АКАДЕМИЯ УПРАВЛЕНИЯ «ТИСБИ»

А.К. Шалабанов, Д.А. Роганов

ПРАКТИКУМ ПО ЭКОНОМЕТРИКЕ С ПРИМЕНЕНИЕМ MS EXCEL

Линейные модели парной и множественной регрессии

КАЗАНЬ 2008

Рекомендовано к печати

Научно-методическим советом

Академии управления «ТИСБИ»

Составители: Шалабанов А.К., Роганов Д.А.

Рецензенты: К.ф-м.н, доц. кафедры теоретической кибернетики Казанского государственного университета Нурмеев Н.Н.

К.т.н. доцент кафедры математики Академии управления «ТИСБИ» Печеный Е.А.

Практикум по эконометрики содержит основные понятия и формулы эконометрики из разделов по парной и множественной регрессии и корреляции. Предназначено для студентов дневного и дистанционного отделения Академии управления «ТИСБИ». Подробно разобраны типовые задачи. Продемонстрирована возможность реализации решения задач в MS Excel. Представлены варианты индивидуальных контрольных заданий.

Содержание

Введение

Успешная работа современного экономиста в любой области экономики тесным образом связана с использованием математических методов и средств вычислительной техники. При решении задач из различных областей человеческой деятельности часто приходится использовать методы, основанные на эконометрических моделях. Эконометрика – одна из базовых дисциплин экономического образования во всем мире, но в России данный предмет только начал входить в учебные планы обучения будущих экономистов, так как прежде в СССР в условиях централизованной плановой экономике эконометрика была попросту не нужна.

Практикум по эконометрики предназначен для студентов дневного и дистанционного отделения Академии управления «ТИСБИ» и содержит в себе подробные примеры решения типовых задач и варианты контрольных заданий. Предлагаемый материал должен способствовать формированию у студентов практических навыков использования эконометрических методов при решении конкретных задач. Предполагается, что студенты ознакомлены с курсами линейной алгебры, математического анализа, теории вероятностей и математической статистики.

Для самостоятельного решения студентам предлагается две задачи. Для большего понимания перед их решением желательно изучить теоретический материал по учебникам, которые приведены в списке литературы, хотя необходимые формулы и методы приведены в методических указаниях. Так же, предлагаемые задачи могут быть решены (частично или полностью) на компьютере с помощью различных пакетов прикладных программ (ППП). В данном пособии приведены примеры решения в MS Excel, т.к. данная программа присутствует в подавляющем большинстве персональных компьютеров.

При решении без использования компьютера рекомендуется производить промежуточные вычисления с точностью до пяти–шести знаков после запятой.

1. Определение эконометрики

Эконометрика – быстроразвивающаяся отрасль науки, цель которой состоит в том, чтобы придать количественные меры экономическим отношениям.

Термин «эконометрика» был впервые введен бухгалтером П. Цьемпой (Австро-Венгрия, 1910 г.). Цьемпа считал, что если к данным бухгалтерского учета применить методы алгебры и геометрии, то будет получено новое, более глубокое представление о результатах хозяйственной деятельности. Это употребление термина, как и сама концепция, не прижилось, но название «эконометрика» оказалось весьма удачным для определения нового направления в экономической науке, которое выделилось в 1930 г.

Слово «эконометрика» представляет собой комбинацию двух слов: «экономика» и «метрика» (от греч. «метрон»). Таким образом, сам термин подчеркивает специфику, содержание эконометрики как науки: количественное выражение тех связей и соотношений, которые раскрыты и обоснованы экономической теорией. И. Шумпетер (1883–1950), один из первых сторонников выделения этой новой дисциплины, полагал, что в соответствии со своим назначением эта дисциплина должна называться «экономометрика». Советский ученый А.Л. Вайнштейн (1892–1970) считал, что название настоящей науки основывается на греческом слове *метрия* (геометрия, планиметрия и т.д.), соответственно по аналогии – эконометрия. Однако в мировой науке общеупотребимым стал термин «эконометрика». В любом случае, какой бы мы термин ни выбрали, эконометрика является наукой об измерении и анализе экономических явлений.

Зарождение эконометрики является следствием междисциплинарного подхода к изучению экономики. Эта наука возникла в результате взаимодействия и объединения в особый «сплав» трех компонент: экономической теории, статистических и математических методов. Впоследствии к ним присоединилось развитие вычислительной техники как условие развития эконометрики.

В журнале «Эконометрика», основанном в 1933 г. Р. Фришем (1895–1973), он дал следующее определение эконометрики: «Эконометрика – это не то же самое, что экономическая статистика. Она не идентична и тому, что мы называем экономической теорией, хотя значительная часть этой теории носит количественный характер. Эконометрика не является синонимом приложений математики к экономике. Как показывает опыт, каждая из трех отправных точек – статистика, экономическая теория и математика – необходимое, но не достаточное условие для понимания количественных соотношений в современной экономической жизни. Это – единство всех трех составляющих. И это единство образует эконометрику».

Таким образом, эконометрика – это наука, которая дает количественное выражение взаимосвязей экономических явлений и процессов.

2. Парная регрессия и корреляция

2.1. Теоретическая справка

Парная (простая) линейная регрессия представляет собой модель, где среднее значение зависимой (объясняемой) переменной рассматривается как функция одной независимой (объясняющей) переменной x, т.е. это модель вида:

$$\hat{y}_x = f(x). \tag{2.1}$$

Так же y называют результативным признаком, а x признаком-фактором. Знак «^» означает, что между переменными x и y нет строгой функциональной зависимости.

Практически в каждом отдельном случае величина y складывается из двух слагаемых:

$$y = \hat{y}_x + \varepsilon, \tag{2.2}$$

где y – фактическое значение результативного признака; $\hat{y}_x$ – теоретическое значение результативного признака, найденное исходя из уравнения регрессии; ε – случайная величина, характеризующая отклонения реального значения результативного признака от теоретического, найденного по уравнению регрессии.

Случайная величина ε называется также возмущением. Она включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Ее присутствие в модели порождено тремя источниками: спецификацией модели, выборочным характером исходных данных, особенностями измерения переменных.

Различают *линейные* и *нелинейные* регрессии.

Линейная регрессия: $y = a + b \cdot x + \varepsilon$.

Нелинейные регрессии делятся на два класса: регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, и регрессии, нелинейные по оцениваемым параметрам. Например:

регрессии, *нелинейные по объясняющим переменным*:

• полиномы разных степеней $y = a + b_1 \cdot x + b_2 \cdot x^2 + \ldots + b_n \cdot x^n + \varepsilon$;

• равносторонняя гипербола $y = a + \frac{b}{x} + \varepsilon$;

регрессии, *нелинейные по оцениваемым параметрам*:

• степенная $y = a \cdot x^b \cdot \varepsilon$;

• показательная $y = a \cdot b^x \cdot \varepsilon$;

• экспоненциальная $y = e^{a+bx+\varepsilon}$.

Построение уравнения регрессии сводится к оценке ее параметров. Для оценки параметров регрессий, линейных по параметрам, используют *метод наименьших квадратов (МНК)*. МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических $\hat{y}_x$ минимальна, т.е.

$$\sum \left(y - \hat{y}_x \right)^2 \to \min \tag{2.3}$$

Для линейных и нелинейных уравнений, приводимых к линейным, решается следующая система относительно a и b:

$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum xy. \end{cases} \tag{2.4}$$

Можно воспользоваться готовыми формулами, которые вытекают непосредственно из решения этой системы:

$$a = \bar{y} - b \cdot \bar{x}, \qquad b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2}, \tag{2.5}$$

где $\operatorname{cov}(x,y)=\overline{y\cdot x}-\overline{y}\cdot\overline{x}$ – ковариация признаков x и y, $\sigma_x^2=\overline{x^2}-\overline{x}^2$ – дисперсия признака x и

$$\overline{x}=\frac{1}{n}\sum x,\ \overline{y}=\frac{1}{n}\sum y,\ \overline{y\cdot x}=\frac{1}{n}\sum y\cdot x,\ \overline{x^2}=\frac{1}{n}\sum x^2.$$

(Ковариация – числовая характеристика совместного распределения двух случайных величин, равная математическому ожиданию произведения отклонений этих случайных величин от их математических ожиданий. Дисперсия – характеристика случайной величины, определяемая как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания. Математическое ожидание – сумма произведений значений случайной величины на соответствующие вероятности.)

Тесноту связи изучаемых явлений оценивает *линейный коэффициент парной корреляции* r_{xy} для линейной регрессии $(-1\le r_{xy}\le 1)$:

$$r_{xy}=b\cdot\frac{\sigma_x}{\sigma_y}=\frac{\operatorname{cov}(x,y)}{\sigma_x\cdot\sigma_y} \tag{2.6}$$

и *индекс корреляции* ρ_{xy} – для нелинейной регрессии $(0\le\rho_{xy}\le 1)$:

$$\rho_{xy}=\sqrt{1-\frac{\sigma^2_{\text{ост}}}{\sigma_y^2}}=\sqrt{1-\frac{\sum(y-\hat{y}_x)^2}{\sum(y-\overline{y})^2}},$$

где $\sigma_y^2=\sum(y-\overline{y})^2$ – общая дисперсия результативного признака y; $\sigma^2_{\text{ост}}=\sum(y-\hat{y}_x)^2$ – остаточная дисперсия, определяемая исходя из уравнения регрессии $\hat{y}_x=f(x)$.

Оценку качества построенной модели даст коэффициент (индекс) детерминации r_{xy}^2 (для линейной регрессии) либо ρ_{xy}^2 (для нелинейной регрессии), а также средняя ошибка аппроксимации.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических:

$$\bar{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right|\cdot 100\%. \quad (2.7)$$

Допустимый предел значений $\bar{A}$ – не более 10%.

Средний коэффициент эластичности $\bar{Э}$ показывает, на сколько процентов в среднем по совокупности изменится результат y от своей средней величины при изменении фактора x на 1% от своего среднего значения:

$$\bar{Э} = f'(x)\frac{\bar{x}}{\bar{y}}. \quad (2.8)$$

После того как найдено уравнение линейной регрессии, проводится *оценка значимости* как уравнения в целом, так и отдельных его параметров.

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

Оценка значимости уравнения регрессии в целом производится на основе F*-критерия Фишера*, которому предшествует дисперсионный анализ. Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной y от среднего значения $\bar{y}$ раскладывается на две части – «*объясненную*» и «*необъясненную*»:

$$\sum(y-\bar{y})^2 = \sum(\hat{y}_x-\bar{y})^2 + \sum(y-\hat{y}_x)^2,$$

где $\sum(y-\bar{y})^2$ – общая сумма квадратов отклонений; $\sum(\hat{y}_x-\bar{y})^2$ – сумма квадратов отклонений, объясненная регрессией (или факторная сумма квадратов отклонений); $\sum\left(y-\hat{y}_x\right)^2$ – остаточная сумма квадратов отклонений, характеризующая влияние неучтенных в модели факторов.

Схема дисперсионного анализа имеет вид, представленный в таблице 1.1 (n – число наблюдений, m – число параметров при переменной x).

