

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

Acronyms and names

- *TSS*: Total Sum of Squares
- *RSS*: Residual Sum of Squares
- *RSE*: Residual Standard Error
- *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

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

$$= \sigma = \sqrt{\frac{RSS}{n-p}}$$

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

Estimates, non-matrix form

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

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

$$\hat{\beta}_0 = \bar{y}$$

then

$$S(\hat{\beta}_0) = TSS$$
$$= S_{yy}$$
$$= \sum_{i=1}^n (y_i - \bar{y})^2$$

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

$$TSS = MSS + RSS$$

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

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

standardised residuals

$$r_i = \frac{\hat{\epsilon}_i}{\sqrt{Var(\hat{\epsilon}_i)}}$$

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

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

pivotal function

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

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

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

prediction interval (for predicted variable)

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

Different/Paralell lines model \ model : $Y_{i,j} = \alpha_i + \beta_i (x_{i,j} + \bar{x}_i) + \epsilon_{i,j}$; \ into matrix notation : $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ *epsilon* \

Different lines

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & (x_{11} - \bar{x}_1) & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_{1n_1} - \bar{x}_1) & 0 & 0 \\ 0 & 0 - \bar{x}_1 & 1 & (x_{21} - \bar{x}_2) \\ 0 & 0 - \bar{x}_1 & 1 & (x_{2n_2} - \bar{x}_2) \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix}$$

Least squares estimate

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{y}_1 \\ \frac{S_{x_1 y_1}}{S_{x_1 x_1}} \\ y_2 \\ \frac{S_{x_2 y_2}}{S_{x_2 x_2}} \end{pmatrix}$$
$$RSS = RSS1 + RSS2$$

95% confidence interval for $(\beta_1 - \beta_2)$

$$\hat{\beta}_1 - \hat{\beta}_2 \pm t(n_1 + n_2 - 4; 0.975) \sqrt{\frac{RSS_1 + RSS_2}{n_1 + n_2 - 4} (\frac{1}{S_{x_1 x_2}} + \frac{1}{S_{x_1 x_2}})}$$

Parallel lines

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & 0 & (x_{11} - \bar{x}_1) \\ \vdots & \vdots & \vdots \\ 1 & 0 & (x_{1n_1} - \bar{x}_1) \\ 0 & 1 & (x_{21} - \bar{x}_2) \\ \vdots & \vdots & \vdots \\ 0 & 1 & (x_{2n_2} - \bar{x}_2) \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \frac{\bar{y}_1}{\bar{y}_2} \\ \frac{S_{x_1 y_1} + S_{x_2 y_2}}{S_{x_1 x_1} + S_{x_2 x_2}} \end{pmatrix}$$

95% confidence interval for $\alpha_1 - \alpha_2 + \beta(\bar{x}_2 - \bar{x}_1)$

$$\mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ \bar{x}_2 - \bar{x}_1 \end{pmatrix}$$

$p = 3$

$$\mathbf{b}^T \hat{\boldsymbol{\beta}} \pm t(n_1 + n_2 - p; 0.975) \sqrt{\frac{RSS}{n_1 + n_2 - p} \mathbf{b}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{b}}$$

$$\mathbf{b}^T \hat{\boldsymbol{\beta}} \pm t(n_1 + n_2 - p; 0.975) \sqrt{\frac{RSS}{n_1 + n_2 - p} \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{x}_2 - \bar{x}_1)^2}{S_{x_1 x_1} + S_{x_2 x_2}} \right)}$$

$$(\mathbf{X}^T \mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & \frac{1}{S_{x_1 x_1} + S_{x_2 x_2}} \end{pmatrix}$$

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)