EXERCISE 4

19th November 2024

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

EXERCISE 1

Assume that $y_1, \dots, y_{200}$ are realizations of independent Gaussian random variables with variance equal to 1 and and mean $\beta_1 + \beta_2 \exp\{z_i\}$ for $i = 1, \dots, 120$, and mean $\beta_1 + \beta_3 \exp\{z_i^2\}$ for $i = 121, \dots, 200$; where the z_i are known constants and $(\beta_1, \beta_2, \beta_3)$ are unknown real parameters.

a) Are the assumptions of a Gaussian linear model satisfied in the above formulation? Motivate the answer.

b) State the parameter space and sample space.

c) Express the model in matrix form: $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, explicitly stating how $\underline{Y}$, X, $\underline{\beta}$, and $\underline{\varepsilon}$ are defined and their dimensions. Write the distribution of $\underline{Y}$ and $\underline{\varepsilon}$.

d) Obtain the expression of the matrix $X^T X$ and the vector $X^T \underline{y}$; state how these elements should be used to obtain the maximum likelihood estimate $\underline{\hat{\beta}}$.

e) Write the distribution of the maximum likelihood estimator $\underline{\hat{B}}$.

f) Let $\underline{e} = \underline{y} - X\underline{\hat{\beta}}$ be the vector of the residuals. State which of the following identities are satisfied and motivate the answer:

$$\sum_{i=1}^{200} e_i = 0 \qquad \sum_{i=1}^{200} e_i z_i = 0 \qquad \sum_{i=1}^{200} e_i z_i^2 = 0$$

$$\sum_{i=1}^{200} e_i \exp\{z_i\} = 0 \qquad \sum_{i=1}^{200} e_i \exp\{z_i^2\} = 0 \qquad \sum_{i=1}^{120} e_i \exp\{z_i\} = 0$$

$Y_i \sim N(\beta_1 + \beta_2 e^{z_i}, 1)$ indep. $i = 1, \dots, 120$

$Y_i \sim N(\beta_1 + \beta_3 e^{z_i^2}, 1)$ indep. $i = 121, \dots, 200$

1.a)

1. Normality, Homoschedesticity and independence
2. The model is linear in β_1, β_2 and β_3 $\quad \mathbb{E}[Y_i] = \beta_1 + \beta_2 e^{z_i} = \beta_1 e^{z_i} / \beta_2$ NOT LINEAR!
3. The covariates are linearly indipendent

   -> We can easily check that $z_1, \dots, z_n$ does not depend on each other

1.b)

sample space $\mathcal{Y} = \mathbb{R}^{200}$ (if $n = 100 \rightarrow \mathcal{Y} = \mathbb{R}^{100}$)

parameter space $\Theta = \mathbb{R}^3$ (space of $(\beta_1, \beta_2, \beta_3)$, σ^2 known)

1.c)

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad i = 1, \dots, 200$$

- $x_{i2} = \begin{cases} e^{z_i} & \text{if } i = 1, \dots, 120 \\ 0 & \text{otherwise} \end{cases}$
- $x_{i3} = \begin{cases} e^{z_i^2} & \text{if } i = 121, \dots, 200 \\ 0 & \text{otherwise} \end{cases}$
- $\varepsilon_i \sim N(0, 1)$

In matrix form we have:

$\underline{Y} = [Y_1, \dots, Y_{120}, Y_{121}, \dots, Y_{200}]^T$ vector of response variables (dim. 200×1)

$\underline{Y} \sim N_{200}(X\underline{\beta}, I_{200})$, $I_{200} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}_{200 \times 200}$

$$X_{200 \times 3} = [\underline{1} \ \underline{x}_2 \ \underline{x}_3] = \begin{bmatrix} 1 & e^{z_1} & 0 \\ 1 & e^{z_2} & \vdots \\ \vdots & \vdots & \vdots \\ 1 & e^{z_{120}} & 0 \\ \vdots & 0 & e^{z_{121}^2} \\ \vdots & \vdots & \vdots \\ 1 & 0 & e^{z_{200}^2} \end{bmatrix} \qquad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}_{3 \times 1}$$ vector of parameters

$\underline{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_{120}, \varepsilon_{121}, \dots, \varepsilon_{200}]$ vector of the error terms

