Exercise 1 (Simple linear model and Simple Gaussian linear model)

22nd October 2024

Luca Danese - l.danese1@campus.uniurb.it

↳ for any questions contact me!

Schedule

1. 22-10-2024 ~ 12.45 - 15.15
2. 29-10-2024 ~ 12.45 - 15.15
3. 5-11-2024 ~ 12.45 - 15.15
4. 19-11-2024 ~ 12.45 - 15.15
5. 26-11-2024 ~ 12.45 - 15.15
6. 3.12.2024 ~ 12.45 - 15.15
7. 5.12.2024 ~ 11.30 - 13.45
8. 10.12.2024 ~ 12.45 - 15.15

Check out the webpage of the course for any changes!

---

1 Mother and Daughter heights data

Let consider a sample of data with $n = 11$ observations (Table 1) with two variables:

- **mother's height** x (independent variable);
- **daughter's height** y (dependent variable).

Table 1: Mother and Daughter heights data: data are expressed in centimeters.

	1	2	3	4	5	6	7	8	9	10	11
x	153.7	156.7	173.5	157.0	161.8	140.7	179.8	150.9	154.4	162.3	166.6
y	163.1	159.5	169.4	158.0	164.3	150.0	170.3	158.9	161.5	160.8	160.6

We would like to find out if there exists a relationship between these two variables.

Exercise 1.1

Starting from the data (in Table 1), write the equation of the simple linear regression model. Compute $\bar{x}$, $\bar{y}$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$ and, then, find the estimates of the linear model parameters.

1.1)

The simple linear model for variables (x, y) has the following form:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \,, \quad i = 1, \dots, n$$

where (β_1, β_2) are the coefficients (sistematic component) and ε_i is the error term (stochastic component).

We should make the following assumptions in order to define the model:

- ASSUMPTIONS on the independent variables
  1. $x_1, \dots, x_n$ fixed and non-stochastic
  2. the x_i's cannot be equal (sample variance $\neq 0$)
- ASSUMPTIONS on the stochastic components
  1. $\mathbb{E}[\varepsilon_i] = 0 \quad i = 1, \dots, n$ → ABSENCE OF SYSTEMATIC ERROR
  2. $\text{Var}[\varepsilon_i] = \sigma^2 > 0 \quad i = 1, \dots, n$ → HOMOSCHEDASTICITY OF THE ERRORS
  3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$ → ERRORS ARE NOT CORRELATED

We want to estimate β_1 and β_2 is such a way that the sum of SQUARED RESIDUALS is minimised.

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{(\beta_1, \beta_2) \in \mathbb{R}^2}{\operatorname{argmin}} \, S(\beta_1, \beta_2) \rightarrow \text{LEAST SQUARE ESTIMATE of } (\beta_1, \beta_2)$$

The least square estimate of (β_1, β_2) is:

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

β_1, β_2 TRUE PARAM.

$\hat{\beta}_1, \hat{\beta}_2$ ESTIMATE

$\hat{\beta}_1, \hat{\beta}_2$ ESTIMATOR

$$\bar{x} = \frac{153.7 + 156.7 + 173.5 + 157 + 161.8 + 140.7 + 179.8 + 150.9 + 154.4 + 162.3 + 166.6}{11} = 159.76$$

$$\bar{y} = \frac{163.1 + 159.5 + 169.4 + 158.0 + 164.3 + 150 + 170.3 + 158.9 + 161.5 + 160.8 + 160.6}{11} = 161.49$$

$$\sum_{i=1}^{n} x_i y_i = 284335$$

$$\sum_{i=1}^{n} x_i^2 = 281941$$

Now we can estimate β_1 and β_2

- $$\hat{\beta}_2 = \frac{284335 - 11 \cdot 159.76 \cdot 161.49}{281941 - 11 \cdot (159.76)^2} = 0.45473$$

- $$\hat{\beta}_1 = 161.49 - 0.45473 \cdot 159.76 = 88.842$$

$$\hat{y}_i = 88.842 + 0.45473 \cdot x_i$$

Exercise 1.2

Given the results of the previous exercise, compute the fitted values for each i. Make a plot involving the observations of the couple (y, x) and the estimated regression line.

