TMA4268 21.01.2019

M3L1: Linear regression

Simple linear regression

one x

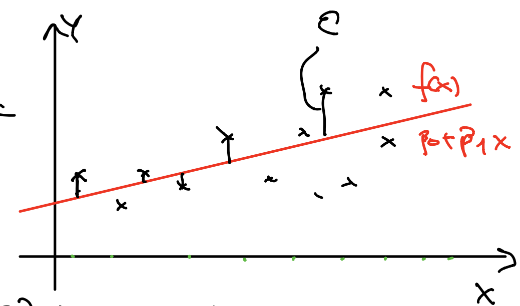
Model: $Y = f(x) + \varepsilon = \underset{\text{intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1} \cdot \underset{\text{covariate}}{x} + \underset{\text{error term}}{\varepsilon}$

Additional assumption:

ε has $E(\varepsilon) = 0$

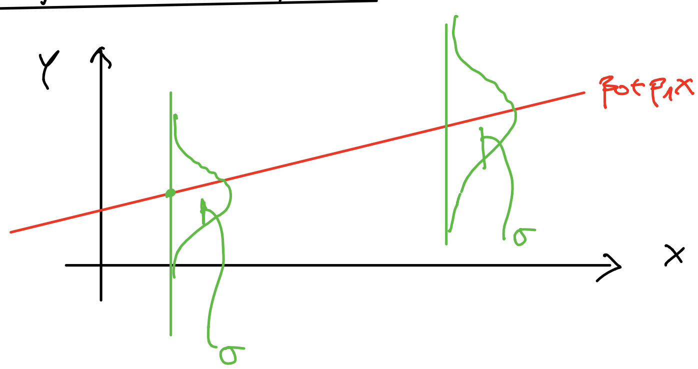
$Var(\varepsilon) = \sigma^2$ ← homoscedastic error

not dependent on x



(if $Var(\varepsilon)$ varies with x ⇒ heteroscedastic errors)

And often $\varepsilon \sim N(0, \sigma^2)$



And we also assume that the pairs (x_i, Y_i) $i = 1, .., n$ are independent

PARAMETER ESTIMATION

For a given dataset (x_i, Y_i), $i=1,..,n$ independent pairs. (observed)

We don't know $\underbrace{\beta_0, \beta_1 \text{ and } \sigma^2}_{\text{unknown parameters}}$, and need to find estimators

$\underbrace{\hat\beta_0, \hat\beta_1}$ and $\hat\sigma^2$

- least squares
- maximum likelihood

restricted maximum likelihood

Let $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. We find $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing

$$\underset{\substack{\uparrow \\ \text{residual}}}{R}SS = \sum_{i=1}^{n} (\underbrace{Y_i - \hat{Y}_i}_{e_i})^2 \qquad (Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i)$$

sum of squares

e_i ‖ residual ↑ predicted value for the error ε_i ← RV

$e_i = \hat{\varepsilon}_i$

$$\left.\begin{aligned} \frac{\partial RSS}{\partial \beta_0} &= 0 \\ \frac{\partial RSS}{\partial \beta_1} &= 0 \end{aligned}\right\} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We also need $\hat{\sigma}^2$ ← the variance of ε's

Remember that $\varepsilon \sim N(0, \sigma^2)$; $\mathrm{Var}(\varepsilon) = E(\varepsilon^2) - \underbrace{E(\varepsilon)^2}_{0}$

Since the e_i's are predictions of the ε_i's, we use them.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{\underset{\substack{\uparrow \\ \#\text{param. estimated } (\hat{\beta}_0, \hat{\beta}_1)}}{n-2}} = \frac{RSS}{n-2}$$

$RSE = \hat{\sigma}$

Distribution of parameter estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{constants}} + \overset{N(0,\sigma^2)}{\varepsilon_i} \Rightarrow Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$E(Y_i) = \beta_0 + \beta_1 x_i + \underbrace{E(\varepsilon_i)}_{0}$$

$$\hat{\beta}_j \sim N(\beta_j, \underbrace{Var(\hat{\beta}_j)}_{c_{jj}\sigma^2})$$

$j = 0, 1$

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\sigma} \sim N(0,1)$$

+ can be shown

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

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\hat{\sigma}} \sim t_{n-2}$$

$\leftarrow$ n − # regr. param. estimated

$(1-\alpha)\cdot 100\%$ CI : $\left[\hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-2} \cdot \sqrt{c_{jj}} \cdot \hat{\sigma}\right]$

confidence interval

Hypothesis test: $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

	reject H_0	not reject H_0
H_0 true	type I error	correct
H_0 false	correct	type II error

guilty criminal go free

crime of justice

$P(\text{Type I error}) \leq \alpha$
0.05

p-value $= P(T_0 \geq |t_0| \mid H_0 \text{ true})$

reject H_0 when p-value $\leq \alpha$.

How good is the regression

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \text{total variability}$$

total   sums of squares

regr. line

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \text{not explained by regression}$$

obs

$$R^2 = \frac{\overbrace{TSS - RSS}^{\text{explained by regression}}}{TSS} = 1 - \frac{RSS}{TSS} \in [0,1]$$

High R^2 is good.

Multiple linear regression (MLR)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p \cdot x_{ip} + \varepsilon_i$$

$$\begin{bmatrix} Y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \varepsilon_1 \\ Y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \varepsilon_2 \\ \vdots \\ Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \varepsilon_n \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{\substack{Y \\ n\times 1 \\ \text{response vector}}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & & x_{2p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & & x_{np} \end{bmatrix}}_{\substack{X \\ n\times(p+1) \\ \text{design matrix}}} \cdot \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{\substack{\beta \\ (p+1)\times 1 \\ \text{vector of regression param}}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\substack{\varepsilon \\ n\times 1 \\ \text{vector of errors}}}$$

Combine independent pairs (x_i, Y_i) and $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$

$$\Rightarrow \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \sim \underbrace{N_n}_{\text{multivariate normal dim } n} \Big(0, \underbrace{\sigma^2 I}_{\begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & \\ & & \ddots & \\ 0 & & & \sigma^2 \end{bmatrix}}\Big)$$

Homework: Distribution of Y?