Регрессионный анализ

Основные понятия и определения

Методы дисперсионного и корреляционного анализа позволяют выявить наличие связи между случайными величинами и оценить силу этой связи.

Регрессионный анализ выявляет конкретный функциональный вид связи между с.в.

Парная (простая) линейная регрессия даёт правила, определяющие линию регрессии, которая лучше других предсказывает наиболее вероятные значения одной переменной на основании другой (переменных всего две).

Множественная регрессия является расширением простой линейной регрессии.

По оси *Y* располагают переменную, которую необходимо предсказать (зависимую), а по оси *X* – переменную, на основе которой будет осуществляться предсказание (независимую).

Регрессионный анализ

Основные понятия и определения

Зависимая переменная – это переменная в регрессии, которую нельзя изменять, её изменение является следствием влияния независимой переменной (переменных).

Независимая переменная – это та переменная в регрессии, которую можно изменять.

Коэффициенты регрессии (β) — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.

Регрессионный анализ

Допущения:

1. Переменные модели должны иметь распределение, близкое к нормальному.

2. Зависимая и независимые переменные должны быть измерены в метрической шкале.

3. Для построения линейных регрессий, зависимая и независимые переменные должны иметь линейную связь.

4. Отсутствие мультиколлинеарности – независимость между собой переменных-предикторов, отсутствие высокой корреляции (для множественной регрессии). Решение: удаление высоко коррелируемых переменных из анализа или центрирование данных (вычитание средних значений из каждого наблюдения по необходимым переменным).

Регрессионный анализ

Допущения:

5. Отсутствие автокорреляции – отсутствие независимости остатков. Выявляется с помощью теста Дурбина-Уотсона (обнаруживает автокорреляцию первого порядка).

– Если d=0 – полная положительная автокорреляция;

– Если d=4 – полная отрицательная автокорреляция;

– Если d=2 – отсутствие автокорреляции.

6. Гомоскедастичность - дисперсия остатков одинакова для каждого значения. Определяется с помощью диаграммы рассеяния.

Регрессионный анализ

В зависимости от вида уравнения регрессии $y = f(x)$ различают линейную ($f(x)$—многочлен первой степени) и нелинейную ($f(x)$ — многочлен степени ≥ 2) регрессии.

Вид функции $f(x)$ выбирается, исходя из особенностей исследуемого явления (процесса), а также из общего графического анализа зависимости между y и x.

Обычно строят линейные регрессионные модели, а при нелинейной зависимости $y = f(x)$ используют различные линеаризующие преобразования переменных y и x.

Регрессионный анализ

Линеаризующие функциональные преобразования
$(y^* = a^* + b^* x^*)$

Исходная зависимость $y = f(x)$	Преобразование переменных		Преобразование коэффициентов	
	y^*	x^*	a^*	b^*
$y = a + \frac{b}{x}$	y	$\frac{1}{x}$	a	b
$y = \frac{a}{b + x}$	$\frac{1}{y}$	x	$\frac{a}{b}$	$\frac{1}{a}$
$y = \frac{ax}{b + x}$	$\frac{1}{y}$	$\frac{1}{x}$	$\frac{b}{a}$	$\frac{1}{a}$
$y = \frac{x}{a + bx}$	$\frac{x}{y}$	x	a	b
$y = ab^x$	$\lg y$	x	$\lg a$	$\lg b$
$y = ax^b$	$\lg y$	$\lg x$	$\lg a$	b
$y = ae^{bx}$	$\ln y$	x	$\ln a$	b
$y = ae^{\frac{b}{x}}$	$\ln y$	1	$\ln a$	b
$y = a + bx^n$	y	x^n	a	b

Регрессионный анализ

Линейный регрессионный анализ

Линейный регрессионный анализ исходит из наличия зависимости

$$y = \alpha + \beta x,$$

где α и β — неизвестные коэффициенты регрессии. Выборочные оценки α и β обозначим *a* и *b* соответственно,

a – константа, определяет точку пересечения прямой с осью *Y*. Как правило, не интерпретируется.

b – угловой коэффициент, характеризует наклон прямой (slope), *b* показывает, на какую величину в среднем изменится результативный признак *y*, если переменная *x* увеличится на единицу своего измерения;

Y – зависимая переменная, X – независимая переменная.

