



ФИНАНСОВЫЙ УНИВЕРСИТЕТ  
ПРИ ПРАВИТЕЛЬСТВЕ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

# *Практикум по количественным методам*

---

Занятие 1

Бурцева Юлия Валентиновна

[burcevajulia67@mail.ru](mailto:burcevajulia67@mail.ru)



# Литература

---

**Теория статистики с основами теории вероятностей:**  
Т33 Учеб. пособие для вузов/И.И. Елисеева, В.С. Князевский, Л.И. Ниворожкина, З.А. Морозова; Под ред. И.И. Елисеевой. — М.: ЮНИТИ-ДАНА, 2001. — 446 с.  
ISBN 5-238-00132-0.

**Кремер Н.Ш.**  
К79 Теория вероятностей и математическая статистика: Учебник для вузов. — 2-е изд., перераб. и доп.— М.: ЮНИТИ-ДАНА, 2004. — 573 с.  
ISBN 5-238-00573-3

**Гмурман В. Е.**  
111 Теория вероятностей и математическая статистика. Изд. 4-е, доп. Учеб. пособие для вузов. М., «Высш. школа», 1972.



# Литература

---

Орлова И.В.

- О 66**      Экономико-математические методы и модели. Выполнение расчетов в среде EXCEL / Практикум: Учебное пособие для вузов. – М.: ЗАО «Финстатинформ», 2000. – 136 с.

ISBN 5-7866-0142-0

- л82      **Луценко А.Г., Поляков В.А.**  
**Аналитические методы и информационные технологии в обработке экономической информации: Методические рекомендации для выполнения контрольных, курсовых и выпускных квалификационных работ.** – Тула: ВЗФЭИ, 2010. – 64 с.



# Литература

---

СБОРНИК ЗАДАНИЙ К КОНТРОЛЬНОЙ РАБОТЕ

по дисциплине

«Анализ данных в Microsoft Excel»

Szd\_Analyzdann\_bBi\_17.pdf



# Регрессионный анализ

---

В зависимости от количества факторов, включенных в регрессию, принято различать регрессию простую и множественную.

**Простая регрессия** представляет собой регрессию между двумя переменными ( $y$  и  $x$ ), т.е. рассматривается модель вида:  $y=f(x)$

**Множественная регрессия** соответственно представляет собой регрессию результирующего признака с двумя и большим числом факторов, т.е. рассматривается модель  $y=f(x_1, x_2, x_3, \dots)$



# Парная регрессия

---

Исследование связи между переменными начинается с теории, устанавливающей связь между явлениями.

Прежде всего из круга факторов, влияющих на результативный признак, необходимо выделить наиболее существенно влияющие факторы. Парная регрессия возможна, если рассматривается доминирующий фактор. Предположим, что выдвигается гипотеза: величина спроса на товар  $A$  находится в обратной зависимости от цены. В этом случае необходимо знать, какие остальные факторы предполагаются неизменными. Возможно, в дальнейшем их придется учесть в модели и от простой регрессии перейти к множественной.



# Парная регрессия

---

В парной регрессии выбор вида математической функции  $\hat{y}_x = f(x)$  может быть осуществлен тремя путями: *графически, аналитически*, т.е. исходя из теории взаимосвязи, и *экспериментально*.

# Парная регрессия

➤ При изучении зависимости между двумя признаками графический метод подбора вида уравнения регрессии достаточно нагляден: на основе поля корреляции. Основные типы кривых, используемые при количественной оценке связей, представлены ниже:

