Р.З. Абдуллин, В.Р. Абдуллин

ЭКОНОМЕТРИКА В MS EXCEL

Практикум

Министерство образования и науки Российской Федерации
Байкальский государственный университет

Р.З. Абдуллин, В.Р. Абдуллин

ЭКОНОМЕТРИКА В MS EXCEL

Практикум

Иркутск
Издательство БГУ
2016

УДК 519.862.6(075.8)
ББК 22.1я7
А13

Печатается по решению редакционно-издательского совета
Байкальского государственного университета

Рецензенты: канд. физ.-мат. наук, доц. О.В. Леонова
канд. физ.-мат. наук, доц. М.П. Базилевский

Абдуллин Р.З.
А13 Эконометрика в MS Excel [Электронный ресурс] : практикум / Р.З. Абдуллин, В.Р. Абдуллин. – Иркутск : Изд-во БГУ, 2016. – 135 с. – Режим доступа: http://lib-catalog.isea.ru.

Содержит указания по выполнению лабораторных (расчетно-графических) работ по описательной статистике, корреляционному и дисперсионному анализу, построению уравнений регрессии, выделению тенденции временного ряда с использованием надстройки «Анализ данных» MS Excel. Включает десять лабораторных работ, каждая из которых сопровождается краткими теоретическими сведениями, разобранным примером выполнения работы и контрольными вопросами по теме работы. Составлен на основании федеральных государственных образовательных стандартов высшего образования по направлениям подготовки для уровня бакалавриата 38.03.01 Экономика, 38.03.02 Менеджмент, 38.03.03 Управление персоналом, 38.03.06 Торговое дело.

Для студентов очной и заочной форм обучения.

УДК 519.862.6(075.8)
ББК 22.1я7

СОДЕРЖАНИЕ

Предисловие

Любая область экономической деятельности связана с необходимостью количественного описания, анализа и прогнозирования экономических явлений и их взаимосвязей на основе реальных данных. Это требует применения современных методов и инструментов обработки социально-экономической информации, знания достижений экономической мысли и понимания современного научного языка. Большинство современных методов основывается на эконометрических моделях, концепциях и приемах; моделях временных рядов. Без знания основ эконометрики и анализа временных рядов научиться использовать их невозможно. Изучение современной экономической литературы также предполагает хорошую подготовку в области математической статистики и эконометрики.

Практикум предназначен для студентов очной и заочной форм обучения и ориентирован на освоение начального курса эконометрики и получение навыков статистического анализа и построения эконометрических моделей с использованием пакета прикладных программ.

В настоящее время существует большое число статистических и эконометрических пакетов программ (STATISTICA, STATGRAPHICS, SPSS, GAUSS, Mesosaur, Econometric Views и другие). Выбор «Пакета анализа» MS Excel обусловлен его широким распространением, доступностью, изучением MS Excel в курсе «Информатика», а также тем, что он обладает достаточным набором средств статистического анализа и математических операций для решения задач, входящих в начальный курс эконометрики.

Практикум состоит из трех частей. Первая часть, включающая три лабораторные работы, посвящена описательной статистике, статистической оценке параметров, проверке статистических гипотез, парной и множественной корреляции и однофакторному дисперсионному анализу. Она является необходимой базой при построении эконометрических моделей.

Вторая часть, включающая пять работ, связана с построением моделей парной линейной и нелинейной регрессии, линейной множественной регрессии, верификацией моделей и оценкой качества моделей, построением точечных и интервальных прогнозов, анализом мультиколлинеарности, исследованием автокорреляции, использованием фиктивных переменных в эконометрических моделях.

Третья часть, включающая две работы, посвящена выделению тенденции временного ряда и его аналитическому выравниванию, а также выявлению сезонных колебаний.

Все лабораторные работы содержат краткие теоретические сведения; содержание и этапы выполнения работы; примеры решения типовых задач в MS Excel с необходимыми пояснениями порядка действий и диалоговых окон; контрольные вопросы по теме работы и разнообразные данные для исследования, охватывающие широкий спектр направлений применения эконометрики. Объединение этих частей дает студентам возможность более глубокого освоения теоретического материала и приобретения практических навыков исполь-

зования эконометрических методов в анализе социально-экономической информации. В практикуме используется двойная нумерация рисунков и формул, первое число указывает номер лабораторной работы, второе – порядковый номер рисунка или формулы в лабораторной работе.

Практикум полезен также и преподавателям при организации и проведении практических занятий по базовому курсу эконометрики с использованием надстройки «Анализ данных» MS Excel.

Лабораторная работа № 1. Описательная статистика

Цель работы. Оценка свойств генеральной совокупности по эмпирическим (наблюдаемым) данным (выборке) путем построения эмпирического распределения, нахождения числовых характеристик выборки, нахождения точечных и интервальных оценок параметров нормального распределения, проверки гипотезы о нормальном распределении генеральной совокупности.

Краткие сведения. *Генеральная совокупность* – все мыслимое множество объектов на которых изучается некоторый признак(и). В математической статистике *генеральная совокупность* – это совокупность всех мыслимых значений изучаемого признака, которые могли быть получены при данном комплексе условий. Генеральная совокупность рассматривается как случайная величина *X* с неизвестным законом распределения.

Выборка (выборочная совокупность) – часть объектов генеральной совокупности, на которой произведены измерения изучаемого признака. Совокупность $x_1, x_2, \cdots, x_n$ измеренных на выборке значений изучаемого признака также называют *выборкой*. Количество *n* произведенных измерений (наблюдений) признака, называется *объемом выборки*.

Сущность выборочного метода состоит в оценке по выборке $x_1, x_2, \cdots, x_n$ свойств генеральной совокупности (свойств распределения случайной величины X).

Выборочные данные, упорядоченные по возрастанию или убыванию, называются *вариационным рядом*. Различные значения исследуемого признака в выборке называются *вариантами*.

Упорядоченная по возрастанию или убыванию последовательность вариант x_i с указанием частот n_i (или относительной частоты $\frac{n_i}{n}$) их повторения в выборке называется *точечным вариационным рядом.* Точечный вариационный ряд представляется таблицей, в первой строке (столбце) которой приводятся упорядоченные по возрастанию варианты x_i, в последующих строках (столбцах) соответствующие им частоты n_i и относительные частоты $\frac{n_i}{n}$. Точечный вариационный ряд представляется в виде следующей таблицы

Таблица 1.1

x_i	x_1	x_2	$\cdots$	x_k
n_i	n_1	n_2	$\cdots$	n_k
$\frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\cdots$	$\frac{n_k}{n}$
m_i	$m_1 = n_1$	$m_2 = n_1 + n_2$	$\cdots$	$m_i = n$

Здесь k – количество различных вариант в выборке, m_i – накопленные частоты вариант x_i.

Точечный вариационный ряд является статистическим аналогом ряда распределения дискретной случайной величины.

При большом количестве вариант или при непрерывном характере исследуемого признака от точечного вариационного ряда переходят к *интервальному вариационному ряду* – выборочным данным сгруппированным по k последовательным интервалам числовой оси. Количество интервалов k определяется как натуральное число $k \approx 1 + log_2 n$. Длины h интервалов группирования находятся как $h = \frac{x_{max} - x_{min}}{k}$, где x_{max} *и* x_{min} наибольшее и наименьшее значения в выборке. Границы C_i интервалов (C_{i-1}, C_i) определяются как

$$C_0 = x_{min}, C_1 = C_0 + h, C_2 = C_1 + h, \dots, C_k = C_{k-1} + h.$$

Интервальный вариационный ряд – это упорядоченная последовательность интервалов (C_{i-1}, C_i) с указанием частоты n_i каждого интервала, равной количеству выборочных данных попавших в рассматриваемый интервал, $n_1 + n_2 + \cdots + n_k = n$. В интервальный вариационный ряд также включают середины $x_i^0 = \frac{C_{i-1} + C_i}{2}$ интервалов, относительные частоты $\frac{n_i}{n}$ интервалов, накопленные частоты интервалов $m_i = n_1 + n_2 + \cdots + n_i$, относительные накопленные частоты интервалов $\frac{m_i}{n}$. Интервальный вариационный ряд представляется следующей таблицей.

Таблица 1.2

Номер интервала	$C_{i-1} - C_i$	x_i^0	n_i	$\frac{n_i}{n}$	m_i	$\frac{m_i}{n}$
1	$C_0 - C_1$	x_1^0	n_1	$\frac{n_1}{n}$	$m_1 = n_1$	$\frac{m_1}{n}$
2	$C_1 - C_2$	x_2^0	n_2	$\frac{n_2}{n}$	m_2=$m_1 + n_2$	$\frac{m_2}{n}$
…	…	…	…	…	…	…
k	$C_{k-1} - C_k$	x_k^0	n_k	$\frac{n_k}{n}$	m_k=n	1

От интервального вариационного ряда можно перейти к точечному вариационному ряду поставив в соответствие серединам интервалов частоты соответствующих интервалов. Столбцы с третьего по шестой интервального вариационного ряда являются точечным вариационным рядом построенном по серединам интервалов, который представляется следующей таблицей.

Таблица 1.3

x_i^0	x_1^0	x_2^0	…	x_k^0
n_i	n_1	n_2	…	n_k
$\frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_k}{n}$
m_i	m_1	m_2	…	m_k

Для графического представления вариационных рядов (эмпирического распределения) используются полигон частот или относительных частот, гистограмма частот или относительных частот, кумулятивная кривая.

Полигон частот строится по точечному вариационному ряду (или точечному вариационному ряду, построенному по серединам интервалов) и представляет собой ломанную, отрезки которой последовательно соединяют точки $(x_1, n_1), (x_2, n_2), \dots, (x_k, n_k)$.

Полигон относительных частот это ломанная, отрезки которой последовательно соединяют точки $(x_1, n_1/n), (x_2, n_2/n), \dots, (x_k, n_k/n)$. Полигон относительных частот является статистической аппроксимацией многоугольника распределения дискретной случайной величины.

Гистограмма частот или относительных частот строится по интервальному вариационному ряду и представляет собой ступенчатую фигуру на плоскости, состоящую из прямоугольников основаниями которых служат интервалы (C_0, C_1), (C_1, C_2), … , (C_{k-1}, C_k), а высоты равны частотам $n_1, n_2, \cdots, n_k$ или относительным частотам $\frac{n_1}{n}, \frac{n_2}{n}, \dots, \frac{n_k}{n}$ этих интервалов.

Кумулятивная кривая (кумулята) – это кривая относительных накопленных частот (или накопленных частот), строится по точеному вариационному ряду (или интервальному вариационному ряду) и представляет собой плавную линию, проходящую через точки $(x_1, m_1), (x_2, m_2), \dots, (x_k, m_k)$ (или точки (C_0, 0), (C_1, m_1/n), (C_2, m_2/n), ... , (C_k, m_k/n)).

Вариационные ряды и их графические изображения представляют эмпирическое (выборочное) распределение исследуемого признака генеральной совокупности. Эмпирическое распределение задается и эмпирической функцией распределения $F^*(x) = \frac{n_x}{n}$, которая для любого числа *x* определяется как отношение количества n_x выборочных данных меньших числа *x* к объему выборки *n*. Эмпирическая функция распределения является статистическим аналогом функции распределения исследуемого признака генеральной совокупности (исследуемой случайной величины *X*). Эмпирическая функция распределения, построенная по точечному вариационному ряду по серединам интервалов, имеет вид

$$F^*(x) = \begin{cases} 0, & x \le x_1^0, \\ n_1/n, & x_1^0 < x \le x_2^0, \\ \dfrac{n_1 + n_2}{n}, & x_2^0 < x \le x_3^0, \\ \dfrac{n_1 + n_2 + n_3}{n}, & x_3^0 < x \le x_4^0 \\ \vdots & \vdots \\ 1, & x_k^0 < x. \end{cases}$$

Числовые характеристики выборки.

Характеристики положения выборки (*средние величины*) определяют положение выборки на числовой оси одним числом, вокруг которого концентрируются выборочные данные. Наиболее распространенными характеристиками положения являются *выборочная средняя* $\bar{x}$, *выборочная средняя геометрическая* $\bar{x}_{геом}$, *мода выборки* $\bar{x}_{mod}$, *медиана выборки* $\bar{x}_{med}$.

Выборочная средняя $\bar{x}$ по исходной выборке определяется как $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, по точечному вариационному ряду как $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$.

Выборочная средняя геометрическая $\bar{x}_{геом}$ по исходной выборке определяется как $\bar{x}_{геом} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n}$, по точечному вариационному ряду как $\bar{x}_{геом} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}$.

Модой выборки $\bar{x}_{mod}$ называется варианта с наибольшей частотой. Она определяется по точечному вариационному ряду. Для выборки, заданной интервальным вариационным рядом, мода выборки $\bar{x}_{mod}$ определяется соотношением

$$\bar{x}_{mod} = \bar{x}_{mod(min)} + h \cdot \frac{n_{mod} - n_{mod-1}}{2 \cdot n_{mod} - n_{mod-1} - n_{mod+1}},$$

где n_{mod} – частота модального интервала, интервала с наибольшей частотой; $\bar{x}_{mod(min)}$ – нижняя граница модального интервала; h – длина модального интервала; n_{mod-1} – частота интервала, предшествующего модальному; *и* n_{mod+1} – частота интервала, следующего за модальным.

Медиана выборки $\bar{x}_{med}$ называется значение признака, приходящее на середину вариационного ряда (выборочных значений, упорядоченных по возрастанию), т.е. медиана выборки делит выборку на две части равные по частоте. Для выборки, заданной интервальным вариационным рядом, медиана выборки $\bar{x}_{med}$ определяется соотношением

$$\bar{x}_{med} = \bar{x}_{med(min)} + h \cdot \frac{n/2 - m_{med-1}}{n_{med}},$$

где n_{med} – частота медианного интервала, первого интервала для которого накопленная частота превышает половину объема выборки (*n/2*); $\bar{x}_{med(min)}$ – нижняя граница медианного интервала; h – длина медианного интервала; m_{med-1} – накопленная частота интервала, предшествующего медианному.

Характеристики вариации (рассеяния) выборки описывают изменчивость значений изучаемого признака. Наиболее распространенными характеристиками вариации являются:

– *вариационный размах* $R = x_{max} - x_{min}$;

– *выборочная дисперсия*, равная среднему арифметическому квадратов отклонений выборочных значений от выборочной средней, определяется по исходной выборке как $D_B = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, а по точечному вариационному ряду как $D_B = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i$;

– *выборочное среднее квадратическое отклонение (стандартное отклонение)* $\sigma_B = \sqrt{D_B}$;

– *исправленная выборочная дисперсия* $S^2 = \frac{n}{n-1} D_B$;

– *исправленное выборочное среднее квадратическое отклонение* $s = \sqrt{S^2}$;

– *коэффициент вариации* $V = \frac{\sigma_B}{\bar{x}} \cdot 100\%$.

Выборочная средняя и выборочная дисперсия обладают такими же свойствами, что и математическое ожидание, и дисперсия случайной величины.

Характеристики формы распределения выборки:

– *выборочный коэффициент асимметрии* $\tilde{A} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^3 \cdot n_i}{n \cdot s^3}$ является мерой отклонения распределения выборки от симметричного, при $\tilde{A} = 0$ распределение выборки (полигон частот, гистограмма частот) симметрично относительно прямой $x = \bar{x}$, при $\tilde{A} > 0$ ($\tilde{A} < 0$) распределение выборки имеет более пологую правую (левую) часть;

– *выборочный эксцесс* $\tilde{E} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 \cdot n_i}{n \cdot s^4} - 3$ является показателем «крутости» распределения выборки по сравнению с нормальным распределением.

Статистическое оценивание параметров распределения генеральной совокупности. Различают:

– *точечные оценки* – оценки в виде одного числа $\tilde{\theta}$, в окрестности которого находится истинное значение θ оцениваемого параметра;

– *интервальные оценки*, представляющие числовой интервал ($\tilde{\theta}_н$, $\tilde{\theta}_в$) с заданной вероятностью γ, близкой к единице, накрывающие истинное значение оцениваемого параметра. Интервал ($\tilde{\theta}_н$, $\tilde{\theta}_в$) называется *доверительным*, а вероятность $\gamma = P(\theta \in (\tilde{\theta}_н, \tilde{\theta}_в))$ называется *доверительной вероятностью* или *надежностью* оценки.

Точечные оценки $\tilde{\theta}$ и границы $\tilde{\theta}_н$ *и* $\tilde{\theta}_в$ доверительного интервала находятся как функции выборки $x_1, x_2, \cdots, x_n$ и являются случайными величинами.

Свойства точечных оценок.

Оценка $\tilde{\theta}$ *параметра* θ *называется несмещенной, если ее математическое ожидание равно оцениваемому параметру* θ*, т.е.* $M\tilde{\theta} = \theta$*. В противном случае оценка называется смещенной.* Несмещенность оценки гарантирует отсутствие систематических ошибок при оценивании.

Оценка $\tilde{\theta}$ *параметра* θ *называется состоятельной, если она сходится по вероятности к оцениваемому параметру.* Состоятельность оценки обеспечивает при достаточно больших объемах выборки, с вероятностью близкой к единице, приближенное равенство $\theta \approx \tilde{\theta}$.

Несмещенная оценка $\tilde{\theta}$ *параметра* θ *называется эффективной, если она имеет наименьшую дисперсию среди всех несмещенных оценок параметра* θ*, построенных по выборкам одного и того же объема n.*

Выборочная средняя $\bar{x}$ является несмещенной, состоятельной оценкой математического ожидания генеральной совокупности.

Выборочная дисперсия D_B является смещенной и состоятельной оценкой дисперсии генеральной совокупности. Исправленная выборочная дисперсия S^2 является несмещенной и состоятельной оценкой дисперсии генеральной совокупности. Выборочная дисперсия D_B и исправленная выборочная дисперсия S^2 являются эффективными оценками дисперсии генеральной совокупности лишь асимптотически (т.е. только при неограниченном возрастании объема выборки).

Интервальные оценки заданной надежности γ *математического ожидания* (a) *и дисперсии* (σ^2) *генеральной совокупности, распределенной по нормальному закону,* определяются неравенствами

$$\bar{x} - t_\gamma \cdot \frac{s}{\sqrt{n}} < a < \bar{x} + t_\gamma \cdot \frac{s}{\sqrt{n}};$$

$$\frac{S^2 \cdot (n-1)}{U_2} < \sigma^2 < \frac{S^2 \cdot (n-1)}{U_1}.$$

Здесь $t_\gamma = t(\frac{1+\gamma}{2}, n-1)$ – квантиль уровня $\frac{1+\gamma}{2}$ распределения Стьюдента с числом степеней свободы равным *n-1*; $U_1 = \chi^2(\frac{1-\gamma}{2}, n-1)$ и $U_2 = \chi^2(\frac{1+\gamma}{2}, n-1)$ квантили соответственно уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$ распределения χ^2 (распределения хи-квадрат) с числом степеней свободы равным *n-1*.

Проверка гипотезы о нормальном распределении генеральной совокупности на уровне значимости α=0,05 по критерию согласия χ^2- Пирсона.

Проверяется нулевая гипотеза H_0: $X \sim N(a, \sigma)$ о нормальном распределении генеральной совокупности (случайно величины *X*). Параметры ***a*** (математическое ожидание) и ***σ*** (среднее квадратическое отклонение) неизвестны. За их значения принимаются их несмещенные и состоятельные оценки: выборочная средняя $\bar{x}$ и исправленное выборочное среднее квадратическое отклонение $s = \sqrt{S^2}$. Таким образом, проверяется нулевая гипотеза H_0: $X \sim N(\bar{x}, s)$. Для проверки этой гипотезы используется статистика критерия

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i},$$

которая является мерой расхождения теоретического распределения $N(\bar{x}, s)$ и эмпирического распределения представленного точечным или интервальным вариационным рядом. В этом критерии k – количество интервалов в интервальном вариационном ряду; n_i – частоты интервалов (C_{i-1}, C_i) (наблюдаемые частоты); n – объем выборки; p_i – теоретические вероятности попадания случайной величины *X* в интервал (C_{i-1}, C_i), подсчитанные по предполагаемому нормальному распределению; np_i – теоретические частоты интервалов. Для эмпирического распределения, представленного интервальным вариационным рядом, теоретические вероятности вычисляются как

$$p_i = P(C_{i-1} \le X \le C_i) = F(C_i, \bar{x}, s) - F(C_{i-1}, \bar{x}, s)$$

или

$$p_i = P(C_{i-1} \le X \le C_i) = \Phi\left(\frac{C_i - \bar{x}}{s}\right) - \Phi\left(\frac{C_{i-1} - \bar{x}}{s}\right),$$

где $F(x, \bar{x}, s)$ функция распределения нормально распределенной случайной величины с математическим ожиданием равным $\bar{x}$ и средне квадратическим отклонением равным *s*, $\Phi(x)$ – функция Лапласа.

Статистика $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$ при $n \to \infty$ имеет распределение χ^2 с числом степеней свободы равным *k-r-1*, где r – число параметров теоретического распределения (для нормального распределения *r=2*).

Если вычисленное по выборке значение критерия χ^2 больше критического значения $\chi^2_{кр}(1-\alpha, k-r-1)$, то нулевая гипотеза H_0: $X \sim N(\bar{x}, s)$ отвергается (гипотеза противоречит выборочным данным). Если вычисленное значение критерия $\chi^2 \le \chi^2_{кр}(1-\alpha, k-r-1)$, то нулевая гипотеза H_0: $X \sim N(\bar{x}, s)$ принимает-

ся на уровне значимости α (гипотеза о нормальном распределении генеральной совокупности с параметрами $a = \bar{x}\, u\, \sigma = s$ согласуется с выборочными данными). Критическое значение $\chi^2_{кр}(1-\alpha, k-r-1)$ это квантиль уровня 1-α распределения χ^2 с числом степеней свободы равным *k-r-1*.

Содержание лабораторной работы (расчетно-графической работы).

По заданной выборке значений изучаемого признака генеральной совокупности:

1. Построить вариационный и точечный вариационный ряды.
2. От точечного вариационного ряда перейти к интервальному вариационному ряду.
3. Построить точечный вариационный ряд по серединам интервалов.
4. По точечному вариационному ряду по серединам интервалов построить полигон частот, по интервальному вариационному ряду построить кумулятивную кривую и гистограмму частот.
5. По точечному вариационному ряду по серединам интервалов построить эмпирическую функцию распределения и ее график.
6. По выборке найти выборочные среднюю, дисперсию, среднее квадратическое отклонение, исправленную выборочную дисперсию, моду, медиану, коэффициент асимметрии и эксцесс.
7. Найти моду и медиану выборки по интервальному вариационному ряду.
8. Привести несмещенные точечные оценки математического ожидания и дисперсии генеральной совокупности.
9. В предположении, что генеральная совокупность имеет нормальное распределение, построить доверительные интервалы надежности γ=0,95 для неизвестных математического ожидания и дисперсии этого распределения.
10. На уровне значимости α=0,05 проверить гипотезу о нормальном распределении генеральной совокупности.
11. Написать заключение о проведенном исследовании свойств генеральной совокупности.

Выполнение работы в MS Excel.

Выполнение работы с использованием MS Excel рассмотрим на примере изучения расхода бензина (литров) на 100 км. в городском цикле для автомобилей одной модели по результатам ста измерений, представленным ниже.

13,28	14,83	16,13	10,01	11,63	9,10	14,77	20,85	9,80	16,47
11,92	10,87	13,22	10,70	9,45	14,92	15,36	13,74	18,05	11,26
8,53	16,44	12,52	11,06	9,95	12,59	8,93	5,00	8,23	5,41
4,23	7,47	13,63	8,08	12,48	7,03	10,34	14,95	13,56	9,01
9,52	12,47	10,73	14,41	13,70	9,63	12,85	12,63	10,31	8,59
18,84	10,85	15,02	11,61	18,07	11,61	14,96	9,29	6,59	11,72
10,05	13,55	7,23	15,65	11,88	13,15	16,45	14,52	15,31	15,54
14,19	11,07	17,25	16,12	9,88	12,43	14,60	11,32	14,56	11,45
9,76	8,22	15,72	11,98	8,75	9,05	14,38	12,59	9,89	13,28
12,52	10,58	14,95	8,67	11,92	10,28	10,51	11,48	7,67	18,33

Ввод данных и построение вариационных рядов.

Открыв Microsoft Excel введем выборочные данные в ячейки **A2 – A101**, в ячейке **A1** поместим заголовок «Выборка». Для построения вариационного ряда скопируем выборку (ячейки **A2 – A101**) в ячейки **C2 – C101**, в ячейке **C1** разместим заголовок «Вариационный ряд». Выделим ячейки **C2 – C101**. В вкладке «***Редактирование***» выберем группу «***Сортировка и фильтр***» и в ней «***Сортировка по возрастанию***», в ячейках **C2 – C101** получим вариационный ряд, который используем для построения точечного вариационного ряда. Точечный вариационный ряд разместим в столбцах **E** и **F**. В ячейке **E1** разместим заголовок «Точечный вариационный ряд», в ячейке **E2** заголовок «x_i», в ячейке **F2** – «n_i». В столбце «x_i» разместим в порядке возрастания различные выборочные значения (варианты), а в столбце «n_i» частоты их повторения в выборке, см. рис. 1.1. В данной выборке большое число различных вариант. Поэтому необходимо перейти к интервальному вариационному ряду.

	A	B	C	D	E	F	G
1	Выборка		Вариационный ряд		Точечный вариационный ряд		
2	13,28		4,23		xi	ni	
3	11,92		5,00		4,23	1	
4	8,53		5,41		5,00	1	
5	4,23		6,59		5,41	1	
6	9,52		7,03		6,59	1	
7	18,84		7,23		7,03	1	
8	10,05		7,47		7,23	1	
9	14,19		7,67		7,47	1	
10	9,76		8,08		7,67	1	
11	12,52		8,22		8,08	1	
12	14,83		8,23		8,22	1	
13	10,87		8,53		8,23	1	
14	16,44		8,59		8,53	1	
15	7,47		8,67		8,59	1	
16	12,47		8,75		8,67	1	

Рис. 1.1. Вариационные ряды

Построение интервального вариационного ряда и точечного вариационного ряда по серединам интервалов.

Число интервалов $k \approx 1 + log_2 n = 1 + log_2 100 \approx 7$, длины h интервалов группирования $h = \frac{x_{max} - x_{min}}{k} = \frac{20{,}85 - 4{,}23}{7} = 2{,}37375$. Для простоты расчетов границ интервалов примем $h=2{,}4$. За начало C_0 первого интервала примем $x_{min}=4{,}23$. Последующие границы интервалов найдем как $C_i=C_{i-1}+h$, $i=1, \ldots, 8$.

Интервальный вариационный ряд расположим в ячейках H10 – O18, см. рис.1.2. Частоты n_i интервалов определим по точечному вариационному ряду, середины интервалов вычислим как $x_i^0 = \frac{C_{i-1} + C_i}{2}, i = 1, \ldots, 8.$

G	H	I	J	K	L	M	N	O

Число интервалов k=7

Длина интервала $h=(x_{max}-x_{min})/k=$ 2,37375

Округленное значение h равно 2,4

Интервальный вариационный ряд

№ интервала	C_{i-1}	C_i	x_{i0}	n_i	n_i/n	m_i	m_i/n
1	4,23	6,63	5,43	4	0,04	4	0,04
2	6,63	9,03	7,834953	13	0,13	17	0,17
3	9,03	11,43	10,23495	26	0,26	43	0,43
4	11,43	13,83	12,63495	28	0,28	71	0,71
5	13,83	16,23	15,03495	20	0,2	91	0,91
6	16,23	18,63	17,43495	7	0,07	98	0,98
7	18,63	21,03	19,83495	2	0,02	100	1
Суммы				100	1		

Рис. 1.2. Интервальный вариационный ряд

В интервальном вариационном ряду столбцы с четвертого по седьмой образуют точечный вариационный ряд, построенный по серединам интервалов. В последней строке интервального вариационного ряда вычислены контрольные суммы: сумма частот интервалов, которая должна равняться объему выборки, и сумма относительных частот интервалов, которая должна равняться единице. Для вычисления этих сумм выделяются ячейки, содержащие суммируемые величины, во вкладке **«Формулы»** выбирается функция **«Автосумма»**, в результате выполнения которой в ячейке, следующей за выделенными, получается значение искомой суммы.

Построение полигона частот, кумулятивной кривой и гистограммы частот.

Построение полигона частот. Выделим свободную ячейку, откроем вкладку **«Вставка»** и выберем **«Вставить точечную (X, Y) или пузырьковую диаграмму»**, среди предлагаемых диаграмм выберем **«Точечная с прямыми отрезками и маркерами»**. При необходимости перемести поле построения диаграммы на свободное место. В открытом окне вкладки **«Конструктор»** выберем **«Выбрать данные»**, в ее открывшемся окне в поле «Диапазон данных для диаграммы» укажем диапазон ячеек содержащих середины интервалов и соответствующие им частоты (в примере это ячейки K11 – L18), в части «Элементы легенды» выберем «Изменить» и поле «Имя ряда» введем «Полигон частот», затем по «ОК» получим полигон частот, см. рис. 1.3.

Рис. 1.3. Полигон частот

Построение кумулятивной кривой (кумуляты). Предварительно построим вспомогательную таблицу, первый столбец которой содержит границы интервалов интервального вариационного ряда, второй – относительные накопленные частоты интервалов, поставленные в соответствие концам интервалов, для начала первого интервала относительная накопленная частота равна нулю, см. рис. 1.4. В рассматриваемом примере эта таблица расположена в ячейках в ячейках H41 – I49. Выделим свободную ячейку, откроем вкладку **«Вставка»** и выберем **«Вставить точечную (X, Y) или пузырьковую диаграмму»**, среди предлагаемых диаграмм выберем **«Точечная с прямыми отрезками и маркерами»**. При необходимости переместим поле построения диаграммы на свободное место. В открытом окне вкладки **«Конструктор»** выберем **«Выбрать данные»**, в ее открывшемся окне в поле «Диапазон данных для диаграммы» укажем диапазон ячеек содержащих границы интервалов и соответствующие им относительные накопленные частоты (в примере это ячейки H41 – I49), в части «Элементы легенды» выберем «Изменить» и поле «Имя ряда» введем «Кумулята», затем по «ОК» получим кумулятивную кривую, см. рис. 1.4.

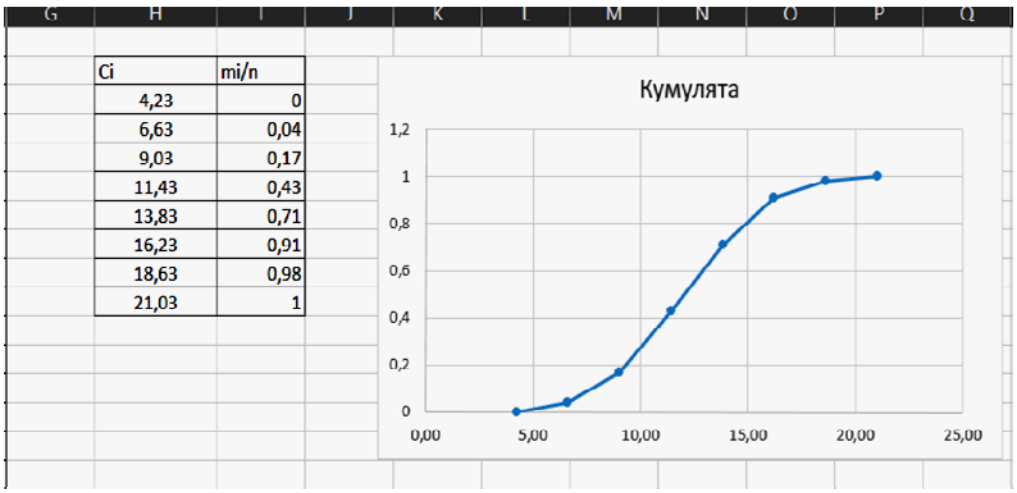


Рис. 1.4. Кумулята

Построение гистограммы частот. Выделим исходную выборку (в рассматриваемом примере ячейки A2-A101). В окне вкладки **«Вставка»** выберем **«Вставка статистической диаграммы»** и в ее окне выберем «Гистограмма». Полученную гистограмму переместим на свободное место на листе. Для рассматриваемого примера построенная гистограмма приведена на рис. 1.5. По оси абсцисс приведены границы интервалов, по оси ординат частоты интервалов.

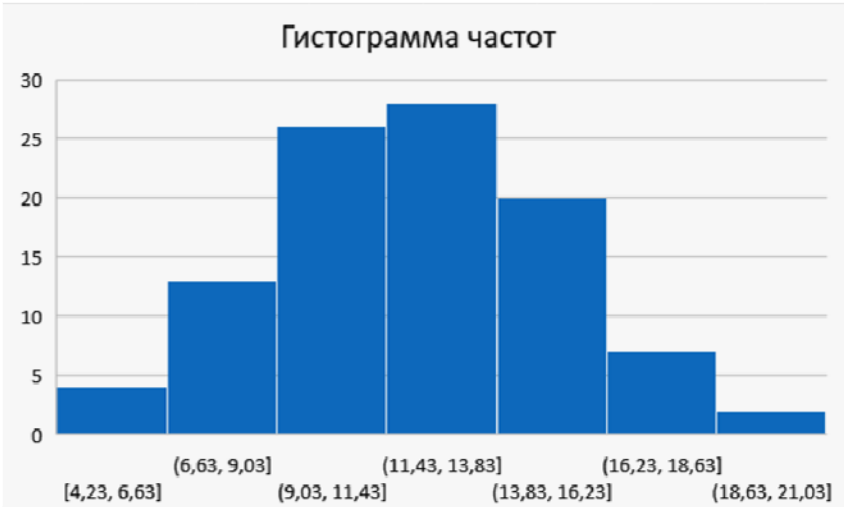


Рис. 1.5. Гистограмма частот

Построение эмпирической функции распределения и ее графика.

Эмпирическую функцию распределения $F^*(x) = \frac{n_x}{n}$, которая для любого числа x определяется как отношение количества n_x выборочных данных меньших числа x к объему выборки n, построим по точечному вариационному ряду по серединам интервалов, используя столбец x_{i_0} – середин интервалов и столбец $\frac{m_i}{n}$ – относительных накопленных частот интервального вариационного ряда, см. рис. 1.2.

$$F^*(x) = \begin{cases} 0, & x \le 5{,}43, \\ 0{,}04, & 5{,}43 < x \le 7{,}835, \\ 0{,}17, & 7{,}835 < x \le 10{,}235, \\ 0{,}43, & 10{,}235 < x \le 12{,}635, \\ 0{,}71, & 12{,}635 < x \le 15{,}035, \\ 0{,}91, & 15{,}035 < x \le 17{,}435, \\ 0{,}98, & 17{,}435 < x \le 19{,}835, \\ 1, & 19{,}835 < x. \end{cases}$$

Нахождение числовых характеристик выборки.

Числовые характеристики выборки (выборочная средняя, выборочная дисперсия, среднее квадратическое отклонение, исправленная выборочная дисперсия, мода выборки, медиана выборки, коэффициент асимметрии и эксцесс) могут быть найдены каждая отдельно или все вместе. Для отдельного нахождения числовых характеристик выборки в окне вкладки **«Формулы»** выбирается **«Другие функции»** и в ее окне выбирается группа **«Статистические»**, затем

выбираются необходимые функции (например, СРЗНАЧ, ДИСП.В, МЕДИАНА, МОДА.ОДН., СКОС и т.д.).

Для одновременного нахождения всех числовых характеристик выборки нужно открыть вкладку **«Данные»**, в ее окне выбрать **«Анализ данных»** и в окне «Инструменты анализа» выбрать **«Описательная статистика»**. Заполнение окна для рассматриваемого примера приведено на рис 1.6.

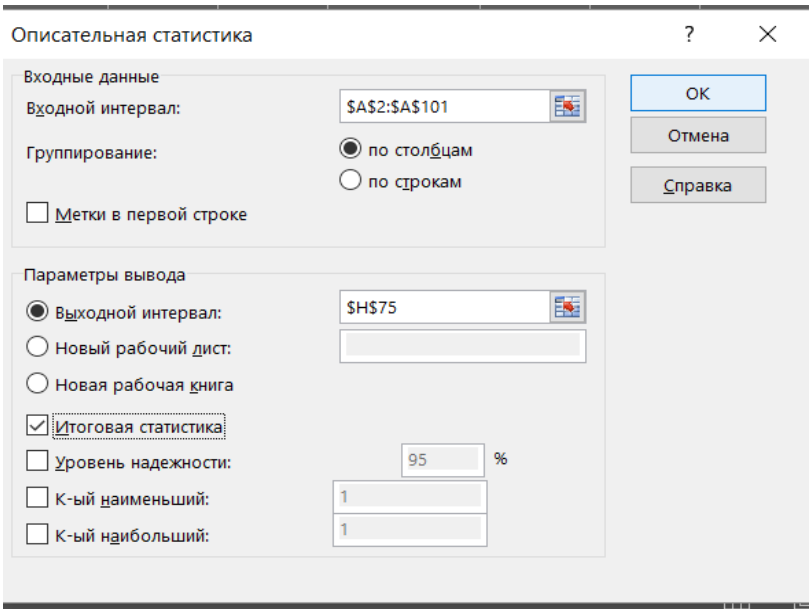


Рис. 1.6. Заполнение окна «Описательная статистика»

В поле *Входной интервал* указываются ячейки в которых располагается выборка, в поле *Группирование* выбирается расположение выборки по столбцу или по строке, в части *Параметры вывода* указывается место вывода результатов и выбираются выводимые результаты. Для рассматриваемого примера результаты выполнения этой операции приведены на рис. 1.7.

	G	H	I	J
74				
75		*Столбец1*		
76				
77		Среднее	12,05899	
78		Стандартная ошибка	0,319654	
79		Медиана	11,90072	
80		Мода	#Н/Д	
81		Стандартное отклонение	3,196539	
82		Дисперсия выборки	10,21786	
83		Эксцесс	-0,10288	
84		Асимметричность	0,10946	
85		Интервал	16,61625	
86		Минимум	4,234953	
87		Максимум	20,8512	
88		Сумма	1205,899	
89		Счет	100	

Рис. 1.7. Числовые характеристики выборки

В этой таблице:

− *среднее* – выборочная средняя;

− *стандартная ошибка* – среднее квадратическое отклонение выборочной средней;

− *медиана* – медиана выборки;

− *мода* – мода выборки;

− *стандартное отклонение* – исправленное выборочное среднее квадратическое отклонение;

− *дисперсия выборки* – исправленная выборочная дисперсия;

− *эксцесс* – эксцесс выборки;

− *асимметричность* – коэффициент асимметрии;

− *интервал* – вариационный размах (разность между наибольшим и наименьшим значением в выборке);

− *минимум* – наименьшее значение в выборке;

− *максимум* – наибольшее значение в выборке;

− *сумма* – сумма всех выборочных данных;

− *счет* – объем выборки.

Нахождение медианы выборки по вариационному и интервальному вариационному ряду. Выборка содержит четное число наблюдений (n = 100), поэтому по вариационному ряду выборочную медиану находим как полусумму пятидесятого и пятьдесят первого значения в вариационном ряду. Для этого, выделив, например, ячейку **K106** и учитывая расположение выборки, в строке формул введем **=(A51+A52)/2**, по **Enter** в этой ячейке получим искомое значение медианы выборки, см. рис. 1.8. Для нахождения медианы выборки по интервальному вариационному ряду по формуле

$$\bar{x}_{med} = \bar{x}_{med(min)} + h \cdot \frac{n/2 - m_{med-1}}{n_{med}},$$

определим медианный интервал (для которого накопленная частота впервые превышает половину объема выборки), в примере это четвертый интервал, см. рис.1.2. Затем выделив, например, ячейку **L107**, и учитывая расположение $\bar{x}_{med(min)}$, m_{med-1}, n_{med} на листе Excel, см. рис.1.2, и значения объема выборки *n* и длины интервала *h=2,4*, в строке формул введем **=I14+2,4*(100/2-N13)/L14**. По Enter в ячейке **L107** получим значение медианы выборки, см. рис. 1.8.

Нахождение моды выборки по точечному вариационному и интервальному вариационному ряду. По точечному вариационному ряду по серединам интервалов, см. рис.1.2, мода выборки равна 12,635. Для нахождения моды выборки по интервальному вариационному ряду по формуле

$$\bar{x}_{mod} = \bar{x}_{mod(min)} + h \cdot \frac{n_{mod} - n_{mod-1}}{2 \cdot n_{mod} - n_{mod-1} - n_{mod+1}}$$

определим модальный интервал (интервал с наибольшей частотой), в примере это четвертый интервал. Выделим, например, ячейку **L105** и учитывая расположение $\bar{x}_{mod(min)}$, n_{mod}, n_{mod-1}, n_{mod+1} на листе Excel, см. рис. 1.2, в

строке формул введем **=I14+2,4*(L14-L13)/(2*L14-L13-L15)**. По Enter в ячейке **L105** получим значение моды выборки, см. рис. 1.8.

	E	F	G	H	I	J	K	L	M
103									
104	Мода выборки по точечному вариационному ряду по серединам интервалов: x_{mod}=11,58						x_{mod}=	12,635	
105	Мода выборки по интервальному вариационному ряду:						x_{mod}=	11,91495	
106	Медиана выборки по вариационному ряду:					x_{med}=	10,50852		
107	Медиана выборки по интервальному вариационному ряду:						x_{med}=	12,03495	
108									
109									
110	Несмещенная оценка математического ожидания нормальной случайной величины=12,05899.								
111	Несмещенная оценка дисперсии нормальной случайной величины=10,21786.								

Рис. 1.8. Выборочные медиана и мода

Несмещенные точечные оценки математического ожидания и дисперсии генеральной совокупности соответственно равны выборочной средней и исправленной выборочной дисперсии, значения которых берутся из таблицы числовых характеристик выборки, рис. 1.7, и также приведены на рис. 1.8.

Доверительные интервалы надежности γ=0,95 для неизвестных математического ожидания a и дисперсии σ^2 нормально распределенной генеральной совокупности определяются неравенствами:

$$\bar{x} - t_\gamma \cdot \frac{s}{\sqrt{n}} < a < \bar{x} + t_\gamma \cdot \frac{s}{\sqrt{n}};$$

$$\frac{S^2 \cdot (n-1)}{U_2} < \sigma^2 < \frac{S^2 \cdot (n-1)}{U_1}.$$

В ячейках **B120**, **B121**, **B122** вычислим соответственно значения квантилей $t_\gamma = t(\frac{1+\gamma}{2}, n-1)$, $U_1 = \chi^2(\frac{1-\gamma}{2}, n-1)$ и $U_2 = \chi^2(\frac{1+\gamma}{2}, n-1)$. Выделив ячейку **B120** в строке формул введем **=СТЬЮДЕНТ.ОБР.2Х(0,05;99)**, по Enter в этой ячейке получим значение квантили уровня $\frac{1+\gamma}{2}$ распределения Стьюдента с числом степеней свободы равным *n-1=99*. Выделив **B121** в строке формул введем **=ХИ2.ОБР(0,025;99)**, по Enter в этой ячейке получим значение квантили уровня $\frac{1-\gamma}{2}$ распределения χ^2 (распределения хи-квадрат) с числом степеней свободы равным *n-1=99*. Выделив **B122** в строке формул введем **=ХИ2.ОБР(0,975;99)**, по Enter в этой ячейке получим значение квантили уровня $\frac{1+\gamma}{2}$ распределения χ^2 (распределения хи-квадрат) с числом степеней свободы равным *n-1=99*. См. рис 1.9.

Для значений границ доверительного интервала математического ожидания используем, например, ячейки **D120** и **F120**. См. рис.1.9. Выбрав ячейку **D120** и учитывая расположение на листе выборочного среднего, исправленного среднего квадратического отклонения, см. рис. 1.7, и t_γ в строке формул введем

=I77-B120*I81/КОРЕНЬ(100)

По Enter в этой ячейке получим нижнюю границу интервальной оценки математического ожидания. Аналогичным образом в ячейке **F120**, введя в строке формул **=I77+B120*I81/КОРЕНЬ(100)**, получим верхнюю границу интервальной оценки математического ожидания генеральной совокупности.

Для значений границ интервальной оценки дисперсии генеральной совокупности выберем, например, ячейки **D122** и **F122**. Выделяя поочередно эти ячейки и вводя в строке формул соответственно **=I82*(100-1)/B122** и **=I82*(100-1)/B121** получим границы интервальной оценки дисперсии генеральной совокупности надежности 0,95, см рис. 1.9.

	A	B	C	D	E	F	G
118							
119				Доверительные интервалы			
120	tγ=	1,984217		11,42473	< a <	12,69326	
121	U1=	73,36108					
122	U2=	128,422		7,87691	< σ2 <	13,7889	
123							
124							

Рис. 1.9. Интервальные оценки

Проверка гипотезы о нормальном распределении генеральной совокупности.

Проверка нулевой гипотезы H_0: $X \sim N(a, \sigma)$ о нормальном распределении исследуемого признака с $a = \bar{x}$ *и* $\sigma = s$ проводится с использованием критерия согласия Пирсона χ^2. Для вычисления значения статистики критерия $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$ используем интервальный вариационный ряд. Для этого к полученному ранее интервальному вариационному ряду добавим справа три столбца: столбец теоретических вероятностей p_i; столбец теоретических частот $n \cdot p_i$; столбец значений слагаемых $\frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$, см. рис 1.10. Расположение значений выборочной средней $\bar{x}$ и выборочного среднего квадратического отклонения s приведено в таблице числовых характеристик, рис.1.7, а расположение границ интервалов на рис. 1.10. Теоретические вероятности p_i интервалов (C_{i-1}, C_i) вычислим как разность значений функции распределения нормального распределения, т.е. как

$$p_i = P(C_{i-1} \le X \le C_i) = F(C_i, \bar{x}, s) - F(C_{i-1}, \bar{x}, s).$$

Для заданного x значение $F(x,\bar{x},s)$ функции распределения нормально распределенной случайной величины, с математическим ожиданием равным $\bar{x}$ и средне квадратическим отклонением равным s, вычисляет функция **НОРМ.РАСПР** категории **«Статистические»**. Для нахождения p_1 выделим ячейку **P11** и в строке формул введем

=НОРМ.РАСП(J11;$B$105;$B$109;1)-НОРМ.РАСП(I11;$B$105;$B$109;1)

По Enter в ячейке **P11** получим значение вероятности p_1. Аналогично вычисляются теоретические вероятности других интервалов. В примере объем выборки n равен 100 и для получения значений теоретических частот $n{\cdot}p_i$ в соответствующих ячейках вычислим значения $n{\cdot}p_i$, см. рис. 1.10.

Вычисление значений слагаемых $\frac{(n_i-n{\cdot}p_i)^2}{n{\cdot}p_i}$, для рассматриваемого примера. Выделим ячейку **R11** и учитывая расположение n_1 и $n{\cdot}p_1$ в строке формул введем **=(L11-Q11)^2/Q11**. По Enter получим значение этого слагаемого. Аналогично вычисляются значения этих слагаемых для других интервалов. На рис.1.10 в последней строке приведены сумма теоретических вероятностей (в примере ячейка P18), сумма теоретических частот (в примере ячейка Q18) и вычисленное значение статистики $\chi^2=\sum_{i=1}^{k}\frac{(n_i-n{\cdot}p_i)^2}{n{\cdot}p_i}$ (в примере ячейка R18). Для нахождения этих сумм нужно, например, выделить ячейки с соответствующими слагаемыми, в вкладке **«Главная»** выбрать **«Редактирование»** и в ее окне функцию «∑ **Автосумма»**. В ячейке под выделенными ячейками получим значение суммы выделенных слагаемых. В примере вычисленное значение статистики $\chi^2=\sum_{i=1}^{k}\frac{(n_i-n{\cdot}p_i)^2}{n{\cdot}p_i}$, равное 0,18709, располагается в ячейке **R18**.

H	I	J	K	L	M	N	O	P	Q	R	S
		Интервальный вариационный ряд						Расчет критерия Пирсона хи-квадрат			
№ интервала	Ci-1	Ci	Xi0	ni	ni/n	mi	mi/n	pi	n·pi	(ni-n·pi)^2/(n·pi)	
1	4,23	6,63	5,43	4	0,04	4	0,04	0,0377	3,767337	0,014368846	
2	6,63	9,03	7,834953	13	0,13	17	0,17	0,1272	12,72025	0,006152523	
3	9,03	11,43	10,23495	26	0,26	43	0,43	0,2505	25,05437	0,035690762	
4	11,43	13,83	12,63495	28	0,28	71	0,71	0,2881	28,81449	0,023023021	
5	13,83	16,23	15,03495	20	0,2	91	0,91	0,1935	19,35379	0,021576375	
6	16,23	18,63	17,43495	7	0,07	98	0,98	0,0759	7,587532	0,04549486	
7	18,63	21,03	19,83495	2	0,02	100	1	0,0173	1,734058	0,040785971	
Суммы				100	1			0,9903	99,03183	0,187092359	
							Хи-квадрат критическое =			9,487729037	

Рис. 1.10. Критерий Пирсона χ^2

Критическое значение $\chi^2_{кр}(1-\alpha, k-r-1)$ для $\alpha = 0{,}05, k = 7, r = 2$ найдем, используя функцию **ХИ2.ОБР** категории **«Статистические»**. Выделим ячейку **R19** и в строке формул введем **=ХИ2.ОБР(0,95;4)**. По Enter в этой ячейке получим критическое значение критерия, в примере оно равно 9,4877.

В примере вычисленное значение статистики $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$, равное 0,18709, меньше критического, равного 9,4877, следовательно, на уровне значимости 0,05 нулевая гипотеза H_0: $X \sim N(\bar{x}, s)$ о нормальном распределении генеральной совокупности с оценками параметров $\hat{a} = \bar{x} = 12{,}059$ и $\hat{\sigma} = s = 3{,}197$, принимается.

Заключение о проведенном исследовании свойств генеральной совокупности.

Проведенное исследование выборки расхода бензина (литров) на 100 км. в городском цикле для автомобилей одной модели показало:

- полигон частот и гистограмма частот близки по форме к кривой нормального распределения, см. рис. 1.3 и рис. 1.5, а кумулятивная кривая к графику функции распределения нормальной случайной величины, см. рис. 1.4;
- выборочная средняя $\bar{x} = 12{,}059$, выборочная медиана $\bar{x}_{med} = 12{,}035$ и выборочная мода $\bar{x}_{mod} = 11{,}915$ имеют приближенно равные значения, следовательно, эмпирическое распределение симметрично относительно выборочного среднего, что свидетельствует в пользу нормального распределения расхода бензина (литров) на 100 км. в городском цикле;
- выборочные коэффициент асимметрии $\tilde{A} = 0{,}1095$ и эксцесс $\tilde{E}$ имеют значения близкие к нулю, что также говорят в пользу нормального распределения исследуемого признака;
- по критерию Пирсона χ^2 статистическая гипотеза о нормальном распределении генеральной совокупности, с математическим ожиданием равным 12,059 и средне квадратическим отклонение равным 3,197, согласуется с выборочными данными на уровне значимости 0,05, так как вычисленное значение статистики χ^2, равное 0,18709, меньше критического, равного 9,4877.

Таким образом, расход бензина (литров) на 100 км. в городском цикле для автомобилей одной модели имеет нормальное распределение с математическим ожиданием a=12,059 и средне квадратическим отклонением σ=3,197. Средний расход бензина на 100 км. в городском цикле составляет 12,059 литров и с вероятностью 0,95 лежит в интервале от 11,425 до 12,693 литров.

Контрольные вопросы

1. Сформулируйте понятия генеральной совокупности и выборки.
2. В чем заключается суть выборочного метода?
3. Какая выборка называется репрезентативной (представительной)?
4. Что представляет собой точечный вариационный ряд?
5. Как строится интервальный вариационный ряд?
6. Что понимается под эмпирическим распределением и с помощью чего оно может быть представлено?

