Exercices sur les modèles de régression

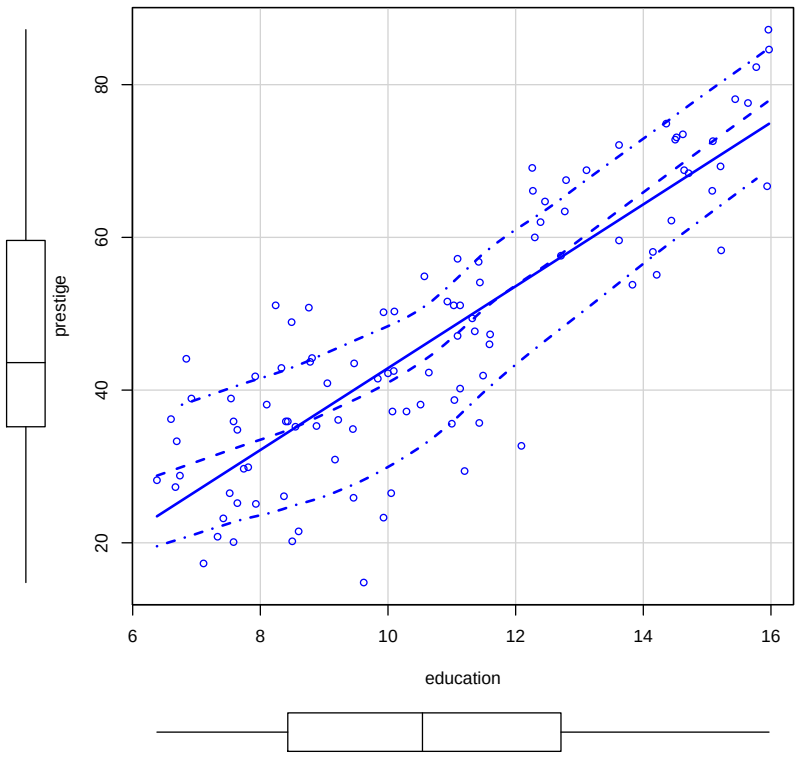
Christophe Chesneau

Université de Caen

Exercices sur les modèles de régression

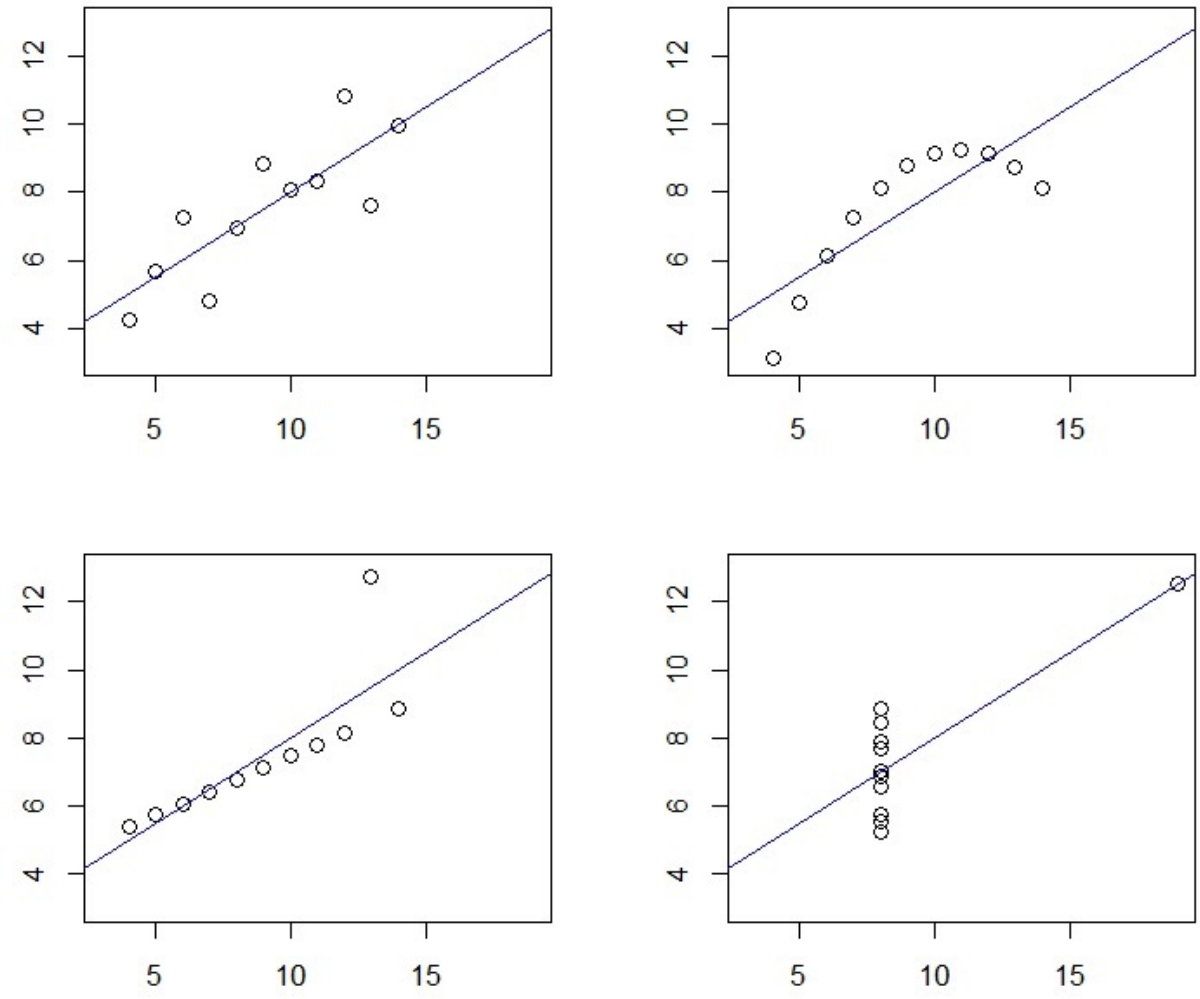
Christophe Chesneau

https://chesneau.users.lmno.cnrs.fr/



Caen, le 22 Février 2022

Exercice 1. On dipose de 4 jeux de données à partir desquels on souhaite expliquer une variable quantitative Y à partir d'une variable quantitative X. Pour chacun d'entre eux, on adopte le modèle de *rls*. Les nuages de points et les droites de régression associées sont :



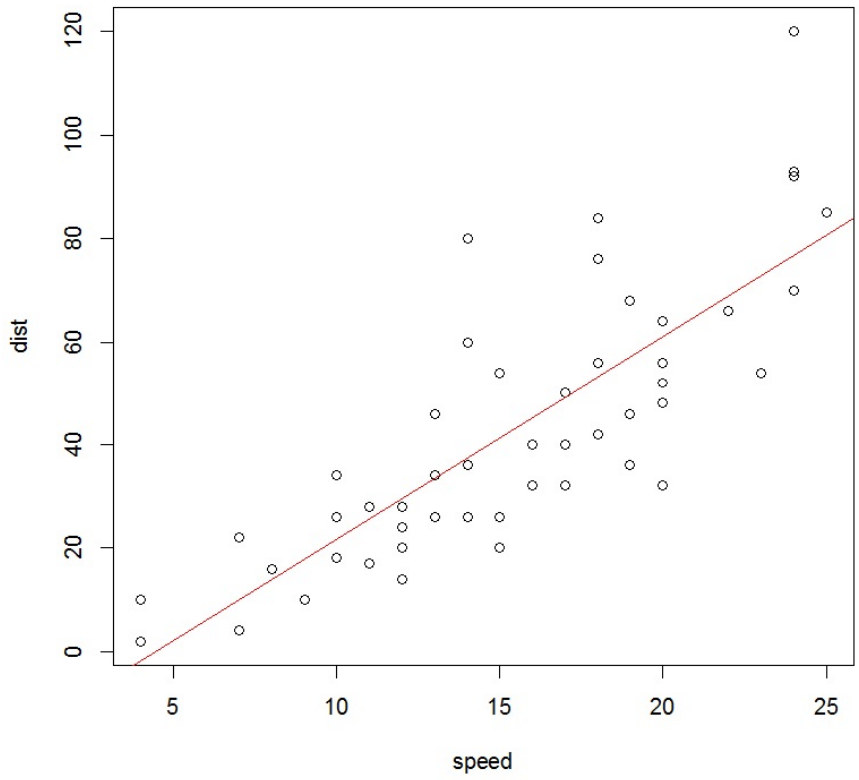
Pour chaque jeu de données :

- dire si le modèle de *rls* est bien adapté aux données,
- proposer une meilleure alternative si besoin est.

Exercice 2. On souhaite expliquer la distance de freinage d'une voiture (variable Y) à partir de sa vitesse (variable X). On adopte le modèle de *rls* : $Y = \beta_0 + \beta_1 X + \epsilon$.

Décrire brièvement l'enjeu des commandes R suivantes :

```
data(cars)
attach(cars)
plot(cars)
reg = lm(dist ~ speed)
abline(reg, col = "red")
```




```
summary(reg)
```

Cela renvoie :

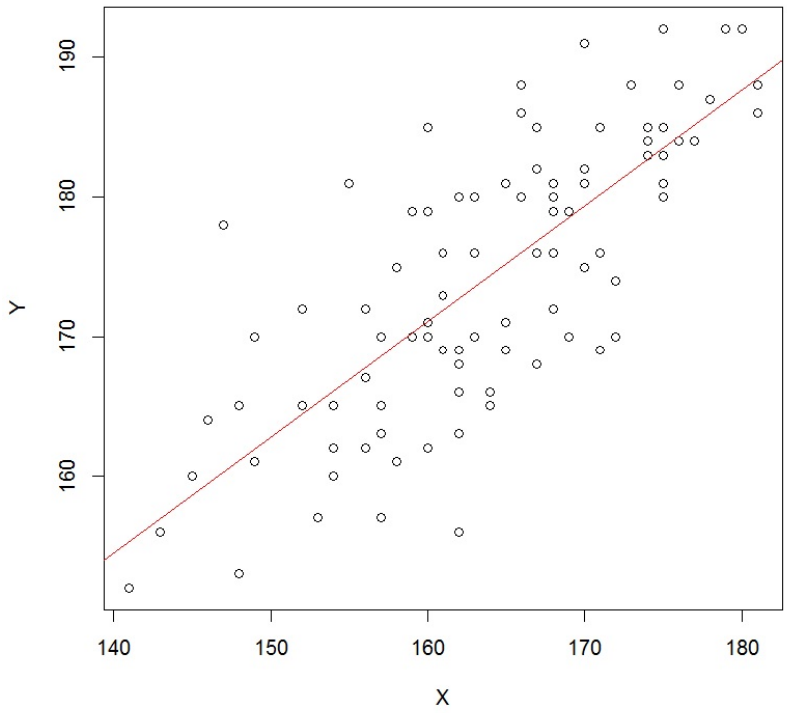
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-17.5791	6.7584	-2.60	0.0123	*
speed	3.9324	0.4155	9.46	0.0000	***

```
Residual standard error:  15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic:  89.57 on 1 and 48 DF, p-value:  1.49e-12
```

Exercice 3. On souhaite expliquer la taille d'un homme marié (variable Y) à partir de la taille de sa femme (variable X). On adopte le modèle de *rls* : $Y = \beta_0 + \beta_1 X + \epsilon$.

