Feuille de Travaux Dirigés n° 2
Correction de Régression linéaire multiple

Exercice II.1 Dans cet exercice, nous n'utiliserons que le logiciel `R` pour faire les calculs des valeurs critiques des quantiles de Fisher.
Question 1. La somme des carrés dûe à la régression pour l'ensemble des trois variables est égale à :

$$981,326 + 190,232 + 129,431 = 1300,989.$$

Nous pouvons également calculer la somme ainsi :

$$1743,281 - 442,292 = 1300,989.$$

Question 2. La proportion de la variation dans le niveau d'anxiété est égale à :

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{1300,989}{1743,281} = 0,746,$$

ou encore $74,60\%$.

Question 3. Pour répondre à cette question, il faudrait s'assurer que les trois hypothèses du modèle sont vérifiées. Malheureusement nous ne pourrons pas le faire ici puisque nous ne connaissons pas les valeurs des observations. Donc nous allons supposer que les trois hypothèses sont vérifiées mais dans la pratique il faudrait les vérifier ABSOLUMENT.

Pour conclure que dans l'ensemble les trois variables ont un effet significatif sur le niveau d'anxiété, il faut faire **un test de Fisher**. Le modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,$$

où ε est la variable résiduelle sur laquelle les trois hypothèses sont faites.
L'hypothèse nulle :

$$\mathcal{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

contre l'hypothèse alternative :

$$\mathcal{H}_1 : \exists j = 1,\ 2,\ \text{ou } 3,\ \beta_j \neq 0.$$

Calculons la statistique du test de Fisher observée qui est égale à :

$$F_{obs} = \frac{SC_{reg}/ddl}{SC_{res}/ddl} = \frac{1300,989/3}{442,292/(22-3-1=18)} \simeq 17,649.$$

Le quantile de la loi de Fisher critique lu dans la table des quantiles de la loi de Fisher à 95% est égal à :

$$F_{c,3,18} = 3,159908.$$

La statistique du test de Fisher observée est plus grande que le quantile de la loi de Fisher critique, à 95%**.** Donc nous sommes dans la zone de rejet de l'hypothèse nulle $\mathcal{H}_0$. Donc nous décidons de refuser l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$, c'est-à-dire :

$$\exists j = 1,\ 2,\ \text{ou } 3,\ \beta_j \neq 0.$$

Question 4.

Source de variation	Somme des carrés	*ddl*
Régression due à X_1	$981,326$	**1**
Résiduelle	$\mathbf{761,955}$	**20**
Totale	$\mathbf{1743,281}$	**21**

Question 5. Même remarque qu'à la question 3 de cet exercice.

a) Le modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

L'hypothèse nulle

$$\mathcal{H}_0 : \beta_1 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \beta_1 \neq 0.$$

Calculons la statistique du test de Fisher observée qui est égale à :

$$F_{obs} = \frac{SC_{reg}/ddl}{SC_{res}/ddl} = \frac{981,326/1}{761,955/(22-1-1=20)} = 25,758.$$

Le quantile de la loi de Fisher critique lu dans la table des quantiles de la loi de Fisher à 95% est égal à :

$$F_{c,1,20} = 4,351244.$$

La statistique du test de Fisher observée est plus grande que le quantile de la loi de Fisher critique. Donc nous sommes dans la zone de rejet de l'hypothèse nulle $\mathcal{H}_0$. Donc nous décidons de refuser l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$, c'est-à-dire :

$$\beta_1 \neq 0.$$

b) Le modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

L'hypothèse nulle

$$\mathcal{H}_0 : \beta_2 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \beta_2 \neq 0.$$

Calculons la statistique du test de Fisher observée qui est égale à :

$$F_{obs} = \frac{SC_{reg}/ddl}{SC_{res}/ddl} = \frac{190,232/1}{571,723/(22-2-1=19)} = 6,332.$$

