

## Première Année Master M.A.E.F. 2024 – 2025

## Econométrie II

Contrôle continu n°2, avril 2025

Examen de 1h30. Tout document ou calculatrice est interdit.

## Exercice théorique (Sur 20 points)

On considère  $p$  suites de réels observés  $(X_i^{(j)})_{i \in \mathbf{N}}$ , et pour  $n \geq p+1$ , on définit la matrice  $Z = (Z_{ij})_{1 \leq i \leq n, 1 \leq j \leq p+1}$  telle que  $Z_{i1} = 1$  et  $Z_{ij} = X_i^{(j-1)}$  pour tout  $1 \leq i \leq n$  et  $1 \leq j \leq p+1$ , et on note  $[Z] = \{Z\beta, \beta \in \mathbf{R}^{p+1}\}$ . Avec  $Y = {}^t(Y_1, \dots, Y_n)$ , on suppose qu'il existe un vecteur réel inconnu  $\theta = (\theta_j)_{0 \leq j \leq p} \in \mathbf{R}^{p+1}$  tel que

$$Y = Z\theta + \varepsilon$$

où  $\varepsilon = {}^t(\varepsilon_1, \dots, \varepsilon_n)$ , la suite  $(\varepsilon_k)_{k \in \mathbf{N}}$  est une suite non observée de variables aléatoires indépendantes et identiquement distribuées (v.a.i.i.d.) centrées de variance  $0 < \sigma_\varepsilon^2 < \infty$ . On suppose que  $Y$  est aussi observé.

1. On note  $\|U\|^2 = \langle U, U \rangle$  où  $\langle U, V \rangle = {}^tUV$  pour  $U, V \in \mathbf{R}^n$ . Montrer que  $\hat{Y} = {}^t(\hat{Y}_1, \dots, \hat{Y}_n) = \operatorname{argmin}_{U \in [Z]} \|Y - U\|$  s'écrit  $\hat{Y} = MY$  avec  $M$  une matrice que l'on précisera en distinguant suivant le rang de  $Z$  (**2pts**).
2. Uniquement dans cette question, on suppose que les suites  $(X_i^{(j)})_{i \in \mathbf{N}}$  sont des suites de v.a.i.i.d., de lois absolument continues telles que  $\mathbb{E}[X_0^{(j)}] = m_j$  et  $\operatorname{var}(X_0^{(j)}) = \sigma_j^2$  avec  $0 < \sigma_j^2 < \infty$  pour  $1 \leq j \leq p$ , les suites  $(X_i^{(j)})_{i \in \mathbf{N}}$  étant aussi indépendantes les unes les autres. On note  $\bar{X}_i^{(j)} = X_i^{(j)} - m_j$  et  $\bar{Z}$  la matrice constituée de la même manière que  $Z$  mais avec les  $\bar{X}_i^{(j)}$  qui remplacent les  $X_i^{(j)}$ . Démontrer que  $Z$  et  $\bar{Z}$  sont presque sûrement de rang  $p+1$  (**1.5pts**). Démontrer que  $\hat{Y} = \bar{Z}({}^t\bar{Z}\bar{Z})^{-1}{}^t\bar{Z}Y$  presque sûrement (**1.5pts**). Démontrer que  $\frac{1}{n}{}^t\bar{Z}\bar{Z}$  converge presque sûrement quand  $n \rightarrow \infty$  vers une matrice (dépendant des  $\sigma_j^2$ ) que l'on précisera (**1pt**).
3. On supposera désormais que  $Z$  est une matrice de nombres réels et que  $\frac{1}{n}{}^tZZ = D$  avec  $D = (d_{ij})_{1 \leq i, j \leq p+1}$  une matrice diagonale définie positive. Préciser  $d_{11}$  (**0.5pts**). On note  $X^{(j)} = (X_i^{(j)})_{1 \leq i \leq n}$ . Déterminer  $\frac{1}{n} \sum_{i=1}^n X_i^{(j)}$  et  $\langle X^{(j)}, X^{(j')} \rangle$  pour  $j, j' \in \{1, \dots, p\}$  (**1pt**). Montrer que  $\hat{\theta}_j = \frac{1}{n d_{j+1, j+1}} \langle X^{(j)}, Y \rangle$  pour  $j \in \{1, \dots, p\}$  (**1pt**) et montrer que  $\hat{\theta}_j$  converge en probabilité vers  $\theta_j$  pour  $j \in \{0, \dots, p\}$  (**1pt**).
4. Rappeler le comportement de  $\hat{\sigma}^2 = \frac{1}{n-p-1} \|Y - \hat{Y}\|^2$  quand  $n \rightarrow \infty$  sous les hypothèses précédentes (**0.5pts**).
5. On veut tester  $H_0 : \theta_j = 0$  contre  $H_1 : \theta_j \neq 0$ , où  $j$  est fixé dans  $\{1, \dots, p\}$ . Montrer que le test de Student s'écrit  $\hat{T}_j = \frac{\langle X^{(j)}, Y \rangle}{\hat{\sigma} \sqrt{\langle X^{(j)}, X^{(j)} \rangle}}$  (**1pt**).
6. On suppose que  $\theta_p \neq 0$  et que l'on a oublié prendre en compte  $X^{(p)}$  dans l'analyse. On effectue donc une régression linéaire en utilisant  $Z'$ , matrice extraite de  $Z$  en supprimant sa colonne  $p+1$ . On note alors  $\hat{\theta}' = (\hat{\theta}'_j)_{0 \leq j \leq p-1}$  l'estimateur par moindres carrés de  $(\theta_j)_{0 \leq j \leq p-1}$  et  $\hat{Y}'$  les valeurs prédites par cette nouvelle régression.

