Exercice II.1 Un bureau de conseil en ressources humaines a effectué une étude sur le niveau d'anxiété Y mesuré sur une échelle de 1 à 50 de cadres d'entreprises au cours d'une période de deux semaines. Nous voulons examiner si les facteurs suivants peuvent influencer sur le niveau d'anxiété des cadres :

- X_1 : pression artérielle systolique
- X_2 : test évaluant les capacités managériales
- X_3 : niveau de satisfaction du poste occupé.

Le tableau d'analyse de la variance indique l'apport de chaque variable introduite dans l'ordre indiqué et ceci pour 22 cadres.

Source de variation	Somme des carrés	*ddl*
Régression due à X_1	$981,326$	1
Régression due à X_2	$190,232$	1
Régression due à X_3	$129,431$	1
Résiduelle	$442,292$	18
Totale	$1743,281$	21

1. Quelle est la somme des carrés due à la régression pour l'ensemble des trois variables explicatives ?
2. Quelle proportion de la variation dans le niveau d'anxiété est expliquée par les trois variables explicatives ?
3. Pouvons-nous conclure que dans l'ensemble les trois variables explicatives ont un effet significatif sur le niveau d'anxiété ? Utiliser un seuil de signification $\alpha = 5\%$. Préciser les hypothèses que nous voulons tester.
4. Si nous ne tenons compte que de la variable explicative X_1, quel serait alors le tableau d'analyse de la variance correspondant ?

Source de variation	Somme des carrés	*ddl*
Régression due à X_1	$981,326$	
Résiduelle		
Totale		

5. Tester les hypothèses nulles suivantes, au seuil de signification $\alpha = 5\%$, en utilisant un rapport F approprié :

   a) $\mathcal{H}_0 : \beta_1 = 0$ dans le modèle $Y = \beta_0 + \beta_1 X_1 + \varepsilon$;
   b) $\mathcal{H}_0 : \beta_2 = 0$ dans le modèle $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$;
   c) $\mathcal{H}_0 : \beta_3 = 0$ dans le modèle $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.

6. Quelle est la valeur du coefficient de détermination R^2 associée à l'estimation de chaque modèle spécifié à la question 5. ?
7. Lequel des trois modèles semble le mieux approprié pour expliquer les fluctuations du niveau d'anxiété des cadres d'entreprises ?

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 :

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

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

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 1.9 (Intervalles de confiance vs Région de confiance)
On considère le modèle de régression linéaire simple $y = \beta_1 + \beta_2 x + \epsilon$. Soit un échantillon $(x_i, y_i)_{1 \leq i \leq 100}$ de statistiques résumées

$$\sum_{i=1}^{100} x_i = 0 \qquad \sum_{i=1}^{100} x_i^2 = 400 \qquad \sum_{i=1}^{100} x_i y_i = 100 \qquad \sum_{i=1}^{100} y_i = 100 \qquad \hat{\sigma}^2 = 1.$$

1. Exprimer les intervalles de confiance à 95% pour β_1 et β_2.
2. Donner l'équation de la région de confiance à 95% de (β_1, β_2). (Rappel : l'ensemble des points (x, y) tels que $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} \leq 1$ est l'intérieur d'une ellipse centrée en (x_0, y_0), dont les axes sont parallèles à ceux des abscisses et des ordonnées, et de sommets $(x_0 \pm a, 0)$ et $(0, y_0 \pm b)$.)
3. Représenter sur un même graphique les résultats obtenus.

Exercice 1.9 (Intervalles de confiance vs Région de confiance)

1. Il sort des statistiques résumées que $\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x} = 1$ et $\hat{\beta}_2 = (\sum x_i y_i)/(\sum x_i^2) = 1/4$. La droite des moindres carrés a donc pour équation $y = 1 + x/4$. Les estimateurs des variances se calculent facilement

$$\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2}{\sum (x_i - \overline{x})^2} \times \frac{\sum x_i^2}{n} = \frac{\hat{\sigma}^2}{n} = \frac{1}{100} \;\Rightarrow\; \hat{\sigma}_1 = \frac{1}{10}$$

tandis que

$$\hat{\sigma}_2^2 = \frac{\hat{\sigma}^2}{\sum (x_i - \overline{x})^2} = \frac{1}{400} \;\Rightarrow\; \hat{\sigma}_2 = \frac{1}{20}.$$