Exercise 1.3

Compute the residuals (e_i) and the unbiased estimate (S^2) of the variance σ^2. Then, find

1.2)

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

$\hat{y}_1 = 88.42 + 0.45473 \cdot 153.7 = 158.734$

$\hat{y}_2 = 160.098$

$\hat{y}_3 = 167.738$

$\hat{y}_4 = 160.235$

$\hat{y}_5 = 162.417$

$\hat{y}_6 = 152.823$

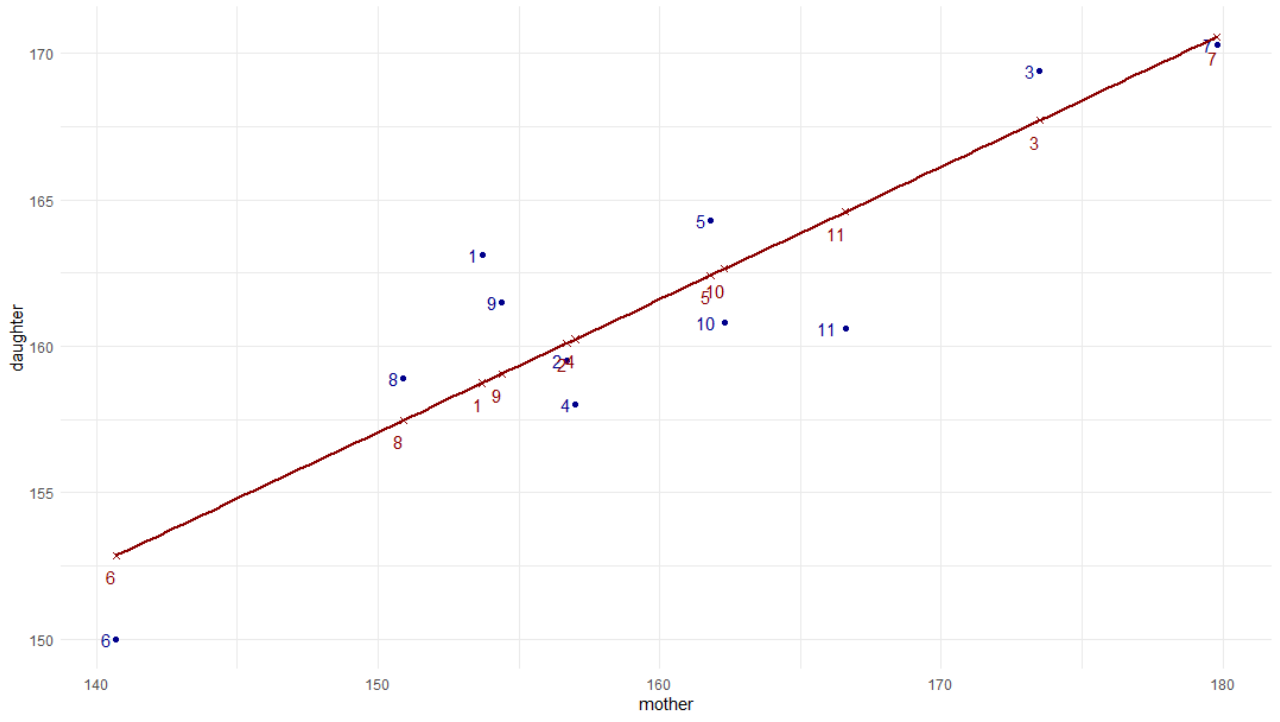
$\hat{y}_7 = 170.102$

$\hat{y}_8 = 157.461$

$\hat{y}_9 = 159.052$

$\hat{y}_{10} = 162.645$

$\hat{y}_{11} = 164.6$



1.3)

$\hat{e}_i = y_i - \hat{y}_i$ → ESTIMATE OF RESIDUAL i

$\hat{e}_1 = 163.1 - 158.734 = 4.366$

$\hat{e}_2 = -0.598$

$\hat{e}_3 = 1.662$

$\hat{e}_4 = -2.235$

$\hat{e}_5 = 1.883$

$\hat{e}_6 = -2.823$

$\hat{e}_7 = -0.332$

$\hat{e}_8 = 1.439$

$\hat{e}_9 = 2.448$

$\hat{e}_{10} = -1.845$

$\hat{e}_{11} = -4$

To get an estimate $\hat{\sigma}^2$ of σ^2 we use the following estimator:

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\cdot\sum_{i=1}^{n}(y_i - \hat{B}_1 - \hat{B}_2 x_i)^2$$

UNBIASED estim. for σ^2

BIASED estimator for σ^2

$$\hat{\sigma}^2 = \frac{1}{11-2}\cdot 66.207 = 7.36$$

We need to get an estimate for $\widehat{var}(\hat{B}_2)$ and $\widehat{var}(\hat{B}_1)$.

$$\widehat{var}(\hat{B}_2) = \frac{S^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{7.36}{472.005} = 0.0063$$

$$\widehat{var}(\hat{B}_1) = S^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) = 7.36 \left(\frac{1}{n} + \frac{25524.42}{472.005} \right) = 160.9582$$

Exercise 1.4

Compute the total sum of squares (SST), the residual sum of squares (SSE) and the regression sum of squares (SSR). Then, find the coefficient of determination R^2. Compute the correlation coefficient r_{xy} and its squared. What happens in this case?

