О.В. Леонова, Н.П. Шерстянкина

ЭКОНОМЕТРИКА
КУРС ЛЕКЦИЙ И МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ РАСЧЕТНО-ГРАФИЧЕСКИХ РАБОТ

Министерство образования и науки Российской Федерации
Байкальский государственный университет

О.В. Леонова, Н.П. Шерстянкина

ЭКОНОМЕТРИКА
КУРС ЛЕКЦИЙ И МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ РАСЧЕТНО-ГРАФИЧЕСКИХ РАБОТ

Иркутск
Издательство БГУ
2017

УДК 330.43 (075.8)
ББК 65в631я7
Л47

Печатается по решению редакционно-издательского совета
Байкальского государственного университета

Рецензенты канд. ф.-м. наук, доц. А.В. Бурдуковская
канд. ф.-м. наук, доц. Н.В. Мамонова

Леонова О.В.
Л47 Эконометрика. Курс лекций и методические указания по выполнению расчетно-графических работ [Электронный ресурс] : учеб. пособие / О.В. Леонова, Н.П. Шерстянкина. – Иркутск: Изд-во БГУ, 2017. – 157 с. – Режим доступа: http://lib-catalog.isea.ru.

Учебное пособие охватывает все разделы программы курса «Эконометрика». Объединенное и взаимосвязанное изложение основ математической статистики и методов эконометрического моделирования способствует цельному и системному их восприятию, а методические указания – самостоятельному выполнению расчетно-графических работ с использованием MS Excel.

Рекомендуется для студентов всех специальностей БГУ.

УДК 330.43 (075.8)
ББК 65в631я7

Оглавление

Раздел 1. Основы математической статистики

Введение

Математическая статистика – наука, которая изучает методы получения, описания и обработки опытных данных с целью изучения закономерностей массовых случайных явлений. Задачи математической статистики в некотором смысле обратны задачам теории вероятностей. В теории вероятностей мы исходим из того, что теоретико-вероятностная модель процесса задана и мы производим расчет возможного реального течения этого процесса. В математической статистике наоборот требуется по статистическим данным подобрать подходящую теоретико-вероятностную модель изучаемого явления или процесса.

Все задачи математической статистики касаются вопросов обработки опытных данных, но в зависимости от характера измеряемой величины и цели исследования могут принимать ту или иную форму. Можно выделить три основных типа задач:

1. Первичная статистическая обработка данных или описательная (дескриптивная) статистика. На этом этапе данные необходимо представить в виде рядов, графиков, вычислить сводные характеристики выборки и получить предварительные сведения о законе распределения изучаемой случайной величины.

2. Статистическое оценивание неизвестных параметров. Предполагается, что изучаемая случайная величина имеет закон распределения определенного вида. Однако, параметры, определяющие этот закон неизвестны, требуется по результатам наблюдений найти приближенные значения этих параметров – оценки параметров либо в виде одного числа, либо в виде интервала.

3. Статистическая проверка гипотез. На разных стадиях исследования возникает необходимость в формулировании и проверке некоторых предположений – гипотез, касающихся либо значений неизвестных параметров, либо вида предполагаемого закона распределения, либо наличия связей (корреляции) между величинами. Процедура обоснования сопоставления выдвинутой гипотезы с имеющимися данными осуществляется с помощью специально сконструированного критерия и называется статистической проверкой гипотез.

1. Дескриптивная (описательная) статистика

Рассмотрим способы представления исходных данных (выборки заданного объема) в компактном виде (в виде рядов, графиков), определение числовых характеристик выборки, а также обсудим эмпирические аналоги теоретических распределений и выводы, которые можно сделать по элементам описательной статистики.

1.1. Эмпирические распределения и их графические представления

Выборка $\{x_1, x_2, \ldots, x_n\}$ объема n имеющихся в нашем распоряжении значений исследуемой случайной величины X является той исходной информацией, на основании которой строятся выводы о свойствах изучаемой генеральной совокупности в целом и, в частности, составляется представление о функции и ряде распределения или плотности анализируемого закона распределения вероятностей.

Упорядоченная по величине последовательность выборочных значений $x_1^{(n)} \le x_2^{(n)} \le \ldots \le x_n^{(n)}$ называется вариационным рядом. Среди членов вариационного ряда могут быть совпадающие между собой значения. Если через $n_1, n_2, \ldots, n_r$ обозначить общее число повторений всех несовпадающих значений выборки, то получим два ряда чисел:

x_i	x_1	x_2	$\ldots$	x_r
n_j	n_1	n_2	$\ldots$	n_r

$$\sum_{j=1}^{r} n_j = n. \qquad (1.1)$$

Первый ряд содержит различные выборочные значения, расположенный в порядке возрастания. Числа второго ряда показывают количество повторений каждого из этих значений в выборке и называются частотами. Ряд (1.1) называется точечным вариационным рядом, что соответствует дискретной вариации признака, или эмпирическим распределением признака по частотам.

От распределения частот (т. е. ряда (1.1)) можно перейти к распределению относительных частот $w_i = \dfrac{n_i}{n}, \sum_{i=1}^{r} w_i = 1,$ заданных в виде доли w_i или в виде процента $w_i \cdot 100\%$ $(\sum_{t=1}^{r} w_i \cdot 100\% = 100\%)$:

x_i	x_1	x_2	$\ldots$	x_r
w_i	w_1	w_2	$\ldots$	w_r

$$, \qquad \sum_{i=1}^{r} w_i = 1. \qquad (1.2)$$

Вариационный ряд (1.2), построенный по относительным частотам, является статистической аппроксимацией ряда распределения вероятностей случайной величины X.

Если объем выборки n велик ($n > 50$) и при этом мы имеем дело с непрерывной случайной величиной (или дискретной, число возможный значений которой достаточно велико), то часто удобнее, с точки зрения дальнейшей статистической обработки результатов наблюдений, перейти к интервальному вариационному ряду или группированной выборке. Этот переход осуществляется следующим образом:

1) отмечаются наименьше $x_{\min}$ и наибольшее $x_{\max}$ значения в выборке;

2) весь диапазон $[x_{\min}; x_{\max}]$ разбивается на k равных интервалов группирования (количество интервалов не должно быть меньше 8-10 и больше 20-25), выбор числа интервалов существенно зависит от объема выборки; для примерной ориентации в выборе k можно пользоваться приближенной формулой $k \approx 1 + \log_2 n$ либо $k \approx 1 + 3{,}32 \ln n$;

3) определяется величина шага или ширина интервала группирования h, для чего вариационный размах $R = x_{\max} - x_{\min}$ делится на число интервалов k:

$$h = \frac{x_{\max} - x_{\min}}{k};$$

4) находятся крайние точки каждого из интервалов: $C_0 = x_{\min}$, $C_1 = C_0 + h$, $C_2 = C_1 + h, ..., C_k = C_{k-1} + h$, а также их середины: $x_1^*, x_2^*, ..., x_k^*$;

5) подсчитываются числа выборочных данных, попавших в каждый из интервалов: $n_1, n_2, ..., n_k$ (очевидно, $n_1 + n_2 + ... + n_k = n$); выборочные данные, попавшие на границы интервалов, либо равномерно распределяются по двум соседним интервалам, либо относятся только к какому-либо из них, например, к левому.

Таким образом, следуя этой методике, от ряда (1.1) или (1.2) можно перейти к интервальному вариационному ряду

$C_i - C_{i+1}$	$C_0 - C_1$	$C_1 - C_2$	$\dots$	$C_{k-1} - C_k$
n_i	n_1	n_2	$\dots$	n_k

$$, \quad \sum_{i=1}^{k} n_i = n. \qquad (1.3)$$

От интервального ряда (1.3) можно вновь перейти к точечному, если в качестве значения случайной величины, соответствующего i-му интервалу, взять его середину $x_i^* = (C_i + C_{i+1})/2$. Получим ряд

x_i^*	x_1^*	x_2^*	$\dots$	x_k^*
n_i	n_1	n_2	$\dots$	n_k

$$, \quad \sum_{i=1}^{k} n_i = n. \qquad (1.4)$$

В некоторых задачах от ряда (1.1) или (1.4) целесообразно перейти к ряду, содержащему кумулятивные или накопленные частоты m_i. **Накопленная** (интегральная или кумулятивная) частота m_i значения x_i, получается суммированием частот значений, предшествующих данному, с частотой n_i, т.е. $m_i = n_1 + n_2 + ... + n_i$. Накопленная частота крайнего правого значения (или максимального элемента выборки) равна объему выборки n. Несмотря на видимую

несхожесть, ряды (1.1) – (1.4) отражают одно и то же фактическое распределение признака.

Замечание. Предложенную процедуру построения вариационных рядов ни в коем случае не следует считать единственно возможной. Количество интервалов, их длины, а также расположение интервалов относительно выборочного материала могут варьироваться по усмотрению исследователя в зависимости от решаемых задач.

Для наглядного представления вариационные ряды изображают в виде графиков. Наиболее распространенными способами представления эмпирических данных является гистограмма, полигон частот или относительных частот и полигон накопленных частот, или кумулятивная кривая – кумулята.

Гистограмма состоит из последовательности примыкающих друг к другу прямоугольников (рис. 1.1). Ширина этих прямоугольников равна ширине интервалов группирования h и откладывается по оси абсцисс, а высота откладывается по оси ординат и пропорциональна частоте n_i или относительной частоте w_i. В первом случае имеем гистограмму частот с высотами прямоугольников, равными n_i / h, и общей площадью, равной объему выборки n. Во втором – гистограмму относительных частот с высотами прямоугольников $n_i / n \cdot h$, и общей площадью, равной 1. Ступенчатая ломаная, ограничивающая в этом случае сверху построенную фигуру, является статистической аппроксимацией кривой распределения или графика функции плотности вероятности $f(x)$ исследуемой случайной величины X. Эту же аппроксимацию мы получим, если через середины верхних оснований прямоугольников проведем плавную линию (пунктир).

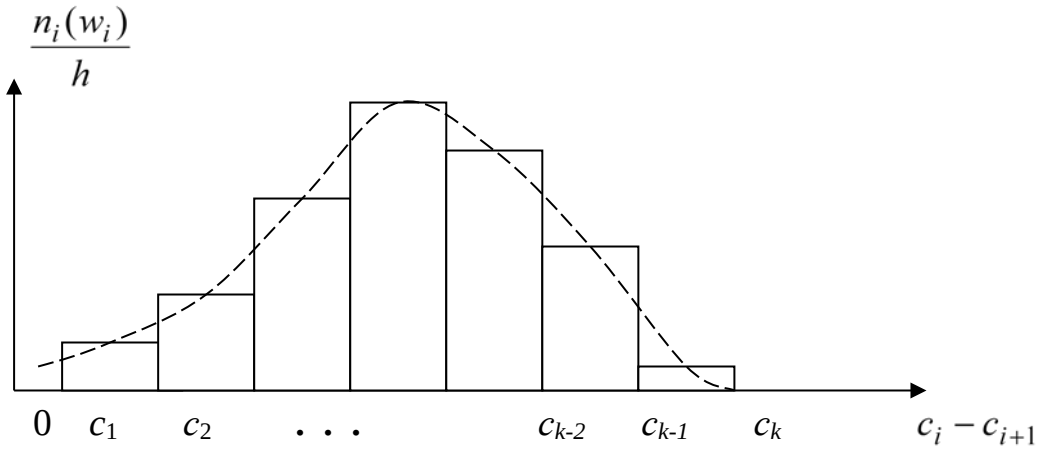


Рис. 1.1. Гистограмма (относительных) частот

Полигон частот или относительных частот представляет собой многоугольник с вершинами в точках (x_i, n_i) или (x_i, w_i) (рис. 1.2).

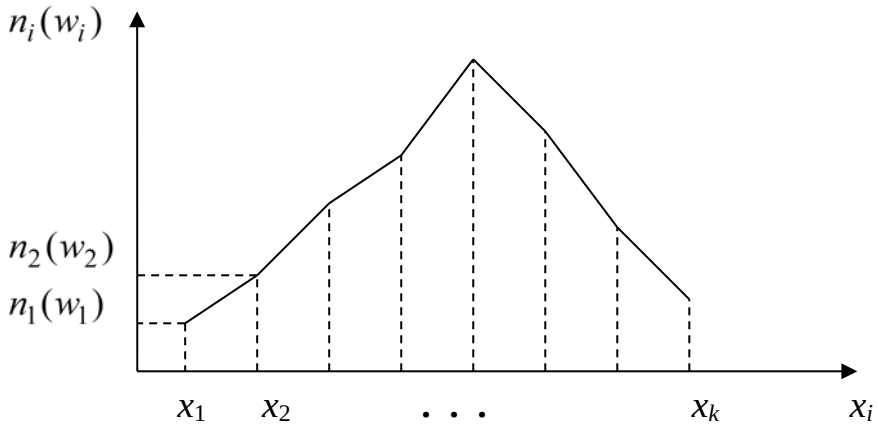


Рис. 1.2. Полигон (относительных) частот

При изображении полигона частот или относительных частот интервального вариационного ряда вершины многоугольника расположены в точках с абсциссами, соответствующими срединным значениям интервалов x_i^*, и ординатами, равными частоте n_i или относительной частоте w_i.

Полигон накопленных частот (кумулята) получается изображением в прямоугольной системе координат вариационного ряда с накопленными частотами. При построении кумуляты дискретного признака на ось абсцисс наносятся значения признака – элементы выборки x_i. Ординатами служат вертикальные отрезки – накопленные частоты m_i (рис. 1.3).

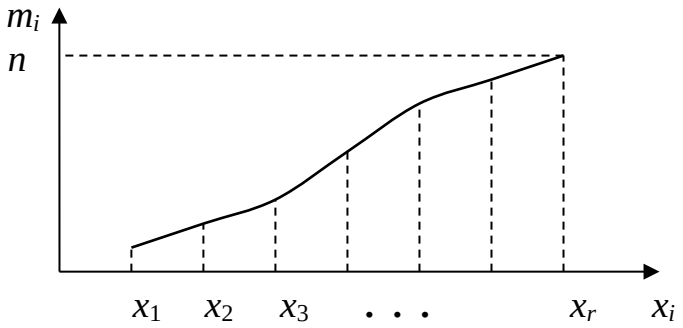


Рис. 1.3. Кумулята

1.2. Эмпирическая или статистическая функция распределения

Пусть n_x – число элементов выборки $\{x_1, x_2, ..., x_n\}$, меньших x. **Статистической** или **эмпирической** функцией распределения называется функция

$$F^*(x) = \frac{n_x}{n}.$$

Иначе, статистическая функция распределения $F^*(x)$ есть относительная частота события $\{X < x\}$ в серии из n независимых измерений случайной величины X. На основании закона больших чисел статистическая функция распределения сходится по вероятности к теоретической функции распределения $F(x)$

генеральной совокупности, когда объем n неограниченно возрастает. Следовательно, $F^*(x)$ является статистической аппроксимацией функции распределения $F(x) = P\{X < x\}$, ее приближенным значением и обладает следующими свойствами:

1) значения $F^*(x)$ принадлежат отрезку [0,1];

2) $F^*(x)$ – неубывающая функция;

3) если $x_{\max}$ – наибольший элемент выборки, а $x_{\min}$ – наименьший, то

$$F^*(x) = \begin{cases} 0, & x \le x_{\min}, \\ 1, & x > x_{\max}; \end{cases}$$

4) $F^*(x)$ непрерывна слева.

Для выборки, представленной рядом (1.1), эмпирическая функция распределения $F^*(x)$ запишется как

$$F^*(x) = \begin{cases} 0, & x \le x_1 \\ \dfrac{n_1}{n}, & x_1 < x \le x_2 \\ \dfrac{n_1 + n_2}{n}, & x_2 < x \le x_3 \\ \dots & \dots \quad \dots \\ 1, & x > x_r \end{cases}.$$

График эмпирической функции распределения представляет собой ступенчатую линию со скачками в точках, определяемых элементами выборки (рис.1.4).

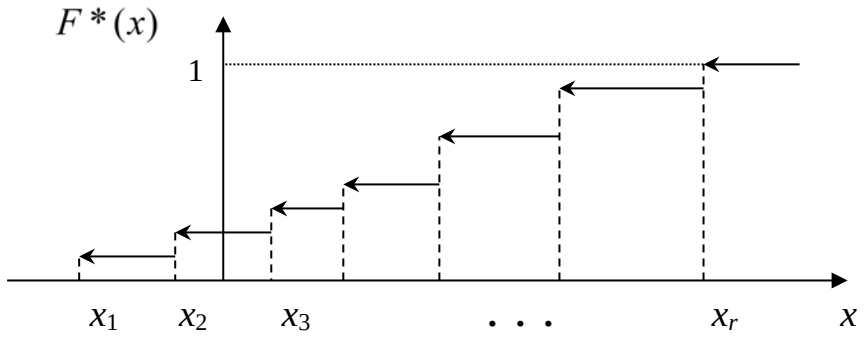


Рис. 1.4. График эмпирической функции распределения

Проиллюстрируем построение вариационных рядов, их графиков, а также эмпирической функции распределения на следующем примере.

Пример 1.1. Для изучения производительности труда X (тыс. руб.) обследовано $n = 100$ предприятий данной отрасли. Результаты наблюдений представлены ниже.

13,9	15,3	15,0	15,4	14,8	14,9	16,0	14,7	15,8	16,4
15,4	13,6	15,4	13,5	15,2	15,0	16,6	15,8	13,8	15,1
16,6	16,9	14,7	13,6	16,0	16,0	16,3	14,1	14,8	13,9
14,7	15,2	15,3	17,1	14,1	14,3	15,6	16,4	16,1	15,9
15,6	15,4	14,9	14,6	15,6	15,2	13,0	13,7	14,2	15,2
16,6	13,6	14,4	15,7	14,8	15,3	15,1	15,0	15,8	16,8
13,6	15,0	15,0	12,5	13,9	13,9	16,4	15,4	15,0	15,1
14,8	15,9	14,5	14,4	15,4	13,8	15,4	15,8	13,4	14,8
15,5	17,3	14,5	13,2	15,4	16,0	14,8	13,8	14,1	15,8
13,8	14,2	14,6	15,6	14,6	16,7	15,5	14,4	14,7	14,1

1. По данным выборки построить точечный вариационный ряд, распределив значения x_i по частотам n_i (ряд 1).

2. От ряда 1 перейти к интервальному ряду (ряд 2).

3. От ряда 2 перейти к точечному ряду, распределив значения по частотам (ряд 3) и относительным частотам в виде доли и в виде процента (ряд 4).

4. Построить:

а) гистограмму относительных частот для ряда 2;

б) полигон частот для ряда 3;

в) кумулятивную кривую для ряда 3.

5. Найти эмпирическую функцию распределения случайной величины *X*, используя ряд 3, и построить ее график.

Решение.

1. Для того чтобы построить точечный вариационный ряд, необходимо расположить наблюдаемые значения x_i в порядке их возрастания и относительно каждого x_i указать частоту n_i, т.е. количество повторений x_i в выборке; при этом сумма всех частот равна объему выборки *n*.

Ряд 1:

x_i n_i	12,5 1	13,0 1	13,2 1	13,4 1	13,5 2	13,6 3	13,7 1	13,8 4	13,9 4	14,1 4
x_i n_i	14,2 2	14,3 1	14,4 3	14,5 2	14,6 4	14,7 4	14,8 5	14,9 2	15,0 6	15,1 3
x_i n_i	15,2 4	15,3 4	15,4 7	15,5 2	15,6 4	15,7 1	15,8 5	15,9 2	16,0 4	16,1 1
x_i n_i	16,3 1	16,4 3	16,6 3	16,7 1	16,8 1	16,9 1	17,1 1	17,3 1		

Здесь объем выборки $n = \sum n_i = 100$, а число различных значений *r* =38.

2. Так как объем выборки велик и число различных значений исследуемого случайного признака также велико, то целесообразно перейти от точечного ряда 1 к интервальному. Такой переход осуществляется по изложенной выше методике следующим образом:

а) отмечаются наименьшее $x_{\min} = 12{,}5$ и наибольшее $x_{\max} = 17{,}3$ значения в выборке;

б) весь обследованный диапазон [12,5;17,3] разбивается на число интервалов k, где $k \approx 1 + \log_2 n$. В нашем примере $n = 100$ и $k = 8$;

в) определяется величина шага или ширина интервала группирования h:

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{17{,}3 - 12{,}5}{8} = \frac{4{,}8}{8} = 0{,}6;$$

г) отмечаются крайние точки каждого из интервалов C_i, C_{i+1} в порядке возрастания, а также подсчитываются числа выборочных данных, попавших в каждый из интервалов $n_1, n_2, \ldots, n_k$ здесь $n_1 + n_2 + \ldots + n_8 = 100$

Ряд 2:

$C_i - C_{i+1}$	12,5 – 13,1	13,1 – 13,7	13,7 – 14,3	14,3 – 14,9
n_i	2	8	15	20
$C_i - C_{i+1}$	14,9 – 15,5	15,5 – 16,1	16,1 – 16,7	16,7 – 17,3
n_i	26	17	8	4

3. Для того чтобы от интервального ряда 2 перейти вновь к точечному, необходимо отметить середины интервалов x_i^* и сопоставить им частоты n_i или относительные частоты w_i. Распределение производительности труда по частотам запишется в виде ряда 3, а распределение по относительным частотам в виде ряда 4:

Ряд 3:

x_i^*	12,8	13,4	14,0	14,6	15,2	15,8	16,4	17,0	$\sum n_i = 100$
n_i	2	8	15	20	26	17	8	4	

Ряд 4:

x_i^*	12,8	13,4	14,0	14,6	15,2	15,8	16,4	17,0	
w_i	0,02	0,08	0,15	0,20	0,26	0,17	0,08	0,04	$\sum w_i = 1$
$w_i \cdot 100$	2	8	15	20	26	17	8	4	$\sum w_i 100\% = 100\%$

Гистограмма относительных частот для ряда 2 изображена на рис. 1.5.

Для построения кумуляты представим ряд 3 по накопленным частотам m_i:

x_i^*	12,8	13,4	14,0	14,6	15,2	15,8	16,4	17,0
m_i	2	10	25	45	71	88	96	100

Тогда кумулятой будет плавная кривая, изображенная на рис. 1.7.

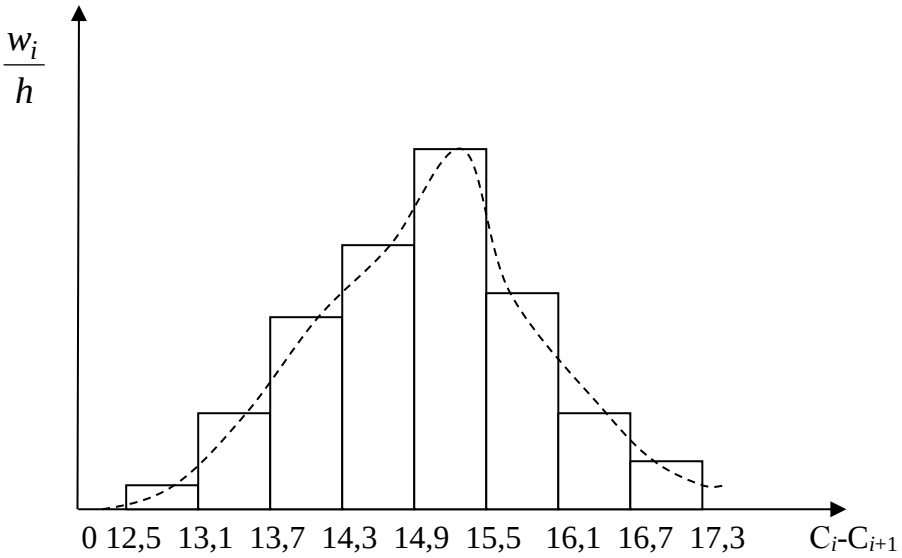


Рис. 1.5. Гистограмма относительных частот

Полигон частот показан на рис.1.6.

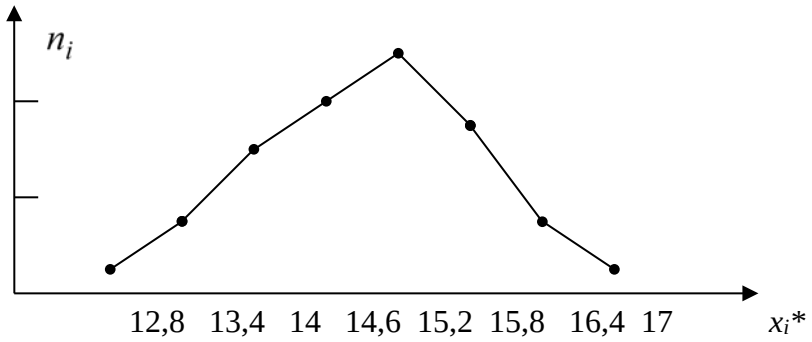


Рис. 1.6. Полигон частот

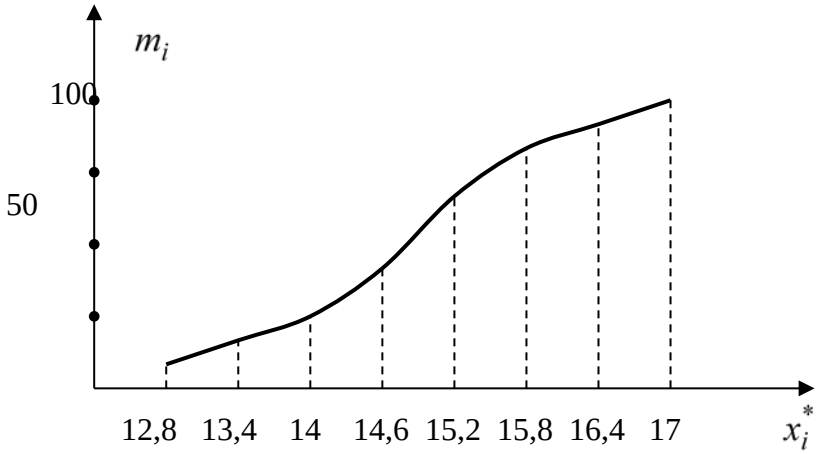


Рис. 1.7. Кумулята

5. Эмпирическая функция распределения для ряда 3 запишется так:

$$F^*(x)=\begin{cases}0, & x\le 12{,}8\\ 0{,}02, & 12{,}8<x\le 13{,}4\\ 0{,}10, & 13{,}4<x\le 14{,}0\\ 0{,}25, & 14{,}0<x\le 14{,}6\\ 0{,}45, & 14{,}6<x\le 15{,}2\\ 0{,}71, & 15{,}2<x\le 15{,}8\\ 0{,}88, & 15{,}8<x\le 16{,}4\\ 0{,}96, & 16{,}4<x\le 17{,}0\\ 1, & x>17{,}0\end{cases}$$

Здесь, например, значение функции $F^*(x)$, равное 0,02, найдено как $\frac{2}{100}$, так как значение X<13,4, а именно $x_1=12{,}8$ наблюдалось 2 раза; значения X<14,0, а именно $x_1=12{,}8$ и $x_2=13{,}4$ наблюдались 2+8 = 10 раз, следовательно, $F^*(x)$= 10/100 =0,10 при 13,4< $x\le$14,0 и т. д.

График $F^*(x)$ изображен на рис. 1.8.

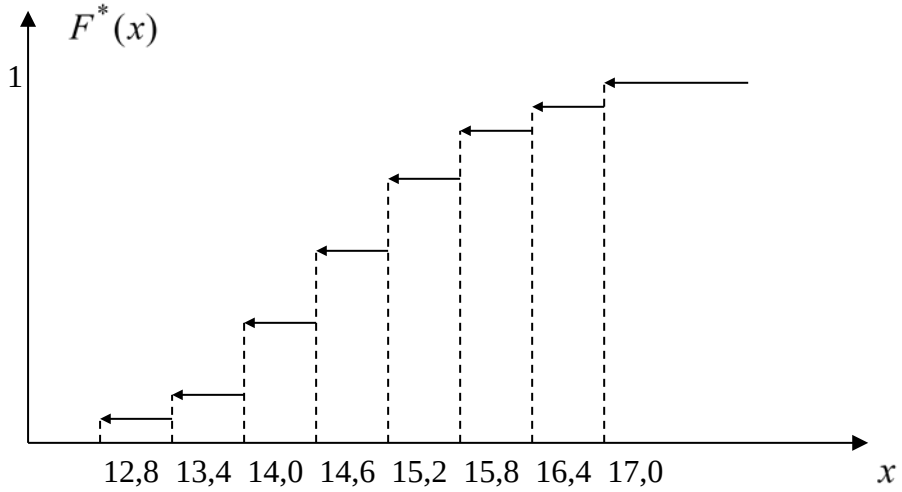


Рис. 1.8. График эмпирической функции распределения

1.3. Числовые характеристики эмпирических распределений

Исчерпывающие сведения об интересующем нас законе распределения вероятностей дают вариационные ряды, их графические представления, а также эмпирическая функция распределения. Однако нередко при практическом изучении генеральной совокупности этого бывает недостаточно, и требуется охарактеризовать имеющуюся совокупность значений некоторыми количественными показателями. К таким показателям или числовым характеристикам_выборки относятся меры положения, меры рассеяния и меры формы.

Характеристики или меры положения. Существует несколько характеристик, применяемых для описания характера расположения распределений:

среднее (арифметическое, геометрическое и гармоническое), медиана, мода, а также выборочные квантили.

Арифметическое (или выборочное) среднее $\bar{x}$ (или $\bar{x}_в$) для не сгруппированной выборки $\{x_1, x_2, \ldots, x_n\}$ объема n определяется формулой

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \,. \tag{1.5}$$

В случае выборки, представляемой рядом вида (1), выборочное среднее равно:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{r} n_i x_i \,. \tag{1.6}$$

Выборочное среднее представляет собой значение, относительно которого может быть «сбалансировано» все эмпирическое распределение (фактически, это абсцисса центра масс гистограммы). Эта характеристика является одной из наиболее употребительных статистических мер: многие средние показатели в экономике подсчитываются по формулам (1.5), (1.6).

Среднее геометрическое $\bar{x}_{геом}$ определяется как

$$\bar{x}_{геом} = \sqrt[n]{x_1 x_2 \ldots x_n} \,,$$

либо

$$\bar{x}_{геом} = \sqrt[n]{x_1^{n_1} x_2^{n_2} \ldots x_r^{n_r}} \,.$$

Геометрическое среднее находит применение при оценке темпов изменения величин, например, при расчетах индексов цен. Геометрическое среднее следует применять прежде всего тогда, когда среднее значение должно быть рассчитано из значений, заданных через некоторые равные промежутки времени; $\bar{x}_{геом}$ применяется, когда переменная меняется во времени с приблизительно постоянным соотношением между измерениями. К этому случаю относятся многообразные явления роста. Прирост населения во времени, изменение числа пациентов или эксплуатационные расходы – вот известные примеры подобного типа явлений.

Геометрическое среднее применяется также тогда, когда отдельные значения в выборке далеко отстоят от остальных значений; это меньше влияет на геометрическое среднее (чем на арифметическое среднее), так как оно дает более правильное представление о среднем.

Среднее гармоническое $\bar{x}_{гарм}$ задается соотношением

$$\bar{x}_{гарм} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}}$$

или

$$\bar{x}_{гарм} = \frac{1}{\frac{1}{n}\sum_{i=1}^{r}\frac{n_i}{x_i}}.$$

Область применения гармонического среднего весьма ограничена. В экономике, в частности, пользуются иногда гармоническим средним при анализе средних норм времени, а также в некоторых видах индексных расчетов, когда суммируемый признак выражен обратной величиной данного признака, т.е. $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}$.

Гармоническое среднее необходимо тогда, когда наблюдения, для которых мы хотим получить арифметическое среднее, заданы обратными значениями, когда эти наблюдения каким-либо образом уже содержат эту обратную зависимость.

Между тремя средними значениями существует следующее соотношение:

$$\bar{x}_{гарм} \le \bar{x}_{геом} \le \bar{x}.$$

Причем равенство справедливо при одинаковых выборочных значениях.

<u>Медиана</u> x_{med} исследуемого признака определяется как его средневероятное значение, т.е. такое значение, для которого

$$P\{X < x_{med}\} = P\{X > x_{med}\} = \frac{1}{2}.$$

При определении выборочного (приближенного) значения медианы имеющиеся в нашем распоряжении наблюдения $x_1, x_2, ..., x_n$ располагают в вариационный ряд и определяют в качестве x_{med} средний (т. е. $\frac{1}{2}(n+1)$-й) член этого ряда, если n нечетно, и любое значение между средними, т.е. $\frac{1}{2}n-$м и $\left(\frac{1}{2}n+1\right)-$м членами этого ряда, если n четно.

При исчислении медианы интервального вариационного ряда вначале находят интервал, содержащий медиану. Медианному интервалу соответствует первая из накопленных частот, превышающая половину объема выборки. Для нахождения медианы при постоянстве плотности внутри интервала, содержащего медиану, используют следующую формулу:

$$x_{med} = x_{med(\min)} + h\frac{n/2 - m_{med-1}}{n_{med}}, \qquad (1.7)$$

где $x_{med(\min)}$ – нижняя граница медианного интервала; h – интервальная разность; m_{med-1} – накопленная частота интервала, предшествующего медианному; n_{med} – частота медианного интервала.

Медиана может быть определена и графически по кумуляте. Для этого последнюю ординату, равную сумме всех частот, т.е. объему выборки n, делят пополам. Из полученной точки восстанавливают перпендикуляр до пересечения с кумулятой. Абсцисса точки пересечения и дает значение медианы (рис. 1.9).

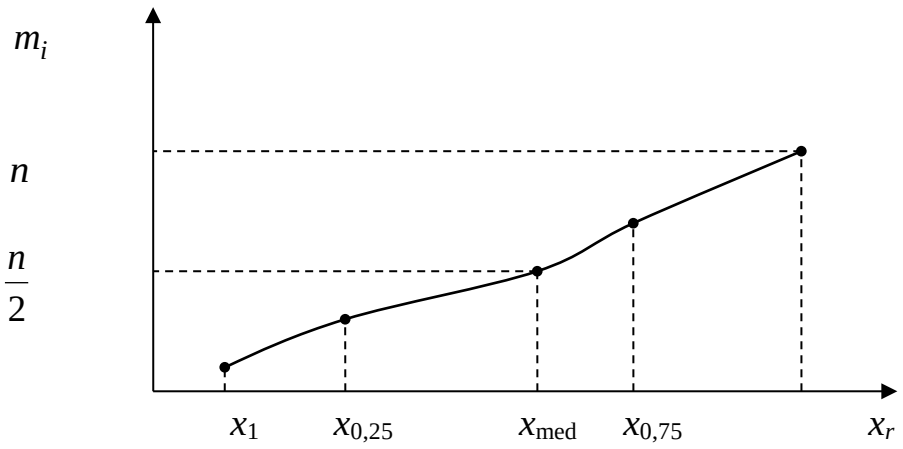


Рис. 1.9. Кумулята

Модальное значение (или просто *мода*) x_{mod} есть такое значение исследуемого признака, которое чаще всего встречается в данном вариационном ряду. Для дискретного ряда мода определяется по частотам и соответствует выборочному значению с наибольшей частотой. В случае интервального распределения с равными интервалами модальный интервал, т. е. содержащий моду, определяется по наибольшей частоте, а при неравных интервалах – по наибольшей плотности. Вычисление моды производится по формуле:

$$x_{\text{mod}} = x_{\text{mod(min)}} + h\frac{n_{\text{mod}} - n_{\text{mod}-1}}{2n_{\text{mod}} - n_{\text{mod}-1} - n_{\text{mod}+1}}, \qquad (1.8)$$

где $x_{\text{mod(min)}}$ – нижняя граница модального интервала; h – интервальная разность; n_{mod} – частота модального интервала; $n_{\text{mod}-1}$ – частота интервала, предшествующего модальному; $n_{\text{mod}+1}$ - частота интервала, следующего за модальным.

В случае симметричной плотности среднее значение $\bar{x}$, мода x_{mod} и медиана x_{med} совпадают между собой.

Выборочной квантилью порядка или уровня p называется абсцисса x_p точки, лежащей на кумулятивной кривой и имеющей ординату p (см. рис. 1.9). Порядок квантили p определяет долю общего числа наблюдений в выборке, результаты которых не превосходят x_p. Квантили порядка 0,25 и 0,75 называют соответственно нижним и верхним квартилями, медиана есть квантиль порядка 0,5, т. е. $x_{0,5} = x_{med}$. Нахождение квартилей осуществляется точно так же, как и определение медианы.

Характеристики или меры рассеяния. Средние величины, характеризующие вариационный ряд одним числом, не учитывают вариацию или разброс значений признака. Для измерения вариации применяется ряд способов.

Вариационный размах R, представляющий собой разность между наибольшим и наименьшим значениями в выборке:

$$R = x_{\max} - x_{\min},$$

применяется в качестве приблизительной оценки вариации. Особенно широко используется размах в ряде отраслей промышленности при статистическом изучении качества продукции.

Одной из наиболее часто используемых характеристик рассеяния данных является выборочное среднее квадратическое (стандартное) отклонение:

$$\sigma_B = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

дающее абсолютный разброс значений признака относительно среднего и определяемое таким образом для несгруппированных данных. Если данные сгруппированы, то

$$\sigma_B = \sqrt{\frac{1}{n}\sum_{i=1}^{r}n_i(x_i - \overline{x})^2}.$$

Квадрат этой величины σ^2 называется выборочной дисперсией и обозначается D_B:

$$D_B = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\overline{x})^2.$$

Выборочная дисперсия также может использоваться для оценки разброса значений исследуемого признака.

Для оценки относительной изменчивости признака используется коэффициент вариации

$$V = \frac{\sigma_B}{\overline{x}} \cdot 100\%,$$

который дает возможность охарактеризовать относительный разброс значений признака вокруг его среднего, выраженный в процентах.

Меры формы. Форма распределения исследуемой случайной величины характеризуется коэффициентами асимметрии и эксцесса, выборочные значения которых определяются формулами

$$A_s = \frac{\mu_3}{\sigma_B^3}, \qquad E_k = \frac{\mu_4}{\sigma_B^4} - 3,$$

где μ_3, μ_4 – центральные эмпирические моменты третьего и четвертого порядков соответственно. Для несгруппированной выборки объема n центральный момент k-го порядка равен:

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k, \quad k = 1,2,3,4,\ldots$$

Для нормального распределения коэффициенты асимметрии и эксцесса равны нулю. Поэтому, если для изучаемого распределения эти коэффициенты имеют небольшие значения, то можно предположить близость эмпирического распределения к нормальному закону. Наоборот, большие значения этих характеристик указывают на значительное отклонение от нормального распределения.

Асимметрия служит для характеристики «скошенности» распределения. Если коэффициент асимметрии положительный, то более пологая часть кривой

распределения расположена правее моды, если отрицательный – левее (рис.1. 10).

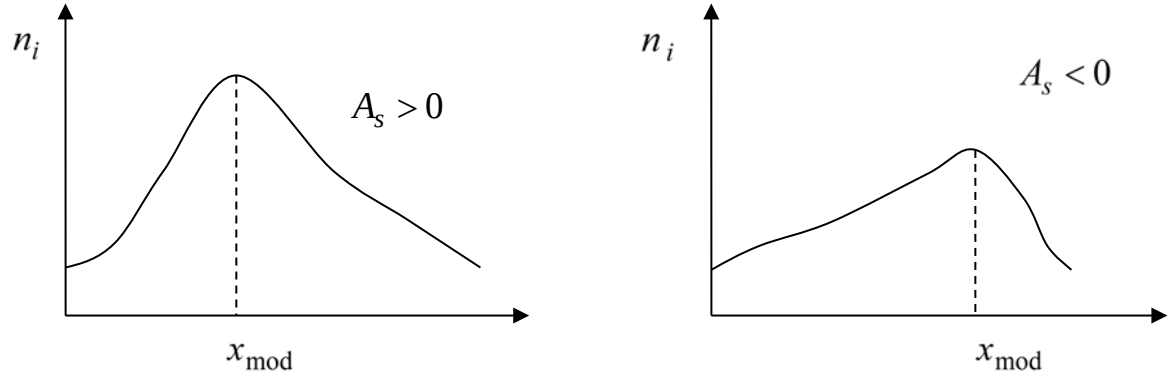


Рис. 1.10. Иллюстрация графиков, имеющих лево- и правостороннюю асимметрию

Для оценки «крутости», т. е. большего или меньшего подъема кривой эмпирического распределения по сравнению с нормальной кривой, используется коэффициент эксцесса. Если $E_k > 0$, то эмпирическая кривая имеет более высокую и «острую» вершину, чем кривая Гаусса; если $E_k < 0$, то сравниваемая кривая имеет более низкую и плоскую вершину (рис. 1.11).

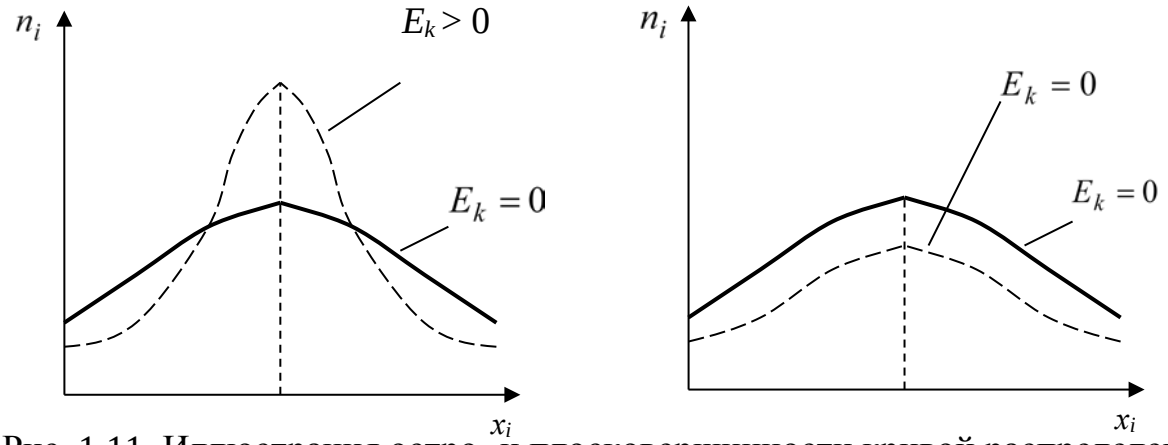


Рис. 1.11. Иллюстрация остро- и плосковершинности кривой распределения

Пример 1.2. При изучении производительности труда X тыс. руб. по данным, представленным в примере 1.1, определить выборочное среднее $\bar{x}$, выборочную дисперсию D_B, выборочное среднее квадратическое отклонение σ_B, коэффициент вариации V, моду x_{mod} и медиану x_{med} по точечному ряду 1 и интервальному ряду 2, а также коэффициенты асимметрии A_s и эксцесса E_k. Проанализировать результаты, полученные в итоге первичной статистической обработки данных, используя решения примеров 1.1 и 1.2.

Решение. Для упрощения вычислений расчет характеристик выборки произведем по ряду 3. Для удобства вычислений составим вспомогательную таблицу (табл. 1.1).

Таблица 1.1

Вспомогательная таблица для расчета характеристик выборки по сгруппированным данным

k	x_i^*	n_i	$x_i^* n_i$	$n_i(x_i^* - \bar{x})$	$n_i(x_i^* - \bar{x})^2$	$n_i(x_i^* - \bar{x})^3$	$n_i(x_i^* - \bar{x})^4$	m_i
1	12,8	2	25,6	- 4,356	9,4874	- 20,6635	45,0051	2
2	13,4	8	107,2	- 12,624	19,9207	- 31,4348	49,6041	10
3	14,0	15	210,0	- 14,670	14,3473	- 14,0316	13,7229	25
4	14,6	20	292,0	- 7,560	2,8577	- 1,0802	0,4083	45
5	15,2	26	396,2	5,772	1,2814	0,2845	0,0632	71
6	15,8	17	268,6	13,974	11,4866	9,4420	7,7613	88
7	16,4	8	131,2	11,376	16,1767	23,0032	32,7106	96
8	17,0	4	68,0	8,088	16,3539	33,0677	66,8628	100
Итог	-	100	1497,8	0,0	91,9117	- 1,4127	216,1383	-

Пользуясь данными табл. 1.1 и формулой (1.6), найдем выборочное среднее

$$\bar{x} = \frac{1}{100}\sum_{i=1}^{8} n_i x_i^* = \frac{1497{,}8}{100} = 14{,}978.$$

Для проверки правильности вычисления $\bar{x}$ полезно убедиться в выполнении условия $\sum n_i(x_i^* - \bar{x}) = 0$.

По данным табл. 1.1 найдем выборочные:

- дисперсию

$$D_B = \frac{1}{100}\sum_{i=1}^{8} n_i (x_i^* - \bar{x})^2 = 0{,}9191;$$

- среднее квадратическое отклонение σ_B=0,9587;
- коэффициент вариации $V = \frac{\sigma_B}{\bar{x}} \cdot 100\% = 6{,}4\%$;
- центральные моменты третьего и четвертого порядков:

$$\mu_3 = \frac{1}{100}\sum_{i=1}^{8} n_i (x_i^* - \bar{x})^3 = -0{,}0141;$$

$$\mu_4 = \frac{1}{100}\sum_{i=1}^{8} n_i (x_i^* - \bar{x})^4 = 2{,}1614;$$

- коэффициент асимметрии:

$$A_s = \frac{\mu_3}{\sigma_B^3} = \frac{-0{,}0141}{0{,}8811} = -0{,}0160;$$

- коэффициент эксцесса:

$$E_k = \frac{\mu_4}{\sigma_B^4} - 3 = \frac{2{,}1614}{0{,}8447} - 3 = -0{,}4412.$$

Определим моду и медиану. Мода исследуемой случайной величины X для заданного эмпирического распределения в виде ряда 1 x_{mod}=15,4, так как частота этого значения наибольшая и равна 7. В случае интервального ряда 2 модальному интервалу соответствует наибольшая частота n_{mod}, равная 26. Следовательно,

$x_{\text{mod(min)}}$=14,9; h =0,6; n_{mod}=26, $n_{\text{mod}-1}$=20, $n_{\text{mod}+1}$=17.

$$x_{\text{mod}} = 14{,}9 + 0{,}9\frac{26-20}{2\cdot 26-20-17} = 15{,}14.$$

Медиану определим как средний член ряда по точеному распределению выборки. В нашем примере n =100, поэтому в качестве медианы берем любое значение между 50-м и 51-м членами ряда 1. Здесь x_{med} =15,0.

Медианному интервалу заданного эмпирического распределения в виде ряда 2 соответствует накопленная частота 71, отсюда $x_{med(\min)}$=14,9; h=0,6; m_{med-1}= 45; n_{med} =26. Используя формулу (7), получим

$$x_{med} = 14{,}9 + 0{,}6\frac{50-45}{26} = 15{,}0154 \approx 15{,}02.$$

Определим медиану графически по кумуляте, представленной на рис. 1.7. Для этого последнюю ординату, равную объему выборки n=100, поделим пополам. Восстановим перпендикуляр до пересечения с кумулятой. Абсцисса точки пересечения $x_{med} \approx 15$ и будет медианой.

Таким образом, средняя производительность труда изученной группы предприятий составила $\bar{x} = 14{,}978$ (тыс. руб.), абсолютный разброс значений показателя X равен σ=0,9587 (тыс. руб.), относительный разброс $V = 6{,}4\%$. Наибольшее число предприятий имеют производительность труда, равную 15,14 (тыс. руб.), а половина – более 15,02 (тыс. руб.)

Построенные вариационный ряды 1-3, их графические изображения (рис. 1.5-1.8) представляют данные в компактном виде. Кроме этого имеется возможность получить сведения о законе распределения вероятностей исследуемой случайной величины. Здесь внешний контур гистограммы (рис. 1.5), графики кумулятивной кривой (рис. 1.7) и эмпирической функции распределения (рис. 1.8) свидетельствуют о близости эмпирического распределения к нормальному закону. К этому же выводу можно прийти, сравнивая значения выборочного среднего, моды, медианы. Так как $\bar{x}$, x_{mod} и x_{med} незначительно отличаются друг от друга ($\bar{x} \approx x_{\text{mod}} \approx x_{med} \approx 15{,}00$), есть основание предполагать, что теоретическое распределение симметрично относительно своего среднего значения, что является еще одним доводом в пользу выбора модели нормального закона. И, наконец, близость значений выборочных коэффициентов асимметрии A_s и эксцесса E_k к нулю также свидетельствует в пользу выбора нормального закона распределения для анализируемой случайной величины.

Следовательно, в результате первичной статистической обработки данных мы получили возможность определить некоторые средние показатели интересующего нас признака, а также считать, что случайная величина X – производительность труда – распределена по нормальному закону. Нахождение приближенных значений параметров этого закона (оценок), и достоверное подтверждение такой гипотезы составляет содержание следующих задач и приемов математической статистики.

1.4. Рекомендации по выполнению расчетно-графической работы по теме «Описательная статистика» в MS Excel

Задание по выборке (объемом $\geq$ 80) (может быть выдана преподавателем или смоделирована студентом самостоятельно):

1. Дать экономическую интерпретацию исходным данным.

2. Построить точечный вариационный ряд, распределив значения по частотам (ряд 1).

3. От точечного ряда перейти к интервальному, взяв число интервалов k (ряд 2).

4. От интервального ряда перейти к точечному сгруппированному ряду (ряд 3), распределив значения: а) по частотам и относительным частотам в виде доли или процента (ряд 4), б) по накопленным частотам (ряд 5).

5. Построить полигон частот для ряда 3 или 4, гистограмму для ряда 2, кумуляту для ряда 5.

6. Построить эмпирическую функцию распределения по ряду 4.

7. Определить числовые характеристики: выборочное среднее, моду, медиану (по точечному, интервальному рядам и графику), выборочную дисперсию, среднеквадратическое отклонение, коэффициент вариации, асимметрию, эксцесс.

8. Сделать вывод о близости к нормальному закону.

Пример 1.3. Изучается случайная величина X – количество пассажиров одного авиарейса «Иркутск-Москва» или «Москва-Иркутск», максимальная вместимость самолета типа «Аэробус А320» 140 человек (табл. 1.2).

Таблица 1.2

117	107	109	104	108	88	109	97	112	115
123	104	103	111	108	108	126	110	109	128
104	123	110	109	126	114	114	108	117	109
103	99	112	100	129	109	94	107	120	106
103	96	113	108	116	104	107	113	107	112
85	109	112	131	95	94	87	122	134	106
100	115	117	101	118	117	108	95	120	118
106	103	119	113	116	131	113	99	115	98
116	108	113	114	111	107	97	131	126	120
95	99	115	111	110	112	91	107	101	100

Построение точечного вариационного ряда 1

Заданную выборку данных набираем (копируем) в MS Excel (рис. 1.12).

Выстраиваем все данные в один столбец А. Затем выделяем этот столбец и сортируем данные по возрастанию значений. В MS Excel это можно сделать с помощью кнопки  , расположенной на вкладке Главная MS Excel, далее вы-

бираем пункт «Сортировка по возрастанию». Если после сортировки столбца А первое значение (в ячейке А1) не является минимальным, то снова выбираем сортировку по возрастанию и убираем галочку в Мои данные содержат заголовки.

В примере 1 получается минимальное значение $x_{\min} = 85$ чел. Оно будет расположено в ячейке А1. Максимальное – $x_{\max} = 134$ чел. Оно будет расположено в ячейке А100. Между ними – все значения выборки, упорядоченные по возрастанию.

Чтобы быстро посчитать повторяющиеся значения, используем формулу =СЧЁТЕСЛИ($A$1:$A$100;A1) в ячейке В1 и затем растягиваем эту ячейку до конца столбца (рис. 1.13). В столбце В появятся значения повторений каждого числа в выборке (рис. 1.13).

	A	B	C	D	E	F	G	H	I	J
1	117	107	109	104	108	88	109	97	112	115
2	123	104	103	111	108	108	126	110	109	128
3	104	123	110	109	126	114	114	108	117	109
4	103	99	112	100	129	109	94	107	120	106
5	103	96	113	108	116	104	107	113	107	112
6	85	109	112	131	95	94	87	122	134	106
7	100	115	117	101	118	117	108	95	120	118
8	106	103	119	113	116	131	113	99	115	98
9	116	108	113	114	111	107	97	131	126	120
10	95	99	115	111	110	112	91	107	101	100
11										

Рис. 1.12. Исходные данные, скопированные в MS Excel из MS Word

Формируем ряд 1 (см. табл. 1.1) на этом же листе. Первая строка (уникальные значения из столбца А) x_i – значения выборки, начиная с $x_1 = x_{\min}$ и заканчивая $x_r = x_{\max}$, где r – число различных значений в выборке, $0 \le r \le n$. Вторая строка n_i – частота значения x_i (количество повторений одного и того же значения x_i, взятые из столбца В). Перенос уникальных значений осуществляем через

1	85	=СЧЁТЕСЛИ($A$1:$A$100;A1)
2	87	1
3	88	1
4	91	1
5	94	2
6	94	2
7	95	3
8	95	3
9	95	3
10	97	1

Рис. 1.13. Использование функции =СЧЁТЕСЛИ($A$1:$A$100;A1) для вычисления частот р.1

специальную вставку в ячейку D2, отмечая пункты «значения» и «транспонировать» (шаг 2 на рис. 1.14): выделяем уникальные значения (столбец A и B одновременно), не забываем, что можно сразу выделить несколько разрывных массивов, зажимая клавишу Ctrl (шаг 1 на рис. 1.14):

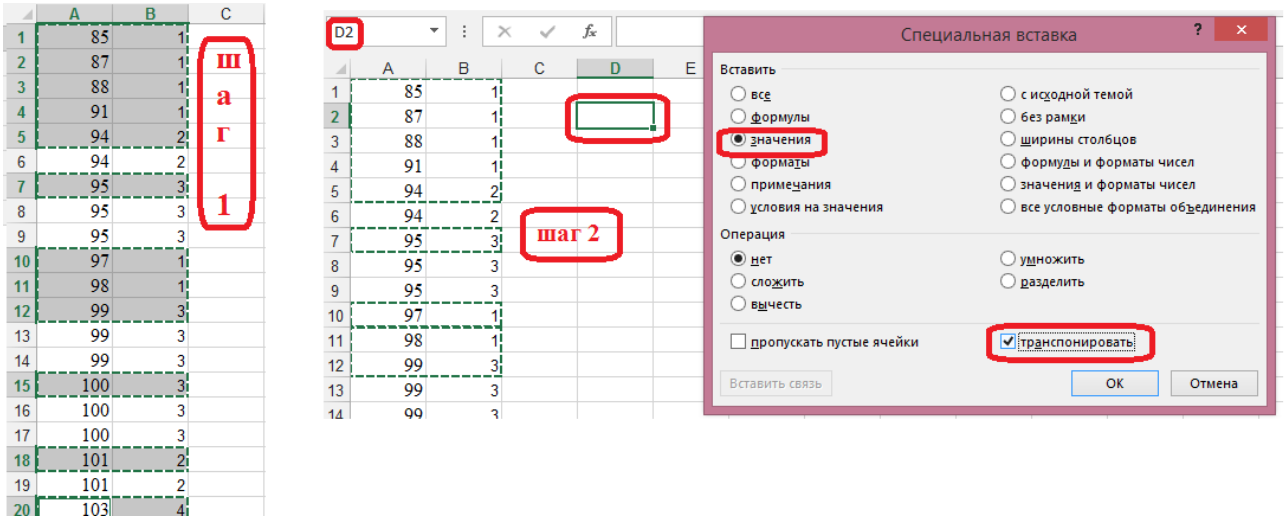


Рис. 1.14. Перенос значений в ряд 1

Оформляем полученный ряд, как на рис. 1.15.