(a) Montrer que  $\hat{Y}' = \left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \mathbb{I}_n + \sum_{j=1}^{p-1} \frac{\langle X^{(j)}, Y \rangle}{n d_{j+1, j+1}} X^{(j)}$  avec  $\mathbb{I}_n = {}^t(1, 1, \dots, 1)$  (**1pt**).

(b) En déduire que  $\|Y - \hat{Y}'\|^2 = \|\theta_p X^{(p)} + U'\|^2$  avec  $U' = (I_n - Z'(Z'Z')^{-1}Z')\varepsilon$  (**2pts**).

(c) Montrer que  $\frac{1}{n^2} \operatorname{var}({}^tX^{(p)}U') \xrightarrow{n \rightarrow \infty} 0$  (**1.5pts**). En déduire que  $\frac{1}{n} \|Y - \hat{Y}'\|^2 \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \theta_p^2 d_{p+1, p+1} + \sigma_\varepsilon^2$  (**2.5pts**).

(d) Si on note respectivement  $\widehat{BIC}$  et  $\widehat{BIC}'$  les critères BIC avec, respectivement, les variables  $X^{(1)}, \dots, X^{(p)}$  et  $X^{(1)}, \dots, X^{(p-1)}$ , que peut-on dire de  $\mathbb{P}(\widehat{BIC}' < \widehat{BIC})$  (**1.5pts**)? Qu'en déduire plus généralement (**0.5pts**)?

## Exercice de TP utilisant le logiciel R (Sur 6 points)

On considère la base de données **Price** des prix (absolus et relatifs) dans de nombreuses villes du monde. Plus précisément, les variables considérées pour chaque ville sont:

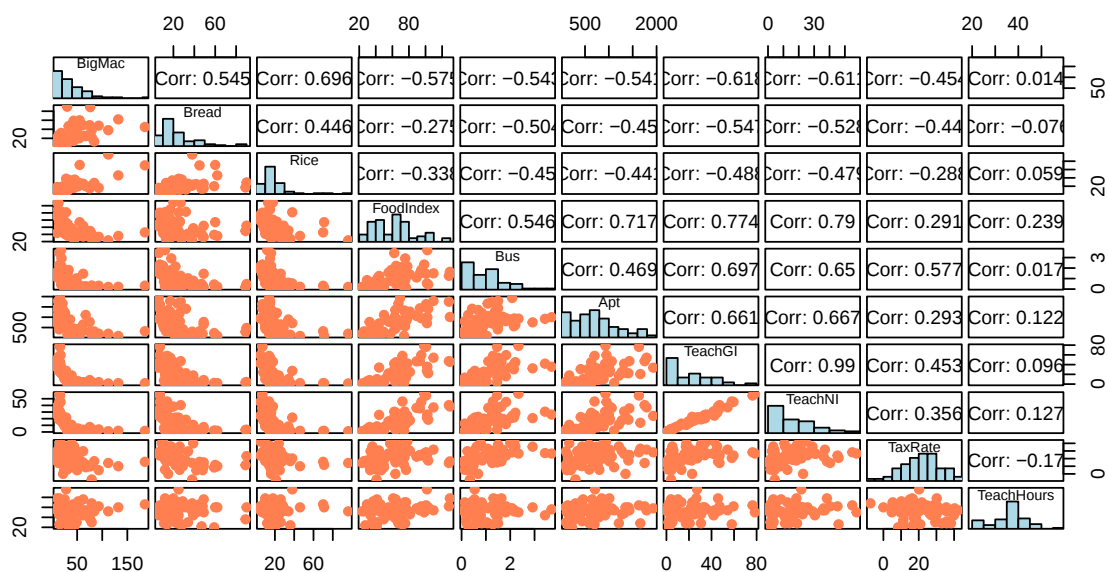
- **BigMac**: Minutes de travail pour acheter un Big Mac;
- **Bread**: Minutes de travail pour acheter 1 kg de pain;
- **Rice**: Minutes de travail pour acheter 1 kg de riz;
- **FoodIndex**: Indice des prix des denrées alimentaires (Zurich=100);
- **Bus**: Coût en dollars US pour un aller simple de 10 km en autobus;
- **Apt**: Loyer normal (en dollars US) d'un appartement de 3 pièces;
- **TeachGI**: Revenu brut d'un enseignant du primaire, en milliers de dollars US;
- **TeachNI**: Revenu net de l'enseignant du primaire, en milliers de dollars US;
- **TaxRate**: Taux d'imposition payé par un enseignant du primaire;
- **TeachHours**: Heures de travail hebdomadaires d'un enseignant du primaire.

On essaie de modéliser au mieux **FoodIndex** en fonction des autres variables.

1. Un premier regard est porté sur la base de données **Price**:

```
head(Price); str(Price); dim(Price)
pairs(Price)
```

Voici le graphe et les résultats numériques obtenus:



```
> head(Price)
      BigMac Bread Rice FoodIndex  Bus Apt TeachGI TeachNI TaxRate TeachHours
Amsterdam    16     9   9      65.9 2.00 890   34.3    20.5 40.2332         39
Athens        21    12  19      63.5 0.61 620   19.5    15.9 18.4615         29
Auckland      19    19   9      55.4 1.57 780   22.0    16.1 26.8182         40
Bangkok       50    42  25      46.4 0.47 120    4.2     4.0  4.7619         35
```

```
> str(Price)
'data.frame': 69 obs. of  10 variables:
 $ BigMac   : int  16 21 19 50 22 15 16 93 54 18 ...
 $ Bread    : int   9 12 19 42 19 7 10 48 20 11 ...
 $ Rice     : int   9 19 9 25 10 7 16 16 28 12 ...
 $ FoodIndex: num  65.9 63.5 55.4 46.4 62.9 98.4 64.4 33.3 36.9 74.7 ...
 $ Bus      : num   2 0.61 1.57 0.47 0.91 2.34 2.21 0.32 0.36 1.5 ...
 $ Apt      : int  890 620 780 120 590 930 630 200 350 590 ...
```

```

$ TeachGI    : num  34.3 19.5 22 4.2 25.5 78.5 45 4.1 4.1 30.6 ...
$ TeachNI    : num  20.5 15.9 16.1 4 20.1 57.6 28.9 3.9 3.3 18.9 ...
$ TaxRate    : num  40.23 18.46 26.82 4.76 21.18 ...
$ TeachHours: int   39 29 40 35 39 35 36 41 40 24 ...

> dim(Price)
[1] 69 10

```

Questions 1: Que peut-on déduire du graphe (0.5pts)?

