадратов Q_y, Q_x, Q_{xy} по формулам 1 и 2

$$\tilde{\beta}_1 = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{Q_{xy}}{Q_x} \qquad (1)$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}. \qquad (2)$$

3. Нанести графики прямых регрессий Y на x и X на y на диаграмму рассеяния.

4. Для линейной регрессии Y на x вычислить остатки $e_i, i = 1,2,..., n.$; остаточную сумму квадратов $Q_e = \sum e_i^2$ оценку дисперсии ошибок наблюдений S^2, коэффициент детерминации R^2 и оценку коэффициента корреляции r.

5. Для линейной регрессии $y = \beta_0 + \beta_1 x$ выписать матрицу A, транспонированную матрицу A^T, информационную матрицу $B=A^TA$ и обратную матрицу к матрице $B= (A^TA)$. Найти оценки β_0 и β_1, используя формулу для расчета оценок в матричном виде.

$$\tilde{\beta} = \left(A^T A\right)^{-1} A^T Y \qquad (3)$$

6. Ввести данные в пакет STATISTICA, выполнить п. 1—4, сравнить результаты расчетов и полученные графики, записать в отчет результаты.

Чтобы показать технику вычислений, рассмотрим пример расчета простой линейной регрессии с небольшим объемом данных.

Пример 1. Пример простой линейной регрессии Y на x. Исходные данные: результаты наблюдений зависимой переменной (y) и фактора (x) следующие:

У	Х
4,0	5,5
5,6	8,1
5,7	8,5
3,6	5,9
4,0	7,8

Решение.

1.По данным примера вычислим суммы квадратов Q_y, Q_x и сумму произведений Q_{xy}, n=5. Предварительно найдем средние значения:

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{5}(5{,}5 + 8{,}1 + 8{,}5 + 5{,}9 + 7{,}8) = 7{,}16;$$

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{5}(4 + 5{,}6 + 5{,}7 + 3{,}6 + 4) = 4{,}58;$$

$$Q_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2 = (5{,}5^2 + 8{,}1^2 + 8{,}5^2 + 5{,}9^2 + 7{,}8^2) - 5(7{,}16)^2 = 263{,}76 - 5*51{,}226 = 7{,}432;$$

$$Q_y = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2 = (4^2 + 5{,}6^2 + 5{,}7^2 + 3{,}6^2 + 4^2) - 5*(4{,}58)^2 =$$

$$= 108{,}81 - 5*20{,}976 = 3{,}928;$$

$$Q_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n*\bar{x}*\bar{y} = (5{,}5*4 + 8{,}1*5{,}6 + 8{,}5*5{,}7 +$$

$$+ 5{,}9*3{,}6 + 7{,}8*4) - 5*7{,}16*4{,}58 = 168{,}25 - 5*7{,}16*4{,}58 = 4{,}289.$$

Оценки параметров линейной регрессии $y = \beta_0 + \beta_1 x$ по формулам (1) и (2) равны:

$$\tilde{\beta}_1 = \frac{Q_{xy}}{Q_x} = \frac{4{,}289}{7{,}432} \approx 0{,}577;$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 * \bar{x} = 4{,}58 - 0{,}577*7{,}16 \approx 0{,}451.$$

Таким образом, уравнение линейной регрессии *Y* на *x* имеет вид

$$y = 0{,}451 + 0{,}577*x.$$

Аналогично, оценки параметров линейной регрессии *X* на *y*.

$$\tilde{\beta}_1' = \frac{Q_{xy}}{Q_y} \approx 1{,}091; \tilde{\beta}_0' = \bar{x} - \tilde{\beta}_1'$$

$$\bar{y} = 7{,}16 - 1{,}091*4{,}58 \approx 2{,}163$$

Уравнение линейной регрессии *X* на *y* имеет вид

$$x = 2{,}163 + 1{,}091y$$

2. Диаграмма рассеяния исходных данных и прямая регрессии *Y* на *x* показана на рис. 1.

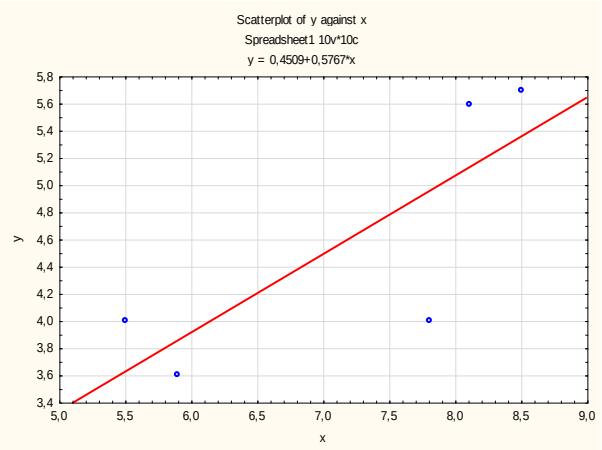


Рис. 1. Диаграмма рассеяния и прямая регрессии Y на x

3. Для линейной регрессии Y на x вычислим остатки:

$$e_i = y_i - \left(\tilde{\beta}_0 + \tilde{\beta}_1 x_i \right), i = 1,2,...,5;$$

$$e_1 = 4 - (0{,}451 + 0{,}577 * 5{,}5) = 0{,}377;$$

$$e_2 = 5{,}6 - (0{,}451 + 0{,}577 * 8{,}1) = 0{,}478;$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$e_5 = 4 - (0{,}451 + 0{,}577 * 7{,}8) = -0{,}949.$$

Остаточная сумма квадратов Q_e:

$$Q_e = (0{,}377)^2 + (0{,}478)^2 + (0{,}35)^2 + (-0{,}25)^2 + (-0{,}949)^2 \approx 1{,}457.$$

Оценка дисперсии ошибок наблюдений

$$S^2 = \frac{Q_e}{n-k} = \frac{1{,}457}{5-2} \approx 0{,}486,$$

где k — число оцениваемых параметров; для простой линейной регрессии $k = 2$.

Коэффициент детерминации R^2:

$$R^2 = 1 - \frac{Q_e}{Q_y} = 1 - \frac{1{,}457}{3{,}928} \approx 0{,}629.$$

Оценка коэффициента корреляции *r*:

$$r = \frac{Q_{xy}}{\sqrt{Q_x Q_y}} = \frac{4{,}286}{\sqrt{7{,}438 * 3{,}928}} \approx 0{,}793$$

4. Вычислим оценки параметров линейной регрессии *Y* на *x* в матричном виде, используя формулу (3):

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

где $\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix}$; *A* — регрессионная матрица:

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} 1 & 5{,}5 \\ 1 & 8{,}1 \\ 1 & 8{,}5 \\ 1 & 5{,}9 \\ 1 & 7{,}8 \end{pmatrix};$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 4 \\ 5{,}6 \\ 5{,}7 \\ 3{,}6 \\ 4 \end{pmatrix}.$$

Последовательно вычисляем:

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5{,}5 & 8{,}1 & 8{,}5 & 5{,}9 & 7{,}8 \end{pmatrix},$$

$$B = A^T A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5{,}5 & 8{,}1 & 8{,}5 & 5{,}9 & 7{,}8 \end{pmatrix}$$

Определитель матрицы *B:*

$$|B| = det(A^T A) = 37{,}6.$$

Обратная матрица к матрице *B:*

$$B^{-1} = \frac{1}{|B|} * B^{*} = \frac{1}{37{,}16} * \begin{pmatrix} 262{,}76 & -35{,}8 \\ -35{,}8 & 5 \end{pmatrix} = \begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix},$$

где B^{*} — присоединенная матрица к матрице *B,* составленная из алгебраических дополнений к элементам матрицы *B*.

Далее вычисляем произведения матриц

$$B^{-1} * A^{-1} = \begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5{,}5 & 8{,}1 & 8{,}5 & 5{,}9 & 7{,}8 \end{pmatrix} =$$

$$= \begin{pmatrix} 1{,}7992 & -0{,}7056 & -1{,}0910 & 1{,}4139 & -0{,}4166 \\ -0{,}2234 & 0{,}1265 & 0{,}1803 & -0{,}1695 & 0{,}0861 \end{pmatrix}.$$

Окончательно

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix} = B^{-1} A^{T} Y =$$

$$= \begin{pmatrix} 1{,}7992 & -0{,}7056 & -1{,}0910 & 1{,}4139 & -0{,}4166 \\ -0{,}2234 & 0{,}1265 & 0{,}1803 & -0{,}1695 & 0{,}0861 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 5{,}6 \\ 5{,}7 \\ 3{,}6 \\ 4 \end{pmatrix} = \begin{pmatrix} 0{,}4509 \\ 0{,}5767 \end{pmatrix}.$$

Сравнивая полученные значения с результатами в п. 2 видно, что расхождение имеется только в третьем десятичном знаке [4].

5. Выполнение задания в пакете STATISTICA.

Откройте новый файл данных. В таблице удалите ненужные столбцы **(Var-Delete)** и строки наблюдений **(Cases-Delete).** Дайте имена переменным: *Y*— зависимая переменная **(Dependent),** *X*— фактор (независимая переменная — **Independent)**. В ячейки таблицы введите данные.

Построим график исходных данных. Для этого можно воспользоваться меню **Graphs — графики** и выбрать необходимый тип графика. В нашем примере мы воспользуемся двумерными диаграммами рассеяния **(Stats 2D Graphs -» Scatterplots).** В диалоговом окне при помощи кнопки **Variables — Переменные** выберите необходимые переменные, которые вы хотите отобразить графически и необходимый тип графика.

В Переключателе модулей (STATISTICA Module Switcher) выберите модуль Множественная регрессия (Multiple Regression). После запуска модуля на экране откроется стартовая панель модуля (рис. 2). Далее выберите переменные для анализа (воспользуйтесь кнопкой **Variables).** В качестве зависимой переменной *(Dependent)* выберите У, в качестве независимой ***(Independent)*** — *X*. После определения зависимых и независимых переменных на стартовой панели нажмите ОК. Появится окно с результатами вычислений (рис. 3).

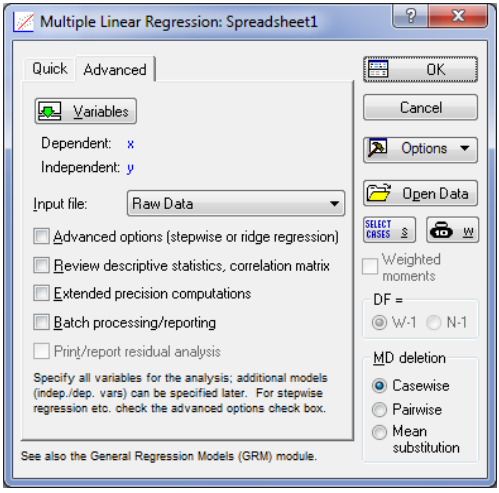


Рис. 2. Стартовая панель модуля Multiply Regression

В диалоговом окне **Результаты Множественной регрессии — Multiple Regression Results** просмотрите результаты оценивания. Результаты можно просмотреть в численном и графическом виде. Окно результатов анализа имеет следующую структуру: верх окна — информационный. Он состоит из двух

частей: в первой части содержится основная информация о результатах оценивания, во второй высвечиваются значимый *стандартизованный* регрессионный коэффициент X-beta=,793; стандартизованный коэффициент регрессии вычисляется по формуле

$$\tilde{\beta}_{01} = \tilde{\beta}_1 * (s_x / s_y),$$

где s_x *и* s_y — оценки среднеквадратических отклонений для переменных *x и y*.

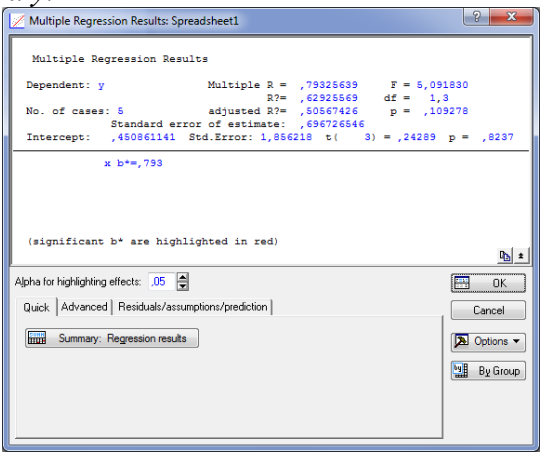


Рис. 3. Окно результатов множественной регрессии

Внизу окна **Результаты множественной регрессии** находятся функциональные кнопки, позволяющие всесторонне рассмотреть результаты анализа.

Рассмотрим вначале информационную часть окна. В ней содержатся краткие сведения о результатах анализа. А именно:

— ***Dep.Var.*** — имя зависимой переменной (*Y*);

—***No. of Cases*** — число наблюдений (объем выборки *n*), по которым построена регрессия (*n* = 5);

—***Multiple R*** — коэффициент множественной корреляции (описывает степень линейной зависимости между *Y* и факторами); в случае простой линейной регрессии равен модулю коэффициента корреляции;

—***R — square — RI*** — квадрат коэффициента множе-

ственной корреляции (коэффициент детерминации). Если регрессионная модель значима, то коэффициент детерминации равен той доле дисперсии ошибок наблюдений, которая объясняется регрессионной моделью.

Коэффициент детерминации, вычисляется по формуле

$$R^2 = 1 - \frac{Q_e}{Q_y};$$

— ***Adjusted R-square: adjusted RI—*** скорректированный коэффициент детерминации

$$R_1^2 = 1 - \frac{Q_e/(n-k)}{Q_y/(n-1)},$$

где n — число наблюдений, а k — число оцениваемых параметров регрессионной модели; для простой линейной регрессии $k = 2$, так как определяются оценки двух параметров β_0 и β_1;

— ***Std. Error of estimate*** — среднее квадратическое отклонение ошибок наблюдений

$$S = \sqrt{S^2} = \sqrt{\frac{Q_e}{(n-k)}};$$

—***Intercept*** — оценка свободного члена регрессии $\left(\tilde{\beta}_0\right)$;

—***Std. Error*** — стандартная ошибка оценки свободного члена $\sqrt{D\left[\tilde{\beta}_0\right]}$;

—***t(n-k) and p-value*** — выборочное значение /-статистики и вычисленного уровня значимости p.

t-статистика используется для проверки гипотезы $H_0: \beta_0 = 0$:

$$t = \frac{\tilde{\beta}_0}{\sqrt{D\left[\tilde{\beta}_0\right]}}.$$

Уровень значимости $p = P\left[T(n-k) > |t_в|\right]$, где *T(n - к)* — случайная величина, имеющая распределение Стьюдента с (я - *к)* степенями свободы, $t_в$ — выборочное значение *t*-статистики.