Для проверки в конце каждой строки n_i можно посчитать сумму частот, выбрав кнопку «автосумма» Σ на панели управления (вкладка Главная). В конце, подсчитав итоговую сумму, сложив все промежуточные суммы, должно получиться значение объема выборки n. В примере 1 получилась сумма, равная 100 (рис. 1.15). По крайней мере, это означает, что при построении ряда 1 никакие значения исходных данных не были потеряны.

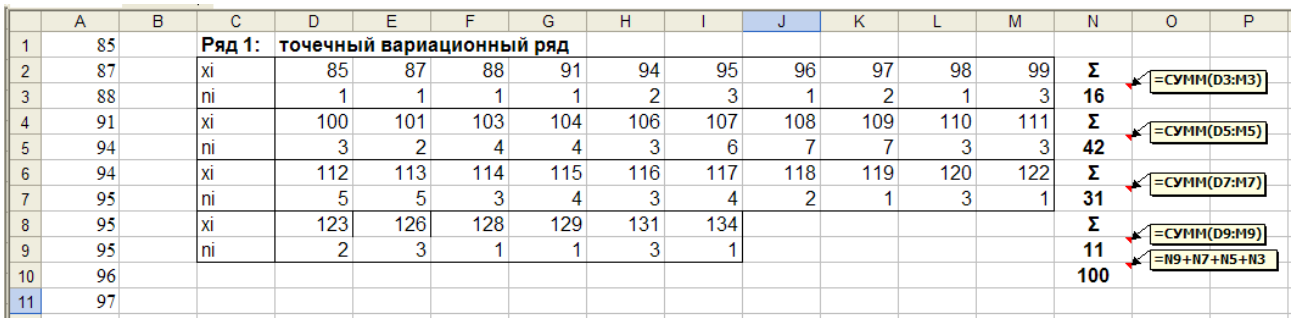


Рис. 1.15. Построение ряда 1 в MS Excel

Построение интервального вариационного ряда 2

Для построения интервального вариационного ряда 2 (табл. 1.2) сделаем вспомогательные расчеты. Для удобства можно в одном столбце набрать необходимые для вычислений показатели n, $x_{\min}$, $x_{\max}$, R – размах вариации (1.1), k – количество интервалов, k округляем сразу до целого значения: =ОКРУГЛ(LOG(n;2)+1;0), h– шаг интервала, вычисленный, C_0, C_1, …, C_k – границы интервалов. В соседнем столбце проводим расчеты перечисленных показателей с помощью формул (рис. 1.16). Важно помнить, что в MS Excel все вычисления надо начинать со знака равенства «=». Для нахождения k используем встроенную функцию = LOG(n;2), где в скобках ставим ссылку на ячейку, в которой записано значение объема выборки n, затем через знак «;» пишем 2 – это основание логарифма. Так как k должно быть целым числом, то сразу применяем функцию =ОКРУГЛ(LOG(n;2)+1;0), где 0 – это число разрядов. При расчете границ интервалов в C_0 ссылаемся на $x_{\min}$, в C_1 пишем формулу (рис. 1.16), не забывая ставить значок $ (значок $ можно поставить 1) вручную, нажав Shift 4 в латинской раскладке клавиатуры, 2) автоматически, нажав клавишу F4) для шага интервала h, чтобы при последующем копировании для автоматического вычисления C_2,…, C_k, это значение было зафиксировано и не менялось.

После того, как границы интервалов найдены, формируем интервальный

	A	B	C	D	E	F
10	97		n=	100		
11	98		x min =	85		
12	99		x max =	134		
13	99		R=	49		
14	99					
15	100		k=	8,00	=ОКРУГЛ(LOG(D10;2)+1;0)	
16	100		h=	6,125	=D13/D15	
17	100		Co=	85	=D11	
18	101		C1=	91,125	=D17+$D$16	
19	101		C2=	97,25		
20	103		C3=	103,375		
21	103		C4=	109,5		
22	103		C5=	115,625		
23	103		C6=	121,75		
24	104		C7=	127,875		
25	104		C8=	134		

Рис. 1.16. Расчеты для построения ряда 2 примера 1.3

ряд 2, используя для этого функцию =СЦЕПИТЬ(). Ссылками указываем границы интервалов, используя функцию =ОКРУГЛ(ссылка на ячейку;2), 2 – число разрядов, исходя из исходных данных, знак «;» между значениями границ интервалов вводим следующим образом (с учетом кавычек и пробела): “; ” (рис. 1.17). Например, для получения первого интервала в ячейке G14 запишем: =СЦЕПИТЬ(ОКРУГЛ(D17;2);"; ";ОКРУГЛ(D18;2)), затем растянем ячейку G14 на k интервалов вниз (рис. 1.17). Полученный столбец для удобства последующего переноса расчетов в MS Word скопируем и вставим в свободное место, используя

специальную вставку: значения и транспонировать (рис. 1.18). Следующим шагом нам нужно посчитать, сколько значений выборки вошло в каждый из интервалов. Для этого в MS Excel необходимо настроить пакет анализа: MS Excel-2013 (и более новые версии) меню Файл (зеленый прямоугольник в левом верхнем углу, затем выбираем Параметры, потом Надстройки, потом в середине внизу будет кнопка Перейти, нажимаем ее, ставим галочку напротив «Пакет анализа», нажимаем ОК, переходим на вкладке Данные в меню Анализ данных (расположен в правом верхнем углу)). Открываем Анализ данных и выбираем пункт Гистограмма: входной интервал: все исходные данные столбца А. Интервал карманов: значения границ интервалов с С1 (а не С0) до Ck. Выходной интервал: свободное пространство на листе (рис. 1.19). Итог имеет вид как на рис. 1.19. Значения столбца Частота через меню «Специальная вставка» – «транспонировать» переносим во вторую строку ряда 2 n_i (рис. 1.20), находим сумму частот, чтобы убедиться в том, что расчет частот для каждого интервала выполнен без потерь данных.

	A	B	C	D	E	F	G	H	I	J	K
12	99		x max =	134			Интервальный ряд 2				
13	99		R=	49			Ci-1; Ci	=СЦЕПИТЬ(ОКРУГЛ(D17;2);"; ";ОКРУГЛ(D18;2))			
14	99						85; 91,13				
15	100		k=	8,00			91,13; 97,25	=СЦЕПИТЬ(ОКРУГЛ(D18;2);"; ";ОКРУГЛ(D19;2))			
16	100		h=	6,125			97,25; 103,38				
17	100		Co=	85			103,38; 109,5				
18	101		C1=	91,125			109,5; 115,63				
19	101		C2=	97,25			115,63; 121,75				
20	103		C3=	103,375			121,75; 127,88				
21	103		C4=	109,5			127,88; 134				
22	103		C5=	115,625							
23	103		C6=	121,75							
24	104		C7=	127,875							
25	104		C8=	134							

Рис. 1.17. Применение функции =СЦЕПИТЬ()

	G	H	I	J	K	L	M	N	O	P	Q
12	Интервальный ряд 2										
13	Ci-1; Ci		Ci-1; Ci	85; 91,13	91,13; 97,25	97,25; 103,38	103,38; 109,5	109,5; 115,63	115,63; 121,75	121,75; 127,88	127,88 ; 134
14	85; 91,13										
15	91,13; 97,25										
16	97,25; 103,38										
17	103,38; 109,5										
18	109,5; 115,63										
19	115,63; 121,75										
20	121,75; 127,88										
21	127,88; 134										

Рис. 1.18. Формирование интервального ряда 2

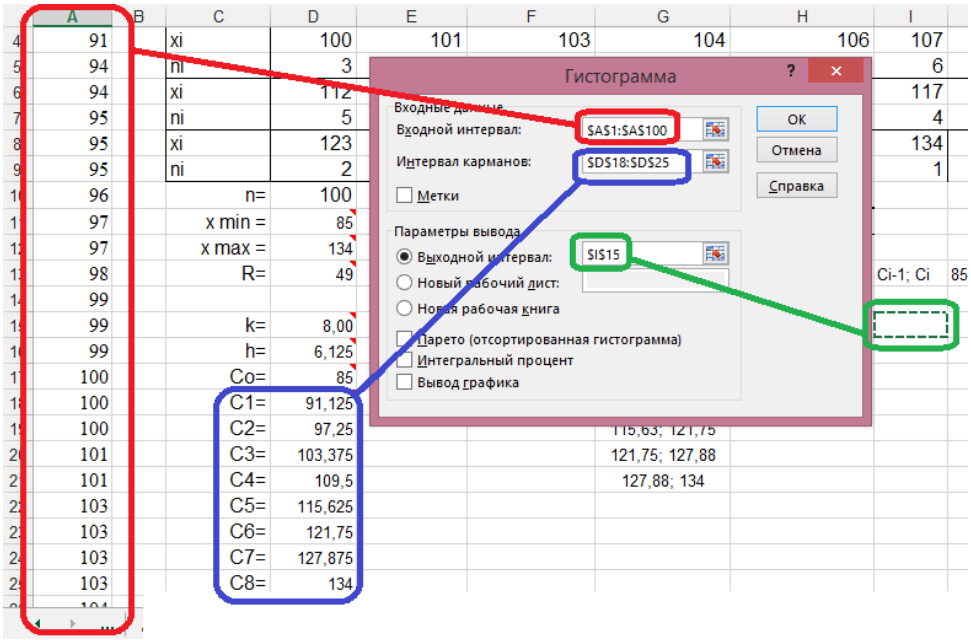


Рис. 1.19. Использование функции гистограмма
для вычисления частот интервального ряда 2

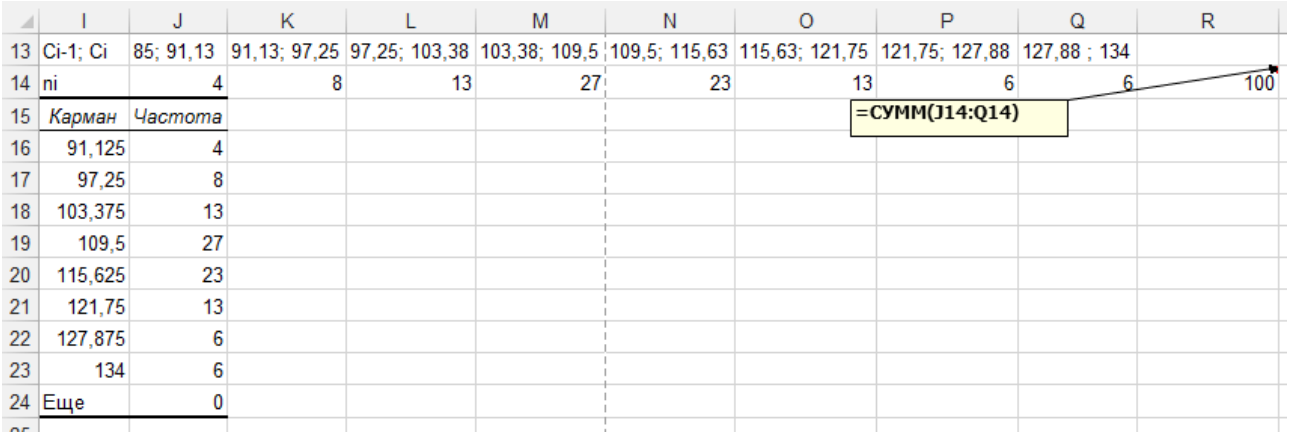


Рис. 1.20. Перенос частот в ряд 2

Построение точечных сгруппированных рядов 3–5

Для построения рядов 3–5, вычисления значений эмпирической функции распределения (1.8) создаем вспомогательную таблицу в MS Excel (рис. 1.21). Во всех ячейках вспомогательной таблицы на рис. 1.21 записываем ссылки или формулы (см. примечания к ячейкам). В столбцах E, G, I, J, K в первой строке таблицы (в MS Excel это строка 27) делаем вычисления по формулам (1.5), (1.6), (1.8), (1.9), а затем их копируем в остальные строки. При заполнении столбцов C_{i-1}, C_i в первой строке ставим ссылки на значения ячеек C_0 и C_1, рассчитанные ранее (рис. 1.16). Во второй строке столбца C_{i-1} ставим ссылку на верхнюю гра-

ницу предыдущего интервала, а в столбце C_i пишем формулу вычисления границ интервала, не забывая фиксировать шаг интервала (рис. 1.21). Затем снова копируем в оставшиеся строки. В столбец n_i ссылки ставим вручную из интервального ряда 2 (рис. 1.20, столбец частота). При расчете накопленных частот в столбце m_i в первой строке ставим ссылку на частоту n_1, во второй записываем значение m_2 по формулам (1.7), в оставшиеся ячейки столбца m_i копируем m_2 (рис. 1.21).

	A	B	C	D	E	F	G	H	I	J	K
22	103		C5=	115,625					127,875	6	
23	103		C6=	121,75					134	6	
24	103		C7=	127,875							
25			C8=	134							F*(x)
26	104		C i-1	C i	x*i	n i	x* i * n i	m i	n i / (n*h)	w i	0
27	104		85	91,125	88,063	4	352,25	4	0,0065	0,04	0,04
28	104		91,125	97,25	94,188	8	753,5	12	0,0131	0,08	0,12
29	104		97,25	103,375	100,313	13	1304,1	25	0,0212	0,13	0,25
30	106		103,375	109,5	106,438	27	2873,8	**52**	0,0441	0,27	0,52
31	106		109,5	115,625	112,563	23	2588,9	75	0,0376	0,23	0,75
32	106		115,625	121,75	118,688	13	1542,9	88	0,0212	0,13	0,88
33	107		121,75	127,875	124,813	6	748,9	94	0,0098	0,06	0,94
34	107		127,875	134	130,938	6	785,6	100	0,0098	0,06	1
35	107					**100**	**10950**			**1**	

Рис. 1.21. Вспомогательные расчеты для построения рядов 3–5 и $F^*(x)$

Ряды 3 – 5 оформляем по табл. 1.3 – 1.5 (рис. 1.22). Для этого копируем соответствующие данные из вспомогательной таблицы и вставляем их в новую таблицу с использованием «специальной вставки» контекстного меню, в которой надо отметить пункты «значения» и «транспонировать».

	X	Y	Z	AA	AB	AC	AD	AE	AF
1	**Ряд 3:**	**точечный сгруппированный ряд (по частотам)**							
2	x*i	88,063	94,188	100,313	106,438	112,563	118,688	124,813	130,938
3	ni	4	8	13	27	23	13	6	6
4									
5	**Ряд 4:**	**точечный сгруппированный ряд (по относительным частотам)**							
6	x*i	88,063	94,188	100,313	106,438	112,563	118,688	124,813	130,938
7	w i	0,04	0,08	0,13	0,27	0,23	0,13	0,06	0,06
8	w i*100%	4%	8%	13%	27%	23%	13%	6%	6%
9									
10	**Ряд 5:**	**точечный сгруппированный ряд (по накопленным частотам)**							
11	x*i	88,063	94,188	100,313	106,438	112,563	118,688	124,813	130,938
12	m i	4	12	25	52	75	88	94	100

Рис. 1.22. Ряды 3–5

Построение графиков

Для построения гистограммы выделяем столбец $n_i/(nh)$ вспомогательной таблицы (рис. 1.23), затем на панели инструментов выбираем вкладку Вставка, нажимаем кнопку «Гистограмма» и выбираем подходящий вид гистограммы (рис. 1.23).

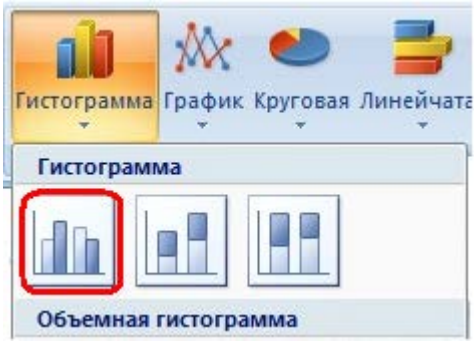


Рис. 1.23. Выбор подходящего вида гистограммы в MS Excel

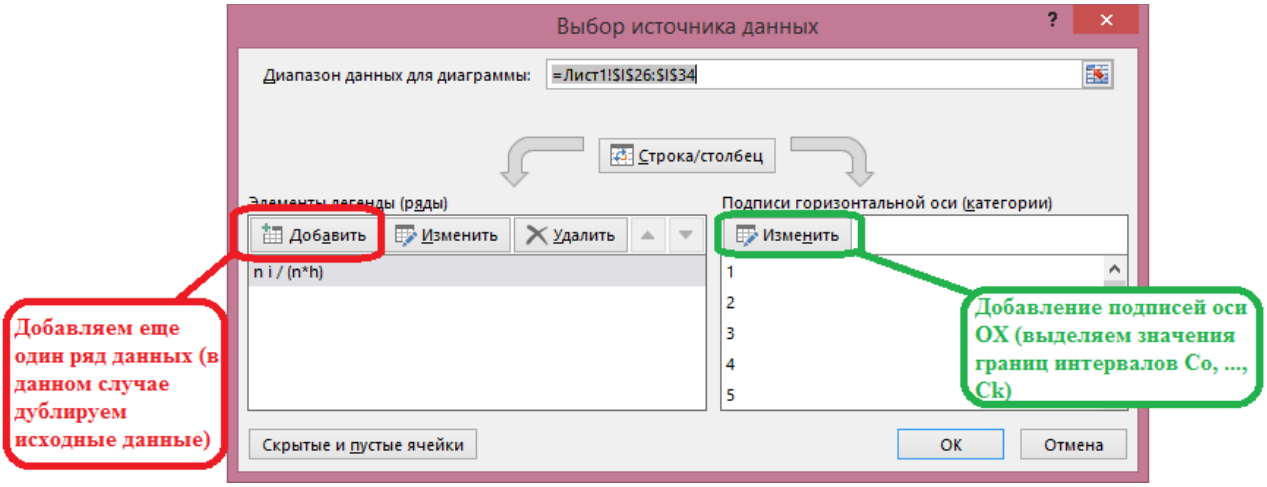


Рис. 1.24. Добавление еще одного ряда данных на график и подписей оси OX

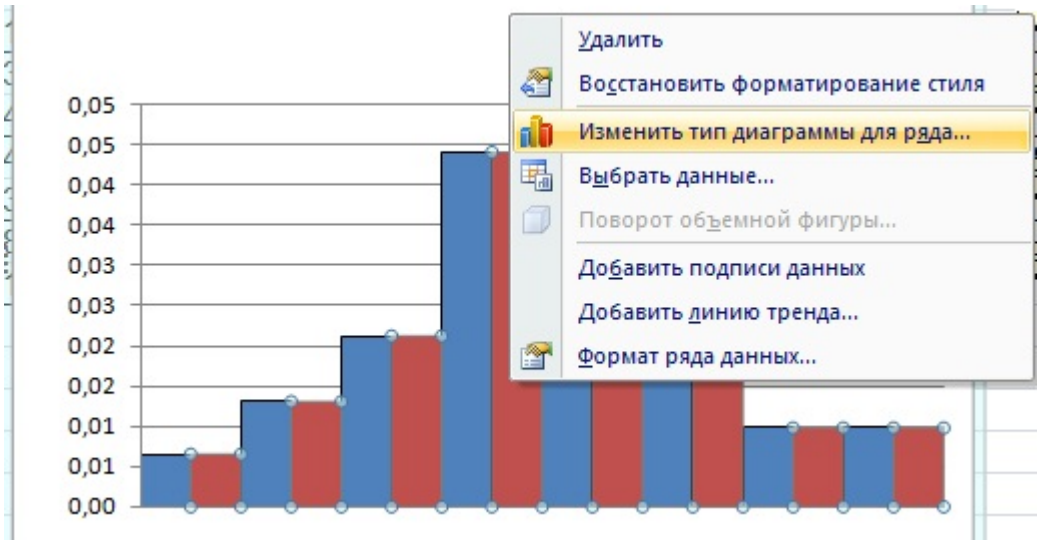


Рис. 1.25. Добавлением на гистограмму ломаной линий

После этого на самом рабочем листе появится гистограмма, которую нужно будет отредактировать: 1) нужно «склеить» столбцы – в контекстном меню выбираем «формат ряда данных», вкладку «Параметры ряда» и устанавливаем ширину зазора, равную 0 (рис. 1.24); 2) чтобы на гистограмму добавить ломаную линию, соединяющую середины прямоугольников, выполняем следующие действия: в контекстном меню выбираем пункт «Выбрать данные», затем

добавляем на график еще одну гистограмму (), и ее переделываем в ломаную линию (рис. 1.25); 3) подписи оси OX делаем искусственно, так как нужным образом, как на рис. 1.1, в MS Excel их не сделаешь. Самым простым способом это можно сделать, добавив подписи оси OX, затем скопировав гистограмму в Paint, и переместить п подписи оси OX влево. (рис. 1.26).

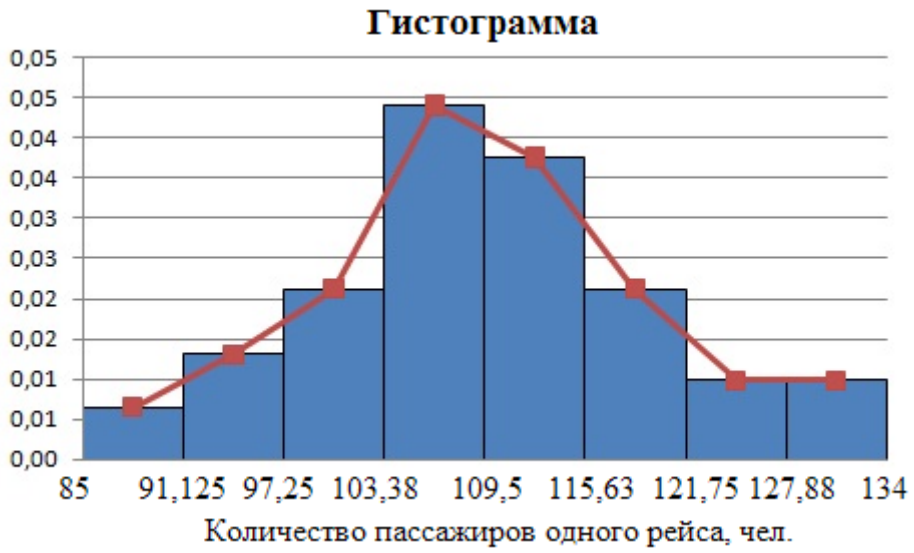


Рис. 1.26. Гистограмма, отредактированная с помощью Paint

Для построения полигона частот (относительных частот) выделяем столбец n_i, затем на вкладке Вставка выбираем кнопку «График» и вид графика (рис. 1.27).

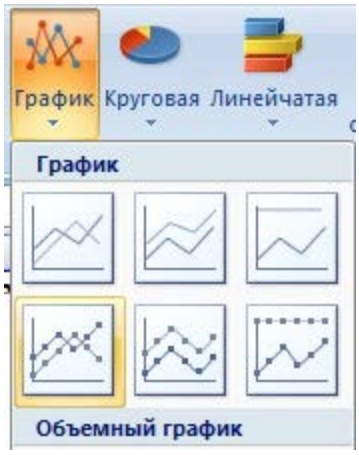


Рис. 1.27. Выбор подходящего вида графика для полигона частот

Затем на построенном графике делаем название диаграммы, подписи оси OX, (рис. 1.28).

Рис. 1.28. Полигон частот

Эмпирическая функция распределения

Значения эмпирической функции распределения $F^*(x)$ уже рассчитаны во вспомогательной таблице (рис. 1.21). Для построения графика функции $F^*(x)$ в MS Excel нет специальной встроенной команды, поэтому график, подобный рис. 1.4, можно построить следующим образом. Выделяем столбец $F^*(x)$ и строим по нему гистограмму. Она получится, как лесенка с поднимающимися вверх ступеньками. Затем «склеиваем» столбики, делаем их фон белого цвета, убираем сетку, закрасив ее линии в белый цвет. Как и при построении гистограммы, делаем подписи оси X (можно их просто оттуда скопировать). Предварительно график $F^*(x)$ будет выглядеть как на рис. 1.29. После этого копируем полученный график функции в Paint. С помощью ластика аккуратно убираем вертикальные линии, потом к оставшимся горизонтальным прямым подрисовываем стрелки с левой стороны. График эмпирической функции распределения $F^*(x)$ готов (рис. 1.30).

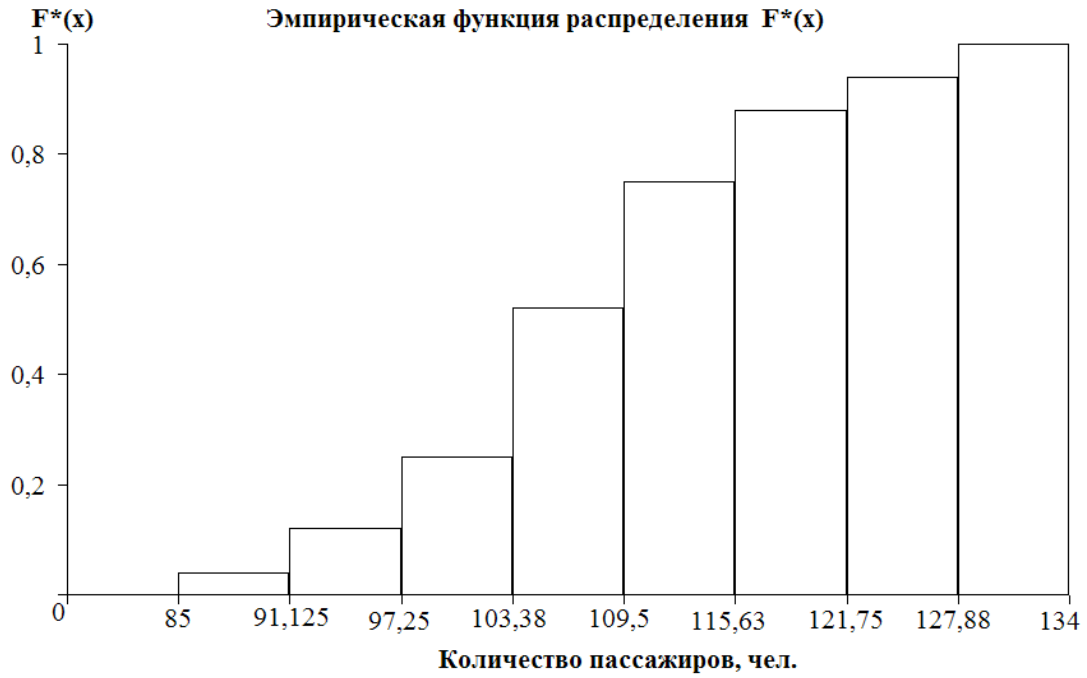
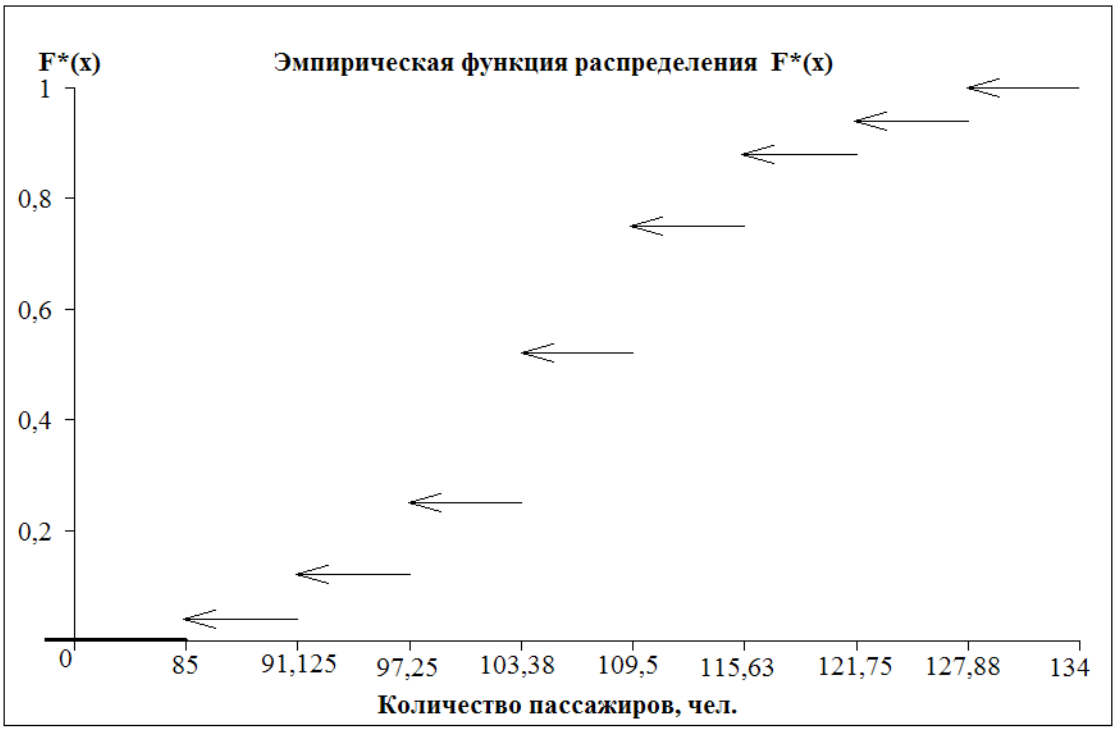


Рис. 1.29. Предварительная подготовка графика $F^*(x)$ в MS Excel



1.30. Эмпирическая функция распределения

Вычисление числовых характеристик

Предварительные расчеты среднего выборочного значения уже сделаны (рис. 1.21), поэтому остается $\overline{x}_в$ вычислить по формуле (1.6). Для расчета моды определяем модальный интервал по наибольшей частоте (на рис. 1.21 он выделен жирным шрифтом), внизу таблицы записываем все компоненты формулы

(1.8) и находим x_{mod}. Аналогично отмечаем медианный интервал по накопленной частоте, превышающей половину объема выборки n. В примере 1 это $m_4 = 52$. Все расчеты x_{mod}, x_{med} и $\bar{x}_в$(x ср) показаны на рис. 1.31.

	B	C	D	E	F	G	H	I	J	K
36		Расчет моды			Расчет медианы			x ср =	109,5	=G35/F35
37		x mod(min)=	103,38	x med(min)=		103,38				
38		n mod =	27	n med=		27				
39		n mod-1=	13	m med-1=		25				
40		n mod+1 =	23	n/2=		50				
41		h=	6,125							
42		x mod=	108,139	x med=		109,046				
43		=D37+D41*(D38-D39)/(2*D38-D39-D40)				=G37+D41*(G40-G39)/G38				
44										

1.31. Вычисления мер положения

Для вычисления мер разброса и мер формы в MS Excel создаем новую вспомогательную таблицу (табл. 1.6., рис. 1.32). В столбце «x*i–xср» ссылку на значение $\bar{x}_в$(x ср) обязательно фиксируем (кнопка F4). В следующих четырех столбцах значок возведения в степень «^» ставится как shift 6 в латинской раскладке клавиатуры.

Обязательно рассчитывается столбец «проверка»: $\sum_{i=1}^{k} n_i (x_i * - \bar{x}_в)^2 = 0$.

	C	D	E	F	G	H	I	J
45				=E48*D48				
46			=C48-$J$36	Проверка:	=D48*E48^2	=D48*C48^2	=D48*E48^3	=D48*E48^4
47	x*i	n i	x*i - x cp	ni*(x*i - x cp)	ni*(x*i - x cp)^2	ni*(x*i)^2	ni*(x*i - x cp)^3	ni*(x*i - x cp)^4
48	88,0625	4	-21,438	-85,750	1838,27	31020,02	-39407,8	844805,1
49	94,1875	8	-15,313	-122,500	1875,78	70970,28	-28722,9	439819,4
50	100,3125	13	-9,188	-119,438	1097,33	130813,77	-10081,7	92626,0
51	106,4375	27	-3,063	-82,688	253,23	305881,42	-775,5	2375,0
52	112,5625	23	3,063	70,438	215,71	291417,28	660,6	2023,2
53	118,6875	13	9,188	119,438	1097,33	183127,39	10081,7	92626,0
54	124,8125	6	15,313	91,875	1406,84	93468,96	21542,2	329864,6
55	130,9375	6	21,438	128,625	2757,40	102867,77	59111,7	1267207,7
56	**Σ**	**100**	**-**	**0,000**	**10541,9**	**1209566,9**	**12408,3**	**3071346,9**
57								
58	**Dв1=**	**105,42**	=G56/D56	**Dв2=**		**105,42**	=H56/D56-J36^2	
59	**σв=**	**10,27**	=D58^0,5					
60	**V=**	**9,38**	=D59/J36*100					
61	μ3=	124,083	=H56/D56					
62	**As=**	**0,115**	=D61/D59^3					
63	μ4=	30713,5	=I56/D56					
64	**Ek=**	**-0,236**	=D63/D59^4-3					

Рис. 1.32. Расчеты мер разброса и формы

Для вычисления среднеквадратического отклонения можно значение дисперсии возвести в степень 0,5 или использовать встроенную в MS Excel функцию = КОРЕНЬ(), в скобках ставится ссылка на соответствующую ячейку.

1.5. Оформление полученных результатов в MS WORD

После проведения всех расчетов темы 1 оформляем полученные результаты в MS Word (если нет такой возможности, то в письменном виде на листах А4 или в тетради). Правила оформления расчетно-графической работы аналогичны правилам оформления курсовой работы. Начинаем с титульного листа, на котором указываем название работы, номер варианта (если таковой был), ФИО исполнителя и номер группы, ФИО преподавателя, проверяющего эту работу. На следующей странице описываем все сделанные в MS Excel расчеты, сопровождая их соответствующими выводами. Все формулы, встречающиеся при выполнении работы, необходимо набирать с помощью встроенного в MS Word редактора формул Microsoft Equation 3.0 (рис. 1.33) или Math Type (Кнопка π Уравнение на вкладке Вставка). Рассмотрим все сказанное выше на данных примера 1.

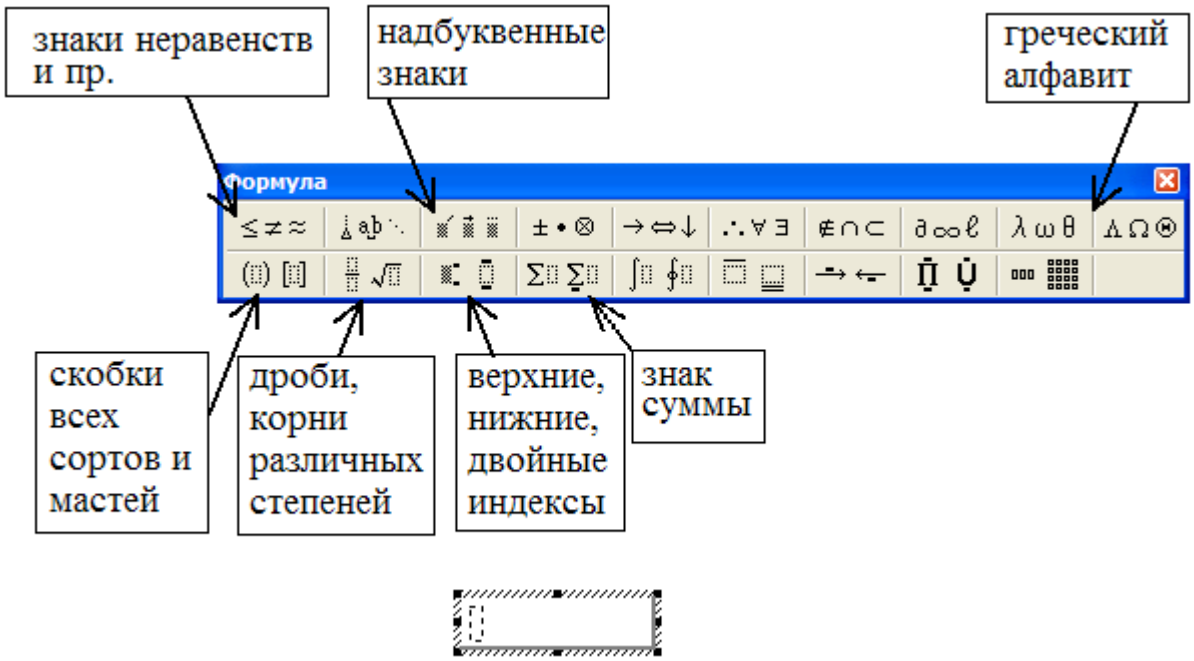


Рис. 1.33. Редактор формул Microsoft Equation 3.0

Пример 1.4. Оформление результатов проведенных расчетов по теме «Описательная статистика». В расчетно-графической работе анализируется случайная величина X – количество пассажиров одного авиарейса «Иркутск-Москва» или «Москва-Иркутск». Проведено 100 наблюдений, результаты которых представлены в таблице:

Результаты наблюдений: количество пассажиров одного авиарейса, чел.

117	107	109	104	108	88	109	97	112	115
123	104	103	111	108	108	126	110	109	128
104	123	110	109	126	114	114	108	117	109
103	99	112	100	129	109	94	107	120	106
103	96	113	108	116	104	107	113	107	112

85	109	112	131	95	94	87	122	134	106
100	115	117	101	118	117	108	95	120	118
106	103	119	113	116	131	113	99	115	98
116	108	113	114	111	107	97	131	126	120
95	99	115	111	110	112	91	107	101	100

2. Для построения точечного вариационного ряда 1, расположим значения x_i по возрастанию и отметим частоту n_i, соответствующую каждому x_i.

Ряд 1. Точечный вариационный ряд.

x_i	85	87	88	91	94	95	96	97	98
n_i	1	1	1	1	2	3	1	2	1
x_i	99	100	101	103	104	106	107	108	109
n_i	3	3	2	4	4	3	6	7	7
x_i	110	111	112	113	114	115	116	117	118
n_i	3	3	5	5	3	4	3	4	2
x_i	119	120	122	123	126	128	129	131	134
n_i	1	3	1	2	3	1	1	3	1

Проверка: $\sum_{n=1}^{36} n_i = 100$. В результате построение ряда 1 получилось 36 различных значений в выборке.

3. Чтобы от ряда 1 перейти к интервальному ряду 2, проводим следующие вспомогательные расчеты:

$x_{\max} = 134$ чел., $x_{\min} = 85$ чел.

Размах вариации $R = 134 - 85 = 49$ чел. Получаем диапазон значений в выборке $[85; 134]$, который для удобства расчетов следует разбить на k интервалов:

$k \approx \log_2 100 + 1 = 7{,}645$. Так как k должно быть целым, тогда берем $k = 8$ интервалам.

Шаг интервала (ширина интервала) $h = \dfrac{85}{8} = 6{,}125$.

Находим границы интервалов:

$C_0 = 85$, $C_1 = 85 + 6{,}125 = 91{,}125$, $C_2 = 91{,}125 + 6{,}125 = 97{,}25$, $C_3 = 103{,}375$, $C_4 = 109{,}5$, $C_5 = 115{,}625$, $C_6 = 121{,}75$, $C_7 = 127{,}875$, $C_8 = 134 = x_{\max}$.

Подсчитываем, сколько значений попало в каждый интервал, и оформляем результаты в виде ряда 2:

Ряд 2. Интервальный ряд.

$C_{i-1} - C_i$	85 – 91,125	91,125 – 97,25	97,25 – 103,375	103,375 – 109,5
n_i	4	8	13	27
$C_{i-1} - C_i$	109,5 –115,625	115,625 – 121,75	121,75 – 127,875	127,875 – 134
n_i	23	13	6	6

Проверка: $\sum_{n=1}^{8} n_i = 100$.

4. Для построения ряда 3 находим середину каждого интервала:

$x_1^* = \frac{85 + 91{,}125}{2} = 88{,}063$, $x_2^* = \frac{91{,}125 + 97{,}25}{2} = 94{,}188$,

$x_3^* = \frac{97{,}25 + 103{,}375}{2} = 100{,}313$, $x_4^* = \frac{103{,}375 + 109{,}5}{2} = 106{,}348$,

$x_5^* = \frac{109{,}5 + 115{,}625}{2} = 112{,}563$, $x_6^* = \frac{115{,}625 + 121{,}75}{2} = 118{,}688$,

$x_6^* = \frac{121{,}75 + 127{,}875}{2} = 124{,}813$, $x_8^* = \frac{127{,}875 + 134}{2} = 130{,}938$.

Ряд 3. Точечный ряд.

x_i^*	88,063	94,188	100,313	106,438	112,563	118,688	124,813	130,938
n_i	4	8	13	27	23	13	6	6

Для ряда 4 находим относительные частоты:

$w_1 = \frac{n_1}{n} = \frac{4}{100} = 0{,}04$, $w_2 = \frac{n_2}{n} = \frac{8}{100} = 0{,}08$, $w_3 = \frac{n_3}{n} = \frac{13}{100} = 0{,}13$,

$w_4 = \frac{n_4}{n} = \frac{27}{100} = 0{,}27$, $w_5 = \frac{n_5}{n} = \frac{23}{100} = 0{,}23$, $w_6 = \frac{n_6}{n} = \frac{13}{100} = 0{,}13$,

$w_7 = \frac{n_7}{n} = \frac{6}{100} = 0{,}06$, $w_8 = \frac{n_8}{n} = \frac{6}{100} = 0{,}06$.

Относительная частота w_i показывает, какую долю занимает данное значение x_i^* в общем объеме выборки. Например, $x_5^* = 112{,}563$ составляет 23% от всех значений в выборке, т.е. в 23 % случаев наполняемость одного авиарейса была примерно 113 пассажиров.

Ряд 4. Точечный ряд, построенный по относительным частотам.

x_i^*	88,063	94,188	100,313	106,438	112,563	118,688	124,813	130,938
w_i	0,04	0,08	0,13	0,27	0,23	0,13	0,06	0,06
w_i,%	4%	8%	13%	27%	23%	13%	6%	6%

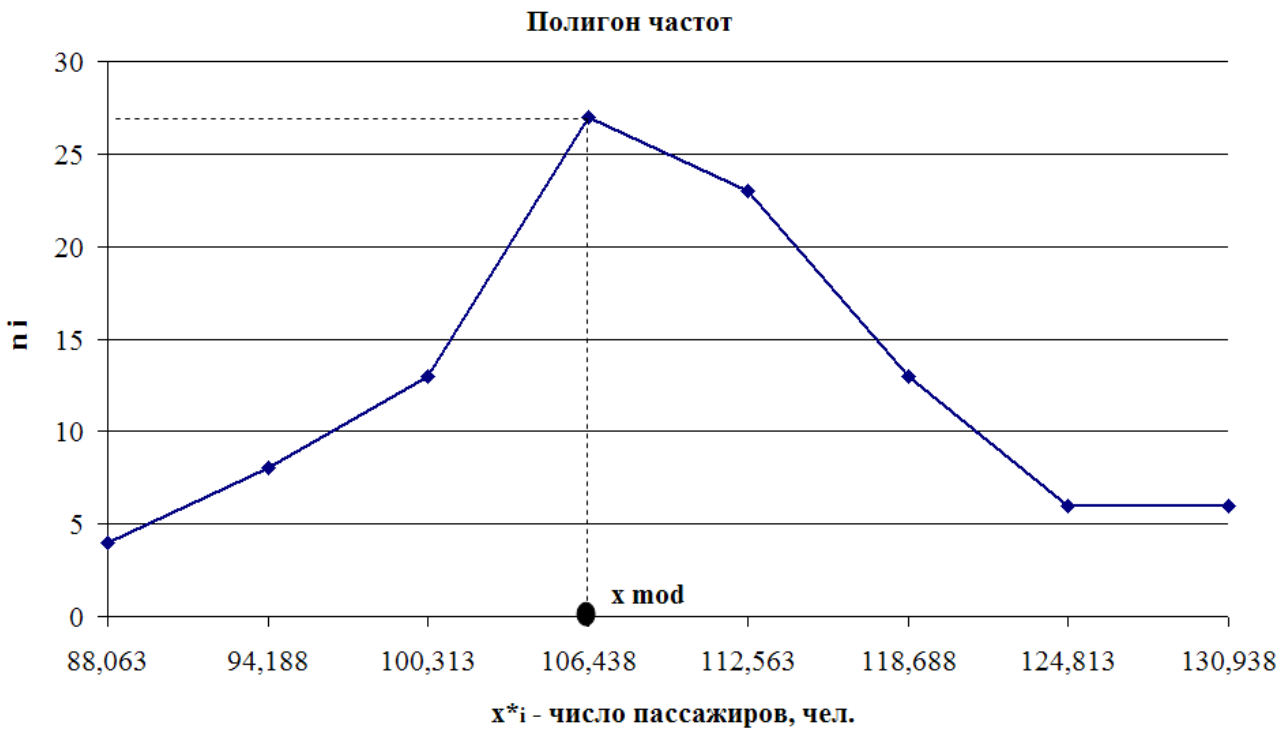
Проверка: $\sum_{n=1}^{8} w_i = 1$, $\sum_{n=1}^{8} w_i \cdot 100\% = 100\%$.

Для ряда 5 рассчитываем накопленные частоты: $m_1 = n_1 = 4$,
$m_2 = n_2 + m_1 = 4 + 8 = 12$, $m_3 = n_3 + m_2 = 13 + 12 = 25$,
$m_4 = n_4 + m_3 = 25 + 27 = 52$, $m_5 = n_5 + m_4 = 52 + 23 = 75$,
$m_6 = n_6 + m_5 = 13 + 75 = 88$, $m_7 = n_7 + m_6 = 6 + 88 = 94$,
$m_8 = n_8 + m_7 = 6 + 94 = 100$.

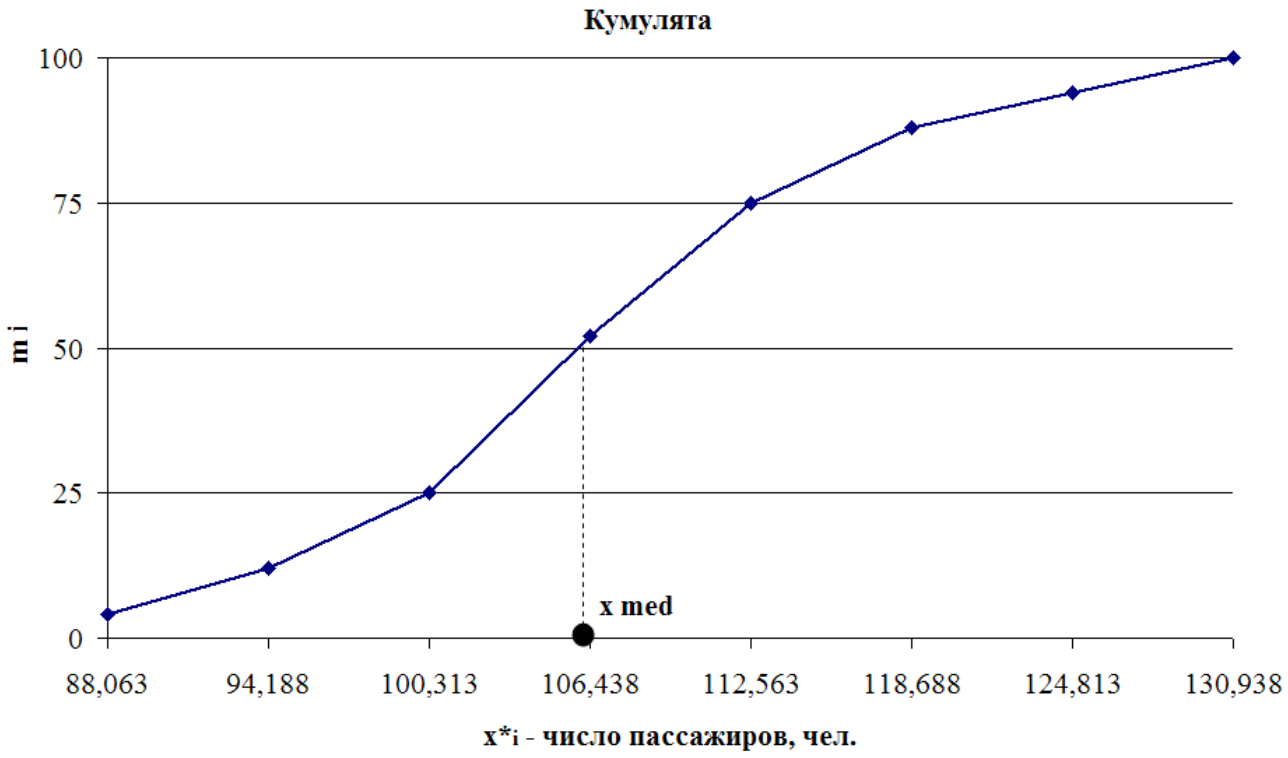
Ряд 5. Точечный ряд, построенный по накопленным частотам.

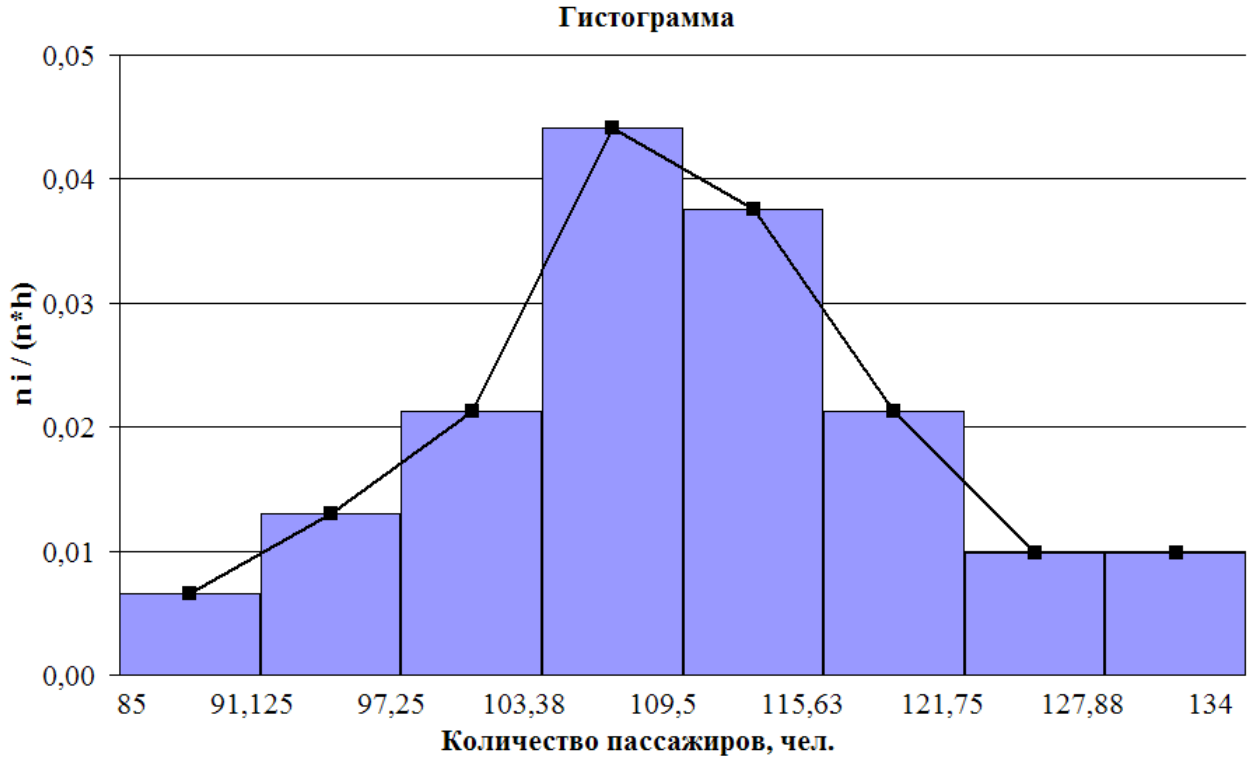
x_i^*	88,063	94,188	100,313	106,438	112,563	118,688	124,813	130,938
m_i	4	12	25	52	75	88	94	100

5. Графики:



По графикам можно определить следующие меры положения: моду x_{mod} – по полигону частот, как значение, соответствующее наибольшей частоте ($x_{\text{mod}} \approx 106$ чел.), медиану x_{med} – по кумуляте, как значение, соответствующее половине выборке, т.е. 50 ($x_{med} \approx 106$ чел.). Это означает, что на авиарейсах «Иркутск–Москва» или «Москва–Иркутск» чаще всего летает 106 пассажиров, средневероятное число пассажиров тоже составляет 106 пассажиров.





6. Эмпирическая функция распределения: $F^*(x)$ – это статистическая аппроксимация функции распределения $F(x) = P(x < X)$. Например, $F^*(x) = 0{,}75$ – это вероятность того, что $x < 118{,}688$, т.е. в 75% случаев число пассажиров в одном рейсе составляло менее 119 чел.

$$F^*(x)=\begin{cases}0, & x\le 88{,}063,\\ 0{,}04, & 88{,}063<x\le 94{,}188,\\ 0{,}12, & 94{,}188<x\le 100{,}313,\\ 0{,}25, & 100{,}313<x\le 106{,}438,\\ 0{,}52, & 106{,}438<x\le 112{,}563,\\ 0{,}75, & 112{,}563<x\le 118{,}688,\\ 0{,}88, & 118{,}688<x\le 124{,}813,\\ 0{,}94, & 124{,}813<x\le 130{,}938,\\ 1, & x>130{,}938.\end{cases}$$

7. Числовые характеристики

Для расчета числовых характеристик составим вспомогательную таблицу:

		расчет $\bar{x}_в$		расчет $D_в$		расчет A_s	расчет E_k
x_i^*	n_i	$x_i^* n_i$	$(x_i^*-\bar{x}_в)n_i$	$n_i(x_i^*-\bar{x}_в)^2$	$(x_i^*)^2 n_i$	$n_i(x_i^*-\bar{x}_в)^3$	$n_i(x_i^*-\bar{x}_в)^4$
88,0625	4	352,25	-21,438	1838,27	31020,02	-39407,8	844805,1
94,1875	8	753,5	-15,313	1875,78	70970,28	-28722,9	439819,4
100,3125	13	1304,1	-9,188	1097,33	130813,8	-10081,7	92626,0
106,4375	27	2873,8	-3,063	253,23	305881,4	-775,5	2375,0
112,5625	23	2588,9	3,063	215,71	291417,3	660,6	2023,2
118,6875	13	1542,9	9,188	1097,33	183127,4	10081,7	92626,0
124,8125	6	748,9	15,313	1406,84	93468,96	21542,2	329864,6
130,9375	6	785,6	21,438	2757,40	102867,8	59111,7	1267207,7
Σ	**100**	**10950**	**0,000**	**10541,9**	**1209567**	**12408,3**	**3071346,9**

А) Меры положения

Среднее выборочное значение:

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{8} n_i x_i^* = \frac{1}{100}(88{,}0625 \cdot 4 + 94{,}1875 \cdot 8 + 100{,}3125 \cdot 13 + 106{,}4375 \cdot 27 +$$

$$+ 112{,}5628 \cdot 23 + 118{,}6875 \cdot 13 + 124{,}8125 \cdot 6 + +130{,}9375 \cdot 6) = \frac{10950}{100} = 109{,}5 \ \textit{чел.}$$

В течение наблюдаемого времени один авиарейс в среднем перевозил 109,5 пассажиров.

