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.

An appropriate regression model is the multiple regression model which can be defined as follows

$$Volume_i = \beta_1 + \beta_2 Diameter_i + \beta_3 Height_i + \varepsilon_i$$

with the following assumptions:

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

(ii) Linear independence among independent variables: the vectors $x_j \in \mathbb{R}^n$, $j = 1, \dots, p$, are linear independent.

(iii) linearity

(i) and (iii) imply $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$, where $\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$

$i = 1, \dots, n$

Answer

The geometry suggests the following model

$$\log(Volume) = \log(\pi/4) + 2\log(Diameter) + \log(Height) + \varepsilon$$

Therefore, we can consider the following model

(a) $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$

where

- $Y_i = \log(\text{Volume}_i)$
- $x_{i2} = \log(\text{Diameter}_i)$
- $x_{i3} = \log(\text{Height}_i)$

Exercise 1.2

Knowing that

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

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

(Previous data refers to the logarithmic case)

The model (a) can be re-written using matrix notation as

$$\underset{31\times 1}{\underline{Y}} = \underset{31\times 3}{X}\ \underset{3\times 1}{\underline{\beta}} + \underset{31\times 1}{\underline{\varepsilon}} \qquad \text{where } \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I)$$

where

$$X = \begin{bmatrix} 1 & x_{12} & x_{13} \\ 1 & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ 1 & x_{31,2} & x_{31,3} \end{bmatrix} \qquad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{31} \end{bmatrix}$$

The estimate of $\underline{\beta}$ corresponds to

$$\underline{\tilde{\beta}} = (X^TX)^{-1}X^T\underline{y}$$

and then

$$\underline{\tilde{\beta}} = \begin{bmatrix} 96.572 & 3.139 & -24.165 \\ 3.139 & 0.849 & -1.227 \\ -24.165 & -1.227 & 6.310 \end{bmatrix} \begin{bmatrix} 101.455 \\ 263.056 \\ 439.896 \end{bmatrix} = \begin{bmatrix} -6.6418 \\ 2.0494 \\ 1.314 \end{bmatrix}$$

The estimated model:

$$\underline{y} = X \begin{bmatrix} -6.6418 \\ 2.0494 \\ 1.314 \end{bmatrix} + \underline{\varepsilon}$$

Using the notation of the equation (a), we have

$\hat{\beta}_1 = -6.6418$ $\hat{\beta}_2 = 2.0494$ $\hat{\beta}_3 = 1.314$

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 $\hat{\boldsymbol{\beta}}$.

To find the unbiased estimate for σ^2, we need to compute

$$S^2 = \frac{1}{n-p}\, \underline{e}^T \underline{e} = \frac{n}{n-p}\, \frac{1}{n}\, \underline{e}^T \underline{e} = \frac{n}{n-p}\, \hat{\sigma}^2$$

where p is the number of coefficients (the intercept is included) and

$$\underset{31\times 1}{\underline{e}} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_{31} - \hat{y}_{31} \end{bmatrix} = \underline{y} - \underline{\hat{y}} = (I - P)\underline{y}$$

$P = X(X^TX)^{-1}X^T$ is the projection matrix and $\underset{31\times 31}{I}$ is the identity matrix

$$\underline{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 1 \end{bmatrix}$$

31×31

Knowing $\hat{\sigma}^2 = 0.00598$, we can find S^2 easily

$$S^2 = \frac{31}{28}\, 0.00598 = 0.0066207$$

We also know the distribution of the estimator $\underline{\hat{\beta}}$:

$$\underline{\hat{\beta}} \sim N_p\left(\underline{\beta}, \sigma^2 (X^TX)^{-1}\right)$$

Where each estimator $\hat{\beta}_i$: $\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 (X^TX)^{-1}_{ii}\right)$ and the variance/covariance matrix is $\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^TX)^{-1}$

Its estimate corresponds to

$$\widehat{Var}(\hat{\beta}) = s^2 (X^T X)^{-1} = 0.0066207 \begin{bmatrix} 96.572 & 3.139 & -24.165 \\ 3.139 & 0.849 & -1.227 \\ -24.165 & -1.227 & 6.310 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.639 & 0.0208 & -0.16 \\ 0.0208 & 0.0056 & -0.0081 \\ -0.16 & -0.0081 & 0.042 \end{bmatrix}$$

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 $H0 : \beta_2 = \beta_3 = 0$.

