Типовой расчет 5.

Обработка данных методами линейного корреляционного анализа.

5.1. ТЕОРЕТИЧЕСКОЕ ВВЕДЕНИЕ.

5.1.1. Двумерный случайный вектор. Линейная корреляция.

Рассмотрим систему двух случайных величин или *двумерный случайный вектор* $(X, Y)^T$ с центром распределения $\begin{pmatrix} M(X) \\ M(Y) \end{pmatrix} = \begin{pmatrix} a_x \\ a_y \end{pmatrix}$ и ковариационной матрицей

$$K = \begin{pmatrix} D(X) & K_{xy} \\ K_{xy} & D(Y) \end{pmatrix}, \tag{5.1}$$

где a_x и a_y – математические ожидания; $D(X) = \sigma_x^2$ и $D(Y) = \sigma_y^2$ – дисперсии случайных величин X и Y соответственно; K_{xy} – ковариация между величинами X и Y, определяется следующим образом:

$$K_{xy} = \operatorname{cov}(X, Y) = M[(X - a_x)(Y - a_y)]. \tag{5.2}$$

В качестве нормированной ковариации вводится *коэффициент корреляции*:

$$\rho_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}, \tag{5.3}$$

который характеризует степень *линейной зависимости* между случайными величинами X и Y.

Свойства коэффициента корреляции следующие.

1. Коэффициент корреляции является безразмерным коэффициентом, не зависящим от начала отсчета величин X и Y.

2. Коэффициент корреляции по абсолютной величине не превышает единицу: $-1 \le \rho_{xy} \le 1$.

3. Если $|\rho_{xy}| = 1$, случайные величины X и Y *связаны линейной функциональной зависимостью.*

4. Если $\rho_{xy} = 0$, случайные величины X и Y *некоррелированы*, т.е. между ними *отсутствует линейная зависимость*.

5. Чем ближе значение $|\rho_{xy}|$ к единице, тем сильнее линейная зависимость между *X* и *Y* . Чем ближе значение $|\rho_{xy}|$ к нулю, тем слабее линейная зависимость между *X* и *Y*.

6. Если $\rho_{xy} > 0$, то с увеличением одной случайной величины математическое ожидание (среднее значение) другой увеличивается; если $\rho_{xy} < 0$, то с увеличением одной случайной величины математическое ожидание (среднее значение) другой уменьшается.

Для случайного вектора $(X, Y)^T$ вводятся *условные математические ожидания* $M(X/Y=y)$ и $M(Y/X=x)$. $M(X/Y=y)$ – это математическое ожидание случайной величины *X* при условии, что *Y* приняло одно из своих возможных значений *y*. Аналогично, $M(Y/X=x)$ – это математическое ожидание случайной величины *Y* при условии, что *X* приняло одно из своих возможных значений *x*.

Функцией регрессии Y на *X* называется зависимость величины $M(Y/X=x)$ от аргумента *x*. Она характеризует зависимость математического ожидания величины *Y* от значения, принимаемого величиной *X*. Аналогично *функцией регрессии X* на *Y* называется зависимость величины $M(X/Y=y)$ от аргумента *y*. Она характеризует зависимость математического ожидания величины *X* от значения, принимаемого величиной *Y*. Если обе функции регрессии *Y* на *X* и *X* на *Y* являются линейными, *корреляционная зависимость* между случайными величинами *X* и *Y* называется *линейной*. В случае линейной корреляционной зависимости уравнения регрессии – *Y* на *X* и *X* на *Y* – называются *уравнениями линейной регрессии*.

Уравнение линейной регрессии *Y* на *X* имеет вид

$$y = a_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x - a_x), \tag{5.4}$$

а уравнение линейной регрессии *X* на *Y* –

$$y = a_y + \frac{1}{\rho_{xy}}\frac{\sigma_y}{\sigma_x}(x - a_x). \tag{5.5}$$

5.1.2. Выборочные характеристики двумерного случайного вектора.

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 $\overline{Y}$ (см. формулу (3.2)), за оценку дисперсий σ_x^2 и σ_y^2 – соответствующие эмпирические

дисперсии S_x^2 и S_y^2, вычисленные по формуле (3.3). Здесь и далее ссылки на формулы с первой цифрой 3 даются на текст типового расчета 10.3.

Несмещенной оценкой ковариации K_{xy} является величина

$$\tilde{K}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}). \qquad (5.6)$$

Для практических расчетов формулу (5.6) удобно преобразовать к виду:

$$\tilde{K}_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}\right). \qquad (5.7)$$

Расчет упрощается, если, как и при нахождении оценок параметров одномерной случайной величины, ввести линейную замену (3.6):

$$X_i = C_1 + h_1 U_i; \quad Y_i = C_2 + h_2 V_i. \qquad (5.8)$$

При такой замене формула (5.7) принимает вид

$$\tilde{K}_{xy} = \frac{h_1 h_2}{n-1}\left(\sum_{i=1}^{n} U_i V_i - n\overline{U}\,\overline{V}\right). \qquad (5.9)$$

Оценку коэффициента корреляции ρ_{xy} находят по формуле

$$\rho_{xy} \approx r = \frac{\tilde{K}_{xy}}{S_x S_y}. \qquad (5.10)$$

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

Уравнение линейной регрессии *Y* на *X*:

$$\frac{y - \overline{Y}}{S_y} = r\frac{x - \overline{X}}{S_x}. \qquad (5.11)$$

Уравнение линейной регрессии *X* на *Y*:

$$\frac{y - \overline{Y}}{S_y} = \frac{1}{r}\frac{x - \overline{X}}{S_x}. \qquad (5.12)$$

Выборочные уравнения прямых регрессии используют для предсказания среднего значения одной переменной по значению другой.