Медиана: $x_{med} = 103{,}375 + 6{,}125 \cdot \frac{50-25}{27} = 109{,}046$ чел.

Медиану также можно определить, как значение случайной величины *X*, расположенное между $x_{n/2}$ и $x_{(n/2)+1}$ при четном *n*. x_{50}, x_{51} определяем по ряду 1, как значения, распложенные напротив накопленных частот 50 и 51:

$$x_{med} = \frac{x_{50} + x_{51}}{2} = \frac{109 + 109}{2} = 109 \text{ чел.}$$

Мода: $x_{\text{mod}} = 103{,}375 + 6{,}125 \cdot \frac{27-13}{27 \cdot 2 - 13 - 23} = 108{,}139$ чел.

По ряду 1 x_{mod} – это значение, соответствующее наибольшей частоте, следовательно, $x_{\text{mod}} = 108$ или 109 (значения, стоящие напротив частот $n_{17,18} = 7$).

Таким образом, наиболее часто встречающееся число пассажиров одного авиарейса составляет 108 чел., средневероятное – 109 чел.

Б) Меры разброса (рассеяния)

Дисперсия:

$$D_в = \frac{1}{100}(1838{,}27 + 1875{,}78 + 1097{,}33 + 253{,}23 + 215{,}71 + 1097{,}33 + 1406{,}84 +$$

$$+ 2757{,}40) = \frac{10541{,}9}{100} = 105{,}419.$$

Дисперсию также можно вычислить по второй формуле:

$$D_в = \frac{1}{100}\sum_{i=1}^{k}(x_i^*)^2 \cdot n_i - (\bar{x}_в)^2 = \frac{1209567}{100} - (109{,}5)^2 = 105{,}419.$$

Среднеквадратическое отклонение: $\sigma_в = \sqrt{D_в} = \sqrt{105{,}419} = 10{,}27$ чел.

Коэффициент вариации: $V_в = \frac{10{,}27}{109{,}5} \cdot 100\% = 9{,}38\%$.

Абсолютное отклонение от среднего значения составляет $\pm 10{,}27$ чел., относительное отклонение от среднего равно 9,38%.

В) Меры формы

Выборочный коэффициент асимметрии: $A_s = \frac{124{,}083}{10{,}27^3} = 0{,}115$, где $\mu_3 = \frac{12408{,}3}{100} = 124{,}083$.

Выборочный коэффициент эксцесса: $E_k = \frac{30713{,}469}{10{,}27^4} - 3 = -0{,}236$, где $\mu_4 = \frac{3071346{,}9}{100} = 30713{,}469$.

Положительное значение коэффициента асимметрии говорит о том, что более длинная часть графика находится справа от вершины. Отрицательное значение коэффициента эксцесса говорит о плосковершинности кривой распределения.

8. Вывод о близости наблюдаемого распределения к нормальному:

1) Полигон частот имеет колоколообразный вид;

2) $\bar{x}_в \approx x_{\text{mod}} \approx x_{med}$: $109{,}5 \approx 109{,}046 \approx 108{,}139$;

3) Значения коэффициентов асимметрии и эксцесса близки к нулю;

4) коэффициент вариации меньше 33%.

Таким образом, на основании проделанных расчетов можно сделать вывод о близости наблюдаемого распределения случайной величины X – числа пассажиров одного авиарейса «Иркутск–Москва» или «Москва–Иркутск» к нормальному.

2. Статистическое оценивание параметров

2.1. Постановка задачи оценивания параметров

Пусть мы располагаем выборкой $\mathrm{X} = \{x_1, \ldots, x_n\}$ объема n из генеральной совокупности. Пусть интересующее нас свойство этой совокупности описывается с помощью уравнения или математической модели вида

$$Y(\mathrm{X};\theta) = 0, \tag{2.1}$$

где X – текущее значение исследуемого в общем случае p – мерного случайного признака, $\theta = (\theta_1, \ldots, \theta_k)$ – k – мерный вектор параметров, значения которых неизвестны до получения выборки.

В качестве модели (2.1) могут рассматриваться модели законов распределения вероятностей, либо модели статистических зависимостей, существующих между анализируемыми показателями.

Например:

1) Пусть ξ – дискретная случайная величина, распределенная по закону Пуассона $\xi \sim \Pi(\lambda)$, в качестве модели (2.1) могут быть рассмотрены вероятности

$$P\{\xi = x;\theta\} = \frac{\theta^x}{x!}e^{-\theta}, \quad \theta = M\xi = D\xi, \quad x > 0, \quad \theta \text{ – неизвестный параметр.}$$

2) Пусть ξ – одномерная непрерывная случайная величина, распределенная по нормальному закону $\xi \sim N(a,\sigma), \quad a = M\xi, \quad \sigma = \sqrt{D\xi}$, то в качестве модели (2.1) можно рассмотреть функцию плотности вероятности

$$f_\xi(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}}e^{\frac{-(x-\theta_1)^2}{2\theta_2}}, \ \theta_1 = M\xi = a, \quad \theta_2 = D\xi = \sigma^2.$$

3) Пусть x – уровень производительности труда, y – уровень заработной платы, тогда $y = \theta_0 + \theta_1 x + \theta_2 x^2$.

В дальнейшем будем считать, что задана теоретико-вероятностная модель, то есть изучаемая случайная величина X распределена по закону $p(x;\theta)$, где θ – один или несколько неизвестных параметров, а $p(\bullet)$ – вероятность, если X – дискретная случайная величина и $p(\bullet)$ – плотность распределения вероятностей, если X – непрерывная случайная величина.

Если выборка X объема n была получена независимым образом, то совместная плотность распределения выборки будет иметь вид:

$$p(x;\theta) = p(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta) = p(x_1,\theta)\cdots p(x_n,\theta). \tag{2.2}$$

2.2. Свойства точечных оценок параметров

Пусть ξ – случайная величина, описываемая законом распределения $p(x;\theta)$, θ – неизвестный параметр.

Задача статистического оценивания параметров состоит в построении такой функции выборки вида

$$\hat{\theta} = T(x_1, \ldots, x_n), \tag{2.3}$$

которая в определенном смысле наиболее точно соответствовала бы истинному значению параметра.

Любая функция выборки называется статистикой. Статистика, принимаемая в качестве приближенного значения неизвестного параметра, называется статистической оценкой. Оценка, найденная в виде одного числа, называется точечной. По содержанию, оценка как функция выборки, является случайной величиной, принимающей различные значения при переходе от одной выборки к другой в рамках одной генеральной совокупности. По этой причине она подвержена разбросу относительно истинного значения параметра. Для того, чтобы разброс был минимальным и оценка наилучшим образом соответствовала бы истинному значению параметра, она должна удовлетворять следующим требованиям:

1) Состоятельность. Оценка $\hat{\theta}$ неизвестного параметра θ называется **состоятельной**, если по мере роста числа наблюдений при $n \to \infty$ она сходится по вероятности к истинному значению параметра, то есть $\forall \varepsilon > 0$ при $n \to \infty$

$$P\{\left|\hat{\theta} - \theta\right| > \varepsilon\} \to 0, n \to \infty.$$

Практически, требование состоятельности означает, что оценка совпадает с истинным значением этого параметра лишь для выборок большого объема.

2) Несмещенность. Оценка $\hat{\theta}$ неизвестного параметра θ называется **несмещенной**, если $\forall n$ результат усреднения выборки по всем возможным значениям выборки объема n совпадает с истинным значением параметра

$$M\hat{\theta} = \theta.$$

Практически, несмещенные оценки не всегда удается построить, поэтому требуется построить хотя бы асимптотически несмещенные оценки, то есть такие, для которых $M\hat{\theta} \to \theta, n \to \infty.$

3) Эффективность. Оценка $\hat{\theta}$ неизвестного параметра θ называется **эффективной**, если она среди всех прочих оценок этого параметра обладает наименьшей мерой случайного разброса относительно истинного значения параметра. В качестве меры разброса оценки принимается ее вариация

$$V(\hat{\theta}) = M(\hat{\theta} - \theta)^2.$$

Если оценка обладает свойством несмещенности, то есть, $M\hat{\theta} = \theta,$ то ее вариация совпадает с дисперсией, $V(\hat{\theta}) = M(\hat{\theta} - M\hat{\theta})^2 = D\hat{\theta}.$ Очевидно, что $V(\hat{\theta}) \geq 0$. Сделать вариацию оценки минимально возможной, то есть равной нулю нельзя, так как для нее существует нижняя граница, достичь которую

можно, но превзойти нельзя. Свойство эффективности характеризует качество оценки.

Если оценка обладает свойством несмещенности и эффективности, то она обязательно будет состоятельной.

2.3. Методы статистического оценивания параметров

1. Метод максимального правдоподобия (ММП).

Пусть независимая выборка $X = \{x_1, \ldots, x_n\}$ объема n извлечена из распределения $p(x;\theta)$, где θ – один или несколько неизвестных параметров. По этой выборке построим функцию (2.2). С точки зрения вероятности, это совместная плотность распределения выборки, x_i – текущие значения случайной величины, параметр θ – фиксирован. С точки зрения математической статистики, наоборот – x_i являются фиксированными, а параметр θ – неизвестная величина.

Функцию вида (2.2) называют функцией правдоподобия и обозначают

$$L(\theta) = \prod_{i=1}^{n} p(x_i;\theta). \qquad (2.4)$$

ММП заключается в том, что в качестве оценки неизвестного параметра $\hat{\theta}$ принимается такой аргумент функции правдоподобия (2.2), при котором она достигает своего максимума

$$\hat{\theta} = \arg\max_{\theta} L(\theta). \qquad (2.5)$$

Для определения ММП-оценки решается уравнение правдоподобия

$$\frac{\partial L(\theta)}{\partial \theta} = 0. \qquad (2.6)$$

Часто для удобства находят максимум не функции правдоподобия а логарифмированной функции правдоподобия $l(\theta) = \ln L(\theta)$. Тогда уравнение правдоподобия будет иметь вид:

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{1}{L(\theta)} \frac{\partial L(\theta)}{\partial \theta} = 0.$$

Свойства ММП-оценок.

1) ММП-оценки являются асимптотически несмещенными.
2) ММП-оценки являются асимптотически эффективными.
3) ММП-оценки являются нормально распределенными.

Пример 2.1. Пусть независимая выборка $X = \{x_1, \ldots, x_n\}$ извлечена из распределения Пуассона, то есть $p(x,\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$, $\lambda = MX$. Найти ММП-оценку параметра λ.

Решение. Составим функцию правдоподобия:

$$L(\lambda) = \prod_{i=1}^{n} p(x_i;\lambda) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \frac{\lambda^{x_2}}{x_2!} e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{x_1! x_2! \ldots x_n!} e^{-n\lambda}.$$

Прологарифмируем эту функцию:

$$l(\lambda) = \ln L(\lambda) = \sum_{i=1}^{n} x_i \ln\lambda - \ln(x_1! \cdots x_n!) - n\lambda .$$

Для нахождения максимума, найдем производную этой функции и приравняем ее нулю

$$\frac{\partial l(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} x_i \frac{1}{\lambda} - n = 0 .$$

Решая это уравнение, найдем оценку параметра

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} ,$$

то есть оценкой неизвестного параметра является выборочное среднее.

Пример 2.2. Пусть изучаемая случайная величина X распределена по нормальному закону с математическим ожиданием, равным a и дисперсией D. Найти ММП-оценки параметров a и D.

Решение. Плотность распределения этой величины зависит от a и D и имеет вид

$$p(x; a, D) = \frac{1}{\sqrt{2\pi D}} e^{\frac{-(x-a)^2}{2D}} .$$

Для определения оценок параметров этого распределения имеем независимую выборку $X = \{x_1, \ldots, x_n\}$. Функция правдоподобия в этом случае будет функцией двух параметров и определится как

$$L(a, D) = \frac{1}{\sqrt{2\pi D}} e^{\frac{-(x_1-a)^2}{2D}} \frac{1}{\sqrt{2\pi D}} e^{\frac{-(x_2-a)^2}{2D}} \cdots \frac{1}{\sqrt{2\pi D}} e^{\frac{-(x_n-a)^2}{2D}} = \frac{1}{(2\pi D)^{n/2}} e^{-\frac{1}{2D}\sum_{i=1}^{n}(x_i-a)^2} .$$

Для простоты будем искать максимум логарифма функции правдоподобия $l(a, D) = \ln L(a, D) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln D - \frac{1}{2D}\sum_{i=1}^{n}(x_i - a)^2$. Составим уравнения правдоподобия:

$$\frac{\partial l(a, D)}{\partial a} = \frac{1}{D}\sum_{i=1}^{n}(x_i - a) = 0 , \quad \frac{\partial l(a, D)}{\partial D} = -\frac{n}{2D} + \frac{1}{2D^2}\sum_{i=1}^{n}(x_i - a)^2 = 0.$$

Из первого уравнения найдем

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.$$

Подставляя это значение вместо a во второе уравнение и решая его относительно D, получим

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = D_B .$$

Таким образом, оценками математического ожидания и дисперсии нормально распределенной случайной величины, являются выборочное среднее и

выборочная дисперсия соответственно. Если проверить найденные оценки на несмещенность, то получится, что оценка математического ожидания в виде выборочного среднего является несмещенной. А оценка дисперсии в виде выборочной дисперсии является смещенной. Обычно смещение в оценке дисперсии устраняют, и в качестве несмещенной оценки используют исправленную выборочную дисперсию

$$S^2 = \frac{n}{n-1} D_B = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

2. Метод моментов (ММ).

Пусть выборка $X = \{x_1, \ldots, x_n\}$ извлечена из генеральной совокупности, вероятностное свойство которой описывается функцией плотности $p(x;\theta) = p(x;\theta_1, \theta_2, \ldots, \theta_s)$, $\theta_1, \theta_2, \ldots, \theta_s$ – неизвестные параметры. Предположим, что первые S начальные теоретические моменты существуют и конечны, теоретический момент r-го порядка определяется по формуле $m_r = MX^r = \int x^r p(x;\theta_1, \theta_2, \ldots, \theta_s) dx = m_r(\theta_1, \ldots, \theta_s)$, $r = \overline{1,s}$. По выборке X найдем выборочные или эмпирические начальные моменты, которые будут несмещенными оценками соответствующих теоретических моментов

$$\hat{m}_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r, \quad r = \overline{1,s}.$$

Метод моментов состоит в том, что оценки неизвестных параметров находятся как решение системы линейных уравнений:

$$\begin{aligned} m_1(\theta_1, \ldots, \theta_s) &= \hat{m}_1 \\ m_2(\theta_1, \ldots, \theta_s) &= \hat{m}_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ m_s(\theta_1, \ldots, \theta_s) &= \hat{m}_s \end{aligned} . \qquad (2.7)$$

Использование начальных моментов не обязательно. Достоинством метода моментов является его простота.

Пример 2.3. При тестировании группы студентов есть основание считать, что средний балл X – равномерно распределенная на отрезке $[a,b]$ случайная величина. Результаты обследований представлены в виде ряда:

x_i	1	2	3	4	5
n_i	12	10	9	9	10

$$n = \sum n_i = 50.$$

Найти методом моментов оценки параметров $\hat{a}$ и $\hat{b}$.

Для равномерного на отрезке $[a,b]$ распределения имеем:

$$p(x;a,b) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}.$$

Найдем теоретические начальные моменты

$$m_1 = MX = \int_a^b x p(x;a,b) dx = \frac{1}{b-a} \int_a^b x dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2} = m_1(a,b);$$

$$m_2 = MX^2 = \int_a^b x^2 p(x;a,b)dx = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3} = m_2(a,b).$$

Эмпирические начальные моменты находятся по заданной выборке объема $n = 50$:

$$\hat{m}_1 = \frac{1}{n}\sum_{i=1}^{5} n_i x_i = \frac{1}{50}(12 \cdot 1 + 10 \cdot 2 + 9 \cdot 3 + 9 \cdot 4 + 10 \cdot 5) = 2{,}9,$$

$$\hat{m}_2 = \frac{1}{n}\sum_{i=1}^{5} n_i x_i^2 = \frac{1}{50}(12 \cdot 1 + 10 \cdot 2^2 + 9 \cdot 3^2 + 9 \cdot 4^2 + 10 \cdot 5^2) = 10{,}54.$$

По методу моментов оценки двух неизвестных параметров $\hat{a}$ и $\hat{b}$ находятся из решения системы уравнений:

$$\begin{cases} m_1(a,b) = \hat{m}_1 \\ m_2(a,b) = \hat{m}_2 \end{cases}, \quad \text{то есть} \quad \begin{cases} \dfrac{a+b}{2} = 2{,}9 \\ \dfrac{a^2 + ab + b^2}{3} = 10{,}54 \end{cases}$$

Из решения этой системы получаем $\hat{a} \approx 0{,}37, \quad \hat{b} \approx 5{,}43.$

2.4. Понятие об интервальном оценивании

Оценка $\hat{\theta}$ неизвестного параметра θ является лишь его приближенным значением, поэтому замена параметра его оценкой может привести к ошибкам. Пусть величина $\Delta > 0$ характеризует точность оценивания

$$\left|\theta - \hat{\theta}\right| < \Delta. \tag{2.8}$$

Так как $\hat{\theta}$ является случайной величиной, то задав точность Δ, мы не можем абсолютно достоверно (с вероятностью равной 1), гарантировать выполнение неравенства (2.8). Можно говорить лишь о практической достоверности (с вероятностью близкой к 1). Необходимо определить интервал вида $(\hat{\theta} - \Delta; \hat{\theta} + \Delta)$, который с заранее заданной вероятностью, близкой к 1, покрывал бы истинное неизвестное значение параметра. Такой интервал называется доверительным интервалом или интервальной оценкой._А выбираемая исследователем вероятность, близкая к 1, называется доверительной вероятностью или надежностью. Доверительная вероятность обозначается γ и определяется по формуле

$$\gamma = P\{\hat{\theta} - \Delta < \theta < \hat{\theta} + \Delta\}. \tag{2.9}$$

Ширина доверительного интервала существенно зависит от объема выборки (уменьшается с ростом n) и от величины доверительной вероятности (увеличивается с приближением вероятности к 1).

Приведем формулы расчета доверительных границ для параметров нормального распределения. Пусть случайная величина $X \sim N(a,\sigma)$.

1. Доверительный интервал для неизвестного математического ожидания a при неизвестной дисперсии σ^2 определяется по формуле

$$\bar{x} - t_\gamma \frac{S}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{S}{\sqrt{n}} \tag{2.10}$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное среднее, являющееся несмещенной оценкой математического ожидания, n – объем выборки, $S = \sqrt{S^2}$ – корень из исправленной дисперсии, t_γ – квантиль распределения Стьюдента находится из таблицы по заданной доверительной вероятности γ и числу степеней свободы $k = n - 1$.

Пример 2.4. Из многочисленного количества сотрудников фирмы случайным образом отобрано $n = 25$ человек. Средняя заработная плата этих сотрудников составила $\bar{x} = 700$ д.е. при среднем квадратическом отклонении $S = 100$ д.е. Требуется с доверительной вероятностью $\gamma = 0{,}95$ определить интервальную оценку для:

1) среднемесячной заработной платы на фирме;

2) суммы затрат фирмы на заработную плату отдела из 520 сотрудников.

Решение.

1) Пусть случайная величина X – размер заработной платы, тогда среднемесячная заработная плата на фирме. Для построения доверительного интервала воспользуемся формулой (2.10), получим

$$700 - 2{,}064\frac{100}{\sqrt{25}} < a < 700 + 2{,}064\frac{100}{\sqrt{25}}.$$

Значение $t_{0,95} = t(0{,}95;24) = 2{,}064$ найдено из таблицы квантилей распределения Стьюдента при уровне $p = \frac{1+\gamma}{2} = 0{,}975$ и числе степеней свободы $k = n - 1 = 24$. Окончательно

$$658{,}72 < a < 741{,}28,$$

то есть с вероятностью 0,95 можно гарантировать, что средняя заработная плата на фирме находится в пределах от 658,72 д.е. до 741,28 д.е.

2) Для определения суммы затрат на заработную плату отдела необходимо найденные доверительные границы умножить на количество сотрудников $N = 520$, получим

$$342534 < Na < 385465.$$

2. Доверительный интервал для неизвестной дисперсии σ^2 при неизвестном математическом ожидании a определяется по формуле

$$\frac{(n-1)S^2}{U_2} < \sigma^2 < \frac{(n-1)S^2}{U_1}, \qquad (2.11)$$

где S^2 – несмещенная оценка дисперсии, найденная по выборке объема n, $U_2 = \chi^2\left(\frac{1+\gamma}{2}; n-1\right)$ и $U_1 = \chi^2\left(\frac{1-\gamma}{2}; n-1\right)$ – квантили распределения χ^2 – квадрат находятся по таблице.

Пример 2.5. При изучении производительности труда X (тыс. руб.) и обследовании $n = 100$ предприятий есть основание считать что величина X является случайной, распределённой по нормальному закону, несмещенные оценки

параметров которого равны $\hat{a} = \overline{x} = 15$ (тыс. руб.) и $S = 0{,}96$ (тыс. руб.). Найти доверительный интервал для неизвестной дисперсии, считая $\gamma = 0{,}95$.

Решение. Найдем квантили распределения «хи-квадрат» $U_1 = \chi^2(0{,}025;99) = 74{,}2$, $U_2 = \chi^2(0{,}975;99) = 129{,}6$. Так как $S^2 = (0{,}96)^2 = 0{,}92$, то мы имеем по формуле (2.11):

$$\frac{99 \cdot 0{,}92}{129{,}6} < \sigma^2 < \frac{99 \cdot 0{,}92}{74{,}2}$$

или

$$0{,}7 < \sigma^2 < 1{,}26.$$

2.5. Рекомендации по выполнению расчетно-графической работы по теме «Статистическое оценивание параметров» в MS Excel

Задание:

1. Указать несмещенные оценки неизвестного математического ожидания и дисперсии случайной величины, выборка которой была представлена в теме 1.

2. Построить доверительные интервалы для неизвестного математического ожидания и дисперсии, в предположении, что выборка из нормальной генеральной совокупности и $\gamma_1 = 0{,}95$, $\gamma_2 = 0{,}9$.

Вычисление в MS Excel несмещенных оценок генеральных средней и дисперсии большой сложности не составляет (рис. 1.28). Для нахождения квантилей распределения Стьюдента воспользуемся встроенными в MS Excel специальными статистическими функциями. Для t_γ нам понадобится функция =СТЬЮДРАСПОБР(вероятность; степени свободы). В качестве вероятности вводится значение $1-\gamma$, степеней свободы – $n-1$. Пересчет $(1+\gamma)/2$ уже встроен в эту функцию, т.е. специально его делать не надо. Для вычисления квантилей распределения χ^2 u_1 и u_2 также находим статистическую функцию =ХИ2ОБР(вероятность; степени свободы). Вероятность для u_1 равна $1-\alpha/2$, для u_2 – $\alpha/2$, степени свободы = $n-1$. Когда все компоненты формул (1.18) и (1.19) уже записаны, можно вычислить доверительные интервалы для неизвестных математического ожидания и дисперсии (рис. 1.23).

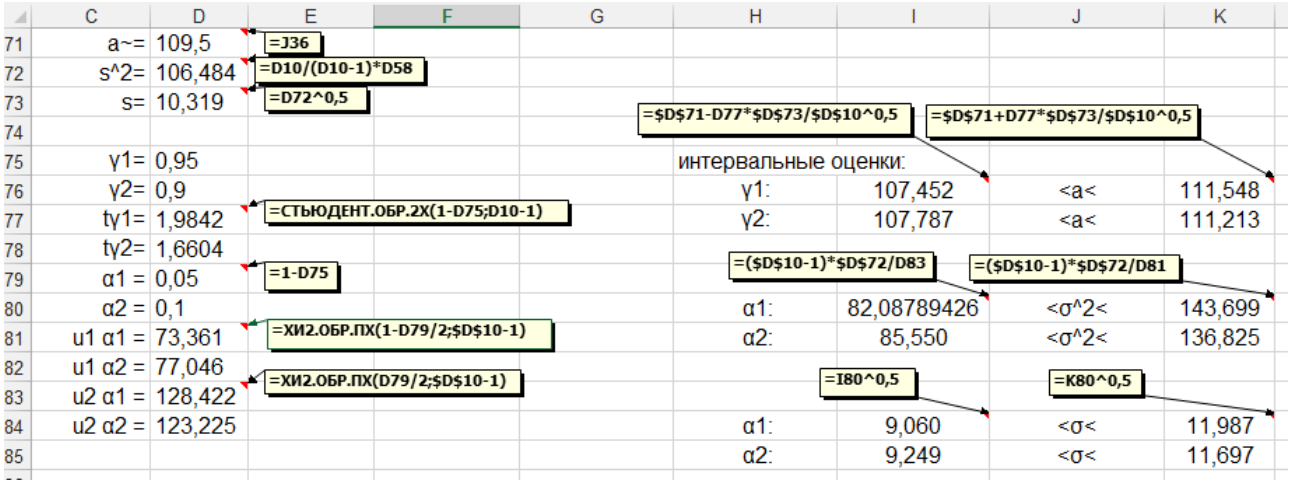


Рис. 1.23. Вычисление точечных и интервальных оценок

2.6. Оформление результатов проведенных расчетов по теме «Статистическое оценивание параметров»

Пример 2.6. 1. Точечные оценки.

Несмещенная оценка неизвестного математического ожидания:

$\tilde{a} = \bar{x}_в = 109{,}5$ чел.

Несмещенная оценка неизвестной дисперсии:

$$s^2 = \frac{n}{n-1} D_в = \frac{100}{99} \cdot 105{,}42 = 106{,}48.$$

$$s = \sqrt{106{,}48} = 10{,}319.$$

2. Интервальные оценки:

а) для неизвестного математического ожидания:

Пусть доверительная вероятность $\gamma_1 = 0{,}95$, тогда $t_{0,95} = t\left(\frac{1+0{,}95}{2}; 100-1\right) = t(0{,}975; 99) = 1{,}9842$, при этом

$$109{,}5 - 1{,}9842\frac{10{,}319}{\sqrt{100}} < a < 109{,}5 + 1{,}9842\frac{10{,}319}{\sqrt{100}},$$

$$107{,}452 < a < 111{,}548.$$

С вероятностью 0,95 можно гарантировать, что среднее число пассажиров одного авиарейса будет в пределах от 107,452 до 111,548 чел. Другими словами, доверительный интервал от 107,452 до 111,548 чел. с вероятностью 0,95 покроет неизвестное значение среднего числа пассажиров одного авиарейса.

Пусть доверительная вероятность $\gamma_2 = 0{,}9$, тогда $t_{0,9} = t\left(\frac{1+0{,}9}{2}; 100-1\right) = t(0{,}95; 99) = 1{,}6604$, при этом

$$109{,}5 - 1{,}6604\frac{10{,}319}{\sqrt{100}} < a < 109{,}5 + 1{,}6604\frac{10{,}319}{\sqrt{100}},$$

$$107{,}787 < a < 111{,}213.$$

С вероятностью 0,9 можно гарантировать, что среднее число пассажиров одного авиарейса будет в пределах от 107,787 до 111,213 чел.

Из расчетов видно, что при меньшей доверительной вероятности ширина доверительного интервала сужается.

Б) для неизвестной дисперсии:

Пусть доверительная вероятность $\gamma_1 = 0{,}95$, тогда

$u_1 = \chi^2\left(\frac{0{,}05}{2};99\right) = \chi^2(0{,}025;99) = 73{,}361$ и

$u_2 = \chi^2(1-0{,}025;99) = \chi^2(0{,}975;99) = 128{,}42$.

$$\frac{99 \cdot 106{,}48}{128{,}42} < \sigma^2 < \frac{99 \cdot 106{,}48}{73{,}361}, \qquad 82{,}088 < \sigma^2 < 143{,}699,$$

$$\sqrt{82{,}088} < \sigma < \sqrt{143{,}699}, \qquad 9{,}060 < \sigma < 11{,}987.$$

Доверительный интервал от 82,088 до 143,699 с вероятностью 0,95 покроет неизвестное значение дисперсии, а доверительный интервал от 9,06 до 11,987 – неизвестное значение среднего квадратического отклонения.

Пусть доверительная вероятность $\gamma_2 = 0{,}9$, тогда

$u_1 = \chi^2\left(\frac{0{,}1}{2};99\right) = \chi^2(0{,}05;99) = 77{,}046$ и

$u_2 = \chi^2(1-0{,}05;99) = \chi^2(0{,}95;99) = 123{,}23$.

$$\frac{99 \cdot 106{,}48}{123{,}23} < \sigma^2 < \frac{99 \cdot 106{,}48}{77{,}046}, \quad 85{,}550 < \sigma^2 < 136{,}825 \ ,$$

$$\sqrt{85{,}550} < \sigma < \sqrt{136{,}825}, \qquad 9{,}249 < \sigma < 11{,}697.$$

Доверительный интервал от 85,550 до 136,825 с вероятностью 0,9 покроет неизвестное значение дисперсии, а доверительный интервал от 9,249 до 11,697 – неизвестное значение среднего квадратического отклонения.

3. Статистическая проверка гипотез

3.1. Постановка задачи

При решении многих практических задач результаты наблюдений используются для проверки предположений или гипотез относительно тех или иных свойств распределения генеральной совокупности.

Любое предположение, касающееся либо значений неизвестных параметров, либо вида предполагаемой модели, либо наличия связей (корреляции), называется статистической гипотезой.

Процедура обоснования сопоставления выдвинутой гипотезы с имеющимися данными (выборкой), называется статистической проверкой гипотезы.

Та гипотеза, относительно которой ведется проверка, называется основной или нулевой и обозначается $H_0: \theta = \theta_0$. Наряду с H_0 выдвигается конкурирующая или альтернативная гипотеза H_1, которая принимается в случае отклонения H_0. Например, $H_1: \theta > \theta_0$, либо $H_1: \theta < \theta_0$, либо $H_1: \theta \neq \theta_0$, либо $H_1: \theta = \theta_1$, где $\theta_0 \neq \theta_1$. Выбор альтернативной гипотезы определяется конкретной формулировкой задачи.

Гипотеза о параметрах случайной величины называется параметрической, все остальные – непараметрические. H_1 формулируется только для параметрических гипотез, во всех остальных случаях, H_1 состоит в отклонении H_0.

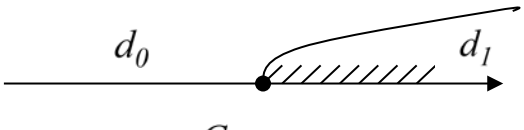
Если гипотеза однозначно определяет либо значение неизвестного параметра, либо вид предполагаемого закона распределения, она называется простой, в противном случае – сложной.

Пример. Имеется предположение, что средний расход топлива автомобиля составляет 10 литров на 100 км. Сформулируем основную гипотезу $H_0: a = 10$ л, при альтернативной гипотезе $H_1: a = 9$ л. – это простая двух альтернативная параметрическая гипотеза. Сформулируем основную гипотезу $H_0: a = 10$ л, при альтернативной гипотезе $H_1: a < 10$л. – это сложная двух альтернативная параметрическая гипотеза.

Пример. Непараметрическая простая гипотеза будет иметь вид: $H_0: X \sim N(0,1)$, H_1 состоит в отклонении H_0. Непараметрическая сложная гипотеза будет иметь вид: $H_0: X \sim N(a, 1), a_1 < a < a_2$, H_1 состоит в отклонении H_0.

Правило проверки статистической гипотезы называется решающим правилом (процедурой) или критерием проверки гипотезы. Критерием называется также специальным образом сконструированная величина, закон распределения которой известен и затабулирован.

Правило проверки статистической гипотезы основано на том, что множество значений этого критерия разбивается на две области d_0 и d_1 ,



d_0 –область, при попадании в которую H_0 принимается (область принятия нулевой гипотезы), d_1 –область, при попадании в которую H_0 отвергается и

принимается H_1 (область отклонения нулевой гипотезы или критическая область), C – критическая точка.

При проверке гипотезы можно прийти либо к правильному решению: принять H_0, если она верна, либо отвергнуть H_0, если она не верна; либо допустить одну из двух ошибок: ошибка 1-го рода – отвергнуть H_0, когда она на самом деле верна (ошибка типа пропуска цели), ошибка 2-го рода – принять H_0, когда она на самом деле не верна (ошибка типа ложной тревоги). Последствия этих ошибок различны, желательно провести проверку таким образом, чтобы минимизировать вероятности ошибок 1 и 2 рода. Однако одновременно уменьшить эти вероятности невозможно, поскольку уменьшение одной вероятности влечет за собой увеличений другой.

Оптимальный критерий был предложен Нейманом и Пирсоном: вероятность ошибки 1-го рода не должна превосходить некоторого заранее заданного числа α, $P(d_1/H_0) \le \alpha$, α – уровень значимости. При этом вероятность ошибки 2-го рода должна быть минимальной $P(d_0/H_1) = \beta \to \min$ или должна быть максимальной мощность критерия $1-\beta = P(d_1/H_1) = \omega \to \max$.

3.2. Общая логическая схема проверки гипотез

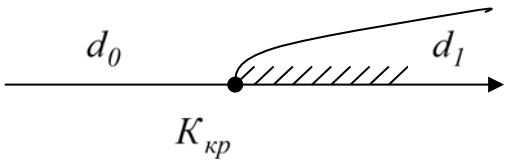
1. Выдвигается основная гипотеза H_0, если H_0 – параметрическая, то выдвигается альтернативная гипотеза H_1.

2. Задается уровень значимости α, значения α стандартные $\alpha = 0{,}1;\ 0{,}05;\ 0{,}01;\ 0{,}001$. Величина α определяет размер критической области, $\alpha = P(d_1/H_0)$.

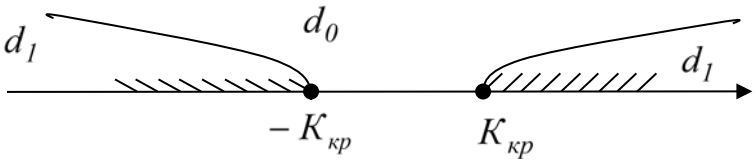
3. Выбирается критерий проверки гипотезы H_0, статистика критерия K является некоторой функцией выборки $\{x_1, x_2, \ldots, x_n\}$ случайной величины, закон распределения которой известен и за табулирован, чаще всего используется стандартный нормальный закон $N(0{,}1)$, распределение Стьюдента или t – распределение, распределение Пирсона или χ^2 – распределение, распределение Фишера или F – распределение.

4. Из таблиц распределения критерия при заданном уровне значимости α выбирается критическая точка $K_{кр}$, которая делит множество значений критерия на области принятия нулевой гипотезы d_0 и критическую область d_1. Размер критической области определяется уровнем значимости α, положение на оси определяется видом альтернативной гипотезы.

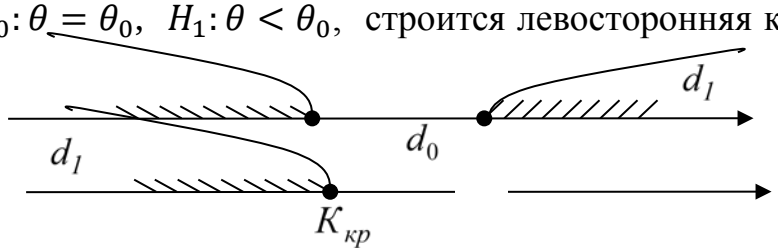
Например, 1) если $H_0: \theta = \theta_0$, $H_1: \theta > \theta_0$, либо $H_1: \theta = \theta_1$, где $\theta_0 \ne \theta_1$, строится правосторонняя критическая область,



2) если $H_0: \theta = \theta_0$, $H_1: \theta \neq \theta_0$, строится двухсторонняя критическая область.



3) если $H_0: \theta = \theta_0$, $H_1: \theta < \theta_0$, строится левосторонняя критическая область.



5. По заданной выборке определяется наблюдаемое значение критерия K_0, которое сравнивается с критической точкой $K_{кр}$. Если K_0 попадает в область d_0, то гипотеза H_0 принимается, если K_0 попадает в область d_1, то гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1, если $K_0 = K_{кр}$, то следует поменять уровень значимости α.

Замечание. В случае двухсторонней критической области, область d_0 совпадает с доверительным интервалом, который с надежностью $\gamma = 1 - \alpha$ покрывает истинное значение параметра θ. Гипотезу H_0 принимают на уровне значимости α, если доверительный интервал содержит гипотетическое значение параметра θ_0. В этом случае доверительный уровень и уровень значимости одно и тоже понятие.

Пример 3.1. Предположим, что введен новый способ управления производством при котором измеряется характеристика X – производительность труда одного сотрудника. Данные наблюдений представлены выборкой, в результате первичной статистической обработки которой получены следующие результаты: объем выборки $n = 100$, выборочное среднее $\bar{x} = 14{,}978$, несмещенная оценка дисперсии $S^2 = 0{,}928$, $S = 0{,}963$. Требуется проверить утверждение, что среднее значение и среднее квадратическое отклонение характеристики X остались неизменными.

Решение.

1. Необходимо проверить гипотезы $H_0: MX = a = 15$, $H_0: DX = \sigma^2 = 1$

Наряду с основными гипотезами нужно сформулировать альтернативные гипотезы. Пусть $H_1: MX = a \neq 15$, $H_1: DX = \sigma^2 < 1$.

Таким образом, проверяемые гипотезы имеет вид:

а) $H_0: a = 15$, $H_1: a \neq 15$ б) $H_0: \sigma^2 = 1$, $H_1: \sigma^2 < 1$.

2. Выбираем уровень значимости α, пусть $\alpha = 0{,}05$ для обоих гипотез.

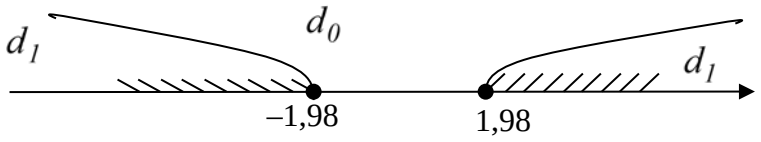
3. Выбираем критерий для проверки гипотез. Так как сравнение математического ожидания случайной величины с гипотетическим средним производится

при неизвестной дисперсии, оценке S^2, которая найдена по выборке, то следует использовать критерий Стьюдента ($t-$критерий), а для проверки гипотезы о дисперсии – критерий χ^2.

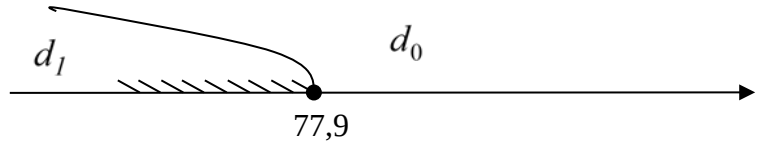
Итак, имеем а) $t=\dfrac{\overline{x}-a_0}{S}\sqrt{n}\sim t(n-1)$, б) $\chi^2=\dfrac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$.

4. Из таблиц распределения выбранных статистик по заданному уровню значимости $\alpha=0{,}05$ найдем критическую точку, которая разделит множество значений критерия на области принятия нулевой гипотезы d_0 и критическую область d_1. Для определения критических точек воспользуемся соответствующими таблицами квантилей.

а) $t_{1-\frac{\alpha}{2}}(n-1)=t_{0{,}975}(99)=1{,}98$, так как конкурирующая гипотеза определяет двухстороннюю критическую область



б) $\chi^2_\alpha(n-1)=\chi^2_{0{,}05}(99)=77{,}9$, так как конкурирующая гипотеза определяет левостороннюю критическую область



5. Найдем наблюдаемое значение критерия

а) $t_0=\dfrac{14{,}978-15}{0{,}963}\sqrt{100}=-0{,}228$, б) $\chi_0^2=\dfrac{(100-1)\cdot 0{,}928}{1}=91{,}872$.

6. Сравним наблюдаемое значение критерия с критической точкой

а) так как $|t_0|<t_{0{,}975}(99)$, то данные наблюдений не противоречат выдвинутой гипотезе, то есть гипотезу $H_0: a=15$ следует принять и считать, что $a=15$;

б) так как $\chi_0^2>\chi_0^2(99)$, то $H_0:\sigma^2=1$ следует принять, данные наблюдений согласуются с предположением о том, что дисперсия случайной величины равна 1.

Таким образом, обе выдвинутые гипотезы приняты. Можно считать, что новый способ управления с точки зрения увеличения производительности труда оказался неэффективным.

3.3. Проверка гипотезы о виде закона распределения

Пусть $\{x_1, x_2, \ldots, x_n\}$ – выборка наблюдений случайной величины *X*. Проверяется гипотеза $H_0 : F(x) = F_0(x)$, где $F_0(x)$ – гипотетически заданное распределение. Одним из наиболее мощных и простых в реализации является критерий согласия Пирсон или критерий χ^2.

В этом критерии за меру расхождения статистического и теоретического законов распределения принимается величина χ^2, выборочное значение которой определяется по формуле

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

где k – число интервалов группирования или различных вариантов, n – объем выборки. В случае, если X – дискретная случайная величина, то p_i – вероятность реализации значения x_i, вычисленная в предположении, что гипотеза $H_0 : F(x) = F_0(x)$ верна, то есть $p_i = P\{X = x_i / H_0\}$. Если X – непрерывная случайная величина, то p_i – вероятность попадания значения x_i в i – й интервал , вычисленная в предположении, что гипотеза $H_0 : F(x) = F_0(x)$ верна, то есть $p_i = P\{x_i < X < x_{i+1} / H_0\}, i = \overline{1,k}$. Очевидно, что $\sum_{i=1}^{k} p_i = 1$.

При $n \to \infty$ закон распределения статистики χ^2 независимо от вида закона распределения величины X стремится к закону $\chi^2(q), q = k - r - 1$, где r – число параметров предполагаемого закона распределения.

Процедура проверки гипотезы $H_0 : F(x) = F_0(x)$ с помощью критерия χ^2, состоит из следующих этапов:

1. По выборке $\{x_1, x_2, \ldots, x_n\}$ наблюдений случайной величины X найти оценки неизвестных параметров предполагаемого закона распределения.

2. Получить эмпирическое распределение случайной величины в виде точечного или интервального вариационных рядов.

3. Определить теоретические вероятности p_i в предположении, что $H_0 : F(x) = F_0(x)$ верна.

4. Вычислить наблюдаемое значение критерия по выборке χ_0^2.

5. Принять статистическое решение: гипотеза $H_0 : F(x) = F_0(x)$ не противоречит выборке при заданном уровне значимости α, если $\chi_0^2 < \chi_{1-\alpha}^2(k - r - 1)$, где $\chi_{1-\alpha}^2$ – квантиль уровня $1 - \alpha$ распределения χ^2 с числом степеней свободы $(k - r - 1)$. Если же $\chi_0^2 > \chi_{1-\alpha}^2(k - r - 1)$, то гипотеза $H_0 : F(x) = F_0(x)$ отклоняется.

По результатам первичной статистической обработки данных делается вывод о принадлежности или нет наблюдаемого распределения к нормальному. Чтобы убедиться в этом окончательно, воспользуемся критерием согласия χ^2:

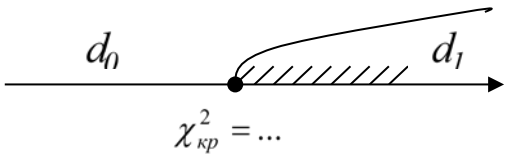
1. Выдвигаем гипотезу H_0 о том, что случайная величина X распределена по нормальному закону: H_0: $X \sim N(a,\sigma)$, где $\hat{a} = \bar{x}_в = ...$, $\hat{\sigma} = s = ...$

2. Пусть уровень значимости $\alpha = 0{,}05$.

3. Для проверки гипотезы используем критерий согласия χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2(k-3).$$

4. Из таблиц квантилей распределения χ^2 найдем критическую точку $\chi^2_{кр} = \chi^2_{0,95}(k-3) = ...$. Критическая область правосторонняя:



5. Для расчета наблюдаемого значения критерия χ^2_0 составим две вспомогательные таблицы (используем интервальный ряд и значения функции Лапласа). Расчет np_i, где

$$p_i = \Phi(Z_i) - \Phi(Z_{i-1}),\ Z_i = (C_i - \bar{x}_в)/s$$

представляется в виде таблицы (табл. 3.1). Расчет $\chi^2_{набл}$ – тоже в виде таблицы (табл. 3.2).

Таблица 3.1

i	C_{i-1}	C_i	Z_{i-1}	Z_i	$\Phi(Z_{i-1})$	$\Phi(Z_i)$	p_i	np_i
1								
2								
…								
k								
$\sum$	—	—	—	—	—	—	≈ 1	$\approx n$

Таблица 3.2

i	n_i	np_i	$(n_i - np_i)^2$	$(n_i - np_i)^2 / np_i$
1				
2				
…				
k				
$\sum$	n	$\approx n$	—	$\chi^2_{набл} =$

Сравниваем наблюдаемое значение критерия $\chi^2_{набл}$=… с критической точкой $\chi^2_{кр} = \ldots$. Если $\chi^2_{набл} < \chi^2_{кр}$, т. е. $\chi^2_{набл}$ принадлежит области принятия нулевой гипотезы, гипотезу о нормальном распределении следует принять. В противоположном случае – отвергнуть, т.е. наблюдаемое распределение не согласуется с нормальным.

3.4. Рекомендации по выполнению расчетно-графических работ по теме «Статистическая проверка гипотез» в MS Excel

Задание:

По результатам первичной статистической обработки данных сделать вывод о принадлежности наблюдаемого распределения к нормальному.

Для выполнения расчетов в MS Excel составляем таблицы 3.1 и 3.2 (рис. 3.1). Значения $\Phi(Z_i)$ и $\Phi(Z_{i-1})$ находим, используя встроенную статистическую функцию в MS Excel 2007: =НОРМСТРАСП(z)–0,5 (от полученного значения отнимаем 0,5), в MS Excel 2010-14: =ГАУСС(z). $\chi^2_{кр}$ находим также с использованием уже известной встроенной функции в MS Excel 2007: =ХИ2ОБР(вероятность; число степеней свободы), в MS Excel 2010-14: =ХИ2.ОБР.ПХ(вероятность; число степеней свободы) (рис. 3.1). Вероятность равна 0,05.

	C	D	E	F	G	H	I	J	K
88	Расчет n*p i	=(C90-$D$71)/$D$73			=ГАУСС(E90)				=I90*$D$10
89	C i-1	C i	Z i-1	Z i	Φ(Z i-1)	Φ(Z i)	p i =Φ(Z i)-Φ(Z i-1)	n*p i	
90	85	91,125	-2,37	-1,78	-0,4912	-0,4625	0,0287	2,869	
91	91,125	97,25	-1,78	-1,19	-0,4625	-0,3824	0,0801	8,011	
92	97,25	103,375	-1,19	-0,59	-0,3824	-0,2236	0,1588	15,881	
93	103,375	109,5	-0,59	0,00	-0,2236	0,0000	0,2236	22,360	
94	109,5	115,625	0,00	0,59	0,0000	0,2236	0,2236	22,360	
95	115,625	121,75	0,59	1,19	0,2236	0,3824	0,1588	15,881	
96	121,75	127,875	1,19	1,78	0,3824	0,4625	0,0801	8,011	
97	127,875	134	1,78	2,37	0,4625	0,4912	0,0287	2,869	
98						Σ	0,9824	98,241	
99	Расчет хи^2 набл.	=J90		=(C101-D101)^2					
100	ni	n* p i	(ni-n*pi)^2	(ni-n*pi)^2/n*pi	=E101/D101				
101	4	2,869	1,279	0,446					
102	8	8,011	0,000	0,000					
103	13	15,881	8,302	0,523					
104	27	22,360	21,533	0,963					
105	23	22,360	0,410	0,018					
106	13	15,881	8,302	0,523					
107	6	8,011	4,043	0,505		=ХИ2.ОБР.ПХ(D79;D15-2-1)			
108	6	2,869	9,803	3,417					
109			хи^2 набл.	6,394		хи^2 кр.=	11,070		

Рис. 3.1. Проверка гипотезы о виде распределения

3.5. Оформление результатов проведенных расчетов по теме «Статистическая проверка гипотез»

Пример 3.2. Оформление результатов проведенных расчетов по Теме 3.

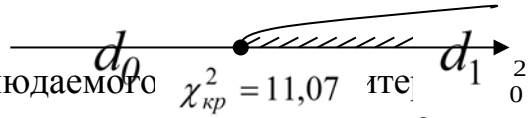
1.Выдвигаем гипотезу H_0 о том, что случайная величина X – число пассажиров одного авиарейса распределена по нормальному закону:

H_0: $X \sim N(a,\sigma)$, где $\tilde{a} = \bar{x}_в = 109{,}5$, $\tilde{\sigma} = s = 10{,}319$.

2. $\alpha = 0{,}05$ – ошибка 1 рода.

3. $\chi^2_{набл} = \sum_{i=1}^{8} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2(8-2-1)$.

4. Найдем из таблиц квантилей распределения χ^2 (или вычисляем в MS Excel) критическую точку $\chi^2_{кр} = \chi^2_{0,95}(8-2-1) = 11{,}07$. Критическая область правосторонняя:



Для расчета наблюдаемого критерия χ^2_0 составим две вспомогательные таблицы (используем интервальный ряд 2 и значения функции Лапласа).

Расчет $n \cdot p_i$

i	C_{i-1}	C_i	Z_{i-1}	Z_i	$\Phi(Z_{i-1})$	$\Phi(Z_i)$	p_i	np_i
1	85	91,125	-2,37	-1,78	-0,4912	-0,4625	0,0287	2,869
2	91,125	97,25	-1,78	-1,19	-0,4625	-0,3824	0,0801	8,0108
3	97,25	103,38	-1,19	-0,59	-0,3824	-0,2236	0,1588	15,881
4	103,375	109,5	-0,59	0,00	-0,2236	0,0000	0,2236	22,36
5	109,5	115,63	0,00	0,59	0,0000	0,2236	0,2236	22,36
6	115,625	121,75	0,59	1,19	0,2236	0,3824	0,1588	15,881
7	121,75	127,88	1,19	1,78	0,3824	0,4625	0,0801	8,0108
8	127,875	134	1,78	2,37	0,4625	0,4912	0,0287	2,869
$\sum$	—	—	—	—	—	—	**0,9824**	**98,24**

Расчет $\chi^2_{набл}$

i	n_i	np_i	$(n_i - np_i)^2$	$(n_i - np_i)^2 / np_i$
1	4	2,869	1,279	0,446
2	8	8,0108	0,000	0,000
3	13	15,881	8,302	0,523
4	27	22,36	21,533	0,963
5	23	22,36	0,410	0,018
6	13	15,881	8,302	0,523
7	6	8,0108	4,043	0,505
8	6	2,869	9,803	3,417
$\sum$	—	—	—	$\chi^2_{набл}$ = **6,394**

Сравниваем наблюдаемое значение критерия $\chi^2_{набл}$=6,394 с критической точкой $\chi^2_{кр} = 11{,}07$. Так как $\chi^2_{набл} < \chi^2_{кр}$, т. е. $\chi^2_{набл}$ принадлежит области принятия нулевой гипотезы, гипотезу о нормальном распределении числа пассажиров одного авиарейса следует принять.

4. Однофакторный дисперсионный анализ

Под дисперсионным анализом понимается статистический метод обработки результатов наблюдений, зависящих от различных одновременно действующих факторов. Его задача состоит в оценке влияния этих факторов и их взаимодействий в изменение некоторой выходной величины, предположительно от них зависящей. Оценка этого влияния производится с некоторой наперед заданной вероятностью, при этом формулируются определенные допущения о выходной величине и самих факторах, которые, как правило, качественной природы. Дисперсионный анализ может быть использован для выявления совместного влияния экономических факторов, не поддающихся количественному измерению, на изучаемый экономический показатель. В зависимости от числа факторов дисперсионный анализ может быть одно-, двух-, трех- и в общем случае многофакторным. Суть метода состоит в том, что общая суммарная дисперсия результирующего показателя разлагается на части, обусловленные действием отдельных факторов и их взаимодействием, и остаточную дисперсию, связанную с действием всех неучтенных в данном наблюдении факторов. Статистическое изучение частей позволяет делать выводы о том, оказывает ли влияние на результирующий показатель тот или иной качественный фактор, и если оказывает, то какова сила влияния каждого уровня этого фактора.

Продемонстрируем основную идею дисперсионного анализа на примере решения следующей однофакторной задачи.

4.1. Теоретические сведения о дисперсионном анализе

В этом случае исследуется влияние на результирующий показатель y одного качественного фактора A, уровням которого $a_1, a_2, .., a_m$ соответствует шкала наименований. Пусть над каждым уровнем фактора A производится серия из n наблюдений. Данные $i-$серии обозначим $y_{i1}, y_{i2}, \ldots, y_{in}$ $(i = 1,2,\ldots m)$. Тогда аддитивная модель однофакторного дисперсионного анализа может быть записана в виде

$$y_{ij} = \mu + a_{ij} + \varepsilon_{ij}, \quad i = 1,2,\ldots m;\ j = 1,2,\ldots,n,$$

где $y_{ij}-$результат измерения изучаемого признака с $i-$м уровнем фактора A в $j-$й серии наблюдений; $\mu-$общее (генеральное) среднее; $a_{ij}-$эффект $i-$го уровня фактора A в $j-$й серии наблюдений; $\varepsilon_{ij}-$случайные ошибки, распределенные по нормальному закону $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Смысл модели однофакторного дисперсионного анализ состоит в том, что она показывает из каких компонент состоит значение интересующего нас признака.

Всего мы располагаем $m \cdot n$ наблюдениями y_{ij}, где $i-$номер уровня качественного фактора A и $j-$номер производимого над ним наблюдения $(i = 1,2,\ldots m;\ j = 1,2,\ldots,n)$. данные наблюдений представляются в виде таблицы (табл. 4.1).

Таблица 4.1

Уровни фактора A	Номер испытания				Сумма по строкам	Внутригрупповые выборочные средние
	1	2	…	n		
a_1	y_{11}	y_{12}	…	y_{1n}	$\sum_{j=1}^{n} y_{1j}$	$\bar{y}_1 = \frac{1}{n}\sum_{j=1}^{n} y_{1j}$
a_2	y_{21}	y_{22}	…	y_{2n}	$\sum_{j=1}^{n} y_{2j}$	$\bar{y}_2 = \frac{1}{n}\sum_{j=1}^{n} y_{2j}$
…	…	…	…	…	…	…
a_m	y_{m1}	y_{m2}	…	y_{mn}	$\sum_{j=1}^{n} y_{mj}$	$\bar{y}_m = \frac{1}{n}\sum_{j=1}^{n} y_{mj}$
Общая выборочная средняя						$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} \bar{y}_i$

Тогда общая сумма квадратов $SS_{общ}$ величины y или полная сумма квадратов отклонений результирующего признака от общего среднего представима в виде суммы

$$SS_{общ} = SS_A + SS_R,$$

где $SS_A = n\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2$ – межгрупповая дисперсия или дисперсия, обусловленная фактором A;

$SS_R = \sum_{j=1}^{n}\sum_{i=1}^{m}(y_{ij} - \bar{y}_i)^2 = SS_{общ} - SS_A$ – внутригрупповая или остаточная дисперсия;

$SS_{общ} = \sum_{j=1}^{n}\sum_{i=1}^{m}(y_{ij} - \bar{y})^2$ – общая дисперсия.

