Линейная парная регрессия. Оценка параметров (спецификация)

Уравнение линейной парной регрессии имеет вид:

$$\hat{y} = a + b \cdot x$$

Коэффициент ***b*** – **коэффициент регрессии**

Он показывает **среднее изменение показателя y с изменением фактора x на единицу**

Уравнение линейной парной регрессии:

$$\hat{y} = a + b \cdot x$$

Необходимо оценить (найти) параметры уравнения: **a** и **b**.

Метод наименьших квадратов

Минимизируется сумма квадратов отклонений фактических значений переменной от теоретических

$$S(a, b) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \varepsilon_i^2 = \sum_i (y_i - a - b \cdot x_i)^2 \to \min$$

Метод наименьших квадратов

Задача сводится к решению системы нормальных уравнений

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum_i y_i + 2na + 2b\sum_i x_i = 0, \\ \dfrac{\partial S}{\partial b} = -2\sum_i y_i x_i + 2a\sum_i x_i + 2b\sum_i x_i^2 = 0 \end{cases}$$

Выкладки на доске

Метод наименьших квадратов

Преобразуя систему уравнений, получаем:

$$\begin{cases} a + b \cdot \bar{x} = \bar{y}, \\ a \cdot \bar{x} + b \cdot \overline{x^2} = \overline{x \cdot y} \end{cases}$$

Выкладки на доске

Метод наименьших квадратов

Решая систему находим параметры модели:

$$b = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{cov}(x, y)}{\sigma_x^2}$$

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

Выкладки на доске

Как количественно оценить линейную связь между переменными?

с помощью линейного коэффициента корреляции:

$$r_{xy} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (x_i - \bar{x})^2}} = \frac{\mathrm{cov}(x, y)}{\sigma_x \cdot \sigma_y} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{y^2} - \bar{y}^2}\sqrt{\overline{x^2} - \bar{x}^2}}$$

Коэффициент корреляции связан с коэффициентом регрессии:

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma_x \cdot \sigma_y} = b \frac{\sigma_x}{\sigma_y}$$

Так как

$$b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2}$$

Свойства линейного коэффициента корреляции:

1) $|r_{xy}| \le 1$

2) $r_{xy} > 0$: между x и y связь **прямая**,

$r_{xy} < 0$: **обратная**

3) $|r_{xy}| > 0{,}7$: связь достаточно сильная

$|r_{xy}| < 0{,}3$: связь слабая

$|r_{xy}| \approx 0$: линейная связь отсутствует

Примерные задания для самостоятельной работы

Пример 1. В некоторой бюрократической стране годовая зарплата Y каждого индивидуума определяется по формуле

$Y = 10\,000 + 500\,S + 200\,T$,

где S – число лет обучения индивидуума, T – его трудовой стаж (в годах), X – возраст. Рассчитайте Cov(X,Y), Cov(X,S) и Cov(X,T)

для выборки из 5 индивидуумов, представленной ниже, и проверьте, что

Cov (X,Y) = 500 Cov (X,S) + 200 Cov (X,T).

Объясните аналитически, почему так происходит.

Индивидуум	Возраст	Годы обучения	Трудовой стаж	Заработная плата
1	18	11	1	15 700
2	29	14	6	18 200
3	33	12	8	17 600
4	35	16	10	20 000
5	45	12	5	17 000

Пример 2. В таблице приведена динамика экономических показателей России: валовой внутренний продукт РФ (в процентах к предыдущему году) – показатель y (ВВП) и капитальные вложения в основные фонды РФ (в процентах к предыдущему году) – фактор x (КВОФ). Требуется оценить влияние фактора x на результативный признак y.

Год	ВВП (y,%)	КВОФ (x,%)
1991	95	85
1992	85,5	60
1993	91,3	88
1994	87,3	76
1995	95,8	90
1996	94	82
1997	100,4	95
1998	95,1	88
1999	104,6	105,3
2000	109,9	117,4
2001	105	108,7
2002	104,3	109,9

Пример 0. Оценить влияние фактора Х на результативный У для ситуации затруднения движения на дорогах (найти коэффициент корреляции):

Фактор Х (баллы пробок на дорогах)	Признак У (число невыбранных направлений движения)
1	15
2	10
3	2
4	2
5	- 4
6	- 10

У=19,2 – 4,76х
Где а=19,2,
в= -4,76

r = - 0,98