Le quantile de la loi de Fisher critique lu dans la table des quantiles de la loi de Fisher à 95% est égal à :

$$F_{c,1,19} = 4,38075.$$

La statistique du test de Fisher observée est plus grande que le quantile de la loi de Fisher critique. Donc nous sommes dans la zone de rejet de l'hypothèse nulle $\mathcal{H}_0$. Donc nous décidons de refuser l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$, c'est-à-dire :

$$\beta_2 \neq 0.$$

c) Le modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

L'hypothèse nulle

$$\mathcal{H}_0 : \beta_3 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \beta_3 \neq 0.$$

Calculons la statistique du test de Fisher observée qui est égale à :

$$F_{obs} = \frac{SC_{reg}/ddl}{SC_{res}/ddl} = \frac{129,431/1}{442,292/(22-3-1=18)} \simeq 5,267.$$

Le quantile de la loi de Fisher critique lu dans la table des quantiles de la loi de Fisher à 95% est égal à :

$$F_{c,1,18} = 4,413873.$$

La statistique du test de Fisher observée est plus grande que le quantile de la loi de Fisher critique. Donc nous sommes dans la zone de rejet de l'hypothèse nulle $\mathcal{H}_0$. Donc nous décidons de refuser l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$, c'est-à-dire :

$$\beta_3 \neq 0.$$

Question 6. La valeur du coefficient R^2 associée à l'estimation du modèle spécifié en 5.a) est égale à :

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{981,326}{1743,281} = 0,563.$$

La valeur du coefficient R^2 associée à l'estimation du modèle spécifié en 5.b) est égale à :

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{1171,558}{1743,281} = 0,672.$$

La valeur du coefficient R^2 associée à l'estimation du modèle spécifié en 5.c) est égale à :

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{1300,989}{1743,281} = 0,746.$$

Question 7. Le modèle qui semble le mieux adapté est le modèle 5.c) car ce modèle a le plus grand coefficient de détermination R^2.

Remarque : Pour l'instant à cette étape, le cours de choix du modèle n'a pas été fait, donc nous ne calculons pas le R^2 adjusté pour voir quel serait le modèle le mieux approprié. Et si nous appliquions le cours du choix de modèle, nous calculerions le coefficient R^2 ajusté du second modèle, c'est-à-dire celui en 5.b) et le coefficient R^2 ajusté du troisième modèle, c'est-à-dire celui en 5.c)

Exercice II.2 Avant de lire le corrigé de cet exercice, il serait préférable de vérifier toutes les hypothèses du modèle, à savoir les trois hypothèses du modèles linéaire gaussien.

Question 1. Quel pourcentage de variation dans la résistance à la rupture est expliquée par chacune des régressions ?

Pour la régression de la résistance à la rupture (Y) en fonction de l'épaisseur (X_1) :

$$R^2_{Y,X_1} = \frac{SC_{reg}}{SC_{tot}} = \frac{980,64}{1420,67} = 0,6903.$$

Pour la régression de la résistance à la rupture (Y) en fonction de la densité (X_2) :

$$R^2_{Y,X_2} = \frac{SC_{reg}}{SC_{tot}} = \frac{643,57}{1420,67} = 0,453.$$

Pour la régression de la résistance à la rupture (Y) en fonction de l'épaisseur (X_1) et de la densité (X_2) :

$$R^2_{Y,X_1,X_2} = \frac{SC_{reg}}{SC_{tot}} = \frac{1204,86}{1420,67} = 0,8481.$$

Question 2. Pour chaque régression, le tableau est le suivant :

	Carré moyen résiduel	Écart-type des résidus
Régression avec X_1	44,003	6,633
Régression avec X_2	77,710	8,815
Régression avec X_1, X_2	23,979	4,897

Question 3. Le tableau d'analyse de variance pour la régression comportant les deux variables explicatives est le suivant :

Source de variation	*ddl*	Somme des carrés	Carrés moyens	F_{obs}
Régression(X_1, X_2)	**2**	**1204,86**	**602,43**	**25,123**
Résiduelle	**9**	**215,81**	**23,979**	
Totale	**11**	**1420,67**		

Question 4. Tester au seuil de signification $\alpha = 5\%$, l'hypothèse nulle $\mathcal{H}_0 : \beta_1 = \beta_2 = 0$ contre l'hypothèse alternative $\mathcal{H}_1$: au moins un des $\beta \neq 0$. Quelle est votre conclusion ?