7. Как определяется эмпирическая функция распределения?

8. Статистическим аналогом чего является полигон относительных частот?

9. Как строится гистограмма частот?

10. Приведите числовые характеристики положения выборки.

11. Сформулируйте понятия выборочной моды и медианы.

12. Как находится выборочная средняя по точечному вариационному ряду?

13. Как определяются выборочные мода и медиана по интервальному вариационному ряду?

14. Приведите числовые характеристики вариации выборки.

15. Как находятся выборочные дисперсия и среднее квадратическое отклонение?

16. Какие бывают оценки параметров распределения генеральной совокупности?

17. В чем заключаются свойства несмещенности, состоятельности и эффективности точечных оценок?

18. Приведите несмещенные и состоятельные оценки математического ожидания и дисперсии генеральной совокупности.

19. Сформулируйте понятие доверительного интервала и доверительной вероятности.

20. Приведите доверительные интервалы заданной надежности для математического ожидания и дисперсии нормально распределенной генеральной совокупности.

21. Как изменяется доверительный интервал с увеличением доверительной вероятности?

22. Какие могут быть ошибки при проверке статистических гипотез?

23. В чем заключается ошибка первого рода?

24. В чем заключается ошибка второго рода?

25. Вероятностью чего является уровень значимости (уровень α)?

26. Приведите общую схему проверки гипотезы о виде закона распределения генеральной совокупности.

27. Как определяется мера расхождения теоретического (предполагаемого) и эмпирического распределения в критерии согласия χ^2- Пирсона?

28. Для проверки каких гипотез применяются критерии согласия?

Лабораторная работа № 2. Корреляционный анализ

Цель работы. Овладение методами исследования корреляционной зависимости между несколькими количественными случайными величинами по выборочным данным в MS Excel 2010.

Краткие сведения. *Корреляционной зависимостью* двух случайных величин Y и X называется функциональная зависимость условного математического ожидания M_xY (или M_yX) одной из них от значения x (или y) другой величины. Корреляционная зависимость может быть представлена в виде *уравнения регрессии* Y по X: $M_xY = \varphi(x)$ или *уравнения регрессии* X по Y: $M_yX = \psi(y)$. Если функция $\varphi(x)$ линейная, то корреляционная зависимость называется линейной, если $\varphi(x)$ нелинейная, то корреляционная зависимость называется нелинейной.

Основная задача корреляционного анализа – выявление наличия, вида и тесноты корреляционной зависимости между случайными величинами путем точечного и интервального оценивания различных (парных, множественных, частных) коэффициентов корреляции по выборке (x_1, y_1), (x_2, y_2), …, (x_n, y_n) из генеральной совокупности, распределенной по многомерному нормальному закону.

Парная корреляция. Для оценки по выборке наличия и тесноты корреляционной связи между двумя случайными величинами Y и X, имеющими совместное нормальное распределение, используются выборочный коэффициент корреляции и эмпирическое корреляционное отношение.

Выборочный коэффициент корреляции r_{yx} используется для оценки наличия и тесноты парной линейной корреляционной зависимости

$$M_xY = a_y + \rho\frac{\sigma_y}{\sigma_x}(x - a_x),$$

где $a_y = MY$, $a_x = MX$, $\rho -$ генеральный (теоретический) коэффициент корреляции, σ_y и $\sigma_x -$ среднеквадратические отклонения величин Y и X. Выборочный коэффициент корреляции r_{yx} является точечной оценкой генерального коэффициента корреляции ρ и вычисляется по формуле

$$r_{yx} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{s_x \cdot s_y},$$

где $\overline{x}$ и $\overline{y}$ – выборочные общие средние X и Y, $\overline{x \cdot y}$ – выборочное общее среднее произведения XY, s_x и s_y – выборочные среднеквадратические отклонения величин X и Y. Выборочный коэффициент корреляции r_{yx} показывает, на сколько величин s_y изменится в среднем зависимая величина Y при увеличении аргумента X на одно s_x и является показателем тесноты парной линейной корреляционной зависимости.

Свойства выборочного коэффициента корреляции:

- $-1 \le r_{yx} \le 1$, чем больше $|r_{yx}|$, тем теснее линейная корреляционная зависимость Y и X, т.е. меньше разброс выборочных значений (x_i, y_i) относительно оцененной по выборке линии регрессии

$$y_x = \overline{y} + r_{yx}\frac{s_y}{s_x}(x - \overline{x}),$$

которая описывает зависимость условной (групповой) средней y_x величины Y от значений *x* величины X;

- $r_{yx} = r_{xy}$;
- если все выборочные значения умножить на одно и то же число, то величина коэффициента корреляции не изменяется;
- при $r_{yx} = \pm 1$ корреляционная зависимость представляет линейную функциональную зависимость между Y и X, т.е. все выборочные значения (x_i, y_i) лежат на оцененной линии регрессии $y_x = \overline{y} + r_{yx}\frac{s_y}{s_x}(x - \overline{x})$;
- при $r_{yx} = 0$ линейная корреляционная зависимость между Y и X отсутствует, их групповые средние y_x и x_y совпадают с их общими средними $\overline{y}$ и $\overline{x}$, а регрессия Y по X принимает вид $y_x = \overline{y}$. Равенство $r_{yx} = 0$ говорит лишь об отсутствии *линейной* корреляционной зависимости между величинами Y и X, но не об отсутствии корреляции или стохастической зависимости между Y и X.

Проверка значимости коэффициента корреляции r_{yx} осуществляется путем проверки гипотезы $H_0: \rho = 0$, т.е. предположения об отсутствии линейной корреляционной зависимости между величинами Y и X. При справедливости этой гипотезы статистика

$$t = \frac{r_{yx}\sqrt{n-2}}{\sqrt{1 - r_{yx}^2}}$$

имеет распределение Стьюдента (t-распределение) с числом степеней свободы n-2, где n – объем выборки. Гипотеза $H_0: \rho = 0$ отвергается при уровне значимости α (т.е. выборочный коэффициент корреляции значимо отличается от нуля), если вычисленное по выборке объема n значение t удовлетворяет неравенству

$$|t| = \frac{|r_{yx}|\sqrt{n-2}}{\sqrt{1-r_{yx}^2}} > t(1-\alpha, n-2),$$

где $t(1-\alpha, n-2)$ – квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы $n-2$.

Доверительный интервал надежности $\gamma = 1-\alpha$ для генерального коэффициента корреляции ρ при значимом выборочном коэффициенте корреляции r строится с помощью z-преобразования Фишера

$$z = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right).$$

Распределение статистики z уже при малых объемах выборки близко к нормальному, что позволяет построить доверительный интервал для ее матема-

тического ожидания и от него перейти к доверительному интервалу для генерального коэффициента корреляции ρ, который имеет вид

$$th\left(z-\frac{t_{1-\alpha}}{\sqrt{n-3}}\right)<\rho<th\left(z+\frac{t_{1-\alpha}}{\sqrt{n-3}}\right),$$

где $th(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$ – гиперболический тангенс x, а $t_{1-\alpha}$ определяется по функции Лапласа из условия $\Phi(t_{1-\alpha})=(1-\alpha)/2$.

Для определения наличия и тесноты любой (линейной или нелинейной) корреляционной связи используется *эмпирическое корреляционное отношение* Y к X

$$\eta_{yx}=\sqrt{\frac{\delta_y^2}{S_y^2}},$$

которое тем больше, чем большее влияние на вариацию Y оказывает изменчивость X по сравнению с неучтенными факторами. Здесь $S_y^2=\frac{1}{n}\sum n_{y_i}(y_i-\bar{y})^2$ – общая выборочная дисперсия величины Y, $\delta_y^2=\frac{1}{n}\sum n_x(\bar{y}_x-\bar{y})^2$ – межгрупповая дисперсия Y. Свойства эмпирического корреляционного отношения:

- $0\le|r_{yx}|\le\eta_{yx}\le 1$;
- $\eta_{yx}\neq\eta_{xy}$;
- при $\eta_{yx}=0$ корреляционная связь между Y и X отсутствует;
- при $\eta_{yx}=1$ между Y и X существует функциональная зависимость;
- при $|r_{yx}|=\eta_{yx}$ между Y и X существует линейная корреляционная зависимость.

Величина η_{yx}^2 называется *эмпирическим коэффициентом детерминации*, она показывает, какая часть общей вариации Y обусловлена вариацией X.

Многомерный корреляционный анализ исследует корреляционную зависимость совокупности случайных величин $X_1,X_2,X_3,\dots,X_p$ имеющих совместное нормальное распределение. *Корреляционная матрица* Q_p, составленная из парных генеральных коэффициентов корреляции ρ_{ij} величин X_i и X_j,

$$Q_p=\begin{bmatrix}1 & \rho_{12} & \cdots & \rho_{1p}\\ \rho_{21} & 1 & \cdots & \rho_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ \rho_{p1} & \rho_{p2} & \cdots & 1\end{bmatrix},$$

характеризует линейную корреляционную зависимость между парами величин X_i и X_j этой совокупности. *Основная задача многомерного корреляционного анализа* состоит в оценке корреляционной матрицы Q_p по выборочным данным. Такой оценкой является *матрица выборочных коэффициентов корреляции*

$$q_p=\begin{bmatrix}1 & r_{12} & \cdots & r_{1p}\\ r_{21} & 1 & \cdots & r_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ r_{p1} & r_{p2} & \cdots & 1\end{bmatrix},$$

в которой r_{ij} – выборочные коэффициенты корреляции между величинами X_i и X_j. Матрицы Q_p и q_p симметричные, поэтому при вычислении матрицы q_p приводятся только элементы, расположенные на главной диагонали и под ней.

Теснота линейной корреляционной связи одной из величин X_i с совокупностью остальных $p-1$ величин $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p$ оценивается *выборочным коэффициентом множественной корреляции*

$$R_{i/1,\dots,i-1,i+1,\dots,p} = \sqrt{1 - \frac{|q_p|}{q_{ii}}},$$

где $|q_p|$ – определитель матрицы q_p, q_{ii} – алгебраическое дополнение элемента r_{ii} матрицы q_p. В частности, для трех величин X_1, X_2, X_3 выборочный коэффициент множественной корреляции $R_{i/jk}$ вычисляется по формуле

$$R_{i/jk} = \sqrt{\frac{r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}r_{jk}}{1 - r_{jk}^2}}.$$

Выборочный коэффициент множественной корреляции принимает значения от 0 до 1. Чем ближе значение $R_{i/1,\dots,i-1,i+1,\dots,p}$ к единице тем теснее линейная корреляционная связь X_i с остальными величинами $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p$. Величина $R^2 = (R_{i/1,\dots,i-1,i+1,\dots,p})^2$ называется *выборочным множественным коэффициентом детерминации*, которая показывает долю вариации переменной X_i объясняемую вариацией остальных переменных. Множественный коэффициент корреляции $R_{i/1,\dots,i-1,i+1,\dots,p}$ значим при уровне значимости α, если вычисленное значение F-статистики

$$F = \frac{R^2(n-p)}{(1-R^2)(p-1)} > F(\alpha, p-1, n-p),$$

где $F(\alpha, p-1, n-p)$ значение -критерия на уровне значимости α при числе степеней свободы $k_1 = p-1$ и $k_2 = n-p$.

Частные коэффициенты корреляции. Если величины из совокупности $X_1, X_2, X_3, \dots, X_p$ коррелируют друг с другом, то на величинах парных коэффициентов корреляции r_{ij} переменных X_i и X_j сказывается влияние и других переменных совокупности, что приводит к искажению значений коэффициентов корреляции r_{ij}. Для оценки линейной корреляционной зависимости между величинами X_i и X_j, очищенной от влияния других величин совокупности, используется *выборочный частный коэффициент корреляции* $r_{ij/1,\dots,p}$. Он определяется соотношением

$$r_{ij/1,\dots,p} = \frac{-q_{ij}}{\sqrt{q_{ii} \cdot q_{jj}}},$$

где q_{ij}, q_{ii} , q_{jj} алгебраические дополнения соответственно элементов r_{ij}, r_{ii}, r_{jj} матрицы выборочных коэффициентов корреляции q_p. Например, для совокупности из трех случайных величин X_1, X_2, X_3 выборочные частные коэффициенты корреляции $r_{ij/k}$ находятся по формуле

$$r_{ij/k} = \frac{r_{ij} - r_{ik} \cdot r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}.$$

Частный коэффициент корреляции принимает значения от -1 до +1. Значимость частного коэффициента корреляции при заданном уровне α определяется также, как и значимость коэффициента корреляции с помощью t-статистики: если

$$|t| = \frac{|r_{ij/1,\dots,p}|\sqrt{n-p}}{\sqrt{1 - r_{ij/1,\cdots,p}^2}} > t(1 - \alpha, n - p),$$

то частный коэффициент корреляции $r_{ij/k}$ значимо отличается от нуля. Матрица частных коэффициентов корреляции является также симметричной, элементы ее главной диагонали равны единице.

Содержание лабораторной работы.

1. Ввод выборочных данных для исследования корреляционной зависимости совокупности величин $X_1, X_2, \dots, X_p$.

2. Построение матрицы выборочных коэффициентов корреляции и оценка наличия и тесноты линейной корреляционной зависимости между парами величин.

3. Проверка значимости наибольшего по модулю коэффициента корреляции при уровне значимости $\alpha = 0{,}05$.

4. Построение доверительного интервала надежности $\gamma = 1 - \alpha$ для генерального коэффициента корреляции ρ между наиболее тесно связанными величинами заданной совокупности.

5. Нахождение выборочного коэффициента множественной корреляции $R_{1/2,\dots,p}$ и выборочного множественного коэффициента детерминации $R^2 = (R_{1/1,\dots,p})^2$.

6. Построение матрицы выборочных частных коэффициентов корреляции и оценка «очищенной» корреляционной зависимости X_1 с другими величинами совокупности.

7. Общее заключение о корреляционной зависимости исследуемых величин.

Выполнение работы в MS Excel.

Проведение корреляционного анализа в MS Excel-2010 приведем на примере исследования корреляционной зависимости трех величин: производительности труда (X_1) рабочих одинаковой квалификации, фондовооруженности (X_2) и энерговооруженности (X_3) их рабочих мест. Результаты выборочного обследования приведены в таблице 2.1, содержащей n = 14 наблюдений.

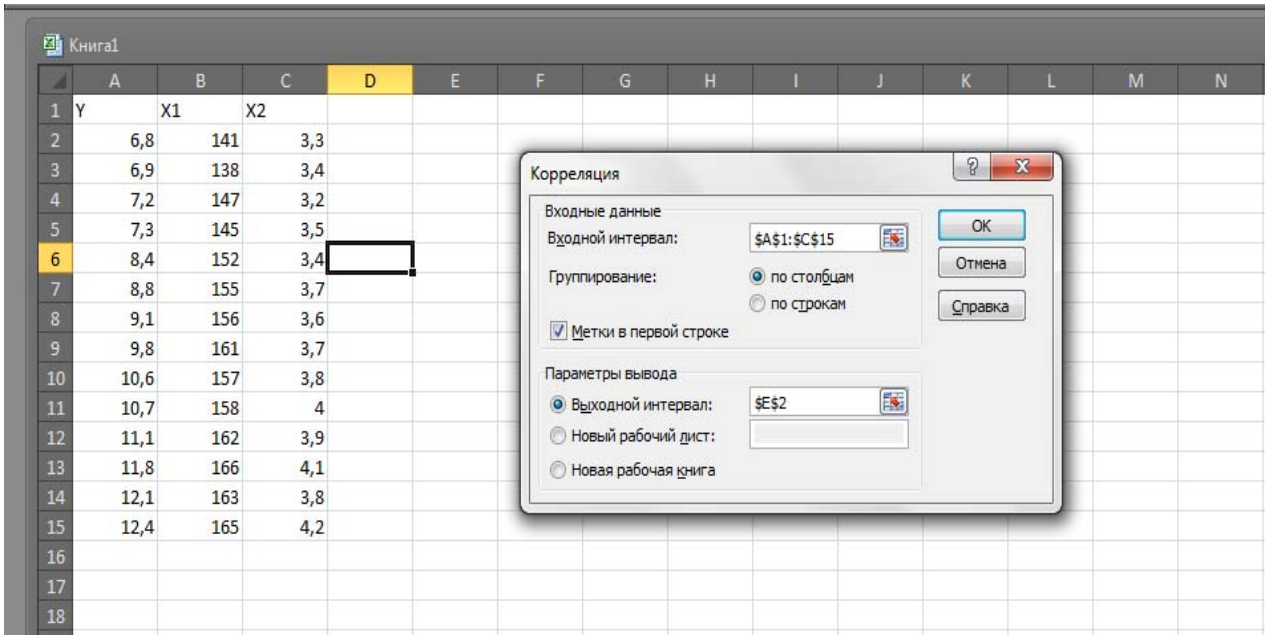
Таблица 2.1

X_1	6,8	6,9	7,2	7,3	8,4	8,8	9,1	9,8	10,6	10,7	11,1	11,8	12,1	12,4
X_2	141	138	147	145	152	155	156	161	157	158	162	166	163	165
X_3	3,3	3,4	3,2	3,5	3,4	3,7	3,6	3,7	3,8	4	3,9	4,1	3,8	4,2

Ввод данных для исследования корреляционной зависимости рассматриваемых величин. Введем данные расположив их по столбцам A, B и C: в первых

ячейках этих столбцов укажем имена переменных; значения X_1 разместим в ячейках **A2-A15**; значения X_2 в **B2-B15**; значения X_3 в ячейках **C2-C15**.

Построение матрицы выборочных коэффициентов корреляции. Откроем вкладку «**Данные**», в группе «**Анализ**» выберем надстройку «**Анализ данных**». В открывшемся окне «**Инструменты анализа**» выберем функцию «**Корреляция**». В части «Входные данные» окна «**Корреляция**», в поле «Входной интервал», укажем: расположение выборочных данных на листе Excel: **A1-C15**; выберем группирование «по столбцам», если значения переменных расположены по столбцам, если значения переменных расположены по строкам, то выбирается «по строкам»; поставим флажок в поле «Метки в первой строке (столбце)», что указывает на то, что в первой строке (столбце) сгруппированных по столбцам (строкам) данных находятся имена переменных. В части «Параметры вывода» выбирается место расположения результатов выполнения функции «**Корреляция**»: «Выходной интервал» – указывается ячейка текущего листа, с которого (вправо и вниз) будет расположена корреляционная матрица q_p; «Новый рабочий лист» – вывод корреляционной матрицы на новый рабочий лист; «Новая рабочая книга» – вывод корреляционной матрицы в новую рабочую книгу. Выберем «Выходной интервал» и ячейку **E2**, с которой будет расположена корреляционная матрица. По «ОК» получим в ячейках **E2-H5** корреляционную матрицу. Заполнение окна «Корреляция» для рассматриваемого примера приведено на рис. 2.1.



	A	B	C
1	Y	X1	X2
2	6,8	141	3,3
3	6,9	138	3,4
4	7,2	147	3,2
5	7,3	145	3,5
6	8,4	152	3,4
7	8,8	155	3,7
8	9,1	156	3,6
9	9,8	161	3,7
10	10,6	157	3,8
11	10,7	158	4
12	11,1	162	3,9
13	11,8	166	4,1
14	12,1	163	3,8
15	12,4	165	4,2

Рис. 2.1. Заполнение окна «Корреляция»

Результаты выполнения функции «Корреляция» приведены на рис. 2.2. На рис. 2.2 кроме корреляционной матрицы также приведены: выборочные данные; результаты проверки значимости парных коэффициентов корреляции; доверительный интервал для генерального коэффициента корреляции между X_1 и X_2; выборочный коэффициент множественной корреляции $R_{1/23}$; выборочный

множественный коэффициент детерминации R^2; матрица частных коэффициентов корреляции. Полученные результаты снабжены краткими поясняющими записями.

Корреляционная матрица расположена в ячейках **F3-H5**: в ячейке **F4** выборочный коэффициент корреляции $r_{21} = 0{,}944$, в ячейке **F5** выборочный коэффициент корреляции $r_{31} = 0{,}919$, в ячейке **G5** выборочный коэффициент корреляции $r_{32} = 0{,}854$. Пустые ячейки корреляционной матрицы заполним исходя из ее симметричности.

Проверка значимости коэффициентов парной корреляции. Для проверки значимости коэффициентов парной корреляции на заданном уровне значимости $\alpha = 0{,}05$ вычислим значения t-статистики, $t = \frac{r_{yx}\sqrt{n-2}}{\sqrt{1-r_{yx}^2}}$. Для вычисления t-статистики для r_{X1X2} выделим, например, ячейку **F10**, в строке формул введем

=F4*(14-2)^0,5/(1-F4^2)^0,5

По «Enter» в ячейке **F10** получим значение t-статистики для r_{X1X2}. Выделив ячейку **F11** и введя в строке формул **=F5*(14-2)^0,5/(1-F5^2)^0,5,** по «Enter» в ячейке **F11** получим значение t-статистики для r_{X1X3}. Выделив ячейку **F12** и введя в строке формул **=G5*(14-2)^0,5/(1-G5^2)^0,5,** по «Enter» в ячейке **F12** получим значение t-статистики для r_{X2X3} (см. рис. 2.2). Для нахождения критической точки $t(1-\alpha, n-2)$ распределения Стьюдента при заданном уровне значимости $\alpha = 0{,}05$ выделим, например, ячейку **F14**. В вкладке «**Формулы**» выберем «**Другие функции**», в группе «**Статистические**» выберем функцию «**СТЬЮДЕНТ.ОБР.2Х**». В окне этой функции в поле «Вероятность» введем значение α, равное 0,05, в поле «Степени свободы» зададим число степеней свободы n-2, равное 12. По «ОК» в ячейке **F14** получим значение $t(1-\alpha, n-2)$, в рассматриваемом примере оно равно 2,1788. Модули t-статистик для всех коэффициентов парной корреляции превышают критическое значение 2,1788 (см. рис. 2.2), следовательно, все коэффициенты парной корреляции значимы.

	A	B	C	D	E	F	G	H	I	J	K	L
1	X1	X2	X3	Корреляционная матрица								
2	6,8	141	3,3			*X1*	*X2*	*X3*				
3	6,9	138	3,4		X1	1	0,944018	0,919392				
4	7,2	147	3,2		X2	0,944018	1	0,853917				
5	7,3	145	3,5		X3	0,919392	0,853917	1				
6	8,4	152	3,4									
7	8,8	155	3,7									
8	9,1	156	3,6									
9	9,8	161	3,7			t-статистики						
10	10,6	157	3,8		rX1X2	9,912769						
11	10,7	158	4		rX1X3	8,096928						
12	11,1	162	3,9		rX2X3	5,684104						
13	11,8	166	4,1									
14	12,1	163	3,8		t(1-α,n-2)=	2,178813		Доверительный интервал для ρ(X1,X2)				
15	12,4	165	4,2		z(X1,X2)=	1,773738		0,828328		0,982491		
16					t(1-α)=	1,959964						

Рис. 2.2. Результаты корреляционного анализа

Построение доверительного интервала $th\left(z-\frac{t_{1-\alpha}}{\sqrt{n-3}}\right)<\rho<th\left(z+\frac{t_{1-\alpha}}{\sqrt{n-3}}\right)$ *надежности* $\gamma=1-\alpha=0{,}95$ *для генерального коэффициента корреляции* ρ_{X1X2}. Проведем z-преобразование Фишера для выборочного коэффициента корреляции $r_{X1X2}=0{,}944$. Для этого выделим, например, ячейку **F15**. В вкладке «**Формулы**» выберем «**Другие функции**», в группе «**Статистические**» выберем функцию «**ФИШЕР**». В окне этой функции в поле «x» введем значение коэффициента корреляции r_{X1X2}. По «ОК» в ячейке **F15** получим значение z, равное в этом примере 1,7736 (см. рис. 2.2). Для вычисления значений $z-\frac{t_{1-\alpha}}{\sqrt{n-3}}$ и $z+\frac{t_{1-\alpha}}{\sqrt{n-3}}$ предварительно найдем значение $t_{1-\alpha}$. Выделим, например, ячейку **F16.** В вкладке «**Формулы**» выберем «**Другие функции**», в группе «**Статистические**» выберем функцию «**НОРМ.СТ.ОБР.**». В окне этой функции в поле «Вероятность» введем значение $1-\alpha/2$, равное 0,975. По «ОК» в ячейке **F16** получим значение $t_{1-\alpha}$, равное 1,9599 (см. рис. 2.2).

Для получения нижней границы доверительного интервала для ρ_{X1X2} используем функцию **ФИШЕРОБР** вычисления гиперболического тангенса $th(x)$. Выделим ячейку **H15** и в строке формул введем

=ФИШЕРОБР(F15-F16/(14-3)^0,5)

По «Enter» в ячейке **H15** получим искомую нижнюю границу доверительного интервала, в этом примере равную 0,8283.

Для получения верхней границы доверительного интервала для ρ_{X1X2} выделим ячейку **J15** и в строке формул введем

=ФИШЕРОБР(F15+F16/(14-3)^0,5)

По «Enter» в ячейке **J15** получим искомую верхнюю границу доверительного интервала, равную в этом примере 0,9825 (см. рис. 2.2). Аналогичным образом могут быть построены доверительные интервалы для других генеральных коэффициентов корреляции.

Для нахождения выборочных коэффициентов множественной корреляции и частных коэффициентов корреляции построим предварительно матрицу алгебраических дополнений q_{ij} элементов выборочной корреляционной матрицы, см. рис. 2.3. Для этого в ячейке **A19** вычислим определитель корреляционной матрицы: выделим эту ячейку и в строке формул, учитывая расположение выборочной корреляционной матрицы, см. рис. 2.2, введем **=МОПРЕД(F3:H5)**. По Enter в **A19** получим значение определителя. Выделим ячейки **A21 – C23** и введем в строке формул **МОБР(F3:H5)**. Нажав **Ctrl+Shift+Enter**, в ячейках **A21 – C23** получим матрицу обратную к корреляционной матрице. Для получения матрицы алгебраических дополнений q_{ij} элементов выборочной корреляционной матрицы необходимо умножить элементы полученной обратной матрицы на определитель корреляционной матрицы. Матрицу алгебраических дополнений q_{ij} элементов выборочной корреляционной матрицы разместим в ячейках **F21 – H23**, см. рис. 2.3. Выделим ячейку **F21** и введя в строке формул **=A21*A19** по «Enter» в ячейке **F21** получим значение q_{11}. Аналогично

вычисляются другие алгебраические дополнения элементов корреляционной матрицы.

	A	B	C	D	E	F	G	H	I	J	K
17											
18	Определитель корреляционной марицы										
19	0,016642										
20	Обратная к корреляционной матрице					Матрица алгебраических дополнений				qij	
21	16,27361	-9,55012	-6,80681			0,270825	-0,15893	-0,11328			
22	-9,55012	9,296875	0,841535			-0,15893	0,154718	0,014005			
23	-6,80681	0,841535	6,539528			-0,11328	0,014005	0,108831			
24											
25											
26	Матрица частных коэффициентов корреляции										
27		X1	X2	X3		Коэффициент множественной корреляции X1 с X2,X3					
28	X1	1	0,776423	0,659825		0,968788					
29	X2	0,776423	1	-0,10793		Множественный коэффициент детерминации					
30	X3	0,659825	-0,10793	1		0,938551					
31											

Рис. 2.3. Частные коэффициенты корреляции

Нахождение выборочного коэффициента множественной корреляции $R_{1/2,3}$ *и выборочного множественного коэффициента детерминации* R^2. Для вычисления выборочного коэффициента множественной корреляции $R_{1/2,3} = \sqrt{1 - \frac{|q_p|}{q_{11}}}$ выделим, например, ячейку **F28** и в строке формул введем

=КОРЕНЬ(1-A19/F21).

По «Enter» в **F28** получим значение выборочного коэффициента множественной корреляции $R_{1/2,3}$, в примере оно равно 0,968788.

Выделив ячейку **F30** и введя в строке формул **=F28^2,** по «Enter» получим в этой ячейке значение множественного коэффициента детерминации R^2, в примере равное 0,938551.

Построение матрицы частных коэффициентов корреляции. Для этой матрицы отведем ячейки **B28-D30**, в ячейках **A28-A30** и **B27-D27** введем имена переменных X1, X2, X3, а над этими ячейками заголовок «Матрица частных коэффициентов корреляции». Для вычисления частных коэффициентов корреляции используем формулу $r_{ij/1,\dots,p} = \frac{-q_{ij}}{\sqrt{q_{ii} \cdot q_{jj}}}$. В примере необходимые алгебраические дополнения q_{ij} находятся в ячейках **F21:H23**. (см. рис. 2.3). В ячейки **B28, C29, D30** введем “1”. В силу симметрии этой матрицы вычислим только элементы, расположенные ниже главной диагонали. Выделив ячейку **B29** и введя в строке формул **=−F22/КОРЕНЬ(F21*G22)** по «ОК» в этой ячейке получим значение $r_{21/3}$. Выделив ячейку **B30** и введя в строке

мул =**−F23/КОРЕНЬ(F21*H23),** по «ОК» в ячейке **B30** получим значение $r_{31/2}$. Выделив ячейку **C30** и введя в строке формул =**−G23/КОРЕНЬ(G22*H23),** по «ОК» в ячейке **C30** получим значение $r_{32/1}$. Остальные элементы матрицы частных коэффициентов корреляции (ячейки **C28, D28, D29**) заполняются исходя из ее симметричности (см. рис. 2.3).

Общее заключение. Значения выборочных парных коэффициентов корреляции $r_{12} = 0{,}944$ и $r_{13} = 0{,}9194$ говорят о сильной линейной корреляционной зависимости производительности труда (X1) от фондовооруженности (X2) и энерговооруженности (X3). Фондовооруженность и энерговооруженность также сильно коррелированы, $r_{23} = 0{,}8539$. Все коэффициенты парной корреляции значимы, о чем свидетельствуют значения их t-статистик $t_{X1X2} = 9{,}913$, $t_{X1X3} = 8{,}097$, $t_{X2X3} = 5{,}684$, модули которых превышают критическое значение t-статистики $t(0{,}95; 12) = 2{,}179$. Для генерального коэффициента корреляции ρ_{X1X2} 95% - й доверительный интервал имеет вид (0,8283; 0,9824), что также говорит о сильной линейной корреляционной связи производительности труда и фондовооруженности. Значение множественного коэффициента корреляции X1 с X2 и X3 равно 0,9688. Значение множественного коэффициента детерминации говорит о том, что 93,86% вариации производительности труда объясняется вариацией фондовооруженности и энерговооруженности. Значения частных коэффициенты корреляции $r_{12/3} = 0{,}776$ и $r_{13/2} = 0{,}66$ и проверка их значимости говорят о значимом влиянии фондовооруженности и энерговооруженности на производительность труда. Проверка значимости частного коэффициента корреляции $r_{23/1} = -0{,}107$ говорит об отсутствии значимой линейной корреляционной зависимости фондовооруженности и энерговооруженности.

Контрольные вопросы

1. Сформулируйте понятия функциональной и стохастической зависимостей.
2. Какая взаимосвязь случайных величин называется корреляционной?
3. В чем заключается основная задача корреляционного анализа?
4. Для оценки какой корреляционной зависимости используется выборочный коэффициент корреляции? Каковы его свойства?
5. Как проверяется значимость коэффициента корреляции?
6. Что показывает интервальная оценка коэффициента корреляции?
7. Что характеризует эмпирическое корреляционное отношение? Каковы его свойства?
8. Что характеризует эмпирический коэффициент детерминации?
9. В чем заключается основная задача многомерного корреляционного анализа?
10. Какие величины являются элементами матрица выборочных коэффициентов корреляции?

11. Для совокупности трех случайных величин X, Y, Z получена матрица выборочных коэффициентов корреляции $\begin{bmatrix} 1 & 0,4 & 0,7 \\ 0,4 & 1 & 0,6 \\ 0,7 & 0,6 & 1 \end{bmatrix}$. Укажите наиболее тесно связанные пары величины.

12. Что оценивает выборочный коэффициент множественной корреляции?

13. Как проверяется значимость множественного коэффициента корреляции?

14. Что характеризует выборочный множественный коэффициент детерминации?

15. Определите выборочный множественный коэффициент детерминации $R^2 = (R_{1/2,3})^2$ по матрице выборочных коэффициентов корреляции, приведенной в 11-м вопросе.

16. Для характеристики какой взаимосвязи используется частный коэффициент корреляции?

17. Определите выборочный частный коэффициент корреляции $r_{13/2}$ по матрице выборочных коэффициентов корреляции, приведенной в 11-м вопросе.

18. Проверьте значимость частного коэффициента корреляции $r_{13/2}$, найденного в предыдущем вопросе, при объеме выборки n=19 и уровне значимости $\alpha = 0,05$.

Лабораторная работа № 3.
Однофакторный дисперсионный анализ

Цель работы. Изучение однофакторного дисперсионного анализа, овладение инструментом однофакторного дисперсионного анализа в MS Excel 2010.

Краткие сведения. *Однофакторный дисперсионный анализ*. При исследовании зависимостей одной из наиболее простых является ситуация, когда рассматривается влияние на изучаемую числовую величину (изучаемый признак) X только одного фактора, принимающего конечное число значений. Фактор, влияние которого изучается, может иметь качественный или количественный характер. Значения фактора называются *уровнем фактора* или *способом обработки*. Значения изучаемого признака называют *откликом*. Пусть фактор F имеет m уровней $F_1, F_2, \dots, F_m$, при каждом уровне фактора произведены несколько измерений отклика, x_{ij} – значение отклика (изучаемого признака) в -ом измерении при уровне фактора F_j, n_j – число измерений отклика при j-ом уровне фактора, общее число наблюдаемых (выборочных) значений изучаемого признака $N = n_1 + n_2 + \cdots + n_m$. Выборочные данные представляются в виде следующей таблицы.

Таблица 3.1

	Уровни фактора			
Результаты измерений отклика	F_1	F_2	$\cdots$	F_m
	x_{11}	x_{12}	$\cdots$	x_{1m}
	x_{21}	x_{22}	$\cdots$	x_{2m}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$x_{n_1 1}$	$x_{n_2 2}$	$\cdots$	$x_{n_m m}$

Выборочные данные, сгруппированные по уровням фактора F, расположены в столбцах табл. 3.1.

При изменении уровня фактора наибольшей изменчивости, как правило, подвержено положение (среднее, медиана) случайной величины (изучаемого признака). В однофакторных задачах предполагается, что выборочные данные x_{1j}, $x_{2j}, \dots, x_{n_j j}$ при разных уровнях фактора принадлежат некоторому сдвиговому семейству распределений. Часто в качестве такого сдвигового семейства распределений рассматривается семейство нормальных распределений. Различие выборочных данных при разных уровнях фактора может быть объяснено действием чистой случайности, что означает принадлежность всех данных одному и тому же распределению. Это предположение именуют *нулевой гипотезой* H_0. Если это предположение верно, то нет влияния фактора на изучаемый признак. В противном случае, при наличии влияния фактора, возникает задача оценки эффектов уровней фактора. Совокупность статистических методов проверки нулевой гипотезы и оценки эффектов уровней фактора, в предположении принадлежности выборочных данных семейству нормальных распределений, называется *дисперсионным анализом*.

Модель однофакторного дисперсионного анализа. Для описания данных табл. 3.1 используется аддитивная модель

$$x_{ij} = a_j + e_{ij}, (j = 1,2, \dots, m; i = 1,2, \dots, n_j), \quad (3.1)$$

которая показывает, на какие компоненты раскладывается значение изучаемого признака. *Здесь* a_j – неслучайная величина, являющаяся результатом действия j-го уровня фактора (эффект j-го уровня фактора); e_{ij} – случайная величина, отражающая внутренне присущую наблюдениям изменчивость (ошибка, вызванная влиянием других неучтенных факторов).

Модель (3.1) может быть представлена в виде

$$x_{ij} = \mu + \eta_j + e_{ij}, (j = 1,2, \dots, m; i = 1,2, \dots, n_j), \quad (3.2)$$

где μ – общая (генеральная) средняя изучаемого признака, η_j – неслучайная величина отражающая эффект j-го уровня фактора, $\eta_j = a_j - \mu$, e_{ij} – случайная ошибка.

Предпосылки однофакторного дисперсионного анализа:

- Математические ожидания ошибок e_{ij} равны нулю, т.е. $M(e_{ij}) = 0$;
- Ошибки e_{ij} взаимно независимы;
- Дисперсии ошибок e_{ij}, следовательно, и величин x_{ij}, одинаковые для любых i $и$ j, т.е. $D(e_{ij}) = \sigma^2$;
- Все ошибки e_{ij} распределены по нормальному закону $N(0, \sigma)$.

При выполнении предпосылок однофакторного дисперсионного анализа нулевая гипотеза об отсутствии влияния фактора на изучаемый признак выражается в равенстве эффектов уровней фактора a_j (групповых средних), и для модели (3.1) имеет вид

$$H_0: a_1 = a_2 = \cdots = a_m,$$

а для модели (3.2)

$$H_0: \eta_1 = \eta_2 = \cdots = \eta_m = 0.$$

Гипотеза о равенстве групповых средних нормальных совокупностей с одинаковыми дисперсиями σ^2 равносильна гипотезе о равенстве факторной (межгрупповой) и остаточной (внутригрупповой) дисперсий. Эти дисперсии в этом случае равны общей (генеральной) дисперсии σ^2. При равенстве групповых средних выборочная факторная $MS_F^2 = \frac{1}{m-1}\sum_{j=1}^{m} n_j(\bar{x}_j - \bar{x})^2$ и выборочная остаточная $MS_R^2 = \frac{1}{N-m}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$ дисперсии являются независимыми несмещенными оценками одной и той же генеральной дисперсии σ^2 и их различие незначимо. Здесь $\bar{x}_j$ – выборочное групповое среднее (среднее при j-ом уровне фактора), $\bar{x}$ – выборочная общая средняя изучаемого признака. Таким образом, проверка нулевой гипотезы H_0, при уровне значимости α, сводится к проверке существенности различия несмещенных выборочных оценок MS_F^2 и MS_R^2 дисперсии σ^2 с помощью F-критерия $F = \frac{MS_F^2}{MS_R^2}$, который имеет F-распределение Фишера-Снедекора с $k_1 = m - 1$ и $k_2 = N - m$ степенями свободы.

Гипотеза H_0 об отсутствии влияния фактора F на исследуемый признак X принимается, если вычисленное значение статистики $F = \frac{MS_F^2}{MS_R^2}$ меньше критического $F_{\text{кр}}(\alpha, m-1, N-m)$. В этом случае выборочные данные при разных уровнях фак-

тора принадлежат одному и тому же нормальному распределению и общая выборочная средняя $\bar{x}$ и выборочная остаточная дисперсия MS_R^2 являются несмещенными оценками математического ожидания и дисперсии σ^2 этого распределения. $F_{\text{кр}}(\alpha, m-1, N-m)$ – квантиль уровня $(1-\alpha)$ F-распределения Фишера-Снедекора с $k_1 = m-1$ и $k_2 = N-m$ степенями свободы.

Если $F = \frac{MS_F^2}{MS_R^2} > F_{\text{кр}}(\alpha, m-1, N-m)$, то гипотеза H_0 отвергается, т.е. фактор F оказывает существенное влияние на исследуемый признак. В этом случае выборочные данные при разных уровнях фактора принадлежат различным нормальным распределениям $N(a_j, \sigma)$, групповые выборочные средние $\overline{x_j}$ являются точечными оценками математических ожиданий a_j этих распределений (т.е. эффектов a_j уровней фактора). Доверительные интервалы для эффектов a_j уровней фактора (математических ожиданий a_j) надежности $\gamma = 1-2\alpha$ имеют вид

$$\overline{x_j} - t_{1-\alpha} \cdot \sqrt{\frac{MS_R^2}{n_j}} < a_j < \overline{x_j} + t_{1-\alpha} \cdot \sqrt{\frac{MS_R^2}{n_j}}, \qquad (3.3)$$

где $t_{1-\alpha}$ – квантиль уровня $(1-\alpha)$ распределения Стьюдента с числом степеней свободы равным $N-m$.

Содержание лабораторной работы.

1. Ввод выборочных данных для исследования влияния качественного фактора на изучаемый признак.

2. Проведение однофакторного дисперсионного анализа при уровне значимости $\alpha = 0{,}05$.

3. При наличии влияния фактора, построение интервальных оценок надежности $\gamma = 0{,}9$ для эффектов a_j уровней фактора.

4. Общее заключение о влиянии фактора на исследуемый признак.

Выполнение работы в MS Excel.

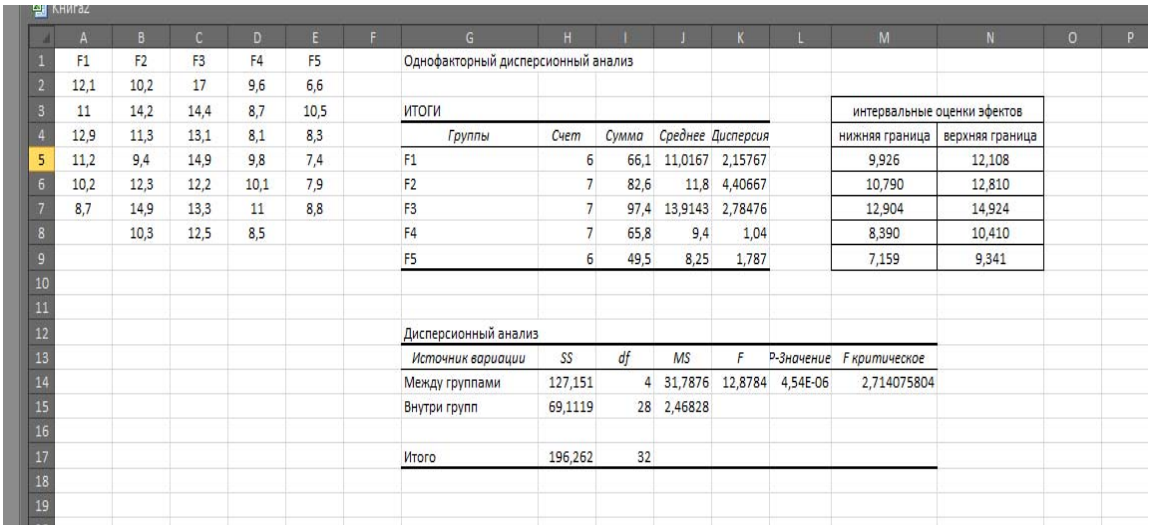
Проведение однофакторного дисперсионного анализа в MS Excel-2010 приведем на примере исследования влияния дня рабочей недели на производительность труда рабочих. На уровне значимости $\alpha = 0{,}05$ необходимо установить влияние различных дней недели на производительность труда. Результаты выборочного обследования производительности труда рабочих (признака X) по разным дням рабочей недели (фактор F) приведены в табл. 3.1.

Таблица 3.1

	Уровни фактора (дни рабочей недели)				
	F_1 (понед.)	F_2 (вторник)	F_3 (среда)	F_4 (четверг)	F_5 (пятница)
Результаты измерений производительности труда (отклика)	12,1	10,2	17, 0	9,6	6,6
	11,0	14,2	14,4	8,7	10,5
	12,9	11,3	13, 1	8,1	8,3
	11,2	9,4	14,9	9,8	7,4
	10,2	12,3	12,2	10,1	7,9
	8,7	14,9	13,3	11,0	8,8
	–	10,3	12,5	8,5	–

Таблица содержит $N = 33$ наблюдения, в понедельник и пятницу проведено по шесть измерений производительности труда, в остальные дни по семь.

Ввод данных для исследования зависимости. Введем, сгруппированные по дням недели, данные расположив их по столбцам A-E, в первых ячейках этих столбцов укажем уровни фактора, см. рис. 3.1.



	A	B	C	D	E
1	F1	F2	F3	F4	F5
2	12,1	10,2	17	9,6	6,6
3	11	14,2	14,4	8,7	10,5
4	12,9	11,3	13,1	8,1	8,3
5	11,2	9,4	14,9	9,8	7,4
6	10,2	12,3	12,2	10,1	7,9
7	8,7	14,9	13,3	11	8,8
8		10,3	12,5	8,5	
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					

Однофакторный дисперсионный анализ

ИТОГИ

Группы	Счет	Сумма	Среднее	Дисперсия
F1	6	66,1	11,0167	2,15767
F2	7	82,6	11,8	4,40667
F3	7	97,4	13,9143	2,78476
F4	7	65,8	9,4	1,04
F5	6	49,5	8,25	1,787

интервальные оценки эфектов	
нижняя граница	верхняя граница
9,926	12,108
10,790	12,810
12,904	14,924
8,390	10,410
7,159	9,341

Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-Значение	F критическое
Между группами	127,151	4	31,7876	12,8784	4,54E-06	2,714075804
Внутри групп	69,1119	28	2,46828			
Итого	196,262	32				

Рис. 3.1. Результаты дисперсионного анализа

Проведение однофакторного дисперсионного анализа MS Excel-2010. Откроем вкладку «**Данные**», в группе «**Анализ**» выберем надстройку «**Анализ данных**». В открывшемся окне «**Инструменты анализа**» выберем функцию «**Однофакторный дисперсионный анализ**». В части «Входные данные» окна этой функции в поле «Входной интервал» укажем:

- расположение выборочных данных на текущем листе Excel: **A1-E8**;
- выберем группирование «по столбцам»;
- поставим флажок в поле «Метки в первой строке», что указывает на то, что в первой строке сгруппированных по столбцам данных находятся имена уровней факторов;
- в поле «Альфа» приводится величина уровня значимости, равная 0,05.

В части «Параметры вывода» выбираем место расположения результатов выполнения функции «**Однофакторный дисперсионный анализ**», а именно, «Выходной интервал» и ячейку **G1**, с которой (вправо и вниз) будет расположена результаты выполнения этой функции. По «ОК» в ячейках **G1-M17** получим результаты дисперсионного анализа см. рис. 3.1.

Результаты однофакторного дисперсионного анализа представлены двумя таблицами: «**ИТОГИ**» и «**Дисперсионный анализ**».

В таблице «ИТОГИ» приведены:

- в столбце «Группы» указаны уровни фактора: $F_1, F_2, \dots, F_5$;
- в столбце «Счет» – количество n_j наблюдаемых значений изучаемого признака для каждого уровня фактора;

- в столбце «Сумма» – сумма $\sum_{i=1}^{n_j} x_{ij}$ наблюдаемых значений изучаемого признака для каждого уровня F_j фактора;
- в столбце «Среднее» – групповые выборочные средние $\bar{x_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$ для каждого уровня фактора;
- в столбце «Дисперсия» – групповые выборочные дисперсии $S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$ для каждого уровня фактора.

Таблица «Дисперсионный анализ» содержит:

- столбец «Источник вариации», с источниками вариации исследуемого признака;
- столбец «SS» содержит: сумму квадратов отклонений групповых средних от общей средней (факторная сумма квадратов отклонений) $SS_F^2 = \sum_{j=1}^{m} n_j(\bar{x_j} - \bar{x})^2$; внутригрупповую (остаточную) сумму квадратов отклонений $SS_R^2 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$ и общую (полную) сумму квадратов отклонений выборочных данных от общего среднего $SS_{\text{общ}}^2 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2$; $SS_{\text{общ}}^2 = SS_F^2 + SS_R^2$;
- столбец «df», содержащий число степеней свободы для каждой из сумм квадратов отклонений, они соответственно равны $m - 1,\ N - m,\ N - 1$;
- столбец «MS», с средним $MS_F^2 = \frac{SS_F^2}{m-1}$ квадратов отклонений групповых средних от общей средней и средним $MS_R^2 = \frac{SS_R^2}{N-m}$ внутригрупповой суммы квадратов отклонений;
- столбец «F» содержит вычисленное значение F-критерия $F = \frac{MS_F^2}{MS_R^2}$;
- столбец «p – значение» содержит уровень значимости α, при котором вычисленное значение F-критерия $F = \frac{MS_F^2}{MS_R^2}$ является критической точкой F-распределение Фишера-Снедекора с $k_1 = m - 1$ и $k_2 = N - m$ степенями свободы, если p – значение меньше заданного уровня значимости α, то нулевая гипотеза об отсутствии влияния фактора на изучаемый признак отвергается; если p – значение больше заданного уровня значимости α, то нулевая гипотеза об отсутствии влияния фактора на изучаемый признак принимается;
- столбец «F – критическое» содержит критическую точку $F_{\text{кр}}(\alpha, m - 1, N - m)$ F-распределения Фишера-Снедекора с $k_1 = m - 1$ и $k_2 = N - m$ степенями свободы для заданного уровня значимости α.

В рассматриваемом примере вычисленное значение *F*-статистики, равное 12,878, больше критического $F_{\text{кр}}$, равного 2,714. Также *P*-значение, равное $4{,}54 \cdot 10^{-6}$, меньше заданного уровня значимости $\alpha = 0{,}05$. Следовательно, гипотеза H_0 об отсутствии влияния фактора (дня рабочей недели) на изучаемый признак (производительность труда) отклоняется на уровне значимости $\alpha = 0{,}05$ и групповые выборочные средние, приведенные в столбце «Среднее» таблицы «ИТОГИ», являются точечными оценками эффектов a_j уровней фактора.

Построение интервальных оценок надежности $\gamma = 0{,}9$ для эффектов a_j уровней фактора.

При отклонении гипотезы H_0, т.е. при влиянии фактора на изучаемый признак, групповые средние $\bar{x}_j$ являются точечными оценками эффектов a_j уровней фактора в модели (3.1). Интервальные оценки эффектов уровней фактора определяются неравенством (3.3). На том же листе Excel построим таблицу «Интервальные оценки эффектов» для доверительных интервалов надежности 0,9 для эффектов уровней фактора. Под эту таблицу отведем ячейки **M3-N9**, см. рис. 3.1. Вычисленные нижние границы $\left(\bar{x}_j - t_{1-\alpha} \cdot \sqrt{\frac{MS_R^2}{n_j}}\right)$ интервальных оценок эффектов расположим в ячейках **M5-M9**, верхние $\left(\bar{x}_j + t_{1-\alpha} \cdot \sqrt{\frac{MS_R^2}{n_j}}\right)$ в ячейках **N5-N9**.

Для вычисления нижней границы интервальной оценки эффекта a_1 выделим ячейку **M5** и (учитывая расположение величин $\bar{x}_1$, n_1 и S_R^2, а также $N - m = 33 - 5 = 28$, $\alpha = 0{,}05$) в строке формул введем

=J5-СТЬЮДЕНТ.ОБР(0,95;28)*КОРЕНЬ(J15/H5).

По «ОК» в ячейке **M5** получим значение нижней границы.

Для вычисления верхней границы интервальной оценки эффекта a_1 выделим ячейку **N5** и в строке формул введем

=J5+СТЬЮДЕНТ.ОБР(0,95;28)*КОРЕНЬ(J15/H5).

По «ОК» в ячейке **N5** получим значение верхней границы.

Нижнюю и верхнюю границы интервальной оценки эффекта a_2 расположим в ячейках **M6** и **N6**, для их вычисления в приведенных выше формулах необходимо заменить **J5** на **J6**, **H5** на **H6**. Аналогичным образом находятся границы интервальных оценок эффектов a_3, a_4, a_5.

Общее заключение о влиянии фактора на исследуемый признак.

P-значение, приведенное в таблице «Дисперсионный анализ» и равное $4{,}54 * 10^{-6}$, меньше заданного уровня значимости $\alpha = 0{,}05$; вычисленное значение F-статистики, равное 12,878 больше критического значения $F_{\text{кр}}(0{,}05; 4; 28) = 2{,}714$. Следовательно, нулевая гипотеза о равенстве групповых средних (об отсутствии влияния дней недели на производительность труда) отвергается. Дни рабочей недели оказывают влияние на производительность труда. Точечные оценки эффектов уровней (средних производительностей труда по дням недели): понедельник – 11,0167; вторник – 11,8; среда – 13,91; четверг – 9,4; пятница – 8,25. Интервальные оценки, надежности 0,9, производительностей труда по дням недели имеют вид: понедельник – (9,926 – 12,108); вторник – (10,79–12,81); среда – (12,904–14,924); четверг – (8,39–10,41); пятница – (7,159–9,341).

