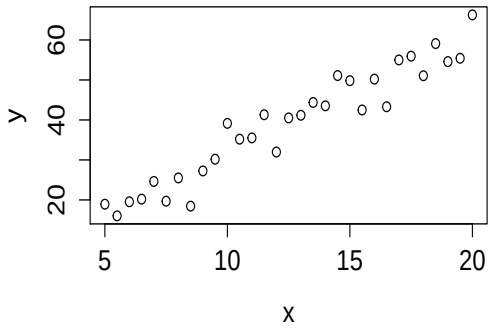
TD Modèle linéaire - Feuille 1

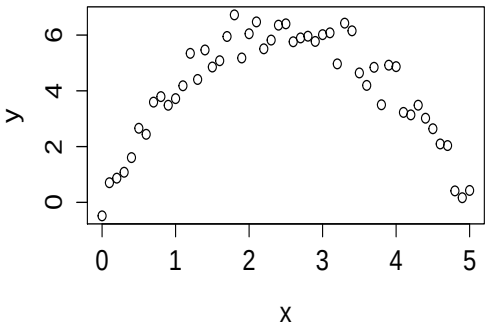
STID2 - IUT Vannes - 2021/2022

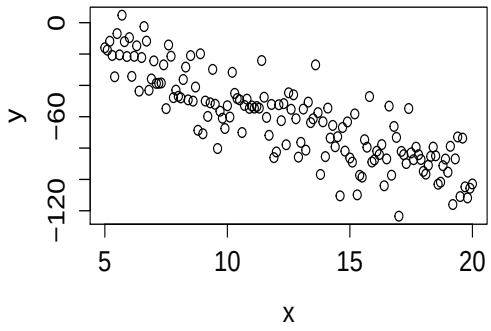
A. Cuzol / A. Poterie

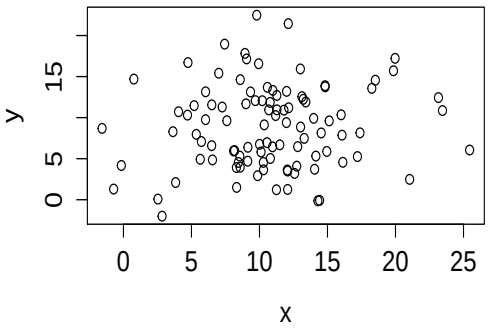
Exercice 1

On observe des nuages de points $(x_i, y_i)_{i=1,\dots,n}$. Pour chaque situation, peut-on supposer l'existence d'une liaison linéaire entre les variables x et y ? Que peut-on dire du coefficient de corrélation dans chaque cas ?









Exercice 2

Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone O3 dans l'air. On cherche en particulier à savoir si on peut expliquer le taux maximal d'ozone de la journée par la température T12 à 12h. Les données sont :

O3	115,4	76,8	113,8	81,6	115,4	125	83,6	75,2	136,8	102,8
T12	23,8	16,3	27,2	7,1	25,1	27,5	19,4	19,8	32,2	20,7

1. Représenter les données.
2. Peut-on supposer l'existence d'une relation linéaire entre la concentration en ozone et la température ?
3. Calculer le coefficient de corrélation empirique ρ entre les variables O3 et T12.
4. La valeur de ce coefficient confirme-t-elle la réponse à la question 2 ?
5. On décide de poser un modèle de régression linéaire simple pour expliquer le lien entre O3 et T12.
   (a) Rappeler l'écriture de ce modèle et rédéfinir tous ses éléments.
   (b) Préciser les hypothèses du modèle.
   (c) Quels sont les paramètres inconnus dans ce modèle ?
   (d) Donner l'écriture vectorielle du modèle. Que contient exactement chaque vecteur ?

Exercice 3

On souhaite maintenant estimer les paramètres du modèle proposé dans l'**Exercice 2** par la méthode des moindres carrés. On note $\hat{\beta}_0$et $\hat{\beta}_1$ les estimateurs de l'ordonnée à l'origine et de la pente.

1. Quelle quantité doit-on minimiser pour obtenir $\hat{\beta}_0$ et $\hat{\beta}_1$?
2. Comment interpréter cette quantité graphiquement ?
3. Rappeler les expressions théoriques des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$.
4. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$ à partir des données de l'**Exercice 2**. Rappeler le lien entre $\hat{\beta}_1$ et ρ.
5. Donner l'équation de la droite des moindres carrés estimée. Superposer cette droite au graphique de l'**Exercice 2**.

Exercice 4

Dans l'**Exercice 3**, on a estimé β_0 et β_1 par $\hat{\beta}_0$ et $\hat{\beta}_1$. Il reste un paramètre à estimer : σ^2 (la variance des erreurs). On estime ce paramètre à partir des résidus.

1. Rappeler ce que sont les résidus du modèle. Faire une illustration graphique.
2. Compléter le tableau suivant :

O3	115,4	76,8	113,8	81,6	115,4	125	83,6	75,2	136,8	102,8
T12	23,8	16,3	27,2	7,1	25,1	27,5	19,4	19,8	32,2	20,7
$\hat{O3}$										
Résidus										

3. Prouver que la somme des résidus est nulle en théorie. En pratique, est-ce bien le cas ici ?
4. Rappeler l'expression de l'estimateur $\hat{\sigma}^2$ de la variance σ^2 des erreurs. Calculer $\hat{\sigma}^2$.

Exercice 5

On souhaite calculer la variance des estimateurs des moindres carrés.

1. Rappeler les expressions de $\hat{V}(\hat{\beta}_0)$ et $\hat{V}(\hat{\beta}_1)$.
2. Calculer ces variances à partir des données de l'**Exercice 2**. On donne les quantités intermédiaires suivantes :

$$\sum_{i=1}^{10} \mathrm{T12}_i^2 = 5242$$

$$\sum_{i=1}^{10} (\mathrm{T12}_i - \overline{\mathrm{T12}})^2 = 442$$

Exercice 6

Nous souhaitons exprimer la hauteur Y d'un arbre (en mètres) en fonction de son diamètre X (en centimètres) à 1m30 du sol. Pour cela, nous avons mesuré 20 couples diamètre-hauteur et les résultats ci-dessous sont disponibles :

$$\bar{x} = 34.9; \quad \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2 = 28.29; \quad \bar{y} = 18.34$$

$$\frac{1}{20}\sum_{i=1}^{20}(y_i - \bar{y})^2 = 2.85; \quad \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y}) = 6.26$$

1. On note $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ l'estimation de la droite de régression. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$.
2. On mesure le diamètre d'un nouvel arbre à 1m30 du sol : 50cm. Quelle valeur pouvez-vous prédire pour la hauteur de cet arbre ?