Если $p > \alpha$, где α — заданный уровень значимости, то гипотеза $H_0: \beta_0 = 0$ принимается.

В данном случае *p* = 0,823749, следовательно гипотеза $H_0: \beta_0 = 0$ принимается.

— *F*—выборочное значение F-статистики, F_B. F-статистика используется для проверки гипотезы $H_0: \beta_0 = 0$.

Если гипотеза $H_0: \beta_0 = 0$ верна, то статистика *F* имеет распределение Фишера с (*к*- 1) и (*n* - *к*) степенями свободы.

Гипотеза H_0 принимается на уровне значимости α, если выборочное значение статистики *F*, F_B, меньше $F_{1-\alpha}(k-1, n-k)$- квантили распределения Фишера порядка 1 - α. Если гипотеза $H_0: \beta_0 = 0$ принимается, то *регрессионная модель незначима.*

—***df***— число степеней свободы F -статистики: (*к*- 1; n-*к*).

—***p*** — вычисленный уровень значимости.

Вычисленный уровень значимости *p: p = P[F(k* - 1; *n* - *k) > F_B],* где F_B — выборочное значение *F* -статистики.

Если *p* < *a,* то гипотеза $H_0: \beta_0 = 0$ отклоняется; если *p*> α, то гипотеза $H_0: \beta_0 = 0$ принимается.

В данном примере *p* = 0,109278, следовательно гипотеза $H_0: \beta_0 = 0$ принимается на уровне значимости α = 0,05. *Регрессионная модель незначима.*

Функциональные кнопки. При нажатии кнопки **Regression Summary — Результаты регрессии** на экране появится следующая таблица с результатами анализа (рис. 4.):

Во втором столбце таблицы **(BETA)** выводится стандартизованный коэффициент регрессии β_{01} :

$$\beta_{01} = \tilde{\beta}_1 * \left(s_x / s_y\right),$$

где s_x *и* s_y — оценки среднеквадратических отклонений для переменных *x и y*.

Regression Summary for Dependent Variable: y (Spreadsheet1)
R= ,79325639 R?= ,62925569 Adjusted R?= ,50567426
F(1,3)=5,0918 p<,10928 Std.Error of estimate: ,69673

N=5	b*	Std.Err. of b*	b	Std.Err. of b	t(3)	p-value
Intercept			0,450861	1,856218	0,242892	0,823749
x	0,793256	0,351542	0,576695	0,255570	2,256508	0,109278

Рис. 4. Результаты регрессии

Стандартизированные коэффициенты регрессии — безразмерные величины[2].

В случае множественной регрессии стандартизованные коэффициенты регрессии используются для сравнения влияния на зависимую переменную факторов, имеющих различную размерность.

В четвертом столбце *(B)* приведены МНК-оценки коэффициентов регрессии: $\tilde{\beta}_0$ и $\tilde{\beta}_1$.

В пятом столбце ***(St.Err. of B)*** — их стандартные отклонения

В шестом столбце — *t*-статистики для проверки гипотезы $H_0: \beta_0 = 0$:

$$t_i = \frac{\tilde{\beta}_i}{\sqrt{D\left[\tilde{\beta}_i\right]}}; i = 0,1.$$

В седьмом столбце — соответствующие уровни значимости

$$p = P\left[T\left(n-k\right) > \left|t_i\right|\right].$$

В данном случае гипотеза $H_0: \beta_1 = 0$ принимается на уровне значимости $\alpha = 0,05$. Вычисленный уровень значимости $p > \alpha$. Это означает, что регрессионная модель незначима. Гипотеза $H_0: \beta_0 = 0$ также принимается при $\alpha = 0,05$.

Чтобы просмотреть и проанализировать остатки, войдите в меню Residual Analysis (анализ остатков), нажав соответствующую кнопку в нижней правой части панели результатов вычислений (рис. 3). Это меню представлено на рис. 5.

Чтобы просмотреть остатки и их график, нажмите в левой нижней части этого меню кнопку **Plots of residuals(A)** (графики остатков (A)). Выбрав опцию **Raw residuals** (значения остатков), получим график остатков, наблюдаемые значения (**observed value**) зависимой переменной *Y,* предсказанные значения *Y* (**predicted**), остатки (**residuals**) и стандартизированные остатки (**Standard Residual**) вычисляемые по формуле $\frac{e_i}{S}, i = 1,2,..., n,$ где S — оценка среднеквадратического отклонения ошибок наблюдений, $S \approx 0,7$ (рис. 6).

Остаточная сумма квадратов Q_e (**Residual**) сумма квадратов, обусловленная регрессией Q_R (**Regress**) и сумма квадратов отклонений зависимой переменной Y *от* среднего Q_y (**Total**) вычисляются при нажатии кнопки **Analysis of Variance** (дисперсионный анализ) на панели результатов вычислений (рис. 3). Результаты дисперсионного анализа приведены на рис. 7.

В этой же таблице приведены соответствующие значения числа степеней свободы *(df),* средние квадраты, F-статистика для проверки гипотезы о незначимости регрессионной модели и вычисленный уровень значимости *p.*

В данном примере гипотеза о незначимости регрессионной модели по F-критерию также принимается, т. к. $p \approx 0,11$, что больше обычно задаваемого уровня значимости $\alpha = 0,05$.

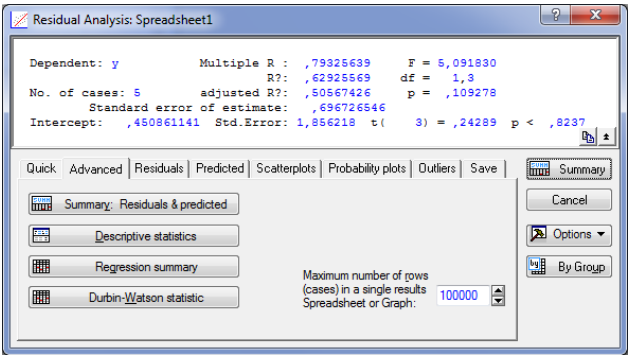
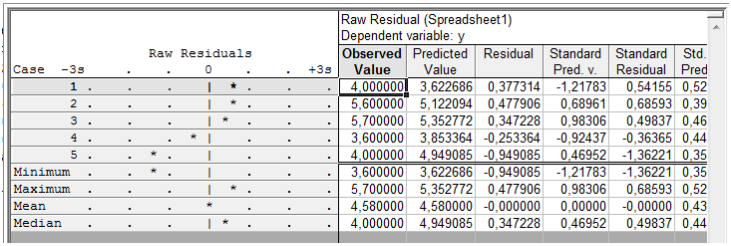


Рис. 5. Меню анализа остатков



Raw Residuals

Case	-3s	0	+3s
1		*	
2		*	
3		*	
4		*	
5		*	
Minimum		*	
Maximum		*	
Mean		*	
Median		*	

Raw Residual (Spreadsheet1)
Dependent variable: y

Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std. Pred
4,000000	3,622686	0,377314	-1,21783	0,54155	0,52
5,600000	5,122094	0,477906	0,68961	0,68593	0,39
5,700000	5,352772	0,347228	0,98306	0,49837	0,46
3,600000	3,853364	-0,253364	-0,92437	-0,36365	0,44
4,000000	4,949085	-0,949085	0,46952	-1,36221	0,35
3,600000	3,622686	-0,949085	-1,21783	-1,36221	0,35
5,700000	5,352772	0,477906	0,98306	0,68593	0,52
4,580000	4,580000	-0,000000	0,00000	-0,00000	0,43
4,000000	4,949085	0,347228	0,46952	0,49837	0,44

Рис.6. График остатков (слева) и их значения (столбец Residual) справа

Analysis of Variance; DV: y (Spreadsheet1)

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	2,471716	1	2,471716	5,091830	0,109278
Residual	1,456284	3	0,485428		
Total	3,928000				

Рис.7. Дисперсионный анализ

Задания для лабораторной работы

1. 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[4]:

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

Определите регрессионную модель по этим данным.

2. 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

3. Автосервисное предприятие имеет следующие данные по стоимости ежегодного технического обслуживания автомобилей определенной марки в зависимости от времени эксплуатации.

Стоимость тех. обслуживания, *y* (тыс. руб.)	5,3	5,2	6,0	5,7	6,6	6,8	8,1	6,9	10,3	4,0	2,5
Время эксплуатации, *x* (лет)	5	4	5	6	7	8	10	8	11	3	2

Определите регрессионную модель для этих данных.

Лабораторная работа № 2
Проверка значимости и адекватности простой линейной регрессии. Прогнозирование. Множественная линейная регрессия

1. Основные понятия

Ковариационная матрица оценок параметров регрессионной модели. Доверительные интервалы для параметров регрессионной модели. Проверка гипотез о равенстве параметров нулю. Разложение суммы квадратов отклонений результатов наблюдений от среднего (Q_y) на сумму квадратов обусловленную регрессией (Q_R) и остаточную на сумму квадратов (Q_e). Смысл тождества: $Q_y = Q_R + Q_e$ и проверка гипотезы о незначимости регрессионной модели. Проверка адекватности по графику остатков. Критерий Дарбина—Уотсона. Проверка гипотезы о нормальном распределении остатков. Проверка

адекватности по повторным наблюдениям. Доверительные интервалы для среднего предсказанного значения и для индивидуального предсказанного значения[4].

2. Задание

По выборке из своего варианта, используя результаты расчетов полученные в работе 1, выполнить следующие расчеты и задания:

1. Вычислить ковариационную матрицу оценок параметров регрессионной модели.

2. Вычислить доверительные интервалы для параметров регрессии и для дисперсии ошибок наблюдений при доверительной вероятности 0,95.

3. Вычислить сумму квадратов, обусловленную регрессией по одной из формул

$$Q_R = \tilde{\beta}_1 Q_{xy} = \tilde{\beta}_1^{\,2} Q_{xy} = \tilde{\beta}_1^{\,2} Q_x = \frac{Q_{xy}^2}{Q_x}$$

4. Проверить тождество: $Q_y = Q_R + Q_e$.

5. Проверить гипотезу о незначимости модели $H_0 : \beta_1 = 0\, H_0$ по F-критерию Фишера и используя доверительный интервал для β_1.

6. Построить график остатков.

7. Вычислить статистику Дарбина—Уотсона.

8. Вычислить доверительные интервалы для среднего предсказанного значения и индивидуального предсказанного значения $\tilde{Y}(x_0)$. В качестве x_0 взять два значения

$$x_{01} = \frac{x_{\min} + x_{\max}}{2} \text{ и } x_{02} = x_{\max} + 2,$$

где x_{min} и x_{max} минимальное и максимальное значение x в заданной выборке.

Границы доверительных интервалов для предсказанных значений нанести на график, содержащий прямую регрессии Y на x и диаграмму рассеяния. Доверительную вероятность взять равной 0,90.

9. Ввести данные в пакет STATISTICA, выполнить п. 1—8. Сравнить результаты расчетов и записать их в отчет.

Пример 1 (продолжение). Продолжим решение примера 1 (**прошлая работа**) по пунктам задания в работе 2.

1. Ковариационная матрица оценок параметров регрессионной модели K вычисляется по формуле

$$K = S^2\left(A^T A\right)^{-1} = S^2 B^{-1} =$$

$$= 0{,}486\begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix} \approx \begin{pmatrix} 3{,}449 & -0{,}468 \\ -0{,}468 & 0{,}065 \end{pmatrix}.$$

Таким образом имеем:

$$D\left[\tilde{\beta}_0\right] = 3{,}449, \quad D\left[\tilde{\beta}_1\right] = 0{,}065, \quad \mathbf{cov}\left[\tilde{\beta}_0, \tilde{\beta}_1\right] = -0{,}468.$$

В пакете STATISTICA выводятся значения стандартных отклонений ***(St. Error of B):***

$$\sqrt{D\left(\tilde{\beta}_0\right)} = \sqrt{3{,}449} \approx 1{,}856 \text{ и } \sqrt{D\left(\tilde{\beta}_1\right)} = \sqrt{0{,}065} \approx 0{,}255$$

(см. рис. Результаты регрессии).

2. Доверительные интервалы для параметров линейной регрессии вычисляются по следующим формулам:

$$\text{для } \beta_0 : \tilde{\beta}_0 \pm t_{1-\frac{\alpha}{2}}(n-k)\sqrt{D\left[\tilde{\beta}_0\right]};$$

$$\text{для } \beta_1 : \tilde{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-k)\sqrt{D\left[\tilde{\beta}_1\right]},$$

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

При доверительной вероятности $1-\alpha = 0{,}95$, $t_{0{,}975}$(5-2) = $t_{0{,}975}(3) = 3{,}182$ (используйте статистический калькулятор!)

Окончательно имеем следующие значения доверительных интервалов:

$$\text{для } \beta_0 : 0{,}451 \pm 3{,}185 * \sqrt{3{,}449} = 0{,}451 \pm 5{,}909,$$

для $\beta_1 : 0{,}577 \pm 3{,}182 * \sqrt{0{,}065} = 0{,}577 \pm 0{,}811.$

Таким образом, оба коэффициента регрессии β_0 и β_1, *незначимы* на уровне значимости α =0,05, т. к. 95%-е доверительные интервалы для β_0 и β_1, включают нуль.