Суммы квадратов, деленные на соответствующие числа степеней свободы $\nu_0 = mn - 1, \nu_A = m - 1, \nu_R = m(n-1)$, дадут три несмещенных оценки дисперсий:

$S_A^2 = MS_A = \frac{SS_A}{m-1}$ – несмещенная оценка дисперсии SS_A, обусловленной действием фактора A или оценка межгрупповой дисперсии;

$S_R^2 = MS_R = \frac{SS_R}{m(n-1)}$ – несмещенная оценка остаточной дисперсии SS_A или оценка внутригрупповой дисперсии,

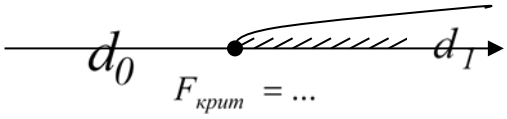
$S_0^2 = MS_{общ} = \frac{SS_{общ}}{mn-1}$ – несмещенная оценка общей дисперсии.

Для числа степеней свободы справедливо условие: $\nu_0 = \nu_A + \nu_R$.

Проверка значимости влияния фактора *A* соответствует проверке основной гипотезы H_0: $a_1 = a_2 = ... = a_m = 0$, где a_i – средний эффект i – го уровня фактора A. Проверка гипотезы о значимости влияния фактора *A* заключается в сравнении оценки S_A^2 и оценки остаточной дисперсии S_R^2. Если гипотеза H_0: $a_1 = a_2 = ... = a_m = 0$ верна, то оценки дисперсий S_A^2 и S_R^2 должны отличаться между собой лишь случайно, то есть незначимо.

Рассчитываем критическое значение статистики критерия Фишера $F_{крит} = F(\alpha;\nu_1;\nu_2)$, используя для этого соответствующее приложение учебника или встроенную в MS Excel функцию «=FРАСПОБР($\alpha;\nu_1;\nu_2$)».

Критическая область правосторонняя:



При этом критерий Фишера $F = \dfrac{S_A^2}{S_R^2}$, обычно применяемый для сравнения дисперсий, покажет их незначимость, если $F_0 < F_{кр}(\alpha, \nu_A, \nu_R)$, где $F_{кр}(\alpha, \nu_A, \nu_R)$ – критическая точка распределения Фишера, найденная по уровню значимости α и числам степеней свободы ν_A, ν_R.

Если же $F_0 > F_{кр}(\alpha, \nu_A, \nu_R)$, то F – критерий указывает на значимое расхождения между S_A^2 и S_R^2, то есть на недопустимость нулевой гипотезы. В таком случае мы имеем основание считать, что фактор *A* оказывает существенное влияние на исследуемый признак, то есть является значимым.

Для удобства результаты дисперсионного анализа вносят в таблицу (табл. 4.2).

Таблица 4.2

Однофакторный дисперсионный анализ

Источник изменчивости	Число степеней свободы	Сумма квадратов SS	Средний квадрат MS	Критерий Фишера F_0	Критическая точка $F_{кр}$	Гипотеза
Фактор *A*	$m-1$	SS_A	$S_A^2 = \dfrac{SS_A}{m-1}$	$F_0 = \dfrac{S_A^2}{S_R^2}$	$F_{кр}(\alpha, \nu_A, \nu_R)$	H_0: $a_1 = a_2 = ... = a_m = 0$
Ошибка	$m(n-1)$	SS_R	$S_R^2 = \dfrac{SS_R}{m(n-1)}$	–	–	–
Итог	$mn-1$	$SS_{общ}$	–	–	–	–

Если фактор *A* оказывается значимым, то следующим этапом процедуры дисперсионного анализа является проверка различий между его уровнями, то

есть процедура множественного сравнения с целью выделения наиболее информативных. Эта процедура ранжирования или упорядочения уровней значимого фактора по силе их влияния может осуществляться с использованием метода Тьюки. Метод состоит из следующих этапов:

1. Определяются все $C_m^2 = \frac{m(m-1)}{2}$ разностей между средними вида $\left|\overline{y}_i - \overline{y}_j\right|$.

2. Все разности нормируются, то есть делятся на $\sqrt{\frac{S_R^2}{n}}$.

3. По таблицам определяется критическая точка – это квантиль уровня $1-\alpha$ для распределения стьюдентизированного размаха с $\nu_1 = m$ и $\nu_2 = m(n-1)$ степенями свободы.

4. Нормированные разности сравниваются с критической точкой и делается вывод о наличии существенных отличий.

Все полученные данные сводятся в таблицу (табл. 4.3).

Таблица 4.3

модули разности $\left|\overline{y}_i - \overline{y}_j\right|$		S_R^2	$\sqrt{\frac{S_R^2}{n}}$	$\left|\overline{y}_i - \overline{y}_j\right| / \sqrt{\frac{S_R^2}{n}}$	Критическая точка	> существенное отличие
$\left|\overline{y}_1 - \overline{y}_2\right|$						
$\left|\overline{y}_1 - \overline{y}_3\right|$						
…						
$\left|\overline{y}_{m-1} - \overline{y}_m\right|$						

4.2. Рекомендации по выполнению расчетно-графической работы по теме «Однофакторный дисперсионный анализ» в MS Excel

Задание: при уровне значимости 0,05, установить значимость влияния фактора *A* методом однофакторного дисперсионного анализа. Дать интерпретацию фактору и его уровню, а также результирующему показателю в терминах экономических величин.

1. Найти или придумать самостоятельно выборку для выполнения расчетно-графической работы. Выборка включает в себя: качественный фактор А, 3-5 уровней фактора А, 4-5 наблюдений по каждому уровню фактора А.

2. Провести вычисления в Excel **двумя способами:** вручную и с помощью функции MS Excel Анализ данных → Однофакторный дисперсионный анализ.

3. Сделать выводы о значимости фактора А.

4. Рассчитать коэффициент детерминации и сделать выводы.

5. Применить метод Тьюки для выявления наиболее существенных уровней фактора А.

Пример 4.1. Известны итоговые результаты в баллах 20 студентов по дисциплине «Теория вероятностей и математическая статистика». Требуется установить, влияет ли базовое (среднее) образование на успеваемость студента.

Уровни фактора *A* – базовое (среднее) образование студента	Номер испытания				
	1	2	3	4	5
Гимназия или лицей	100	100	98	99	95
Школа с углубленным изучением предметов	100	97	99	91	89
Обычная школа	65	87	84	71	56
Техникум	41	68	56	47	71

Решение. Для более удобного вычисления в MS Excel составим вспомогательные таблицы (табл. 4.4–4.5).

Таблица 4.4

Таблица для вычисления средних

Уровни фактора *A*	Номер испытания				Сумма по строкам	Внутригрупповые выборочные средние
	1	2	…	n		
a_1	x_{11}	x_{12}	$\cdots$	x_{1n}	$\sum_{j=1}^{n} x_{1j}$	$\bar{x}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{1j}$
a_2	x_{21}	x_{22}	$\cdots$	x_{2n}	$\sum_{j=1}^{n} x_{2j}$	$\bar{x}_2 = \frac{1}{n}\sum_{j=1}^{n} x_{2j}$
…	…	…	…	…	…	…
a_m	x_{m1}	x_{m2}	$\cdots$	x_{mn}	$\sum_{j=1}^{n} x_{mj}$	$\bar{x}_m = \frac{1}{n}\sum_{j=1}^{n} x_{mj}$
Общая выборочная средняя						$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_i$

Таблица 4.5

Таблица для вычисления сумм квадратов (дисперсий)

SS_A	$(\bar{x}_i - \bar{x})^2$	$(\bar{x}_1 - \bar{x})^2$	$(\bar{x}_2 - \bar{x})^2$	…	$(\bar{x}_m - \bar{x})^2$	Σ	–
SS_R	$(x_{1j} - \bar{x}_1)^2$	$(x_{11} - \bar{x}_1)^2$	$(x_{12} - \bar{x}_1)^2$	…	$(x_{1n} - \bar{x}_1)^2$	Σ	Σ
	$(x_{2j} - \bar{x}_2)^2$	$(x_{21} - \bar{x}_2)^2$	$(x_{22} - \bar{x}_2)^2$	…	$(x_{22} - \bar{x}_2)^2$	Σ	
	…	…	…	…	…	Σ	
	$(x_{mj} - \bar{x}_m)^2$	$(x_{m1} - \bar{x}_m)^2$	$(x_{m2} - \bar{x}_m)^2$	…	$(x_{mn} - \bar{x}_m)^2$	Σ	
$SS_{общ}$	$(x_{1j} - \bar{x})^2$	$(x_{11} - \bar{x})^2$	$(x_{12} - \bar{x})^2$	…	$(x_{1n} - \bar{x})^2$	Σ	Σ
	$(x_{2j} - \bar{x})^2$	$(x_{21} - \bar{x})^2$	$(x_{22} - \bar{x})^2$	…	$(x_{2n} - \bar{x})^2$	Σ	
	…	…	…	…	…	Σ	
	$(x_{mj} - \bar{x})^2$	$(x_{m1} - \bar{x})^2$	$(x_{m2} - \bar{x})^2$	…	$(x_{mn} - \bar{x})^2$	Σ	

Заносим исходные данные в MS Excel, как в таблице для вычисления средних (рис. 4.1). Вычисляем суммы квадратов (дисперсий), не забывая максимально автоматизировать при этом все расчеты (рис. 4.2).

	A	B	C	D	E	F	G	H	I	J
1	уровни фактора А	номер испытания					сумма по строкам	групповые выборочные средние	=СУММ(B3:F3)	
2		1	2	3	4	5			=G3/5	
3	a1	100	100	98	99	95	492	**98,4**		
4	a2	100	97	99	91	89	476	**95,2**		
5	a3	65	87	84	71	56	363	**72,6**		
6	a4	41	68	56	47	71	283	**56,6**		
7				общая выборочная средняя				**80,7**	=СУММ(H3:H6)/4	
8										

Рис. 4.1. Расчет средних в MS Excel

	A	B	C	D	E	F	G	H	I	J	K
9					=(B3-$H3)^2	=СУММ(E11:E14)					
10											
11	SS A		SS R	(x1j - x cp1)^2	2,56	2,56	0,16	0,36	11,56		
12	(xi cp - x cp)^2	=(H3-$H$7)^2		(x2j - x cp2)^2	23,04	3,24	14,44	17,64	38,44		
13	313,29			(x3j - x cp3)^2	57,76	207,36	129,96	2,56	275,56	=СУММ(E15:I15)	
14	210,25			(x4j - x cp4)^2	243,36	129,96	0,36	92,16	207,36		
15	65,61		сумма по столбцам		326,72	343,12	144,92	112,72	532,92	1460,4	=SS R
16	580,81	=СУММ(A13:A16)	SS общ.	(x1j-x cp)^2	372,49	372,49	299,29	334,89	204,49		
17	**1169,96**			(x2j-x cp)^2	372,49	265,69	334,89	106,09	68,89		
18	**5849,8**	=SS A		(x3j-x cp)^2	246,49	39,69	10,89	94,09	610,09	=СУММ(E20:I20	
19				(x4j-x cp)^2	1576,09	161,29	610,09	1135,69	94,09		
20		=A17*5	сумма по столбцам		2567,56	839,16	1255,16	1670,76	977,56	7310,2	=SS общ
21											
22					=(B3-$H$7)^2	=СУММ(E16:E19)					
23	Проверка:	m*SS A + SS R = SS общ									
24	SS общ=	5*1169,96 + 1460,4 =	7310,2	=SS общ	Верно! ☼☺☼						

Рис. 4.2. Расчет сумм квадратов

Теперь, когда вычислены дисперсии и сделана проверка, приступаем к расчету значений критерия Фишера. Определяем критическую точку. Для этого находим число степеней свободы:

$$\nu_A = m-1 = 4-1 = 3,\ \nu_R = m(n-1) = 4(5-1) = 16,\ \nu_{общ} = mn-1 = 4\cdot 5-1 = 19.$$

Проверка: $\nu_{общ} = \nu_A + \nu_R = 3+16 = 19$.

Все расчеты заносим в таблицу в MS Excel (рис. 4.3).

	A	B	C	D	E	F	G	H	I
29									
30	Источник изменчивости	Число степеней свободы	Сумма квадратов	Средний квадрат	Критерий Фишера	F крит	Гипотеза		
31	Фактор А (между группами)	3	=A18 5849,80	=C31/B31 1949,93	=D31/D32 **21,36**	=FРАСПОБР(0,05;B31;B32) **3,239**	отклонить		
32	Остаток		=J15	=C32/B32					
33	(внутри групп)	16	1460,4	91,28	–	–	–		
34	Итог	19	7310,20	–	–	–	–		
35			=J20						
36									

Рис. 4.3. Расчет значений критерия Фишера

	A	B	C	D	E	F	G	H	I
37					=B40/$D$42				
38									
39	модули разности (МР)		MSост	(MSост/n)^0,5	МР/(MSост/n)^0,5	Крит.т.	> существенное отличие		
40	|хср1-хср2|	3,2			0,749		нет		
41	|хср1-хср3|	25,8		=D32	6,038		да		
42	|хср1-хср4|	41,8	91,28	4,273	9,783	3,52	да		
43	|хср2-хср3|	22,6			5,290		да		
44	|хср2-хср4|	38,6		=(C42/5)^0,5	9,034		да		
45	|хср3-хср4|	16			3,745		да		
46									
47									
48		За отклонение гипотезы ответственны практически все факторы							

Рис. 4.4. Расчет метода Тьюки

4.3. Однофакторный дисперсионный анализ с помощью Анализа данных

После проделанных расчетов можно выполнить проверку результата с помощью пакета анализа (пункт меню «данные», затем выбираем «пакет анализа») (рис. 4.5).

При введении данных в раздел пакет анализа «однофакторный дисперсионный анализ» необходимо обратить внимание на расположение исходных данных в строках или столбцах (рис. 4.6). Для расчетов входной интервал выделяем вместе с названиями уровней фактора, не забывая при этом поставить отметку в графе «метки в первом столбце». Когда все данные правильно введены, нажимаем кнопку «ОК» и переходим на тот лист, на который будут выведены результаты расчетов (рис. 4.7).

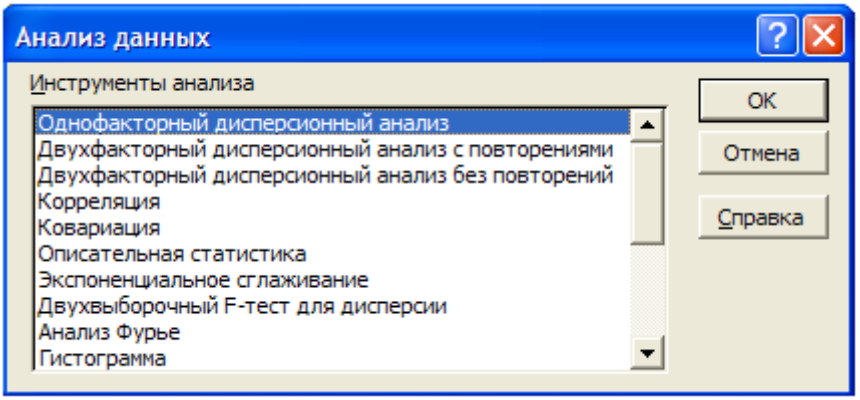


Рис. 4.5. Выбор однофакторного дисперсионного анализа

	A	B	C	D	E	F
1	уровни фактора А	номер испытания				
2		1	2	3	4	5
3	a1	100	100	98	99	95
4	a2	100	97	99	91	89
5	a3	65	87	84	71	56
6	a4	41	68	56	47	71
7						
8						
9						
10						
11	SS A					
12	(xi ср - x ср)^2					
13	313,29					
14	210,25					
15	65,61					
16	580,81					
17	**1169,96**					
18	**5849,8**	=SS A				
19						
20						
21						

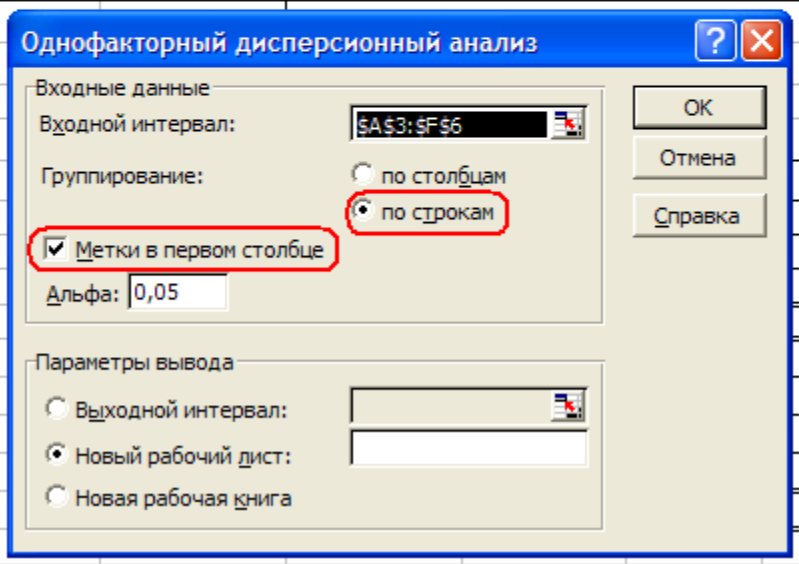


Рис. 4.6. Ввод данных для расчета однофакторного дисперсионного анализа

По рис. 4.7. видно, что все результаты, сделанные вручную, совпадают с результатами, рассчитанными автоматически. Таким образом, они верны.

	A	B	C	D	E	F	G
1	Однофакторный дисперсионный анализ						
2							
3	ИТОГИ						
4	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
5	a1	5	492	98,4	4,3		
6	a2	5	476	95,2	24,2		
7	a3	5	363	72,6	168,3		
8	a4	5	283	56,6	168,3		
9				80,7			
10							
11	Дисперсионный анализ						
12	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
13	Между группами	5849,8	3	1949,933333	**21,36327947**	0,0000077	**3,238866952**
14	Внутри групп	1460,4	16	91,275			
15							
16	Итого	7310,2	19				
17							

Рис. 4.7. Результаты расчетов значений критерия Фишера с помощью пакета анализа

4.4. Оформление полученных результатов

1. Выдвигаем гипотезу H_0 о том, что фактор A – незначим, т.е. базовое (среднее) образование студентов не оказывает существенного влияния на успеваемость студента по дисциплине «теория вероятностей и математическая статистика», т.е. H_0: $a_1 = a_2 = a_3 = a_4$.

Альтернативная гипотеза H_1 о том, что фактор A – значим, т.е. оказывает существенное влияние на результирующий признак, т.е. H_1: $a_1 \neq a_2 \neq a_3 \neq a_4$.

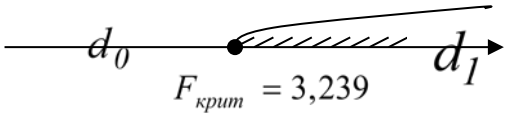
2. Задаем уровень значимости $\alpha = 0{,}05$.

3. Рассчитываем критическое значение статистики критерия Фишера $F_{крит} = F(\alpha; k_1; k_2)$, используя для этого соответствующее приложение учебника или встроенную в MS Excel функцию «=FРАСПОБР($\alpha; k_1; k_2$)».

$k_1 = \nu_A = m - 1 = 4 - 1 = 3$, $k_2 = \nu_R = m(n-1) = 4(5-1) = 16$.

$F_{крит} = F(0{,}05; 3; 16) = 3{,}239$

Критическая область правосторонняя:



4. Вычисляем наблюдаемое значение статистики критерия Фишера:

$$F_0 = \frac{S_A^2}{S_R^2}$$

где $S_A^2 = \dfrac{SS_A}{m-1}$ – несмещенная оценка дисперсии SS_A, обусловленной действием фактора A; $S_R^2 = \dfrac{SS_R}{m(n-1)}$ – несмещенная оценка остаточной дисперсии SS_A.

Проведем все расчеты с использованием вспомогательных таблиц.

Таблица 1. Расчет выборочных средних

Уровни фактора A	Номер испытания					Сумма по строкам	Внутригрупповые выборочные средние
	1	2	3	4	5		
a_1	100	100	98	99	95	$\sum_{j=1}^{n} x_{1j} = 492$	$\bar{x}_1 = \frac{492}{5} = 98{,}4$
a_2	100	97	99	91	89	$\sum_{j=1}^{n} x_{2j} = 476$	$\bar{x}_2 = \frac{476}{5} = 95{,}2$
a_3	65	87	84	71	56	$\sum_{j=1}^{n} x_{3j} = 363$	$\bar{x}_3 = \frac{363}{5} = 72{,}6$
a_4	41	68	56	47	71	$\sum_{j=1}^{n} x_{4j} = 283$	$\bar{x}_4 = \frac{283}{5} = 56{,}6$

Общая выборочная средняя	$\bar{x} = \frac{1}{4}(98{,}4 + 95{,}2 + 72{,}6 + 56{,}6) = 80{,}7$

Таблица 2. Расчет дисперсий

SS_A		$(x_{1j} - \bar{x}_1)^2$	2,56	2,56	0,16	0,36	11,56	
$(\bar{x}_i - \bar{x})^2$	SS_R	$(x_{2j} - \bar{x}_2)^2$	23,04	3,24	14,44	17,64	38,44	
313,29		$(x_{3j} - \bar{x}_3)^2$	57,76	207,36	129,96	2,56	275,56	
210,25		$(x_{4j} - \bar{x}_4)^2$	243,36	129,96	0,36	92,16	207,36	Итог:
65,61	сумма по столбцам		326,72+	343,12+	144,92+	112,72+	532,92=	**1460,4**
580,81		$(x_{1j} - \bar{x})^2$	372,49	372,49	299,29	334,89	204,49	
1169,96	$SS_{общ}$	$(x_{2j} - \bar{x})^2$	372,49	265,69	334,89	106,09	68,89	
5849,8		$(x_{3j} - \bar{x})^2$	246,49	39,69	10,89	94,09	610,09	
		$(x_{4j} - \bar{x})^2$	1576,09	161,29	610,09	1135,69	94,09	
	сумма по столбцам		2567,56+	839,16+	1255,16+	1670,76+	977,56=	**7310,2**

Дисперсии SS_A и SS_R находим по формулам:

$SS_A = n\sum_{i=1}^{m}(\bar{x}_i - \bar{x})^2 = 5 \cdot (313{,}29 + 210{,}25 + 65{,}61 + 580{,}81) = 5849{,}8$ – межгрупповая дисперсия или дисперсия, обусловленная фактором *A*;

$SS_R = \sum_{j=1}^{n}\sum_{i=1}^{m}(x_{ij} - \bar{x}_i)^2 = 2{,}56 + 2{,}56 + ... + 92{,}12 + 207{,}36 = 1460{,}4$ – внутригрупповая или остаточная дисперсия;

$SS_{общ} = \sum_{j=1}^{n}\sum_{i=1}^{m}(x_{ij} - \bar{x})^2 = 372{,}49 + 372{,}49 + ... + 1135{,}69 + 94{,}09 = 7310{,}2$ – общая дисперсия;

Проверка вычислений:

$SS_{общ} = SS_A + SS_R = 5849{,}8 + 1460{,}2 = 7310{,}2$ – верно.

Таблица 3. Расчет значений критерия Фишера

Источник изменчивости	Число степеней свободы	Сумма квадратов SS	Средний квадрат MS	Критерий Фишера F_0	$F_{крит}$	Гипотеза
Фактор *A* (между группами)	$m-1=3$	5849,8	$MS_A = S_A^2 =$ =1949,93	$F_0 = \frac{S_A^2}{S_R^2}$ =21,36	3,329	отклонить
Остаток (внутри групп)	$m(n-1)=$ 16	1460,2	$MS_R = S_R^2 =$ =91,28	–	–	–
Итог	$mn-1=19$	7310,2	–	–	–	–

Таким образом, $F_0 > F_{крит}$ (21,36 > 3,329), следовательно, с вероятностью 0,95 гипотеза H_0 отклоняется, т.е. базовое (среднее) образование студента оказывает существенное влияние на его успеваемость по дисциплине «теория вероятностей и математическая статистика».

Так как фактор *A* значим, то находим коэффициент детерминации:

$$R^2 = \frac{SS_A}{SS_{общ}} \cdot 100\% = \frac{5849{,}8}{7310{,}2} \cdot 100\% = 80{,}02\%.$$

Коэффициент детерминации показывает, что уровни фактора *A* объясняют вариацию результирующего показателя на 80,02%, т.е. базовое образование объясняет изменчивость успеваемости студента на 80,02%, оставшиеся 19,98% приходятся на другие факторы.

Метод Тьюки

модули разности $\lvert\bar{x}_i - \bar{x}_j\rvert$		MS_R	$\sqrt{\frac{MS_R}{n}}$	$\lvert\bar{x}_i - \bar{x}_j\rvert / \sqrt{\frac{MS_R}{n}}$	Критическая точка	> существенное отличие
$\lvert\bar{x}_1 - \bar{x}_2\rvert$	3,2	91,28	4,273	0,749	3,12	нет
$\lvert\bar{x}_1 - \bar{x}_3\rvert$	25,8			6,038		да
$\lvert\bar{x}_1 - \bar{x}_4\rvert$	41,8			9,783		да
$\lvert\bar{x}_2 - \bar{x}_3\rvert$	22,6			5,290		да
$\lvert\bar{x}_2 - \bar{x}_4\rvert$	38,6			9,034		да
$\lvert\bar{x}_3 - \bar{x}_4\rvert$	16			3,745		да

Между всеми групповыми средними значениями есть существенное отличие, кроме $\bar{x}_1$ и $\bar{x}_2$ (между гимназией/лицеем и школой с углубленным изучением предметов).

Критическую точку находим по таблице квантилей стьюдентизированного размаха (см. ниже) при $\alpha = 0{,}05$, $k = m - 1 = 3$, $\nu = \nu_R = m(n-1) = 16$.

Таблица VI. $\alpha = 0{,}10$

ν \ k	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	8,93	13,4	16,4	18,5	20,2	21,5	22,6	23,6	24,5	25,2	25,9	26,5	27,1	27,6	28,1	28,5	29,0	29,3	29,7
2	4,13	5,73	6,77	7,54	8,14	8,63	9,05	9,41	9,72	10,0	10,3	10,5	10,7	10,9	11,1	11,2	11,4	11,5	11,7
3	3,33	4,47	5,20	5,74	6,16	6,51	6,81	7,06	7,29	7,49	7,67	7,83	7,98	8,12	8,25	8,37	8,48	8,58	8,68
4	3,01	3,96	4,59	5,03	5,39	5,68	5,93	6,14	6,33	6,49	6,65	6,78	6,91	7,02	7,13	7,23	7,33	7,41	7,50
5	2,85	3,72	4,26	4,66	4,98	5,24	5,46	5,65	5,82	5,97	6,10	6,22	6,34	6,44	6,54	6,63	6,71	6,79	6,86
6	2,75	3,56	4,07	4,44	4,73	4,97	5,17	5,34	5,50	5,64	5,76	5,87	5,98	6,07	6,16	6,25	6,32	6,40	6,47
7	2,68	3,45	3,93	4,28	4,55	4,78	4,97	5,14	5,28	5,41	5,53	5,64	5,74	5,83	5,91	5,99	6,06	6,13	6,19
8	2,63	3,37	3,83	4,17	4,43	4,65	4,83	4,99	5,13	5,25	5,36	5,46	5,56	5,64	5,72	5,80	5,87	5,93	6,00
9	2,59	3,32	3,76	4,08	4,34	4,54	4,72	4,87	5,01	5,13	5,23	5,33	5,42	5,51	5,58	5,66	5,72	5,79	5,85
10	2,56	3,27	3,70	4,02	4,26	4,47	4,64	4,78	4,91	5,03	5,13	5,23	5,32	5,40	5,47	5,54	5,61	5,67	5,73
11	2,54	3,23	3,66	3,96	4,20	4,40	4,57	4,71	4,84	4,95	5,05	5,15	5,23	5,30	5,38	5,45	5,51	5,57	5,63
12	2,52	3,20	3,62	3,92	4,16	4,35	4,51	4,65	4,78	4,89	4,99	5,08	5,16	5,24	5,31	5,37	5,44	5,49	5,55
13	2,50	3,18	3,59	3,88	4,12	4,30	4,46	4,60	4,72	4,83	4,93	5,02	5,10	5,18	5,25	5,31	5,37	5,43	5,48
14	2,49	3,16	3,56	3,85	4,08	4,27	4,42	4,56	4,68	4,79	4,88	4,97	5,05	5,12	5,19	5,26	5,32	5,37	5,43
15	2,48	3,14	3,54	3,83	4,05	4,23	4,39	4,52	4,64	4,75	4,84	4,93	5,01	5,08	5,15	5,21	5,27	5,32	5,38
16	2,47	3,12	3,52	3,80	4,03	4,21	4,36	4,49	4,61	4,71	4,81	4,89	4,97	5,04	5,11	5,17	5,23	5,28	5,33
17	2,46	3,11	3,50	3,78	4,00	4,18	4,33	4,46	4,58	4,68	4,77	4,86	4,93	5,01	5,07	5,13	5,19	5,24	5,30
18	2,45	3,10	3,49	3,77	3,98	4,16	4,31	4,44	4,55	4,65	4,75	4,83	4,90	4,98	5,04	5,10	5,16	5,21	5,26
19	2,45	3,09	3,47	3,75	3,97	4,14	4,29	4,42	4,53	4,63	4,72	4,80	4,88	4,95	5,01	5,07	5,13	5,18	5,23
20	2,44	3,08	3,46	3,74	3,95	4,12	4,27	4,40	4,51	4,61	4,70	4,78	4,85	4,92	4,99	5,05	5,10	5,16	5,20
24	2,42	3,05	3,42	3,69	3,90	4,07	4,21	4,34	4,44	4,54	4,63	4,71	4,78	4,85	4,91	4,97	5,02	5,07	5,12
30	2,40	3,02	3,39	3,65	3,85	4,02	4,16	4,28	4,38	4,47	4,56	4,64	4,71	4,77	4,83	4,89	4,94	4,99	5,03
40	2,38	2,99	3,35	3,60	3,80	3,96	4,10	4,21	4,32	4,41	4,49	4,56	4,63	4,69	4,75	4,81	4,86	4,90	4,95
60	2,36	2,96	3,31	3,56	3,75	3,91	4,04	4,16	4,25	4,34	4,42	4,49	4,56	4,62	4,67	4,73	4,78	4,82	4,86
120	2,34	2,93	3,28	3,52	3,71	3,86	3,99	4,10	4,19	4,28	4,35	4,42	4,48	4,54	4,60	4,65	4,69	4,74	4,78
∞	2,33	2,90	3,24	3,48	3,66	3,81	3,93	4,04	4,13	4,21	4,28	4,35	4,41	4,47	4,52	4,57	4,61	4,65	4,69

Стьюдентизированный размах

Раздел 2. Основы эконометрики

Введение

Эконометрика – это наука, в которой с помощью статистических методов устанавливаются количественные взаимосвязи между экономическими переменными. То есть эконометрика – набор математико-статистических средств, позволяющих верифицировать модельные соотношения между анализируемыми экономическими показателями и оценивать неизвестные значения параметров в этих соотношениях на основе исходных экономических данных.

Задача эконометрического исследования – это оценка и проверка эконометрической модели. Эконометрическая модель, как правило, основана на теоретических предположениях о наборе взаимосвязанных переменных и характере связи между ними. При стремлении к «наилучшему» описанию связей приоритет отдается качественному анализу. Поэтому в качестве этапов эконометрического исследования можно указать:

- Постановка проблемы и спецификация модели;
- Сбор и анализ качества данных об экономике или ее секторе в зависимости от объекта моделирования и целей исследования;
- Оценивание неизвестных параметров модели, проверка разнообразных гипотез и верификация модели;
- Интерпретация параметров модели;
- Прогнозирование на основе построенной и верифицированной модели.

Эконометрическое исследование также включает решение проблем, связанных с анализом мультиколлинеарности факторов, оценку ее статистической значимости, введение фиктивных переменных, выявление автокорреляции, тренда и специальные вопросы, связанные с построением систем одновременных переменных.

5. Двумерная регрессионная модель

На практике во время проведения экономических исследований при анализу двух переменных, из которых одна не является случайной, пытаются определить кривую, дающую наилучшее приближение к исходным данным. Метод такого приближения получил называется регрессионным анализом. Модели и методы регрессионного анализа занимают центральное место в эконометрике.

Задачами регрессионного анализа являются установление формы зависимости между переменными, оценка функции регрессии, построение прогноза зависимой переменной.

В экономике в большинстве случаев между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует множество значений другой переменной. Такая зависимость называется статистической или стохастической (вероятностной).

Построение регрессионной модели включает в себя 6 этапов:

1. Спецификация модели.
2. Оценка неизвестных параметров модели.
3. Оценка значимости полученных коэффициентов регрессии.
4. Верификация модели.
5. Экономическая интерпретация полученных коэффициентов регрессии.
6. Построение точечного и интервального прогнозов.

Рассмотрим каждый из этапов построения модели подробно.

5.1. Спецификация модели парной линейной регрессии (1 этап)

Рассмотрим простой случай, когда экономическая модель состоит из одного уравнения, которое содержит только две переменные. Обозначив переменные через y и x, мы предположим между ними зависимость $y = f(x)$. На первом шаге мы лишь **идентифицировали** переменную x, как оказывающую воздействие на другую переменную y. Второй шаг состоит в **спецификации** формы связи между y и x – выборе формы уравнения и набора соответствующих переменных. Содержательные соображения или положения экономической теории могут привести к конкретному виду этой связи, однако простейшим соотношением является линейная как по независимой или объясняющей переменной x, так и по неизвестным параметрам a и b модель

$$y = a + bx. \qquad (5.1)$$

Возможны и другие формы связи между переменными x и y:

$$y = ae^{bx}, \qquad y = ax^b, \qquad y = a + b \cdot \frac{1}{x}.$$

Третье из этих соотношений линейно относительно a и b (линейно относительно y и $\frac{1}{x}$), а первое и второе могут быть сведены к линейной форме для преобразованных переменных, если взять логарифмы от обеих частей

$$\ln y = \ln a + bx \quad \text{и} \quad \ln y = \ln a + b \ln x.$$

Если ввести $y' = \ln y$ и $x' = \ln x$, то мы получим линейную зависимость вида (5.1). Подробнее вопрос о построении таких моделей мы рассмотрим в п. 5.7.

Таким образом, в модели (5.1) a и b – постоянные, а x и y могут непосредственно или после логарифмических или иных преобразований представлять экономические переменные, например такие, как цены или спрос. Очевидно, что при таком подходе охватывается широкая область функциональных взаимосвязей между исходными экономическим переменными.

Задача построения модели (5.1) состоит в определении значений неизвестных параметров a и b – их оценок – по имеющимся в нашем распоряжении данным так, чтобы полученное соотношение «наилучшим» образом описывало зависимость y от x. Имея набор значений двух переменных x_i, y_i, $i = 1, \ldots, n$; и изображая пары (x_i, y_i) точками на координатной плоскости $X\,\mathrm{O}\,Y$ (рис. 5.1), мы получаем разброс этих точек относительно реальной линии связи – **поле корреляции** или **диаграмму рассеяния**.

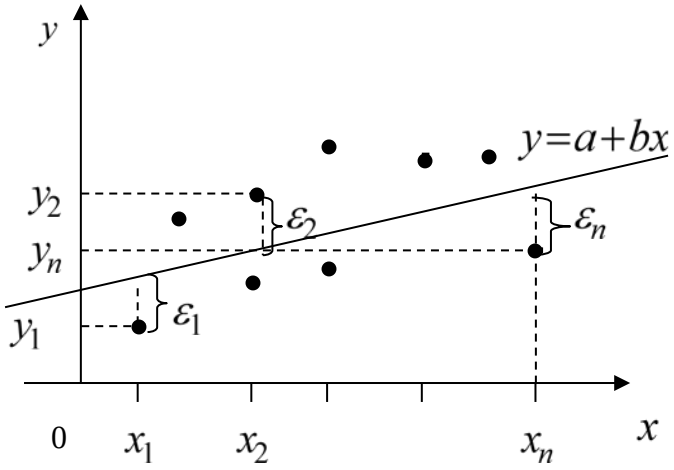


Рис. 5.1. Диаграмма рассеяния и теоретическая линия связи

Эта идея находит свое формальное воплощение в новой гипотезе о характере линейной зависимости:

$$y = a + bx + \varepsilon, \tag{5.2}$$

где ε – случайная (или стохастическая) переменная, способная принимать и положительные, и отрицательные значения.

Запишем уравнение зависимости (5.2) для n наблюдений x_i, y_i:

$$y_i = a + bx_i + \varepsilon_i \qquad i = 1, \ldots, n. \tag{5.3}$$

Здесь x_i – неслучайная (детерминированная) величина, а y_i, ε_i – случайные величины; y_i – объясняемая (зависимая) переменная, x_i – объясняющая (независимая) переменная, **фактор** или **регрессор**. Уравнение (5.3) называется также регрессионным уравнением или линейной регрессионной моделью с двумя переменными (**моделью парной линейной регрессии**).

Какова природа случайной составляющей или **ошибки** ε_i? Источниками ошибок могут быть разные причины:

1. **Пропущенные объясняющие переменные**. Линейное соотношение между y и x является простым описанием. В действительности существуют другие факторы, также оказывающие влияние на y, но не учтенные в модели (5.1). Влияние этих факторов приводит к тому, что наблюдаемые точки лежат вне прямой (см. рис. 5.1). Не все факторы можно включить в модель по разным причинам, одной из которых является невозможность измерить переменную, другой – слабое влияние фактора, которое не стоит учитывать. Объединив все эти составляющие, мы и получаем то, что обозначено через ε.

2. **Агрегирование переменных**. Часто рассматриваемая зависимость – это объединённый из некоторого числа микроэкономических соотношений экономический показатель, например, такой как, валовой внутренний продукт – это денежная стоимость всех готовых товаров и услуг, произведенных в стране в течение определенного периода времени.

3. **Неправильное описание структуры модели**. Структура модели может быть описана неправильно или не совсем корректно. Например, зависимость некоторого показателя y от времени t. Если ожидаемое и фактическое значения тесно связаны, то будет казаться, что между y и t существует зависимость, но это будет лишь аппроксимация, и расхождение вновь будет связано с наличием случайной величины ε.

4. **Неправильная функциональная спецификация**. Функциональное соотношение между y и x математически может быть определено неправильно, т. е. сам вид функциональной зависимости выбран неверно. Например, мы рассматриваем зависимость между курсом доллара и стоимостью барреля нефти, используя линейную функцию, а истинная зависимость может быть более сложной, нелинейной.

5. **Ошибки измерения**. Ошибки могут сопровождать любые наблюдения или измерения экономических показателей. Например, данные по расходам семьи на питание составляются на основании записей участников опросов, которые, как предполагается, тщательно фиксируют свои ежедневные расходы. Разумеется, при этом возможны ошибки. В данном случае источниками ошибок являются особенности собранного материала (присущ элемент случайности).

Таким образом, можно считать, что случайные величины ε_i являются суммарным проявлением всех этих факторов.

Сформулируем теперь те основные предпосылки или гипотезы, которые лежат в основе линейной регрессионной модели с двумя переменными.

Основные гипотезы:

1. $y_i = a + bx_i + \varepsilon_i, i = 1,\ldots,n,\ n > 2$, – спецификация модели.

2. $x_1,\ldots,x_n$ – детерминированные величины, линейно не связанные между собой, т.е. вектор $(x_1,\ldots,x_n)^T$ не коллинеарен вектору $(1,\ldots,1)^T$.

3. $\varepsilon_1,\ldots,\varepsilon_n$ – случайные величины, для которых

3a. $M\varepsilon_i = 0,\ M\left(\varepsilon_i^2\right) = D(\varepsilon_i) = \sigma^2$ – не зависит от i.

3b. $M\left(\varepsilon_i \varepsilon_j\right) = 0$ при $i \neq j$, т.е. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ – некоррелированны для разных наблюдений.

3c. $\varepsilon_i \sim N\left(0, \sigma^2\right)$, т. е. ε_i – нормально распределенные случайные величины с математическим ожиданием или средним значением, равным нулю, и дисперсией σ^2.

Гипотезы 1–3c определяют нормальную линейную модель парной регрессии. Для такой модели условие 3b. эквивалентно условию статистической независимости ошибок $\varepsilon_i, \varepsilon_j$ при $i \neq j$.

Обсудим предпосылки, лежащие в основе построения линейной модели.

1. Спецификация модели отражает наше представление о механизме зависимости y_i от x_i и сам выбор объясняющей переменной x; на линейный характер связи может указывать и разброс точек на диаграмме рассеивания.

2. Величины $x_1, \ldots, x_n$ являются неслучайными или детерминированными, линейно не связанными между собой. Если же в реальной ситуации их значения также представляются результатами измерений, то предполагается, что ошибки таких измерений пренебрежимо малы.

3a. Условие $M(\varepsilon_i) = 0$ означает отсутствие систематических ошибок, ошибки носят только случайный характер. Условие независимости дисперсий ошибок от номера наблюдений $M\left(\varepsilon_i^2\right) = D(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$, или однородности наблюдений называется также **гомоскедастичностью**; случай, когда $M\left(\varepsilon_i^2\right) = \sigma_i^2$, т. е. условие гомоскедастичности не выполняется. называется **гетероскедастичностью**. Ниже на рис. 5.2 приведен пример типичного разброса точек для случая гомоскедастичности ошибок; на рис. 5.3 – пример данных с гетероскедастичными ошибками.

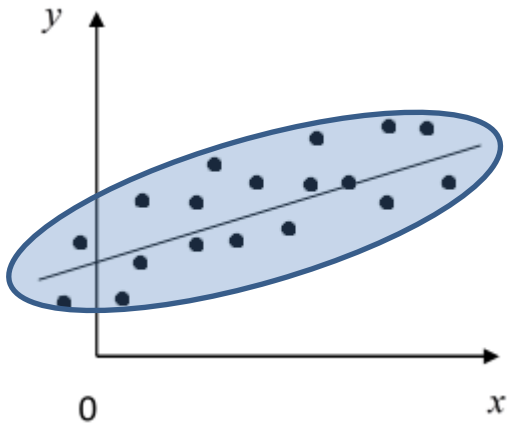


Рис. 5.2. Однородные наблюдения ($M\varepsilon_i^2 = \sigma^2$, $i = 1, \ldots, n$)

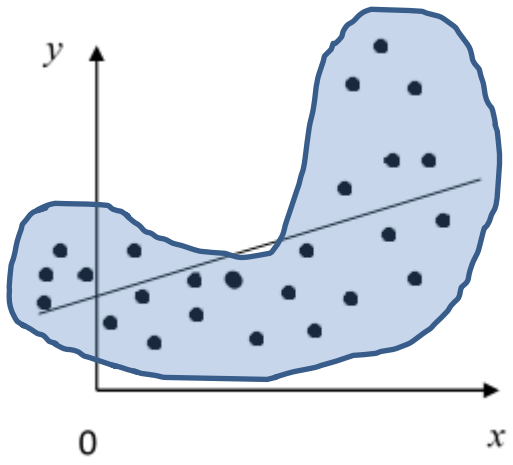


Рис. 5.3. Неоднородные наблюдения ($M\varepsilon_i^2 = \sigma_i^2, i = 1,\ldots,n$)

3b. Условие $M(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$, указывает на некоррелированность ошибок, а в случае нормальной модели и на независимость для разных наблюдений. Это требование оказывается вполне естественным в широком классе реальных ситуаций, особенно, если речь идет о пространственных данных (значения анализируемых переменных регистрируются на различных объектах: индивидуумах, семьях, предприятиях, банках, регионах и т. п.). Однако условие часто нарушается, когда наши данные являются временными рядами. В случае, когда это условие не выполняется, говорят об **автокорреляции ошибок**.

3c. Так как можно считать, что случайная составляющая ε_i в различных наблюдениях обусловлена суммарным аддитивным эффектом большого числа независимых случайных факторов, ни один из которых не является доминирующим, то обращение к центральной предельной теореме служит достаточным обоснованием выбора нормального распределения для нее.

5.2. Оценивание неизвестных параметров модели: метод наименьших квадратов (2этап)

Рассмотрим задачу «наилучшей» аппроксимации набора наблюдений (x_i, y_i), $i = 1,\ldots,n$, линейной функцией $y = a + bx$ в смысле минимизации величины

$$R = \sum_{i=1}^{n} \left(y_i - a - bx_i \right)^2 . \tag{5.4}$$

Нахождение оценок $\hat{a}$ и $\hat{b}$ в соответствии с этим условием называется методом наименьших квадратов (МНК). Запишем необходимые условия экстремума:

$$\begin{cases} \dfrac{\partial R}{\partial a} = -2\sum\limits_{i=1}^{n} \left(y_i - a - bx_i \right) = 0 \\ \dfrac{\partial R}{\partial b} = -2\sum\limits_{i=1}^{n} \left(y_i - a - bx_i \right) x_i = 0. \end{cases}$$

Решение этой **системы нормальных уравнений** дает нам явный вид оценок

$$\hat{a} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$\hat{b} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad (5.5)$$

(для краткости индексы суммирования у знака суммы $\sum$ опущены).

Если $\hat{b}$ найдено по формуле (5.5), то $\hat{a} = \bar{y} - \hat{b}\bar{x}$, где $\bar{x} = \frac{1}{n}\sum x_i$, $\bar{y} = \frac{1}{n}\sum y_i$.

Уравнение прямой линии $y = \hat{a} + \hat{b}x$, полученное в результате минимизации величины (5.4), проходит через точку $(\bar{x}, \bar{y})$. Единственность МНК-оценок (5.5) обеспечивается предпосылкой 2.

Из общей теории МНК при сделанных выше предпосылках 3a, 3b следуют свойства МНК-оценок (подробнее эти свойства мы обсудим в разделе 6.2): 1) линейная зависимость от *y*, 2) несмещенность, 3) эффективность, поскольку в классе линейных несмещенных оценок МНК-оценки обладают наименьшей возможной дисперсией (теорема Гаусса-Маркова).

Несмещенные оценки дисперсий и ковариаций оценок $\hat{a}$ и $\hat{b}$ определяются по формулам

$$\hat{D}(\hat{a}) = \frac{\sum x_i^2}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{R_{\min}}{n-2}; \qquad (5.6)$$

$$\hat{D}(\hat{b}) = \frac{n}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{R_{\min}}{n-2}; \qquad (5.7)$$

$$\operatorname{cov}(\hat{a}, \hat{b}) = \frac{-\sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{R_{\min}}{n-2}, \qquad (5.8)$$

где $R_{\min} = \sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{b}x_i\right)^2$ – остаточная сумма квадратов и под $\hat{a}$ и $\hat{b}$ понимаются их значения, найденные по формулам (5.5).

Несмещенной оценкой дисперсии ошибок наблюдений будет $S^2 = \hat{\sigma}^2 = \frac{R_{\min}}{n-2}$.

Остатки регрессии e_i определяются из уравнения

$$y_i = \hat{y}_i + e_i = \hat{a} + \hat{b}x_i + e_i. \qquad (5.9)$$

Не следует путать остатки регрессии с ошибками регрессии в уравнении модели $y_i = a + bx_i + \varepsilon_i$. Разница состоит в том, что **остатки** e_i в отличие от **ошибок** ε_i вычисляются. С учетом введенного обозначения для остатков можно записать несмещенную оценку дисперсии σ^2: $S^2 = \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$.

Пример использования формул (5.5) – (5.7) мы рассмотрим ниже в п. 5.7 с тем, чтобы проиллюстрировать все этапы построения и анализа линейной модели и задачу прогнозирования на ее основе.

5.3. Оценка значимости коэффициентов регрессии (3 этап)

Этап 3а. Построение доверительных интервалов. С помощью формул (5.5) мы можем получить по данным наблюдений над величинами *x, y* лишь **оценки** неизвестных параметров линейной модели. Поэтому возникает вопрос о точности и надежности найденных оценок. В математической статистике он решается построением **доверительных интервалов** для истинных значений параметров, которые по сути представляют собой множество всех возможных гипотетических значений, не противоречащих результатам экспериментов.

Если выполнено условие 3с. нормальной линейной регрессионной модели, т. е. $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, то y_i будут также нормально распределены. Отсюда и МНК-оценки коэффициентов регрессии $\hat{a}$ и $\hat{b}$ имеют совместное нормальное распределение как линейные функции от y_i.

Если гипотеза нормальности ошибок не выполняется, то нормальность оценок, вообще говоря, неверна. Однако при некоторых условиях регулярности на поведение x_i при росте n, оценки $\hat{a}$ и $\hat{b}$ имеют асимптотически нормальное распределение, т. е. $\hat{a} \sim N(a, \hat{D}(\hat{a}))$, $\hat{b} \sim N(b, \hat{D}(\hat{b}))$ при $n \to \infty$.

В этих условиях справедливы формулы интервальных оценок или доверительных интервалов:

$$\begin{aligned} \hat{a} - t_\gamma \sqrt{\hat{D}(\hat{a})} < a < \hat{a} + t_\gamma \sqrt{\hat{D}(\hat{a})}, \\ \hat{b} - t_\gamma \sqrt{\hat{D}(\hat{b})} < b < \hat{b} + t_\gamma \sqrt{\hat{D}(\hat{b})}, \end{aligned} \tag{5.10}$$

где $t_\gamma = t\left(\frac{1+\gamma}{2}, n-2\right)$ – квантиль t-распределения (распределения Стьюдента) уровня $\frac{1+\gamma}{2}$ и числа степеней свободы $n-2$. Здесь γ – доверительная вероятность или надежность:

$$P\left(\hat{a} - t_\gamma \sqrt{\hat{D}(\hat{a})} < a < \hat{a} + t_\gamma \sqrt{\hat{D}(\hat{a})}\right) = \gamma,$$

это вероятность того, что построенный нами доверительный интервал покроет истинное значение параметра a. Аналогично можно определить γ и для параметра b. Обычно значения доверительной вероятности стандартизованы и принимаются равными 0,9; 0,95; 0,99; 0,999.

Доверительный интервал для неизвестной дисперсии ошибок наблюдений σ^2:

$$\frac{(n-2)S^2}{u_2} < \sigma^2 < \frac{(n-2)S^2}{u_1}, \tag{5.11}$$

где $u_1 = \chi^2\left(\frac{1-\gamma}{2}, n-2\right)$ и $u_2 = \chi^2\left(\frac{1+\gamma}{2}, n-2\right)$ – квантили χ^2-распределения.

Этап 3б. Критерий Стьюдента. При статистическом исследовании реальной ситуации возникает необходимость не только оценить неизвестные параметры модели, но и проверить по отношению к ним некоторые гипотезы. Например, можно ли считать потребление пропорционально зависящим от дохода $(a=0)$? Будет ли предельная склонность к потреблению больше половины $\left(b > \frac{1}{2}\right)$? И, наконец, служит ли линейная зависимость адекватным отражением эмпирических данных?

Статистики, которые использовались для построения доверительных интервалов, могут использоваться и для проверки или тестирования гипотез о параметрах модели.

Так, для проверки гипотезы $H_0: a = a_0$ против альтернативной гипотезы $H_1: a \neq a_0$ используется статистика

$$t = \frac{\hat{a} - a_0}{\sqrt{\hat{D}(\hat{a})}} = \frac{\hat{a} - a_0}{S_{\hat{a}}} \sim t(n-2), \tag{5.12}$$

распределенная по закону Стьюдента с $(n-2)$ степенями свободы.

Аналогично для гипотезы $H_0: b = b_0$ и $H_1: b \neq b_0$ используется критерий, статистика которого

$$t = \frac{\hat{b} - b_0}{\sqrt{\hat{D}(\hat{b})}} = \frac{\hat{b} - b_0}{S_{\hat{b}}} \sim t(n-2). \tag{5.13}$$

Мы отвергаем гипотезу H_0 (и принимаем H_1) с уровнем значимости $\alpha = 1-\gamma$, если $|t_0| > t_{\frac{1+\gamma}{2}}$ (или $|t_0| > t_{1-\frac{\alpha}{2}}$), t_0 – наблюдаемое или экспериментальное значение t-статистики, в противном случае гипотезу H_0 следует принять, т. е. считать, что результаты наблюдений согласуются с гипотезой H_0, не противоречат ей.

Для такого вида альтернативной гипотезы H_1 область принятия H_0 совпадает с доверительным интервалом для соответствующего неизвестного параметра: гипотеза H_0 принимается на уровне значимости α, если построенный доверительный интервал для a (или b) в форме (5.9) (или (5.10)) покрывает гипотетическое значение параметра a_0 (или b_0).

Для проверки гипотезы $H_0 : \sigma^2 = \sigma_0^2$, против $H_1 : \sigma^2 \neq \sigma_0^2$ может использоваться доверительный интервал (5.11). Гипотезу H_0 принимаем с уровнем $\alpha = 1 - \gamma$, если интервал покрывает значение σ_0^2.

При использовании современных статистических пакетов программ не требуется искать нужные квантили t-распределения (или χ^2-распределения), поскольку в них (пакетах) рассчитывается уровень ошибки, с которой можно отвергнуть нулевую гипотезу и, если он меньше желаемого значения, либо равен ему, то нулевая гипотеза отвергается.

5.4. Верификация модели (4 этап)

Пригодность построенной модели $\hat{y} = \hat{a} + \hat{b}x$ или ее верификация, а также качество оценивания регрессии может быть проверено двумя равноценными способами: дисперсионным анализом в регрессии и с использованием элементов теории корреляции.

Этап 4а. Дисперсионный анализ в регрессии. Суть метода, как уже отмечалось в главе 4, заключается в разложении общей суммарной дисперсии выходной величины *y* на составляющие, обусловленные действием входных переменных-факторов, и остаточную дисперсию, обусловленную ошибкой или всеми неучтенными в данной модели переменными. Фактор оказывает несущественное влияние на *y*, если соответствующая ему дисперсия и дисперсия ошибок статистически незначимы. Для проверки гипотезы о равенстве таких дисперсий используется уже известный нам критерий Фишера (F-критерий).

Рассмотрим $SS_{общ.} = \Sigma(y_i - \bar{y})^2$ – величину, характеризующую разброс значений y_i относительно среднего значения $\bar{y}$. Разобьем эту сумму на две части: объясненную регрессионным уравнением и не объясненную (т. е. связанную с ошибками ε_i).

Обозначим через $\hat{y}_i = \hat{a} + \hat{b}x_i$ предсказанное по модели значение y_i, тогда $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$ (см. рис. 5.4).