Контрольные вопросы

1. Для исследования какой взаимосвязи используется однофакторный дисперсионный анализ?

2. Что такое фактор и его уровни?

3. Сформулируйте аддитивную модель однофакторного дисперсионного анализа.

4. Какое из слагаемых аддитивной модели однофакторного дисперсионного анализа определяет эффект уровня фактора.

5. Приведите предпосылки однофакторного дисперсионного анализа.

6. Сформулируйте нулевую гипотезу однофакторного дисперсионного анализа.

7. С помощью какого статистического критерия осуществляется проверка нулевой гипотезы однофакторного дисперсионного анализа?

8. Поясните таблицу дисперсионного анализа.

9. Если нулевая гипотеза однофакторного дисперсионного анализа принимается, то какому распределению принадлежат выборочные данные? Каковы значения параметров этого распределения?

10. Приведите точечные и интервальные оценки эффектов уровней фактора при отклонении нулевой гипотезы однофакторного дисперсионного анализа.

11. Какому распределению принадлежат выборочные данные при принятии нулевой гипотезы однофакторного дисперсионного анализа?

Лабораторная работа № 4. Парная линейная регрессия

Цель работы. Освоение построения по выборочным данным модели парной линейной регрессии, оценки точности и надежности параметров и всей модели, построения прогнозов значений зависимой переменной в MS Excel 2010. Интерпретация модели.

Краткие сведения. Модель парной линейной регрессии описывает зависимость условного среднего M_xY зависимой случайной величины Y в виде линейной функции значений x объясняющей переменной (фактора) X: $M_xY = a + bx$. Наблюдаемые в выборке $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ значения y_i зависимой переменной Y описываются в виде суммы детерминированной $a + bx_i$ и случайной ε_i составляющих:

$$y_i = a + bx_i + \varepsilon_i. \quad (4.1)$$

Случайная величина ε_i, называемая *ошибкой регрессии*, отражает влияние пропущенных объясняющих переменных, неправильной структуры и функциональной спецификации модели, агрегирования переменных, ошибки измерений.

Основные предпосылки парной линейной регрессии.

1. Связь значений зависимой величины от значений фактора задается соотношением (4) (эта зависимость называется *спецификацией модели*).

2. $x_1, x_2, \dots, x_n$ – детерминированные величины, линейно не связанные между собой, т.е. векторы $(x_1, x_2, \dots, x_n)$ и (1, 1, …, 1) не коллинеарные.

3. Ошибки регрессии ε_i – случайные величины с $M\varepsilon_i = 0$ и $D\varepsilon_i = \sigma^2$ для всех i.

4. Ошибки регрессии ε_i и ε_j (или переменные y_i и y_j) не коррелированы в разных наблюдениях, т.е. $Cov(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i, \varepsilon_j) = 0$.

5. Ошибки регрессии ε_i распределены по нормальному закону с нулевой средней и дисперсией σ^2, т.е. $\varepsilon_i \sim N(0, \sigma^2)$, соответственно $y_i \sim N(a + bx_i, \sigma^2)$.

Модель парной линейной регрессии содержит три неизвестных параметра: коэффициенты a и b уравнения регрессии и дисперсию σ^2 ошибок регрессии ε_i. Оценки коэффициентов a и b находятся из условия минимизации по a и b суммы квадратов

$$S = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

отклонений наблюдаемых значений y_i от вычисленных по уравнению регрессии $\hat{y}_i = a + bx_i$. Эти оценки называются *оценками метода наименьших квадратов* и определяются соотношениями

$$\hat{b} = \frac{Cov\,(X,Y)}{Var\,X} = \frac{r_{yx} \cdot s_y}{s_x}, \; \hat{a} = \bar{y} - \hat{b} \cdot \bar{x},$$

где $Cov(X, Y)$ – выборочная ковариация величин X и Y, $Var\,X$ – выборочная дисперсия X, s_x и s_y – выборочные среднеквадратические отклонения величин X и Y, $\bar{y}$ и $\bar{x}$ – выборочные средние Y и X.

Согласно теоремы Гаусса-Маркова, при выполнении предпосылок 1–4, эти оценки обладают наименьшей дисперсией в классе всех линейных несмещенных оценок.

Величины $e_i = y_i - \hat{a} - \hat{b}x_i$ называются *остатками регрессии*.

Несмещенной оценкой дисперсии σ^2 ошибок регрессии ε_i является величина

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

Оценки дисперсий оценок $\hat{b}$ и $\hat{a}$ определяются как

$$\widehat{Var}(\hat{b}) = \frac{s^2}{n \cdot VarX}, \widehat{Var}(\hat{a}) = \frac{s^2}{n}\left[1 + \frac{(\bar{x})^2}{VarX}\right].$$

Стандартные отклонения коэффициентов уравнения регрессии определяются соотношениями $s_b = \sqrt{\widehat{Var}(\hat{b})}$ и $s_a = \sqrt{\widehat{Var}(\hat{a})}$.

Интервальные оценки параметров уравнения регрессии надежности $\gamma = 1 - \alpha$:

$$\begin{gathered} \hat{a} - t_\gamma \cdot s_a < a < \hat{a} + t_\gamma \cdot s_a, \\ \hat{b} - t_\gamma \cdot s_b < b < \hat{b} + t_\gamma \cdot s_b, \\ \frac{(n-2)s^2}{U_2} < \sigma^2 < \frac{(n-2)s^2}{U_1}, \end{gathered} \qquad (4.2)$$

где α заданный уровень значимости, $t_\gamma = t\left(\frac{1+\gamma}{2}, n-2\right)$ – квантиль уровня $\frac{1+\gamma}{2}$ распределения Стьюдента (t-распределения) с числом степеней свободы $n-2$, $U_1 = \chi^2\left(\frac{1-\gamma}{2}, n-2\right)$ и $U_2 = \chi^2\left(\frac{1+\gamma}{2}, n-2\right)$ – квантили соответственно уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$ распределения χ^2 с числом степеней свободы $n-2$.

Оцененное уравнение регрессии Y на X имеет вид $\hat{y} = \hat{a} + \hat{b}x$. *Статистическая значимость параметров уравнения регрессии (их значимое отличие от нуля)* определяется путем проверки принадлежности нулевых значений доверительным интервалам. Если доверительный интервал надежности $\gamma = 1 - \alpha$ содержит ноль, то нулевая гипотеза о равенстве параметра нулю принимается с уровнем значимости α. Проверка значимого отличия от нуля параметров a и b уравнения регрессии осуществляется также путем проверки нулевых гипотез $H_0: a = 0$ и $H_0: b = 0$ против альтернативных гипотез $H_1: a \neq 0$ и $H_1: b \neq 0$. Для проверки этих гипотез используются t- статистики $t_a = \frac{\hat{a}}{s_a}$ и $t_b = \frac{\hat{b}}{s_b}$, распределенные по закону Стьюдента с $(n-2)$ степенями свободы. Если вычисленные значения статистик $t_a = \frac{\hat{a}}{s_a}$ или $t_b = \frac{\hat{b}}{s_b}$ превышают по модулю критическое значение $t(1-\alpha, n-2)$, то нулевая гипотеза отвергается и принимается альтернативная гипотеза (параметр значимо отличается от нуля). Если вычисленное значения t-статистики по модулю меньше критического значения $t(1-\alpha, n-2)$, то нулевая гипотеза принимается (параметр незначимо отличается от нуля) при заданном уровне α. Критическое значение $t(1-\alpha, n-2)$ определяется как квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы $n-2$. Принятие нулевой гипотезы $H_0: b = 0$ говорит об отсутствии значимой линейной корреляционной зависимости величин Y и X.

Верификация и оценка качества модели. Верификация модели парной линейной регрессии означает проверку соответствия модели эмпирическим данным и заключается в установлении значимости уравнения регрессии, т.е. в значимости влияния фактора X на условную среднюю зависимой величины Y. Проверка значимости уравнения регрессии заключается в проверке нулевой гипотезы $H_0: b = 0$, об отсутствии влияния фактора X на зависимую величину Y, против альтернативной гипотезы $H_1: b \neq 0$, о значимом влиянии фактора X на Y. Значимость уравнения регрессии может быть проверена двумя равноценными способами: с использованием дисперсионного анализа; с использование теории корреляции.

Дисперсионный анализ в линейной регрессии основывается на том, что общая сумма квадратов отклонений y_i от их общего среднего $\bar{y}$, $SS_{\text{общ.}} = \sum(y_i - \bar{y})^2$, разлагается на сумму квадратов отклонений, объясняемых регрессией, $SS_R = \sum(\hat{y}_i - \bar{y})^2$, и остаточную сумму квадратов отклонений $SS_{\text{ост.}} = \sum(y_i - \hat{y}_i)^2$. $SS_{\text{общ.}} = SS_R + SS_{\text{ост.}}$. При справедливости нулевой гипотезы $H_0: b = 0$ средние квадраты $MS_R = \frac{SS_R}{m-1}$ и $MS_{\text{ост.}} = \frac{SS_{\text{ост.}}}{n-m}$ являются независимыми несмещенными оценками одной и той же генеральной дисперсии σ^2 зависимой переменной Y и их различие незначимо. Проверка нулевой гипотезы $H_0: b = 0$, при уровне значимости α, сводится к проверке существенности различия несмещенных выборочных оценок $MS_{\text{ост}}$ и MS_R дисперсии σ^2 с помощью F-критерия $F = \frac{MS_R}{MS_{\text{ост}}}$, который имеет F-распределение Фишера-Снедекора с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы, где m число коэффициентов в уравнении регрессии, а n объем выборки. Гипотеза H_0 об отсутствии влияния фактора X на исследуемый признак Y принимается, если вычисленное значение статистики $F = \frac{MS_R}{MS_{\text{ост}}}$ меньше критического $F_{\text{кр}}(\alpha, m-1, n-m)$. Если $F = \frac{MS_R}{MS_{\text{ост}}} > F_{\text{кр}}(\alpha, m-1, n-m)$, то гипотеза H_0 отвергается и принимается гипотеза $H_1: b \neq 0$, т.е. фактор X оказывает влияние на исследуемый признак Y. $F_{\text{кр}}(\alpha, m-1, n-m)$ – квантиль уровня $(1-\alpha)$ F-распределения Фишера-Снедекора с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы.

Использование элементов теории корреляции при проверке значимости уравнения регрессии основано на соотношении $\hat{b} = r_{yx} \cdot \frac{s_y}{s_x}$ и заключается в проверке значимого отличия от нуля коэффициента корреляции r_{yx}, следовательно, и значимости коэффициента регрессии b. Проверка нулевой гипотезы $H_0: r_{yx} = 0$, т.е. предположения об отсутствии линейной корреляционной зависимости между величинами Y и X, производится с помощью статистики $t = \frac{r_{yx}\sqrt{n-2}}{\sqrt{1-r_{yx}^2}}$, которая при справедливости нулевой гипотезы имеет распределение Стьюдента (t-распределение) с числом степеней свободы $n - 2$. Гипотеза $H_0: r_{yx} = 0$ отвергается при уровне значимости α (т.е. оцененное уравнение линейной регрессии значимо), если вычисленное по выборке объема n значение t-статистики удовлетворяет неравенству

$$|t| = \frac{|r_{yx}|\sqrt{n-2}}{\sqrt{1-r_{yx}^2}} > t(1-\alpha, n-2), \quad (4.3)$$

где $t(1-\alpha, n-2)$ – квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы $n-2$. Если нулевая гипотеза $H_0: r_{yx} = 0$ принимается, то оцененное уравнение линейной регрессии незначимо – зависимая величина *Y* и фактор *X* не связаны линейной корреляционной зависимостью.

Для парной линейной регрессии оба способа проверки значимости уравнения регрессии равнозначны, а *F*-критерий и *t*-критерий связаны равенством $F = t^2$.

Мерой качества уравнения регрессии и характеристикой прогностической силы регрессионной модели является *коэффициент детерминации*

$$R^2 = 1 - \frac{SS_{\text{ост.}}}{SS_{\text{общ.}}},$$

который показывает, какая доля вариации зависимой переменной объясняется вариацией фактора. $0 \le R^2 \le 1$. Значимое отличие от нуля коэффициента детерминации R^2 устанавливается также с помощью приведенного выше *F*-критерия. *F*-критерий и коэффициент детерминации R^2 связаны равенством $F = (n-2)\frac{R^2}{1-R^2}$. Для парной линейной регрессии коэффициент детерминации $R^2 = r_{yx}^2$.

Качество подгонки построенной линейной регрессии к выборочным данным характеризуется средней ошибкой аппроксимации

$$A = \frac{1}{n}\sum\left|\frac{e_i}{y_i}\right| \cdot 100\%.$$

Прогнозирование по уравнению регрессии. *Точечный прогноз* $\hat{y}_x$ среднего зависимой величины Y для заданного значения x вычисляется по уравнению регрессии $\hat{y}_x = \hat{a} + \hat{b}x$ и является наилучшей несмещенной линейной оценкой теоретического условного среднего $M_x Y = a + bx$. *Доверительный интервал* надежности γ прогноза условного среднего для заданного значения x задается неравенством

$$\hat{y}_x - t_\gamma \cdot \sqrt{\hat{D}(\hat{y}(x))} < M_x Y < \hat{y}_x + t_\gamma \cdot \sqrt{\hat{D}(\hat{y}(x))}. \quad (4.4)$$

Здесь t_γ – квантиль уровня $\frac{1+\gamma}{2}$ распределения Стьюдента с числом степеней свободы $n-2$, $\hat{D}(\hat{y}(x)) = s^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{n \cdot VarX}\right)$ – оценка дисперсии прогноза условного среднего величины *Y*, $VarX$ – выборочная дисперсия независимой переменной *X*. Графики нижней и верхней границ доверительного интервала называются *доверительными кривыми* надежности γ. Уравнение линейной регрессии может быть записано в виде $\hat{y} = \bar{y} + \hat{b}(x - \bar{x})$. Отсюда следует, что линия регрессии проходит через точку $(\bar{x}, \bar{y})$ и при $x = \bar{x}$ доверительные кривые наиболее близко подходят к линии регрессии.

Содержание лабораторной работы.

1. Ввести выборочные данные и построить диаграмму рассеяния.
2. Оценить параметры уравнения парной линейной регрессии.

3. Проверить значимость коэффициента корреляции, параметров уравнения регрессии и самого уравнения регрессии при уровне значимости $\alpha = 0{,}05$.

4. Оценить точность построенной модели. Построить 95%-й доверительный интервал для дисперсии σ^2 ошибки регрессии.

5. Построить точечные и интервальные, надежности $\gamma = 0{,}95$, прогнозы среднего зависимой переменной для выборочных значений независимой переменной. Построить линию регрессии и 95%-е доверительные кривые.

6. Дать общее заключение об оцененной модели и ее интерпретацию.

Выполнение работы в MS Excel. Выполнение работы в Excel рассмотрим на примере построения регрессионной зависимости $y = a + bx$ совокупных расходов на жилье (y, млрд дол.) от располагаемого совокупного личного дохода (x, млрд дол.) (функции спроса на жилье в зависимости от располагаемого дохода), используя данные для США за 1959–1970 г., приведенные в книге К. Доугерти «Введение в эконометрику». Эти данные приведены на рис. 4.1.

Ввод данных и построение диаграммы рассеяния. Выборочные данные по расходам на жилье и располагаемому личному доходу разместим по столбцам: в ячейке **A1** имя независимой переменной x, в ячейках **A2–A13** ее наблюдаемые значения; в ячейке **B1** имя зависимой переменной y, в ячейках **B2–B13** ее наблюдаемые значения, соответствующие значениям независимой переменной.

Для построения диаграммы рассеяния выберем вкладку «**Вставка**», в группе «**Диаграммы**» выберем «**Точечная**», в ее окне выберем тип диаграммы «**Точечная с маркерами**». Далее во вкладке «**Работа с диаграммами**» откроем вкладку «**Конструктор**» и в группе «Макеты диаграмм» выберем «**Макет 1**», а в группе «**Данные**» откроем «**Выбрать данные**». В открывшемся окне «Выбор источника данных» в поле «Диапазон данных для диаграммы» введем диапазон ячеек с данными для диаграммы, в рассматриваемом примере $A$1:$B$13. Внимание! В первом столбце (строке) должны находится значения независимой переменной. По «ОК» на открытом листе Excel получим диаграмму рассеяния. В соответствующих полях введем необходимые названия осей координат и название диаграммы. Диаграмма рассеяния представлена на рис. 4.1.

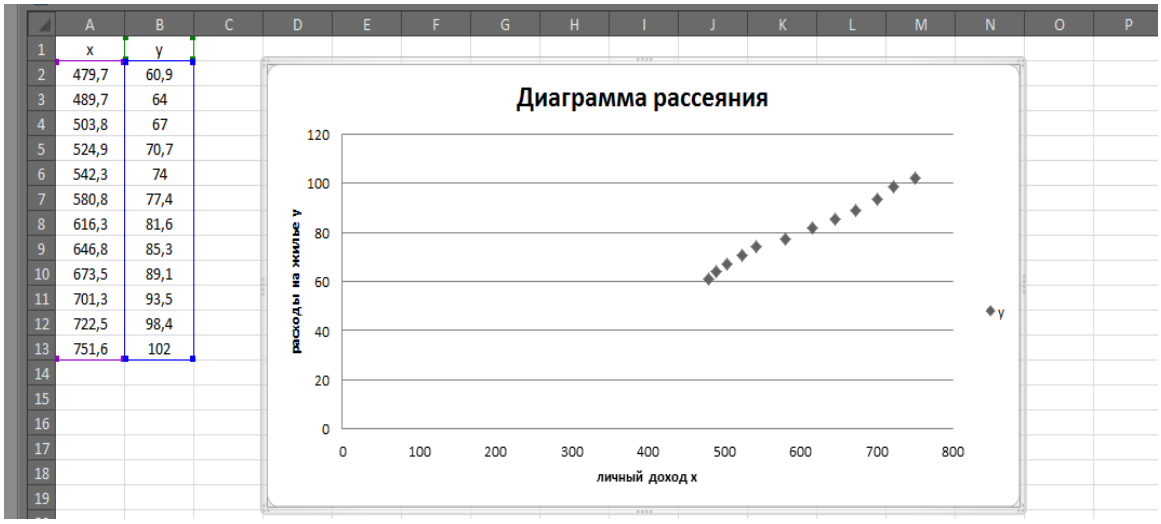


Рис. 4.1. Данные и диаграмма рассеяния

Оценка уравнения парной линейной регрессии. Откроем вкладку «**Данные**», в группе «**Анализ**» выберем надстройку «**Анализ данных**». В открывшемся окне «**Инструменты анализа**» выберем функцию «**Регрессия**». В появившемся окне «Регрессия» укажем входные данные для оценки параметров регрессии, выводимые результаты и их расположение. Заполнение окна «Регрессия» для рассматриваемого примера приведено на рис. 4.2. В части «Входные данные» в поле ввода «Входной интервал Y» указываем диапазон ячеек, содержащий значения зависимой переменной, в нашем примере это **B1:B13**; в поле ввода «Входной интервал X» – диапазон ячеек, содержащий значения независимой переменной, в примере это **A1:A13**. В поле «Метки» устанавливаем флажок v, он указывает на то, что первые строки диапазонов данных содержат имена этих данных (заголовки). В «Константа-ноль» флажок не устанавливаем, в этом случае строится регрессия $y = a + bx$; при установке флажка строится регрессия $y = bx$ без постоянной a. При установке флажка v в левом поле «Уровень надежности», наряду с используемым по умолчанию стандартным уровнем надежности 95% ($\gamma = 0{,}95$), можно задать и другое его значение, в этом случае будут выведены интервальные оценки параметров регрессии для двух уровней надежности.

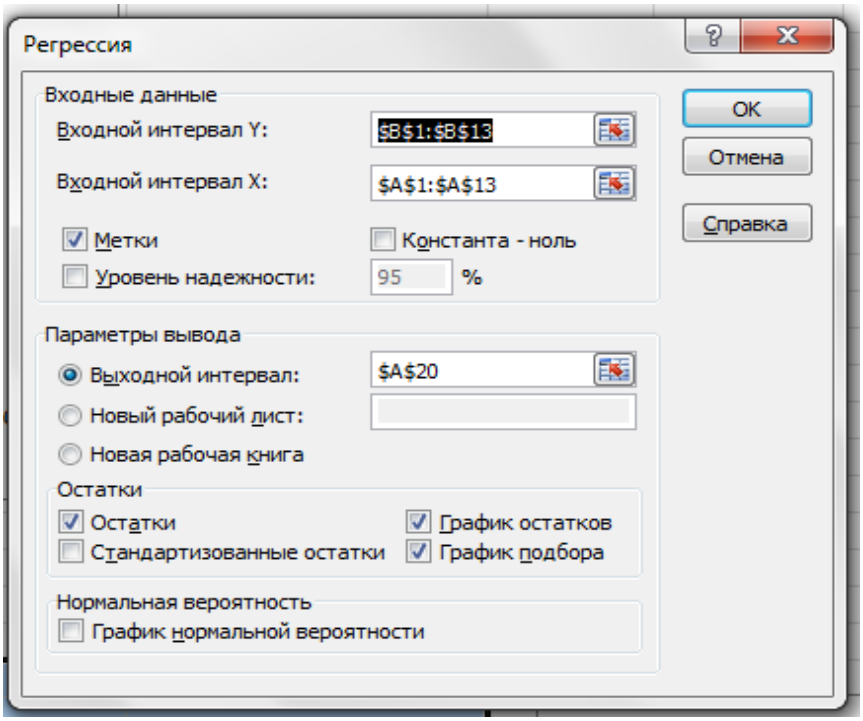


Рис. 4.2. Заполнение окна регрессия

В части «Параметры вывода» указывается одно из мест расположения выводимых результатов:

- «Выходной интервал» – для помещения результатов на текущем рабочем листе, положение результатов указывается заданием верхней левой ячейки, начиная с которой располагаются результаты;

• «Новый рабочий лист» – для расположения результатов на новом рабочем листе;

• «Новая книга» – для помещения результатов в новой книге.

В нашем примере выбран «Выходной интервал» и ячейка A20. Далее, выставляя флажки, указываем какую дополнительную информацию, предлагаемую функцией «Регрессия», мы хотим иметь в результатах:

• «Остатки» – для выдачи прогнозов $\hat{y}_i = \hat{a} + \hat{b} \cdot x_i$ и остатков регрессии $e_i = y_i - \hat{y}_i$;

• «Стандартизованные остатки» – для вывода нормированных остатков e_i / s;

• «График нормальной вероятности» – для вывода таблицы, в которой указывается какими перцентилями являются наблюдаемые значения зависимой переменной *Y*, и построения соответствующего графика;

• «График остатков» – для вывода точечной диаграммы остатков e_i;

• «График подбора» – для вывода наложенных на диаграмму рассеяния точек $(x_i, \hat{y}_i)$ линии регрессии $\hat{y}_i = \hat{a} + \hat{b}x_i$.

В примере выбраны «Остатки», «График остатков» и «График подбора». По ОК получаем результаты регрессии, которые включают в себя таблицу регрессионной статистики, таблицу дисперсионного анализа, таблицу коэффициентов регрессии, таблицу остатков и графики остатков и подбора. Результаты регрессии приведены на рис. 4.3–4.4. В действительности на экране несколько иная картина, что обусловлено тем, что заголовки некоторых строк и столбцов таблиц не умещаются в ячейках и выводимые графики наложены друг на друга и расположены в правой верхней части экрана. Проведем коррекцию представления полученных результатов.

	A	B	C	D	E	F	G	H	I	J
19										
20	ВЫВОД ИТОГОВ									
21										
22	*Регрессионная статистика*			Дисперсионный анализ						
23	Множественный R	0,995739			*df*	*SS*	*MS*	*F*	*Значимость F*	
24	R-квадрат	0,991497		Регрессия	1	2018,748969	2018,748969	1165,994979	1,09804E-11	
25	Нормированный R-квадрат	0,990646		Остаток	10	17,31353056	1,731353056			
26	Стандартная ошибка	1,315809		Итого	11	2036,0625				
27	Наблюдения	12								
28										
29										
30		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*	
31	Y-пересечение	-4,78851	2,5213615	-1,89917	0,08674	-10,4064486	0,829438539	-10,4064486	0,829438539	
32	x	0,141205	0,0041352	34,14667	1,1E-11	0,131990838	0,150418624	0,131990838	0,150418624	
33										

Рис. 4.3. Таблицы итогов регрессии

Прежде всего, отформатируем ячейки содержащие заголовки для получения их полного текста. Для этого, выделив ячейку, щелкнем на ней правой клавишей мышки и в появившемся меню выберем ***Формат ячейки***, затем в окне формата ячейки щелкнем ***Выравнивание*** и в его окне установим флажок в позиции ***перенос по словам***, щелкнув *ОК* получим полный текст заголовка. Разнесем графики подбора и остатков, разместив их рядом с таблицей остатков.

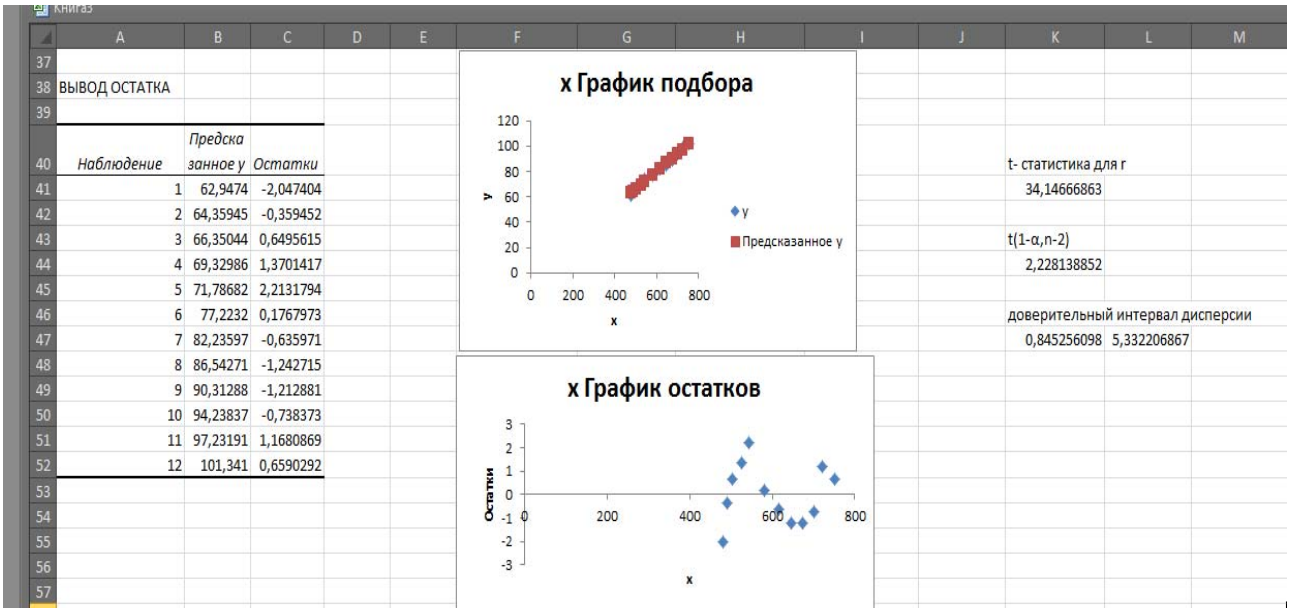


Рис. 4.4. Остатки и графики результатов регрессии

Пояснения к таблице «Регрессионная статистика»:

- Множественный R – множественный коэффициент корреляции между y и $\hat{y}$, для парной линейной регрессии значение выборочного коэффициента корреляции r_{yx};
- R-квадрат – коэффициент детерминации R^2;
- Нормированный R-квадрат – скорректированный коэффициент детерминации $\bar{R}^2 = R^2 - \frac{m-1}{n-m}(1-R^2)$, где m число коэффициентов в модели регрессии;
- Стандартная ошибка – оценка s среднеквадратического отклонения σ ошибок регрессии ε_i, т.е. $s = \sqrt{s^2}$;
- Наблюдений – объем выборки n.

Пояснения к таблице «Дисперсионный анализ»:

- df – число степеней свободы.
- SS – сумма квадратов.
- MS – средние квадраты.
- F – вычисленное значение критерия Фишера (F-статистики).
- Значимость F – уровень значимости, при котором вычисленное значение критерия Фишера является критической точкой распределения Фишера. Нулевая гипотеза о незначимости уравнения регрессии ($H_0: b = 0$) отклоняется, если это значение меньше заданного уровня значимости α.

• В строке «Регрессия» приведены число степеней свободы равное $m-1$, сумма квадратов отклонений SS_R объясняемых регрессией, средний квадрат MS_R, значение F и значимость F.

• В строке «Остаток» приведены число степеней свободы равное $n-m$, остаточная сумма квадратов отклонений $SS_{\text{ост.}}$, остаточный средний квадрат $MS_{\text{ост.}}$

• В строке «Итого» приведены число степеней свободы $n-1$ и общая сумма квадратов отклонений $SS_{\text{общ.}}$.

Следующая таблица содержит МНК-оценки коэффициентов уравнения регрессии, их стандартные ошибки, значения t-статистик для проверки нулевых гипотез $H_0: a = 0$ и $H_0: b = 0$, P-значения и границы доверительных интервалов для коэффициентов уравнения регрессии для заданных надежностей.

В строке с именем «Y-пересечение» приводятся:

- оценка $\hat{a}$ коэффициента a;
- ее стандартная ошибка s_a;
- вычисленное значение t-статистики, равное $\hat{a}/s_a$;
- P-значение – вероятность того, что случайная величина имеющая распределение Стьюденте (t-распределение) с числом степеней свободы n-2 примет значение по абсолютной величине больше, чем модуль вычисленного значения t-статистики, т.е. P-значение это уровень значимости, при котором вычисленное значение t-статистики является критической точкой, следовательно, нулевая гипотеза $H_0: a = 0$ отклоняется, если P-значение меньше заданного уровня значимости, и принимается в противном случае;
- нижняя и верхняя границы 95 % доверительного интервала для a.

В строке с именем «X» приводятся аналогичные данные для коэффициента b уравнения регрессии.

Таблица «Вывод остатка» содержит порядковые номера наблюдений (i), предсказанные (прогнозные) значения среднего зависимой переменной $\hat{y}_i = \hat{a} + \hat{b}x_i$ и остатки регрессии $e_i = y_i - \hat{y}_i$.

На графике подбора выводится диаграмма рассеяния и точки $(x_i, \hat{y}_i)$ линии регрессии $\hat{y}_i = \hat{a} + \hat{b}x_i$. На графике остатков представлены остатки e_i для наблюдаемых значений x_i.

Таким образом, в рассматриваемом примере выполнив функцию «Регрессия» мы получили:

- уравнение регрессии $\hat{y}_x = -4{,}788 + 0{,}141x$;
- оценку среднеквадратического отклонения ошибок регрессии $s = 1{,}316$ и оценку дисперсии ошибок $s^2 = 1{,}731$;
- 95%-е доверительные интервалы для коэффициентов уравнения регрессии $-10{,}406 < a < 0{,}829$ и $0{,}131 < b < 0{,}15$;
- значение t-статистики для коэффициента a, $t_a = -1{,}899$, и ее P-значение, равное $0{,}0867$. P-значение больше заданного уровня значимости $\alpha = 0{,}05$ поэтому принимаем гипотезу $H_0: a = 0$, коэффициент a незначимо отличается от нуля;

• значение t-статистики для коэффициента b, $t_b = 34{,}147$, и ее P-значение равное $1{,}1 \cdot 10^{-11}$, что значительно меньше заданного уровня значимости 0,05, поэтому отклоняем гипотезу $H_0: b = 0$, следовательно, уравнение регрессии значимо;

• коэффициент детерминации $R^2 = 0{,}9915$; вычисленное значение F-статистики, $F = 1165{,}99$ и ее уровень значимости, равный $1{,}098 \cdot 10^{-11}$, что значительно меньше заданного уровня значимости 0,05, это позволяет отклонить нулевую гипотезу о незначимости коэффициента детерминации R^2 и сделать вывод о значимости уравнения регрессии;

• выборочный коэффициент корреляции, совпадающий со значением «Множественный R» таблицы «Регрессионная статистика», т.е. $r_{yx} = 0{,}9957$;

• прогнозные значения $\hat{y}_x$ среднего зависимой переменной и остатки регрессии e_x для наблюдаемых значений x;

• линию регрессии, наложенную на диаграмму рассеяния и график остатков.

Проверка значимости коэффициента корреляции. Проверка значимости коэффициента корреляции, гипотезы $H_0: r_{yx} = 0$, заключается в проверке неравенства (4.3). В ячейке **B23** находится значение коэффициента корреляции, объем выборки $n = 12$. Для вычисления t- статистики $t = \frac{r_{yx}\sqrt{n-2}}{\sqrt{1-r_{yx}^2}}$ для коэффициента корреляции выделим, например, ячейку **К41** и в строке формул введем **=B23*(12-2)^0,5/(1-B23^2)^0,5.** По «Enter» в ячейке **К41** получим значение t-статистики равное 34,147. Для нахождения критической точки $t(1-\alpha, n-2)$ распределения Стьюдента при заданном уровне значимости $\alpha = 0{,}05$ выделим, например, ячейку **К44**. В вкладке «**Формулы**» выберем «**Другие функции**», в группе «**Статистические**» выберем функцию «**СТЬЮДЕНТ.ОБР.2Х**». В окне этой функции в поле «Вероятность» введем значение α, равное 0,05, в поле «Степени свободы» зададим число степеней свободы n-2, равное 10. По «ОК» в ячейке **К44** получим значение $t(1-\alpha, n-2)$, в рассматриваемом примере оно равно 2,228 (см. рис. 4.4). Модуль t-статистики для коэффициента корреляции превышает критическое значение 2,228. Следовательно, коэффициент корреляции значимо отличается от нуля и построенное уравнение регрессии $\hat{y}_x = -4{,}788 + 0{,}141x$ значимо.

Построение 95%-о доверительного интервала для дисперсии σ^2 ошибки регрессии. Доверительный интервал надежности $\gamma = 1 - \alpha$ дисперсии σ^2 определяется неравенством (4.2). В примере величина s находится в ячейке **B26**, объем выборки равен 12, $\gamma = 0{,}95$. Функция ***ХИ2.ОБР*** находит односторонние критические точки распределения χ^2 при заданном уровне значимости и числе степеней свободы. Для нахождения $\frac{s^2(n-2)}{U_2}$ выделим ячейку **К47** и в строке формул введем **=B26^2*(12-2)/ХИ2.ОБР(0,975;10).** По «Enter» в этой ячейке получим 0,845. Для нахождения $\frac{s^2(n-2)}{U_1}$ выделим ячейку **L47** и в строке формул введем **=B26^2*(12-2)/ХИ2.ОБР(0,025;10)**. По «Enter» в этой ячейке получим

5,33 (см. рис.4.4). Следовательно, 95%-й доверительный интервал для σ^2 имеет вид $0{,}845 < \sigma^2 < 5{,}33$.

Построение интервальных прогнозов, надежности $\gamma = 1 - \alpha = 0{,}95$*, среднего зависимой переменной для выборочных значений независимой переменной и построение линии регрессии и 95%-х доверительных кривых.* Доверительный интервал для среднего зависимой переменной *y*, при заданном значении *x* объясняющей переменной, определяется неравенством (4.4). Для построения интервальных прогнозов условного среднего *Y* для выборочных значений x_i создадим, например, в ячейках **A60-D72** дополнительную таблицу. В ячейки **A61:A72** скопируем наблюдаемые значения x_i, в ячейки **B61:B72** скопируем прогнозные значения $\hat{y}(x_i)$. Эти значения должны быть упорядочены по возрастанию x_i. В столбцах ***C*** и ***D*** будем размещать нижние и верхние границы доверительных интервалов. Вычислим выборочную среднюю $\bar{x}$. Для этого выделим ячейку **A75** и в строке формул введем **=СРЗНАЧ(A61:A72),** по «Enter» получим в этой ячейке значение $\bar{x}$, равное 602,77. Вычислим выборочную дисперсию $Varx$. Для этого выделим ячейку **B75** и в строке формул введем **=ДИСП.В(A61:A72),** по «Enter» получим в этой ячейке значение $Varx$, равное 9204,31 (см. рис. 4.5). Для вычисления нижней границы $\hat{y}_x - t_\gamma \cdot \sqrt{\widehat{D}(\hat{y}(x))}$ при $x = x_1$ выделим ячейку ***C61*** и (учитывая *n=12, n-2=10* и расположение *s* в ячейке B26) в строке формул введем

=B61-СТЬЮДЕНТ.ОБР.2Х(0,05;10)*B26*(1/12+(A61-A75)^2/(12*B75))^0,5.

По «Enter» в ячейке ***C61*** получим искомое значение нижней границы при $x = x_1$. Для вычисления верхней границы $\hat{y}_x + t_\gamma \cdot \sqrt{\widehat{D}(\hat{y}(x))}$ для $x = x_1$ выделим ячейку ***D61*** и в строке формул введем

=B61+СТЬЮДЕНТ.ОБР.2Х(0,05;10)*B26*(1/12+(A61-A75)^2/(12*B75))^0,5.

По Enter в ячейке ***D61*** получим искомое значение верхней границы для $x = x_1$. Вычисление нижних и верхних границ доверительного интервала среднего зависимой переменной для других значений x_i производится аналогичным образом, нужно заменить в приведенных формулах ячейки ***B61*** и ***A61*** на имена ячеек содержащих соответствующие значения $\hat{y}(x_i)$ и x_i. Границы 95%-х доверительных интервалов прогнозов среднего приведены на рис. 4.5.

Используя полученные результаты, построим верхнюю и нижнюю доверительные кривые, а также линию регрессии. Для этого выделим, например, ячейку F62; в вкладке «**Вставка**» в группе «**Диаграммы**» выберем «**Точечная**» и среди типов диаграмм выберем «**Точечная с гладкими кривыми**». В открывшейся вкладке «**Конструктор**» в группе «**Макеты диаграмм**» выберем «**Макет1**», после чего в группе «**Данные**» щелкнем по «**Выбрать данные**». В открывшемся окне «**Выбор источника данных**» в поле «**Диапазон данных**» укажем положение данных, для рассматриваемого примера укажем ***A60:D72***. По ОК получим нужные графики, после чего скорректируем заголовок диаграммы и наименования осей координат. 95%-е доверительные кривые и линия регрессии приведены на рис. 4.5.

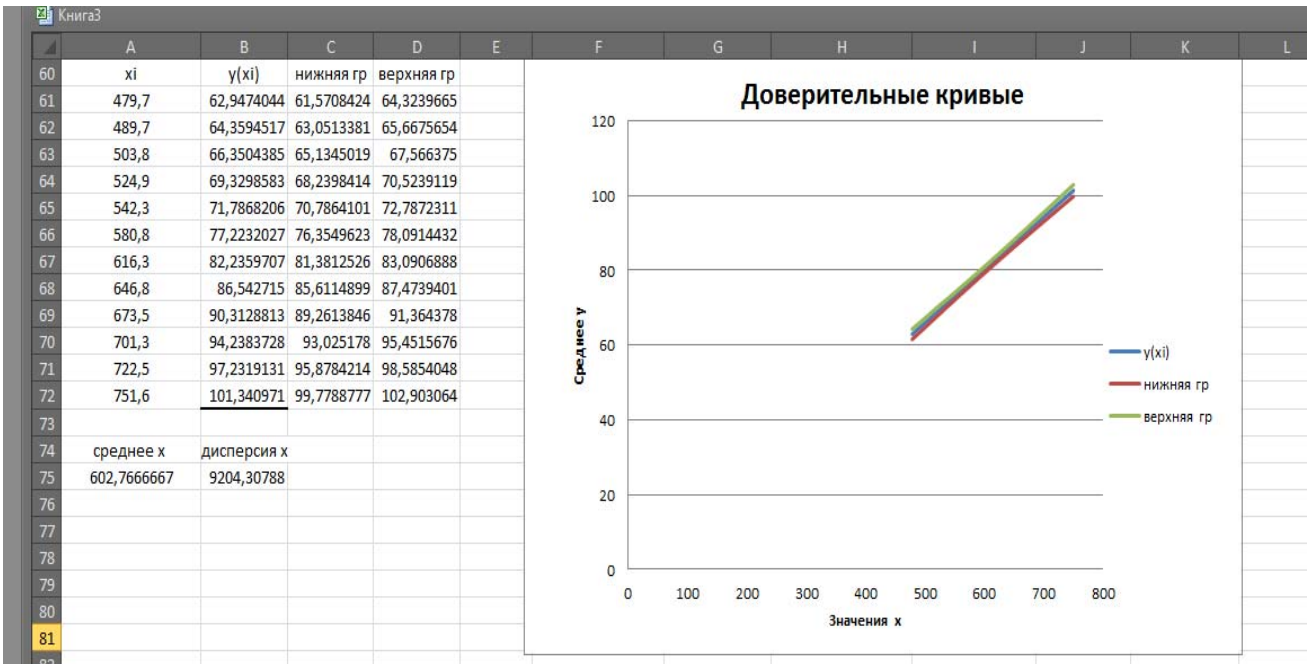


Рис. 4.5. Доверительные кривые и линия регрессии

Общее заключение о оцененной модели и ее интерпретация.

Построенная модель $\hat{y}_x = -4{,}788 + 0{,}141x$ достаточно хорошо согласуется с имеющейся выборкой. Об этом свидетельствует высокое значение коэффициента детерминации $R^2 = 0{,}9915$, т.е. 99,15% вариации совокупных расходов на жилье *Y* относительно среднего объясняется изменением располагаемого совокупного личного дохода *X*. Большое значение F-статистики, $F = 1165{,}99$, и ее уровень значимости, равный $1{,}098 \cdot 10^{-11}$, свидетельствует о наличии значимой линейной корреляционной зависимости совокупных расходов на жилье *Y* от располагаемого совокупного личного дохода *X*. Об этом также говорит значение коэффициента корреляции $r_{yx} = 0{,}9957$ и его t-статистика, $t = 34{,}147$, значительно превышающая критическое значение при заданном уровне значимости $\alpha = 0{,}05$. Оценка $s = 1{,}316$ среднеквадратического отклонения σ ошибок регрессии ε_i мала по сравнению с $\bar{y} = 80{,}25$, что свидетельствует о малом разбросе выборочных данных относительно линии регрессии.

Оценка регрессионной зависимости проводилась для значений объясняющей переменной *X* из промежутка от 479 до 752, поэтому построенная модель может быть использована для прогнозов среднего объясняемой переменной на этом промежутке и для значений *x* близких к этому промежутку.

Интерпретация модели. Согласно модели, затраты на жилье увеличиваются линейно с ростом располагаемых доходов. Отрицательность свободного члена и значительное смещение вправо от нуля промежутка наблюдаемых значений *x* исключают возможность содержательной его интерпретации. Интерпретация коэффициента регрессии $\hat{b}$: в рамках построенной модели увеличение располагаемого совокупного личного дохода на 1 млрд дол. влечет увеличение совокупных расходов на жилье в среднем на 0,141 млрд дол. в ценах 1972 г., т.е. предельный спрос на жилье по располагаемому доходу, согласно модели, равен 0,141.

Контрольные вопросы

1. Какая зависимость называется корреляционной?
2. Что описывает уравнение регрессии?
3. Запишите модель парной линейной регрессии и объясните ее компоненты.
4. Каковы источники ошибки регрессии?
5. В чем сущность метода наименьших квадратов оценивания параметров линейного уравнения регрессии?
6. Каковы предпосылки парной линейной регрессии?
7. Приведите оценки метода наименьших квадратов для параметров уравнения парной линейной регрессии.
8. Сформулируйте свойства несмещенности, состоятельности и эффективности оценок параметров.
9. Сформулируйте теорему Гаусса-Маркова.
10. В чем различие ошибок и остатков регрессии?
11. Как оценивается значимость параметров уравнения регрессии?
12. Как оценивается значимость уравнения регрессии?
13. Как связан коэффициент регрессии с коэффициентом корреляции?
14. Что характеризует коэффициент детерминации?
15. Сформулируйте нулевые гипотезы о значимости параметров уравнения регрессии. Как осуществляется проверка этих гипотез?
16. Сформулируйте понятие доверительного интервала и его надежности.
17. Как определяются доверительные интервалы для параметров уравнения парной линейной регрессии?
18. Что влияет на величину доверительного интервала прогноза среднего зависимой величины?
19. Что представляют доверительные кривые? Как они изменяются с увеличением надежности?

Лабораторная работа № 5. Нелинейная парная регрессия

Цель работы. Освоение способов перехода от нелинейной взаимосвязи зависимой и объясняющей переменной к линейной модели; освоение построения по выборочным данным нелинейной модели парной регрессии; оценка значимости построенной модели и ее прогностических свойств; оценка точности и надежности параметров модели; построение прогнозов значений зависимой переменной в MS Excel 2010. Интерпретация модели.

Краткие сведения. В парной линейной регрессии взаимосвязь наблюдаемых в выборке значений y_i зависимой переменной Y и значений x_i фактора X описывается линейной по x_i и линейной по параметрам зависимостью $y_i = a + bx_i + \varepsilon_i$**.** Она не всегда наилучшим образом отражает существующую взаимосвязь Y и X. Поэтому наряду с парной линейной регрессией рассматривают и нелинейные модели парной регрессии, в которых исследуются нелинейные взаимосвязи y_i и x_i.

Различают два класса нелинейных моделей регрессии:

- *регрессии, нелинейные относительно объясняющих переменных, но линейные относительно параметров модели*, например, $y_i = a + b\sqrt{x_i} + \varepsilon_i$, $y_i = a + bx_i + cx_i^2 + \varepsilon_i$ или $y_i = a + b\frac{1}{x_i} + \varepsilon_i$;
- *регрессии нелинейные по параметрам*, например, $y_i = a \cdot x_i^b \cdot \varepsilon_i$ или $y_i = e^{a+bx} + \varepsilon_i$.

1. ***Нелинейные по объясняющим переменным, но линейные по параметрам***, уравнения регрессии введением новых объясняющих переменных сводятся к линейным регрессионным уравнениям и оцениваются методом наименьших квадратов. Например: нелинейная модель $y_i = a + b\sqrt{x_i} + \varepsilon_i$ введением новой переменной $z_i = \sqrt{x_i}$ приводится к линейной модели парной регрессии $y_i = a + bz_i + \varepsilon_i$; нелинейная модель $y_i = a + b\frac{1}{x_i} + \varepsilon_i$ введением новой объясняющей переменной $z_i = \frac{1}{x_i}$ также приводится к линейной модели парной регрессии $y_i = a + bz_i + \varepsilon_i$; нелинейная модель $y_i = a + bx_i + cx_i^2 + \varepsilon_i$ введением новой объясняющей переменной $z_i = x_i^2$ приводится к линейной модели множественной регрессии $y_i = a + bx_i + cz_i + \varepsilon_i$. Параметры нелинейных по объясняющим переменным, но линейных по параметрам, уравнений регрессии совпадают с параметрами преобразованного линейного уравнения регрессии. Ошибки регрессии ε_i преобразованной линейной модели должны удовлетворять предпосылкам линейной регрессии. Оценки метода наименьших квадратов параметров преобразованной линейной модели являются оценками параметров исходной нелинейной модели. Например, для нелинейной модели $y_i = a + b\frac{1}{x_i} + \varepsilon_i$ оцененным уравнением регрессии будет $\hat{y}_x = \hat{a} + \hat{b} \cdot \frac{1}{x}$, где $\hat{a}$ и $\hat{b}$ оценки метода наименьших квадратов линейной модели $y_i = a + bz_i + \varepsilon_i$ с $z_i = \frac{1}{x_i}$.

2. ***Регрессии нелинейные по параметрам*** разделяются на *внутренне линейные* и *внутренне нелинейные* модели. Внутренне линейные модели с помощью соответствующих преобразований приводятся к линейному виду и затем

оцениваются методом наименьших квадратов. При этом нужно помнить, что свойства оценок параметров и модели (получаемые, например, с помощью функции «Регрессия» в MS Excel) относятся к преобразованной линейной модели, а не к исходной нелинейной модели. Внутренне нелинейные модели не могут быть приведены к линейному виду. Внутренне нелинейные модели оцениваются специальными методами. В работе рассматриваются только внутренне линейные модели.

Рассмотрим наиболее широко применяемые при моделировании социально-экономических процессов внутренне линейные модели регрессии и их преобразования к линейному виду.

Мультипликативная модель (степенная с постоянной эластичностью $E_x(y) = b$*)*

$$y_i = a \cdot x_i^b \cdot \varepsilon_i,$$

ε_i – мультипликативная случайная ошибка регрессии. Эта модель нелинейная относительно оцениваемых параметров *a* и *b*. Прологарифмировав это уравнение, получим

$$\ln y_i = \ln a + b \cdot \ln x_i + \ln \varepsilon_i.$$

Введя новые величины $y_i^* = \ln y_i$, $x_i^* = \ln x_i$, $a^* = \ln a$ и $\varepsilon_i^* = \ln \varepsilon_i$, получим линейное уравнение

$$y_i^* = a^* + b \cdot x_i^* + \varepsilon_i^*,$$

в котором ошибки регрессии ε_i^* должны удовлетворять предпосылкам линейной регрессии. Получив МНК оценки $\widehat{a^*}$ и $\hat{b}$ параметров линеаризованной модели, одновременно получаем оценку параметра *b* нелинейной модели, а оценка параметра *a* находится как $\hat{a} = e^{\widehat{a^*}}$.

Экспоненциальная модель (с постоянным темпом прироста b)

$$y_i = e^{a+bx_i} \cdot \varepsilon_i \text{ (или } y_i = A \cdot e^{bx_i} \cdot \varepsilon_i).$$

Прологарифмировав получим $\ln y_i = a + b \cdot x_i + \ln \varepsilon_i$. Введя новые величины $y_i^* = \ln y_i$ и $\varepsilon_i^* = \ln \varepsilon_i$, получим линейное уравнение

$$y_i^* = a + b \cdot x_i + \varepsilon_i^*,$$

в котором ошибки регрессии ε_i^* должны удовлетворять предпосылкам линейной регрессии. МНК оценки $\hat{a}$ и $\hat{b}$ параметров линеаризованной модели являются оценками параметров исходной нелинейной модели.

Экспоненциальная модель

$$y_i = e^{a+\frac{b}{x_i}+\varepsilon_i} \text{ (или } y_i = A \cdot e^{b/x_i+\varepsilon_i}).$$

Прологарифмировав получим $\ln y_i = a + b/x_i + \varepsilon_i$. Введя новую зависимую переменную $y_i^* = \ln y_i$ и новую объясняющую переменную $x_i^* = 1/x_i$, получим линейное уравнение

$$y_i^* = a + b \cdot x_i^* + \varepsilon_i.$$

МНК оценки $\hat{a}$ и $\hat{b}$ параметров линеаризованной модели являются оценками параметров исходной нелинейной модели.

Показательная модель (с постоянным темпом прироста равным ln*b)*

$$y_i = a \cdot b^{x_i} \cdot \varepsilon_i$$

логарифмированием приводится к виду $\ln y_i = \ln a + \ln b \cdot x_i + \ln \varepsilon_i$ и введением новых величин $y_i^* = \ln y_i$, $a^* = \ln a$, $b^* = \ln b$ и $\varepsilon_i^* = \ln \varepsilon_i$ преобразуется в линейную модель $y_i^* = a^* + b^* \cdot x_i + \varepsilon_i^*$. Ошибки регрессии ε_i^* должны удовлетворять предпосылкам линейной регрессии. По МНК оценкам $\widehat{a^*}$ и $\widehat{b^*}$ параметров линеаризованной модели оценки параметров a и b показательной модели находятся как $\hat{a} = e^{\widehat{a^*}}$ и $\hat{b} = e^{\widehat{b^*}}$.