C'est pour cette question qu'il est important de regarder si les hypothèses sont vérifiées.

```
> model12<-lm(Y~ X1 + X2,data=Exo2TD8)
> shapiro.test(residuals(model12))
Shapiro-Wilk normality test
data: residuals(model12)
W = 0.9408, p-value = 0.5082
```

Calculons la statistique du test de Fisher observée qui est égale à :

$$F_{obs} = \frac{SC_{reg}/ddl}{SC_{res}/ddl} = \frac{1204,86/2}{215,81/(12-2-1=9)} \simeq 25,123.$$

Le quantile de loi de Fisher critique lu dans la table des quantiles de la loi de Fisher à 95% est égal à :

$$F_{c,2,9} = 4,256495.$$

La statistique du test de Fisher observée est plus grande que le quantile de loi de Fisher critique. Donc nous sommes dans la zone de rejet de l'hypothèse nulle $\mathcal{H}_0$. Donc nous décidons de refuser l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$, c'est-à-dire :

$$\exists j = 1, \text{ ou } 2, \quad \beta_j \neq 0.$$

Question 5. Dans le cas du modèle de régression ne comportant que l'épaisseur du matériau comme variable explicative, déterminer un intervalle de confiance à 95% pour β_1.

C'est également pour cette question qu'il est important de regarder si les hypothèses sont vérifiées.

```
> model1<-lm(Y~ X1,data=Exo2TD8)
> shapiro.test(residuals(model1))
Shapiro-Wilk normality test
data: residuals(model1)
W = 0.9219, p-value = 0.3019
```

Un intervalle de confiance à 95% pour β_1, d'après le complément du cours 8 est égal à :

$$[6,036 - 2,228 * 1,279; 6,036 + 2,228 * 1,279] = [3,187; 8,885],$$

où le quantile de la loi de Student critique lu dans une table des quantiles de la loi de Student à 95% est égal à :

$$t_{c,95\%} = 2,228.$$

Remarque : Le logiciel `R` nous donne également un intervalle de confiance pour β_1 en tapant les lignes suivantes :

```
> model1<-lm(Y~ X1,data=Exo2TD8)
> confint(model1)
```

```
2.5 % 97.5 %
(Intercept) -6.242858 13.28806
X1 3.187036 8.88479
```

Question 6. Avec l'intervalle de confiance calculé à la question 5.), pouvons-nous affirmer, au seuil de signification $\alpha = 5\%$, que la régression est significative entre la résistance à la rupture et l'épaisseur du matériau ? Justifier votre conclusion ?

C'est aussi pour cette question qu'il est important de regarder si les hypothèses sont vérifiées.

La régression est significative entre la résistance à la rupture et l'épaisseur du matériau si le test de Student qui teste si $\beta_1 = 0$ n'est pas vérifié. Calculons la statistique du test de Student observée :

$$t_{obs} = \frac{6,036}{1,279} = 4,721.$$

Le quantile de la loi de Student critique lu dans une table des quantiles de la loi de Student à 95% est égal à :

$$t_{c,95\%} = 2,228.$$

La statistique du test de Student observée est plus grande que le quantile de la loi de Student critique. Par conséquent nous sommes dans la zone de rejet de l'hypothèse nulle $\mathcal{H}_0$. Donc nous décidons de refuser l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$. **Donc la régression est significative entre la résistance à la rupture et l'épaisseur du matériau.**

Remarque : Nous pouvons répondre plus rapidement en disant que l'intervalle de confiance calculé à la question précédente ne contient pas 0. Par conséquent **la régression est significative entre la résistance à la rupture et l'épaisseur du matériau.**

Remarque : Si nous avons les sorties de `R` à notre disposition, nous pouvons conclure directement en regardant la $p-$valeur de X_1.

```
> model1<-lm(Y~ X1,data=Exo2TD8)
> summary(model1)
Call:
lm(formula = Y ~ X1, data = Exo2TD8)
Residuals:
Min 1Q Median 3Q Max
-8.266 -4.887 -1.209 3.232 10.770
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.523 4.383 0.804 0.440237
X1 6.036 1.279 4.721 0.000816 ***
--
```

```
Residual standard error: 6.633 on 10 degrees of freedom
Multiple R-Squared: 0.6903, Adjusted R-squared: 0.6593
F-statistic: 22.29 on 1 and 10 DF, p-value: 0.0008155
```

Cette $p-$valeur est égale à `0.000816`, qui est inférieur à 5%. Donc même conclusion qu'en faisant les calculs à la main précédents.