Le quantile d'ordre 0.975 d'une Student à 98 degrés de liberté est à peu près le même que celui d'une Student à 100 degrés de liberté, c'est-à-dire environ 1.984 que l'on va arrondir à 2. L'intervalle de confiance à 95% pour β_1 est donc

$$IC(\beta_1) = [\hat{\beta}_1 - 2\hat{\sigma}_1, \hat{\beta}_1 + 2\hat{\sigma}_1] = [0.8; 1.2]$$

et pour β_2

$$IC(\beta_2) = [\hat{\beta}_2 - 2\hat{\sigma}_2, \hat{\beta}_2 + 2\hat{\sigma}_2] = [0.15; 1.35]$$

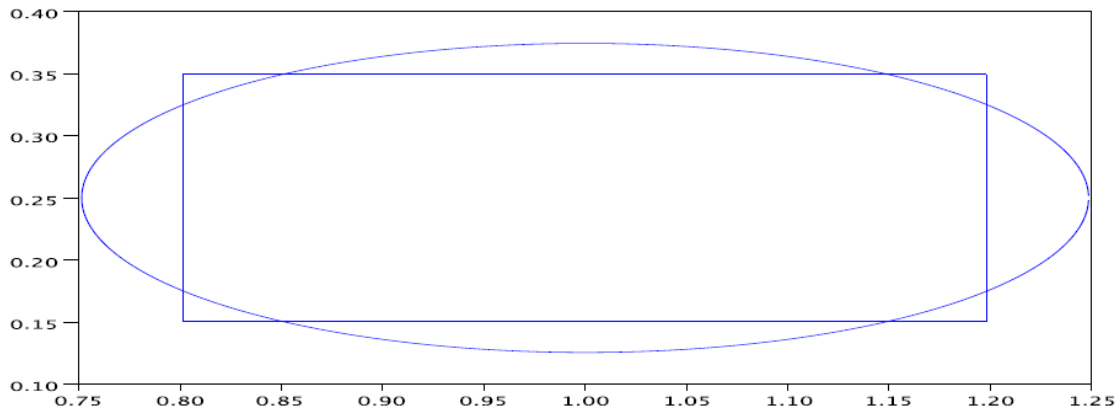

FIGURE 1.12 – Intervalles de confiance vs Région de confiance.

2. Avec les notations du cours, la région de confiance simultanée à 95% est l'ensemble des points (β_1, β_2) tels que

$$\frac{1}{2\hat{\sigma}^2}\left(n(\beta_1-\hat{\beta}_1)^2+2n\bar{x}(\beta_1-\hat{\beta}_1)(\beta_2-\hat{\beta}_2)+\sum x_i^2(\beta_2-\hat{\beta}_2)^2\right)\leq f^2_{n-2}(0.95).$$

Le quantile d'ordre 0.95 d'une loi de Fisher à (2,100) degrés de liberté étant égal à 3.09, nous arrondirons à nouveau et prendrons $f^2_{98}(0.95)\approx 3$, de sorte que nous obtenons comme région de confiance l'ensemble des points (β_1, β_2) tels que

$$\frac{1}{2}\left(100(\beta_1-1)^2+400(\beta_2-1/4)^2\right)\leq 3 \;\Leftrightarrow\; \frac{(\beta_1-1)^2}{\left(\frac{\sqrt{6}}{10}\right)^2}+\frac{(\beta_2-1/4)^2}{\left(\frac{\sqrt{6}}{20}\right)^2}\leq 1.$$

La région de confiance est donc l'intérieur d'une ellipse de centre $(\hat{\beta}_1, \hat{\beta}_2) = (1, 1/4)$ et de sommets $(1\pm\sqrt{6}/10, 0)$ et $(0, 1/4\pm\sqrt{6}/20)$, c'est-à-dire $(1.24, 0)$, $(0, 0.37)$, $(0.76, 0)$, $(0, 0.13)$.

