Парная регрессия и корреляция в эконометрических исследованиях. Смысл и оценка параметров

$$y = f(x) + \varepsilon$$

часть значения y, которая объяснена уравнением регрессии

необъясненная часть значения y (или возмущение)

Экономический смысл ε

- *Невключение объясняющих переменных в уравнение.* На самом деле на переменную *Y* влияет не только переменная *X*, но и ряд других переменных, которые не учтены в модели по следующим причинам:
  - мы знаем, что другая переменная влияет, но не можем ее учесть, потому как не знаем, как измерить (психологический фактор, например);
  - существуют факторы, которые мы знаем, как измерить, но влияние их на Y так слабо, что их не стоит учитывать;
  - существенные переменные, но из-за отсутствия опыта или знаний мы их таковыми не считаем.
- *Неправильная функциональная спецификация.* Функциональное соотношение между *Y* и *X* может быть определено неправильно. Например, мы предположили линейную зависимость, а она может быть более сложной.
- *Ошибки наблюдений и измерений.*

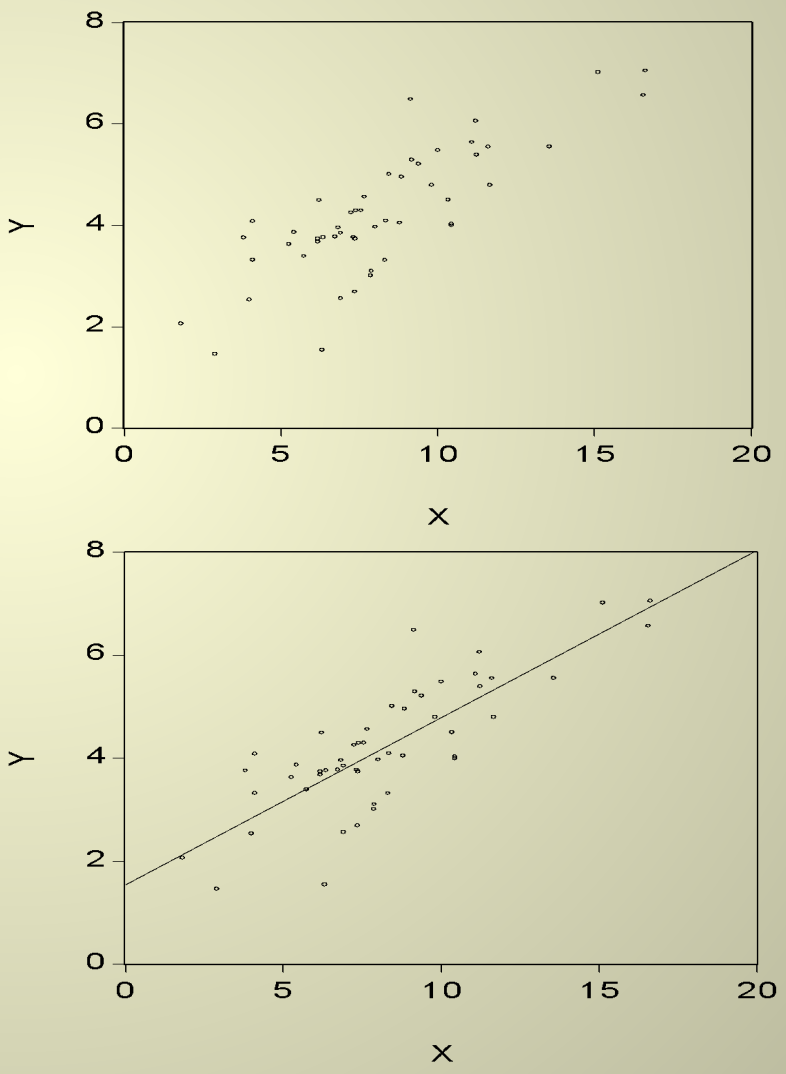
Построение уравнения регрессии
1. Постановка задачи