Question 7. Quel est l'apport marginal de X_2 lorsqu'elle est introduite à la suite de X_1 ?

L'apport marginal de la variable explicative X_2 lorsqu'elle est introduite à la suite de la variable explicative X_1 est égal à :

$$1204,858 - 980,635 = 224,223.$$

Remarque : Nous retrouvons cette valeur en utilisant le logiciel `R` :

```
> model12<-lm(Y~ X1 + X2,data=Exo2TD8)
> anova(model12)
Analysis of Variance Table
Response: Y
Df Sum Sq Mean Sq F value Pr(>F)
X1 1 980.63 980.63 40.8959 0.000126 ***
X2 1 224.22 224.22 9.3509 0.013617 *
Residuals 9 215.81 23.98 --
```

Question 8. Est-ce que la contribution marginale de la variable « densité du matériau », lorsqu'elle est introduite à la suite de la variable « épaisseur du matériau » est significative au seuil $\alpha = 5\%$? Utiliser les deux façons équivalentes d'effectuer ce test.

C'est aussi pour cette question qu'il est important de regarder si les hypothèses sont vérifiées.

Pour répondre à la question : « Est-ce que la contribution marginale de la variable « densité du matériau », lorsqu'elle est introduite à la suite de la variable « épaisseur du matériau » est significative au seuil $\alpha = 5\%$? », **il suffit de faire un test, soit un test de Fisher, soit un test de Student.**

« F partiel »	F_c	t_{obs}	t_c
$\mathbf{9,350}$	$\mathbf{5,120}$	$\mathbf{3,058}$	$\mathbf{2,262}$

Remarque : Pour obtenir la valeur $9,350$, nous calculons la statistique du test de Fisher :

$$F_{obs} \text{ partiel } = \frac{224,223/1}{215,809/9} = 9,350.$$

Pour obtenir la valeur $5,120$, nous lisons dans une table des quantiles de la loi de Fisher :

$$F_{c,1,9} = 5,120.$$

Comme $F_{obs} > F_{c,1,9}$, nous en déduisons que nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et par conséquent nous décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$ à savoir la contribution marginale de la variable « densité du matériau », lorsqu'elle est introduite à la suite de la variable « épaisseur du matériau » est significative au seuil $\alpha = 5\%$.

Remarque : En regardant les sorties de `R` qui sont affichées à la question précédente, nous retrouvons cette valeur de `9.3509` ainsi que la $p-$valeur qui est égale à `0.013617` qui nous permet de conclure directement sans passer par la valeur critique. Bien sûr, nous retrouvons la même conclusion que nous venons d'établir.

Pour obtenir la valeur $3,058$, nous calculons la statistique du test de Student :

$$t_{obs} = \frac{11,072}{3,621} = 3,058.$$

Pour obtenir la valeur $2,262$, nous lisons dans une table des quantiles de la loi de Student :

$$t_{c,9} = 2,262.$$

Comme $t_{obs} > t_{c,9}$, nous en déduisons que nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et par conséquent nous décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$ à savoir la contribution marginale de la variable « densité du matériau », lorsqu'elle est introduite à la suite de la variable « épaisseur du matériau » est significative au seuil $\alpha = 5\%$.