1. Décrire brièvement l'enjeu des commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/taille-couple.txt",
header = T)
attach(w)
plot(X, Y)
reg = lm(Y ~ X)
abline(reg, col = "red")
```




```
summary(reg)
```

Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	38.3572	12.0738	3.18	0.0020	**
X	0.8294	0.0736	11.27	0.0000	***

```
Residual standard error:  6.498 on 93 degrees of freedom
Multiple R-squared:  0.5772, Adjusted R-squared:  0.5727
F-statistic:  127 on 1 and 93 DF, p-value:  < 2.2e-16
```

2. Commenter la qualité du modèle.

3. Commenter la commande R suivante :

```
predict(reg, data.frame(X = 176))
```

   Cela renvoie 184.3365.

4. Commenter la commande R suivante :

```
confint(reg, level = 0.95)
```

   Cela renvoie :

	2.5 %	97.5 %
(Intercept)	14.3809381	62.3333650
X	0.6832576	0.9755984

5. Commenter la commande R suivante :

```
predict(reg, data.frame(X = 168), interval = "confidence")
```

   Cela renvoie :

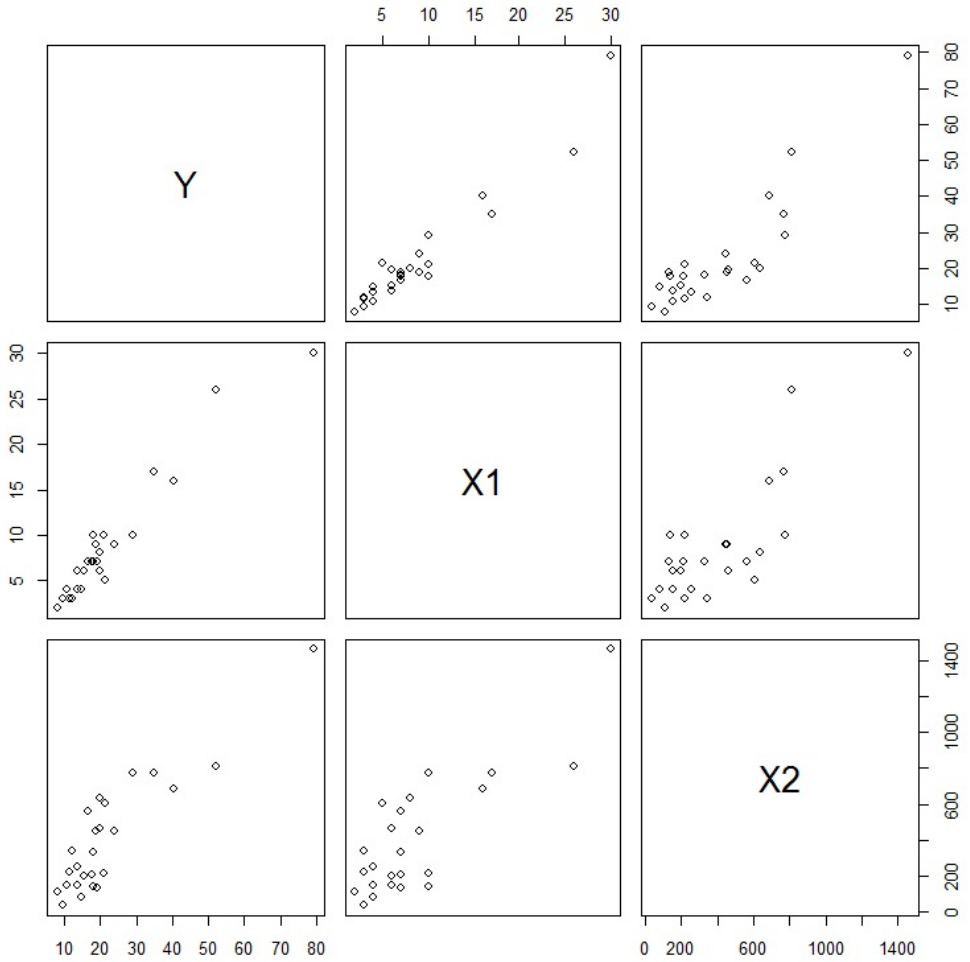
fit	lwr	upr
177.7011	176.2405	179.1616

Exercice 4. On souhaite expliquer le temps en minutes (variable Y) qu'un employé met pour approvisionner un réseau de distributeurs de boissons à partir du nombre de caisses de bouteilles à charger (variable $X1$) et de la distance parcourue en mètres (variable $X2$). On adopte le modèle de *rlm* :

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \epsilon.$$

1. Décrire brièvement l'enjeu des commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/boisson.txt", header
= T)
attach(w)
pairs(w)
```



2. Décrire brièvement l'enjeu des commandes R suivantes :

```
reg = lm(Y ~ X1 + X2)
summary(reg)
```

Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.3412	1.0967	2.13	0.0442	$*$
X1	1.6159	0.1707	9.46	0.0000	$***$
X2	0.0144	0.0036	3.98	0.0006	$***$

```
Residual standard error:  3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic:  261.2 on 2 and 22 DF, p-value:  4.687e-16
```

3. Commenter la qualité du modèle.
4. Commenter la commande R suivante :

```
predict(reg, data.frame(X1 = 5, X2 = 278))
```

Cela renvoie 14.41975.

5. Commenter la commande R suivante :

```
confint(reg, level = 0.95)
```

Cela renvoie :

	2.5 %	97.5 %
(Intercept)	0.066751987	4.61571030
X1	1.261824662	1.96998976
X2	0.006891745	0.02187791

6. Commenter la commande R suivante :

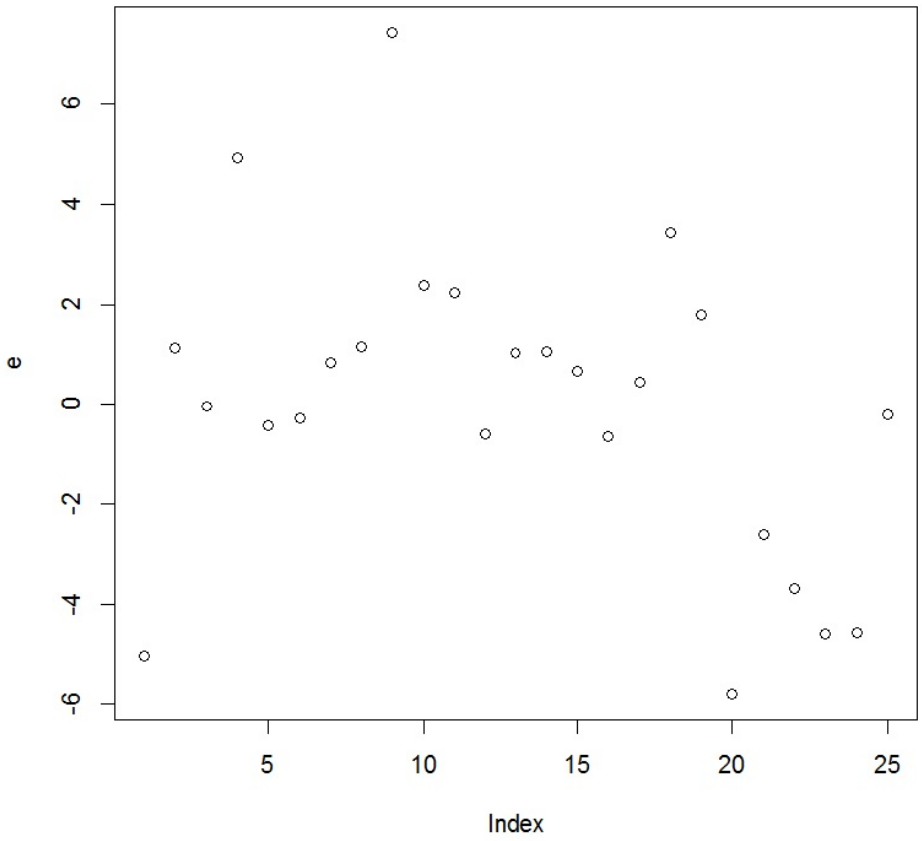
```
predict(reg, data.frame(X1 = 3, X2 = 431), interval = "confidence")
```