Коэффициент эластичности (Э) показывает, на сколько процентов в среднем изменится *y* при изменении *x* на 1%. Для линейной регрессии: $Э = b\frac{\tilde{x}}{\tilde{y}}$

Регрессионный анализ

В основе регрессионного анализа лежит принцип наименьших квадратов: в качестве уравнения регрессии $y = f(x)$ выбирается функция, доставляющая минимум сумме квадратов разностей

$$s = \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

Количественной мерой рассеяния значений y_i вокруг регрессии $f(x)$ является дисперсия

$$D = \frac{1}{n-k} \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

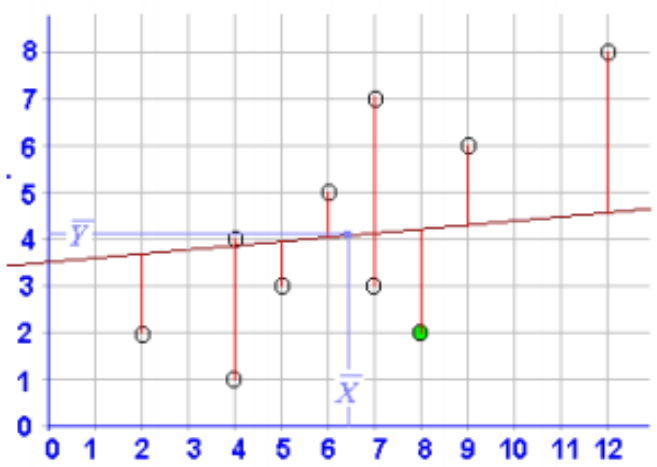
где k — число коэффициентов, входящих в аналитическое выражение регрессии (например, если $f(x)$ — многочлен степени s, то $k = s + 1$).

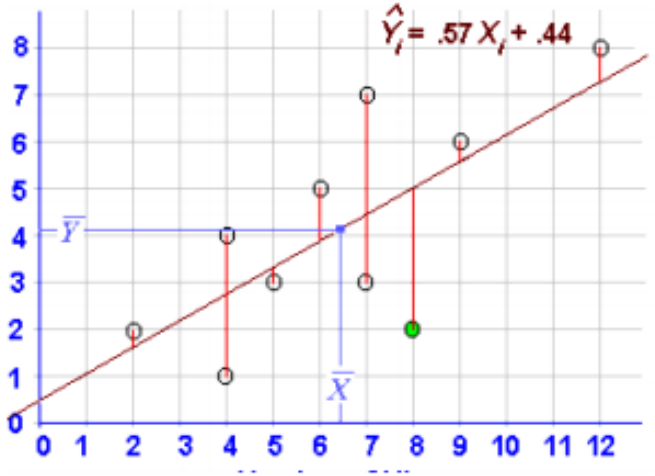
Регрессионный анализ

Используют **метод наименьших квадратов** – подбирают такую линию регрессии чтобы общая сумма квадратов отклонений (Residuals) значений зависимой переменной была наименьшей.

$$\sum e_i = 0$$

$$\sum e_i^2 \quad \text{- минимальна}$$





Регрессионный анализ

Оценка коэффициентов регрессии

Пусть x_i, y_i – наблюдаемые значения с.в. x и y

$$\begin{cases} \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} (\alpha + \beta x_i) = 0; \\ \sum\limits_{i=1}^{n} y_i x_i - \sum\limits_{i=1}^{n} (\alpha + \beta x_i) x_i = 0, \end{cases}$$

$$\begin{cases} n\alpha + \beta \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i; \\ \alpha \sum\limits_{i=1}^{n} x_i + \beta \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i x_i. \end{cases}$$

