EXERCISE 3

5th November 2024

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

1 Exercise 1

Among 100 elementary school children, data about daily time spent in front of the TV (TV variable), gender (G variable) and time spent answering to a logic-mathematics question (T variable) were collected.

Exercise 1.1

Specify an appropriate regression model for the response variable T.

Exercise 1.2

The model specified at ex 1.1 provides the following values: $SST = 985$, $R^2 = 0.51$, $S.E.(\hat{B}_2) = 0.9$, $S.E.(\hat{B}_3) = 2.3$, where $\hat{B}_2$ and $\hat{B}_3$ are the maximum-likelihood estimators of the regression coefficients for TV variable and G variable, and $\hat{\rho}_{(\hat{B}_2,\hat{B}_3)} = 0.68$.
Perform a statistical test to check the goodness of fit of our model by employing the p-value (specify also the null hypothesis).

Exercise 1.3

Identify all the elements of the matrix $(X^TX)^{-1}$ which can be computed within the available data (specified in the above exercises).

1.1

The regression model for variable T corresponds to:

$$T_i = \beta_1 + \beta_2 \cdot TV_i + \beta_3 D_i + \varepsilon_i \quad i = 1, \dots, 200$$

where D_i is a dummy variable such that:

$$D_i = \begin{cases} 1, & \text{male} \\ 0, & \text{female} \end{cases}$$

- $T_i = \beta_1 + \beta_2 TV_i + \beta_3 D_i + \varepsilon_i$ — MODEL FOR MALE
- $T_i = \beta_1 + \beta_2 TV_i + \varepsilon_i$ → here $D_i = 0$ — MODEL FOR FEMALE

The assumptions related to the multiple linear regression model wit p covariates, is:

i) Linearity: conditionally on $X_{i1} = x_{i1}, \dots, X_{ip} = x_{ip}$ the model is

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i \quad , \quad i = 1, \dots, n$$

ii) $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad i = 1, \dots, n$

iii) The vectors $X_j \in \mathbb{R}^n$, $j = 1, \dots, p$, are linearly independent

$\Rightarrow$ Absence of multicollinearity of the covariates

$\Rightarrow$ rank $(X) = p$ (full rank)

1.2

The statistical test of goodness of fit is based on the following system of hypothesis

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

The test statistic is the following:

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

We need to find SSR and SSE.

$$SSR = SST \cdot R^2 = 985 \cdot 0.51 = 502.35$$

$$SSE = SST - SSR = SST\,(1 - R^2) = 482.65$$

Hence,

$$F^{obs} = \frac{502.35 \,/\, (3-1)}{482.65 \,/\, (100-3)} = 50.479$$

then we can compute the p-value

$$\alpha^{obs} = \mathbb{P}\left(F_{2,97} \geq 50.479\right) = 8.88 \cdot 10^{-16} \leq 0.01$$

$\Rightarrow$ we reject H_0, we do not reject the model

$$(X^TX)^{-1} = \begin{bmatrix} C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,1} & C_{3,2} & C_{3,3} \end{bmatrix}$$

- $C_{2,2} = \dfrac{\hat{Var}(\hat{B}_2)}{S^2} = \dfrac{(0.9)^2}{482.65/97} \cdot \dfrac{0.81}{4.9758} = 0.1628$

$S^2 = SSE / (n-p)$

- $C_{3,3} = \dfrac{\hat{Var}(\hat{B}_3)}{S^2} = \dfrac{(2.3)^2}{4.9758} = 1.0631$

- $C_{2,3} = C_{3,2} = \dfrac{\hat{Cov}(\hat{B}_2, \hat{B}_3)}{S^2} = \dfrac{0.68\ (0.9 \cdot 2.3)}{4.9758} = 0.2829$

$$\rho(\hat{B}_2, \hat{B}_3) = \frac{\hat{Cov}(\hat{B}_2, \hat{B}_3)}{S.E.(\hat{B}_2)\, S.E.(\hat{B}_3)} \Rightarrow \hat{Cov}(\hat{B}_2, \hat{B}_3) = \rho(\hat{B}_2, \hat{B}_3) \cdot S.E.(\hat{B}_2)\, S.E.(\hat{B}_3)$$

$$(X^TX)^{-1} = \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & 0.1628 & 0.2829 \\ \cdot & 0.2829 & 1.0631 \end{bmatrix}$$

we don't have elements about β_1, so the fins row and the first column are empty

2 Exercise 2

Among 100 households in northern Italy, the variables Y = monthly expenditures for foods (in hundreds of euros), X_1 = monthly household income (in hundreds of euros), X_2 = number of household members, and X_3 = type of diet (divided in *"vegetarian"*, *"vegan"* and *"other"*) were collected.