5.1.3. Построение доверительного интервала для коэффициента корреляции. Проверка гипотезы о существовании линейной зависимости.

Будем предполагать, что заданная двумерная выборка имеет двумерное нормальное распределение. Тогда доверительный интервал для коэффициента корреляции можно найти по номограммам. В Приложении (см. [1])cvприведены такие номограммы (рис. П1) для доверительной вероятности $P = 0{,}95$. 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 по 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.

Эти же графики можно использовать для проверки гипотезы H_0 об отсутствии линейной зависимости между величинами X и Y, т.е. о том, что истинный коэффициент корреляции $\rho_{xy} = 0$ при альтернативной гипотезе H_1: $\rho_{xy} \neq 0$. Гипотеза H_0 принимается, т.е. линейная зависимость между величинами не существует (с уровнем значимости $\alpha = 1 - P$), если значение $\rho_{xy} = 0$ принадлежит найденному доверительному интервалу. Здесь P – доверительная вероятность при определении доверительного интервала. Гипотеза H_0 отвергается, т.е. принимается альтернативная гипотеза H_1 (линейная зависимость между величинами существует), если значение $\rho_{xy} = 0$ не принадлежит найденному доверительному интервалу.

Для проверки гипотезы H_0: $\rho_{xy} = 0$ при альтернативной гипотезе H_1: $\rho_{xy} \neq 0$ можно использовать другой критерий. Гипотеза H_0 принимается с уровнем значимости α, т.е. линейная зависимость между величинами не существует, если

$$|r| < \frac{t_{1-\alpha/2}(n-2)}{\sqrt{n-2+t^2_{1-\alpha/2}(n-2)}}, \qquad (5.13)$$

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

Если принята гипотеза о существовании линейной зависимости между случайными величинами, то, зная доверительный интервал для коэффициента корреляции, можно сделать вывод о силе взаимосвязи между X и Y. Если доверительный интервал примыкает

к единице или минус единице, то говорят, что связь сильная. Если доверительный интервал примыкает к нулю, то говорят, что связь слабая. Если доверительный интервал расположен примерно посередине интервала (–1; 0) или (0; 1), то говорят, что связь средней величины.

5.2. Содержание типового расчета.

Заданы результаты *n* экспериментов, в каждом из которых измерены значения двух величин *x* и *y*, т.е. задана выборка объема *n*, извлеченная из двумерной нормальной генеральной совокупности (*X, Y*). По приведенным исходным данным требуется:

– найти оценки характеристик наблюдаемого двумерного случайного вектора;

– найти оценку коэффициента корреляции;

– записать эмпирические уравнения линейной регрессии;

– проверить гипотезу об отсутствии линейной зависимости между величинами *X* и *Y*;

– сделать вывод о силе и характере связи между величинами *X* и *Y*.

5.3. Порядок выполнения типового расчета. Примеры.

1. Нахождение оценок числовых характеристик двумерного случайного вектора. Расчет оценки коэффициента корреляции.

Необходимо определить оценки числовых характеристик двумерного случайного вектора. За оценку математических ожиданий a_x и a_y принимаются средние арифметические $\overline{X}$ и $\overline{Y}$, рассчитанные по формуле (3.23), за оценку дисперсий σ_x^2 и σ_y^2 – соответствующие эмпирические дисперсии S_x^2 и S_y^2, вычисленные по формуле (3.3). Несмещенная оценка ковариации $\widetilde{K}_{xy}$ определяется по формуле (5.6).

Для упрощения расчетов и последующего контроля правильности вычислений следует провести кодировку данных по формуле (5.8). Оценки определяются по формулам (3.7), (3.8), (3.4), (5.9).

Для контроля правильности вычислений весь расчет необходимо повторить при другом начале отсчета. Результаты этих расчетов должны совпасть с точностью до величины возможных ошибок округления. Если результаты расчетов совпадают, определяется оценка коэффициента корреляции по формуле (5.10).

2. Нахождение уравнений линейной регрессии.

На этом этапе расчетов требуется записать выборочные уравнения линейной регрессии *Y* на *X* и *X* на *Y*. На одном чертеже построить прямые регрессии и нанести все

экспериментальные точки. Выборочные уравнения линейной регрессии записываются в соответствии с формулами (5.11), (5.12).

3. Построение доверительного интервала для коэффициента корреляции ρ. Проверка гипотезы о существовании линейной зависимости между величинами X и Y.