Таблица 2.1

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum\left(y-\bar{y}\right)^2$	$n-1$	$S^2_{\text{общ}}=\dfrac{\sum\left(y-\bar{y}\right)^2}{n-1}$
Факторная	$\sum\left(\hat{y}_x-\bar{y}\right)^2$	m	$S^2_{\text{факт}}=\dfrac{\sum\left(\hat{y}_x-\bar{y}\right)^2}{m}$
Остаточная	$\sum\left(y-\hat{y}_x\right)^2$	$n-m-1$	$S^2_{\text{ост}}=\dfrac{\sum\left(y-\hat{y}_x\right)^2}{n-m-1}$

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду (напомним, что степени свободы – это числа, показывающие количество элементов варьирования, которые могут принимать произвольные значения, не изменяющие заданных характеристик). Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину F-критерия Фишера:

$$F=\frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}.$$

Фактическое значение F-критерия Фишера сравнивается с табличным значением $F_{\text{табл}}\left(\alpha;k_1;k_2\right)$ при уровне значимости α и степенях свободы $k_1=m$ и $k_2=n-m-1$. При этом, если фактическое значение F-критерия больше табличного, то признается статистическая значимость уравнения в целом.

Для парной линейной регрессии $m=1$, поэтому

$$F=\frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}=\frac{\sum\left(\hat{y}_x-\bar{y}\right)^2}{\sum\left(y-\hat{y}_x\right)^2}\cdot\left(n-2\right).$$

Величина F-критерия связана с коэффициентом детерминации r_{xy}^2, и ее можно рассчитать по следующей формуле:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n-2). \tag{2.9}$$

Для оценки *статистической значимости параметров регрессии и корреляции* рассчитываются t-*критерий Стьюдента* и *доверительные интервалы* каждого из показателей. Оценка значимости коэффициентов регрессии и корреляции с помощью t-критерия Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_b = \frac{b}{m_b}; \quad t_a = \frac{a}{m_a}; \quad t_r = \frac{r_{xy}}{m_r}. \tag{2.10}$$

Стандартные ошибки параметров линейной регрессии и коэффициента корреляции определяются по формулам:

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{S_{\text{ост}}^2}{n \cdot \sigma_x^2}};$$

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{(n-2)} \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{S_{\text{ост}}^2 \frac{\sum x^2}{n^2 \sigma_x^2}}; \tag{2.11}$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}}.$$

Сравнивая фактическое и критическое (табличное) значения t-статистики – $t_{\text{табл}}$ и $t_{\text{факт}}$ – делаем вывод о значимости параметров регрессии и корреляции. Если $t_{\text{табл}} < t_{\text{факт}}$ то параметры a, b и r_{xy} не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x. Если $t_{\text{табл}} > t_{\text{факт}}$, то признается случайная природа формирования a, b или r_{xy}.

Для расчета доверительного интервала определяем *предельную ошибку* Δ для каждого показателя:

$$\Delta_a = t_{\text{табл}} m_a, \qquad \Delta_b = t_{\text{табл}} m_b.$$

Формулы для расчета *доверительных интервалов* имеют следующий вид:

$$\gamma_a = a \pm \Delta_a; \qquad \gamma_{a_{\min}} = a - \Delta_a; \quad \gamma_{a_{\max}} = a + \Delta_a;$$

$$\gamma_b = b \pm \Delta_b; \qquad \gamma_{b_{\min}} = b - \Delta_b; \quad \gamma_{b_{\max}} = b + \Delta_b;$$

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное, и отрицательное значения.

Связь между F-критерием Фишера и t-статистикой Стьюдента выражается равенством

$$|t_r| = |t_b| = \sqrt{F}\,. \qquad (2.12)$$

В прогнозных расчетах по уравнению регрессии определяется *предсказываемое индивидуальное значение* y_0 как точечный прогноз при $x = x_0$, т.е. путем подстановки в линейное уравнение $\hat{y}_x = a + b \cdot x$ соответствующего значения x. Однако точечный прогноз явно нереален, поэтому он дополняется расчетом стандартной ошибки

$$m_{\hat{y}_0} = \sqrt{S^2_{\text{ост}}\left(1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum (x - \overline{x})^2}\right)} = \sqrt{S^2_{\text{ост}}\left(1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{n \cdot \sigma_x^2}\right)}, \qquad (2.13)$$

где $S^2_{\text{ост}} = \dfrac{\sum (y - \hat{y}_x)^2}{n-2}$, и построением *доверительного интервала* прогнозного значения y_0^*:

$$\hat{y}_0 - m_{\hat{y}_0} \cdot t_{\text{табл}} \le y_0^* \le \hat{y}_x + m_{\hat{y}_x} \cdot t_{\text{табл}}$$

2.2. Решение типовой задачи

Пример. По территориям региона приводятся данные за 199X г.

Таблица 2.2

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	78	133
2	82	148
3	87	134
4	79	154
5	89	162
6	106	195
7	67	139
8	88	158
9	73	152
10	87	162
11	76	159
12	115	173

Требуется:

1. Построить линейное уравнение парной регрессии y по x.

2. Рассчитать линейный коэффициент парной корреляции, коэффициент детерминации и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F-критерия Фишера и t-критерия Стьюдента.

4. Выполнить прогноз заработной платы y при прогнозном значении среднедушевого прожиточного минимума x, составляющем 107% от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

6. На одном графике отложить исходные данные и теоретическую прямую.

Решение

1. Для расчета параметров уравнения линейной регрессии строим расчетную таблицу 2.3.

Таблица 2.3

№	x	y	$y \cdot x$	x^2	y^2	$\hat{y}_x$	$y-\hat{y}_x$	$(y-\hat{y}_x)^2$	A_i
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	78	133	10374	6084	17689	148,78	–15,78	249,01	11,86
2	82	148	12136	6724	21904	152,46	–4,46	19,89	3,01
3	87	134	11658	7569	17956	157,06	–23,06	531,76	17,21
4	79	154	12166	6241	23716	149,70	4,30	18,49	2,79
5	89	162	14418	7921	26244	158,90	3,10	9,61	1,91
6	106	195	20670	11236	38025	174,54	20,46	418,61	10,49
7	67	139	9313	4489	19321	138,66	0,34	0,12	0,24
8	88	158	13904	7744	24964	157,98	0,02	0,00	0,01
9	73	152	11096	5329	23104	144,18	7,82	61,15	5,14
10	87	162	14094	7569	26244	157,06	4,94	24,40	3,05
11	76	159	12084	5776	25281	146,94	12,06	145,44	7,58
12	115	173	19895	13225	29929	182,82	–9,82	96,43	5,68
Итого	1027	1869	161808	89907	294377	1869,08	–0,08	1574,91	68,97
Среднее значение	85,58	155,75	13484,0	7492,25	24531,4	155,76	–	131,24	5,75
σ	12,97	16,53	–	–	–	–	–		–
σ^2	168,31	273,34	–	–	–	–	–		–

По формулам (2.5) находим параметры регрессии

$$b = \frac{\overline{y \cdot x} - \overline{y} \cdot \overline{x}}{\overline{x^2} - \overline{x}^2} = \frac{13484 - 155,75 \cdot 85,58}{7492,25 - 85,58^2} = \frac{154,915}{168,31} = 0,92;$$

$$a = \overline{y} - b \cdot \overline{x} = 155,75 - 0,92 \cdot 85,58 = 77,02.$$

Получено уравнение регрессии:

$$y = 77,02 + 0,92 \cdot x.$$

Параметр регрессии позволяет сделать вывод, что с увеличением среднедушевого прожиточного минимума на 1 руб. среднедневная заработная плата возрастает в среднем на 0,92 руб. (или 92 коп.).

После нахождения уравнения регрессии заполняем столбцы 7–10 таблицы 2.3.

2. Тесноту линейной связи оценит коэффициент корреляции (2.6):

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 0,92 \cdot \frac{12,97}{16,53} = 0,722;$$

Т.к. значение коэффициента корреляции больше 0,7, то это говорит о наличии весьма тесной линейной связи между признаками.

Коэффициент детерминации:

$$r_{xy}^2 = 0,521.$$

Это означает, что 52% вариации заработной платы (y) объясняется вариацией фактора x – среднедушевого прожиточного минимума.

Качество модели определяет средняя ошибка аппроксимации (2,7):

$$\overline{A} = \frac{1}{n}\sum A_i = \frac{68,97}{12} = 5,75\%.$$

Качество построенной модели оценивается как хорошее, так как $\overline{A}$ не превышает 10%.

3. Оценку статистической значимости уравнения регрессии в целом проведем с помощью F-критерия Фишера. Фактическое значение F-критерия по формуле (2.9) составит

$$F_{\text{факт}} = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n-2) = \frac{0,521}{1 - 0,521} \cdot 10 = 10,88.$$

Табличное значение критерия при пятипроцентном уровне значимости и степенях свободы $k_1 = 1$ и $k_2 = 12 - 2 = 10$ составляет $F_{\text{табл}} = 4,96$. Так как $F_{\text{факт}} = 10,41 > F_{\text{табл}} = 4,96$, то уравнение регрессии признается статистически значимым.

Оценку статистической значимости параметров регрессии и корреляции проведем с помощью t-статистики Стьюдента и путем расчета доверительного интервала каждого из параметров.

Табличное значение t-критерия для числа степеней свободы $df = n - 2 = 12 - 2 = 10$ и уровня значимости $\alpha = 0,05$ составит $t_{\text{табл}} = 2,23$.

Определим стандартные ошибки m_a, m_b, $m_{r_{xy}}$ (остаточная дисперсия на одну степень свободы $S^2_{\text{ост}} = \frac{\sum\left(y-\hat{y}_x\right)^2}{n-2} = \frac{1574{,}91}{10} = 157{,}49$):

$$m_a = \sqrt{S^2_{\text{ост}} \frac{\sum x^2}{n^2 \sigma_x^2}} = \sqrt{157{,}49 \cdot \frac{89907}{12^2 \cdot 164{,}94}} = 24{,}42;$$

$$m_b = \sqrt{\frac{S^2_{\text{ост}}}{n \cdot \sigma_x^2}} = \sqrt{\frac{157{,}49}{12 \cdot 164{,}94}} = 0{,}282;$$

$$m_{r_{xy}} = \sqrt{\frac{1-r_{xy}^2}{n-2}} = \sqrt{\frac{1-0{,}521}{12-2}} = 0{,}219.$$

Тогда

$$t_a = \frac{a}{m_a} = \frac{77{,}02}{24{,}42} = 3{,}15;$$

$$t_b = \frac{b}{m_b} = \frac{0{,}92}{0{,}282} = 3{,}26;$$

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{0{,}722}{0{,}219} = 3{,}30.$$

Фактические значения t-статистики превосходят табличное значение:

$$t_a = 3{,}26 > t_{\text{табл}} = 2{,}3;\ t_b = 3{,}16 > t_{\text{табл}} = 2{,}3;\ t_{r_{xy}} = 3{,}25 > t_{\text{табл}} = 2{,}3,$$

поэтому параметры a, b и r_{xy} не случайно отличаются от нуля, а статистически значимы.