Exercice 7

Nous observons les poids de 12 pères et de leurs fils aînés respectifs :

Père	65	63	67	64	68	62	70	66	68	67	69	71
Fils	68	66	68	65	69	66	68	65	71	67	68	70

Nous disposons des résultats numériques suivants :

$$\sum_{i=1}^{12} p_i = 800; \quad \sum_{i=1}^{12} p_i^2 = 53418; \quad \sum_{i=1}^{12} p_i f_i = 54107; \quad \sum_{i=1}^{12} f_i = 811; \quad \sum_{i=1}^{12} f_i^2 = 54849.$$

1. On note $\hat{f} = \hat{\beta}_0 + \hat{\beta}_1 p$ l'estimation de la droite de régression du poids des fils en fonction du poids des pères. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$.
2. Avec le logiciel R, on a ajusté le modèle de régression linéaire :

```
p=c(65, 63, 67, 64, 68, 62, 70, 66, 68, 67, 69, 71)
f=c(68, 66, 68, 65, 69, 66, 68, 65, 71, 67, 68, 70 )

reg=lm(f~p)
summary(reg)
```

et on obtient la sortie suivante :

```
lm(formula = f ~ p)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2658 -0.8494  0.2106  0.6752  2.7815

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.8248    10.1780   3.520  0.00554 **
p             0.4764     0.1525   3.123  0.01082 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.404 on 10 degrees of freedom
Multiple R-squared:  0.4937,Adjusted R-squared:  0.4431
F-statistic: 9.752 on 1 and 10 DF,  p-value: 0.01082
```

Retrouver dans cette sortie R les valeurs estimées pour :

- l'ordonnée à l'origine $\hat{\beta}_0$;
- la pente $\hat{\beta}_1$;
- la variance $\hat{\sigma}^2$ des erreurs ;
- la variance $\hat{V}(\hat{\beta}_0)$ de l'estimateur de l'ordonnée à l'origine ;
- la variance $\hat{V}(\hat{\beta}_1)$ de l'estimateur de la pente ;
- le coefficient de détermination R^2.

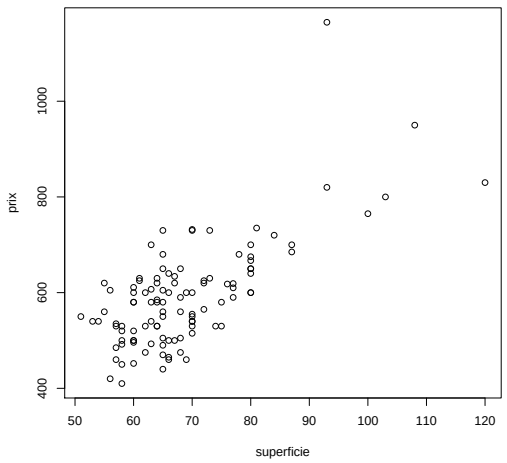
TD Modèle linéaire - Feuille 2

STID2 - IUT Vannes - 2021/2022

A. Cuzol / A. Poterie

Exercice 1

On a relevé en juin 2005 dans les petites annonces les superficies (variable "superficie", en m^2) et les prix (variable "prix", en euros) de 108 appartements de type T3 à louer dans l'agglomération de Rennes. Les données sont représentées sur la figure ci-dessous :



On donne les écarts-types estimés des variables "superficie" et "prix" : $s_{superficie} = 11.55$ et $s_{prix} = 109.20$.

Une régression linéaire simple expliquant le prix en fonction de la superficie a été effectuée avec le logiciel R, et a donné le résultat suivant :

```
Residuals:
    Min      1Q  Median      3Q     Max
-133.68  -54.58  -13.62   47.88  411.55

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 134.3450    45.4737   2.954  0.00386 **
superficie     ?         0.6525  10.203  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 77.93 on 106 degrees of freedom
Multiple R-squared: 0.4955,     Adjusted R-squared: 0.4907
F-statistic: 104.1 on 1 and 106 DF,  p-value: < 2e-16
```

