EXERCISE 2

29TH October 2024

Luca Donese - l.donese1@campus.unimib.it

2 Computer repair data

A computer repair company is interested in knowing the relationship between the duration of interventions (measured in minutes) and the number of electronic components to be replaced or repaired. Therefore, a simple linear regression model was considered to explain the duration in minutes of interventions (y) as a function of the number of units (x) to be replaced.
A sample of 14 interventions provided the following data: $\bar{y} = 95.768$, $\bar{x} = 6$, $\sum_{i=1}^{14}(y_i - \bar{y})^2 = 31108.357$, and $\sum_{i=1}^{14}(x_i - \bar{x})^2 = 114$. The model provides a coefficient of determination $R^2 = 0.984$.

Exercise 2.1

Starting from the data, compute the maximum likelihood estimates of β_1 and β_2. Then, write the equation of the estimated linear regression model.

Exercise 2.2

Find the estimate for the variance σ^2 using the decomposition of the total sum of squares. Through a valid test, verify the goodness of fit at 5% significance level.

Exercise 2.3

Given the standard errors (S.E.) of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$, which correspond to $\sqrt{\hat{Var}(\hat{\beta}_1)} = 4.014$ and $\sqrt{\hat{Var}(\hat{\beta}_2)} = 0.604$. Through a valid test (at 5 % significance level), verify if the coefficients β_1 and β_2 are significant (you can use the following t-table for computing p-values).

Exercise 2.4

Given the ex. 2.2, is there any statistical test in the exercise 2.3 that might be unnecessary?

2.1) In the Gaussian linear model we have that the errors are normally distributed:

$$\varepsilon_i \sim N(0, \sigma^2)$$

Moreover, as a consequence, the $y_1, \dots, y_n$ are normally distibuted

$$y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$$

We have that $y_1, \dots, y_n$ are independent, but not identically distributed.

Since we have distributive assumptions we can estimate $\theta = (\beta_1, \beta_2, \sigma^2)$ via maximum likelihood estimation.

By maximizing the log-likelihood function for $\theta \in \Theta = \mathbb{R}^2 \times (0, +\infty)$ we can get the following maximum likelihood estimate for β_1 and β_2:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \quad \text{and} \quad \hat{\beta}_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Since we don't have $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ we make use of the following relationship:

$$R^2 = r_{xy}^2 = \left(\frac{S_{xy}}{S_x S_y} \right)^2$$

$$\Rightarrow 0.984 = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

$$S_{xy} = \sqrt{0.984} \cdot \sqrt{\frac{114}{14} \cdot \frac{31108.357}{14}} = \underline{133.43} \qquad S_x = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = S_{xy} \cdot n = 133.43 \cdot 14 = 1868.05$$

Hence,

$$\hat{\beta}_2 = \frac{1868.05}{114} = 16.386$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = 95.786 - 16.386 \cdot 6 = -2.532$$

Thus the estimated model is:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i = -2.532 + 16.386 x_i$$

2.2)

The unbiased estimate S^2 of σ^2 can be computed by using:

$$S^2 = \frac{SSE}{n-2}.$$

Using the decomposition we know

$$SST = SSR + SSE,$$

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

Then we have

$$SSR = SST \cdot R^2 = 31108.357 \cdot 0.984 = 30610.623$$

$$SSE = SST - SSR = 31108.357 - 30610.623 = 497.734$$

and finally we can compute

$$S^2 = \frac{497.734}{14-2} = 41.478$$

We can test the goodness of fit of the model with the following test

$$\begin{cases} H_0: R^2 = 0 \\ H_1: R^2 > 0 \end{cases}$$

where the test statistic is:

$$F = \frac{SSR / 1}{SSE / (n-2)} = \frac{30610.623}{41.478} = 737.999$$

The p-value is:

$$\alpha^{obs} = P\left(F_{1,12} > 737.999\right) = 3.81 \cdot 10^{-12} < 0.05$$

$\Rightarrow$ we reject $H_0: R^2 = 0$. The model has a good fit

2.3) $\sqrt{\hat{var}(\hat{\beta}_1)} = 4.014 \quad \sqrt{\hat{var}(\hat{\beta}_2)} = 0.604$

