

Для двухфакторной модели данная система будет иметь вид:

$$\begin{cases} na + b_1 \sum x_1 + b_2 \sum x_2 = \sum y, \\ a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 = \sum y x_1, \\ a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 = \sum y x_2. \end{cases} \quad (3.3)$$

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

$$\begin{aligned} b_1 &= \frac{\sigma_y}{\sigma_{x_1}} \cdot \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}; \\ b_2 &= \frac{\sigma_y}{\sigma_{x_2}} \cdot \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}; \\ a &= \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2. \end{aligned} \quad (3.4)$$

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

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

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \dots + \beta_m t_{x_m} + \varepsilon, \quad (3.5)$$

где $t_y, t_{x_1}, \dots, t_{x_m}$ – *стандартизированные переменные*: $t_y = \frac{y - \bar{y}}{\sigma_y}$,

$t_{x_i} = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$, для которых среднее значение равно нулю: $\bar{t}_y = \bar{t}_{x_i} = 0$, 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

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

где r_{yx_j} и $r_{x_jx_j}$ – коэффициенты парной и межфакторной корреляции.

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

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

$$\bar{\Theta}_{yx_j} = b_j \frac{\bar{x}_j}{\bar{y}}, \quad (3.8)$$

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*:

$$R_{yx_1x_2\dots x_m} = \sqrt{1 - \frac{\sigma_{y_{\text{ост}}}^2}{\sigma_y^2}}. \quad (3.9)$$

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

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

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

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

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:

$$R_{yx_1x_2\dots x_m} = \sqrt{\sum \beta_i \cdot r_{yx_i}}, \quad (3.11)$$

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 $R_{yx_1x_2\dots x_m}^2$.

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

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

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

$$r_{yx_i \cdot x_1x_2\dots x_{i-1}x_{i+1}\dots x_m} = \sqrt{1 - \frac{1 - R_{yx_1x_2\dots x_i\dots x_m}^2}{1 - R_{yx_1x_2\dots x_{i-1}x_{i+1}\dots x_m}^2}}, \quad (3.13)$$

или по рекуррентной формуле:

$$r_{y x_i \cdot x_1 x_2 \dots x_{i-1} x_{i+1} \dots x_m} = \frac{r_{y x_i \cdot x_1 x_2 \dots x_{i-1} x_{i+1} \dots x_{m-1}} - r_{y x_m \cdot x_1 x_2 \dots x_{m-1}} \cdot r_{x_i x_m \cdot x_1 x_2 \dots x_{i-1} x_{i+1} \dots x_{m-1}}}{\sqrt{(1 - r_{y x_m \cdot x_1 x_m \dots x_{m-1}}^2)(1 - r_{x_i x_m \cdot x_1 x_2 \dots x_{i-1} x_{i+1} \dots x_{m-1}}^2)}} \quad (3.14)$$

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:

$$r_{y x_1 \cdot x_2} = \sqrt{1 - \frac{1 - R_{y x_1 x_2}^2}{1 - r_{y x_2}^2}}; \quad r_{y x_2 \cdot x_1} = \sqrt{1 - \frac{1 - R_{y x_1 x_2}^2}{1 - r_{y x_1}^2}}.$$

$$r_{y x_1 \cdot x_2} = \frac{r_{y x_1} - r_{y x_2} \cdot r_{x_1 x_2}}{\sqrt{(1 - r_{y x_2}^2) \cdot (1 - r_{x_1 x_2}^2)}}; \quad r_{y x_2 \cdot x_1} = \frac{r_{y x_2} - r_{y x_1} \cdot r_{x_1 x_2}}{\sqrt{(1 - r_{y x_1}^2) \cdot (1 - r_{x_1 x_2}^2)}}.$$

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

$$F_{x_i} = \frac{R_{y x_1 \dots x_i \dots x_m}^2 - R_{y x_1 \dots x_{i-1} x_{i+1} \dots x_m}^2}{1 - R_{y x_1 \dots x_i \dots x_m}^2} \cdot \frac{n - m - 1}{1} \quad (3.16)$$

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

$$t_{b_i} = \frac{b_i}{m_{b_i}} \quad (3.17)$$

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

$$m_{b_i} = \frac{\sigma_y \cdot \sqrt{1 - R_{yx_1 \dots x_m}^2}}{\sigma_{x_i} \cdot \sqrt{1 - R_{x_i x_1 \dots x_m}^2}} \cdot \frac{1}{\sqrt{n - m - 1}}, \quad (3.18)$$