2. On tape ensuite les commandes:

```

regBM=lm(FoodIndex~BigMac,data=Price)
summary(regBM); plot(regBM,1)

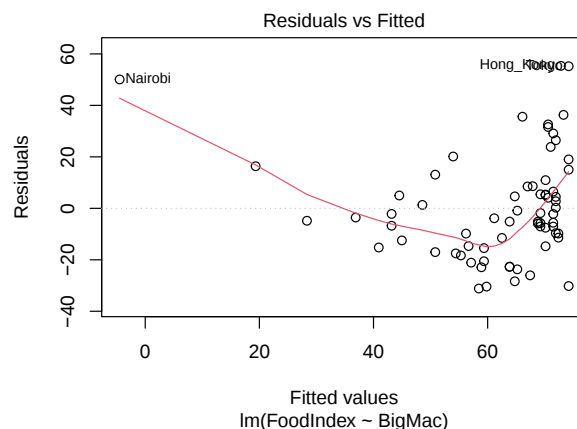
```

Voici les résultats numériques et la figure obtenue:

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )     |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 78.68879 | 3.80273    | 20.693  | < 2e-16 ***  |
| BigMac      | -0.44958 | 0.07824    | -5.746  | 2.44e-07 *** |

Residual standard error: 20.27 on 67 degrees of freedom  
Multiple R-squared: 0.3301, Adjusted R-squared: 0.3201  
F-statistic: 33.02 on 1 and 67 DF, p-value: 2.435e-07



Questions 2: Comment a-t-on obtenu explicitement la valeur -5.746? Pourquoi a-t-on 67 DF? Que conclure? (1pt)

3. On a ensuite tapé les commandes:

```

reg=lm(FoodIndex ~.,data=Price)
summary(reg); plot(reg)

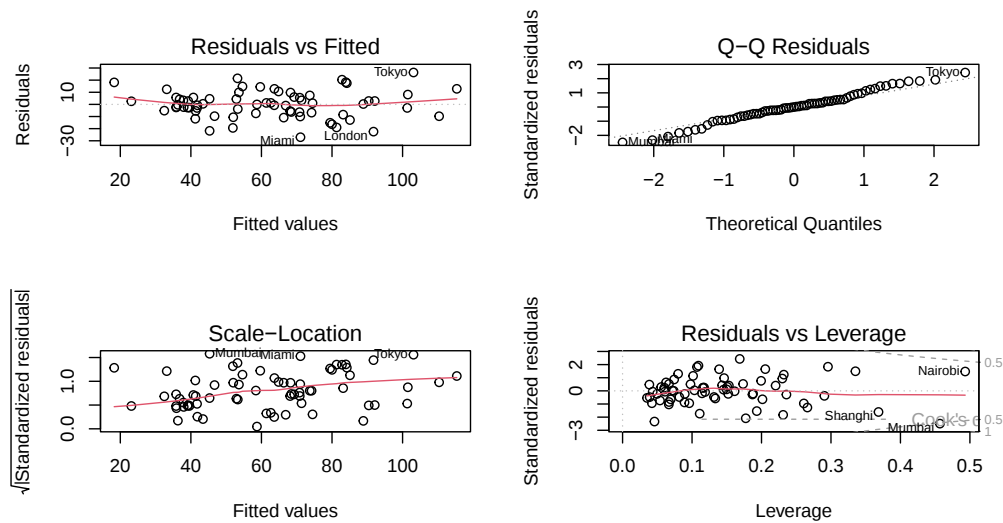
```

Et à nouveau les résultats obtenus:

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )     |
|-------------|----------|------------|---------|--------------|
| (Intercept) | -1.09968 | 11.19872   | -0.098  | 0.9221       |
| BigMac      | -0.20569 | 0.07798    | -2.638  | 0.0107 *     |
| Bread       | 0.44383  | 0.10564    | 4.201   | 9.11e-05 *** |
| Rice        | 0.26881  | 0.13597    | 1.977   | 0.0527 .     |
| Bus         | 3.59014  | 2.83317    | 1.267   | 0.2101       |
| Apt         | 0.01825  | 0.00434    | 4.204   | 9.02e-05 *** |
| TeachGI     | -0.97768 | 0.86750    | -1.127  | 0.2643       |
| TeachNI     | 2.22275  | 1.13819    | 1.953   | 0.0556 .     |
| TaxRate     | 0.26530  | 0.25724    | 1.031   | 0.3066       |
| TeachHours  | 0.48015  | 0.20478    | 2.345   | 0.0224 *     |

Residual standard error: 11.86 on 59 degrees of freedom  
 Multiple R-squared: 0.7981, Adjusted R-squared: 0.7673  
 F-statistic: 25.91 on 9 and 59 DF, p-value: < 2.2e-16