В пакете STATISTICA (см. рис. 8) вычисляются значения *t*-статистик для проверки гипотезы $H_0 : \beta_0 = 0$

Regression Summary for Dependent Variable: y (Spreadsheet1)
R= ,79325639 R?= ,62925569 Adjusted R?= ,50567426
F(1,3)=5,0918 p<,10928 Std.Error of estimate: ,69673

N=5	b*	Std.Err. of b*	b	Std.Err. of b	t(3)	p-value
Intercept			0,450861	1,856218	0,242892	0,823749
x	0,793256	0,351542	0,576695	0,255570	2,256508	0,109278

Рис.8. Результаты регрессии

$$t = \frac{\tilde{\beta}_0}{\sqrt{D\left[\tilde{\beta}_0\right]}} = \frac{0{,}451}{1{,}856} \approx 0{,}243$$

и для проверки гипотезы $H_0 : \beta_1 = 0$

$$t = \frac{\tilde{\beta}_1}{\sqrt{D\left[\tilde{\beta}_1\right]}} = \frac{0{,}577}{0{,}255} \approx 2{,}256.$$

Обе гипотезы принимаются на уровне значимости соответственно:

$$p = 0{,}824 \text{ и } p = 0{,}109.$$

Доверительный интервал для дисперсии ошибок наблюдений определяется по формуле

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

где $\chi^2_{1-\frac{\alpha}{2}}(n-k)$ и $\chi^2_{1-\frac{\alpha}{2}}(n-k)$ — квантили распределения χ^2 с $(n-k)$ степенями свободы. При доверительной вероятности

$1-\alpha = 0{,}95$ имеем (используйте статистический калькулятор!) при *n = 5 и k=2:*

$$\chi^2_{0{,}975}(3) = 7{,}81, \quad \chi^2_{0{,}025}(3) = 0{,}216$$

Таким образом доверительный интервал для дисперсии ошибок наблюдений имеет вид

$$\frac{(5-2)*0{,}486}{7{,}81} < \sigma^2 < \frac{(5-2)*0{,}486}{2{,}216}$$ или окончательно

$$0{,}187 < \sigma^2 < 6{,}75.$$

3. Сумма квадратов, обусловленная регрессией Q_R:

$$Q_R = \tilde{\beta}_1 Q_{xy} = 0{,}577 * 4{,}286 \approx 2{,}472$$

(сравните результаты расчета с результатами дисперсионного анализа, рис. 9).

	Analysis of Variance; DV: y (Spreadsheet1)				
Effect	**Sums of Squares**	df	Mean Squares	F	p-value
Regress.	2,471716	1	2,471716	5,091830	0,109278
Residual	1,456284	3	0,485428		
Total	3,928000				

Рис.9. Результаты дисперсионного анализа

4. Проверяем тождество $Q_y = Q_R + Q_e$: 3,928 ≈ 2,472 + 1,457 =3,929.

5. Проверим гипотезу $H_0 : \beta_1 = 0$ о незначимости регрессионной модели по критерию Фишера.

Выборочное значение статистики Фишера *F* равно

$$F_{\text{в}} = \frac{Q_R/(k-1)}{Q_e/(n-k)} = \frac{2{,}472/1}{1{,}457/3} = 5{,}091.$$

Так как F_B меньше квантили распределения Фишера $F_{1-\alpha}(k-1, n-k) = F_{0{,}95}(1{,}3) = 10{,}13$, то гипотеза $H_0 : \beta_1 = 0$ не отклоняется: регрессионная модель *незначима* (сравните этот результат со значениями *F*-статистики и *p*-уровня на рис. 2).

Тот же результат получим используя 95%-й доверительный интервал для β_1: (-0,235; 1,387).

Так как 95%-й доверительный интервал для p, накрывает 0, гипотеза $H_0 : \beta_1 = 0$ принимается на уровне значимости α =0,05.

6. График остатков. В данном примере число остатков очень мало *(n* = 5) поэтому сделать какие-либо выводы о выполнении предположений регрессионного анализа по остаткам нельзя. Более того, так как регрессионная модель незначима, то проверка этих предложений лишена смысла.

7. Вычислим статистику Дарбина—Уотсона

$$d = \sum_{i=2}^{n} \frac{(e_i - e_{i-1})}{Q_e} =$$

$$\frac{(0{,}48 - 0{,}378)^2 + (0{,}35 - 0{,}48)^2 + (-0{,}25 - 0{,}35)^2 + (-0{,}947 + 0{,}25)^2}{1{,}457} \approx$$

$$\approx \frac{0{,}873}{1{,}457} \approx 0{,}5989$$

Для n - 5 критических значений статистики Дарбина—Уотсона в таблице нет. Поэтому проверить гипотезу о некоррелированности остатков при столь малом числе наблюдений нельзя.

8. Вычислим доверительные интервалы для предсказанных значений. Здесь надо иметь в виду, что если регрессионная модель незначима и неадекватна результатам наблюдений, как это имеет место в данном примере, то эту модель использовать для прогноза нельзя. Мы приведем соответствующие расчеты, чтобы продемонстрировать только технику вычислений.

Найдем предсказанное значение Y в точках:

$$x_{01} = \frac{x_{\min} + x_{\max}}{2} = \frac{5{,}5 + 8{,}5}{2} = 7,$$

$$x_{02} = x_{\max} + 2 = 8{,}5 + 2 = 10{,}5,$$

$$\tilde{y}(x_{01}) = \tilde{\beta}_0 + \tilde{\beta}_1 x_{01} = 0{,}454 + 0{,}576 * 7 = 4{,}486,$$

$$\tilde{y}(x_{02}) = 0{,}454 + 0{,}576 * 10{,}5 = 6{,}502.$$

Границы доверительного интервала для *среднего предсказанного значения* (**confidence limit**) вычисляются по формуле

$$\tilde{y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-2) S \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{Q_x}}$$

или по более общей формуле:

$$\tilde{y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-2) S \sqrt{a^T(x_0) B^{-1} a(x_0)},$$

где $a^{\mathrm{T}}(x)$ — вектор-строка регрессионной матрицы *A;* в случае простой линейной регрессии: $a^T(x) = (1, x)$.

В данном примере, при доверительной вероятности $1 - \alpha = 0{,}90$ имеем при $x_{01} = 7, \left(t_{0,95}(3) = 2{,}353\right)$:

$$4{,}486 \pm 2{,}353\sqrt{0{,}486}\sqrt{\frac{1}{5} + \frac{(7-7{,}16)^2}{7{,}458}} = 4{,}486 \pm 0{,}740..$$

По более общей формуле

$$a^T(x_0) B^{-1} a(x_0) = (1;7)\begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix}\begin{pmatrix} 1 \\ 7 \end{pmatrix} = 0{,}231.$$

Таким образом, доверительный интервал для среднего предсказанного значения равен

$$4{,}486 \pm 2{,}353 * \sqrt{0{,}486}\sqrt{0{,}231} = 4{,}486 \pm 0{,}789.$$

Чтобы вычислить *доверительный интервал для индивидуального предсказанного значения* (**prediction limit**) оценка дисперсии $D\left[\tilde{y}(x_0)\right]$ должна включать еще один источник вариации — разброс относительно линии регрессии, определяемый дисперсией S^2. Таким образом, доверительный интервал для индивидуального значения вычисляется по формуле

$$\tilde{y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-2) S \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{Q_x}}$$

или, в общем случае:

$$\tilde{y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-2) S \sqrt{a^T(x_0) B^{-1} a(x_0)}.$$

В рассматриваемом примере для индивидуального предсказанного значения Y при $x_{01} = 7$, получим следующие значения границ доверительного интервала

$$4{,}486 \pm 2{,}353^* \sqrt{0{,}486} \sqrt{1 + \frac{1}{5} + \frac{(7 - 7{,}16)^2}{7{,}458}} = 4{,}486 \pm 1{,}780$$

или по общей формуле

$$4{,}486 \pm 2{,}353^* \sqrt{0{,}486} \sqrt{1 + 0{,}231} = 4{,}486 \pm 1{,}820.$$

Аналогично вычисляются значения границ доверительных интервалов для среднего и индивидуального предсказанного значения Y при $x = 10{,}5$. Соответственно, имеем:

$$6{,}502 \pm 2{,}353^* \sqrt{0{,}486} \sqrt{\frac{1}{5} + \frac{(10{,}5 - 7{,}16)^2}{7{,}458}} = 6{,}502 \pm 2{,}136;$$

$$6{,}502 \pm 2{,}353^* \sqrt{0{,}486} \sqrt{1 + \frac{1}{5} + \frac{(10{,}5 - 7{,}16)^2}{7{,}458}} = 6{,}502 \pm 2{,}693.$$

Выполнение задания в пакете STATISTICA

Основные моменты статистического анализа результатов расчетов для простой линейной регрессии в пакете STATISTICA мы уже прокомментировали[4].

Рассмотрим вычисление предсказанных значений и доверительных интервалов для них.

Вычисления выполняются при нажатии кнопки **Predict dependent variable** (предсказанное значение зависимой переменной) в окне **Multiple Regression Results** (рис. 10).

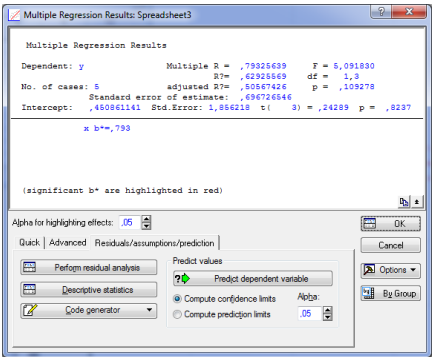


Рис. 10. Окно результатов множественной регрессии

Предварительно надо задать уровень значимости α и вид вычисляемого доверительного интервала: **Confidence limits** — доверительный интервал для среднего предсказанного значения; или **Prediction limits** — доверительный интервал для индивидуального предсказанного значения.

Нажав кнопку и задав значение независимой переменной, например, $x_{01} = 7,0$, в таблице результатов (рис. 11) получим предсказанное значение: $\tilde{y}(x_0) \approx 4{,}488$ и 90%-е доверительные интервалы для среднего предсказанного значения: (3,748; 5,227).

Predicting Values for (Spreadsheet3)
variable: y

Variable	b-Weight	Value	b-Weight * Value
x	0,576695	7,000000	4,036868
Intercept			0,450861
Predicted			4,487729
-90,0%CL			3,748167
+90,0%CL			5,227290

Рис. 11. Вычисление предсказанного значения

Множественная регрессия. Пример 2.

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

Y	X_1	X_2	X_3
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 (зависимая переменная) — общий доход от проданных билетов, млн руб.; x_1, — средства на развитие компаний в регионе, млн руб.; x_2 — число конкурирующих компаний; x_3 — процент пассажиров, летавших бесплатно.

Найти уравнение множественной регрессии. Проверить значимость и адекватность регрессионной модели. Существенно ли влияет на доход число пассажиров, летавших бесплатно? Какой доход (в среднем) может ожидать компания, вложившая в развитие 2,5 млн руб., если число конкурирующих компаний в регионе равно десяти, а число пассажиров, летавших бесплатно по разным причинам, составляет 3 %. Принять уровень значимости а = 0,05.

Решение в пакете STATISTICA. Проведите те же операции в модуле **Multiple Regression,** что и в работе 1: введите данные: ***Variables: dependent var- Y, independent var-Xl, X2, X3,*** **OK -> Regression Summary.** Результаты регрессионного анализа приведены на рис. 12.

Уравнение множественной регрессии имеет вид: $Y = 170{,}76 + 25{,}42x_1 - 13x_2 - 2{,}7x_3$.

Из данной таблицы видно, что гипотеза $H_0 : \beta_3 = 0$ принимается на уровне значимости p=0,267, так как $p > \alpha$ = 0,05. Остальные коэффициенты регрессионной модели значимы.

Regression Summary for Dependent Variable: Var1 (Spreadsheet3)
R= ,95117377 R?= ,90473155 Adjusted R?= ,87874924
F(3,11)=34,821 p<,00001 Std.Error of estimate: 27,798

N=15	b*	Std.Err. of b*	b	Std.Err. of b	t(11)	p-value
Intercept			**170,7600**	**52,0862**	**3,2784**	**0,0074**
Var2	**0,7313**	**0,1399**	**25,4233**	**4,8642**	**5,2266**	**0,0003**
Var3	**-0,4151**	**0,1190**	**-13,0035**	**3,7278**	**-3,4883**	**0,0051**
Var4	-0,145960	0,124879	-2,7059	2,31510	-1,16881	0,267186

Рис.12. Результаты регрессионного анализа

Проверим гипотезу о незначимости регрессионной модели. Для этого используем опцию **Analysis of Variance** (дисперсионный анализ).

Результаты дисперсионного анализа приведены в таблице (рис. 13). Из таблицы видно, что статистика критерия Фишера, вычисляемая по формуле

$$F = \frac{Q_R / (k-1)}{Q_e / (n-k)}$$

равна F(3,ll) =34,821, так как p = 0,000007, что меньше, чем а = 0,05, то гипотеза о незначимости модели отклоняется.

Так как коэффициент p_3 незначим, пересчитаем уравнение множественной регрессии используя два фактора x_1 и x_2. Результаты регрессионного анализа (**Regression Summary for Dependent Variable**)приводятся на рис. 14.

Уравнение множественной регрессии имеет вид: $Y = 159{,}86 + 22{,}39x_1 - 12{,}53x_2$

Коэффициенты регрессионной модели $\beta_0, \beta_1, \beta_2$ значимы (соответствующие уровни значимости равны соответственно: 0,009; 0,00017; 0,0059).

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	**80720,3874**	**3,0000**	**26906,7958**	**34,8211**	**0,0000**
Residual	8499,88	11	772,72		
Total	89220,26				

Рис. 13. Таблица дисперсионного анализа

Regression Summary for Dependent Variable: Var1 (Spreadsheet3)
R= ,94493383 R?= ,89289994 Adjusted R?= ,87504993
F(2,12)=50,022 p<,00000 Std.Error of estimate: 28,219

N=15	b*	Std.Err. of b*	b	Std.Err. of b	t(12)	p-value
Intercept			159,8629	52,0209	3,0731	0,0097
Var2	0,6440	0,1201	22,3882	4,1754	5,3620	0,0002
Var3	-0,4001	0,1201	-12,5316	3,7619	-3,3311	0,0060

Рис. 14. Результаты регрессионного анализа

Регрессионная модель значима: F= 50,022, уровень значимости p = 0,000002.