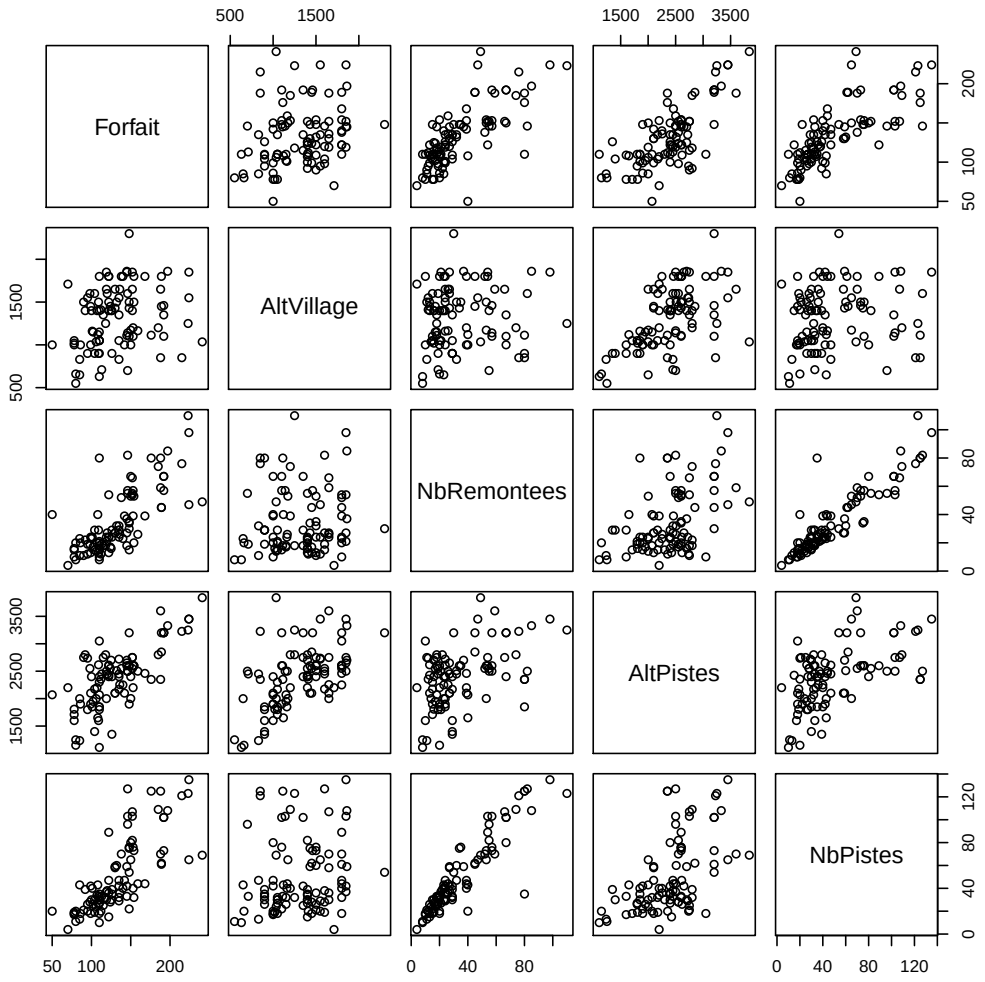
1. Écrire le modèle de régression linéaire qui a été posé pour étudier la relation entre le prix des appartements et leur superficie, et préciser ses hypothèses.
2. Donner une estimation du coefficient de corrélation entre le prix et la superficie d'un appartement de type T3.
3. A quoi correspond la quantité " ?" dans la sortie ci-dessus ? Donner sa valeur en justifiant précisément le calcul.
4. Donner un intervalle de confiance de niveau $1-\alpha$ (avec $\alpha = 5\%$) pour le paramètre de pente β_1. On donne le quantile de niveau 97.5% pour la loi de Student à 106 degrés de liberté : $t_{106}(0.975) \approx 1.98$.
5. Donner une estimation de la variance du terme d'erreur du modèle de régression.
6. Que vaut la somme des résidus de la régression ? Que vaut la somme des carrés des résidus de la régression ?
7. On réalise un test de Student pour savoir si la superficie joue un rôle sur le prix des appartements de type T3.
   (a) Quels sont les hypothèses de test H_0 et H_1 ?
   (b) Quelle est la statistique de test ?
   (c) Quelle est sa loi sous l'hypothèse H_0 ?
   (d) Quelle est la conclusion du test au niveau 5% ? Peut-on faire le lien avec la question 4 ?

Exercice 2 *(extrait d'examen)*

On a observé, pour 97 stations de ski françaises :
- le prix du forfait semaine ("Forfait") ;
- l'altitude de la station village ("AltVillage") ;
- le nombre de remontées ("NbRemontees") ;
- l'altitude du sommet des pistes ("AltPistes") ;
- le nombre de pistes ("NbPistes").

La matrice des corrélations et la matrice des nuages de points entre toutes les variables sont fournies ci-après.

	Forfait	AltVillage	NbRemontees	AltPistes	NbPistes
Forfait	1.00	0.25	0.74	0.72	0.77
AltVillage	0.25	1.00	0.07	0.56	0.15
NbRemontees	0.74	0.07	1.00	0.48	0.92
AltPistes	0.72	0.56	0.48	1.00	0.55
NbPistes	0.77	0.15	0.92	0.55	1.00



On décide d'ajuster un modèle de régression simple pour expliquer le prix du forfait semaine par le nombre de pistes dans la station. La sortie donnée par le logiciel R est la suivante :

```
lm(formula = Forfait ~ NbPistes, data = Ski)

Residuals:
    Min      1Q  Median      3Q     Max
-56.406 -13.992  -1.785  12.165  91.208

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 87.19944    4.60760   18.93   <2e-16 ***
NbPistes     0.90714    0.07787   11.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.54 on 95 degrees of freedom
Multiple R-squared: ? ,     Adjusted R-squared: ?
F-statistic: 135.7 on 1 and 95 DF,  p-value: < 2.2e-16
```

1. Quelles hypothèses doit-on vérifier sur ce modèle de régression avant de construire des intervalles de confiance et des tests sur les paramètres ?
2. Donner un intervalle de confiance à 95% pour la pente (on donne le quantile de niveau 97.5% pour la loi de Student à 95 degrés de liberté : $t_{95}(0.975) \approx 1.98$). Que peut-on en conclure ?
3. Le nombre de pistes dans la station a-t-il un effet significatif sur le prix du forfait ? Répondre par un test statistique.
4. Que vaut le coefficient R^2 de ce modèle ?
5. Ecrire l'équation du modèle ajusté.
6. Donner une prédiction du prix du forfait semaine pour une station ayant 70 pistes.

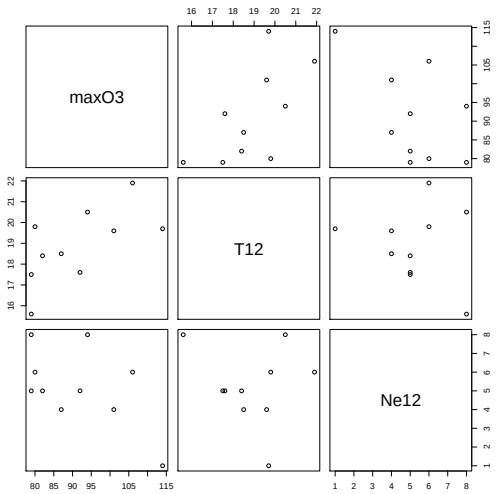
Exercice 3

On s'intéresse de nouveau à la concentration d'ozone O3 dans l'air. On avait montré que la concentration en ozone peut être expliquée par la variable de température à 12h (T12). On souhaite maintenant ajuster un modèle de régression multiple pour expliquer l'ozone par deux variables explicatives : la température à 12h (T12) et la nébulosité à 12h (Ne12).