Questions 3: Expliquez ce qui a été fait et ce que l'on peut en conclure (1pt)

4. On poursuit avec les commandes:

```
library(MASS)
BX=boxcox(FoodIndex~.,plotit = TRUE,lambda = seq(-5,5,0.01),data=Price)
ind=which(BX$y==max(BX$y)); lambda=BX$x[ind]; lambda
regBC=lm(I(FoodIndex^0.5)~.,data=Price); summary(regBC)
1-sum((FoodIndex-regBC$fit^2)^2)/sum((FoodIndex-mean(FoodIndex))^2)
```

Et à nouveau les résultats obtenus:

```
> lambda
[1] 0.37
```

```
> summary(regBC)
```

Coefficients:

|             | Estimate   | Std. Error | t value | Pr(> t )     |
|-------------|------------|------------|---------|--------------|
| (Intercept) | 3.8724016  | 0.6943306  | 5.577   | 6.43e-07 *** |
| BigMac      | -0.0141949 | 0.0048351  | -2.936  | 0.00474 **   |
| Bread       | 0.0291718  | 0.0065501  | 4.454   | 3.82e-05 *** |
| Rice        | 0.0146537  | 0.0084305  | 1.738   | 0.08740 .    |
| Bus         | 0.2692093  | 0.1756594  | 1.533   | 0.13073      |
| Apt         | 0.0011580  | 0.0002691  | 4.304   | 6.42e-05 *** |
| TeachGI     | -0.0458402 | 0.0537860  | -0.852  | 0.39751      |
| TeachNI     | 0.1143680  | 0.0705689  | 1.621   | 0.11042      |
| TaxRate     | 0.0161516  | 0.0159490  | 1.013   | 0.31534      |
| TeachHours  | 0.0294259  | 0.0126966  | 2.318   | 0.02396 *    |

Residual standard error: 0.7354 on 59 degrees of freedom  
 Multiple R-squared: 0.8025, Adjusted R-squared: 0.7724  
 F-statistic: 26.64 on 9 and 59 DF, p-value: < 2.2e-16

```
[1] 0.8058518
```

Questions 4: Expliquez ce qui a été fait, ce que représente 0.37, comment on calcule 0.8058518, et ce que l'on peut conclure de tout ceci (1.5pts)

5. On a ensuite tapé les commandes:

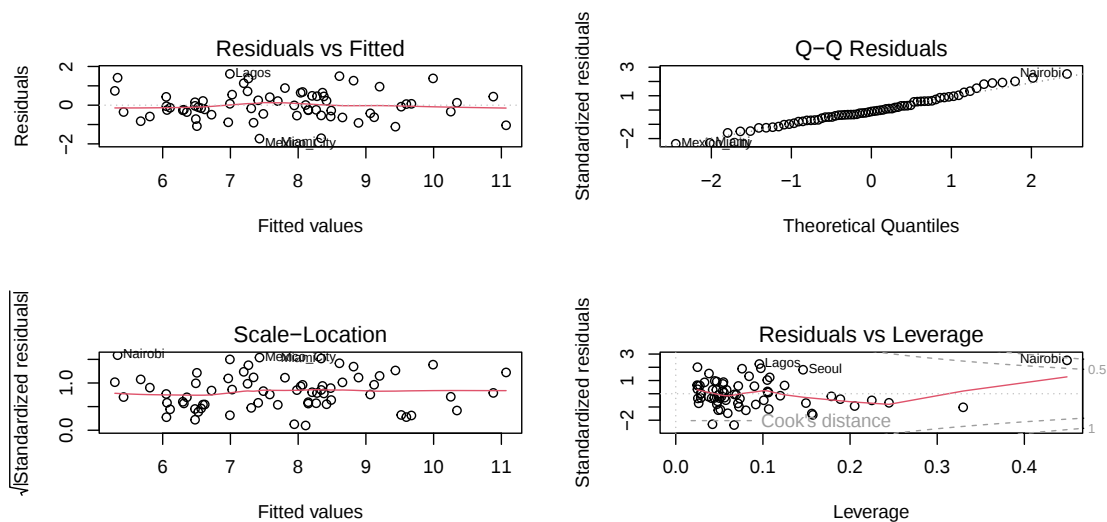
```
Reg=stepAIC(regBC,k=log(60), trace=FALSE)
summary(Reg); par(mfrow=c(2,2)); plot(Reg)
```