Рассчитаем доверительные интервалы для параметров регрессии a и b. Для этого определим предельную ошибку для каждого показателя:

$$\Delta_a = t_{\text{табл}} \cdot m_a = 2{,}23 \cdot 24{,}42 = 54{,}46;$$

$$\Delta_b = t_{\text{табл}} \cdot m_b = 2{,}23 \cdot 0{,}282 = 0{,}63.$$

Доверительные интервалы

$$\gamma_a = a \pm \Delta_a = 77{,}02 \pm 54{,}46 \text{ и } 22{,}56 \le a^* \le 131{,}48;$$

$$\gamma_b = b \pm \Delta_b = 0,92 \pm 0,63 \text{ и } 0,29 \le b^* \le 1,55$$

Анализ верхней и нижней границ доверительных интервалов приводит к выводу о том, что с вероятностью $p = 1 - \alpha = 0,95$ параметры a и b, находясь в указанных границах, не принимают нулевых значений, т.е. являются статистически значимыми и существенно отличны от нуля.

4. Полученные оценки уравнения регрессии позволяют использовать его для прогноза. Если прогнозное значение прожиточного минимума составит: $x_0 = \bar{x} \cdot 1,07 = 85,6 \cdot 1,07 = 91,6$ руб., тогда индивидуальное прогнозное значение заработной платы составит: $\hat{y}_0 = 77,02 + 0,92 \cdot 91,6 = 161,29$ руб.

5. Ошибка прогноза составит:

$$m_{\hat{y}_0} = \sqrt{S^2_{\text{ост}} \left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{n \cdot \sigma_x^2}\right)} = \sqrt{157,49 \cdot \left(1 + \frac{1}{12} + \frac{\left(91,6 - 85,6\right)^2}{12 \cdot 164,94}\right)} = 13,17.$$

Предельная ошибка прогноза, которая в 95% случаев не будет превышена, составит:

$$\Delta_{\hat{y}_0} = t_{\text{табл}} \cdot m_{\hat{y}_0} = 2,23 \cdot 13,17 = 29,37.$$

Доверительный интервал прогноза:

$$\gamma_{\hat{y}_0} = \hat{y}_0 \pm \Delta_{\hat{y}_0} = 161,29 \pm 29,37 \text{ и } 131,92 \le y_0^* \le 190,66.$$

Выполненный прогноз среднемесячной заработной платы является надежным ($p = 1 - \alpha = 1 - 0,05 = 0,95$) и находится в пределах от 131,92 руб. до 190,66 руб.

6. В заключение решения задачи построим на одном графике исходные данные и теоретическую прямую (рис. 2.1):

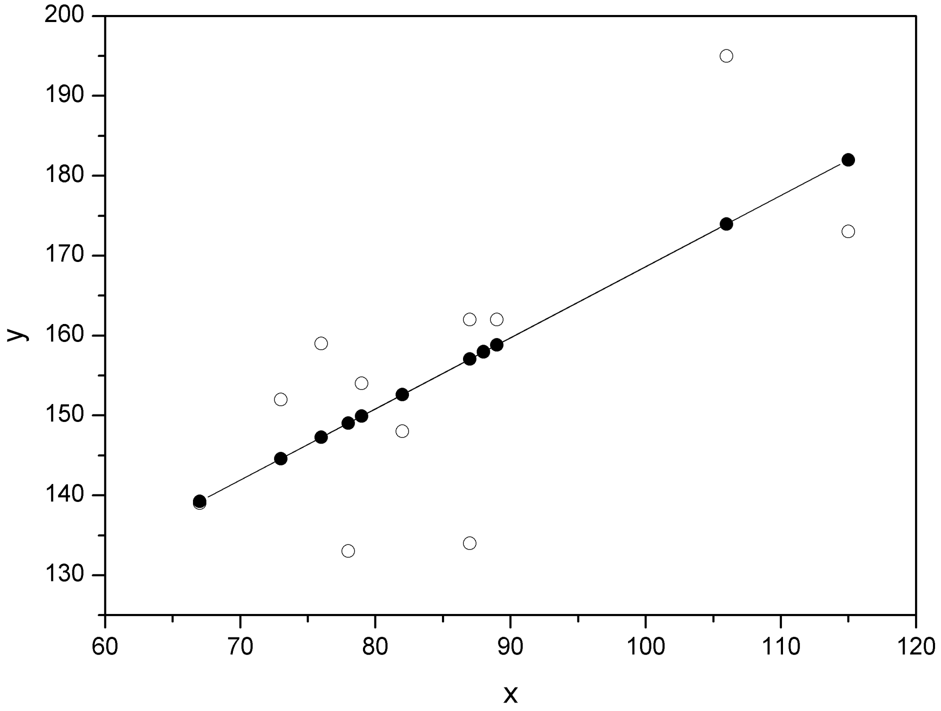


Рис. 2.1.

2.3. Решение типовой задачи в MS Excel

С помощью инструмента анализа данных **Регрессия** можно получить результаты регрессионной статистики, дисперсионного анализа, доверительных интервалов, остатки и графики подбора линии регрессии.

Если в меню сервис еще нет команды **Анализ данных**, то необходимо сделать следующее. В главном меню последовательно выбираем **Сервис→Надстройки** и устанавливаем «флажок» в строке **Пакет анализа** (рис. 2.2):

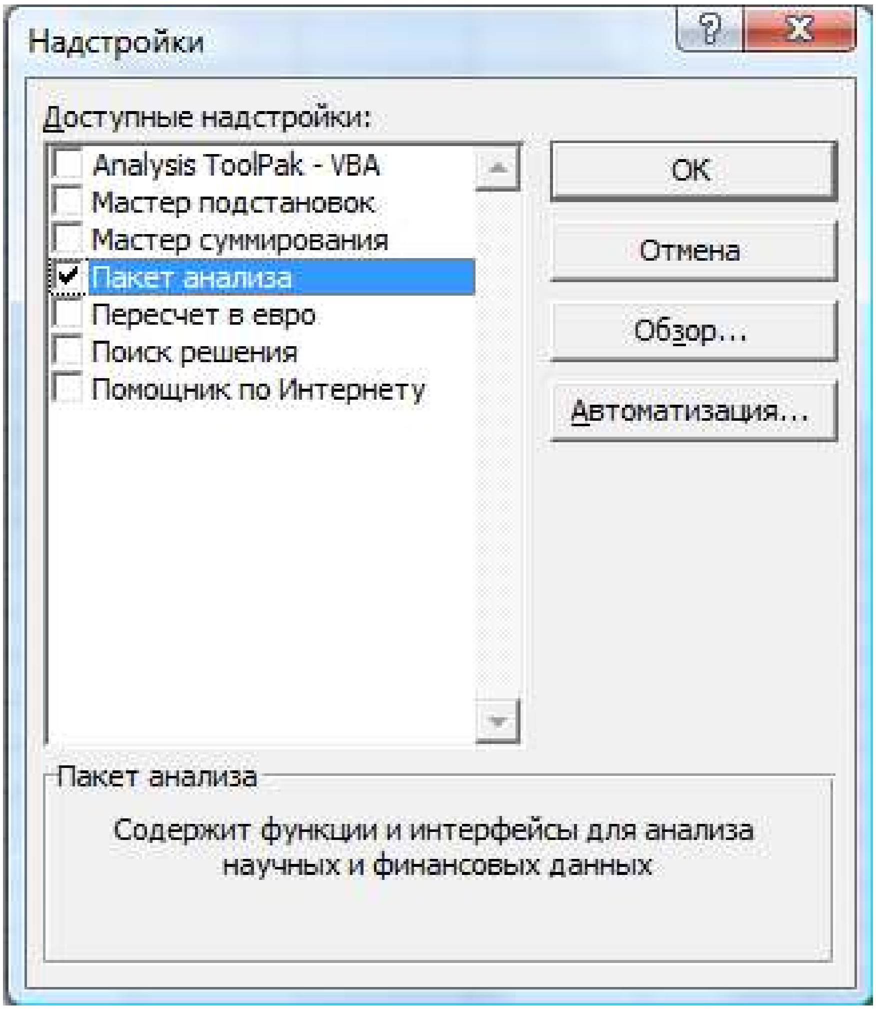


Рис. 2.2

Далее следуем по следующему плану.

1. Если исходные данные уже внесены, то выбираем **Сервис→Анализ данных→Регрессия**.

2. Заполняем диалоговое окно ввода данных и параметров вывода (рис. 2.3):

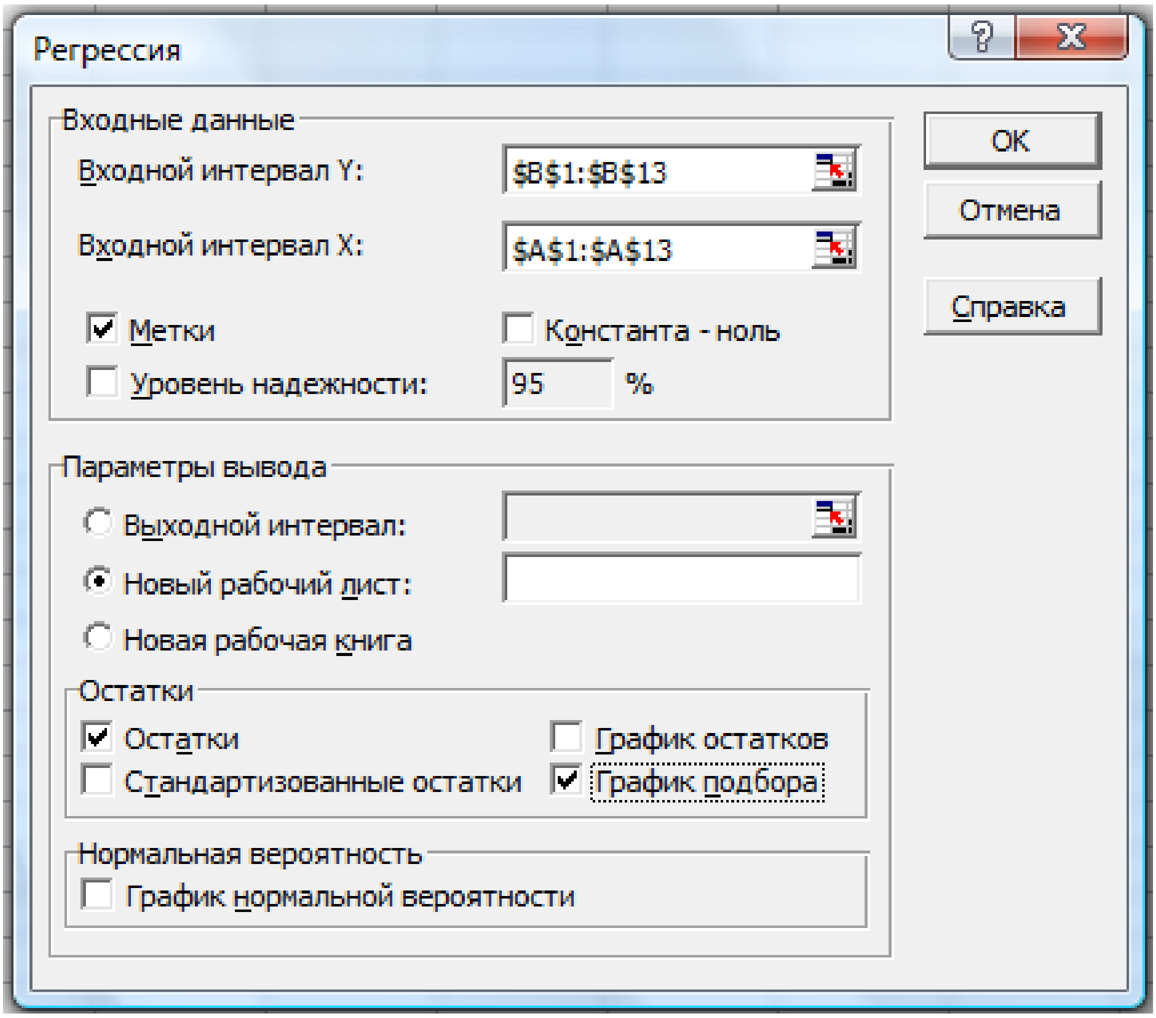


Рис. 2.3

Здесь:

Входной интервал Y – диапазон, содержащий данные результативного признака;

Входной интервал X – диапазон, содержащий данные признака-фактора;

Метки – «флажок», который указывает, содержи ли первая строка названия столбцов;

Константа – ноль – «флажок», указывающий на наличие или отсутствие свободного члена в уравнении;

Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона;

Новый рабочий лист – можно указать произвольное имя нового листа (или не указывать, тогда результаты выводятся на вновь созданный лист).

Получаем следующие результаты для рассмотренного выше примера:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ВЫВОД ИТОГОВ												
2													
3	*Регрессионная статистика*												
4	Множественный R	0,721025214											
5	R-квадрат	0,519877359											
6	Нормированный R-кв	0,471865095											
7	Стандартная ошибка	12,5495908											
8	Наблюдения	12											
9													
10	Дисперсионный анализ												
11		*df*	*SS*	*MS*	*F*	*Значимость F*							
12	Регрессия	1	1705,327706	1705,328	10,82801	0,008142							
13	Остаток	10	1574,922294	157,4922									
14	Итого	11	3280,25										
15													
16		*Коэффициенты*	*Стандартная*	*t-статисти*	*P-Значени*	*Нижние 95*	*Верхние 9*	*Нижние 95*	*Верхние 95,0%*				
17	Y-пересечение	76,9764852	24,21156138	3,179328	0,009831	23,02976	130,9232	23,02976	130,9232				
18	x	0,920430553	0,279715587	3,290594	0,008142	0,297185	1,543676	0,297185	1,543676				
19													
20													
21													
22	ВЫВОД ОСТАТКА												
23													
24	*Наблюдение*	*Предсказанное y*	*Остатки*										
25	1	148,7700683	-15,77006831										
26	2	152,4517905	-4,45179052										
27	3	157,0539433	-23,05394328										
28	4	149,6904989	4,309501138										
29	5	158,8948044	3,105195612										
30	6	174,5421238	20,45787622										
31	7	138,6453322	0,354667771										
32	8	157,9743738	0,025626164										
33	9	144,1679155	7,832084455										
34	10	157,0539433	4,946056717										
35	11	146,9292072	12,0707928										
36	12	182,8259988	-9,825998758										

Рис. 2.4

Откуда выписываем, округляя до 4 знаков после запятой и переходя к нашим обозначениям:

Уравнение регрессии:

$$\hat{y}_x = 76{,}9765 + 0{,}9204x\,.$$

Коэффициент корреляции:

$$r_{xy} = 0{,}7210\,.$$

Коэффициент детерминации:

$$r_{xy}^2 = 0{,}5199\,.$$

Фактическое значение F -критерия Фишера:

$$F = 10,8280$$

Остаточная дисперсия на одну степень свободы:

$$S^2_{\text{ост}} = 157,4922.$$

Корень квадратный из остаточной дисперсии (стандартная ошибка):

$$S_{\text{ост}} = 12,5496.$$

Стандартные ошибки для параметров регрессии:

$$m_a = 24,2116, \qquad m_b = 0,2797.$$

Фактические значения t -критерия Стьюдента:

$$t_a = 3,1793, \qquad t_b = 3,2906.$$

Доверительные интервалы:

$$23,0298 \le a^* \le 130,9232,$$

$$0,2972 \le b^* \le 1,5437.$$

Как видим, найдены все рассмотренные выше параметры и характеристики уравнения регрессии, за исключением средней ошибки аппроксимации (значение t-критерия Стьюдента для коэффициента корреляции совпадает с t_b). Результаты «ручного счета» от машинного отличаются незначительно (отличия связаны с ошибками округления).

3. Множественная регрессия и корреляция

3.1.Теоретическая справка

Множественная регрессия – это уравнение связи с несколькими независимыми переменными:

$$y = f\left(x_1, x_2, ..., x_m\right) + \varepsilon,$$

где y – зависимая переменная (результативный признак); $x_1, x_2, ..., x_m$ – независимые переменные (признаки-факторы).

Для построения уравнения множественной регрессии чаще используются следующие функции:

• линейная – $y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_m \cdot x_m + \varepsilon$;

• степенная – $y = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot ... \cdot x_m^{b_m} \cdot \varepsilon$;

• экспонента – $y = \mathrm{e}^{a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_m \cdot x_m + \varepsilon}$;

• гипербола – $y = \dfrac{1}{a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_m \cdot x_m + \varepsilon}$

Можно использовать и другие функции, приводимые к линейному виду.

Для оценки параметров уравнения множественной регрессий применяют *метод наименьших квадратов* (МНК). Для линейных уравнений

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_m \cdot x_m + \varepsilon \tag{3.1}$$

строится следующая система нормальных уравнений, решение которой позволяет получить оценки параметров регрессии:

$$\begin{cases} \sum y = na + b_1 \sum x_1 + b_2 \sum x_2 + ... + b_m \sum x_m, \\ \sum yx_1 = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + ... + b_m \sum x_m x_1, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \sum yx_m = a \sum x_m + b_1 \sum x_1 x_m + b_2 \sum x_2 x_m + b_m \sum x_m^2. \end{cases} \tag{3.2}$$

Для двухфакторной модели данная система будет иметь вид:

$$\begin{cases} na + b_1\sum x_1 + b_2\sum x_2 = \sum y, \\ a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 = \sum yx_1, \\ a\sum x_2 + b_1\sum x_1x_2 + b_2\sum x_2^2 = \sum yx_2. \end{cases} \tag{3.3}$$

Так же можно воспользоваться готовыми формулами, которые являются следствием из этой системы:

$$b_1 = \frac{\sigma_y}{\sigma_{x_1}} \cdot \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r_{x_1x_2}^2};$$

$$b_2 = \frac{\sigma_y}{\sigma_{x_2}} \cdot \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r_{x_1x_2}^2}; \tag{3.4}$$

$$a = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 .$$

В линейной множественной регрессии параметры при x называются *коэффициентами «чистой» регрессии*. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизмененном значении других факторов, закрепленных на среднем уровне.

Метод наименьших квадратов применим и к уравнению множественной регрессии в *стандартизированном масштабе*:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + ... + \beta_m t_{x_m} + \varepsilon, \tag{3.5}$$

где t_y, t_{x_1}, ..., t_{x_m} – *стандартизированные переменные*: $t_y = \dfrac{y - \overline{y}}{\sigma_y}$, $t_{x_i} = \dfrac{x_i - \overline{x}_i}{\sigma_{x_i}}$, для которых среднее значение равно нулю: $\overline{t}_y = \overline{t}_{x_i} = 0$, а среднее квадратическое отклонение равно единице: $\sigma_{t_y} = \sigma_{t_{x_i}} = 1$; β_i – *стандартизированные коэффициенты регрессии*.

В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии β_i можно сравнивать между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных коэффициентов регрессии в отличие от коэффициентов «чистой» регрессии, которые несравнимы между собой.

Применяя МНК к уравнению множественной регрессии в стандартизированном масштабе, получим систему нормальных уравнений вида

$$\begin{cases} r_{yx_1} = \beta_1 \qquad + \beta_2 r_{x_1x_2} + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ r_{yx_2} = \beta_1 r_{x_1x_2} + \beta_2 \qquad + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{yx_m} = \beta_1 r_{x_1x_m} + \beta_2 r_{x_2x_m} + \beta_3 r_{x_3x_m} + \ldots + \beta_m, \end{cases} \tag{3.6}$$

где r_{yx_i} и $r_{x_ix_j}$ – коэффициенты парной и межфакторной корреляции.

Коэффициенты «чистой» регрессии b_i связаны со стандартизованными коэффициентами регрессии β_i следующим образом:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}} \ \left(\beta_i = b_i \frac{\sigma_{x_i}}{\sigma_y} \right). \tag{3.7}$$

Поэтому можно переходить от уравнения регрессии в стандартизованном масштабе (3.5) к уравнению регрессии в натуральном масштабе переменных (3.1), при этом параметр a определяется как $a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_m\bar{x}_m$.

Рассмотренный смысл стандартизованных коэффициентов регрессии позволяет их использовать при отсеве факторов – из модели исключаются факторы с наименьшим значением β_i.

Средние коэффициенты эластичности для линейной регрессии рассчитываются по формуле

$$\bar{Э}_{yx_j} = b_j \frac{\bar{x}_j}{\bar{y}}, \tag{3.8}$$

которые показывают на сколько процентов в среднем изменится результат, при изменении соответствующего фактора на 1%. Средние показатели эластичности можно сравнивать друг с другом и соответственно ранжировать факторы по силе их воздействия на результат.

Тесноту совместного влияния факторов на результат оценивает *индекс множественной корреляции:*

$$R_{yx_1x_2\dots x_m} = \sqrt{1 - \frac{\sigma^2_{y_{\text{ост}}}}{\sigma^2_y}}\,. \tag{3.9}$$

Значение индекса множественной корреляции лежит в пределах от 0 до 1 и должно быть больше или равно максимальному парному индексу корреляции:

$$R_{yx_1x_2\dots x_m} \ge r_{yx_i} \quad \left(i = \overline{1, m}\right).$$

При линейной зависимости *коэффициент множественной корреляции* можно определить через матрицы парных коэффициентов корреляции:

$$R_{yx_1x_2\dots x_m} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}\,, \tag{3.10}$$

где

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_m} \\ r_{yx_1} & 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{yx_2} & r_{x_2x_1} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_m} & r_{x_mx_1} & r_{x_mx_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_mx_1} & r_{x_mx_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы межфакторной корреляции.

Так же при линейной зависимости признаков формула коэффициента множественной корреляции может быть также представлена следующим выражением:

$$R_{yx_1x_2...x_m} = \sqrt{\sum \beta_i \cdot r_{yx_i}}\,, \tag{3.11}$$

где β_i – стандартизованные коэффициенты регрессии; r_{yx_i} – парные коэффициенты корреляции результата с каждым фактором.