Чтобы проверить выполнение предположений регрессионного анализа и адекватность модели рассмотрим остатки. Для этого используем опцию Residual Analysis (анализ остатков).

Начнем с проверки гипотезы о том, что все сериальные корреляции в последовательности остатков равны нулю (гипотеза H_0). Для проверки этой гипотезы используется критерий Дарбина—Уотсона (рис.15).

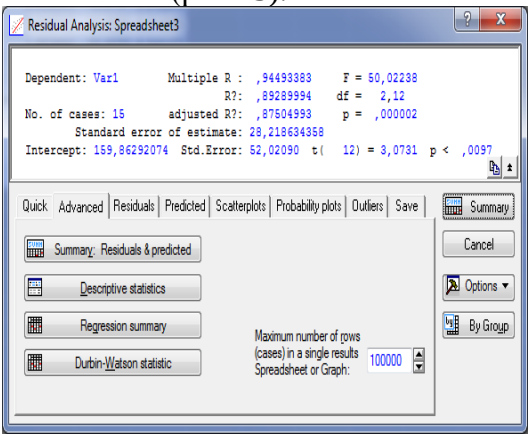


Рис.15. Окно для вычисления статистика Дарбина -Уотсона

Чтобы проверить гипотезу H_0, в окне **Multiple Regression Results** выберите опцию **Residual Analysis** (рис. 15), а затем — **Durbin-Watson stat.** Результат приводится на рис. 16.

Durbin-Watson d (Spreadsheet and serial correlation of residu

	Durbin-Watson d	Serial Corr.
Estimate	1,896936	-0,058644

Рис.16. Статистика Дарбина-Уотсона

В данном случае статистика Дарбина—Уотсона d = 1,8969, что больше табличного значения d_2 = 1,75 (см. Приложение 2), следовательно, гипотеза H_0: все сериальные корреляции равны нулю принимается на уровне значимости 2α = 0,1.

Построим график остатков. Для этого в окне **Residual Analysis** нужно выбрать опцию **Casewise plot of residual.** Результаты приводятся на рис. 17.

Все остатки укладываются в симметричную относительно нулевой линии полосу шириной ±2S. Это означает, что, по-видимому, дисперсии ошибок наблюдений постоянны.

Raw Residual (Spreadsheet3)
Dependent variable: Var1

Case	Raw Residuals −3s … 0 … +3s	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1		79,3000	90,5178	-11,2178	-1,04146	-0,39753	12,91724	2,000232	-14,1915	0,017666
2		200,1000	182,7455	17,3545	0,18116	0,61500	8,30152	0,278301	18,9988	0,013077
3		163,2000	143,8134	19,3866	-0,33495	0,68702	14,44502	2,735197	26,2705	0,075702
4		200,1000	249,0087	-48,9087	1,05958	-1,73321	10,87774	1,147002	-57,4447	0,205264
5		146,0000	176,0291	-30,0291	0,09212	-1,06416	8,74939	0,412563	-33,2230	0,044419
6		177,7000	179,6345	-1,9345	0,13992	-0,06855	19,31221	5,623892	-3,6388	0,002596
7		30,9000	54,2606	-23,3606	-1,52211	-0,82784	13,72636	2,379251	-30,6013	0,092752
8		291,9000	298,6988	-6,7988	1,71830	-0,24093	15,00633	3,025845	-9,4797	0,010638
9		160,0000	149,1632	10,8368	-0,26403	0,38403	11,76768	1,501325	13,1181	0,012527
10		339,4000	312,1317	27,2682	1,89637	0,96632	16,05110	3,596329	40,3107	0,220082
11		159,6000	145,1508	14,4492	-0,31722	0,51204	10,31924	0,938863	16,6797	0,015574
12		88,3000	76,6488	11,6512	-1,22532	0,41289	11,88165	1,548712	14,1620	0,014885
13		237,5000	219,0027	18,4973	0,66180	0,65550	12,65852	1,883894	23,1572	0,045172
14		107,2000	146,4883	-39,2883	-0,29949	-1,39228	7,79312	0,134440	-42,5322	0,057756
15		155,0000	112,9060	42,0940	-0,74467	1,49171	9,91241	0,794153	48,0192	0,119103
Minimum		30,9000	54,2606	-48,9087	-1,52211	-1,73321	7,79312	0,134440	-57,4447	0,002596
Maximum		339,4000	312,1317	42,0940	1,89637	1,49171	19,31221	5,623892	48,0192	0,220082
Mean		169,0800	169,0800	-0,0000	0,00000	-0,00000	12,24797	1,866667	0,6403	0,063148
Median		160,0000	149,1632	10,8368	-0,26403	0,38403	11,88165	1,548712	13,1181	0,044419

Рис.17. График остатков

Теперь проверим гипотезу о нормальности распределения остатков. Для этого в том же окне (**Residual Analysis**) необходимо выбрать опцию **Normal Probability Plot of Residuals.** Результаты выполнения процедуры представлены на специальном графике (рис. 18).

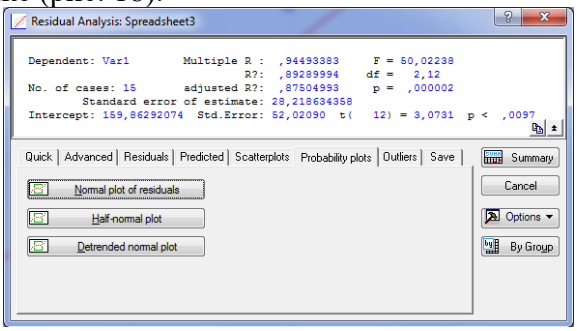


Рис. 18. Выбор опции опцию Normal Probability Plot of Residuals

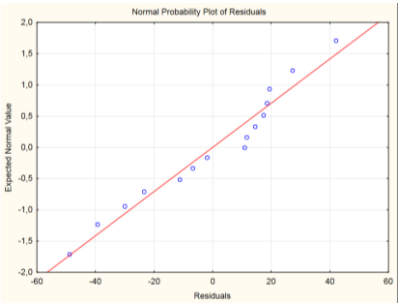


Рис.19. Остатки на графике Normal Probability Plot

Из графика (рис. 19) видно, что точки расположены близко к прямой, значит, можно предположить, что остатки распределены по нормальному закону. Гипотезу о нормальном распределении остатков можно также проверить по критерию χ^2 или критерию Колмогорова—Смирнова.

Таким образом, можно считать, что предположения регрессионного анализа выполняются. Распределение остатков на рис. 17 (случайное, без каких-либо закономерностей) показывает, что регрессионная модель адекватна результатам наблюдений и может быть использована для прогнозирования. Для выполнения прогноза в окне Multiple Regression Results нужно выбрать опцию **Predict Dependent Var**, в появившемся окне нужно ввести значения факторов x_1, x_2 и задать уровень значимости α = 0,05.

В появившемся окне (рис. 20) приведены результаты прогноза: при $x_1 = 2{,}5$, $x_2 = 10$: в первом столбце приведены оценки параметров регрессии $\tilde{\beta}_i = 1, 2$; во втором — значения факторов x_i.

	Predicting Values for (Spreadshee variable: Var1		
Variable	**b-Weight**	Value	b-Weight * Value
Var2	22,3882	2,50000	55,970
Var3	-12,5316	10,00000	-125,316
Intercept			159,863
Predicted			90,518
-95,0%PL			22,899
+95,0%PL			158,136

Рис. 20. Результаты прогноза

Предсказываемое значение *Y* выведено в строке **Predicted,** ниже вычислены 95 % доверительные интервалы для среднего предсказанного значения *Y*=90,518.

Задания для самостоятельного решения

Решите следующие задачи и проведите полный статистический анализ результатов.

Задачи.

1. Используя приведенные ниже данные, найдите уравнение множественной регрессии и ответьте на следующие вопросы:

- каковы оценки коэффициентов регрессии и стандартные ошибки этих оценок?
- каков коэффициент детерминации?
- 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

2. Используя приведенные ниже данные, найдите уравнение множественной регрессии и ответьте на следующие вопросы: каковы оценки коэффициентов регрессии и стандартные ошибки этих оценок? Каков коэффициент детерминации? Каков 95%-й доверительный интервал для предсказанного среднего 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

3. 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 при таком прогнозе могут быть использованы следующие факторы. Данные об этих факторах и количестве налоговых деклараций приведены ниже[4]:

Экономический индекс, $\mathbf{x_1}$	Население в радиусе 1 км от фирмы, x_2, тыс. чел.	Средний доход в районе, x_3, тыс. руб.	Количество деклараций на период с 1 марта по 15 апреля, Y, тыс.
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

4. Определите уравнение множественной регрессии для этих данных.

5. Какой процент дисперсии данных описывается этим уравнением?

Лабораторная работа №3
Кластерный анализ

Пусть $X_1, X_2, ..., X_n$ — исходная совокупность объектов, каждый из которых задан набором p признаков. Например, объектами могут быть пациенты клиники, а признаками- фи-

зические данные (вес, давление и т. д.) и результаты амбулаторного обследования каждого пациента (содержание сахара в крови, уровень гемоглобина и т. д.)

Задача кластерного анализа состоит в разбиении исходной совокупности объектов на группы схожих, близких между собой объектов. Эти группы называют кластерами или таксонами. Другими словами, кластерный анализ это один из способов классификации объектов по их признакам [6].

Одна из концепций состоит в построении разбиения исходного множества объектов доставляющего оптимальное значение определенной целевой функции. Большая группа методов кластеризации использует в качестве целевой функции внутригрупповую сумму квадратов: разбиение каждого множества должно быть таково, чтобы оно минимизировало внутригрупповые суммы квадратов. Эти методы используют евклидову метрику и называются методами минимальной дисперсии.

Пусть $X_1, X_2, ..., X_n$ - объекты, каждый из которых задан набором *p* признаков. Распределения объектов по кластерам на однородные в некотором смысле группы должно удовлетворять критерию оптимальности, который выражается в терминах расстояния $\rho(X_i X_j)$ между любой парой объектов рассматриваемой совокупности.

В качестве расстояния (метрики) может быть взята любая неотрицательная действительная функция $\rho(X_i X_j)$, определенная на множестве $X_1, X_2, ..., X_n$ и удовлетворяющая следующим условиям:

а) $\rho(X_i X_j) = 0$ тогда и только тогда, когда $X_i = X_j$;

б) $\rho(X_i X_j) = \rho(X_j, X_i)$;;

в) $\rho(X_i X_j) \le \rho(X_j, X_i) + \rho(X_k, X_j)$;.

Выбор расстояния между объектами неоднозначен и в этом состоит основная сложность.

Наиболее популярной метрикой является евклидова. Эта метрика отвечает интуитивным представлениям близости.

При этом на расстояние между объектами могут сильно влиять изменения масштабов (единиц измерения) по осям. Например, если один из признаков измерен в метрах, а затем его значение переведены в сантиметры (т.е. умножены на 100), то евклидово расстояние между объектами сильно изменится и это приведет к тому, что результаты кластерного анализа могут значительно отличаться от предыдущих.

Если признаки измерены в разных единицах измерения, то требуется их предварительная нормировка — такое преобразование исходных данных, которое переводит их в безразмерные величины.

Наиболее известные способы нормировки следующие:

$$z_1 = \frac{x - \bar{x}}{\sigma};\ z_2 = \frac{x}{\bar{x}};\ z_3 = \frac{x}{x'},\ z_4 = \frac{x}{x_{\max}},\ z_5 = \frac{x - \bar{x}}{x_{\max} - x_{\min}},$$

где z_i, , i= l, 2, ..., 5 — нормированное значение; x — исходное значение, $\bar{x}$ и σ — соответственно среднее и среднее квадратическое отклонение *x, x′* — эталонное (нормативное) значение, $x_{\max}$ и $x_{\min}$ — наибольшее и наименьшее значение *x.*

В пакете STATISTICA нормировка любой переменной выполняется по формуле $\frac{x - \bar{x}}{\sigma}$. Для этого нужно щелкнуть правой кнопкой мыши на имени переменной и в открывшемся меню выбрать: **Fill/Standardize Block ->Standardize Columns.**

Нормировка, особенно по формуле- $\frac{x - \bar{x}}{\sigma}$, сильно искажает геометрию исходного пространства, что может изменить результаты кластеризации.

Выбор метрики для каждой задачи должен производиться с учетом целей кластеризации, свойств признаков анализируемых объектов, вероятностной структуры данных и т. п. [7].

Наиболее употребительные метрики следующие (в скобках указано английское обозначение некоторых метрик,

используемых в пакете STATISTICA в опции ***Distance measure*).**

1. Евклидова метрика ***(Euclidean distance):***

$$\rho_E\left(X_i, X_j\right)=\left[\sum_{k=1}^{p}\left(X_{ki}-X_{kj}\right)^2\right]^{\frac{1}{2}},$$

где X_{ki} — значение *k-mo* признака *i*-го объекта.

2. «Взвешенная» евклидова метрика:

$$\rho_{\text{вЕ}}\left(X_i, X_j\right)=\left[\sum_{k=1}^{p}W_k\left(X_{ki}-X_{kj}\right)^2\right]^{\frac{1}{2}},$$

где W_k — «вес» *k-го* признака. Применяется в тех случаях, когда каждому признаку можно приписать «вес», пропорциональный степени важности данного признака в задаче классификации.

3. Хеммингово расстояние

$$\rho_H\left(X_i, X_j\right)=\frac{\textit{число случаев}: X_{ki}\neq X_{kj}}{p}.$$

ρ_H используется для признаков измеряемых в номинальной шкале и принимающих два значения. В пакете STATISTICA используется связанная с $p_н$ метрика: процент несогласия ***(Percent disagreement).***

4. Метрика Махаланобиса, определяемая формулой

$$\rho_0\left(X_i, X_j\right)=\left[\left(X_i-X_j\right)^T\Lambda^T\sum\nolimits_X^{-1}\Lambda\left(X_i-X_j\right)\right]^{\frac{1}{2}},$$

где $\sum_X$ — ковариационная матрица генеральной совокупности, из которой извлекаются объекты X_i и X_j, Λ — симметричная неотрицательно-определенная матрица весовых коэффициентов, выбираемая обычно диагональной.