Cela renvoie :

fit	lwr	upr
13.38881	10.82736	15.95026

7. Commenter les commandes R suivantes :

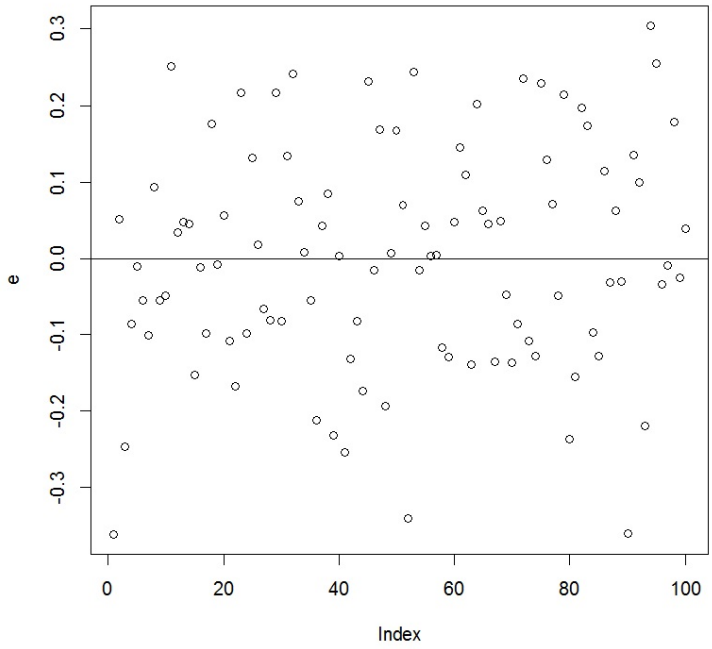
```
e = residuals(reg)
plot(e)
```



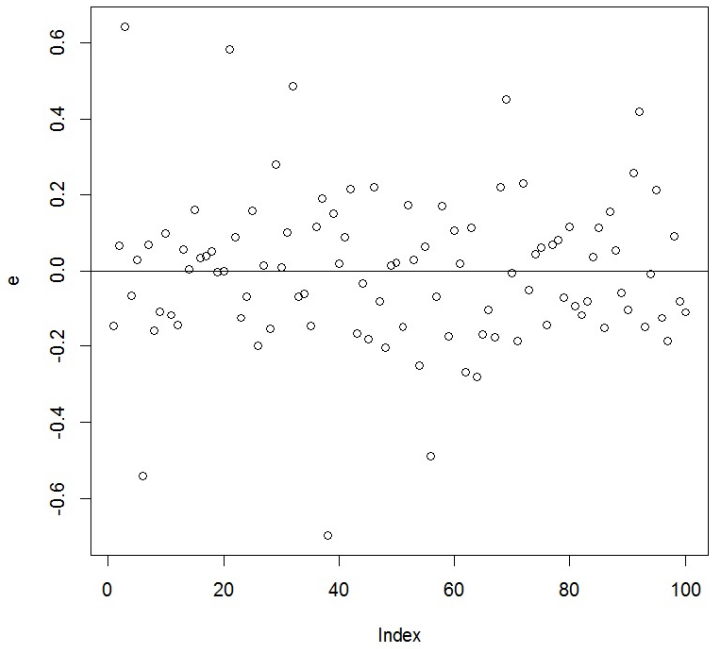
Que pensez-vous de ce nuage de points ? Est-ce que les hypothèses standards du modèle *rlm* semblent vérifiées ?

Exercice 5. Les graphiques ci-dessous représentent les résidus de modèles de *rlm*. Pour chacun d'entre eux, dire si les hypothèses standards du modèle de *rlm* semblent être validées.

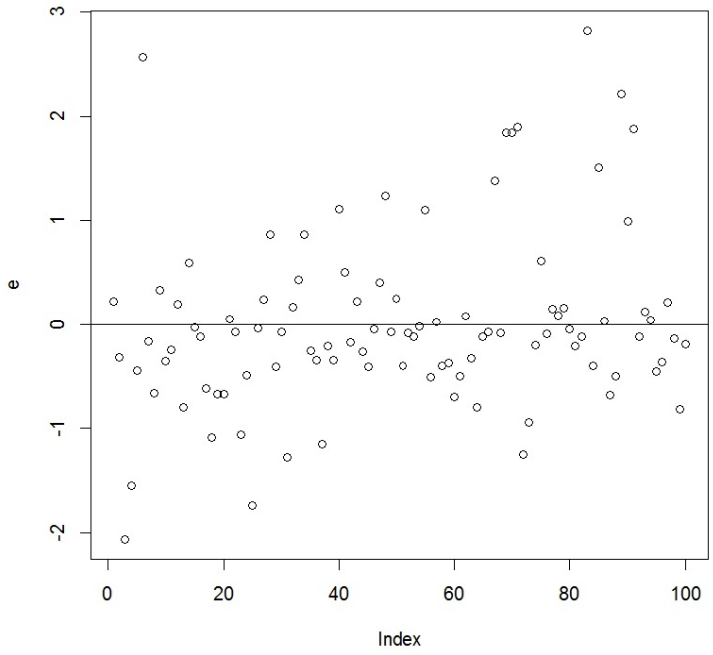
Graphique des résidus 1 :



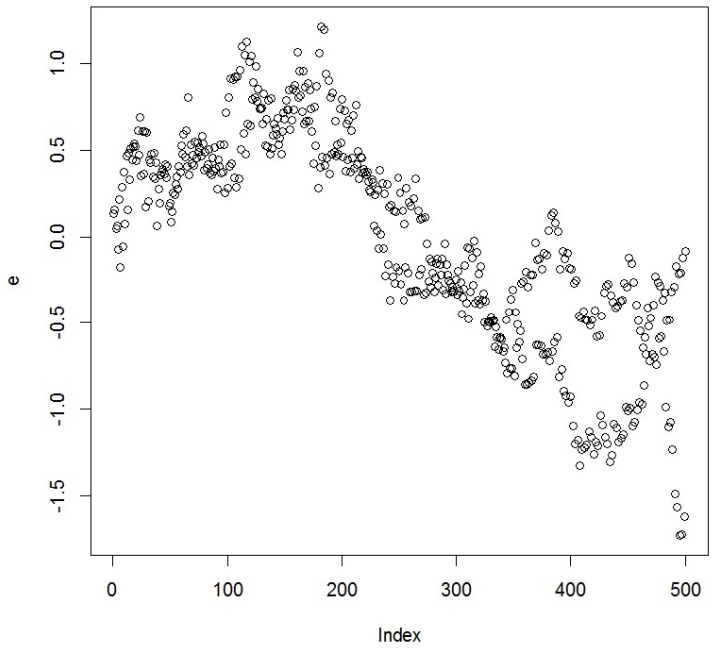
Graphique des résidus 2 :



Graphique des résidus 3 :



Graphique des résidus 4 :



Exercice 6. On souhaite expliquer une variable quantitative Y à partir d'une variable quantitative $X1$. Pour ce faire, on considère le modèle de *rls* :

$$Y = \beta_0 + \beta_1 X1 + \epsilon,$$

où $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres β_0, β_1 et σ sont des réels inconnus. On les estime alors à l'aide de n observations de $(Y, X1)$ par la méthode des *mco*. Le tableau de ce modèle de *rls* renvoyé par la commande `summary` est donné ci-dessous :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	•	19.4194	16.15	0.0000
X1	30.5364	2.0103	•	0.0000

```
Residual standard error:  35.12 on 17 degrees of freedom
Multiple R-squared:  0.9314, Adjusted R-squared:  •
F-statistic:  • on 1 and 17 DF, p-value:  2.535e-11
```

Les points • représentent des informations volontairement effacées.

1. Quelle est la valeur de n ?
2. Donner une estimation ponctuelle de la valeur prédite de Y lorsque $X = 2$.
3. Est-ce que la régression est significative ? Si oui, est-elle "seulement" significative ?
4. Peut-on affirmer que $\beta_1 \neq 30.40$ au risque 5% ?
5. Donner un intervalle de confiance pour β_0 au niveau 95%.
6. Donner la valeur du R^2 ajusté.
7. Donner une estimation ponctuelle de σ^2.
8. Donner la valeur du f_{obs} du test global de Fisher.
9. Retrouver les résultats numériques précédents avec des commandes R adéquates, le jeu de données étant disponible ici :

`https://chesneau.users.lmno.cnrs.fr/scores.txt`

Exercice 7. On souhaite expliquer une variable quantitative Y à partir de 10 variables quantitatives $X1, \ldots, X10$. Pour ce faire, on considère le modèle de *rlm* :

$$Y = \beta_0 + \beta_1 X1 + \ldots + \beta_{10} X10 + \epsilon,$$

où $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Les paramètres $\beta_0, \beta_1, \ldots, \beta_{10}$ et σ sont des réels inconnus. On les estime alors avec n observations de $(Y, X1, \ldots, X10)$ par la méthode des *mco*. Le tableau de ce modèle de *rlm* renvoyé par la commande `summary` est donné ci-dessous :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	•	•	4.14	0.0005	***
X1	-0.0033	0.0011	•	0.0061	**
X2	-0.0427	0.0149	-2.87	0.0092	**
X3	0.0497	0.0678	0.73	0.4722	
X4	-0.5389	•	•	0.1709	
X5	0.1362	0.0707	1.93	0.0676	.
X6	-0.4224	•	-0.39	0.7004	
X7	0.0459	0.6801	0.07	0.9468	
X8	•	0.1520	-0.25	0.8041	
X9	-0.3626	0.5593	-0.65	•	
X10	-0.5980	0.4966	-1.20	0.2419	

```
Residual standard error:  0.5537 on 21 degrees of freedom
Multiple R-squared:  0.6903, Adjusted R-squared:  •
F-statistic:  • on 10 and 21 DF, p-value:  •
```

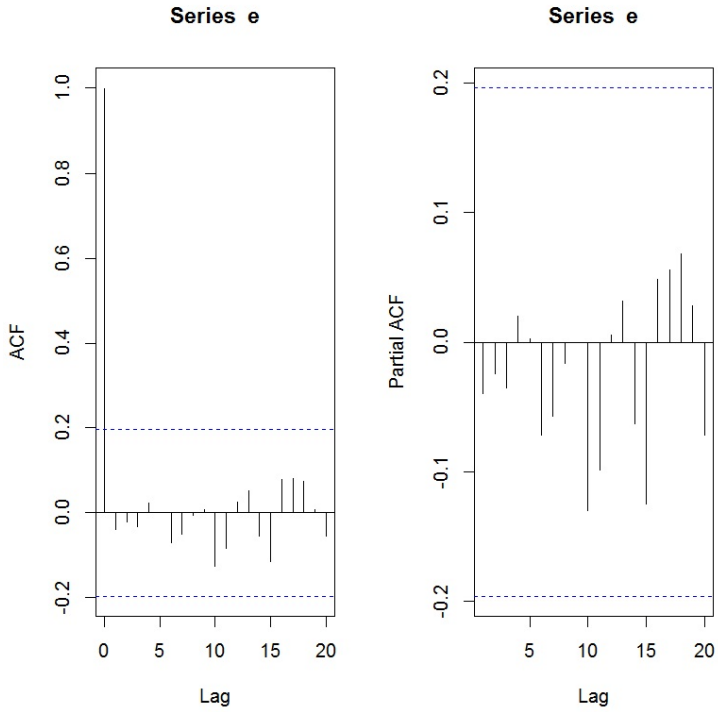
Les points • représentent des informations volontairement effacées.