Avec pour résultats obtenus:

Coefficients:

|             | Estimate   | Std. Error | t value | Pr(> t )     |
|-------------|------------|------------|---------|--------------|
| (Intercept) | 4.6712231  | 0.5402114  | 8.647   | 2.65e-12 *** |
| BigMac      | -0.0122607 | 0.0040152  | -3.054  | 0.003311 **  |
| Bread       | 0.0271351  | 0.0064574  | 4.202   | 8.49e-05 *** |
| Apt         | 0.0011364  | 0.0002756  | 4.123   | 0.000111 *** |
| TeachNI     | 0.0608001  | 0.0098193  | 6.192   | 4.99e-08 *** |
| TeachHours  | 0.0290592  | 0.0126035  | 2.306   | 0.024434 *   |

Residual standard error: 0.7563 on 63 degrees of freedom  
 Multiple R-squared: 0.7769, Adjusted R-squared: 0.7592  
 F-statistic: 43.88 on 5 and 63 DF, p-value: < 2.2e-16



Questions 5: Expliquez pourquoi est écrit 5 and 63 DF. Expliquez ce qui a été fait et ce que l'on peut en conclure (1pt)

6. On termine avec les commandes:

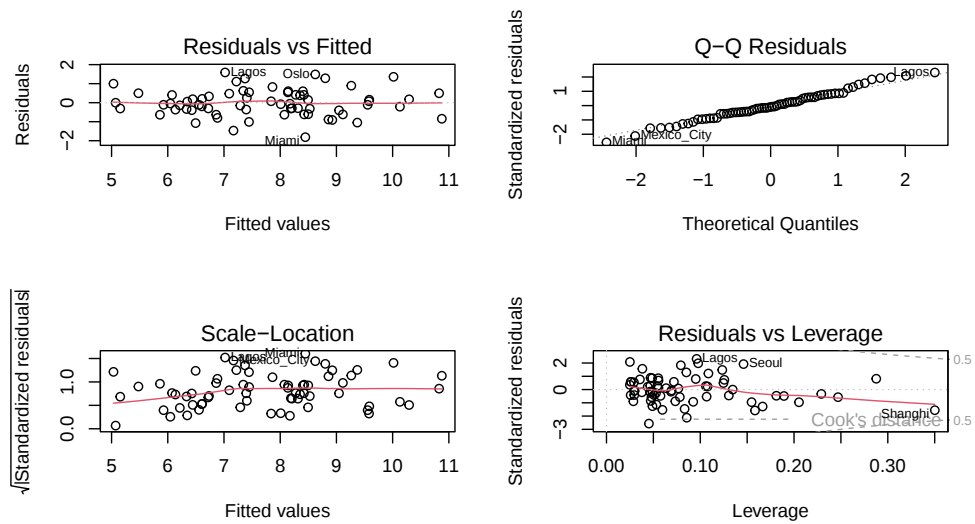
```
Price2=Price[-46,]
RegBC2=lm(I(FoodIndex^0.5) ~ ., data=Price2)
Reg2=stepAIC(RegBC2,k=log(59))
summary(Reg2); plot(Reg2)
```

Avec pour résultats obtenus:

Coefficients:

|             | Estimate   | Std. Error | t value | Pr(> t )     |
|-------------|------------|------------|---------|--------------|
| (Intercept) | 5.0933991  | 0.5403822  | 9.426   | 1.39e-13 *** |
| BigMac      | -0.0208690 | 0.0050328  | -4.147  | 0.000104 *** |
| Bread       | 0.0296561  | 0.0062441  | 4.749   | 1.25e-05 *** |
| Apt         | 0.0010405  | 0.0002659  | 3.913   | 0.000229 *** |
| TeachNI     | 0.0556508  | 0.0095837  | 5.807   | 2.36e-07 *** |
| TeachHours  | 0.0276745  | 0.0120556  | 2.296   | 0.025095 *   |

Residual standard error: 0.7228 on 62 degrees of freedom  
 Multiple R-squared: 0.7983, Adjusted R-squared: 0.7821  
 F-statistic: 49.08 on 5 and 62 DF, p-value: < 2.2e-16



Questions 6: Expliquez ce qui a été fait et ce que l'on peut en conclure (1pt)