19. Свойства коэффициентов уравнения парной линейной регрессии.

Модель парной линейной регрессии имеет вид:

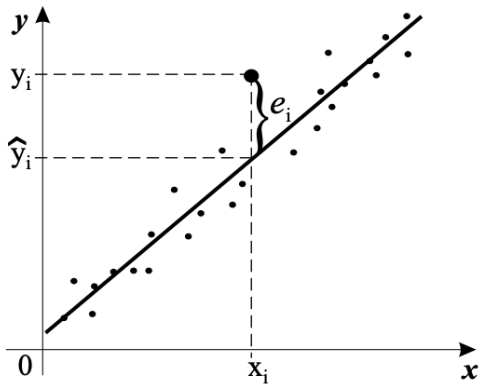
$
y = \beta_0 + \beta_1 x + \varepsilon
$

где:

- y — зависимая переменная,
- x — независимая переменная,
- β_0, β_1 — неизвестные параметры (коэффициенты),
- ε — случайная ошибка

Метод наименьших квадратов (МНК)

$
y = \beta_0 + \beta_1 x + \varepsilon
$
будем аппроксимировать $\hat{y} = b_0 + b_1 x$



Величина $\widehat{y}_i$ — это расчетное значение переменной y_i, соответствующее x_i. Наблюдаемые значения y_i не лежат в точности на линии регрессии, т.е. не равны $\widehat{y}_i$.

Определим остаток e_i в i-м наблюдении как разность между фактическим и расчетным значениями зависимой переменной:

$$e_i = y_i - \widehat{y}_i \tag{7.16}$$

Неизвестные значения (b_0, b_1) определим **методом наименьших квадратов** (МНК), суть которого заключается в минимизации суммы квадратов остатков:

$$Q = \sum e_i^2 = \sum (y_i - \widehat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2 \to \min. \tag{7.17}$$

Здесь (x_i, y_i) — известные значения наблюдения (числа), b_0, b_1 — неизвестные.

Запишем необходимые условия экстремума — частные производные от Q по неизвестным параметрам b_0 и b_1 должны быть равны нулю:

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} &= -2\sum(y_i - b_0 - b_1 x_i) = 0; \\ \dfrac{\partial Q}{\partial b_1} &= -2\sum(y_i - b_0 - b_1 x_i)x_i = 0. \end{cases}$$

В результате получим систему нормальных уравнений

$$\begin{cases} b_0 n - b_1 \Sigma x_i = \Sigma y_i, \\ b_0 \Sigma x_i + b_1 \Sigma x_i^2 = \Sigma x_i y_i, \end{cases} \text{ или } \begin{cases} b_0 + b_1 \dfrac{1}{n}\Sigma x_i = \dfrac{1}{n}\Sigma y_i; \\ b_0 \dfrac{1}{n}\Sigma x_i + b_1 \dfrac{1}{n}\Sigma x_i^2 = \dfrac{1}{n}\Sigma x_i y_i. \end{cases}$$

Заметим, что $\dfrac{1}{n}\sum x_i = \overline{x}$, $\dfrac{1}{n}\sum y_i = \overline{y}$, $\dfrac{1}{n}\sum x_i^2 = \overline{(x^2)}$ и $\dfrac{1}{n}\sum x_i y_i = \overline{xy}$. Тогда система примет простой вид:

$$\begin{cases} b_0 + b_1 \overline{x} = \overline{y}, \\ b_0 \overline{x} + b_1 \overline{(x^2)} = \overline{xy}. \end{cases}$$

Решая эту систему относительно b_0 и b_1, получим:

$$b_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2}, \qquad b_0 = \overline{y} - b_1 \overline{x}. \tag{7.18}$$

Проверим достаточное условие существования минимума функции $Q(b_0, b_1)$ при найденных значениях b_0 b_1 (**??**). Напомним, что для функции двух переменных в некоторой точке наблюдается минимум, если выполнены условия:

$$\Delta = \begin{vmatrix} A & B \\ B & C \end{vmatrix} > 0 \qquad \text{и} \qquad A > 0$$

где $A = \dfrac{\partial^2 Q}{\partial b_0^2}$, $B = \dfrac{\partial^2 Q}{\partial b_0 \partial b_1}$ и $C = \dfrac{\partial^2 Q}{\partial b_1^2}$.

Вычислим вторые частные производные:

$$A = \frac{\partial^2 Q}{\partial b_0^2} = \frac{\partial}{\partial b_0}\left[-2\Sigma(y_i - b_0 - b_1 x_i)\right] = 2n,$$
$$B = \frac{\partial^2 Q}{\partial b_0 \partial b_1} = \frac{\partial}{\partial b_1}\left[-2\Sigma(y_i - b_0 - b_1 x_i)\right] = 2\sum x_i,$$
$$C = \frac{\partial^2 Q}{\partial b_1^2} = \frac{\partial}{\partial b_1}\left[-2\Sigma(y_i - b_0 - b_1 x_i)x_i\right] = 2\sum x_i^2.$$

Составим требуемый определитель и проверим выполнение достаточных условий существования минимума:

$$\Delta = \begin{vmatrix} 2n & 2\Sigma x_i \\ 2\Sigma x_i & 2\Sigma x_i^2 \end{vmatrix} = 2n \begin{vmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{vmatrix} = 2n(\overline{x^2} - (\overline{x})^2) = 2n\widehat{D}(x) > 0, \quad A = 2n > 0.$$

Значит при значениях b_0 b_1, определяемых по формулам (**??**), функция $Q(b_0, b_1)$ достигает своего минимума.