Решение

$$b = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2}; \qquad a = \frac{\sum\limits_{i=1}^{n} y_i - b \sum\limits_{i=1}^{n} x_i}{n}.$$

Регрессионный анализ

Проверка правильности вычислений

$$\bar{y} = a + b\bar{x}, \quad \text{где} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i; \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Если интервалы между значениями независимой переменной *x* постоянны, $x_{i+1} - x_i = l = \text{const}$

$$\tilde{x}_i = \frac{x_i - x_1}{l} + 1$$

т.е. $x_1, x_2, \ldots, x_n$ преобразуется в $i = 1, 2, \ldots, n$

$y = a + bx$ преобразуется в $y = \tilde{a} + \tilde{b}i.$

$$\tilde{b} = \frac{12\sum_{i=1}^{n} i y_i - 6(n+1)\sum_{i=1}^{n} y_i}{n(n^2-1)}; \quad \tilde{a} = \bar{y} - \tilde{b}\bar{x} = \frac{\sum_{i=1}^{n} y_i - \tilde{b}\sum_{i=1}^{n} i}{n}.$$

Регрессионный анализ

Важно! Регрессия *y* по *x*: $y = \alpha + \beta x$ не эквивалентна в общем случае регрессии *x* по *y*: $-x = \alpha^* + \beta^* y$

Если S_y и S_x — С.К.О. совокупностей значений *y* и *x* соответственно

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2;\ S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

то регрессии $y = f(x)$ и $x = \varphi(y)$

$$y = \bar{y} + r\frac{S_y}{S_x}(x - \bar{x});\quad x = \bar{x} + r\frac{S_x}{S_y}(y - \bar{y}),$$

где *r* — коэффициент корреляции.

Регрессии *y* по *x* и *x* по *y* совпадают только в одном случае, когда существует абсолютная корреляция между *y* и *x*, т.е. когда $|r| = 1$. При $r = 0$ прямые регрессии *y* по *x* и *x* по *y* перпендикулярны. Тогда

$$\beta = r\frac{S_y}{S_x};\ \beta^* = r\frac{S_x}{S_y}.$$

При $S_y = S_x$ коэффициенты корреляции и регрессии совпадают.

Регрессионный анализ

Пример. В результате наблюдений за зависимостью $y = f(x)$ получены следующие данные:

y_i	2	3	7	10	11	13	18	21	25	31
x_i	8	11	14	18	4	26	31	32	34	41

Найти оценку коэффициентов регрессии *y* по x методом наименьших квадратов.

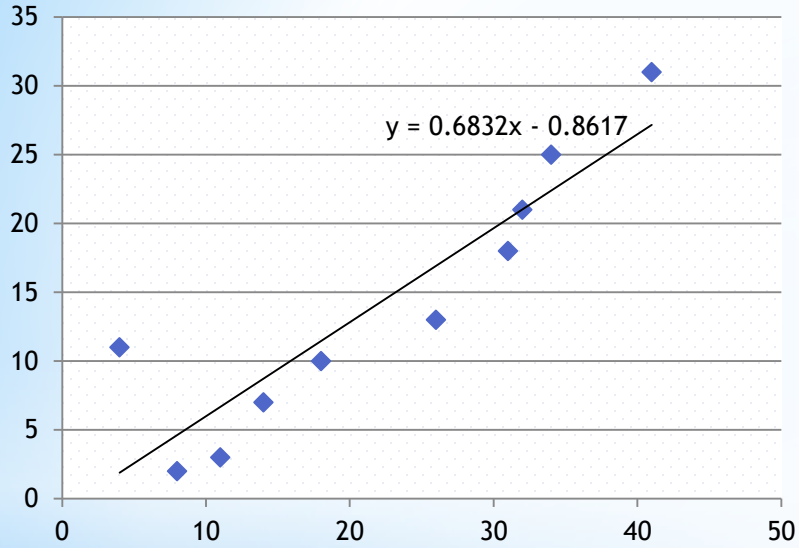
$$\sum_{i=1}^{10} x_i = 219;\ \sum_{i=1}^{10} y_i = 141;\ \sum_{i=1}^{10} x_i^2 = 6219;\ \sum_{i=1}^{10} x_i y_i = 4060.$$