$\underline{\varepsilon} \sim N_{200}(0, I_{200})$

1.d)

$$X^T X = \begin{bmatrix} \underline{1}^T \\ \underline{x}_2^T \\ \underline{x}_3^T \end{bmatrix} [\underline{1} \ \underline{x}_2 \ \underline{x}_3] = \begin{bmatrix} \underline{1}^T\underline{1} & \underline{1}^T\underline{x}_2 & \underline{1}^T\underline{x}_3 \\ \underline{x}_2^T\underline{1} & \underline{x}_2^T\underline{x}_2 & \underline{x}_2^T\underline{x}_3 \\ \underline{x}_3^T\underline{1} & \underline{x}_3^T\underline{x}_2 & \underline{x}_3^T\underline{x}_3 \end{bmatrix} = \begin{bmatrix} 200 & \sum_{i=1}^{120} e^{z_i} & \sum_{i=121}^{200} e^{z_i^2} \\ \sum_{i=1}^{120} e^{z_i} & \sum_{i=1}^{120} e^{2z_i} & 0 \\ \sum_{i=121}^{200} e^{z_i^2} & 0 & \sum_{i=121}^{200} e^{2z_i^2} \end{bmatrix}$$

$$X^T y = \begin{bmatrix} \underline{1}^T \\ \underline{x}_2^T \\ \underline{x}_3^T \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_{200} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{200} y_i \\ \sum_{i=1}^{124} e^{z_i} y_i \\ \sum_{i=124}^{200} e^{z_i^2} y_i \end{bmatrix}_{3 \times 1}$$

$\rightarrow$ the MLE $\hat{\underline{\beta}}$ is equal to $\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{y}$

1.e)

$$\hat{\underline{\beta}}(\underline{y}) \sim N_3(\underline{\beta}, (X^T X)^{-1})$$

↳ since $\sigma^2 = 1$

1.f)

$$\underline{e} = \underline{y} - X\hat{\underline{\beta}}$$

- $\sum_{i=1}^{200} e_i = 0$ $\Rightarrow$ the model includes the intercept, then the sum of the residuals is equal to zero
- $\sum_{i=1}^{200} e_i z_i \neq 0$ $\Rightarrow$ since $[z_1, \ldots, z_{200}] \notin X^T$ and $X^T \underline{e} = 0$
- $\sum_{i=1}^{200} e_i z_i^2 \neq 0$
- $\sum_{i=1}^{200} e_i e^{z_i} \neq 0$
- $\sum_{i=1}^{200} e_i e^{z_i^2} \neq 0$
- $\sum_{i=1}^{120} e_i e^{z_i} = 0$

EXERCISE 2

Consider the following models, for $i = 1, \dots, n$

1. $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 \log_{10} x_{i,3} + \beta_4 x_{i,4}^2 + \varepsilon_i$ and $\varepsilon_i \sim N(0, \sigma^2)$ independent.
2. $Y_i = \dfrac{\beta_1 + \beta_2 x_{i,2}}{\beta_3 x_{i,1}} + \varepsilon_i$ and $\varepsilon_i \sim N(0, \sigma^2)$ independent.
3. $\log(Y_i) = \dfrac{\beta_2 x_{i,1} + \beta_3 \log(x_{i,3})}{x_{i,2}} + \varepsilon_i$ and $\varepsilon_i \sim N(0, \sigma^2)$ independent.
4. $Y_i = \beta_1 \, x_{i,2}{}^{\beta_2} \exp\{\varepsilon_i\}$ and $\varepsilon_i \sim N(0, 1)$ independent.

Answer the following questions:

a) For each model, indicate whether it is a linear regression model. If it is not, explain why and whether it can be expressed in the form $Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i$ by a suitable transformation and write explicitly such transformation.

b) Consider model 4 appropriately transformed, denoting with Y^*, $x^*_{i,2}$, (β_1^*, β_2^*) and ε_i^* the transformed quantities. Express it in the matrix form $\underline{Y}^* = X^* \underline{\beta}^* + \underline{\varepsilon}^*$, explicitly stating $\underline{Y}^*$ (and its distribution), X^*, $\underline{\beta}^*$, and $\underline{\varepsilon}^*$.

c) Write the expression of the maximum likelihood estimator $\underline{\hat{B}}^*$ and its exact distribution.