- β_1

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases} \qquad \beta_1 > 0 \text{ (one-sided test)}$$

This is a two-sided test because $H_1: \beta_1 \neq 0$. We have the following test statistic:

$$T = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \overset{H_0}{\sim} t_{n-2}, \text{ where } t^{obs} = \frac{-2.532}{4.014} = -0.631$$

The p-value is:

$$\alpha^{obs} = P(|T_{n-2}| > |t^{obs}|) = 2 \cdot P(T_{12} \leq -0.631) = 2 \cdot 0.5399 > 0.05$$

$\Rightarrow$ we don't reject the hypothesis that β_1 is not significant

- β_2

$$\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$$

Here we have the same procedure

$$t^{obs} = \frac{\hat{\beta}_2}{\sqrt{\hat{var}(\hat{\beta}_2)}} = \frac{16.386}{0.604} = 27.129$$

The p-value is:

$$P(|T_{n-2}| \geq |t^{obs}|) = 2 \cdot P(T_{12} \leq -27.129) = 2 \cdot 1.9 \cdot 10^{-12} < 0.05$$

$\Rightarrow$ we reject the hypothesis that β_2 is not significant

2.4)

Given the result of the test in 2.2) we could avoid the test for β_2 in 2.3).
In case of a simple linear model in fact testing $R^2=0$ and $\beta_2=0$ is the same

3 Bacteria mortality data

Suppose we want to analyze bacterial mortality (y) as a function of radiation exposure (x). The output of a linear regression of y as a function of x is partially summarized in the table below:

Table 1: Output of a linear regression.

Variable	Coefficients	S.E.	T-value	P-value
Constant	49.162	22.76		
Exposure (x)	-19.46		-7.79	<0.0001

where $n = 15$, $R^2 = 0.823$ and $S = 41.83$.

Exercise 3.1

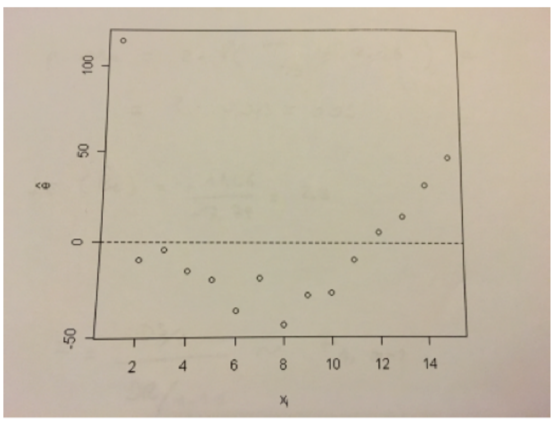
Complete the Table 1 writing the equations you should use.

Exercise 3.2

Through a valid statistical test, evaluate the goodness of fit of the model.

Exercise 3.3

Discuss about the hypothesis related to the model, looking the following residual plot.



3.1)

Variable	Coefficients	S.E.	T-value	P-value
Constant	49.162	22.76	2.16	0.05
Exposure (x)	-19.46	2.498	-7.79	<0.0001

~ β_1

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

- $$t_1^{obs} = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} = \frac{49.162}{22.76} = 2.16$$

- p-value

$$\alpha^{obs} = P(|T_{13}| > |t_1^{obs}|) = 2 \cdot P(T_{13} \leq -2.16) = 0.05$$

- for 1% significance level we don't reject H_0
- for 10% significance level we reject H_0

~ β_2

$$\hat{SE}(\hat{\beta}_2) = \sqrt{\hat{Var}(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{t_2^{obs}} = \frac{-19.46}{-7.79} = 2.498$$

3.2)