Данные наблюдений

	x	y
1	x_1	y_1
2	x_2	y_2
…	…	…
n	x_n	y_n

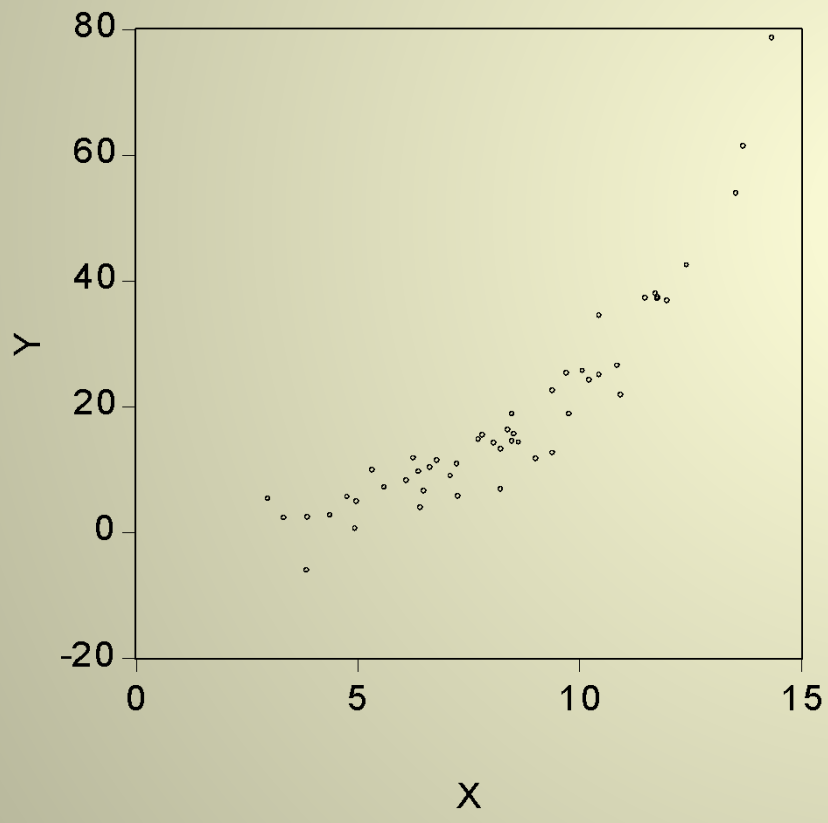
Зависимости $\hat{y} = f(x)$ соответствует некоторая кривая на плоскости. И по форме облака наблюдений можно определить вид регрессионной функции.