Тогда $SS_{общ.}$ представляется в виде суммы двух слагаемых:

$$\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2; \qquad SS_{общ.} = SS_R + SS_{ост.}. \tag{5.14}$$

Здесь через $SS_R = \Sigma(\hat{y}_i - \bar{y})^2$ обозначена сумма квадратов, объясненная регрессией, и $SS_{ост.} = \Sigma(y_i - \hat{y}_i)^2$ – остаточная сумма квадратов, обусловленная ошибкой.

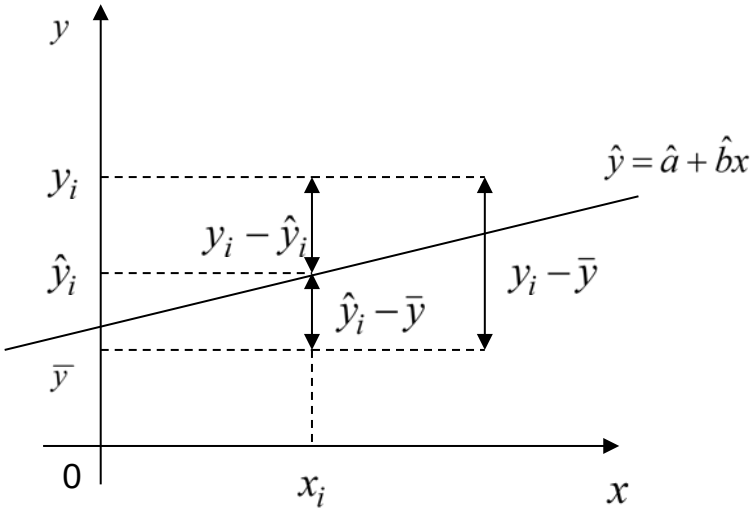


Рис. 5.4. Графическое представление выражения $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

Заметим, что вектор остатков регрессии ортогонален константе, т.е. $s^T e = \Sigma e_i = 0$, вообще говоря, только в том случае, когда константа включена в число объясняющих параметров регрессии. Поэтому (5.14) справедливо только в этом случае.

Коэффициент детерминации или доля объясненной дисперсии *y*, определяется по формуле

$$R^2 = 1 - \frac{SS_{ост.}}{SS_{общ.}} = \frac{SS_R}{SS_{общ.}}. \qquad (5.15)$$

В силу определения $0 \le R^2 \le 1$. Если $R^2 = 0$, то это значит, что регрессия ничего не дает, т. е. фактор *x* не улучшает качество предсказания y_i по сравнению с тривиальным предсказанием $\hat{y}_i = \bar{y}$.

Другой крайний случай $R^2 = 1$ означает точную подгонку: все наблюдаемые значения (x_i, y_i) лежат на регрессионной прямой (все остатки $e_i = 0$).

Чем ближе к 1 значение R^2, тем лучше качество подгонки или качество регрессии, $\hat{y}$ более точно аппроксимирует *y*.

Используя критерий Фишера, проверяем гипотезу об отсутствии линейной функциональной связи между *x* и *y* H_0 : $b = 0$. Наблюдаемое значение критерий Фишера определяется по формуле:

$$F_0 = \frac{MS_R}{MS_{ост.}} = \frac{SS_R/1}{SS_{ост.}/(n-2)} \sim F(1, n-2), \qquad (5.16)$$

как величина, распределенная по закону Фишера со степенями свободы $(1, n-2)$.

Используя коэффициент детерминации (5.15), получим для F -статистики

$$F_0 = (n-2)\frac{R^2}{1-R^2}. \qquad (5.17)$$

Вычисления, необходимые для дисперсионного анализа уравнения регрессии, обычно сводят в таблицу (табл. 5.1).

Таблица 5.1

Дисперсионный анализ одномерной регрессии

Источник дисперсии	Число степеней свободы	Сумма квадратов SS	Средний квадрат MS	Критерий Фишера F	Критическая точка	Гипотеза $H_0: b=0$
Регрессор x	1	SS_R	$MS_R = \frac{SS_R}{1}$	$F_0 = \frac{MS_R}{MS_{ост.}}$	$F_{кр.} = F(\alpha;1,n-2)$	Сравниваем F_0 и $F_{кр}$
Ошибка	$n-2$	$SS_{ост.} = SS_{общ} - SS_R$	$MS_{ост.} = \frac{SS_{ост.}}{n-2}$	–	–	–
Общая дисперсия (итог)	$n-1$	$SS_{общ.}$	–	–	–	–

Если при заданном уровне значимости α наблюдаемое значение F-статистики больше критической точки $F_0 > F(\alpha;1,n-2)$, то гипотеза $H_0: b=0$ отвергается, то есть связь между x и y есть, и результаты наблюдений не противоречат предположению о ее линейности. В противном случае $H_0: b=0$ принимается и постулируется отсутствие значимой линейной функциональной связи между x и y. Исходя из соотношения (5.16), малым значениям F-статистики будут соответствовать и малые значения коэффициента детерминации R^2 (плохая аппроксимация данных).

Этап 4б. Использование элементов теории корреляции. Другой способ верификации линейной модели состоит в использовании элементов теории корреляции. Мерой линейной связи двух величин является коэффициент корреляции, выборочное значение которого

$$r_B = \hat{r} = r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{[n\Sigma x_i^2 - (\Sigma x_i)^2][n\Sigma y_i^2 - (\Sigma y_i)^2]}} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_x \cdot \sigma_y} \qquad (5.18а)$$

будет его несмещенной оценкой.

Кроме формулы (5.18а) коэффициент корреляции можно вычислить следующим образом:

$$r_B = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_x \cdot \sigma_y}, \qquad (5.18б)$$

где $\overline{xy} = \frac{\sum x_i y_i}{n}$, $\sigma_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$, $\sigma_y = \sqrt{\frac{\sum (y_i - \overline{y})^2}{n}} = \sqrt{\frac{SS_{общ}}{n}}$.

$$r_B = \hat{b} \cdot \frac{\sigma_x}{\sigma_y}, \qquad (5.18в)$$

Формулы (5.18а)–(5.18в) являются математически равноценными, но (5.18в) – самая простая с вычислительной точки зрения.

Коэффициент корреляции обладает следующими свойствами (рис. 5.5):

1. Значения коэффициента корреляции принадлежат промежутку [–1; 1], т.е. $-1 \le r_B \le 1$. Чем больше его абсолютное значение $|r_B|$ к 1, тем теснее связь между признаками. Положительная величина коэффициента корреляции свидетельствует о прямой связи между ними, отрицательная – об обратной.

2. При $|r_B|$ = 1 связь между x и y представляет собой линейную функциональную зависимость, при которой все точки располагаются на прямой линии.

3. При $r_B = 0$ линейная функциональная связь отсутствует. При этом линия регрессии параллельна оси OX.

4. При $|r_B| \in (0;0{,}3)$ – линейная связь практически отсутствует, $|r_B| \in [0{,}3;0{,}5)$ – слабая, $|r_B| \in [0{,}5;0{,}7)$ – средняя, $|r_B| \in [0{,}7;0{,}9)$ – сильная, $|r_B| \in [0{,}9;1)$ – очень сильная.

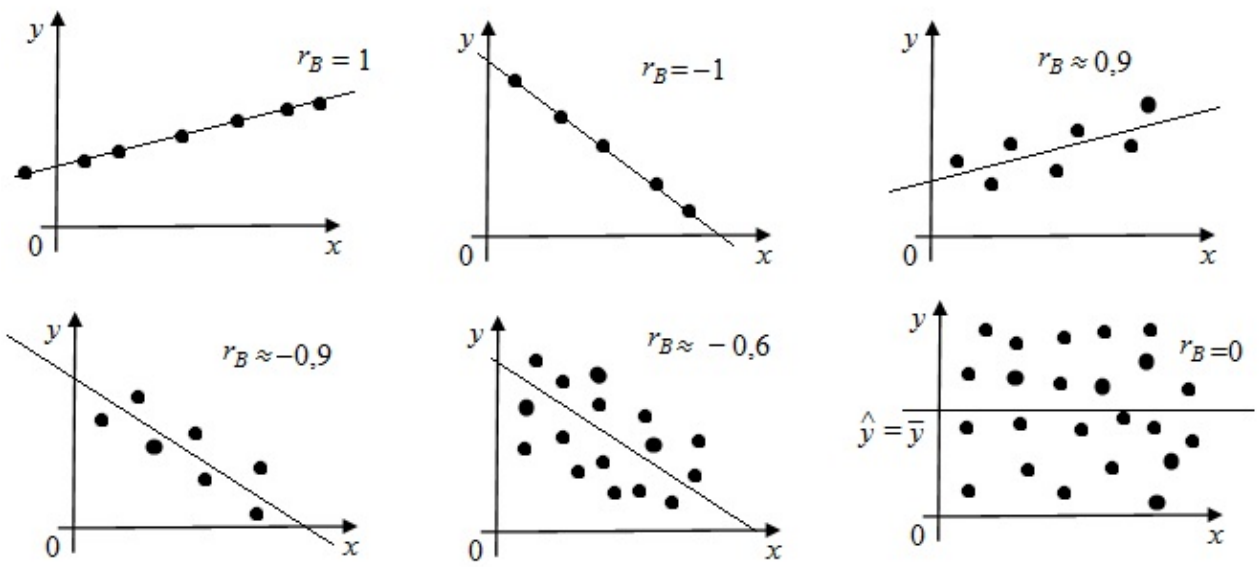


Рис. 5.5. Иллюстрация свойств коэффициента корреляции

Гипотеза об отсутствии линейной функциональной связи между x и y может быть записана как $H_0 : r = 0$. Для проверки H_0 используется критерий, статистика которого

$$t = \frac{r_B\sqrt{n-2}}{\sqrt{1-r_B^2}} \sim t(n-2) \qquad (5.19)$$

распределена по закону Стьюдента с $(n-2)$ степенями свободы.

Вывод о значимости корреляции между x и y может быть сделан, если $|t_0| > t_{1-\frac{\alpha}{2}}$, где $t_{1-\frac{\alpha}{2}} = t\left(1-\frac{\alpha}{2}, n-2\right)$ – квантиль t-распределения, α – уровень значимости.

Коэффициент детерминации $R^2 = r_B^2$ (чаще всего выражаемый в %). Если $r_B = 0{,}9$, то это значит, что линейная регрессия *y* на *x* объясняет 81% дисперсии *y*. Остальные 19% приходятся на долю прочих факторов, не учтенных в уравнении регрессии.

Этап 4в. Средняя ошибка аппроксимации. Фактические значения зависимой переменной *y* отличаются от теоретических (предсказанные) $\hat{y}$. Чем меньше эти отличия, тем ближе теоретические значения $\hat{y}$ к эмпирическим *y*, тем лучше качество модели. Величина отклонений $(y_i - \hat{y}_i), i = 1,2,...,n$ по каждому наблюдению представляет собой ошибку аппроксимации. В отдельных случаях ошибка аппроксимации может быть равной нулю. Отклонения $(y_i - \hat{y}_i)$ несравнимы между собой, кроме равенства нулю. Если $(y_i - \hat{y}_i)$ = 6 в одном случае и $(y_i - \hat{y}_i)$= 12, в другом, то это не означает, что во втором случае модель дает вдвое худший результат. Кроме этого ошибка аппроксимации $(y_i - \hat{y}_i)$ может быть как положительной, так и отрицательной. Поэтому для сравнения разных результатов между собой используют ее относительное абсолютное значение

$$A = \frac{1}{n} \cdot \sum \frac{|y_i - \hat{y}_i|}{y_i} \cdot 100 = \frac{1}{n} \cdot \sum \frac{|e_i|}{y_i} \cdot 100 \ . \qquad (5.20)$$

Средняя ошибка аппроксимации, рассчитанная по формуле (5.20), характеризует качество подбора модели, чем она ниже, тем меньше отклонений эмпирических данных от теоретических. Из практики известно, что значение *A* не должно превышать 10-15%.

5.5. Интерпретация уравнения регрессии (5 этап)

Существуют два этапа интерпретации уравнения регрессии. Первый этап состоит в словесном толковании уравнения так, чтобы это было понятно человеку, не являющемуся специалистом в области статистики или эконометрики. На втором этапе необходимо решить, следует ли ограничиться этим или провести более детальное исследование зависимости, например, проверить по отношению к исследуемым переменным некоторые статистические гипотезы, либо улучшить качество и предсказательные свойства модели.

Представим простой способ интерпретации коэффициентов линейного уравнения регрессии $\hat{y} = \hat{a} + \hat{b}x$, когда *x* и *y* – переменные с простыми, естественными единицами измерения.

Во-первых, можно сказать, что увеличение *x* на одну единицу измерения приведет к увеличению *y* в среднем на $\hat{b}$ единиц (в единицах измерения и переменной *x* и переменной *y*). Здесь коэффициент регрессии $\hat{b}$ есть абсолютный показатель силы связи, характеризующий среднее абсолютное изменение результата *y* при изменении фактора *x* на единицу своего измерения. Вторым шагом является проверка, каковы действительно единицы измерения *x* и *y*, и замена слова «единица» фактическим количеством.

Постоянная $\hat{a}$ дает прогнозируемое значение y (в единицах y), если $x = 0$. Это может иметь или не иметь ясного смысла в зависимости от конкретной ситуации. Если $x = 0$ находится достаточно далеко от выборочных значений переменной x, то буквальная интерпретация может привести к неверным результатам; даже если линия регрессии довольно точно описывает значения наблюдаемой выборки, мы не можем гарантировать, что это ее свойство сохранится при экстраполяции влево или вправо. В случае, когда интерпретация $\hat{a}$ не имеет никакого смысла, эта константа выполняет единственную функцию: она позволяет определить положение линии регрессии на графике.

При интерпретации уравнения регрессии важно помнить о трех вещах. Во-первых, $\hat{a}$ является лишь оценкой a, а $\hat{b}$ – оценкой параметра b. Поэтому вся интерпретация в действительности представляет собой лишь оценку. Во-вторых, уравнение регрессии отражает только общую тенденцию для выборки. При этом каждое отдельное наблюдение подвержено воздействию случайностей. В-третьих, верность интерпретации зависит от правильности спецификации уравнения.

Для линейного уравнения $y = a + bx$ можно вычислить коэффициент эластичности

$$E = f'(x)\frac{x}{y} = \frac{\hat{b}x}{y}. \qquad (5.21)$$

Эластичность E приближенно показывает на сколько процентов в среднем изменится значение зависимой переменной y при изменении независимой переменной x на 1% от своего среднего значения. При этом, если $|E| > 1$, то y эластичен, если $|E| < 1$, то y неэластичен, если $|E| = 1$, то y нейтрален. При интерпретации уравнения регрессии значение эластичности в любой точке будет зависеть не только от значения $\hat{b}$, но также и от значений y и x в данной точке.

5.6. Прогноз на основе линейной модели (6 этап)

Построенная адекватная модель может использоваться для прогнозирования. Оценка прогнозируемых величин в регрессионном анализе получается подстановкой в регрессию значения независимой переменной. Прогноз на основе уравнения регрессии звучит так: «если независимая переменная равна такой-то величине, то зависимая переменная составит такую-то величину».

Рассмотрим подробнее задачу прогноза на основе линейной модели. **Прогноз среднего** значения y, соответствующего некоторому заданному значению x_0, которое может лежать как между выборочными наблюдениями от x_1 до x_n, так и вне соответствующего интервала, может быть **точечным** или **интервальным**.

В случае точечного прогноза мы определяем прогнозное значение x_0, исходя из собственных предположений или опираясь на мнения аналитиков. Кроме этого, мы может строить прогноз по трем сценариям развития ситуации: благо-

приятная (оптимистический прогноз), без особых изменений (умеренный прогноз) и негативная (пессимистический прогноз). Следует отметить, что для каждой уникальной модели рост независимого фактора x не всегда является благоприятным сценарием для зависимой переменной y, поэтому при прогнозировании всегда надо помнить какой тип связи между переменными (прямая или обратная) и какой у них экономический смысл. Определив значение x_0 по модели находим прогнозное значение y_0:

$$y_0 = a + bx_0. \tag{5.22}$$

Точечный прогноз не всегда удобен для последующего анализа экономической ситуации, поэтому находятся промежутки изменения зависимой переменной y, т.е. строим интервальный прогноз:

$$y_0 - t_\gamma \cdot \Delta \le y_0 \le y_0 + t_\gamma \cdot \Delta, \tag{5.23}$$

где значение Δ – стандартная ошибка предсказываемого среднего значения y, находится по формуле:

$$\Delta = \hat{\sigma} \cdot \sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\Sigma(x_i - \bar{x})^2}\right)}, \tag{5.24}$$

где $t_\gamma = t\left(\frac{1+\gamma}{2}, n-2\right)$, γ – доверительная вероятность, $\hat{\sigma}^2 = \frac{R^2_{\min}}{n-2} = MS_R$.

Если значение x_0 существенно отличается от $\bar{x}$, т.е. $|x_0 - \bar{x}| > \varepsilon$, то величина (ширина) доверительного интервала увеличивается, что может исказить прогнозные значения. При этом когда $x_0 = \bar{x}$ ширина минимальна (рис. 5.6). Таким образом, прогноз значений зависимой переменной y по формуле (5.24) оправдан, если значения независимой переменной x не выходят за пределы выборки (причем прогноз тем более точный, чем ближе x_0 к $\bar{x}$).

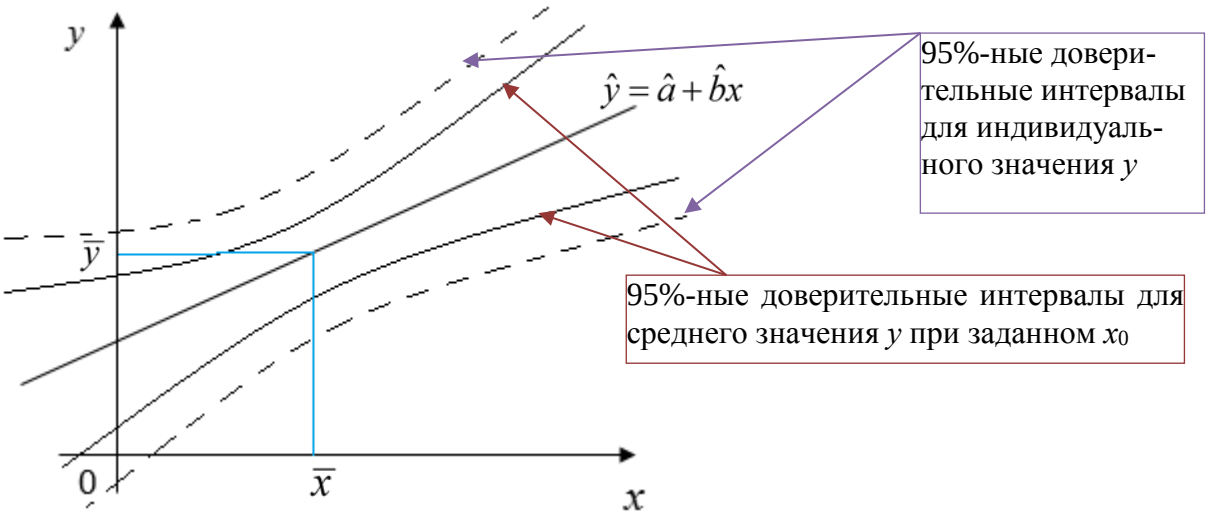


Рис. 5.6. Доверительные интервалы линии регрессии

При построении доверительного интервала индивидуальных значений $\hat{y}_0$ зависимой переменной необходимо учитывать еще один источник вариации –

рассеяние вокруг линии регрессии, т.е. в оценку Δ включить величину $\hat{\sigma}^2$. В результате оценки дисперсии индивидуальных значений y_0 при $x = x_0$ будет равна:

$$\Delta^* = \hat{\sigma} \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\Sigma(x_i - \bar{x})^2}\right)}, \qquad (5.25)$$

а соответствующий доверительный интервал для прогнозов индивидуальных значений $\hat{y}_0$ будет определяться по формуле:

$$y_0 - t_\gamma \cdot \Delta^* \le y_0 \le y_0 + t_\gamma \cdot \Delta^*. \qquad (5.26)$$

5.7. Рекомендации по выполнению расчетно-графической работы по теме «Линейная парная регрессия» в MS Excel

Расчетно-графическая работа выполняется по выборке, выданной преподавателем или собранной студентом самостоятельно. Все расчеты выполняются в двух вариантах: полностью вручную в MS Excel и с помощью функции Регрессия пакета анализа.

Задание:

0. Определить (!) какой из заданных показателей является зависимой переменной, а какой – независимой.

1. Построить поле корреляции.

2. Найти точечные и интервальные оценки параметров модели $y = a + bx$ ($\tilde{a}, \tilde{b}, \tilde{\sigma}^2, \tilde{D}(a), \tilde{D}(b)$).

3. Оценить значимость коэффициентов регрессии при $\gamma = 0{,}95$, используя:

а) *t*-критерий Стьюдента;

б) доверительные интервалы истинных значений параметров.

4. Верифицировать полученную модель, используя:

а) дисперсионный анализ в регрессии;

б) элементы теории корреляции;

в) средняя ошибка аппроксимации.

5. Интерпретировать полученные результаты.

6. В случае пригодной линейной модели построить точечные и интервальные прогнозы зависимой переменной (при α = 0,05).

7. Для заданных показателей подобрать наиболее подходящую нелинейную модель, вычислить ее основные характеристики и провести анализ.

8. По результатам выполненных расчетов сделать вывод о том, какая модель наилучшим образом аппроксимирует исходные данные.

Пример 5.1. Известны следующие данные по одному из субъектов Российской Федерации:

Совокупные доходы физ. лиц, млн. руб.	14855,3	18745,1	20268,7	20319,3	20174,8	22524,5	21805,8
Вклады физ. лиц в банках, тыс. руб.	36 643	38 297	38 993	40 394	41 090	42 691	43 916

Совокупные доходы физ. лиц, млн. руб.	21571,3	22902,8	23928,4	23741,8	30271,9	30481,9	33088,0
Вклады физ. лиц в банках, тыс. руб.	43 988	44 684	43 721	44 198	46 465	47 481	48 438
Совокупные доходы физ. лиц, млн. руб.	32133,7	34915,7	33377,5	34923,4	32558,7	33149,4	
Вклады физ. лиц в банках, тыс. руб.	49 632	53 506	52 559	53 461	49 484	48 387	

1 этап. Спецификация модели

Определим, какой из заданных показателей будет зависимой переменной, а какой – независимой. Так как сбережения в банках – это часть дохода, то совокупные доходы физических лиц обозначим в качестве независимой переменной x, а вклады в банках – y.

	A	B	C	D	E	F	G
1							
2	Совокупные доходы физ. лиц, млн. руб.	Вклады физ. лиц в банках, тыс. руб.					
3	xi	yi					
4	14855,30	36 643					
5	18745,10	38 297					
6	20268,70	38 993					
7	20319,30	40 394					
8	20174,80	41 090					
9	22524,50	42 691					
10	21805,80	43 916					
11	21571,30	43 988					
12	22902,80	44 684					
13	23928,40	43 721					
14	23741,80	44 198					
15	30271,90	46 465					
16	30481,90	47 481					
17	33088,00	48 438					
18	32133,70	49 632					
19	34915,70	53 506					
20	33377,50	52 559					
21	34923,40	53 461					
22	32558,70	49 484					
23	33149,40	48 387					
24							

Рис. 5.7. Исходные данные в MS Excel

Занесем исходные данные в MS Excel в виде таблицы, состоящей из двух столбцов, в **первом** расположены значения **независимой** переменной x, а во **втором** – **зависимой** переменной y (рис. 5.7). Чтобы определить характер зависимости – построим поле корреляции. Для этого выделим оба столбца (**данные *x*** должны быть в **первом** столбце, ***y* – во втором**), заходим в меню «Вкладка» и выбираем точечную диаграмму (рис. 5.8). Кнопка **+** в правом верхнем углу позволяет добавить на диаграмму названия осей координат.

Следующим шагом наносим на поле корреляции прямую $y = a + bx$: в контекстном меню (правая кнопка мыши) выбираем пункт «добавить линию тренда»

(предварительно подсветив график). В появившемся меню справа выделяем линейную модель и отмечаем галочкой «показывать уравнения на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации (R^2)» (рис. 5.9). Поле корреляции примет вид как на рис. 5.10.

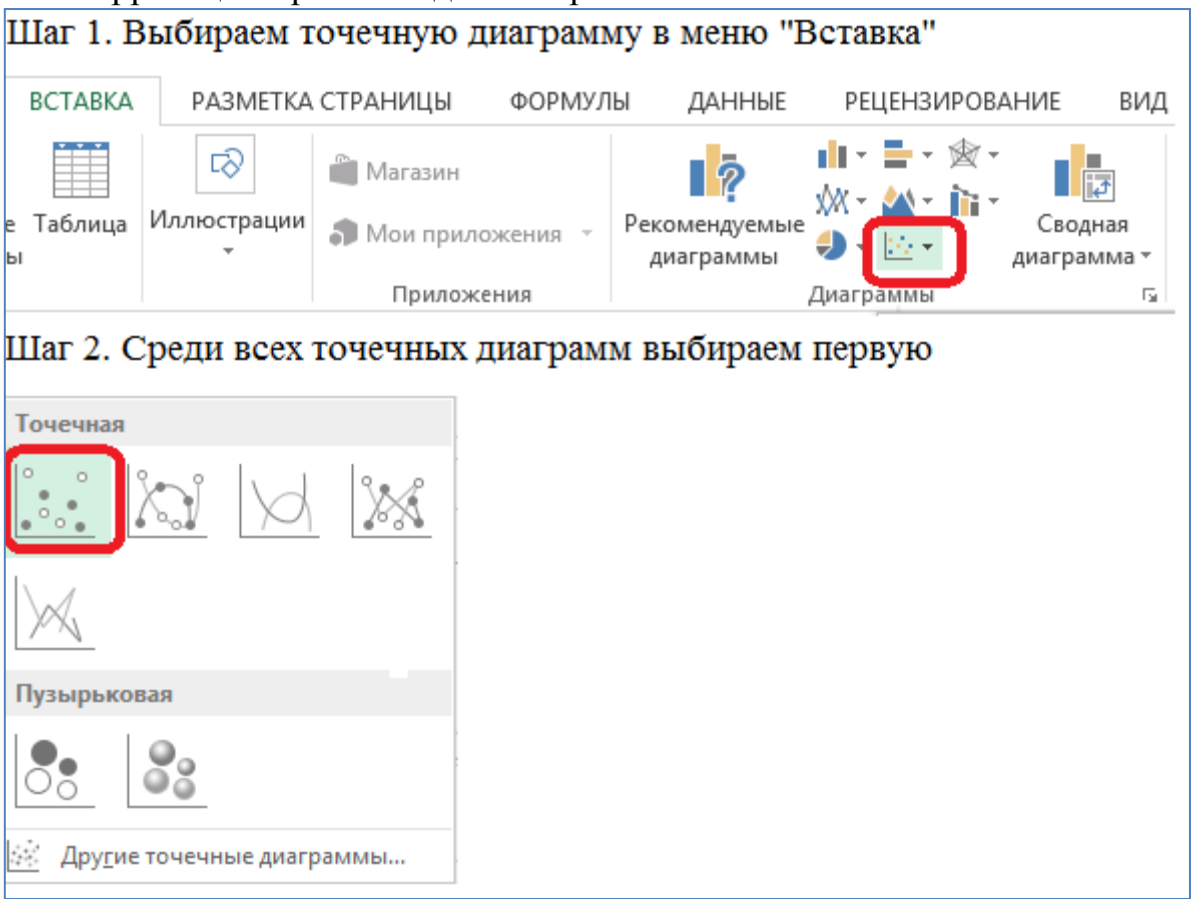


Рис. 5.7. Построение поля корреляции с использованием точечной диаграммы

Рис. 5.8. Поле корреляции

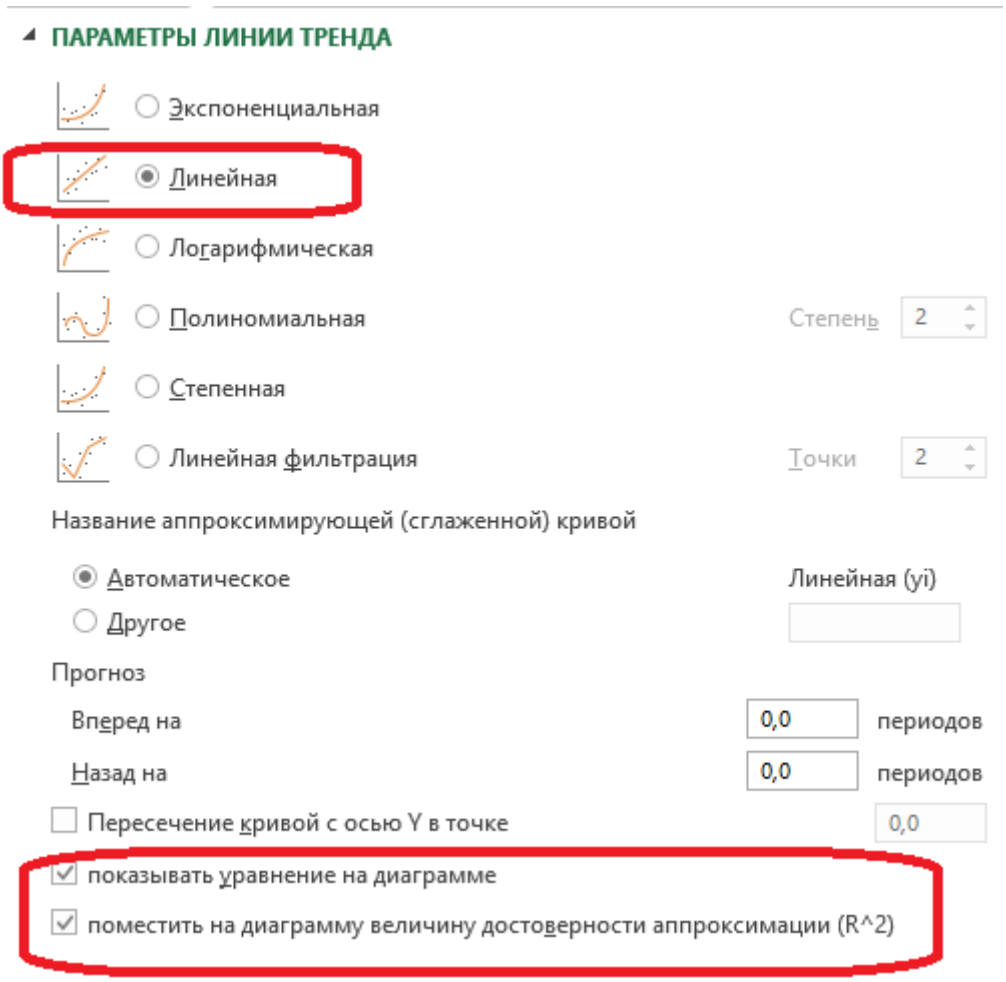


Рис. 5.9. Добавление на диаграмму прямой $y = a + bx$

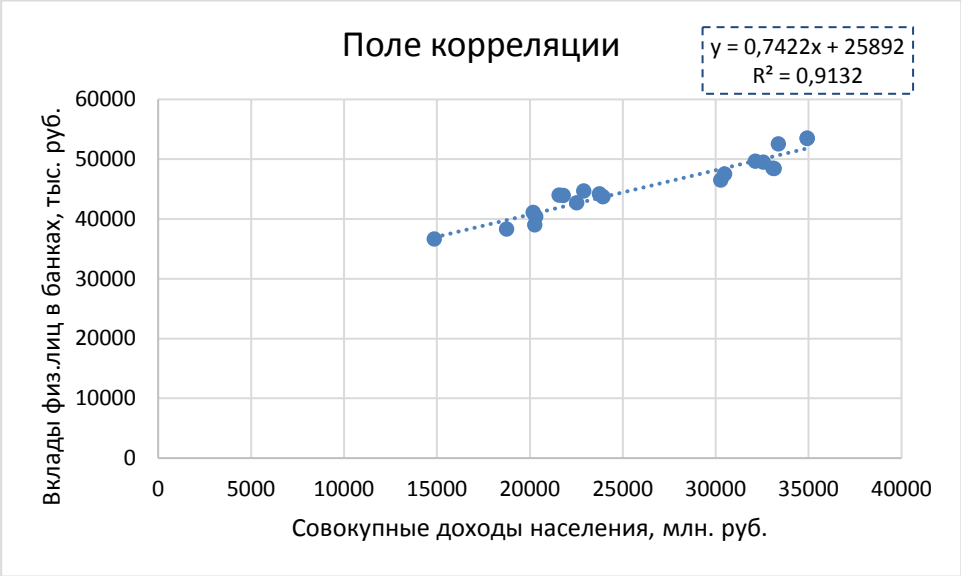


Рис. 5.10. Поле корреляции подготовлено для анализа

Ручной расчет в Excel

2 этап. Нахождение оценок параметров по МНК

Формируем вспомогательную таблицу (рис. 5.11). Затем очень аккуратно находим значения оценок параметров a и b по формулам (5.5). **Не забудьте после**

вычислений сверить их значения с теми, что получились в уравнении на диаграмме, так как MS Excel для расчетов также использует метод наименьших квадратов (рис. 5.10). Для удобства вычислений знаменатель формул (5.5) лучше найти отдельно (рис. 5.12), так как также он встречается в других формулах:

$$k = n\sum x_i^2 - \left(\sum x_i\right)^2 . \qquad (5.26)$$

	A	B	C	D	E	F
2	номер	Совокупные доходы физ. лиц, млн. руб.	Вклады физ. лиц в банках, тыс. руб.	=B4^2	=C4^2	=C4*B4
3	наблюдения	xi	yi	xi ^2	yi ^2	xi*yi
4	1	14855,30	36643	220679938	1342720222	544344942
5	2	18745,10	38297	351378774	1466649792	717878545
6	3	20268,70	38993	410820200	1520487505	790346114
7	4	20319,30	40394	412873952	1631634923	820767665
8	5	20174,80	41090	407022555	1688359830	828975592
9	6	22524,50	42691	507353100	1822546071	961599917
10	7	21805,80	43916	475492914	1928573160	957613111
11	8	21571,30	43988	465320984	1934933675	948875777
12	9	22902,80	44684	524538248	1996690420	1023396548
13	10	23928,40	43721	572568327	1911561692	1046183387
14	11	23741,80	44198	563673067	1953487690	1049346653
15	12	30271,90	46465	916387930	2159036092	1406596820
16	13	30481,90	47481	929146228	2254481257	1447322616
17	14	33088,00	48438	1094815744	2346240135	1602716643
18	15	32133,70	49632	1032574676	2463358652	1594867318
19	16	34915,70	53506	1219106106	2862908409	1868204786
20	17	33377,50	52559	1114057506	2762417156	1754278076
21	18	34923,40	53461	1219643868	2858080232	1867040446
22	19	32558,70	49484	1060068946	2448686643	1611141418
23	20	33149,40	48387	1098882720	2341308930	1604002471
24	сумма	525738,00	908029,51	14596405782	41694162487	24445498846
25		=СУММ(B4:B23)				
26						

Рис. 5.11. Вспомогательная таблица для расчета оценок параметров

	A	B	C	D	E	F	G
22	19	32558,70	49484	1060068946	2448686643	1611141418	
23	20	33149,40	48387	1098882720	2341308930	1604002471	
24	сумма	525738,00	908029,51	14596405782	41694162487	24445498846	
25							
26		=СУММ(B4:B23)					
27			=A23*D24-(B24)^2 формула (5.26)				
28	k=	15527670992					
29	a~=	25891,81	=(D24*C24-B24*F24)/B28 (5.5) ! a~ совпадает с уравнением на диаграмме				
30	b~=	0,742	=(A23*F24-B24*C24)/B28 (5.5) ! b~ совпадает с уравнением на диаграмме				
31							

Рис. 5.12. Вычисление оценок параметров *a* и *b*

Далее вычисляем значения остатков e_i, остаточной суммы квадратов, дисперсии коэффициентов регрессии, ковариацию, несмещенные оценки дисперсии

ошибок наблюдений (формулы (5.6)–(5.8)) (рис. 5.13). Можно также сразу вычислить для дисперсионного анализа в регрессии сумму квадратов SS_R и $SS_{общ}$

Уравнение модели:
yi^ = 25892,81 + 0,7422 * xi
формула (5.9)

=C4-I4 =(C4-$O$6)^2 =(I4-$O$6)^2 =B24/A23 =$B$29+$B$30*B4 =D24/B28*O16 (5.6) =A23/B28*O16 (5.7) =K24/(A23-2)

yi^	ei=yi - yi^	(ei)^2	(yi-y cp)^2	(yi^-y cp)^2
36917,15	-274,00	75077	76708321	71983804
39804,09	-1507,22	2271724	50475507	31330749
40934,88	-1941,45	3769218	41063062	19950509
40972,43	-578,93	335160	25079810	19616439
40865,19	224,47	50387	18591789	20577928
42609,09	82,20	6756	7345117	77974
42075,68	1839,84	3385007	2208055	110608
41901,64	2086,24	4352391	1998250	12248832
42889,86	1794,48	3220174	514281	63082
43651,04	70,37	4952	2822621	30640
43512,55	685,73	470224	1447687	3568046
48359,07	-1893,64	3585879	1131997	8747368
48514,93	-1033,55	1068227	4325994	9693591
50449,13	-2011,13	4044628	9220498	2547880
49740,86	-108,63	11801	17899316	1883030
51805,62	1700,54	2891829	65685795	41013002
50663,99	1894,71	3589930	51225889	27694066
51811,33	1649,69	2721464	64956191	41086231
50056,29	-572,09	327283	16668687	21667318
50494,70	-2107,63	4442084	8913798	25940925
908029,51	0,00000	40624194	468282663	427658468
	(5.14)	◄SS ост▲	▲SS общ	▲SS R ► (5.14)
		R min▲	468282663	

x cp= 26286,9
y cp= 45401,48
D^(a^)= 2121543
D^(b^)= 0,0029
cov(a^,b^) -76,41
▲(5.8)
S^2= 2256900

Проверка: если значения yi^ найдены верно, то сумма остатков = 0,000

=K24+M24 ! Проверка SS R+SS ост= SS общ

Рис. 5.13. Расчет всех значений, необходимых для анализа полученной модели по формулам (5.14).

3 этап. Оценивание коэффициентов регрессии с использованием доверительных интервалов и критерия Стьюдента

А) Оценка значимости коэффициентов регрессии с использованием доверительных интервалов.

Расчет доверительных интервалов проводим по формулам (5.10). Для этого нам нужно будет найти значение квантиля распределения Стьюдента t_γ при $\gamma = 0{,}95$. Все расчеты в MS Excel представлены на рис. 5.14.

=A23*D24-(B24)^2 формула (5.26)

k=	1,5528E+10
a^=	25891,81
b^=	0,7422

=(D24*C24-B24*F24)/B28 (5.5) ! a^ совпадает с уравнением на диаграмме

=(A23*F24-B24*C24)/B28 (5.5) ! b^ совпадает с уравнением на диаграмме

Рис. 5.14. Нахождение доверительных интервалов и критерий Стьюдента

Б) Критерий Стьюдента.

Находим для проверки гипотез H_{0a} и H_{0b} наблюдаемые значения статистики критерия Стьюдента по формулам (5.12) и (5.13) (рис. 5.14).

4 этап. Верификация модели

А) Дисперсионный анализ в регрессии. Так как суммы квадратов уже рассчитаны, остается составить таблицу критерия Фишера, найти значение коэффициента детерминации R^2 (5.15), наблюдаемого значения критерия Фишера F_0 (5.16) и критического значения критерия Фишера $F_{крит}$ (рис. 5.15).

	A	B	C	D	E	F	G
41	Дисперсионный анализ в регрессии						
42							
43	Источник	Число степ.	SS	MS	Критерий	Крит.	Гипотеза
44	дисперсии	свободы			Фишера	точка	H0: b=0
45	Регрессор	1	427658468	427658468,4	189,49	4,41	отклонить
46	x						
47	Ошибка	18	40624194	2256899,694			
48	Итог	19	468282663				
49							
50	R^2=	0,9132					

=FРАСПОБР(0,05;1;A23-2)

=C45/C48

Рис. 5.15. Дисперсионный анализ в регрессии

	I	J	K	L	M	N	O	P	Q
39			=(A23*F24-B24*C24)/КОРЕНЬ(B28*(A23*E24-C24^2))						
40	Элементы теории корреляции:						=(B4-$O$5)^2		=ABS(J4)/C4
41							Вспомогательная таблица:		
42	(5.18а) r в=	0,9556		=(J49-J47*J48)/(J45*J46)				(xi - xср)^2	A
43	(5.18б) r в=	0,9556	=B30*J45/J4					130681479	0,0075
44	(5.18в) r в=	0,9556	проверка:	(r в) ^2=	0,9132	=J44^2		56878747	0,0394
45	sx=	6230,50			Верно!			36218731	0,0498
46	sy=	4838,82	=КОРЕНЬ(P63/A23					35612250	0,0143
47			=КОРЕНЬ(L24/A23)					37357766	0,0055
48								14155654	0,0019
49	(xy)ср=	1222274942,29	=F24/A23					20080257	0,0419
50								22236883	0,0474
51	H0: r=0		=J44*(A23-2)^0,5/(1-M44)^0,5					11452133	0,0402
52	t0r=	13,76551	(5.19)					5562522	0,0016
53								6477534	0,0155
54								15880225	0,0408
55	Средняя ошибка аппроскимации							17598025	0,0218
56								46254961	0,0415
57	A=	2,62	=Q63/A23*100					34185070	0,0022
58		(5.20)						74456189	0,0318
59								50276608	0,0360
60								74589132	0,0309
61								39335475	0,0116
62								47093906	0,0436
63							Сумма	776383550	0,5250

Рис. 5.16. Вычисление коэффициента корреляции по формулам (5.18а–в) и средней ошибки аппроксимации по формуле (5.20)

Б) Элементы теории корреляции. Находим значение коэффициента корреляции по трем формулам (5.18а–в), предварительно вычислив во вспомогательной таблице столбец $(x_i - \bar{x})^2$. Сразу делаем проверку $(r_B)^2 = R^2$. Затем проверяем гипотезу о незначимости полученного коэффициента корреляции по формуле (5.16). Все расчеты представлены на рис. 5.16.

В) Средняя ошибка аппроксимации. Для нахождения средней ошибки аппроксимации по формуле (5.20) во вспомогательной таблице создадим столбец А для пошагового вычисления отношения модуля остатков к y_i (рис. 5.16). Функция =ABS() позволяет находить модуль числа.

5 этап. Интерпретация коэффициентов регрессии

На данном этапе из расчетной части нужно вычислить коэффициент эластичности по формуле (5.21) и сделать вывод (рис. 5.17).

6 этап. Построение прогноза

Пусть в заданном примере доходы населения увеличатся на 12% в случае оптимистического прогноза. Точечные прогнозные значения переменных находим по формуле (5.22), интервальные – (5.23) и (5.24). Все расчеты представлены на рис. 5.17.

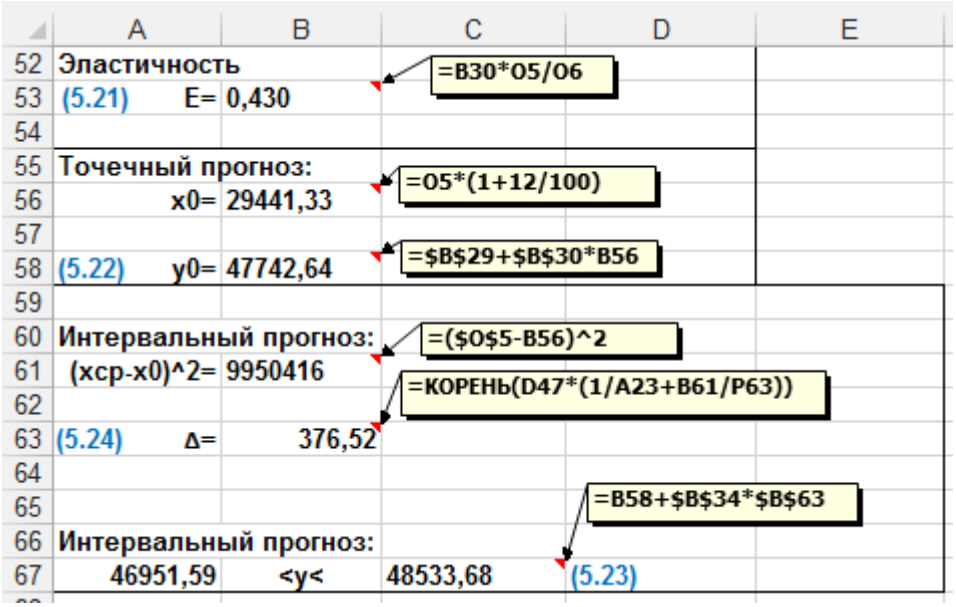


Рис. 5.17. Вычисление коэффициента эластичности и построение точечного и интервального прогнозов

Автоматический расчет с помощью Анализа данных

Проверка всех результатов расчетов проводится с использованием пакета анализа: меню Данные – Анализ данных – регрессия (рис. 5.18). Галочку метки ставим только в том случае, если выделяем исходные данные вместе с названиями столбцов x_i и y_i. Галочку остатки ставим для вычисления средней ошибки аппроксимации.

Результаты расчетов, проведенных с помощью Пакета анализа, представлены на рис. 5.19. Множественный R – это значение коэффициента корреляции, *R*-квадрат – коэффициент детерминации. В таблице Дисперсионный анализ *F* – это значение $F_{набл}$. Если значимость *F* меньше значения 0,05, то гипотеза об отсутствии линейной связи отклоняется. В следующей таблице в столбце Коэффициенты и строке Y-пересечение – это значение $\hat{a}$, в строке xi – это $\hat{b}$. В столбце Стандартная ошибка стоят значения $S_a = \sqrt{\hat{D}(\hat{a})}$ и $S_b = \sqrt{\hat{D}(\hat{b})}$ соответственно. В столбце t-статистика – значения t_{0a} и t_{0b} соответственно. Если в столбце *P*-

значение стоит число меньшее, чем 0,05, то соответствующий коэффициент регрессии стастически значим с вероятностью 0,95. Нижние и верхние 95% – это границы доверительных интервалов при $\gamma = 0{,}95$. В последней таблице Вывод остатка даны теоретические или предсказанные значение зависимой переменной, рассчитанные по построенной модели. **Полное описание всех результатов расчетов приведено ранее и в оформлении работы.**

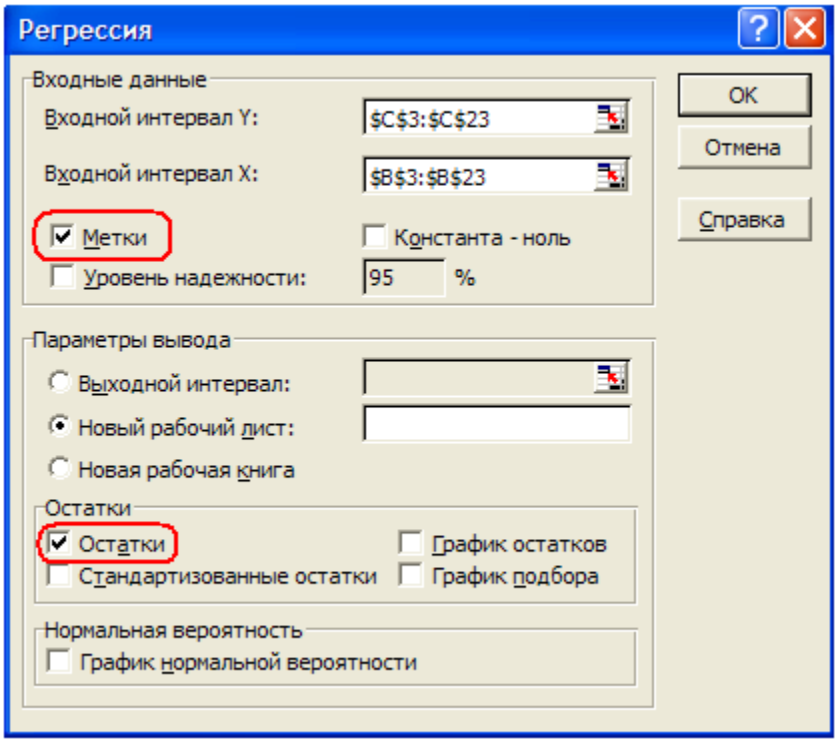


Рис. 5.18. Использование меню Анализа данных в MS Excel

	A	B	C	D	E	F	G	H
1	ВЫВОД ИТОГОВ							
2	*Регрессионная статистика*							
3	Множественный R	0,9556	**коэффициент корреляции**					
4	R-квадрат	0,9132	**коэффициент детерминации**					
5	Нормированный R-квадрат	0,9084						
6	Стандартная ошибка	1502,30	**√(Rmin/(n-2)**					
7	Наблюдения	20						
8	Дисперсионный анализ		**SS R**	**SS ост**	**SS общ**	**F0**		
9		*df*	*SS*	*MS*	*F*	*Значимость F*		
10	Регрессия	1	427658468,4	427658468,4	189,49	0,00000		
11	Остаток	18	40624194,5	2256899,694				
12	Итого	19	468282662,9					
13								
14		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	
15	Y-пересечение	25891,81	1456,5518	17,7761	0,00000	22831,71	28951,92	
16	xi	0,7422	0,0539	13,7655	0,00000	0,6289	0,8555	
17		**a^** **b^**	**√D^(a^)** **√D^(b^)**	**t0a** **t0b**		**доверительные интервалы**		
18	ВЫВОД ОСТАТКА							
19		**y^**	**ei**					
20	*Наблюдение*	*Предсказанное yi*	*Остатки*					
21	1	36917,15	-274,00					
22	2	39804,09	-1507,22					
23	3	40934,88	-1941,45					

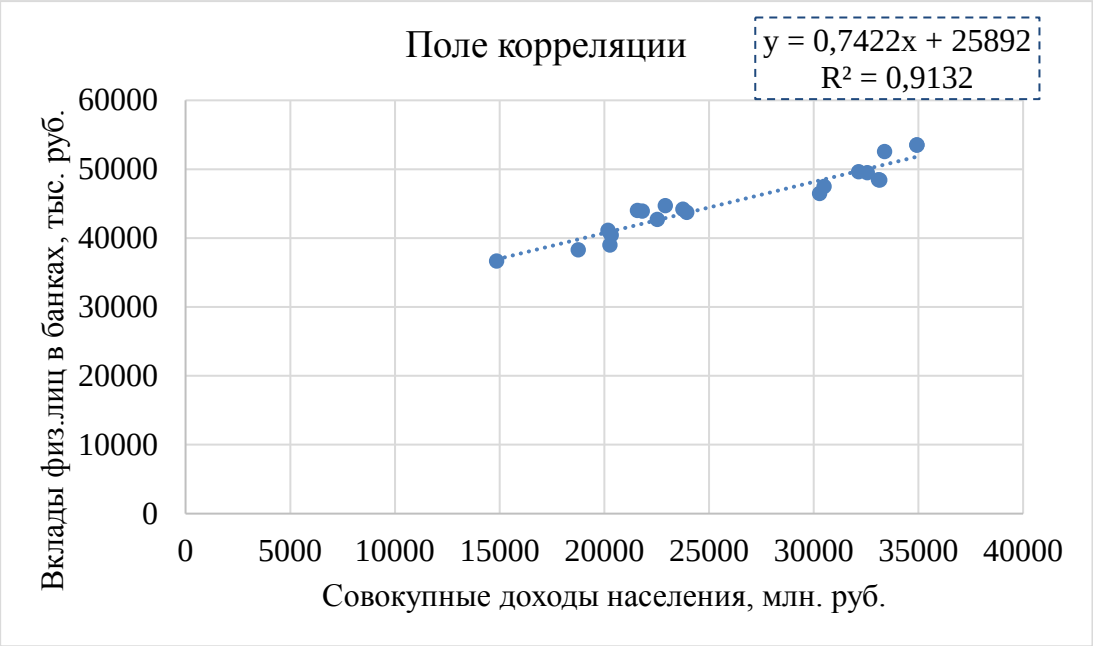
Рис. 5.19. Результаты расчетов линейной модели $y = a + bx$

5.8. Оформление результатов расчетов

Исходные данные (20 наблюдений):

Совокупные доходы физ. лиц, млн. руб.	14855,3	18745,1	20268,7	20319,3	20174,8	22524,5	21805,8
Вклады физ. лиц в банках, тыс. руб.	36 643	38 297	38 993	40 394	41 090	42 691	43 916
Совокупные доходы физ. лиц, млн. руб.	21571,3	22902,8	23928,4	23741,8	30271,9	30481,9	33088,0
Вклады физ. лиц в банках, тыс. руб.	43 988	44 684	43 721	44 198	46 465	47 481	48 438
Совокупные доходы физ. лиц, млн. руб.	32133,7	34915,7	33377,5	34923,4	32558,7	33149,4	
Вклады физ. лиц в банках, тыс. руб.	49 632	53 506	52 559	53 461	49 484	48 387	

1 этап. Спецификация модели. В качестве независимой переменной x возьмем совокупные доходы физических лиц, млн. руб., зависимой переменной будет y – вклады физических лиц в банках, тыс. руб. Очевидно, что доля денежных средств, идущая на сбережения напрямую зависит от получаемых доходов. Построим поле корреляции, чтобы определить характер зависимости. Из графика видно, что точки распределены практически однородно относительно прямой, поэтому можно сказать, что условие гомоскедастичности выполняется.



2 этап. Построение модели. Найдем оценки параметров модели $y = a + bx$ с помощью метода наименьших квадратов. Для этого составляем и заполняем первую вспомогательную таблицу:

№	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	14855,30	36643	220679938	1342720222	544344942
2	18745,10	38297	351378774	1466649792	717878545

3	20268,70	38993	410820200	1520487505	790346114
4	20319,30	40394	412873952	1631634923	820767665
5	20174,80	41090	407022555	1688359830	828975592
6	22524,50	42691	507353100	1822546071	961599917
7	21805,80	43916	475492914	1928573160	957613111
8	21571,30	43988	465320984	1934933675	948875777
9	22902,80	44684	524538248	1996690420	1023396548
10	23928,40	43721	572568327	1911561692	1046183387
11	23741,80	44198	563673067	1953487690	1049346653
12	30271,90	46465	916387930	2159036092	1406596820
13	30481,90	47481	929146228	2254481257	1447322616
14	33088,00	48438	1094815744	2346240135	1602716643
15	32133,70	49632	1032574676	2463358652	1594867318
16	34915,70	53506	1219106106	2862908409	1868204786
17	33377,50	52559	1114057506	2762417156	1754278076
18	34923,40	53461	1219643868	2858080232	1867040446
19	32558,70	49484	1060068946	2448686643	1611141418
20	33149,40	48387	1098882720	2341308930	1604002471
Итого	**525738,00**	**908029,51**	**14596405782**	**41694162487**	**24445498846**

Оценки параметров модели находятся из условия

$$R = \sum_{i=1}^{20} (y_i - a - bx_i)^2 \to \min .$$

Тогда

$$\hat{a} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{14596405782 \cdot 908029{,}51 - 525738 \cdot 24445498846}{20 \cdot 14596405782 - (525738)^2} =$$

$$= 25891{,}81;$$

$$\hat{b} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{20 \cdot 24445498846 - 525738 \cdot 908029{,}51}{20 \cdot 14596405782 - (525738)^2} = 0{,}0539 .$$

Уравнение прямой линии примет вид: $y = 25891{,}1 + 0{,}0539 \cdot x$. При этом уравнение модели запишем в виде: $y_i = 25891{,}1 + 0{,}0539 \cdot x_i + \varepsilon_i$.