1. Quelle est la valeur de n ?
2. Donner une estimation ponctuelle de β_8.
3. Est-ce que la régression est hautement significative en $X2$?
4. Donner un intervalle de confiance pour β_5 au niveau 95%.
5. Peut-on affirmer que $\beta_{10} \neq -0.51$ au risque 5% ?
6. Donner la valeur du R^2 ajusté.
7. Donner une estimation ponctuelle de σ^2.
8. Donner une estimation ponctuelle de l'écart-type de $\widehat{\beta}_6$ (*emco* de β_6).
9. Donner les valeurs du f_{obs} et de la p-valeur du test global de Fisher. Quelle est l'hypothèse nulle associée à ce test statistique ?
10. Retrouver les résultats numériques précédents avec des commandes R adéquates, le jeu de données étant disponible ici :

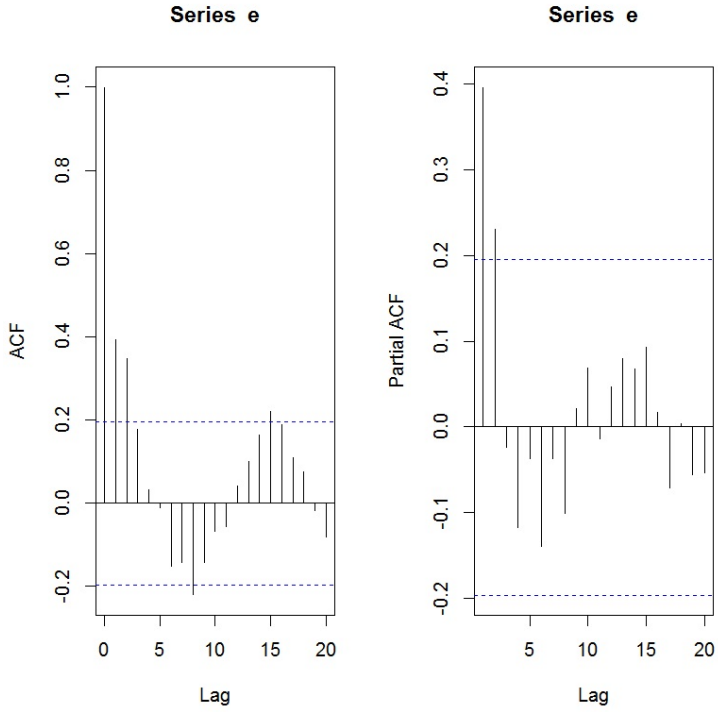
`https://chesneau.users.lmno.cnrs.fr/chenilles.txt`

Exercice 8. Les graphiques ci-dessous représentent le corrélogramme et le corrélogramme partiel des résidus de modèles de *rlm*. Commentez les.

Acf et pacf 1 :



Acf et pacf 2 :

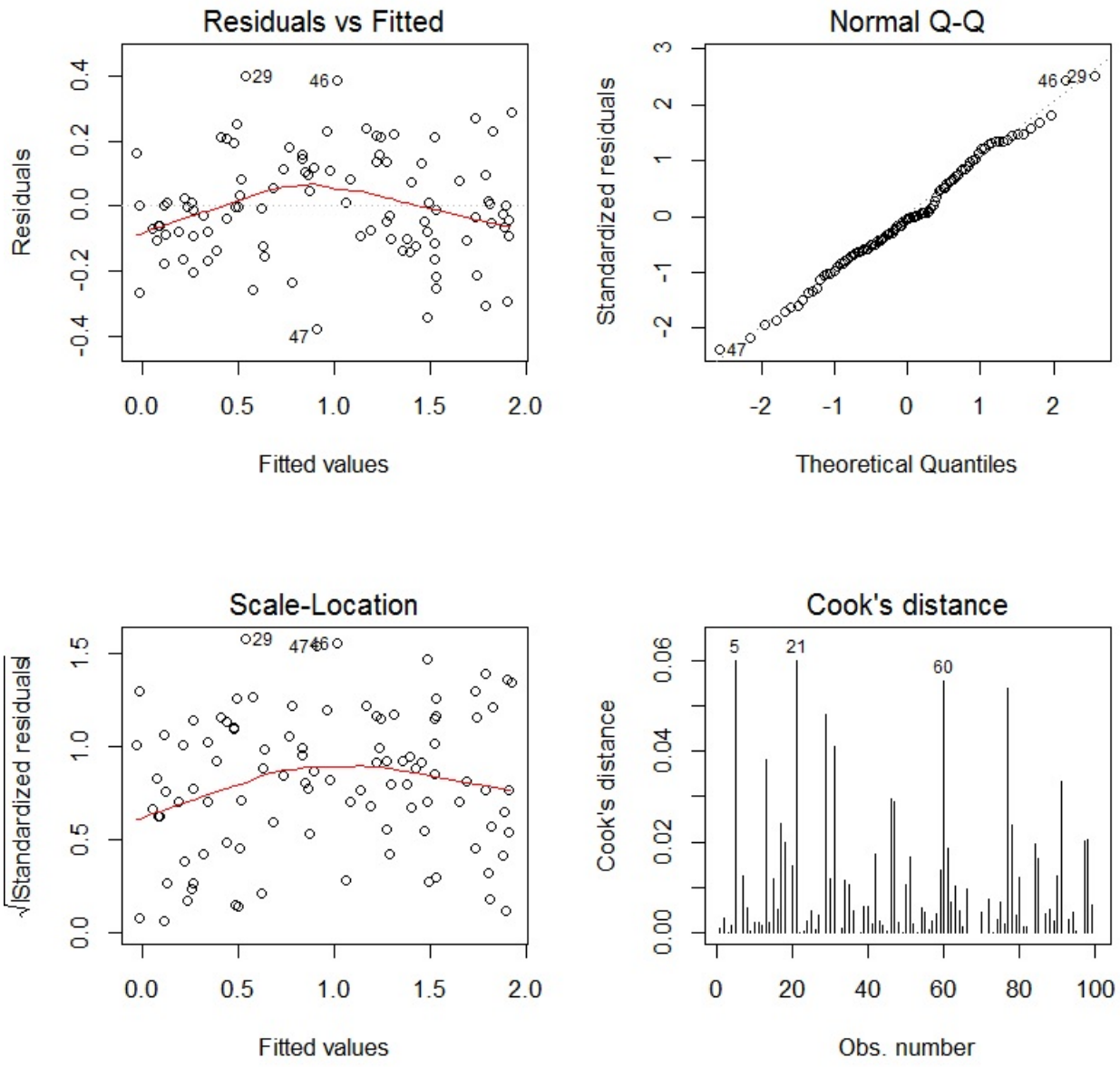


Exercice 9. On désigne par `reg` un modèle de *rlm*. Les graphiques ci-dessous sont les résultats des commandes R suivantes :

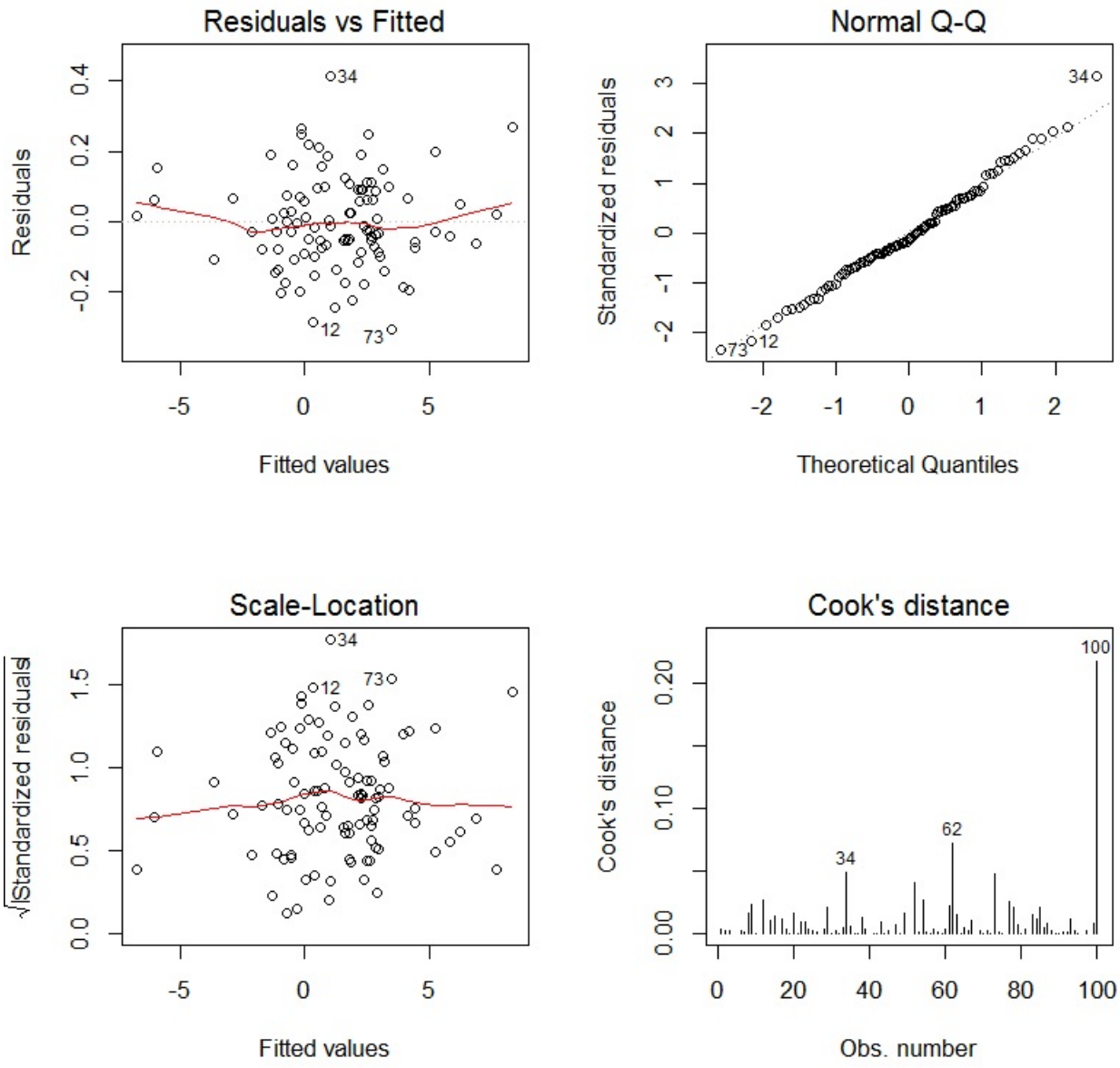
```
par(mfrow = c(2, 2))
plot(reg, 1:4)
```

Pour chacun d'entre eux, dire si les hypothèses standards du modèle de *rlm* semblent être validées.