Поле корреляции



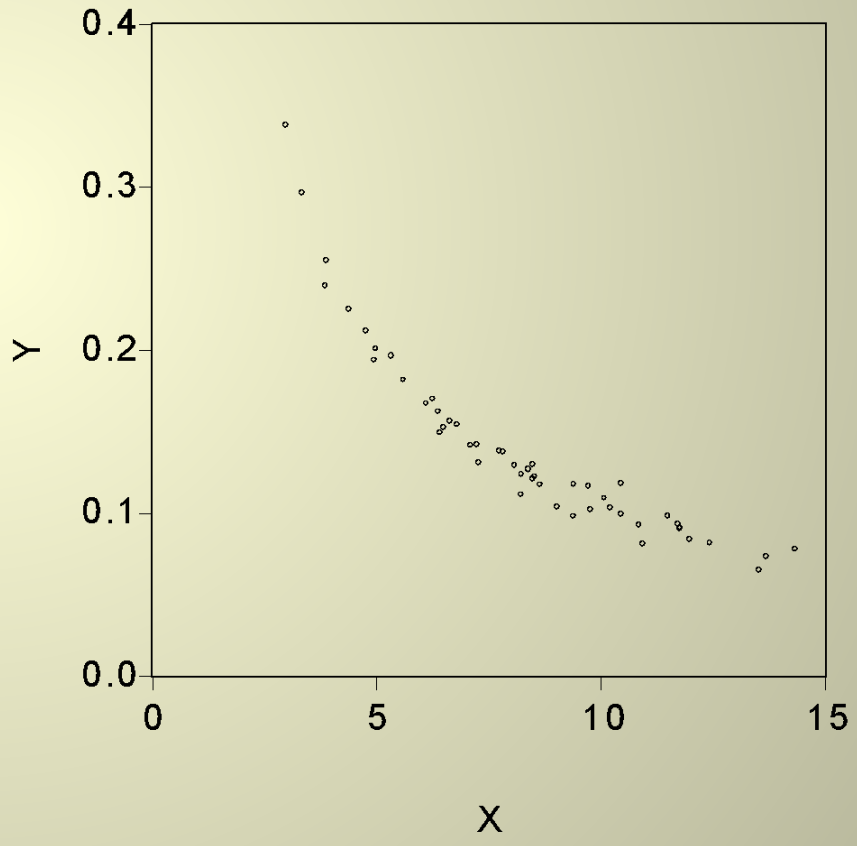
Степенная

$$Y = \alpha e^{\beta X} \varepsilon$$



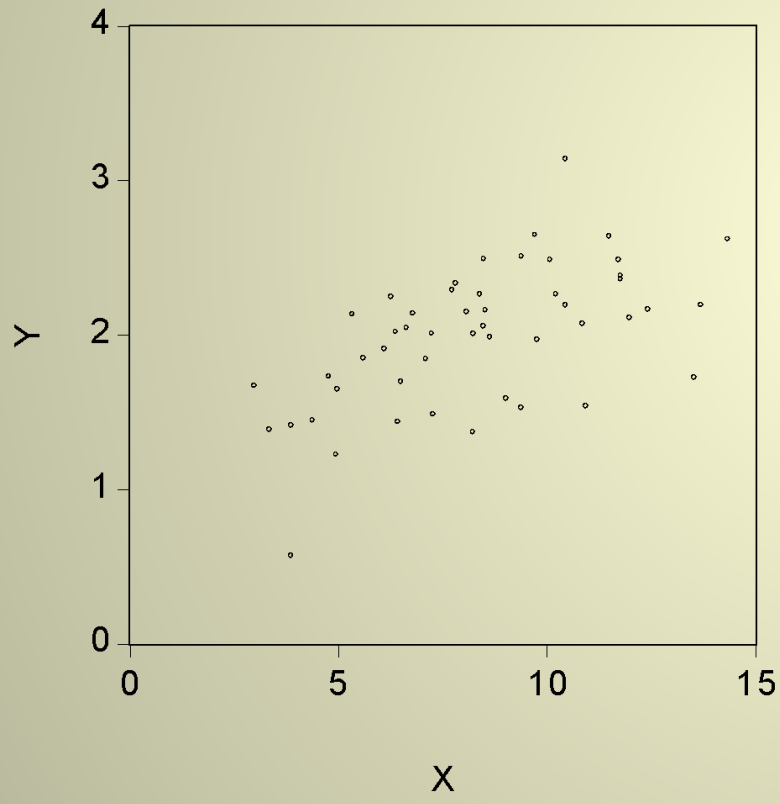
Гиперболическая

$$Y = \alpha + \frac{\beta}{X} + \varepsilon$$

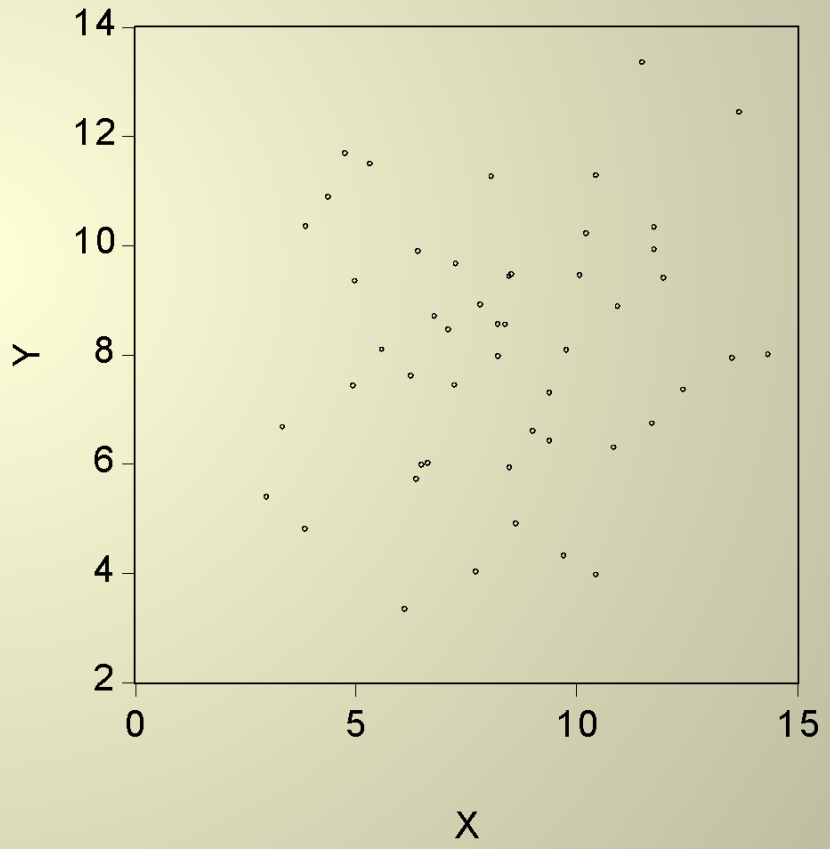


Показательная

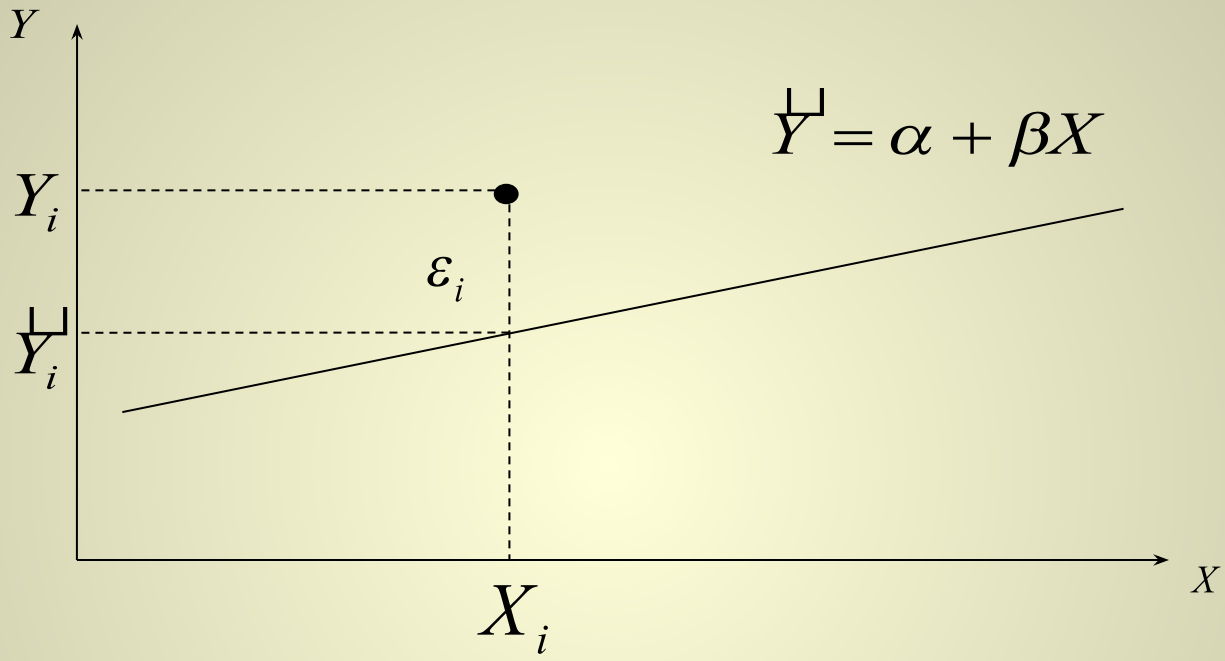
$$Y = \alpha X^{\beta} \varepsilon$$



X и Y независимы



Парная линейная регрессионная модель



Для формализации рассмотрим разность между расчетными (теоретическими) и наблюдаемыми значениями y:

$$\varepsilon_i = y_i - \hat{y}_i$$

Наилучшей считается такая зависимость, для которой сумма квадратов отклонений принимает минимальное значение, т. е.

$$S = \sum (y_i - \hat{y}_i)^2 \to \min$$

2. Спецификация модели

В парной регрессии выбор вида аналитической зависимости может быть осуществлен тремя методами:

– *графическим* (на основе анализа поля корреляции);

– *аналитическим* (на основе изучения теоретической природы связи между исследуемыми признаками);

– *экспериментальным* (построение нескольких моделей различного вида с выбором наилучшей, согласно применяемому критерию качества).

