Linear Regression Formulas *Linear Regression Equation* $y = \alpha + \beta x + \epsilon$
Matrix form: $\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\epsilon}$
with $E(Y) = \boldsymbol{\beta} X$
Acronyms and names

- *TSS*: Total Sum of Squares
- *RSS*: Residual Sum of Squares
- *MSS*: Mean Sum of Squares
- $S_x x$: Corrected sum of squares of x
- $S_y y$: Corrected sum of squares of y
- $S_x y$: Corrected sum of products of xy

Least square estimates (matrix form) Sum of squares function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i + X_i \boldsymbol{\beta}$$

Estimates, matrix from

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}$$
$$RSS = S(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y^T Y} - \boldsymbol{Y^T X} \widehat{\boldsymbol{\beta}}$$
$$\sigma^2 = \frac{RSS}{n-p}$$

where σ^2 is the *variance*. $n - p$ is the *degrees of freedom*.
Estimates, non-matrix form

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$
$$S_{yy} = \sum_{i=1}^{n} (i_i - \overline{y})^2$$
$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(i_i - \overline{y})$$
$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}$$
$$\widehat{\alpha} = \overline{y} - \frac{S_{xy}}{Sxx} \overline{x}$$

Correlation

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$
$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

RSS,TSS, MSS

$$\widehat{\beta_0} = \overline{y}$$
$$then$$
$$S(\widehat{\beta_0}) = TSS \qquad = S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

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

$$TSS = MSS + RSS$$

R^2 **(standard and adjusted)**

$$R^2 = 1 - \frac{RSS}{TSS}$$

In the case of a *simple* linear regression (one explanatory variable) $R^2 = r^2$.

$$R^2(adj) = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}}$$
$$R^2(adj) = 1 - (1 - R^2) \frac{n-1}{n-p-1}$$

Assumptions of a linear model

- **A**: the deterministic part of the model captures all the non-random structure in the data.
- **B**: the scale of the variability of the errors is constant at all values of the explanatory variables.
- **C**: errors are independent.
- **D**: errors are normally distributed.
- **E**: the values of the explanatory variables are recorded without error.

Residuals

$$\widehat{\epsilon_i} = y_i - \widehat{(}y_i)$$

standardised residuals

$$r_i = \frac{\widehat{\epsilon_i}}{\sqrt{Var(\widehat{\epsilon_i})}}$$