CHAPTER 1: Simple linear regression

mathematical model: $y = \beta_0 + \beta_1 x$

statistical model: $Y = \beta_0 + \beta_1 x + \varepsilon.$

$$y = X\beta + \varepsilon. \quad \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \underbrace{\underset{n\times(n+1)}{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & & \\ 1 & x_{i1} & \cdots & x_{in} \\ \vdots & \vdots & & \\ 1 & x_{n1} & \cdots & x_{nn} \end{bmatrix}} \underset{(n+1)\cdot 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}}}_{n \times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y_i = \beta_0 + x_{i1}\cdot\beta_1 + \cdots + x_{in}\cdot\beta_n + \varepsilon_i$$

Ordinary least squares (OLS)

$$\min \quad S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{in}\beta_n)^2$$

Simple linear regression

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

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Matrix form

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X'X)^{-1} X'y = \beta + (X'X)^{-1} \cdot X' \cdot \varepsilon$$

proof: $e = y - X\hat{\beta}$ $\qquad$ $\circledast$ $\varepsilon = y - X\beta$. $(e \neq \varepsilon)$

$$e'e = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - \hat{\beta}'X' \cdot y - y'X\hat{\beta} + \hat{\beta}'X' \cdot X\hat{\beta}$$

$$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X \cdot \hat{\beta}$$

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \Rightarrow X'X\hat{\beta} = X'y$$

$$\Rightarrow \hat{\beta} = (X'X)^{-1} \cdot X'y.$$

$$\hat{\beta} = (X'X)^{-1} X'y = (X'X)^{-1} X'(X\beta + \varepsilon) = (X'X)^{-1} X'X\beta + (X'X)^{-1} X' \cdot \varepsilon$$

$$= \beta + \underbrace{(X'X)^{-1} \cdot X'}_{A} \cdot \varepsilon = \beta + A \cdot \varepsilon$$

$$H = \begin{bmatrix} \dfrac{\partial^2 S}{\partial \beta_0^2} & \dfrac{\partial^2 S}{\partial \beta_0 \, \partial \beta_1} \\ \dfrac{\partial^2 S}{\partial \beta_1 \partial \beta_0} & \dfrac{\partial^2 S}{\partial \beta_1^2} \end{bmatrix} = \begin{bmatrix} 2n & 2n\bar{X} \\ 2n\bar{X} & 2\sum_{i=1}^{n} X_i^2 \end{bmatrix}$$

(1) : regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

(2) residuals: $e_i = y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})$

$$\sum_{i=1}^{n} e_i = 0. \quad \left(\because \sum_{i=1}^{n} (y_i - \bar{y}) = 0, \ \sum_{i=1}^{n} (x_i - \bar{x}) = 0 \right)$$

(3) prediction: $\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$

(4) for a model $y_i = \beta_1 x_i + \varepsilon_i$, $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

$\sum_{i=1}^{n} e_i \neq 0$ in this case.

Theorem Connection between errors and estimators.

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} a_{ni} \varepsilon_i$$

$$\hat{\beta}_0 = \beta_0 + \sum_{i=1}^{n} b_{ni} \varepsilon_i$$

$$a_{ni} = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad b_{ni} = \frac{1}{n} - \bar{x} \cdot a_{ni}.$$

Gauss-Markov

$E(\varepsilon_i) = 0.$

$var(\varepsilon_i) = \sigma^2, \quad \forall i$

$cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0. \quad \forall i \neq j$

Theorem: mean and variance of estimators.

$E(\hat{\beta_0}) = \beta_0. \qquad E(\hat{\beta_1}) = \beta_1.$

$$var(\hat{\beta_0}) = \sigma^2_{\hat{\beta_0}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$var(\hat{\beta_1}) = \sigma^2_{\hat{\beta_1}} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$cov(\hat{\beta_0}, \hat{\beta_1}) = -\bar{x} \cdot \sigma^2_{\hat{\beta_1}}$$

(*) Among all linear unbiased estimators of y, the ordinary least squares estimators have minimum variance.

The unbiased and convergent estimator of σ^2

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 , \quad e_i = y_i - \hat{y}_i$$

remplace σ^2 by s^2,

$$\hat{\sigma}^2_{\hat{\beta}_0} = s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{s^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