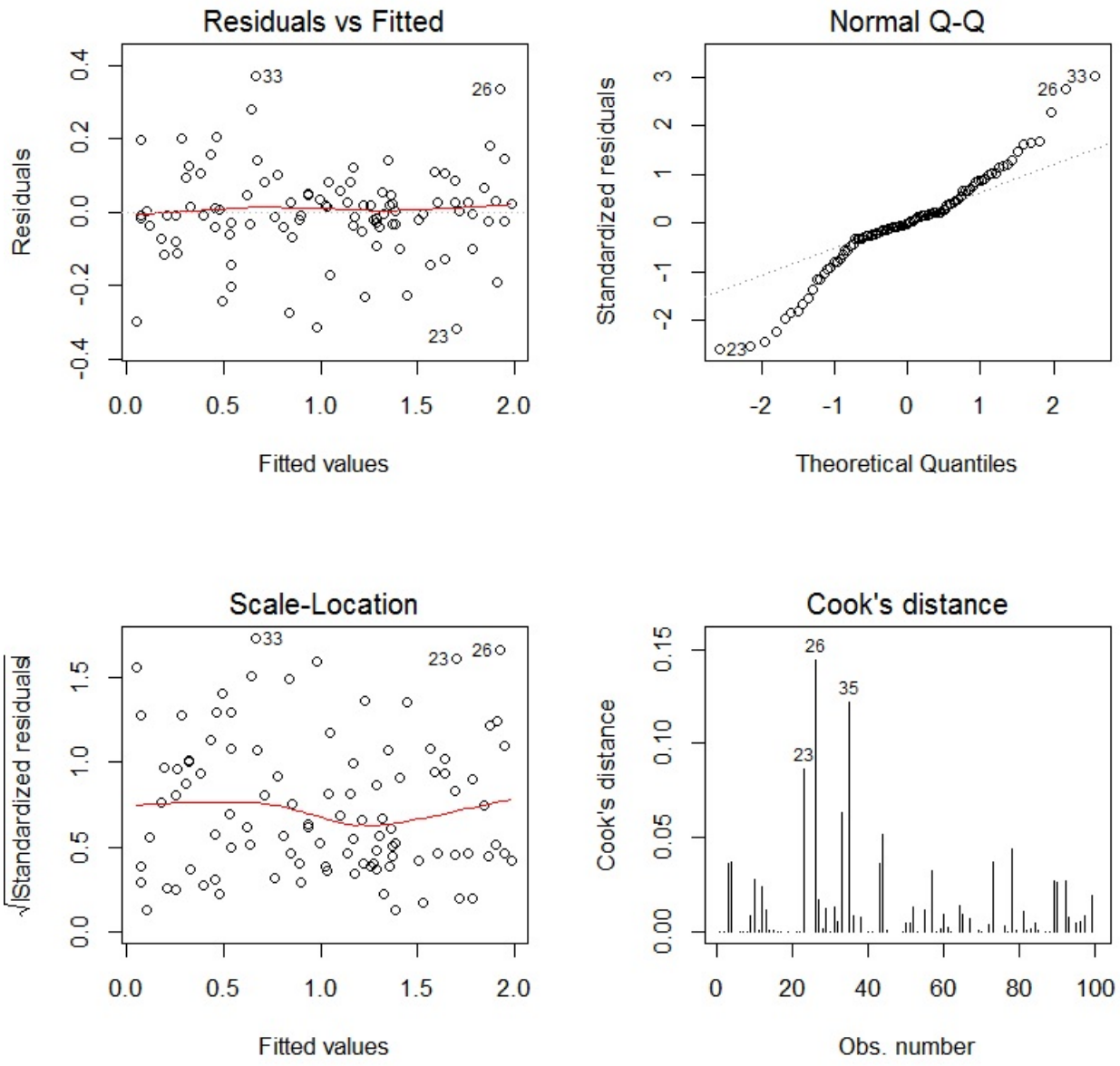
Graphique 1 :



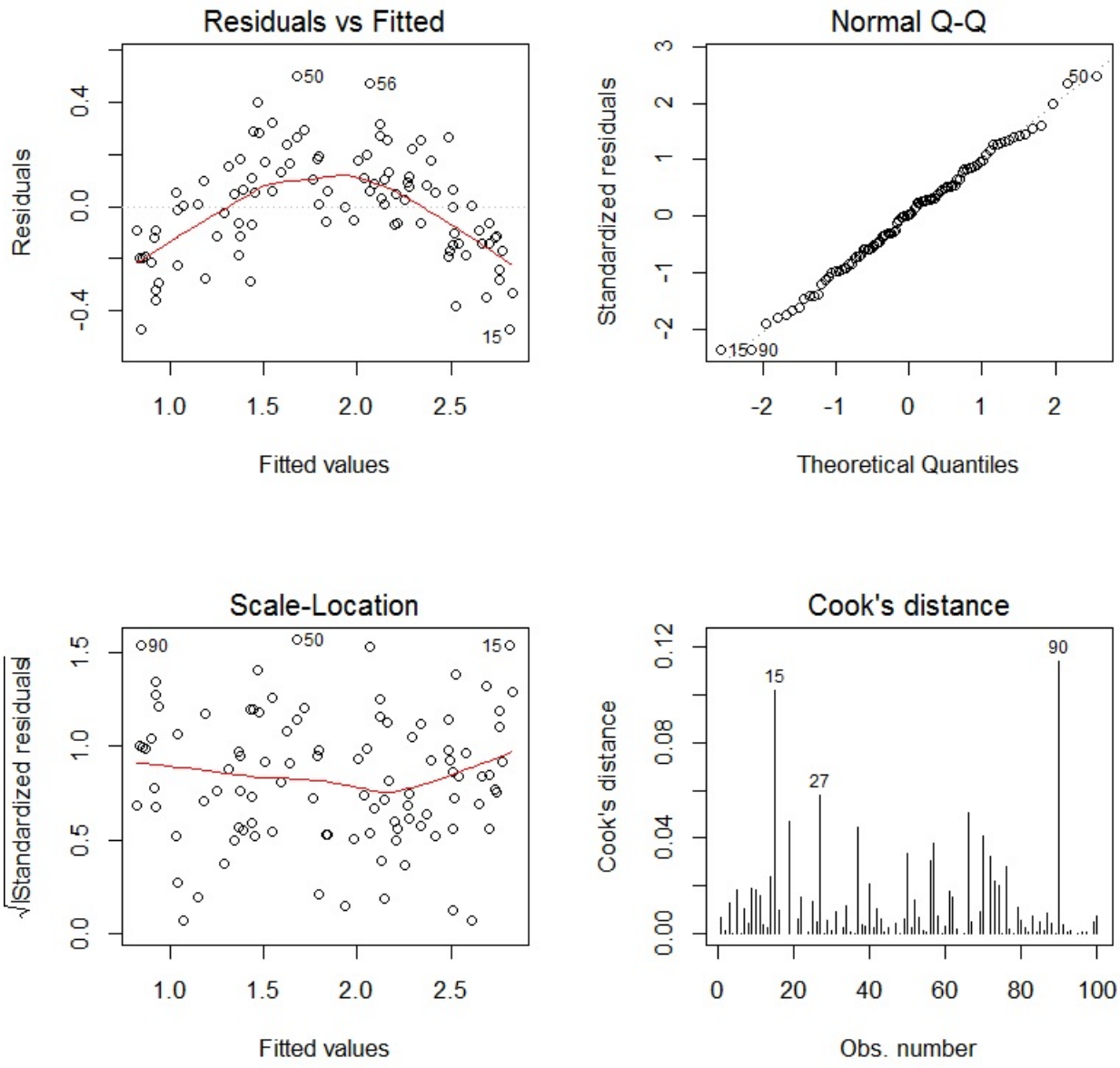
Graphique 2 :


Residuals vs Fitted
Residuals
Fitted values
34
12
73
Normal Q-Q
Standardized residuals
Theoretical Quantiles
34
73
12
Scale-Location
√|Standardized residuals|
Fitted values
34
12
73
Cook's distance
Cook's distance
Obs. number
100
62
34

