Exercises: Multiple Linear Regression Part I

Valentina Zangirolami - valentina.zangirolami@unimib.it

December 5-14, 2023

(Thanks to Dr. Roberto Ascari for providing these exercises)
(Referring to the theoretical parts: 9, 10, 12, 13, 14, 15, 16, 17, 18, 19)

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.

The regression model for the variable T corresponds to

$$T_i = \beta_1 + \beta_2 TV_i + \beta_3 D_i + \varepsilon_i \quad , \; i = 1, \dots, 100$$

where D_i is a dummy variable s.t.

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

The assumptions related to the multiple linear regression model with p covariates are:

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

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

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

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

$\Rightarrow$ Absence of multicollinearity of the covariates or, equivalently, $\text{rank}(X) = p$ (full rank)

Exercise 1.2

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

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

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

and we can re-write our null hypothesis also as

$H_0: \beta_2 - \beta_3 = 0$

Then, the test statistic corresponds to

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

(p is the number of our coefficients)

To compute the test statistic we need to find the values of SSR and SSE.

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

and then $SSE = SST - SSR = SST(1-R^2) = 482.65$

Hence,

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

and the p-value is

$$\alpha^{obs} = \mathbb{P}(F_{2,97} > 50.479) = 8.881784 \cdot 10^{-16}$$

(By using R, 1- pf(50.479, 2, 97))

$\alpha^{oss} < 0.01$ => we reject H_0, which means we are not rejecting the model

FOR THE EXAM

During the exam, you can find the p-value by choosing among ~~quantiles~~ quantiles we will provide.
Given the value of the quantiles, you can understand the ~~which level of alpha~~ value of $f_{x^{obs}}$. (Some examples in the next ex)

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).

(NOTES PART 13)

Let consider

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

and we know

(a) $\text{Var}(\underline{\hat{\beta}}) = \sigma^2 (X^TX)^{-1}$

(b) $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}$

Hence,

$$c_{2,2} = \frac{\widehat{\text{Var}}(\hat{\beta}_2)}{S^2} \overset{(*)}{=} \frac{(0.9)^2}{482.65/97} = \frac{0.81}{4.9758} = 0.1628$$

(*) is obtained by computing $S^2 = SSE/(n-p)$

and equivalently

$$c_{3,3} = \frac{\widehat{\text{Var}}(\hat{\beta}_3)}{S^2} = \frac{(2.3)^2}{4.9758} = 1.0631$$

Using (b), we can also find c_{ij}. In a sense, we need the value of $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$. However, we don't know this value and the text provides ρ. Thus, we can exploit the following formula

$$\rho = \text{Corr}(\hat{\beta}_2, \hat{\beta}_3) = \frac{c_{ij}}{(c_{ii}c_{jj})^{1/2}}$$

$$c_{ij} = \rho (c_{ii} c_{jj})^{1/2} = \frac{\rho \sqrt{\widehat{Var}(\hat{\beta}_2)} \sqrt{\widehat{Var}(\hat{\beta}_3)}}{S^2} = 0.2829$$

Finally, our matrix is

$$(X^T X)^{-1} = \begin{bmatrix} - & - & - \\ - & 0.1028 & 0.2829 \\ - & 0.2829 & 1.0631 \end{bmatrix}$$

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.

Let consider the same assumptions of ex. 1.1. The multiple regression model corresponds to

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

where

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

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

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 $\hat{\beta}_2$.