5. Коэффициент корреляции Пирсона (Pearson *r*)*:*

$$\rho_k \frac{\sum_{k=1}^{p}\left(X_{ki}-\overline{X}_i\right)\left(X_{kj}-\overline{X}_j\right)}{\sqrt{\sum_{k=1}^{p}\left(X_{ki}-\overline{X}_i\right)^2\sum_{k=1}^{p}\left(X_{kj}-\overline{X}_j\right)^2}},$$

где $\overline{X}_i=\frac{1}{p}\sum_{k=1}^{p}X_{ki}$, $\overline{X}_j=\frac{1}{p}\sum_{k=1}^{p}X_{kj}$.

Процедуры классификации на основе методов кластерного анализа используют расстояния между множествами объектов. Эти расстояния можно ввести различными способами. Пусть S_i-i-ый класс (группа, кластер), n_i — число элементов в $i-$**м** классе, $\overline{X}(i)$— «центр тяжести» *i*-го класса. Компоненты вектора $\overline{X}(i)$ вычисляются по формуле

$$\overline{X}_j(i)=\frac{1}{n_i}\sum_{k=1}^{n_i}X_{jk},\ j=1,2,\ldots p.$$

Наиболее употребительные меры расстояния между классами следующие.

1. Расстояние, измеряемое по принципу «ближайшего соседа»

$$\rho_{\min}(S_l,S_m)=\min_{X_i\in S_l;Y_j\in S_m}\rho(X_i,Y_j).$$

В этом методе расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. В результате кластеры представляют длинные «цепочки».

2.Расстояние, измеряемое по принципу «дальнего соседа»:

$$\rho_{\max}(S_l,S_m)=\max_{X_i\in S_l;Y_j\in S_m}\rho(X_i,Y_j).$$

Расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т. е. «наиболее удаленными соседями»). Метод обычно работает очень хорошо, если кластеры не имеют удлиненную форму.

3. Расстояние, измеряемое по «центрам тяжести» классов

$$\rho(S_l, S_m) = \rho\left(\overline{X}(l), \overline{Y}(m)\right).$$

4. Расстояние, измеряемое по принципу «средней связи»

$$\rho_{cp}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{Y_j \in S_m} \rho\left(X_i, Y_j\right).$$

5. Обобщенное расстояние (по А. Н. Колмогорову):

$$\rho_r(S_l, S_m) = \left[\frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{Y_j \in S_m} \rho^r\left(X_i, Y_j\right)\right]^{\frac{1}{r}}.$$

Обобщенное расстояние ρ_r при $r = 1$, $\rho_1(S_l, S_m) = \rho_{cp}(S_l, S_m)$ называется средним расстоянием между классами S_l и S_m, соответствующим данной метрике ρ.

В процедурах кластеризации, использующих последовательное объединение элементов и классов, применяется следующая формула для пересчета расстояния между классом S_l, и классом $S_{m,q} = S_m \bigcup S_q$, являющимся объединением двух классов S_m и S_q:

$$\rho_r = \left(S_l, S_{m,q}\right) = \left\{\frac{n_m\left[\rho_r(S_l, S_m)\right]^r + n_q\left[\rho_r(S_l, S_q)\right]^r}{n_m + n_q}\right\}^{\frac{1}{r}},$$

где n_m и n_q — число элементов соответственно в классах S_m и S_q. С этой же целью используют также следующую формулу

$$\rho\left(S_l, S_{m,q}\right) = \alpha\rho_{lm} + \beta\rho_{lq} + \gamma\rho_{mq} + \delta\left|\rho_{lm} - \rho_{lq}\right|,$$

где α, β, γ, и δ — числовые коэффициенты, значения которых определяет выбор той или иной меры расстояния между классами. Например, при $\alpha = \beta = -\delta = 1/2, \gamma = 0$ — расстояние определяется по принципу ближайшего соседа; при $\alpha = \beta = \delta = 1/2, \gamma = 0$ — расстояние определяется по принципу дальнего соседа; при $\alpha = n_m/\left(n_m + n_q\right), \beta = n_q/\left(n_m + n_q\right)$, $\gamma = \delta = 0$ получим расстояние между классами, определяемое как среднее из расстояний между всеми парами элементов, из

которых один берется из одного класса, а второй из другого класса.

6. Статистическое расстояние между классами

$$\rho_S(S_l, S_m) = \frac{n_l n_m}{n_l + n_m}\left(\overline{X} - \overline{Y}\right)^T\left(\overline{X} - \overline{Y}\right).$$

Использование меры ρ_S обосновывается следующим образом. Рассмотрим класс, содержащий n элементов: $X_1, X_2, ..., X_n$, причем размерность каждого элемента равна p, а «центр тяжести» класса равен $\overline{X}$.

В качестве меры рассеяния элементов используют матрицу рассеяния $S_X = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^T$.

След trS_x матрицы S_x (т. е. сумму диагональных элементов) называют статистическим рассеянием множества элементов $X_1, X_2, ..., X_n$ или внутри-групповой суммой квадратов

$$trS_X = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p}\left(X_{ki} - \overline{X}_k\right)^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^T\left(X_i - \overline{X}\right).$$

Можно показать, что

$$trS_X = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_i - X_j\right)^T\left(X_i - X_j\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_E^2\left(X_i, X_j\right),$$

где суммирование проводится для значений $i<j$.

Таким образом, использование trS_x в качестве меры рассеяния класса связано с евклидовой метрикой. Определитель матрицы рассеяния $|S_X|$ называют обобщенной дисперсией множества элементов $X_1, X_2, ..., X_n$.

Рассмотрим два класса S_l, и S_m и их объединение $S_l \bigcup S_m = S(l, m)$.

Тогда матрица рассеяния для класса $S(l, m)$ равна

$$S_X(l, m) = \sum_{i=1}^{n_i}\left(X_i - M\right)\left(X_i - M\right)^T + \sum_{j=1}^{n_m}\left(Y_j - M\right)\left(Y_j - M\right)^T,$$

где $X_i \in S_l, i = 1,..., n_l$; $Y_j \in S_m$, $j = 1,..., n_m$;

$$M = \frac{1}{n_l + n_m}\left(\sum_{i=1}^{n_i} X_i + \sum_{j=1}^{n_m} Y_j\right) = \frac{1}{n_l + n_m}\left(n_l \overline{X} + n_m \overline{Y}\right).$$

Матрицу рассеяния S_X для класса $S(l, m)$ можно преобразовать так

$$S_X(l,m) = S_X(l) + S_X(m) + \frac{n_l n_m}{n_l + n_m}\left(\overline{X} - \overline{Y}\right)\left(\overline{X} - \overline{Y}\right)^T$$

Матрица $\frac{n_l n_m}{n_l + n_m}\left(\overline{X} - \overline{Y}\right)\left(\overline{X} - \overline{Y}\right)^T$ называется матрицей межгруппового рассеяния, а след этой матрицы есть статистическое расстояние между классами S_l и S_m:

$$tr\frac{n_l n_m}{n_l + n_m}\left(\overline{X} - \overline{Y}\right)\left(\overline{X} - \overline{Y}\right)^T = \frac{n_l n_m}{n_l + n_m}\left(\overline{X} - \overline{Y}\right)^T\left(\overline{X} - \overline{Y}\right) = \rho_S(S_l, S_m).$$

Меры расстояния между классами могут быть также построены на основе вероятностных распределений каждого из классов.

Методы кластерного анализа в пакете STATISTICA

В модуле ***Cluster Analysis*** пакета STATISTICA реализуются следующие методы кластеризации[7]:

- соединения (древовидная кластеризация), ***Joining (tree clustering);***
- метод *K* -средних ***(K-means clustering);***
- двухвходовое объединение ***(Two-way joining).***

Иерархические алгоритмы

Первая опция ***(Joining)*** представляет группу так называемых иерархических алгоритмов кластеризации. В основе этих алгоритмов лежит идея последовательной кластеризации. Пусть исходное множество содержит *n* объектов $X_1, X_2,..., X_n$.

В качестве расстояния между объектами X_t и Xj выбирается некоторая метрика р. Выбор метрики необходимо сделать в опции *distance measure* панели *Joining*.

1. **Метод одиночной связи *(Single Linkage).*** Кластеры

объединяются исходя из расстояния, измеряемого по методу «ближайшего соседа». Группы, между которыми расстояния самые маленькие, объединяются. Каждое объединение уменьшает число групп на единицу. Расстояние между группами определяется как расстояние между ближайшими членами групп. Метод приводит к «цепным» кластерам.

2. **Метод полной связи *(Complete Linkage).*** Расстояние между группами определяется как расстояние измеряемое по принципу «дальнего соседа». Расстояние между объединяемыми кластерами равно диаметру наименьшей сферы, содержащей оба кластера. Метод создает компактные кластеры в виде гиперсфер, которые плохо объединяются с другими кластерами. Если кластеры имеют удлиненную форму, то метод не работает.

3. **Метод невзвешенного попарного среднего *(Unweightedpair-group average).*** Расстояние между кластерами определяется по принципу «средней связи».

4. **Метод взвешенного попарного среднего (*Weighted pair-group average).*** Расстояние между кластерами определяется по принципу «средней связи», но с учетом в качестве весов числа объектов, содержащихся в кластерах.

5. **Невзвешенный центроидный метод *(Unweighted pair-group centroid).*** Расстояния между кластерами определяется как расстояние между их «центрами тяжести»

$$\rho(S_l, S_m) = \rho(\overline{X}, \overline{Y}).$$

6. **Взвешенный центроидный метод (*Weighted pair-group centroid).*** Расстояние между классами определяется как расстояние между их «центрами тяжести», но с учетом весов, определяемых по количеству объектов в каждом кластере (т. е. с учетом размеров кластеров).

7. **Метод Уорда (Ward's metod). В** этом методе в качестве целевой функции используется сумма квадратов расстояний между каждым элементом и «центром тяжести» класса, содержащего этот элемент. Кластеризация представляет последовательную процедуру, на каждом шаге которой объе-

диняются два таких класса, при объединении которых происходит минимизация статистического расстояния между классами ρ_S, вычисляемого по формуле

$$\rho_S = \frac{n_l n_m}{n_l + n_m}\left(\overline{X} - \overline{Y}\right)^T \left(\overline{X} - \overline{Y}\right).$$

Рассмотрим работу иерархического алгоритма кластеризации на простом примере.

Пример. 3. Провести кластеризацию четырех объектов методами одиночной связи (*Single Linkage*) и полной связи (*Complete Linkage*). Каждый объект определяется двумя признаками

Признак	Объект			
	1	2	3	4
x_i	0	-1	1	4
y_i	-2	0	2	0

Решение

Расстояние между объектами X_i и X_j определим как квадрат евклидовой метрики

$$\rho_{ij} = \left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2, \quad i, j = 1,2,3,4;\ i \neq j$$

Таким образом, расстояние между первым и вторым объектами равно:

$\rho_{12} = (0 + 1)^2 + (-2 - 0)^2 = 5$, а между первым и третьим объектами

$\rho_{13} = (0 - 1)^2 + (-2 - 2)^2 = 17$ и т. д.

На **первом шаге** матрица расстояний между объектами D_1, имеет вид

$$D_1 = \begin{pmatrix} 0 & 5 & 17 & 20 \\ 5 & 0 & 8 & 25 \\ 17 & 8 & 0 & 13 \\ 20 & 25 & 13 & 0 \end{pmatrix}$$

Наиболее близки первый и второй объекты: $\rho_{12} = 5$, следовательно, эти объекты объединяются в один кластер.

На втором шаге имеем следующие кластеры:

Кластеры	1	2	3
Объекты	(1,2)	3	4

Определяем расстояние между кластерами по методу «ближайшего соседа»:

$$\rho_{12}^{(2)} = \rho_{(1,2),3} = \min\left(\rho_{13};\rho_{23}\right) = \min(17;8)=8$$

$$\rho_{13}^{(2)} = \rho_{(1,2),4} = \min\left(\rho_{14};\rho_{24}\right) = \min(20;25)=20;$$

$$\rho_{23}^{(2)} = \rho_{3,4} = 13$$

Для вычисления расстояния между кластерами можно воспользоваться формулой (1):

$$\rho\left(S_l, S_{(m,q)}\right) = \alpha\rho_{lm} + \beta\rho_{lq} + \gamma\rho_{mq} + \delta\left|\rho_{lm} - \rho_{lq}\right|.$$

Расстояние по принципу «ближайшего соседа» определяется при $\alpha = \beta = \frac{1}{2}, \gamma = 0, \delta = -\frac{1}{2}$. Таким образом, например, расстояние между первым и вторым кластерами $\rho_{12}^{(2)}$ по формуле (1) равно

$$\rho_{12}^{(2)} = \rho_{(1,2)3} = \frac{1}{2}\rho_{13} + \frac{1}{2}\rho_{23} - \frac{1}{2}\left|\rho_{12} - \rho_{23}\right| =$$

$$\frac{1}{2}17 + \frac{1}{2}8 - \frac{1}{2}(17-8) = 8.$$

Матрица расстояний D_2 между тремя кластерами на втором шаге имеет вид

$$D_2 = \begin{matrix} 0 & 8 & 20 \\ 8 & 0 & 13 \\ 20 & 13 & 0 \end{matrix}$$

Как следует из матрицы расстояний D_2 наиболее близки первый и второй кластеры: $\rho_{12}^{(2)} = 8$, следовательно, эти кластеры объединяются в один кластер.

Таким образом, на **третьем шаге** имеем следующие кластеры:

Номера кластеров	1	2
Состав кластеров (в скобках номера кластеров на втором шаге)	(1,2)	3
Состав кластеров (в скобках указаны номера исходных объектов)	(1,2,3)	4

$$\rho_{12}^{(3)} = \rho_{(1,2),3}^{2} = \min\left(\rho_{13}^{(2)}; \rho_{23}^{(2)}\right) = \min(20;13) = 13.$$

Матрица расстояний D_3 между двумя кластерами на третьем шаге

$$D_3 = \begin{pmatrix} 0 & 13 \\ 13 & 0 \end{pmatrix}$$

На последнем, **четвертом шаге,** оба кластера объединяются.