$$b = \frac{10 \cdot 4060 - 219 \cdot 141}{10 \cdot 6219 - 219^2} = 0{,}68318; \quad a = \frac{141 - 0{,}68318 \cdot 219}{10} = -0{,}86164.$$

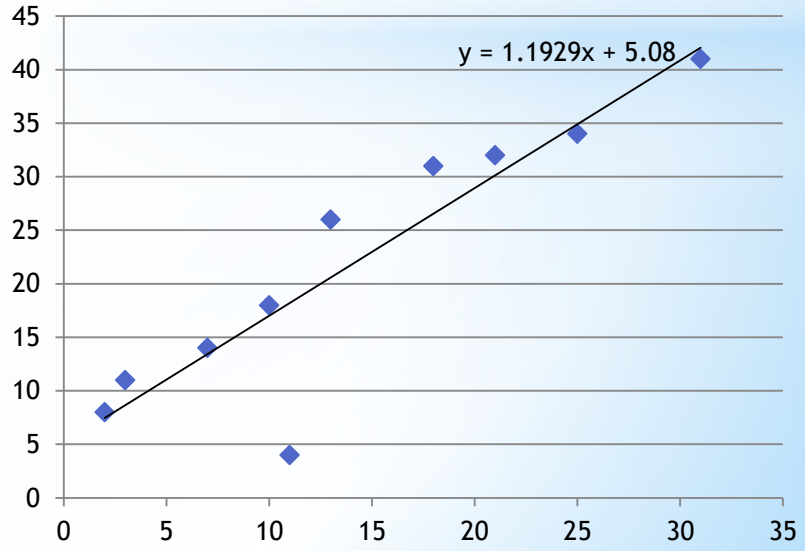
Уравнение регрессии *y* по x

$$y = -0{,}86164 + 0{,}68318 \cdot x.$$

Регрессионный анализ



Уравнение регрессии y по x



Уравнение регрессии x по y

Регрессионный анализ

Пример. В результате наблюдений за зависимостью $y = f(x)$ получены следующие данные:

y_i	13	18	24	21	25	31	36	41	35	41	48	56	61	60	70
x_i	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30

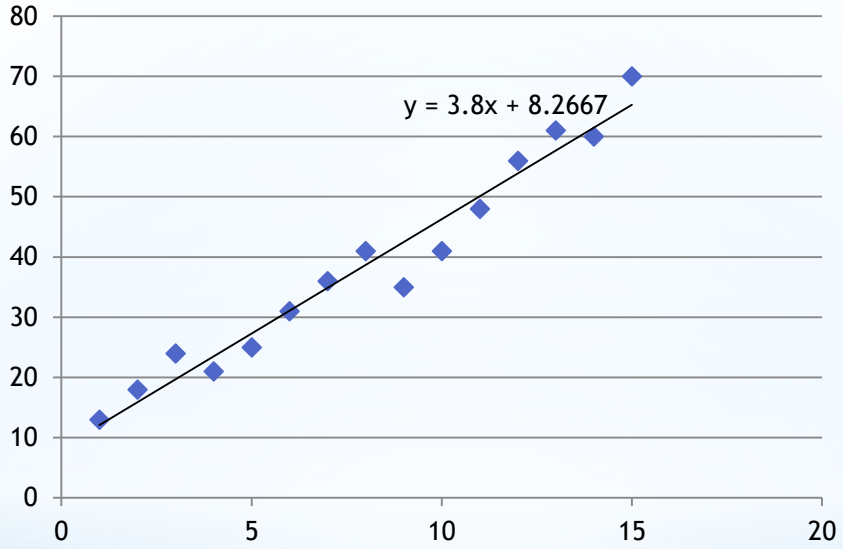
Найти оценку коэффициентов регрессии y по x методом наименьших квадратов.