3. Оценка параметров модели

3.1. Оценка параметров линейной парной регрессии – метод наименьших квадратов (МНК)

$$S = \sum (y_i - \hat{y}_i)^2 \to \min \quad \text{или} \quad \sum \varepsilon^2 \to \min$$

$$S = \sum (y_i - \hat{y}_i)^2 = \sum (y - a - bx)^2$$

$$S'_a = -2\sum y + 2na + 2b\sum x = 0$$

$$S'_b = -2\sum yx + 2a\sum x + 2b\sum x^2 = 0$$

Отсюда получаем систему уравнений:

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

Разделим оба уравнения на n:

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

$$a = \frac{\sum y}{n} - \frac{b\sum x}{n} = \bar{y} - b\bar{x}$$

Подставляем во второе уравнение:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

3.2. Оценка параметров нелинейных моделей

Зависимость	Формула	Линеаризующее преобразование	Зависимость между параметрами
Гиперболическая	$y = a + \frac{b}{x}$	**y1=y** **X=1/x**	**a1=a** **b1=b**
Логарифмическая	$y = a + b \times \ln x$	**y1=y** **X=ln x**	**a1=a** **b1=b**
Экспоненциальная	$y = e^{a+bx}$	**Y=ln y** **x1=x**	**a1=a** **b1=b**
Степенная	$y = a \times x^{b}$	**Y=ln y (Y=lg y)** **X=ln x (X=lg x)**	**ln a=C (lg a=C)** **b1=b**
Показательная	$y = a \times b^{x}$	**Y=ln y (Y=lg y)** **x1=x**	**ln a=C (lg a=C)** **ln b=B (lg b=B)**

Оценка параметров внутренне нелинейных моделей:

1. Задаются некоторые «правдоподобные» начальные (исходные) значения параметров a и b.
2. Вычисляются теоретические значения $\hat{y}i = f(xi)$ с использованием этих значений параметров.
3. Вычисляются остатки $ei = \hat{y}i - yi$ и сумма квадратов остатков S.
4. Вносятся изменения в одну или более оценку параметров.
5. Вычисляются новые теоретические значения $\hat{y}i$, остатки ei и S.
6. Если произошло уменьшение S, то новые значения оценок используются в качестве новой отправной точки.
7. Шаги 4, 5 и 6 повторяются до тех пор, пока не будет достигнута ситуация, когда величину S невозможно будет улучшить (в пределах заданной точности).
8. Полученные на последнем шаге значения параметров a и b являются оценками параметров нелинейного уравнения регрессии.

4. Проверка качества уравнения регрессии

H0: уравнение статистически не значимо

$$yi = \hat{y}i + \varepsilon i$$

$$D(y) = D(\hat{y}) + D(\varepsilon)$$

$$\frac{1}{n}\sum(y-\bar{y})^2 = \frac{1}{n}\sum(\hat{y}-\bar{y})^2 + \frac{1}{n}\sum(y-\hat{y})^2$$

полная (общая) сумма квадратов отклонений = **сумма квадратов отклонений, объясненная регрессией** + **(остаточная) сумма квадратов отклонений, не объясненная регрессией**

F-критерий Фишера:

$$F = \frac{\frac{D(\hat{y})}{k}}{\frac{D(\varepsilon)}{n-m-1}} \quad \textit{или} \quad \frac{R^2}{1-R^2} \times \frac{n-m-1}{m}$$

где m – число независимых переменных в уравнении регрессии (для парной регрессии m = 1);

n – число единиц совокупности.

Если **Fфакт > Fтабл**, то *H0* о случайной природе связи отклоняется и признается статистическая значимость и надежность уравнения.

Если **Fфакт < Fтабл**, то *H0* не отклоняется и признается статистическая незначимость уравнения регрессии.

Уровень значимости (α) – вероятность отвергнуть верную гипотезу (ошибка первого рода). Уровень значимости α обычно принимает значения 0,05 и 0,01, что соответствует вероятности совершения ошибки первого рода 5% и 1%.