Для анализа полученной модели рассчитываем теоретические значения объясняемой переменной: $\hat{y}_i = 25891{,}1 + 0{,}0539 \cdot x_i$. Также найдем значение остатков $e_i = y_i - \hat{y}_i$ и минимальное значение функции R. Для этого составим вторую вспомогательную таблицу:

№	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$\sum_{i=1}^{n} (e_i)^2 = (y_i - \hat{y}_i)^2$
1	36643	36917,15	-274,00	75077
2	38297	39804,09	-1507,22	2271724

3	38993	40934,88	-1941,45	3769218
4	40394	40972,43	-578,93	335160
5	41090	40865,19	224,47	50387
6	42691	42609,09	82,20	6756
7	43916	42075,68	1839,84	3385007
8	43988	41901,64	2086,24	4352391
9	44684	42889,86	1794,48	3220174
10	43721	43651,04	70,37	4952
11	44198	43512,55	685,73	470224
12	46465	48359,07	-1893,64	3585879
13	47481	48514,93	-1033,55	1068227
14	48438	50449,13	-2011,13	4044628
15	49632	49740,86	-108,63	11801
16	53506	51805,62	1700,54	2891829
17	52559	50663,99	1894,71	3589930
18	53461	51811,33	1649,69	2721464
19	49484	50056,29	-572,09	327283
20	48387	50494,7	-2107,63	4442084
Итого	**908029,51**	**908029,5**	**0,00000**	$R_{\min}$ = **40624194**

Остаточная сумма квадратов: $R_{\min} = \sum_{i=1}^{n} (e_i)^2 = 40624194$.

Вычислим несмещенные оценки дисперсий и ковариаций оценок $\hat{a}$ и $\hat{b}$:

$$\hat{D}(\hat{a}) = \frac{\sum x_i^2}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{R_{\min}}{n-2} = \frac{14596405782}{15527670992} \cdot \frac{40624194}{20-2} = 2121543;$$

$$\hat{D}(\hat{b}) = \frac{n}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{R_{\min}}{n-2} = \frac{20}{15527670992} \cdot \frac{40624194}{20-2} = 0{,}0029;$$

$$\operatorname{cov}(\hat{a},\hat{b}) = \frac{-\sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{R_{\min}}{n-2} = \frac{-525738}{15527670992} \cdot \frac{40624194}{20-2} = -76{,}41.$$

Несмещенная оценка дисперсии ошибок наблюдений:

$$S^2 = \hat{\sigma}^2 = \frac{R_{\min}}{n-2} = \frac{40624194}{20-2} = 2256900.$$

3 этап. Оценка значимости коэффициентов регрессии при $\gamma = 0{,}95$ с помощью:

а) *доверительных интервалов истинных значений параметров*

Для нахождения интервальных оценок полученных коэффициентов регрессии предварительно вычислим:

– квантиль распределения Стьюдента

$t_{0,95} = t_{0,95}(20-2) = t\left(\frac{1+0{,}95}{2}, 20-2\right) = 2{,}101$(значение находим аналогично теме 2).

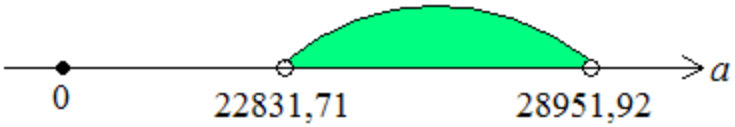
$\sqrt{\hat{D}(\hat{a})} = 1456{,}55$, $\sqrt{\hat{D}(\hat{b})} = 0{,}0539$.

Доверительный интервал для параметра a:

$$\hat{a} - t_\gamma \sqrt{\hat{D}(\hat{a})} < a < \hat{a} + t_\gamma \sqrt{\hat{D}(\hat{a})},$$

$$25891{,}81 - 2{,}101 \cdot 1456{,}55 < a < 25891{,}81 + 2{,}101 \cdot 1456{,}55,$$

$$22831{,}71 < a < 28951{,}92.$$

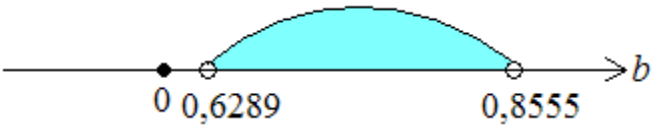


Доверительный интервал для параметра b:

$$\hat{b} - t_\gamma \sqrt{\hat{D}(\hat{b})} < b < \hat{b} + t_\gamma \sqrt{\hat{D}(\hat{b})},$$

$$0{,}7422 - 2{,}101 \cdot 0{,}0539 < b < 0{,}7422 + 2{,}101 \cdot 0{,}0539,$$

$$0{,}6289 < b < 0{,}8555.$$



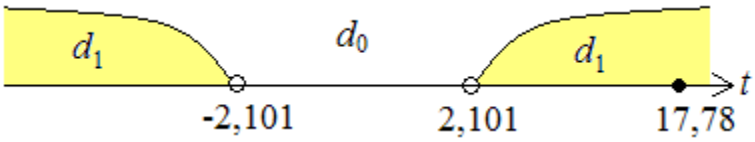
Как мы видим, оба доверительных интервала для коэффициентов регрессии **не содержат нулевых значений**, значит оба коэффициента считаются статистически **значимыми**.

б) *t-критерий Стьюдента*

Проверяем гипотезу $H_0 : a = 0$ против альтернативной гипотезы $H_1 : a \neq 0$, используя при этом статистику $t_{0a} = \dfrac{\hat{a} - 0}{\sqrt{\hat{D}(\hat{a})}} \sim t_\gamma$.

$t_{0a} = \dfrac{25891{,}81 - 0}{1456{,}55} = 17{,}78$ – наблюдаемое или экспериментальное значение t-статистики.

Критическая область двухсторонняя:

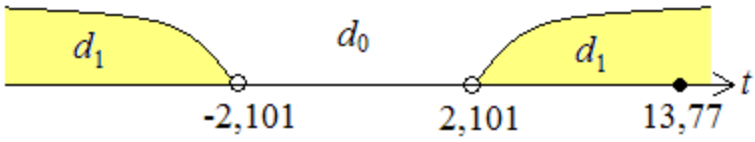


Гипотеза H_0 отвергается с вероятностью 0,95, следовательно, принимается гипотеза H_1, так как $|t_{0a}| > t_\gamma$, т.е. 17,78 > 2,101. Это означает, что параметр a – значим.

Проверяем гипотезу $H_0 : b = 0$ против конкурирующей гипотезы $H_1 : b \neq 0$, используя тот же критерий, только для параметра b: $t_{0b} = \dfrac{\hat{b} - 0}{\sqrt{\hat{D}(\hat{b})}} \sim t_\gamma$.

$t_{0b} = \dfrac{0{,}7422 - 0}{0{,}0539} = 13{,}77$ – наблюдаемое или экспериментальное значение t-статистики.

Критическая область двухсторонняя:



Гипотеза H_0 отвергается с вероятностью 0,95, следовательно, принимается гипотеза H_1, так как $|t_{0b}| > t_\gamma$, т.е. 13,78 > 2,101. Это означает, что параметр b – значим.

4 этап. Верификация модели. Пригодность построенной модели $\tilde{y} = \tilde{a} + \tilde{b}x$ или ее верификация, а также качество оценивания регрессии может быть проверено двумя равноценными способами: дисперсионным анализом в регрессии и с использованием элементов теории корреляции.

а) ***Дисперсионный анализ в регрессии.*** Суть метода заключается в разложении общей суммарной дисперсии вкладов физических лиц в банках на составляющие, обусловленные действием доходов населения, и остаточную дисперсию, обусловленную ошибкой или всеми неучтенными в данной модели переменными. Для проверки гипотезы о равенстве таких дисперсий используется критерий Фишера (F-критерий). Поскольку для оценок дисперсий используются суммы квадратов SS отклонений значений данной переменной от ее средней величины, то можно говорить о разложении общей суммы квадратов $SS_{общ.}$ на составляющие. Найдем эти суммы. Сначала вычислим среднее значение зависимой переменной:

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{908029{,}51}{20} = 45401{,}48$ тыс. руб. – средняя сумма вкладов физических лиц в банках в течение наблюдаемого периода.

Для расчета сумм квадратов, ошибки аппроксимации и коэффициента корреляции составим третью вспомогательную таблицу:

№	y_i	$\hat{y}_i$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(x_i - \bar{x}_i)^2$	$\left\lvert \frac{e_i}{y_i} \right\rvert$
1	36643	36917,15	76708321	71983804	130681479	0,007
2	38297	39804,09	50475507	31330749	56878747	0,039
3	38993	40934,88	41063062	19950509	36218731	0,050
4	40394	40972,43	25079810	19616439	35612250	0,014
5	41090	40865,19	18591789	20577928	37357766	0,005
6	42691	42609,09	7345117	7797416	14155654	0,002
7	43916	42075,68	2208055	11060889	20080257	0,042
8	43988	41901,64	1998250	12248832	22236883	0,047
9	44684	42889,86	514281	6308224	11452133	0,040
10	43721	43651,04	2822621	3064026	5562522	0,002
11	44198	43512,55	1447687	3568046	6477534	0,016
12	46465	48359,07	1131997	8747368	15880225	0,041
13	47481	48514,93	4325994	9693591	17598025	0,022
14	48438	50449,13	9220498	25478806	46254961	0,042

№	y_i	$\hat{y}_i$	$(y_i-\bar{y})^2$	$(\hat{y}_i-\bar{y})^2$	$(x_i-\bar{x}_i)^2$	$\left\lvert\frac{e_i}{y_i}\right\rvert$
15	49632	49740,86	17899316	18830300	34185070	0,002
16	53506	51805,62	65685795	41013002	74456189	0,032
17	52559	50663,99	51225889	27694066	50276608	0,036
18	53461	51811,33	64956191	41086231	74589132	0,031
19	49484	50056,29	16668687	21667318	39335475	0,012
20	48387	50494,7	8913798	25940925	47093906	0,044
Итого	**908029**	**908029,5**	**46828266**	**427658468**	**776383550**	**0,525**

$SS_{общ.}=\Sigma(y_i-\bar{y})^2=468282663$– величина, характеризующая разброс значений y_i относительно среднего значения $\bar{y}$. Разобьем эту сумму на две части: объясненную регрессионным уравнением и не объясненную (т. е. связанную с ошибками ε_i):

$SS_R=\Sigma(\hat{y}_i-\bar{y})^2=427658468$ – сумма квадратов, объясненная регрессией,

$SS_{ост.}=\Sigma(y_i-\hat{y}_i)^2=40624194$ – остаточная сумма квадратов, обусловленная ошибкой.

Проверка: $SS_{общ.}=SS_R+SS_{ост.}=427658468+40624194=468282663$ (верно).

Найдем коэффициент детерминации, или долей объясненной дисперсии y, называется

$$R^2=\frac{SS_R}{SS_{общ.}}=\frac{427658468}{468282663}=0{,}9132.$$

Значение коэффициента детерминации близко к 1. Это означает, 91,32% линейная регрессия y на x объясняет дисперсию y, т.е. 91,32% общей вариации вкладов физических лиц в банках объясняется их доходами. При этом остальные 8,68% приходятся на долю прочих факторов, не учтенных в уравнении регрессии. Например, к этим факторам можно отнести: уровень безработицы, инфляции, объем промышленного производства и т.д.

Далее при заданном уровне значимости $\alpha=0{,}05$ проверяем гипотезу об отсутствии линейной функциональной связи между x и y H_0: $b=0$, используя статистику критерия Фишера:

$$F_0=\frac{MS_R}{MS_{ост.}}=\frac{SS_R/1}{SS_{ост.}/(n-2)}\sim F_{кр}(\alpha,1,n-2)$$

Число степеней свободы $(1,n-2)$. MS_R и $MS_{ост.}$ обозначены средние квадраты (от англ. mean of squares), которые дают несмещенные оценки соответствующих теоретических дисперсий.

$F_{кр.}=F(\alpha;1,20-2)=4{,}41$ – находим также как в теме 4.

Составим четвертую вспомогательную таблицу

Дисперсионный анализ одномерной регрессии

Источник дисперсии	Число степеней свободы	Сумма квадратов SS	Средний квадрат MS	Критерий Фишера F_0	Критическая точка $F_{кр.} = F(0{,}05;1{,}18)$	Гипотеза $H_0 : b = 0$
Регрессор x	1	427658468	427658468	189,49	$F_{кр.} = 4{,}41$	Откло-нить
Ошибка (остаток)	18	40624194	2256899,69	–	–	–
Общая дисперсия (итог)	19	468282663	–	–	–	–

Критическая область правосторонняя:

Если при заданном уровне значимости α=0,05 наблюдаемое значение F-статистики больше критической точки $F_0 > F_{кр}$, т.е. 189,49 > 4,41. Гипотеза $H_0 : b = 0$ отвергается, то есть линейная связь между x и y есть, и результаты наблюдений не противоречат предположению о ее линейности. Полученную модель $\hat{y} = 25891{,}81 + 0{,}7422x$ в целом можно считать пригодной для дальнейшего использования.

б) *Использование элементов теории корреляции*

Другой способ верификации линейной модели состоит в использовании элементов теории корреляции. Мерой линейной связи двух величин является коэффициент корреляции. По формуле (5.18а):

$$r_B = \hat{r}_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{[n\Sigma x_i^2 - (\Sigma x_i)^2][n\Sigma y_i^2 - (\Sigma y_i)^2]}} =$$

$$= \frac{20 \cdot 24445498846 - 525738 \cdot 908029{,}51}{\sqrt{20 \cdot 14596405782 - (525738)^2} \cdot \sqrt{20 \cdot 41694162487 - (908029{,}51)^2}} = 0{,}9556$$

По формуле (5.18б):

$$r_B = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} = \frac{1222274942{,}29 - 26286{,}9 \cdot 45401{,}47565}{6230{,}50 \cdot 4838{,}82} = 0{,}9556,$$

где $\overline{xy} = \frac{\sum x_i y_i}{n} = \frac{24445498846}{20} = 1222274942{,}29$,

$$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{\frac{776383550}{20}} = 6230{,}50,$$

$$\sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}} = \sqrt{\frac{SS_{общ}}{n}} = \sqrt{\frac{468282662{,}9}{20}} = 4838{,}82.$$

По формуле (5.18в):

$$r_B = \hat{b} \cdot \frac{\sigma_x}{\sigma_y} = 0{,}7422 \cdot \frac{6230{,}50}{4838{,}82} = 0{,}9556 \cdot$$

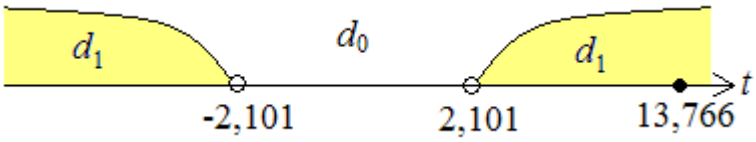
Проверка $R^2 = r_B^2$. $r_B^2 = (0{,}9556)^2 = 0{,}9132 = R^2$ (верно).

Значение коэффициента корреляции 0,9556 говорит о том, что линейная связь между доходами физических лиц и их вкладами в банках очень тесная и прямая, т.е. рост доходов населения приводит к увеличению вкладов в банках.

Проверяем гипотезу об отсутствии линейной связи между x и y $H_0 : r_B = 0$ с помощью критерия Стьюдента $t_{0r} = \frac{r_B\sqrt{n-2}}{\sqrt{1-r_B^2}} \sim t_\gamma$.

$$t_{0r} = \frac{0{,}9556 \cdot \sqrt{20-2}}{\sqrt{1-0{,}9132}} = 13{,}766$$

Критическая область двухсторонняя:



С вероятностью 0,95 гипотезу $H_0 : r_B = 0$ отвергаем, так как $|t_{0r}| > t_\gamma$, т.е. 13,766 > 2,101. Это означает, что коэффициент корреляции статистически значим.

в) *Средняя ошибка аппроксимации*

По формуле (5.20) получаем:

$A = \frac{1}{20} \cdot 0{,}525 \cdot 100 = 2{,}62\% < 10\%$ – это свидетельствует о хорошем подборе линейной модели к исходным данным.

5 этап. Интерпретация полученных показателей.

Значение коэффициента регрессии $\hat{a} = 25891{,}81$ тыс. руб. говорит о том, что если доходы физических лиц будут равны нулю, то их вклады в банках составят 25891,91 тыс. руб. Если сравнить это значение со средним 45401,48 тыс. руб., то при отсутствии доходов у населения, вклады в банках существенно сократятся.

Коэффициент регрессии $\hat{b} = 0{,}7422$ показывает, на сколько единиц увеличатся (уменьшатся) вклады физических лиц в банках при увеличении (сокращении) их доходов на 1 млн. руб.

Найдем коэффициент эластичности для данной модели:

$E = Э = \frac{\hat{b}\bar{x}}{\bar{y}} = \frac{0{,}7422 \cdot 26286{,}9}{45401{,}48} = 0{,}43\%$. (полученное значение будет сразу в %). $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{525738}{20} = 26286{,}9$ млн. руб.

Значение коэффициента эластичности приближенно показывает, что значение величины вкладов физических лиц в банках изменится на 0,43% при изменении их доходов на 1 % от среднего значения.

6 этап. Прогноз на основе линейной модели.

Точечный прогноз: $y_0 = 25891{,}81 + 0{,}7422x_0$.

Пусть прогнозное значение независимой переменной изменяется на 12% от среднего, то найдем новое значение x_0 независимой переменной:

$$x_0 = \bar{x} \cdot \left(1 + \frac{P\%}{100\%}\right) = 26286{,}9 \cdot 1{,}12 = 29441{,}33 \text{ млн. руб.}$$

При этом сумма вкладов в банках в оптимистичном случае составит:

$$y_0 = 25891{,}81 + 0{,}7422 \cdot 29441{,}33 = 47742{,}64 \text{ тыс. руб.},$$

Интервальный прогноз значения y_0 :

$$\hat{y}_0 - t_\gamma \cdot \Delta < y < \hat{y}_0 + t_\gamma \cdot \Delta .$$

Выпишем все необходимые для расчета значения:

$t_\gamma = 2{,}101$,

$$\hat{\sigma}^2 = \frac{R_{\min}}{n-2} = 2256900,$$

$$(\bar{x} - x_0)^2 = (26286{,}9 - 47742{,}64)^2 = (26286{,}9 - 43060{,}32)^2 = 9950416.$$

$$\Delta = \sqrt{2256900 \cdot \left[\frac{1}{20} + \frac{9950416}{776383550}\right]} = \sqrt{2551864{,}28} = 376{,}52.$$

$$47742{,}64 - 2{,}101 \cdot 376{,}52 < y < 47742{,}64 + 2{,}101 \cdot 376{,}52,$$

Если доходы физических лиц увеличатся на 12% от среднего уровня, то с вероятностью 95% вклады физических лиц в банках будут в пределах:

$$44386{,}50 < y < 51098{,}77.$$

5.9. Нелинейная регрессия

Многие экономические процессы наилучшим образом описываются нелинейными соотношениями, например, нелинейными функциями спроса и производственными функциями. Здесь мы рассмотрим нелинейные модели, которые с помощью преобразования переменных, сводятся к линейным, и потому для их построения могут использоваться описанные выше приемы.

В случае простого регрессионного анализа (линейного однофакторного) речь идет об уравнениях вида

$$y = a + bx, \tag{5.27}$$

состоящих из постоянной величины (которая может и отсутствовать), независимой переменной, умноженной на некоторый коэффициент, и случайной составляющей (ошибки), которой мы можем временно пренебречь. В общем случае линейное уравнение выглядит так

$$y = a + b_1 x_1 + b_2 x_2 + \dots \ . \tag{5.28}$$

Уравнения вида

$$y = a + \frac{b}{x}, \tag{5.29}$$

$$y = ax^b \tag{5.30}$$

являются нелинейными. Их графические изображения для выбранных значений a и b будут представлены кривыми.

Заметим, что уравнение (5.29) является линейным по неизвестным параметрам a и b и нелинейным по переменной x. Поэтому оценки параметров могут быть найдены по формулам (5.5) (с заменой $z_i = \frac{1}{x_i}$). Уравнение (5.29) примет вид $y = a + bz$.

Нелинейность по переменным всегда можно обойти путем использования соответствующих определений. Например, для модели вида

$$y = a + b_1 x_1^2 + b_2 \sqrt{x_2} + \dots$$

можно определить $z_1 = x_1^2$, $z_2 = \sqrt{x_2}$ и т. д., тогда модель или соотношение примет вид

$$y = a + b_1 z_1 + b_2 z_2 + \dots$$

и теперь оно является линейным как по переменным, так и по параметрам. Такой тип преобразований является лишь косметическим, он не меняет свойств оценок, полученных для линейных моделей, и обычно уравнения регрессии записываются с нелинейными выражениями относительно переменных. Это позволяет избежать лишних обозначений.

Уравнение (5.30) является нелинейным как по параметрам, так и по переменной x. Такое соотношение может быть преобразовано в линейное уравнение путем логарифмирования:

$$\ln y = \ln a + b \ln x. \tag{5.31}$$

Если обозначить $y' = \ln y$, $z = \ln x$ и $a' = \ln a$, то уравнение (5.31) можно переписать в следующем виде

$$y' = a' + bz. \qquad (5.31)$$

Процедура оценивания регрессии теперь будет следующей. Сначала вычислим y' и z для каждого наблюдения путем взятия логарифмов от исходных значений. Затем оценим регрессионную зависимость y' от z. Коэффициент при z будет представлять собой непосредственно оценку $\hat{b}$. Постоянный член является оценкой $\hat{a}'$, т. е. $\hat{\ln a}$. Для получения оценки a необходимо взять антилогарифм, т. е. вычислить $\exp(a')$.

Функции вида (5.30) часто встречаются в эконометрическом моделировании. Для таких функций эластичность y по x равна b. Действительно, если соотношение между y и x имеет вид (5.30), то эластичность

$$E = f'(x)\frac{x}{y} = abx^{b-1}\frac{x}{ax^b} = b.$$

Оценка этого коэффициента по результатам наблюдений будет показывать, на сколько процентов в среднем изменится значение y при изменении x на 1% от своего среднего значения.

5.10. Рекомендации по выполнению расчетно-графической работы по теме «Нелинейная регрессия» в MS Excel

Задание (исходными данными для расчетно-графической работы служит выборка по предыдущей теме):

1. Аппроксимировать данные темы «Линейная парная регрессия» нелинейной моделью.

2. Интерпретировать оценки параметра $\hat{b}$.

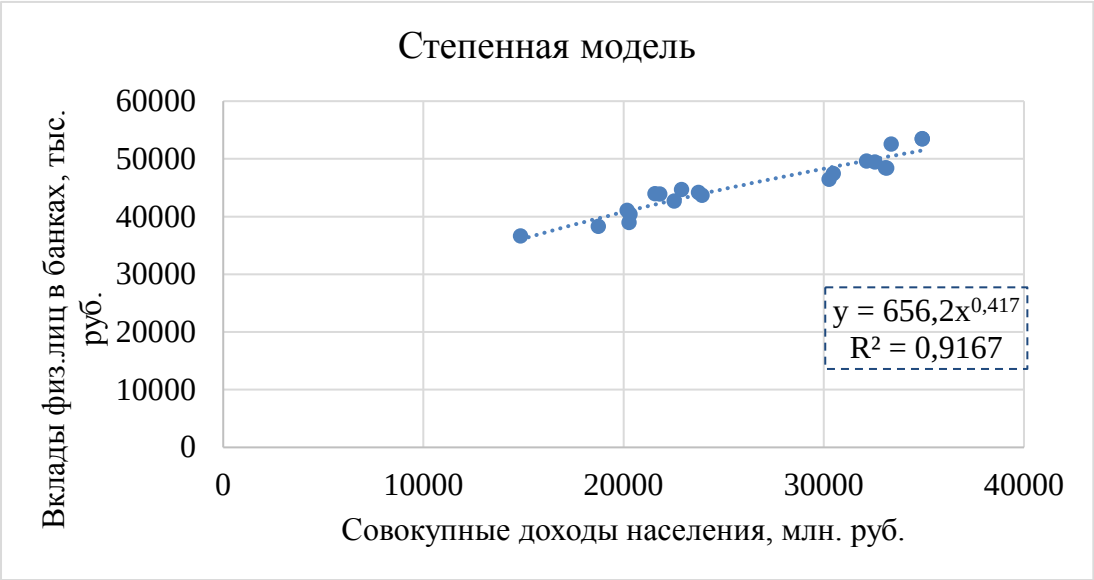
3. Определить индекс корреляции, пояснить смысл и охарактеризовать различие с линейным коэффициентом корреляции.

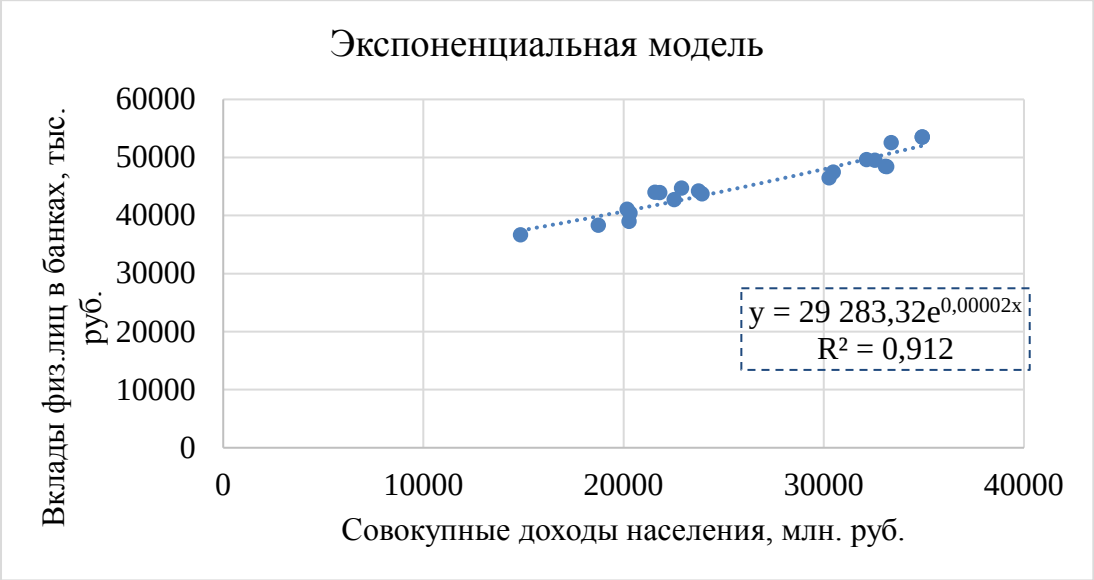
4. Рассчитать среднюю ошибку аппроксимации и коэффициент эластичности.

5. Сделать вывод о том, какая модель (линейная или нелинейная) лучше аппроксимирует исходные данные.

Пример 5.2. Используя исходные данные примера 5.1, построим графики (поле корреляции) для различных видов нелинейных моделей (рис. 5.20). Графики строятся аналогично полю корреляции. При выборе линии тренда следуют отмечать мышкой экспоненциальную, степенную или логарифмическую модели (рис. 5.20).

Следующим шагом, используя метод наименьших квадратов, найдем оценки неизвестных параметров моделей. Расчет в MS Excel здесь не приводится, так как он аналогичен вычислениям в теме «Линейная парная регрессия».





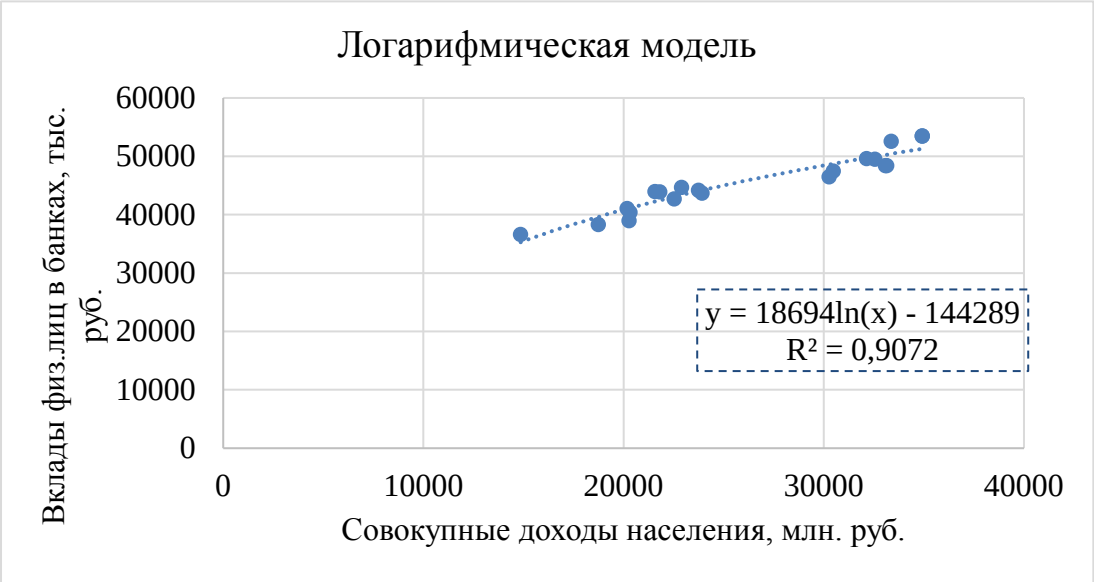


Рис. 5.20. Построение поля корреляции для нелинейных моделей

А) Степенная модель $y = ax^b$

Для нахождения оценок параметров степенной модели приведем ее к линейному виду с помощью логарифмирования:

$\ln y = \ln(ax^b) = \ln a + \ln x^b = \ln a + b\ln x.$

Неизвестные коэффициенты линеаризованной модели найдем по формулам:

$$\ln\hat{a} = \frac{\sum\left(\ln x_i\right)^2 \sum \ln y_i - \sum \ln x_i \sum \ln x_i \ln y_i}{k}, \qquad \hat{b} = \frac{n\sum \ln x_i \ln y_i - \sum \ln x_i \sum \ln y_i}{k},$$

где $k = n\sum\left(\ln x_i\right)^2 - \left(\sum \ln x_i\right)^2$.

Значение параметра a для степенной модели найдем по формуле: $\hat{a} = e^{\ln\hat{a}}$.

Составим вспомогательную таблицу:

номер наблюдения	Совокупные доходы физ. лиц, млн. руб.	Вклады физ. лиц в банках, тыс. руб.					
	x_i	y_i	$\ln x_i$	$\ln y_i$	$(\ln x_i)^2$	$(\ln y_i)^2$	$\ln x_i \cdot \ln y_i$
1	14855,30	36643	9,61	10,51	92,28	110,44	100,95
2	18745,10	38297	9,84	10,55	96,80	111,37	103,83
3	20268,70	38993	9,92	10,57	98,34	111,75	104,83
4	20319,30	40394	9,92	10,61	98,39	112,50	105,21
5	20174,80	41090	9,91	10,62	98,25	112,86	105,30
6	22524,50	42691	10,02	10,66	100,45	113,67	106,86
7	21805,80	43916	9,99	10,69	99,80	114,28	106,79
8	21571,30	43988	9,98	10,69	99,58	114,31	106,69
9	22902,80	44684	10,04	10,71	100,78	114,65	107,49
10	23928,40	43721	10,08	10,69	101,66	114,18	107,74
11	23741,80	44198	10,07	10,70	101,51	114,41	107,77
12	30271,90	46465	10,32	10,75	106,46	115,49	110,88
13	30481,90	47481	10,32	10,77	106,60	115,95	111,18
14	33088,00	48438	10,41	10,79	108,30	116,38	112,27
15	32133,70	49632	10,38	10,81	107,70	116,91	112,21
16	34915,70	53506	10,46	10,89	109,43	118,54	113,89
17	33377,50	52559	10,42	10,87	108,49	118,15	113,21
18	34923,40	53461	10,46	10,89	109,43	118,52	113,88
19	32558,70	49484	10,39	10,81	107,97	116,84	112,32
20	33149,40	48387	10,41	10,79	108,34	116,36	112,28
сумма	**525738,00**	**908029,51**	**202,95**	**214,35**	**2060,56**	**2297,56**	**2175,59**

Вычислим значения $k, \ln\hat{a}, \hat{b}, \hat{a}$: $\quad k = 20 \cdot 2060{,}56 - (202{,}95)^2 = 24{,}313$.

$$\ln\hat{a} = \frac{2060{,}56 \cdot 214{,}35 - 202{,}95 \cdot 2175{,}59}{24{,}313} = 6{,}486,$$

$$\hat{b} = \frac{20 \cdot 2175{,}59 - 202{,}95 \cdot 214{,}35}{24{,}313} = 0{,}417. \qquad \hat{a} = e^{6{,}486} = 656{,}2.$$

Степенная модель, описывающая зависимость между вкладами физических лиц в банках и совокупными доходами населения, примет вид:

$$y = 656{,}2 \cdot x^{0{,}417}.$$

Коэффициент эластичности:

$E = \frac{y'}{y} \cdot x = \frac{(a \cdot x^b)'}{a \cdot x^b} \cdot x = \frac{a \cdot b \cdot x^{b-1}}{a \cdot x^b} \cdot x = \hat{b} = 0{,}417$, показывает, что если совокупные доходы населения изменятся на 1%, то вклады физических лиц в банках изменятся на 0,417%.

Индекс корреляции:

$$\eta_{степ} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_{i,степ} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

Составим вспомогательную таблицу для расчета индекса корреляции. Значения $\tilde{y}_i$ найдем по формуле: $\tilde{y}_i = 656{,}2 \cdot x_i^{0{,}417}$.

$\bar{y} = 45401{,}48$ тыс. руб. (найдено в примере 5.1).

	$\hat{y}_{i,степ}$	$e_i = y_i - \hat{y}_{i,степ}$	$\left(\hat{y}_{i,степ} - \bar{y}\right)^2$	$\left\lvert \frac{e_i}{y_i} \right\rvert$
	36023,2	619,98	87952679	0,02
	39691,6	-1394,72	32602901	0,04
	41006,2	-2012,77	19318434	0,05
	41048,9	-655,35	18945300	0,02
	40926,9	162,77	20021989	0,00
	42850,8	-159,51	6505952	0,00
	42275,3	1640,22	9772977	0,04
	42085,1	1902,75	10998112	0,04
	43149,4	1534,91	5071722	0,03
	43944,8	-223,43	2121783	0,01
	43801,6	396,65	2559529	0,01
	48472	-2006,61	9428355	0,04
	48612	-1130,59	10307238	0,02
	50303,6	-1865,61	24030910	0,04
	49693,5	-61,26	18421446	0,00
	51444,1	2062,07	36513156	0,04
	50486,7	2072,04	25859119	0,04
	51448,8	2012,20	36570344	0,04
	49966,5	-482,29	20839434	0,01
	50342,5	-1955,44	24413824	0,04
сумма	**907573**	**–**	**422255205**	**0,53**

$\sum_{i=1}^{n}\left(\hat{y}_{i,степ} - \bar{y}\right)^2 = 422255205.$

$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 468282663$ (значение найдено на четвертом этапе в теме «Линейная парная регрессия»).

$\eta_{степ} = \sqrt{\frac{422255205}{468282663}} = 0{,}9496$ – степенная связь между вкладами физических лиц в банках и совокупными доходами населения сильная.

По формуле (5.20) получаем:

$A = \frac{1}{20} \cdot 0{,}53 \cdot 100 = 2{,}65\% < 7\%$ – это свидетельствует о хорошем подборе степенной модели к исходным данным.

Б) Экспоненциальная модель $y = ae^{bx}$

Для нахождения оценок параметров экспоненциальной модели приведем ее к линейному виду с помощью логарифмирования:

$\ln y = \ln(ae^{bx}) = \ln a + \ln e^{bx} = \ln a + bx.$

Неизвестные коэффициенты линеаризованной модели найдем по формулам:

$$\ln \hat{a} = \frac{\sum x_i^2 \sum \ln y_i - \sum x_i \sum x_i \ln y_i}{k}, \qquad \hat{b} = \frac{n\sum x_i \ln y_i - \sum x_i \sum \ln y_i}{k},$$

где $k = n\sum x_i^2 - \left(\sum x_i\right)^2$.

Значение параметра a для степенной модели найдем по формуле: $\hat{a} = e^{\ln \hat{a}}$.

Составим вспомогательную таблицу:

№	x_i	y_i	$\ln y_i$	x_i^2	$(\ln y_i)^2$	$x_i \cdot \ln y_i$
1	14855,30	36643	10,51	220679938,1	110,44	156114,08
2	18745,10	38297	10,55	351378774,0	111,37	197819,35
3	20268,70	38993	10,57	410820199,7	111,75	214263,44
4	20319,30	40394	10,61	412873952,5	112,50	215515,11
5	20174,80	41090	10,62	407022555,0	112,86	214327,22
6	22524,50	42691	10,66	507353100,3	113,67	240150,59
7	21805,80	43916	10,69	475492913,6	114,28	233104,51
8	21571,30	43988	10,69	465320983,7	114,31	230633,21
9	22902,80	44684	10,71	524538247,8	114,65	245228,95
10	23928,40	43721	10,69	572568326,6	114,18	255689,15
11	23741,80	44198	10,70	563673067,2	114,41	253952,76
12	30271,90	46465	10,75	916387929,6	115,49	325315,88
13	30481,90	47481	10,77	929146227,6	115,95	328231,93
14	33088,00	48438	10,79	1094815744,0	116,38	356954,67
15	32133,70	49632	10,81	1032574675,7	116,91	347442,28
16	34915,70	53506	10,89	1219106106,5	118,54	380146,50
17	33377,50	52559	10,87	1114057506,3	118,15	362802,94
18	34923,40	53461	10,89	1219643867,6	118,52	380200,86

№	x_i	y_i	$\ln y_i$	x_i^2	$(\ln y_i)^2$	$x_i \cdot \ln y_i$
19	32558,70	49484	10,81	1060068945,7	116,84	351940,30
20	33149,40	48387	10,79	1098882720,4	116,36	357582,18
сумма	**525738,00**	**908029,51**	**214,35**	**14596405781,8**	**2297,56**	**5647415,90**

Вычислим значения $k, \ln\hat{a}, \hat{b}, \hat{a}$:

$$k = 20 \cdot 14596405781{,}8 - (525738{,}00)^2 = 15527670992.$$

$$\ln\hat{a} = \frac{14596405781{,}8 \cdot 214{,}35 - 525738{,}0 \cdot 5647415{,}9}{15527670992} = 10{,}285,$$

$$\hat{b} = \frac{20 \cdot 5647415{,}9 - 525738{,}0 \cdot 214{,}35}{15527670992} = 0{,}000016,$$

$$\hat{a} = e^{10{,}285} = 29283{,}32.$$

Экспоненциальная модель, описывающая зависимость между вкладами физических лиц в банках и совокупными доходами населения, примет вид:

$$y = 29283{,}32 \cdot e^{0{,}000016x}.$$

Коэффициент эластичности:

$E = \frac{(a \cdot e^{bx})'}{a \cdot e^{bx}} \cdot x = \frac{a \cdot b \cdot e^{bx}}{a \cdot e^{bx}} \cdot x = \hat{b}\bar{x} = 0{,}000016 \cdot 26286{,}9 = 0{,}433$, показывает, что если совокупные доходы населения изменятся на 1%, то вклады физических лиц в банках изменятся на 0,433%.

Индекс корреляции:

$$\eta'_{эксп} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_{i,эксп} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

Составим вспомогательную таблицу для расчета индекса корреляции.

Значения $\hat{y}_{i,эксп}$ найдем по формуле: $\hat{y}_{i,эксп} = 29283б32 \cdot e^{0{,}000016x_i}$.

$\bar{y} = 45401{,}48$ тыс. руб.

$\hat{y}_{i,эксп}$	$e_i = y_i - \hat{y}_{i,эксп}$	$\left(\hat{y}_{i,эксп} - \bar{y}\right)^2$	$\left\lvert\frac{e_i}{y_i}\right\rvert$
37397,37	-754,22	64065711	0,02
39870,76	-1573,90	30588814	0,04
40883,57	-1890,14	20411483	0,05
40917,64	-524,14	20104757	0,01
40820,41	269,24	20986147	0,01
42430,54	260,75	8826451	0,01
41931,42	1984,10	12041258	0,05
41769,84	2218,04	13188749	0,05
42695,64	1988,70	7321533	0,04
43422,71	298,70	3915500	0,01

	43289,51	908,76	4460386	0,02
	48203,13	-1737,70	7849270	0,04
	48370,08	-888,70	8812621	0,02
	50490,70	-2052,69	25900171	0,04
	49703,59	-71,36	18508208	0,00
	52033,14	1473,01	43979029	0,03
	50731,93	1826,77	28413753	0,03
	52039,74	1421,27	44066570	0,03
	50052,60	-568,40	21632999	0,01
	50541,76	-2154,69	26422565	0,04
сумма	**907596,12**	**–**	**431495974**	**0,548**

$\sum_{i=1}^{n}\left(\hat{y}_{i,эксп}-\bar{y}\right)^2 = 431495974.$

$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 = 468282663$ (значение найдено на четвертом этапе в теме «Линейная парная регрессия»).

$\eta_{эксп} = \sqrt{\frac{431495974}{468282663}} = 0{,}9599$ – экспоненциальная связь между вкладами физических лиц в банках и совокупными доходами населения сильная.

По формуле (5.20) получаем:

$A = \frac{1}{20}\cdot 0{,}548\cdot 100 = 2{,}74\% < 15\%$ – это свидетельствует о хорошем подборе экспоненциальной модели к исходным данным.

В) Логарифмическая модель $y = a + b\ln x$

Неизвестные коэффициенты логарифмической модели найдем по формулам:

$$\hat{a} = \frac{\sum\left(\ln x_i\right)^2\sum y_i - \sum\ln x_i\sum\left(\ln x_i\right)\cdot y_i}{k}, \qquad \hat{b} = \frac{n\sum\left(\ln x_i\right)\cdot y_i - \sum\ln x_i\sum y_i}{k},$$

где $k = n\sum\left(\ln x_i\right)^2 - \left(\sum\ln x_i\right)^2$.

Составим вспомогательную таблицу:

№	x_i	y_i	$\ln x_i$	$(\ln x_i)^2$	$\ln x_i \cdot y_i$
1	14855,30	36643	9,61	92,28	351998,17
2	18745,10	38297	9,84	96,80	376790,88
3	20268,70	38993	9,92	98,34	386691,33
4	20319,30	40394	9,92	98,39	400676,32
5	20174,80	41090	9,91	98,25	407288,46
6	22524,50	42691	10,02	100,45	427867,41
7	21805,80	43916	9,99	99,80	438713,06
8	21571,30	43988	9,98	99,58	438960,30
9	22902,80	44684	10,04	100,78	448586,76

10	23928,40	43721	10,08	101,66	440835,16
11	23741,80	44198	10,07	101,51	445297,31
12	30271,90	46465	10,32	106,46	479429,14
13	30481,90	47481	10,32	106,60	490239,93
14	33088,00	48438	10,41	108,30	504090,71
15	32133,70	49632	10,38	107,70	515066,48
16	34915,70	53506	10,46	109,43	559711,38
17	33377,50	52559	10,42	108,49	547432,38
18	34923,40	53461	10,46	109,43	559251,00
19	32558,70	49484	10,39	107,97	514180,48
20	33149,40	48387	10,41	108,34	503650,40
сумма	**525738,00**	**908029,51**	**202,95**	**2060,56**	**9236757,07**

Вычислим значения $k, \hat{a}, \hat{b}$:

$$k = 20 \cdot 2060{,}56 - (202{,}95)^2 = 24{,}313.$$

$$\hat{a} = \frac{2060{,}56 \cdot 908029{,}51 - 202{,}95 \cdot 9236757{,}07}{24{,}313} = -144288{,}7,$$

$$\hat{b} = \frac{20 \cdot 9236757{,}07 - 202{,}95 \cdot 908029{,}51}{24{,}313} = 18693{,}7.$$

Логарифмическая модель, описывающая зависимость между вкладами физических лиц в банках и совокупными доходами населения, примет вид:

$$y = -144288{,}7 + 18693{,}7 \ln x.$$

Коэффициент эластичности:

$E = \frac{(a + b\ln x)'}{a + b\ln x} \cdot x = \frac{b \cdot \frac{1}{x}}{a + b\ln x} \cdot x = \frac{\hat{b}}{\bar{y}} = 0{,}412$, показывает, что если совокупные доходы населения изменятся на 1%, то вклады физических лиц в банках изменятся на 0,412%.

Индекс корреляции:

$$\eta_{лог} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_{i,лог} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

Составим вспомогательную таблицу для расчета индекса корреляции.

Значения $\hat{y}_i$ найдем по формуле: $y_i = -144288{,}7 + 18693{,}7 \cdot \ln x_i$.

$\bar{y} = 45401{,}48$ тыс. руб.

$\hat{y}_{i,лог}$	$e_i = y_i - \hat{y}_{i,лог}$	$\left(\hat{y}_{i,лог} - \bar{y}\right)^2$	$\left\lvert \frac{e_i}{y_i} \right\rvert$
35285,00	1358,15	102343149	0,04
39632,69	–1335,83	33278848	0,03
41093,52	–2100,09	18558484	0,05

	41140,13	–746,63	18159070	0,02
	41006,72	82,94	19313919	0,00
	43066,19	–374,90	5453579	0,01
	42459,99	1455,53	8652315	0,03
	42257,87	1730,01	9882238	0,04
	43377,54	1306,80	4096317	0,03
	44196,45	-475,04	1452083	0,01
	44050,10	148,18	1826212	0,00
	48592,34	-2126,91	10181639	0,05
	48721,58	-1240,20	11023070	0,03
	50255,16	-1817,16	23558271	0,04
	49708,08	-75,85	18546880	0,00
	51260,24	2245,91	34325178	0,04
	50418,01	2140,69	25165607	0,04
	51264,37	2196,65	34373495	0,04
	49953,71	-469,50	20722804	0,01
	50289,82	-1902,74	23895900	0,04
сумма	**–**	**–**	**424809058**	**0,555**

$\sum_{i=1}^{n}\left(\hat{y}_{i,лог} - \bar{y}\right)^2 = 424809058.$

$SS_{общ} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 468282663$ (значение найдено на четвертом этапе в теме «Линейная парная регрессия»).

$\eta_{лог} = \sqrt{\frac{424809058}{468282663}} = 0{,}952$ – логарифмическая связь между вкладами физических лиц в банках и совокупными доходами населения сильная.

По формуле (5.20) получаем:

$A = \frac{1}{20} \cdot 0{,}555 \cdot 100 = 2{,}77\% < 15\%$ – это свидетельствует о хорошем подборе логарифмической модели к исходным данным.

Сравнение моделей:

Показатель	Линейная	Степенная	Логарифмическая	Экспоненциальная
Коэффициент (индекс) корреляции	0,955	0,9496	0,952	0,9599
Средняя ошибка аппроксимации	2,72	2,65	2,77	2,74
Эластичность	0,43	0,417	0,412	0,433

Из сравнительной таблицы видно, что все четыре модели хорошо аппроксимируют исходный данные.

6. Многомерная регрессионная модель

6.1. Спецификация модели

Построение модели множественной регрессии начинается с решения вопроса о спецификации модели. Суть проблемы спецификации включает в себя два вопроса: отбор факторов и выбор вида уравнения регрессии. Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям:

1. Они должны быть количественно измеримы. Если есть необходимость включить в модель качественный фактор, то ему нужно придать количественную определенность.

2. Факторы не должны быть интеркоррелированы (корреляция между объясняющими переменными) и тем более находиться в функциональной связи.

Отбор факторов производится на основе качественного теоретико-экономического анализа. Однако теоретический анализ часто не позволяет однозначно ответить на вопрос о количественной взаимосвязи рассматриваемых признаков и целесообразности включения какого-либо фактора в модель. Поэтому отбор факторов обычно осуществляется в две стадии: на первой подбираются факторы исходя из сущности проблемы; на второй – на основе матрицы коэффициентов корреляции определяют t-статистики для параметров регрессии. Стандартные компьютерные программы обработки регрессионного анализа позволяют перебрать различные функции (линейные и нелинейные) и выбрать ту из них, для которой остаточная дисперсия и ошибка аппроксимации минимальны, а коэффициент детерминации максимален. Если исследователя не устраивает предлагаемый стандартной программой набор функций, то он может использовать любые другие функции, предварительно линеаризовав их с помощью замены переменных или логарифмических преобразований.

6.2. Линейная модель множественной регрессии

Экономические явления, как правило, определяются большим числом одновременно и совокупно действующих факторов. В связи с этим часто возникает задача исследования зависимости одной переменной Y от нескольких объясняющих переменных $X_1, X_2, \ldots, X_n$. Обозначим $i-$е наблюдение зависимой переменной y_i, объясняющие переменные $- x_{i1}, x_{i2}, \ldots, x_{ip}$. Тогда модель множественной линейной регрессии можно записать в виде

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + \varepsilon_i, \tag{6.1}$$

где $i = \overline{1,n}$, ε_i удовлетворяет следующим предпосылкам:

1. $M\varepsilon_i = 0, \quad M\left(\varepsilon_i^2\right) = D\left(\varepsilon_i\right) = \sigma^2 \quad \forall i = \overline{1,n},$
2. $M(\varepsilon_i \cdot \varepsilon_j) = 0 \quad \forall i \neq j,$
3. $\varepsilon_i \sim N(0, \sigma^2)$.

Включение в регрессионную модель новых объясняющих переменных усложняет вычисления. Это приводит к целесообразности использования матричных обозначений. Введем следующие обозначения:

$Y = (y_1 \dots y_n)^t$ – матрица-столбец или вектор значений зависимой переменной,

$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$ – матрица значений объясняющих переменных,

$b = (b_0 \dots b_p)^t$ – матрица-столбец или вектор параметров,

$\varepsilon = (\varepsilon_1 \dots \varepsilon_n)^t$ – матрица-столбец или вектор случайных компонент, тогда в матричной форме модель (1) примет вид:

$$Y = Xb + \varepsilon \qquad (6.2),$$

которая удовлетворяет следующим предпосылкам:

1. ε – случайный вектор, X – неслучайная (детерминированная) матрица,
2. $M(\varepsilon) = 0, \quad V(\varepsilon) = M(\varepsilon \cdot \varepsilon^t) = \sigma^2 E,$ где E – единичная матрица размерности $n \times n$,
3. ε – нормально распределенный случайный вектор, то есть $\varepsilon \sim N(0, \sigma^2 E),$
4. $r(X) = p + 1 < n.$

Модель (6.2), удовлетворяющая предпосылкам 1–4, называется классической нормальной линейной моделью множественной регрессии, если не выполняется 3-я предпосылка – то классической линейной моделью множественной регрессии.

6.3. Оценка параметров линейной модели

Как и в случае регрессионного уравнения с одной переменной для оценки вектора неизвестных параметров $b = (b_0 \dots b_p)^t$ используется метод наименьших квадратов. Условие минимизации остаточной суммы квадратов запишется в виде:

$$R = (Y - Xb)^t (Y - Xb) \to \min.$$

Необходимые условия экстремума дают систему нормальных уравнений:

$$X^t Y = X^t X b.$$

Решением этого уравнения является вектор

$$\hat{b} = (X^t X)^{-1} X^t Y \qquad (6.3)$$

Полученные оценки обладают свойствами несмещенность, состоятельность и эффективность. Для определения статистической значимости этих оценок потребуется матрица вариации оценок

$$V(\hat{b}) = \sigma^2 (X^t X)^{-1}$$

и несмещенная оценка дисперсии

$$\hat{\sigma}^2 = S^2 = \frac{R_{\min}}{n-p}.$$

Эти формулы позволяют записать оценку матрицы вариаций

$$V(\hat{b}) = (X^t X)^{-1} \frac{R_{\min}}{n-p}$$

и оценки дисперсий МНК-оценок неизвестных параметров модели

$$\hat{D}(\hat{b}_i) = \hat{V}_{ii} = (X^t X)^{-1}_{ii} \frac{R_{\min}}{n-p}.$$

МНК-оценки (6.3) обладают также наименьшей дисперсией в классе линейных несмещенных оценок, т.е. являются наиболее эффективными (теорема Гаусса-Маркова).

6.4. Построение доверительных интервалов и проверка статистических гипотез

Статистический анализ значимости коэффициентов регрессии для нормальной модели проводят с помощью построения доверительных интервалов и проверок статистических гипотез. Для построения доверительных интервалов неизвестных параметров можно использовать формулы:

$$\hat{b}_i - t_\gamma \sqrt{\hat{D}(\hat{b}_i)} < b_i < \hat{b}_i + t_\gamma \sqrt{\hat{D}(\hat{b}_i)}, \qquad (6.4)$$

где $\hat{b}_i$ – точечная оценка неизвестного параметра b_i, $t_\gamma = t(\gamma; n-p)$ – квантиль распределения Стьюдента определяется по таблице и зависит от доверительной вероятности γ, числа наблюдений n и числа параметров модели p. Эта интервальная оценка покроет истинное неизвестное значение параметра b_i с доверительной вероятностью или надежностью $\gamma = 1 - \alpha$.

Гипотеза $H_0 : b_i = b_{i0}$ будет принята на уровне значимости α, если соответствующий доверительный интервал содержит гипотетически заданное значение b_{i0}. Эту гипотезу можно проверить также с помощью критерия Стьюдента. Для этого нужно сравнить наблюдаемое значение критерия $t_0 = \frac{\hat{b}_i - b_{i0}}{\sqrt{\hat{D}(\hat{b}_i)}}$ с критической точкой $t_{кр} = t(1-\alpha; n-p)$.

Проверка значимости коэффициентов регрессии или значимости влияния регрессоров – это проверка гипотез $H_0 : b_i = 0$. Регрессор считается статистически незначимым, если это гипотеза принимается, т.е. доверительный интервал для соответствующего коэффициента регрессии покрывает нуль.

6.5. Верификация модели (проверка ее пригодности и адекватности)

Качество построенной множественной регрессии можно оценить с помощью дисперсионного анализа и коэффициента детерминации R^2.

Общая сумма квадратов $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ разбивается на две части: ESS – дисперсия, которая объясняется уравнением регрессии и RSS – остаточная дисперсия, которая объясняется ошибками измерений или другими неучтенными в модели факторами, эти дисперсии определяются по формулам

$$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \; RSS = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2. \qquad (6.5)$$

Гипотеза об отсутствии линейной функциональной связи между объясняемой переменной y и регрессорами $x_1,\ldots,x_p$ записывается в виде $H_0 : b_1 = \ldots = b_p = 0$. Для проверки этой гипотезы используется критерий, статистика которого

$$F_0 = \frac{\sum(\hat{y}_i - \bar{y})^2 / (p-1)}{\sum(y_i - \hat{y}_i)^2 / (n-p)}$$

имеет распределение Фишера с числами степеней свободы $p-1$ и $n-p$, где n – количество наблюдений, p – количество оцениваемых параметров модели. Если $F_0 > F_{кр}(1-\alpha; p-1; n-p)$, то гипотеза $H_0 : b_1 = \ldots = b_p = 0$ отвергается на уровне значимости α, уравнение в целом значимо и оцененная линейная множественная регрессия $\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \ldots + \hat{b}_p x_p$ адекватна наблюдениям, т.е. пригодна для описания зависимости между y и $x_1,\ldots,x_p$.