$$\tilde{x}_i = \frac{x_i - x_1}{l} + 1 = \frac{x_i - 2}{2} + 1.$$

$$\tilde{b} = \frac{12 \cdot \sum\limits_{i=1}^{15} i \cdot y_i - 6 \cdot (15+1) \cdot \sum\limits_{i=1}^{15} y_i}{15 \cdot (225-1)} = \frac{12 \cdot 5704 - 6 \cdot 16 \cdot 580}{15 \cdot 224} = 3{,}8;$$

$$\tilde{a} = \frac{580}{15} - \frac{3{,}8 \cdot (15+1)}{2} = 8{,}266.$$

Регрессионный анализ



Связь коэффициента корреляции и уравнения регрессии

$$b_{f(x)} = r_{xy}\frac{\sigma_y}{\sigma_x};\ b_{\varphi(y)} = r_{yx}\frac{\sigma_x}{\sigma_y}$$

Регрессионный анализ

Простейшие оценки коэффициентов регрессии

Метод Барлетта-Кенуя

Пары наблюдений (y_i, x_i) упорядочиваются по x и разбиваются на 3 примерно равные группы (первая и последняя должны быть равного объема), в каждой группе находятся суммы по x и по y: Y_1, Y_2, Y_3 и X_1, X_2, X_3. Тогда

$$\tilde{b} = \frac{Y_3 - Y_1}{X_3 - X_1} \quad \text{с ошибкой} \quad S_{\tilde{b}} = \frac{0{,}8s\sqrt{n}}{X_3 - X_1},$$

где

$$s = \frac{8}{9}\sum_{i=1}^{n}\frac{|y_i - y_{i+1}|}{n}; \; \tilde{a} = y - \tilde{b}\bar{x}.$$

Если n пар наблюдений разбить на 4 группы, содержащие 1/6, 1/3, 1/3 и 1/6 часть наблюдения, то

Регрессионный анализ

Простейшие оценки коэффициентов регрессии

Метод Барлетта-Кенуя

$$\tilde{b} = \frac{3Y_1 + Y_2 - Y_3 - 3Y_4}{3X_1 + X_2 - X_3 - 3X_4} \quad \text{с ошибкой} \quad S_{\tilde{b}} = \frac{1{,}9s\sqrt{n}}{3X_1 + X_2 - X_3 - 3X_4}.$$

Эти оценки применимы для больших выборок при $n \geq 200$.

Метод Керрича

Для частного случая зависимости $y = bx$ ($a = 0$)

$$d_i = \lg y_i - \lg x_i; \quad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{и} \quad S_{\bar{d}} = \left\{ \frac{1}{n(n-1)} \sum_{i=1}^{n} (d_i - \bar{d})^2 \right\}^{\frac{1}{2}}.$$

Когда $S_{\bar{d}} / \bar{d} << 1$, оценкой lg b является величина d. Тогда оценка будет равна $\bar{b} = 10^{\bar{d}}$.

Регрессионный анализ

Простейшие оценки коэффициентов регрессии

Пример.

. *Для совокупности значений*

x_i:	1	2	3	4	5	6	7	8	9	10;
y_i:	3	8	6	16	15	18	21	29	28	32

найти оценку коэффициента регрессии методом Керрича.

Вычисляем последовательность значений $d_i = \lg y_i - \lg x_i$:

d_i: 0,477; 0,602; 0,301; 0,602; 0,477; 0,477; 0,559; 0,559; 0,493; 0,505.

Далее $\tilde{d} = 0{,}497$ и $\tilde{b} = 10^{0{,}497} = 3{,}14$, что близко к обычной оценке $\tilde{b} = 3{,}23$.

