ПРАКТИЧЕСКОЕ ЗАНЯТИЕ 14.

ЛИНЕЙНЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ.

Пусть (X_i, Y_i), $i = 1, 2, \ldots n$ – выборка объема n из наблюдений случайного двумерного вектора $(X, Y)^T$. Определим оценки числовых характеристик этого вектора. За оценку математических ожиданий a_x и a_y принимаются средние арифметические $\overline{X}$ и $\overline{Y}$ (1.3), за оценку дисперсий σ_x^2 и σ_y^2 – соответствующие эмпирические дисперсии S_x^2 и S_y^2, вычисленные по формуле (1.5). Несмещенной оценкой ковариации K_{xy} является величина:

$$\tilde{K}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}). \qquad (1.92)$$

Для практических расчетов формулу (1.84) удобно преобразовать к виду:

$$\tilde{K}_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}\right). \qquad (1.93)$$

Расчет упрощается, если, как и при нахождении оценок параметров одномерной случайной величины, ввести линейную замену (1.8):

$$X_i = C_1 + h_1 U_i;\ Y_i = C_2 + h_2 V_i. \qquad (1.94)$$

При такой замене формула (1.93) принимает вид

$$\tilde{K}_{xy} = \frac{h_1 h_2}{n-1}\left(\sum_{i=1}^{n} U_i V_i - n\overline{U}\,\overline{V}\right). \qquad (1.95)$$

Оценку коэффициента корреляции ρ_{xy} находят по формуле

$$\rho_{xy} \approx r = \frac{\tilde{K}_{xy}}{S_x S_y}. \qquad (1.96)$$

Уравнения оценочных (выборочных) прямых регрессии получают по следующим формулам.

Уравнение линейной регрессии *Y* на *X*:

$$\frac{y - \overline{Y}}{S_y} = r\frac{x - \overline{X}}{S_x}. \qquad (1.97)$$

Уравнение линейной регрессии *X* на *Y*:

$$\frac{y - \overline{Y}}{S_y} = \frac{1}{r}\frac{x - \overline{X}}{S_x}. \qquad (1.98)$$

Задача 1.28. В первом столбце табл. 1.13 записаны измеренные значения величины X – изменения содержания азота в стали при выпуске из конвертера по сравнению с начальным содержанием) [10^{-4} % ·]; во втором – величины Y (значения начальной концентрации углерода в этой же стали [%]). Найти оценку коэффициента корреляции по этой двумерной выборке. Вычислить выборочные параметры линейной регрессии Y на X и X на Y.

Таблица 1.13

Исходные данные и результаты расчетов к задаче 1.28.

№	X	Y	U	V	U^2	V^2	UV
1	–2,0	0,11	–4	1	16	1	–4
2	0,5	0,09	1	–1	1	1	–1
3	–1,5	0,13	–3	3	9	9	–9
4	–5,5	0,11	–11	1	121	1	–11
5	3,5	0,06	7	–4	49	16	–28
6	–1,0	0,12	–2	2	4	4	–4
7	2,0	0,08	4	–2	16	4	–8
8	0,0	0,11	0	1	0	1	0
9	1,5	0,07	3	–3	9	9	–9
Σ	–	–	–5	–2	225	46	–74

Решение

Вводим линейную замену (1.94), выбирая $C_1 = 0$, $h_1 = 0{,}5$; $C_2 = 0{,}10$, $h_2 = 10^{-2}$. Вычисляем оценки математических ожиданий (1.9):

$$\overline{U} = -\frac{5}{9} \approx 0,556; \quad \overline{X} = -0,5 \cdot 0,556 = -0,278;$$

$$\overline{V} = -\frac{2}{9} \approx -0,22; \quad \overline{Y} = 0,10 - 0,22 \cdot 10^{-2} = 0,0978.$$

Несмещенные оценки дисперсий находим по формуле (1.10):

$$S_x^2 = \frac{(0,5)^2}{8}\left(225 - 9\left(-\frac{5}{9}\right)^2\right) \approx 6,94; \quad S_x \approx 2,63;$$

$$S_y^2 = \frac{10^{-4}}{8}\left(46 - 9\left(-\frac{2}{9}\right)^2\right) \approx 5,69 \cdot 10^{-4}; \quad S_y = 2,39 \cdot 10^{-2}.$$

Расчет оценки ковариации проводим по формуле (1.95):

$$\tilde{K}_{xy} = \frac{0{,}5 \cdot 10^{-2}}{8}\left(-74 - 9\left(-\frac{5}{9}\right)\left(-\frac{2}{9}\right)\right) \approx -4{,}69 \cdot 10^{-2}.$$

Оценку коэффициента корреляции находим по формуле (1.96)

$$r = \frac{-4{,}69 \cdot 10^{-2}}{2{,}69 \cdot 2{,}39 \cdot 10^{-2}} \approx -0{,}746.$$