The system of hypothesis is

$$\begin{cases} H0: \beta_2 = 0 \\ H1: \beta_2 \neq 0 \end{cases}$$

and the test statistic

$$T_2 = \frac{\hat{\beta}_2}{\sqrt{\hat{Var}(\hat{\beta}_2)}} \overset{H0}{\sim} t_{m-p} \qquad (p = \#\text{coefficients})$$

where

$$\sqrt{\hat{Var}(\hat{\beta}_2)} = \sqrt{S^2 c_{22}}$$

- $S^2 = \frac{SSE}{m-p} = \frac{300}{120-5} = 2.6087$
- $\sqrt{\hat{Var}(\hat{\beta}_2)} = \sqrt{2.6087 \cdot 0.02} = 0.2284$
- $t_2^{obs} = \frac{0.5}{0.2284} = 2.189$

Hence, the p-value is

$$\alpha^{obs} = 2 \min \{ \mathbb{P}(t_{115} \leq -2.189), \mathbb{P}(t_{115} > 2.189) \} =$$

$$= 2 \cdot 0.0153 = 0.0307$$

We reject H0 at 5% significance level while we cannot reject H0 at 1% significance level.

How to compute p-value during the exam

If we provide the value of quantile $t_{115,0.0153} = -2.189$, then you know that the p-value is

$$\alpha^{obs} = 2 \cdot 0.0153 = 0.0304$$

Interpretation of β_2: The mean of monthly expenditures for foods increases by 50 euros as the monthly household income increases by 100 euros with X_2, D_1 and D_2 being constant.

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).

If we assume the residuals follow a Gaussian distribution,

$$\hat{\beta}_2 - \hat{\beta}_3 \sim N(\beta_2 - \beta_3,\ Var(\hat{\beta}_2) + Var(\hat{\beta}_3) - 2\,Cov(\hat{\beta}_2, \hat{\beta}_3)) =$$
$$\overset{(*)}{=} N(\beta_2 - \beta_3,\ \sigma^2\, 0.13)$$

(*) We want to obtain the variance, then we know

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

Using the previous values,

$$\sigma^2 \cdot c_{2,2} + \sigma^2 \cdot c_{3,3} - 2\sigma^2 c_{2,3} = \sigma^2 (0.02 + 0.07 + 0.04) = \sigma^2\, 0.13$$

TEST

System of hypothesis

$$\begin{cases} H0: \beta_2 = \beta_3 \\ H1: \beta_2 \neq \beta_3 \end{cases} \qquad \begin{cases} H0: \beta_2 - \beta_3 = 0 \\ H1: \beta_2 - \beta_3 \neq 0 \end{cases}$$

Knowing the distribution of $\hat{\beta}_2 - \hat{\beta}_3$, we can use a t-test

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

$$t_{23}^{obs} = \frac{(0.5 - 0.8)}{\sqrt{s^2 \cdot 0.13}} = \frac{(0.5 - 0.8)}{\sqrt{\frac{300}{115}\, 0.13}} = \frac{-0.3}{\sqrt{2.6087 \cdot 0.13}} = -0.5152$$

and the p-value corresponds to

$$\alpha^{obs} = 2 \cdot \mathbb{P}(t_{115} \leq -0.5152) = 0.6074$$ We cannot reject H0

EXAM

Given $t_{115, 0.3037} = -0.5152$, the p-value is $\alpha^{obs} = 2 \cdot 0.3037 = 0.6074$

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, compute and interpret the partial coefficient of determination and decide the best model through an appropriate test (specify hypothesis, test statistic and p-value).

Let mod 1 be the model defined in the ex. 2.1 and the mod 2 is defined as follows

(mod 2): $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 D_{i1} + \beta_5 D_{i2} + \beta_6 D_{i1} x_{i1} + \varepsilon_i$

THIS PART IS NOT INCLUDED IN THE FINAL EXAM (i.e. you don't need to know PCD)

The partial coefficient of determination (PCD) is

$$PCD = \frac{SSE_{mod1} - SSE_{mod2}}{SSE_{mod1}} = \frac{R^2_{mod2} - R^2_{mod1}}{1 - R^2_{mod1}} = \frac{300 - 282}{300} = 0.06$$

Adding the interaction term into the model leads to explain 6% more of the deviance than mod 1.

We can consider a test which involves PCD with the following system of hypothesis

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

It corresponds to the test about a subset of β (see notes part 16) which encompasses "nested models".