(Restricted Model SSE = 8.309)

Inference on $\hat{\beta}_2$

System of hypothesis

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

Test statistic:

$$t_2^{obs} = \frac{\hat{\beta}_2 - 0}{\sqrt{\widehat{var}(\hat{\beta}_2)}} = \frac{2.0494}{\sqrt{0.0056}} = 27.38626$$

$$\alpha^{obs} = 2 \min\{\mathbb{P}(t_{28} \leq 27.38626), \mathbb{P}(t_{28} \geq 27.38626)\} \approx 0$$

We reject H_0.

$$\begin{cases} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \exists j \in \{2,3\} \text{ s.t } \beta_j \neq 0 \end{cases}$$

$$f^{obs} = \frac{\frac{8.309 - 0.1855}{3 - 1}}{\frac{0.1855}{31 - 3}} = 613.094$$

$\alpha^{obs} = \mathbb{P}(f_{2,28} \geq 613.094) \approx 0$ We reject H_0.

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).

Data contain three variables:

- Weight is a quantitative variable
- Smoking is a qualitative variable
- Gest is a quantitative variable

The multiple regression model can be represented by

$$Weight_i = \beta_1 + \beta_2 Gest_i + \beta_3 D_i + \varepsilon_i \qquad \varepsilon_i \overset{ii}{\sim} N(0, \sigma^2) \quad i = 1, \dots, 32$$

where

$$D_i = \begin{cases} 1, & \text{if Smoking} = Y \\ 0, & \text{otherwise} \end{cases}$$

and using matrix notation:

$$\underset{32\times 1}{\underline{Weight}} = \underset{32\times 3}{X}\ \underset{3\times 1}{\underline{\beta}} + \underset{32\times 1}{\underline{\varepsilon}} \qquad \underline{\varepsilon} \sim N_3(\underline{0}, \sigma^2 I)$$

$$\Rightarrow \underline{Weight} \sim N_3(X\underline{\beta}, \sigma^2 I) \quad \text{where } I \text{ is an identity matrix}$$

$$X = \begin{bmatrix} 1 & Gest_1 & Smoking_1 \\ 1 & Gest_2 & Smoking_2 \\ 1 & Gest_3 & Smoking_3 \\ \vdots & \vdots & \vdots \\ 1 & Gest_{32} & Smoking_{32} \end{bmatrix} \qquad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{32} \end{bmatrix}$$

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.

$\hat{\beta}_3 = -2635 + 2390 = -245$

Referring to the previous notation (ex.2), the two equations can be represented by

$$\widehat{Weight^{(n)}} = \hat{\beta}_1 + \hat{\beta}_2 Gest \qquad (Di = 0)$$
$$\widehat{Weight^{(y)}} = (\hat{\beta}_1 + \hat{\beta}_3) + \hat{\beta}_2 Gest \qquad (Di = 1)$$

In the second equation, the intercept increases by $\hat{\beta}_3$ due to $\hat{Di} = 1$ (Smoking = Yes).

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.

(Smoking = No) $\widehat{Weight}^{(N)} = -2546.138 + 147.207\ Gest$

(Smoking = Yes) $\widehat{Weight}^{(Y)} = -2474.564 + 139.029\ Gest$

Considering the estimated regression model for Smoking = NO, on average the birth weight of baby increases by 147.207 grams as Gest increases by one week.

Instead, considering smokers: On average the birth weight of baby increases by 139.029 grams as Gest increases by one-week.

