Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, \dots, n$$

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

$\hat{\beta}_0 = ?$ $\qquad \hat{\beta}_1 = ?$

- Least squares Estimation
  Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that
  the minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = f(\beta_0, \beta_1)$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (1)$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \quad (2)$$

Using (1)

$$\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

$$\beta_0 + \beta_1 \bar{x} = \bar{y} \quad (1)$$

Using (2)

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$(2) \quad \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

solving this system we have :

- $$\hat{\beta}_1 = \frac{\left(n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

- $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

with

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

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

We also get :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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

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

$$MSE = \frac{SSE}{n-2} = \hat{\sigma}^2$$

$$E(\hat{\sigma}^2) = \sigma^2$$

<u>Hypothesis testing</u> :

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

$$T = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}}{\sqrt{SS_{xx}}}}$$

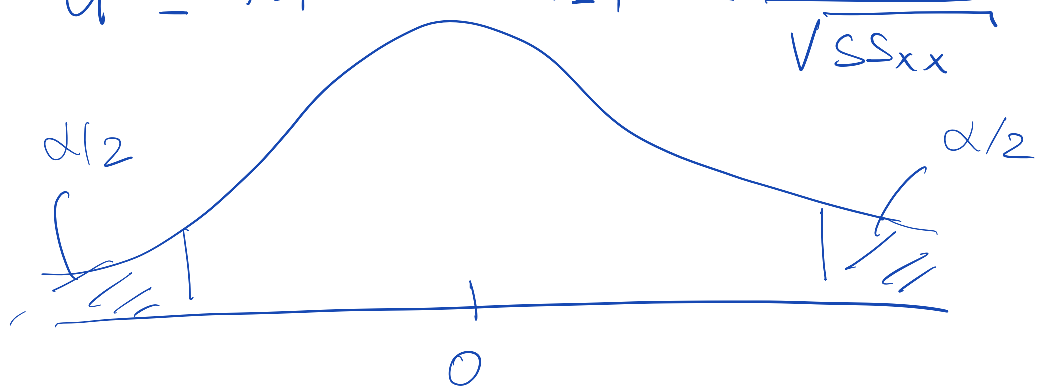
Under H_0
$T \sim$ student-t with $n-2$ d.f.

* Confidence interval for β_1

At α level, the $(1-\alpha) \times 100\%$ C.I. (L, U) can be obtained with

$$L = \hat{\beta}_1 - t_{\alpha/2, n-2} \frac{\hat{\sigma}}{\sqrt{SS_{xx}}}$$

$$U = \hat{\beta}_1 + t_{\alpha/2, n-2} \frac{\hat{\sigma}}{\sqrt{SS_{xx}}}$$



Coefficient of determination

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

Confidence interval for the mean response

Interval for x^*

$\mu_{y|x^*} = \beta_0 + \beta_1 x^*$

$\hat{\mu}_{y|x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^*$

Confidence interval $(1-\alpha) \times 100\%$

$$\hat{\mu}_{y|x^*} \pm t_{\alpha/2,\, n-2} \cdot se_{\hat{\mu}_{y|x^*}}$$

$$se_{\hat{\mu}_{y|x^*}} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

"Prediction interval" for a new observation at x^*

$y^{new} = \beta_0 + \beta_1 x^* + \varepsilon^{new}$

Point prediction: $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x^*$

The $(1-\alpha) \times 100\%$ interval

$$\hat{y}_{new} \pm t_{\alpha/2,\, n-2}\, se_{\hat{y}_{new}}$$

$$se_{\hat{y}_{new}} = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$