Les données sont les suivantes (extraites du jeu de données complet) :

O3	87	82	92	114	94	80	79	79	101	106
T12	18,5	18,4	17,6	19,7	20,5	19,8	15,6	17,5	19,6	21,9
Ne12	4	5	5	1	8	6	8	5	4	6

On donne la matrice des nuages de points ainsi que la matrice de corrélation entre toutes les variables :



	maxO3	T12	Ne12
maxO3	1.0000000	0.6409338	-0.5369661
T12	0.6409338	1.0000000	-0.1443160
Ne12	-0.5369661	-0.1443160	1.0000000

1. D'après les deux matrices ci-dessus, semble-t-il pertinent d'utiliser les variables T12 et Ne12 pour chercher à expliquer la concentration en ozone ?
2. Donner l'écriture du modèle de régression multiple expliquant O3 par T12 et Ne12.
3. Rappeler les hypothèses sur ce modèle.

4. L'hypothèse de non colinéarité des variables explicatives semble-t-elle respectée ?
5. On rappelle que l'écriture du modèle sous forme matricielle est la suivante :

$$O3 = X\beta + e$$

   (a) Donner la dimension de $O3$, X, β et e

   (b) Donner explicitement le contenu de $O3$, X, β et e.

6. On donne les quantités suivantes :

$$(X'X)^{-1} = \begin{pmatrix} 14.557 & -0.702 & -0.226 \\ -0.702 & 0.036 & 0.005 \\ -0.226 & 0.005 & 0.027 \end{pmatrix}$$

$$(X'X)^{-1}X' = \begin{pmatrix} 0.660 & 0.503 & 1.065 & 0.496 & -1.651 & -0.706 & 1.790 & 1.135 & -0.113 & -2.181 \\ -0.020 & -0.019 & -0.048 & 0.009 & 0.070 & 0.036 & -0.106 & -0.052 & 0.019 & 0.111 \\ -0.034 & -0.008 & -0.011 & -0.111 & 0.083 & 0.026 & 0.061 & -0.012 & -0.029 & 0.035 \end{pmatrix}$$

   (a) Rappeler l'expression de l'estimateur des moindres carrés $\hat{\beta}$.

   (b) Calculer $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$.

   (c) Vérifier le résultat en le comparant à celui donné dans la sortie R de la régression (les valeurs peuvent être légèrement différentes en raison des arrondis dans les matrices ci-dessus) :

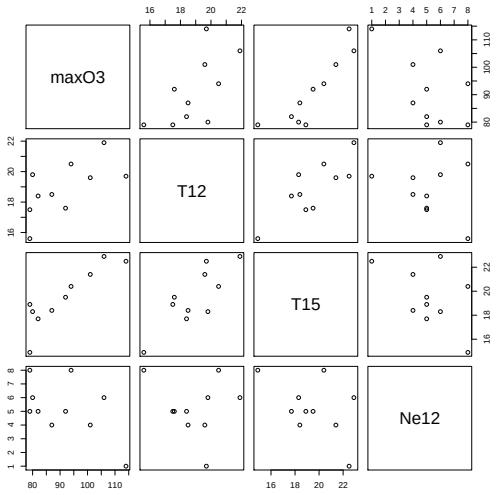
```
lm(formula = maxO3 ~ T12 + Ne12, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-12.761  -7.003   3.741   5.175   8.477

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   30.173     33.205   0.909   0.3937
T12            3.991      1.649   2.420   0.0461 *
Ne12          -2.738      1.434  -1.909   0.0979 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.703 on 7 degrees of freedom
Multiple R-squared: 0.6126,     Adjusted R-squared: 0.5019
F-statistic: 5.533 on 2 and 7 DF,  p-value: 0.0362
```

Exercice 4

On décide d'ajouter une nouvelle variable explicative au modèle précédent pour expliquer l'ozone : la température à 15h (T15). Le nouveau jeu de données est le suivant :

O3	87	82	92	114	94	80	79	79	101	106
T12	18,5	18,4	17,6	19,7	20,5	19,8	15,6	17,5	19,6	21,9
T15	18.4	17.7	19.5	22.5	20.4	18.3	14.9	18.9	21.4	22.9
Ne12	4	5	5	1	8	6	8	5	4	6

On donne la matrice des nuages de points ainsi que la matrice de corrélation entre toutes les variables :



	maxO3	T12	T15	Ne12
maxO3	1.0000000	0.6409338	0.8886707	-0.5369661
T12	0.6409338	1.0000000	0.8198490	-0.1443160
T15	0.8886707	0.8198490	1.0000000	-0.4884783
Ne12	-0.5369661	-0.1443160	-0.4884783	1.0000000

1. Que constate-t-on ?
2. Quelle solution proposez-vous pour résoudre le problème posé ?

TD Modèle linéaire - Feuille 3

STID2 - IUT Vannes - 2021/2022

A. Cuzol / A. Poterie

Exercice 1 *(Extrait de l'examen 2019-2020)*

Le jeu de données "Boston" contient des informations sur 506 quartiers de la ville de Boston et alentours. Nous disposons des variables suivantes :

- "medv" : valeur médiane des prix des maisons (en dizaines de milliers de dollars) ;
- "rm" : nombre moyen de pièces ;
- "age" : proportion de maisons construites avant 1940 ;
- "lstat" : proportion de ménages ayant un niveau de vie peu élevé ;
- "crim" : taux de criminalité ;
- "chas" : proximité de la rivière Charles (modalité "1" si le quartier est au bord de la rivière, modalité "0" sinon).

La matrice des corrélations entre toutes les variables quantitatives du jeu de données est la suivante :

	medv	rm	age	lstat	crim
medv	1.00	0.70	-0.38	-0.74	-0.39
rm	0.70	1.00	-0.24	-0.61	-0.22
age	-0.38	-0.24	1.00	0.60	0.35
lstat	-0.74	-0.61	0.60	1.00	0.46
crim	-0.39	-0.22	0.35	0.46	1.00

1. On lance l'ajustement du modèle de régression multiple expliquant le prix médian des maisons par toutes les variables quantitatives disponibles. Le résultat est le suivant :

```
lm(formula = medv ~ rm + age + lstat + crim)

Residuals:
    Min      1Q  Median      3Q     Max
-18.105  -3.501  -1.143   1.968  28.180

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.34910    3.17079  -0.741  0.45913
rm           5.11625    0.45083  11.349  < 2e-16 ***
age             ?       0.01116   1.129     ??
lstat       -0.61258    0.05642 -10.857  < 2e-16 ***
crim        -0.10639    0.03216  -3.308  0.00101 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.488 on 501 degrees of freedom
Multiple R-squared:  0.6468,Adjusted R-squared:  0.6439
F-statistic: 229.3 on 4 and 501 DF,  p-value: < 2.2e-16
```

(a) Que vaut le coefficient estimé $\hat{\beta}_{age}$?