The interpretation of intercept does not provide relevant information, given that no makes sense facing with the case of Gest = 0

Exercise 2.3

Consider the following table where for the two models $\mathcal{M}_0$ (the restricted model) 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 ($\mathcal{M}_0$)	30	839951.03
Unconstrained ($\mathcal{M}_4$)	28	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.

Degree of freedom (first table)

M_0 : $n - p_0 = 32 - 2 = 30$

M_4 : $n - p = 32 - 4 = 28$

where p_0 are the number of regression coefficients for M_0 and p are the number of regression coefficients for M_4.

Given the following system of hypothesis

$$\begin{cases} H_0 : \mu_N = \mu_Y \\ H_1 : \mu_N \neq \mu_Y \end{cases}$$

TEST STATISTIC

$$F^{obs} = \frac{839951.03 - 384391.46 / (4-2)}{384391.46 / 32-4} = 16.59$$

P-VALUE

$$\alpha^{obs} = \mathbb{P}(F_{2,28} > 16.59) = 1 - \mathbb{P}(F_{2,28} < 16.59) \approx 0$$

We reject the null hypothesis. The coefficients β_3 and β_4 are significant.

The pairwise system of hypothesis can be rewritten as

$$\begin{cases} H_0: \beta_3 = \beta_4 = 0 \\ H_1: \exists j \in \{3,4\}: \beta_j \neq 0 \end{cases}$$

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.

System of hypothesis

$$\begin{cases} H_0: \beta_4 = 0 \\ H_1: \beta_4 \neq 0 \end{cases}$$

TEST STATISTIC:

$$f^{obs} = \frac{387069.83 - 384391.6/1}{384391.6/28} = 0.20$$

P-VALUE

$$\alpha^{obs} = \mathbb{P}(F_{1,28} \geq 0.20) = 0.6621 \qquad (f_{1,28;0.3379} \approx 0.20)$$

We cannot reject H_0.

Exercise 2.5 Knowing the SSE = 3735789.2 of the model $\underline{Y} = \beta_1 \underline{1} + \underline{\varepsilon}$.
Compute R^2 and adjusted-R^2 related to $\mathcal{M}_4$. Explain the difference.

Let consider

M_0: $\underline{Y} = \beta_1 \underline{1} + \underline{\varepsilon}$ where $\tilde{\sigma}^2 = \frac{e_0^T e_0}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$

and

M_4: $y_i = \beta_1 + \beta_2 Gest_i + \beta_3 Smoke_i + \beta_4 Gest_i \times Smoke_i + \varepsilon_i$

with $\hat{\sigma}^2 = \frac{e_4^T e_4}{n}$ where e_4 refers to residuals of M_4

Then, we know

$$\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n} e_i^2} - 1 = \frac{1}{1-R^2} - 1 = \frac{R^2}{1-R^2}$$

Exploiting this equivalence, we can find our R^2 (*)

$$\frac{3735789.2}{384391.46} - 1 = 8.7187$$

$$\frac{R^2}{1-R^2} = 8.7187 \Rightarrow R^2 + 8.7187 R^2 = 8.7187 \Rightarrow R^2 = \frac{8.7187}{9.7187} = 0.8971$$

While the adjusted-R^2 (R^2_{adj})

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p} = 1 - (1 - 0.8971)\frac{31}{28} = 0.886075$$

In the case of multiple regression model, the R^2_{adj} is a penalized R^2 with respect to the number of covariates $(p-1)$ into the model.

R^2_{adj} penalizes model with many variables.

The values However, R^2 and R^2_{adj} reach an highest value, near to 1, which means that our model is good.

(*) You can also consider directly

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{384391.46}{3735789.2} = 0.8971$$

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.

We are dealing with ~~ANOVA~~ one-way ANOVA.