Обратная модель

$$y_i = \frac{1}{a + b \cdot x_i + \varepsilon_i}$$

приводится к линеаризованному виду $y_i^* = a + b \cdot x_i + \varepsilon_i$ с помощью замены $y_i^* = 1/y_i$. Более сложная обратная модель (*логистическая*)

$$y_i = \frac{1}{1 + e^{a+bx_i+\varepsilon_i}}$$

приводится к линеаризованному виду $y_i^* = a + b \cdot x_i + \varepsilon_i$ с помощью замены $y_i^* = \ln(\frac{1}{y_i} - 1)$. В обеих случаях МНК оценки $\hat{a}$ и $\hat{b}$ параметров линеаризованной модели являются оценками параметров обратной модели.

Примеры внутренне нелинейных моделей (не приводимых к линейным по параметрам зависимостям): $y_i = a \cdot x_i^b + \varepsilon_i$, $y_i = a + x_i^b + \varepsilon_i$, $y_i = e^{a+bx_i} + \varepsilon_i$, $y_i = a \cdot b^{x_i} + \varepsilon_i$, $y_i = \frac{x_i}{a+b \cdot x_i+\varepsilon_i}$, $y_i = \frac{1}{1+e^{a+bx_i+\varepsilon_i}}$.

Линеаризация многофакторных нелинейных регрессионных моделей производится с использованием тех же приемов. Например, производственная функция Кобба-Дугласа $Y = A \cdot K^{\alpha} \cdot L^{\beta}$ логарифмированием приводится к виду $\ln Y = \ln A + \alpha \ln K + \beta \ln L$ и заменами $y^* = \ln Y$, $A^* = \ln A$, $K^* = \ln K$, $L^* = \ln L$ преобразуется к линейному по параметрам уравнение $y^* = A^* + \alpha K^* + \beta L^*$.

Для сопоставления различных линейных и нелинейных регрессионных моделей по их прогностическому качеству используются *индекс корреляции (корреляционное отношение)*

$$\eta = \sqrt{1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}}$$

и *средняя ошибка аппроксимации*

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%.$$

Здесь $\hat{y}_i = \varphi(x_i)$ – рассчитанные по уравнению регрессии значения зависимой переменной, $\bar{y}$ – выборочная средняя значений зависимой переменной, n – объем выборки. Индекс корреляции характеризует разброс выборочных значений относительно линии регрессии $\hat{y}_x = \varphi(x)$. $0 \le \eta \le 1$. Чем больше значе-

ние η, тем меньше разброс выборочных значений вокруг линии регрессии (тем лучше качество подгонки уравнения регрессии к выборочным данным). Если η равно или близко к нулю, то оцененная модель непригодна, она не объясняет изменение зависимой переменной изменением объясняющей переменной, т. е. построенная модель не лучше модели $\hat{y}_x = \bar{y}$. Квадрат индекса корреляции называется *коэффициентом детерминации* и обозначается R^2, т.е. $R^2 = \eta^2$. Он показывает долю вариации зависимой переменной *Y* объясняемую вариацией фактора *X* в построенной модели регрессии. Средняя ошибка аппроксимации *A* характеризует среднее относительное отклонение выборочных значений от построенной линии регрессии $\hat{y}_x = \varphi(x)$.

Проверка статистической значимости уравнения нелинейного регрессии (гипотезы $H_0: R^2 = 0$) проводится по *F*-критерию Фишера $F = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m-1}$, который имеет *F*-распределение Фишера-Снедекора с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы, где m – число коэффициентов в уравнении регрессии, а *n* – объем выборки. Оцененное уравнение нелинейной регрессии статистически незначимо, если вычисленное значение *F*-критерия меньше критического $F_{\text{кр}}(\alpha, m-1, n-m)$. Если $F > F_{\text{кр}}(\alpha, m-1, n-m)$, то оцененное уравнение нелинейной регрессии статистически значимо, т.е. влияние фактора *X* на исследуемый признак *Y* может быть описано оцененным уравнением нелинейной регрессии. $F_{\text{кр}}(\alpha, m-1, n-m)$ – квантиль уровня $(1-\alpha)$ *F*-распределения Фишера-Снедекора с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы.

Содержание лабораторной работы.

1. Ввод данных и построение диаграммы рассеяния для подбора подходящей нелинейной по фактору или внутренне линейной регрессионной модели.
2. Построение линейной модели парной регрессии и нахождение для нее средней ошибки аппроксимации.
3. Выбор нелинейной модели, и приведение ее к линейному виду, преобразование выборочных данных.
4. Оценка линеаризованной модели и ее значимости, нахождение оценок параметров нелинейной модели и запись нелинейной модели.
5. Построение прогнозов среднего зависимой переменной по оцененной нелинейной модели регрессии для выборочных значений фактора (регрессора), построение линии регрессии наложенной на диаграмму рассеяния, нахождение индекса корреляции и средней ошибки аппроксимации.
6. Проверка статистической значимости уравнения нелинейной регрессии.
7. Сравнение линейной и нелинейной регрессионных моделей.
8. Интерпретация модели и общее заключение.

Выполнение работы в MS Excel.

Построение в MS Excel парной нелинейной регрессии и сопоставление ее с линейной регрессией проведем на примере построения регрессионной зависимости себестоимости добычи единицы объема газа Y (центы) от процента жидкости в добываемом газе X. Данные наблюдений по десяти скважинам приведены в табл. 5.1.

Таблица 5.1

X	13,3	16,9	19,9	23,2	26,3	28,7	30,1	35,1	37,4	42,6
Y	3,4	5,1	4,8	6,7	6,0	6,3	9,5	9,9	11,6	13,8

Ввод данных. В ячейках **A1-A11** расположим значения X, а в ячейках **B1-B11** значения Y. Построение диаграммы рассеяния осуществляется также как в работе 4. По диаграмме рассеяния (рис. 5.1) можем предположить, что имеющиеся данные могут быть описаны линейной регрессией $y_i = a + bx_i + \varepsilon_i$ или нелинейной регрессией (например, мультипликативной или экспоненциальной моделью). На рис. 5.1 приведены также значения преобразованных переменных $x^* = lnx$ и $y^* = lny$ для мультипликативной и $x\,, y^* = lny$ для экспоненциальной моделей и диаграммы рассеяния для новых переменных.

Линейную регрессионную модель оценим, следуя работе 4. Результаты приведены на рис. 5.2. Согласно которым линейная регрессия значима и имеет вид $y = -0{,}965 + 0{,}311x$. Для вычисления средней ошибки аппроксимации $A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \cdot 100\%$ к таблице «Вывод остатков» добавим столбец для значений $\left|\frac{e_i}{y_i}\right| = \left|\frac{y_i-\hat{y}_i}{y_i}\right|$. Этот столбец занимает ячейки **D54-D64**, см. рис. 5.2. Для вычисления $\left|\frac{e_1}{y_1}\right|$ воспользуемся функцией **ABC** нахождения модуля числа в группе «**Математические**» вкладки «**Формулы**». Выделим ячейку **D55** и, учитывая расположение на листе величин e_1 и y_1, в строке формул введем **=ABS(C55/B2)**. По Enter в ячейке **D55** получим искомое значение. Копируя эту формулу в ячейки **D56-D64**, получим остальные значения $\left|\frac{e_i}{y_i}\right|$. Среднюю ошибку аппроксимации найдем, используя операцию нахождения среднего значения **СРЗНАЧ** в группе «**Статистические**» вкладки «**Формулы**». Для этого выделим ячейку **E55** и в строке формул введем **=СРЗНАЧ(D55:D64)***100. По «ОК» получим искомое значение. В примере 14,96%, см. рис. 5.2.

Выбор нелинейной модели, ее линеаризация и оценка параметров нелинейной модели. Для мультипликативной, $y_i = a \cdot x_i^b \cdot \varepsilon_i$, и экспоненциальной, $y_i = e^{a+bx_i} \cdot \varepsilon_i$, моделей проведем необходимые преобразования переменных. Мультипликативная модель сводится к линейной введением новых переменных $y_i^* = \ln y_i$ и $x_i^* = \ln x_i$; а экспоненциальная заменой $y_i^* = \ln y_i$. Значения $x_i^* = \ln x_i$ разместим в ячейках **C1-C11**, а значения $y_i^* = \ln y_i$ в ячейках **D1-D11**. Построим также диаграммы рассеяния для преобразованных переменных, близость выборочных точек к некоторой прямой свидетельствует о приемлемости рассматриваемой нелинейной регрессии, см. рис. 5.1. Оценки параметров линеаризованных моделей проведем также, как и в работе 4. Результаты приведены на рис. 5.3 и 5.4.

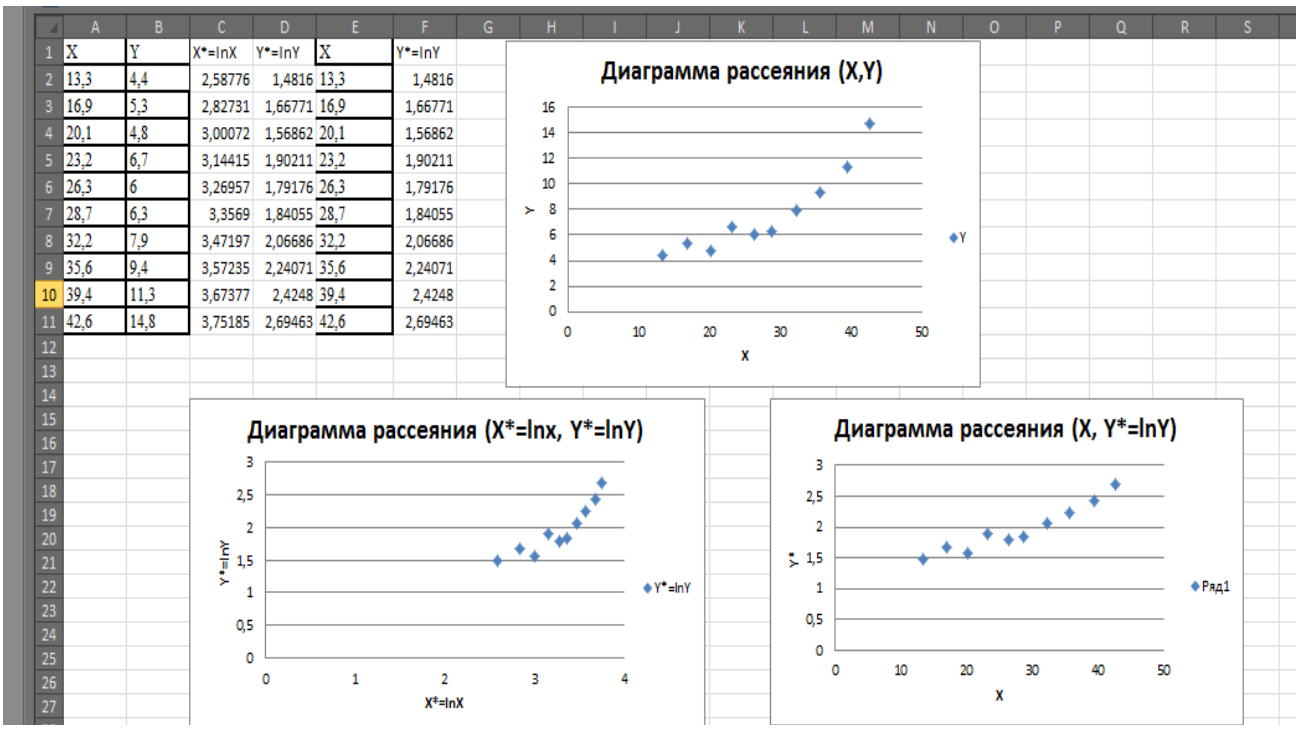


Рис. 5.1. Преобразования переменных и диаграммы рассеяния

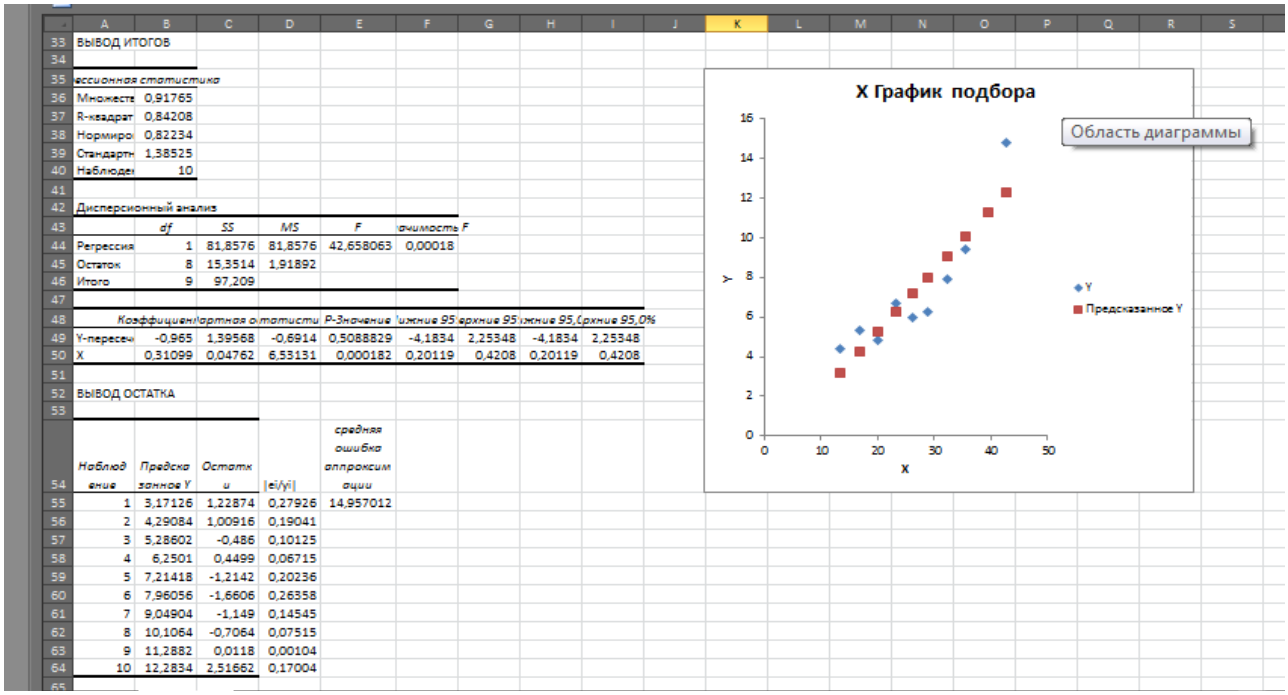


Рис. 5.2. Линейная регрессия

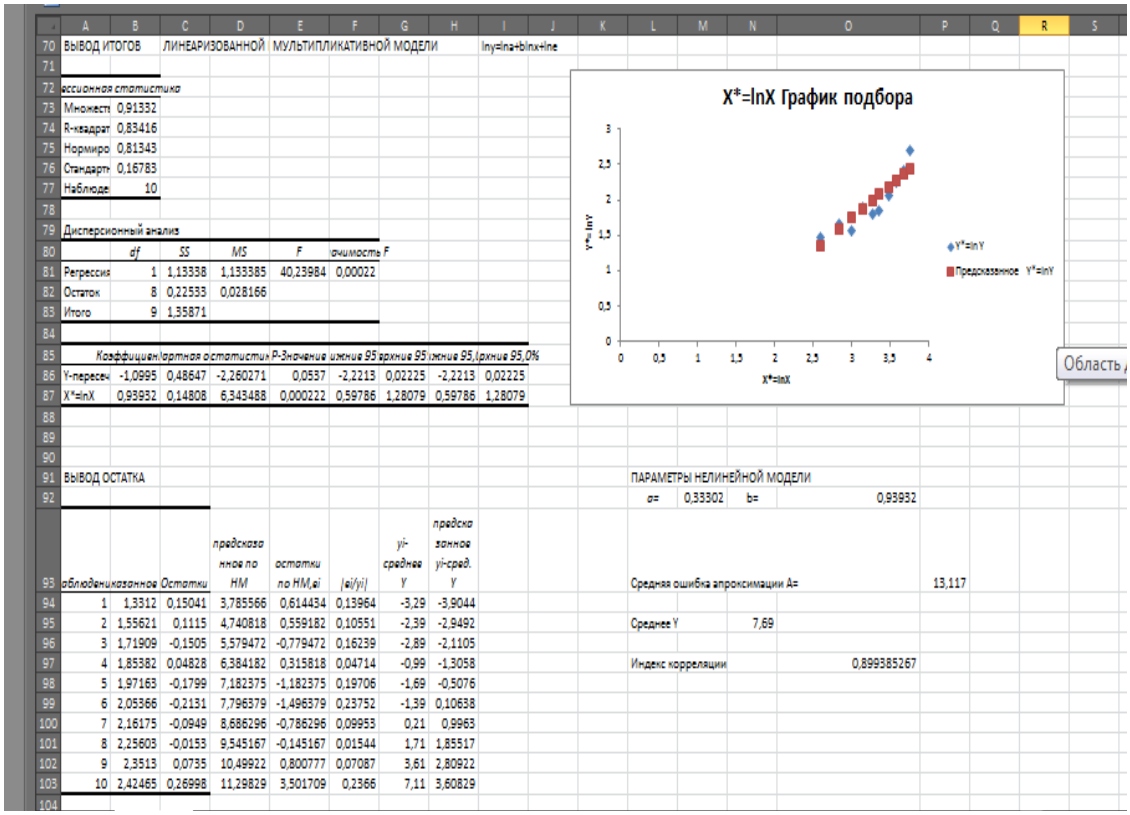
Мультипликативная модель. Результаты регрессии для линеаризованной модели $y_i^* = a^* + b \cdot x_i^*$ показывают значимость оцененного уравнения и его параметров. Для линеаризованной модели получены следующие оценки параметров: $\hat{a}^* = -1{,}0995$, $\hat{b} = 0{,}93932$. Для мультипликативной модели $y_i = a \cdot x_i^b$ найдем оценки ее параметров: для нахождения оценки параметра $a = e^{a^*}$ выделим, например, ячейку **M92** и в строке формул введем **=EXP(B86)**, в этой ячей-

ке получим искомое значение $\hat{a} = 0{,}33302$; оценка параметра b мультипликативной модели совпадает с его оценкой для линеаризованной модели. Оцененная мультипликативная модель имеет вид

$$y = 0{,}33302 \cdot x^{0{,}93932}.$$

Нахождение прогнозных значений зависимой переменной по мультипликативной модели, вычисление средней ошибки аппроксимации и индекса корреляции. К таблице «Вывод остатка», расположенной в примере в ячейках **A93-C103**, добавим следующие столбцы. В ячейках **D93-D103** столбец «предсказанное по НМ» – прогнозных значений $\hat{y}_i = 0{,}33302 \cdot x_i^{0{,}93932} = e^{y_i^*}$. В ячейках **E93-E103** столбец «остатки по НМ» – значений остатков $e_i = y_i - \hat{y}_i$ нелинейной модели. В ячейках **F93-F103** столбец $\left|\frac{e_i}{y_i}\right|$ – относительных ошибок аппроксимации. В ячейках **G93-G103** столбец «y_i – среднее Y » – отклонений выборочных значений y_i от их среднего $\bar{y}$. В ячейках **H93-H103** столбец «предсказанное y_i – сред.Y» – отклонений предсказанных значений $\hat{y}_i$ от среднего $\bar{y}$.

Вычисление прогнозного значения $\hat{y}_1 = 0{,}33302 \cdot x_1^{0{,}93932} = e^{y_1^*}$. Выделим ячейку **D94** и в строке формул введем **=EXP(B94)**. По «Enter» в этой ячейке получим искомое значение $\hat{y}_1$. Аналогично вычисляются другие прогнозные значения $\hat{y}_i$. Для вычисления остатков нелинейной модели $e_i = y_i - \hat{y}_i$ выделим ячейку **E94** и в строке формул введем **=B2-D94**, по «Enter» в этой ячейке получим искомое значение e_1. Скопировав эту формулу в ячейки **E95-E103**, получим значения других остатков.



ВЫВОД ИТОГОВ ЛИНЕАРИЗОВАННОЙ МУЛЬТИПЛИКАТИВНОЙ МОДЕЛИ lny=lna+blnx+ine

Множественный	0,91332
R-квадрат	0,83416
Нормированный	0,81343
Стандартная	0,16783
Наблюдения	10

Дисперсионный анализ

	df	SS	MS	F	значимость F
Регрессия	1	1,13338	1,133385	40,23984	0,00022
Остаток	8	0,22533	0,028166		
Итого	9	1,35871			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	нижние 95%	верхние 95%	нижние 95,0%	верхние 95,0%
Y-пересеч	-1,0995	0,48647	-2,260271	0,0537	-2,2213	0,02225	-2,2213	0,02225
X*=lnX	0,93932	0,14808	6,343488	0,000222	0,59786	1,28079	0,59786	1,28079

ВЫВОД ОСТАТКА

ПАРАМЕТРЫ НЕЛИНЕЙНОЙ МОДЕЛИ: a= 0,33302 b= 0,93932

Наблюдения	Предсказанное	Остатки	предсказанное по НМ	остатки по НМ,ei	|ei/yi|	yi-среднее Y	предсказанное yi-сред. Y
1	1,3312	0,15041	3,785566	0,614434	0,13964	-3,29	-3,9044
2	1,55621	0,1115	4,740818	0,559182	0,10551	-2,39	-2,9492
3	1,71909	-0,1505	5,579472	-0,779472	0,16239	-2,89	-2,1105
4	1,85382	0,04828	6,384182	0,315818	0,04714	-0,99	-1,3058
5	1,97163	-0,1799	7,182375	-1,182375	0,19706	-1,69	-0,5076
6	2,05366	-0,2131	7,796379	-1,496379	0,23752	-1,39	0,10638
7	2,16175	-0,0949	8,686296	-0,786296	0,09953	0,21	0,9963
8	2,25603	-0,0153	9,545167	-0,145167	0,01544	1,71	1,85517
9	2,3513	0,0735	10,49922	0,800777	0,07087	3,61	2,80922
10	2,42465	0,26998	11,29829	3,501709	0,2366	7,11	3,60829

Средняя ошибка апроксимации A= 13,117

Среднее Y 7,69

Индекс корреляции 0,899385267

Рис. 5.3. Оценка мультипликативной модели

Для вычисления относительных ошибок аппроксимации $\left|\frac{e_i}{y_i}\right|$ выделим ячейку **F94** и в строке формул введем **=ABS(E94/B2)**, по «Enter» в этой ячейке получим искомое значение $\left|\frac{e_1}{y_1}\right|$. Скопировав эту формулу в ячейки **F95-F103**, получим остальные значения $\left|\frac{e_i}{y_i}\right|$.

Для вычисления отклонений $y_i - \bar{y}$ найдем выборочное среднее $\bar{y}$. Выделив ячейку **N95** и введя в строке формул **=СРЗНАЧ(B2:B11)**, получим значение выборочного среднего $\bar{y}$, в примере равное 7,69. Выделив ячейку **G94** и введя в строке формул **=B2-7,69**, по «Enter» в этой ячейке получим $y_1 - \bar{y}$. Скопировав эту формулу в ячейки **G95-G103**, получим остальные значения $y_i - \bar{y}$. Для нахождения отклонений предсказанных значений $\hat{y}_i$ от среднего $\bar{y}$ выделим ячейку **H94** и в строке формул введем **=D94-7,69**, по Enter в этой ячейке получим $\hat{y}_1 - \bar{y}$. Скопировав эту формулу в ячейки **H95-H103**, получим остальные значения $\hat{y}_i - \bar{y}$.

Среднюю ошибку аппроксимации мультипликативной модели найдем, используя вычисленные относительные ошибки аппроксимации $\left|\frac{e_i}{y_i}\right|$ и функцию **СРЗНАЧ** вычисления выборочного среднего. В ячейку **P93** введем **=СРЗНАЧ (F94:F103)*100**. По Enter получим значение средней ошибки аппроксимации мультипликативной модели, равное в примере 13,117%.

Нахождение индекса корреляции мультипликативной модели. Выделим ячейку **O97**, в строке формул введем

=КОРЕНЬ(1-(СУММКВ(E94:E103))/СУММКВ(G94:G103)).

По «Enter» получим в этой ячейке значение индекса корреляции, равное в примере 0,8996.

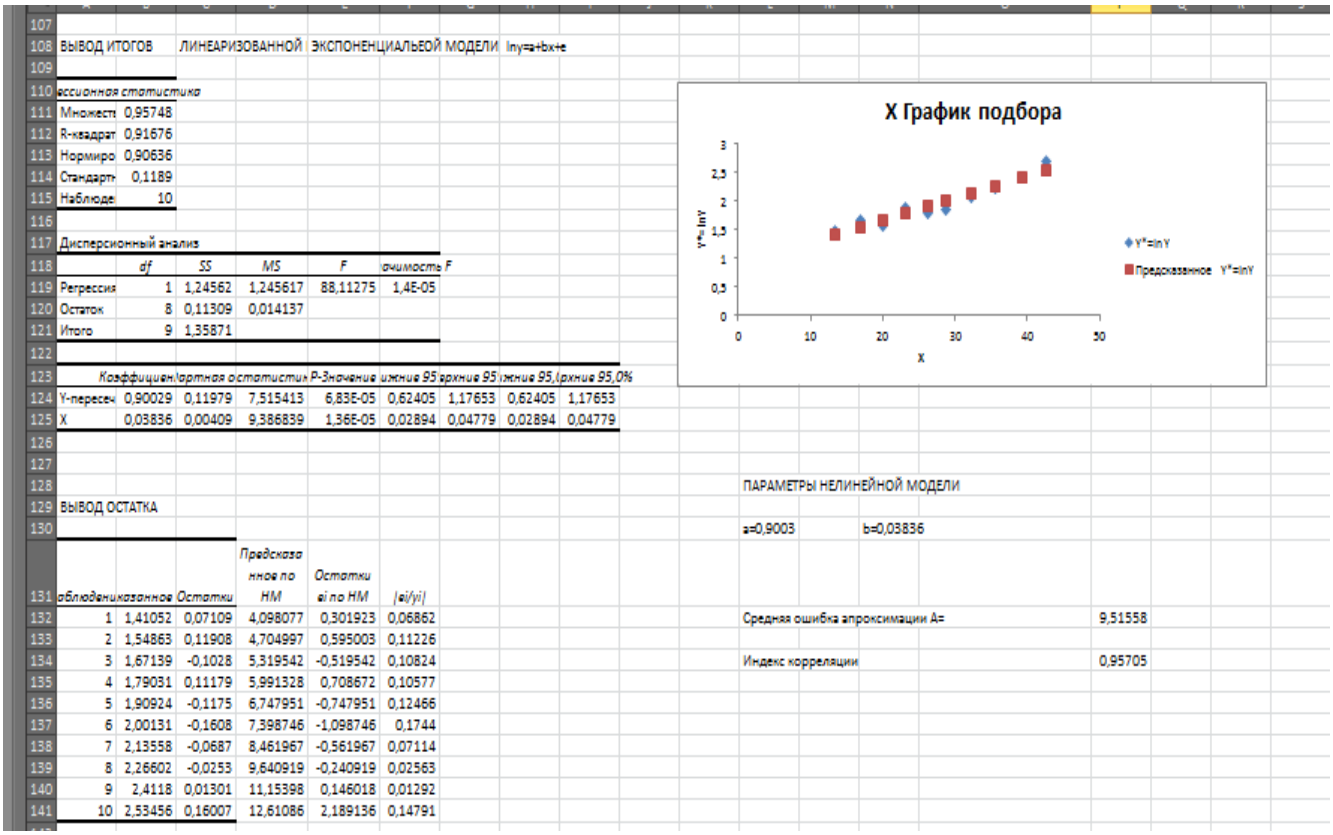
Экспоненциальная модель. Результаты регрессии для линеаризованной модели $y_i^* = a + b \cdot x_i$ приведены на рис. 5.4. Они показывают значимость оцененного уравнения и его параметров. Оценки $\hat{a} = 0{,}9003,\ \hat{b} = 0{,}03836$ параметров линеаризованной модели являются также оценками параметров экспоненциальной модели $y_i = e^{a+bx_i}$. Оцененная экспоненциальная модель имеет вид

$$y = e^{0{,}9003+0{,}03836x}.$$

Для нахождения прогнозных значений зависимой переменной по экспоненциальной модели, вычисления средней ошибки аппроксимации и индекса корреляции к таблице «Вывод остатка», расположенной в примере в ячейках **A131-C141**, добавим следующие столбцы. В ячейках **D131-D141** столбец «предсказанное по НМ» – прогнозных значений $\hat{y}_i = e^{0{,}9003+0{,}03836 \cdot x_i}$. В ячейках **E131-E141** столбец «остатки по НМ» – значений остатков $e_i = y_i - \hat{y}_i$ экспоненциальной модели. В ячейках **F131-F141** столбец $\left|\frac{e_i}{y_i}\right|$ – относительных ошибок аппроксимации.

Вычисление прогнозных значения $\hat{y}_i = e^{0{,}9003+0{,}03836 \cdot x_i}$. Выделим ячейку **D132** и введем в строке формул **=EXP(B124+B125*A2)**. По «Enter» в этой ячей-

ке получим искомое значение $\hat{y}_1$. Аналогично вычисляются другие значения $\hat{y}_i$. Для вычисления остатков регрессии нелинейной модели, $e_i = y_i - \hat{y}_i$, выделим ячейку **E132** и в строке формул введем **=B2-D132**, по «Enter» в этой ячейке получим искомое значение e_1. Скопировав эту формулу в ячейки **E133-E141**, получим значения других остатков. Для вычисления относительных ошибок аппроксимации $\left|\frac{e_i}{y_i}\right|$, выделим ячейку **F132** и в строке формул введем **=ABS (E132/B2)**, по «Enter» в этой ячейке получим искомое значение $\left|\frac{e_1}{y_1}\right|$, копируя эту формулу в ячейки **F95-F103**, получим остальные значения $\left|\frac{e_i}{y_i}\right|$.



108	ВЫВОД ИТОГОВ		ЛИНЕАРИЗОВАННОЙ	ЭКСПОНЕНЦИАЛЬЕОЙ МОДЕЛИ	lny=a+bx+e				
110	ессионная статистика								
111	Множеств	0,95748							
112	R-квадрат	0,91676							
113	Нормиро	0,90636							
114	Стандарт	0,1189							
115	Наблюде	10							
117	Дисперсионный анализ								
118		df	SS	MS	F	ачимость F			
119	Регрессия	1	1,24562	1,245617	88,11275	1,4E-05			
120	Остаток	8	0,11309	0,014137					
121	Итого	9	1,35871						
123		Коэффициен	артная остатистик	P-Значение	ижние 95	рхние 95	ижние 95,0	рхние 95,0%	
124	Y-пересеч	0,90029	0,11979	7,515413	6,83E-05	0,62405	1,17653	0,62405	1,17653
125	X	0,03836	0,00409	9,386839	1,36E-05	0,02894	0,04779	0,02894	0,04779
129	ВЫВОД ОСТАТКА								
131	аблюдениказанное	Остатки	Предсказанное по НМ	Остатки ei по НМ	|ei/yi|				
132	1	1,41052	0,07109	4,098077	0,301923	0,06862			
133	2	1,54863	0,11908	4,704997	0,595003	0,11226			
134	3	1,67139	-0,1028	5,319542	-0,519542	0,10824			
135	4	1,79031	0,11179	5,991328	0,708672	0,10577			
136	5	1,90924	-0,1175	6,747951	-0,747951	0,12466			
137	6	2,00131	-0,1608	7,398746	-1,098746	0,1744			
138	7	2,13558	-0,0687	8,461967	-0,561967	0,07114			
139	8	2,26602	-0,0253	9,640919	-0,240919	0,02563			
140	9	2,4118	0,01301	11,15398	0,146018	0,01292			
141	10	2,53456	0,16007	12,61086	2,189136	0,14791			

ПАРАМЕТРЫ НЕЛИНЕЙНОЙ МОДЕЛИ

a=0,9003 b=0,03836

Средняя ошибка апроксимации A= 9,51558

Индекс корреляции 0,95705

Рис. 5.4. Оценка экспоненциальной модели

Среднюю ошибку аппроксимации экспоненциальной модели найдем, используя вычисленные относительные ошибки аппроксимации $\left|\frac{e_i}{y_i}\right|$. Выделим ячейку **P132** и в строке формул введем **=СРЗНАЧ(F132:F141)*100**. По «Enter» получим значение средней ошибки аппроксимации экспоненциальной модели, равное в примере 9,516 %.

Нахождение индекса корреляции экспоненциальной модели. Выделим ячейку **P97** и в строке формул введем

=КОРЕНЬ(1-(СУММКВ(E132:E141))/СУММКВ(G94:G103)).

По «Enter» получим в этой ячейке значение индекса корреляции, равное в примере 0,957.

Построение графиков нелинейной регрессии и диаграммы рассеяния. Для совмещения диаграммы рассеяния с графиками мультипликативной и экспоненциальной регрессий скопируем в ячейки **A146-156** выборочные значения x_i, в ячейки **B146-156** выборочные значения y_i, в ячейки **C146-156** вычисленные по мультипликативной модели значения зависимой переменной $\hat{y}_i = 0{,}33302 \cdot x_i^{0{,}93932}$, в ячейки **D146-156** вычисленные по экспоненциальной модели значения зависимой переменной $\hat{y}_i = e^{0{,}9003+0{,}03836 \cdot x_i}$. В вкладке «**Вставка**» в группе «**Диаграммы**» выберем вид «**Точечная**», в группе «**Макеты диаграмм**» выберем «**Макет 1**», в группе «**Данные**» выберем «**Выбрать данные**». В открывшемся окне «**Выбор источника данных**» в поле «**Диапазон данных для диаграммы**» укажем ячейки с данными **A147:D156.** По «**ОК**» получим необходимую диаграмму и графики, см. рис. 5.5. Средствами MS Excel проводится корректировка заголовка и надписей осей на диаграмме.

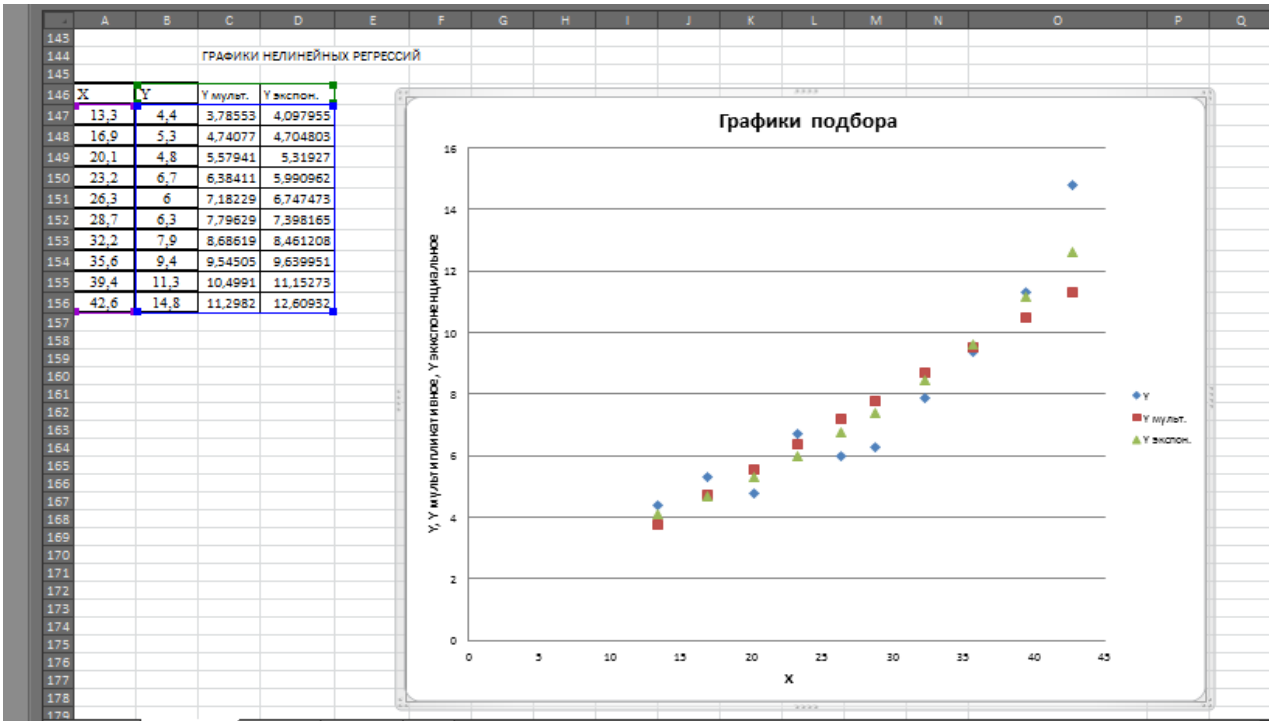


Рис. 5.5. Графики нелинейных регрессий и диаграмма рассеяния

Интерпретация модели и общее заключение. Построенная мультипликативная модель $y = 0{,}33302 \cdot x^{0{,}93932}$ значима и согласуется с выборочными данными. Об этом свидетельствуют значение индекса корреляции $\eta = 0{,}8993$, средняя ошибка аппроксимации $A = 13{,}117\%$ и значение F-критерия, $F = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m-1}$, равное 33,7 и превышающее $F_{\text{кр}}(0{,}05; 1; 8) = 5{,}32$. Коэффициент детерминации $R^2 = \eta^2 = 0{,}8082$, т.е. 80,8% вариации себестоимости добычи единицы объема газа объясняется в этой модели вариацией процента жидкости в добываемом газе. Среднее относительное отклонение выборочных данных от линии регрессии составляет 13,12%. Мультипликативная модель обладает постоянной эластичностью, равной параметру b. В построенной мультипликативной модели

b=0,9393, следовательно, однопроцентное увеличение содержания жидкости приводит в среднем к увеличению себестоимости добычи газа на 0,9393%.

Экспоненциальная модель $y = e^{0,9003+0,03836 \cdot x}$ также значима и лучше, чем мультипликативная модель, согласуется с выборочными данными. Об этом свидетельствуют значение индекса корреляции $\eta = 0,957$, средняя ошибка аппроксимации $A = 9,526\%$ и значение F-критерия, $F = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m-1}$, равное 87 и превышающее $F_{кр}(0,05; 1; 8) = 5,32$. Коэффициент детерминации $R^2 = \eta^2 = 0,9158$, т.е. 91,58% вариации себестоимости добычи единицы объема газа объясняется в экспоненциальной модели вариацией процента жидкости в добываемом газе. Среднее относительное отклонение выборочных данных от линии регрессии составляет 9,53%, что является приемлемой ошибкой аппроксимации. Экспоненциальная модель обладает постоянным темпом прироста, равным параметру *b*. В построенной экспоненциальной модели *b=0,03836*, следовательно, увеличение содержания жидкости в добываемом газе на 1% приводит к увеличению себестоимости добычи газа на 3,836%.

Линейная регрессионная модель $y = -0,965 + 0,311 \cdot x$ дает среднюю ошибка аппроксимации 14,96% и коэффициент детерминации $R^2 = 0,842$, т.е. линейная модель объясняет только 84,2% вариации себестоимости добычи газа вариацией процента содержания в нем жидкости.

Таким образом, из рассмотренных регрессионных зависимостей лучшими свойствами обладает экспоненциальная модель $y = e^{0,9003+0,03836 \cdot x}$.

Контрольные вопросы

1. В чем отличие регрессионных моделей нелинейных только по факторам от нелинейных по параметрам?
2. В чем отличие внутренне линейных и внутренне нелинейных регрессионных моделей?
3. Приведите примеры моделей нелинейных по объясняющей переменной, но линейных по параметрам.
4. Приведите примеры внутренне линейных моделей. Как осуществляется их линеаризация?
5. Как осуществляется линеаризация логистической модели?
6. Как оцениваются параметры внутренне линейных моделей?
7. Приведите примеры внутренне нелинейных моделей.
8. Какие показатели корреляции используются при анализе нелинейных взаимосвязей?
9. Как определяется и что характеризует средняя ошибка аппроксимации?
10. Как определяется и что характеризует индекс корреляции?
11. По каким критериям отбираются нелинейные регрессионные модели?
12. Как определяется и что характеризует коэффициент эластичности?
13. Как определяется темп прироста и что он характеризует?
14. С помощью какого критерия проверяется значимость нелинейного уравнения регрессии?

Лабораторная работа № 6. Множественная регрессия

Цель работы. Освоение построения по выборочным данным модели множественной линейной регрессии, оценки точности и надежности параметров и всей модели, построения прогнозов значений зависимой переменной в MS Excel 2010. Интерпретация модели.

Краткие сведения. Модель множественной регрессии описывает зависимость условного среднего M_xY зависимой случайной величины Y в виде функции значений $x_2,\ x_3, \dots,\ x_p$ нескольких объясняющих переменных (факторов) $X_2, X_3, \dots, X_p$:

$$M_xY = f(x_2, x_3, \dots, x_p, b),$$

где b – вектор параметров модели, который оценивается по выборочным данным $(x_{12}, x_{13}, \dots, x_{1p}, y_1), (x_{22}, x_{23}, \dots, x_{2p}, y_2), \dots, (x_{n2}, x_{n3}, \dots, x_{np}, y_n)$. Здесь x_{ij} – значение j-го фактора в i-ом измерении. Чаще всего при построении множественной регрессии рассматриваются следующие уравнения:

- линейное по факторам и параметрам $y_i = b_1 + b_2x_{i2} + b_2x_{i3} + \dots + b_px_{ip} + \varepsilon_i$;
- уравнение линейное по параметрам и нелинейное по факторам $y_i = b_1 + b_2\varphi_1(x_{i2}, x_{i3}, \dots, x_{ip}) + b_2\varphi_2(x_{i2}, x_{i3}, \dots, x_{ip}) + \dots + b_2\varphi_k(x_{i2}, x_{i3}, \dots, x_{ip)} + \varepsilon_i$;
- степенное уравнение $y_i = b_1 \cdot x_{i2}^{b_2} \cdot x_{i3}^{b_3} \cdot \dots \cdot x_{ip}^{b_p} \cdot \varepsilon_i$;
- экспоненциальное уравнение $y_i = e^{b_1 + b_2x_{i2} + b_3x_{i3} + \dots + b_px_{ip} + \varepsilon_i}$;
- гиперболическое уравнение $y_i = \frac{1}{b_1 + b_2x_{i2} + b_3x_{i3} + \dots + b_px_{ip} + \varepsilon_i}$.

Случайная величина ε_i, называемая *ошибкой регрессии*, отражает влияние пропущенных объясняющих переменных, неправильной структуры и функциональной спецификации модели, агрегирования переменных, ошибки измерений.

Нелинейные многофакторные уравнения линеаризуются такими же преобразованиями, что и соответствующие нелинейные однофакторные уравнения, см. работу №5.

В данной работе рассмотрим только линейную по параметрам и факторам модель множественной регрессии

$$y_i = b_1 + b_2x_{i2} + b_2x_{i3} + \dots + b_px_{ip} + \varepsilon_i,\ i = 1,2, \dots, n, \qquad (6.1)$$

в которой значения y_i зависимой переменной представлены в виде суммы детерминированной $b_1 + b_2x_{i2} + b_3x_{i3} + \dots + b_px_{ip}$ и случайной ε_i составляющих.

Оценки параметров модели (8) находятся *методом наименьших квадратов*, из условия минимизации суммы квадратов остатков регрессии $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, где $\hat{y}_i = b_1 + b_2x_{i2} + b_2x_{i3} + \dots + b_px_{ip}$ – вычисляемые по уравнению регрессии (прогнозные) значения зависимой переменной y. Введем следующие векторы и матрицу:

$X_1 = \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}, X_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \cdots \\ x_{n2} \end{pmatrix}, X_3 = \begin{pmatrix} x_{13} \\ x_{23} \\ \cdots \\ x_{n3} \end{pmatrix}, \cdots, X_p = \begin{pmatrix} x_{1p} \\ x_{2p} \\ \cdots \\ x_{np} \end{pmatrix}$ – векторы наблюдаемых значений факторов; $\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}$ – вектор наблюдаемых значений зависимой переменной; $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}$ – вектор ошибок регрессии; $\boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_p \end{pmatrix}$ – вектор параметров уравнения регрессии; $\boldsymbol{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1p} \\ 1 & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix}$ – матрица значений факторов.

Используя вектора $\boldsymbol{Y}, \boldsymbol{\varepsilon}, \boldsymbol{b}$ и матрицу $\boldsymbol{X}$, линейные зависимости (6.1) выборочных значений зависимой переменной от значений факторов можно представить в виде

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{b} + \boldsymbol{\varepsilon} . \qquad (6.2)$$

Вектор X_1 и соответствующий ему первый столбец из «1» в матрице X учитывает наличие в уравнении регрессии (6.1) свободного члена b_1.

Основные предпосылки линейной множественной регрессии.

1. Связь значений y_i зависимой величины от значений $x_{i2}, x_{i3}, \dots, x_{ip}$ факторов задается соотношением (6.1) или в матричной форме (6.2). (Эта зависимость называется *спецификацией модели*).

2. $x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}$ – детерминированные величины, векторы $X_1, X_2, X_3, \dots, X_p$ линейно независимы, т.е. матрица $\boldsymbol{X}$ детерминированная и ее ранг равен p.

3. Ошибки регрессии ε_i – случайные величины с $M\varepsilon_i = 0$ и $D\varepsilon_i = \sigma^2$ для всех i, т.е. ошибки регрессии не имеют систематической составляющей и имеют одинаковую дисперсию.

4. Ошибки регрессии ε_i и ε_j (или величины y_i и y_j) не коррелированы в разных наблюдениях, т.е. $Cov(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i, \varepsilon_j) = 0$.

5. Ошибки регрессии ε_i распределены по нормальному закону с нулевой средней и дисперсией σ^2, т.е. $\varepsilon_i \sim N(0, \sigma^2)$.

При выполнении этих предпосылок модель (6.1) называется *классической нормальной регрессией*. Эта модель множественной линейной регрессии содержит p неизвестных параметров регрессии $b_1, b_2, \dots, b_p$ и неизвестную дисперсию σ^2 ошибок регрессии ε_i. Оценка $\widehat{\boldsymbol{b}}$ вектора параметров регрессии находятся из условия минимизации по $b_1, b_2, \dots, b_p$ суммы квадратов остатков $e_i = y_i - \hat{y}_i$, т.е. величины

$$\boldsymbol{S}^2 = \boldsymbol{e}^{\boldsymbol{T}} \cdot \boldsymbol{e} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - b_1 - b_2 x_{i2} - b_3 x_{i3} - \cdots - b_p x_{ip}\right)^2,$$

где $\boldsymbol{e^T} = (e_1, e_2, \dots, e_n)$ – вектор остатков регрессии. Эти оценки $\widehat{\boldsymbol{b}}$ называются *оценками метода наименьших квадратов* и определяются соотношением

$$\widehat{\boldsymbol{b}} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \cdots \\ \hat{b}_p \end{pmatrix} = (\boldsymbol{X}^T \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}^T \cdot \boldsymbol{Y}. \tag{6.3}$$

Теорема Гаусса-Маркова. При выполнении предпосылок 1-4 оценка (6.3) метода наименьших квадратов обладают наименьшей дисперсией в классе линейных по $\boldsymbol{Y}$ несмещенных оценок.

Несмещенной оценкой дисперсии σ^2 ошибок регрессии ε_i является величина $s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$, где $e_i = y_i - \hat{y}_i$ – остатки регрессии.

При выполнении предпосылок 1-5 оценки $\hat{b}_i$ параметров уравнения множественной линейной регрессии имеют нормальное распределение со средним b_i и дисперсией $\sigma^2(\boldsymbol{X}^T \cdot \boldsymbol{X})_{ii}^{-1}$, где $(\boldsymbol{X}^T \cdot \boldsymbol{X})_{ii}^{-1}$ – *i*-й элемент главной диагонали матрицы $(\boldsymbol{X}^T \cdot \boldsymbol{X})^{-1}$, *i=1,2,…,p*. *Оценки* $\widehat{Var}(\hat{b}_i)$ *дисперсий оценок* $\hat{b}_i$ параметров уравнения множественной линейной регрессии определяются как

$$\widehat{Var}(\hat{b}_i) = s^2(\boldsymbol{X}^T \cdot \boldsymbol{X})_{ii}^{-1}. \tag{6.4}$$

Стандартные отклонения s_{b_i} оценок коэффициентов b_i уравнения регрессии определяются соотношениями $s_{b_i} = \sqrt{\widehat{Var}(\hat{b}_i)}$, *i=1,2,…,p*.

Доверительные интервалы надежности $\gamma = 1 - \alpha$ для значений параметров b_i уравнения (6.1) имеют вид

$$\hat{b}_i - t_\gamma s_{b_i} < b_i < \hat{b}_i + t_\gamma s_{b_i}, \tag{6.5}$$

где $t_\gamma = t\left(\frac{1+\gamma}{2}, n-p\right)$ – квантиль уровня $\frac{1+\gamma}{2}$ распределения Стьюдента с числом степеней свободы $n-p$. Доверительный интервал надежности γ для дисперсии σ^2 ошибок регрессии определяется неравенством

$$\frac{(n-p)s^2}{U_2} < \sigma^2 < \frac{(n-p)s^2}{U_1},$$

где $U_1 = \chi^2\left(\frac{1-\gamma}{2}, n-p\right)$ и $U_2 = \chi^2\left(\frac{1+\gamma}{2}, n-p\right)$ – квантили соответственно уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$ распределения χ^2 с числом степеней свободы $n-p$.

Проверка значимости влияния факторов $X_2, X_3, \dots, X_p$ *на зависимую величину* Y на уровне значимости α, при выполнении предпосылок 1 – 5, осуществляется также, как и для парной линейной регрессии, проверкой гипотез H_0: $b_i = b_{i0}$, с $b_{i0} = 0$, с помощью *t*-статистики $t = \frac{\hat{b}_i - b_{i0}}{s_{b_i}}$. Если вычисленное (при $b_{i0} = 0$) значение *t*-статистики $t = \frac{\hat{b}_i}{s_{b_i}}$ по модулю больше критического значения $t_\gamma = t(1 - \frac{\alpha}{2}, n-p)$, то нулевая гипотеза отклоняется на уровне значимости $\alpha = 1 - \gamma$, т.е. влияние фактора X_i статистически значимо. При $\left|\frac{\hat{b}_i}{s_{b_i}}\right| < t(1 -$

$\frac{\alpha}{2}, n-p$) нулевая гипотеза принимается и влияние фактора X_i на зависимую величину Y статистически незначимо. Следует отметить, гипотеза $H_0: b_i = b_{i0}$ принимается на уровне значимости $\alpha = 1-\gamma$, если соответствующий доверительный интервал (6.5) надежности γ покрывает значение b_{i0}. Так, если доверительный интервал (6.5) покрывает нуль, то влияние фактора X_i на зависимую величину Y статистически незначимо.

Качество модели множественной линейной регрессии (6.1), как и в случае парной линейной регрессии, оценивается с помощью дисперсионного анализа и коэффициента детерминации.