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

Регрессионный анализ

Статистическое оценивание регрессии

Оценка значимости уравнения регрессии в целом дается с помощью *F-критерия Фишера*. Согласно *F*-критерию Фишера, выдвигается «нулевая» гипотеза H_0 о статистической незначимости уравнения регрессии и показателя тесноты связи.

Статистика *F*-критерия

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

где r – коэффициент корреляции, m – число независимых переменных в уравнении регрессии (для парной регрессии $m = 1$); n – число единиц совокупности.

$F_{табл} = F(\alpha, f_1, f_2)$, где $f_1 = m$, $f_2 = n - m - 1$

Регрессионный анализ

Статистическое оценивание регрессии

Если нулевая *гипотеза* H_0 *справедлива*, то факторная и остаточная дисперсии не отличаются друг от друга (т. е. отличие величины F от нуля статистически незначимо).

Если нулевая *гипотеза* H_0 ***не*** *справедлива*, то факторная дисперсия превышает остаточную в несколько раз.

Табличное значение F-критерия – это максимальная величина отношения дисперсий, которая может иметь место при случайном расхождении их для данного уровня вероятности наличия нулевой гипотезы.

Если $F_{факт} > F_{табл}$, то H_0 о случайной природе связи отклоняется и признается статистическая значимость и надежность уравнения.

Если $F_{факт} < F_{табл}$, то H_0 не отклоняется и признается статистическая незначимость уравнения регрессии.

Регрессионный анализ

Статистика t-критерия

$$t_\beta = \frac{b - \beta}{S_\beta}$$

где

$$S_\beta = \frac{S}{S_x\sqrt{n-1}}; \quad S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2;$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i;$$

β - истинное значение коэффициента регрессии, b – выборочная оценка коэффициента регрессии.

Статистика t_β при справедливости нулевой гипотезы H_0: $\beta = b$ имеет распределение Стьюдента с $f = n - 2$ степенями свободы.

Регрессионный анализ

Значение коэффициента β регрессии является значимым с достоверностью α, если $|b| > t_{\frac{1+\alpha}{2}} S_\beta$

Гипотеза о равенстве коэффициента β заданному значению β_0 принимается, если $|\beta - b| > t_{\frac{1+\alpha}{2}} S_\beta$

Двусторонний $\alpha \cdot 100\%$-й доверительный интервал для β

$$b - t_{\frac{1+\alpha}{2}} S_\beta \le \beta \le b + t_{\frac{1+\alpha}{2}} S_\beta$$

Статистические выводы относительно коэффициента α могут быть получены с помощью статистики

$$t_\alpha = \frac{a - \alpha}{S_\alpha}, \quad \text{где} \quad S_\alpha = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\,S_x^2}},$$

Регрессионный анализ

Здесь a, α — соответственно выборочная оценка и истинное значение коэффициента α.

При H_0: a = α статистика t_α имеет распределение Стьюдента с $f = n - 2$ степенями свободы. Проверка гипотез о значениях коэффициента *a* и построение доверительных интервалов для него выполняются по аналогии с коэффициентом β.

Пример. Для совокупности данных

x_i	1,2	2,4	2,8	4,2	5,9	6,8	8,1	9,2	10,1	11
Y_i	7	12	17	24	29	38	46	45	54	68

найти оценки коэффициентов α и β регрессии $y = a + \beta x$ и провести их статистический анализ при доверительной вероятности α = 0,95

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

Регрессионный анализ

$$\sum_{i=1}^{10} x_i = 61{,}7; \quad \left(\sum_{i=1}^{10} x_i\right)^2 = 3806{,}89;$$

$$\sum_{i=1}^{10} x_i^2 = 486{,}99; \quad \sum_{i=1}^{10} y_i = 340; \quad \sum_{i=1}^{10} x_i y_i = 2695{,}1.$$

$$b = \frac{10 \cdot 2695{,}1 - 61{,}7 \cdot 340}{10 \cdot 486{,}99 - 3806{,}89} = 5{,}6189; \quad a = \frac{\sum_{i=1}^{10} y_i - b \cdot \sum_{i=1}^{10} x_i}{n} = \frac{340 - 5{,}6189 \cdot 61{,}7}{10} = -0{,}668.$$