Число степеней свободы связано с числом единиц совокупности n и с числом определяемых по ней констант: *k1 = m, k2 = n □- m □-1*

3. Значения *F*-критерия Фишера на уровне значимости $\alpha = 0{,}05$

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

t-критерий Стьюдента

H_0: a=0; b=0

Стандартные ошибки параметров регрессии и коэффициента корреляции:

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

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{n-2} \times \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{S^2_{ост} \frac{\sum x^2}{n^2 \sigma^2_x}} = S_{ост} \frac{\sqrt{\sum x^2}}{n \sigma_x}$$

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

2. Значения критических уровней $t_{\alpha,k}$ в зависимости от k степеней свободы и заданного уровня значимости α для распределения Стьюдента

k \ α	0,1	0,05	0,025	0,02	0,01	0,005	0,001
1	6,31	12,71	25,45	31,82	63,66	127,32	636,62
2	2,92	4,30	6,21	6,96	9,92	14,09	31,60
3	2,35	3,18	4,18	4,54	5,84	7,45	12,92
4	2,13	2,78	3,50	3,75	4,60	5,60	8,61
5	2,02	2,57	3,16	3,36	4,03	4,77	6,87
6	1,94	2,45	2,97	3,14	3,71	4,32	5,96
7	1,89	2,36	2,84	3,00	3,50	4,03	5,41
8	1,86	2,31	2,75	2,90	3,36	3,83	5,04
9	1,83	2,26	2,69	2,82	3,25	3,69	4,78
10	1,81	2,23	2,63	2,76	3,17	3,58	4,59
12	1,78	2,18	2,56	2,68	3,05	3,43	4,32
14	1,76	2,14	2,51	2,62	2,98	3,33	4,14
16	1,75	2,12	2,47	2,58	2,92	3,25	4,01
18	1,73	2,10	2,45	2,55	2,88	3,20	3,92
20	1,72	2,09	2,42	2,53	2,85	3,15	3,85
22	1,72	2,07	2,41	2,51	2,82	3,12	3,79
24	1,71	2,06	2,39	2,49	2,80	3,09	3,75
26	1,71	2,06	2,38	2,48	2,78	3,07	3,71
28	1,70	2,05	2,37	2,47	2,76	3,05	3,67
30	1,70	2,04	2,36	2,46	2,75	3,03	3,65
32	1,69	2,04	2,35	2,45	2,74	3,01	3,62
34	1,69	2,03	2,35	2,44	2,73	3,00	3,60
36	1,69	2,03	2,34	2,43	2,72	2,99	3,58
38	1,69	2,02	2,33	2,43	2,71	2,98	3,57
40	1,68	2,02	2,33	2,42	2,70	2,97	3,55
45	1,68	2,01	2,32	2,41	2,69	2,95	3,52
50	1,68	2,01	2,31	2,40	2,68	2,94	3,50
∞	1,64	1,96	2,24	2,33	2,58	2,81	3,29

Оценка значимости параметров уравнения и коэффициента корреляции проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_b = \frac{b}{m_b}; \quad t_a = \frac{a}{m_a}; \quad t_r = \frac{r}{m_r}$$

Если **tфакт > tтабл**, то H_0 отклоняется, т. е. a, b, r не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x.

Если **tфакт < tтабл**, то H_0 не отклоняется и признается случайная природа формирования a, b, r.

Доверительные интервалы – это пределы, в которых лежит точное значение определяемого показателя с заданной вероятностью.

Доверительные интервалы для параметров a и b уравнения линейной регрессии определяются соотношениями:

$$\gamma_a = a \pm t_{табл} \cdot m_a; \quad \gamma_{a_{\min}} = a - t_{табл} \cdot m_a \quad \gamma_{a_{\max}} = a + t_{табл} \cdot m_a$$

$$\gamma_b = b \pm t_{табл} \cdot m_b; \quad \gamma_{b_{\min}} = b - t_{табл} \cdot m_b \quad \gamma_{b_{\max}} = b + t_{табл} \cdot m_b$$

Точечный и интервальный прогноз по уравнению линейной регрессии

Точечный прогноз заключается в получении прогнозного значения *y*, которое определяется путем подстановки в уравнение регрессии соответствующего (прогнозного) значения *x*.

Интервальный прогноз заключается в построении доверительного интервала прогноза.

При построении доверительного интервала прогноза используется *стандартная ошибка прогноза*:

$$m_{\hat{y}_p} = \sigma_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

Строится *доверительный интервал прогноза*:

$$\gamma_{\hat{y}_p} = \hat{y}_p \pm t_{табл} \cdot m_{\hat{y}_p}$$