Свойства коэффициентов уравнения парной линейной регрессии

Рассматриваемая нами модель регрессии (7.14)

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i. \tag{1}$$

Её оценкой по имеющимся данным является выборочное уравнение регрессии (7.15)

$$\widehat{y_i} = b_0 + b_1 \cdot x_i. \tag{2}$$

МНК-оценки коэффициентов β_0 и β_1, найденные нами ранее:

$$b_0 = \overline{y} - b_1\overline{x}, \tag{3.a}$$

$$b_1 = \frac{\operatorname{cov}(x,y)}{D(x)}. \tag{3.b}$$

Рассмотрим выражение (3.b). Подставим в него вместо y его выражение из (1):

$$b_1 = \frac{\operatorname{cov}(x,y)}{D(x)} = \frac{\operatorname{cov}(x, \beta_0 + \beta_1 \cdot x + \varepsilon)}{D(x)} = \frac{\operatorname{cov}(x,\beta_0) + \beta_1 \operatorname{cov}(x,x) + \operatorname{cov}(x,\varepsilon)}{D(x)} =$$
$$= \frac{0 + \beta_1 D(x) + \operatorname{cov}(x,\varepsilon)}{D(x)} = \beta_1 + \frac{\operatorname{cov}(x,\varepsilon)}{D(x)}.$$

Здесь учтено, что $\operatorname{cov}(x,\beta_0) = 0$, т.к. $\beta_0 = \text{const}$, и то, что $\operatorname{cov}(x,x) = D(x)$.

Рассмотрим выражение (3.b), подставив в него вместо b_1 полученное только что выражение, а вместо $\overline{y}$ — выражение $\frac{1}{n}\sum y_i$, где y_i определяется в (1):

$$b_0 = \overline{y} - b_1\overline{x} = \frac{1}{n}\sum(\beta_0 + \beta_1 \cdot x_i + \varepsilon_i) - \overline{x} \cdot \left[\beta_1 + \frac{\operatorname{cov}(x,\varepsilon)}{D(x)}\right] =$$
$$= \frac{1}{n}\sum \beta_0 + \beta_1 \frac{1}{n}\sum x_i + \frac{1}{n}\sum \varepsilon_i - \overline{x}\beta_1 - \frac{\overline{x} \cdot \operatorname{cov}(x,\varepsilon)}{D(x)} =$$
$$= \frac{1}{n} n \cdot \beta_0 + \beta_1\overline{x} + \frac{1}{n}\sum \varepsilon_i - \overline{x}\beta_1 - \frac{\overline{x} \cdot \operatorname{cov}(x,\varepsilon)}{D(x)} = \beta_0 + \left[\frac{1}{n}\sum \varepsilon_i - \frac{\overline{x} \cdot \operatorname{cov}(x,\varepsilon)}{D(x)}\right].$$

Получили, что коэффициенты b_0 и b_1 можно выразить через «истинные» коэффициенты β_0, β_1 и значения независимой переменной x:

$$b_1 = \beta_1 + \frac{\operatorname{cov}(x,\varepsilon)}{D(x)}, \tag{4}$$

$$b_0 = \beta_0 + \left[\frac{1}{n}\sum \varepsilon_i - \frac{\overline{x} \cdot \operatorname{cov}(x,\varepsilon)}{D(x)}\right]. \tag{5}$$

Несмещённость и дисперсии оценок

Покажем, что оценки b_0 и b_1 коэффициентов β_0 и β_1 являются несмещенными и дисперсии их равны следующим величинам:

$$D(b_0) = \sigma_{b_0}^2 = \frac{\overline{x^2} \cdot \sigma_\varepsilon^2}{n \cdot D(x)}; \qquad D(b_1) = \sigma_{b_1}^2 = \frac{\sigma_\varepsilon^2}{n \cdot D(x)}$$

Предварительно оценим (учитывая, что $\overline{\varepsilon} = 0$)

$$\operatorname{cov}(x, \varepsilon) = \frac{1}{n}\sum_i (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon}) = \frac{1}{n}\sum_i (x_i - \overline{x})\varepsilon_i - \frac{\overline{\varepsilon}}{n}\sum_i (x_i - \overline{x}) = \frac{1}{n}\sum_i (x_i - \overline{x})\varepsilon_i$$