(b) On donne le quantile de niveau 97.5% pour la loi de Student à 501 degrés de liberté : $t_{501}(0.975) \approx 1.96$. Donner l'intervalle de confiance à 95% pour le paramètre β_{age}.

(c) Quel sera le résultat du test de Student pour le coefficient β_{age} ?

2. On décide de retirer la variable "age" du modèle, on relance l'estimation et on obtient le résultat suivant :

```
lm(formula = medv ~ rm + lstat + crim)

Residuals:
    Min      1Q  Median      3Q     Max
-17.925  -3.566  -1.157   1.906  29.024

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.56225    3.16602  -0.809  0.41873
rm           5.21695    0.44203  11.802  < 2e-16 ***
lstat       -0.57849    0.04767 -12.135  < 2e-16 ***
crim        -0.10294    0.03202  -3.215  0.00139 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.49 on 502 degrees of freedom
Multiple R-squared:  0.6459,Adjusted R-squared:  0.6437
F-statistic: 305.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

(a) Pourquoi le R^2 de ce nouveau modèle est inférieur au R^2 du modèle précédent ?

(b) En se basant sur le critère du R^2 ajusté, ce nouveau modèle est-il meilleur que le précédent ?

(c) Que nous indique le résultat du test de Fisher global présenté sur la dernière ligne de la sortie ?

(d) Par rapport au contexte de l'exercice, comment interpréter le fait que le coefficient associé à la variable "rm" soit positif ? Inversement, comment interpréter le fait que les coefficients associés aux variables "lstat" et "crim" soient négatifs ?

Exercice 2

On cherche à expliquer le lien entre l'indice de masse graisseuse d'un individu (variable "Fat") et des variables anatomiques :

- la circonférence de la cuisse ("Thigh") ;
- la circonférence mesurée au niveau du triceps ("Triceps") ;
- la circonférence mesurée au niveau de l'avant-bras ("Midarm").

On peut construire plusieurs modèles différents à partir de ces 3 variables explicatives. Dans cet exercice on souhaite comparer deux modèles :

- Un modèle de régression multiple expliquant l'indice de masse graisseuse par les 3 variables anatomiques disponibles ;
- Un modèle de régression simple expliquant l'indice de masse graisseuse par la circonférence de la cuisse uniquement.

1. Pourrait-on utiliser le critère du R^2 pour choisir entre ces deux modèles ? Le critère du R^2 ajusté ?
2. On décide de mettre en place un test de Fisher emboîté. Ecrire explicitement les hypothèses de test H_0 et H_1 (on notera β_{Th}, β_{Tr} et β_{Mi} les coefficients associés aux variables Thigh, Triceps et Midarm).
3. Rappeler l'expression de la statistique de test du test de Fisher emboîté, et sa loi sous H_0.

La sortie du test est la suivante :

```
Analysis of Variance Table

Model 1: Fat ~ Thigh
Model 2: Fat ~ Thigh + Triceps + Midarm
  Res.Df     RSS  Df  Sum of Sq   F  Pr(>F)
1     18 113.424
2     16  98.405  2    15.019     ?  0.321
```

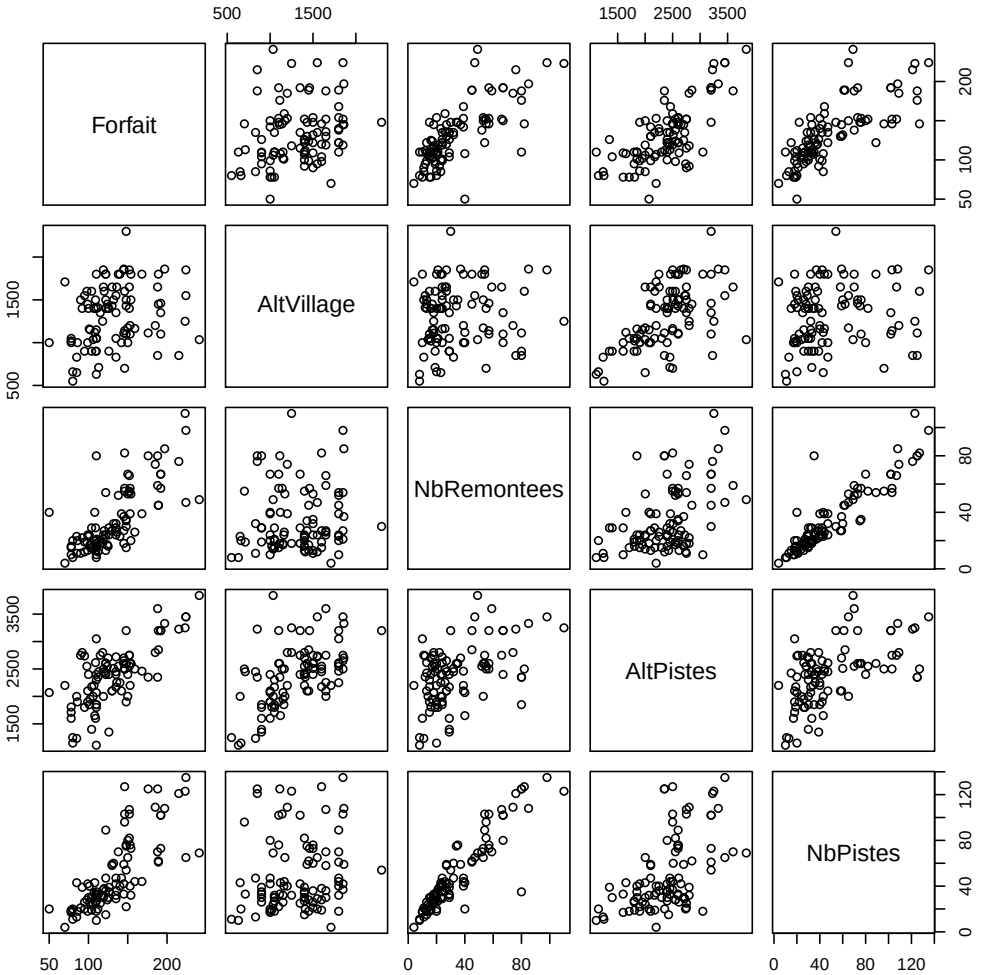
4. Calculer la valeur prise par la statistique de test (" ?" dans la sortie) et conclure le test au niveau 5% en utilisant le quantile à 95% de la loi de Fisher$(2\,;16)$: $f_{2;16}^{0,95} \approx 3,6$.
5. Conclure : quel modèle proposez-vous finalement de conserver ?

Exercice 3

On a observé, pour 97 stations de ski françaises :

- le prix du forfait semaine ("Forfait") ;
- l'altitude de la station village ("AltVillage") ;
- le nombre de remontées ("NbRemontees") ;
- l'altitude du sommet des pistes ("AltPistes") ;
- le nombre de pistes ("NbPistes").

La matrice des nuages de points ainsi que la matrice de corrélation entre toutes les variables sont fournies ci-après.



	Forfait	AltVillage	NbRemontees	AltPistes	NbPistes
Forfait	1.00	0.25	0.74	0.72	0.77
AltVillage	0.25	1.00	0.07	0.56	0.15
NbRemontees	0.74	0.07	1.00	0.48	0.92
AltPistes	0.72	0.56	0.48	1.00	0.55
NbPistes	0.77	0.15	0.92	0.55	1.00

On souhaite proposer un modèle de régression multiple pour expliquer le prix du forfait semaine.

1. D'après les informations données au début de l'exercice, est-il judicieux de mettre en place un modèle de régression incluant toutes les variables explicatives disponibles ?

2. On décide de comparer deux modèles de régression multiple :
   — un modèle expliquant la variable Forfait par NbRemontees et AltPistes ;
   — un modèle expliquant la variable Forfait par AltVillage, NbRemontees et AltPistes.

   Pour cela on a utilisé la fonction anova de R qui donne la sortie suivante :