На этом этапе расчетов требуется найти доверительный интервал для коэффициента корреляции и проверить гипотезу об отсутствии линейной зависимости между величинами X и Y. Уровень значимости α при проверке гипотезы задает преподаватель. Доверительная вероятность $P = 1 - \alpha$.

5.4. Пример выполнения типового расчета.

В первом столбце табл. 5.1 содержатся измеренные значения величины X – изменения содержания азота в стали при ее выпуске из конвертера по сравнению с начальным содержанием [$\times 10^{-4}$, %]; во втором столбце – значения 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 5.1.

Номер эксперимента	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 (5.8), выбирая $C_1 = 0$, $h_1 = 0{,}5$; $C_2 = 0{,}10$, $h_2 = 10^{-2}$. Вычисляем оценки математических ожиданий по формуле (3.7):

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

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

Несмещенные оценки дисперсий находим по формуле (3.8):

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

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

Расчет оценки ковариации проводим по формуле (5.9):

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

Оценку коэффициента корреляции находим по формуле (5.10):

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

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

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

Прямые регрессии представлены на рис. 5.1, там же приведены экспериментальные точки.

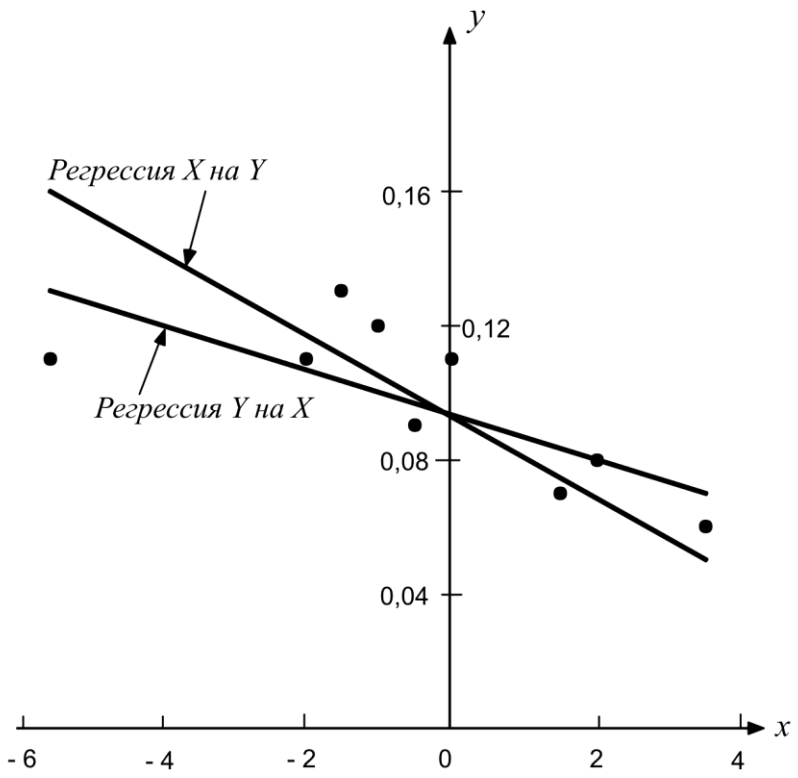

Рис. 5.1. Зависимость изменения концентрации азота в стали (*y*)
при выпуске из конвертера от начальной концентрации углерода (*x*)

По номограммам (см. рис. П1 приложения в [1]) для значения $r = -0{,}746$ ($n = 9$) находим интервал: $-0{,}95 < \rho < -0{,}14$. Так как значение $\rho = 0$ не принадлежит найденному доверительному интервалу, гипотеза о существовании линейной зависимости не противоречит экспериментальным данным с уровнем значимости $\alpha = 0{,}05$.

Проверим гипотезу об отсутствии линейной зависимости между величинами *X* и *Y* с помощью критерия (5.13). По таблице квантилей распределения Стьюдента находим $t_{0{,}975}(7) = 2{,}365$. Вычислим

$$\frac{t_{1-\alpha/2}(n-2)}{\sqrt{n-2+t^2_{1-\alpha/2}(n-2)}} = \frac{2{,}365}{\sqrt{7+2{,}365^2}} = 0{,}667.$$

Так как $|r| = 0{,}746 > 0{,}667$, принимаем гипотезу о существовании линейной зависимости между величинами *X* и *Y*.

Полученные результаты позволяют сделать вывод о том, 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.

5.5. Оформление отчета.

В отчете по типовому расчету должны быть представлены все проведенные расчеты, уравнения выборочных прямых регрессии. На чертеже должны быть представлены уравнения прямых регрессии, там же должны быть проставлены все экспериментальные точки. В выводах сформулировать полученный результат проверки гипотезы о наличии (отсутствии) линейной взаимосвязи между случайными величинами. Если принята гипотеза о наличии линейной взаимосвязи, сделать вывод о силе и характере связи между величинами X и Y.

Точность расчетов оценок математического ожидания – запасной знак по сравнению с исходными данными, оценок дисперсий, средних квадратических отклонений, ковариации – три значащие цифры, оценки коэффициента корреляции – три знака после запятой.