To evaluate the goodness of fit, we can use the F-test

$$\begin{cases} H_0: R^2 = 0 \\ H_1: R^2 > 0 \end{cases}$$

We have the following test statistic

$$F = \frac{SSR/1}{SSE/(n-2)} \overset{H_0}{\sim} F_{1,n-2}$$

To find f^{obs} we need to compute SSR and SSE

$$SSE = (n-2)\,s^2 = 13 \cdot (41.83)^2 = 22746.7357$$

$$SST = \frac{SSE}{1-R^2} = \frac{22746.7357}{0.177} = 128512.634$$

$$SSR = SST - SSE = 128512.6311 - 22746.7357 = 105765.8954$$

Hence,

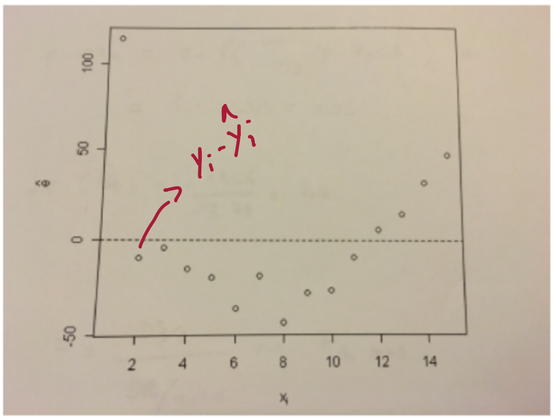
$$f^{obs} = \frac{105765.8954}{22746.7357/15-2} = 60.446$$

and the p-value is:

$$\alpha^{obs} = P(F_{1,13} > 60.446) = 3.05 \cdot 10^{-6} < 0.01$$

=> we reject $H_0: R^2 = 0$ for any significance levels

3.3)



- We can observe that the residuals are not centered around 0, while their average should be 0

- The plot suggest a systematic behaviour (a relationship among the residuals)

  -> By including transformations of the covariates we can solve this issue

4 Grades data

In 2011 among 62 adolescents, the variables x "daily hours spent on average to video games" and y "average report card grade" were observed. We proposed a gaussian simple linear model with y, as response variable, and obtained the following estimates: $\hat{\beta}_1 = 7.4$, $\hat{\beta}_2 = -0.48$, $SSE = 223$, $\frac{Var(\hat{\beta}_1)}{\sigma^2} = 3.43$ and $\frac{Var(\hat{\beta}_2)}{\sigma^2} = 0.07$.

Exercise 4.1

Compute the OLS estimates for β_1 and β_2 and provide an explanation.

Exercise 4.2

Through a valid test, use p-values to evaluate if the coefficients are significant (~~you can use the t-table below for computing p-values~~). Then, evaluate the goodness of fit using p-values.

4.1)

With the assumption of normality for the regression error term, OLS (ordinary least square) estimates correspond to ML (maximum likelihood) estimates

$$\Rightarrow \hat{\beta}_1^{OLS} = \hat{\beta}_1^{ML} = 7.4$$

$$\hat{\beta}_2^{OLS} = \hat{\beta}_2^{ML} = -0.48$$

4.2)

- Inference on β_1

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases} \rightarrow \text{it is a two-sided test}$$

$$T_1 = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \overset{H_0}{\sim} t_{n-2}$$

$$t_1^{obs} = \frac{7.4}{\hat{SE}(\hat{\beta}_1)}$$

$$\frac{Var(\hat{\beta}_1)}{\sigma^2} \cdot S^2$$

Since we know that $\frac{\hat{Var}(\hat{B}_1)}{\hat{\sigma}^2} = 3.43$, we can get $\hat{Var}(\hat{B}_1)$ by:

$$\frac{\hat{Var}(\hat{B}_1)}{\hat{\sigma}^2} \cdot \hat{\sigma}^2$$ we get this estimate with S^2

$$S^2 = \frac{SSE}{n-2} = \frac{223}{60} = 3.7163$$

Then we have

$$t_1^{obs} = \frac{7.4}{\sqrt{3.7163 \cdot 3.43}} = 2.0726$$

We can compute the p-value:

$$\alpha^{obs} = P_{H_0}\left(|T_{n-2}| > |t_1^{obs}|\right) = 2 \cdot P\left(T_{n-2} \leq -2.0726\right)$$
$$= 0.0425 \leq 0.05$$