```
Model 1: Forfait ~ NbRemontees + AltPistes
Model 2: Forfait ~ AltVillage + NbRemontees + AltPistes
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     94 38525
2     93 37900  1    624.66      ?  0.2188
```

   (a) Quel test a été utilisé pour répondre au problème posé ?
   (b) Donner précisément les hypothèses de test H_0 et H_1.
   (c) Calculer la valeur prise par la statistique de test F.
   (d) Expliquer par un graphique annoté et commenté ce que représente la p-value de ce test.
   (e) Conclure le test au niveau 5%.
   (f) Quel autre test aurait conduit à la même conclusion ? Pourquoi ?

3. On ajuste le modèle expliquant le prix des forfaits par le nombre de remontées dans la station et l'altitude du sommet des pistes. La sortie donnée par le logiciel R est la suivante :

```
lm(formula = Forfait ~ NbRemontees + AltPistes, data = Ski)

Residuals:
    Min      1Q  Median      3Q     Max
-76.416  -8.571  -0.695   9.226  48.679

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.494035   9.413797   2.496   0.0143 *
NbRemontees  0.858483   0.104148   8.243 9.85e-13 ***
AltPistes    0.033132   0.004331   7.650 1.72e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.24 on 94 degrees of freedom
Multiple R-squared: 0.7226,     Adjusted R-squared: 0.7167
F-statistic: 122.4 on 2 and 94 DF,  p-value: < 2.2e-16
```

   (a) Tous les paramètres sont ils significativement différents de 0 ?
   (b) Le test de Fisher global permet-il également de répondre à cette question ?

TD Modèle linéaire - Feuille 4

STID2 - IUT Vannes - 2021/2022

A. Cuzol / A. Poterie

Exercice 1

On dispose d'un jeu de données donnant les caractéristiques de 18 voitures. On s'intéresse aux variables suivantes :

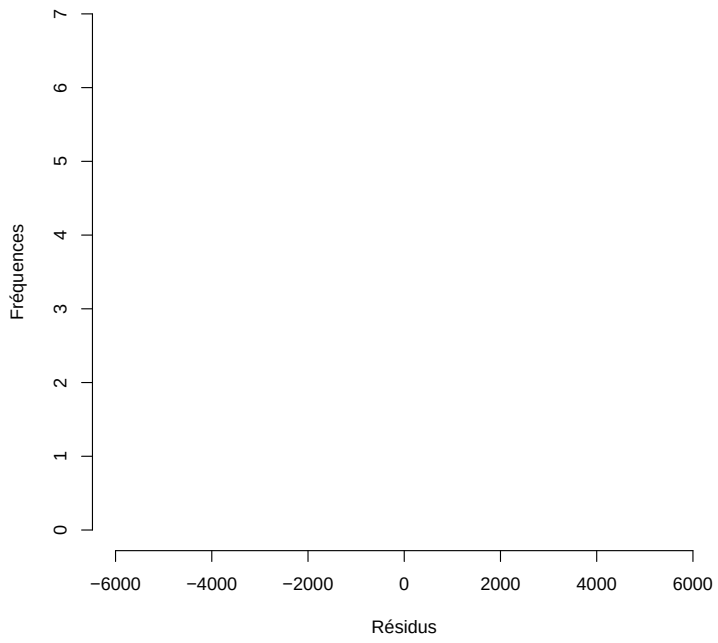
- le prix ("prix") ;
- la largeur ("larg") ;
- la vitesse ("vit") ;
- la finition ("finit").

On ajuste le modèle de régression multiple de l'exercice précédent expliquant le prix de la voiture par la largeur, la vitesse et la finition.

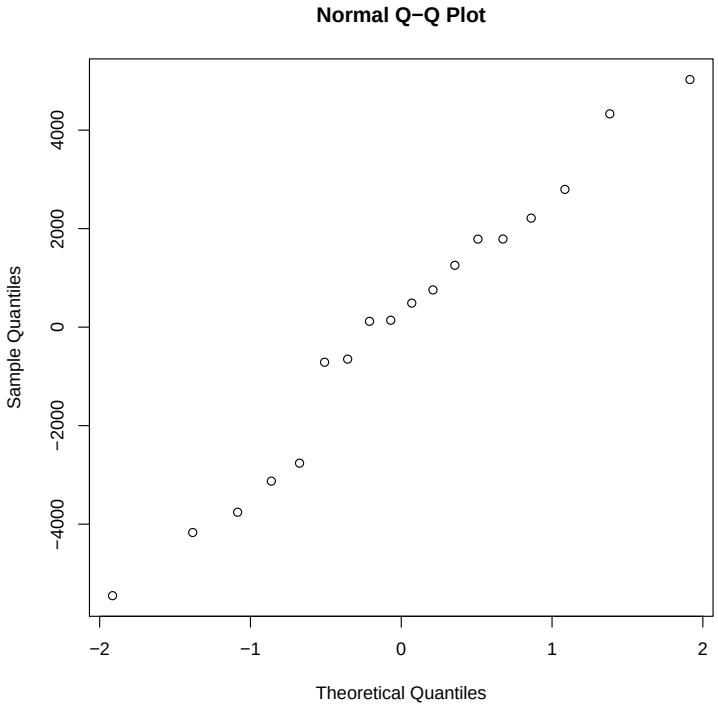
On donne les résidus de la régression pour les $n = 18$ voitures :

-3758	118	2797	1787	489	-4170	-649	1256	5027
1790	756	-5452	4330	141	2213	-712	-3126	-2761

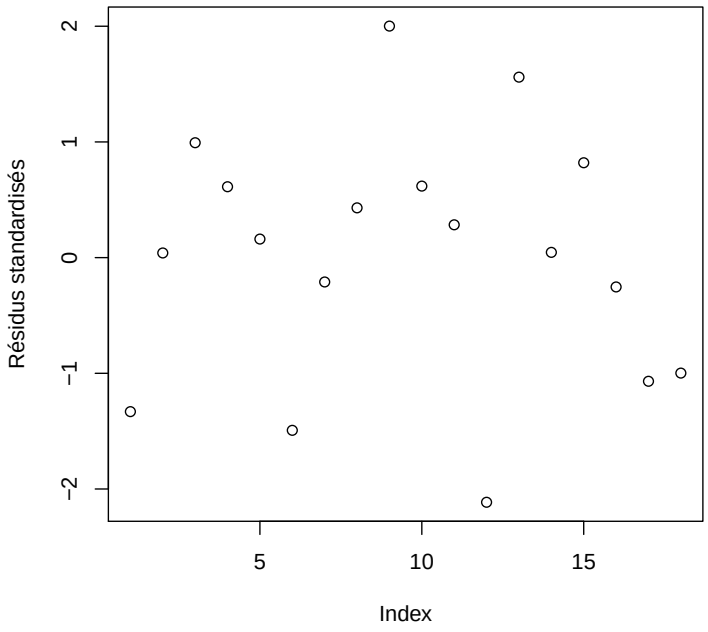
1. Tracer l'histogramme des résidus sur le graphique ci-dessous. L'hypothèse gaussienne sur les erreurs semble-t-elle respectée ?

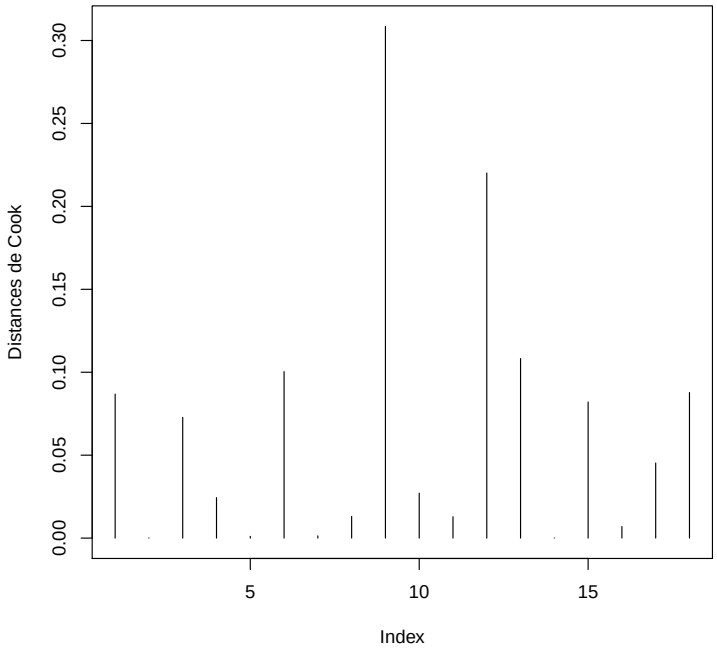


2. Que peut-on conclure en observant le graphique ci-dessous construit à partir des résidus ?



3. On donne ci-dessous les graphiques des résidus standardisés et des distances de Cook pour cette régression. On donne également les quantiles : $f_{3;15}^{0,1} = 0,19$ et $f_{3;15}^{0,5} = 0,83$. Commenter ces résultats.




Distances de Cook
0.00
0.05
0.10
0.15
0.20
0.25
0.30
5
10
15
Index

Exercice 2 *(Extrait d'examen)*

On considère des données qui viennent du Groupe d'Etude et de Réflexion Inter-régional (GERI). Elles décrivent quatre grands thèmes : la démographie, l'emploi, la fiscalité directe locale et la criminalité. Voici la liste des variables disponibles :