Для удобства результаты дисперсионного анализа вносят в таблицу

Таблица 6.1

Источник дисперсии	Число степеней свободы	Сумма квадратов	Критерий Фишера	Критическая точка	Гипотеза
Факторы $x_1,\ldots,x_p$	$p-1$	ESS	$F_0 = \frac{ESS/(p-1)}{RSS/(n-p)}$	$F_{кр}(1-\alpha; p-1; n-p)$	$H_0 : b_1 = \ldots = b_p = 0$
Ошибки (остатки)	$n-p$	RSS	-	-	
Общая дисперсия (итог)	$n-1$	TSS	-	-	-

Коэффициент детерминации определяется по формуле

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}, \qquad (6.6)$$

он принимает значения от 0 до 1 и показывает качество подгонки модели регрессии к значениям y_i. Если R^2 = 0, то регрессия ничего не дает в плане предска-

зания по сравнению с тривиальным $\bar{y}$. Если $R^2 = 1$, то это означает точную подгонку, т.е. все наблюдаемые значения лежат на регрессионной плоскости. Коэффициент детерминации R^2 характеризует долю зависимой переменной, обусловленной регрессией или изменчивостью объясняющих переменных, чем ближе R^2 к 1, тем лучше качество модели, т.е. регрессия лучше описывает зависимость между факторами и результирующим показателем.

Вместе с тем использование только одного коэффициента детерминации для выбора наилучшего уравнения регрессии может оказаться недостаточным. Недостатком коэффициента детерминации является то, что он увеличивается при добавлении новых факторов в модель. При соответствующем соотношении количества факторов и числа наблюдений можно добиться того, что $R^2 = 1$, однако это вовсе не будет означать хорошее качество модели. Для того, чтобы устранить этот эффект используют скорректированный коэффициент детерминации

$$\overline{R}^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}.$$

Он может уменьшаться при введении в модель новых факторов, не оказывающих существенного влияния на результирующий показатель. Однако даже увеличение скорректированного коэффициента детерминации при введении в модель новой переменной не всегда означает, что ее коэффициент регрессии значим. То есть увеличение скорректированного коэффициента детерминации не означает улучшение качества регрессионной модели.

F – статистика может быть найдена с учетом коэффициента детерминации по формуле $F = \dfrac{R^2}{1-R^2} \cdot \dfrac{n-p}{p-1}$.

6.6. Интерпретация коэффициентов множественной регрессии

Множественный регрессионный анализ позволяет разграничить влияние независимых переменных, допуская при этом возможность их коррелированности. Коэффициент регрессии при каждой переменной x_i дает оценку ее влияния на величину y в случае неизменности влияния на нее всех остальных факторов. Так, например, в оцененной линейной регрессии $\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2$ коэффициенты $\hat{b}_1$ и $\hat{b}_2$ являются показателями силы связи, характеризующими абсолютное (в натуральных единицах измерения) изменение объясняемой переменной y при изменении каждого из факторов x_1 и x_2 соответственно на единицу своего измерения при фиксированном влиянии второй переменной.

Пример 6.1. Зависимость расходов на продукты питания по совокупности семей характеризуется уравнением: $\hat{y} = 0{,}5 + 0{,}35x_1 + 0{,}73x_2$, где y – расходы семьи за месяц на продукты питания (тыс. руб.), x_1 – месячный доход на одного члена семьи (тыс. руб.), x_2 – размер семьи (чел.). Анализ этого уравнения позво-

ляет сделать выводы: с ростом дохода на одного члена семьи на 1 тыс. руб. расходы на питание возрастут в среднем на 350 руб. при том же составе семьи. Т.е. 35% дополнительных семейных расходов тратится на продукты питания. Увеличение размера семьи при тех же доходах на 1 человека приведет к увеличению расходов на 730 руб.

При изучении вопросов потребления коэффициенты регрессии рассматриваются как характеристики предельной склонности к потреблению. Например, если функция потребления имеет вид

$$C_t = b_0 + b_1 R_t + b_2 R_{t-1} + \varepsilon,$$

то потребление C_t в период времени t зависит от дохода того же периода R_t и дохода предыдущего периода R_{t-1}. Соответственно коэффициент b_1 характеризует эффект единичного возрастания дохода R_t при неизменном уровне предыдущего дохода. Коэффициент b_1 обычно называют краткосрочной предельной склонностью к потреблению. Общим эффектом возрастания как текущего, так и предыдущего дохода будет рост потребления на $b = b_1 + b_2$. Коэффициент b рассматривается как долгосрочная склонность к потреблению. Так как коэффициенты $b_1, b_2 > 0$, то долгосрочная склонность к потреблению b должна превосходить краткосрочную b_1. Например, за период 1905-1951 гг. (за исключением военных лет) М. Фридман построил для США следующую функцию потребления: $C_t = 53 + 0{,}58 R_t + 0{,}32 R_{t-1}$ с краткосрочной предельной склонностью к потреблению 0,58 и долгосрочной склонностью к потреблению 0,9.

Функция потребления может рассматриваться также в зависимости от прошлых привычек потребителя, т.е. от предыдущего уровня потребления C_{t-1}: $C_t = b_0 + b_1 R_t + b_2 C_{t-1} + \varepsilon.$

В этом уравнении параметр b_1 так же характеризует краткосрочную предельную склонность к потреблению, т.е. влияние на потребление единичного роста доходов того же периода R_t. Долгосрочную предельную склонность к потреблению здесь измеряет выражение $\frac{b_1}{1-b_2}$. Например, если уравнение регрессии имеет вид: $C_t = 23{,}4 + 0{,}46 R_t + 0{,}2 C_{t-1}$, то краткосрочная предельная склонностью к потреблению равна 0,46, а долгосрочная $\frac{b_1}{1-b_2} = \frac{0{,}46}{1-0{,}2} = 0{,}575.$

Относительными показателями силы связи в уравнении множественной регрессии являются частные коэффициенты эластичности:

$$E_{yx_i} = \hat{b}_i \frac{\bar{x}_i}{\bar{y}}, \tag{6.7}$$

где $\bar{x}_i$ и $\bar{y}$ – выборочные средние величины объясняющей переменной x_i и результирующего показателя y соответственно, значения которых подсчитаны в результате статистического анализа регрессионной модели. Эластичность E_{yx_i}

показателя y по переменной x_i примерно определяет, на сколько процентов изменится значение y от своего среднего уровня при увеличении переменной x_i на 1% от своего среднего уровня.

Пример. 6.2. По ряду регионов была построена множественная регрессия величины импорта на определенный товар y относительно отечественного производства этого товара x_1, изменения запасов x_2 и потребления на внутреннем рынке x_3: $\hat{y} = -66{,}028 + 0{,}135x_1 + 0{,}476x_2 + 0{,}343x_3$. При этом были рассчитаны показатели частной эластичности $E_{yx_1} = 1{,}053\%$, $E_{yx_2} = 0{,}056\%$, $E_{yx_3} = 1{,}987\%$. Т.е. с ростом величины отечественного производства на 1% размер импорта в среднем по совокупности регионов возрастет на 1,053% при неизменных запасах и потреблении семей; с ростом изменения запасов на 1% при неизменном производстве и внутреннем потреблении величина импорта увеличивается в среднем на 0,056%; при неизменном объеме производства и величины запасов с увеличением внутреннего потребления на 1% импорт товара возрастает в среднем по совокупности регионов на 1,987%.

Показатели эластичности можно сравнивать друг с другом и ранжировать факторы по силе их воздействия на результирующий показатель. В рассматриваемом примере наибольшее воздействие на величину импорта оказывает размер внутреннего потребления товара x_3, а наименьшее – изменение запасов x_2.

6.7. Прогноз на основе множественной линейной регрессии

Прогноз на основе модели множественной линейной регрессии может быть точечным и интервальным. Если задан набор объясняющих переменных $x^0 = (x_1^0, x_2^0, \ldots, x_k^0)$, то точечный прогноз получится подстановкой прогнозных значений регрессоров в уравнение модели. Для получения интервального прогноза сначала нужно рассчитать оценку дисперсии оценки прогнозируемой величины $\hat{D}(\hat{y}_0) = \hat{\sigma}^2(1 + x^0(X^t X)^{-1} x^{0^t})$, где $\hat{\sigma}^2 = S^2 = \dfrac{R_{\min}}{n-p}$.

С надежность γ можно утверждать, что истинное прогнозируемое значение y_0 покроется доверительным интервалом, который можно построить по формуле

$$\hat{y}_0 - t_\gamma \sqrt{\hat{D}(\hat{y}_0)} < y_0 < \hat{y}_0 + t_\gamma \sqrt{\hat{D}(\hat{y}_0)}, \qquad (6.8)$$

где $t_\gamma = t\left(\dfrac{1+\gamma}{2}, n-p\right)$ – квантиль распределения Стьюдента.

6.8. Мультиколлинеарность факторов

При построении эконометрической модели предполагается, что независимые переменные воздействуют на зависимую изолированно, т.е. влияние отдельной переменной на результативный признак не связано с влиянием других переменных. В реальной экономической ситуации все явления в той или иной мере

связаны, поэтому добиться выполнения этого предположения практически невозможно.

Различают функциональные и стохастические связи между объясняющими переменными. В первом случае говорят об ошибках спецификации модели, которые должны быть исправлены. Функциональная связь столбцов матрицы X приведет к невозможности нахождения единственной МНК-оценки вектора $\hat{b}$, что формально следует из вырожденности матрицы $X^t X$ и невозможности решить систему нормальных уравнений.

Более часто между объясняющими переменными наблюдается стохастическая связь, что приводит к уменьшению определителя матрицы $X^t X$: чем сильнее связь, тем меньше будет определитель. Это приводит к росту МНК-оценок параметров и их стандартных ошибок.

Корреляционная связь может существовать как между двумя объясняющими переменными (интеркорреляция), так и между несколькими (мультиколлинеарность). Существует несколько признаков, указывающих на наличие мультиколлинеарности факторов:

- Не соответствующие положениям экономической теории знаки коэффициентов регрессии;
- Значительные изменения параметров модели при небольшом изменении количества наблюдений;
- Незначимость параметров модели, обусловленная высокими значениями стандартных ошибок параметров при высоких значениях коэффициента детерминации и $F-$статистики.

Существование корреляционной связи между независимыми переменными может быть выявлено с помощью парных коэффициентов корреляции $r_{x_i x_j}$. Высокие значения парных коэффициентов корреляции $|r_{x_i x_j}| > 0{,}75$ указывают на наличие интеркорреляции (линейной связи между двумя факторами).

Наличие мультиколлинеарности можно подтвердить, если вычислить определитель матрицы парных коэффициентов корреляции. Если этот определитель равен 0, то строки (столбцы) матрицы находятся в линейной зависимости, что говорит о наличии мультиколлинеарности факторов. Если этот определитель равен 1, то связь между независимыми переменными полностью отсутствует, что говорит об отсутствии в построенной модели мультиколлинеарности факторов.

Для устранения мультиколлинеарности используют следующие возможности:

- Анализ связей между объясняющими переменными с целью отбора только тех, которые слабо связаны друг с другом;
- Функциональные преобразования тесно связанных между собой переменных;
- Использование специальных методов корректировки моделей: ридж-регрессия (гребневая регрессия), метод главных компонент.

6.9. Частная корреляция

Как известно, для оценки тесноты связи между переменными используется коэффициент корреляции. Если переменные коррелируют друг с другом, то на значении коэффициента корреляции частично сказывается влияние других переменных. В связи с этим возникает необходимость исследовать частную корреляцию между переменными при исключении (элиминировании) влияния одной или нескольких переменных. Например, при наличии двумерной регрессионной модели $\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2$ можно определить тесноту связи между переменными y и x_1, исключая влияние переменной x_2. Для этого требуется вычислить коэффициент частной корреляции по формуле

$$r_{yx_1/x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2) \cdot (1 - r_{x_1x_2}^2)}}, \qquad (6.9)$$

которая связывает коэффициенты частной и парной корреляции.

Между переменными y и x_2, исключая влияние переменной x_1, коэффициент частной корреляции вычисляется по формуле

$$r_{yx_2/x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_1}^2) \cdot (1 - r_{x_1x_2}^2)}}. \qquad (6.10)$$

Между переменными x_1 и x_2, исключая влияние переменной y, коэффициент частной корреляции вычисляется по формуле

$$r_{x_1x_2/y} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1 - r_{yx_1}^2) \cdot (1 - r_{yx_2}^2)}}. \qquad (6.11)$$

Коэффициенты частной корреляции принимают значения от –1 до 1. Если $r_{yx_i/x_j} = 0$, то это означает отсутствие прямого (линейного влияния) переменной x_i на y. Если вычислить все коэффициенты частной корреляции, то из них можно составить матрицу коэффициентов частной корреляции, которая имеет вид (для двумерной регрессии):

	y	x_1	x_2
y	1		
x_1	r_{yx_1/x_2}	1	
x_2	r_{yx_2/x_1}	$r_{x_1x_2/y}$	1

Коэффициенты частной корреляции r_{yx_i/x_j} позволяют ранжировать факторы по степени влияния их на результативный признак и находят применение в процедуре отбора факторов для включения их в уравнение регрессии (учитываются факторы со значимыми коэффициентами частной корреляции). Статистическая значимость коэффициентов частной корреляции определяется по аналогии с обычным коэффициентом корреляции с помощью критерия Стьюдента.

Пример 6.3. По 20 предприятиям региона изучалась зависимость выработки продукции на одного работника y (млн. руб.) от ввода в действие новых

основных фондов x_1 (%) и удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%).

По результатам вычислений было составлено уравнение множественной регрессии $\hat{y} = 1{,}8353 + 0{,}9459x_1 + 0{,}0856x_2$. Оценка $\hat{b}_0 = 1{,}8353$ показывает агрегированное влияние прочих (кроме x_1 и x_2) факторов на результирующий показатель y. Оценки $\hat{b}_1 = 0{,}9459$ и $\hat{b}_2 = 0{,}0856$ указывают, что с увеличением x_1 и x_2 на 1% выработка продукции увеличивается соответственно на 0,9459 и 0,0856 млн. руб. Рассчитаны соответствующие $t-$статистики $t_{\hat{b}_0} = 3{,}9$, $t_{\hat{b}_1} = 4{,}45$, $t_{\hat{b}_2} = 1{,}42$, $t_0 = 1{,}74$, так как $t_0 > t_{\hat{b}_2}$, то коэффициент b_2 объявляется статистически незначимым, значит фактор x_2 оказывает несущественное влияние на y.

По результатам вычислений подсчитаны коэффициенты парной корреляции, из которых составлена матрица

	y	x_1	x_2
y	1		
x_1	0,9699	1	
x_2	0,9408	0,9428	1

Значения коэффициентов парной корреляции указывают на весьма тесную связь выработки продукции y как с коэффициентом обновления основных фондов $-x_1$, так и с долей рабочих высокой квалификации $-x_2$. Но в то же время межфакторная связь $r_{x_1x_2} = 0{,}9428$ весьма тесная и превышает $r_{yx_2} = 0{,}9408$, т.е. в построенной модели имеет место мультиколлинеарность факторов. Для устранения этого эффекта можно исключить из модели фактор x_2 как малоинформативный и оказывающий несущественное влияние на y.

Значения коэффициентов частной корреляции сведены в таблицу:

	y	x_1	x_2
y	1		
x_1	0,7335	1	
x_2	0,3247	0,3679	1

Коэффициенты частной корреляции дают более точную характеристику тесноты зависимости двух признаков, чем коэффициенты парной корреляции, так как избавляют парную зависимость от взаимодействия данной пары признаков с другими признаками, входящими в модель.

Наиболее тесно связаны y и x_1 так как $r_{yx_1/x_2} = 0{,}7335$, связь y и x_2 гораздо слабее так как $r_{yx_2/x_1} = 0{,}3247$, а межфакторная зависимость x_1 и x_2 выше, чем парная частная y и x_2, так как $r_{x_1x_2/y} = 0{,}3679 > r_{yx_2/x_1} = 0{,}3247$. Эти показатели приводят к выводу о необходимости исключить фактор x_2 – доля высококвалифицированных рабочих из уравнения множественной регрессии.

Если сравнить коэффициенты парной и частной корреляции, то можно увидеть, что из-за высокой межфакторной зависимости коэффициенты парной корреляции дают завышенные оценки тесноты связи:

$$r_{yx_1} = 0{,}9699 \sim r_{yx_1/x_2} = 0{,}7335,$$
$$r_{yx_2} = 0{,}9408 \sim r_{yx_2/x_1} = 0{,}3247.$$

Именно по этой причине рекомендуется при наличии мультиколлинеарности факторов исключать из модели тот фактор, у которого теснота парной зависимости меньше, чем теснота межфакторной связи $\begin{pmatrix} 0{,}9408 < 0{,}9428 \\ 0{,}3247 < 0{,}3679 \end{pmatrix}$.

6.10. Фиктивные переменные

В моделях, рассмотренных ранее, в качестве факторов рассматривались экономические переменные, принимающие количественные значения в некотором интервале. Вместе с тем может оказаться необходимым включить в модель фактор, имеющий два или более качественных уровня. Это могут быть разного рода атрибутивные признаки: профессия, пол, образование, климатические условия, принадлежность к региону и т.п. Чтобы ввести такие переменные в регрессионную модель, им требуется присвоить те или иные цифровые метки, т.е. качественные переменные должны быть преобразованы в количественные. Такого рода сконструированные переменные в эконометрике называют фиктивными переменными. Однако надо понимать, что это такие же равноправные переменные как и любые регрессоры. Их фиктивность состоит только в том, что они количественным образом описывают качественный признак.

Качественное различие можно формализовать с помощью любой переменной, принимающей два значения. В эконометрической практике почти всегда используют фиктивные переменные типа «0-1», поскольку в этом случае интерпретация выглядит наиболее просто. Например, при исследовании зависимости заработной платы от различных факторов может возникнуть вопрос, влияет ли на ее размер наличие или отсутствие у сотрудника высшего образования, и если влияет то в какой степени.

Пример 6.4. Пусть y – размер заработной платы сотрудника (тыс. руб.), $x = (x_1, x_2, \ldots, x_k)^t$ – набор объясняющих переменных, от которых может зависеть величина y (трудовой стаж, категория оплаты, бонусы и т.д.). В действительности, y и x_i – это логарифмы соответствующих характеристик, так ка связь между зарплатой и признаками имеет степенной характер. Линеаризация степенной зависимости позволяет перейти к линейной модели

$$y = b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} + \varepsilon_i,$$

где y_i – размер зарплаты i – го сотрудника, $i = \overline{1,n}$.

Теперь введем в рассмотрение такой фактор как наличие или отсутствие у сотрудника высшего образования.

Для этого введем новую бинарную переменную d, полагая при этом, что

$$d_i = \begin{cases} 1, & \textit{если в } i - \textit{м наблюдении сотрудник} \\ & \textit{имеет высшее образование,} \\ 0, & \textit{в противном случае.} \end{cases}$$

В результате получаем новую модель с фиктивной переменной

$$y = b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} + c d_i + \varepsilon_i, \quad i = \overline{1,n}.$$

Таким образом, мы считаем, что средняя зарплата есть $X^t B$ (в матричном обозначении) при отсутствии высшего образования и $X^t B + c$ – при его наличии. Величина c здесь интерпретируется как среднее изменение зарплаты при переходе из одной категории (без высшего образования) в другую (с высшим образованием) при неизменном уровне других факторов. К модели с фиктивной переменной можно применить метод наименьших квадратов и получить оценки неизвестных параметров. Если протестировать гипотезу $H_0 : c = 0$, то можно проверить предположение о несущественном различии в зарплате между категориями.

Фиктивные переменные можно вводить не только в линейные, но и в нелинейные модели, приводимые путем преобразований к линейному виду.

Пример 6.5. Было проведено обследование рынка недвижимости в Москве, для этого студентами РЭШ были собраны данные и по 464 наблюдениям построена логарифмическая модель

$$\ln y = 7{,}106 + 0{,}670 \ln x_1 + 0{,}431 \ln x_2 + 0{,}147 \ln x_3 - 0{,}114 \ln x_4 - 0{,}068 d_1 +$$
$$+ 0{,}134 d_2 + 0{,}042 d_3 + 0{,}114 d_4 + 0{,}214 d_5 + 0{,}140 d_6 + 0{,}164 d_7 + 0{,}169 d_8,$$

где

y – цена квартиры ($ США),

x_1 – жилая площадь (кв. м.),

x_2 – площадь нежилых помещений (кв. м.),

x_3 – площадь кухни (кв. м.),

x_4 – расстояние от квартиры до центра Москвы (км.).

Фиктивные переменные:

$$d_1 = \begin{cases} 1, & \textit{если квартира на } 1 - \textit{м или последнем этаже,} \\ 0, & \textit{в противном случае.} \end{cases}$$

$$d_2 = \begin{cases} 1, & \textit{если квартира в кирпичном доме,} \\ 0, & \textit{в противном случае.} \end{cases}$$

$$d_3 = \begin{cases} 1, & \textit{если в квартире есть балкон}, \\ 0, & \textit{в противном случае.} \end{cases}$$

$$d_4 = \begin{cases} 1, & \textit{если в доме есть лифт}, \\ 0, & \textit{в противном случае.} \end{cases}$$

$$d_5 = \begin{cases} 1, & \textit{если квартира однокомнатная}, \\ 0, & \textit{в противном случае.} \end{cases}$$

$$d_6 = \begin{cases} 1, & \textit{если квартира двухкомнатная}, \\ 0, & \textit{в противном случае.} \end{cases}$$

$$d_7 = \begin{cases} 1, & \textit{если квартира трехкомнатная}, \\ 0, & \textit{в противном случае.} \end{cases}$$

$$d_8 = \begin{cases} 1, & \textit{если квартира четырехкомнатная}, \\ 0, & \textit{в противном случае.} \end{cases}$$

Из анализа $t-$статистик получено, что все коэффициенты регрессии, кроме коэффициентов при d_5 и d_6, значимы при доверительной вероятности $\gamma = 0{,}95$.

Коэффициент при $\ln x_1$, равный 0,67, означает, что при увеличении жилой площади квартиры на 1% ее цена увеличивается на 0,67%. То есть эластичность цены квартиры по жилой площади равна 0,67.

Отрицательный коэффициент при $\ln x_4$, равный –0,114, означает, что при увеличении расстояния до центра на 1% цена квартиры уменьшается на 0,114%.

Интерпретация коэффициентов при фиктивных переменных:

Отрицательный коэффициент при d_1, равный -0,068, означает, что квартира на 1-м или последнем этаже стоит на 6,8% дешевле аналогичной квартиры на других этажах. Квартира в кирпичном доме стоит на 13,4% дороже квартиры в панельном доме, наличие в доме лифта увеличивает стоимость квартиры на 11,4%, а наличие балкона – на 4,2%.

Переменные d_5, d_6, d_7, d_8 были включены в модель, чтобы учесть возможные различия в структуре рынка жилья для квартир с разным количеством комнат. В выборке были представлены 5,6 и даже 8-ми комнатные квартиры, поэтому переменные $d_5 + d_6 + d_7 + d_8 \neq 1$(в сумме не дают константу, что означает отсутствие полной коллинеарности факторов).

Коэффициенты при d_6, d_7, d_8 можно считать равными. Квартиры с числом комнат от 2 до 4 стоят дороже многокомнатных, а однокомнатные – еще дороже (при прочих равных условиях).

6.11. Множественная регрессия в нелинейных моделях

В качестве моделей множественной регрессии кроме линейной часто используются нелинейные:

полином $k-$й степени

$$y = b_0 + b_{11}x_1 + b_{12}x_1^2 + \cdots + b_{1k}x_1^k + \cdots + b_{p1}x_p + b_{p2}x_{p2}^2 + \cdots + b_{pk}x_p^k + \varepsilon,$$

обратная функция $y = \dfrac{1}{b_0 + b_1x_1 + \cdots + b_px_p + \varepsilon}$,

степенная функция $y = b_0x_1^{b_1} \cdot \cdots \cdot x_p^{b_p}\varepsilon$,

показательная функция $y = b_0b_1^{x_1} \cdot \cdots \cdot b_p^{x_p}\varepsilon$,

полулогарифмическая функция $y = b_0 + b_1\ln x_1 + \cdots + b_p\ln x_p + \varepsilon$.

Нелинейные функции могут представлять собой также «смешанные» модели. Например, можно построить уравнение множественной регрессии с тремя независимыми переменными $y = b_0x_1^{b_1}x_2^{b_2}b_3^{x_3}\varepsilon$. В этом уравнении использованы две функции: степенная (для учета влияния переменных x_1 и x_2) и показательная степенная (для учета влияния переменной x_3).

При выборе формы регрессии необходимо учитывать ряд обстоятельств.

Во-первых, нужно принимать во внимание теоретические предпосылки построения модели; выводы, сформированные в экономической теории о характере взаимосвязи показателей, ограничениях, налагаемых на параметры модели.

Например, форму степенной функции имеет производственная функция Кобба-Дугласа

$$P = b_0L^{b_1}K^{b_2}\varepsilon,$$

где P – объем продукции, L – затраты труда, K – величина капитала, b_0, b_1, b_2 – параметры модели, ε – случайная компонента. На параметры производственной функции может быть наложено линейное ограничение вида $b_1 + b_2 = 1$. Тогда

$$P = b_0L^{b_1}K^{1-b_1}\varepsilon, \text{ или } \frac{P}{K} = b_0\left(\frac{L}{K}\right)^{b_1}\varepsilon.$$

Во-вторых, при выборе функции регрессии следует учитывать возможность оценки ее параметров, простоту их интерпретации. Все перечисленные ранее нелинейные функции являются линеаризуемыми, т.е. их можно преобразовать в линейную форму. Оценка параметров таких моделей производится путем применения МНК к линеаризованной форме нелинейной функции.

Для функций, линейных по параметрам, линеаризация производится с помощью замены переменных. Например, функция

$$y = b_0 + b_1x_1^2 + b_2\sqrt{x_2} + \varepsilon$$

является линейной по параметрам b_0, b_1, b_2 и нелинейной по переменным x_1 и x_2. Для определения МНК-оценок параметров этой модели можно воспользоваться стандартной формулой для множественной линейной регрессии, формально можно обозначить $z_1 = x_1^2$ и $z_2 = \sqrt{x_2}$, тогда линеаризованная модель будет иметь вид

$$y = b_0 + b_1z_1 + b_2z_2 + \varepsilon.$$

Если случайная составляющая ε удовлетворяет предпосылкам, лежащим в основе метода наименьших квадратов, то свойства МНК-оценок линеаризованной модели будут совпадать со свойствами МНК-оценок параметров линейной модели.

Нелинейность по параметрам является более серьезной проблемой. Линеаризация функций, в которых переменные связаны мультипликативно, заключается в логарифмировании правой и левой частей уравнения по любому основанию, чаще всего – по натуральному. Например, степенная функция

$$y = b_0 x_1^{b_1} x_2^{b_2} \varepsilon$$

после линеаризации путем логарифмирования примет вид $\ln y = \ln b_0 + b_1 \ln x_1 + b_2 \ln x_2 + \ln \varepsilon$. Если ввести обозначения $y' = \ln y$, $b_0' = \ln b_0$, $x_1' = \ln x_1$, $x_2' = \ln x_2$, $\varepsilon' = \ln \varepsilon$ получим линейную по параметрам модель

$$y' = b_0' + b_1 x_1' + b_2 x_2' + \varepsilon'.$$

В степенной функции $y = b_0 x_1^{b_1} \cdot \dots \cdot x_p^{b_p} \varepsilon$ коэффициенты $b_1, b_2, \dots, b_p$ являются коэффициентами эластичности зависимой переменной y по соответствующим независимым переменным $x_1, x_2, \dots, x_p$.

В показательной функции $y = b_0 b_1^{x_1} \cdot \dots \cdot b_p^{x_p} \varepsilon$ коэффициенты $b_1, b_2, \dots, b_p$ показывают во сколько раз в среднем изменится зависимая переменная при изменении соответствующей независимой переменной $x_1, x_2, \dots, x_p$ на единицу при неизменных значениях других переменных, включенных в уравнение регрессии.

В случае нелинейных моделей степень концентрации распределения наблюдаемых значений вблизи регрессионной плоскости показывает корреляционное отношение или индекс корреляции

$$\eta = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

где $\hat{y}_i$ – рассчитанные по модели значения переменной y, y_i – наблюдаемые значения переменной y, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ – среднее значение переменной y. По определению $0 \le \eta \le 1$, если $\eta = 1$, то имеет место функциональная зависимость по которой рассчитаны значения $\hat{y}_i$, если $\eta = 0$, то построенная модель непригодна для анализа и прогнозирования.

Коэффициент детерминации для нелинейной модели можно определить через индекс корреляции $R^2 = \eta^2$, интерпретация коэффициента детерминации дается в процентах. Как и в случае линейной модели, он показывает долю вариации величины y, которая объясняется вариацией факторов $x_1, x_2, \dots, x_p$, включенных в уравнение регрессии.

6.12. Рекомендации по выполнению расчетно-графической работы на тему «Линейная модель множественной регрессии»

Задание:

1. Самостоятельно собрать данные для расчетно-графической работы по указанной теме. Количество основных переменных не менее 7: одна зависимая и не менее 6 независимых, включая фиктивные. Число наблюдений не менее 40.

2. Построить матрицу парных и частных коэффициентов корреляции. Проанализировать модель на наличие мультиколлинеарности факторов.

3. Найти точечные и интервальные оценки параметров модели $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots$.

4. Оценить значимость коэффициентов регрессии при $\gamma = 0{,}95$, используя:

   а) *t*-критерий Стьюдента;

   б) доверительные интервалы истинных значений параметров.

5. Верифицировать полученную модель, используя:

   а) дисперсионный анализ в регрессии;

   б) элементы теории корреляции;

   в)среднюю ошибку аппроксимации.

6. Найти наилучшую модель методом последовательного исключения наименее значимого фактора.

7. Рассчитать частные коэффициенты корреляции. Сделать выводы.

8. Интерпретировать полученные результаты.

9. В случае пригодной линейной модели построить точечные и интервальные прогнозы зависимой переменной (при $\alpha = 0{,}05$).

Пример 6.6. Изучаемый объект: рынок жилой недвижимости г. Иркутска (по данным на апрель 2010 года).

Переменные:

зависимая: y – цена объекта жилой недвижимости, тыс. руб.

независимые:

x_1 – площадь, кв. м; x_2 – количество комнат; $x_3 = \begin{cases} 1, \text{кирпичный} \\ 0, \text{другой} \end{cases}$ – материал, из которого построен дом; $x_4 = \begin{cases} 1, \text{не первый и не последний} \\ 0, \text{первый или последний} \end{cases}$ – этаж, на котором расположен изучаемый объект; x_5 – площадь кухни, кв. м; $x_6 = \begin{cases} 1, \text{раздельный} \\ 0, \text{совмещенный} \end{cases}$ – санузел; $x_7 = \begin{cases} 1, \text{есть} \\ 0, \text{нет} \end{cases}$ – балкон (наличие в рассматриваемой квартире); x_8 – удаленность от центра города, км (центр – квадрат улиц: Ленина, Карла Маркса, Тимирязева, Декабрьских событий), расстояние определялось по 2ГИС;

$x_9 = \begin{cases} 1, \text{есть} \\ 0, \text{нет} \end{cases}$ – лифт (наличие в рассматриваемом доме); $x_{10} = \begin{cases} 1, \text{является} \\ 0, \text{нет} \end{cases}$ – новостройка; $x_{11} = \begin{cases} 1, \text{ремонт не ттребуетс} \\ 0, \text{ремонт ттребуетс} \end{cases}$ – необходимость в проведении ремонта в рассматриваемой квартире.

Количество наблюдений равно 121.

Исходные данные:

y	x_1	x_2				x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
		x_{2_1}	x_{2_2}	x_{2_3}	x_{2_4}									
830	18	1	0	0	0	1	1	5	0	1	8	0	0	0
900	18	1	0	0	0	0	1	3	0	1	3	0	0	1
950	24	1	0	0	0	0	1	4,5	0	1	3	0	0	0
950	19	1	0	0	0	0	1	4	0	1	3	0	0	1
1000	18	1	0	0	0	0	1	4	0	1	3	0	0	1
1150	29	1	0	0	0	1	1	5	0	1	8	1	1	1
1180	30	1	0	0	0	1	1	6	0	0	7	0	0	0
1220	34	1	0	0	0	0	1	7	1	1	16	0	0	0
1250	30	1	0	0	0	0	1	6	0	1	10	0	0	0
1250	30,5	1	0	0	0	0	0	6,1	0	0	14	0	0	1
1320	18,3	1	0	0	0	0	0	3	0	1	3	0	0	1
1350	37	1	0	0	0	1	1	8	0	1	2	1	1	0
1400	30	1	0	0	0	0	1	6	0	1	6	0	0	0
1450	30	1	0	0	0	0	1	9	1	1	13	1	1	1
1450	30	1	0	0	0	0	1	6	0	1	4	0	0	0
1450	30	1	0	0	0	0	0	6	0	0	1	0	0	0
1500	30	1	0	0	0	0	1	6	0	1	9	0	0	0
1500	30	1	0	0	0	1	0	6	0	0	1	0	0	0
1500	33,8	1	0	0	0	1	1	7	0	1	2	0	0	1
1520	32	1	0	0	0	1	1	6,5	0	1	7	0	0	1
1550	30	1	0	0	0	0	0	6	1	1	6	0	0	1
1550	40	1	0	0	0	0	1	9	1	1	1	1	0	0
1600	45	1	0	0	0	0	1	10	1	1	16	1	0	0
1600	30	1	0	0	0	1	1	6	0	0	4	0	0	0
1600	30	1	0	0	0	0	1	6	0	1	4	0	0	1
1600	30,4	1	0	0	0	1	1	5	0	1	0,5	0	0	0
1600	32	1	0	0	0	1	1	8	0	1	3	1	1	0
1650	30	1	0	0	0	0	1	5	0	1	1	0	0	1
1700	30	1	0	0	0	0	1	6	0	1	6	0	0	1
1799	31	1	0	0	0	0	1	6	1	1	8	0	0	1
1800	41,4	1	0	0	0	1	1	10,7	1	1	1	0	0	0
1800	30	1	0	0	0	1	1	6	0	1	1	0	0	0
1900	36	1	0	0	0	1	1	10	0	1	4	1	1	0
1900	31	1	0	0	0	1	0	10	0	1	6	0	1	1
1900	47	1	0	0	0	1	0	9	0	1	2	1	1	0
2200	48	1	0	0	0	0	1	6	0	1	10	0	1	1
2200	33	1	0	0	0	1	1	6	0	1	0	0	0	1
2200	52	1	0	0	0	1	1	14	0	1	0,5	1	1	0
2250	47	1	0	0	0	1	0	9	0	1	2	1	1	0
2300	41	1	0	0	0	0	1	10	0	1	4	1	1	0
2300	41	1	0	0	0	1	1	9	0	1	3	1	1	1

y	x_1	x_2				x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
		x_{2_1}	x_{2_2}	x_{2_3}	x_{2_4}									
2350	37	1	0	0	0	1	1	9	1	1	6	1	1	1
2500	53	1	0	0	0	1	1	9	0	1	5	1	1	0
2700	47	1	0	0	0	0	1	11	0	1	6	1	1	1
1290	43,2	0	1	0	0	1	0	6	1	1	14	0	0	0
1470	45	0	1	0	0	0	1	6	1	1	14	0	0	0
1550	44,7	0	1	0	0	0	1	6,2	1	1	13	0	0	0
1650	48	0	1	0	0	1	0	11	1	1	5	0	0	0
1700	42	0	1	0	0	1	1	6	1	1	6	0	0	0
1700	44	0	1	0	0	0	0	9	0	1	14	1	0	0
1750	72	0	1	0	0	1	1	13	1	1	16	1	1	1
1780	44	0	1	0	0	0	1	6	0	1	15	0	0	0
1900	44	0	1	0	0	1	1	6	1	1	6	0	0	1
1950	47	0	1	0	0	1	1	8	0	1	17	0	0	1
2000	42	0	1	0	0	0	1	6	1	1	1	0	0	0
2100	42	0	1	0	0	0	0	6	0	1	3	0	0	1
2150	43	0	1	0	0	1	0	6	0	1	2	0	0	1
2150	46	0	1	0	0	0	1	6	0	1	0,5	0	0	0
2150	44	0	1	0	0	0	0	6	0	0	1	0	0	1
2200	45	0	1	0	0	1	0	6	1	1	2	0	0	0
2250	45	0	1	0	0	0	0	6	1	1	8	0	0	1
2300	44	0	1	0	0	0	0	6	0	1	6	0	0	1
2300	43	0	1	0	0	0	1	6	1	1	3	0	0	0
2350	68	0	1	0	0	0	1	12	1	1	7	1	1	0
2400	50	0	1	0	0	1	1	6	1	1	4	0	0	0
2450	46	0	1	0	0	0	1	6	0	1	6	0	0	0
2500	32	0	1	0	0	0	1	9	1	1	12	1	0	1
2500	48	0	1	0	0	0	1	8	1	1	8	0	0	1
2500	50	0	1	0	0	1	0	6	1	1	3	0	0	0
2500	42	0	1	0	0	1	1	6	0	0	0,1	0	0	1
2600	56	0	1	0	0	1	0	10	1	1	9	1	1	1
2600	64,5	0	1	0	0	1	1	9	1	1	4	1	1	1
2600	48	0	1	0	0	0	1	8	1	1	2	1	0	1
2650	48	0	1	0	0	0	1	9	1	1	4	1	0	1
2750	48	0	1	0	0	0	0	8	1	1	6	1	0	1
3300	45	0	1	0	0	1	1	6	0	0	0,1	0	0	0
3600	55	0	1	0	0	1	1	12	1	0	0,5	0	1	1
3900	77	0	1	0	0	1	1	9	1	1	1	0	0	1
4000	74	0	1	0	0	1	1	15	0	1	3	1	1	0
4000	69	0	1	0	0	1	1	11	1	0	0,1	0	0	0
6700	90	0	1	0	0	1	1	28	1	1	2	1	1	1
2100	59	0	0	1	0	0	0	6	0	0	5	0	0	0
2200	65	0	0	1	0	0	1	6,2	1	1	15	0	0	1
2300	55	0	0	1	0	0	1	6	0	1	13	0	0	0
2400	58,5	0	0	1	0	1	1	6	0	1	2	0	0	0
2400	77	0	0	1	0	0	0	9,5	1	1	6	1	0	0
2650	67	0	0	1	0	0	1	8	1	1	8	1	0	0
2650	58,5	0	0	1	0	1	0	8	1	1	0,1	0	0	0
2800	57	0	0	1	0	0	0	8	1	1	8	1	0	0
2900	63	0	0	1	0	0	0	9	1	1	16	1	0	0
2900	80	0	0	1	0	1	1	13	1	1	2	1	1	1
2950	59	0	0	1	0	0	1	6	0	1	5	0	0	1

y	x_1	x_2				x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
		x_{2_1}	x_{2_2}	x_{2_3}	x_{2_4}									
3100	100	0	0	1	0	1	1	14	0	1	8	1	1	1
3200	66	0	0	1	0	0	1	9	1	1	7	1	0	1
3200	86	0	0	1	0	1	1	10	1	1	13	1	0	1
3200	60	0	0	1	0	0	1	6	0	1	1	0	0	0
3200	64	0	0	1	0	1	1	9	1	1	0,5	0	0	0
3500	67	0	0	1	0	0	0	8	1	1	6	1	0	1
3500	74	0	0	1	0	1	1	8	1	1	7	0	0	1
3500	57	0	0	1	0	1	0	6	0	1	1	0	0	1
3550	62	0	0	1	0	0	1	6	1	1	6	0	0	1
3590	68,4	0	0	1	0	1	0	11,1	1	1	4	1	1	1
3700	83	0	0	1	0	1	1	10	1	1	0,5	1	1	1
3800	56	0	0	1	0	1	1	8	0	1	0,1	0	0	0
4200	68	0	0	1	0	1	1	9	1	1	2	0	0	0
5150	86	0	0	1	0	1	1	11	1	1	1	1	1	1
5200	87	0	0	1	0	0	1	9	1	1	2	1	0	1
5300	100	0	0	1	0	1	1	25	1	1	6	1	1	0
6550	75	0	0	1	0	1	1	8	1	0	0,1	0	0	0
8900	83	0	0	1	0	1	0	10	1	1	0,1	0	0	1
3090	83	0	0	0	1	0	1	10	1	1	13	0	0	0
3100	62	0	0	0	1	1	0	7,6	1	1	6	0	0	1
3200	96	0	0	0	1	1	1	9	1	1	12	0	0	0
4000	98	0	0	0	1	0	1	19	1	1	8	1	1	1
4500	87	0	0	0	1	0	1	11,2	1	1	4	1	0	0
5100	88	0	0	0	1	0	1	10	1	1	6	1	0	0
5300	100	0	0	0	1	1	1	25	1	1	7	1	0	0
6000	72	0	0	0	1	0	1	6	1	1	6	0	0	1
5200	87	0	0	0	0	0	1	9	1	1	2	1	0	1
5400	90	0	0	0	0	0	0	9	1	1	6	1	0	0
5800	90	0	0	0	0	0	0	9	1	1	7	1	0	1

1 этап: анализ взаимосвязей

Вычисления в MS Excel (здесь и далее в оформлении расчетно-графической работы промежуточные вычисления приводить не надо**)**

Строим матрицу парных коэффициентов корреляции: Данные => Анализ данных => корреляция. Выделяем исходные данные (рис. 6.1) вместе с обозначениями переменных, выводим результаты на новый рабочий лист (рис. 6.2). Для удобства коэффициенты корреляции можно выделить разными цветами с помощью условного форматирования (рис. 6.3).

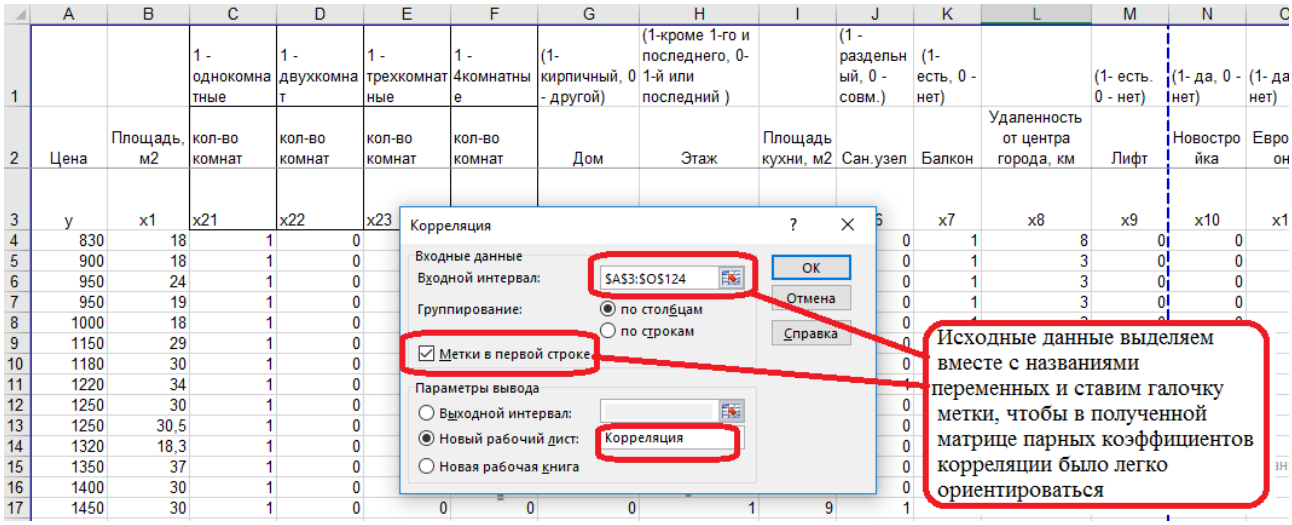
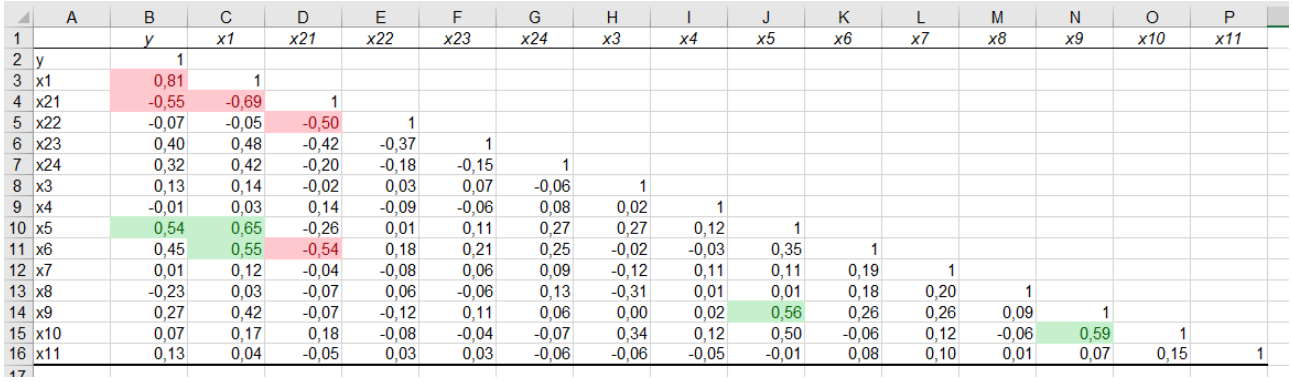


Рис. 6.1. Вычисление парных коэффициентов корреляции



	y	x1	x21	x22	x23	x24	x3	x4	x5	x6	x7	x8	x9	x10	x11
y	1														
x1	0,81	1													
x21	-0,55	-0,69	1												
x22	-0,07	-0,05	-0,50	1											
x23	0,40	0,48	-0,42	-0,37	1										
x24	0,32	0,42	-0,20	-0,18	-0,15	1									
x3	0,13	0,14	-0,02	0,03	0,07	-0,06	1								
x4	-0,01	0,03	0,14	-0,09	-0,06	0,08	0,02	1							
x5	0,54	0,65	-0,26	0,01	0,11	0,27	0,27	0,12	1						
x6	0,45	0,55	-0,54	0,18	0,21	0,25	-0,02	-0,03	0,35	1					
x7	0,01	0,12	-0,04	-0,08	0,06	0,09	-0,12	0,11	0,11	0,19	1				
x8	-0,23	0,03	-0,07	0,06	-0,06	0,13	-0,31	0,01	0,01	0,18	0,20	1			
x9	0,27	0,42	-0,07	-0,12	0,11	0,06	0,00	0,02	0,56	0,26	0,26	0,09	1		
x10	0,07	0,17	0,18	-0,08	-0,04	-0,07	0,34	0,12	0,50	-0,06	0,12	-0,06	0,59	1	
x11	0,13	0,04	-0,05	0,03	0,03	-0,06	-0,06	-0,05	-0,01	0,08	0,10	0,01	0,07	0,15	1

Рис. 6.2. Матрица парных коэффициентов корреляции

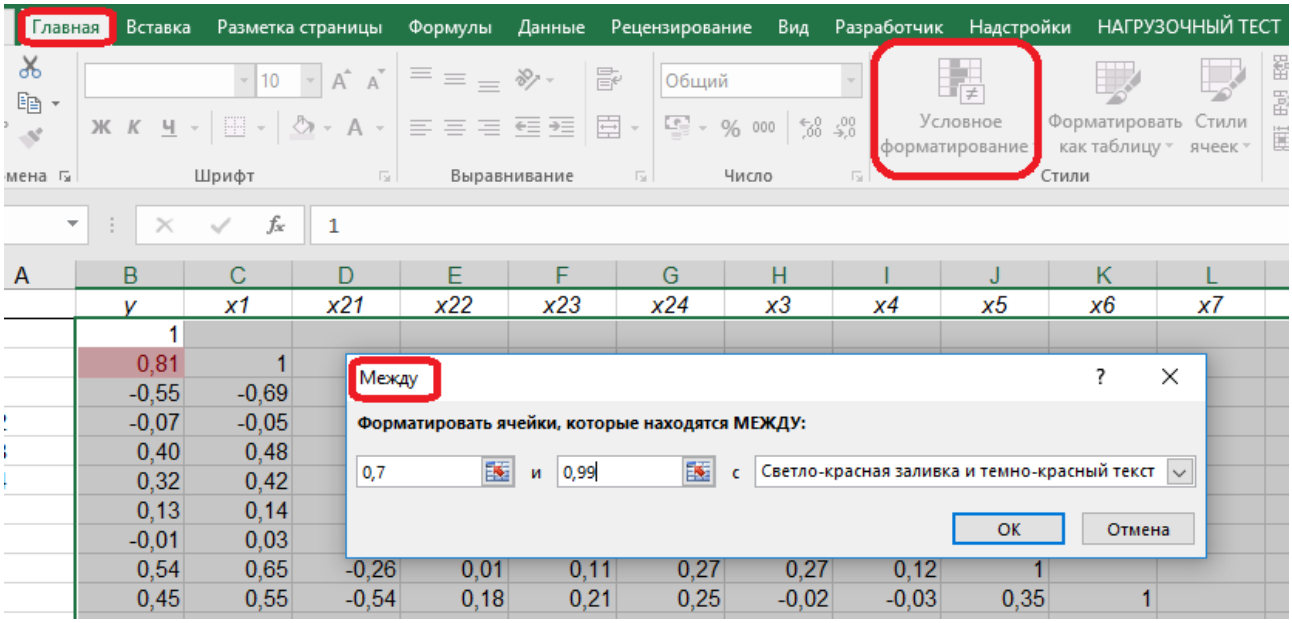


Рис. 6.3. Условное форматирование парных коэффициентов корреляции

Матрицу парных коэффициентов корреляции копируем в MS Word и проводим анализ взаимосвязей между переменными (табл. 6.2).

Таблица 6.2

Матрица парных коэффициентов корреляции

	y	*x1*	*x21*	*x22*	*x23*	*x24*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*
y	1														
x1	0,81	1													
x21	-0,55	-0,69	1												
x22	-0,07	-0,05	-0,50	1											
x23	0,40	0,48	-0,42	-0,37	1										
x24	0,32	0,42	-0,20	-0,18	-0,15	1									
x3	0,13	0,14	-0,02	0,03	0,07	-0,06	1								
x4	-0,01	0,03	0,14	-0,09	-0,06	0,08	0,02	1							
x5	0,54	0,65	-0,26	0,01	0,11	0,27	0,27	0,12	1						
x6	0,45	0,55	-0,54	0,18	0,21	0,25	-0,02	-0,03	0,35	1					
x7	0,01	0,12	-0,04	-0,08	0,06	0,09	-0,12	0,11	0,11	0,19	1				
x8	-0,23	0,03	-0,07	0,06	-0,06	0,13	-0,31	0,01	0,01	0,18	0,20	1			
x9	0,27	0,42	-0,07	-0,12	0,11	0,06	0,00	0,02	0,56	0,26	0,26	0,09	1		
x10	0,07	0,17	0,18	-0,08	-0,04	-0,07	0,34	0,12	0,50	-0,06	0,12	-0,06	0,59	1	
x11	0,13	0,04	-0,05	0,03	0,03	-0,06	-0,06	-0,05	-0,01	0,08	0,10	0,01	0,07	0,15	1

Из таблицы 6.2 видно, что с зависимой переменной тесно связаны переменные: площадь квартиры, количество комнат, средняя связь с переменными: площадь кухни, санузел. С остальными независимыми переменными цена квартиры связана слабо.

Между независимыми переменными тесная связь наблюдается у:
- площадью квартиры и количеством комнат, площадью кухни и типом санузла;
- площадью кухни и наличием лифта в доме, новостройками;
- наличием лифта и новостройками.

Анализ мультиколлинеарности. Нужно рассчитать в MS Excel определитель матрицы $X^T X$.

Вычисление определителя матрицы $X^T X$ в MS Excel

В матрице X первый столбец состоит из единиц (столбец, отвечающий за свободный коэффициент) (рис. 6.4). Матрица X имеет размерность (количество наблюдений х количество переменных). В данной работе размерность матрицы X (121 х 15).

Для того чтобы записать матрицу X^T выделяем данные матрицы X, копируем в буфер и вставляем, используя меню «специальная вставка»/«транспонировать» (рис. 6.5). Матрица X^T в результате копирования через «специальную ставку» (транспонирования) примет вид как на рис. 6.6.

Произведение матриц $X^T X$ находим с помощью функции =МУМНОЖ(массив X^T; массив X) (рис. 6.7). С ячейки, в которой записана функция (на рис. 6.7 это P18), выделяем массив размером 15х15, нажимаем клавишу

F2, затем последовательно, не отпуская клавиш, нажимаем Ctrl Shift Enter, и в выделенном массиве появится результат умножения матриц $X^T X$.