Exercise 2.1

Specify a multiple linear regression model for the response variable Y.

Exercise 2.2

Let $c_{j,h}$ be the elements of the matrix $(X^TX)^{-1}$, where $c_{2,2} = 0.02$, $c_{3,3} = 0.07$, $c_{2,3} = -0.02$. Let also consider that $\hat{\beta}_2 = 0.5$, $\hat{\beta}_3 = 0.8$ and $SSE = 300$.
Evaluate the significance of β_2 and try to interpret the value of $\hat{\beta}_2$.

Exercise 2.3

Find the probability distribution of $\hat{\beta}_2 - \hat{\beta}_3$ and build a statistical test based on the null hypothesis $H0 : \beta_2 = \beta_3$ (at 1% significance level).

Exercise 2.4

Knowing $SSE = 282$ for a model which includes an interaction between the dummy variable (referred to the type *"vegeterian"*) and X_1, decide the best model through an appropriate test (specify hypothesis, test statistic and p-value).

2.1

Let's consider the following model (with the assumptions of ex. 1.1)

$$Y_i = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 D_{1i} + \beta_5 D_{2i} + \varepsilon_i ,$$

where

$$D_{1i} = \begin{cases} 1 & \text{if } X_{3i} = \text{"Vegeterian"} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{if } X_{3i} = \text{"Vegan"} \\ 0 & \text{otherwise} \end{cases}$$

X_{3i}	D_{1i}	D_{2i}
VEGETERIAN	1	0
VEGAN	0	1
OTHER	0	0

2.2

The system of hypothesis is:

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

and the test statistic is:

$$T_2 = \frac{\hat{\beta}_2}{s.e.(\hat{\beta}_2)} \overset{H_0}{\sim} t_{n-p}$$

$$\rightsquigarrow s.e.(\hat{\beta}_2) = S^2 \cdot C_{22}$$

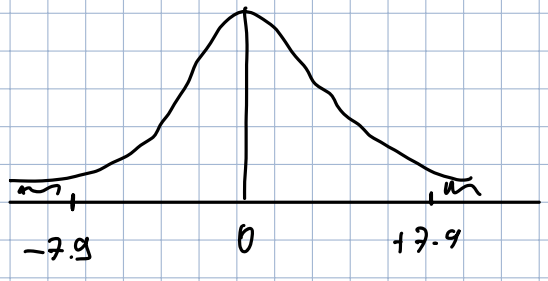
$$S^2 = SSE/(n-p) = 300/(100-5) = 3.15$$

$$s.e.(\hat{\beta}_2) = 3.15 \cdot 0.02 = 0.063$$

$$\Rightarrow t_2^{obs} = \frac{0.5}{0.063} = 7.936$$

Then the p-value is:

$$\alpha^{obs} = 2 \cdot P(t_{99} \leq -7.936)$$



-> Should we reject H_0?

Interpretation of β_2:

The mean of monthly expenditures for food increases by 50 euros as the monthly household income increases by 100 euros with X_2, D_1 and D_2 being constant

2.3

If we assume that the residuals have a Gaussian distribution,

$$\hat{\beta}_2 - \hat{\beta}_3 \sim N\left(\beta_2 - \beta_3, Var(\hat{\beta}_2) + Var(\hat{\beta}_3) - 2Cov(\hat{\beta}_2, \hat{\beta}_3)\right)$$

Since we know that

- $Var(\hat{\beta}_2) = \sigma^2 \cdot C_{2,2}$
- $Var(\hat{\beta}_3) = \sigma^2 \cdot C_{3,3}$
- $Cov(\hat{\beta}_2, \hat{\beta}_3) = \sigma^2 \cdot C_{2,3}$

we have that

$$Var(\hat{\beta}_2) + Var(\hat{\beta}_3) - 2Cov(\hat{\beta}_2, \hat{\beta}_3) = \sigma^2 \cdot c_{2,2} + \sigma^2 \cdot c_{3,3} - 2 \cdot \sigma^2 c_{2,3}$$