Дисперсионный анализ во множественной регрессии. При наличии в модели свободного члена, в уравнении регрессии (6.1) присутствует свободный член b_1, общая сумма квадратов отклонений выборочных значений y_i от их общего среднего $\bar{y}$, $SS_{\text{общ.}} = \sum(y_i - \bar{y})^2$, разлагается на сумму квадратов отклонений, объясняемых регрессией, $SS_R = \sum(\hat{y}_i - \bar{y})^2$, и остаточную сумму квадратов отклонений $SS_{\text{ост.}} = \sum(y_i - \hat{y}_i)^2$. $SS_{\text{общ.}} = SS_R + SS_{\text{ост.}}$. Гипотеза об отсутствии линейной зависимости Y от факторов $X_2, X_3, \dots, X_p$ имеет вид

$$H_0: b_2 = b_3 = \ \dots = b_p = 0.$$

При справедливости нулевой гипотезы H_0 средние квадраты $MS_R = \frac{SS_R}{p-1}$ и $MS_{\text{ост.}} = \frac{SS_{\text{ост.}}}{n-p}$ являются независимыми несмещенными оценками одной и той же генеральной дисперсии σ^2 зависимой переменной Y и их различие статистически незначимо. Проверка нулевой гипотезы H_0, при уровне значимости α, сводится к проверке существенности различия двух несмещенных выборочных оценок $MS_{\text{ост}}$ и MS_R дисперсии σ^2 с помощью F-критерия $F = \frac{MS_R}{MS_{\text{ост}}}$, который имеет F-распределение Фишера-Снедекора с $k_1 = p-1$ и $k_2 = n-p$ степенями свободы, где p число коэффициентов в уравнении регрессии (6.1). Если вычисленное значение статистики $F = \frac{MS_R}{MS_{\text{ост}}}$ меньше критического $F_{\text{кр}}(\alpha, p-1, n-p)$, то гипотеза H_0 об отсутствии влияния факторов $X_2, X_3, \dots, X_p$ на исследуемый признак Y принимается, т.е. оцененное линейное регрессионное уравнение незначимо.

Если $F = \frac{MS_R}{MS_{\text{ост}}} > F_{\text{кр}}(\alpha, p-1, n-p)$, то гипотеза H_0 отвергается, т.е. факторы $X_2, X_3, \dots, X_p$ *оказывают влияние на исследуемый признак Y и оцененное уравнение регрессии* $\hat{y} = \hat{b}_1 + \hat{b}_2 x_2 + \hat{b}_2 x_3 + \cdots + \hat{b}_p x_p$ значимо. Здесь $F_{\text{кр}}(\alpha, p-1, n-p)$ – квантиль уровня $(1-\alpha)$ F-распределения Фишера-Снедекора с $k_1 = p-1$ и $k_2 = n-p$ степенями свободы.

Качество построенной регрессионной модели характеризуется *коэффициентом детерминации* $R^2 = 1 - \frac{SS_{\text{ост.}}}{SS_{\text{общ.}}}$. Он показывает, долю вариации зависимой переменной относительно ее средней объясняемую вариацией факторов $X_2, X_3, \dots, X_p$. $0 \le R^2 \le 1$.

F-критерий и коэффициент детерминации R^2 связаны равенством

$$F = \frac{n-p}{p-1} \cdot \frac{R^2}{1-R^2}.$$

Если $R^2 = 0$, то построенная регрессия $\hat{y} = \hat{b}_1 + \hat{b}_2 x_2 + \hat{b}_2 x_3 + \cdots + \hat{b}_p x_p$ не улучшает качество предсказания y по сравнению с тривиальным предсказанием $y = \bar{y}$. При $R^2 = 1$ имеет место точная подгонка, все выборочные данные $(x_{i2}, x_{i3}, \ldots, x_{ip}, y_i)$ лежат на плоскости регрессии

$$\hat{y} = \hat{b}_1 + \hat{b}_2 x_2 + \hat{b}_3 x_3 + \cdots + \hat{b}_p x_p.$$

Качество подгонки уравнения регрессии к выборочным данным и его прогностическое свойство характеризуется также *средней ошибкой аппроксимации*

$$A = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%.$$

При неизменном объеме выборки коэффициент детерминации R^2 в общем случае возрастает с увеличением числа факторов. При числе факторов p равном числу выборочных данных n выполняется равенство $R^2 = 1$. Но при этом полученное уравнение регрессии не будет иметь содержательной экономической интерпретации. Рекомендуется, чтобы объем выборки превышал число параметров модели не менее чем в пять раз. Для устранения эффекта роста коэффициента детерминации с ростом числа факторов в модели вводится *скорректированный коэффициент детерминации*

$$\bar{R}^2 = 1 - \frac{MS_{\text{ост.}}}{MS_{\text{общ.}}} = 1 - \frac{SS_{\text{ост}}/(n-p)}{SS_{\text{общ}}/n-1}.$$

Скорректированный коэффициент детерминации $\bar{R}^2$ связан с коэффициентом детерминации R^2 соотношением $\bar{R}^2 = R^2 - \frac{p-1}{n-p} \cdot (1 - R^2)$ и удовлетворяет неравенству $\bar{R}^2 \leq R^2$. С ростом числа факторов $\bar{R}^2$ уменьшается относительно R^2.

Интерпретация коэффициентов уравнения множественной линейной регрессии. Множественный регрессионный анализ позволяет приближенно оценить влияние каждого фактора на зависимую переменную, допуская при этом коррелированность факторов. Оценка $\hat{b}_i$ коэффициента регрессии при i-м факторе X_i показывает, на сколько единиц приближенно изменится среднее зависимой переменной при изменении этого фактора на одну единицу при неизменных значениях других факторов. Относительное изменение зависимой переменной, вызванное изменением фактора X_i, характеризуется *частным коэффициентом эластичности*, который для линейной модели определяется как

$$E_{x_i}(y) = \hat{b}_i \cdot \frac{\bar{x}_i}{\bar{y}},$$

где $\bar{x}_i$ – выборочное среднее значений фактора X_i, $\bar{y}$ – общее среднее зависимой переменной. Частный коэффициент эластичности $E_{x_i}(y)$ показывает приближенно, на сколько процентов относительно среднего $\bar{y}$ измениться зависимая переменная y при изменении значения фактора X_i на 1% относительно его среднего $\bar{x}_i$.

Точечный и интервальный прогноз по модели множественной регрессии. Точечный прогноз $\hat{y}_x$ среднего зависимой величины Y для заданных значений $x_2, x_3, \dots, x_p$ факторов $X_2, X_3, \dots, X_p$ вычисляется непосредственно по оцененному уравнению регрессии

$$\hat{y}_x = \hat{b}_1 + \hat{b}_2 x_2 + \hat{b}_3 x_3 + \cdots + \hat{b}_p x_p.$$

При выполнении предпосылок 1 – 5 эта оценка имеет наименьшую дисперсию в классе линейных по Y несмещенных оценок. Оценка $\widehat{D}(\hat{y}_x)$ дисперсии прогноза $\hat{y}_x$ для заданного вектора $x = (1, x_2, x_3, \dots, x_p)$ значений факторов определяется как

$$\widehat{D}(\hat{y}_x) = s^2(1 + x(\boldsymbol{X^T X})^{-1} x^T).$$

Доверительный интервал надежности γ для среднего y_x зависимой величины при заданном векторе $x = (1, x_2, x_3, \dots, x_p)$ значений факторов определяется неравенством

$$\hat{y}_x - t_\gamma \sqrt{\widehat{D}(\hat{y}_x)} < y_x < \hat{y}_x + t_\gamma \sqrt{\widehat{D}(\hat{y}_x)},$$

где $t_\gamma = t\left(\frac{1+\gamma}{2}, n-p\right)$ – квантиль уровня $\frac{1+\gamma}{2}$ распределения Стьюдента с числом степеней свободы $n - p$.

Содержание лабораторной работы.

1. Ввести выборочные данные.
2. Построить корреляционную матрицу.
3. Оценить параметры уравнения множественной линейной регрессии.
4. Проверить значимость коэффициентов уравнения регрессии и самого уравнения регрессии при уровне значимости $\alpha = 0{,}05$.
5. Оценить качество построенной модели.
6. Построить точечный и интервальный, надежности $\gamma = 0{,}95$, прогнозы среднего зависимой переменной для значений факторов равных их выборочным средним, т.е. для $x_2 = \bar{x}_2,\ x_3 = \bar{x}_3, \dots,\ x_p = \bar{x}_p$.
7. Дать общее заключение об оцененной модели и ее интерпретацию.

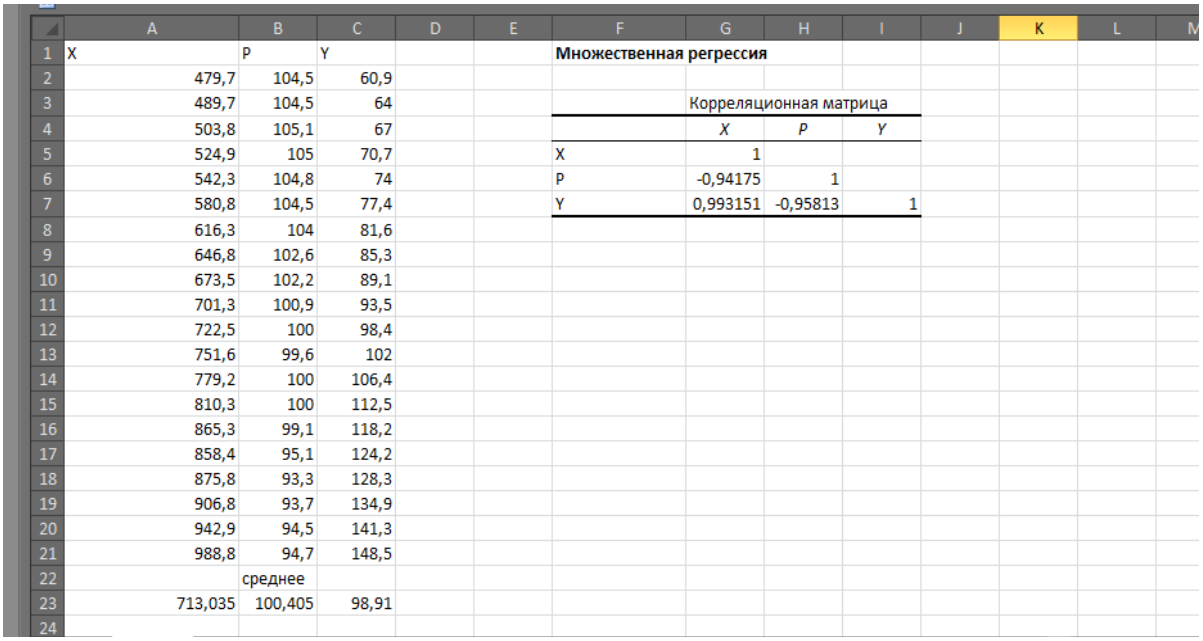
Выполнение работы в MS Excel.

Порядок выполнения работы рассмотрим на примере построения линейной регрессионной зависимости расходов на жилье (Y, млрд. дол.) от располагаемого личного дохода (X, млрд. дол.) и индекса реальных цен (P) относительно 1972 г. по данным США за 1959–1978 г. Эти данные приведены на рис. 6.1.

Ввод данных. В ячейках **A1-A21** расположим имя фактора X (располагаемый личный доход) и его значения, в ячейках **B1-B21** имя фактора P (индекса реальных цен) и его значения, в ячейках **C1-C21** имя зависимой переменной Y (расходы на жилье) и его значения.

Построение корреляционной матрицы. Следуя работе №2, построим корреляционную матрицу для величин X, P, Y. Расположим ее в ячейках **G5-I7**, см. рис. 6.1. Парные коэффициенты корреляции $r_{yx} = 0{,}993$, $r_{yp} = -0{,}958$, $r_{px} = -0{,}942$

говорят о тесной парной линейной корреляционной зависимости рассматриваемых величин.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	X	P	Y			Множественная регрессия							
2	479,7	104,5	60,9										
3	489,7	104,5	64				Корреляционная матрица						
4	503,8	105,1	67				*X*	*P*	*Y*				
5	524,9	105	70,7			X	1						
6	542,3	104,8	74			P	-0,94175	1					
7	580,8	104,5	77,4			Y	0,993151	-0,95813	1				
8	616,3	104	81,6										
9	646,8	102,6	85,3										
10	673,5	102,2	89,1										
11	701,3	100,9	93,5										
12	722,5	100	98,4										
13	751,6	99,6	102										
14	779,2	100	106,4										
15	810,3	100	112,5										
16	865,3	99,1	118,2										
17	858,4	95,1	124,2										
18	875,8	93,3	128,3										
19	906,8	93,7	134,9										
20	942,9	94,5	141,3										
21	988,8	94,7	148,5										
22		среднее											
23	713,035	100,405	98,91										
24													

Рис. 6.1. Множественная регрессия

Оценка уравнения множественной линейной регрессии $y = b_1 + b_2 x + b_3 p$. Откроем вкладку «**Данные**», в группе «**Анализ**» выберем надстройку «**Анализ данных**». В открывшемся окне «**Инструменты анализа**» выберем функцию «**Регрессия**». В появившемся окне «Регрессия» укажем входные данные для оценки параметров уравнения регрессии, выводимые результаты и их расположение. Заполнение окна «Регрессия» для рассматриваемого примера приведено на рис. 6.2. В части «Входные данные» в поле ввода «Входной интервал Y» указываем диапазон ячеек, содержащий значения зависимой переменной, в нашем примере это **C1:C21.** В поле ввода «Входной интервал X» – диапазон ячеек, содержащий значения независимых переменных, в примере это **A1:B21**. Значения объясняющих переменных должны располагаться в последовательных столбцах. В поле «Метки» устанавливаем флажок ν, он указывает на то, что первые строки диапазонов данных содержат имена этих данных (заголовки). В «Константа-ноль» флажок не устанавливаем. При установке флажка ν в левом поле «Уровень надежности», наряду с используемым по умолчанию стандартным уровнем надежности 95% ($\gamma = 0{,}95$), можно задать и другое его значение, в этом случае будут выведены интервальные оценки параметров регрессии для двух уровней надежности.

В части «Параметры вывода» выбираем «Выходной интервал» – для помещения результатов на текущем рабочем листе, положение результатов на листе указываем заданием верхней левой ячейки, начиная с которой располагаются результаты, в нашем примере выбрана ячейка **A25**. Далее, выставляя флажки, указываем какую дополнительную информацию, предлагаемую функцией «Регрессия», мы хотим иметь в результатах:

- «Остатки» – для выдачи прогнозов $\hat{y}_i$ и остатков регрессии $e_i = y_i - \hat{y}_i$;
- «График остатков» – для вывода точечной диаграммы остатков e_i;
- «График подбора» – для вывода наложенных на диаграмму рассеяния точек линии регрессии. По «ОК» получаем результаты регрессии, которые включают в себя таблицу регрессионной статистики, таблицу дисперсионного анализа, таблицу коэффициентов регрессии, таблицу остатков и графики остатков и подбора. Результаты регрессии приведены на рис. 6.3–6.4. Пояснения к выводимым результатам см. в работе № 4.

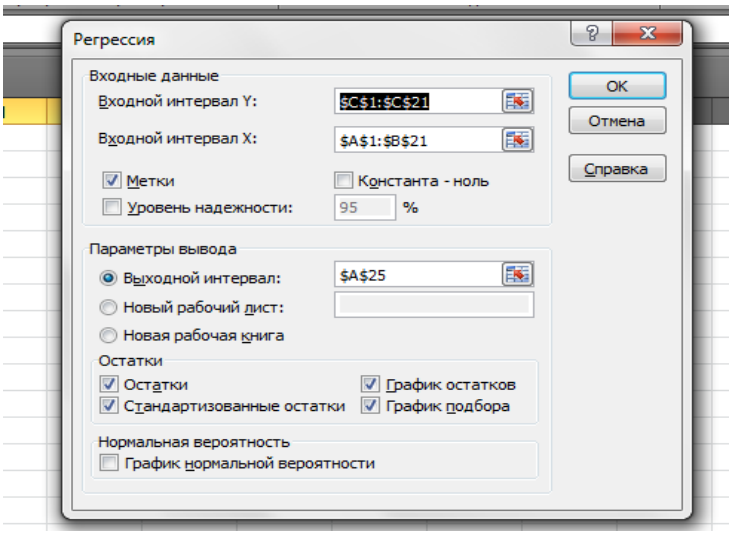


Рис. 6.2. Заполнение окна «Регрессия»

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
24																
25	ВЫВОД ИТОГОВ										ВЫВОД ОСТАТКА					
26																
27	*Регрессионная статистика*										*Наблюдение*	*Предсказанное Y*	*Остатки*			
28	Множественный R	0,99547									1	62,2976	-1,39764			
29	R-квадрат	0,99096									2	63,6365	0,36352			
30	Нормированный R-квадрат	0,98989									3	64,7371	2,26291			
31	Стандартная ошибка	2,71399									4	67,6933	3,00675			
32	Наблюдения	20									5	70,2852	3,71477			
33											6	75,8334	1,56663			
34	Дисперсионный анализ										7	81,2422	0,35775			
35		*df*	*SS*	*MS*	*F*	*Значимость F*					8	87,1624	-1,86244			
36	Регрессия	2	13723,3	6861,64	931,563	4,2484E-18					9	91,2619	-2,16194			
37	Остаток	17	125,217	7,36573							10	96,6895	-3,18945			
38	Итого	19	13848,5								11	100,709	-2,30855			
39											12	105,129	-3,12937			
40		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*		13	108,3	-1,89981			
41	Y-пересечение	135,171	52,5044	2,57447	0,01969	24,39642944	245,945	24,3964	245,945		14	112,464	0,03638			
42	X	0,13388	0,01143	11,7103	1,5E-09	0,109762894	0,15801	0,10976	0,15801		15	121,008	-2,80802			
43	P	-1,31194	0,44571	-2,94346	0,00908	-2,252313203	-0,37157	-2,25231	-0,37157		16	125,332	-1,13198			
44											17	130,023	-1,72306			
45						ty=	2,10982				18	133,649	1,25129			
46											19	137,432	3,86761			
47											20	143,315	5,18469			

Рис. 6.3. Итоги регрессии

Из таблицы коэффициентов регрессии имеем следующие МНК-оценки параметров уравнения регрессии $\hat{b}_1 = 135{,}71$, $\hat{b}_2 = 0{,}13388$, $\hat{b}_3 = -1{,}31194$.

Их стандартные ошибки равны $s_{b_1} = 52{,}5044$, $s_{b_2} = 0{,}01143$, $s_{b_3} = 0{,}4457$. 95%-е доверительные интервалы коэффициентов регрессии: $24{,}3964 < b_1 < 245{,}945$; $0{,}10976 < b_2 < 0{,}15801$; $-2{,}2523 < b_3 < -0{,}37157$. В таблице «Регрессионная статистика» величина «Стандартная ошибка» является оценкой стандартного отклонения σ зависимой переменной (ошибки регрессии), т.е. $s = \hat{\sigma} = 2{,}71399$.

Построенное уравнение регрессии: $\hat{y} = 135{,}71 + 0{,}1339 \cdot x - 1{,}3119 \cdot p$.

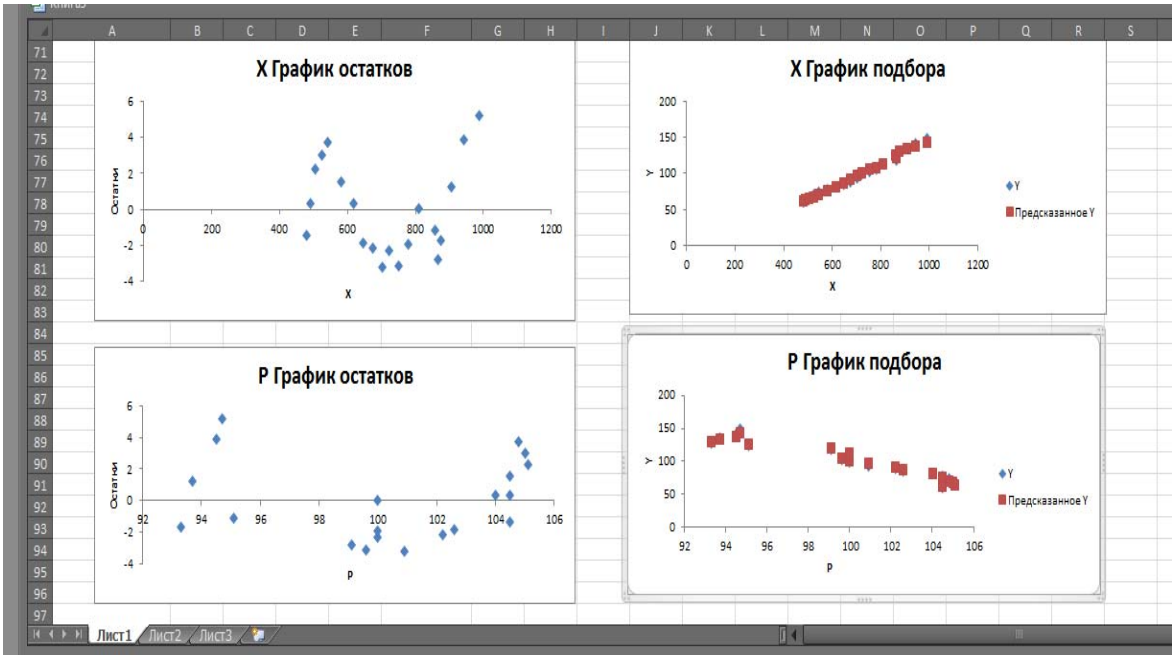


Рис. 6.4. Графики остатков и подбора множественной регрессии

Графики остатков и подбора в множественной регрессии в MS Excel выдаются отдельно по каждому фактору. Приведенные на рис. 6.4 графики остатков регрессии имеют колебательный характер, а графики подбора говорят о хорошем качестве подгонки построенной модели к наблюдаемым данным.

Верификация модели. Проверка значимости коэффициентов b_1, b_2, b_3 уравнения регрессии (значимого влияния располагаемого личного дохода X и индекса реальных цен P на совокупные расходы на жилье Y) путем проверки нулевых гипотез $H_0: b_i = 0$ с помощью *t*-статистик $t_{b_i} = \frac{\hat{b}_i}{s_{b_i}}$. Для вычисления критического значения $t_\gamma = t\left(1 - \frac{\alpha}{2}, n - p\right)$ при *n=20, p=3* и $\alpha = 0{,}05$ выделим ячейку **G45**, в вкладке «**Формулы**» выберем «**Другие функции**», в группе «**Статистические**» выберем функцию «**СТЬЮДЕНТ.ОБР.2Х**». В окне этой функции в поле «Вероятность» введем значение $\alpha = 1 - \gamma$, равное 0,05, в поле «Степени свободы» зададим число степеней *n-p*, равное 17. По «ОК» в ячейке **G45** получим значение t_γ, в рассматриваемом примере оно равно 2,1098 (см. рис. 6.3). Значения *t*-статистик для коэффициентов b_1, b_2, b_3 уравнения регрессии соответственно равны $t_{b_1} = 2{,}574$, $t_{b_2} = 11{,}71$, $t_{b_3} = -2{,}943$ и превышают по модулю критическое значение $t_\gamma = t\left(1 - \frac{\alpha}{2}, n - p\right) = 2{,}1098$. Следовательно, при уровне

значимости $\alpha = 0{,}05$ коэффициенты b_1, b_2, b_3 уравнения регрессии значимо отличаются от нуля. О значимом влиянии располагаемого личного дохода X и индекса реальных цен P на расходы на жилье Y говорят также *p*-значения, которые меньше заданного уровня значимости $\alpha = 0{,}05$, а также доверительные интервалы для коэффициентов уравнения регрессии, которые не содержат нуля.

Большие значения скорректированного коэффициента детерминации $\bar{R}^2 = 0{,}98989$ и F-статистики, $F = 931{,}56$, уровень значимости (Значимость F) которой $4{,}248*10^{-18}$ существенно меньше заданного уровня значимости $\alpha = 0{,}05$, говорят о статистической значимости построенного уравнения регрессии и хорошем качестве подгонки модели к выборочным данным. 98,99% вариации зависимой переменной объясняется вариацией объясняющих переменных.

Построение точечного и интервального прогнозы среднего зависимой переменной. Построение прогноза среднего зависимой переменной для значений факторов равных их выборочным средним, т.е. для $X = \bar{x},\ P = \bar{p}$. Для нахождения выборочных средних факторов и зависимой переменной последовательно выделяя, например, ячейки A23, B23, C23 и вводя соответственно в строке формул **=СРЗНАЧ(A2:A21)**, **=СРЗНАЧ(B2:B21)**, **=СРЗНАЧ(C2:C21)**, получим значения выборочных средних $\bar{x} = 713{,}035$, $\bar{p} = 100{,}405$, $\bar{y} = 98{,}91$. Для вычисления прогноза среднего $\hat{y} = \hat{b}_1 + \hat{b}_2 \cdot \bar{x} + \hat{b}_3 \cdot \bar{p}$ при заданных значениях факторов выделим, например, ячейку **H102** и, учитывая расположение значений факторов и коэффициентов уравнения регрессии, в строке формул введем **=B41+B42*A23+B43*B23**. По ОК в **H102** получим искомое значение $\hat{y} = 98{,}91$, совпадающее с выборочным средним $\bar{y}$, см. рис. 6.5.

Построение интервальной оценки среднего зависимой величины надежности $\gamma = 0{,}95$. Доверительный интервал надежности γ для среднего y_x зависимой величины при заданном векторе значений факторов $x = (1, x_2, x_3, \dots, x_p)$ определяется неравенством

$$\hat{y}_x - t_\gamma \sqrt{\widehat{D}(\hat{y}_x)} < y_x < \hat{y}_x + t_\gamma \sqrt{\widehat{D}(\hat{y}_x)},$$

где $t_\gamma = t\left(1 - \frac{\alpha}{2}, n - p\right)$ – квантиль уровня $1 - \frac{\alpha}{2}$ распределения Стьюдента с числом степеней свободы $n - p$. Оценка $\widehat{D}(\hat{y}_x)$ дисперсии прогноза $\hat{y}_x$ для заданного вектора значений факторов $x = (1, x_2, x_3, \dots, x_p)$ определяется как $\widehat{D}(\hat{y}_x) = s^2(1 + x(\boldsymbol{X^T X})^{-1} x^T)$. Предварительно вычислим оценку $\widehat{D}(\hat{y}_x)$ дисперсии прогноза $\hat{y}_x$ (см. рис. 6.5).

Для этого в ячейках **A100-C119** создадим матрицу $\boldsymbol{X}$, первый столбец которой состоит из единиц, второй – из значений фактора X, третий – из значений фактора P. В ячейках **B123-U125** разместим транспонированную матрицу $\boldsymbol{X^T}$. В ячейках **F98-H98** расположим вектор $(1, \bar{x}, \bar{p})$=(1;713,035;100,405) значений факторов для которых вычисляется интервальный прогноз. В ячейках **K98-K100** расположим транспонированный вектор $(1, \bar{x}, \bar{p})^T$. Выделим ячейку **H105** и, учитывая расположение величины *s* в ячейке **B31**, в строке формул введем **=B31^2*(1+МУМНОЖ(МУМНОЖ(F98:H98;МОБР(МУМНОЖ(B123:U125; A100:C119)));K98:K100))**. По «Enter» в этой ячейке получим искомое значение

оценки дисперсии среднего, равное 7,734. Выделим под нижнюю границу доверительного интервала ячейку **К108** и в строке формул введем

=Н102-СТЬЮДЕНТ.ОБР(0,975;17)*КОРЕНЬ(Н105)

По Enter в ячейке **К108** получим значение нижней границы доверительного интервала, равное 93,0426. Аналогично, выделив ячейку **N108** и введя в строке формул

=Н102+СТЬЮДЕНТ.ОБР(0,975;17)*КОРЕНЬ(Н105),

получим в ней значение верхней границы доверительного интервала, равное 104,777. Таким образом, доверительный интервал, надежности $\gamma = 0{,}95$, для среднего y_x зависимой величины Y (расходов на жилье) для значений факторов $x = 713{,}035$ и p= 100,405 задается неравенством $93{,}0426 < \hat{y}_x < 104{,}777$.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
98	Матрица X				(1,x,p)=	1	713,035	100,405		1	1											
99										x	713,035											
100	1	479,7	104,5							p	100,405											
101	1	489,7	104,5																			
102	1	503,8	105,1			прогноз среднего y		98,91														
103	1	524,9	105																			
104	1	542,3	104,8																			
105	1	580,8	104,5		оценка дисперсии прогноза			7,73401														
106	1	616,3	104																			
107	1	646,8	102,6							нижняя граница			верхняя граница									
108	1	673,5	102,2		Доверительный интервал прогноза среднего Y						93,0426			104,777								
109	1	701,3	100,9																			
110	1	722,5	100																			
111	1	751,6	99,6																			
112	1	779,2	100																			
113	1	810,3	100																			
114	1	865,3	99,1																			
115	1	858,4	95,1																			
116	1	875,8	93,3																			
117	1	906,8	93,7																			
118	1	942,9	94,5																			
119	1	988,8	94,7																			
120																						
121	Транспонированная матрица X																					
122																						
123		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
124		479,7	489,7	503,8	524,9	542,3	580,8	616,3	646,8	673,5	701,3	722,5	751,6	779,2	810,3	865,3	858,4	875,8	906,8	942,9	988,8	
125		104,5	104,5	105,1	105	104,8	104,5	104	102,6	102,2	100,9	100	99,6	100	100	99,1	95,1	93,3	93,7	94,5	94,7	
126																						

Рис. 6.5. Построение точечного и интервального прогноза среднего $\hat{y}_x$

Общее заключение об оцененной модели и ее интерпретация.

Построенная модель линейной множественной регрессии средних расходов на жилье Y (млрд. дол.) от располагаемого личного дохода X (млрд. дол.) и индекса реальных цен P (%)

$$\hat{y} = 135{,}71 + 0{,}13388 \cdot x - 1{,}31194 \cdot p$$

хорошо согласуется с имеющейся выборкой. Об этом свидетельствует высокое значение нормированного коэффициента детерминации $R^2 = 0{,}98989$, т.е. 98,99% вариации Y относительной ее средней объясняется изменениями X и P. Большое значение F-статистики, $F = 931{,}56$, и ее уровень значимости, равный $4{,}248 \cdot 10^{-18}$, свидетельствует о наличии значимой линейной корреляционной зависимости Y от X и P. Оценка $s = 2{,}714$ среднеквадратического отклонения σ ошибок регрессии ε_i мала по сравнению с $\bar{y} = 98{,}91$, что также свидетельствует о малом разбросе выборочных данных относительно плоскости регрессии. Значения t-статистик коэффициентов b_1, b_2, b_3 уравнения регрессии и их p-значения, равные соответственно 0,0197, $1{,}5 \cdot 10^{-9}$ и 0,00091 говорят об их зна-

чимом отличии от нуля. Следовательно, располагаемый личный доход и индекс цен значимо влияют на расходы на жилье и построенная регрессионная модель статистически значима.

Интерпретация построенной модели $\hat{y} = 135{,}71 + 0{,}13388 \cdot x - 1{,}31194 \cdot p$. Коэффициент при индексе цен p имеет отрицательный знак, что согласуется с теоретическим положением о снижении спроса на жилье с ростом цен. Коэффициент при величине располагаемых доходов x положительный, что согласуется с положением о росте спроса с ростом доходов. Значения коэффициентов при x и p говорят о возрастании расходов на жилье в среднем на 0,13388 млрд дол. при росте располагаемых личных доходов на 1 млрд дол. и сокращении расходов на жилье на 1,31194 млрд дол при росте индекса цен на 1%.

Выборочный коэффициент корреляции располагаемого личного дохода X и индекса цен P близок по модулю к единице, $r_{XP} = -0{,}94175$, это говорит о сильной коррелированности рассматриваемых факторов и о необходимости проведения дополнительных исследований на мультиколлинеарность. Кроме того, выборочные данные являются временными рядами, поэтому возможна автокорреляции остатков. Следовательно, необходимо исследовать построенную модель на автокорреляцию остатков.

Контрольные вопросы

1. В чем заключается спецификация модели множественной регрессии?
2. Что характеризует множественный коэффициент корреляции?
3. Как находятся оценки параметров линейной множественной регрессии?
4. Может ли быть линейная множественная регрессия быть нелинейной по объясняющим переменным?
5. Сформулируйте критерии значимости параметров множественной регрессии.
6. Приведите предпосылки линейной множественной регрессии.
7. Сформулируйте Теорему Гаусса-Маркова.
8. С помощью каких критериев проверяется значимость линейного уравнения множественной регрессии?
9. В чем отличие ошибок регрессии от остатков регрессии?
10. Что характеризует скорректированный коэффициент детерминации?
11. Как определяется средняя ошибка аппроксимации, что она характеризует?
12. Как интерпретируются коэффициенты линейной множественной регрессии?
13. Что характеризует частный коэффициент эластичности для линейной множественной регрессии?
14. В чем заключается прогноз значений зависимой переменной? Как определяется дисперсия прогноза?
15. Как строится интервальный прогноз среднего зависимой переменной?
16. С увеличением надежности интервального прогноза он увеличивается или уменьшается?

Лабораторная работа № 7.
Анализ мультиколлинеарности и авторегрессии в модели множественной регрессии

Цель работы. Освоение методов выявления мультиколлинеарности и автокорреляции ошибок в множественной регрессии с использованием пакета анализа MS Excel 2010.

Краткие сведения. Здесь используются обозначения, принятые в кратких сведениях к лабораторной работе № 6.

Мультиколлинеарность.

Одной из предпосылок классической линейной регрессии является предположение о линейной независимости объясняющих переменных. Это означает линейную независимость векторов-столбцов X_1, X_2,…, $X_\text{р}$ значений факторов, что равносильно тому, что определитель матрицы $\boldsymbol{X}^T \cdot \boldsymbol{X}$ не равен нулю (ранг этой матрицы и матрицы X равен p). В этом случае существует обратная матрица $(\boldsymbol{X}^T \cdot \boldsymbol{X})^{-1}$ и оценки коэффициентов уравнения регрессии однозначно определяются соотношением $\widehat{\boldsymbol{b}} = (\boldsymbol{X}^T \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}^T \cdot \boldsymbol{Y}$.

Под *мультиколлинеарностью* понимается высокая взаимная коррелированность факторов, выбранных в качестве объясняющих переменных в модели множественной регрессии. Мультиколлинеарность может проявляться в *функциональной* (полной, явной) и *стохастических* формах.

При *функциональной форме мультиколлинеарности* между факторами (между векторами-столбцами X_1, X_2,…, $X_\text{р}$ значений факторов) существует линейная функциональная зависимость, т.е. нарушается предположение о линейной независимости объясняющих переменных. В этом случае определители матрицы $\boldsymbol{X}^T \cdot \boldsymbol{X}$ и матрицы $q_\text{ф} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$ выборочных коэффициентов корреляции между факторами X_1, X_2,…, $X_\text{р}$ равны нулю, что не позволяет получить однозначные оценки коэффициентов уравнения регрессии. Если факторы линейно независимы, то $q_\text{ф} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$ и $detq_\text{ф} = 1$. Если хотя бы два фактора линейно зависимы, например, X_1 и X_2, то $q_\text{ф} = \begin{bmatrix} 1 & 1 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$ и $detq_\text{ф} = 0$.

В *стохастической* форме мультиколлинеарности между хотя бы двумя объясняющими переменными существует тесная линейная корреляционная зависимость. В этом случае определитель матрицы $\boldsymbol{X}^T \cdot \boldsymbol{X}$ отличен от нуля, но может принимать очень маленькие по модулю значения, а $detq_\text{ф} < 1$ и близок к нулю. Это приводит к большим значениям элементов обратной матрицы $(\boldsymbol{X}^T \cdot \boldsymbol{X})^{-1}$. Следовательно, оценки $s_{b_i}^2 = \widehat{Var}(\hat{b}_i) = s^2(\boldsymbol{X}^T \cdot \boldsymbol{X})_{ii}^{-1}$ дисперсий коэффициентов регрессии принимают большие значения и не имеют смысла, в

силу малости вычисленных значений *t*-статистик коэффициентов регрессии. Кроме того, в этом случае вычислительные погрешности приводят к значительным ошибкам в оценках коэффициентов регрессии и их дисперсий. При этом уравнение линейной множественной регрессии может оказаться в целом значимым по *F*-критерию $F = \frac{MS_R}{MS_{\text{ост}}}$ при незначимости некоторых коэффициентов уравнения регрессии. Стохастическая форма мультиколлинеарности факторов частое явление в экономических исследованиях.

Точных количественных критериев выявления наличия или отсутствия мультиколлинеарности в стохастической форме не существует. Для выявления мультиколлинеарности используют его следующие характерные признаки.

- Анализ корреляционной матрицы. Если модули некоторых парных коэффициентов корреляции между факторами превышает 0,75, то имеет место мультиколлинеарность.
- Определители матрицы $q_{\text{ф}}$, выборочных коэффициентов корреляции между факторами, близок к нулю.
- Небольшие изменения выборочных данных (например, отбрасывание небольшой части выборочных данных или добавление небольшого количества новых данных) приводят к существенному изменению оценок коэффициентов уравнения регрессии.
- Оценки коэффициентов уравнения регрессии имеют большие стандартные отклонения и малую значимость, а само уравнение в целом значимо, о чем свидетельствуют большие значения коэффициента детерминации R^2 и *F*-статистики.
- Интерпретация коэффициентов уравнения регрессии не согласуется с положениями экономической теории (например, коэффициенты имеют знаки или значения, не согласующиеся с теорией).

Наличие мультиколлинеарности требует доработки модели с целью ее устранения или уменьшения. Эти методы рассматриваются в более широких курсах эконометрики.

Автокорреляция ошибок регрессии.

При построении регрессионных моделей по временным рядам (упорядоченным данным за последовательные моменты или промежутки времени) предположение о некоррелированности ошибок регрессии ε_i и ε_j (или y_i и y_j) в разных наблюдениях не выполняется, т.е. $Cov(\varepsilon_i, \varepsilon_j) \neq 0$. Это объясняется тем, что значения изучаемых величин в момент времени *t* в значительной степени зависят от их значений в предшествующие моменты времени. В этом случае говорят об *автокорреляции* данных и строят авторегрессионные модели, учитывающие автокорреляцию данных. Простейшим примером такой модели является *авторегрессионный процесс первого порядка*: $\varepsilon_t = \rho \cdot \varepsilon_{t-1} + \vartheta_t$, который описывает ошибку регрессии ε_t в момент времени *t* как линейную функцию от ошибки ε_{t-1} в момент времени *t-1* и случайной ошибки ϑ_t распределенной по нормальному закону с нулевым средним и постоянной дисперсией для всех *t*. Величина ρ называется *коэффициентом авторегрессии*. Авторегрессионные модели рассматриваются в более подробных курсах эконометрики.

Применение метода наименьших квадратов для оценивания множественной регрессии $y_i = b_1 + b_2 x_{i2} + b_2 x_{i3} + \cdots + b_p x_{ip} + \varepsilon_i$ при наличии корреляции ошибок дает несмещенные и состоятельные оценки $\hat{b}_i$ коэффициентов регрессии, но оценки $s_{b_i}^2$ их дисперсий несостоятельные и смещенные (как правило, в сторону занижения). Это приводит к тому, что результаты тестирования гипотез о значимости коэффициентов регрессии по *t*-критерию $t = \frac{\hat{b}_i}{s_{b_i}}$ оказываются недостоверными и оцененная модель дает более оптимистическую картину регрессии, чем есть на самом деле.

Различают *положительную автокорреляцию* при положительном коэффициенте авторегрессии ρ, что геометрически выражается в чередовании зон с положительными и отрицательными значениями остатков регрессии e_i. *Отрицательная автокорреляция* имеет место при отрицательном коэффициенте авторегрессии ρ, что геометрически выражается в том, что последовательные значения остатков регрессии имеют разные знаки. Таким образом, о наличии автокорреляции можно судить по графику остатков.

Для выявления наличия автокорреляции первого порядка (зависимости значения ε_t только от предшествующего значения ε_{t-1}) используется *критерий Дарбина-Уотсона*. Этот критерий основан на простой идеи: если корреляция есть в ошибках регрессии ε_t, то она присутствует и в остатках регрессии e_t получаемых после применения метода наименьших квадратов. Критерий Дарбина-Уотсона основан на статистике

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2},$$

которая принимает значения от 0 до 4. Нулевая гипотеза об отсутствии автокорреляции (H_0: $\rho = 0$) принимается или отклоняется при попадании наблюдаемого (вычисленного) значения *d* в промежутки в соответствии с рис. 7.1.

H_0 отвергается, положительная автокорреляция	Зона неопределенности	H_0 принимается, автокорреляция отсутствует	Зона неопределенности	H_0 отвергается, отрицательная автокорреляция	
0	$d_н$	$d_в$	$4-d_в$	$4-d_н$	4

Рис. 7.1. Критерий Дарбина-Уотсона

При $d \in (d_в, 4 - d_в)$ автокорреляция отсутствует; при $d \in (d_н, d_в)$ или $d \in (4 - d_в, 4 - d_н)$ ничего о наличии или отсутствии автокорреляции сказать нельзя (зона неопределенности критерия); при $d \in (0, d_н)$ нулевая гипотеза отвергается и имеет место положительная автокорреляция; при $d \in (4 - d_н, 4)$ нулевая гипотеза отвергается и имеет место отрицательная автокорреляция. Верхние $d_в$ и нижние $d_н$ границы критического значения статистики d критерия Дарбина-Уотсона для уровня значимости $\alpha = 0{,}05$ приводятся в специальных таблицах, например, в приложении 2.4 учебника Эконометрика / И.И. Елисеева, С.В. Курышева, Т.В. Костеева и др.

Содержание лабораторной работы.

1. Исследовать на мультиколлинеарность линейную множественную регрессию, построенную в лабораторной работе №6, используя:

- парные коэффициенты корреляции факторов;
- определитель матрицы $q_ф$ выборочных коэффициентов корреляции между факторами;
- оценку уравнения регрессии на укороченной выборке;
- сопоставление значимости коэффициентов и самого уравнения регрессии;
- согласованность интерпретации коэффициентов с положениями теории.

2. По критерию Дарбина-Уотсона исследовать автокорреляцию ошибок регрессии, построенной в лабораторной работе № 6.

3. Общее заключение о мультиколлинеарности и автокорреляции ошибок построенного в работе №6 линейного уравнения множественной регрессии.

Выполнение работы в MS Excel.

Анализ на мультиколлинеарности и автокорреляцию ошибок проведем на примере линейной множественной регрессии, построенной в лабораторной работе № 6. В этой работе была построена линейная регрессия

$$\hat{y} = 135{,}71 + 0{,}13388 \cdot x - 1{,}31194 \cdot p$$

зависимости расходов на жилье (Y, млрд дол.) от располагаемого личного дохода (X, млрд дол.) и индекса реальных цен (P). Исходная выборка является временным рядом.

Анализ мультиколлинеарности по матрице парных коэффициентов корреляции. Парные коэффициенты корреляции факторов X, P и зависимой переменной Y, приведенные на рис. 6.1 работы № 6, равны $r_{yx} = 0{,}993$, $r_{yp} = -0{,}958$, $r_{px} = -0{,}942$ Это говорит о тесной корреляционной зависимости Y с каждым из факторов X и P, а также о тесной корреляционной зависимости самих факторов. Что свидетельствует о наличии мультиколлинеарности.

Анализ мультиколлинеарности по определителю матрицы $q_ф$ выборочных коэффициентов корреляции между факторами X и P. Матрицу $q_ф$ определим, используя корреляционную матрицу, приведенную на рис. 6.1 работы №6. Матрица $q_ф$ располагается в ячейках **G5-H6**, т.е. $q_ф = \begin{bmatrix} 1 & -0{,}94175 \\ -0{,}94175 & 1 \end{bmatrix}$. Ее определитель, равный 0,113, близок к нулю. Это говорит о наличии мультиколлинеарности.

Анализ мультиколлинеарности по регрессии на укороченной выборке. Укоротим выборку, например, отбросив последние (или первые) четыре наблюдения и, следуя работе № 6, построим линейную множественную регрессию по укороченной выборке. Результаты этой регрессии приведены на рис. 7.2. Полученное уравнение регрессии

$$\hat{y} = 98{,}95 + 0{,}1308 \cdot x - 0{,}9234 \cdot p$$

значимо и обладает хорошими аппроксимационными качествами, $R^2 = 0{,}99$. Оценки свободного члена и коэффициента регрессии при факторе p уравнения регрессии, полученного по укороченной выборке, значительно отличаются от

оценок, полученных по полной выборке. Следовательно, имеет место мультиколлинеарность.

Сопоставление значимости коэффициентов и самого уравнения регрессии. Уравнения регрессии, построенные по полной и укороченной выборкам, значимы. Оценки свободного члена и коэффициента регрессии при факторе p уравнения регрессии, полученного по укороченной выборке, имеют большие (относительно самих оценок) стандартные отклонения, $s_{b_1} = 51{,}05$ и $s_{b_3} = 0{,}442$; их p-значения превышают заданный уровень значимости $\alpha = 0{,}05$. Следовательно, эти коэффициенты статистически не значимы. Таким образом, значимость оцененного уравнения регрессии противоречит незначимости части его коэффициентов, что свидетельствует о наличии мультиколлинеарности.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
128	определитель X^T·X=	369458754			Регрессия по укороченной выборке									
129														
130	ВЫВОД ИТОГОВ										Статистика Дарбина-Уотсона			
131														
132	*Регрессионная статистика*										d=	0,441637		
133	Множественный R	0,996352												
134	R-квадрат	0,9927173												
135	Нормированный R-квад	0,9915969												
136	Стандартная ошибка	1,8158262												
137	Наблюдения	16												
138														
139	Дисперсионный анализ													
140		*df*	*SS*	*MS*	*F*	*Значимость F*								
141	Регрессия	2	5842,866	2921,433	886,0279	1,2732E-14								
142	Остаток	13	42,86392	3,297225										
143	Итого	15	5885,73											
144														
145		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*					
146	Y-пересечение	98,95406	51,0544	1,938208	0,074623	-11,34227269	209,2504	-11,3423	209,2504					
147	X	0,1308261	0,009692	13,49823	5,04E-09	0,10988759	0,151765	0,109888	0,151765					
148	P	-0,942343	0,441777	-2,13307	0,052555	-1,896744684	0,012059	-1,89674	0,012059					

Рис. 7.2. Результаты регрессии по укороченной выборке

Согласованность интерпретации коэффициентов с положениями теории. В обеих построенных моделях коэффициенты регрессии при факторах x и p имеют знаки, не противоречащие экономической теории (расходы на жилье растут с ростом располагаемого личного дохода и убывают с ростом индекса реальных цен). Значения свободного члена не поддаются интерпретации в виду больших значений факторов в выборке. С укорочением выборки оценка коэффициента регрессии при факторе p значительно изменяется (на 30 %), что не позволяет приближенно оценить изменение среднего спроса на жилье при росте цен на 1 %. Оценка коэффициента регрессии при факторе x изменяется менее чем на 2,3 % при укорочении выборки, что позволяет говорить об увеличении среднего спроса на жилье приблизительно на 0,131 млрд дол. при росте располагаемого дохода на 1 млрд. дол. Невозможность однозначной интерпретации параметров уравнения регрессии говорит о мультиколлинеарности.

Анализ автокорреляции ошибок регрессии по критерию Дарбина-Уотсона. Вычислим статистику $d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$ для уравнения регрессии построенного в лабораторной работе №6. Сумма квадратов остатков $\sum_{t=1}^{n} e_t^2$ вычисляется при построении уравнения регрессии, она располагается в таблице «Дисперсионный анализ» на пересечении строки «Остаток» и столбца «SS», см. рис. 6.3. В рассматриваемом примере эта сумма располагается в ячейке **C37** и равна 125,217, а значения остатков e_t в ячейках **M28-M47**, см. рис. 6.3 в работе № 6. Для вычисления $\sum_{t=2}^{n}(e_t - e_{t-1})^2$ используем функцию ***СУММКВРАЗН*** в группе «Математические». Выделим под статистку d, например, ячейку **L132** и в строке формул введем **=СУММКВРАЗН(M29:M47;M28:M46)/C37**. Полученное значение статистики *d* равное 0,4416. Верхнюю $d_{\text{в}}$ и нижнюю $d_{\text{н}}$ границы критического значения статистики d критерия Дарбина-Уотсона для уровня значимости $\alpha = 0{,}05$, числа объясняющих переменных *p*=2 и объема выборки равного 20 найдем по приложению 2.4 учебника «Эконометрика» под редакцией И.И. Елисеевой. $d_{\text{в}} = 1{,}54$, $d_{\text{н}} = 1{,}10$. Сопоставим найденное значение *d*=0,4416 с границами промежутков принятия или отклонения гипотезы об отсутствии автокорреляции, см. рис.7.1. В рассматриваемом примере эти промежутке равны (0; 1,10), (1,10; 1,54), (1,54; 2,46), (2,46; 2,9), (2,9; 4). Найденное значение d принадлежит первому промежутку, что говорит о наличии положительной автокорреляции первого порядка ошибок регрессии. Об этом свидетельствуют также графики остатков, в которых промежутки положительных остатков регрессии чередуются с промежутками их отрицательных значений, см. рис. 6.4 работы № 6.

Общее заключение о мультиколлинеарности и автокорреляции ошибок уравнения множественной регрессии $\hat{y} = 135{,}71 + 0{,}13388 \cdot x - 1{,}31194 \cdot p$. В построенной модели имеет место мультиколлинеарность. В пользу этого утверждения говорят: высокая коррелированность факторов, $r_{px} = -0{,}942$; определитель матрицы $q_{\text{ф}}$ выборочных коэффициентов корреляции между факторами, равный 0,113, близок к нулю; значительное изменение оценок коэффициентов регрессии при изменении выборки; значимость регрессии построенной по укороченной выборке при незначимости коэффициента регрессии при факторе p и свободного члена. Проведенный анализ показал наличие положительной автокорреляция ошибок регрессии. Об этом свидетельствуют графики остатков регрессии и критерий Дарбина-Уотсона.

Результаты анализа модели на мультиколлинеарность и автокорреляцию говорят о необходимости доработки модели в плане выбора объясняющих переменных и структуры модели.

Контрольные вопросы

1. Что понимается под мультиколлинеарностью в модели множественной регрессии?
2. В чем заключается функциональная форма мультиколлинеарности?
3. В чем заключается стохастическая форма мультиколлинеарности?

4. К чему приводит мультиколлинеарность факторов?

5. Перечислите характерные признаки проявления мультиколлинеарности.

6. Какая предпосылка классической линейной регрессии нарушается при наличии мультиколлинеарности факторов?

7. Матрица $q_{\text{ф}}$, выборочных коэффициентов корреляции между факторами, равна $\begin{pmatrix} 1 & -0{,}91 \\ -0{,}91 & 1 \end{pmatrix}$. Имеет ли место мультиколлинеарность факторов?

8. В чем заключается проблема автокорреляции?

9. Каковы последствия автокорреляции?

10. Какая предпосылка классической линейной регрессии нарушается при наличии автокорреляции?

11. Как на графиках остатков регрессии проявляется наличие положительной и отрицательной автокорреляции?

12. Сформулируйте и поясните модель авторегрессионного процесса первого порядка.

13. С помощью какого критерия выявляется наличие автокорреляции первого порядка?

14. Как вычисляется статистика критерия Дарбина-Уотсона?

15. На какие области разбивается множество значений статистики Дарбина-Уотсона?

Лабораторная работа № 8.
Линейные регрессионные модели переменной структуры, фиктивные переменные

Цель работы. Освоение построения линейных регрессионных моделей переменной структуры с использованием фиктивных переменных в пакете анализа MS Excel, интерпретация параметров модели с фиктивными переменными.

Краткие сведения. В модели линейной множественной регрессии в качестве объясняющих переменных (факторов, регрессоров) могут рассматриваться не только непрерывные количественные признаки, принимающие значения на некотором числовом промежутке, но и качественные признаки, имеющие два или несколько уровней. К таким признакам можно отнести: пол (мужской, женский); сезон года (зима, весна, лето, осень); расположение квартиры (районы города); уровень образования (общее среднее, среднее техническое, высшее); используемая технология производства и тому подобное.

