OLS - ALGORITHM

PREDICTOR ⟶ OUTCOME

(x) (y)

THE THEORY

$$y = \alpha + \beta x + \text{error}$$

error → ϵ

Assumptions:

1. Average value of error is zero [mean(ϵ)=0]
2. There is no link between x and ϵ

   [covariance $(x, \epsilon) = 0$]

THE THEORY

$$\hat{\beta} = \frac{sd_y}{sd_x} \times \text{correlation}(x, y)$$

$$\hat{\alpha} = \text{mean}(y) - \hat{\beta} \times \text{mean}(x)$$

sd_y : standard deviation of y

sd_x : standard deviation of x

$\hat{\alpha}$, $\hat{\beta}$: sample estimates

THE THEORY

the error term is represented by residual

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

$\hat{\epsilon}_i \neq \epsilon_i$ (sample $\neq$ population)

We choose $\hat{\alpha}$, $\hat{\beta}$ such that

$\sum \epsilon_i^2$ is as small as possible.

That's why we call it OLS.

THE THEORY

There are two assumptions that we need to make w.r.t. the errors :

1. the residuals are normally distributed.
2. the error ϵ has the same variance across the values of x

→ PLOT HISTOGRAM OF $\hat{\epsilon}$

THE THEORY

$\hat{\beta}$: association between x and y

$\hat{\alpha}$: value of y when $x = 0$

THE THEORY

- We test the hypothesis that x is associated with y

H_0: there is no association between x and y

$$\beta = 0$$

H_A : there is a link between x and y

$$\beta \neq 0$$

THE THEORY

- We test the hypothesis that x is associated with y

- We perform a t-test

$$\text{T-statistic} = \frac{\hat{\beta} - \text{null value}}{\text{standard error}}$$

What's the null value?
it's zero (0)

$$\text{T-statistic} = \frac{\hat{\beta}}{se(\hat{\beta})}$$

THE THEORY

- We test the hypothesis that x is associated with y
- We perform a t-test
- We compute the p-value.
- We estimate the confidence interval:

$$\hat{\beta} \pm 1.96 \times se(\hat{\beta})$$

THE THEORY

How good our model really is?

— A quick way to check that is to use this metric known as the R^2

$$R^2 = \left[\text{Correlation}(x, y) \right]^2$$

RECAP :

MODEL : $y = \alpha + \beta x + \epsilon$

WHAT DO WE WANT TO ESTIMATE? $\hat{\alpha}, \hat{\beta}$

HOW DO WE GET THE RESIDUALS? $\epsilon_i = y_i - \hat{y}_i$

WHILE ESTIMATING $\hat{\alpha}$ and $\hat{\beta}$, WHAT TEST DO WE PERFORM? t-test

HOW GOOD OUR MODEL IS? R^2

In R:

$$lm(y \sim x, data)$$

gets you the results of the regression model.

LET'S DO THIS FOR "EVAL" DATASET

y = teaching score

x = beauty score

BEFORE YOU RUN ANY MODEL, PLEASE DO THIS

1. SUMMARIZE THE DATA USING skim( )
2. VISUALISE x and y
3. COMPUTE CORRELATION BETWEEN x and y

The model that we estimate is:

$$\widehat{\text{teaching score}} = 3.88 + 0.067 \times \text{beauty score}$$

$\hat{\beta} = 0.067$: teaching score goes up by 0.067 when beauty score goes up by 1 pt

$\hat{\alpha} = 3.88$: average teaching score when beauty score $= 0$

CHECK: $\hat{\beta} = \dfrac{SD_{\text{teaching score}}}{SD_{\text{beauty score}}}$ | $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

The model that we estimate is:

$$\widehat{\text{teaching score}} = 3.88 + 0.067 \times \text{beauty score}$$

Recall that the residuals (ϵ) can be computed using

$$\epsilon = y - \hat{y} \quad \Big| \begin{array}{l} \text{teaching score} = 4 \\ \text{beauty score} = 3.9 \end{array}$$

$$\epsilon = 4 - (3.88 + 0.067 \times 3.9)$$

$$\epsilon = -0.14$$

The model that we estimate is:

$$\widehat{\text{teaching score}} = 3.88 + 0.067 \times \text{beauty score}$$

- Residuals help us evaluate how well the model fits the data.
- Draw the residual plot (residual vs beauty score)
  - the plot should be roughly horizontal

The model that we estimate is:

$$\widehat{\text{teaching score}} = 3.88 + 0.067 \times \text{beauty score}$$

- Let's compute R^2
- Recall that $R^2 = [\text{correlation}(x, y)]^2$

$$\text{correlation}(\text{teaching score}, \text{beauty score}) = 0.187$$

$$R^2 = (0.187)^2 = 0.035$$

What does this mean? the model explains 3.5% variation in the teaching score.

The model that we estimate is:

$$\widehat{\text{teaching score}} = 3.88 + 0.067 \times \text{beauty score}$$

$\hat{\beta} = 0.067$ | $se(\hat{\beta}) = 0.0163$

- let's focus on estimating the confidence interval for $\hat{\beta}$.

- $CI = \hat{\beta} \pm 1.96 \times se(\hat{\beta})$

$= 0.067 \pm 0.03136$

$= \{0.035, 0.099\}$