In a sense, we know

$$\underbrace{\sum_{j=1}^{3}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2}_{SST} = \underbrace{\sum_{j=1}^{3}(n_j-1)s_j^2}_{SSE} + \underbrace{\sum_{j=1}^{3} n_j(\bar{y}_j-\bar{y})^2}_{SSR}$$

where $j=\{1,2,3\}$ refers to each group and n_j is the number of observations for each group (see the table).

In this case, SSE corresponds to the deviance between groups while SSR is the deviance within groups.

We can obtain the average of survival time by

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{3} n_j\bar{y}_j = \frac{(4\cdot 0.347 + 3\cdot 3.067 + 8\cdot 1.764)}{15} = 1.646733$$

Then, let's compute SSE and SSR

$$SSE = 3\cdot(0.2055)^2 + 2(0.4618)^2 + 7(0.5007)^2 = 2.308113$$

$$SSR = 4(0.347-1.646733)^2 + 3(3.067-1.646733)^2 + 8(1.764-1.646733)^2 = 12.91871$$

We are ready to find SST

$$SST = SSE + SSR = 2.308113 + 12.91871 = 15.22682$$

Exercise 3.1

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

System of hypothesis

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \exists j \text{ s.t. } \mu_j = \mu_i \ (i \neq j, i,j \in \{1,2,3\}) \end{cases}$$

test statistic

$$f^{obs} = \frac{SSR/(J-1)}{SSE/(n-J)} \overset{H_0}{\sim} F_{J-1,\, n-J}$$

$$f^{obs} = \frac{12.92/2}{2.308/12} = 33.59$$

p-value

$$\alpha^{obs} = \mathbb{P}(F_{2,12} > 33.59) \approx 0$$ We reject null hypothesis

Exercise 3.2

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

An appropriate model can be

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad \text{assuming } \varepsilon_i \sim N(0, \sigma^2),\ i = 1, \dots, 15$$

and which implies $Y_i \sim N(\mu_i, \sigma^2)$, where

$$\mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$$

Therefore, it also implies:

(a)
$$\begin{cases} Y_i \sim N(\beta_1, \sigma^2) & i = 1, \dots, 4 \quad \text{(Group A)} \\ Y_i \sim N(\beta_1 + \beta_2, \sigma^2) & i = 5, \dots, 7 \quad \text{(Group B)} \\ Y_i \sim N(\beta_1 + \beta_3, \sigma^2) & i = 8, \dots, 15 \quad \text{(Group C)} \end{cases}$$

considering the covariates as follows

$$x_{i2} = \begin{cases} 1 & \text{if the antidote is B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if the antidote is C} \\ 0 & \text{otherwise} \end{cases}$$

Basically, (a) guarantees

$$\mu_A = \beta_1 \ ; \ \mu_B = \beta_1 + \beta_2 \ ; \ \mu_C = \beta_1 + \beta_3$$

and if we want to find the maximum likelihood estimates, the property of invariance we know that they satisfy the following conditions:

$$\hat{\beta}_1 = \hat{\mu}_A = \bar{Y}_A$$

$$\hat{\beta}_1 + \hat{\beta}_2 = \hat{\mu}_B = \bar{Y}_B \qquad \text{(for equivariance property)}$$

$$\hat{\beta}_1 + \hat{\beta}_3 = \hat{\mu}_C = \bar{Y}_C$$

Hence, we can find the estimates as follows

$$\hat{\beta}_1 = \bar{y}_A$$
$$\hat{\beta}_2 = \bar{y}_B - \bar{y}_A$$
$$\hat{\beta}_3 = \bar{y}_C - \bar{y}_A$$

Now, we are ready to compute our values

$$\hat{\beta}_1 = 0.347$$
$$\hat{\beta}_2 = 3.067 - 0.347 = 2.72$$
$$\hat{\beta}_3 = 1.764 - 0.347 = 1.417$$

The estimated regression model can be written as

$$\hat{y}_i = 0.347 + 2.72\, x_{i2} + 1.417\, x_{i3}$$

Interpretation of regression coefficients:

- $\hat{\beta}_1 = 0.347$ is the average of the survival time (Y) when the group is A
- $\hat{\beta}_2 = 2.72$ is the average increase in Y when the group is B
- $\hat{\beta}_3 = 1.417$ is the average increase in Y when the group is C

Exercise 3.3

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

To test if the means of each group are homogeneous, we can use equivalently the following system of hypothesis

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

Therefore, the restricted model (or null model) is

$$M_0: y_i = \beta_1 + \varepsilon_i$$

and the unconstrained model (or full model) is

$$M_2: y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

And we can point out: $\tilde{\sigma}^2$ as the estimated variance of the null model while $\hat{\sigma}^2$ of the estimated variance of the full model, where

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{SST}{n} = \frac{15.23}{15} = 1.015$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 = \frac{SSE}{n} = \frac{2.308}{15} = 0.1539$$

Hence, the test statistic corresponds to

$$f^{obs} = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/(J-1)}{\hat{\sigma}^2/(n-J)} = \frac{SSR/(J-1)}{SSE/(n-J)} = 33.59 \text{ as ex. 3.1}$$

We can find R^2 by exploiting

$$R^2 = \frac{SSR}{SST} = 1 - \frac{2.308}{15.23} = 0.848457$$

The model explains 84.8% of the total variability of the response variable.

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.

Equation of linear regression model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \qquad i = 1, \dots, 18$$

where

$$x_{i2} = \begin{cases} 1, & \text{if Brand} = B \\ 0, & \text{otherwise} \end{cases} \qquad x_{i3} = \begin{cases} 1, & \text{if Brand} = C \\ 0, & \text{otherwise} \end{cases}$$

As in the previous exercise, we can find the ML estimates as follows

$$\hat{\beta}_1 = \bar{y}_A \,, \quad \hat{\beta}_2 = \bar{y}_B - \bar{y}_A \,, \quad \hat{\beta}_3 = \bar{y}_C - \bar{y}_A$$

$$\hat{\beta}_1 = 10, \quad \hat{\beta}_2 = 11 - 10 = 1 \,, \quad \hat{\beta}_3 = 12 - 10 = 2$$

- $\hat{\beta}_1 = 10$ represents the average of tar contents (in milligrams) when the Brand is A
- $\hat{\beta}_2$ is the average change in tar contents when the Brand is B
- $\hat{\beta}_3$ is the average change in tar contents when the Brand is C

Exercise 4.1

Complete the following table.

Sources of variation	D.o.f.	Deviance
Total	17	13.3748
Regression	2	12
Residual	15	1.3748

Then, perform the statistical test to evaluate if the means (of each group) are homogeneous.

We can compute SST, SSR and SSE as follows (you can also check the previous execcise)

$$SSR = \sum_{j=1}^{3} n_j (\bar{y}_j - \bar{y})^2 = 6 \cdot (10-11)^2 + 6 \cdot (0)^2 + 6 \cdot (1)^2 = 12$$

$$SSE = \sum_{j=1}^{3} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 = 0.33 + 0.6648 + 0.38 = 1.3748$$

$$SST = SSR + SSE = 12 + 1.3748 = 13.3748$$

The degrees of freedom for SSR is equal to $J-1 = 3-1 = 2$

" " SSE is equal to $n-J = 18-3 = 15$

" " SST is equal to $n-1 = 18-1 = 17$

System of hypothesis

$$\begin{cases} H_0: \mu_A = \mu_B = \mu_C \\ H_1: \exists j \text{ s.t. } \mu_j \neq \mu_i \ (i \neq j, \ i,j \in \{A,B,C\}) \end{cases}$$

test statistic:

$$f^{obs} = \frac{SSR/(J-1)}{SSE/(n-J)} = \frac{12/2}{1.3748/15} = 65.46407$$

P-value

$$\alpha^{obs} = \mathbb{P}(F_{2,15} > 65.46407) \approx 0 \quad \text{we reject } H_0$$