Принимая предположение теоремы Гаусса-Маркова о том, что x - неслучайная (детерминированная) величина, оценим математическое ожидание и дисперсию коэффициентов, используя их представление (4) и (5). Из (4) оценим:

$$M[\operatorname{cov}(x, \varepsilon)] = M\left[\frac{1}{n}\sum_i (x_i - \overline{x})\varepsilon_i\right] = \frac{1}{n}\sum_i (x_i - \overline{x}) M[\varepsilon_i] = 0, \tag{6}$$

$$D[\operatorname{cov}(x, \varepsilon)] = D\left[\frac{1}{n}\sum_i (x_i - \overline{x})\varepsilon_i\right] = \frac{1}{n^2}\sum (x_i - \overline{x})^2 D(\varepsilon_i) = \frac{D(x)}{n}\sigma_\varepsilon^2. \tag{7}$$

Из (5) оценим:

$$M\left[\frac{1}{n}\sum \varepsilon_i\right] = \frac{\sum M(\varepsilon_i)}{n} = 0; \tag{8}$$

$$D\left[\frac{1}{n}\sum \varepsilon_i\right] = \frac{\sum D(\varepsilon_i)}{n^2} = \frac{\sigma_\varepsilon^2 \cdot n}{n^2} = \frac{\sigma_\varepsilon^2}{n}; \tag{9}$$

$$M\left[\frac{\overline{x}\ \operatorname{cov}(x, \varepsilon)}{D(x)}\right] = \frac{\overline{x}}{D(x)} M[\operatorname{cov}(x, \varepsilon)]_{(6)} = 0; \tag{10}$$

$$D\left[\frac{\overline{x}\ \operatorname{cov}(x, \varepsilon)}{D(x)}\right] = \frac{(\overline{x})^2}{D^2(x)} D[\operatorname{cov}(x, \varepsilon)]_{(7)} = \frac{(\overline{x})^2}{D^2(x)} \frac{D(x)}{n}\sigma_\varepsilon^2 = \frac{(\overline{x})^2 \sigma_\varepsilon^2}{nD(x)}. \tag{11}$$

Используя полученные оценки, найдем метематическое ожидание и дисперсию коэффициентов b_0 и b_1.

$$M(b_1) = M\left[\beta_1 + \frac{\operatorname{cov}(x, \varepsilon)}{D(x)}\right] = \beta_1 + \underbrace{\frac{M[\operatorname{cov}(x, \varepsilon)]}{D(x)}}_{(6)} = \beta_1$$

$$D(b_1) = M\left[\beta_1 + \frac{\operatorname{cov}(x, \varepsilon)}{D(x)}\right] = \underbrace{\frac{D[\operatorname{cov}(x, \varepsilon)]}{D^2(x)}}_{(7)} = \frac{D(x)}{n}\sigma_\varepsilon^2 \frac{1}{D^2(x)} = \frac{\sigma_\varepsilon^2}{n \cdot D(x)}$$

$$M(b_0) = M[\beta_0] + \underbrace{M\left[\frac{1}{n}\sum \varepsilon_i\right]}_{(8)} - \underbrace{M\left[\frac{\overline{x}\ \operatorname{cov}(x, \varepsilon)}{D(x)}\right]}_{(10)} = \beta_0;$$

$$D(b_0) = \underbrace{D\left[\frac{1}{n}\sum \varepsilon_i\right]}_{(9)} + \underbrace{D\left[\frac{\overline{x}\ \operatorname{cov}(x, \varepsilon)}{D(x)}\right]}_{(11)} = \frac{\sigma_\varepsilon^2}{n} + \frac{(\overline{x})^2 \sigma_\varepsilon^2}{nD(x)} = \frac{\sigma_\varepsilon^2}{n}\left[1 + \frac{(\overline{x})^2}{D(x)}\right] =$$

$$= \frac{\sigma_\varepsilon^2 \overline{x^2}}{nD(x)} \qquad \text{так как} \quad 1 + \frac{(\overline{x})^2}{D(x)} = \frac{D(x) + (\overline{x})^2}{D(x)} = \frac{(\overline{x^2} - (\overline{x})^2) + \overline{x})^2}{D(x)} = \frac{\overline{x^2}}{D(x)}$$

Оценкой дисперсии ошибки σ_ε^2 по выборке служит величина

$$S_e^2 = \frac{Q_e}{n-2} = \frac{\sum (y_i - \widehat{y}_i)^2}{n-2}.$$

Тогда для коэффициентов b_0 и b_1 выборочного уравнения регрессии имеем:

$$M(b_0) = \beta_0, \qquad M(b_1) = \beta_1.$$

Это означает, что найденные оценки являются несмещенными.

Дисперсии и средние квадратические отклонения коэффициентов b_0 и b_1:

$$D(b_0) = \frac{S_e^2 \overline{x^2}}{nD(x)}, \qquad S_{b_0} = \frac{S_e \sqrt{\overline{x^2}}}{\sqrt{n}\sigma_x},$$

$$D(b_1) = \frac{S_e^2}{nD(x)}, \qquad S_{b_1} = \frac{S_e}{\sqrt{n}\sigma_x}.$$