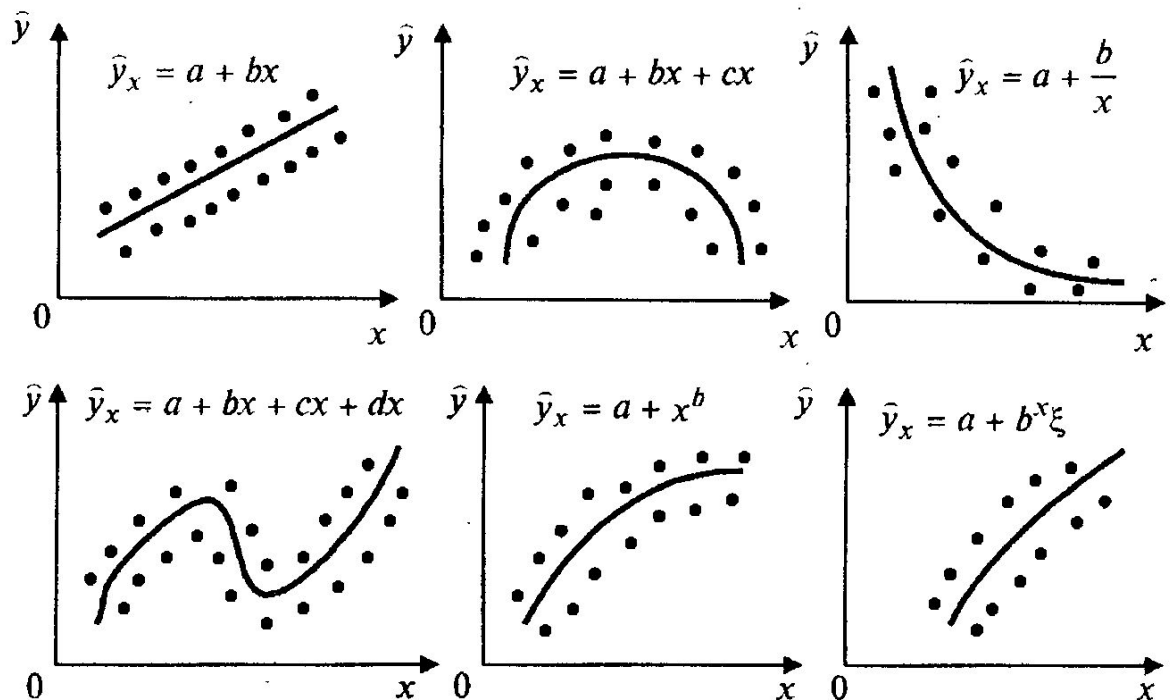


Рис. 14.1. Основные типы кривых, используемые при количественной оценке связей





# Парная регрессия

---

Класс математических функций для описания связи двух переменных достаточно широк. Кроме уже указанных, используются и другие типы кривых:

$$\hat{y} = \frac{1}{a + bx}; \quad \hat{y} = a + bx + c \frac{1}{x}; \quad \hat{y} = a + b \lg x;$$

$$\hat{y} = \frac{1}{a + bx + cx^2}; \quad \hat{y} = \frac{a}{1 + be^{-cx}};$$

$$\lg \hat{y} = a + bx + cx^2.$$



# Парная регрессия

---

В практических исследованиях, как правило, имеет место некоторое рассеяние точек относительно линии регрессии. Оно обусловлено влиянием прочих, не учитываемых в уравнении регрессии факторов, т.е. имеют место отклонения фактических данных от теоретических ( $y - y_x$ ).

Значение этих отклонений лежит в основе расчета остаточной вариации:

$$D_{\text{ост}} = \frac{1}{n} \sum (y - \hat{y}_x)^2.$$

Чем меньше значение остаточной дисперсии, тем в меньшей мере наблюдается влияние прочих, не учитываемых в уравнении регрессии факторов, тем лучше уравнение регрессии подходит к исходным данным. При машинной обработке статистических данных перебираются разные математические функции и в автоматическом режиме выбирается та из них, для которой остаточная дисперсия является наименьшей.



# Парная регрессия

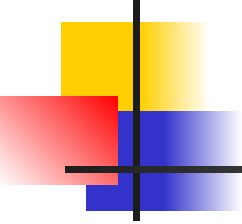
---

Если остаточная дисперсия оказывается примерно одинаковой для нескольких функций, то на практике предпочтение отдается более простым видам функций, так как они в большей степени поддаются интерпретации и требуют меньшего объема наблюдений.

Результаты многих исследований подтверждают, что число наблюдений должно в 6—7 раз превышать число рассчитываемых параметров при переменной  $x$ . Это означает, что искать линейную регрессию, имея менее 7 наблюдений, вообще не имеет смысла. Если вид функции усложняется, то требуется увеличение объема наблюдений: каждый параметр при  $x$  должен содержать хотя бы 7 наблюдений. Значит, если мы выбираем параболу второй степени:

$$\hat{y} = a + bx + cx^2,$$

то требуется объем информации уже не менее 14 наблюдений.