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

$$m_{b_1} = \frac{\sigma_y \cdot \sqrt{1 - R_{yx_1 x_2}^2}}{\sigma_{x_1} \cdot \sqrt{1 - r_{x_1 x_2}^2}} \cdot \frac{1}{\sqrt{n - 3}}; \quad (3.19)$$

$$m_{b_2} = \frac{\sigma_y \cdot \sqrt{1 - R_{yx_1 x_2}^2}}{\sigma_{x_2} \cdot \sqrt{1 - r_{x_1 x_2}^2}} \cdot \frac{1}{\sqrt{n - 3}}. \quad (3.20)$$

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

$$|t_{b_i}| = \sqrt{F_{x_i}}. \quad (3.21)$$

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:

$$z = \begin{cases} 1 & \text{— мужской пол,} \\ 0 & \text{— женский пол.} \end{cases} \quad (3.22)$$

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:

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

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

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

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:

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

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

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

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

$$b_1 = \frac{\sigma_y}{\sigma_{x_1}} \cdot \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2} = \frac{2,396}{1,890} \cdot \frac{0,970 - 0,941 \cdot 0,943}{1 - 0,943^2} = 0,946;$$

$$b_2 = \frac{\sigma_y}{\sigma_{x_2}} \cdot \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2} = \frac{2,396}{6,642} \cdot \frac{0,941 - 0,970 \cdot 0,943}{1 - 0,943^2} = 0,0856;$$

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

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:

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

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):

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

Вычисляем:

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

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 у 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:

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

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

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):

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

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

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

3. Нескорректированный коэффициент множественной детерминации $R_{yx_1x_2}^2 = 0,947$ 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

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

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):

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

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

Фактические значения t -критерия Стьюдента:

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

Табличное значение критерия при уровне значимости $\alpha = 0,05$ и числе степеней свободы $k = 17$ составит $t_{\text{табл}}(\alpha = 0,05; k = 17) = 2,11$. 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:

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

и

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

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):

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

Найдем $R_{yx_1}^2$ и $R_{yx_2}^2$:

$$R_{yx_1}^2 = r_{yx_1}^2 = 0,970^2 = 0,941;$$

$$R_{yx_2}^2 = r_{yx_2}^2 = 0,941^2 = 0,885.$$

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 $F_{x_2} = 0,89 < F_{\text{табл}}(\alpha = 0,05; k_1 = 1; k_2 = 17) = 4,45$.

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:

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

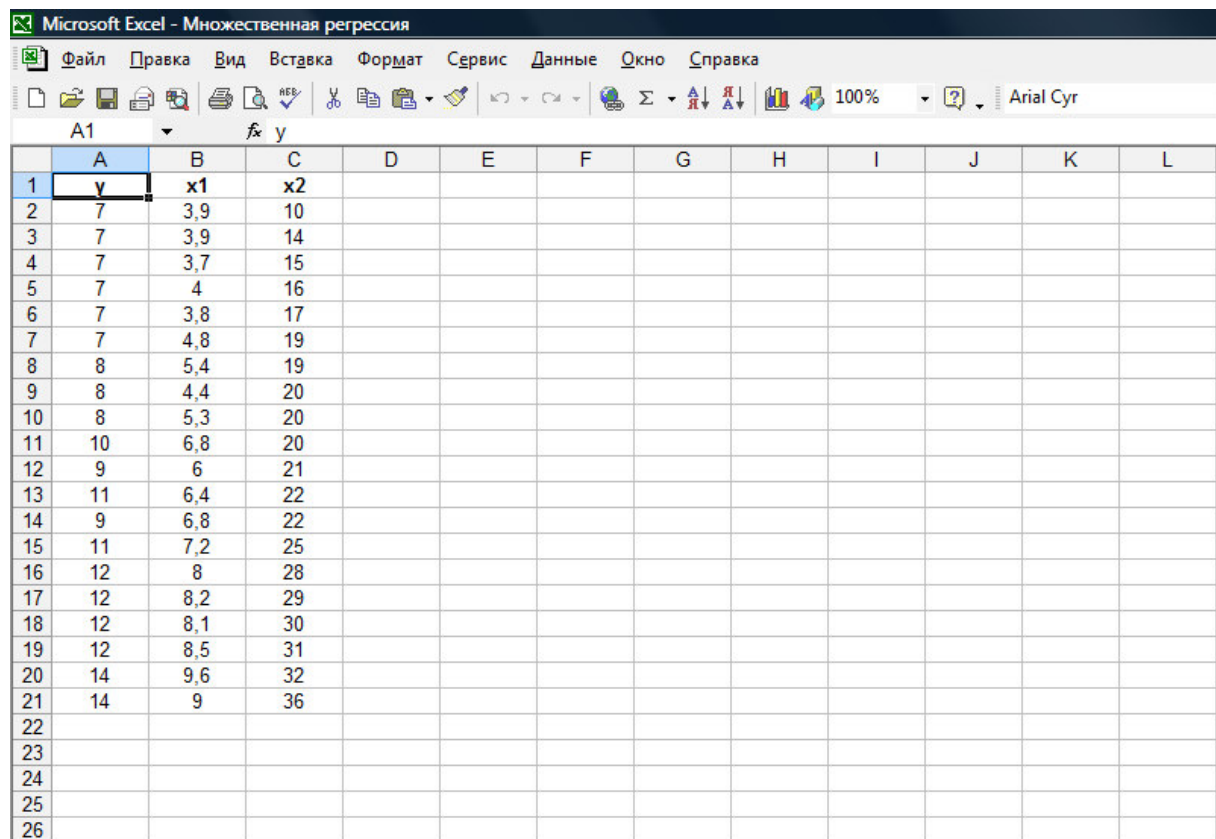
$$\alpha = \bar{y} - \beta \cdot \bar{x} = 9,6 - 1,23 \cdot 6,19 = 1,99.$$