Качественные признаки могут существенно влиять на зависимую переменную, что приводит к изменению параметров уравнения регрессии при переходе от данных полученных для одного уровня качественного фактора к данным полученным для другого уровня этого же фактора. Например, при исследовании производительности труда *y* от стоимости рабочего места *x* для работников одной квалификации в зависимости от пола были получены уравнения регрессии: для мужчин $y = 0{,}34 + 0{,}22x$, для женщин $y = 0{,}36 + 0{,}19x$. Коэффициенты этих уравнений отличаются и говорят о том, производительность труда женщин с ростом стоимости рабочего места растет медленнее, чем у мужчин.

Влияние стоимости рабочего места и пола работника на производительность труда можно описать и одним уравнением регрессии введя дополнительную объясняющую переменную *z*, принимающую значение «1» для мужчин и «0» для женщин. Объединив выборки для мужчин и женщин в одну выборку можно рассматривать регрессионную модель, включающую переменную *z*:

$$y = b_0 + b_1 x + b_2 z. \qquad (8.1)$$

При *z=1* (для мужчин) это уравнение имеет вид $y = (b_0 + b_2) + b_1 x$. При *z=0* (для женщин) $y = b_0 + b_1 x$. В модели (8.1) качественный фактор «пол работника» влияет только на свободный член уравнения регрессии (для мужчин он равен $(b_0 + b_2)$, для женщин b_0) и не влияет на коэффициент регрессии при факторе x. Для учета влияния качественного фактора «пол работника» на коэффициент регрессии при факторе x рассматривается модель

$$y = b_0 + b_1 \cdot x + b_2 \cdot z + b_3 \cdot (x \cdot z). \qquad (8.2)$$

Произведение $x \cdot z$ рассматривается как новая объясняющая переменная. По этой модели для мужчин имеем регрессию $y = (b_0 + b_2) + (b_1 + b_3) \cdot x$, для женщин $y = b_0 + b_1 \cdot x$. В модели (8.2) качественный фактор «пол работника» влияет и на свободный член уравнения регрессии (для мужчин он равен $(b_0 + b_2)$, для женщин b_0), и на коэффициент регрессии при факторе x (для

мужчин он равен $(b_1 + b_3)$, для женщин b_1). В моделях (8.1) и (8.2) с изменением уровня качественного признака «пол работника» происходит изменение структуры зависимости, поэтому говорят о модели переменной структуры.

Для введения в уравнение регрессии качественных признаков как объясняющих переменных их необходимо преобразовать в количественные. Для этого различным уровням качественного фактора присваиваются цифровые метки. Введенные таким образом новые количественные переменные называют *фиктивными* (манекенными). Переменные, принимающие только два значения, называются *бинарными* (булевыми).

Если уровней фактора всего два, то одному уровню присваивается значение «1», другому «0», и качественный фактор рассматривается как количественный принимающий только два значения 0 или 1.

Если качественный признак имеет k уровней (k>2), то эти уровни можно оцифровать натуральными числами 1, 2, …, k и рассматривать его как количественную переменную, принимающую k различных числовых значения. Например, для признака «время года» можно ввести фиктивную переменную z, принимающую для «зимы» значение 1, для «весны» – 2, для «лета» – 3, для «осени» – 4. Но из-за трудностей содержательной интерпретации параметров модели с такими фиктивными переменными так не поступают. В этом случае вводят (k-1) бинарных фиктивных переменных $z_1, z_2, \dots, z_{k-1}$, каждый из которых для одного из уровней качественного признака принимает значение «1», а для других уровней «0». Введение k таких фиктивных переменных приводит к их линейной зависимости, так как в этом случае имело бы место равенство $z_1 + z_2 + \dots + z_k = 1$, т.е. к полной мультиколлинеарности факторов и невозможности применения метода наименьших квадратов для оценивания модели с такими фиктивными переменными. Например, для качественного признака «время года» с четырьмя уровнями нужно ввести три фиктивные бинарные переменные

$$z_1 = \begin{cases} 1, & \text{для «зимы»,} \\ 0, & \text{для других периодов,} \end{cases}$$
$$z_2 = \begin{cases} 1, & \text{для «весны»,} \\ 0, & \text{для других периодов,} \end{cases} \qquad (8.3)$$
$$z_3 = \begin{cases} 1, & \text{для «лета»,} \\ 0, & \text{для других периодов.} \end{cases}$$

Нулевые значения этих фиктивных переменных соответствуют «осени». С использованием этих фиктивных переменных зависимость спроса y на некоторый товар от дохода x, учитывающая сезонные изменения спроса, может быть описана уравнением

$$y = b_0 + b_1 x + b_2 z_1 + b_3 z_2 + b_4 z_3. \qquad (8.4)$$

В этой модели склонность к потреблению, коэффициент b_1, одинакова для всех периодов года, а сводные члены различны. Для «осени» он равен b_0, для «зимы» – $(b_0 + b_2)$, для «весны» – $(b_0 + b_3)$, для «лета» – $(b_0 + b_4)$. При одинаковом доходе x средний спрос по сезонам составляет: «зимой» – $y =$

$(b_0 + b_2)+b_1x$, «весной» – $y = (b_0 + b_3)+b_1x$, «летом» – $y = (b_0 + b_4)+b_1x$ «осенью» – $y = b_0+b_1x$. Таким образом, при одинаковом доходе отличия среднего спроса по сезонам определяются коэффициентами при фиктивных переменных. Например, средний спрос зимой отличается от среднего спроса осенью на величину b_2, а от спроса весной на величину $b_2 - b_3$. При изменении склонности к потреблению по временам года в модель (8.4) необходимо ввести фиктивные переменные $x \cdot z_1,\ x \cdot z_2,\ x \cdot z_3$.

Следует отметить, что в некоторых случаях для учета изменчивости свободного члена необходимо вводить одну группу фиктивных переменных, а для учета изменчивости коэффициентов регрессии другую группу фиктивных переменных. Например, при изучении влияния сезонности и социальной группы на спрос для учета влияния социальной группы на склонность к потреблению (коэффициент при доходе x) нужно ввести новую группу фиктивных переменных, которые описывают уровни качественного признака «социальная группа». Регрессионная модель может содержать в качестве объясняющих переменных только фиктивные переменные. Например, модель $y = b_0 + b_2z_1+b_3z_2 + b_4z_3$ с фиктивными переменными (8.3) описывает изменение среднего спроса в зависимости от времени года. Фиктивные переменные могут использоваться и в нелинейных моделях.

Регрессионные модели с фиктивными переменными оцениваются классическим методом наименьших квадратов. Оценка значимости полученного уравнения и его коэффициентов, оценка качества модели, построение прогнозов производится также, как и для модели обычной множественной регрессии (см. работу № 6).

Критерий Чоу. На практике не редки случаи, когда имеются две выборки значений, объясняющих $X_1, X_2, \cdots, X_p$ и зависимой переменной Y, $(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i)$ объемами n_1 и n_2 полученные при различных условиях (например, при разных уровнях некоторого качественного признака). Возникает вопрос об однородности этих выборок в регрессионном смысле, т. е. можно ли их объединить в одну выборку объема $n = n_1 + n_2$ и оценивать регрессионную модель по объединенной выборке. В критерии Чоу по каждой из выборок строятся линейные модели

$$y_i = b_0^* + \sum_{j=1}^{p} b_j^* x_{ij} + \varepsilon_i^* \ , \ \ i = 1, \dots, n_1;$$

$$y_i = b_0^{**} + \sum_{j=1}^{p} b_j^{**} x_{ij} + \varepsilon_i^{**} \ , \quad i = 1, \dots, n_2.$$

Гипотеза об однородности выборок имеет вид $H_0: b_j^* = b_j^{**},\ j = 0, 1, \dots, p; D(\varepsilon_i^*) = D(\varepsilon_i^{**})$. Если нулевая гипотеза верна (параметры этих уравнений регрессии отличаются незначимо), то выборки объединяются в одну и по объединенной выборке оценивается модель

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + \varepsilon_i \ , \quad i = 1, \dots, n.$$

Для проверки нулевой гипотезы в критерии Чоу используется F-статистика

$$F = \frac{(\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n_1} e_i^2 - \sum_{i=1}^{n_2} e_i^2)(n-2p-2)}{(\sum_{i=1}^{n_1} e_i^2 + \sum_{i=1}^{n_2} e_i^2)(p+1)}, \quad (8.5)$$

где $\sum_{i=1}^{n} e_i^2$, $\sum_{i=1}^{n_1} e_i^2$, $\sum_{i=1}^{n_2} e_i^2$ – суммы квадратов остатков регрессий, построенных соответственно по объединенной, первой и второй выборкам, p – количество объясняющих переменных. Если вычисленное значение F больше критического $F(\alpha, p+1, n-2p-2)$, то нулевая гипотеза отвергается на уровне значимости α. Если нулевая гипотеза принимается, то можно рассматривать уравнение регрессии, полученное по объединенной выборке.

Критерий Чоу может быть использован при построении моделей регрессии при наличии воздействия качественных факторов. Выборка делится на группы по уровням качественного признака, и рассматриваются гипотезы об их однородности, которые проверяются критерием Чоу. Если гипотезы об однородности принимаются, то качественный признак не оказывает влияния на исследуемую зависимую переменную и рассматривается регрессия по полной выборке без фиктивной переменной. Если гипотеза об однородности отклоняется, то качественный признак влияет на зависимую переменную и строится регрессия с включением фиктивных переменных по полной выборке или регрессии по каждой группе.

Содержание лабораторной работы.

1. В соответствии с поставленной задачей исследования определить необходимые фиктивные переменные и спецификацию модели.

2. Сформировать и ввести выборочные данные с учетом фиктивных переменных.

3. Методом наименьших квадратов оценить параметры модели (следуя работе № 6).

4. Верификация модели (проверка значимости коэффициентов уравнения регрессии и всего уравнения в целом при уровне значимости α=0,05), оценка качества построенной модели.

5. По критерию Чоу проверить гипотезу о однородности выборок для разных значений качественного признака».

6. Интерпретация модели и общее заключение о проведенном исследовании.

Выполнение работы в MS Excel. Построение регрессионной модели с фиктивными переменными рассмотрим на примере исследования зависимости заработной платы Y (тыс. р.) от возраста X (лет) и пола работника по данным приведенным в табл. 8.1

Таблица 8.1

№ п/п	*Y*	*X*	*Пол*	№ п/п	*Y*	*X*	*Пол*
1	30	29	Ж	11	25	28	Ж
2	40	40	М	12	35	30	М
3	30	36	Ж	13	20	25	М
4	32	32	Ж	14	40	48	М
5	20	23	М	15	22	30	Ж
6	35	45	Ж	16	32	40	М
7	35	38	Ж	17	39	40	М

№ п/п	*Y*	*X*	*Пол*	№ п/п	*Y*	*X*	*Пол*
8	40	40	М	18	36	38	М
9	38	50	М	19	26	29	Ж
10	40	47	М	20	25	25	М

Определение необходимой фиктивной переменной и спецификации модели. В данной задаче «Пол» является качественным признаком, принимающим два значения. Признак представим фиктивной бинарной переменной z принимающей значение 1 для мужчин и 0 для женщин. Этот признак может оказывать влияние на среднюю зарплату при одинаковом возрасте мужчин и женщин, а также приводить к разному изменению зарплаты с изменением возраста. Поэтому рассмотрим две модели с фиктивной переменной:

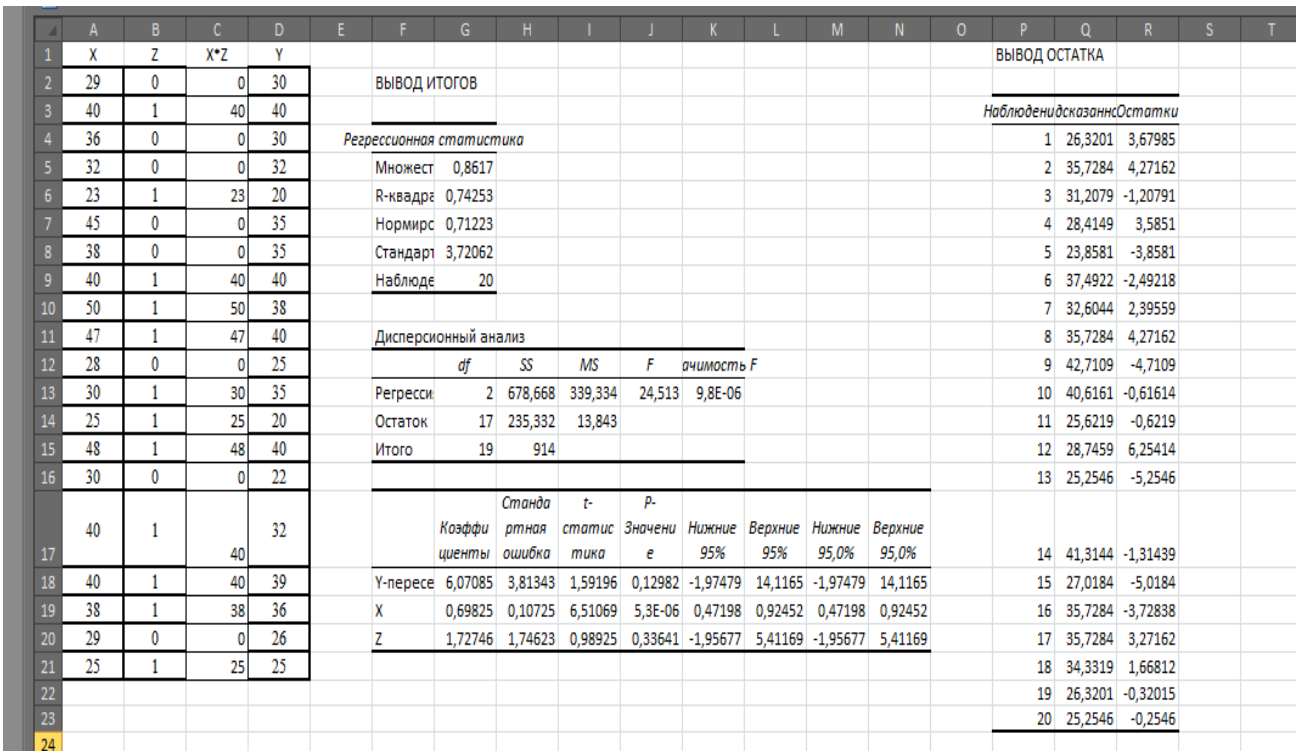
$$y = b_0 + b_1 x + b_2 z; \tag{8.6}$$

$$y = b_0 + b_1 \cdot x + b_2 \cdot z + b_3 \cdot (x \cdot z). \tag{8.7}$$

Первая из них позволяет проанализировать различие средних зарплат мужчин и женщин одинакового возраста. Вторая позволяет также проанализировать влияние пола работника на изменение средней зарплаты с увеличением возраста работника. Переменная $x \cdot z$ во второй модели представляет новую количественную объясняющую переменную.

Формирование и ввод выборочных данных с учетом фиктивных переменных. Для обеих моделей построим в MS Excel единую матрицу значений объясняющих переменных. В ячейках **A1-A21** расположим имя фактора X (возраст) и его выборочные значения, в ячейках **B1-B21** имя фиктивной переменной (пол) и его оцифрованные значения, в ячейках **C1-C21** имя новой переменной ($x \cdot z$) и его значения, в ячейках **D1-D21** имя зависимой переменной Y (зарплата) и его выборочные значения (см. рис. 8.1).

Оценка параметров модели. Оценка параметров обеих моделей производится с помощью функции «Регрессия» также как в лабораторной работе № 6. Результаты регрессии для первой модели (8.6) приведены на рис. 8.1, для второй модели (8.7) на рис. 8.2.



X	Z	X*Z	Y
29	0	0	30
40	1	40	40
36	0	0	30
32	0	0	32
23	1	23	20
45	0	0	35
38	0	0	35
40	1	40	40
50	1	50	38
47	1	47	40
28	0	0	25
30	1	30	35
25	1	25	20
48	1	48	40
30	0	0	22
40	1	40	32
40	1	40	39
38	1	38	36
29	0	0	26
25	1	25	25

ВЫВОД ИТОГОВ

Регрессионная статистика

Множест	0,8617
R-квадра	0,74253
Нормирс	0,71223
Стандарт	3,72062
Наблюде	20

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*ачимость F*
Регресси	2	678,668	339,334	24,513	9,8E-06
Остаток	17	235,332	13,843		
Итого	19	914			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересе	6,07085	3,81343	1,59196	0,12982	-1,97479	14,1165	-1,97479	14,1165
X	0,69825	0,10725	6,51069	5,3E-06	0,47198	0,92452	0,47198	0,92452
Z	1,72746	1,74623	0,98925	0,33641	-1,95677	5,41169	-1,95677	5,41169

ВЫВОД ОСТАТКА

Наблюдени	*дсказанн*	*Остатки*
1	26,3201	3,67985
2	35,7284	4,27162
3	31,2079	-1,20791
4	28,4149	3,5851
5	23,8581	-3,8581
6	37,4922	-2,49218
7	32,6044	2,39559
8	35,7284	4,27162
9	42,7109	-4,7109
10	40,6161	-0,61614
11	25,6219	-0,6219
12	28,7459	6,25414
13	25,2546	-5,2546
14	41,3144	-1,31439
15	27,0184	-5,0184
16	35,7284	-3,72838
17	35,7284	3,27162
18	34,3319	1,66812
19	26,3201	-0,32015
20	25,2546	-0,2546

Рис. 8.1. Регрессия с фиктивной переменной z

Оценка значимости коэффициентов уравнений регрессии и уравнений в целом, оценка качества построенных моделей. Из приведенных результатов следует, что оцененная модель (8.6) имеет вид

$$y = 6{,}071 + 0{,}698 \cdot x + 1{,}727 \cdot z,$$

а модель (8.7)

$$y = 8{,}864 + 0{,}614 \cdot x - 1{,}858 \cdot z + 0{,}105 \cdot (xz).$$

По F-критерию оба уравнения статистически значимы, но коэффициенты при фиктивной переменной z и новой переменной xz незначимо отличаются от нуля при уровне значимости α=0,05. Скорректированные коэффициенты детерминации, равные для первой модели $\bar{R}^2 = 0{,}712$ и для второй $\bar{R}^2 = 0{,}697$, говорят об удовлетворительном качестве подгонки. Следуя работе № 4, построим линейное уравнение регрессии без учета фактора «Пол». Полученное уравнение $y = 6{,}226 + 0{,}723 \cdot x$ значимо, так как значимость F для F-статистики, равная $1{,}8 \cdot 10^{-6}$, меньше заданного уровня значимости α=0,05. Скорректированный коэффициент детерминации равен 0,713. Результаты этой регрессии приведены на рис. 8.3.

	A	B	C	D	E	F	G	H	I
51	ВЫВОД ИТОГОВ								
52									
53	*ессионная статистика*								
54	Множеств	0,863058							
55	R-квадрат	0,744869							
56	Нормиров	0,697032							
57	Стандарт	3,817634							
58	Наблюде	20							
59									
60	Дисперсионный анализ								
61		*df*	*SS*	*MS*	*F*	*ачимость F*			
62	Регрессия	3	680,8107	226,9369	15,571	5,27E-05			
63	Остаток	16	233,1893	14,57433					
64	Итого	19	914						
65									
66		*Коэффициен*	*тартная ои*	*татистик*	*t-Значение*	*ижние 95%*	*ерхние 95%*	*ижние 95,0*	*рхние 95,0%*
67	Y-пересеч	8,864172	8,269797	1,071873	0,299679	-8,66701	26,39536	-8,66701	26,39536
68	X	0,614557	0,244462	2,513919	0,02302	0,096321	1,132792	0,096321	1,132792
69	Z	-1,85636	9,51749	-0,19505	0,847812	-22,0325	18,31982	-22,0325	18,31982
70	X*Z	0,104964	0,273767	0,383406	0,706465	-0,4754	0,685324	-0,4754	0,685324

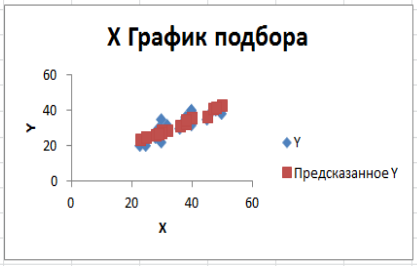


Рис. 8.2. Регрессия с переменными z и xz

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
100	ВЫВОД ИТОГОВ										ВЫВОД ОСТАТКА			
101														
102	*ессионная статистика*										*Наблюдени*	*дсказанно*	*Остатки*	
103	Множеств	0,85306									1	27,1923	2,80767	
104	R-квадрат	0,7277									2	35,1449	4,85513	
105	Нормиров	0,71258									3	32,253	-2,25304	
106	Стандарт	3,71841									4	29,3612	2,6388	
107	Наблюде	20									5	22,8546	-2,85458	
108											6	38,7597	-3,75966	
109	Дисперсионный анализ										7	33,699	1,30105	
110		*df*	*SS*	*MS*	*F*	*ачимость F*					8	35,1449	4,85513	
111	Регрессия	1	665,121	665,121	48,1045	1,8E-06					9	42,3744	-4,37445	
112	Остаток	18	248,879	13,8266							10	40,2056	-0,20557	
113	Итого	19	914								11	26,4694	-1,46937	
114											12	27,9153	7,08471	
115		*Коэффициен*	*тартная ои*	*татистик*	*t-Значение*	*ижние 95%*	*ерхние 95%*	*ижние 95,0*	*рхние 95,0%*		13	24,3005	-4,3005	
116	Y-пересеч	6,22655	3,80792	1,63516	0,11938	-1,77359	14,2267	-1,77359	14,2267		14	40,9285	-0,92853	
117	X	0,72296	0,10424	6,93574	1,8E-06	0,50396	0,94195	0,50396	0,94195		15	27,9153	-5,91529	
118											16	35,1449	-3,14487	
119											17	35,1449	3,85513	
120											18	33,699	2,30105	
121											19	27,1923	-1,19233	
122											20	24,3005	0,6995	
123														

Рис. 8.3. Регрессия Y на X

Проверка гипотезы об однородности выборок по критерию Чоу. Исходную выборку разделим на две части по значению признака «Пол». Результаты этого разбиения приведены на рис. 8.4. На каждой из выборок оценим линейную регрессию Y на X. Их результаты приведены на рис. 8.5, 8.6.

	A	B	C	D	E	F	G	H	I	J	K	L
124												
125												
126												
127												
128	Пол - мужской				Пол - женский							
129												
130	X	Y	Z		X	Y	Z					
131	40	40	1		29	30	0					
132	23	20	1		36	30	0					
133	40	40	1		32	32	0					
134	50	38	1		45	35	0					
135	47	40	1		38	35	0					
136	30	35	1		28	25	0					
137	25	20	1		30	22	0					
138	48	40	1		29	26	0					
139	40	32	1									
140	40	39	1									
141	38	36	1									
142	25	25	1									
143												

Рис. 8.4. Выборки, сгруппированные по признаку «Пол»

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
144																						
145																						
146	ВЫВОД ИТОГОВ			Пол - мужской									ВЫВОД ИТОГОВ			Пол - женский						
147																						
148	ессионная статистика												Регрессионная статистика									
149	Множеств	0,86354											Множеств	0,7687								
150	R-квадрат	0,74571											R-квадрат	0,5909								
151	Нормиров	0,72028											Нормиров	0,52272								
152	Стандарт	4,11608											Стандарт	3,26007								
153	Наблюден	12											Наблюден	8								
154																						
155	Дисперсионный анализ												Дисперсионный анализ									
156		df	SS	MS	F	ачимость F								df	SS	MS	F	ачимость F				
157	Регрессия	1	496,829	496,829	29,3251	0,00029							Регрессия	1	92,1067	92,1067	8,66637	0,02582				
158	Остаток	10	169,421	16,9421									Остаток	6	63,7683	10,6281						
159	Итого	11	666,25										Итого	7	155,875							
160																						
161	Коэффициен	гартная о	татисти	²-Значение	ижние 95%	ерхние 95%	жние 95,0	рхние 95,0%					Коэффициен	гартная о	татисти	²-Значение	ижние 95%	ерхние 95%	жние 95,0	рхние 95,0%		
162	Y-пересеч	7,00782	5,07923	1,3797	0,19775	-4,30942	18,325	-4,30942	18,325				Y-пересеч	8,86417	7,062	1,25519	0,25608	-8,41591	26,1443	-8,41591	26,1443	
163	X	0,71952	0,13287	5,41527	0,00029	0,42347	1,01557	0,42347	1,01557				X	0,61456	0,20876	2,94387	0,02582	0,10374	1,12537	0,10374	1,12537	
164																						

Рис. 8.5. Регрессии по отдельным выборкам

Вычислим значение F-статистики критерия Чоу и критическое значение F-статистики. В рассматриваемом примере объем выборки $n = 20$, число объясняющих переменных p=1, уровень значимости $\alpha = 0{,}05$. Для нахождения сумм квадратов остатков используем функцию **СУММКВ** в группе «Математические» вкладки «Формулы». Выделим ячейку **F183** и, учитывая формулу (8.5) и расположение остатков регрессий по полной выборке и ее частям, в строке формул введем

=(СУММКВ(M103:M122)-СУММКВ(C170:C181)- СУММКВ(O170:O177))* (20-2-2)/((СУММКВ(C170:C181)+СУММКВ(O170:O177))*(1+1))

По «Enter» в ячейке **F183** получим вычисленное значение *F*-статистики равное 0,538. Для нахождения критического значения *F*-статистики используем функцию **F.ОБР.ПХ** группы «Статистические». Выделим ячейку **H183** и, учитывая объем выборки n=20 и число объясняющих переменных p=1, в строке формул введем

=F.ОБР.ПХ(0,05;2;16)

По «Enter в ячейке» **H183** получим критическое значение *F*-статистики равное 3,634, см. рис. 8.6. Вычисленное значение *F*-статистики меньше критического, следовательно, нулевая гипотеза об однородности отдельных выборок принимается.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
165																	
166																	
167	ВЫВОД ОСТАТКА		Пол - мужской										ВЫВОД ОСТАТКА		Пол - женский		
168																	
169	*аблюдени*	*дсказанн*	*Остатки*										*Наблюдени*	*дсказанн*	*Остатки*		
170	1	35,78864	4,211358										1	26,68631	3,313685		
171	2	23,55679	-3,55679										2	30,98821	-0,98821		
172	3	35,78864	4,211358										3	28,52998	3,470015		
173	4	42,98385	-4,98385										4	36,51922	-1,51922		
174	5	40,82529	-0,82529										5	32,21732	2,782676		
175	6	28,59344	6,406565										6	26,07176	-1,07176		
176	7	24,99583	-4,99583										7	27,30087	-5,30087		
177	8	41,54481	-1,54481										8	26,68631	-0,68631		
178	9	35,78864	-3,78864														
179	10	35,78864	3,211358														
180	11	34,3496	1,650399														
181	12	24,99583	0,004168				Критерий Чоу										
182						F-статистика		F-критическое									
183						0,538257		3,633723									

Рис. 8.6. Проверка однородности выборок по тесту Чоу

Интерпретация модели и общее заключение о проведенном исследовании.

Все три построенных уравнения регрессии заработной платы (y) от возраста (x) с учетом и без учета пола (z) работников

$$y = 6{,}071 + 0{,}698 \cdot x + 1{,}727 \cdot z,$$

$$y = 8{,}864 + 0{,}614 \cdot x - 1{,}858 \cdot z + 0{,}105 \cdot (xz),$$

$$y = 6{,}226 + 0{,}723 \cdot x$$

статистически значимы при заданном уровне значимости $\alpha = 0{,}05$ и имеют примерно одинаковые аппроксимационные свойства. Их нормированные коэффициенты детерминации равны соответственно 0,712, 0,697, 0,712. Во всех построенных уравнениях регрессии возраст x работников значимо влияет на зара-

ботную плату. Так как *P*-значения для коэффициентов регрессии при x, равные соответственно $5{,}3 \cdot 10^{-6};\ 0{,}023;\ 1{,}8 \cdot 10^{-6}$, меньше заданного уровня значимости α. В первом и втором уравнениях *P*-значения для коэффициентов регрессии при поле работников и факторе xz значительно превышают заданный уровень значимости. Следовательно, пол z работников и фактор xz не оказывают значимого влияния на заработную плату работников.

Согласно первой модели при одинаковом возрасте средняя заработная плата работников мужчин на 1,727 тыс. р. больше чем у женщин. С увеличением возраста на один год средняя заработная плата возрастает примерно на 0,723 тыс. р. по третьей модели и на 0,698 тыс. р. по первой. Уравнение, включающее произведение факторов «Возраст» и «Пол», имеет несколько худшее качество подгонки.

Анализ однородности отдельных выборок для мужчин и женщин по тесту Чоу показал их однородность. Поэтому эти выборки можно объединить в одну и использовать уравнение регрессии $y = 6{,}226 + 0{,}723 \cdot x$, построенное по объединенной выборке.

Контрольные вопросы

1. Для учета влияния каких факторов используются фиктивные переменные в моделях регрессии?
2. Какие значения может принимать бинарная фиктивная переменная?
3. Сколько фиктивных переменных следует ввести в модель для учета региональных различий, если данные собраны по пяти регионам?
4. Как используются фиктивные переменные для моделирования сезонного фактора?
5. Какие из перечисленных факторов учитываются в регрессии с помощью фиктивных переменных: 1) профессия, 2) курс доллара, 3) численность населения, 4) размер среднемесячных потребительских расходов, 5) местоположение пункта продажи?
6. С помощью фиктивных переменных напишите уравнение, соответствующее наличию двух структурных изменений в моменты времени t_0 и t_1, $t_0 < t_1$.
7. Может ли уравнение регрессии в качестве объясняющих переменных содержать только фиктивные переменные?
8. Каким методом осуществляется оценка моделей регрессии с фиктивными переменными?
9. Как формулируется гипотеза об однородности двух выборок в регрессионном смысле?
10. Как осуществляется проверка на однородность в регрессионном смысле двух выборок по критерию Чоу?
11. Как учитывается влияние качественного фактора на коэффициент регрессии?

Лабораторная работа № 9. Выделение тенденции временного ряда: скользящая средняя; экспоненциальное сглаживание

Цель работы. Освоение основных понятий анализа одномерных временных рядов, методов выделения тенденции временного ряда с использованием пакета анализа MS Excel 2010.

Краткие сведения. *Временной ряд* – это совокупность значений $(y_1, y_2, y_3, \dots, y_n)$ некоторого числового показателя за несколько последовательных моментов или периодов времени t, характеризующая состояние и изменение изучаемого явления. Моменты времени t на оси времени располагаются через одинаковые промежутки, а периоды времени одинаковой длины. Значения y_t показателя в момент *t* или период времени t называются *уровнями временного ряда. Моментный временной ряд* – уровни временного ряда характеризуют изучаемое явление в конкретные последовательные (равноотстоящие) моменты времени. *Интервальный временной ряд* – уровни временного ряда характеризуют изучаемое явление в последовательные равные промежутки времени. Существенным отличием временных рядов от пространственных данных является статистическая зависимость значений показателя в момент времени t от его значений в предшествующие моменты времени.

Уровни изучаемого временного ряда должны быть:

− однородными по экономическому содержанию и отражать существо изучаемого явления и цель исследования;

− измеренными по единой методике и в единых единицах измерения;

− не содержать аномальных (значительно отличающихся от других) наблюдений.

Уровни y_t временного ряда формируются из следующих компонент:

− ***тенденции (тренда) T*** – характеризует изменение явления (процесса), происходящее в некотором направлении в течение значительного промежутка времени (описывает чистое влияние долговременных факторов);

− ***циклической компоненты C*** – отражает повторяемость экономических процессов в течение длительных периодов, представляет собой более быстрые, чем тенденция квазипериодические колебания изучаемого признака;

− ***сезонных колебания S*** – отражает регулярную повторяемость экономических процессов в течение не очень длительных промежутков времени. Связаны, например, со сменой времен года и ритмами человеческой активности;

− ***случайной компоненты ε*** – отражающей влияние не поддающихся учету и регистрации случайных факторов.

Циклическая и сезонная компоненты характеризуют колебания уровней временного ряда относительно основной тенденции, а случайная компонента – случайный разброс уровней относительно тенденции и сезонной (и/или циклической) составляющей. Тенденция, циклическая и сезонные компоненты, при их наличии, определяют детерминированную часть уровней временного ряда.

Основные описательные статистики временных рядов средняя и дисперсия рассчитываются по обычным формулам:

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t; \qquad s^2 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2.$$

Для характеристики корреляционной зависимости между последовательными уровнями временного ряда, отстоящими друг от друга на k промежутка времени, используется выборочный коэффициент автокорреляции k-го порядка

$$r_k = \frac{\sum_{t=k+1}^{n}(y_t-\bar{y}_1)(y_{t-k}-\bar{y}_2)}{\sqrt{\sum_{t=k+1}^{n}(y_t-\bar{y}_1)^2 \cdot \sum_{t=k+1}^{n}(y_{t-k}-\bar{y}_2)^2}},$$

$$\bar{y}_1 = \frac{1}{n-k}\sum_{t=k+1}^{n} y_t, \quad \bar{y}_2 = \frac{1}{n-k}\sum_{t=k+1}^{n} y_{t-k}.$$

Последовательность значений r_1 ,r_2, r_3, … выборочных коэффициентов автокорреляции называют *выборочной автокорреляционной функцией* (*коррелограммой*) временного ряда, аргументом которой является величина k.

Основные задачи анализа временных рядов заключаются:

− в определении его структуры (из каких компонент состоят уровни временного ряда);

− в выделении и придании количественного описания каждой его компоненте;

− построение математической модели процесса, представленного временным рядом;

− в построении прогноза будущих значений временного ряда.

В зависимости от характера колебаний уровней относительно тренда временного ряда различают:

аддитивную модель тренда и сезонности $y_t = T_t + S_t + \varepsilon_t$. Она применяется при приблизительно одинаковой амплитуде периодических колебаний уровней, обусловленных наличием сезонной и/или циклической компоненты, вокруг тренда;

мультипликативную модель тренда и сезонности $y_t = T_t \cdot S_t \cdot \varepsilon_t$. Применяется при возрастающей или убывающей амплитуде сезонных или циклических колебаний уровней вокруг тренда.

Построение аддитивной и мультипликативной модели сводится к оценке значений их компонент T_t, S_t, ε_t для каждого уровня ряда.

Методы распознавания наличия тренда и его типа.

Графический метод. Графическое изображение временного ряда часто позволяет установить наличие тренда и его тип: линейный, нелинейный (параболический, степенной, экспоненциальный, логарифмический, гиперболический, логистический).

Методы сглаживания и фильтрации предназначены для преобразования временных рядов с целью удаления из них высокочастотных или сезонных колебаний уровней ряда относительно тренда и выделения основной тенденции (тренда). В этом подходе различают *методы скользящей средней* и *аналитического выравнивания*.

В методах скользящих средних наблюдаемые значения уровней временного ряда заменяются средними их значениями, вычисляемыми на отрезке (за несколько последовательных моментов времени) скользящем вдоль временного

ряда. Отклонения уровней от тренда, вызванные сезонными и высокочастотными колебаниями, имеют разные знаки, и усреднение позволяет исключить эти колебания из уровней временного ряда и выделить тренд.

Аналитическое выравнивание заключается в оценивании тренда как некоторой явной функции времени t, т.е. $T_t = f(t)$.

Метод корреляционного анализа. Основывается на анализе выборочной автокорреляционной и частной автокорреляционной функций временного ряда.

Метод проверки статистических гипотез о типе тренда. Этот метод основывается на вычислении средних характеристик динамики процесса (абсолютного прироста, абсолютного ускорения, темпа роста, темпа прироста, эластичности) на отдельных непересекающихся частях временного ряда и проверки гипотезы о незначимости их различия. Если такая гипотеза принимается, то принимается и решение о наличии соответствующего тренда. Например, принятие гипотезы о равенстве средних абсолютных приростов $\Delta y_t = y_{t+1} - y_t$ на разных частях временного ряда говорит о наличии линейного тренда $y_t = a + b \cdot t$, а принятие гипотезы о равенстве средних темпов прироста говорит о наличии экспоненциального тренд $y_t = \exp(a + b \cdot t)$.

В данной работе рассматривается выделение тренда временного ряда с использованием метода скользящей средней и экспоненциального сглаживания.

Простая скользящая средняя. Среднее уровней ряда, попавших в отрезок (окно) скольжения, вычисляется как обычное выборочное среднее. Отрезок скольжения, на котором вычисляется текущее среднее уровней временного ряда, может содержать нечетное 2k+1 или четное 2k количество уровней ряда.

При нечетной 2k+1 длине окна скольжения, вычисленное среднее определяет значение тренда в средней (k+1)-й точке этого окна. Т.е. значения тренда T_t для моментов времени $k+1 \le t \le n-k$ определяется как $T_t = \frac{1}{2k+1}\sum_{j=-k}^{k} y_{t+j}$. Например, при длине окна скольжения равном 5 (k=2) среднее $\frac{1}{5}(y_1 + y_2 + y_3 + y_4 + y_5)$ пяти первых уровней принимается за значение тренда для момента времени $t = 3$. Затем окно сдвигается вправо на один момент времени и среднее $\frac{1}{5}(y_2 + y_3 + y_4 + y_5 + y_6)$ уровней ряда, попавших в это окно, принимается за значение тренда для момента времени $t = 4$. Последовательно сдвигая окно скольжения на один шаг вправо, получают значения тренда для следующих моментов времени. Последнее значение тренда для момента времени $t = n - 2$ определяется как $T_{n-2} = \frac{1}{5}(y_{n-4} + y_{n-3} + y_{n-2} + y_{n-1} + y_n)$.

При четной 2k длине окна скольжения (например, для квартальных данных имеющих сезонную компоненту), вычисленное среднее определяет значение тренда для середины промежутка времени между k-м и (k+1)-м моментами времени, вошедшими в окно скольжения. Эти промежуточные моменты времени не присутствуют во временном ряде. Поэтому, после нахождения значений тренда в промежуточных точках, значения тренда для моментов времени рассматриваемых во временном ряде определяются как среднее двух значений тренда в соседних промежуточных точках. Например, при длине окна скольжения равном четырем (k=2) значения тренда сначала вычисляются для промежуточных моментов времени 2,5; 3,5; 4,5; …; n-2,5. Затем вычисляется значение

тренда для $t = 3$ как среднее значений тренда для моментов времени 2,5 и 3,5; значение тренда для $t = 4$ как среднее значений тренда для моментов времени 3,5 и 4,5 и так далее.

Взвешенная скользящая средняя. В этом случае значение тренда для момента времени соответствующего середине окна скольжения с нечетной длиной определяется как взвешенное среднее уровней ряда, попавших в окно. Т.е. значения T_t тренда для моментов времени $k + 1 \le t \le n - k$ определяется как

$$T_t = \frac{1}{2k+1}\sum_{j=-k}^{k} \beta_j \cdot y_{t+j}.$$

Весовые коэффициенты β_j определяются в зависимости от длины 2k+1 окна скольжения и степени p полинома $\sum_{i=0}^{p} \alpha_i \cdot \tau^i$, $(\tau = -k, \dots, 0, \dots, k)$, используемого для аппроксимации уровней ряда внутри окна скольжения. При длине окна скольжения равной 5 и степени p аппроксимирующего полинома равной 2 или 3 весовые коэффициенты имеют значения $\beta_1 = -\frac{3}{35}$, $\beta_2 = \frac{12}{35}$, $\beta_3 = \frac{17}{35}$, $\beta_4 = \frac{12}{35}, \beta_5 = -\frac{3}{35}$. Для длины окна скольжения равной 7 и степени p аппроксимирующего полинома равной 2 или 3 весовые коэффициенты имеют значения $\beta_1 = \frac{2}{21}$, $\beta_2 = \frac{3}{21}, \beta_3 = \frac{6}{21}, \beta_4 = \frac{7}{21}, \beta_5 = \frac{6}{21}, \beta_6 = \frac{3}{21}, \beta_7 = -\frac{2}{21}$.

Метод простой скользящей средней дает хорошие результаты для временных рядов с линейной тенденцией. Для рядов с нелинейной тенденцией необходимо применять метод взвешенной скользящей средней.

Метод экспоненциального сглаживания. При построении прогноза S_{t+1} уровня временного ряда для момента времени *t+1* часто используется взвешенная сумма его предшествующих уровней $S_{t+1} = \sum_{j=0}^{N} \alpha_j \cdot y_{t-j}$. Очевидно, что более поздние наблюдения y_i играют большую роль при формировании уровня ряда для момента времени *t+1,* чем более ранние наблюдения. Более ранние наблюдения содержат «старую» тенденцию, а более поздние наблюдения отражают «новую» тенденцию. Один из способов уменьшения роли более ранних наблюдений заключается в том, что коэффициенты α_j образуют убывающую геометрическую прогрессию, $\alpha_j = \alpha \cdot \beta^j$, и прогноз S_{t+1} определяется соотношением

$$S_{t+1} = \alpha \cdot (y_t + \beta \cdot y_{t-1} + \beta^2 \cdot y_{t-2} + \dots + \beta^N \cdot y_{t-N}),$$

где $\alpha \in (0; 1), \beta = 1 - \alpha$. Правая часть этого соотношения называется *экспоненциально взвешенным скользящим средним*, а построение прогнозов по этому соотношению называют *экспоненциальным сглаживанием* временного ряда, α – *параметр сглаживания*.

Путем несложных преобразований экспоненциальное сглаживание сводится к рекуррентной формуле

$$S_{t+1} = S_t + \alpha \cdot (y_t - S_t),$$

выражающей значение экспоненциального среднего S_{t+1} как суммы экспоненциального среднего S_t предыдущего момента времени и доли α разницы текущего наблюдения y_t и S_t. Применение экспоненциального сглаживания связано с выбором значения параметра сглаживания α и начального значения S_0. При α

близком к единице веса предшествующих наблюдений быстро убывают и на прогноз S_{t+1} оказывают большое влияние только последние наблюдения. Это приводит к малым расхождениям сглаженных значений от наблюдаемых значений ряда. При α близком к нулю в прогнозе S_{t+1} веса предшествующих наблюдений убывают медленно, что приводит к отфильтровыванию случайных колебаний уровней временного ряда. В практических расчетах значение параметра сглаживания определяют, как $\alpha = \frac{2}{N+1}$ или как $\alpha = \frac{2}{n+1}$, где N – число наблюдений, входящих в интервал сглаживания, n – число наблюдений во временном ряде. За начальное сглаженное значение S_0 принимается первое значение y_1 временного ряда.

Использование экспоненциального сглаживания для выравнивания временного ряда оправдано для временных рядов с незначительным сезонным эффектом.

Прогноз уровней ряда методом экспоненциального сглаживания для будущих моментов времени: $\hat{y}_{n+1} = S_{n+1} = S_n + \alpha \cdot (y_n - S_n)$, $\hat{y}_{n+2} = S_{n+2} = S_{n+1} + \alpha \cdot (\hat{y}_{n+1} - S_{n+1}) = S_{n+1} = \hat{y}_{n+1}$. Таким образом, прогноз будущих значений для моментов времени *n+2, n+3* и так далее совпадает с прогнозом $\hat{y}_{n+1}$.

Содержание лабораторной работы.

1. Ввод временного ряда, построение его графика, анализ структуры ряда по его графику.

2. Методом простой скользящей средней выделить тренд временного ряда для значений длины окна скольжения равных 4 и 5. Построить графики выравненных уровней ряда и исходного ряда.

3. Выделить тенденцию методом взвешенной скользящей средней для длины окна сглаживания равной 5, построить графики выравненных уровней ряда и исходного ряда.

4. Выделить тренд методом экспоненциального сглаживания для различных значений параметра сглаживания. Построить графики выравненных уровней ряда и исходного ряда. Вычислить прогноз уровня временного ряда для двух последующих моментов времени.

5. Построить временной ряд отклонений уровней исходного ряда от его тренда, построенного одним из рассмотренных способов.

6. Сформулировать общее заключение о структуре временного ряда, характере его тенденции и характере колебаний уровней ряда относительно тренда.

Выполнение работы в MS Excel.

Рассмотрим анализ временного ряда $\{y_t\}$ квартальных объемов выпуска продукции предприятием за ряд лет в сопоставимых ценах (млн р.) по данным приведенным в следующей таблице 9.1.

Таблица 9.1

Квартал	Год				
	1	2	3	4	5
I	601,2	666,8	826,2	754,1	688,8
II	639,7	665,3	812,0	711,8	658,3
III	647,9	678,4	752,3	656,0	635,8
IV	623,0	693,7	774,1	690,0	660,0

Ввод временного ряда, построение его графика, анализ структуры ряда по его графику. Расположим временной ряд в ячейках A1-A21, ряд содержит 20 уровней, *n=20*. Для построения графика временного ряда выделим ячейки A2-A21. Во вкладке **«Вставка»**, выбрав вид графика **«точечная с маркерами»**, получим график временного ряда. Из графика видно, что уровни ряда в начальном периоде от *t=1* до *t=10* имеют тенденцию возрастания, а затем тенденцию убывания. Уровни ряда содержат колебания, но их характер не говорит о наличии сезонной компоненты связанной с кварталами года, см. рис. 9.1.

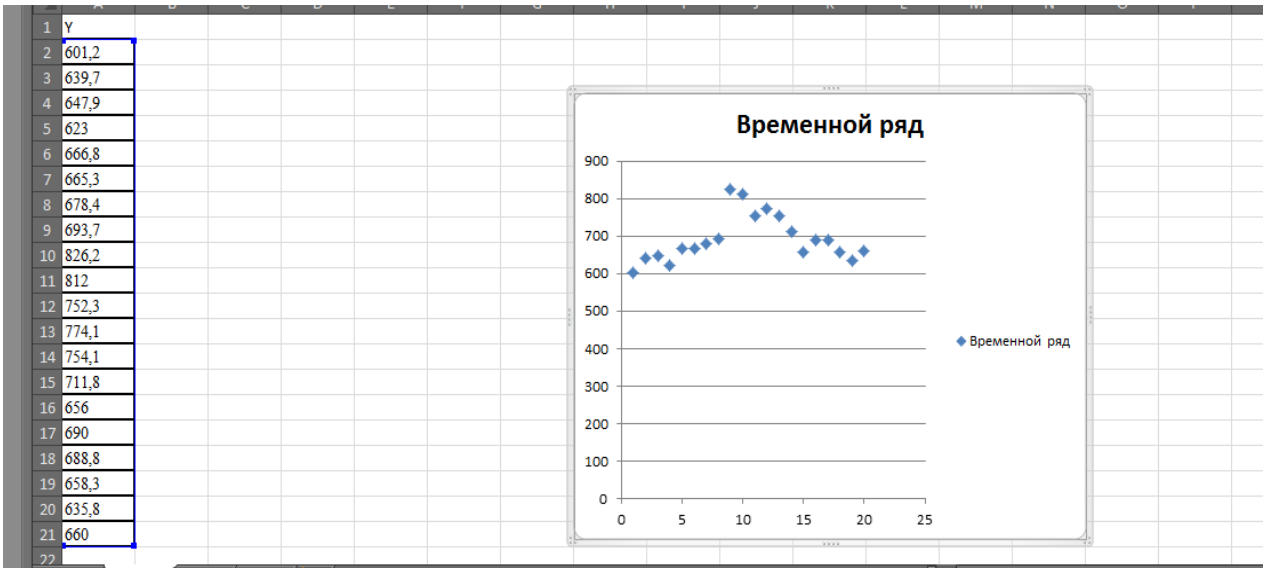


Рис. 9.1. Временной ряд и его график

Выделение тренда временного ряда методом простой скользящей средней с длиной окна скольжения равной 4. По вкладке **«Данные»** выберем **«Анализ данных»**, в окне «Анализа данных» выберем **«Скользящее среднее»**. Заполнение полей окна «Скользящее среднее»:

− в поле *входной интервал* укажем ячейки **A1-A21**, содержащие временной ряд;

− выберем метки в первой строке, так как в ячейке **A1** имя временного ряда;

− в поле *интервал* укажем длину окна скольжения равную 4;

− в поле *выходной интервал*, для удобства построения графиков ряда и скользящих средних, укажем ячейку **I2**;

− выберем вывод графика; при необходимости можно также выбрать «Стандартные погрешности», в этом случае наряду со сглаженным рядом выводятся и вычисленные по окну скольжения средние квадратические отклонения уровней ряда от сглаженных значений $e_t = \sqrt{\frac{\sum_{i=0}^{T-1}(y_{t-i}-\bar{y}_{t-i})^2}{T}}$, где T – длина окна скольжения, $\bar{y}_{t-i}$ – скользящее среднее для момента времени *t-i*.

По «ОК» в ячейках **I2-I21** получим сглаженные значения временного ряда, в ячейках **J2-J21** значения средних квадратических отклонений e_t, см. рис. 28. Следует учесть, что процедура сглаживания относит сглаженные зна-

чения уровней, попавших в окно, к последнему моменту времени в окне скольжения. Поэтому в данном случае сглаженные значения располагаются начиная с ячейки **I5**. Необходимо рассчитать сглаженные значения для моментов времени находящихся в середине окна скольжения. В данном случае длина окна скольжения четная и равна 4. Значения скользящей средней для третьего момента времени в окне скольжения находятся как среднее арифметическое скользящего среднего для данного и следующего положения окна скольжения на временной оси. Для этого выделим ячейку **B4** и в строке формул введем **=СРЗНАЧ(I5:I6)**, по «ОК» получим значение скользящей средней для момента времени *t=3*. Аналогично в ячейках B5-B19 получаем значения скользящей средней для последующих моментов времени, до момента времени *n-2* включительно, см. рис. 9.2.

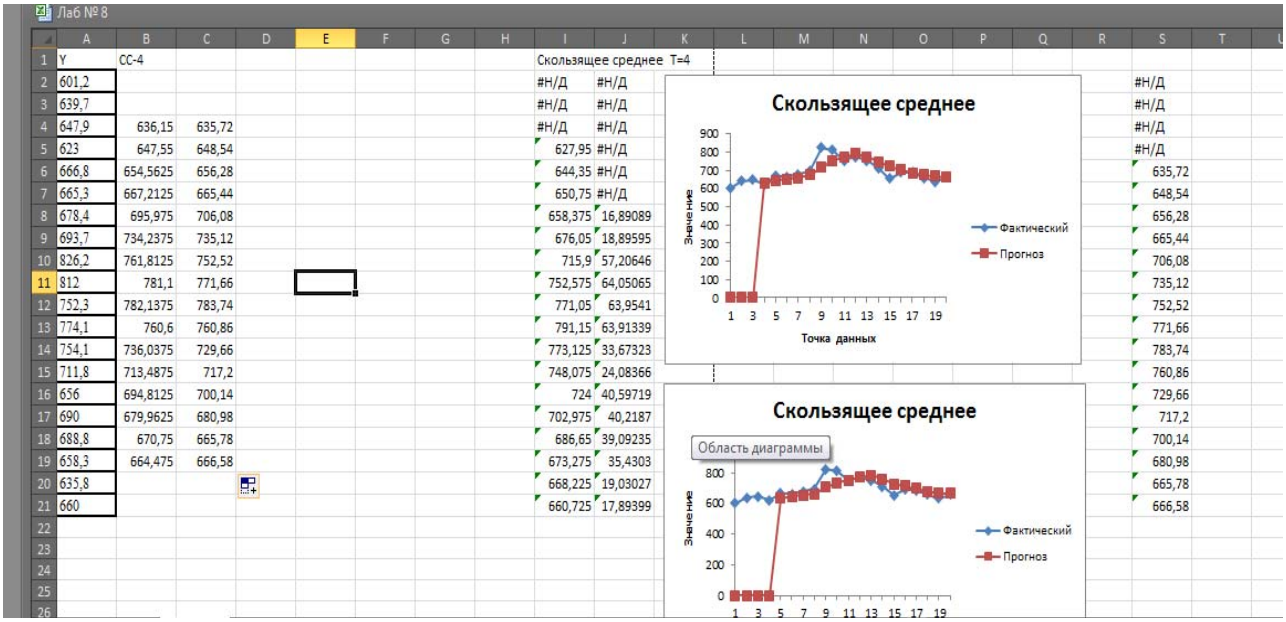


Рис. 9.2. Построение скользящей средней с длиной окна равной 4 и 5

Построение скользящей средней, с нечетной длиной окна скольжения равной 5, производится аналогичным образом, результат приведен в ячейках **S2-S21**, см. рис. 9.2. Среднее значение уровней, попавших в интервал сглаживания, приписывается последнему моменту времени, вошедшему в интервал сглаживания. Поэтому их нужно переместить в середины соответствующих интервалов. Для этого выделим ячейку **C4** и в строке формул введем **=S6**, тем самым среднее первых пяти уровней относим к моменту времени *t=3*. Аналогично в ячейках **C5-C19** получаем значения скользящей средней для последующих моментов времени, см. рис. 9.2.