# Парная линейная регрессия и корреляция

---

Линейная регрессия сводится к нахождению уравнения вида:

$$\hat{y}_x = a + bx .$$

Уравнение вида  $\hat{y}_x = a + bx$  позволяет по заданным значениям фактора  $x$  иметь теоретические значения результативного признака, подставляя в него фактические значения фактора  $x$ . На графике эти теоретические значения представляют линию регрессии. На практике построение линейной регрессии сводится к оценке ее параметров:  $a$  и  $b$ .

# Парная линейная регрессия и корреляция

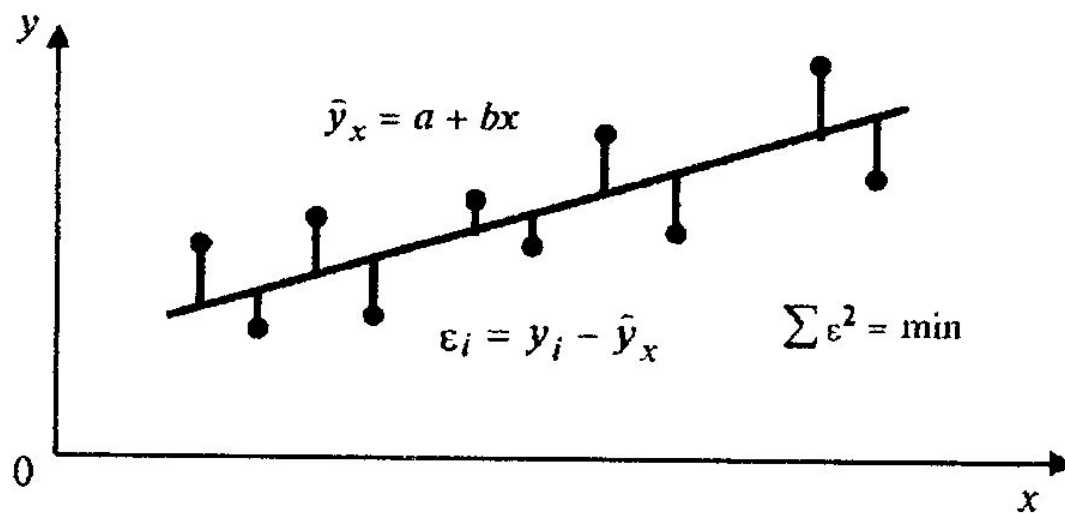
При классическом подходе оценивание параметров линейной регрессии проводится *методом наименьших квадратов*.

Метод наименьших квадратов позволяет получить такие оценки параметров  $a$  и  $b$ , при которых сумма квадратов отклонений фактических значений результативного признака  $y$  от расчетных, теоретических ( $\hat{y}_x$ ) была бы минимальной, т.е.

$$\sum_{i=1}^n (y_i - \hat{y}_{ix})^2 = \min .$$

# Парная линейная регрессия и корреляция

Иными словами, среди множества точек корреляционного поля линия на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была минимальной (рис. 12.3).



**Рис. 14.3. Оценка параметров линейной регрессии методом наименьших квадратов**





# Парная линейная регрессия и корреляция

---

Чтобы найти минимум функции, находятся частные производные по каждому из параметров  $(a, b)$  и приравниваются нулю.

Обозначим  $\sum \varepsilon^2$  через  $S$ . Тогда  $S = \sum (y_i - y_x)^2 = \sum (y - a - bx)^2$ ;

$$\frac{dS}{da} = -2\sum y_x + 2na + 2b\sum x = 0;$$

$$\frac{dS}{db} = -2\sum y_x + 2a\sum x + 2b\sum x^2 = 0.$$

Получим систему нормальных уравнений для оценки параметров:

$$\begin{cases} na + b\sum x = \sum y; \\ a\sum x + b\sum x^2 = \sum yx. \end{cases}$$

# Парная линейная регрессия и корреляция

Решая данную систему нормальных уравнений либо методом последовательного исключения переменных, либо методом определителей, найдем оценки искомых параметров  $a$  и  $b$ . Можно воспользоваться готовыми формулами:  $a = \bar{y} - b\bar{x}$  (вытекает из первого уравнения системы нормальных уравнений, если все его члены разделить на  $n$ ):

$$b = \frac{\text{cov}(x, y)}{\sigma_x^2},$$

где  $\text{cov}(x, y)$  — ковариация признаков;

$\sigma_x^2$  — дисперсия признака  $x$ .