- CRIM : taux de criminalité (en $^0/_{00}$)
- TXCR : taux de croissance de la population sur la période 1982-1990
- ETRA : part des étrangers dans la population totale
- URBR : taux d'urbanisation
- JEUN : part des 0-19 ans dans la population totale
- AGE : part des plus de 65 ans dans la population totale
- CHOM : taux de chômage
- FISC : produit, en francs constants 1990 et par habitant, des quatre taxes directes locales (professionnelle, habitation, foncier bâti, foncier non bâti)
- FE90 : taux de fécondité (en $^0/_{00}$)

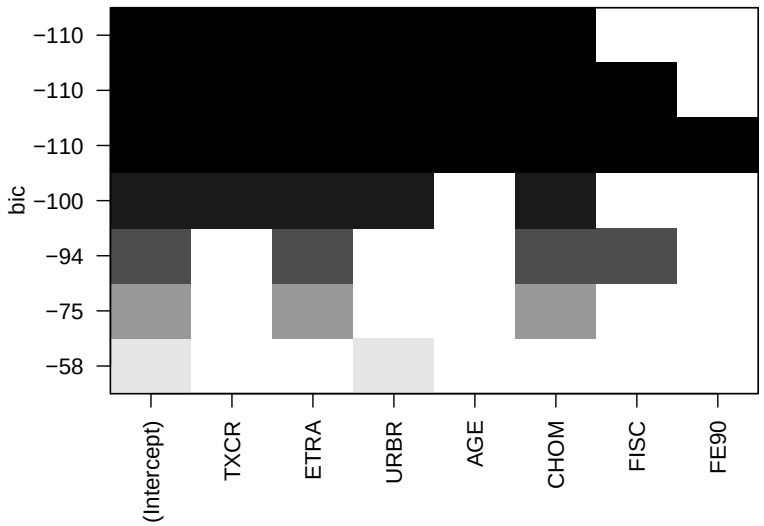
On dispose d'observations sur 95 départements français pour l'année 1990. On cherche à étudier la variable CRIM en fonction des autres variables.

1. On estime le modèle de régression multiple expliquant la variable CRIM par toutes les variables disponibles. On calcule les VIF des variables explicatives et on obtient le résultat suivant avec R :

TXCR	ETRA	URBR	JEUN	AGE	CHOM	FISC	FE90
1.470026	3.681228	4.078686	11.949079	12.425848	1.364732	2.040640	4.287334

   Que nous indique ce résultat ?

2. On a mis en place une procédure de sélection de modèle basée sur une recherche exhaustive utilisant le critère BIC. On obtient le résultat suivant :



   (a) Expliquer (en deux lignes maximum !) comment a fonctionné la procédure de sélection.
   (b) Quel modèle conserve-t-on finalement ?

TD Modèle linéaire - Feuille 5

STID2 - IUT Vannes - 2021/2022

A. Cuzol / A. Poterie

Exercice 1

Un éleveur souhaite acheter de nouvelles vaches pour sa production laitière. Il possède trois races différentes de vaches et se pose donc la question de savoir si la race est importante pour son choix. Les productions de lait journalières de chacune de ses bêtes suivant sa race sont données dans le tableau ci-dessous.

Race A	Race B	Race C
20.1	22.6	31.2
19.8	24.1	31.6
21.3	23.8	31
20.7	22.5	32.1
	23.4	31.4
	24.5	
	22.9	

1. Quelle est la variable à expliquer ? Quelle est la variable explicative ? Ces variables sont-elles quantitatives ou qualitatives ?

2. Combien de modalités a la variable qualitative ? On note ce nombre I.

3. On note n_i le nombre d'observations par modalité et n le nombre total d'observations. Donner la valeur de n et des n_i pour $i = 1, \ldots, I$.

4. On note $\bar{y}_i$ la moyenne empirique de la variable à expliquer pour la modalité i.
   — Donner l'expression de $\bar{y}_i$;
   — Calculer $\bar{y}_i$ pour $i = 1, \ldots, I$.

5. On note $\bar{y}$ la moyenne empirique générale de la variable à expliquer.
   — Donner l'expression de $\bar{y}$ en fonction des moyennes $\bar{y}_i$;
   — Calculer $\bar{y}$.

6. Donner l'écriture du modèle d'analyse de la variance à un facteur, en notant μ_i la moyenne théorique dans chaque groupe i.

7. On souhaite tester l'influence de la variable "Race" sur la production de lait. On donne la sortie R suivante :