Качество построенной модели в целом оценивает коэффициент (индекс) детерминации. *Коэффициент множественной детерминации* рассчитывается как квадрат индекса множественной корреляции $R^2_{yx_1x_2...x_m}$.

Для того чтобы не допустить преувеличения тесноты связи, применяется *скорректированный индекс множественной детерминации*, который содержит поправку на число степеней свободы и рассчитывается по формуле

$$\hat{R}^2 = 1 - \left(1 - R^2\right)\frac{(n-1)}{(n-m-1)}\,, \tag{3.12}$$

где n – число наблюдений, m – число факторов. При небольшом числе наблюдений нескорректированная величина коэффициента множественной детерминации R^2 имеет тенденцию переоценивать долю вариации результативного признака, связанную с влиянием факторов, включенных в регрессионную модель.

Частные коэффициенты (или индексы) корреляции, измеряющие влияние на y фактора x_i, при элиминировании (исключении влияния) других факторов, можно определить по формуле

$$r_{yx_i \cdot x_1x_2...x_{i-1}x_{i+1}...x_m} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2...x_i...x_m}}{1 - R^2_{yx_1x_2...x_{i-1}x_{i+1}...x_m}}}\,, \tag{3.13}$$

или по рекуррентной формуле:

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m} = \frac{r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}} - r_{yx_m \cdot x_1 x_2 \ldots x_{m-1}} \cdot r_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}}{\sqrt{\left(1 - r^2_{yx_m \cdot x_1 x_m \ldots x_{m-1}}\right)\left(1 - r^2_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}\right)}} \quad (3.14)$$

Рассчитанные по рекуррентной формуле частные коэффициенты корреляции изменяются в пределах от –1 до +1, а по формулам через множественные коэффициенты детерминации – от 0 до 1. Сравнение их друг с другом позволяет ранжировать факторы по тесноте их связи с результатом. Частные коэффициенты корреляции дают меру тесноты связи каждого фактора с результатом в чистом виде.

При двух факторах формулы (3.12) и (3.13) примут вид:

$$r_{yx_1 \cdot x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2}}{1 - r^2_{yx_2}}}; \; r_{yx_2 \cdot x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2}}{1 - r^2_{yx_1}}}.$$

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right) \cdot \left(1 - r^2_{x_1 x_2}\right)}}; \; r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{x_1 x_2}\right)}}.$$

Значимость уравнения множественной регрессии в целом оценивается с помощью F-критерия Фишера:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}. \quad (3.15)$$

Частный F-критерий оценивает статистическую значимость присутствия каждого из факторов в уравнении. В общем виде для фактора x частный F-критерий определится как

$$F_{x_i} = \frac{R^2_{yx_1 \ldots x_i \ldots x_m} - R^2_{yx_1 \ldots x_{i-1} x_{i+1} \ldots x_m}}{1 - R^2_{yx_1 \ldots x_i \ldots x_m}} \cdot \frac{n - m - 1}{1} \quad (3.16)$$

Фактическое значение частного F-критерия сравнивается с табличным при уровне значимости α и числе степеней свободы: $k_1 = 1$ и $k_2 = n - m - 1$. Если фактическое значение F_{x_i} превышает $F_{\text{табл}}(\alpha,\ k_1,\ k_2)$, то дополнительное включение фактора x_i в модель статистически оправданно и коэффициент чистой регрессии b_i при факторе x_i статистически значим.

Если же фактическое значение F_{x_i} меньше табличного, то дополнительное включение в модель фактора x_i не увеличивает существенно долю объясненной вариации признака y, следовательно, нецелесообразно его включение в модель; коэффициент регрессии при данном факторе в этом случае статистически незначим.

Оценка значимости коэффициентов чистой регрессии проводится по t-критерию Стьюдента. В этом случае, как и в парной регрессии, для каждого фактора используется формула

$$t_{b_i} = \frac{b_i}{m_{b_i}} \tag{3.17}$$

Для уравнения множественной регрессии (3.1) средняя квадратическая ошибка коэффициента регрессии может быть определена по формуле:

$$m_{b_i} = \frac{\sigma_y \cdot \sqrt{1 - R^2_{yx_1 \dots x_m}}}{\sigma_{x_i} \cdot \sqrt{1 - R^2_{x_i x_1 \dots x_m}}} \cdot \frac{1}{\sqrt{n - m - 1}}, \tag{3.18}$$

где $R^2_{x_i x_1 \dots x_m}$ – коэффициент детерминации для зависимости фактора x_i со всеми другими факторами уравнения множественной регрессии. Для двухфакторной модели ($m = 2$) имеем:

$$m_{b_1} = \frac{\sigma_y \cdot \sqrt{1 - R^2_{yx_1x_2}}}{\sigma_{x_1} \cdot \sqrt{1 - r^2_{x_1x_2}}} \cdot \frac{1}{\sqrt{n - 3}}; \tag{3.19}$$

$$m_{b_2} = \frac{\sigma_y \cdot \sqrt{1 - R^2_{yx_1x_2}}}{\sigma_{x_2} \cdot \sqrt{1 - r^2_{x_1x_2}}} \cdot \frac{1}{\sqrt{n - 3}}. \tag{3.20}$$

Существует связь между t-критерием Стьюдента и частным F-критерием Фишера:

$$\left| t_{b_i} \right| = \sqrt{F_{x_i}}. \tag{3.21}$$

Уравнения множественной регрессии могут включать в качестве независимых переменных качественные признаки (например, профессия, пол, образование, климатические условия, отдельные регионы и т.д.). Чтобы ввести такие переменные в регрессионную модель, их необходимо упорядочить и присвоить им те или иные значения, т.е. качественные переменные преобразовать в количественные.

Такого вида сконструированные переменные принято в эконометрике называть *фиктивными переменными*. Например, включать в модель фактор «пол» в виде фиктивной переменной можно в следующем виде:

$$z = \begin{cases} 1 \; - \; \text{мужской пол,} \\ 0 \; - \; \text{женский пол.} \end{cases} \tag{3.22}$$

Коэффициент регрессии при фиктивной переменной интерпретируется как среднее изменение зависимой переменной при переходе от одной категории (женский пол) к другой (мужской пол) при неизменных значениях остальных параметров.

3.2. Решение типовой задачи

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%).

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7,0	3,9	10,0	11	9,0	6,0	21,0
2	7,0	3,9	14,0	12	11,0	6,4	22,0
3	7,0	3,7	15,0	13	9,0	6,8	22,0
4	7,0	4,0	16,0	14	11,0	7,2	25,0
5	7,0	3,8	17,0	15	12,0	8,0	28,0
6	7,0	4,8	19,0	16	12,0	8,2	29,0
7	8,0	5,4	19,0	17	12,0	8,1	30,0
8	8,0	4,4	20,0	18	12,0	8,5	31,0
9	8,0	5,3	20,0	19	14,0	9,6	32,0
10	10,0	6,8	20,0	20	14,0	9,0	36,0

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью t-критерия оценить статистическую значимость коэффициентов чистой регрессии.

6. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

7. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Решение

Для удобства проведения расчетов поместим результаты промежуточных расчетов в таблицу:

Таблица 3.1

№	y	x_1	x_2	yx_1	yx_2	x_1x_2	x_1^2	x_2^2	y^2
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	7,0	3,9	10,0	27,3	70,0	39,0	15,21	100,0	49,0
2	7,0	3,9	14,0	27,3	98,0	54,6	15,21	196,0	49,0
3	7,0	3,7	15,0	25,9	105,0	55,5	13,69	225,0	49,0
4	7,0	4,0	16,0	28,0	112,0	64,0	16,0	256,0	49,0
5	7,0	3,8	17,0	26,6	119,0	64,6	14,44	289,0	49,0
6	7,0	4,8	19,0	33,6	133,0	91,2	23,04	361,0	49,0
7	8,0	5,4	19,0	43,2	152,0	102,6	29,16	361,0	64,0
8	8,0	4,4	20,0	35,2	160,0	88,0	19,36	400,0	64,0
9	8,0	5,3	20,0	42,4	160,0	106,0	28,09	400,0	64,0
10	10,0	6,8	20,0	68,0	200,0	136,0	46,24	400,0	100,0
11	9,0	6,0	21,0	54,0	189,0	126,0	36,0	441,0	81,0
12	11,0	6,4	22,0	70,4	242,0	140,8	40,96	484,0	121,0
13	9,0	6,8	22,0	61,2	198,0	149,6	46,24	484,0	81,0
14	11,0	7,2	25,0	79,2	275,0	180,0	51,84	625,0	121,0
15	12,0	8,0	28,0	96,0	336,0	224,0	64,0	784,0	144,0
16	12,0	8,2	29,0	98,4	348,0	237,8	67,24	841,0	144,0
17	12,0	8,1	30,0	97,2	360,0	243,0	65,61	900,0	144,0
18	12,0	8,5	31,0	102,0	372,0	263,5	72,25	961,0	144,0
19	14,0	9,6	32,0	134,4	448,0	307,2	92,16	1024,0	196,0
20	14,0	9,0	36,0	126,0	504,0	324,0	81,0	1296,0	196,0
Сумма	192	123,8	446	1276,3	4581	2997,4	837,74	10828,0	1958,0
Ср. знач.	9,6	6,19	22,3	63,815	229,05	149,87	41,887	541,4	97,9

Найдем средние квадратические отклонения признаков:

$$\sigma_y = \sqrt{\overline{y^2} - \bar{y}^2} = \sqrt{97,9 - 9,6^2} = \sqrt{5,74} = 2,396;$$

$$\sigma_{x_1} = \sqrt{\overline{x_1^2} - \overline{x}_1^2} = \sqrt{41,887 - 6,19^2} = \sqrt{3,571} = 1,890;$$

$$\sigma_{x_2} = \sqrt{\overline{x_2^2} - \overline{x}_2^2} = \sqrt{541,4 - 22,3^2} = \sqrt{44,11} = 6,642.$$

1. Для нахождения параметров линейного уравнения множественной регрессии

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

необходимо решить систему линейных уравнений относительно неизвестных параметров a, b_1, b_2 (3.3) либо воспользоваться готовыми формулами (3.4).