Exercice I.3. On utilise le modèle de régression linéaire multiple :

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

1. Compléter le tableau d'analyse de variance suivant :

Source de variation	Somme des carrés	*ddl*	Carrés moyens	F_{obs}
Régression	1 504, 4			
Résiduelle			19, 6	
Totale	1 680, 8			

2. Tester l'hypothèse nulle

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

contre l'hypothèse alternative

$$\mathcal{H}_1 : \text{ au moins un des } \beta \neq 0.$$

3. Quel est le coefficient de détermination R^2 du modèle ?
4. Donner une estimation de la variance de ε.

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**	**752,2**	**38,37**
Résiduelle	**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.

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 2.4 (Deux variables explicatives)
On examine l'évolution d'une variable réponse y_i en fonction de deux variables explicatives x_i et z_i. Soit $X = (\mathbb{1}\ x\ z)$ la matrice $n \times 3$ du plan d'expérience.

1. Nous avons obtenu les résultats suivants :

$$X'X = \begin{pmatrix} 25 & 0 & 0 \\ ? & 9.3 & 5.4 \\ ? & ? & 12.7 \end{pmatrix} \qquad (X'X)^{-1} = \begin{pmatrix} 0.04 & 0 & 0 \\ 0 & 0.1428 & -0.0607 \\ 0 & -0.0607 & 0.1046 \end{pmatrix}.$$

   (a) Donner les valeurs manquantes.

   (b) Que vaut n ?

   (c) Calculer le coefficient de corrélation linéaire empirique entre x et z.

2. La régression linéaire de Y sur $(\mathbb{1}, x, z)$ donne

$$Y = -1.61\mathbb{1} + 0.61x + 0.46z + \hat{\varepsilon}, \qquad SCR = \|\hat{\varepsilon}\|^2 = 0.3.$$

   (a) Déterminez la moyenne empirique $\bar{y}$.

   (b) Calculer la somme des carrés expliquée (SCE), la somme des carrés totale (SCT), le coefficient de détermination et le coefficient de détermination ajusté.

Exercice 2.4 (Deux variables explicatives)
On examine l'évolution d'une variable y en fonction de deux variables exogènes x et z. On dispose de n observations de ces variables. On note $X = [\mathbb{1}\ \boldsymbol{x}\ \boldsymbol{z}]$ où $\mathbb{1}$ est le vecteur constant et $\boldsymbol{x}$, $\boldsymbol{z}$ sont les vecteurs des variables explicatives.

1. Nous avons obtenu les résultats suivants :

$$X'X = \begin{bmatrix} 25 & 0 & 0 \\ ? & 9.3 & 5.4 \\ ? & ? & 12.7 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} 0.04 & 0 & 0 \\ 0 & 0.1428 & -0.0607 \\ 0 & -0.0607 & 0.1046 \end{bmatrix}.$$

   (a) Les 3 valeurs manquantes se déduisent de la symétrie de la matrice $X'X$.

   (b) Puisque $X = [\mathbb{1}\ \boldsymbol{x}\ \boldsymbol{z}]$, il vient $n = (X'X)_{1,1} = 25$.

(c) Le coefficient de corrélation linéaire empirique entre x et z se déduit lui aussi de la matrice $X'X$. On remarque tout d'abord que les moyennes empiriques sont nulles puisque

$$\bar{x} = \frac{(X'X)_{1,2}}{n} = 0 = \frac{(X'X)_{1,3}}{n} = \bar{z}$$

Par conséquent

$$r_{x,z} = \frac{\sum(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{(x_i - \bar{x})^2}\sqrt{(z_i - \bar{z})^2}} = \frac{\sum x_i z_i}{\sqrt{\sum x_i^2}\sqrt{\sum z_i^2}} = \frac{(X'X)_{2,3}}{\sqrt{(X'X)_{2,2}}\sqrt{(X'X)_{3,3}}}$$

ce qui donne

$$r_{x,z} = \frac{5.4}{\sqrt{9.3}\sqrt{12.7}} \approx 0.5$$

2. La régression linéaire de Y sur $(\mathbb{1}, \boldsymbol{x}, \boldsymbol{z})$ donne

$$Y = -1.6\mathbb{1} + 0.61\boldsymbol{x} + 0.46\boldsymbol{z} + \hat{\boldsymbol{\varepsilon}}, \qquad SCR = \|\hat{\boldsymbol{\varepsilon}}\|^2 = 0.3.$$