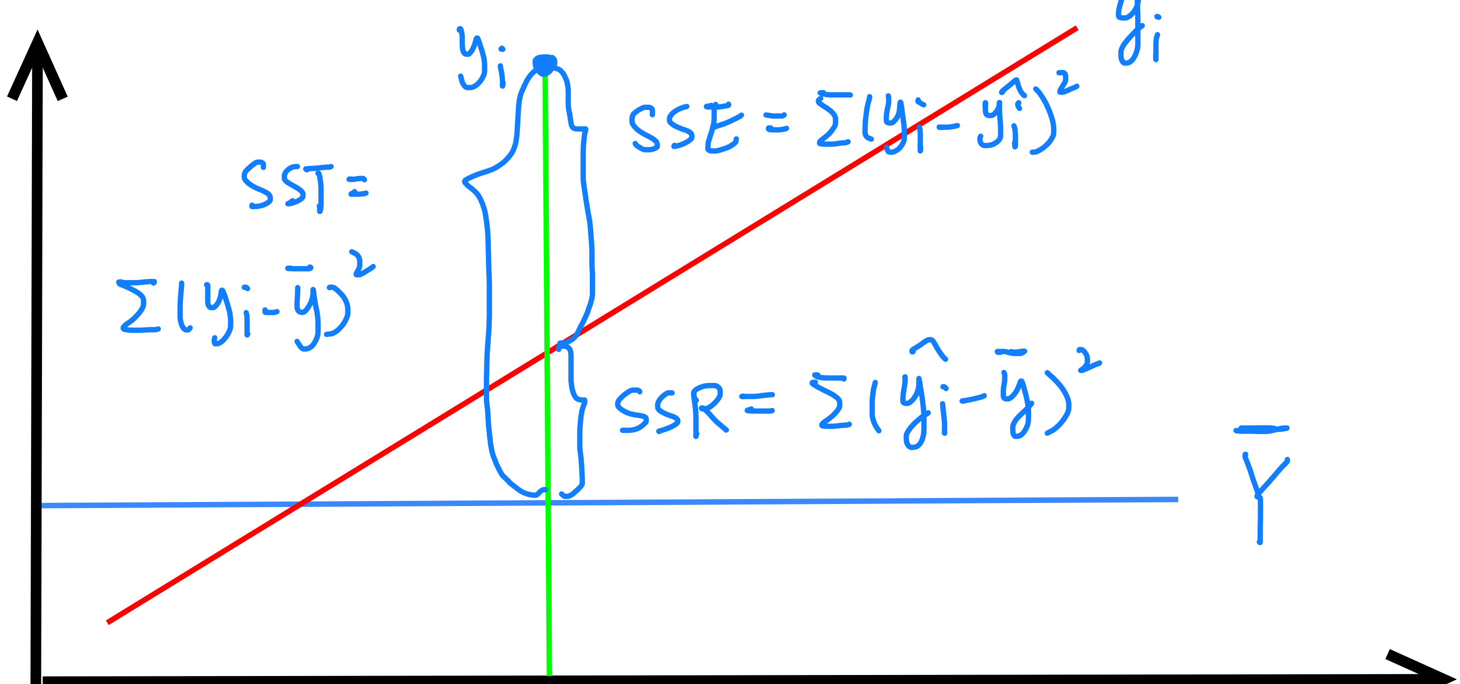
$$\widehat{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x} \cdot \hat{\sigma}^2_{\hat{\beta}_1}$$

the estimation of standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$s.e(\hat{\beta}_0) = \hat{\sigma}_{\hat{\beta}_0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$s.e(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

SST, SSE and SSR.



Sum of Squares regression (SSR) $= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Sum of Squares error (SSE) $= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Sum of squares total (SST) $= \sum_{i=1}^{n} (y_i - \bar{y})^2$

SST = SSR + SSE

Coefficient of determination (R^2)

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

It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

Standard ANOVA.

k: number of independant vars

Source	DF	Sum of squares	Mean square
model	k	SSR $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	MSR $\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{k}$
ERROR	$n-k-1$	SSE $\sum_{i=1}^{n}(y_i-\hat{y})^2$	MSE $\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n-k-1}$
Total	$n-1$	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	

F-test ANOVA.

$$F=\frac{SSB \mid df_b}{SSW \mid df_w}$$

$SSB=\sum(g_i-\bar{x})^2$

g_i : average of each group.

$SSW=\sum(\underline{x_i-g})^2$

distance between each observed value within the group x from the group-mean g.

Confidence intervals and hypothesis tests.

Theorem: Distribution of estimators: known variance.

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\right)$$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$i = 1, \ldots, n.$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

$$\hat{\beta} \sim \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N(\beta, \underbrace{\sigma^2 \Omega}_{\mathrm{cov}(\hat{\beta})})$$

$$\Omega = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{V} & \frac{-\bar{x}}{V} \\ \frac{-\bar{x}}{V} & \frac{1}{V} \end{bmatrix}$$

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

$(\hat{\beta}_0, \hat{\beta}_1)$ et s^2 sont indépendants.