=> we reject $H_0: \beta_1 = 0$

- Inference on β_2

$$\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases} \qquad t_2^{obs} = \frac{-0.48}{\sqrt{3.7167 \cdot 0.07}} = -0.941$$

compute the p-value

$$\alpha^{obs} = P\left(|T_{60}| \geq |t_2^{obs}|\right) = 2 \cdot P\left(T_{60} \leq -0.941\right)$$
$$= 0.35$$

=> we cannot reject $H_0: \beta_2 = 0$ for any significance levels

- Inference on R^2

We can use the following equations to compute SSE and SSR and then f^{obs}.

We know that

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sigma^2}{n \cdot S_x^2}$$

Then we have that

$$\frac{Var(\hat{\beta}_2)}{\sigma^2} = \frac{1}{n \cdot S_x^2} = 0.07$$

Now exploiting the relationship among the t-test and F-test (only for simple linear model):

$$t_2^2 = f^{obs} \qquad \frac{\hat{\beta}_2}{\hat{Var}(\hat{\beta}_2)} = \frac{SSR}{SSE/(n-2)}$$

$$SSR = \hat{\beta}_2 \cdot \frac{SSE}{n-2} \cdot \frac{1}{\hat{Var}(\hat{\beta}_2)} = \hat{\beta}_2 \cdot \frac{n \cdot \hat{\sigma}^2}{(n-2)} \cdot \frac{n \cdot S_x^2}{S^2} =$$

$$SSE = \sum_{i=1}^{n}(y_i-\bar{y})^2 = n \cdot \hat{\sigma}^2$$

$$\frac{Var(\hat{\beta}_2)}{\sigma^2} = \frac{1}{n \cdot S_x^2} \Rightarrow Var(\hat{\beta}_2) = \frac{\boxed{S^2}}{n \cdot S_x^2}$$

estimator for σ^2

$$= \hat{\beta}_2 \cdot \frac{n \cdot \hat{\sigma}^2}{(n-2)} \cdot \frac{n \cdot S_x^2}{S^2} \qquad S^2 = \frac{SSE}{(n-2)} = \frac{n \cdot \hat{\sigma}^2}{(n-2)}$$

$$= \hat{\beta}_2 \cdot \frac{n \cancel{\hat{\sigma}^2}}{\cancel{(n-2)}} \cdot \frac{n \cdot s^2_x \cdot \cancel{(n-2)}}{n \cdot \cancel{\hat{\sigma}^2}} = n \cdot s_x^2 \cdot \hat{\beta}_2 =$$

$$= (-0.48)^2 \cdot (0.07)^{-1} = 3.291$$

$$f^{obs} = \frac{3.291}{223/60} = 0.885$$

compute the p-value.

$\alpha^{obs} = P(F_{1,60} > 0.885) = 0.3506 \Rightarrow$ we cannot reject $H_0 : R^2 = 0$

Finally we can also find SST:

$$SST = SSE + SSR = 3.291 + 223 = 226.291$$

5 Additional exercise

A linear regression model was estimated on 82 units. Complete the tables below and specify the hypothesis, the test statistic and p-value for inference.

Table 2: Analysis of variance.

Deviance	Sum of squares	d.o.f.	F	p-value
Regression	3589.6		10.21	
Residual		-	-	-
Total		-	-	-

where d.o.f. means degree of freedom.

Table 3: Output of a linear regression.

Variable	Coefficients	S.E.	T-value	P-value
Constant	12.7		0.82	
x_1	-19.3			0.002

Table 2: Analysis of variance.