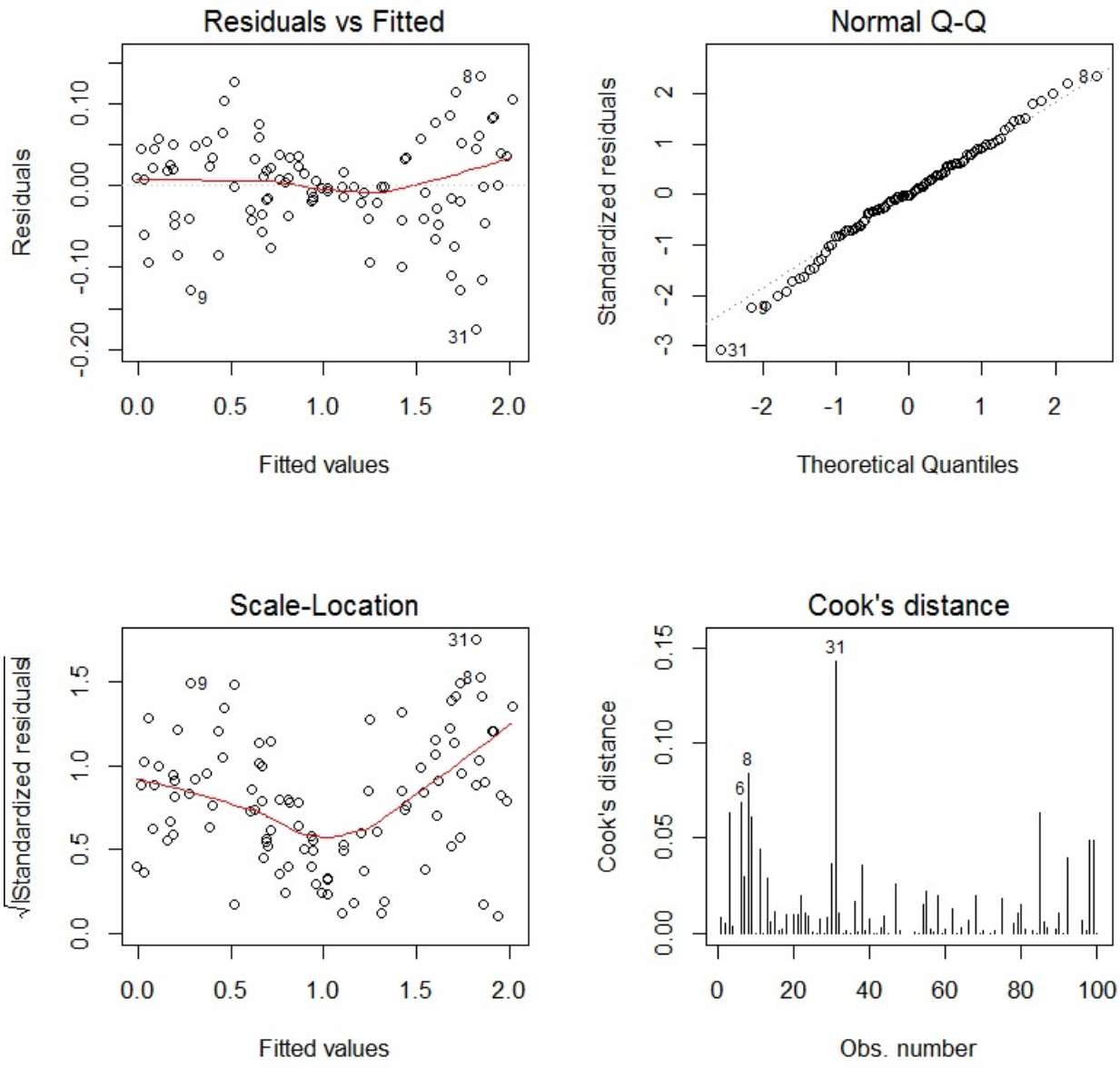
Graphique 3 :


Residuals vs Fitted
Residuals
Fitted values
33
26
23
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. number
35

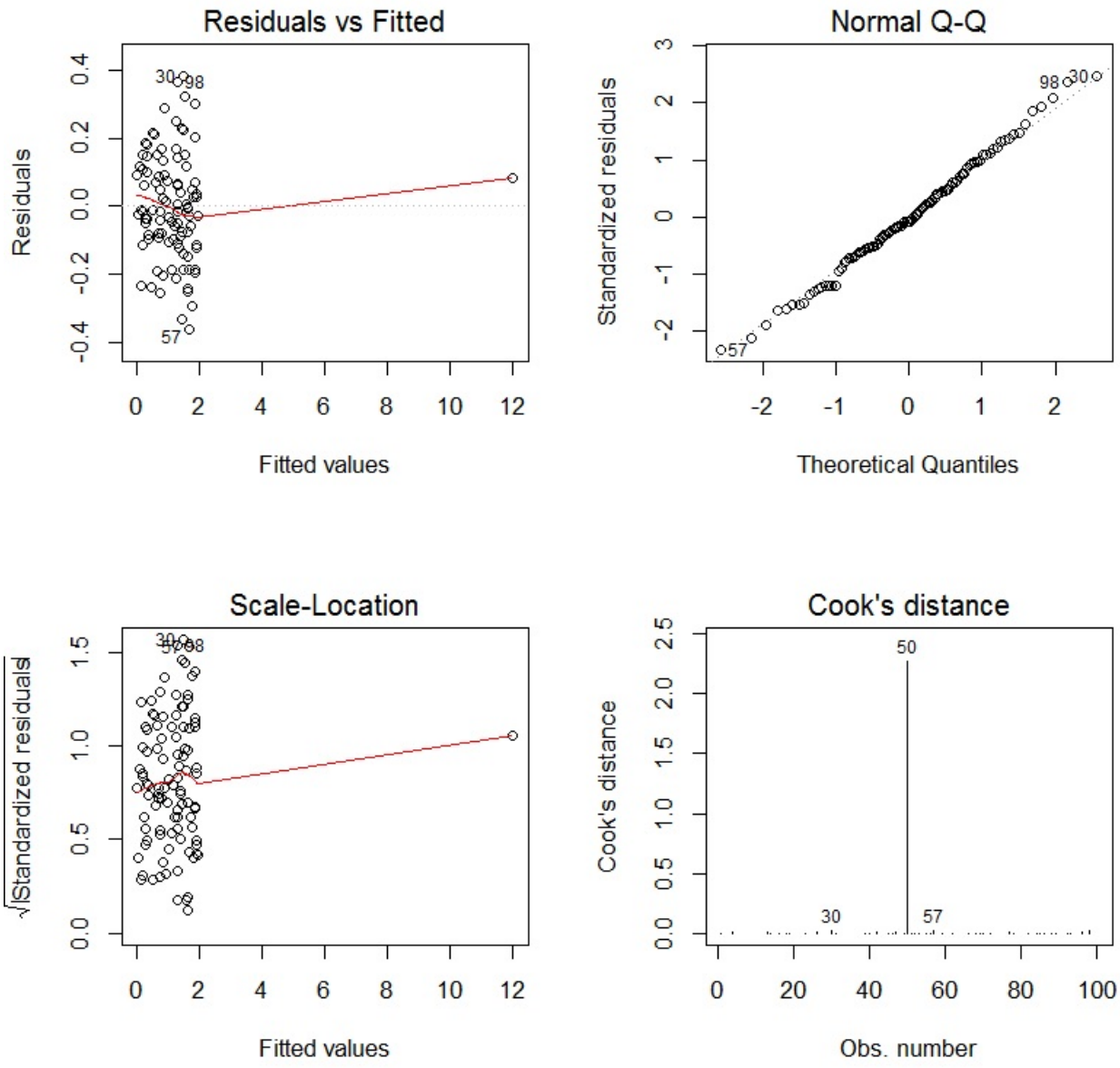
Graphique 4 :


Residuals vs Fitted
Residuals
Fitted values
50
56
15
Normal Q-Q
Standardized residuals
Theoretical Quantiles
50
15
90
Scale-Location
√|Standardized residuals|
Fitted values
90
50
15
Cook's distance
Cook's distance
Obs. number
90
15
27

Graphique 5 :


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. number

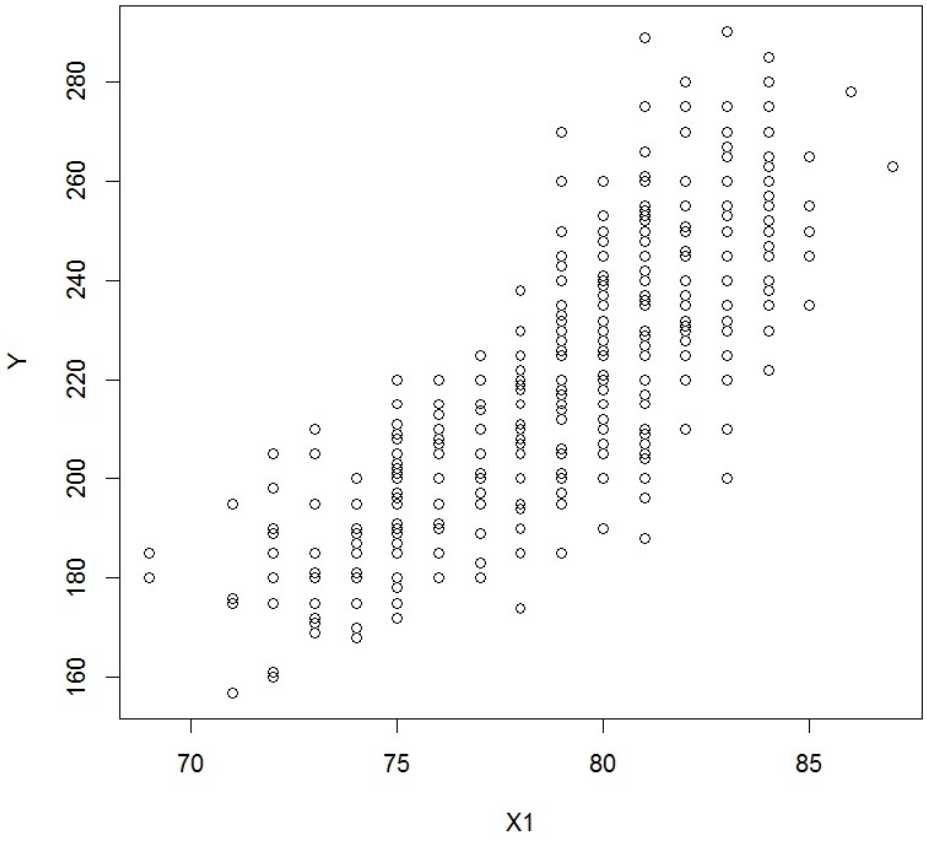
Graphique 6 :


Residuals vs Fitted
Residuals
Fitted values
30
98
57
Normal Q-Q
Standardized residuals
Theoretical Quantiles
98
30
57
Scale-Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. number
50
30
57

Exercice 10. On souhaite expliquer le poids d'un basketteur professionnel de la NBA (variable Y) à partir de sa taille (variable $X1$).

