ЛЕКЦИЯ 8. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ЛИНЕЙНЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ.

1. Двумерный случайный вектор, его выборочные характеристики.

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

$$K = \begin{pmatrix} D(X) & K_{xy} \\ K_{xy} & D(Y) \end{pmatrix}, \qquad (1.87)$$

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:

$$K_{xy} = \operatorname{cov}(X,Y) = M[(X - a_x)(Y - a_y)], \qquad (1.88)$$

В качестве нормированной ковариации вводится коэффициент корреляции

$$\rho_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}, \qquad (1.89)$$

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

Свойства коэффициента корреляции:

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

2. $-1 \le \rho_{xy} \le 1$.

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*, то есть между ними *отсутствует линейная зависимость*.

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

Случайный вектор $(U_1, U_2)^T$ имеет *стандартное нормальное распределение*, если его координаты U_1 и U_2 взаимно независимы и имеют стандартное нормальное распределение. Его центр распределения совпадает с началом координат, а матрица ковариаций является единичной матрицей.

Случайный вектор $(X, Y)^T$ имеет *двумерное нормальное распределение*, если его можно представить в виде

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} a_x \\ a_y \end{pmatrix} + B \begin{pmatrix} U_1 \\ U_2 \end{pmatrix},$$

где $(a_x, a_y)^T$ – числовой вектор, $(U_1, U_2)^T$ – случайный вектор, имеющий стандартное нормальное распределение, B – невырожденная матрица второго порядка. Центром распределения вектора $(X, Y)^T$ является вектор $(a_x, a_y)^T$, матрица ковариаций равна $K = B \cdot B^T$.

Для нормального закона распределения выполняется следующее свойство: если компоненты двумерного нормального вектора некоррелированы ($\rho_{xy} = 0$), то они и независимы. Для других типов распределения это утверждение может и не выполняться, то есть из некоррелированности случайных величин, в общем случае, не следует их независимость.

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

$$y = a_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x - a_x), \qquad (1.90)$$

а уравнение линейной регрессии X на Y –

$$y = a_y + \frac{1}{\rho_{xy}}\frac{\sigma_y}{\sigma_x}(x - a_x). \qquad (1.91)$$

Если случайные величины X и Y имеют двумерное нормальное распределение, корреляционная зависимость между ними может быть только линейной. Для других типов распределения корреляционные зависимости могут быть нелинейными.

Пусть (X_i, Y_i), $i = 1, 2, \dots n$ – 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:

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

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

2. 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 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.

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

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

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 $k = n-2$.

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.