For instance, we can also consider the following system of hypothesis

$$\begin{cases} H_0: \beta_6 = 0 \\ H_1: \beta_6 \neq 0 \end{cases}$$ (In this case, you can also use t-test)

Let p be the number of regression coefficients of mod 2 and p_0 be the number of regression coefficients of mod 1

The test statistic corresponds to

$$F = \frac{(SSE_{mod1} - SSE_{mod2})/(p - p_0)}{SSE_{mod2}/n - p} \overset{H_0}{\sim} F_{p-p_0, n-p}$$

$$f^{obs} = \frac{(300 - 282)/1}{282/114} = 7.278$$

The p-value is

$\alpha^{obs} = \mathbb{P}(F_{1,114} > 7.278) = 0.008 < 0.05$ We reject H_0 at 5% significance level

→ We prefer to include interaction and thus choose the model 2

[p-value for the exam]

If we provide the quantile $f_{1,114;\,0.992} = 7.278$

[You can find the p-value as $\alpha^{obs} = 1 - 0.992 = 0.008$]

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.

Let consider

- Y = Final Score
- X_1 = Number of books read monthly by each child
- X_2 = Number of books read monthly by their parents

Therefore, the multiple regression model represents an appropriate linear regression model which can be expressed as

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$$

where $\underline{Y}$ and $\underline{\varepsilon}$ are vectors with dimension 33×1,
$\underline{\beta}$ is a vector of dimension 3×1 and X is a matrix 33×3.

$\Rightarrow$ The design matrix X must be deterministic and have full rank (e.g. rank$(X) = 3$ in this case).

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	3.409	0.0018
X_2	0.605	0.22	2.75	0.01

(1) t-value for β_2

$$t_2^{obs} = \frac{\hat{\beta}_2}{\sqrt{\hat{Var}(\hat{\beta}_2)}} = \frac{1.5}{0.44} = 3.409$$

(2) p-value for β_2

$$\alpha^{obs} = 2 \cdot \mathbb{P}(T_{30} \leq -3.409) = 0.0018 \qquad (t_{30;0.0009} \approx -3.4)$$

(3) t-value for β_3

$0.01 = 2\mathbb{P}(T_{30} \geq |t\text{-value}|) \Rightarrow t\text{-value} = q_{0.995,30} = 2.75$

Even t-value = -2.75 provides the same p-value.

However, a negative t-value implies a negative value for $\hat{\beta}_3$ estimate. And it makes less sense (i.e. the mean of the total score decreases as the number of books read by the parents increases)

(4) Estimate of β_3

$$\hat{\beta}_3 = t\text{-value} \cdot \sqrt{\hat{Var}(\hat{\beta}_3)} = 2.75 \cdot 0.22 = 0.605$$

- Both the regression coefficients are significantly different from zero (at 5% significance level).

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 $SSE = 1500$

Justify your answer.

Considering the value of deviances related to a model which involves Y, X_1, X_2

$SSR = SST \cdot R^2 = 2980 \cdot 0.59 = 1758.2$

$SSE = SST - SSR = 2980 - 1758.2 = 1221.8$

(1) The case "$SSR_{(YX_1)} = 1800$ and $SSE_{(YX_1)} = 1180$" cannot may be not compatible. Our expectation should be

$$SSR_{(YX_1)} \leq SSR_{(YX_1X_2)}$$

because the model with two covariates should explain more variability than the model with a covariate.

(2) The case "$SSR_{(YX_1)} = 1500$ and $SSE_{(YX_1)} = 1500$" cannot be compatible because in this case $SST = 3000$ instead of 2980. The total deviance (SST) must be the same.

Exercise 3.4

Knowing the SSR is equal to 1440 (for a regression model which just includes X_1), try to evaluate if it is better to include the second independent variable through an appropriate statistical test (specify hypothesis, test statistic and p-value).

