EX 08
Simple linear regression

Simple linear regression – **one** explanatory covariate.

<u>Model assumptions</u>

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \sim N(0, \sigma^2), \forall i$.

$E(Y_i) = E(\beta_0 + \beta_1 x + e_i) = \beta_0 + \beta_1 x_i$.

We are interested to find $\vec{\beta}$, understand how 'good' the model is (is it significant?) and we have a variety of tools to check the assumptions of the model and to extend it.

In a matrix form:

$$Y = X\vec{\beta} + \vec{e}$$

A good model is a model with low error –

$$Y - X\vec{\beta} = \vec{e}$$

So, we want to find a model such that some norm of this vector is small (in linear regression – 2 norm).

$$\min_{\beta} ||Y - X\vec{\beta}||^2$$

Let's practice some algebra in high dimensions:

$$\left|\left|X\vec{\beta} - Y\right|\right|^2 = \left(X\vec{\beta} - Y\right)^T\left(X\vec{\beta} - Y\right) = Y^TY - Y^TX\vec{\beta} - \vec{\beta}^TX^TY + \vec{\beta}^TX^TX\vec{\beta}$$

And the derivative with respect to $\vec{\beta}$ (gradient):

$$2Y^TX - 2\vec{\beta}^TX^TX$$

Setting the gradient to zero –

$$Y^TX - \vec{\beta}^TX^TX = 0$$

$$\vec{\beta} = (X^TX)^{-1}X^TY$$

There are many ways to solve those equations, usually using matrix decompositions (QR and Cholesky).

The solution for one explanatory variable –

$$\boxed{\beta_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \equiv \frac{SS_{xy}}{SS_x} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \beta_0 = \bar{y} - b_1\bar{x}}$$

Q1 – an engineer examines how the temperature affects the amount of product in the factory -

190	180	170	160	150	140	130	120	110	100	**Temperature**
89	85	78	74	70	66	61	54	51	45	**Product**

a. Calculate $\vec{\boldsymbol{\beta}}$

$n = 10 \qquad \sum_{x_i} x_i = 1450 \qquad \sum_{y_i} y_i = 673 \qquad \sum_{x_i} x_i^2 = 218500 \qquad \sum x_i y_i = 101570$

$\bar{x} = 145 \qquad \bar{y} = 67.3$

$SS_{xy} = \sum x_i y_i - n\bar{x} \cdot \bar{y} = 101570 - 10 \cdot 67.3 \cdot 145 = 3985$

$SS_x = \sum_{x_i} x_i^2 - n\bar{x}^2 = 218500 - 10 \cdot 145^2 = 8250$

$\boxed{b_1 = \frac{SS_{xy}}{SS_x} = \frac{3985}{8250} = 0.48303} \qquad \boxed{b_0 = \bar{y} - b_1\bar{x} = 67.3 - 0.48303 \cdot 145 = -2.73939}$

And the result -

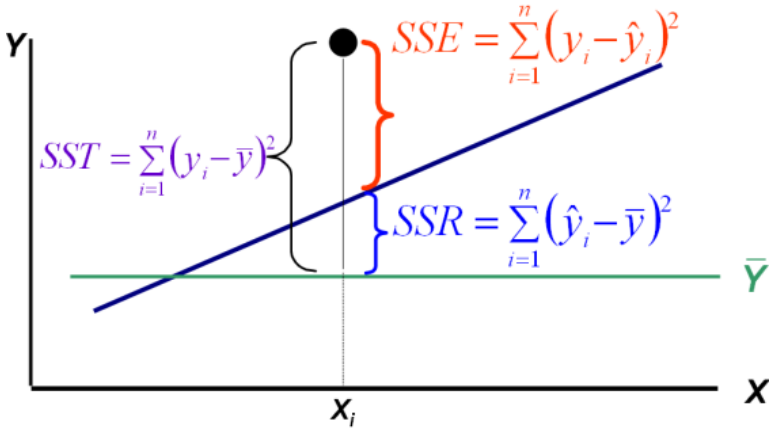
$$\hat{\boldsymbol{y}} = -\mathbf{2.74} + \mathbf{0.48} \cdot \boldsymbol{x}$$

Analysis of variance

$$\boldsymbol{SST = SSR + SSE}$$

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

SST - Variance of the observations
SSR – Sum of squared residuals
SSE – sum of squared errors