Рассчитаем сначала парные коэффициенты корреляции:

$$r_{yx_1} = \frac{\operatorname{cov}(y,\, x_1)}{\sigma_y \cdot \sigma_{x_1}} = \frac{63,815 - 6,19 \cdot 9,6}{1,890 \cdot 2,396} = 0,970;$$

$$r_{yx_2} = \frac{\operatorname{cov}(y,\, x_2)}{\sigma_y \cdot \sigma_{x_2}} = \frac{229,05 - 22,3 \cdot 9,6}{6,642 \cdot 2,396} = 0,941;$$

$$r_{x_1x_2} = \frac{\operatorname{cov}(x_1,\, x_2)}{\sigma_{x_1} \cdot \sigma_{x_2}} = \frac{149,87 - 6,19 \cdot 22,3}{1,890 \cdot 6,642} = 0,943.$$

Находим по формулам (3.4) коэффициенты чистой регрессии и параметр a:

$$b_1 = \frac{\sigma_y}{\sigma_{x_1}} \cdot \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r_{x_1x_2}^2} = \frac{2,396}{1,890} \cdot \frac{0,970 - 0,941 \cdot 0,943}{1 - 0,943^2} = 0,946;$$

$$b_2 = \frac{\sigma_y}{\sigma_{x_2}} \cdot \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r_{x_1x_2}^2} = \frac{2,396}{6,642} \cdot \frac{0,941 - 0,970 \cdot 0,943}{1 - 0,943^2} = 0,0856;$$

$$a = \overline{y} - b_1 \overline{x}_1 - b_2 \overline{x}_2 = 9,6 - 0,946 \cdot 6,19 - 0,0856 \cdot 22,3 = 1,835.$$

Таким образом, получили следующее уравнение множественной регрессии:

$$\hat{y} = 1,835 + 0,946 \cdot x_1 + 0,0856 \cdot x_2.$$

Уравнение регрессии показывает, что при увеличении ввода в действие основных фондов на 1% (при неизменном уровне удельного веса рабочих высокой квалификации) выработка продукции на одного рабочего увеличивается в среднем на 0,946 тыс. руб., а при увеличении удельного веса рабочих высокой квалификации в общей численности рабочих на 1% (при неизменном уровне ввода в действие новых основных фондов) выработка продукции на одного рабочего увеличивается в среднем на 0,086 тыс. руб.

После нахождения уравнения регрессии составим новую расчетную таблицу для определения теоретических значений результативного признака, остаточной дисперсии и средней ошибки аппроксимации.

Таблица 3.2

№	y	x_1	x_2	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	A_i, %
1	7,0	3,9	10,0	6,380	0,620	0,384	8,851
2	7,0	3,9	14,0	6,723	0,277	0,077	3,960
3	7,0	3,7	15,0	6,619	0,381	0,145	5,440
4	7,0	4,0	16,0	6,989	0,011	0,000	0,163
5	7,0	3,8	17,0	6,885	0,115	0,013	1,643
6	7,0	4,8	19,0	8,002	-1,002	1,004	14,317
7	8,0	5,4	19,0	8,570	-0,570	0,325	7,123
8	8,0	4,4	20,0	7,709	0,291	0,084	3,633
9	8,0	5,3	20,0	8,561	-0,561	0,315	7,010
10	10,0	6,8	20,0	9,980	0,020	0,000	0,202
11	9,0	6,0	21,0	9,309	-0,309	0,095	3,429
12	11,0	6,4	22,0	9,773	1,227	1,507	11,158
13	9,0	6,8	22,0	10,151	-1,151	1,325	12,789
14	11,0	7,2	25,0	10,786	0,214	0,046	1,944
15	12,0	8,0	28,0	11,800	0,200	0,040	1,668
16	12,0	8,2	29,0	12,075	-0,075	0,006	0,622
17	12,0	8,1	30,0	12,066	-0,066	0,004	0,547
18	12,0	8,5	31,0	12,530	-0,530	0,280	4,413
19	14,0	9,6	32,0	13,656	0,344	0,118	2,459
20	14,0	9,0	36,0	13,431	0,569	0,324	4,067
Сумма	192	123,8	446	191,992	0,008	6,093	95,437
Ср. знач.	9,6	6,19	22,3	9,6	–	0,305	4,77

Остаточная дисперсия:

$$\sigma^2_{\text{ост}} = \frac{\sum(y-\hat{y})^2}{n} = \frac{6,093}{20} = 0,305.$$

Средняя ошибка аппроксимации:

$$\bar{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \cdot 100\% = \frac{95,437\%}{20} = 4,77\%.$$

Качество модели, исходя из относительных отклонений по каждому наблюдению, признается хорошим, т.к. средняя ошибка аппроксимации не превышает 10%.

Коэффициенты β_1 и β_2 стандартизованного уравнения регрессии $\hat{t}_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \varepsilon$, находятся по формуле (3.7):

$$\beta_1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} = 0,946 \cdot \frac{1,890}{2,396} = 0,746;$$

$$\beta_2 = b_2 \frac{\sigma_{x_2}}{\sigma_y} = 0,0856 \cdot \frac{6,642}{2,396} = 0,237.$$

Т.е. уравнение будет выглядеть следующим образом:

$$\hat{t}_y = 0,746 \cdot t_{x_1} + 0,237 \cdot t_{x_2}.$$

Так как стандартизованные коэффициенты регрессии можно сравнивать между собой, то можно сказать, что ввод в действие новых основных фондов оказывает большее влияние на выработку продукции, чем удельный вес рабочих высокой квалификации.

Сравнивать влияние факторов на результат можно также при помощи средних коэффициентов эластичности (3.8):

$$\bar{Э}_i = b_i \cdot \frac{\bar{x}_i}{\bar{y}_{x_i}}.$$

Вычисляем:

$$\bar{Э}_1 = 0,946 \cdot \frac{6,19}{9,6} = 0,61; \ \bar{Э}_2 = 0,0856 \cdot \frac{22,3}{9,6} = 0,20.$$

Т.е. увеличение только основных фондов (от своего среднего значения) или только удельного веса рабочих высокой квалификации на 1% увеличивает в среднем выработку продукции на 0,61% или 0,20% соответственно. Таким образом, подтверждается большее влияние на результат y фактора x_1, чем фактора x_2.

2. Коэффициенты парной корреляции мы уже нашли:

$$r_{yx_1} = 0{,}970;\ r_{yx_2} = 0{,}941;\ r_{x_1x_2} = 0{,}943.$$

Они указывают на весьма сильную связь каждого фактора с результатом, а также высокую межфакторную зависимость (факторы x_1 и x_2 явно коллинеарны, т.к. $r_{x_1x_2} = 0{,}943 > 0{,}7$). При такой сильной межфакторной зависимости рекомендуется один из факторов исключить из рассмотрения.

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при элиминировании (устранении влияния) других факторов, включенных в уравнение регрессии.

При двух факторах частные коэффициенты корреляции рассчитываются следующим образом:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right) \cdot \left(1 - r^2_{x_1x_2}\right)}} = \frac{0{,}970 - 0{,}941 \cdot 0{,}943}{\sqrt{\left(1 - 0{,}941^2\right) \cdot \left(1 - 0{,}943^2\right)}} = 0{,}734;$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{x_1x_2}\right)}} = \frac{0{,}941 - 0{,}970 \cdot 0{,}943}{\sqrt{\left(1 - 0{,}970^2\right) \cdot \left(1 - 0{,}943^2\right)}} = 0{,}325.$$

Если сравнить коэффициенты парной и частной корреляции, то можно увидеть, что из-за высокой межфакторной зависимости коэффициенты парной корреляции дают завышенные оценки тесноты связи. Именно по этой причине рекомендуется при наличии сильной коллинеарности (взаимосвязи) факторов исключать из исследования тот фактор, у которого теснота парной зависимости меньше, чем теснота межфакторной связи.

Коэффициент множественной корреляции определить через матрицы парных коэффициентов корреляции (3.9):

$$R_{yx_1x_2} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}},$$

где

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{yx_1} & 1 & r_{x_1x_2} \\ r_{yx_2} & r_{x_2x_1} & 1 \end{vmatrix}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} \\ r_{x_2x_1} & 1 \end{vmatrix}$$

– определитель матрицы межфакторной корреляции.

Находим:

$$\Delta r = \begin{vmatrix} 1 & 0,970 & 0,941 \\ 0,970 & 1 & 0,943 \\ 0,941 & 0,943 & 1 \end{vmatrix} = 1 + 0,8607 + 0,8607 -$$

$$-0,8855 - 0,8892 - 0,9409 = 0,0058;$$

$$\Delta r_{11} = \begin{vmatrix} 1 & 0,943 \\ 0,943 & 1 \end{vmatrix} = 1 - 0,8892 = 0,1108.$$

Коэффициент множественной корреляции:

$$R_{yx_1x_2} = \sqrt{1 - \frac{0,0058}{0,1108}} = 0,973.$$

Аналогичный результат получим при использовании формул (3.8) и (3.10):

$$R_{yx_1x_2} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1 - \frac{0,305}{5,74}} = 0,973;$$

$$R_{yx_1x_2} = \sqrt{\sum \beta_i \cdot r_{yx_i}} = \sqrt{0,746 \cdot 0,970 + 0,237 \cdot 0,941} = 0,973;$$

Коэффициент множественной корреляции указывает на весьма сильную связь всего набора факторов с результатом.

3. Нескорректированный коэффициент множественной детерминации $R^2_{yx_1x_2} = 0,947$ оценивает долю дисперсии результата за счет представленных в уравнении факторов в общей вариации результата. Здесь эта доля составляет $94,7\%$ и указывает на весьма высокую степень обусловленности вариации результата вариацией факторов, иными словами – на весьма тесную связь факторов с результатом.

Скорректированный коэффициент множественной детерминации

$$\hat{R}^2 = 1 - \left(1 - R^2\right)\frac{(n-1)}{(n-m-1)} = 1 - \left(1 - 0,947\right)\frac{20-1}{20-2-1} = 0,941$$

определяет тесноту связи с учетом степеней свободы общей и остаточной дисперсий. Он дает такую оценку тесноты связи, которая не зависит от числа факторов и поэтому может сравниваться по разным моделям с разным числом факторов. Оба коэффициента указывают на весьма высокую (более 94%) детерминированность результата y в модели факторами x_1 и x_2.

4. Оценку надежности уравнения регрессии в целом и показателя тесноты связи $R_{yx_1x_2}$ дает F-критерий Фишера:

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

В нашем случае фактическое значение F-критерия Фишера:

$$F_{\text{факт}} = \frac{0,973^2}{1-0,973^2} \cdot \frac{20-2-1}{2} = 151,88.$$

Получили, что $F_{\text{факт}} = 151,88 > F_{\text{табл}} = 3,59$ (при $n = 20$), т.е. вероятность случайно получить такое значение F-критерия не превышает допустимый уровень значимости 5%. Следовательно, полученное значение не случайно, оно сформировалось под влиянием существенных факторов, т.е. подтверждается статистическая значимость всего уравнения и показателя тесноты связи $R_{yx_1x_2}$.