1.4)

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 306.8091$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 240.5676$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 66.24151$$

-> To check if the numbers are correct we can see if $SST = SSE + SSR$

The coefficient R^2 is the proportion of variability of the dependent variable that is predicted by the covariate

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 0.7841$$

The model explains 78% of the total variability of y.

The correlation coefficient measures the strength of the linear relationship between two variables.

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

We need to compute the following quantities:

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = 30.68091$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 117.2005$$

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = 53.0984,$$

so we have

$$r_{xy} = \frac{53.0984}{\sqrt{30.68091}\cdot\sqrt{117.2005}} = 0.88549$$

In the simple linear model $R^2 = r_{xy}^2$, in fact $0.88549^2 = 0.7841$

Exercise 1.5

Starting from the data (in Table 1), write the equation of the gaussian simple linear regression model together with the associated assumptions. Explain the difference from simple linear regression (make a comparison between the assumptions of ex. 1.1 and this case).

1.5)

Given the above two variables we can define the SIMPLE GAUSSIAN LINEAR MODEL in the following way:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad i = 1,\dots,n$$

$$\text{where } \varepsilon_i \sim N(0,\sigma^2)$$

The SIMPLE GAUSSIAN LINEAR MODEL is based on the following assumptions:

1) Absence of systematic error $\mathbb{E}[\varepsilon_i] = 0 \quad i=1,\dots,n$
2) Homoschedasticity of the errors $\text{Var}[\varepsilon_i] = \sigma^2 \quad i=1,\dots,n$
3) Uncorrelated errors $\text{Cov}(\varepsilon_i, \varepsilon_k) = 0 \quad i \neq k$
4) Gaussian distribution of the errors

Although the assumptions are similar to the one of the simple linear model, the simple gaussian linear model requires that the errors are normally distributed.

Exercise 1.6

Let consider the following system of hypothesis

$$\begin{cases} \text{H0: } \beta_2 = 1 \\ \text{H1: } \beta_2 \neq 1 \end{cases}$$

Compute the t-test and the p-value.

1.6)

To test the statistical hypothesis we need a pivotal quantity for β_2

Remark: a pivotal quantity is a transformation of the data (and of the parameter) whose distribution does not depend on the parameter (hence it is completely known)

For parameter β_2 we consider the following pivotal quantity:

$$\frac{\hat{B}_2 - \beta_2}{\sqrt{Var(\hat{B}_2)}} \overset{H_0}{\sim} N(0,1) \quad \text{where} \quad \hat{B}_2 \sim N(\beta_2, Var(\hat{\beta}_2))$$

Since $Var(\hat{B}_2) = \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})$, so it includes σ^2 that is unknown, we use the following estimator:

$$\hat{Var}(\hat{B}_2) = S^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

and then we get the following pivotal quantity

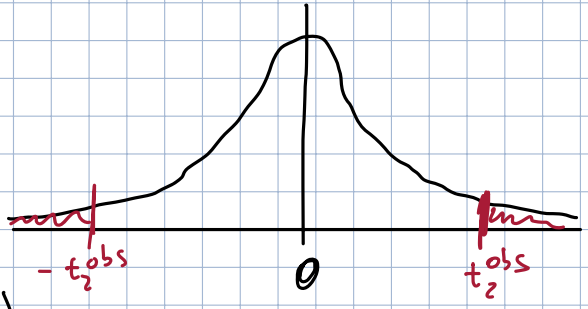
$$T_2 = \frac{\hat{B}_2 - \beta_2}{\sqrt{\hat{V}(\hat{B}_2)}} \overset{H_0}{\sim} t_{n-2}$$

$$t_2^{obs} = \frac{\hat{B}_2 - 1}{\sqrt{\hat{Var}(\hat{B}_2)}} = \frac{0.45473 - 1}{\sqrt{0.00595}} = -7.069$$