Теперь рассмотрим работу алгоритма, в случае, когда расстояние между кластерами определяется по принципу «дальней связи» - *Complete Linkage.*

На **первом шаге** объединяются наиболее близкие первый и второй объекты: $\rho_{12} = 5..$

На **втором шаге** имеем следующие кластеры:

Кластеры	1	2	3
Объекты	(1,2)	3	4

Далее определяем расстояние между объектами по принципу «дальней связи»:

$$\rho_{12}^{(2)} = \rho_{(1,2),3} = \max\left(\rho_{13}; \rho_{23}\right) = \max(17;8) = 17.$$

$$\rho_{13}^{(2)} = \rho_{(1,2),4} = \max\left(\rho_{14}; \rho_{24}\right) = \max(20;25) = 25.$$

$$\rho_{23}^{(2)} = \rho_{3,4} = 13$$

Матрица расстояний D_2 между тремя кластерами на втором шаге имеет вид

$$D_2 = \begin{pmatrix} 0 & 17 & 25 \\ 17 & 0 & 13 \\ 25 & 13 & 0 \end{pmatrix}.$$

Таким образом наиболее близки второй и третий кластеры: $\rho_{32}^{(2)} = 13$. Эти кластеры объединяются и на **третьем шаге** имеем следующие кластеры:

Номера кластеров	1	2

Состав кластеров (в скобках указаны номера кластеров на втором шаге)	1	(2,3)
Состав кластеров (в скобках указаны номера исходных объектов)	(1,2)	(3,4)

Определяем расстояние между кластерами

$$\rho_{12}^{(2)} = \rho_{1,(2,3)} = \max\left(\rho_{12}^2; \rho_{23}^2\right) = \max(17;25) = 25.$$

Матрица расстояний

$$D_3 = \begin{pmatrix} 0 & 25 \\ 25 & 0 \end{pmatrix}$$

На последнем, **четвертом шаге,** оба кластера объединяются.

Выполнение иерархических процедур в пакете STATISTICA

Для реализации любого метода кластеризации из группы иерархических процедур ***Joining (tree clustering)*** необходимо сделать следующие установки:

1) выбрать переменные для анализа ***(Variables);***

2) определить вид входных данных ***(Input):*** можно вводить таблицу с координтами объектов ***(Raw data)*** либо сразу матрицу расстояний между объектами *(Distance matrix);*

3) определить объекты кластеризации: это могут быть переменные (столбцы) ***(Variables (columns)*** либо наблюдения (строки) — ***Cases (rows).*** **В** последнем случае каждая строка таблицы исходных данных есть объект;

4) выбрать метрику, определяющую расстояние между кластерами — ***Amalgamation (linkage) rule****;*

5) выбрать метрику, определяющую расстояние между объектами — ***Distance measure.***

Результаты кластеризации имеют следующий вид:

1) строится горизонтальная или вертикальная дендрограмма — график, на котором определены расстояния между объектами и кластерами при их последовательном объединении. Древовидная структура графика позволяет определить кластеры в зависимости от выбранного порога — заданного расстояния между кластерами;

2) выводится матрица расстояний между исходными

объектами ***(Distance matrix);***

3) выводятся средние и среднеквадратичные отклонения для каждого исходного объекта ***(Distiptive statistics)***.

Рассмотрим решение **примера** в пакете STATISTICA.

Нажмите кнопку ***Module Switcher*** на панели инструментов, в появившемся окне выберете модуль *Cluster Analysis,* а затем *Joining (tree clustering).* В новом окне выполните следующие настройки:

а) нажмите на кнопку ***Variables*** и введите имена двух переменных *x* и *y,* в которых записаны исходные данные примера ;

б) в разделе *Input* введите *Raw data (исходные данные);*

в) в разделе ***Cluster*** выберите ***Cases (rows).*** При этой установке объекты кластеризации — двумерные наблюдения с координатами x_i и y_i , i=1,2, 3, 4;

г) в разделе ***Amalgamation (linkage) rule*** выберите ***Single Linkage*** *(метододиночной связи);*

д) в разделе ***Distance measure*** выберите ***Squared Euclidean distances*** *(квадрат евклидовой метрики)* и нажмите ОК.

В появившемся окне нажмите на кнопку ***Vertical icicle plot.*** На экране появится дендрограмма (рис. 21), показывающая объединение объектов, расстояние между которыми является наименьшим, в кластеры.

На вертикальной оси дендрограммы откладываются расстояния между объектами и между объектами и кластерами. Так, расстояние между объектами С, и C_2 равно 5 (см. матрицу расстояний D_2, в примере 3). Эти объекты объединяются в один кластер на первом шаге.

Расстояние между этим кластером и объектом C_3 равно 8 (см. матрицу расстояний D_2). Объект C_3 объединяется с кластером (C_1, C_2) на втором шаге. Наконец, расстояние между объектом C_4 и кластером (C_1, C_2, C_3) равно 13 (см. матрицу расстояний D_3).

Таким образом, горизонтальные отрезки дендрограммы проводятся на уровнях, соответствующих пороговым значениям расстояний, выбираемым для данного шага кластеризации.

Кластеризация методом одиночной связи (ближайшего соседа) приводит к образованию одного кластера (пороговое расстояние равно 13).

Далее последовательно нажмите ***Continue...*** и ***Cancel*** и в окне установок процедуры в разделе ***Amalgamation...*** выберите ***Complete Linkage***. После выполнения процедуры появится следующая дендрограмма (рис. 22).

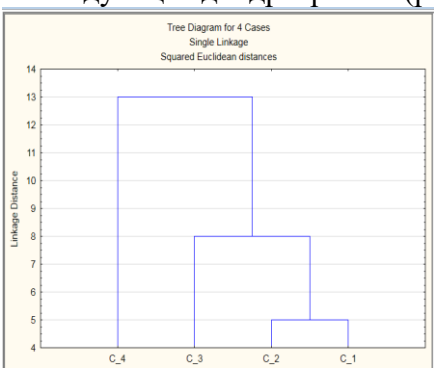


Рис.21. Дендрограмма при методе одиночной связи

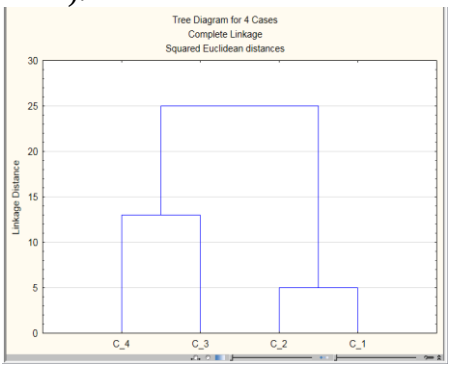


Рис. 22. Дендрограмма при методе полной связи

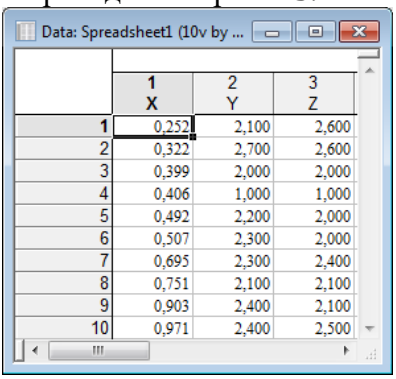
Вторая дендрогамма показывает, что кластеризация методом полной связи (дальнего соседа) при таком же пороговом расстоянии равным 13, приводит к образованию двух кластеров. Сравните полученные дендрограммы с результатами решения примера 3.

Метод *K*-средних

Метод *K*-средних относится к группе так называемых эталонных методов кластерного анализа. Число кластеров *K* задается пользователем. Процедура состоит в следующем. На первом шаге определяют *K* кластеров — эталонов (это могут быть, например, первые *K* объектов). Далее каждый объект присоединяется к ближайшему эталону. В качестве критерия используется минимальное расстояние внутри кластера отно-

сительно среднего. Как только объект включается в кластер, среднее пересчитывается. После пересчета эталона объекты снова распределяются по ближайшим кластерам и т. д. Процедура заканчивается при стабилизации процесса, т. е. при стабилизации центров тяжести.

Пример 4. Провести классификацию $n = 10$ объектов, каждый из которых характеризуется тремя признаками: *x, y* и *z*. Таблица данных приведена на рис. 23.


Data: Spreadsheet1 (10v by ...

	1 X	2 Y	3 Z
1	0,252	2,100	2,600
2	0,322	2,700	2,600
3	0,399	2,000	2,000
4	0,406	1,000	1,000
5	0,492	2,200	2,000
6	0,507	2,300	2,000
7	0,695	2,300	2,400
8	0,751	2,100	2,100
9	0,903	2,400	2,100
10	0,971	2,400	2,500

Рис. 23. Данные для примера 4

Решение в пакете STATISTICA

1. Визуализация данных (в трехмерном случае).

В меню ***Graphs*** выберите ***3D XYZ Graphs.*** В выпадающем меню выберите команду *Scatterplots*, в появившемся окне нажмите на кнопку ***Variables*** и задайте *X, Y, Z* (рис. 4). Затем нажмите на кнопку ***Options 1*** и включите опцию ***Display Case labels*** (имена наблюдений), нажмите ОК.

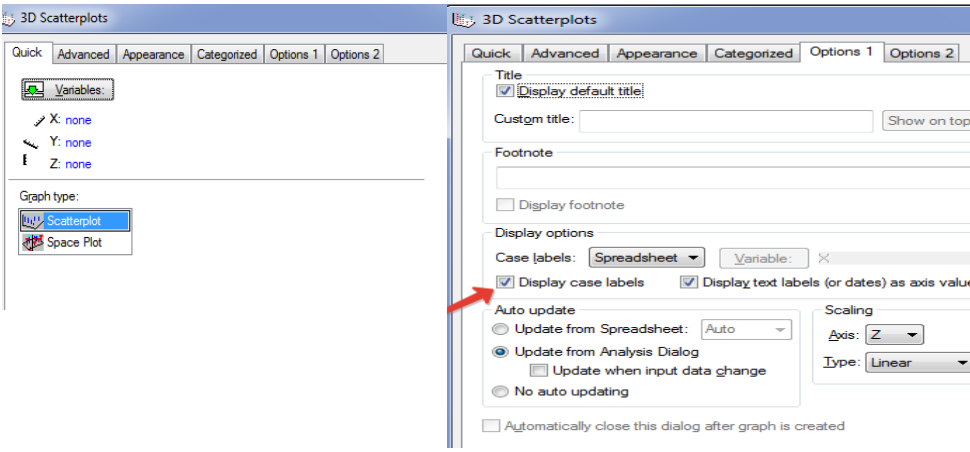


Рис. 24. Ввод данных для построения диаграммы рассеяния

На экране появится диаграмма рассеяния для исходных данных (рис. 25). По диаграмме видно, что объекты образуют три кластера.

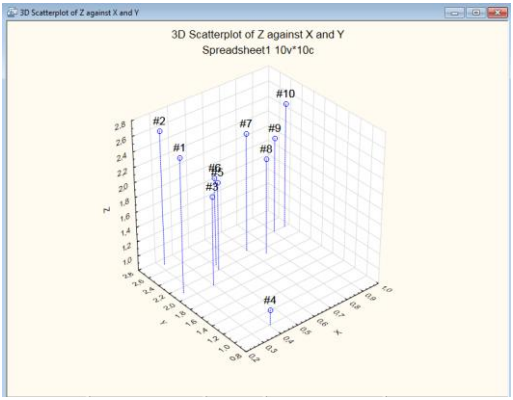
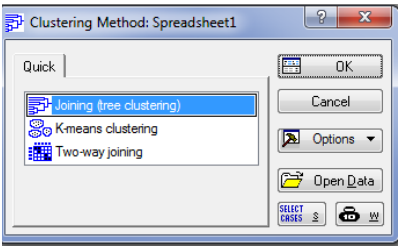


Рис. 25. Диаграмма рассеяния для примера 4.

2. Проведем кластерный анализ с помощью метода *K*-средних ***(K-means clustering)***. На панели инструментов нажмите кнопку *Module Switcher*, в по явившемся окне выберите *Cluster Analysis* и в стартовой панели модуля выберете *K-means clustering*. В новом окне выполните следующие настройки (рис.26):



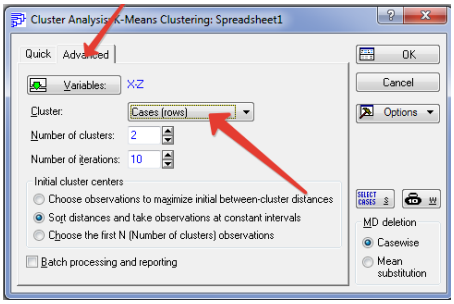


Рис.26. Настройки метода *K*-средних

Далее:

а) нажмите на кнопку *Variables* и введите переменные *X, Y, Z;*

б) в разделе *Cluster* выберите *Cases (rows);*

в) в разделе *Number of clusters* задайте число кластеров, равное трем;

г) задайте число итераций;

д) выберете один из трех методов для начального определения центров кластеров (эталонов): либо выбираются первые *K* объектов, либо выбираются объекты наиболее отстоящие друг от друга, либо отстоящие друг от

друга на одинаковом расстоянии. После выбора установок нажмите ОК.

3.Результаты кластеризации.

а. ***Analysis of variance*** — результаты дисперсионного анализа по каждому признаку *X, Y, Z* (рис. 7): выводятся суммы квадратов отклонения объектов от центров кластеров ***(SS Within)*** и суммы квадратов отклонений между центрами кластеров ***(SS Between)***, значения *K*-статистики и уровни значимости *p*.

Analysis of Variance (Spreadsheet1)

Variable	Between SS	df	Within SS	df	F	signif. p
X	0,035456	2	0,513437	7	0,2417	0,791584
Y	1,537500	2	0,287500	7	18,7174	0,001552
Z	1,941500	2	0,039500	7	172,0320	0,000001

Рис. 27. Результаты дисперсионного анализа

б. Выводятся координаты центров и матрицы расстояний между центрами (рис. 28).