$$= \sigma^2 (0.02 + 0.07 + 0.04) = \sigma^2 \cdot 0.13$$

TEST

$$\begin{cases} H_0: \beta_2 = \beta_3 \\ H_1: \beta_2 \neq \beta_3 \end{cases} \Rightarrow \begin{cases} H_0: \beta_2 - \beta_3 = 0 \\ H_1: \beta_2 - \beta_3 \neq 0 \end{cases}$$

we can use a t-test

$$T_{23} = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - 0}{\sqrt{\sigma^2 \cdot 0.13}} \overset{H_0}{\sim} t_{n-p}$$

$$t_{23}^{obs} = \frac{(0.5 - 0.8)}{\sqrt{S^2 \cdot 0.13}} = \frac{(0.5 - 0.8)}{\sqrt{\frac{300}{95} \cdot 0.13}} = \frac{-0.3}{\sqrt{0.4095}} = \frac{-0.3}{0.63} = -0.47$$

The p-value is:

$$\alpha^{obs} = 2\mathbb{P}(t_{99} \leq -0.47) \times 0.5 \Rightarrow \text{we cannot reject } H_0$$

2.4

Let mod1 be the model in ex. 1.1 and mod2 the following model

INTERACTION TERM

$$Y_i = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 D_{1i} + \beta_5 D_{2i} + \boxed{\beta_6 D_{1i} X_{1i}} + \varepsilon_i$$

We need a test for "nested models":

$$\begin{cases} H_0: \beta_6 = 0 \\ H_1: \beta_6 \neq 0 \end{cases}$$

In general:

- p: number of coefficients in the main model
- p_0: number of coefficients in the nested model

Let p be the number of coefficients in mod2 and p_0 the number of coefficients in mod1.

$$F = \frac{(SSE_{mod_1} - SSE_{mod_2})/(p-p_0)}{SSE_{mod_2}/(n-p)} \overset{H_0}{\sim} F_{p-p_0,\, n-p}$$

$$f^{obs} = \frac{(300-282)/1}{282/94} = 6$$

$$\alpha^{obs} = \mathbb{P}(F_{1,93} \geq 6) < 0.01 \qquad 1 - qf(6, 1, 93)$$

=> we resect H_0

=> we prefer mod2 with the interaction

3 Exercise 3

To assess the verbal skills of 33 children, a test was conducted by collecting: the final score, the number of books read monthly by each child, and the number of books read monthly by their parents.

Exercise 3.1

Choose an appropriate response variable together with an appropriate linear regression model. Then, specify the related assumptions and the dimension of the design matrix X.

Exercise 3.2

Complete the following table and provide an interpretation of the estimates of the significant regression coefficients.

	Estimates	S.E.	t-value	p-value
X_1	1.5	0.44		
X_2		0.22		0.01

Exercise 3.3

Knowing the SST is equal to 2980 and $R^2 = 0.59$, decide if one of the two below options are compatible with the previous data (considering that the below options are based on a regression model with just one independent variable X_1):

- $SSR = 1800$ and $SSE = 1180$
- $SSR = 1500$ and $SSE = 1500$

ARE REFERRED TO A MODEL $Y|X_1$

Justify your answer.

3.1

Let's consider the following variables

- Y = final score
- X_1 = number of books read monthly by each child
- X_2 = number of books read monthly by their parents

We have the following linear regression model:

$$\underline{Y} = X \cdot \underline{\beta} + \underline{\epsilon},$$

where $\underline{Y}$ and $\underline{\epsilon}$ are 33×1 vectors, $\underline{\beta}$ is a 3×1 and X a 33×3 matrix.

Remember that X must be deterministic and have full rank: rank(X) = 3.

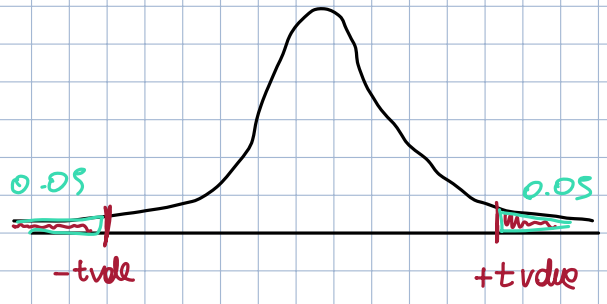
3.2

	ESTIMATES	S.E	t-value	p-value
X_1	1.5	0.44	3.409	0.0018
X_2		0.22	2.75	0.01

- $t_2^{obs} = \dfrac{\hat{\beta}_2}{s.\hat{\epsilon}.(\hat{\beta}_2)} = \dfrac{1.5}{0.44} = 3.409$