We compute the p-value:

$$\alpha^{obs} = P_{H_0}(|T_2| \geq |t_2^{obs}|)$$

$$= P_{H_0}(T_2 \leq -t_2^{obs}) + P_{H_0}(T_2 \geq t_2^{obs})$$



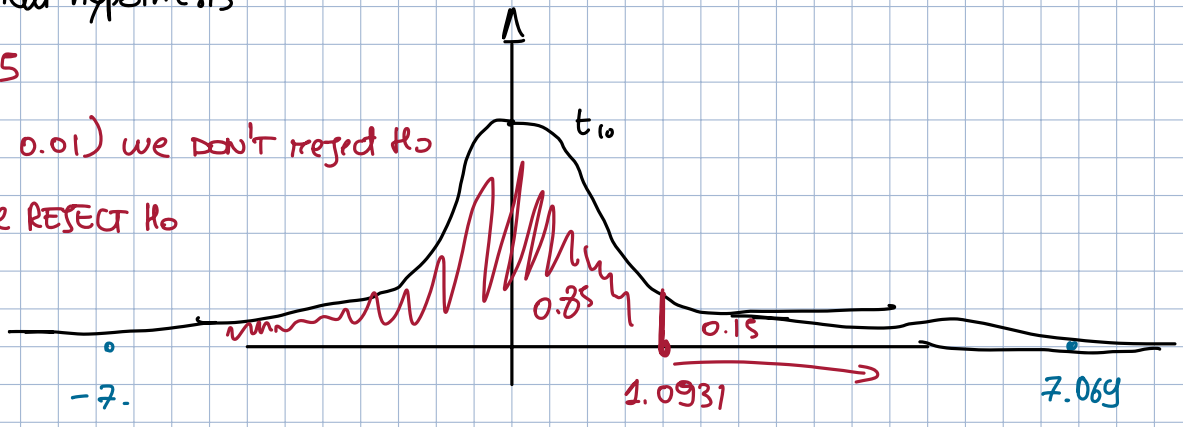
$$= 2\,\mathbb{P}_{H_0}(T_2 \leq -7.069)$$
$$= 2 \cdot 1 \cdot 10^{-5}$$

$1-\alpha$	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995
$t_{1;p}$	1	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567
$t_{2;p}$	0.8165	1.0607	1.3862	1.8856	2.92	4.3027	6.9646	9.9248
$t_{3;p}$	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409
$t_{4;p}$	0.7407	0.941	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041
$t_{5;p}$	0.7267	0.9195	1.1558	1.4759	2.015	2.5706	3.3649	4.0321
$t_{6;p}$	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074
$t_{7;p}$	0.7111	0.896	1.1192	1.4149	1.8946	2.3646	2.998	3.4995
$t_{8;p}$	0.7064	0.8889	1.1081	1.3968	1.8595	2.306	2.8965	3.3554
$t_{9;p}$	0.7027	0.8834	1.0997	1.383	1.8331	2.2622	2.8214	3.2498
$t_{10;p}$	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693
$t_{11;p}$	0.6974	0.8755	1.0877	1.3634	1.7959	2.201	2.7181	3.1058

=> Since α^{obs} is below any conventional significance level (0.1, 0.05, 0.01) we REJECT the null hypothesis

If $\alpha^{obs} = 0.075$

- for (0.05, 0.01) we DON'T reject H_0
- for 0.1 we REJECT H_0



Exercise 1.7 Let consider the following system of hypothesis

$$\begin{cases} \text{H0: } R^2 = 0 \\ \text{H1: } R^2 > 0 \end{cases}$$

Compute the F-test and the p-value. In this case, does an equivalent test exists? Specify the hypothesis, the test statistic and compute the p-value.

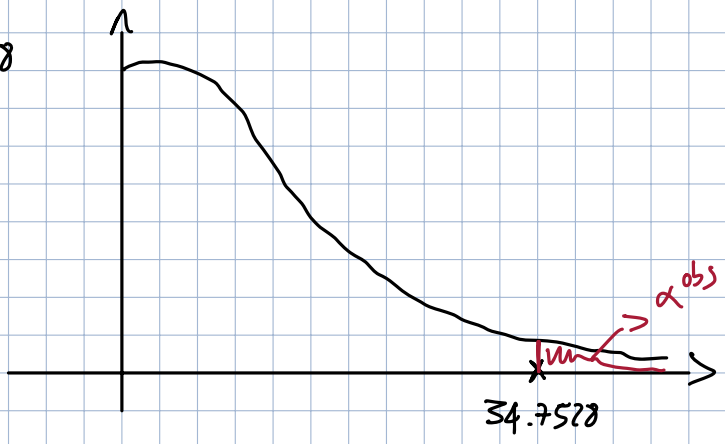
With this system of hypothesis we can test the hypothesis that the model does not explain the variability of y against the hypothesis that it does.

