TÉCNICO LISBOA

Simple Linear Regression (Review)

Regression analysis is a statistical technique used to describe relationships among variables

the simplest case : one explanatory variable

observations: $(x_i, y_i), i = 1, \dots, n$

SLR Model : $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \dots, n$

$Y_i \in \mathbb{R}$ is the real-value response for the i-th observation (dependent variable)

β_0 and β_1 (parameters) – regression coefficients

$x_i \in \mathbb{R}$ is the predictor or explanatory variable (independent variable)

ε_i random variables (error term) unobserved

Assumption (random error) : $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, $\forall i$

$Cov(\varepsilon_i, \varepsilon_j) = 0$ $\forall i \neq j$ (non-correlated)

Usually we also assume that $\varepsilon_i \sim N(0, \sigma^2)$ Gaussian error term

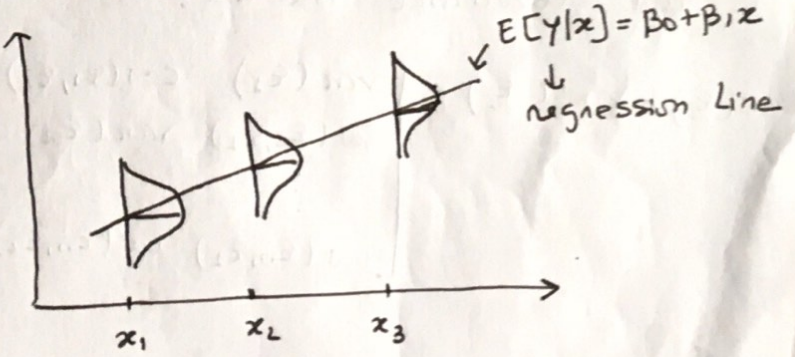
iid (independent and identically distributed)

- the model is simple because we have only one predictor
- the model is linear because Y_i is a linear function of the parameters β_0 and β_1
- the model is a regression model because we are modeling a response variable (Y) as a function of a predictor variable (x)

$E[Y|x_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i] = \beta_0 + \beta_1 x_i$

$Var(Y|x_i) = Var(\beta_0 + \beta_1 x_i + \varepsilon_i) = Var(\varepsilon_i) = \sigma^2$, $\forall i$ homogeneity of variance

$\Rightarrow (Y|x_i) \overset{\text{indep}}{\sim} N(\beta_0 + \beta_1 x_i ; \sigma^2)$



$E[Y|x] = \boxed{\beta_0 + \beta_1 x}$ regression Line

β_0 - intercept, β_1 - slope } unknown parameters

and $\text{var}(Y|x) = \text{var}(\varepsilon) = \sigma^2$? (σ^2 unknown)

Matrix notation:

$$Y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$
$$\vdots$$
$$Y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

$$\boxed{\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}}$$

Notation: Matrix and vectors: in boldface letters

$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$ $(n \times 1)$ vector of response random variables

$\underset{\sim}{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ $(n \times 2)$ matrix called the design matrix

$\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ (2×1) vector of unknown parameters

$\underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ $(n \times 1)$ vector of random errors

$\underset{\sim}{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & 1 \end{bmatrix}$ → Indentity matrix

$\underset{\sim}{\varepsilon} \sim N_n(\underset{\sim}{0}, \sigma^2 \underset{\sim}{I})$ is an n-dimensional normal distribution with:

$$E(\underset{\sim}{\varepsilon}) = E\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \underset{\sim}{0}_{n \times 1}$$

and covariance matrix:

$$\text{cov}(\underset{\sim}{\varepsilon}) = \text{var}(\underset{\sim}{\varepsilon}) = \begin{bmatrix} \text{var}(\varepsilon_1) & \text{cov}(\varepsilon_1, \varepsilon_2) & \cdots & \text{cov}(\varepsilon_1, \varepsilon_n) \\ \text{cov}(\varepsilon_2, \varepsilon_1) & \text{var}(\varepsilon_2) & \cdots & \text{cov}(\varepsilon_2, \varepsilon_n) \\ \vdots & & & \\ \text{cov}(\varepsilon_n, \varepsilon_1) & \text{cov}(\varepsilon_n, \varepsilon_2) & \cdots & \text{var}(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

$$= \sigma^2 \underset{\sim}{I}_{n \times n}$$

$\text{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) - E(\varepsilon_i) E(\varepsilon_j)$

