

1. Методические указания по выполнению контрольной работы

Простейшая линейная модель регрессии

Оценивание параметров регрессии

Пусть на основании анализа исследуемого явления предполагается, что в среднем y есть линейная функция от x , т. е. имеет место уравнение регрессии

$$\tilde{y} = M\left(\frac{y}{x}\right) = a x + b, \quad \text{где } M\left(\frac{y}{x}\right) \text{ — условное математическое ожидание}$$

случайной величины y при заданном x . Объясняющая переменная x рассматривается как неслучайная величина; a и b - неизвестные параметры генеральной совокупности, которые подлежат оценке по результатам выборочных наблюдений.

Предположим, что для оценки параметров a и b двумерной генеральной совокупности (x, y) взята выборка объемом n , где (x_i, y_i) результат i -го наблюдения ($i = 1, 2, \dots, n$). В этом случае линейная модель регрессии имеет вид $y_i = a x_i + b + \varepsilon_i$, где ε_i - взаимно независимые случайные величины с нулевым математическим ожиданием и дисперсией σ^2 , т. е. $M\varepsilon_i = 0; D\varepsilon_i = M\varepsilon_i^2 = \sigma^2$ для всех $i = 1, 2, \dots, n$ и $M\varepsilon_i \varepsilon_j = 0$ при $i \neq j$ - (условие гомоскедастичности, постоянства остаточной дисперсии;) и равными 0 при $i \neq j$ - (условие взаимной некоррелированности регрессионных остатков.)

Согласно методу наименьших квадратов в качестве оценок неизвестных параметров a и b , следует брать такие значения выборочных характеристик, которые минимизируют сумму квадратов отклонений значений результативного признака y_i от условного математического ожидания $\tilde{y}_i$, т. е.

$$Q = \sum_{i=1}^n (y_i - \tilde{y}_i)^2 = \sum_{i=1}^n (y_i - a x_i - b)^2 = \sum_{i=1}^n \varepsilon_i^2$$

Так как Q дифференцируема по a и b , то для отыскания минимума функции найдем частные производные по a и b :

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^n (y_i - a x_i - b)$$

$$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^n (y_i - a x_i - b) x_i$$

Приравняв производные нулю, получим

$$a n + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Данная система уравнений называется *системой нормальных уравнений*. Решая систему относительно a и b , получим:

$$b = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2};$$

$$a = \frac{1}{n} \sum_{i=1}^n y_i - b * \frac{1}{n} \sum_{i=1}^n x_i$$

Перейдя к средним, будем иметь

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

$$a = \bar{y} - b \bar{x}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i \quad \overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2 \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$S_x^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 \quad S_y^2 = \frac{1}{n} \sum_j (y_j - \bar{y})^2$$

Если двумерное распределение Y и X нормально, то линия регрессии Y по X является прямой с уравнением

$$\bar{y} = m_Y + r \frac{\sigma_Y}{\sigma_X} (x - m_X).$$

Определим коэффициент корреляции

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

Он оценивается по формуле

$$\hat{r} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{S_X \cdot S_Y}$$

Расчетные формулы для общего случая(когда есть повторяющиеся наблюдения

Возможные значения	$y_1 \dots y_j \dots y_m$	Всего
x_1	$n_{11} \dots n_{1j} \dots n_{1m}$	$n_{1\bullet}$
...		...
x_i	$n_{i1} \dots n_{ij} \dots n_{im}$	$n_{i\bullet}$
...		...
x_l	$n_{l1} \dots n_{lj} \dots n_{lm}$	$n_{l\bullet}$
Всего	$n_{\bullet 1} \dots n_{\bullet j} \dots n_{\bullet m}$	n

где n_{ij} - частота, с которой пара (x_i, y_j) встретилась в выборке;

для непрерывных распределений в качестве x_i и y_j берут середины интервалов группировки.

Эта таблица является основой для всех последующих вычислений.

Методы корреляционного анализа дают хорошие результаты в том случае, когда данные эксперимента можно считать выбранными из совокупности, распределённой по многомерному нормальному закону.

$\overline{x \cdot y} = \frac{1}{N} \sum_{i,j} x_i y_j n_{ij}$ – выборочное среднее произведение величин X и Y ,

$\bar{x} = \frac{1}{N} \sum_i x_i n_{i\bullet}$ – выборочное среднее значение случайной величины X ,

$\bar{y} = \frac{1}{N} \sum_j y_j n_{\bullet j}$ – выборочное среднее значение случайной величины Y ,

$$S_x^2 = \frac{1}{N} \sum_i (x_i - \bar{x})^2 n_{i\bullet} - \text{выборочная дисперсия случайной величины } X$$

$$S_Y^2 = \frac{1}{N} \sum_j (y_j - \bar{y})^2 n_{\bullet j} - \text{выборочная дисперсия случайной величины } Y$$

$$N = \sum_{i,j} n_{ij}$$