5. Оценим статистическую значимость параметров чистой регрессии с помощью t-критерия Стьюдента. Рассчитаем стандартные ошибки коэффициентов регрессии по формулам (3.19) и (3.20):

$$m_{b_1}=\frac{\sigma_y\cdot\sqrt{1-R^2_{yx_1x_2}}}{\sigma_{x_1}\cdot\sqrt{1-r^2_{x_1x_2}}}\cdot\frac{1}{\sqrt{n-3}}=\frac{2,396\cdot\sqrt{1-0,973^2}}{1,890\cdot\sqrt{1-0,943^2}}\cdot\frac{1}{\sqrt{20-3}}=0,2132;$$

$$m_{b_2}=\frac{\sigma_y\cdot\sqrt{1-R^2_{yx_1x_2}}}{\sigma_{x_2}\cdot\sqrt{1-r^2_{x_1x_2}}}\cdot\frac{1}{\sqrt{n-3}}=\frac{2,396\cdot\sqrt{1-0,973^2}}{6,642\cdot\sqrt{1-0,943^2}}\cdot\frac{1}{\sqrt{20-3}}=0,0607.$$

Фактические значения t-критерия Стьюдента:

$$t_{b_1}=\frac{b_1}{m_{b_1}}=\frac{0,946}{0,2132}=4,44,\ t_{b_2}=\frac{b_2}{m_{b_2}}=\frac{0,0856}{0,0607}=1,41.$$

Табличное значение критерия при уровне значимости $\alpha=0,05$ и числе степеней свободы $k=17$ составит $t_{\text{табл}}(\alpha=0,05;\quad k=17)=2,11$. Таким образом, признается статистическая значимость параметра b_1, т.к. $t_{b_1}>t_{\text{табл}}$, и случайная природа формирования параметра b_2, т.к. $t_{b_2}<t_{\text{табл}}$.

Доверительные интервалы для параметров чистой регрессии:

$$b_1-m_{b_1}\cdot t_{\text{табл}}\le b_1^*\le b_1+m_{b_1}\cdot t_{\text{табл}},\ 0,496\le b_1^*\le 1,396$$

и

$$b_2-m_{b_2}\cdot t_{\text{табл}}\le b_2^*\le b_2+m_{b_2}\cdot t_{\text{табл}},\ -0,0425\le b_1^*\le 0,2137.$$

6. С помощью частных F-критериев Фишера оценим целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1 при помощи формул (3.16):

$$F_{x_1} = \frac{R^2_{yx_1x_2} - R^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot \frac{n - m - 1}{1}; \; F_{x_2} = \frac{R^2_{yx_1x_2} - R^2_{yx_1}}{1 - R^2_{yx_2x_2}} \cdot \frac{n - m - 1}{1}.$$

Найдем $R^2_{yx_1}$ и $R^2_{yx_2}$:

$$R^2_{yx_1} = r^2_{yx_1} = 0,970^2 = 0,941;$$

$$R^2_{yx_2} = r^2_{yx_2} = 0,941^2 = 0,885.$$

Имеем:

$$F_{x_1} = \frac{0,947 - 0,885}{1 - 0,947} \cdot \frac{20 - 2 - 1}{1} = 19,89;$$

$$F_{x_2} = \frac{0,947 - 0,941}{1 - 0,947} \cdot \frac{20 - 2 - 1}{1} = 1,924.$$

Получили, что $F_{x_2} = 0,89 < F_{\text{табл}}\left(\alpha = 0,05;\ k_1 = 1;\ k_2 = 17\right) = 4,45$. Следовательно, включение в модель фактора x_2 после того, как в модель включен фактор x_1 статистически нецелесообразно: прирост факторной дисперсии за счет дополнительного признака x_2 оказывается незначительным, несущественным; фактор x_2 включать в уравнение после фактора x_1 не следует.

Если поменять первоначальный порядок включения факторов в модель и рассмотреть вариант включения x_1 после x_2, то результат расчета частного F-критерия для x_1 будет иным. $F_{x_1} = 17,86 > F_{\text{табл}} = 4,45$, т.е. вероятность его случайного формирования меньше принятого стандарта $\alpha = 0,05$ (5%). Следовательно, значение частного F-критерия для дополнительно включенного фактора x_1 не случайно, является статистически значимым, надежным, достоверным: прирост факторной дисперсии за счет дополнительного фактора x_1 является существенным. Фактор x_1 должен присутствовать в уравнении, в том числе в варианте, когда он дополнительно включается после фактора x_2.

7. Общий вывод состоит в том, что множественная модель с факторами x_1 и x_2 с $R^2_{yx_1x_2} = 0,947$ содержит неинформативный фактор x_2. Если исключить фактор x_2, то можно ограничиться уравнением парной регрессии:

$$\hat{y}_{x_1} = \alpha + \beta x_1.$$

Найдем его параметры:

$$\beta = \frac{\operatorname{cov}(y, \quad x_1)}{\sigma_{x_2}} = \frac{63,815 - 6,19 \cdot 9,6}{3,571} = 1,23;$$

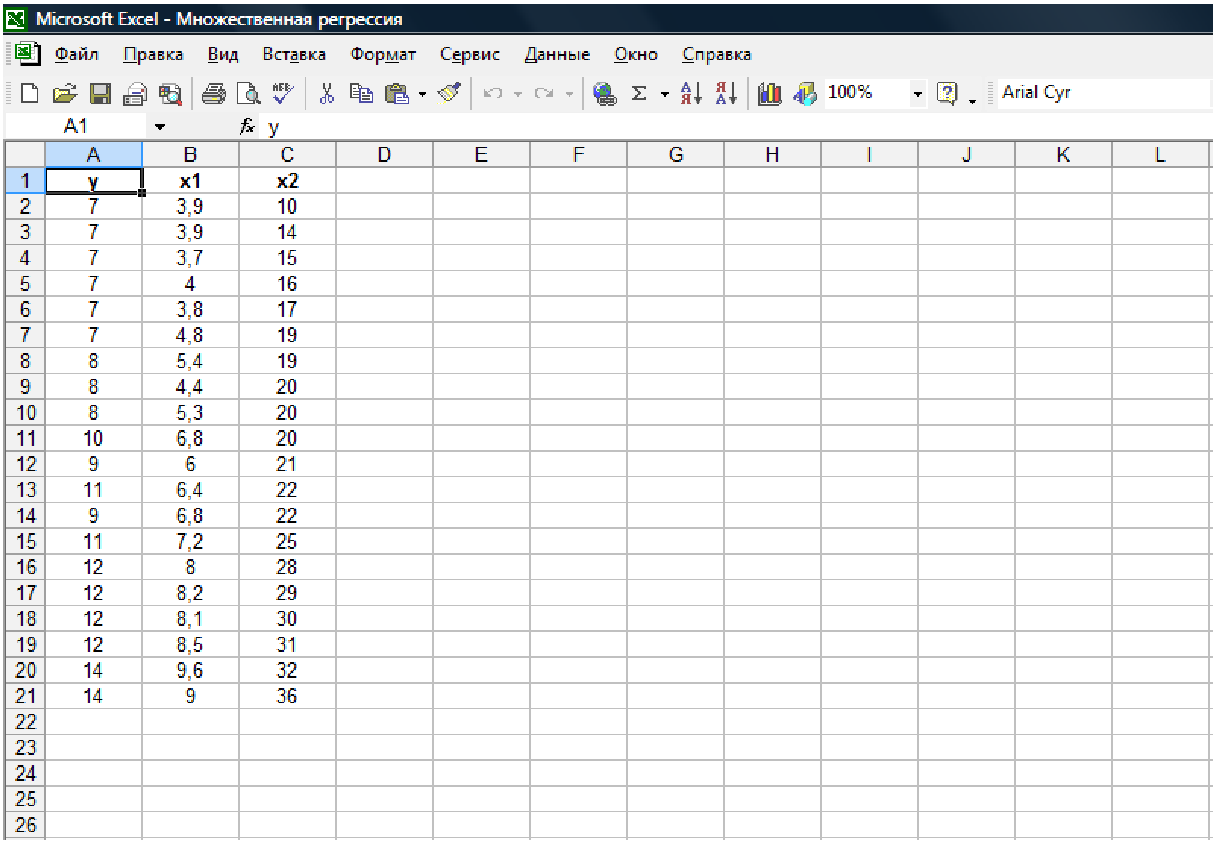
$$\alpha = \overline{y} - \beta \cdot \overline{x} = 9,6 - 1,23 \cdot 6,19 = 1,99.$$

Таким образом,

$$\hat{y}_{x_1} = 1,99 + 1,23 \cdot x_1, \qquad r^2_{yx_1} = 0,941.$$

3.3. Решение типовой задачи в MS Excel

Вносим исходные данные в таблицу **MS Excel**:



	A	B	C
1	y	x1	x2
2	7	3,9	10
3	7	3,9	14
4	7	3,7	15
5	7	4	16
6	7	3,8	17
7	7	4,8	19
8	8	5,4	19
9	8	4,4	20
10	8	5,3	20
11	10	6,8	20
12	9	6	21
13	11	6,4	22
14	9	6,8	22
15	11	7,2	25
16	12	8	28
17	12	8,2	29
18	12	8,1	30
19	12	8,5	31
20	14	9,6	32
21	14	9	36

Рис. 3.1.

Найдем матрицу парных коэффициентов корреляции (**Сервис→Анализ данных→Корреляция**):

