Exercises: Multiple Linear Regression Part II

Valentina Zangirolami - valentina.zangirolami@unimib.it

December 14, 2023

(Referring to the theoretical parts: 9, 10, 12, 13, 14, 15, 16, 17, 18, 19)

1 Exercise 1

Among $n = 31$ cherry trees, data about *Volume*, *Diameter* and *Height* of each tree were collected.

Exercise 1.1
Specify an appropriate regression model for the response variable *Volume* and the related assumptions. Propose a transformation for all variables by thinking about the geometric relationship of our variables.

Exercise 1.2

Knowing that

$$(X^T X)^{-1} = \begin{bmatrix} 96.572 & 3.139 & -24.165 \\ 3.139 & 0.849 & -1.227 \\ -24.165 & -1.227 & 6.310 \end{bmatrix} \qquad X^T \boldsymbol{y} = \begin{bmatrix} 101.455 \\ 263.056 \\ 439.896 \end{bmatrix}$$

Find the estimate for $\boldsymbol{\beta}$ and write the estimated model.

Exercise 1.3

Knowing the maximum likelihood estimate $\hat{\sigma}^2 = 0.00598$, find the unbiased estimate $\hat{s}^2$. Then, calculate the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$.

Exercise 1.4

Perform a statistical test to evaluate if the regression coefficient related to the *Diameter* is significant. Specify the system of hypothesis, the test statistic and the p-value. Then, perform another statistical test with the following null hypothesis $H0 : \beta_2 = \beta_3 = 0$ (Resticted model SSE= 8.309).

2 Exercise 2

Among n=32 births, the following three variables were collected:

- *Weight*: birth weight in grams of baby
- *Smoking*: Smoking status of mother (yes or no)
- *Gest*: length of gestation in weeks

Describe the equation of the multiple regression model specifying the nature of each variable (without including the interaction term). Then, express the model in matrix form: $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, explicitly stating $\underline{Y}$, X, $\underline{\beta}$, and $\underline{\varepsilon}$ (and its distribution).

Exercise 2.1

Knowing the following values

$$\widehat{Weight}^{(n)} = -2390 + 143\ Gest \qquad (Smoking = 0)$$
$$\widehat{Weight}^{(y)} = -2635 + 143\ Gest \qquad (Smoking = 1)$$

What is the estimated value of the regression coefficient associated to the dummy variable? Interpret the results and explain, theoretically, why the two equations are different.

Exercise 2.2

After adding an interaction term, we obtained the following estimated regression model

$$\widehat{Weight} = -2546.138 + 147.207\ Gest + 71.574\ Smoke - 8.178\ Gest \times Smoke$$

Provide the estimated regression model for each group (Smoke: yes or no) and interpret the results.

Exercise 2.3

Consider the following table where for the two models $\mathcal{M}_0$ (the restricted model, i.e. the model which just involves *Gest*) and $\mathcal{M}_4$ (corresponds to the model specified in the ex. 2.2) are expressed the residual sum of squares (SSE). Complete the table.

Model	**D.o.f.**	**SSE**
Resticted ($\mathcal{M}_0$)		839951.03
Unconstrained ($\mathcal{M}_4$)		384391.46

Further, perform a statistical test with the following system of hypothesis (where Smoking has two classes: Y and N)

$$\begin{cases} \text{H0: } \mu_N = \mu_Y \\ \text{H1: } \mu_N \neq \mu_Y \end{cases}$$

Compute the test statistic and the p-value.

Exercise 2.4

Now, we want just to test the regression coefficient related to the interaction. The SSE of the model without the interaction is equal to 387069.83. Perform a valid test and discuss about the results.

Exercise 2.5

Let $SSE = 3735789.2$ be the Residual Sum of Squares of the model $Y_i = \beta_1 + \epsilon_i$. Compute R^2 and adjusted-R^2 related to $\mathcal{M}_4$. Explain the difference.

3 Exercise 3

Among $n = 15$ poisoned lab rats, data about survival time (Y) and antidote/treatment (T) were collected. The researchers consider three kind of treatment: A, B and C, where

Group	n_j	$\bar{y}_j$	s_j
A ($j = 1$)	4	0.347	0.2055
B ($j = 2$)	3	3.067	0.4618
C ($j = 3$)	8	1.764	0.5007

Compute the total deviance (total sum of squares - SST) and explain its components.

Exercise 3.1

Perform a statistical test to evaluate if the means (of each group) are homogeneous. Specify the system of hypothesis, the test statistic and the p-value.

Exercise 3.2

Specify an appropriate linear regression model. Given the previous data, find the estimates for each coefficient and interpret them.

Exercise 3.3

Propose an alternative test for ex. 3.1 specifying the system of hypothesis. Then, compute R^2 of our model.

4 Exercise 4

Lab *Precise* got some measurements to see whether the tar contents (in milligrams) for three different brands of cigarettes are different. The measurements are showed in the following table.

Sample	**Brand A**	**Brand B**	**Brand C**
1	10.21	11.32	11.60
2	10.25	11.20	11.90
3	10.24	11.40	11.80
4	9.80	10.50	12.30
5	9.77	10.68	12.20
6	9.73	10.90	12.20

Write the equation of the linear regression model. Find and interpret the estimates of regression coefficients.

Exercise 4.1

Complete the following table.

Sources of variation	**D.o.f.**	**Deviance**
Total		
Regression		
Residual		

Perform the statistical test to evaluate if the means (of each group) are homogeneous.