Remarque : Si nous avons à notre disposition les sorties de `R`, nous obtenons la même valeur, à savoir `3.058` et la même conclusion, en regardant la $p-$valeur associée à X_2 qui est égale à `0.013617`, que celle que nous venons d'établir.

```
> summary(model12)
Call:
lm(formula = Y ~ X1 + X2, data = Exo2TD8)
Residuals:
Min 1Q Median 3Q Max
-6.897 -2.135 -1.126 1.714 10.122
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -30.081 11.455 -2.626 0.027542 *
X1 4.905 1.014 4.838 0.000923 ***
X2 11.072 3.621 3.058 0.013617 *
--
Residual standard error: 4.897 on 9 degrees of freedom
Multiple R-Squared: 0.8481, Adjusted R-squared: 0.8143
F-statistic: 25.12 on 2 and 9 DF, p-value: 0.0002075
```

Question 9. Nous voulons obtenir diverses estimations et prévisions de la résistance à la rupture. Quelle est, en moyenne, la résistance à la rupture de jouets dont l'épaisseur du matériau utilisé et la densité du matériau sont ceux indiqués dans le tableau suivant ?

C'est aussi pour cette question qu'il est important de regarder si les hypothèses sont vérifiées.

Epaisseur X_1	Densité X_2	Estimation de la résistance moyenne	Écart-type de l'estimation
4	3,8	**31,612**	2,10
3	3,4	**22,278**	1,43
4	2,9	**21,647**	2,57

Remarque : Si nous avons le logiciel `R` à notre disposition ou si nous avons les sorties de ce dernier, nous obtenons les mêmes réponses :

```
> predict(model12,data.frame(X1 = 4, X2 = 3.8),se.fit=TRUE)
$fit
[1] 31.61175
$se.fit
[1] 2.100369
> predict(model12,data.frame(X1 = 3, X2 = 3.4),se.fit=TRUE)
[1] 22.27821
$se.fit
[1] 1.431540
> predict(model12,data.frame(X1 = 4, X2 = 2.9),se.fit=TRUE)
[1] 21.64689
$se.fit
[1] 2.573244
```

Question 10. Entre quelles valeurs peut se situer la résistance moyenne à la rupture, pour des jouets dont l'épaisseur du matériau est $X_1 = 4$ et de densité $X_2 = 3,8$, si l'entreprise utilise un niveau de confiance à 95% ?

C'est pour cette question qu'il est important de regarder si les hypothèses sont vérifiées.

Un intervalle de confiance à 95% est égal à :

$$[31,612 - 4,751\,;\, 31,612 + 4,751] = [26,861\,;\, 36,363].$$

Remarque : Si nous avons les sorties du logiciel `R` à disposition, les calculs sont moins fastidieux (- ;

```
> data.frame(X1 = 4, X2 = 3.8)
X1X2
1 4 3.8
> predict(model12,data.frame(X1 = 4, X2 = 3.8),interval="confidence")
fit lwr upr
[1,] 31.61175 26.86038 36.36311
```

Question 11. Quelle est la marge d'erreur dans l'estimation effectuée à la question 10. ?

La marge d'erreur dans l'estimation effectuée à la question 10. est égale à

$$36,363 - 26,861 = 2 \times 4,751.$$

Question 12. Nous désirons un intervalle de prévision de la résistance à la rupture pour un jouet ayant comme épaisseur de matériau et de densité ceux précisés en 10. Quel est cet intervalle au niveau 95% ?

Remarque : Les calculs étant tellement fastidieux, que le logiciel `R` est indispensable pour répondre à ce type de question, à savoir le calcul des intervalles de prévision.

```
> predict(model12,data.frame(X1 = 4, X2 = 3.8),interval="predict")
fit lwr upr
[1,] 31.61175 19.55839 43.6651
```

Un intervalle de prévision à 95% **est égal à :**

$$[19,55839\,;\,43,6651].$$

Exercice II.3 Question 1. Complétons le tableau d'ANOVA :

Source de variation	Somme des carrés	*ddl*	Carrés moyens	F_{obs}
Régression	$1\,504,4$	**2**	$\mathbf{752,2}$	$\mathbf{38,37}$
Résiduelle	$\mathbf{176,4}$	**9**	$19,6$	
Totale	$1\,680,8$	**11**		

Question 2. Pour répondre à cette question, il faudrait s'assurer que les trois hypothèses du modèle sont vérifiées. Malheureusement nous ne pourrons pas le faire ici puisque nous ne connaissons pas les valeurs des observations. Donc nous allons supposer que les trois hypothèses sont vérifiées mais dans la pratique il faudrait les vérifier ABSOLUMENT.