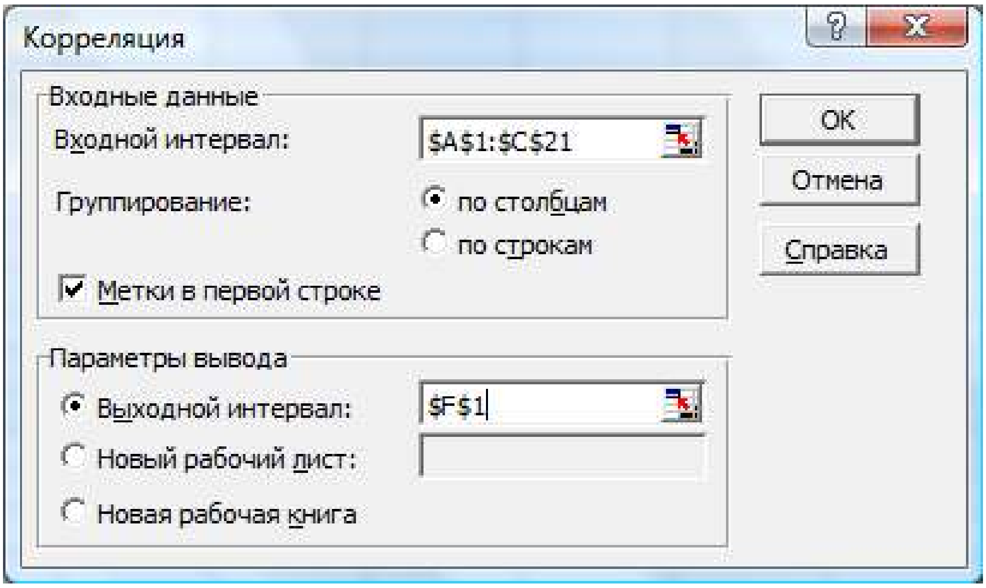


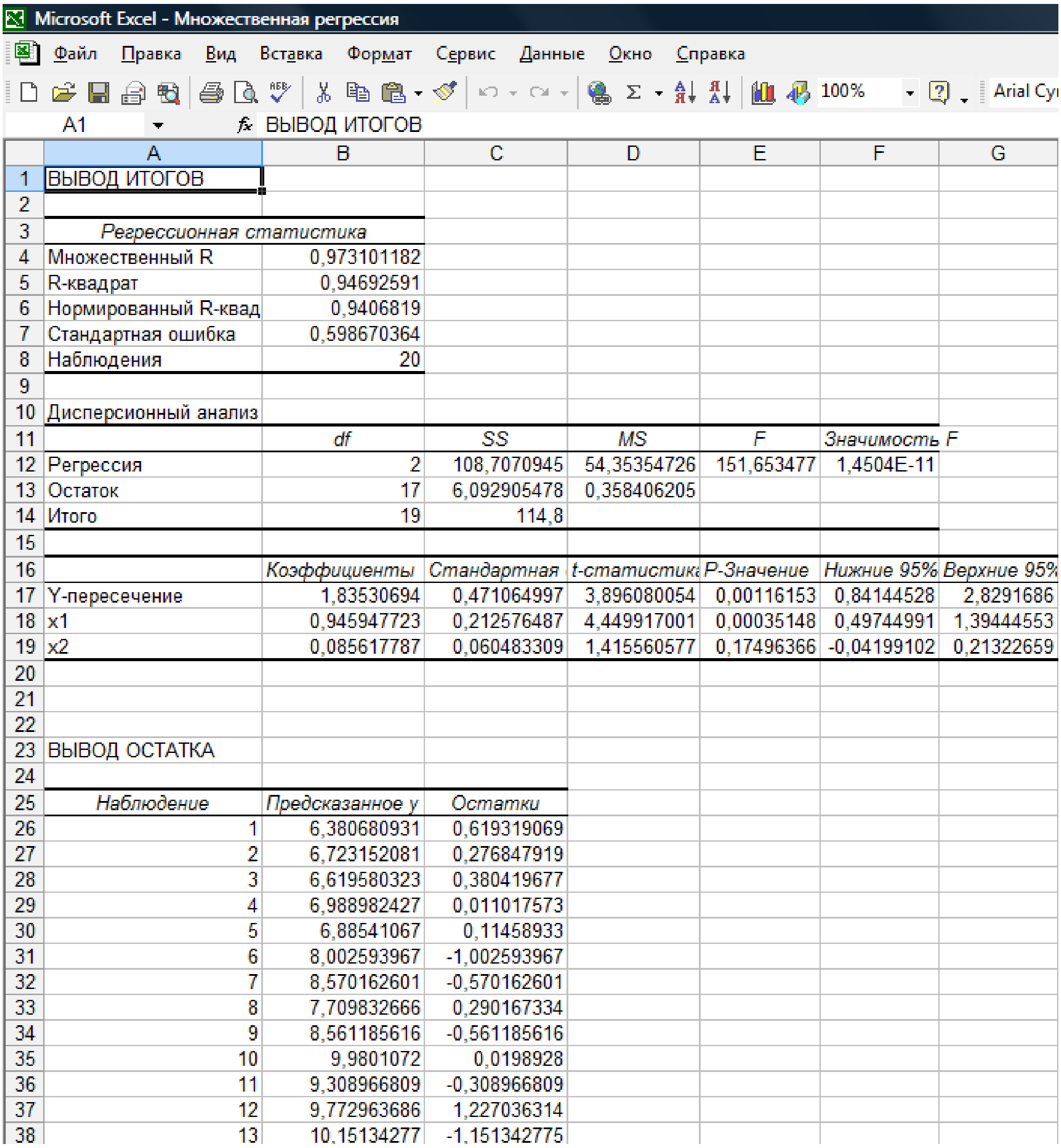
Рис. 3.2.

Получаем следующий результат:

$$\begin{pmatrix} 1 & & \\ 0{,}9699 & 1 & \\ 0{,}9408 & 0{,}9428 & 1 \end{pmatrix}$$

т.е. $r_{yx_1} = 0{,}9699$; $r_{yx_2} = 0{,}9408$; $r_{x_1x_2} = 0{,}9428$.

С помощью инструмента **Регрессия** (**Сервис→Анализ данных→Регрессия**) получаем следующие результаты:



	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	*Регрессионная статистика*						
4	Множественный R	0,973101182					
5	R-квадрат	0,94692591					
6	Нормированный R-квад	0,9406819					
7	Стандартная ошибка	0,598670364					
8	Наблюдения	20					
9							
10	Дисперсионный анализ						
11		*df*	*SS*	*MS*	*F*	*Значимость F*	
12	Регрессия	2	108,7070945	54,35354726	151,653477	1,4504E-11	
13	Остаток	17	6,092905478	0,358406205			
14	Итого	19	114,8				
15							
16		*Коэффициенты*	*Стандартная*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
17	Y-пересечение	1,83530694	0,471064997	3,896080054	0,00116153	0,84144528	2,8291686
18	x1	0,945947723	0,212576487	4,449917001	0,00035148	0,49744991	1,39444553
19	x2	0,085617787	0,060483309	1,415560577	0,17496366	-0,04199102	0,21322659
20							
21							
22							
23	ВЫВОД ОСТАТКА						
24							
25	*Наблюдение*	*Предсказанное y*	*Остатки*				
26	1	6,380680931	0,619319069				
27	2	6,723152081	0,276847919				
28	3	6,619580323	0,380419677				
29	4	6,988982427	0,011017573				
30	5	6,88541067	0,11458933				
31	6	8,002593967	-1,002593967				
32	7	8,570162601	-0,570162601				
33	8	7,709832666	0,290167334				
34	9	8,561185616	-0,561185616				
35	10	9,9801072	0,0198928				
36	11	9,308966809	-0,308966809				
37	12	9,772963686	1,227036314				
38	13	10,15134277	-1,15134277				

Рис. 3.3.

Уравнение регрессии:

$$\hat{y} = 1,8353 + 0,9459x_1 + 0,0856x_2 .$$

Множественный коэффициент корреляции:

$$R = 0,9731 .$$

Коэффициент детерминации:

$$R^2 = 0,9469 .$$

Скорректированный коэффициент детерминации:

$$\hat{R}^2 = 0,9407 .$$

Фактическое значение F-критерия Фишера:

$$F = 151,653 .$$

Фактические значения t-критерия Стьюдента:

$$t_{b_1} = 4,450, \quad t_{b_2} = 1,416 .$$

Доверительные интервалы для параметров регрессии:

$$0,4974 \le b_1^* \le 1,3944 ,$$

$$-0,0420 \le b_2^* \le 0,2132 .$$

Значения частного F-критерия Фишера можно найти как квадрат соответствующего значении t-критерия Стьюдента:

$$F_{x_1} = 4,450^2 = 19,803 , \; F_{x_2} = 1,416^2 = 2,005 .$$

Оставшиеся характеристики можно найти, используя известные формулы и полученные здесь результаты.

Задания для контрольной работы

Задача 1. По территориям региона приводятся данные за 199X г. (p_1 – число букв в полном имени, p_2 – число букв в фамилии):

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	78+ p_1	133+ p_2
2	80+ p_2	148
3	87	135+ p_1
4	79	154
5	106	157+ p_1
6	106+ p_1	195
7	67	139
8	98	158+ p_2
9	73+ p_2	152
10	87	162
11	86	146+ p_2
12	110+ p_1	173

Требуется:

1. Построить линейное уравнение парной регрессии y по x.

2. Рассчитать линейный коэффициент парной корреляции, коэффициент детерминации и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F-критерия Фишера и t-критерия Стьюдента.

4. Выполнить прогноз заработной платы y при прогнозном значении среднедушевого прожиточного минимума x, составляющем 107% от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

6. На одном графике отложить исходные данные и теоретическую прямую.

7. Проверить вычисления в **MS Excel**.

Задача 2. По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%) (p_1 – число букв в полном имени, p_2 – число букв в фамилии).

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7,0	$3{,}6+0{,}1p_1$	11,0	11	9,0	$6{,}0+0{,}1p_2$	21,0
2	7,0	3,7	13,0	12	11,0	6,4	22,0
3	7,0	3,9	15,0	13	9,0	6,9	22,0
4	7,0	4,0	17,0	14	11,0	7,2	25,0
5	7,0	$3{,}8+0{,}1p_1$	18,0	15	12,0	$8{,}0-0{,}1p_2$	28,0
6	7,0	4,8	19,0	16	12,0	8,2	29,0
7	8,0	5,3	19,0	17	12,0	8,1	30,0
8	8,0	5,4	20,0	18	12,0	8,6	31,0
9	8,0	$5{,}6-0{,}1p_1$	20,0	19	14,0	9,6	32,0
10	10,0	6,8	21,0	20	14,0	$9{,}0+0{,}1p_2$	36,0

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью t-критерия Стьюдента оценить статистическую значимость параметров чистой регрессии.

6. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

7. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

8. Проверить вычисления в **MS Excel**.

Рекомендации к выполнению контрольной работы

Практические задания по курсу «Эконометрика» следует выполнять в тетради или на листах бумаги формата А4 (листы скрепляются и заполняются с одной стороны). Работа обязательно должна содержать титульный лист с указанными на нем фамилии, полного имени и номера группы студента. Данные каждого варианта определяется параметрами p_1, p_2. При выполнении контрольных заданий студент должен подставить там, где это необходимо, вместо буквенных параметров индивидуальные анкетные характеристики: p_1 – число букв в полном имени студента; p_2 – число букв в фамилии студента.

Приложение 1.

Таблица значений F -критерия Фишера при уровне значимости $\alpha = 0,05$

k_1 / k_2	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1

Приложение 2.

Критические значения t -критерия Стьюдента при уровне значимости 0,10, 0,05, 0,01 (двухсторонний).

Число степеней свободы d.f.	α			Число степеней свободы d.f.	α		
	00,10	0,05	0,01		00,10	0,05	0,01
1	6,3138	12,706	63,657	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1825	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,5041	21	1,7207	2,0796	2,8314
5	2,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7530	2,1315	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2,8982	∞	1,6449	1,9600	2,5758

Список литературы

Основная:

1. Эконометрика: Учебник / Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2006. – 576 с.

2. Практикум по эконометрике: Учебн. пособие / Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2006. – 344 с.

3. Эконометрика: Учебно-методическое пособие / Шалабанов А.К., Роганов Д.А. – Казань: ТИСБИ, 2004. – 198 с.

4. Доугерти К. Введение в эконометрику: Пер. с англ. – М.: ИНФРА-М, 1999. – 402 с.

Дополнительная:

5. Кремер Н.Ш., Путко Б.А. Эконометрика: Учебник для вузов / Под ред. проф. Н.Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2002. – 311 с.

6. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учебник. – М.: Дело, 2001. – 400 с.

7. Катышев П.К., Магнус Я.Р., Пересецкий А.А. Сборник задач к начальному курсу эконометрики. – М.: Дело, 2002. – 208 с.

8. Эконометрика: Учебник / Тихомиров Н.П., Дорохина Е.Ю. – М.: Издательство «Экзамен», 2003. – 512 с.

9. Сборник задач по эконометрике: Учебное пособие для студентов экономических вузов / Сост. Е.Ю. Дорохина, Л.Ф. Преснякова, Н.П. Тихомиров. – М.: Издательство «Экзамен», 2003. – 224 с.

10. Гмурман В.Е. Теория вероятностей и математическая статистика: Учебн. пособие для вузов. – М.: Высш. шк., 2002. – 479 с.