Cluster Number	Euclidean Distances between Clusters (Spr Distances below diagonal Squared distances above diagonal No. 1	No. 2	No. 3
No. 1	0,000000	1,413322	0,854460
No. 2	1,188832	0,000000	0,089463
No. 3	0,924370	0,299104	0,000000

Рис. 28. Матрица расстояний между центрами классов

в. График распределения центров кластеров (рис. 29):

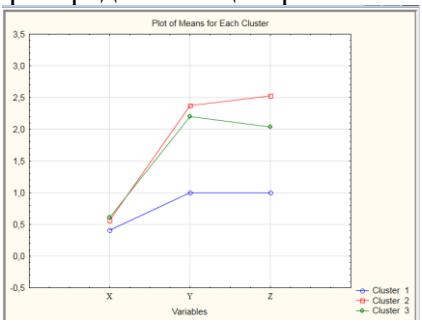


Рис. 29. График координат центров классов

г. Статистики для каждого кластера по координатам *X, Y, Z:* средние центров, стандартные отклонения и дисперсии.

д. Номера объектов, входящих в каждый кластер и расстояния объектов до центра каждого кластера.

В данном примере объекты распределились следующим образом: кластер 1: {4}; кластер 2: {1, 2, 7, 10}; кластер 3: {3, 5, 6, 8, 9}.

Задания для самостоятельного решения

1. Используя иерархические алгоритмы проведите кластеризацию данных из примера 4. Сравните результаты с результатами в примере 4.

2. На предприятии существует 16 научно-производственных отделов, занятых выпуском различной продукции, работ, услуг. Поскольку виды деятельности, количество работающих, рентабельность отделов, существенно различаются между собой, *было решено сгруппировать отделы в несколько однородных групп, а затем для каждой группы разработать свою систему премирования.*

После тщательного анализа выбрали четыре признака, с помощью которых описывались важные (для указанной цели) параметры каждого отдела: X_1 — стоимость активной части основных производственных фондов, тыс. руб.; X_2 — среднемесячный объем работ отдела, тыс. руб.; X_3 — удельный вес работ/услуг отдела по внутрифирменной кооперации, %; X_4 — среднемесячная прибыль отдела, тыс. руб.

Исходные данные по отделам приведены ниже.

Проведите кластеризацию отделов используя иерархические алгоритмы (Joining): а) используя исходные данные; используя стандартизованные данные, т. е. данные, преобразованные по формуле $\frac{X_{ij} - \overline{X_j}}{S_j}$, где $X_{ij} - i - \mathbf{e}$ значение $j - \mathbf{го}$ признака, $i = 1,2,...,16;\ j = 1,2,3,4;$

№ отдела	Значения признаков			
	X_1	X_2	X_3	X_4
1	699	190	53	11
2	532	211	19	42
3	650	152	46	14
4	768	216	67	17
5	67	106	0	32
6	322	397	26	52
7	736	180	49	18
8	501	239	11	60
9	293	391	16	66
10	300	396	29	87

$\overline{X}_j = \frac{1}{16}\sum_{i=1}^{16} x_{ij} -$ оценка среднего для $j - \mathbf{го}$ признака;

$S_j = \sqrt{\frac{1}{15}\sum_{i=1}^{16}\left(x_{ij} - \bar{x}_j\right)^2}$ – оценка среднего квадратического отклонения для $j-$**го** признака. Процедуру стандартизации данных можно выполнить непосредственно в таблице, используя следующую последовательность действий: курсор на имени переменной → нажать правую кнопку мыши → в выпадающем меню выбрать **File/Standardize Block** → **Standardize Columns** → ОК.

Сравните результаты кластеризации. По результатам кластеризации определите число кластеров и их состав. Найдите статистические характеристики каждого кластера.

Проведите кластеризацию используя метод *K*-средних (число кластеров задайте равным 4). Сравните результаты (составы кластеров).

3. Ниже приведены значения основных факторов сельскохозяйственного производства для 20 районов: x_1 — число тракторов 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 (т/га); x_5 — количество хим. средств защиты растений, расходуемых на гектар (ц/га).

Районы	Факторы				
	x_1	x_2	x_3	x_4	x_5
1	1,59	0,26	2,05	0,32	0,14
2	0,34	0,28	0,46	0,59	0,66
3	2,53	0,31	2,46	0,30	0,31
4	4,63	0,40	6,44	0,43	0,59
5	2,16	0,26	2,16	0,39	0,16
6	2,16	0,30	2,69	0,32	0,17
7	0,68	0,29	0,73	0,42	0,23
8	0,35	0,26	0,42	0,21	0,08
9	0,52	0,24	0,49	0,20	0,08
10	3,42	0,31	3,02	1,37	0,73

1) проведите кластеризацию районов используя несколько иерархических алгоритмов (Joining): используя исходные данные; используя стандартизованные данные;

2) проведите кластеризацию используя метод K-средних (число кластеров задайте равным

3). Сравните составы кластеров и их характеристики.

Лабораторная работа №4
Временные ряды
Основные понятия

Автокорреляционная функция. Сериальные корреляции. Аддитивная и мультипликативная модели временного ряда. Метод скользящих средних. Сезонные индексы. Случайная составляющая временного ряда [5].

Типовое задание

Используя выборку данных из своего варианта (24 значения), выполнить следующие задания:

1. Построить график ряда.

2. Вычислить сглаженные ряды, используя простые скользящие средние по: а) трем точкам; б) четырем точкам (после сглаживания провести центрирование); в) пяти точкам.

Сглаженные ряды нанести на три отдельных графика вместе с исходными данными.

3. Рассчитать четыре сезонных индекса для исходного ряда по аддитивной модели ряда. Построить на одном графике:

а) исходные данные y_t;

б) центрированные скользящие средние (оценка тренда) u_t;

в) сезонные индексы = S_1, S_2, S_3, S_4;

г) данные без сезонной составляющей $V_t = y_t - S_i$;

д) остатки $e_t = V_t - u_t = y_t - S_i - u_t$.

4. Повторить расчеты из пункта 3 для мультипликативной модели ряда и построить графики а) и д).

5. Найти дисперсии остатков для обеих моделей ряда. Сравнить результаты и выбрать подходящую модель.

6. Ввести данные в пакет STATISTICA. Выполнить все расчеты в п. 1)—5), сравнить результаты и записать их в отчет.

Выполнение задания в пакете STATISTICA

Для того, чтобы войти в модуль Анализ временных рядов и прогнозирование необходимо в **Переключателе модулей (Statistica Module Switcher)** выбрать модуль **Time series/Forecasting.** Общее назначение модуля — построить простую модель, описывающую ряд, сгладить его, спрогнозировать будущие значения временного ряда на основе наблюдаемых до данного момента, построить регрессионные зависимости одного ряда от другого, провести спектральный или Фурье-анализ ряда и т. д.

В этом модуле реализованы следующие методы анализа временных рядов: ARIMA-АРПСС: модель авторегрессии и проинтегрированного скользящего среднего; Interrupted time series analysis — анализ прерванного временного ряда (модели интервенции для АРПСС); Exponential smoothing & forecasting — экспоненциальное сглаживание и прогнозирование; Seasonal decomposition (Census 1) — сезонная декомпозиция; Xll (Census 2) — monthly — quarterly — Xll метод — месячно — квартально — специальный метод сезонной декомпозиции; distributed lags analysis — анализ распределенных лагов; Spectral (Fourier) analysis — спектральный (Фурье) анализ.

Рассмотрим простейшие методы анализа временных рядов.

1. Выделение тренда методом скользящих средних

В качестве примера возьмем следующие данные: 8,8; 13,5; 18,9; 15,0; 9,8; 16,0; 22,1; 16,9; 10,9; 17,8; 24,4; 18,5; 12,3; 20,2; 27,8; 20,2; 13,5; 23,1; 33,1; 21,9; 13,7; 24; 33,5; 22,1.

Введите данные в пакет STATISTICA.

Выберите в головном меню **Advanced Linear/Linear Models->Time Series/Forecasting**

Для начала анализа необходимо вызвать стартовую панель модуля, для этого войдите в меню **Analysis** — **Анализ** и выберите команду **Startup Panel.** Далее выберите переменную для

анализа (воспользуйтесь кнопкой Variables). Кнопка **Variables — Переменные,** расположенная в левом верхнем углу, открывает диалоговое окно выбора переменных из файла данных. Нажмите ее и откройте диалоговое окно **Select the variables for the time series analysis — Выбрать переменные для анализа временных рядов. В** окне можно выбрать переменные для анализа (максимальное число переменных 20). Выбор можно осуществить либо высвечивая имена в верхней части окна, либо задавая номера переменных в нижней строке.

После определения переменной на экране появится следующее окно (рис.30).

После того как файл открыт и выбраны переменные для анализа, в информационной части панели в поле **Variable — Переменная** появятся имена переменных, расширенные имена автоматически отображаются в графе **Long variable (series) name.**

Слева от имени анализируемых переменных стоит значок **L** в графе **Lock,** означающий, что переменные закрыты на ключ и не могут быть удалены без прерывания анализа.

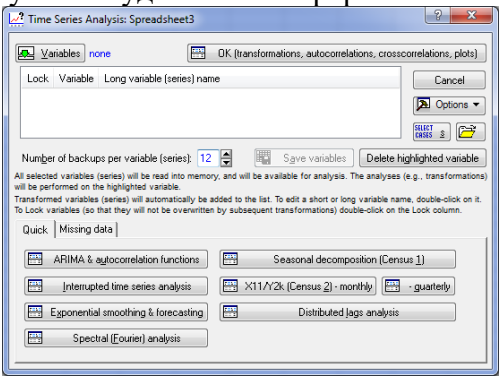


Рис. 30. Стартовая панель модуля

Весь дальнейший диалог происходит именно с этими переменными, которые можно преобразовывать, анализировать, но нельзя удалить из текущего анализа.

В процессе работы ряды многократно преобразовываются, однако, не все преобразования необходимы; чтобы не хранить лишнюю информацию, их следует удалить из диалога. Для

этого служит кнопка **Delete highlighted variable — Удалить высвеченные переменные.**

Напротив, некоторые переменные нужно сохранить для дальнейшего анализа, например, для того чтобы применить альтернативный способ обработки.

Кнопка **Save variables — Сохранить переменные** сохраняет высвеченные переменные в файле данных STATISTICA. Сохраненную таким образом переменную можно проанализировать впоследствии в любом модуле STATISTICA.

В верхней части стартовой панели расположена опция **Number of backups per variables (series) — Число резервов для переменных (рядов),** которая определяет число преобразований ряда в текущем диалоге. Если число преобразований превысит указанное в опции число, то система сделает запрос: сохранять очередное преобразование?

Построим график исходных данных. Для этого в окне Стартовой панели необходимо нажать на кнопку ОК, в результате чего на экране появится следующее окно (рис. 31).

Обратите внимание на опцию **Plot variables (series) after each transformation — Построить график переменных (ряда) после каждого преобразования в данном окне.** После того как вы установите ее, система будет автоматически выдавать график преобразованных данных после каждого преобразования ряда. Опция **Display/plot subset of cases only** позволяет просмотреть численно или построить график только для определенного подмножества данных.

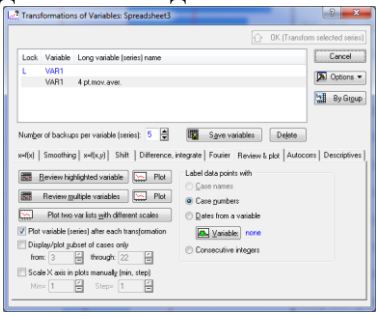


Рис. 31. Окно для выбора преобразования переменных

Следует обратить внимание на кнопки:

- **Review highlighted variable**, эта функция выводит на экран значения выделенной переменной;
- **Plot** (верхняя), выводит график выделенной переменной;
- **Review multiple variables**, эта функция выводит численные результаты для нескольких выделенных переменных представляющих результаты преобразования временного ряда (включая исходные данные);
- **Plot** (нижняя), выводит графики выделенных переменных.

Построим график исходных данных (рис. 32).

Чтобы определить тренд методом скользящих средних необходимо в этом же окне еще раз нажать на кнопку ОК. В результате появится следующее окно (рис. 33).

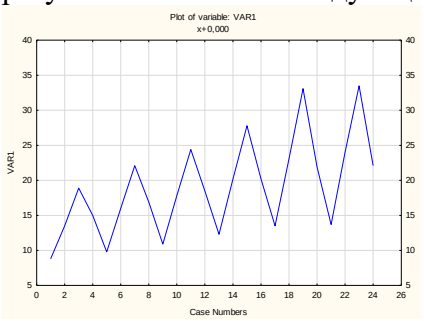


Рис. 32. График исходных данных

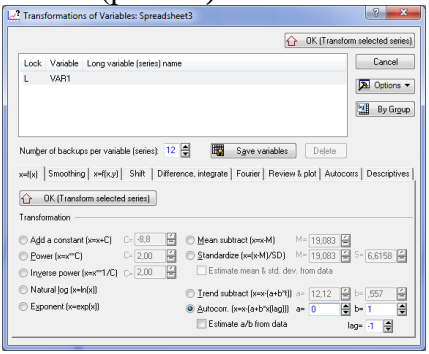


Рис. 33. Окно для преобразования ряда

В левой нижней части окна, в поле **Smoothing — Сглаживание** необходимо выделить опцию **N-pts mov. averg. — сглаживание по методу скользящих средних.** Данная функция позволяет произвести сглаживание по двум, трем и более точкам *(N=2*, 3, ...). Установите *N=4* (сглаживание по четырем точкам) и нажмите на кнопку ОК. На экране появится график сглаженного ряда. Для того, чтобы просмотреть одновременно исходные данные и результаты процедуры простого скользящего среднего по четырем точкам нужно нажать кнопку **Continue...** и воспользоваться функцией **Review multiple variables** в окне преобразования переменных (рис. 34). Функция Plot

позволит просмотреть скользящие средние на графике одновременно с исходными данными.

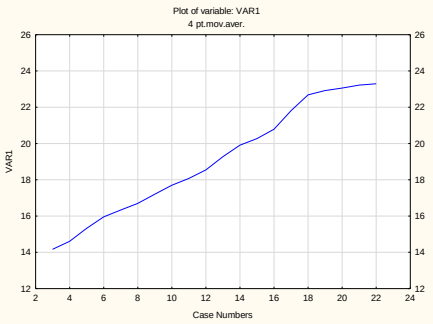


Рис. 34. Исходные данные и результаты сглаживания по четырем точкам

Если необходимо просмотреть результаты сглаживания численно, нужно нажать кнопку **Review multiple variables**, выбрать переменные и нажать ОК.