Таким образом,

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

3.3. Решение типовой задачи в MS Excel

Вносим исходные данные в таблицу MS Excel:



	A	B	C	D	E	F	G	H	I	J	K	L
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									
22												
23												
24												
25												
26												

Рис. 3.1.

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

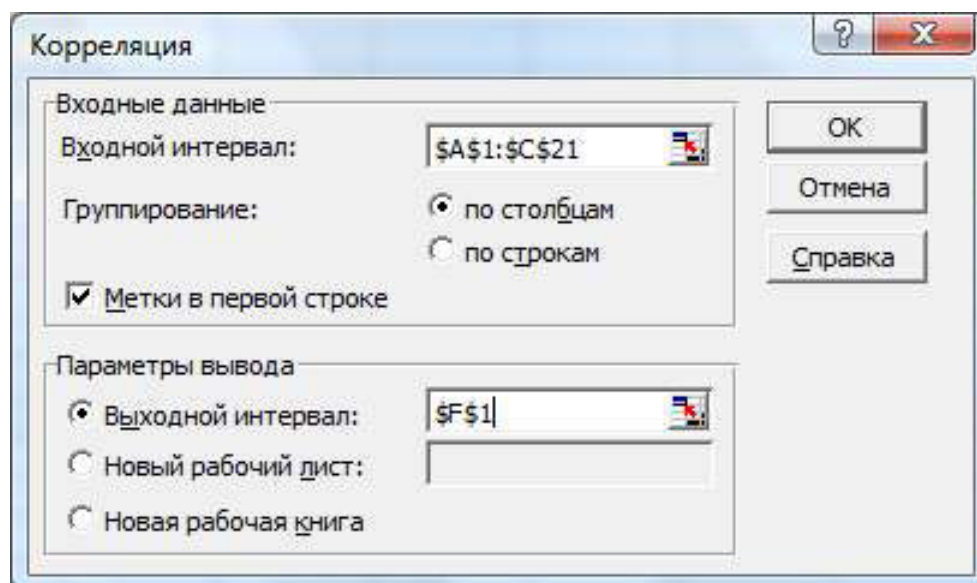


Рис. 3.2.

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:

Microsoft Excel - Множественная регрессия							
Файл Правка Вид Вставка Формат Сервис Данные Окно Справка							
A1 fx ВЫВОД ИТОГОВ							
	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	<i>Регрессионная статистика</i>						
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	<i>Дисперсионный анализ</i>						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>	
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		<i>Коэффициенты</i>	<i>Стандартная</i>	<i>t-статистика</i>	<i>P-Значение</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>
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	<i>ВЫВОД ОСТАТКА</i>						
24							
25	<i>Наблюдение</i>	<i>Предсказанное y</i>	<i>Остатки</i>				
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	-1,151342775				

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:

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

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

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

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

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**.

Приложение 1.

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

$k_1 \backslash k_2$	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