Testons l'hypothèse nulle

$$\mathcal{H}_0 : \quad \beta_1 = \beta_2 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \quad \exists j = 1, \text{ ou } 2, \quad \beta_j \neq 0.$$

Nous avons trouvé d'après le tableau d'ANOVA :

$$F_{obs} = 38,37.$$

Nous lisons dans la table des quantiles de la loi de Fisher, à 95%, pour $\nu_1 = 2$ et $\nu_2 = 9$:

$$F_{c,2,9} = 4,256495.$$

Comme $F_{obs} > F_{c,2,9}$, nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et par conséquent nous décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$, c'est-à-dire :

$$\exists j = 1 \text{ ou } 2, \quad \beta_j \neq 0.$$

Remarque : À cette étape, et avec un test de Fisher, nous ne savons pas dire qu'elle est la ou les variable(s) qu'il faut conserver dans le modèle.

Question 3. Calculons le coefficient de détermination R^2 du modèle :

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{1\,504,4}{1\,680,8} = 0,895.$$

Question 4. Donnons une estimation de la variance de la variable résiduelle ε :

$$s^2 = \frac{||y - \hat{y}||^2}{n-p} = \frac{SC_{res}}{n-p} = \frac{176,4}{9} = 19,6.$$

Exercice II.4 Cet exercice a été traité avec le logiciel `R`, car il demande beaucoup de calculs qui ne sont pas réalisables avec une simple calculette. Le corrigé est mis à part. Il faut alors consulter le document intitulé « Exo4-TD8-Estimation.R ».

Exercice II.5 Pour répondre aux deux questions qui vont suivre, il faudrait s'assurer que les trois hypothèses du modèle sont vérifiées. Malheureusement nous ne pourrons pas le faire ici puisque nous ne connaissons pas les valeurs des observations. Donc nous allons supposer que les trois hypothèses sont vérifiées mais dans la pratique il faudrait les vérifier ABSOLUMENT.

Question 1. Est-ce que la régression est significative dans son ensemble ? Utiliser $\alpha = 0,05$.

Pour cela, nous réalisons un test de Fisher. Nous testons l'hypothèse nulle

$$\mathcal{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \exists j = 1, 2, 3 \text{ ou } 4, \quad \beta_j \neq 0.$$

Nous calculons la statistique du test de Fisher :

$$F_{obs} = \frac{21\,392,50}{95,07} = 225,03.$$

Nous lisons dans une table des quantiles de la loi de Fisher, à 95%, avec $\nu_1 = 4$ et $\nu_2 = 15$

$$F_{c,4,15} = 3,055568.$$

Comme $F_{obs} > F_{c,4,15}$, nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et d'accepter l'hypothèse alternative $\mathcal{H}_1$. Donc la régression est très significative dans son ensemble.

Question 2. Est-ce que l'affirmation de votre collègue est vraisemblable au seuil de signification $\alpha = 0,05$? Effectuer le test approprié.
Pour cela, nous effectuons un test de Fisher. Nous testons l'hypothèse nulle

$$\mathcal{H}_0 : \beta_3 = \beta_4 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \exists j = 3 \text{ ou } 4, \quad \beta_j \neq 0.$$

Nous calculons la statistique du test de Fisher partiel :

$$\begin{aligned} F_{obs} &= \frac{(SC(H_1)_{reg} - SC(H_0)_{reg})/(p-1-k)}{(SC(H_1)_{res})/(n-p)} \\ &= \frac{85\,570 - 62\,983}{1\,426} \times \frac{20-5}{5-1-2} = 118,79. \end{aligned}$$

Nous lisons dans une table des quantiles de la loi de Fisher, à 95%, avec $\nu_1 = 2$ et $\nu_2 = 15$

$$F_{c,2,15} = 3,68232.$$

Comme $F_{obs} > F_{c,2,15}$, nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et d'accepter l'hypothèse alternative $\mathcal{H}_1$. L'affirmation de notre collègue n'est donc pas vraisemblable au seuil de signification $\alpha = 5\%$.