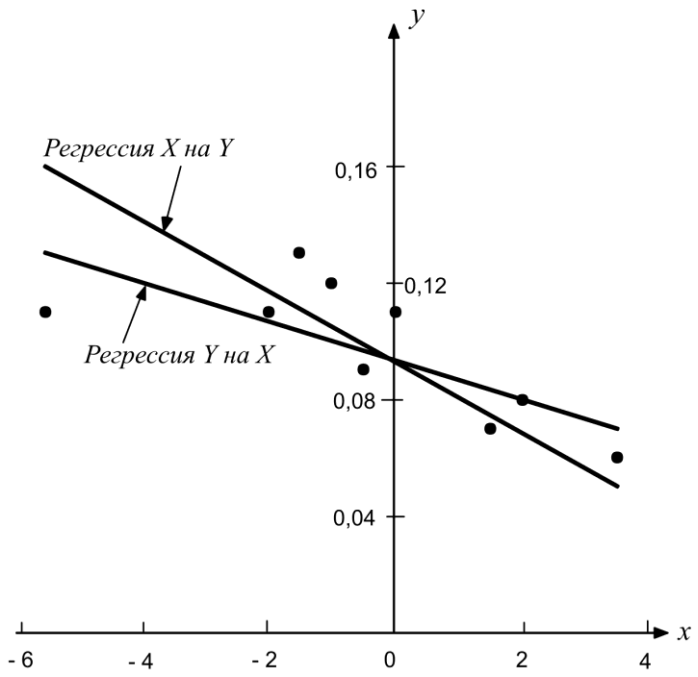

Рис. 1.5. Зависимость изменения концентрации азота в стали (y) при выпуске из конвертера от начальной концентрации углерода (x)

Выборочное уравнение линейной регрессии Y на X:

$$\frac{y - 0{,}0978}{2{,}39 \cdot 10^{-2}} = -0{,}746 \cdot \frac{x + 0{,}278}{2{,}63}$$

или $$y - 0{,}0978 = -0{,}00678\left(x + 0{,}278\right).$$

Выборочное уравнение линейной регрессии X на Y:

$$\frac{y - 0{,}0978}{2{,}39 \cdot 10^{-2}} = -\frac{1}{0{,}746} \cdot \frac{x + 0{,}278}{2{,}63}$$

или $$y - 0{,}0978 = -0{,}0122\left(x + 0{,}278\right).$$

Прямые регрессии представлены на рис. 1.5, там же приведены экспериментальные точки.

Будем предполагать, что заданная двумерная выборка имеет двумерное нормальное распределение. Тогда доверительный интервал для коэффициента корреляции можно найти по номограммам. В Приложении приведены такие номограммы (рис.П1) для доверительной вероятности $\mathcal{P}$ = 0,95 и $\mathcal{P}$ = 0,99. По горизонтальной оси номограммы отложены значения выборочного коэффициента корреляции *r*, по вертикальной оси – значения истинного коэффициента корреляции ρ_{xy}, числа над кривыми указывают объемы выборок *n*. Отложив на горизонтальной оси вычисленное значение выборочного коэффициента корреляции, следует подняться над этой точкой вертикально вверх и найти две точки пересечения с кривыми, соответствующими объему заданной выборки. Ординаты этих двух точек являются границами доверительного интервала истинного коэффициента корреляции.

Для проверки гипотезы H_0: $\rho_{xy} = 0$ при альтернативной гипотезе H_1: $\rho_{xy} \neq 0$ будем использовать следующий критерий. Гипотеза H_0 принимается с уровнем значимости α, то есть линейная зависимость между величинами не существует, если $|r| < T$, где T – значение критерия:

$$T = \frac{t_{1-\alpha/2}(n-2)}{\sqrt{n-2+t^2_{1-\alpha/2}(n-2)}}, \qquad (1.99)$$

в противном случае принимается гипотеза H_1, то есть предполагается, что линейная зависимость между величинами существует. $t_{1-\alpha/2}(n-2)$ – квантиль распределения Стьюдента с числом степеней свободы $k = n-2$.

Задача 1.29. Найти доверительный интервал для коэффициента корреляции ($\mathcal{P}$ = 0,95) и проверить гипотезу об отсутствии линейной зависимости с уровнем значимости α = 0.05 между величинами *X* и *Y* для данных задачи 1.28.

Решение.

По номограммам Приложения для значения *r* = – 0,746; *n* = 9 находим: $-0,95 < \rho < -0,14$. Так как значение ρ = 0 не принадлежит найденному доверительному интервалу, гипотеза о существовании линейной зависимости не противоречит экспериментальным данным с уровнем значимости α = 0,05.

Проверим гипотезу об отсутствии линейной зависимости между величинами *X* и *Y* с помощью критерия *T* (1.99). По таблице квантилей распределения Стьюдента находим $t_{0,975}(7) = 2,365$.

$$T = \frac{t_{1-\alpha/2}(n-2)}{\sqrt{n-2+t^2_{1-\alpha/2}(n-2)}} = \frac{2,365}{\sqrt{7+2,365^2}} = 0,666.$$

Так как $|r| = 0{,}746 > 0{,}666$, принимаем гипотезу о существовании линейной зависимости между величинами *X* и *Y*.

Полученные результаты позволяют сделать вывод, что с увеличением одной из величин среднее значение другой величины уменьшается. Так как коэффициент корреляции значим, можно пользоваться уравнениями выборочных прямых регрессии для предсказания среднего значения одной переменной по значению другой. При этом следует иметь в виду, что предсказанное значение переменной может иметь значительную погрешность. Это следует из очень широкого доверительного интервала для ρ_{xy}. Для получения более точных результатов надо иметь большее число экспериментов.