# Парная линейная регрессия и корреляция

---

Ввиду того, что  $\text{cov}(x, y) = \overline{yx} - \bar{y} \cdot \bar{x}$ , а  $\sigma_x^2 = \overline{x^2} - (\bar{x})^2$ , получим формулу расчета оценки параметра  $b$ :

$$b = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - (\bar{x})^2}.$$

Параметр  $\bar{b}$  называется *коэффициентом регрессии*.

# Парная линейная регрессия и корреляция



---

Параметр  $a$  может не иметь экономического смысла, если его нельзя толковать как значение результативного показателя при  $x = 0$ . Может оказаться, что  $x = 0$  может не быть. Попытки экономически интерпретировать параметр  $a$  могут привести к абсурдам, особенно при  $a < 0$ .

# Парная линейная регрессия и корреляция

Предположим, по группе предприятий, выпускающих один и тот же вид продукции, рассматривается функция издержек:  $y = a + bx + c$ . Информация, необходимая для расчета оценок параметров  $a$  и  $b$ , представлена в табл. 12.1.

<i>Номер предприятия</i>	<i>Выпуск продукции <math>x</math>, тыс. ед</i>	<i>Затраты на производство <math>y</math>, млн руб.</i>
1	1	30
2	2	70
3	4	150
4	3	100
5	5	170
6	3	100
7	4	150
<b>Итого</b>	<b>22</b>	<b>770</b>

# Парная линейная регрессия и корреляция

№	x	y	x*y	x^2	y^2	Y(x)
1	1	30	30	1	900	31,05263
2	2	70	140	4	4900	67,89474
3	4	150	600	16	22500	141,5789
4	3	100	300	9	10000	104,7368
5	5	170	850	25	28900	178,4211
6	3	100	300	9	10000	104,7368
7	4	150	600	16	22500	141,5789
<b>сумма:</b>	<b>22</b>	<b>770</b>	<b>2820</b>	<b>80</b>	<b>99700</b>	<b>770</b>

**среднее:      3,142857                  110      402,8571      11,42857      14242,86                  110**

# Парная линейная регрессия и корреляция

$$\begin{cases} na + b \sum x = \sum y \\ a \sum x + b \sum x^2 = \sum yx \end{cases} \Rightarrow \begin{cases} a + b \frac{\sum x}{n} = \frac{\sum y}{n} \\ a \frac{\sum x}{n} + b \frac{\sum x^2}{n} = \frac{\sum yx}{n} \end{cases} \Rightarrow \begin{cases} a + b\bar{x} = \bar{y} \\ a\bar{x} + b\bar{x}^2 = \bar{y}\bar{x} \end{cases}$$

$$\begin{cases} b = \frac{\bar{y} \cdot \bar{x} - \bar{y}\bar{x}}{\bar{x} \cdot \bar{x} - \bar{x}^2} \\ a = \bar{y} - b \cdot \bar{x} \end{cases} \Rightarrow \begin{cases} b = 36,84211 \\ a = -5,78947 \end{cases}$$

*Определив коэффициенты получим уравнение :*

$$y(x) = -5,79 + 36,84 \cdot x$$

# Парная линейная регрессия и корреляция

Значение параметра  $a$  в данном примере не имеет экономического смысла. Интерпретировать можно лишь знак при параметре  $a$ . Если  $a > 0$ , то относительное изменение результата идет медленнее, чем изменение фактора. Иными словами, вариация результата меньше вариации фактора: коэффициент вариации по фактору  $x$  выше коэффициента вариации для результата  $y$  ( $V_x > V_y$ ). Для доказательства данного положения сравним относительные изменения фактора  $x$  и результата  $y$ :

$$\frac{dy}{y} < \frac{dx}{x} \quad \text{или} \quad \frac{dy}{dx} < \frac{y}{x}; \quad \frac{b \cdot dx}{dx} < \frac{a + bx}{x}, \quad bx < a + bx,$$

откуда  $0 < a$  и  $a > 0$ .