Exercice II.6 Question 1. Dans quelle proportion, notée P, la variation non expliquée par X_1 est réduite avec l'ajout de X_2 dans l'équation de régression ?

Il faut d'abord calculer la proportion de la variation non expliquée par X_1. Elle est égale à :

$$(1 - 0,548) \times 100 = 45,2\%.$$

Ensuite il faut calculer la proportion de la variation non expliquée par X_1 et par X_2. Elle est égale à :

$$(1 - 0,683) \times 100 = 31,7\%.$$

Ensuite nous résolvons une équation à une inconnue :

$$45,2 - (45,2 \times P) = 31,7\%.$$

En résolvant cette équation, on obtient :

$$P = 29,86\%.$$

Donc la proportion P cherchée est égale à $29,86\%$.

Question 2. Déterminer la somme des carrés résiduelle lorsque les variables explicatives X_1 et X_2 sont dans l'équation de régression.

La somme de carrés résiduelle lorsque les variables explicatives X_1 et X_2 sont dans l'équation de régression est égale à :

$$SC_{res} = \mathrm{s}^2 \times (n - p) = 1,6352^2 \times (20 - 3) = 45,45.$$

Question 3. Quelle est la somme de carrés de régression attribuable à X_3 lorsqu'on ajoute cette variable à la suite de X_1 et X_2 ?

Pour répondre à cette question, introduisons quelques notations.

Accroissement de la variation expliquée par l'ajout de la variable explicative X_3 à la suite de la variable explicative X_1 et de la variable explicative X_2 :

$$SC_{reg}(X_1, X_2, X_3) - SC_{reg}(X_1, X_2) = SC_{reg}(X_3|X_1, X_2),$$

soit dans une proportion de

$$\frac{SC_{reg}(X_1, X_2, X_3) - SC_{reg}(X_1, X_2)}{SC_{res}(X_1, X_2)} = \frac{SC_{reg}(X_3|X_1, X_2)}{SC_{res}(X_1, X_2)} = r^2_{Y_{3.1,2}}$$

qui peut également s'écrire, si on divise chaque membre par SC_{tot}

$$\frac{\frac{SC_{reg}(X_1, X_2, X_3)}{SC_{tot}} - \frac{SC_{reg}(X_1, X_2)}{SC_{tot}}}{\frac{SC_{res}(X_1, X_2)}{SC_{tot}}} = \frac{R^2_{Y.1,2,3} - R^2_{Y.1,2}}{1 - R^2_{Y.1,2}} = r^2_{Y_{3.1,2}}.$$

Cette formule donne le coefficient de détermination partielle entre la variable expliquée Y et la variable explicative X_3, étant donné que les variables explicatives X_1 et X_2 sont déjà dans l'équation de régression.

On peut maintenant calculer la somme de carrés de régression attribuable à la variable explicative X_3 lorsqu'on ajoute cette variable à la suite des variables explicatives X_1 et X_2. On a :

$$\begin{aligned} SC_{reg}(X_3|X_1,X_2) &= r^2_{Y_{3.1,2}} \times SC_{res}(X_1,X_2) \\ &= \frac{R^2_{Y.1,2,3} - R^2_{Y.1,2}}{1 - R^2_{Y.1,2}} \times SC_{res}(X_1,X_2) \\ &= \frac{0,940 - 0,683}{1 - 0,683} \times (1,6352)^2(20-3) \\ &= 36,8523. \end{aligned}$$

Question 4. Quelle est la somme des carrés de régression attribuable à X_4 lorsqu'on ajoute cette variable à la suite de X_1, X_2 et X_3 ?

On procède de la même manière que précedemment. On a :

$$\begin{aligned} SC_{reg}(X_4|X_1,X_2,X_3) &= r^2_{Y_{4.1,2,3}} \times SC_{res}(X_1,X_2,X_3) \\ &= \frac{R^2_{Y.1,2,3,4} - R^2_{Y.1,2,3}}{1 - R^2_{Y.1,2,3}} \times SC_{res}(X_1,X_2,X_3) \\ &= \frac{0,959 - 0,940}{1 - 0,940} \times (0,7349)^2(20-4) \\ &= 2,7364. \end{aligned}$$