1. Décrire brièvement l'enjeu des commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/nba.txt",
header = T, sep = ",")
attach(w)
plot(X1, Y)
```



2. On exécute les commandes R suivantes :

```
reg = lm(Y ~ X1)
summary(reg)
```

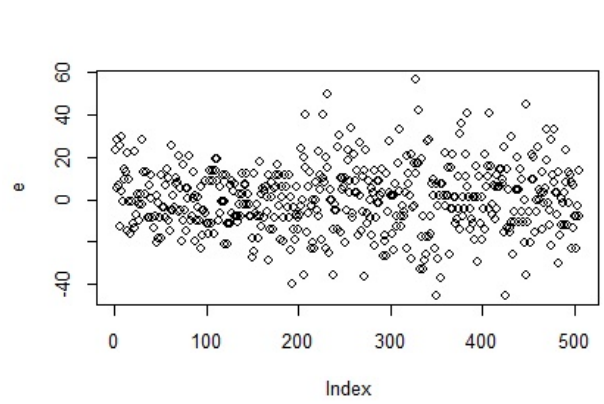
Cela renvoie :

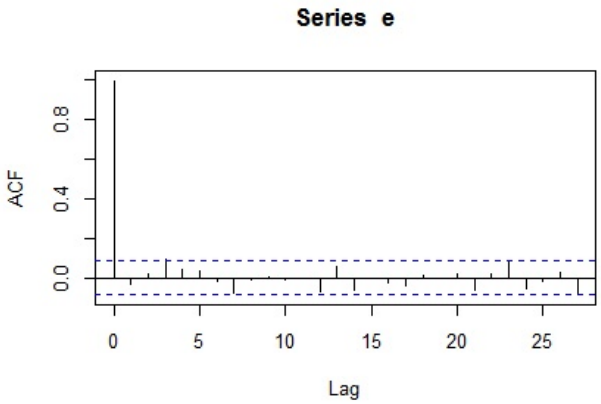
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-279.8693	15.5512	-18.00	0.0000	***
X1	6.3307	0.1965	32.22	0.0000	***

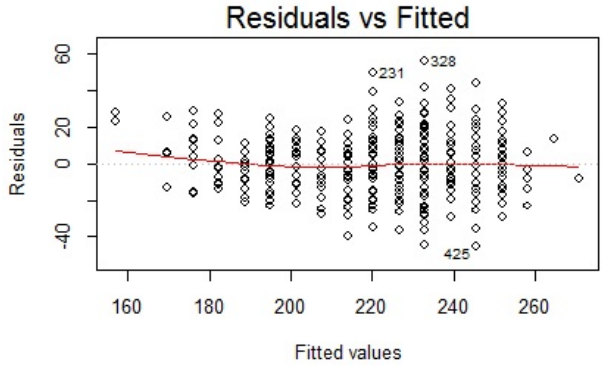
```
Residual standard error:  15.24 on 503 degrees of freedom
Multiple R-squared:  0.6736, Adjusted R-squared:  0.6729
F-statistic:  1038 on 1 and 503 DF, p-value:  < 2.2e-16
```

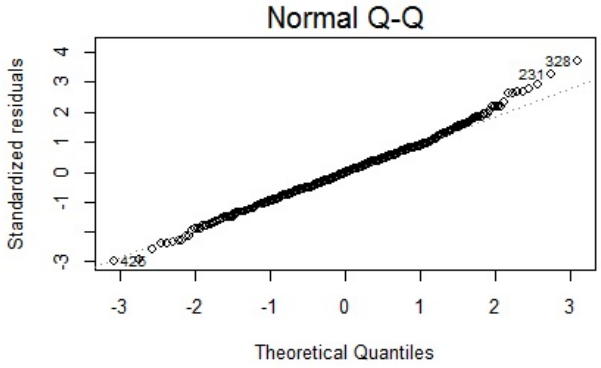
Puis :

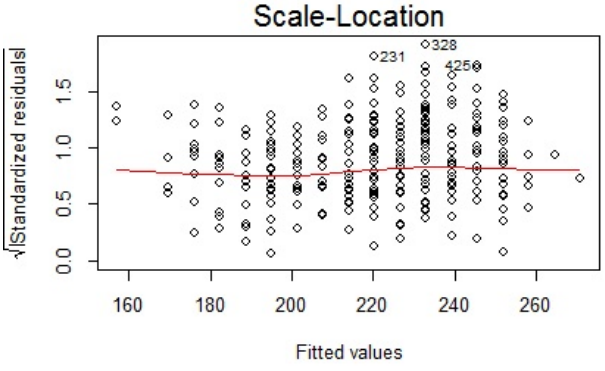
```
e = residuals(reg)
par(mfrow = c(3, 2))
plot(e)
acf(e)
plot(reg, 1:4)
```

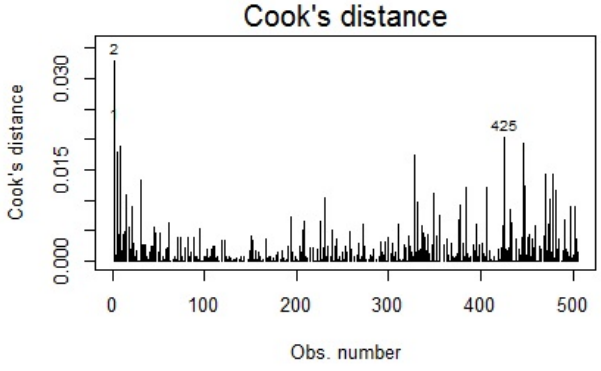












Quel point demande une analyse plus poussée ? Quel test statistique utiliseriez-vous pour cela ?

3. On exécute les commandes R suivantes :

```
shapiro.test(e)
```

Cela renvoie : p-valeur = 0.08593. Est-ce que cela est réjouissant ?

4. Décrire brièvement l'enjeu des commandes R suivantes :

```
library(car)
reg2 = powerTransform(reg)
reg2
```

Cela renvoie :

```
Estimated transformation parameters
-0.1916357
```

On continue :

```
reg3 = lm(bcPower(Y, coef(reg2)) ~ X1)
summary(reg3)
```

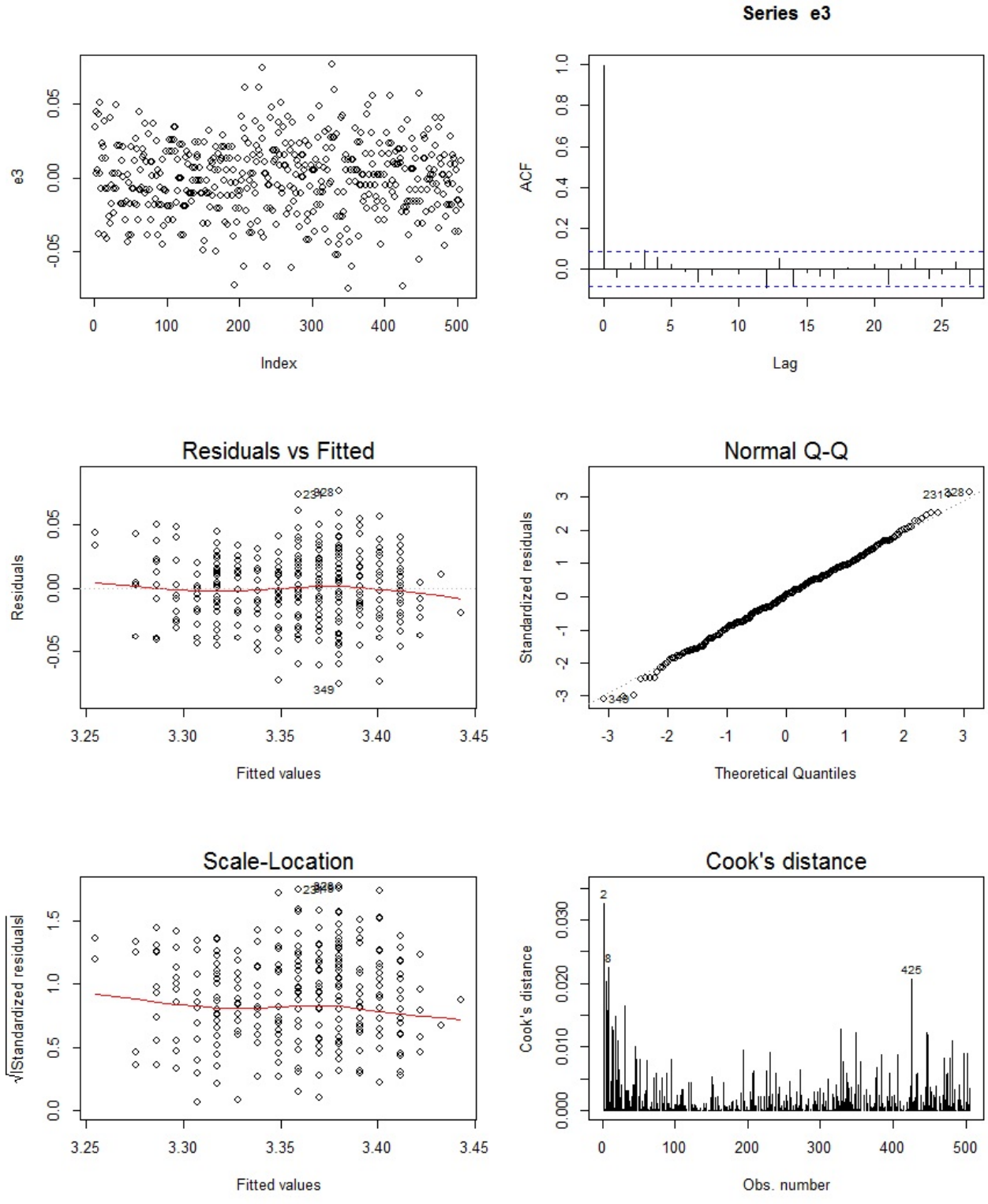
Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.5344	0.0248	102.29	0.0000	***
X1	0.0104	0.0003	33.35	0.0000	***

```
Residual standard error:  0.02428 on 503 degrees of freedom
Multiple R-squared:  0.6886, Adjusted R-squared:  0.688
F-statistic:  1112 on 1 and 503 DF, p-value:  < 2.2e-16
```

5. On exécute les commandes R suivantes :