Графики скользящих средних и исходного ряда. Для их построения в вкладке **«Вставка»** выберем вид графика **«График с маркерами»** и по «Вводу данных» в окне «Выбор источника данных» укажем **A2-C21**. Средствами MS Excel изменим названия легенд рядов. На рис. 9.3 приведены графики исходного временного ряда и скользящих средних с длиной окна скольжения 4 и 5.

Графики скользящих средних не содержат колебаний и показывают изменение тенденции с возрастания на убывание в момент времени *t=11*.

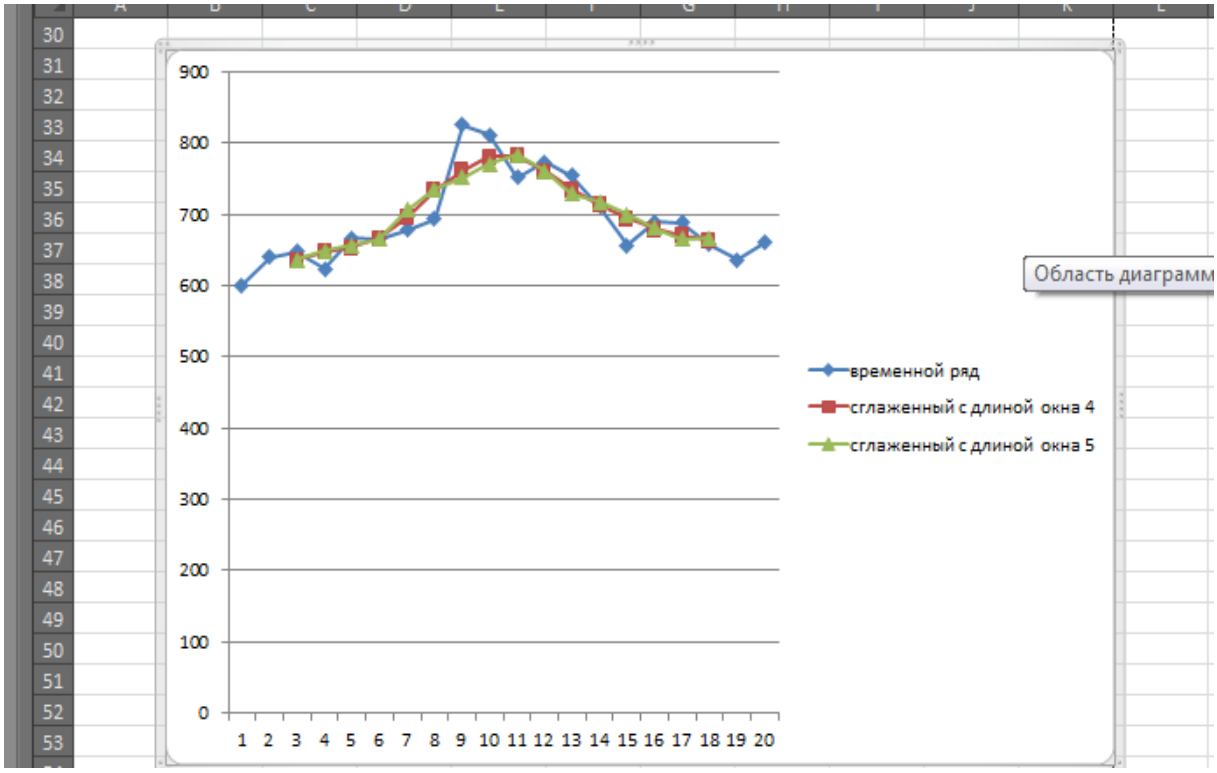
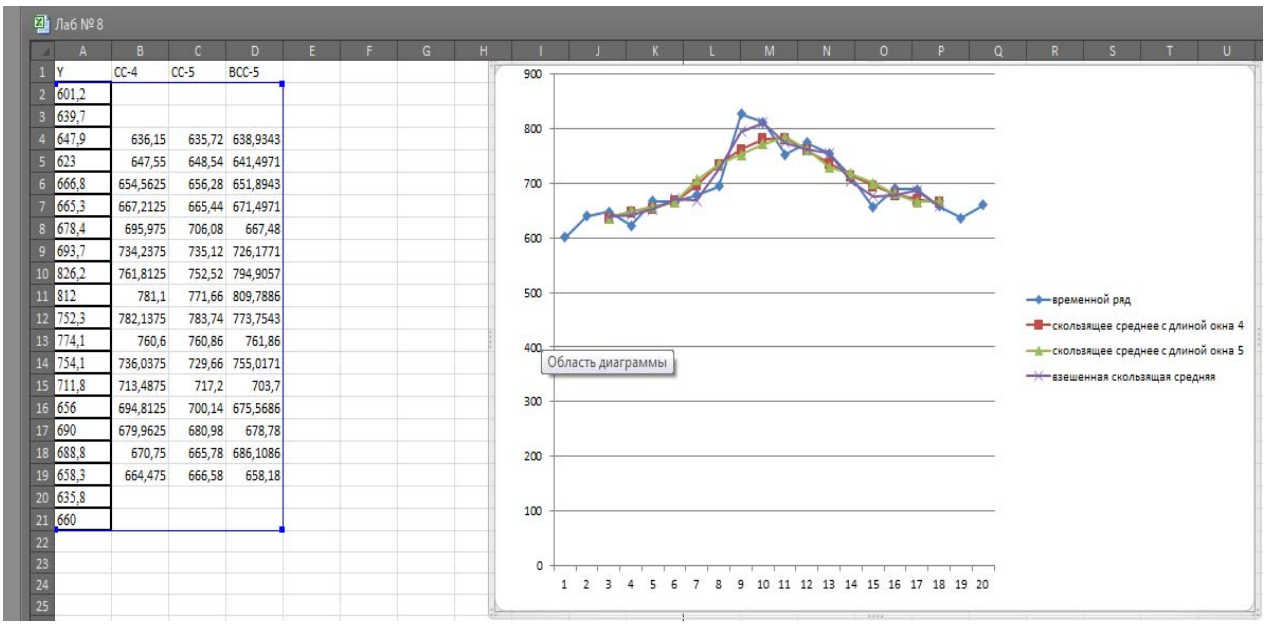


Рис. 9.3. Графики временного ряда и скользящих средних

Выделение тенденции ряда методом взвешенной скользящей средней для длины окна сглаживания равной 5 с аппроксимирующим полиномом второй степени. Выделим ячейку **D4** и введя в строке формул

=(-3*A2+12*A3+17*A4+12*A5-3*A6)/35

получим в этой ячейке первое значение взвешенной скользящей средней. В ячейках **D5-D19** аналогичным образом получаем значения взвешенной скользящей средней при смещении окна скольжения вдоль оси времени, см. рис. 9.4. График взвешенной скользящей средней вместе с предыдущими графиками приведен на рис. 9.4. Взвешенная скользящая средняя лучше, чем простые скользящие средние, согласуется с исходным временным рядом, показывает изменение тенденции в момент времени *t=11* и хуже сглаживает колебания уровней временного ряда.



	A	B	C	D
1	Y	СС-4	СС-5	ВСС-5
2	601,2			
3	639,7			
4	647,9	636,15	635,72	638,9343
5	623	647,55	648,54	641,4971
6	666,8	654,5625	656,28	651,8943
7	665,3	667,2125	665,44	671,4971
8	678,4	695,975	706,08	667,48
9	693,7	734,2375	735,12	726,1771
10	826,2	761,8125	752,52	794,9057
11	812	781,1	771,66	809,7886
12	752,3	782,1375	783,74	773,7543
13	774,1	760,6	760,86	761,86
14	754,1	736,0375	729,66	755,0171
15	711,8	713,4875	717,2	703,7
16	656	694,8125	700,14	675,5686
17	690	679,9625	680,98	678,78
18	688,8	670,75	665,78	686,1086
19	658,3	664,475	666,58	658,18
20	635,8			
21	660			

Рис. 9.4. Графики скользящих средних

Экспоненциальное сглаживание, учитывая колебания уровней относительно выделенной ранее тенденции, проведем для значений параметра сглаживания α равной 0,1 и 0,3. В окне анализа данных выберем инструмент анализа **«Экспоненциальное сглаживание»**. В полях окна этого инструмента введем:

− в поле *входной интервал* укажем ячейки **A1-A21**, содержащие временной ряд;

− выберем метки в первой строке, так как в ячейке **A1** имя временного ряда;

− в поле *фактор затухания* зададим число равное 1-α (в примере соответственно 0,9 и 0,7);

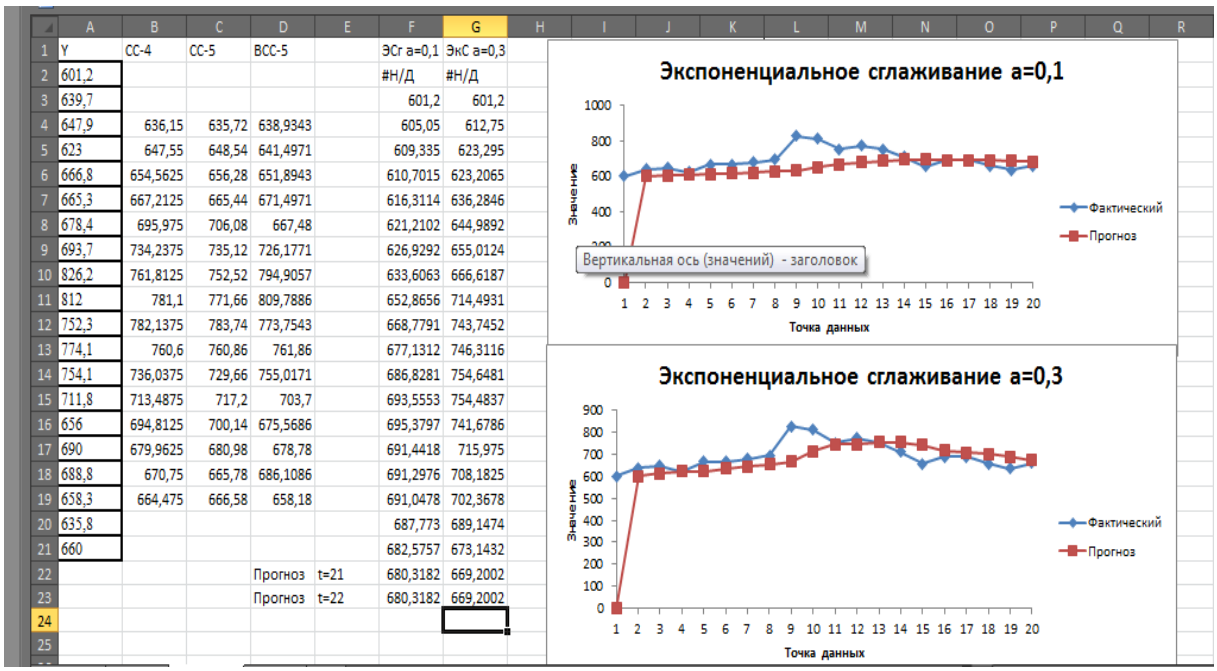
− в поле *выходной интервал* укажем, например, ячейку **F2** для α=0,1 и **G2** для α=0,3;

− выберем вывод графика.

При необходимости можно также выбрать «Стандартные погрешности», в этом случае наряду со сглаженным рядом выводятся и средние квадратические отклонения $e_t = \sqrt{\frac{\sum_{i=1}^{3}(y_{t-i}-\bar{y}_{t-i})^2}{3}}$ уровней ряда от сглаженных значений, где $\bar{y}_{t-i}$ – экспоненциальное скользящее среднее для момента времени *t-i*.

Результаты экспоненциального сглаживания приведены на рис. 9.5.

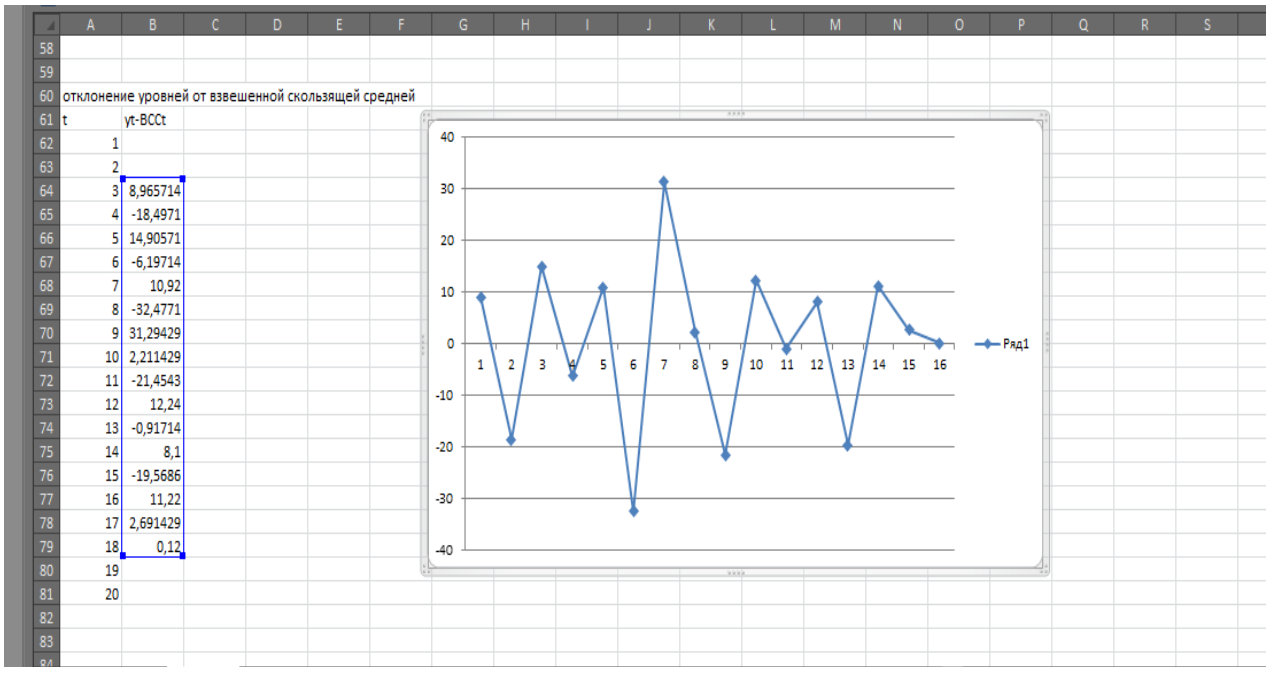
Прогноз уровней временного ряда по формуле экспоненциального сглаживания $S_{t+1} = S_t + \alpha \cdot (y_t - S_t)$, для последующих моментов времени *t=21* и *t=22* рассчитаем в ячейках **F22-F23** и **G22-G23**, см. рис. 9.5.



	Y	СС-4	СС-5	ВСС-5		ЭСг a=0,1	ЭкС a=0,3
	601,2					#Н/Д	#Н/Д
	639,7					601,2	601,2
	647,9	636,15	635,72	638,9343		605,05	612,75
	623	647,55	648,54	641,4971		609,335	623,295
	666,8	654,5625	656,28	651,8943		610,7015	623,2065
	665,3	667,2125	665,44	671,4971		616,3114	636,2846
	678,4	695,975	706,08	667,48		621,2102	644,9892
	693,7	734,2375	735,12	726,1771		626,9292	655,0124
	826,2	761,8125	752,52	794,9057		633,6063	666,6187
	812	781,1	771,66	809,7886		652,8656	714,4931
	752,3	782,1375	783,74	773,7543		668,7791	743,7452
	774,1	760,6	760,86	761,86		677,1312	746,3116
	754,1	736,0375	729,66	755,0171		686,8281	754,6481
	711,8	713,4875	717,2	703,7		693,5553	754,4837
	656	694,8125	700,14	675,5686		695,3797	741,6786
	690	679,9625	680,98	678,78		691,4418	715,975
	688,8	670,75	665,78	686,1086		691,2976	708,1825
	658,3	664,475	666,58	658,18		691,0478	702,3678
	635,8					687,773	689,1474
	660					682,5757	673,1432
				Прогноз	t=21	680,3182	669,2002
				Прогноз	t=22	680,3182	669,2002

Рис. 9.5. Экспоненциальное сглаживание

Временной ряд и график отклонений уровней y_t исходного ряда от его тренда, построенного, например, методом взвешенной скользящей средней, приведен на рис. 9.6.



t	yt-ВССt
1	
2	
3	8,965714
4	-18,4971
5	14,90571
6	-6,19714
7	10,92
8	-32,4771
9	31,29429
10	2,211429
11	-21,4543
12	12,24
13	-0,91714
14	8,1
15	-19,5686
16	11,22
17	2,691429
18	0,12
19	
20	

Рис. 9.6. Ряд отклонений уровней ряда от взвешенной скользящей средней

Общее заключение. Проведенное сглаживание уровней временного ряда квартальных объемов выпуска продукции говорит о тенденции возрастания в начальном периоде и тенденции убывания в конечном периоде. Изменение тенденции происходит в момент времени *t=10*. Графики сглаженных уровней по-

казывают, что простые скользящие средние с окном скольжения 4 и 5 хорошо отражают тенденцию возрастания квартальных объемов выпуска продукции в начальном периоде и тенденцию их убывания в конечном периоде, но сдвигают на шаг вправо точку изменения тенденции.

Взвешенная скользящая средняя лучше согласуется с исходным временным рядом, также сдвигает вправо на шаг точку изменения тенденции и хуже, чем простые скользящие средние сглаживает колебания уровней временного ряда.

Экспоненциальное сглаживания с параметром сглаживания $\alpha=0,1$ занижает тенденцию возрастания в начальном периоде и тенденцию убывания в конечном периоде, а также значительно сдвигает вправо точку изменения характера тенденции. Экспоненциальное сглаживания с параметром сглаживания $\alpha=0,3$ также занижает тенденцию возрастания в начальном периоде и значительно сдвигает вправо точку изменения характера тенденции. В обеих случаях экспоненциальное сглаживание хорошо сглаживает колебания уровней временного ряда.

Из рассмотренных методов сглаживания и фильтрации тенденцию временного ряда квартальных объемов выпуска продукции лучше представляют простые скользящие средние с окном скольжения 4 или 5.

Колебания уровней относительно тенденции в начальный период, от $t=1$ до $t=8$, имеют сезонный характер с возрастающей амплитудой колебаний. При t большем 8 характер колебаний изменяется, поэтому нельзя говорить о наличии сезонных колебаний во всем временном ряде квартальных объемов выпуска продукции.

Контрольные вопросы

1. Дайте определение временного ряда, приведите примеры моментных и интервальных временных рядов.
2. Из каких компонент формируются уровни временного ряда?
3. Что характеризует автокорреляция уровней временного ряда? Как она вычисляется?
4. Какие требования предъявляются к временным рядам?
5. Перечислите основные задачи анализа временных рядов.
6. Приведите аддитивную и мультипликативную модели временного ряда, содержащего тенденцию и сезонные колебания.
7. Перечислите способы выявления тренда временного ряда.
8. Как определяется простая скользящая средняя временного ряда?
9. Как определяется взвешенная скользящая средняя временного ряда?
10. Осуществим ли прогноз уровней временного ряда для будущих моментов времени при использовании простых скользящих средних?
11. В чем заключается экспоненциальное сглаживание временного ряда?
12. Как значение параметра сглаживания влияет на экспоненциальную скользящую среднюю?
13. Каким будет прогноз уровней временного ряда с помощью экспоненциального сглаживания для будущих моментов времени?

Лабораторная работа № 10. Аналитическое выравнивание временного ряда

Цель работы. Изучение основных типов тенденций (трендов) уровней временного ряда и их представление функциями времени, освоение построения аналитических моделей тренда в пакете анализа MS Excel 2010.

Краткие сведения. *Аналитическое выравнивание временного ряда – способ представления тенденции временного ряда* $(y_1, y_2, \dots, y_n)$ *некоторой функцией времени t, т.е. тренд представляется в виде* $T_t = f(t, b)$*, где b вектор параметров*. Тип функции $f(t, b)$ определяется одним из следующих способов:

− путем качественного анализа изучаемого процесса;

− по графику временного ряда или графику его тенденции, выделенной методом скользящей средней;

− расчетом основных показателей динамики временного ряда (абсолютного прироста, абсолютного ускорения, темпа роста, темпа прироста);

− вычислением коэффициентов автокорреляции различных порядков;

− перебором различных форм тренда.

Тип функции $f(t, b)$ выбирают исходя из возможности оценивания его параметров методом наименьших квадратов, т.е. функция $f(t, b)$ должна быть линейной по времени *t* и параметрам *b*, или линейной по параметрам и нелинейной по времени, или внутренне линейной функцией (см. работы № 4 и № 5). Функция $f(t, b)$ может быть линейной комбинацией нескольких функций времени. Например, $f(t, b) = b_1 + b_2 \cdot \varphi_1(t) + b_3 \cdot \varphi_2(t)$. В этом случае вводят новые временные ряды $z_t = \varphi_1(t)$ *и* $z_t^* = \varphi_2(t)$ и методом наименьших квадратов оценивают линейную множественную регрессию $T_t = b_1 + b_2 \cdot z_t + b_3 \cdot z_t^*$. Оцененным уравнением тренда будет функция $T_t = b_1 + b_2 \cdot \varphi_1(t) + b_3 \cdot \varphi_2(t)$.

Значения фактора времени *t* могут быть выбраны одним из следующих способов:

− для первого момента времени принимается *t = 1*, в каждом последующем по времени наблюдении время *t* увеличивается на 1, т.е. *t* принимает значения 1, 2, …, *n*, где *n* объем выборки (длина временного ряда);

− начало отсчета времени выбирается в середине рассматриваемого временного промежутка, т.е. время *t* принимает значения …, –2, –1, 0, 1, –2, … При этом нужно учитывать вид нелинейной функции $f(t, b)$, чтобы обеспечить вычислимость значений этой функции.

Исследуемый временной ряд $(y_1, y_2, \dots, y_n)$, наряду с тенденцией, может содержать циклическую и сезонную компоненты. Поэтому критерии качества подгонки уравнения регрессии $T_t = f(t, b)$ к выборочным данным $(y_1, y_2, \dots, y_n)$, такие как коэффициент детерминации и средняя относительная ошибка аппроксимации, могут иметь плохие значения. Качество подгонки улучшается при оценке уравнения тренда по предварительно сглаженному (например, методом скользящей средней) временному ряду.

Разность $y_t - T_t$, наблюдаемых y_t уровней временного ряда и вычисленных значений тренда $T_t = f(t, b)$, дает временной ряд отклонений изучаемой

величины от тенденции. Временной ряд ($y_1-T_1, y_2-T_2, ..., y_n-T_n$) содержит колебательную (циклическую и/или сезонную) и случайную составляющие уровней исходного временного ряда. Проводя анализ ряда отклонений от тренда можно оценить сезонную составляющую S_t исходного временного ряда. Ряды разностей y_t-T_t или $y_t-(T_t+S_t)$ различных признаков используются для анализа корреляционной и регрессионной зависимости этих признаков.

Основные показатели, характеризующие тенденцию временного ряда.

Цепной абсолютный прирост (разность первого порядка):

$$\Delta y_t = y_{t+1} - y_t.$$

Абсолютное ускорение (разность второго порядка):

$$\Delta^2 y_t = \Delta y_{t+1} - \Delta y_t = y_{t+2} - 2y_{t+1} + y_t.$$

Цепной темп роста:

$$k_t = \frac{y_{t+1}}{y_t}.$$

Темп прироста:

$$\tau_t = \frac{\Delta y_t}{y_t} = k_t - 1.$$

Основные типы тенденций и уравнений тренда. При аналитическом выравнивании временных рядов, описывающих динамику экономических процессов, чаще всего используются следующие уравнения тренда.

Линейный тренд: $T_t = a + b \cdot t$. Свойства линейного тренда.

− *Цепной абсолютный прирост* $\Delta T_t = T_{t+1} - T_t = b$ не зависит от времени *t*, является величиной постоянной.

− *Абсолютное ускорение* $\Delta^2 T_t = \Delta T_{t+1} - \Delta T_t = 0$

− *Цепной темп роста* $k_t = \frac{T_{t+1}}{T_t} = \frac{a+b\cdot(t+1)}{a+b\cdot t} = 1 + \frac{b}{a+b\cdot t}$ убывает с ростом времени *t*.

− *Темп прироста* $\tau_t = \frac{\Delta T_t}{T_t} = k_t - 1 = \frac{b}{a+b\cdot t}$ убывает с ростом времени *t*.

Параболический тренд: $T_t = a + b \cdot t + c \cdot t^2$. Его свойства.

− *Цепной абсолютный прирост* $\Delta T_t = T_{t+1} - T_t = (b + c) + 2c \cdot t$ зависит от времени *t*. Он равномерно возрастает при $c > 0$ и равномерно убывает при $c < 0$ с ростом времени *t*.

− *Абсолютное ускорение* $\Delta^2 T_t = \Delta T_{t+1} - \Delta T_t = 2c$ величина постоянная.

− *Цепной темп роста* $k_t = \frac{T_{t+1}}{T_t} = 1 + \frac{b+c+2c\cdot t}{a+b\cdot t+c\cdot t^2}$ может убывать или возрастать с ростом времени *t* приближаясь к 1.

− *Темп прироста* $\tau_t = \frac{\Delta T_t}{T_t} = k_t - 1 = \frac{b+c+2c\cdot t}{a+b\cdot t+c\cdot t^2}$ может убывать или возрастать приближаясь к 0.

Гиперболический тренд: $T_t = a + \frac{b}{t}$. Его свойства.

− *Цепной абсолютный прирост* $\Delta T_t = T_{t+1} - T_t = -\frac{b}{t^2+t}$ зависит от времени *t*. При $b > 0$ возрастает, оставаясь отрицательной, и стремиться к нулю с

ростом времени *t*. При $b < 0$ убывает, оставаясь положительной, и стремиться к нулю с ростом времени *t*.

– *Абсолютное ускорение* $\Delta^2 T_t = \Delta T_{t+1} - \Delta T_t = \frac{2b}{t\cdot(t+1)\cdot(t+2)}$ убывает по абсолютной величине и стремится к нулю с ростом времени *t*, оставаясь положительной при $b > 0$ и отрицательной при $b < 0$.

– *Цепной темп роста* $k_t = \frac{T_{t+1}}{T_t} = 1 - \frac{b}{(at+b)\cdot(t+1)}$ стремится к 1, убывая или возрастая с ростом времени *t*.

– *Темп прироста* $\tau_t = \frac{\Delta T_t}{T_t} = k_t - 1 = -\frac{b}{(at+b)\cdot(t+1)}$ стремится к 0, убывая или возрастая с ростом времени *t*.

Экспоненциальный тренд: $T_t = e^{a+bt}$. Его свойства.

– *Цепной абсолютный прирост* $\Delta T_t = T_{t+1} - T_t = e^{a+bt}(e^b - 1)$ зависит от времени *t* и пропорционален уровням тренда. С ростом времени *t* при $b > 0$ неограниченно возрастает, оставаясь положительным. При $b < 0$ убывает по модулю, оставаясь отрицательным, и стремиться к нулю с ростом времени *t*.

– *Абсолютное ускорение* $\Delta^2 T_t = \Delta T_{t+1} - \Delta T_t = e^{a+bt}(e^b - 1)^2$ пропорционально уровням тренда. При $b > 0$ с ростом времени *t* неограниченно возрастает, оставаясь положительной. При $b < 0$ убывает, оставаясь положительной, и стремиться к нулю с ростом времени *t*.

– *Цепной темп роста* $k_t = \frac{T_{t+1}}{T_t} = e^b$ постоянная положительная величина.

– *Темп прироста* $\tau_t = \frac{\Delta T_t}{T_t} = k_t - 1 = e^b - 1$ постоянная величина положительная при *b>0* и отрицательная при *b<0*.

Показательный тренд: $T_t = a \cdot b^t$ с $a > 0$ и *b>0* соответствует экспоненциальному тренду $T_t = e^{lna + t\cdot lnb}$. Его свойства.

– *Цепной абсолютный прирост* $\Delta T_t = T_{t+1} - T_t = ab^t(b - 1)$ зависит от времени *t* и пропорционален уровням тренда. С ростом времени *t* при $b > 1$ неограниченно возрастает, оставаясь положительным. При $b < 1$ убывает по модулю, оставаясь отрицательным, и стремиться к нулю с ростом времени *t*.

– *Абсолютное ускорение* $\Delta^2 T_t = \Delta T_{t+1} - \Delta T_t = ab^t(b - 1)^2$ пропорционально уровням тренда. При $b > 1$ с ростом времени *t* неограниченно возрастает, оставаясь положительной величиной. При $b < 1$ убывает, оставаясь положительной величиной, и стремиться к нулю с ростом времени *t*.

– *Цепной темп роста* $k_t = \frac{T_{t+1}}{T_t} = b$ постоянная положительная величина.

– *Темп прироста* $\tau_t = \frac{\Delta T_t}{T_t} = k_t - 1 = b - 1$ постоянная величина, положительная при *b>1* и отрицательная при *b<1*.

Логарифмический тренд: $T_t = a + b \cdot lnt$. Его свойства.

– *Цепной абсолютный прирост* $\Delta T_t = T_{t+1} - T_t = b \cdot ln(1 + \frac{1}{t})$ зависит от времени *t*. С ростом времени *t* уменьшается по модулю и стремится к нулю, оставаясь положительным при $b > 0$ и отрицательным при $b < 0$.

– *Абсолютное ускорение* $\Delta^2 T_t = \Delta T_{t+1} - \Delta T_t = b \cdot (\ln\left(1 + \frac{1}{t+1}\right) - \ln(1 + \frac{1}{t}))$ с ростом времени *t* убывает по модулю и стремится к нулю, оставаясь положительным при $b < 0$ и отрицательным при $b > 0$.

– *Цепной темп роста* $k_t = \frac{T_{t+1}}{T_t} = \frac{a+b\cdot\ln(t+1)}{a+b\cdot lnt}$. При $b > 0$ с ростом времени *t* уменьшается и стремится к 1, оставаясь больше 1. При $b < 0$ с ростом времени *t* возрастает и стремится к 1, оставаясь меньше 1.

– *Темп прироста* $\tau_t = \frac{\Delta T_t}{T_t} = k_t - 1 = \frac{b\cdot\ln(1+\frac{1}{t})}{a+b\cdot lnt}$. При $b > 0$ с ростом времени *t* уменьшается и стремится к 0, оставаясь больше 0. При $b < 0$ с ростом времени *t* возрастает и стремится к 0, оставаясь меньше 0.

Логистический тренд $T_t = \frac{1}{e^{a+bt}+1}$ используется, если диапазон изменения уровней временного ряда ограничен нулем и единицей, т.е. исследуемый признак является относительной величиной и $0 < y_t < 1$. При $a > 0$ и $b < 0$ с ростом времени *t* происходит логистический рост тренда (в начале периода абсолютные приросты положительны и ускоренно растут, в середине периода абсолютные приросты становятся постоянными, в конце периода абсолютные приросты ускоренно уменьшаются). При $a < 0$ и $b > 0$ с ростом времени *t* происходит логистическое уменьшение тренда (в начале периода отрицательные абсолютные приросты ускоренно растут по абсолютной величине, в середине периода абсолютные приросты становятся постоянными, в конце периода абсолютные приросты ускоренно уменьшаются по абсолютной величине).

Содержание лабораторной работы.

1. Из работы № 9 ввести исследуемый временной ряд и сглаженный ряд отражающий его тенденцию.

2. По сглаженному ряду вычислить основные показатели, характеризующие тенденцию временного ряда (цепной абсолютный прирост, абсолютное ускорение, цепной темп роста, темп прироста).

3. Построить графики показателей, характеризующих тенденцию временного ряда.

4. Используя графики исходного временного ряда и сглаженного ряда, а также проведя анализ показателей, характеризующих тенденцию временного ряда, выбрать тип тренда рассматриваемого временного ряда.

5. По исходному временному ряду, следуя работам 4, 5 или 6, оценить линейный тренд и тренд, выбранный в предыдущем пункте.

6. По построенным моделям тренда просчитать прогноз тенденции исследуемого признака на два шага вперед, т.е. для *t = n + 1* и *t = n + 2*.

7. Общее заключение о характере тенденции временного ряда.

Выполнение работы в MS Excel.

Проведем аналитическое выравнивание временного ряда квартальных объемов продаж y_t (млрд дол.) за 1990–1994 годы. Для первого квартала 1990 года примем *t = 1*

Ввод данных. В ячейках **A2-A21** значения времени *t* от 1 до 20. В ячейках **B2-B21** значения y_t уровней временного ряда. В ячейках **C4-C19** находятся значения взвешенной скользящей средней с длиной окна сглаживания равной 5, см. рис. 10.1.

Вычисление основных показателей, характеризующих тенденцию временного ряда, производится по сглаженному временному ряду.

Для упрощения построения графиков основных показателей тенденции скопируем столбец значений времени *t*, ячейки **A1-A21**, в столбцы **E2-E22, G2-G22, I2-I22, K2-K22**. В ячейках **F5-F19** вычислим значения цепного абсолютного прироста $\Delta T_t = T_{t+1} - T_t$, в ячейках **H6-H19** вычислим значения абсолютного ускорения $\Delta^2 T_t = \Delta T_{t+1} - \Delta T_t$, в ячейках **J5-J19** вычислим значения цепного темпа роста $k_t = \frac{T_{t+1}}{T_t}$, в ячейках **L5-L19** вычислим значения темпа прироста $\tau_t = k_t - 1$. См. рис. 10.1.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	t	y(t)	ВСС-5		характеристики тенденции временного ряда								
2	1	1444,1			t	Δyt	t	Δ2yt	t	kt	t	τt	
3	2	1472,2			1		1		1		1		
4	3	1473,7	1470,597		2		2		2		2		
5	4	1444,4	1443,851		3	-26,7457	3		3	0,981813	3	-0,01819	
6	5	1416,3	1422,206		4	-21,6457	4	5,1	4	0,985008	4	-0,01499	
7	6	1427,8	1427,046		5	4,84	5	26,48571	5	1,003403	5	0,003403	
8	7	1448,4	1443,934		6	16,88857	6	12,04857	6	1,011835	6	0,011835	
9	8	1456,4	1464,26		7	20,32571	7	3,437143	7	1,014077	7	0,014077	
10	9	1482,2	1470,474		8	6,214286	8	-14,1114	8	1,004244	8	0,004244	
11	10	1464,5	1473,431		9	2,957143	9	-3,25714	9	1,002011	9	0,002011	
12	11	1478,8	1472,174		10	-1,25714	10	-4,21429	10	0,999147	10	-0,00085	
13	12	1496,4	1513,423		11	41,24857	11	42,50571	11	1,028019	11	0,028019	
14	13	1565,9	1540,229		12	26,80571	12	-14,4429	12	1,017712	12	0,017712	
15	14	1537,3	1553,757		13	13,52857	13	-13,2771	13	1,008783	13	0,008783	
16	15	1560,1	1553,843		14	0,085714	14	-13,4429	14	1,000055	14	5,52E-05	
17	16	1591,8	1604,846		15	51,00286	15	50,91714	15	1,032824	15	0,032824	
18	17	1662,9	1642,92		16	38,07429	16	-12,9286	16	1,023725	16	0,023725	
19	18	1651,7	1662,303		17	19,38286	17	-18,6914	17	1,011798	17	0,011798	
20	19	1669,6			18		18		18		18		
21	20	1704,3			19		19		19		19		
22					20		20		20		20		
23													

Лист1 Лист2 Лист3

Рис. 10.1. Временной ряд и характеристики его тенденции

Построение графиков показателей, характеризующих тенденцию временного ряда.

Используя функцию построения точечных графиков в MS Excel, построим графики характеристик тенденции временного ряда, см. рис. 10.2.

Выбор типа тренда. Из рассмотрения временных рядов характеристик тенденции временного ряда y_t и их графиков следует.

1) Цепной абсолютный прирост имеет тенденцию возрастания и абсолютное ускорение имеют большой разброс значений относительно своего положительного среднего и тенденцию возрастания. Следовательно, линейный, логарифмический, логистический и гиперболический тренды не подходят для ана-

литического выражения тенденции, а параболический, экспоненциальный и показательный тренды подходят.

2) Цепной темп роста и темп прироста имеют тенденцию возрастания, что говорит в пользу параболического тренда. Цепной темп роста и темп прироста колеблются относительно своих положительных средних, что оставляет необходимость рассмотрения и экспоненциального и показательного трендов.

3) логистический тренд не подходит для описания тенденции исследуемого временного ряда, так как не наблюдается ускоренного роста его уровней в начале периода и замедления роста в конце периода.

Исходя из перечисленного, тенденция временного ряда может быть описана параболическим или экспоненциальным трендом.

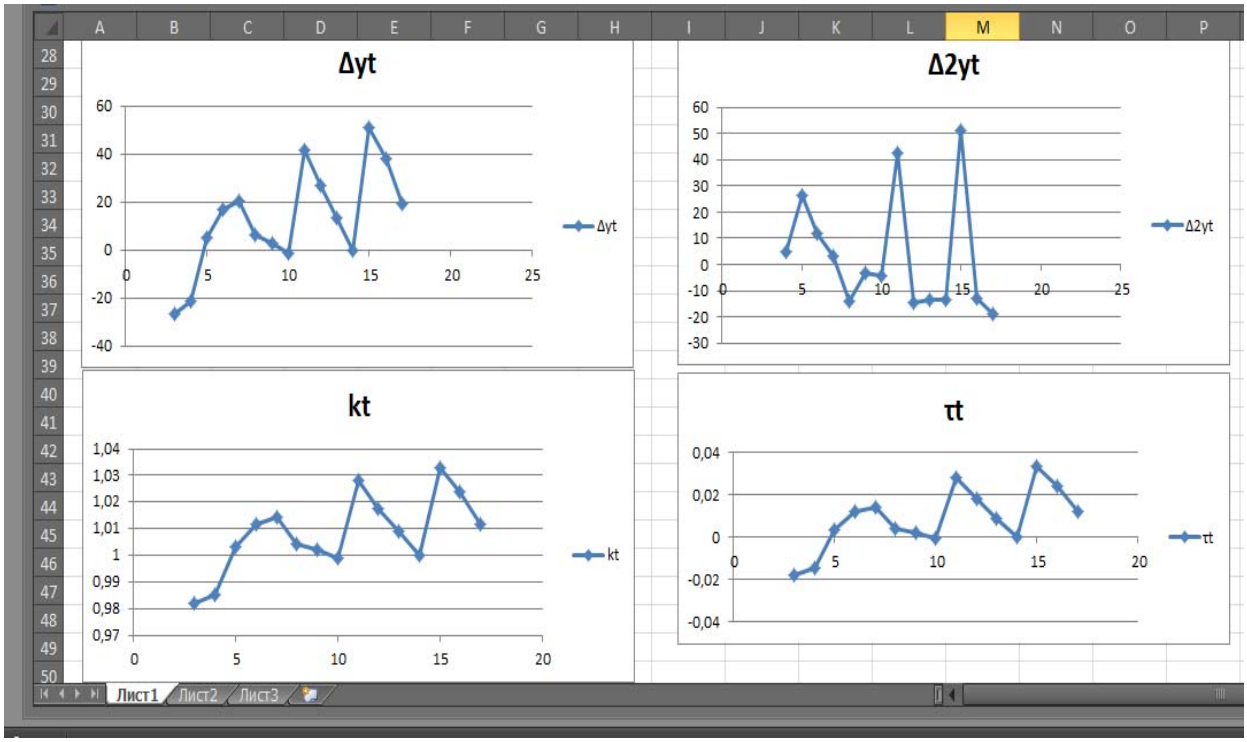
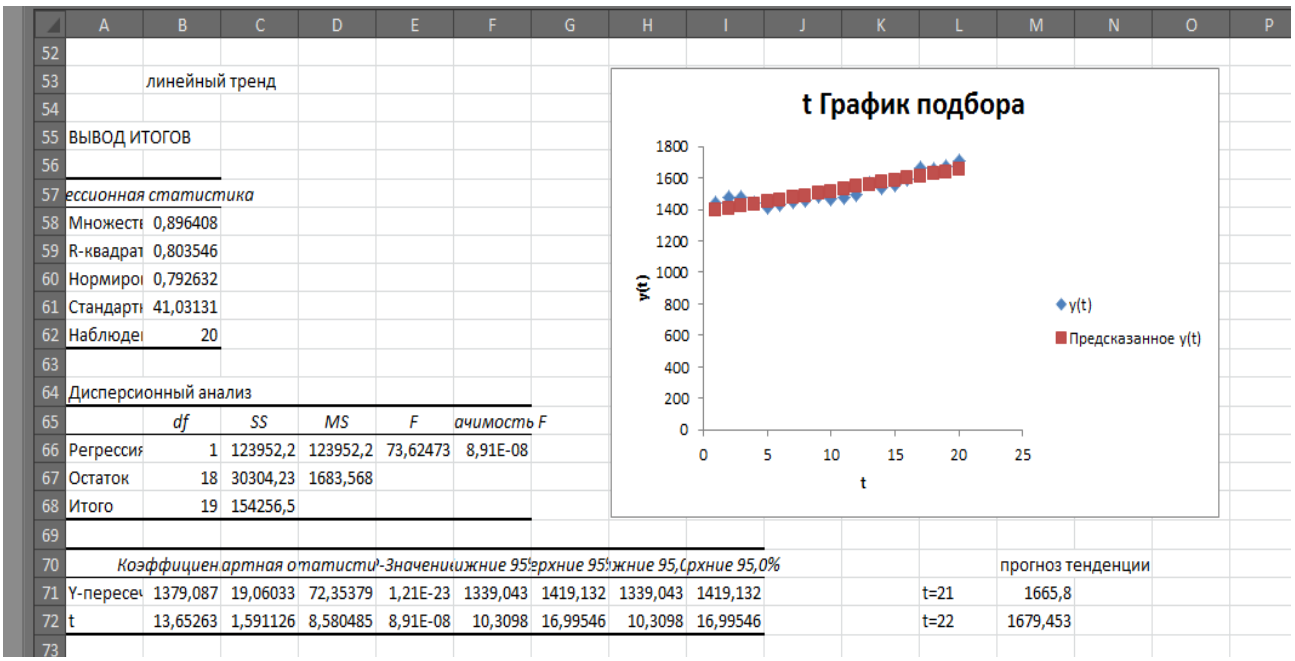


Рис. 10.2. Графики характеристик тенденции временного ряда

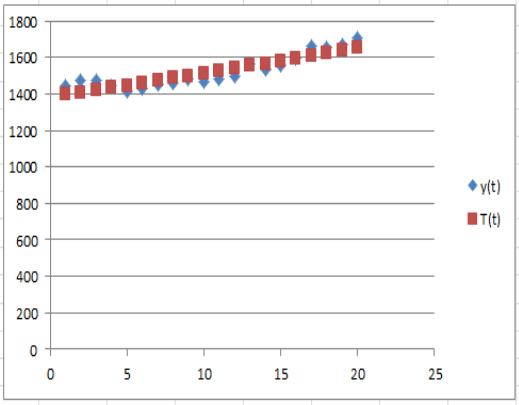
Оценим, например, экспоненциальный и линейный тренды по исходному временному ряду. Построение линейной и экспоненциальной регрессии приведено в работах 4 и 5. Здесь приведем только результаты оценивания, см. рис. 10.3–10.4. Линейный тренд $T_t = 1379{,}087 + 13{,}653 * t$ и экспоненциальный тренд $T_t = e^{7{,}234+0{,}00882*t}$ хорошо согласуются с тенденцией исходного временного ряда, что следует из их графиков подгонки.

Построение прогноза тенденции временного ряда. Прогноз тенденции исследуемого временного ряда для моментов времени *t=21* и *t=22* (на первый и второй кварталы 1995 г.) производится непосредственным вычислением по уравнениям трендов. Результаты прогноза тенденции по обеим моделям тренда приведены на рис. 10.3 и рис. 10.4. По линейному тренду получаем $T_{21} = 1665{,}8$ и $T_{22} = 1679{,}45$. По экспоненциальному тренду имеем $T_{21} = 1667{,}7$ и $T_{22} = 1682{,}5$.



		линейный тренд													
ВЫВОД ИТОГОВ															
ессионная статистика															
Множеств	0,896408														
R-квадрат	0,803546														
Нормиро	0,792632														
Стандарт	41,03131														
Наблюде	20														
Дисперсионный анализ															
	df	*SS*	*MS*	*F*	*ачимость F*										
Регрессия	1	123952,2	123952,2	73,62473	8,91E-08										
Остаток	18	30304,23	1683,568												
Итого	19	154256,5													
Коэффициен	*артная о*	*татистик*	*Значени*	*ижние 95%*	*ерхние 95%*	*жние 95,0*	*рхние 95,0%*					прогноз тенденции			
Y-пересеч	1379,087	19,06033	72,35379	1,21E-23	1339,043	1419,132	1339,043	1419,132			t=21	1665,8			
t	13,65263	1,591126	8,580485	8,91E-08	10,3098	16,99546	10,3098	16,99546			t=22	1679,453			

Рис. 10.3. Линейный тренд



		экспоненциальный тренд		T(t)=exp(7,234+0,00882*t)													
t	y(t)	lny(t)															
1	1444,1	7,275242		ВЫВОД ИТОГОВ													
2	1472,2	7,294513															
3	1473,7	7,295532		*Регрессионная статистика*													
4	1444,4	7,275449		Множеств	0,898951												
5	1416,3	7,255803		R-квадрат	0,808114												
6	1427,8	7,26389		Нормиро	0,797453												
7	1448,4	7,278215		Стандарт	0,026119												
8	1456,4	7,283723		Наблюде	20												
9	1482,2	7,301283															
10	1464,5	7,289269		Дисперсионный анализ													
11	1478,8	7,298986			*df*	*SS*	*MS*	*F*	*ачимость F*								
12	1496,4	7,310818		Регрессия	1	0,051716	0,051716	75,80552	7,19E-08								
13	1565,9	7,356216		Остаток	18	0,01228	0,000682										
14	1537,3	7,337783		Итого	19	0,063996											
15	1560,1	7,352505															
16	1591,8	7,372621		*Коэффициен*	*артная о*	*татистик*	*Значени*	*ижние 95%*	*ерхние 95%*	*жние 95,0*	*рхние 95,0%*			*прогноз тенденции*			
17	1662,9	7,416318		Y-пересеч	7,233853	0,012133	596,2022	4,06E-40	7,208362	7,259344	7,208362	7,259344		t=21	1667,732		
18	1651,7	7,40956		t	0,008819	0,001013	8,706637	7,19E-08	0,006691	0,010947	0,006691	0,010947		t=22	1682,507		
19	1669,6	7,420339															
20	1704,3	7,44091															

Рис. 10.4. Экспоненциальный тренд

Общее заключение о характере тенденции временного ряда. Квартальные объемы продаж имеют тенденцию возрастания. Анализ характеристик тенденции временного ряда показал: цепной абсолютный прирост имеет тенденцию возрастания и абсолютное ускорение имеют большой разброс значений относительно своего положительного среднего и тенденцию возрастания. Следовательно, линейный, логарифмический, логистический и гиперболический тренды не подходят для аналитического выражения тенденции, а параболический, экспоненциальный и показательный тренды подходят.

Цепной темп роста имеет малый разброс значений вокруг своей средней, также и темп прироста имеет малый разброс значений вокруг своей средней. Это говорит в пользу экспоненциального тренда. Эта тенденция хорошо описывается экспоненциальным трендом $T_t = e^{7,234+0,00882*t}$. Прогнозы тенденции на два квартала вперед по линейному и экспоненциальному трендам дают близкие результаты. Прогноз тенденции по экспоненциальному тренду растет быстрее. Цепной темп роста тенденции квартальных объемов продаж, согласно экспоненциальной модели, составляет 1,00886 (100,886%).

Контрольные вопросы

1. Что понимается под аналитическим выравниванием уровней временного ряда?
2. Какими способами может быть определен тип функции тренда при аналитическом выравнивании?
3. Приведите основные характеристики тенденции временного ряда.
4. По какому ряду производят вычисление основных характеристик тенденции временного ряда?
5. Приведите свойства линейного тренда и параболического тренда.
6. Каким методом производится оценка параметров модели тренда?
7. Как изменяется с течением времени цепной абсолютный прирост для гиперболического тренда?
8. Приведите свойства экспоненциального и показательного тренда. Есть ли существенное различие этих свойств?
9. В каких случаях используется логистический тренд?
10. Какой следует выбрать тип тренда, если уровни ряда абсолютных ускорений имеют малый разброс вокруг некоторого отличного от нуля числа? Напишите уравнение этого тренда.
11. Приведите модель логарифмического тренда и перечислите свойства его свойства.
12. Приведите графики основных типов трендов.

Список рекомендуемой литературы

1. Кремер Н.Ш. Эконометрика / Н.Ш. Кремер, Б.А. Путко. – М. : ЮНИТИ, 2003.

2. Ежова Л.Н. Эконометрика. Начальный курс с основами теории вероятностей и математической статистики / Л.Н. Ежова. – Иркутск : БГУЭП, 2008.

3. Магнус Я. Эконометрика. Начальный курс / Я. Магнус, П. Катышев, А. Пересецкий. – М. : Дело, 2004.

4. Практикум по эконометрике: учеб. пособие / И.И. Елисеева, С.В. Курышева, Н.М. Гордиенко и др.; под ред. И.И. Елисеевой. – М. : Финансы и статистика, 2003.

5. Эконометрика: учебник / И.И. Елисеева, С.В. Курышева, Т.В. Костеева и др.; под ред. И.И. Елисеевой. – 2-е изд., перераб. и доп. – М . : Финансы и статистика, 2008.

6. Айвазян С.А. Основы эконометрики / С.А. Айвазян. – М. : ЮНИТИ, 2001. Т. 1, 2.

7. Доугерти К. Введение в эконометрику / Доугерти К. – М. : Инфа-М, 1997.

8. Эконометрика: учебник / под. ред. В.С. Мхитаряна. – М.: Проспект, 2008.

9. Кремер Н.Ш. Теория вероятностей и математическая статистика: учебник для вузов / Н.Ш. Кремер. – М. : ЮНИТИ-ДАНА, 2001.

10. Афанасьев В.Н. Анализ временных рядов и прогнозирование: учебник / В.Н. Афанасьев, М.М. Юзбашев. – М. : Финансы и статистика, 2001.

11. Тимофеев В.С. Эконометрика: учебник для бакалавров / В.С. Тимофеев, А.В. Фаддеенков, В.Ю. Щеколдин. – 2-е изд., перераб. и доп. – М. : Изд-во Юрайт, 2013.

12. Новиков А.И. Эконометрика: учеб. пособие / А.И. Новиков. – 2-е изд., испр. и доп. – М. : ИНФРА-М, 2007.