- $\alpha^{obs} = 2\,\mathbb{P}(t_{30} \leq -3.409) = 0.0018$

- $q_{0.995,30} = 2.75$



- $\hat{\beta}_3 = t\text{-value} \cdot s.\hat{\epsilon}.(\hat{\beta}_3) = 2.75 \cdot 0.22 = 0.605$

Since both p-values for β_2 and β_3 are smaller than 0.05, we reject with a 5% significance level the hypothesis that $\beta_2 = 0$ and that $\beta_3 = 0$.

3.3

Considering a model with just x_1 as independent variable we have:

$$SSR_{(y|x_1,x_2)} = SST_{(y|x_1,x_2)} \cdot R^2 = 2980 \cdot 0.59 = 1758.2$$

$$SSE_{(y|x_1,x_2)} = SST_{(y|x_1,x_2)} - SSR_{(y|x_1,x_2)} = 2980 - 1758.2 = 1221.8$$

- $SSR_{(y|x_1)} = 1800 \quad SSE_{(y|x_1)} = 1180$

  -> This is not compatible because we expect $SSR_{(y|x_1)} \leq SSR_{(y|x_1,x_2)}$

- $SSR_{(y|x_1)} = 1500$ $SSE_{(y|x_1)} = 1500$

  → This is not compatible because $SSR_{(y|x_1)} + SSE_{(y|x_1)} = 3000 \neq 2980$

4 Exercise 4

Considering 84 business company in northern Italy, we estimated the following regression model

$$\hat{y} = 12.7 + 9.3x_1 + 1.9x_2 - 1.6x_3$$

where Y = monthly turnover (in thousands), X_1 = sector (1 = manufacturing, 0 = trade), X_2 = number of employees, and X_3 = decrease in investment advertising compared to the previous year (in hundreds of euros). Further, $SSE = 2308$ and $R^2 = 0.62$.

Exercise 4.1

Interpret the estimate of β_2 ($\hat{\beta}_2 = 9.3$).

Exercise 4.2

Complete the table below and show the formula we should use.

	Estimates	S.E.	t-value	p-value
X_3				0.02

Exercise 4.3

Evaluate the goodness of fit through a valid test (thus, using the p-value).

4.1

If we move from the trade sector to the manufacturing sector the mean monthly turnover of a company increases by 9300€, subject to a fixed number of employees and a fixed decrease in investment advertising compared to the previous year

4.2

	Estimates	S.E.	t-value	p-value
X_3	-1.6	0.674	-2.3739	0.02

$0.02 = 2 \cdot P(t_{f_0} < -|t\text{-value}|) \Rightarrow q_{0.01,30} = -2.3739$

$\hat{S.E.}(\hat{\beta}_3) = \dfrac{\hat{\beta}_4}{-2.3739} = -1.6 / -2.3739 = 0.674$

=> we reject the hypothesis that $\beta_4 = 0$ with a 5% significance level

4.3

We use the following system of hypothesis

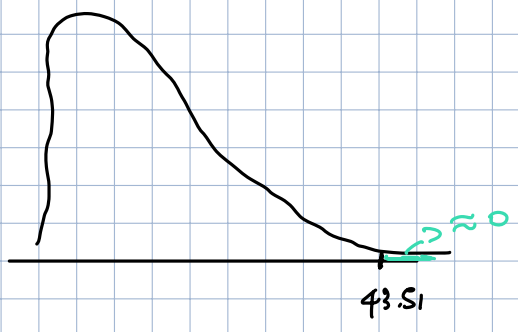
$$\begin{cases} H_0: R^2 = 0 \\ H_1: R^2 > 0 \end{cases} \qquad F = \frac{SSR/(p-1)}{SSE/(n-p)} \overset{H_0}{\sim} F_{p-1,n-p}$$

$$SST = \frac{SSE}{1-R^2} = \frac{2308}{1-0.62} = 6073.684$$

$$SSR = SST - SSE = 6073.684 - 2308 = 3765.684$$

$$f^{obs} = (3765.684/3)/(2308/80) = 43.51$$

$$\alpha^{obs} = P(F_{3,80} \geq 43.51) \approx 0$$



=> we reject H_0

=> the model is good enough