$\text{cov}(\varepsilon_i, \varepsilon_i) = E(\varepsilon_i \varepsilon_i) - E(\varepsilon_i) E(\varepsilon_i)$

$= E(\varepsilon_i^2) - E^2(\varepsilon_i) = \text{var}(\varepsilon_i)$

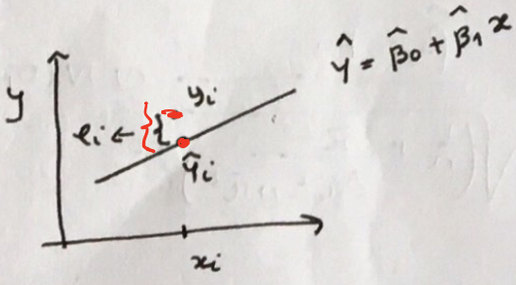
TÉCNICO LISBOA

• Estimation of SLR model:

$E[Y|x] = \beta_0 + \beta_1 x$

$\hat{E}[Y|x_i] = \hat{y}_i = \hat{\mu}_{Y|x_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\min x_i \leq x_i \leq \max x_i$



Residuals: $y_i - \hat{y}_i = e_i = r_i$

Least square Method: find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the sum of square Residuals (it is also called the sum of squared errors, SSE)

$$\min_{\hat{\beta}_0, \hat{\beta}_1 \in \mathbb{R}} SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\begin{cases} \frac{\partial SSE}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial SSE}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$ Normal equations

Solution:

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

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

the Least square estimate of σ^2 is $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \dots =$

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

↓ computational formula

• Inferences in SLR

$\varepsilon_i \sim N(0, \sigma^2)$ iid. $\Rightarrow$ $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $\forall i$ indep

Parameter	Estimator	Expected value:	Variance
Intercept: β_0	$\hat{\beta}_0$	$E(\hat{\beta}_0) = \beta_0$	$Var(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2} \right) \sigma^2$
Slope: β_1	$\hat{\beta}_1$	$E(\hat{\beta}_1) = \beta_1$	$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2}$
$E[Y\|x_0] = \beta_0 + \beta_1 x_0$	$\hat{\mu}_{Y\|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$	$E[\hat{\mu}_{Y\|x_0}] = \beta_0 + \beta_1 x_0$	$Var(\hat{\mu}_{Y\|x_0}) = \left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2} \right) \sigma^2$

* Inference on β_0:

Pivotal quantity: $\hat{\beta}_0 \sim N\left(\beta_0; \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}\right)\sigma^2\right)$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}\right)\sigma^2}} \sim N(0,1)$$

σ^2 → unknown

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

$\hat{\sigma}^2$ estimator

* β_1: Pivotal $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \sim t(n-2)$

* $\mu_{Y|x_0}$: Pivotal ~~variable~~ quantity: $T = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t(n-2)$

$= E[Y|x_0] = \beta_0 + \beta_1 x_0$

* Prediction Intervals for a new or future observation of Y (Y_0)

$(Y|x_0) \sim N(\beta_0 + \beta_1 x_0; \sigma^2)$

$Y|x_0 = Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon = E[Y|x_0] + \varepsilon$

$\hat{Y}|x_0 = \hat{E}[Y|x_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$\hat{Y}_0 - Y_0 = (\hat{\beta}_0 + \hat{\beta}_1 x_0 - \beta_0 - \beta_1 x_0 - \varepsilon)$

$E(\hat{Y}_0 - Y_0) = 0$

$Var(\hat{Y}_0 - Y_0) = Var(\hat{Y}_0) + Var(Y_0) = Var(\hat{E}[Y|x_0]) + \sigma^2$

$= \sigma^2\left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2} + 1\right]$

Pivotal Quantity: $T = \frac{(\hat{Y}_0 - Y_0) - E(\hat{Y}_0 - Y_0)}{\widehat{Var}(\hat{Y}_0 - Y_0)} =$

$$= \frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2}\right)}} \sim t(n-2)$$