d) Let $\underline{e} = \underline{y}^* - X^* \underline{\hat{\beta}}^*$ be the vector of the residuals. State which of the following identities are satisfied and motivate the answer:

$$\sum_{i=1}^n e_i = 0 \qquad \sum_{i=1}^n e_i x_{i,2} = 0$$

$$\sum_{i=1}^n e_i \log(x_{i,2}) = 0 \qquad \sum_{i=1}^n e_i \log(x_{i,2}^2) = 0$$

2.a)

- model 1 is linear
- model 2 is NOT linear and cannot be transformed
- model 3 is linear

$$\underbrace{\log y_i}_{y_i^*} = \beta_2 \underbrace{\frac{x_{i2}}{x_{i2}}}_{x_{i,2}^*} + \beta_3 \underbrace{\frac{\log x_{i3}}{x_{i2}}}_{x_{i,3}^*} + \varepsilon_i$$

- model 4 is NOT a linear model, moreover the error term is not additive, but we can apply the following transformation

$$\log(y_i) = \log\left(\beta_1 x_{i,2}{}^{\beta_2} \exp(\varepsilon_i)\right)$$

$$\underbrace{\log(y_i)}_{y_i^*} = \underbrace{\log(\beta_1)}_{\beta_1^*} + \underbrace{\beta_2}_{\beta_2^*}\underbrace{\log(x_{i,2})}_{x_{i,2}^*} + \underbrace{\varepsilon_i}_{\varepsilon_i^* \sim N(0,1)}$$

2.b)

$$y_i^* = \beta_1^* + \beta_2^* x_{i,2}^* + \varepsilon_i^*$$

$$\underline{y}^* = X^* \underline{\beta}^* + \underline{\varepsilon}^*$$

- $\underline{y}^*$ is an n-dimensional vector of random variables

$$\underline{y}^* = [y_1^*, \dots, y_n^*]^\top = [\log y_1, \dots, \log y_n]^\top \qquad \underline{y}^* \sim N_n(X^*\underline{\beta}^*, I)$$

- X^* is an $(n \times 2)$ matrix of known constants

$$X^* = [\underline{1}_n \ \underline{x}_2^*]$$

- $\underline{\beta}^*$ is a 2-dimensional vector of unknown parameters

$$\underline{\beta}^* = \begin{bmatrix} \beta_1^* \\ \beta_2^* \end{bmatrix}$$

- $\underline{\varepsilon}^*$ is a n-dimensional vector of random variables $\varepsilon_i^* \sim N(0,1)$

$$\underline{\varepsilon}^* \sim N_n(\underline{0}, I_n)$$

2.c)

$$\hat{\underline{\beta}}^* = (X^{*\top} X^*)^{-1} X^{*\top} \underline{y}^* \quad \text{where} \quad \hat{\underline{\beta}}^* \sim N_2(\underline{\beta}^*, (X^{*\top} X^*)^{-1})$$

2.d)

- $\sum_{i=1}^{n} e_i = 0 \Rightarrow$ the model includes the intercept
- $\sum_{i=1}^{n} e_i x_{i2} \neq 0$ · $\sum_{i=1}^{n} e_i \log(x_{i2}) = 0$
- $\sum_{i=1}^{n} e_i \log(x_{i,2}^2) = \sum_{i=1}^{n} e_i \cdot 2 \cdot \log(x_{i,2}) = 2 \underbrace{\sum_{i=1}^{n} e_i \log(x_{i,2})}_{=0} = 0$

EXERCISE 3

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

3.1)

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 model is the multiple linear regression model

$$\text{Volume}_i = \beta_1 + \beta_2 \text{Diameter}_i + \beta_3 \text{Height}_i + \varepsilon_i,$$

with the following assumptions

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

ii) linear independence among the covariates

iii) linearity of $(\beta_1, \beta_2, \beta_3)$

i) and iii) imply that $y_i \overset{\perp\!\!\!\perp}{\sim} N(\mu_i, \sigma^2)$ where $\mu_i = \beta_1 + \beta_2 x_{1,i} + \beta_3 x_{2,i}$

Let's consider the following model to compute the volume of a tree

$$\text{Volume}_i = \pi \left(\frac{\text{Diameter}_i}{2}\right)^2 \text{Height}_i + \varepsilon_i,$$

we can apply the following transformation:

$$\log(\text{Volume}_i) = \log(\pi) + 2 \cdot \log\left(\frac{\text{Diameter}}{2}\right) + \log(\text{Height}) + \log(\varepsilon_i)$$

Then we can consider the following model

$$y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,2} + \varepsilon_i, \text{ where}$$

- $x_{i,1} = \log\left(\frac{\text{Diameter}_i}{2}\right)$
- $y_i = \log(\text{Volume}_i)$
- $x_{i,2} = \log(\text{Height}_i)$

3.2)

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.