Проверим теперь значимость полученных коэффициентов (существенность их отклонения от нуля).

$$\bar{x} = 6{,}17; \quad S_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = 11{,}8112; \quad (S_x = 3{,}3467)$$

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \text{ где } \hat{y}_i = a + b \cdot x_i$$

$$\hat{y}_i \to 6{,}075;\ 12{,}818;\ 15{,}065;\ 22{,}932;\ 32{,}484;\ 37{,}541;\ 44{,}846;\ 51{,}027;\ 56{,}084;\ 61{,}141.$$

Регрессионный анализ

$$S^2 = \frac{1}{8} \cdot \sum_{i=1}^{10} (y_i - \hat{y}_i)^2 =$$

$$= 0{,}125 \cdot \left[(7 - 6{,}075)^2 + (12 - 12{,}818)^2 + \ldots + (68 - 61{,}141)^2 \right] = 13{,}4755;$$

$$S_\beta = \frac{S}{S_x \cdot \sqrt{n-1}} = \frac{\sqrt{13{,}4755}}{3{,}3467 \cdot 3} = 0{,}3656;$$

$$S_\alpha = S \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1) \cdot S_x^2}} = 3{,}671 \cdot \sqrt{\frac{1}{10} + \frac{6{,}17^2}{9 \cdot 11{,}8112}} = 2{,}486.$$

$$t_{\frac{1+0{,}95}{2}}(n-2) = t_{0{,}975}(8) = 2{,}306.$$

$$|b| = 5{,}619 > t_{0{,}975}(8) \cdot S_\beta = 2{,}306 \cdot 0{,}3656 = 0{,}843$$

следовательно, с достоверностью 0,95 делаем вывод о значимости коэффициента регрессии.

Проверяем гипотезу *H*: $\beta = \beta_0 = 5$ (о равенстве коэффициента регрессии $\beta_0 = 5$):

$$|5{,}619 - 5| = 0{,}619 < t_{0{,}975}(8) \cdot S_\beta = 0{,}843,$$

т. е. гипотеза о равенстве $\beta = 5$ не отклоняется.

Регрессионный анализ

Доверительный интервал для β равен

$$5{,}9 - 2{,}306 \cdot 0{,}3656 = 4{,}776 \leqslant \beta \leqslant 6{,}462 = 5{,}619 + 2{,}306 \cdot 0{,}3656.$$

Аналогично для коэффициента α

$$H_0 : \alpha = 0 \, |a| = 0{,}668 < t_{0{,}975} \cdot S_\alpha = 2{,}306 \cdot 2{,}485 = 5{,}73.$$

Следовательно, коэффициент α с вероятностью 0,95 не отличается значимо от нуля, т. е. его значение может быть приравнено к нулю.

Двусторонний доверительный интервал для α имеет вид

$$\hat{a} - t_{0{,}975} \cdot S_\alpha \leqslant a \leqslant \hat{a} + t_{0{,}975} \cdot S_\alpha;$$
$$-0{,}668 - 2{,}306 \cdot 2{,}485 = -6{,}398 \leqslant a \leqslant 5{,}602 = -0{,}668 + 2{,}306 \cdot 2{,}485.$$

Таким образом, уравнение регрессии *y* по *x* адекватно отображается уравнением

$$y = 5{,}619 \cdot x$$

Регрессионный анализ

Замечания.

1. Любая регрессионная модель позволяет обнаружить только *количественные* зависимости, которые не обязательно отражают причинные зависимости, т.е. влияние одного фактора на другой.

2. Гипотезы о причинной связи признаков должны дополнительно обосновываться с помощью теоретического анализа, содержательно объясняющего изучаемое явление или процесс.