We have the following test statistic:

$$F = \frac{R^2}{1-R^2}(n-2) \overset{H_0}{\sim} F_{1,n-2}$$

$$f_{obs} = \frac{0.79}{1-0.79} \cdot 9 = 34.7528$$

$$\alpha^{obs} = \mathbb{P}_{H_0}(F \geq f_{obs})$$
$$= \mathbb{P}_{H_0}(F \geq 34.7528)$$
$$\approx 1 \cdot 10^{-5}$$



Since α^{obs} is below the conventional significance levels, we reject the null hypothesis

In the simple linear model we have the following relationship:

$$F = \frac{R^2}{1-R^2}(n-2) = T_2^2 = \frac{\hat{B}_2}{\hat{V}(\hat{B}_2)},$$

under the system of hypothesis:

$$\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{cases} \qquad y_i = \beta_1 + \beta_2 x_i$$

$$\text{If } R^2 = 0 \Rightarrow \beta_2 = 0$$

If we compute

$$\alpha^{obs} = \mathbb{P}_{H_0}(|T_2^2| \geq |t_2^{obs\,2}|) = \mathbb{P}_{H_0}(T_2^2 \leq -t_2^{obs\,2}) + \mathbb{P}_{H_0}(T_2^2 \geq t_2^{2\,obs})$$

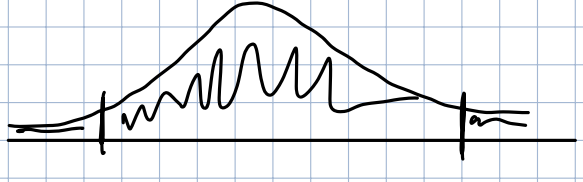
$$\simeq 2 \cdot 1 \cdot 10^{-5} \Rightarrow \text{we reject the null hypothesis}$$

Exercise 1.8

Provide the confidence intervals for β_r, $r = 1, 2$ at level $1 - \alpha = 0.95$.

A confidence interval for β_r can be obtained by considering the stat. test

$$T_r = \frac{\hat{B}_r - \beta_r}{\sqrt{\hat{V}(\hat{B}_r)}} \sim t_{n-2} \qquad r = 1, 2$$



$$1 - \alpha = \mathbb{P}\left(\hat{B}_r - t_{n-2,1-\frac{\alpha}{2}} \cdot \sqrt{\hat{V}(\hat{B}_r)} < \beta_r < \hat{B}_r + t_{n-2,1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{B}_r)}\right)$$

The formula for the confidence interval is:

$$\hat{C} = \hat{B}_n \pm \sqrt{\hat{V}(\hat{B}_n)} \cdot t_{n-2,\,1-\frac{\alpha}{2}}$$

$t_{9,\,0.975} = 2.2622$

$1-\alpha$	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995
$t_{1;p}$	1	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567
$t_{2;p}$	0.8165	1.0607	1.3862	1.8856	2.92	4.3027	6.9646	9.9248
$t_{3;p}$	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409
$t_{4;p}$	0.7407	0.941	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041
$t_{5;p}$	0.7267	0.9195	1.1558	1.4759	2.015	2.5706	3.3649	4.0321
$t_{6;p}$	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074
$t_{7;p}$	0.7111	0.896	1.1192	1.4149	1.8946	2.3646	2.998	3.4995
$t_{8;p}$	0.7064	0.8889	1.1081	1.3968	1.8595	2.306	2.8965	3.3554
$t_{9;p}$	0.7027	0.8834	1.0997	1.383	1.8331	2.2622	2.8214	3.2498
$t_{10;p}$	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693
$t_{11;p}$	0.6974	0.8755	1.0877	1.3634	1.7959	2.201	2.7181	3.1058

$$\hat{C}(\beta_1) = (88.842 - 2.262 \cdot \sqrt{152.423},\ 88.842 + 2.262 \cdot \sqrt{152.423}) = (60.915, 116.769)$$

$$\hat{C}(\beta_2) = (0.45473 - 2.262 \cdot \sqrt{0.00595},\ 0.45473 + 2.262 \cdot \sqrt{0.00595}) = (0.280, 0.629)$$

With a confidence of 95%. β_1 lies in $(60.915, 116.769)$ and β_2 in $(0.280, 0.629)$

Exercise 1.9

During the theoretical lectures, you exploited the relationship between R^2 and the t-test to prove the equivalence among two statistical tests in the case of simple linear regression. Provide the formula and verify that it holds with the data. (In the exercise 1.4, you have already computed the R^2 and the components of deviance decomposition.)

(see theory)