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

where

- $\underline{\varepsilon} \sim N(0, \sigma^2 I)$
- $X = \begin{bmatrix} 1 & X_{1,2} & X_{1,3} \\ 1 & X_{2,2} & X_{2,3} \\ \vdots & \vdots & \vdots \\ 1 & X_{31,2} & X_{31,3} \end{bmatrix} \quad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{31} \end{bmatrix}$

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

$$\hat{\underline{\beta}} = (X^TX)^{-1}X^Ty$$

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

Then, the estimated model is:

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

3.3)

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 $\hat{\beta}$.

- To compute the unbiased estimate of σ^2, we need to do the following

$$\hat{S}^2 = \frac{1}{n-p}\,\underline{e}^T\underline{e} = \frac{n}{n-p}\,\frac{1}{n}\,\underline{e}^T\underline{e} = \frac{n}{n-p}\cdot\hat{\sigma}^2 = \frac{31}{28}\cdot 0.00598 = 0.0066207$$

- We know that $\hat{\underline{B}}$ is distributed in the following way:

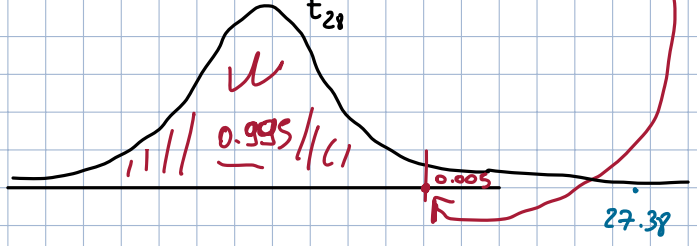
$$\hat{\underline{B}} \sim N_p\left(\underline{\beta}, \sigma^2 (X^T X)^{-1}\right)$$

$$\widehat{Var}\left(\hat{\underline{B}}\right) = \hat{S}^2 \cdot (X^T X)^{-1} = 0.0066207 \begin{bmatrix} 96.572 & 3.139 & -24.165 \\ & 0.849 & -1.227 \\ & & 6.310 \end{bmatrix} = \begin{bmatrix} 0.639 & 0.028 & -0.16 \\ & 0.0056 & -0.0081 \\ & & 0.042 \end{bmatrix}$$

3.4)

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).

1 - α	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995
$t_{22;p}$	0.6858	0.8583	1.0614	1.3212	1.7171	2.0739	2.5083	2.8188
$t_{23;p}$	0.6853	0.8575	1.0603	1.3195	1.7139	2.0687	2.4999	2.8073
$t_{24;p}$	0.6848	0.8569	1.0593	1.3178	1.7109	2.0639	2.4922	2.7969
$t_{25;p}$	0.6844	0.8562	1.0584	1.3163	1.7081	2.0595	2.4851	2.7874
$t_{26;p}$	0.684	0.8557	1.0575	1.315	1.7056	2.0555	2.4786	2.7787
$t_{27;p}$	0.6837	0.8551	1.0567	1.3137	1.7033	2.0518	2.4727	2.7707
$t_{28;p}$	0.6834	0.8546	1.056	1.3125	1.7011	2.0484	2.4671	2.7633
$t_{29;p}$	0.683	0.8542	1.0553	1.3114	1.6991	2.0452	2.462	2.7564
$t_{30;p}$	0.6828	0.8538	1.0547	1.3104	1.6973	2.0423	2.4573	2.75



- $\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$ $\quad t_2^{obs} = \dfrac{\hat{\beta}_2 - 0}{\sqrt{\hat{Var}(\hat{\beta}_2)}} = \dfrac{2.0494}{\sqrt{0.0056}} = 27.38626$