Deviance	Sum of squares	d.o.f.	F	p-value
Regression	3589.6		10.21	0.0014
Residual	28126.15	-	-	-
Total	31715.75	-	-	-

$\alpha^{obs} = P(F_{1,80} > 10.21) = 0.0014 \Rightarrow$ we can reject $H_0: R^2 = 0$

$$f^{obs} = 10.21 = \frac{SSR/1}{SSE/(n-2)} = \frac{3589.6}{SSE/80} \Rightarrow SSE = \frac{3589.6 \cdot 80}{10.21} = 28126.15$$

$SST = 3589.6 + 28126.15 = 31715.75$

Table 3: Output of a linear regression.

	Variable	Coefficients	S.E.	T-value	P-value
β_1	Constant	12.7	15.4878	0.82	0.41
β_2	x_1	-19.3	6.04	-3.195	0.002

- Inference on the constant

$$t_1^{obs} = 0.82 = \frac{\hat{\beta}_1}{\hat{S.E.}(\hat{\beta}_1)} \Rightarrow S.E(\hat{\beta}_1) = \frac{\hat{\beta}_1}{0.82} = \frac{12.7}{0.82} = 15.4878$$

$$\alpha^{obs} = P(|T_{80}| \geq |0.82|) = 2P(T_{80} < -0.82) \approx 0.41$$

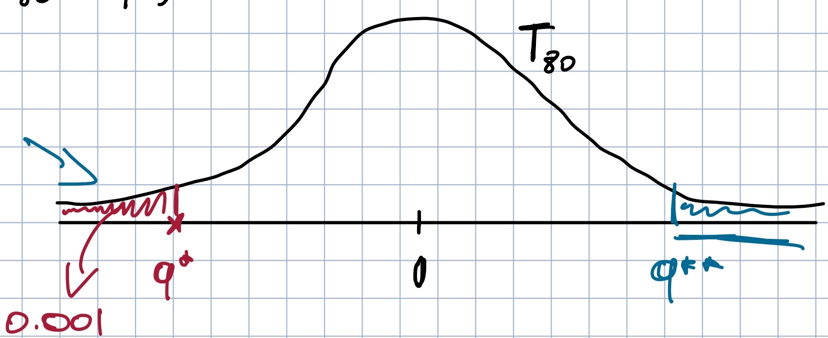
$\Rightarrow$ we cannot reject $H_0: \beta_1 = 0$

- Inference on β_2

$$\alpha^{obs} = 0.002 = 2 \cdot P(T_{80} \leq q^*)$$

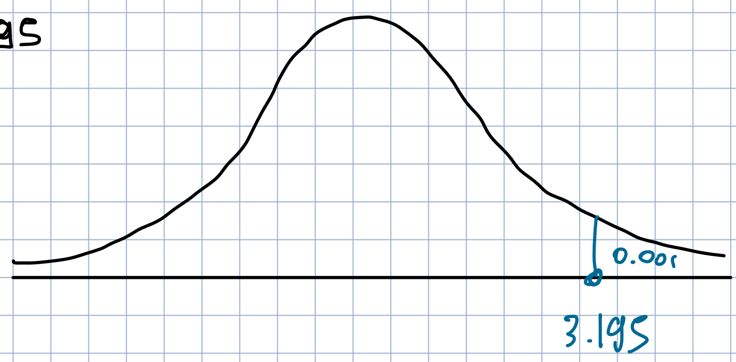
We need to look on the table for $t_{80,0.001}$ because

$$q^*: P(T_{80} < q^*) = 0.001$$



$$P(T_{80} > q^{**}) = 0.99$$

$t_{80;\,0.001} = -3.195$



$1-\alpha = 0.999$

$t_{80,p}$ 3.195

$$T = \frac{\hat{\beta}_2}{\hat{SE}(\hat{\beta}_2)}$$

$$\Rightarrow \hat{SE}(\hat{\beta}_2) = \frac{-19.3}{-3.195} = 6.04$$