Линейная регрессия

- Y — зависимая (объясняемая) случайная величина
- X — независимая (объясняющая) случайная величина

- Предполагаемая линейная зависимость:

$$y = a + bx$$

- Реальные наблюдения (всего n наблюдений):

$$y_i = a + bx_i + \varepsilon_i, \qquad i \in \overline{1,n}$$

- Предположение модели: ошибки имеют идентичное нормальное распределение:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ т. е. } \mathrm{E}(\varepsilon_i) = 0 \text{ и } \mathrm{Var}(\varepsilon_i) = \sigma^2$$

Метод наименьших квадратов

- $\hat{a}$ и $\hat{b}$ — оценки параметров на основе имеющихся данных

- $\hat{y}_i = \hat{a} + \hat{b}x_i$ — значения y_i, предсказываемые моделью
- $\varepsilon_i = y_i - \hat{y}_i$ — i-ая ошибка модели, отклонение реального значения y_i от предсказываемого $\hat{y}_i$

- Значения $\hat{a}$ и $\hat{b}$ подбираются так, чтобы минимизировать сумму квадратов ошибок:

$$SSR = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{a} + \hat{b}x_i\right)\right)^2 \to min$$

- $y = \hat{a} + \hat{b}x$ — полученная прямая наименьших квадратов
- Свойство. ПНК проходит через точку $(\bar{x}; \bar{y})$: $\bar{y} = \hat{a} + \hat{b}\bar{x}$.

Три суммы квадратов

- $\bar{y}$ — среднее значение зависимой случайной величины Y в реальных наблюдениях
- Общая сумма квадратов (Sum of Squares Total):

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Сумма квадратов регрессии (Sum of Squares Explained):

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Сумма квадратов ошибок (Sum of Squared Residuals):

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- $SST = SSE + SSR$

Коэффициент детерминации

- Коэффициент детерминации регрессии:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- Чем больше R^2, тем больше объясняющая сила выбранной модели $y = a + bx$.

- Выборочный коэффициент корреляции:

$$r = \pm\sqrt{R^2}, \qquad \operatorname{sgn}(r) = \operatorname{sgn}(\hat{b})$$

- $R^2 \in [0; 1]$; $r \in [-1; 1]$
- Низкое значение $|r|$ означает либо незначительность зависимости величин X и Y, либо нелинейность этой зависимости (например, $y = a + bx + cx^2$).

Выборочная корреляция

- Выборочные дисперсии:

$$s_x{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i)^2 - n\bar{x}^2\right)$$

$$s_y{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(y_i)^2 - n\bar{y}^2\right)$$

- Выборочная ковариация:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}\right)$$

- Выборочный коэффициент корреляции:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i)^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}(y_i)^2 - n\bar{y}^2}}$$

Коэффициенты регрессионной модели

- Оптимальные значения коэффициентов $\hat{a}$ и $\hat{b}$:

$$\hat{b}_{\text{МНК}} = \frac{s_{xy}}{{s_x}^2} = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n}(x_i)^2 - n\bar{x}^2}$$

$$\hat{a}_{\text{МНК}} = \bar{y} - \hat{b}_{\text{МНК}}\bar{x}$$

- Полученная прямая наименьших квадратов:

$$y = \hat{a}_{\text{МНК}} + \hat{b}_{\text{МНК}}x$$