$\alpha^{obs} = 2 \cdot \mathbb{P}(t_{28} \geq 27.38626) \approx 2 \cdot 0 = 0 \Rightarrow$ we reject H_0

- $\begin{cases} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \exists j \in \{2,3\} \text{ s.t. } \beta_j \neq 0 \end{cases}$ $\quad f_{obs} = \dfrac{\frac{8.309 - 0.1855}{3-1}}{\frac{0.1855}{31-3}} = 613.094$

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

	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995
$f_{1,24;p}$	1.3898	1.7367	2.2116	2.9271	4.2597	5.7166	7.8229	9.5513
$f_{2,24;p}$	1.4695	1.7224	2.0553	2.5383	3.4028	4.3187	5.6136	6.6609
$f_{1,25;p}$	1.387	1.7328	2.2057	2.9177	4.2417	5.6864	7.7698	9.4753
$f_{2,25;p}$	1.4661	1.7176	2.0487	2.5283	3.3852	4.2909	5.568	6.5982
$f_{1,26;p}$	1.3845	1.7292	2.2004	2.9091	4.2252	5.6586	7.7213	9.4059
$f_{2,26;p}$	1.4629	1.7133	2.0425	2.5191	3.369	4.2655	5.5263	6.5409
$f_{1,27;p}$	1.3821	1.7258	2.1954	2.9012	4.21	5.6331	7.6767	9.3423
$f_{2,27;p}$	1.46	1.7093	2.0369	2.5106	3.3541	4.2421	5.4881	6.4885
$f_{1,28;p}$	1.38	1.7227	2.1908	2.8938	4.196	5.6096	7.6356	9.2838
$f_{2,28;p}$	1.4573	1.7056	2.0317	2.5028	3.3404	4.2205	5.4529	6.4403
$f_{1,29;p}$	1.378	1.7199	2.1866	2.887	4.183	5.5878	7.5977	9.2297
$f_{2,29;p}$	1.4547	1.7022	2.0268	2.4955	3.3277	4.2006	5.4204	6.3958
$f_{1,30;p}$	1.3761	1.7172	2.1826	2.8807	4.1709	5.5675	7.5625	9.1797
$f_{2,30;p}$	1.4524	1.699	2.0223	2.4887	3.3158	4.1821	5.3903	6.3547

EXERCISE 4

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).

-> Multiple linear regression model

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

- $D_i = \begin{cases} 1 & \text{if } Smoking_i = \text{"yes"} \\ 0 & \text{if } Smoking_i = \text{"no"} \end{cases}$

-> Matrix form:

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

- $Y = \begin{bmatrix} Weight_1 \\ \vdots \\ Weight_{32} \end{bmatrix}$
- $X = \begin{bmatrix} 1 & Gest_1 & D_1 \\ \vdots & \vdots & \vdots \\ 1 & Gest_{32} & D_{32} \end{bmatrix}$
- $\underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{32} \end{bmatrix}$

4.1)

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.

We have the following relation

$$\hat{Weight}^{(n)} = \hat{\beta}_1 + \hat{\beta}_2\ Gest \qquad (D_i = 0)$$

$$\hat{Weight}^{(y)} = (\hat{\beta}_1 + \hat{\beta}_3) + \hat{\beta}_2\ Gest \qquad (D_i = 1)$$

If $\hat{\beta}_1 = -2390$ and $\hat{\beta}_2 = 143$

$$\Rightarrow \hat{\beta}_1 + \hat{\beta}_3 = -2635 \Rightarrow \hat{\beta}_3 = -2635 + 2390 = -245$$

$$Weight_i = \beta_1 + \beta_2 Gest_i + \beta_3 Smoking_i$$

By keeping constant the time of gestation, the expected weight of a baby born from a smoking mother is smaller than 245 grams than a baby born from a non-smoking mother

4.2)

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") $Weight_i = -2546.138 + 147.207\ Gest_i$

(Smoking = "yes") $Weight_i = -2546.138 + 147.207\ Gest_i + 71.574 - 8.178\ Gest_i$

$$= -2474.564 + 139.029\ Gest_i$$

4.3)

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.