(a) Puisque la constante fait partie du modèle, la moyenne empirique des résidus est nulle : $\bar{\hat{\varepsilon}} = 0$. On en déduit que

$$\bar{y} = -1.6 + 0.61\bar{x} + 0.46\bar{z} + \bar{\hat{\varepsilon}} = -1.6$$

(b) Puisque la constante fait partie du modèle, la somme des carrés expliquée par le modèle est

$$SCE = \|\hat{Y} - \bar{y}\mathbb{1}\|^2 = \sum(\hat{y}_i - \bar{y})^2 = \sum(0.61x_i + 0.46z_i)^2$$

c'est-à-dire

$$SCE = \|\hat{Y} - \bar{y}\mathbb{1}\|^2 = 0.61^2 \sum x_i^2 + 2 \times 0.61 \times 0.46 \sum x_i z_i + 0.46^2 \sum z_i^2$$

ce qui se calcule à nouveau grâce à la matrice $X'X$:

$$SCE = \|\hat{Y} - \bar{y}\mathbb{1}\|^2 = 0.61^2(X'X)_{2,2} + 2 \times 0.61 \times 0.46(X'X)_{2,3} + 0.46^2(X'X)_{3,3} = 9.18$$

La somme des carrés totale est alors immédiate, en vertu de la sacro-sainte formule de décomposition de la variance :

$$SCT = SCE + SCR = 9.18 + 0.3 = 9.48$$

Le coefficient de détermination vaut donc

$$R^2 = \frac{SCE}{SCT} \approx 0.968$$

Autrement dit, 97% de la variance des données est expliquée par ce modèle de régression. Le coefficient de détermination ajusté est à peine différent :

$$R_a^2 = 1 - \frac{n-1}{n-p}(1 - R^2) \approx 0.965$$

et on vérifie bien la relation générale selon laquelle $R_a^2 < R^2$.

Exercice I.5. Dans une étude de régression linéaire multiple comportant quatre variables explicatives X_1, X_2, X_3, X_4, on a obtenu le tableau d'analyse de la variance suivant, et ceci pour 20 observations.

Source de variation	Somme des carrés	*ddl*	Carrés moyens
Régression (X_1, X_2, X_3, X_4)	85 570	4	21 392,50
Résiduelle	1 426	15	95,07
Totale	86 996	19	

1. Est-ce que la régression est significative dans son ensemble ? Utiliser $\alpha = 5\%$.
2. Une de vos collègues mentionne que les variables X_3 et X_4 sont inutiles dans le modèle de régression linéaire. Une autre analyse de régression linéaire ne

comportant cette fois que les variables explicatives X_1 et X_2 conduit au tableau d'analyse de la variance suivant.

Source de variation	Somme des carrés	*ddl*	Carrés moyens
Régression (X_1, X_2)	62 983	2	31 491,50
Résiduelle	24 013	17	1 412,53
Totale	86 996	19	

Est-ce que l'affirmation de votre collègue est vraisemblable au seuil de signification $\alpha = 5\%$? Effectuer le test approprié.

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 I.6. Dans une étude de régression multiple comportant quatre variables explicatives et 20 observations, on a introduit, dans l'ordre, les variables X_1, X_2, X_3, X_4.

Variables explicatives dans l'équation	Variable additionnelle	Écart-type des résidus	Proportion de la variation expliquée
Aucune		$s_Y = 2,7456$	
X_1	X_1	1,8968	0,5480
X_1, X_2	X_2	1,6352	0,6830
X_1, X_2, X_3	X_3	0,7349	0,9400
X_1, X_2, X_3, X_4	X_4	0,6281	0,9590

1. Dans quelle proportion la variation non expliquée par X_1 est réduite avec l'ajout de X_2 dans l'équation de régression ?
2. Quelle est la corrélation partielle entre Y et X_3 après s'être débarrassé de l'influence des variables explicatives X_1 et X_2 sur Y et X_3 ? On considère que le coefficient de régression $\widehat{\beta}_3$ est négatif.
3. Déterminer la somme de carrés résiduelle lorsque les variables explicatives X_1 et X_2 sont dans l'équation de régression.
4. 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 ?
5. Quelle est la somme de carrés de régression attribuable à X_4 lorsqu'on ajoute cette variable à la suite de X_1, X_2 et X_3 ?

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 :

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