INSTITUTO SUPERIOR TÉCNICO

Campus do Taguspark

Regression Sum-of-Squares

- Sum of Squares total: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Sum of Squares Regression: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Sum of Squares Error (Squares of residuals): $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\underbrace{y_i - \hat{y}_i}_{a} + \underbrace{\hat{y}_i - \bar{y}}_{b})^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)}_{=0 \text{ under the validity of the model}}$$

$$SST = SSR + SSE$$

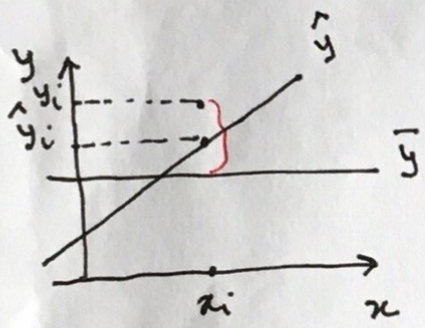
Coefficient of determination $R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

gives the Amount of variation in y's that is explained by the linear relationship with x_i

$0 \le r^2 \le 1$

$r^2 \to 1$ (model good for this data)

$r^2 \to 0$ (not good model for this data)



$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\underbrace{y_i - \hat{y}_i}_{a} + \underbrace{\hat{y}_i - \bar{y}}_{b})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{=0 \text{ under the validity of the model}}$$

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

$$\boxed{SST = SSE + SSR}$$

(total) (residuals) (regression)

Av. Prof. Dr. Cavaco Silva
2744-016 PORTO SALVO
Tel.: 21 423 32 00
Fax: 21 423 32 68
http://www.tagus.ist.utl.pt

Cont. nº 501 507 930
CAE 80 300

TÉCNICO LISBOA

Example: Pizza Data

The owner of a Pizza restaurant chain believes that if the restaurant is located near a college campus then there is a linear relationship between sales and the size of the student population.

Sample of 10 Pizza restaurants located near college campus

Population (1000's): x	2	6	8	8	12	16	20	20	22	26
Sales ($1000's): y	58	105	88	118	117	137	157	169	149	202

$\sum_{i=1}^{10} x_i = 140$; $\sum_{i=1}^{10} y_i = 1300$; $\sum_{i=1}^{10} x_i^2 = 2528$; $\sum_{i=1}^{10} y_i^2 = 184730$; $\sum_{i=1}^{10} x_i y_i = 21040$

a) Estimate the least-squares regression line

$\bar{x} = 14$; $\bar{y} = 130$; $\sum_{i=1}^{10} x_i^2 - n\bar{x}^2 = 2528 - 10 \times 14^2 = 568$

$\sum_{i=1}^{10} x_i y_i - n\bar{x}\bar{y} = 21040 - 10 \times 14 \times 130 = 2840$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{2840}{568} = 5$$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 130 - 5 \times 14 = 60$ $\quad \hat{y} = 60 + 5x$

b) $\hat{E}[Y|x=9] = 60 + 5 \times 9 = 105$; $\hat{E}[Y|x=27]$ → extrapolation

c) Estimate the variance of the sales variable

$$\hat{Var}(Y) = \hat{\sigma}^2 = \frac{1}{8}\left[\left(\sum y_i^2 - n\bar{y}^2\right) - 5^2 \times 568\right] = \frac{15730 - 5^2 \times 568}{8} = \frac{1530}{8}$$

$$= 191.25$$

$\sum y_i^2 - n\bar{y}^2 = 184730 - 10 \times 130^2 = 15730$

d) Estimate the residual of the second observation

$y_2 = 105$ $\quad \hat{y}_2 = 60 + 5 \times 6 = 90$ $\quad e_2 = 105 - 90 = 15$

e) construct a 90% CI for β_1

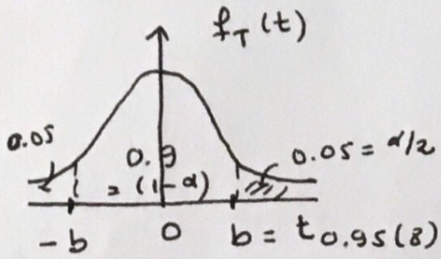
$CI_{90\%}(\beta_1) = ?$

Pivotal quantity variable: $T = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \sim t_{(8)}$