- $D.o.f.\ (M_0) = n - p_0 = 32 - 2 = 30$
  $D.o.f.\ (M_4) = n - p_4 = 32 - 4 = 28$

- $\begin{cases} H_0: \mu_N = \mu_Y \\ H_1: \mu_N \neq \mu_Y \end{cases}$ $\qquad F^{obs} = \dfrac{(839951.03 - 384391.46)/2}{(384391)/32-4} = 16.59$

$\alpha^{obs} = P(f_{2,28} \geq 16.59) \approx 0 \Rightarrow$ we reject H_0

4.4)

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.

$$\begin{cases} H_0: \beta_4 = 0 \\ H_1: \beta_4 \neq 0 \end{cases} \qquad f^{obs} = \frac{SSE_{NO\,INTER} - SSE_{INTER.} / (P_{INT} - P_{NO\,INT})}{SSE_{INTER.} / (n - P_{NO\,INT})} = \frac{(387069.83 - 384391.6)/1}{384391.6/28} = 0.20$$

	0.01	0.02	0.05	0.1	0.15	0.2	0.3	0.325	0.35	0.4
$f_{1,24;p}$	2e-04	6e-04	0.004	0.0161	0.0365	0.0656	0.1521	0.1802	0.2112	0.2824
$f_{2,24;p}$	0.0101	0.0202	0.0514	0.1058	0.1636	0.2252	0.362	0.3996	0.4386	0.5219
$f_{1,25;p}$	2e-04	6e-04	0.004	0.0161	0.0365	0.0656	0.1519	0.18	0.2109	0.2821
$f_{2,25;p}$	0.0101	0.0202	0.0514	0.1058	0.1636	0.2251	0.3618	0.3993	0.4383	0.5214
$f_{1,26;p}$	2e-04	6e-04	0.004	0.0161	0.0365	0.0655	0.1518	0.1798	0.2107	0.2818
$f_{2,26;p}$	0.0101	0.0202	0.0514	0.1058	0.1635	0.2251	0.3616	0.399	0.438	0.521
$f_{1,27;p}$	2e-04	6e-04	0.004	0.0161	0.0365	0.0655	0.1517	0.1797	0.2106	0.2816
$f_{2,27;p}$	0.0101	0.0202	0.0514	0.1058	0.1635	0.225	0.3614	0.3988	0.4377	0.5206
$f_{1,28;p}$	2e-04	6e-04	0.004	0.0161	0.0364	0.0654	0.1516	0.1795	0.2104	0.2813
$f_{2,28;p}$	0.0101	0.0202	0.0514	0.1058	0.1635	0.2249	0.3613	0.3986	0.4375	0.5203
$f_{1,29;p}$	2e-04	6e-04	0.004	0.0161	0.0364	0.0654	0.1514	0.1794	0.2102	0.2811
$f_{2,29;p}$	0.0101	0.0202	0.0514	0.1057	0.1634	0.2249	0.3611	0.3984	0.4372	0.5199
$f_{1,30;p}$	2e-04	6e-04	0.004	0.0161	0.0364	0.0653	0.1513	0.1793	0.2101	0.2809
$f_{2,30;p}$	0.0101	0.0202	0.0514	0.1057	0.1634	0.2248	0.3609	0.3982	0.437	0.5196

$$\alpha^{obs} = P(f_{1,28} \geq 0.20) \approx 0.65 \Rightarrow \text{we don't reject } H_0$$

4.5)

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.

Consider

$$\mathcal{M}_0: \underline{Y} = \beta_1 \underline{1} + \underline{\mathcal{E}} \quad \text{where } \tilde{\sigma}^2 = \frac{\underline{e}_0^T \underline{e}_0}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

($\underline{e}_0$: residuals of $\mathcal{M}_0$)

$$\mathcal{M}_4: y_i = \beta_0 + \beta_2 \text{Guest}_i + \beta_3 \text{Smoke}_i + \beta_4 \text{Gest}_i \times \text{Smoke}_i + \mathcal{E}_i$$

$$\text{where } \hat{\sigma}^2 = \frac{\underline{e}_4^T \underline{e}_4}{n}$$

($\underline{e}_4$: residuals of $\mathcal{M}_4$)

We know then:

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

We can find R^2:

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

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

And the R^2_{adj}:

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

R^2 is an index that quantify the quality of model. One drawback of this index is that it increases if the number of covariates increases.

R^2_{adj} solve this problem by adjusting R^2 with respect to the number of covariates.

In our problem we can see that with the adjustment R^2 slightly decreases.