To compute these two models, we can use the partial coefficient of determination (PCD). → (NO IN THE exam, see ex. 2.4)

Let consider

(i) mod 1: $y_i = \beta_1 + \beta_2 x_{i1} + \varepsilon_i$

(ii) mod 2: $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \varepsilon_i$

NOT INCLUDED IN THE EXAM

The PCD is defined as

$$PCD = \frac{SSE_{mod1} - SSE_{mod2}}{SSE_{mod1}} = \frac{(SST_{mod} - SSR_{mod1}) - SSE_{mod1}}{SSE_{mod1}}$$

$$= \frac{(2980 - 1440) - 1221.8}{(2980 - 1440)} = 0.207$$

mod 2 explains 20% of the deviance more than mod 1

Therefore, the system of hypothesis can be

$$\begin{cases} H_0: PCD = 0 \\ H_1: PCD \neq 0 \end{cases} \quad \text{or} \quad \begin{cases} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{cases}$$

and the test statistic is

$$f^{obs} = \frac{(SSE_{mod1} - SSE_{mod2})/(p - p_0)}{SSE_{mod2}/(n - p)} = \frac{1540 - 1221.8}{1221.8/30} = 7.813$$

$(F \overset{H_0}{\sim} F_{1,30})$

The p-value corresponds to

$$\alpha^{obs} = \mathbb{P}(F_{1,30} > 7.813) = 0.009 < 0.01$$

We reject H_0: we prefer mod 2

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 ($\hat{\beta}_2 = 9.3$).

Given that X_1 is a dummy variable, we can interpret β_1 as follows:

(a) • Moving from the trade sector to a manufacturing sector leads to an increase in monthly turnover of 9300 euros on average, subject to a fixed values for the other covariates (number of employees, decrease in investment advertising compared to the previous year).

(b) • The average difference between the turnover of two companies (in the two different sector) with same number of employees and the same amount of decrease advertising investment decrease compared to the previous year is 9300 €.

Exercise 4.2

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

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

① t-value

$0.02 = 2 \cdot \mathbb{P}(T_{80} \leq -|\text{t-value}|) \Rightarrow \text{t-value} = q_{0.01,30} = -2.3739$

② $\hat{\beta}_4 = -1.6$ (From the text of previous ex)

③ $\sqrt{\widehat{\text{Var}}(\hat{\beta}_4)} = \dfrac{\hat{\beta}_4}{t_4^{\text{obs}}} = \dfrac{-1.6}{-2.3739} = 0.674$

$\Rightarrow \beta_4$ is significantly different from zero at 5% significance level (but no at 1%)

Exercise 4.3

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

We can use the following system of hypothesis

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

and the test statistic is

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

Then, we need to find SSR (given SSE and R^2 values)

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

$$SSR = SST - SSE = 3765.684$$

and

$$f^{obs} = \frac{3765684/3}{2308/80} = 43.51$$

The p-value is

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

We reject H_0: ~~cancel~~ the model is good enough

Exercise 4.4

We would like to assess the added explanatory contribution of the variable "number of employees" compared with the model that does not contain it. Knowing that the regression sum of squares of the model without this variable is equal to 1978, compute an appropriate index/coefficient and interpret the result.

To compare two models, we can use the partial coefficient of determination.

Briefly, we are considering

mod 1 $\quad y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{3i} + \varepsilon_i \qquad i = 1, \dots, 84$

mod 2 $\quad y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + \varepsilon_i \qquad i = 1, \dots, 84$

and we know $SSR_{mod1} = 1978$. Therefore, the residual deviance is

$SSE_{mod1} = SST - SSR_{mod1} = 6073.684 - 1978 = 4095.684$

$$PCD = \frac{SSE_{mod1} - SSE_{mod2}}{SSE_{mod1}} = \frac{4095.684 - 2308}{4095.684} = 0.4365$$

The 43.65% of the deviance not explained by X_1 and X_3 is explained by X_2.