TÉCNICO LISBOA

$P(a \leq T \leq b) = 0.9$

symmetrical interval : $a = -b$



$= \text{quantile } 0.95$

$= 1.860$

$$P\left(-1.860 \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \leq 1.860\right) = 0.95$$

$$\vdots$$

$$P\left(\underbrace{\hat{\beta}_1} - 1.860 \times \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}} \leq \beta_1 \leq \hat{\beta}_1 + 1.860 \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}\right) = 0.9$$

random variable

$CI(\beta_1) = [3.921; 6.079]$ 90%

d) test: $H_0: \beta_1 = 0$ (not significant) vs $H_1: \beta_1 \neq 0$ (significant) significance level 10%.

null hypothesis alternative hypothesis

Very important test : If the null hypothesis is not rejected the model can be use since the variable x doesn't explain the expected value of Y.

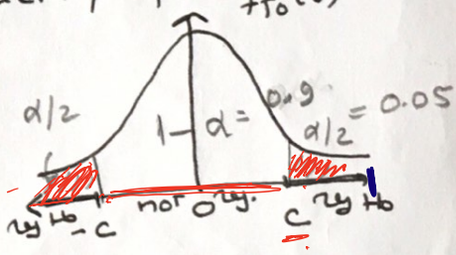
test the significance of the regression model

two sided test

Pivotal quantity: $T = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \sim t_{(8)}$

test statistic: under the validity of H_0

$$t_0 = T|_{H_0} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \sim t_{(8)}$$



α = significance level $= P(\text{reject } H_0 \mid H_0 \text{ is true}) = P(|T_0| > c) = 0.1$

$\Rightarrow c = t_{0.95(8)} = F^{-1}_{t(8)}(0.95) = 1.860$

critical region: $C.R =]-\infty, -1.860[\cup]1.860; +\infty[$

test statistic observed value: $t_0 = \dfrac{5 - 0}{\sqrt{\frac{191.25}{568}}} \simeq 8.617$

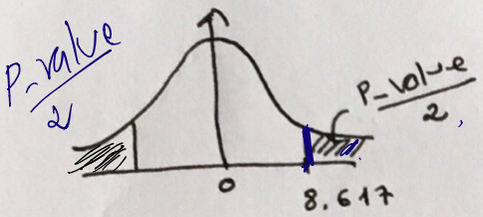
Decision: Since $t_0 \in C.R$, for $\alpha = 0.1$, reject H_0, i.e, for this sample it looks like that the size of the student population explain the expected value of sale (there is a linear relationship between sale, and the student population size)

TÉCNICO LISBOA

Alternative: not fix significance level

P-value of the test:

statistic observed value: $t_0 \cong 8.617$



$$P\text{-value} = 2\left(P(T_0 \geq 8.617)\right) = 2\left(1 - F_{T_0}(8.617)\right) =$$

table: $F_{t(8)}(5.041) = 0.9995$

$$0.9995 < F_{t_8}(8.617) < 1 \iff 0 < 1 - F_{T_0}(8.617) < (1 - 0.9995)$$

$$\iff 0 < P\text{-value} < 0.001$$

$\Rightarrow \forall \alpha \geq 0.001$ rej H_0 $\Rightarrow$ rej H_0 $\forall \alpha$ usual

e) test: $H_0: \beta_0 = 75$ vs $H_1: \beta_0 < 75$ (one sided test)
Left tailed test

use $\alpha = 0.05$

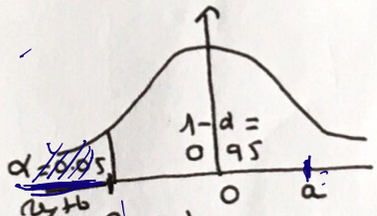
test statistic: $T_0 = \dfrac{\hat{\beta}_0 - 75}{\sqrt{\widehat{var}(\hat{\beta}_0)}} \sim t(8)$ $\qquad \widehat{var}(\hat{\beta}_0) = 85.1197$

observed value $t_0 = \dfrac{60 - 75}{\sqrt{85.1197}} \approx -1.626$

critical region:

$C.R =]-\infty, -1.860[$

since $t_0 \notin C.R.$ not rej. H_0



$a = t_{0.05}(8) = -t_{0.95}(8) = -1.860$