```
Analysis of Variance Table

Response: Production
          Df  Sum Sq  Mean Sq  F value  Pr(>F)
Race       2  307.92  153.959     ?     4.338e-12 ***
Residuals 13   5.60   0.431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Donner les hypothèses de test H_0 et H_1 du test de Fisher qui a été mis en place.

(b) On rappelle l'expression de la statistique de test pour ce test de Fisher :

$$F = \frac{n-I}{I-1} \frac{\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik} - \bar{y}_i)^2}$$

Quelle est sa loi sous l'hypothèse H_0 ?

(c) Remplir le " ?" dans la sortie.

(d) Conclure le test : la race a-t-elle une influence sur la production de lait ?

8. On complète l'analyse avec le test de Tukey pour savoir quelles races ont des productions moyennes significativement différentes. On donne la sortie R correspondante :

```
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = vaches$Production ~ vaches$Race)

$`vaches$Race`
      diff      lwr       upr    p adj
B-A  2.925 1.838834  4.011166 2.21e-05
C-A 10.985 9.822521 12.147479 0.00e+00
C-B  8.060 7.045305  9.074695 0.00e+00
```

(a) Quelles conclusions obtient-on ?

(b) Que conseillez-vous finalement à l'éleveur ?

Exercice 2 *(Extrait d'examen)*

On dispose des variables suivantes, pour 63 stations de ski françaises :

- le prix du forfait semaine ("Forfait") ;
- le département dans lequel se situe la station ("Departement"), ayant 4 modalités :
  - "Hautes-Alpes" ;
  - "Haute-Savoie" ;
  - "Isere" ;
  - "Savoie".

On se demande si le prix du forfait semaine est influencé par le département dans lequel se situe la station. On décide de mettre en place une analyse de la variance. On estime le modèle et on observe les sorties R suivantes :

```
lm(formula = Forfait ~ Departement, data = Ski)

Residuals:
    Min      1Q  Median      3Q     Max
-62.389 -18.896  -1.389  15.923  93.611

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            153.000      9.381  16.309  < 2e-16 ***
DepartementHaute-Savoie  -5.611     11.908  -0.471  0.63923
DepartementIsere        -33.923     12.747  -2.661  0.01001 *
DepartementSavoie       -33.286     11.581  -2.874  0.00562 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.11 on 59 degrees of freedom
Multiple R-squared: 0.2008,     Adjusted R-squared: 0.1601
F-statistic:  4.94 on 3 and 59 DF,  p-value: 0.003973
```

```
Analysis of Variance Table

Response: Forfait
             Df   Sum Sq   Mean Sq   F value   Pr(>F)
Departement   3    14348    4782.6      4.94      ?
Residuals    59    57119     968.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Quel est le prix moyen du forfait semaine dans chacun des départements "Hautes-Alpes", "Haute-Savoie", "Isere" et "Savoie" ?

2. Donner la valeur de "?" dans la deuxième sortie R ci-dessus.

3. Le quantile de niveau $1-\alpha$ de la loi de Fisher de paramètres $(3, 59)$ vaut 2,76 pour $\alpha = 0,05$. Conclure le test au niveau $\alpha = 0,05$: le département où se situe la station de ski a-t-il un effet sur le prix du forfait ?