Чтобы вернуться в стартовую панель модуля нажмите **Exit (Выход).**

2. Преобразования временных рядов. Вычисление коррелограммы

В левой части окна **Time series Transformations** — преобразования временного ряда (рис. 33) имеется ряд опций позволяющих выполнить различные преобразования исходного временного ряда: добавление константы, возведение в целую и др. Преобразования можно выполнять последовательно. Одна из целей преобразований состоит в том, чтобы сделать ряд стационарным. В случае линейного тренда это можно сделать с помощью опции Trend subtract (удаление тренда). Параметры a и b линейного тренда $a + bt$ могут задаваться или оцениваться по исходным данным. В правом нижнем углу размещена опция **Differencing** — взятие разностей определенного порядка. Чтобы вычислить коррелограмму надо переключиться в окно **Transformations of Variables** — преобразования переменных и выбрать опцию **Autocorrelations** (автокорреляции) в правой верхней части окна. Число вычисляемых сериальных корреляций задается в окне **Number of lags**. На рис. 35 приведена коррелограмма для исходного ряда.

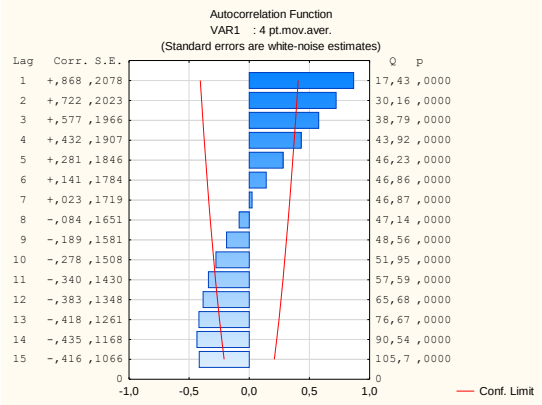


Рис. 35. Коррелограмма

3. Сезонная декомпозиция

Seasonal decomposition — Сезонная декомпозиция позволяет выделить в ряде сезонную компоненту, обозначаемую ***S***, тренд-циклическую компоненту — ***TC*** и нерегулярную (случайную) составляющую — ***I***. Модель может быть мультипликативной или аддитивной.

Нажмите кнопку **Seasonal decomposition (Census 1) — Сезонная декомпозиция** на стартовой панели модуля (рис. 36) и откройте диалог **Сезонная декомпозиция** (рис.37).

В центральной части панели находятся опции, позволяющие задать модель ряда. Эти опции объединены в группу **Seasonal model — Сезонная модель: Additive** — Аддитивная; **Multiplicative** — Мультипликативная.

В опции **Seasonal lag** — **Сезонный лаг** задается число сезонных индексов. Следующая группа опций позволяет определить следующие составляющие: **Moving averages** — Скользящие средние; **Ratios/Differences** — Отношения/Разности (если модель мультипликативная, берется отношение, если аддитивная — разность исходного ряда и тренда); **Seasonal factors** — Сезонные индексы; **Seasonal adj. series** — Ряд без сезонной составляющей; **Smoothed trend cycle** — Сглаженная тренд — циклическая компонента; **Irregular components** — Нерегулярная (случайная) составляющая.

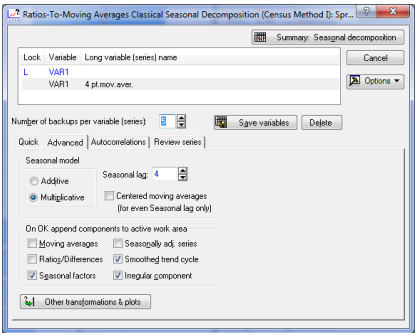


Рис. 36. Панель для установки параметров сезонной декомпозиции

Запустите процедуру сезонной декомпозиции, нажав кнопку **OK (Perform seasonal decomposition) — Да (Выполнить сезонную декомпозицию).** Результаты расчетов для мультипликативной модели выводится в виде таблицы (рис. 37).

Для того чтобы на один график вывести графики нескольких компонент надо отметить их на левой части панели (рис. 36), нажать ОК и затем выбрать их с помощью кнопки **Review Multiple Variables** и нажать **Plot.**

Таблица, приведенная на рис. 8, получилась в результате следующей последовательности преобразований временного ряда *(Y)*. Во втором столбце (**Moving Averages**) приведены простые скользящие средние по четырем точкам временного ряда (без центрирования), причем в таблице значения скользящих средних смещены вниз на 0,5 строки: среднее арифметическое четырех первых точек $\frac{8{,}8+13{,}5+18{,}9+15}{4}=14{,}05$ определяется как значение скользящего среднего в третьей точке и т. д.

В третьем и четвертом столбцах (Rations и Seasonal Factors) вычисляются соответственно отношения элементов исходного ряда к скользящему среднему (в процентах) и скорректированные сезонные индексы.

В пятом столбце (Adjusted Series) вычисляется ряд скорректированный на сезонные индексы, т. е. ряд без сезонной составляющей (вычисляется делением элементов исходного рада (*Y*) на сезонные индексы и умножением результата на 100).

В шестом столбце вычисляется сглаженная тренд-циклическая составляющая (Smoothed Trend-c), т. е. приводятся результаты сглаживания ряда скорректированного на сезонные индексы. Сглаживание выполняется с помощью процедур скользящего среднего по пяти точкам с весами 1, 2, 3, 2, 1. Например, значение в третьей точке вычисляется по формуле

$$\frac{1*13{,}86+2*13{,}40+3*13{,}85+2*15{,}10+1*15{,}44}{9}\approx 14{,}02$$

Значения сглаженного ряда в первых двух точках и в последних двух точках вычисляются по специальным формулам.

Case	VAR1	Moving Averages	Ratios	Seasonal Factors	Adjusted Series	Smoothed Trend-c.	Irreg. Compon.
1	8,80000			63,4753	13,86366	13,45559	1,030327
2	13,50000			100,7384	13,40105	13,70591	0,977757
3	18,90000	14,05000	134,5196	136,4323	13,85302	14,20655	0,975115
4	15,00000	14,30000	104,8951	99,3540	15,09753	14,79562	1,020405
5	9,80000	14,92500	65,6616	63,4753	15,43908	15,36992	1,004499
6	16,00000	15,72500	101,7488	100,7384	15,88273	15,89231	0,999397
7	22,10000	16,20000	136,4198	136,4323	16,19851	16,33243	0,991800
8	16,90000	16,47500	102,5797	99,3540	17,00988	16,81367	1,011670
9	10,90000	16,92500	64,4018	63,4753	17,17203	17,21753	0,997358
10	17,80000	17,50000	101,7143	100,7384	17,66953	17,63905	1,001728
11	24,40000	17,90000	136,3128	136,4323	17,88433	18,08692	0,988799
12	18,50000	18,25000	101,3699	99,3540	18,62029	18,67847	0,996885
13	12,30000	18,85000	65,2520	63,4753	19,37762	19,30423	1,003802
14	20,20000	19,70000	102,5381	100,7384	20,05194	19,84617	1,010369
15	27,80000	20,12500	138,1366	136,4323	20,37640	20,28239	1,004635
16	20,20000	20,42500	98,8984	99,3540	20,33134	20,80730	0,977125
17	13,50000	21,15000	63,8298	63,4753	21,26811	21,66288	0,981777
18	23,10000	22,47500	102,7809	100,7384	22,93069	22,46936	1,020531
19	33,10000	22,90000	144,5415	136,4323	24,26111	22,84231	1,062113
20	21,90000	22,95000	95,4248	99,3540	22,04239	22,73006	0,969746
21	13,70000	23,17500	59,1154	63,4753	21,58320	22,81089	0,946180
22	24,00000	23,27500	103,1149	100,7384	23,82409	23,11482	1,030685
23	33,50000	23,32500	143,6227	136,4323	24,55430	23,54069	1,043058
24	22,10000			99,3540	22,24369	23,75363	0,936433

Рис. 37. Результаты сезонной декомпозиции

В последнем, седьмом столбце, приводится остаточная (случайная) компонента ряда (Irreg. Compon.). Остаточная компонента вычисляется делением значений скорректированного ряда (пятый столбец) на значение сглаженного ряда (шестой столбец).

Задачи для самостоятельного решения

1. Количество персональных компьютеров в Университете за последние 6 лет составило соответственно:

Год	1	2	3	4	5	6
Количество PC	50	100	350	1020	1950	3710

(а) Постройте график по этим данным.

(б) Найдите уравнение линейного и квадратичного трендов.

(в) Оцените количество PC, которое будет использоваться в Университете на седьмой год, используя уравнения линейного и квадратичного трендов.

2. Следующая таблица описывает изменение почтовых тарифов в течение одиннадцати лет.

Год	1	2	3	4	5	6	7	8	9	10	11
Тариф	5	5	8	8	10	13	15	18	20	22	25

(а) Постройте график по этим данным.

(б) Найдите уравнение линейного и квадратичного трендов.

3. Компания, специализирующаяся на производстве очистительных устройств, зафиксировала следующий объем продаж за последние 9 лет.

Год	1	2	3	4	5	6	7	8	9
Количество (× 1 000 000 руб.)	13	15	19	21	27	35	47	49	57

(а) Постройте график по этим данным.

(б) Найдите уравнение линейного и квадратичного трендов.

(в) Какая из оценок тренда будет наиболее точной?

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Многомерное признаковое пространство. Особенности обработки многомерных статистических данных.

2. Какие предпосылки проведения регрессионного анализа должны выполнятся?

3. Какие задачи решаются методами регрессионного анализа?

4. Какие оценки параметров уравнения регрессии называются несмещенными?

5. Что характеризует парный, частный и множественный коэффициенты корреляции? Сформулируйте их основные свойства.

6. В чем состоят отрицательные последствия мультиколлинеарности и как можно избавиться от этого негативного явления?

7. Приведите практический пример задачи регрессионного анализа.

8. В чём состоит задача парной линейной регрессии?

9. Сформулируйте и запишите постановку простейшей задачи парной линейной регрессии.

10. Что такое остатки регрессии?

11. Как можно использовать линейную регрессию для построения прогнозов?

12. Запишите выражение для суммы квадратов отклонений от линии регрессии, когда искомая функциональная зависимость – многочлен второй степени.

13. Решите аналитически простейшую задачу парной линейной регрессии методом наименьших квадратов.

14. Запишите выражения для точечных оценок параметров регрессии в простейшей задаче парной линейной регрессии.

15. Запишите выражение для стандартного отклонения регрессии в простейшей задача парной линейной регрессии.

16. Запишите выражение для стандартного отклонения остатков в простейшей задача парной линейной регрессии.

17. Приведите общую постановку задачи регрессионного анализа.

18. Приведите формулировку задачи о построении доверительных интервалов для коэффициентов регрессии.

19. Опишите порядок вычислений при построении доверительных интервалов.

20. Как влияет на ширину доверительного интервала дополнительная информация о дисперсии остатков?

21. В чем состоит задача компонентного анализа, как интерпретировать главные компоненты и определить их вклад в суммарную дисперсию?

22. Что означает построить доверительный интервал для параметра регрессионной модели?

23. Для чего в анализе качества модели применяется F–критерий и t–критерий?

24. Как осуществить точечный прогноз по уравнению множественной линейной регрессии?

25. Что такое классификация? Чем она отличается от кластеризации?

26. Какие задачи решает кластерный анализ? В чем особенности иерархических кластер-процедур?

27. Кластерный анализ как метод многомерной классификации. Методы определения расстояний между объектами исследования.

28. Определение расстояния между классами в кластерном анализе.

29. Характеристики близости объектов и показателей в кластерном анализе. Функционалы качества разбиения.

30. Иерархические кластер-процедуры.

31. Метод многомерных средних.

32. Области применения анализа временных рядов.

33. Что характеризует автокорреляционная функция.

34. Какие корреляции называются сериальными.

35. Аддитивная и мультипликативная модели временного ряда.

36. В чём заключается метод скользящих средних?

37. Что такое сезонные индексы.

38. Как определяется случайная составляющая временного ряда.

39. Как вычисляются ошибки прогноза.

40. В чём особенности прогнозирования на основе метода экспоненциального сглаживания.

ЗАКЛЮЧЕНИЕ

В результате обучения компетенции должны быть сформированы компетенции в области владения основными методами, способами и средствами получения, хранения и переработки информации (ОК1-2). Также в рамках базовой части общепрофессионального цикла дисциплин навыками профессиональных дисциплин студент дорлжен владеть навыками измерений и обработки экспериментальных данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

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

2.Айвазян С.А. Прикладная статистика. Классификация и снижение размерности / С.А. Айвазян, Бухштабер В.М., Енюков И.С, Мешалкин Л.Д. М.: Финансы и статистика, 1989.

3.Боровиков В.П. Прогнозирование в системе Statistica / В.П. Боровиков, Г.И. Ивченко М.: Финансы и статистика, 2000.

4.Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL/Э.А. Вуколов. Учебное пособие. – 3-е изд., испр. и доп. – М.: Форум, 2008. – 464 с.

5. Кендалл М. Многомерный статистический анализ и временные ряды/М. Кендалл, А. Стьюарт. М.: Наука, 1976.

6. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988.

7.Тюрин Ю. Н., Анализ данных на компьютере/ Ю.Н. Тюрин, А.А. Макаров. М.: Инфра-М, 2003.

СОДЕРЖАНИЕ

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к лабораторным работам № 1-4 по дисциплине
«Современные технологии обработки информации»
для бакалавров направления 152200 «Наноинженерия»
профиль «Инженерные нанотехнологии
в приборостроении» очной формы обучения

Составитель
Разинкин Константин Александрович

В авторской редакции

Подписано к изданию 28.11.2014
Уч.-изд. л. 4,6875

ФГБОУ ВПО «Воронежский государственный
технический университет»
394026 Воронеж, Московский просп., 14