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}$$

Theorem: Distribution of estimators. (unknown variance)

(a)
$$T = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} = \frac{\hat{\beta}_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{V}}} \sim t_{n-2}$$

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

(b)
$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{s\sqrt{\frac{1}{V}}} \sim t_{n-2}$$

(c)
$$F = \frac{1}{2s^2} (\hat{\beta} - \beta)' \cdot \Omega^{-1} (\hat{\beta} - \beta) \sim F_{2,n-2}$$

DF numerator   DF denominator

proof: (a)
$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{V}} \sim N(0,1)$$

$$K = \frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}$$

$$T = \frac{Z}{\sqrt{K/(n-2)}} = \frac{\hat{\beta}_1 - \beta}{s / \sqrt{V}} \sim t_{n-2}$$

(b) similarly. $$T = \frac{\hat{\beta}_0 - \beta}{s / \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{V}}} \sim t_{n-2}$$

(c) $$\hat{\beta} - \beta \sim N_2(0, \sigma^2 \Omega)$$

$$Z = \frac{\hat{\beta} - \beta}{\sigma \sqrt{\Omega}} \sim I_2 \Rightarrow \frac{\Omega^{-\frac{1}{2}}(\hat{\beta} - \beta)}{\sigma} \sim I_2.$$

Rappel: if $u \sim N_p(0, I_p)$, then $u'u \sim \chi^2_p$

$A = PDP'$ $\quad A^k = PD^kP'$

$$K_1 = z'z = \frac{(\hat{\beta} - \beta)' \Omega^{-\frac{1}{2}} \Omega^{-\frac{1}{2}} (\hat{\beta} - \beta)}{\sigma^2} = \frac{(\hat{\beta} - \beta)' \Omega^{-1} (\hat{\beta} - \beta)}{\sigma^2} \sim \chi^2_2$$

$$K_2 = (n-2)\frac{s^2}{\sigma^2} \sim \chi^2_{n-2} \quad \text{ind.}$$

$$F = \frac{K_1 / V_1}{K_2 / V_2} = \frac{K_1 / 2}{K_2 / (n-2)} = F_{2, n-2}$$

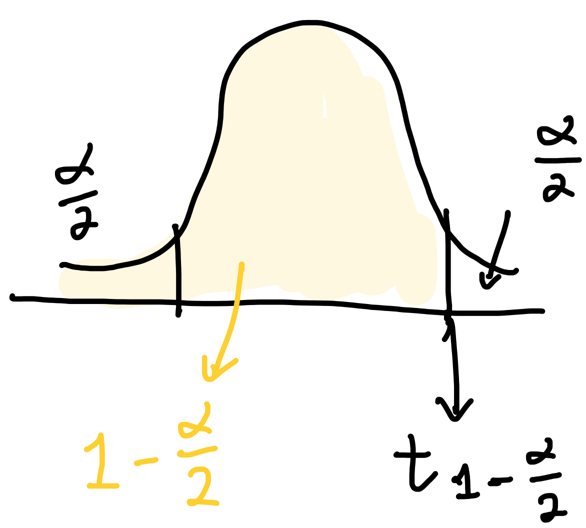
$$\Rightarrow F = \frac{(\hat{\beta} - \beta)' \Omega^{-1} (\hat{\beta} - \beta)}{2 s^2} \sim F_{2, n-2}$$

Confidence interval.

$$\left[\hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}} \cdot S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{V}}\ ,\ \hat{\beta}_0 + t_{n-2,\frac{\alpha}{2}} \cdot S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{V}} \right]$$

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

$$\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}} \cdot S\sqrt{\frac{1}{V}}\ ,\ \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{1}{V}} \right]$$



Confidence region of $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

$$n(\hat{\beta}_0 - \beta_0)^2 + 2n\bar{x}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + \sum_{i=1}^{n} x_i^2 (\hat{\beta}_1 - \beta_1)^2 \leq$$

$$2S^2 \cdot F_{2,n-2}(\alpha)$$

Test for Significance of Regression.

1: $H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$

H_1: At least a $\beta_j \neq 0$

$$F = \frac{MSR}{MSE} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / K}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-K-1)} \sim F_{K,\, n-K-1}$$

decision rule: reject H_0 if $F_{observed} \geq F_{K,\, n-K-1}(\alpha)$

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

$$T = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} \sim t_{n-2}$$

reject H_0 if $T_{observed} \geq t_{n-2, \frac{\alpha}{2}}$

Prediction.

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$E(\hat{y}_0) = \beta_0 + \beta_1 x_0$$

$$\mathrm{var}(\hat{y}_0) = \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{V}\right], \quad V = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

future observation $\quad y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$

$$E(\varepsilon_0) = 0, \quad E(\varepsilon_0\varepsilon_i) = 0. \quad \mathrm{var}(\varepsilon_0) = \sigma^2$$

$$E(y_0 - \hat{y}_0) = 0.$$

$$\mathrm{var}(y_0 - \hat{y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{V}\right]$$

Confidence interval of prediction.

$$\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{V}}$$