# Парная линейная регрессия и корреляция

В рассматриваемом примере:

$$\bar{x} = 3,14; \quad x = 1,25; \quad V_x = 39,8\% ;$$

$$\bar{y} = 110; \quad y = 46,29; \quad V_y = 42,1\% .$$

В уравнении регрессии  $a < 0$ , что соответствует опережению изменения результата над изменением фактора:  $V_y > V_x$ .

Если переменные  $x$  и  $y$  выразить через отклонения от средних уровней, то линия регрессии на графике пройдет через начало координат:  $y' = bx'$ , где  $y' = y - \bar{y}$  и  $x' = x - \bar{x}$ . Оценка коэффициента регрессии при этом не изменится.

# Парная линейная регрессия и корреляция

Уравнение регрессии всегда дополняет линейный коэффициент корреляции  $r_{yx}$ . Существуют разные модификации формулы линейного коэффициента корреляции. Некоторые из них приведены ниже:

$$r_{yx} = b \frac{\sigma_x}{\sigma_y} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{y\bar{x} - \bar{y} \cdot \bar{x}}{\sigma_x \cdot \sigma_y}.$$

Как известно, линейный коэффициент корреляции находится в границах:  $-1 \leq r_{yx} \leq 1$ .

Если коэффициент регрессии  $b > 0$ , то  $0 < r_{yx} < 1$ , и наоборот: при  $b < 0$   $-1 \leq r_{yx} \leq 0$ .



# Парная линейная регрессия и корреляция

---

В рассматриваемом примере *по данным табл. 14.1* значение линейного коэффициента корреляции составило 0,991, что достаточно близко к 1 и означает наличие очень тесной зависимости затрат на производство от объема выпущенной продукции.

# Парная линейная регрессия и корреляция

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции  $r_{yx}^2$ , называемый коэффициентом детерминации.

*Коэффициент детерминации* характеризует долю дисперсии результативного признака  $y$ , объясняемую регрессией, в общей дисперсии. Соответственно величина  $(1 - r^2)$  характеризует долю дисперсии  $y$ , вызванную влиянием остальных не учтенных в модели факторов.

В нашем примере  $r^2 = 0,982$ . Следовательно, уравнением регрессии объясняется 98,2% дисперсии результативного признака, а на долю прочих факторов приходится лишь 1,8% ее дисперсии. Коэффициент детерминации служит одним из критериев оценки качества линейной модели.

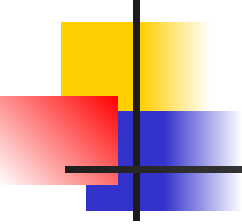


# Оценка существенности параметров линейной регрессии и корреляции

---

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

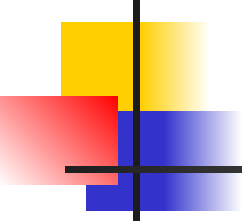
Оценка значимости уравнения регрессии в целом дается с помощью  $F$ -критерия Фишера. При этом выдвигается нулевая гипотеза, что коэффициент регрессии равен нулю:  $b = 0$  и, следовательно, фактор  $x$  не оказывает влияния на результат  $y$ .



# Оценка существенности параметров линейной регрессии и корреляции

---

Если нулевая гипотеза справедлива, то факторная и остаточная дисперсии не отличаются друг от друга. Для опровержения ее необходимо, чтобы факторная дисперсия превышала остаточную в несколько раз. Разработаны (английским статистиком Снедекором) таблицы критических значений  $F$ -отношений при разных уровнях существенности нулевой гипотезы и различном числе степеней свободы. Табличное значение  $F$ -критерия — это максимальное значение отношения дисперсий, которое может иметь место при случайном их расхождении для данного уровня вероятности наличия нулевой гипотезы. Вычисленное значение  $F$ -отношения признается достоверным (отличным от 1), если оно больше табличного. В этом случае отбрасывается нулевая гипотеза об отсутствии связи признаков и делается вывод о существенности этой связи.



# Оценка существенности параметров линейной регрессии и корреляции

---

.....