13. Балдин К.В. Эконометрика: учеб. пособие для вузов / К.В. Балдин, О.Ф. Быстров, М.М. Соколов. – М. : ЮНИТИ-ДАНА, 2004.

Приложение
«Данные для выполнения лабораторных работ»

Данные для лабораторной работы № 1

Вариант 1. Оценить закон распределения и числовые характеристики генеральной совокупности – времени выполнения одной операции (мин.) рабочими по приведенным результатам 100 наблюдений. N(9,2).

8,6	6,4	10,9	9,3	8,5	11,0	10,3	8,4	7,4	10,8
8,9	13,8	10,2	7,3	9,1	10,2	10,8	10,8	6,1	9,0
9,9	8,8	10,7	6,9	8,2	13,2	9,9	13,6	7,7	10,7
9,4	8,0	9,5	13,6	10,4	10,2	10,0	9,7	11,3	6,4
11,4	8,6	9,5	5,7	7,9	9,7	6,6	12,3	8,4	7,8
12,1	8,6	8,4	10,1	8,9	10,0	6,7	9,2	8,9	7,4
6,9	10,7	8,9	10,9	9,5	11,4	9,1	10,5	10,5	8,8
7,5	9,8	7,4	9,8	9,4	9,7	9,5	9,9	8,2	10,3
10,7	8,7	7,5	9,5	7,6	10,8	8,5	11,1	6,9	10,8
8,4	9,8	12,8	8,2	12,0	7,5	11,3	11,0	3,3	8,2

Вариант 2. Оценить закон распределения и числовые характеристики генеральной совокупности – суточных надоев молока (литров.) от одной коровы по приведенным результатам 100 наблюдений. N(15,3).

14,21	15,42	20,42	10,98	11,74	15,91	10,67	17,50	15,98	12,70
16,51	18,67	12,46	14,64	23,51	11,75	16,27	14,67	16,15	18,87
16,74	14,98	15,55	15,41	16,29	12,44	16,80	5,21	18,03	11,27
14,52	17,88	16,85	16,35	15,97	10,10	17,78	14,56	12,59	9,06
17,26	18,66	14,19	15,35	12,85	13,18	18,39	14,90	15,79	13,97
11,74	14,43	15,39	12,48	12,97	12,04	16,26	20,78	14,27	14,85
18,39	15,19	14,59	9,80	12,65	15,91	11,37	20,90	12,15	19,91
17,79	11,26	17,07	17,06	12,64	8,86	18,67	11,42	15,36	18,83
11,77	12,29	10,74	14,87	13,16	15,32	18,11	11,81	13,86	11,25
13,27	15,74	12,44	14,97	12,21	12,06	15,60	15,16	16,99	14,62

Вариант 3. Оценить закон распределения и числовые характеристики генеральной совокупности – времени (мин.) затрачиваемого на поездку от дома до работы по приведенным результатам 110 наблюдений.

21	28	21	22	22	19	24	22	20	23	34
27	27	19	30	30	24	22	19	23	30	30
20	22	19	31	27	29	29	25	21	24	28
28	19	23	26	23	19	27	33	25	26	24
26	25	32	25	26	28	28	22	28	27	25
31	31	25	22	22	25	32	22	29	33	26
27	22	28	29	25	21	25	25	34	27	32
20	26	23	23	25	24	27	24	25	24	23
29	29	31	24	24	28	31	30	20	28	19
23	22	27	24	22	20	18	22	25	28	22

Вариант 4. Оценить закон распределения и числовые характеристики генеральной совокупности – месячной заработной платы (тыс. р.) рабочих одной профессии по приведенным результатам 110 наблюдений. N(30,6).

37,6	28,4	26,5	21,8	19,5	22,8	29,0	36,0	32,2	29,7	32,2
29,9	29,2	29,2	36,2	22,9	30,8	21,9	33,0	30,8	28,7	25,5
35,8	29,9	39,6	44,3	30,6	24,7	47,5	31,6	36,2	33,9	32,8
26,7	25,5	26,2	21,9	22,1	29,4	27,0	31,1	27,1	28,2	34,6
29,1	28,3	40,4	26,3	20,4	31,2	32,7	34,8	15,4	17,6	27,5
21,0	27,5	23,2	32,7	31,5	30,2	38,1	30,6	33,3	33,7	36,3
29,3	24,7	33,7	41,8	28,1	40,1	21,0	25,9	32,7	28,3	18,4
29,0	47,1	28,7	39,3	26,7	24,0	39,2	24,0	38,3	30,4	39,2
34,2	31,8	38,1	33,9	38,4	34,4	39,1	21,2	25,2	34,9	31,5
29,0	33,9	24,4	37,7	44,9	26,9	34,9	32,8	28,2	28,4	36,6

Вариант 5. Оценить закон распределения и числовые характеристики генеральной совокупности – расход топлива (литров) на 100 км самосвалом «КАМАЗ» по приведенным результатам 100 наблюдений. N(50, 7).

55,6	48,1	52,7	55,7	49,1	51,7	57,8	43,2	54,6	61,4
71,2	53,8	62,7	43,4	54,6	42,9	46,2	50,8	47,7	61,8
51,4	61,2	47,6	49,2	50,5	60,7	62,2	43,4	62,6	39,0
47,4	51,5	59,4	51,2	54,5	42,9	57,5	39,1	26,0	60,0
49,1	48,8	45,0	42,1	45,3	48,5	55,6	47,8	43,9	53,1
49,8	51,6	45,3	59,4	45,8	51,6	51,7	41,3	38,4	46,0
45,2	50,8	34,7	44,7	37,1	52,1	48,8	50,8	53,4	53,2
47,7	46,9	49,0	54,4	46,9	42,4	49,9	56,9	42,5	60,3
43,0	54,3	41,7	40,4	39,1	49,1	40,5	43,1	42,7	39,5
53,1	54,5	48,6	43,4	48,9	46,9	70,2	38,3	36,6	57,2

Вариант 6. Оценить закон распределения и числовые характеристики генеральной совокупности – стоимости (тыс. р.) квадратного метра жилья на вторичном рынке по приведенным результатам 100 наблюдений. N(40, 10).

44,28	46,98	31,11	26,32	38,14	55,70	39,00	31,62	29,21	49,36
29,48	48,73	26,83	55,61	29,23	54,95	35,67	33,01	47,11	34,58
44,98	40,80	39,43	65,95	40,42	38,85	31,42	47,86	47,91	34,26
44,06	46,00	54,05	28,73	49,86	27,35	43,80	47,45	37,03	21,37
47,65	41,81	42,60	50,73	17,07	54,42	32,25	43,52	26,06	38,47
37,36	45,04	44,00	45,11	28,69	38,12	24,03	23,73	40,34	36,36
61,11	20,20	34,56	36,39	32,22	45,54	37,11	47,73	28,77	36,85
33,68	38,35	31,57	43,02	23,99	55,57	53,59	30,04	41,36	52,14
45,34	21,82	52,48	33,65	38,19	30,79	30,13	37,31	54,05	39,44
49,08	27,76	38,64	21,79	41,89	43,31	33,57	51,63	21,70	36,57

Вариант 7. Оценить закон распределения и числовые характеристики генеральной совокупности – количества клиентов (чел.) обслуживаемых за смену одним мастером по приведенным результатам 100 наблюдений. N(30, 5).

34	23	42	28	31	27	35	32	29	31
23	26	29	34	23	30	37	34	28	24
28	37	34	35	20	38	29	31	35	31
31	21	33	31	22	37	38	34	40	26
26	22	25	29	34	27	30	26	28	30
24	28	26	32	27	33	28	22	34	25
30	37	28	31	30	27	32	38	25	39
32	30	26	25	29	20	33	31	35	31

28	28	34	33	35	28	31	36	25	33
30	29	27	25	23	27	24	29	31	29

Вариант 8. Оценить закон распределения и числовые характеристики генеральной совокупности – объема недельных продаж бензина АИ-92 (тонн) по приведенным результатам 110 наблюдений. N(13, 2).

13,64	15,53	12,77	11,47	13,72	11,85	17,04	9,07	11,83	10,00	12,62
12,06	13,64	12,93	12,72	10,26	13,48	11,07	14,21	11,60	11,55	17,22
16,81	14,29	11,47	14,74	12,68	12,72	8,29	12,79	12,29	11,27	8,80
13,27	9,70	12,73	13,22	11,81	10,52	14,68	12,21	12,87	15,89	13,65
12,78	8,39	11,26	13,52	14,15	11,71	12,26	10,91	14,23	12,50	13,70
12,95	12,46	9,49	15,59	10,52	13,31	14,49	14,25	14,01	16,13	13,66
13,48	10,84	14,08	11,60	13,06	12,15	11,08	12,10	13,32	17,71	12,91
15,54	13,67	11,28	17,39	15,55	13,55	14,84	9,22	11,94	13,70	11,70
15,78	11,43	12,75	13,20	12,87	13,56	12,95	12,50	14,02	12,54	16,26
11,33	8,84	12,87	14,64	12,70	14,90	13,78	13,51	11,75	11,20	11,91

Вариант 9. Оценить закон распределения и числовые характеристики генеральной совокупности – производительности труда (тыс. р.) рабочих одной профессии по приведенным результатам 100 наблюдений. N(30, 4).

37,6	28,4	34,8	28,8	29,9	26,6	34,7	34,1	31,5	34,7
31,0	28,8	30,9	33,0	33,6	28,1	31,1	24,0	22,2	31,3
23,0	33,2	28,2	22,5	40,1	29,8	24,3	29,2	21,7	25,6
26,1	25,7	38,6	27,2	29,8	30,0	33,4	28,7	31,2	34,7
31,9	29,0	31,6	29,3	25,1	26,5	29,9	35,1	31,0	34,3
25,3	21,6	30,7	31,1	23,5	30,2	32,3	30,1	28,6	32,8
28,2	28,2	23,4	33,0	36,1	33,5	36,5	29,7	28,7	31,2
26,5	29,2	30,3	35,5	37,4	27,1	30,5	29,5	30,1	31,3
30,7	26,8	37,0	33,9	34,2	26,8	28,0	31,1	32,8	29,1
32,1	33,0	30,0	32,0	31,6	32,3	31,2	29,4	25,6	28,5

Вариант 10. Оценить закон распределения и числовые характеристики генеральной совокупности – суточного объема продаж (тыс. р.) цветочных киосков по приведенным результатам 100 наблюдений. N(15, 1).

14,41	15,06	15,97	13,77	14,61	14,40	14,42	16,16	15,15	14,22
15,29	14,94	16,59	15,74	15,43	16,45	14,80	15,19	15,11	15,77
15,21	15,35	15,91	14,47	13,94	15,55	14,70	15,43	15,05	16,41
12,83	15,17	16,73	16,16	15,29	14,99	16,05	13,60	15,93	13,79
14,70	15,22	15,30	13,60	14,69	13,99	14,71	15,15	15,52	16,23
15,67	14,96	15,86	16,01	14,76	14,53	14,84	13,80	15,73	14,63
14,68	14,31	13,06	14,70	14,73	14,54	14,93	14,86	15,78	14,87
14,78	13,88	15,41	16,41	15,32	15,20	16,55	13,46	15,40	15,21
15,63	16,43	14,79	14,72	15,27	14,30	16,12	14,04	12,78	15,95
14,86	15,84	13,32	15,17	14,67	15,68	14,69	13,96	14,65	16,94

Данные для лабораторной работы № 2

Вариант 1. Исследовать корреляционную зависимость себестоимости 1 т. литья (*Y*, тыс. р.), выпуска продукции на одного работающего (*X*, т.), процента брака литья (*Z*, %) по следующим данным.

Y	239	254	262	251	158	101	259	186	204	198	170	173
X	14,6	13,5	21,5	17,4	44,8	111,9	20,1	28,1	22,3	25,3	56	40,2
Z	4,2	6,7	5,5	7,7	1,2	2,2	8,4	1,4	4,2	0,9	1,8	1,8

Y	241	258	254	180	195	184
X	14,1	17,5	30,5	56,2	25,1	39,2
Z	5,4	6,1	4,5	5,3	2,8	1,4

Вариант 2. Исследовать корреляционную зависимость прибыли предприятия (*Y*, млн р.), выпуска продукции на одного работающего (*X*, единиц), доли продукции, производимой на экспорт (*Z*, %) по следующим данным.

Y	20	19	31	80	69	52	41	60	72	71	66	26
X	11	10	12	18	15	13	13	15	16	17	16	12
Z	3	2	4	10	11	6	5	7	10	12	11	4

Y	20	25	34	81	74	61	47	51
X	11	11	13	17	16	14	13	14
Z	3	3	5	11	11	9	6	7

Вариант 3. Исследовать корреляционную зависимость душевого дохода *Y ($)*, индекса человеческого развития *X1* и индекса человеческой бедности *X2* по данным ряда стран.

Страна	*Душевой доход Y*	*Индекс развития X1*	*Индекс бедности X2*
ОАЭ	1600	0,866	14,9
Таиланд	7100	0,833	11,7
Уругвай	6750	0,883	11,7
Ливия	6130	0,801	18,8
Колумбия	6110	0,848	10,7
Иордания	4190	0,73	10,9
Египет	3850	0,514	34,8
Марокко	3680	0,566	41,7
Перу	3650	0,717	22,8
Шри-Ланка	3280	0,711	20,7
Филиппины	2680	0,672	17,7
Боливия	2600	0,589	22,5
Китай	2600	0,625	17,5
Зимбабве	2200	0,513	17,3
Пакистан	2150	0,445	46,8
Уганда	1370	0,328	41,3
Нигерия	1350	0,393	41,6
Индия	1350	0,446	36,7

Вариант 4. Исследовать корреляционную зависимость выработки продукции на одного рабочего *Y* (тыс. р.), ввода новых основных фондов *X1* (% от стоимости основных фондов на конец года) и удельного веса рабочих высокой квалификации в общей численности рабочих *X2* (%) по следующим данным.

Номер предприятия	*X1*	*X2*	*Y*	Номер предприятия	*X1*	*X2*	*Y*
1	3,9	10	7	11	6,0	21	9
2	3,9	14	7	12	6,4	22	11
3	3,7	15	7	13	6,8	22	9
4	4,0	16	8	14	7,2	25	11
5	3,8	17	7	15	8,0	28	12
6	4,8	19	7,5	16	8,2	29	12
7	5,4	19	8	17	8,1	30	12,5
8	4,4	20	8	18	8,5	31	12
9	5,3	20	8,5	19	9,6	32	14
10	6,8	20	10	20	9,0	36	15

Вариант 5. Исследовать корреляционную зависимость валового дохода *Y* торговых предприятий (млн р.), стоимости их основных *X1* (млн р.) и оборотных *X2* (млн р.) средств по приведенным в таблице данным по 20 предприятиям.

Y	203	63	45	113	121	88	110	56	80	237	160	75
X1	118	28	17	50	56	102	116	124	114	154	115	98
X2	105	56	54	63	28	50	54	42	36	106	88	46

Y	130	75	54	102	117	96	99	83
X1	73	33	25	46	53	112	109	120
X2	76	67	55	58	44	56	52	48

Вариант 6. Исследовать корреляционную зависимость чистого дохода *Y* (млрд долл.), оборота капитала *OK* (млрд дол.), использованного капитала *K* (млрд дол.), численности занятых *L* (тыс. чел.), используя данные о деятельности 20 компания США за 1996 год, приведенные в таблице.

№ п/п	*Y*	*OK*	*K*	*L*
1	6,6	6,9	83,6	222,0
2	3,0	18	6,5	32,0
3	6,5	107,9	50,4	82,0
4	3,3	16,7	15,4	45,2
5	0,1	79,6	29,6	299,3
6	3,6	16,2	13,3	41,6
7	1,5	5,9	5,9	17,8
8	5,5	53,1	27,1	151,0
9	2,4	18,8	11,2	82,3
10	3,0	35,3	16,4	103,0
11	4,2	71,9	32,5	225,4
12	2,7	93,6	25,4	675,0
13	1,6	10,0	6,4	43,8
14	2,4	31,5	12,5	102,3
15	3,3	36,7	14,3	105,0
16	1,8	13,8	6,5	49,1
17	2,4	64,8	22,7	50,4
18	1,6	30,4	15,8	480,0
19	1,4	12,1	9,3	71,0
20	0,9	31,3	18,9	43,0

Вариант 7. Исследовать корреляционную зависимость объема месячных продаж пива *Y* (усл. ед.), расходов на рекламу *R* (усл. ед.), числа туристов *T* (тыс. чел.), индекса цен *P* (%) используя месячные данные, приведенные в таблице.

№ п/п	*Y*	*R*	*T*	*P*
1	8646,9	416,0	1741,5	114,1
2	11758,5	327,7	2060,0	116,0
3	11867,2	160,6	1777,8	116,6
4	9577,6	403,1	1378,9	122,6
5	10898,4	269,7	1253,3	119,5
6	9638,6	280,5	794,0	130,6
7	9203,9	335,1	1384,4	125,0
8	9231,1	169,3	1392,5	124,2
9	7334,5	206,0	2484,3	130,7
10	7467,0	216,1	2777,5	131,6
11	7839,6	322,2	3301,9	133,4
12	9787,0	285,5	3635,9	139,1
13	9600,3	79,2	3415,9	142,3
14	7199,9	336,3	2606,8	139,9
15	9547,7	293,1	2508,0	144,5
16	10187,5	238,5	2834,1	143,9
17	9661,2	255,4	2481,8	148,0
18	9189,2	383,6	1474,4	149,3

Вариант 8. Исследовать корреляционную зависимость численности населения N, среднемесячной зарплаты Z, стоимости основных фондов K, оборота розничной торговли OT по городам Приволжского федерального округа по приведенным в таблице данным.

Города	N, тыс. чел.	Z, р.	K, млн р.	OT, млн р.
Уфа	1062,3	22089,5	595304,5	46811,7
Йошкар-Ола	259,2	15098,8	69051,4	8303,8
Саранск	297,4	14152,1	89430,2	6866,1
Казань	1143,5	19410	620442,4	70231,8
Ижевск	628,1	17255,2	233387,0	22437
Чебоксары	453,6	15936,4	105409,9	13406,9
Пермь	991,5	22678,8	472475,3	49157,2
Киров	473,7	17722,4	175144,9	17378,2
Нижний Новгород	1250,6	21821,2	576344,7	83069,5
Оренбург	547,0	18990,0	516399,6	22428,2
Пенза	517,1	16704,4	196632,1	16839,8
Самара	1164,9	20690,5	698362,0	80844,5
Саратов	837,8	18107,0	426384,2	36618,4
Ульяновск	613,8	16191,4	176427,7	26620,7

Вариант 9. Исследовать корреляционную зависимость среднегодовой численности работников *NP*, среднемесячной зарплаты *Z*, инвестиций в основной капитал *IK*, оборота розничной торговли *OT* по городам Приволжского федерального округа по приведенным в таблице данным.

Города	*NP*, тыс. чел.	*Z*, р.	*IK*, млн р.	*OT*, млн р.
Уфа	324,9	22089,5	36968,9	46811,7
Йошкар-Ола	79,9	15098,8	5323,4	8303,8
Саранск	106,9	14152,1	16466,5	6866,1
Казань	343,1	19410	84005,8	70231,8

Города	*NP*, тыс. чел.	*Z*, р.	*IK*, млн р.	*OT*, млн р.
Ижевск	217,4	17255,2	15943,7	22437
Чебоксары	143,0	15936,4	12771,6	13406,9
Пермь	307,0	22678,8	59720,8	49157,2
Киров	150,9	17722,4	17786,3	17378,2
Нижний Новгород	441,6	21821,2	45729,1	83069,5
Оренбург	163,7	18990,0	17267,2	22428,2
Пенза	152,6	16704,4	8251	16839,8
Самара	396,3	20690,5	45894	80844,5
Саратов	251,5	18107,0	21708	36618,4
Ульяновск	182,7	16191,4	11675,1	26620,7

Вариант 10. Исследовать корреляционную зависимость численности населения *N*, среднемесячной зарплаты *Z*, объема строительных работ *STR*, ввода в действие жилых домов *G* по городам Приволжского федерального округа по приведенным в таблице данным.

Города	*N*, тыс. чел.	*Z*, р.	*STR*, млн р.	*G*, тыс. м^2.
Уфа	1062,3	22089,5	52237,5	654,6
Йошкар-Ола	259,2	15098,8	2971,4	96,3
Саранск	297,4	14152,1	5724,3	156,1
Казань	1143,5	19410	29475,4	771,8
Ижевск	628,1	17255,2	5514,5	233,3
Чебоксары	453,6	15936,4	4614,9	254,4
Пермь	991,5	22678,8	16977,0	364,2
Киров	473,7	17722,4	3248,8	250,0
Нижний Новгород	1250,6	21821,2	17594,8	363,1
Оренбург	547,0	18990,0	7079,6	162,8
Пенза	517,1	16704,4	4269,2	588,9
Самара	1164,9	20690,5	12549,4	589,1
Саратов	837,8	18107,0	6177,3	665,3
Ульяновск	613,8	16191,4	4917,6	291,0

Данные для лабораторной работы № 3

Вариант 1. В течение шести лет использовались пять различных технологий по выращиванию пшеницы. Данные по урожайности (в ц/га) приведены в таблице. При уровне значимости $\alpha = 0{,}05$ установить значимость влияния различных технологий на урожайность пшеницы.

Номер наблюдения	*Технология (фактор)*				
	A1	*A2*	*A3*	*A4*	*A5*
1	24,1	12,2	18,2	33,5	20,2
2	22	22,3	12,3	28,2	27,8
3	20,2	16,1	16	26,3	22,4
4	26	14,4	19,8	29,7	18,1
5	21,8	13,6	20,3	24,1	23,8
6	16,3	18,2	22,1	25,8	30,1

Вариант 2. Наблюдаемые в течение пяти лет урожайности (в ц/га) четырех сортов пшеницы приведены в таблице. При уровне значимости $\alpha = 0{,}05$ установить значимость влияния различных сортов пшеницы на урожайность.

Номер наблюдения	Сорт пшеницы (фактор)			
	A1	A2	A3	A4
1	28,7	24,5	23,2	29
2	26,7	28,5	24,7	28,7
3	21,6	27,7	20	22,5
4	25	28,7	24	28
5	28,2	32,5	24	27
6	24,3	29,6	22,5	25,4

Вариант 3. Предприятие использует четыре линии по выпуску одной и той же продукции. На уровне значимости $\alpha = 0{,}05$ установить значимость влияния различных линий на процент брака продукции. Результаты измерений процента брака на различных линиях приведены в таблице.

Номер наблюдения	*Линия (фактор)*			
	L1	L2	L3	L4
1	0,6	0,2	0,8	0,7
2	0,2	0,2	0,6	0,7
3	0,4	0,4	0,2	0,3
4	0,5	0,3	0,4	0,3
5	0,8	0,3	0,9	0,2
6	0,6	0,6	1,1	0,8
7	0,7	0,8	0,8	0,6
8	0,3	0,5	0,7	0,2

Вариант 4. На тридцати студентах, обладающих примерно одинаковыми физическими данными и примерно одинаковым уровнем освоения текстового редактора, был проведен «слепой» эксперимент по определению влияния потребления кофеина на скорость их работы. Для этого они были разбиты на три группы по десять человек, первая группа получила 0 мг кофеина, вторая группа – 100 мг кофеина третья – 200 мг кофеина. После чего каждый набрал один и тот же текст. Время (в сек.), затраченное на набор текста, по группам (уровням фактора) приведено в таблице.

Группа (фактор)	*Наблюдения (время на набор текста студентом)*									
1-я группа	242	245	244	248	247	248	242	244	246	242
2-я группа	248	246	245	247	248	250	247	246	243	244
3-я группа	246	248	250	252	248	250	246	248	245	250

На уровне значимости $\alpha = 0{,}05$ установить значимость влияния уровня потребления кофеина на скорость ввода текста.

Вариант 5. Кирпичный завод использует для производства кирпича глину с трех различных карьеров. Результаты испытаний на прочность кирпичей из глины разных карьеров приведены в таблице.

Номер наблюдения	*Источник глины (фактор)*		
	Карьер 1	*Карьер 2*	*Карьер 3*
1	77	60	69
2	63	86	100
3	70	67	98
4	60	92	85
5	78	95	91
6	84	82	81
7		78	92

На уровне значимости $\alpha = 0{,}05$ установить значимость влияния глины на прочность кирпича.

Вариант 6. Данные производительности труда случайно выбранных рабочих одинаковой профессии на четырех предприятиях приведены в таблице.

Номер наблюдения	*Предприятия (фактор)*			
	Предприятие №1	*Предприятие №2*	*Предприятие №3*	*Предприятие №4*
1	1,3	1,4	1,44	1,27
2	1,27	1,3	1,4	1,05
3	1,21	1,28	1,28	1,24
4	1,09	1,27	1,28	1,22
5	1,03	1,24	1,06	
6	1,01	1,08		
7	1,09			

На уровне значимости $\alpha = 0{,}05$ установить существенно ли отличаются производительности труда рабочих одинаковой квалификации на этих предприятиях?

Вариант 7. Торговая фирма имеет четыре магазина расположенные в различных районах города. Объемы недельных продаж (тыс. р.) на квадратный метр торговой площади по этим магазинам приведены в нижеследующей таблице.

Номер наблюдения	*Объемы недельных продаж по магазинам*			
	Магазин №1	*Магазин №2*	*Магазин №3*	*Магазин №4*
1	51,3	52	55,9	54
2	59	57,8	56	58,1
3	53,5	66,3	57,8	62,4
4	58,6	69	58	64,1
5	63	70,1	70	66
6	69,1	72,6	74,5	67
7	72	74,2	78,3	69

На уровне значимости $\alpha = 0{,}05$ установить существенно ли отличаются объемы продаж в зависимости от района расположения магазинов?

Вариант 8. Четыре предприятия производят один и тот же прибор. Были произведены испытания приборов этих предприятий на время безотказной работы (в часах). Результаты испытаний приведены в таблице.

Номер испытания	*Время безотказной работы прибора (в часах) для различных предприятий*			
	Предприятие №1	*Предприятие №2*	*Предприятие №3*	*Предприятие №4*
1	92	105	83	81
2	102	75	84	70
3	104	87	94	72
4	113	89	86	81
5	109	93	90	85
6	120	96	97	83
7	128	102	100	99
8	143	106	113	89

На уровне значимости $\alpha = 0{,}05$ установить существенно ли отличаются приборы различных производителей по времени безотказной работы.

Вариант 9. На уровне значимости $\alpha = 0{,}05$ методом дисперсионного анализа исследовать влияние ежемесячного душевого дохода (тыс. р.) на еженедельные расходы (р.) на продуктовые товары семьи из трех человек по приведенным данным.

Номер наблюдения	Еженедельные расходы на продуктовые товары (р.)			
	Доход от 5 до 10 тыс. р.	*Доход от 10 до 15 тыс. р.*	*Доход от 15 до 20 тыс. р.*	*Доход от 20 до 30 тыс. р.*
1	1680	1600	2100	1850
2	1550	1820	1860	2100
3	1470	1780	1900	1600
4	1700	1900	1600	1860
5	1300	2100	1470	2200
6	1440	1850	2200	2300
7	1360	2200	1800	1950
8	1800		1750	

Вариант 10. На уровне значимости $\alpha = 0{,}05$ методом дисперсионного анализа исследовать влияние стажа работы по специальности на время выполнения одной и той же операции по приведенным данным.

Номер наблюдения	Время выполнения операции (мин.)			
	Стаж от 2 до 5 лет	*Стаж от 5 до 10 лет*	*Стаж от 10 до 15 лет*	*Стаж свыше 15 лет*
1	7,3	7,7	7,1	7,3
2	7,6	7,2	6,8	7,4
3	8,3	7,7	7,7	8,3
4	8,3	7,8	7,3	7,9
5	8,4	8,1	6,9	7,5
6	7,9	7,1	7,8	8,8
7	8,1	6,8	8,1	8,2
8	7,8	7,6	7,0	7,6

Данные для лабораторной работы № 4

Вариант 1. Построить линейную регрессию зависимости валового дохода Y торговых предприятий (млн руб.) от стоимости их основных фондов *X1* (млн р.) по приведенным в таблице данным по 12 предприятиям.

Y	203	63	45	113	121	88	110	56	80	237	160	75
X1	118	28	17	50	56	102	116	124	114	154	115	98

Вариант 2. Построить линейную регрессионную зависимость совокупных личных расходов Y (доллары) от личного дохода X (доллары) по данным США за 1972–1983 годы, приведенным в таблице.

X	951,4	1008	1005	1011	1056	1105	1162	1201	1209	1249	1254	1285
Y	737	769	764	780	823	864	903	928	932	951	963	1009

Вариант 3. Построить линейную регрессию зависимости себестоимости Y (тыс. р.) единицы продукции предприятия от объема выпуска продукции X (тыс. шт.) по данным приведенным в таблице.

X	1,5	1,7	1,9	2,0	2,2	2,3	2,6	2,8	3,1	3,2	3,5	3,7
Y	4,5	4,6	4,3	4,1	4,0	4,1	3,6	3,6	3,3	3,3	3,4	3,2

Вариант 4. Построить линейную регрессию зависимости цены Y (в центах) галлона бензина от стоимости X (доллары) барреля сырой нефти по данным США за 1975–1988 годы приведенным в таблице.

X	7,7	8,2	8,57	9,0	12,64	21,6	31,77	28,52	26,2	25,88	24,1	12,5	15,4	12,57
Y	57	59	62	63	86	119	133	122	116	113	112	86	90	90

Вариант 5. Построить линейную регрессионную зависимость расходов на питание Y (доллары) от личного дохода X (доллары) по данным США за 1972–1983 годы, приведенным в таблице.

X	951,4	1008	1005	1011	1056	1105	1162	1201	1209	1249	1254	1285
Y	132,4	129,4	128,1	132,3	139,7	145,2	146,1	149,3	153,2	153	154,6	161,2

Вариант 6. Построить линейную регрессионную зависимость ежемесячных затрат на техническое обслуживание Y (тыс. р.) автомобиля от его ежемесячного пробега L (тыс. км) по приведенным данным.

Пробег L, км	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Затраты Y, тыс. р.	13	16	15	20	19	21	26	21	30	32	30	35	34	40	39

Вариант 7. Построить линейную регрессионную зависимость оборота розничной торговли *OT* от численности населения *N* по данным городов Приволжского федерального округа приведенным в таблице.

N, тыс. чел.	1062,3	259,2	297,4	1143,5	628,1	453,6	991,5	473,7	1250,6
OT, млн. р.	46812	8304	6866	70232	22437	13407	49157	17378	83069

N, тыс. чел.	547,0	517,1	1164,9	837,8	613,8
OT, млн. р.	22428	16840	80845	36618	26621

Вариант 8. Построить линейную регрессионную зависимость оборота розничной торговли *OT* от среднемесячной заработной платы *Z* по данным городов Приволжского федерального округа приведенным в таблице.

Z	22089	15099	14152	19410	17255	15936	22679	17722	21821
OT, млн. р.	46812	8304	6866	70232	22437	13407	49157	17378	83069

Z	18990	16704	20690	18107	16191
OT, млн. р.	22428	16840	80845	36618	26621

Вариант 9. Построить линейную регрессионную зависимость численности населения *N* от стоимости основных фондов *K* по данным городов Приволжского федерального округа приведенным в таблице.

N, тыс. чел.	1062,3	259,2	297,4	1143,5	628,1	453,6	991,5	473,7
K, млн. р.	595304	69051	89430	620442	233387	105410	472475	175145

N, тыс. чел.	1250,6	547,0	517,1	1164,9	837,8	613,8
K, млн. р.	576345	516399	196632	698362	426384	176428

Вариант 10. Построить линейную регрессионную зависимость годовой прибыли предприятия (Y, млн р.) от суточного выпуска продукции на одного работающего (X, тыс. р.) по следующим данным.

Y	54	21	20	19	31	80	69	52	41	60	72	71
X	14	11	11	10	12	18	15	13	13	15	16	17

Данные для лабораторной работы № 5

Вариант 1. Построить нелинейную регрессию $y = a + b \cdot x^2$ зависимости наполняемости гостиниц Y(%) приморского курорта от расстояния L (км) до пляжа по следующим данным.

Расстояние L, км	0,1	0,1	0,2	0,3	0,4	0,4	0,5	0,6	0,7	0,7	0,8	0,8	0,9	0,9
Наполняемость Y, %	92	95	94	90	89	86	90	83	84	80	74	76	72	71

Вариант 2. Построить нелинейную регрессию $y = \frac{1}{a+b \cdot x}$ зависимости ежемесячных затрат на техническое обслуживание Y (тыс. руб.) автомобиля от его ежемесячного пробега L (тыс. км) по следующим данным.

Пробег L, км	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Затраты Y, тыс. р.	14	15	15	16	17	19	21	21	22	23	26	31	37	41	40

Вариант 3. Построить нелинейную регрессию $y = a + b \cdot \frac{1}{\cdot x}$ зависимости годового потребления бананов Y (в фунтах) и годового дохода X (в десятках тыс. дол.) по результатам обследования 10 семей США. Результаты обследования приведены в таблице.

X (тыс. дол)	1	2	3	4	5	6	7	8	9	10
Y (фунты)	1,93	7,13	8,78	9,69	10,09	10,42	10,62	10,71	10,79	11,13

Вариант 4. Построить нелинейную регрессию $y = \frac{1}{1+e^{a+bx}}$ зависимости доли Y годовых доходов семьи из трех человек направляемых на отдых от совокупного годового дохода X (млн р.) по результатам обследования 12 семей приведенным в таблице.

X (млн р.)	0,5	0,6	0,7	0,8	0,9	1,0	1,2	1,4	1,5	1,7	1,8	2,0
Y	0,02	0,025	0,02	0,035	0,032	0,041	0,054	0,07	0,076	0,101	0,13	0,16

Вариант 5. Построить нелинейную регрессию $y = e^{a+b \cdot x}$ зависимости производительности труда Y (тыс. р./день) рабочих одинаковой квалификации от фондовооруженности X (тыс. р.) по приведенным в таблице данным.

X	108	114	125	137	142	150	152	154	158	161	162	163	165	168
Y	6,9	6,8	7,3	7,2	8,4	8,8	9,1	10,6	10,7	9,8	11,1	12,1	12,4	12,8

Вариант 6. Построить нелинейную регрессию $y = a + b \cdot lnx$ зависимости прочности нити на разрыв Y (кг) от процента примесей X (%) в материале нити по приведенным в таблице данным.

X	1,2	1,4	1,5	1,7	1,8	2	2,2	2,5	2,9	3,4
Y	16,5	15,4	14,1	12,4	10,1	9,2	8,1	7,0	6,3	5,8

Вариант 7. Построить нелинейную (логистическую) регрессию $y = \frac{1}{1+e^{a+bx}}$ зависимости средней доли *Y* заполнения базы отдыха от температуры *T* (в градусах) воды в соседнем озере по приведенным в таблице данным.

T	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Y	0,16	0,18	0,28	0,45	0,53	0,64	0,7	0,77	0,82	0,87	0,9	0,95	0,98	0,96	0,98

Вариант 8. Построить нелинейную регрессию $y = e^{a+b \cdot x}$ зависимости себестоимости Y (тыс. р.) единицы продукции предприятия от объема выпуска продукции X (тыс. шт.) по данным приведенным в таблице.

X	1,5	1,7	1,9	2,0	2,2	2,3	2,6	2,8	3,1	3,2	3,5	3,7
Y	45,1	45,6	42,9	40,8	40,0	41,1	35,9	36,1	33,2	33,1	33,8	32,7

Вариант 9. Построить нелинейную регрессию $y = a + b \cdot lnx$ зависимости годового оборота розничной торговли Y (млрд р.) от совокупного месячного дохода X (млрд р.) всех работников по данным для городов Приволжского федерального округа приведенным в таблице.

Город	Совокупный месячный доход, *X*	Оборот розничной торговли, *Y*
Уфа	7,18	46,81
Йошкар-Ола	1,21	8,30
Саранск	1,51	6,87
Казань	6,66	70,23
Ижевск	3,75	22,44
Чебоксары	2,28	13,41
Пермь	6,96	49,16
Киров	2,67	17,38
Нижний Новгород	9,64	83,07
Оренбург	3,11	22,43
Пенза	2,55	16,84
Самара	8,20	80,84
Саратов	4,55	36,62
Ульяновск	2,96	26,62

Вариант 10. Построить нелинейную регрессию $y = a + b \cdot \frac{1}{x}$ зависимости среднемесячной зарплаты *Y* (тыс. р.) от средней фондовооруженности *X* (тыс. р.) по данным приведенным в таблице.

X	1,83	0,86	0,84	1,81	1,07	0,74	1,54	1,16	1,31	1,29	1,76	1,69	0,97
Y	22,1	15,1	14,1	19,4	17,3	15,9	22,7	17,7	21,8	16,7	20,1	18,1	16,2

Данные для лабораторной работы № 6

Вариант 1. Используя данные лабораторной работы № 2 построить линейную множественную регрессию зависимости себестоимости 1 т. литья (*Y*, тыс. р.) от выпуска продукции на одного работающего (*X*, т.) и процента брака литья (*Z*, %).

Вариант 2. Используя данные лабораторной работы № 2 построить линейную множественную регрессию зависимости прибыли предприятия (*Y*, млн р.), выпуска продукции на одного работающего (*X*, единиц) и доли продукции, производимой на экспорт (*Z*, %).

Вариант 3. Используя данные лабораторной работы № 2 построить линейную множественную регрессию зависимости индекса человеческой бедности *X2* от душевого дохода *Y* (дол.) и индекса человеческого развития *X1*.

Вариант 4. Используя данные лабораторной работы № 2 построить линейную множественную регрессию зависимости выработки продукции на одного рабочего *Y* (тыс. р.) от ввода новых основных фондов *X1* (% от стоимости основных фондов на конец года) и удельного веса рабочих высокой квалификации в общей численности рабочих *X2*.

Вариант 5. Используя данные лабораторной работы № 2 построить линейную множественную регрессию зависимости валового дохода *Y* торговых предприятий (млн р.) от стоимости их основных *X1* (млн р.) и оборотных *X2* (млн р.) средств.

Вариант 6. Используя данные лабораторной работы № 2 построить линейную регрессионную зависимость чистого дохода *Y* (млрд дол.) от оборота капитала *OK* (млрд дол.), использованного капитала *K* (млрд дол.) и численности занятых *L* (тыс. чел.).

Вариант 7. Используя данные лабораторной работы № 2 построить линейную регрессионную зависимость объема месячных продаж пива *Y* (условных ед.) от расходов на рекламу *R* (условных ед.), числа туристов *T* (тыс. чел.) и индекса цен *P* (%).

Вариант 8. Используя данные лабораторной работы № 2 построить линейную регрессионную зависимость оборота розничной торговли *OT* от численности населения *N* и среднемесячной зарплаты *Z*.

Вариант 9. Используя данные лабораторной работы № 2 построить линейную регрессионную зависимость оборота розничной торговли *OT* от среднегодовой численности работников *NP* и среднемесячной зарплаты *Z*.

Вариант 10. Используя данные лабораторной работы № 2 построить линейную регрессионную зависимость ввода в действие жилых домов *G* от численности населения *N*, среднемесячной зарплаты *Z*.

Данные для лабораторной работы № 7

Исследовать на мультиколлинеарность и автокорреляцию ошибок модель множественной регрессии, построенную в лабораторной работе № 6.

Данные для лабораторной работы № 8

Вариант 1. Исследовать зависимость заработной платы *Y* (тыс. р.) от стажа работа *X* (лет) и пола работника по приведенным данным.

Y	29	42	30	30	20	35	35	43	38	42
X	9	21	16	12	3	25	18	20	30	27
Пол	Ж	М	Ж	Ж	М	Ж	Ж	М	Ж	М

Y	25	35	20	40	25	32	41	39	26	25
X	8	10	5	28	10	20	20	18	9	5
Пол	Ж	М	М	М	Ж	М	М	М	Ж	М

Вариант 2. Исследовать зависимость общего количества баллов *Y* набранных студентом за семестр от количества баллов *X* полученных по математике и пола, обучающегося по приведенным данным.

Y	476	457	540	551	575	698	545	574	645	556	634	637	390	562	560	510
X	58	48	65	70	81	95	73	83	100	98	77	91	44	81	90	62
Пол	Ж	М	М	Ж	М	Ж	Ж	М	Ж	Ж	М	М	Ж	М	Ж	М

Вариант 4. Две одинаковые по численности и оснащению бригады лесорубов производят заготовку леса. Исследовать зависимость недельной заготовки леса *Y* ($м^3$) от расстояния *X (км)* до делянки и бригады лесорубов по приведенным в таблице данным.

Y	412	476	457	540	551	575	698	545	574	645	556	634	637	390	462	560
X	25	23	24	18	19	20	12	21	22	14	23	15	12	24	23	22
Бр.	1	2	1	1	2	2	2	1	1	2	1	2	1	2	1	2

Вариант 5. Исследовать зависимость стоимости *Y* (тыс. долл.) трехкомнатной квартиры на вторичном рынке от ее площади *X* ($м^2$), района расположения (центральный – Ц, другие – Д) и типа дома (кирпичный – К, панельный – П) по приведенным данным.

Y	15,5	38	30	24	32,5	43	17,8	28	32,7	31	33
X	68,1	107	100	71	98	100	58	75	85	66	81
Район	Ц	Ц	Ц	Ц	Ц	Д	Д	Д	Д	Д	Д
Тип	П	П	П	К	П	К	П	К	П	П	П

Y	28	21,5	15,3	21	35,5	22	29	16	22	23	19,5
X	76,4	55	53,7	57	62	74	70	80	62	69,7	79
Район	Д	Д	Д	Д	Д	Д	Д	Д	Д	Д	Д
Тип	П	К	К	П	П	П	П	П	К	П	К

Вариант 6. Исследовать зависимость стоимости квадратного метра *Y* (тыс. долл.) на вторичном рынке от полезной площади квартиры *X* ($м^2$), района расположения (центральный – Ц, другие – Д) и типа дома (кирпичный – К, панельный – П) по приведенным в таблице данным.

Y	0,228	0,352	0,51	0,336	0,332	0,43	0,31	0,373	0,384	0,47	0,41
X	44	58	58	52	51	45	39	40	59	48	52
Район	Ц	Ц	Ц	Ц	Ц	Д	Д	Д	Д	Д	Д
Тип	П	П	П	К	П	К	П	К	П	П	П

Y	0,37	0,4	0,283	0,37	0,572	0,3	0.414	0.2	0,355	0,333	0,2575
X	49	40,5	37,6	38	52	47	45	54	37	42	50,3
Район	Д	Д	Д	Д	Д	Д	Д	Д	Д	Д	Д
Тип	П	К	К	П	П	П	П	П	К	П	К

Вариант 7. Исследовать зависимость ВВП страны от года t (t=1, 2, …, 17) и квартала года по приведенным квартальным данным.

Квартал	Год	t	ВВП	Квартал	Год	t	ВВП
IV	1994	1	225	I	1999	18	901
I	1995	2	235	II		19	1102
II		3	325	III		20	1373
III		4	421	IV		21	1447
IV		5	448	I	2000	22	1527
I	1996	6	425	II		23	1697
II		7	469	III		24	2038
III		8	549	IV		25	2044
IV		9	565	I	2001	26	1922
I	1997	10	513	II		27	2120
II		11	555	III		28	2536
III		12	634	IV		29	2461
IV		13	641	I	2002	30	2268
I	1998	14	551	II		31	2523
II		15	602	III		32	3074
III		16	676	IV		33	2998
IV		17	801	I	2003	34	2893

Вариант 8. Исследовать зависимость стоимости квартиры *Y* (тыс. дол.) на рынке жилья от полезной площади квартиры *X* ($м^2$) и количества комнат по приведенным в таблице данным.

Y	27	21,1	28,7	27,2	28,3	28	45	34,4	30,8	29	27,7
X	68,4	54,7	74,7	71,7	74,5	53	86	62,6	56,4	67,5	69,1
Количество комнат	3	2	3	3	3	2	3	2	2	3	3

Y	34,1	37,7	41,9	36,7	26,4	34,2	35,6	34	58,5	46,6	58,5
X	68,1	75,3	83,7	68,6	48,6	68,5	71,1	68	117	93,2	117

Количество комнат	2	2	3	2	2	2	2	3	3	2	3

Вариант 9. Исследовать зависимость месячного потребления чая *Y* (кг.) от его стоимости *X* (тыс. р. за кг.), пола потребителей (М, Ж) и региона (северный, южный) по приведенным в таблице данным.

№ п/п	Потребление чая	Цена чая	Пол	Регион
1	0,19	1,10	М	Северный
2	0,45	1,00	Ж	Северный
3	0,45	0,81	М	Северный
4	0,58	0,80	Ж	Северный
5	0,52	0,64	М	Северный
6	0,61	0,60	Ж	Северный
7	0,60	0,51	Ж	Северный
8	0,65	0,52	М	Северный
9	0,62	0,32	М	Северный
10	0,69	0,30	Ж	Северный
11	0,53	1,0	М	Южный
12	0,60	1,0	Ж	Южный
13	0,71	0,91	М	Южный
14	0,89	0,82	Ж	Южный
15	0,91	0,65	М	Южный
16	1,10	0,60	Ж	Южный
17	1,05	0,53	Ж	Южный
18	1,24	0,50	М	Южный
19	1,26	0,35	М	Южный
20	1,42	0,38	Ж	Южный

Вариант 10. Исследовать урожайность *Y* (ц/га) двух сортов пшеницы (сорт А и сорт В) в зависимости от двух используемых технологий (Тех. 1 и Тех. 2) и количества внесенных органических удобрений *X* (т/га) по данным приведенным в таблице.

№ п/п	Сорт	Технология	Удобрения, *X*	Урожайность, *Y*
1	А	Тех. 1	15	16
2	А	Тех. 1	10	14
3	А	Тех. 1	22	24
4	А	Тех. 1	19	23
5	А	Тех. 1	18	26
6	А	Тех. 1	16	24
7	А	Тех. 2	24	29
8	А	Тех. 2	17	24
9	А	Тех. 2	20	31
10	А	Тех. 2	19	33
11	А	Тех. 2	23	34
12	А	Тех. 2	17	28
13	В	Тех. 1	13	16
14	В	Тех. 1	16	19
15	В	Тех. 1	22	25
16	В	Тех. 1	19	22
17	В	Тех. 1	18	26
18	В	Тех. 1	17	24

№ п/п	Сорт	Технология	Удобрения, X	Урожайность, Y
19	В	Тех. 2	22	28
20	В	Тех. 2	15	18
21	В	Тех. 2	20	30
22	В	Тех. 2	21	33
23	В	Тех. 2	23	35
24	В	Тех. 2	14	17

Данные для лабораторных работ № 9 и 10

Вариант 1. Объем продаж y_t (тыс. р.) нового товара по неделям t.

t	1	2	3	4	5	6	7	8	9	10
y	27,3	41,8	42,8	56,2	72,5	56	79	74,9	103,3	111,3
t	11	12	13	14	15	16	17	18	19	20
y	125,2	189,3	169,1	193,5	207,4	221,2	267,2	264,0	273,8	321

Вариант 2. Квартальные совокупные потребительские расходы населения США за 1977–1982 гг. (в млрд дол. 1972 г.)

Квартал	Год	t	Расходы
I	1977	1	201,8
II		2	214,5
III		3	216,6
IV		4	230,9
I	1978	5	211,6
II		6	224,8
III		7	226,3
IV		8	240,0
I	1979	9	221,0
II		10	229,0
III		11	231,2
IV		12	245,7
I	1980	13	224,2
II		14	230,2
III		15	231,7
IV		16	244,1
I	1981	17	229,4
II		18	239,3
III		19	240,0
IV		20	247,5
I	1982	21	232,8
II		22	243,4
III		23	242,1
IV		24	251,5

Вариант 3. Среднечасовая заработная плата Y в экономике США, в сопоставимых ценах 1982 г. (дол.), в 1971–1990 гг.

Год	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Y	8,21	8,53	8,55	8,28	8,12	8,24	8,36	8,40	8,17	7,78

Год	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Y	7,69	7,68	7,79	7,80	7,77	7,81	7,73	7,69	7,64	7,53

Вариант 4. Среднегодовые цены *Y* на каучук в Нью-Йорке, центы за фунт.

Год	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Y	21,1	18,0	18,1	35,1	39,7	29,8	39,5	41,5	49,9	64,2	73,4	56,9	45,3
Год	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Y	56,1	49,6	41,8	41,2	44,1	48,8	48,7	50,2	47,6	46,6	47,3	48,9	56,7

Вариант 5. Среднегодовые цены *Y* на говядину в Нью-Йорке, центы за фунт.

Год	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Y	42	49	64	53	44	52	51	71	92	87
Год	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Y	86	99	96	97	89	77	81	82	87	94

Вариант 6. Квартальные потребительские расходы на питание населения США за 1977–1982 гг. (в млрд дол. 1972г.)

Квартал	Год	t	расходы
I	1977	1	37,4
II		2	42,7
III		3	44,5
IV		4	45,8
I	1978	5	40,7
II		6	42,9
III		7	43,6
IV		8	44,5
I	1979	9	40,6
II		10	43,8
III		11	44,9
IV		12	46,6
I	1980	13	42,8
II		14	45,5
III		15	45,0
IV		16	47,3
I	1981	17	40,2
II		18	45,0
III		19	45,5
IV		20	50,9
I	1982	21	40,9
II		22	46,0
III		23	47,2
IV		24	49,8

Вариант 7. Объем платных услуг (млн р.) населению региона по кварталам 1996–1999 гг.

Квартал	Год	t	расходы
I	1996	1	2428
II		2	2010
III		3	2981
IV		4	3074
I	1997	5	2893
II		6	3198

III		7	3250
IV		8	3495
I	1998	9	3528
II		10	3838
III		11	3916
IV		12	4142
I	1999	13	4441
II		14	5583
III		15	6230
IV		16	6467

Вариант 8. ВВП страны по кварталам за 1994–2003 годы

Квартал	Год	t	ВВП	Квартал	Год	t	ВВП
IV	1994	1	225	I	1999	18	901
I	1995	2	235	II		19	1102
II		3	325	III		20	1373
III		4	421	IV		21	1447
IV		5	448	I	2000	22	1527
I	1996	6	425	II		23	1697
II		7	469	III		24	2038
III		8	549	IV		25	2044
IV		9	565	I	2001	26	1922
I	1997	10	513	II		27	2120
II		11	555	III		28	2536
III		12	634	IV		29	2461
IV		13	641	I	2002	30	2268
I	1998	14	551	II		31	2523
II		15	602	III		32	3074
III		16	676	IV		33	2998
IV		17	801	I	2003	34	2893

Вариант 9. Условное душевое потребление молока по кварталам (литров в месяц).

Квартал	Год	t	Потребление молока
I	1	1	4,3
II		2	7,4
III		3	8,4
IV		4	4,5
I	2	5	5,1
II		6	8,9
III		7	9,6
IV		8	5,1
I	3	9	5,8
II		10	9,4
III		11	9,9
IV		12	6
I	4	13	6,3
II		14	9,4
III		15	10
IV		16	6,1
I		17	6,9

II	5	18	9,9
III		19	10,1
IV		20	6,5

Вариант 10. Душевое потребление *Y* масла в январе за 1971–1986 годы (в граммах).

Год	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981
Y	249	230	320	460	300	239	415	516	529	543	570
Год	1982	1983	1984	1985	1986						
Y	517	512	497	551	510						

Учебное издание

Абдуллин Рафаэль Зинатович
Абдуллин Владимир Рафаэлевич

ЭКОНОМЕТРИКА В MS EXCEL

Издается в авторской редакции

ИД № 06318 от 26.11.01.
Подписано в пользование 06.04.16.

Издательство Байкальского государственного университета.
664003, г. Иркутск, ул. Ленина, 11.