	A	B	C	D	E	F	G	H	I	J	K
1			*x1*	*x21*	*x22*	*x23*	*x24*	*x3*	*x4*	*x5*	*x6*
2	**X**	1	18	1	0	0	0	1	1	5	
3		1	18	1	0	0	0	0	1	3	
4		1	24	1	0	0	0	0	1	4,5	
5		1	19	1	0	0	0	0	1	4	
6		1	18	1	0	0	0	0	1	4	
7		1	29	1	**Первый столбец матрицы X состоит из единиц!!!**					5	
8		1	30	1	0	0	0	1	1	6	
9		1	34	1	0	0	0	0	1	7	
10		1	30	1	0	0	0	0	1	6	
11		1	30,5	1	0	0	0	0	0	6,1	
12		1	18,3	1	0	0	0	0	0	3	
13		1	37	1	0	0	0	1	1	8	
14		1	30	1	0	0	0	0	1	6	
15		1	30	1	0	0	0	0	1	9	

Рис. 6.4. Запись матрицы X в MS Excel

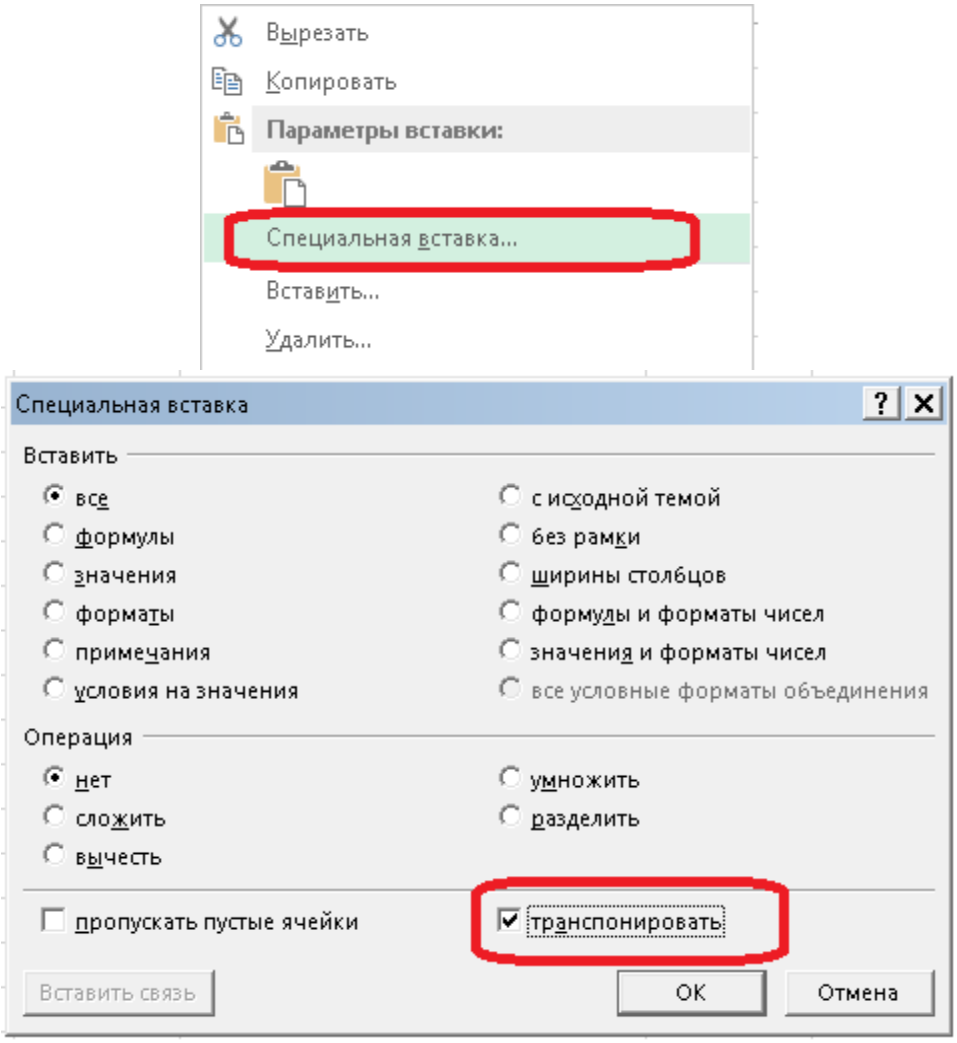


Рис. 6.5. Транспонирование в MS Excel

XT										
1	1	1	1	1	1	1	1	1	1	1
18	18	24	19	18	29	30	34	30	30,5	18,3
1	1	1	1	1	1	1	1	1	1	1
1	0	0	0	0	1	1	0	0	0	0
1	1	1	1	1	1	1	1	1	0	0
5	3	4,5	4	4	5	6	7	6	6,1	3
0	0	0	0	0	0	0	1	0	0	0
1	1	1	1	1	1	0	1	1	0	1
8	3	3	3	3	8	7	16	10	14	3
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0
0	1	0	1	1	1	0	0	0	1	1

Рис. 6.6. Транспонированная матрица X^T

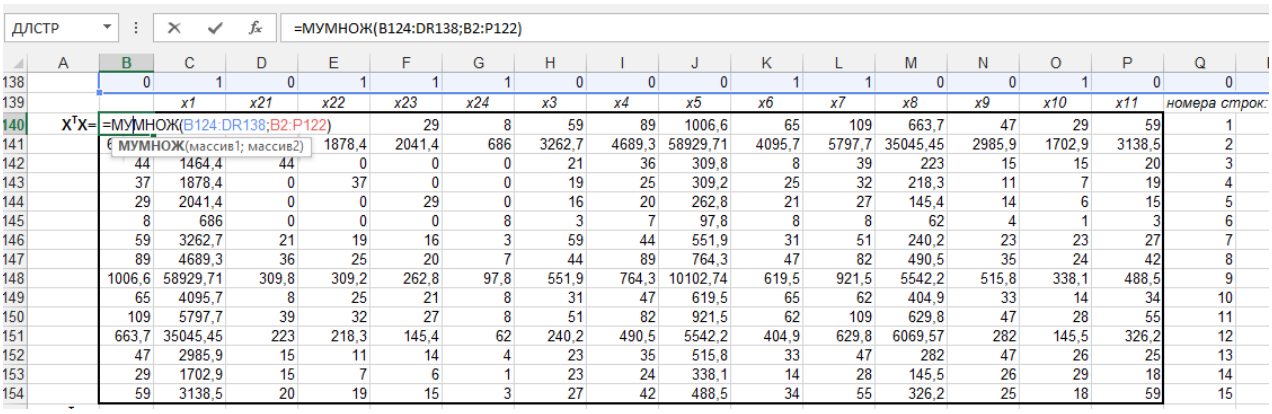


Рис. 6.7. Умножение матриц в MS Excel

После нахождения матрицы $X^T X$ вычисляем ее определитель, используя функцию =МОПРЕД(массив) (рис. 6.8).

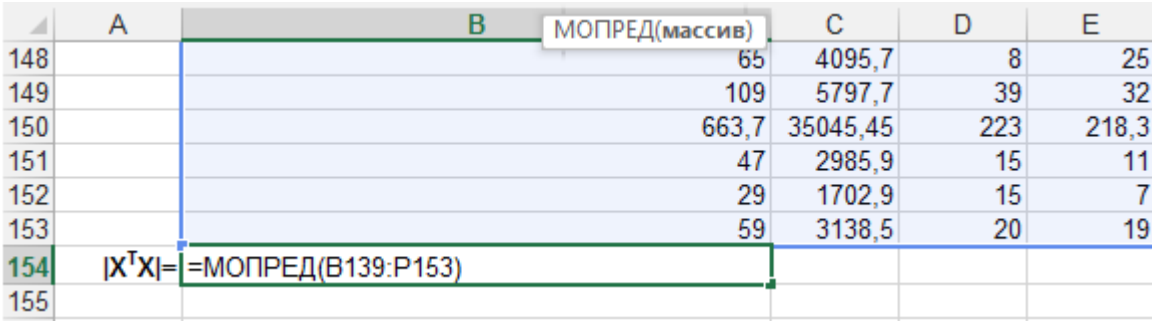


Рис. 6.8. Вычисление определителя матрицы $X^T X$

2 этап. Построение первой модели для всех независимых переменных

Построим **первую модель** $y_i = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_{11} x_{11}$, используя все независимые переменные, меню Анализ данных => регрессия (рис. 6.9). Все исходные данные выделяются вместе с обозначениями переменных (галочка «Метки»), результат выводится на новый рабочий лист. Чтобы легко ориентироваться в листах файла, удобно называть лист по количеству переменных в модели, например, 14. Также ставим галочку «Остатки», чтобы найти ошибку аппроксимации *A* по формуле (5.20). Результаты расчетов представлены в таблице 6.3.

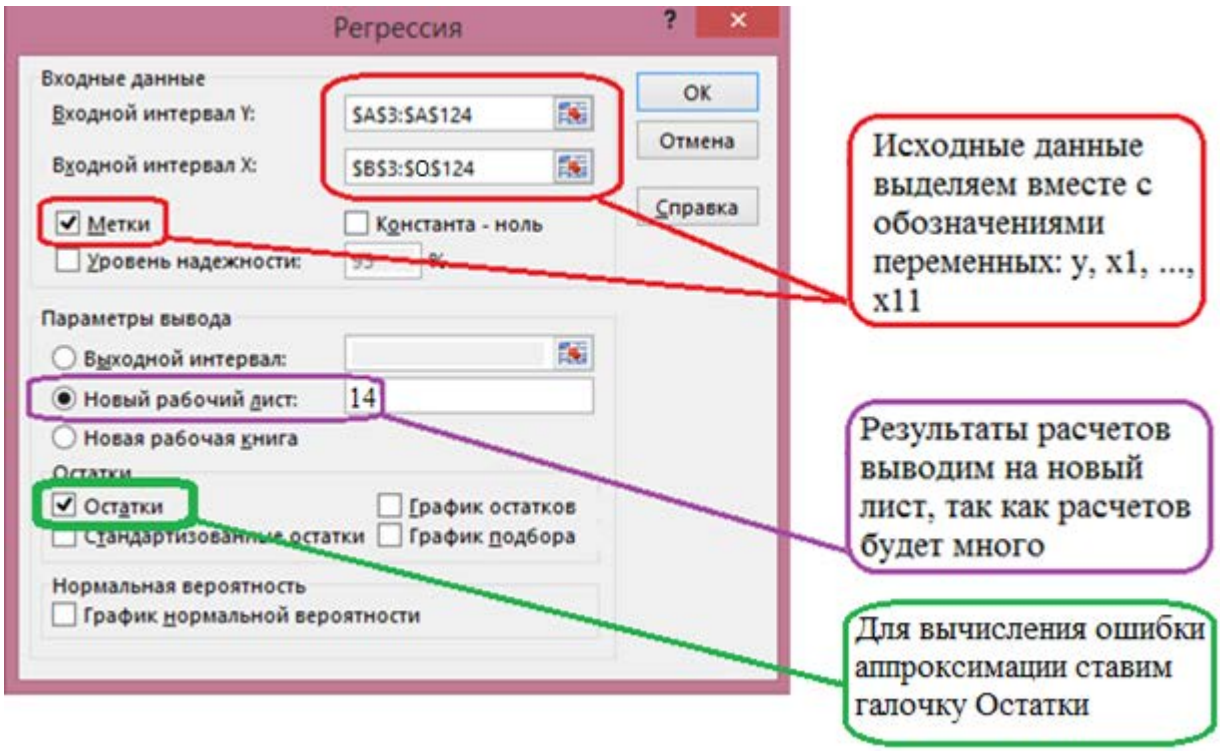


Рис. 6.9. Ввод исходных данных в меню Регрессия

Таблица 6.3

Результаты расчетов

Регрессионная статистика						
Множеств. R	0,8729					
R-квадрат	0,7619					
Нормир. R-квадрат	0,7305					
Станд. ошибка	717,3755					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Значимость F*	
Регрессия	14	174601130	12471509,28	24,23	0,00	
Остаток	106	54550530	514627,65			
Итого	120	229151660				
	Коэффициенты	*Станд. ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	1754,98	889,05	1,97	0,05	-7,64	3517,61
x1	**39,39**	**10,07**	**3,91**	**0,00**	**19,42**	**59,37**
x21	-1325,54	722,62	-1,83	0,07	-2758,20	107,12
x22	**-1334,63**	**604,04**	**-2,21**	**0,03**	**-2532,21**	**-137,05**
x23	**-1061,84**	**503,31**	**-2,11**	**0,04**	**-2059,70**	**-63,98**
x24	-1007,98	528,59	-1,91	0,06	-2055,96	40,00
x3	-70,65	159,22	-0,44	0,66	-386,32	245,01
x4	-8,07	157,93	-0,05 **min**	0,96	-321,18	305,04
x5	**69,96**	**30,87**	**2,27**	**0,03**	**8,77**	**131,15**

x6	176,05	173,05	1,02	0,31	-167,05	519,14
x7	-217,97	236,76	-0,92	0,36	-687,37	251,44
x8	**-78,32**	**15,93**	**-4,92**	**0,00**	**-109,90**	**-46,75**
x9	-211,05	209,63	-1,01	0,32	-626,66	204,57
x10	-252,60	249,72	-1,01	0,31	-747,70	242,50
x11	**335,03**	**136,31**	**2,46**	**0,02**	**64,78**	**605,29**

Полученная на первой итерации модель имеет вид (коэффициенты берем с учетом знака из столбца Коэффициенты таблицы 6.3, внизу в скобках стоят значения *t*-статистик):

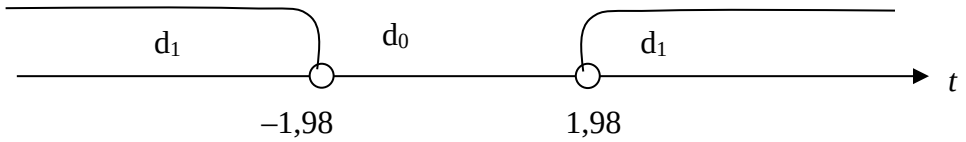
$$y = 1754{,}98 + 39{,}39x_1 - 1325{,}54x_{2_1} - 1334{,}63x_{2_2} - 1061{,}84x_{2_3} - 1007{,}98x_{2_4} - 70{,}65x_3 -$$

$$(t_{b_0} = 1{,}97) \quad (3{,}91) \quad (-1{,}83) \quad (-2{,}21) \quad (-2{,}11) \quad (-1{,}91) \quad (-0{,}44)$$

$$-8{,}07x_4 + 69{,}96x_5 + 176{,}05x_6 - 217{,}97x_7 - 78{,}32x_8 - 211{,}05x_9 - 252{,}60x_{10} + 335{,}03x_{11}$$

$$(-0{,}05) \quad (2{,}27) \quad (1{,}02) \quad (-0{,}92) \quad (-4{,}92) \quad (-1{,}01) \quad (-1{,}01) \quad (2{,}46)$$

Найдем критическую точку $t_{кр.} = t(\alpha, n-k)$ при уровне значимости $\alpha = 1 - 0{,}95 = 0{,}05$, которая зависит от числа степеней свободы, равного $(n-k) = 121 - 15 = 109$, где n = 121 – число наблюдений, k =14+1 – число оцененных параметров модели (14 независимых и 1 зависимая переменная). Вычислим $t_{кр}$ с помощью MS Excel: =СТЬЮДРАСПОБР(0,05;121–15), $t_{кр}$ = 1,98:



Оценка значимости полученных коэффициентов регрессии представлена в таблице 6.4.

Таблица 6.4

Оценка значимости коэффициентов регрессии

Коэффициент	t-статистика	Сравнение с $t_{кр}$ 1,98	Гипотеза Но: b_i=0	Доверительный интервал (ДИ)	0 входит или нет в ДИ	Вывод о значимости
1754,98	889,05	<	да	(-7,64; 3517,61)	да	не значим
39,39	**10,07**	>	нет	**(19,42; 59,37)**	нет	значим
-1325,54	722,62	< (по модулю)[1]	да	(-2758,2; 107,12)	да	не значим
-1334,63	**604,04**	>	нет	**(-2532,21; -137,05)**	нет	значим
-1061,84	**503,31**	>	нет	**(-2059,7; -63,98)**	нет	значим
-1007,98	528,59	<	да	(-2055,96; 40)	да	не значим
-70,65	159,22	<	да	(-386,32; 245,01)	да	не значим
-8,07	157,93	<	да	(-321,18; 305,04)	да	не значим
69,96	**30,87**	>	нет	**(8,77; 131,15)**	нет	значим
176,05	173,05	<	да	(-167,05; 519,14)	да	не значим
-217,97	236,76	<	да	(-687,37; 251,44)	да	не значим
-78,32	**15,93**	>	нет	**(-109,9; -46,75)**	нет	значим

[1] Отрицательные значения t-статистики всегда смотрим по абсолютной величине (по модулю).

-211,05	209,63	<	да	(-626,66; 204,57)	да	не значим
-252,60	249,72	<	да	(-747,7; 242,5)	да	не значим
335,03	**136,31**	>	нет	**(64,78; 605,29)**	нет	значим

По результатам расчетов мы видим, что всего 6 коэффициентов регрессии из 14 являются статистически значимыми с вероятностью 95%, остальные 8 – статистически не значимыми.

Верификация модели:

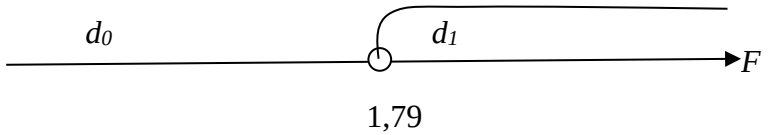
а) дисперсионный анализ.

С помощью критерия Фишера проверим гипотезу *Н*о: $b_1 = b_2 = ... = b_{11} = 0$ (гипотеза об отсутствии линейной функциональной связи).

Найдем критическое значение критерия:

$F_{кр} = F(\alpha; k-1, n-k) = F(0{,}05; 15-1; 121-15) = F(0{,}05; 14; 106) = 1{,}79.$

С помощью MS Excel: =FРАСПОБР(0,05;14;106).



Из полученных расчетов (вывод итогов) наблюдаемое значение критерия $F_0 = 24{,}23 \in d_1$, следовательно, гипотеза Но отклоняется и принимается H_1, т.е. линейная функциональная связь между ценой квартиры и одиннадцатью независимыми переменными существует.

Коэффициент детерминации $R^2 = 0{,}7619 \cdot 100\% = 76{,}19\%$ (см. вывод итогов), т.е. общая вариация (изменчивость) цены на квартиру на 76,19% объясняется вариацией одиннадцати независимых переменных, при этом остальные 23,81% приходятся на неучтенные в модели факторы. Значение коэффициента детерминации достаточно высокое, что свидетельствует о неплохом качестве подгонки.

Общий вывод по модели. Оценивая полученную модель по всем критериям, можно сказать, что ее использование для построения прогноза непригодно, так как из 14 коэффициентов регрессии 8 статистически незначимы, несмотря на хорошие результаты по *F*-статистике. По данной модели можно сказать, что здесь присутствует эффект мультиколлинеарности факторов: большинство *t*-статистик по абсолютной величине меньше критического значения при высоком значении *F*-статистики. Также, если изучить значения парных коэффициентов корреляции между независимыми переменными, то мы видим, что тесная связь наблюдается между: площадью квартиры и количеством комнат $r_{x_1x_2} = 0{,}84$, площадью квартиры и площадью кухни $r_{x_1x_5} = 0{,}65$, площадью квартиры и разновидностью санузла $r_{x_1x_6} = 0{,}55$, наличием лифта и новостройкой (является ли дом новым или нет) $r_{x_{10}x_9} = 0{,}59$.

Последовательно исключая переменные с наименьшей *t*-статистикой, под модулю не превышающей $t_{кр}$ (т.е. наименее значимые факторы), определим какие факторы оказывают наибольшее влияние на стоимость квартир.

Первый фактор, исключаемый из модели – это x_4 (t-статистика = –0,05).

3 этап: Получение второй модели без x_4

Для того чтобы исключать данные из модели быстро и удобно, создадим в рабочем файле копию листа исходных данных (рис. 6.10). На полученном листе – копии исходных данных будем последовательно удалять целиком столбцы, содержащие исключаемые переменные (рис. 6.11).

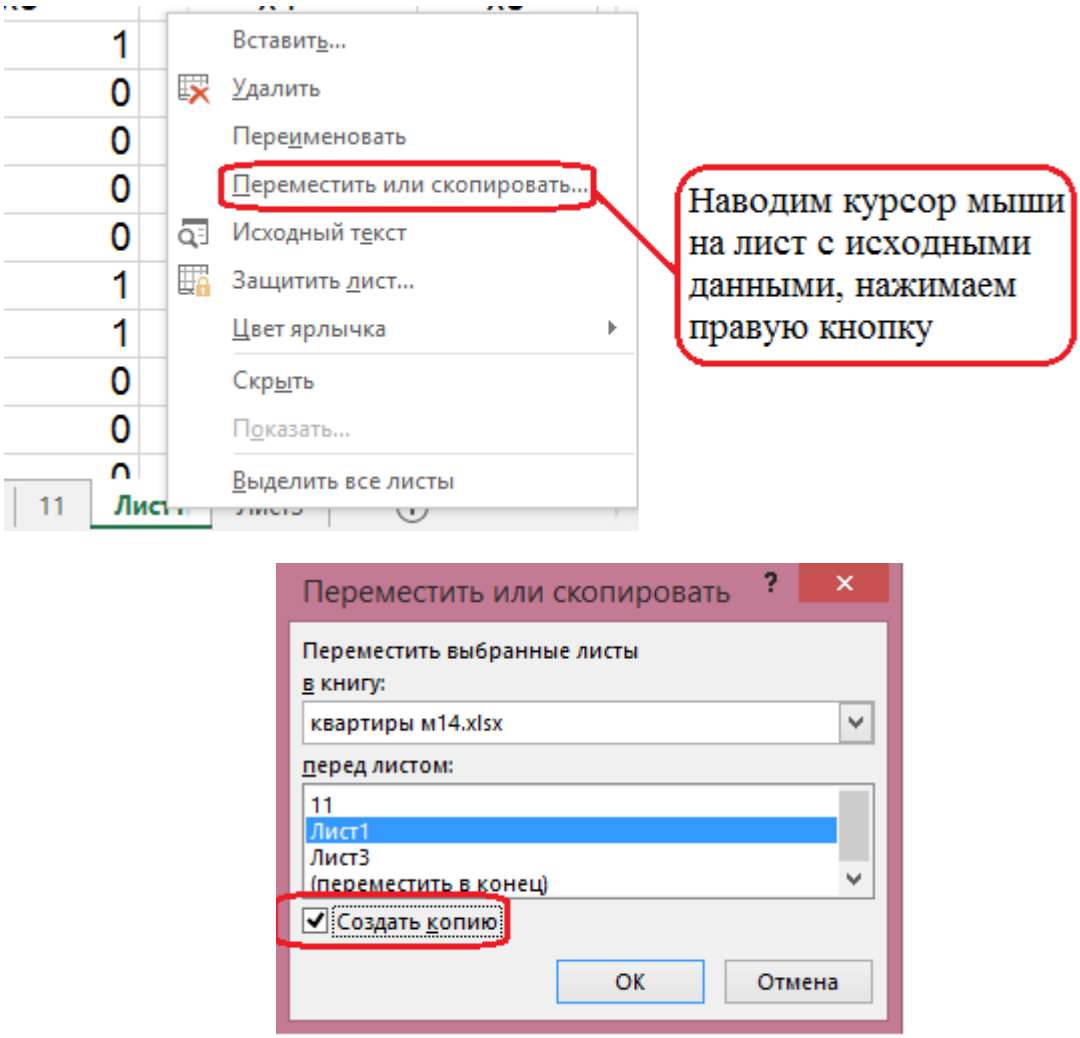


Рис. 6.10. Создание копии листа в MS Excel

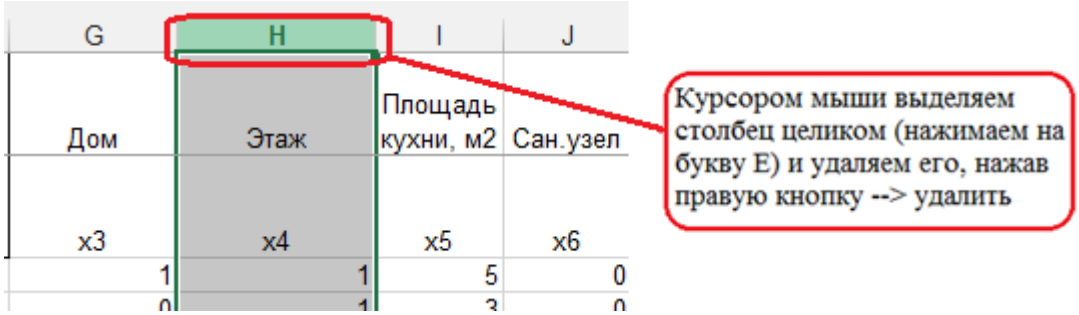


Рис. 6.11. Исключение незначимого фактора

Запускаем Анализ данных => Регрессия, получаем расчеты по второй модели (таблица 6.5)

Таблица 6.5

Вторая модель

Регрессионная статистика						
Множественный R	0,8729					
R-квадрат	0,7619					
Нормированный R-квадрат	0,7330					
Стандартная ошибка	714,0242					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	13	174599786	13430753	26,34	0,00	1,81
Остаток	107	54551874	509831			
Итого	120	229151660				
	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	1761,22	876,52	2,01	0,05	23,61	3498,82
x1	**39,29**	**9,84**	**3,99**	**0,00**	**19,79**	**58,79**
x21	-1334,53	697,60	-1,91	0,06	-2717,44	48,37
x22	**-1340,83**	**588,97**	**-2,28**	**0,02**	**-2508,40**	**-173,26**
x23	**-1066,32**	**493,32**	**-2,16**	**0,03**	**-2044,26**	**-88,37**
x24	-1012,27	519,45	-1,95	0,05	-2042,02	17,49
x3	-70,08	158,08	-0,44	0,66	-383,44	243,29
x5	**69,97**	**30,72**	**2,28**	**0,02**	**9,07**	**130,87**
x6	176,10	172,24	1,02	0,31	-165,35	517,55
x7	-219,59	233,53	-0,94	0,35	-682,53	243,36
x8	**-78,31**	**15,85**	**-4,94**	**0,00**	**-109,74**	**-46,89**
x9	-209,91	207,49	-1,01	0,31	-621,24	201,41
x10	-252,79	248,53	-1,02	0,31	-745,47	239,89
x11	**335,41**	**135,48**	**2,48**	**0,01**	**66,83**	**603,99**

Вычислим $t_{кр}$ с помощью MS Excel: =СТЬЮДРАСПОБР(0,05;121–13), 14 – оцененных параметров, $t_{кр}$ = 1,98:

При этом значение *F*-статистики выросло (26,34), значение коэффициента детерминации осталось на прежнем уровне R^2=0,7619 (76,19%), т.е. влияние фактора x_4 на стоимость квартиры несущественно.

Второй фактор, исключаемый из модели – это x_3 (t-статистика = –0,44).

4 этап: Получение третьей модели без x_3

Удаляем в MS Excel столбец, содержащий переменную x_3. Запускаем Анализ данных => регрессия, в строке входной интервал *X* заменяем: вместо $B$3:$N$124 пишем $B$3:$M$124, получаем расчеты, представленные в таблице 6.6.

Таблица 6.6

Третья модель

Регрессионная статистика						
Множественный R	0,8726					
R-квадрат	0,7615					
Нормированный R-квадрат	0,7350					
Стандартная ошибка	711,3633					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	12	174499591	14541633	28,74	0,00	1,84
Остаток	108	54652070	506038			
Итого	120	229151660				
	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	1770,41	873,01	2,03	0,05	39,94	3500,87
x1	**38,85**	**9,75**	**3,99**	**0,00**	**19,52**	**58,17**
x21	-1367,59	691,01	-1,98	0,05	-2737,30	2,12
x22	**-1371,10**	**582,82**	**-2,35**	**0,02**	**-2526,35**	**-215,85**
x23	**-1092,35**	**487,98**	**-2,24**	**0,03**	**-2059,62**	**-125,08**
x24	-1021,88	517,07	-1,98	0,05	-2046,79	3,04
x5	**68,71**	**30,47**	**2,25**	**0,03**	**8,31**	**129,12**
x6	172,47	171,41	1,01	0,32	-167,29	512,23
x7	-213,86	232,30	-0,92	0,36	-674,33	246,60
x8	**-76,44**	**15,22**	**-5,02**	**0,00**	**-106,60**	**-46,27**
x9	-184,09	198,40	-0,93	0,36	-577,36	209,18
x10	-287,15	235,25	-1,22	0,22	-753,46	179,15
x11	**342,32**	**134,08**	**2,55**	**0,01**	**76,55**	**608,10**

$t_{кр}$=СТЬЮДРАСПОБР(0,05;121-12)=1,98, 13 – оцененных параметров. При этом значение *F*-статистики выросло (28,74), значение коэффициента детерминации уменьшилось R^2 = 0,7615. Доля влияния фактора x_3 очень мала. Исключаемая переменная – x_7.

5 этап: Получение четвертой модели без x_7

Третий фактор, исключаемый из модели – это x_7 (t-статистика = –0,92). Результаты расчетов представлены в таблице 6.7.

Таблица 6.7

Четвертая модель

Регрессионная статистика						
Множественный R	0,8716					
R-квадрат	0,7596					
Нормированный R-квадрат	0,7354					
Стандартная ошибка	710,8656					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	11	174070709	15824610	31,32	0,00	1,88
Остаток	109	55080952	505330			
Итого	120	229151660				
	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	1567,40	844,12	1,86	0,07	-105,61	3240,41
x1	**39,42**	**9,72**	**4,06**	**0,00**	**20,16**	**58,69**
x21	-1345,61	690,12	-1,95	0,05	-2713,40	22,19
x22	**-1345,94**	**581,77**	**-2,31**	**0,02**	**-2499,00**	**-192,89**
x23	**-1088,06**	**487,62**	**-2,23**	**0,03**	**-2054,51**	**-121,61**
x24	-1033,34	516,55	-2,00	0,05	-2057,13	-9,55
x5	**69,31**	**30,45**	**2,28**	**0,02**	**8,96**	**129,65**
x6	152,55	169,92	0,90	0,37	-184,22	489,32
x8	**-78,53**	**15,04**	**-5,22**	**0,00**	**-108,34**	**-48,73**
x9	-215,23	195,36	-1,10	0,27	-602,43	171,97
x10	-295,68	234,90	-1,26	0,21	-761,25	169,90
x11	**333,07**	**133,61**	**2,49**	**0,01**	**68,26**	**597,88**

$t_{кр}$ =СТЬЮДРАСПОБР(0,05;121-11)=1,98, 12 – оцененных параметров. При этом значение F-статистики выросло (31,32), значение коэффициента детерминации уменьшилось R^2 = 0,7596 (75,96%). Доля влияния фактора x_7 не велика. Исключаем фактор x_6

6 этап: Получение пятой модели без x_6

Пятый фактор, исключаемый из модели – это x_6 (t-статистика = 0,90). Результаты расчетов представлены в таблице 6.8.

Таблица 6.8

Пятая модель

Регрессионная статистика						
Множественный R	0,8705					
R-квадрат	0,7579					
Нормированный R-квадрат	0,7358					
Стандартная ошибка	710,2385					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	10	173663403	17366340	34,43	0,00	1,92
Остаток	110	55488257	504439			
Итого	120	229151660				
	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	1542,85	842,93	1,83	0,07	-127,63	3213,34
x1	**40,67**	**9,61**	**4,23**	**0,00**	**21,62**	**59,72**
x21	-1357,89	689,37	-1,97	0,05	-2724,07	8,29
x22	**-1314,74**	**580,22**	**-2,27**	**0,03**	**-2464,60**	**-164,88**
x23	**-1078,79**	**487,08**	**-2,21**	**0,03**	**-2044,07**	**-113,51**
x24	-1014,52	515,67	-1,97	0,05	-2036,46	7,43
x5	**71,08**	**30,35**	**2,34**	**0,02**	**10,93**	**131,24**
x8	**-76,63**	**14,87**	**-5,15**	**0,00**	**-106,11**	**-47,16**
x9	-182,15	191,68	-0,95	0,34	-562,02	197,73
x10	-341,49	229,09	-1,49	0,14	-795,50	112,51
x11	**344,98**	**132,83**	**2,60**	**0,01**	**81,73**	**608,22**

$t_{кр}$ =СТЬЮДРАСПОБР(0,05;121–10)=1,98, 11 – оцененных параметров. При этом значение *F*-статистики выросло (34,43), значение коэффициента детерминации уменьшилось R^2=0,7579. (75,79%). Доля влияния фактора x_6 не велика. Исключаемый фактор – x_9.

7 этап: Получение шестой модели без x_9

Шестой фактор, исключаемый из модели – это x_9 (t-статистика = –0,95). Результаты расчетов представлены в таблице 6.9.

Таблица 6.9

Шестая модель

Регрессионная статистика						
Множественный R	0,8694					
R-квадрат	0,7559					
Нормированный R-квадрат	0,7361					
Стандартная ошибка	709,9279					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	9	173207919	19245324	38,19	0,00	1,97
Остаток	111	55943741	503998			
Итого	120	229151660				

	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	1468,23	838,89	1,75	0,08	-194,10	3130,55
x1	**40,20**	**9,60**	**4,19**	**0,00**	**21,18**	**59,21**
x21	-1242,63	678,32	-1,83	0,07	-2586,78	101,51
x22	**-1187,89**	**564,41**	**-2,10**	**0,04**	**-2306,31**	**-69,48**
x23	**-972,75**	**473,92**	**-2,05**	**0,04**	**-1911,85**	**-33,65**
x24	-885,75	497,33	-1,78	0,08	-1871,25	99,74
x5	**65,03**	**29,67**	**2,19**	**0,03**	**6,25**	**123,81**
x8	**-79,17**	**14,63**	**-5,41**	**0,00**	**-108,15**	**-50,19**
x10	**-437,31**	**205,62**	**-2,13**	**0,04**	**-844,75**	**-29,86**
x11	**347,51**	**132,75**	**2,62**	**0,01**	**84,46**	**610,57**

$t_{кр}$ =СТЬЮДРАСПОБР(0,05;121–9)=1,98, 10 – оцененных параметров. При этом значение F-статистики выросло (38,19), значение коэффициента детерминации осталось на прежнем уровне R^2=0,7559. (75,59%). Доля влияния фактора x_9 не велика. Исключаемый фактор – x_{2_4}.

8 этап: Получение седьмой модели без x_{2_4}

Седьмой фактор, исключаемый из модели – это x_{2_4} (t-статистика = –1,78). Результаты расчетов представлены в таблице 6.10.

Таблица 6.10

Седьмая модель

Регрессионная статистика						
Множественный R	0,8654					
R-квадрат	0,7489					
Нормированный R-квадрат	0,7309					
Стандартная ошибка	716,7786					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	8	171609245	21451156	41,75	0,00	2,02
Остаток	112	57542415	513772			
Итого	120	229151660				
	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	665,50	714,36	0,93	0,35	-749,91	2080,91
x1	**43,55**	**9,50**	**4,58**	**0,00**	**24,73**	**62,38**
x21	-467,36	525,25	-0,89	0,38	-1508,07	573,35
x22	-458,14	391,91	-1,17	0,24	-1234,65	318,37
x23	-302,71	290,99	-1,04	0,30	-879,27	273,84
x5	54,00	29,29	1,84	0,07	-4,04	112,04
x8	**-81,16**	**14,72**	**-5,51**	**0,00**	**-110,33**	**-51,98**
x10	**-450,20**	**207,47**	**-2,17**	**0,03**	**-861,28**	**-39,11**
x11	**364,60**	**133,68**	**2,73**	**0,01**	**99,74**	**629,47**

$t_{кр}$ =СТЬЮДРАСПОБР(0,05;121–8)=1,98, 9 – оцененных параметров. При этом значение *F*-статистики выросло (41,75), значение коэффициента детерминации уменьшилось R^2=0,7489 (74,89%). Доля влияния фактора x_{2_4} не велика. Исключаемый фактор – x_{2_1}.

9 этап: Получение восьмой модели без x_{2_1}

Восьмой фактор, исключаемый из модели – это x_{2_1} (*t*-статистика = –0,89). Результаты расчетов представлены в таблице 6.11.

Таблица 6.11

Регрессионная статистика						
Множественный R	0,8644					
R-квадрат	0,7471					
Нормированный R-квадрат	0,7314					
Стандартная ошибка	716,1177					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	7	171202481	24457497	47,69	0,00	2,09
Остаток	113	57949179	512825			
Итого	120	229151660				
	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	58,50	211,76	0,28	0,78	-361,03	478,03
x1	**50,72**	**5,03**	**10,09**	**0,00**	**40,76**	**60,68**
x22	-138,38	156,21	-0,89	0,38	-447,85	171,09
x23	-114,99	200,23	-0,57	0,57	-511,68	281,70
x5	45,40	27,63	1,64	0,10	-9,33	100,13
x8	**-79,82**	**14,63**	**-5,45**	**0,00**	**-108,81**	**-50,83**
x10	**-531,58**	**186,05**	**-2,86**	**0,01**	**-900,18**	**-162,97**
x11	**369,73**	**133,43**	**2,77**	**0,01**	**105,38**	**634,09**

$t_{кр}$ =СТЬЮДРАСПОБР(0,05;121-7)=1,98, 8 – оцененных параметров. При этом значение F-статистики выросло (47,69), значение коэффициента детерминации уменьшилось R^2 = 0,7471 (74,71%). Доля влияния фактора x_{2_1} не велика. Исключаемый фактор – x_{2_3}.

10 этап: Получение девятой модели без x_{2_3}

Девятый фактор, исключаемый из модели – это x_{2_3} (*t*-статистика = –0,57). Результаты расчетов представлены в таблице 6.12.

Таблица 6.12

Регрессионная статистика						
Множественный R	0,8639					
R-квадрат	0,7464					
Нормированный R-квадрат	0,7330					
Стандартная ошибка	714,0097					
Наблюдения	121					

Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	6	171033337	28505556	55,91	0,00	2,18
Остаток	114	58118324	509810			
Итого	120	229151660				
	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	66,14	210,72	0,31	0,75	-351,28	483,57
x1	**49,17**	**4,22**	**11,65**	**0,00**	**40,81**	**57,53**
x22	-102,41	142,68	-0,72	0,47	-385,06	180,24
x5	49,25	26,72	1,84	0,07	-3,68	102,18
x8	**-79,22**	**14,55**	**-5,44**	**0,00**	**-108,05**	**-50,39**
x10	**-526,71**	**185,31**	**-2,84**	**0,01**	**-893,82**	**-159,61**
x11	**367,44**	**132,98**	**2,76**	**0,01**	**104,01**	**630,88**

$t_{кр}$ =СТЬЮДРАСПОБР(0,05;121–6)=1,98, 7 – оцененных параметров. При этом значение *F*-статистики выросло (55,91), значение коэффициента детерминации уменьшилось R^2 = 0,7464 (74,64%). Доля влияния фактора x_{2_3} не велика. Исключаемый фактор – x_{2_2}.

11 этап: Получение десятой модели без x_{2_2}

Десятый фактор, исключаемый из модели – это x_{2_2} (*t*-статистика = –0,72). Результаты расчетов представлены в таблице 6.13.

Таблица 6.13

Регрессионная статистика						
Множественный R	0,8633					
R-квадрат	0,7452					
Нормированный R-квадрат	0,7342					
Стандартная ошибка	712,5031					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	5	170770688	34154138	67,28	0,00	2,29
Остаток	115	58380972	507661			
Итого	120	229151660				
	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	38,10	206,63	0,18	0,85	-371,18	447,39
x1	**49,48**	**4,19**	**11,81**	**0,00**	**41,18**	**57,78**
x5	47,14	26,50	1,78	0,08	-5,35	99,64
x8	**-79,79**	**14,50**	**-5,50**	**0,00**	**-108,51**	**-51,07**
x10	**-510,72**	**183,58**	**-2,78**	**0,01**	**-874,36**	**-147,08**
x11	**361,54**	**132,44**	**2,73**	**0,01**	**99,19**	**623,89**

$t_{кр}$ =СТЬЮДРАСПОБР(0,05;121–5)=1,98, 6 – оцененных параметров. При этом значение *F*-статистики выросло (67,28), значение коэффициента детерминации уменьшилось R^2 = 0,7452 (74,52%). Доля влияния фактора x_{2_2} не велика. Исключаемый фактор – x_5.

12 этап: Получение одиннадцатой модели без x_5

Одиннадцатый фактор, исключаемый из модели – это x_5 (t-статистика = 1,78). Результаты расчетов представлены в таблице 6.14.

Таблица 6.14

Одиннадцатая модель

Регрессионная статистика						
Множественный R	0,8592					
R-квадрат	0,7382					
Нормированный R-квадрат	0,7292					
Стандартная ошибка	719,1188					
Наблюдения	121					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Знач. F*	*F крит.*
Регрессия	4	169164373	42291093	81,78	0,00	2,45
Остаток	116	59987287	517132			
Итого	120	229151660				
	Коэффициенты	*Стандартная ошибка*	*t-стат.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	142,37	199,98	0,71	0,48	-253,72	538,45
x1	**54,42**	**3,17**	**17,18**	**0,00**	**48,14**	**60,69**
x8	**-78,97**	**14,63**	**-5,40**	**0,00**	**-107,94**	**-49,99**
x10	**-338,47**	**157,41**	**-2,15**	**0,03**	**-650,24**	**-26,69**
x11	**327,95**	**132,31**	**2,48**	**0,01**	**65,89**	**590,00**

$t_{кр}$ =СТЬЮДРАСПОБР(0,05;121–4)=1,98, 5 – оцененных параметров. При этом значение F-статистики выросло (81,78), значение коэффициента детерминации уменьшилось $R^2 = 0,7382$ (73,82%). Доля влияния фактора x_5 не велика.

Все коэффициенты статистически значимы с вероятностью 95%, за исключением коэффициента b_0.

13 этап: Получение двенадцатой и тринадцатой моделей

Посмотрим, какими будут результаты расчетов при исключении оставшихся переменных из модели (табл. 6.15).

Таблица 6.15

Сравнение 12-й и 13-й моделей

Переменные	*Коэффициенты*	*Станд. ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
12-я модель: при исключении x_{10}						
y-пересечение	130,11	202,97	0,64	0,52	-271,86	532,08
x1	53,27	3,17	16,81	0,00	46,99	59,55
x8	-76,90	14,82	-5,19	0,00	-106,25	-47,55
x11	286,39	132,90	2,15	0,03	23,19	549,60
13-я модель: при исключении x_{11}						

y-пересечение	254,67	197,55	1,29	0,20	-136,53	645,86
x1	53,53	3,22	16,65	0,00	47,16	59,90
x8	-76,62	15,05	-5,09	0,00	-106,42	-46,83

Из таблицы видно, что коэффициенты при двух переменных x_1 и x_8 неизменны в случае, когда в модели две переменных или три (включая x_{11}) ,при добавлении в модель переменной x_{10} значения коэффициентов изменяются. Следует обратить внимание на следующий факт: при сравнении знаков коэффициентов корреляции оставшихся четырех переменных со знаками коэффициентов регрессии в трех моделях, мы видим, что у коэффициента b_{10}=–338,47 коэффициента $r_{yx_{10}}$ = 0,07 знаки разные. Поэтому целесообразно оставить модель с тремя переменными: x_1, x_8 и x_{11}.

Полученная модель по критерию Фишера является пригодной (F=104,27, $F_{крит}$ = 2,68, R^2 = 0,7278). Уравнение модели имеет вид:

$$y = 130{,}11 + 53{,}27x_1 - 76{,}9x_8 + 286{,}39x_{11}$$

$$(t_{b_0} = 0{,}64)\ (16{,}81)\quad(-5{,}19)\quad(2{,}15)$$

Средняя ошибка аппроксимации

На полученный лист (для трех переменных) копируем значения переменной y рядом с колонкой «остатки» (рис. 6.12).

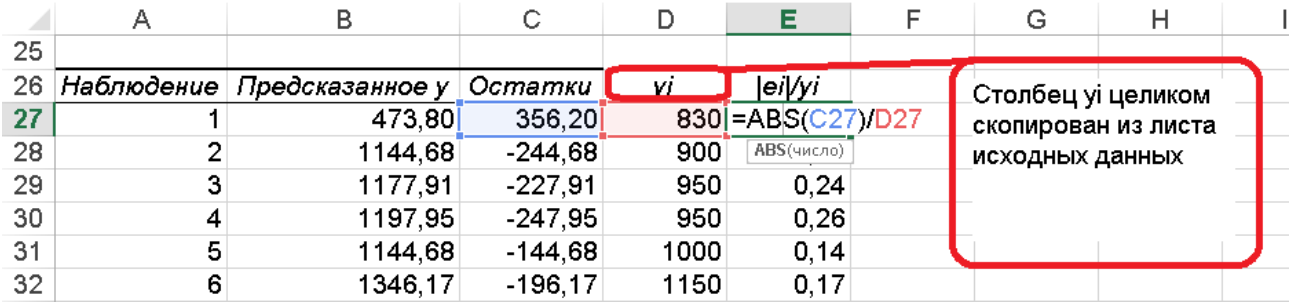


Рис. 6.12. Вычисление ошибки аппроксимации

Вычисляем:

$A = \frac{1}{n}\sum\frac{|e_i|}{y_i}\cdot 100\% = \frac{1}{121}\cdot 20{,}88\cdot 100\% = 17{,}26\%$ – свидетельствует об удовлетворительном качестве аппроксимации, так как A больше 15%.

14 этап: Интерпретация коэффициентов регрессии

b_0 = 130,11 тысяч рублей интерпретации не подлежит, так как не значим;

b_1 = 53,27 тысяч рублей означает, что если площадь квартиры увеличится на 1 квадратный метр, то стоимость квартиры увеличится на 53 270 рублей (можно считать, что стоимость квадратного метра на вторичном рынке жилья г. Иркутска составляет 53 270 рублей);

b_8 = –76,9 тысяч рублей означает, что если удаленность от центра города на 1 км уменьшает стоимость квартиры в среднем на 76 900 рублей;

b_{11} = 286,39 тысяч рублей означает, что если квартира в хорошем состоянии (т.е. ремонт не нужен), то ее стоимость будет выше на 286 390 рублей, и соответственно, если в квартире требуется ремонт, то она будет дешевле на эту сумму.

Рассчитаем коэффициенты частной эластичности по формуле $Э_i = \frac{b_i}{\bar{y}_i} \cdot \bar{x}_i$ [2]:

$Э_1 = \frac{53{,}27}{2637{,}926} \cdot 52{,}374 = 1{,}06\%$, где $\bar{x}_1 = \frac{\sum x_1}{n} = 52{,}374\,м^2$, $\bar{y} = \frac{\sum y}{n} = 2637{,}926$ тыс. руб.

$Э_8 = \frac{-76{,}9}{2637{,}926} \cdot 5{,}485 = -0{,}16\%$, где $\bar{x}_8 = \frac{\sum x_8}{n} = 5{,}485$ км.

Цена квартиры увеличится на 1,06%, если площадь увеличится на 1%, и на уменьшится 0,16%, если расстояние увеличится на 1% от центра города.

R^2 = 0,7278, показывает, что вариация цены на квартиру на 72,78% объясняется изменчивостью площади квартиры, удаленности от центра города и наличия ремонта, при этом оставшиеся 27,22% приходятся на неучтенные в модели факторы, из них всего лишь 2,57% на 11 исключенных факторов.

15 этап: частная корреляция

Рассчитаем коэффициенты частной корреляции, чтобы оценить чистое влияние переменных x_1, x_8 и x_{11} на стоимость квартиры:

$$r_{yx_1/x_8} = \frac{r_{yx_1} - r_{yx_8} r_{x_1x_8}}{\sqrt{\left(1 - r^2_{yx_8}\right)\left(1 - r^2_{x_1x_8}\right)}}; \quad r_{yx_1/x11} = \frac{r_{yx_1} - r_{yx_{11}} r_{x_1x_{11}}}{\sqrt{\left(1 - r^2_{yx_{11}}\right)\left(1 - r^2_{x_1x_{11}}\right)}};$$

$$r_{yx_8/x_1} = \frac{r_{yx_8} - r_{yx_1} r_{x_1x_8}}{\sqrt{\left(1 - r^2_{yx_1}\right)\left(1 - r^2_{x_1x_8}\right)}}; \quad r_{yx_8/x_{11}} = \frac{r_{yx_8} - r_{yx_{11}} r_{x_{11}x_8}}{\sqrt{\left(1 - r^2_{yx_{11}}\right)\left(1 - r^2_{x_{11}x_8}\right)}};$$

$$r_{yx_{11}/x_1} = \frac{r_{yx_{11}} - r_{yx_1} r_{x_1x_{11}}}{\sqrt{\left(1 - r^2_{yx_1}\right)\left(1 - r^2_{x_1x_{11}}\right)}}; \quad r_{yx_{11}/x_8} = \frac{r_{yx_{11}} - r_{yx_8} r_{x_{11}x_8}}{\sqrt{\left(1 - r^2_{yx_8}\right)\left(1 - r^2_{x_{11}x_8}\right)}}.$$

Матрица парных коэффициентов корреляции для указанных переменных имеет вид:

	y	*x1*	*x8*	*x11*
y	1			
x1	0,809	1		
x8	-0,229	0,025	1	
x11	0,133	0,038	0,010	1

[2] Средние значения переменных y, x_1 и x_8 рассчитываем в MS Excel с помощью функции =СРЗНАЧ().

Вычисления в MS Excel[3]:

	A	B	C	D	E	F	G	H	I	J
15	ryx1	0,809		ryx1/x8	0,837	<=(B15-B16*B18)/((1-B16^2)*(1-B18^2))^0,5				
16	ryx8	-0,229		ryx1/x11	0,812	<=(B15-B17*B19)/((1-B17^2)*(1-B19^2))^0,5				
17	ryx11	0,133		ryx8/x1	-0,424	<=(B10-B15*B18)/КОРЕНЬ((1-B15^2)*(1-B18^2))				
18	rx1x8	0,025		ryx8/x11	-0,232	<=(B16-B17*B20)/КОРЕНЬ((1-B17*B17)*(1-B20*B20))				
19	rx1x11	0,038		ryx11/x1	0,173	<=(B17-B15*B19)/КОРЕНЬ((1-B15*B15)*(1-B19*B19))				
20	rx8x11	0,010		ryx11/x8	0,139	<=(B17-B16*B20)/КОРЕНЬ((1-B16*B16)*(1-B20*B20))				

$r_{yx_1/x_8} = \dfrac{0{,}809-(-0{,}229)\cdot 0{,}025}{\sqrt{\left(1-(-0{,}229)^2\right)\left(1-0{,}025^2\right)}} = 0{,}837 > 0{,}809$ – имеется небольшое влияние фактора x_8.

$r_{yx_1/x11} = \dfrac{0{,}809-0{,}133\cdot 0{,}038}{\sqrt{\left(1-0{,}133^2\right)\left(1-0{,}038^2\right)}} = 0{,}812 > 0{,}809$ – влияние фактора x_{11} фактически отсутствует.

$r_{yx_8/x_1} = \dfrac{-0{,}229-0{,}809\cdot 0{,}025}{\sqrt{\left(1-0{,}809^2\right)\left(1-0{,}025^2\right)}} = -0{,}424 \Rightarrow |-0{,}424| > |-0{,}229|$ – имеется существенное влияние фактора x_1.

$r_{yx_8/x_{11}} = \dfrac{-0{,}229-0{,}133\cdot 0{,}010}{\sqrt{\left(1-0{,}133^2\right)\left(1-0{,}010^2\right)}} = -0{,}232 \Rightarrow |-0{,}232| > |-0{,}229|$ – влияние фактора x_{11} фактически отсутствует.

$r_{yx_{11}/x_1} = \dfrac{0{,}133-0{,}809\cdot 0{,}038}{\sqrt{\left(1-0{,}809^2\right)\left(1-0{,}038^2\right)}} = 0{,}137 > 0{,}133$ – влияние фактора x_1 фактически отсутствует.

$r_{yx_{11}/x_8} = \dfrac{0{,}133-(-0{,}229)*0{,}010}{\sqrt{\left(1-(-0{,}229)^2\right)\left(1-0{,}010^2\right)}} = 0{,}139 > 0{,}133$ влияние фактора x_8 фактически отсутствует.

Таким образом, среди шести рассчитанных коэффициентов частной корреляции, существенное изменение коэффициента в большую сторону принесло исключение влияния переменной x_1 из коэффициента корреляции между y и x_8.

16 этап: точечный и интервальный прогноз

Рассчитаем точечный прогноз, предположив значения переменных следующими:

пусть площадь объекта жилой недвижимости составляет 80 квадратных метров, расположен объект на остановке Байкальский микрорайон (6 км) и не требует

[3] в расчетах специально приведены разные способы вычисления одной и той же формулы (корень или степень 0,5, квадрат или умножить число само на себя).

ремонта (1). Вектор $x^0 = (1; 80; 6; 1)$. Подставим его в уравнение модели. Найдем стоимость квартиры:

$$\hat{y}_0 = 130{,}11 + 53{,}27 \cdot 80 - 76{,}9 \cdot 6 + 286{,}39 \cdot 1 = 4216 \text{ тыс. руб.}$$

Для расчета $\hat{D}(\hat{y}_0)$ по формуле (6.8) последовательно вычислим: $(X^t X)$, $(X^t X)^{-1}$, $x^0 (X^t X)^{-1}$, $x^0 (X^t X)^{-1} x^{0^t}$. Матрица X – это матрица, в которой первый столбец состоит из 1, второй столбец – значения переменной x_1, третий – x_8, четвертый – x_{11}. Используя уже известный по 1 этапу алгоритм, получим:

$(X^t X) =$	121	6337,2	663,7	59
	6337,2	385078,3	35045,45	3138,5
	663,7	35045,45	6069,57	326,2
	59	3138,5	326,2	59

Для вычисления матрицы $(X^t X)^{-1}$ воспользуемся функцией =МОБР(массив), дальнейшие вычисления обратной матрицы аналогичны функции =МУМНОЖ(массив):

$(X^t X)^{-1} =$	0,0773	-0,0010	-0,0021	-0,0144
	-0,0010	0,0000	0,0000	0,0000
	-0,0021	0,0000	0,0004	0,0000
	-0,0144	0,0000	0,0000	0,0331

$x^0 (X^t X)^{-1}$	-0,0267	0,0005	0,0001	0,0161

$$x^0 (X^t X)^{-1} x^{0^t} = 0{,}31.$$

$$\hat{\sigma}^2 = MS_{ост} = 533146{,}8.$$

$$\hat{D}(\hat{y}_0) = 533146{,}8 \cdot (1 + 0{,}31) = 549421{,}31.$$

По формуле (6.8) получаем:

$$4216 - 1{,}98 \cdot \sqrt{549421{,}31} < y_0 < 4216 + 1{,}98 \cdot \sqrt{549421{,}31}$$

$$2748{,}03 < y_0 < 5683{,}97.$$

Таким образом, с вероятностью 95% квартира в 6 км от центра города площадью 80 м^2 и не требующая ремонта, будет стоить в пределах от 2748, 03 до 5683,97 тыс. руб.

Список рекомендуемой литературы

1. Ежова Л.Н. Эконометрика. Начальный курс с основами теории вероятностей и математической статистики / Л.Н. Ежова. – Иркутск: Изд-во БГУЭП, 2008. – 287 с.

2. Доугерти К. Введение в эконометрику / К. Доугерти. – М.: ИНФРА-М, 1999. – 402 с.

3. Колемаев В.А. Эконометрика: учебник / В.А. Колемаев. – М.: ИНФРА-М, 2009. – 160 с.

4. Кремер Н.Ш., Путко Б.А. Эконометрика: учебник / Н.Ш. Кремер, Б.А. Путко. – М.: ЮНИТИ-ДАНА, 2006. – 310 с.

5. Магнус Я.Р. Эконометрика. Начальный курс: учебник. / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – М.: Дело, 2004. – 576 с.

6. Эконометрика: учебник / И.И. Елисеева [и др.]; под ред. И. И. Елисеевой. М.: Финансы и статистика, 2001. – 344 с.

7. Практикум по эконометрике: учебное пособие / И.И. Елисеева [и др.]; под ред. И. И. Елисеевой. М.: Финансы и статистика, 2001. – 192 с.

8. Эконометрика: учебник для магистров / И.И. Елисеева [и др.]; под ред. И. И. Елисеевой. М.: Юрайт, 2012. – 453 с.

Учебное издание

Леонова Ольга Васильевна
Шерстянкина Нина Павловна

Эконометрика
Курс лекций и методические указания по выполнению расчетно-графических работ

Учебное пособие

Издается в авторской редакции

ИД № 06318 от 26.11.01.
Подписано в пользование 03.07.17.
Издательство Байкальского государственного университета.
664003, г. Иркутск, ул. Ленина, 11.
http://bgu.ru.