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
- *n* : number of observations
- *p* : number of parameters (does not include intercept)

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

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

where σ^2 is the *variance*. $n-p$ is the *degrees of freedom*.

Estimates, non-matrix form

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

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

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

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

$$TSS = MSS + RSS$$

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

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

Inference for Regression Coefficients estimated standard error (e.s.e./s.e.)

$$e.s.e.(\widehat{\boldsymbol{\beta}}) = e.s.e(\widehat{\boldsymbol{\beta}}) = \sqrt{\frac{RSS}{n-p}\boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{b}}$$

pivotal function

$$\frac{\boldsymbol{b}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{b}^T\boldsymbol{\beta}}{e.s.e(\widehat{\beta})} \sim t(n-p; \frac{1+c}{2})$$

with c : confidence level ($c = 0.95 \implies (c+1)/2 = 0.975$)
prediction interval

$$\boldsymbol{b}^T\widehat{\boldsymbol{\beta}} \pm t(n-p; \frac{1+c}{2})e.s.e.(\widehat{\boldsymbol{\beta}})$$
$$\boldsymbol{b}^T\widehat{\boldsymbol{\beta}} \pm t(n-p; \frac{1+c}{2})\sqrt{\frac{RSS}{n-p}\boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{b}}$$

ANOVA table

Component	df	SS	MS	F value
Model	$p-1$	MSS	$\frac{MSS}{p-1}$	$\frac{\frac{MSS}{p-1}}{\frac{RSS}{n-p}}$
Residual	$n-p$	RSS	$\frac{RSS}{n-p}$	-
Total	$n-1$	TSS	-	-

- df : degrees of freedom
- SS : Sum of Squares
- MS : Mean Squares

F statistic (H_0 : all parameters = 0) (large F values $\implies$ rejection of H_0)