The more the regression line fits the observations $\frac{SSE}{SST}$ is getting smaller.
So we can derive a metric for goodness of fit of the regression -

$$\%\ \boldsymbol{of\ explained\ variance} = \boldsymbol{R^2} \equiv \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

b. Calculate R^2:

$$SSR = b_1^2 \cdot SS_x = 0.48303^2 \cdot 8250 = 1924.876, \qquad SST = SS_y = 1932.1$$

$$\Rightarrow R^2 = \frac{SSR}{SST} = \frac{1924.876}{1932.1} = 0.996261$$

Estimating σ^2 and $Var(\vec{\beta})$:

$$s^2 = \hat{\sigma}_e^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SST - SSR}{n-2} = \frac{SS_y - b_1^2 SS_x}{n-2}$$

$$V(\widehat{\beta_1}) = \frac{\hat{\sigma}_e^2}{\sqrt{SS_x}}, V(\widehat{\beta_0}) = \hat{\sigma}_e^2 \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$$

Knowing the mean and variance of $\vec{\beta}$, we can compute confidence intervals, (and hypothesis testing):

$$\beta_0 \in \left(b_0 - s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}\, t_{1-\frac{\alpha}{2}}^{(n-2)}, b_0 + s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}\, t_{1-\frac{\alpha}{2}}^{(n-2)} \right)$$

$$\beta_1 \in \left(b_1 - \frac{s}{\sqrt{SS_x}} t_{1-\frac{\alpha}{2}}^{(n-2)}, b_1 + \frac{s}{\sqrt{SS_x}} t_{1-\frac{\alpha}{2}}^{(n-2)} \right)$$

c. **Compute confidence interval for $\boldsymbol{\beta_1}$**

$$SS_y = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 1932.1 \Rightarrow s^2 = \frac{SS_y - b_1^2 SS_x}{n-2} = \frac{1932.1 - 0.48^2 \cdot 8250}{8} = 0.903$$

$$\beta_1 \in \left(b_1 \pm \frac{s}{\sqrt{SS_x}} t_{1-\frac{\alpha}{2}}^{n-2} \right) \Rightarrow 0.483 \pm 0.0105 \cdot 2.306 \Rightarrow \beta_1 \in (0.459, 0.507)$$

Hypothesis testing for statistical significance of the regression

We can do it in two ways:

1. **T test on $\boldsymbol{\beta_1}$**:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Test statistic

$$T_{b_1} = \frac{b_1 - 0}{S_{b_1}} = \frac{b_1}{\frac{s}{\sqrt{SS_x}}} \sim t(n-2)$$

Reject if:

$$|T_{b_1}| > t_{1-\frac{\alpha}{2}}^{n-2}$$

2. **F test:**

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}}$$

Reject if:

$$F > f_{1-\alpha}^{1,n-2}$$

d. **Test for significance of the regression:**

$$T_{b_1} = \frac{b_1 - 0}{\frac{s}{\sqrt{SS_x}}} = \frac{0.48303}{\frac{s}{\sqrt{8250}}}$$

$$s = \sqrt{\frac{SS_y - b_1^2 SS_x}{n-2}} = \sqrt{\frac{SS_y - 0.48303^2 \cdot 8250}{8}}$$

$$SS_y = \sum_{y_i} y_i^2 - n\bar{y}^2 = 47225 - 10 \cdot 67.3^2 = 1932.1$$

$$\Rightarrow s = 0.950279 \Rightarrow T_{b_1} = 46.16897$$

$$t_{1-\frac{\alpha}{2}}^{n-2} = t_{0.975}^{8} = 2.306$$

So, the regression if significant in $\alpha = 95\%$

Second way -

$$F = \frac{MSR}{MSE} = \frac{SSR/_1}{SSE/_{n-2}} = \frac{SSR}{s^2} = \frac{1924.876}{0.90303} = 2131.574$$

$$f_{1-\alpha}^{1,n-2} = f_{0.95}^{1,8} = 5.32 \Rightarrow F > f_{1-\alpha}^{1,n-2}$$

Confidence interval for the $E(y_p|X = x_p)$ (not proven)

$$E(y_{p)} \in \left((b_0 + b_1 \cdot x_p) \pm t_{1-\frac{\alpha}{2}}^{n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}} \right)$$

Confidence interval for the prediction

$$y_p \in \left((b_0 + b_1 \cdot x_p) \pm t_{1-\frac{\alpha}{2}}^{n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}} \right)$$

Why do we have +1 inside the squared root?

e. **Calculate $E(Y/X = x_p)$**

$$E(Y/X = x_p) \in \left((b_0 + b_1 x_p) \pm t_{1-\frac{\alpha}{2}}^{n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}} \right)$$

$$E(Y/X = 210) \in \left((-2.739 + 0.48303 \cdot 210) \pm t_{0.975}^{8} \cdot 0.95 \cdot \sqrt{\frac{1}{10} + \frac{(210 - 145)^2}{8250}} \right)$$

$$E(Y/X = 210) \in 98.7 \pm 2.306 \cdot 0.95 \cdot \sqrt{0.6121} \Rightarrow E(Y/X = 210) \in 98.7 \pm 1.714$$