Если же значение  $F$ -критерия окажется меньше табличного, то вероятность нулевой гипотезы выше заданного уровня (например, 0,05) и она не может быть отклонена без серьезного риска сделать неправильный вывод о наличии связи. В этом случае уравнение регрессии считается статистически незначимым.



# Оценка существенности параметров линейной регрессии и корреляции

В рассматриваемом примере:

$$\begin{aligned}\sum_{(i)} \left( y_i - \bar{y}_x \right)^2 &= \sum y^2 - n(\bar{y})^2 = 99700 - 7 \cdot 110^2 = \\ &= 15\,000 \text{ (общая сумма квадратов);}\end{aligned}$$

$$\begin{aligned}\sum_{(i)} \left( \hat{y}_x - \bar{y} \right)^2 &= b^2 \cdot \sum (x - \bar{x})^2 = (36,84)^2 \cdot [80 - 7 \cdot (22 : 7)]^2 = \\ &= 14\,735 \text{ (факторная сумма квадратов);}\end{aligned}$$

$$\sum \left( y - \hat{y}_x \right)^2 = 15\,000 - 14\,735 = 265;$$

$$D_{\text{факторная}} = 14\,735;$$

$$D_{\text{остаточная}} = 265 : 5 = 53;$$

$$F_{\text{фактическое}} = 14\,735 : 53 = 278.$$

Критические значения  $F$ -критерия для уровней значимости  $\alpha = 0,005$  и  $\alpha = 0,01$ :

для  $\alpha = 0,05$   $F_{1,5} = 6,61$ ; для  $\alpha = 0,01$   $F_{1,5} = 16,26$ .

# Оценка существенности параметров линейной регрессии и корреляции

Т а б л и ц а 4. Критерий Фишера  $F$  при  $\alpha=0,05$

$f_2$	Критические значения $F$ при значениях $f_1$ от 1 до $\infty$											
	1	2	3	4	5	6	7	8	9	10	20	$\infty$
1	161	199	216	225	230	234	237	239	241	242	248	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,66	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,80	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,56	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,87	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,44	3,22
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,15	2,93
9	5,12	4,26	3,84	3,63	3,48	3,37	3,29	3,23	3,18	3,14	2,94	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,77	2,54
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,12	1,84
$\infty$	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,57	1,00

# Оценка существенности параметров линейной регрессии и корреляции

Т а б л и ц а 5. Критерий Фишера  $F$  при  $\alpha=0,1$

$f_2$	Критические значения $F$ при значениях $f_1$ от 1 до $\infty$											
	1	2	3	4	5	6	7	8	9	10	20	$\infty$
1	40	49	54	56	57	58	59	59	60	60	62	63
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,44	9,49
3	5,53	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,18	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,84	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,29	3,21	3,10
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,93	2,84	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,59	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,42	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,41	2,30	2,16
10	3,28	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,20	2,06
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,79	1,61
$\infty$	2,71	2,30	2,08	1,89	1,85	1,77	1,72	1,67	1,63	1,59	1,42	1,00





# Оценка существенности параметров линейной регрессии и корреляции

Поскольку  $F_{\text{фактическое}}$  превышает табличные значения при 5- и 1%-м уровне значимости, то можно сделать вывод о значимости уравнения регрессии (связь доказана).

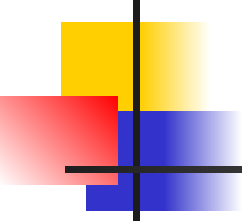
Значение  $F$ -критерия связано с коэффициентом детерминации  $r$ . Факторную сумму квадратов отклонений можно представить как  $r^2 \sigma_y^2 n$ , а остаточную сумму квадратов — как  $(1 - r^2) \cdot \sigma_y^2 = n$ .

Тогда значение  $F$ -критерия можно получить исходя из формулы:

$$F = \frac{r^2}{1 - r^2} \cdot (n - 2).$$

В нашем примере  $r^2 = 0,982$ . Тогда  $F = \frac{0,982}{1 - 0,982} \cdot (7 - 2) = 273$

(некоторое несовпадение результатов связано с ошибками округления).



# Оценка существенности параметров линейной регрессии и корреляции

---