```
e3 = residuals(reg3)
par(mfrow = c(3, 2))
plot(e3)
acf(e3)
plot(reg3, 1:4)
```



Puis :

```
shapiro.test(e3)
```

Cela renvoie : p-valeur = 0.5881.

Est-ce que tout est satisfaisant ?

6. Est-ce que le modèle de *rlm* avec transformation logarithmique de Y :

$$\log(Y) = \beta_0 + \beta_1 X1 + \epsilon,$$

aurait été aussi un choix judicieux pour améliorer le modèle initial quant à la normalité des variables d'erreurs ?

Pour appuyer les réponses, on exécute les commandes R suivantes :

```
reg = lm(log(Y) ~ X1)
summary(reg)
```

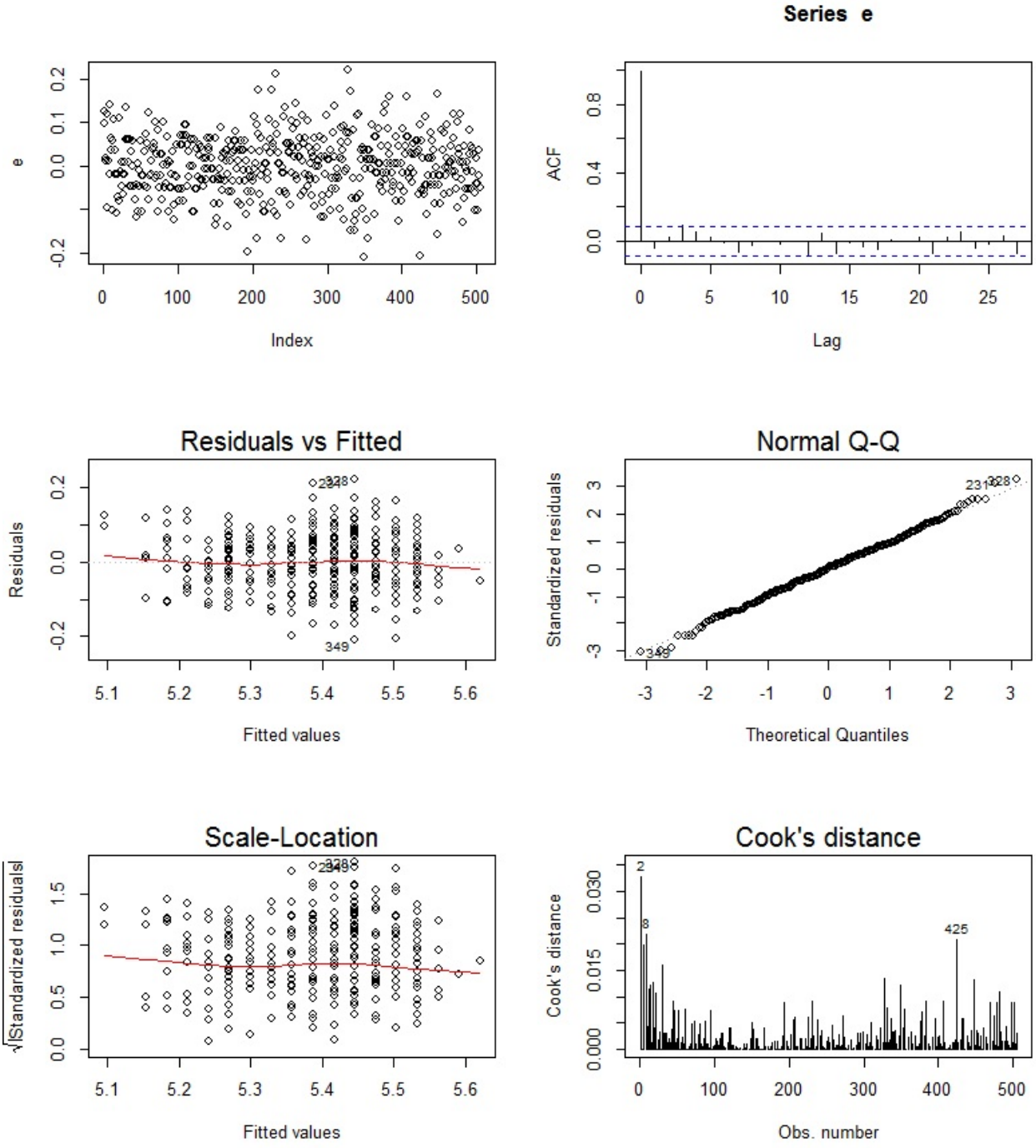
Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.0781	0.0696	44.21	0.0000	***
X1	0.0292	0.0009	33.22	0.0000	***

```
Residual standard error:  0.06823 on 503 degrees of freedom
Multiple R-squared:  0.6869, Adjusted R-squared:  0.6863
F-statistic:  1104 on 1 and 503 DF, p-value:  < 2.2e-16
```

Puis :

```
e = residuals(reg)
par(mfrow = c(3, 2))
plot(e)
acf(e)
plot(reg, 1:4)
```



De plus, on exécute :

```
shapiro.test(e)
```

Cela renvoie : p-valeur = 0.679.

Exercice 11. On souhaite expliquer le poids d'un basketteur professionnel de la NBA à partir de plusieurs autres caractères. Ainsi, pour 505 basketteurs de la NBA, on dispose :

- de leur poids (variable Y),
- de leur taille (variable $X1$),
- de leur rôle sur le terrain (variable $X2$ qualitative à 3 modalités : G, F et C),
- de leur âge (variable $X3$).

On souhaite expliquer Y à partir de $X1$, $X2$ et $X3$.

Pour se donner une idée plus précise des données, on exécute les commandes R suivantes :

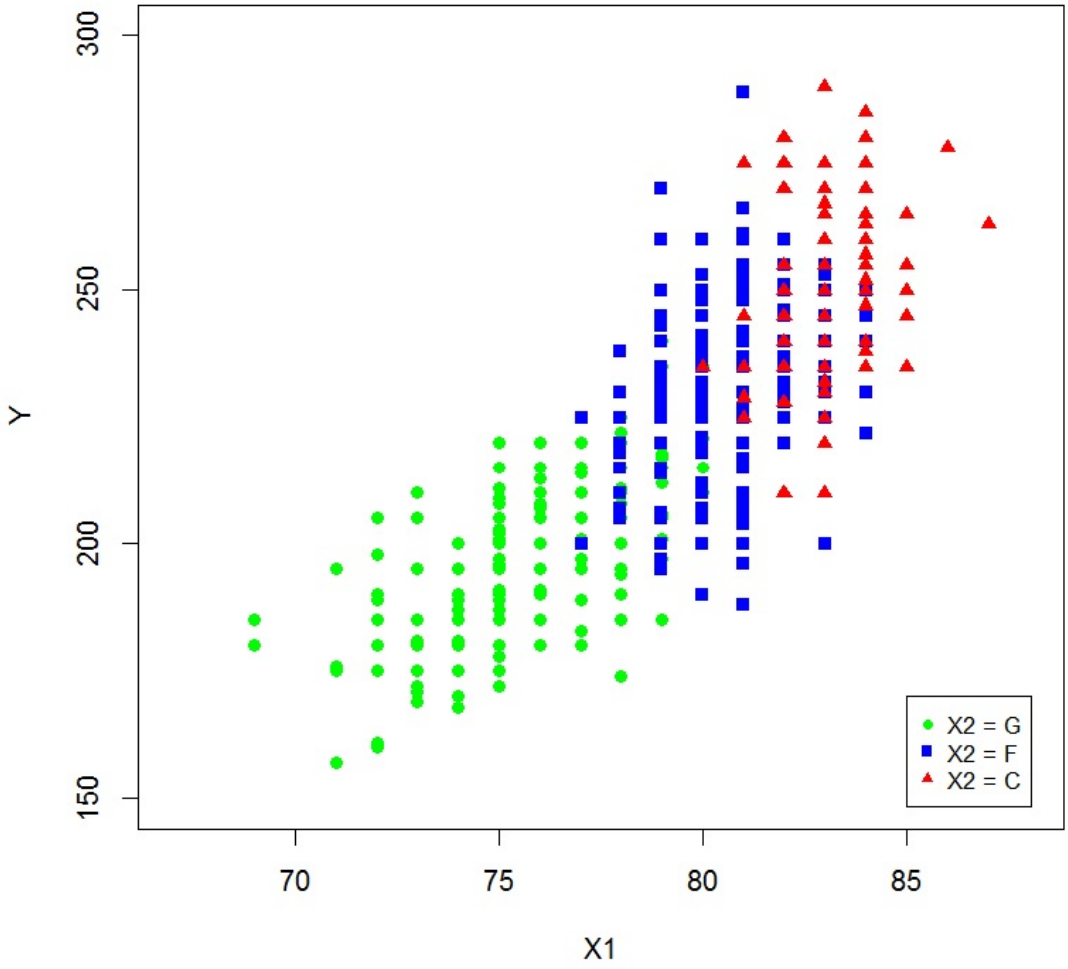
```
w = read.table("https://chesneau.users.lmno.cnrs.fr/nba.txt",
header = T, sep = ",")
attach(w)
head(w)
```

Cela renvoie :

	Joueur	X2	X1	Y	X3
1	Nate Robinson	G	69	180	29
2	Isaiah Thomas	G	69	185	24
3	Phil Pressey	G	71	175	22
4	Shane Larkin	G	71	176	20
5	Ty Lawson	G	71	195	25
6	John Lucas III	G	71	157	30

Puis :

```
plot(X1[X2 == "G"], Y[X2 == "G"], pch = 16, ylab = "Y", xlab = "X1",
xlim = c(67, 88), ylim = c(150, 300), col = "green")
points(X1[X2 == "F"], Y[X2 == "F"], pch = 15, col = "blue")
points(X1[X2 == "C"], Y[X2 == "C"], pch = 17, col = "red")
legend(x = 85, y = 170, c("X2 = G", "X2 = F", "X2 = C"), cex = 0.8, col
= c("green", "blue", "red"), pch = c(16, 15, 17))
```



On exécute les commandes R suivantes :

```
reg = lm(log(Y) ~ X1 + X2 + X3)
summary(reg)
```

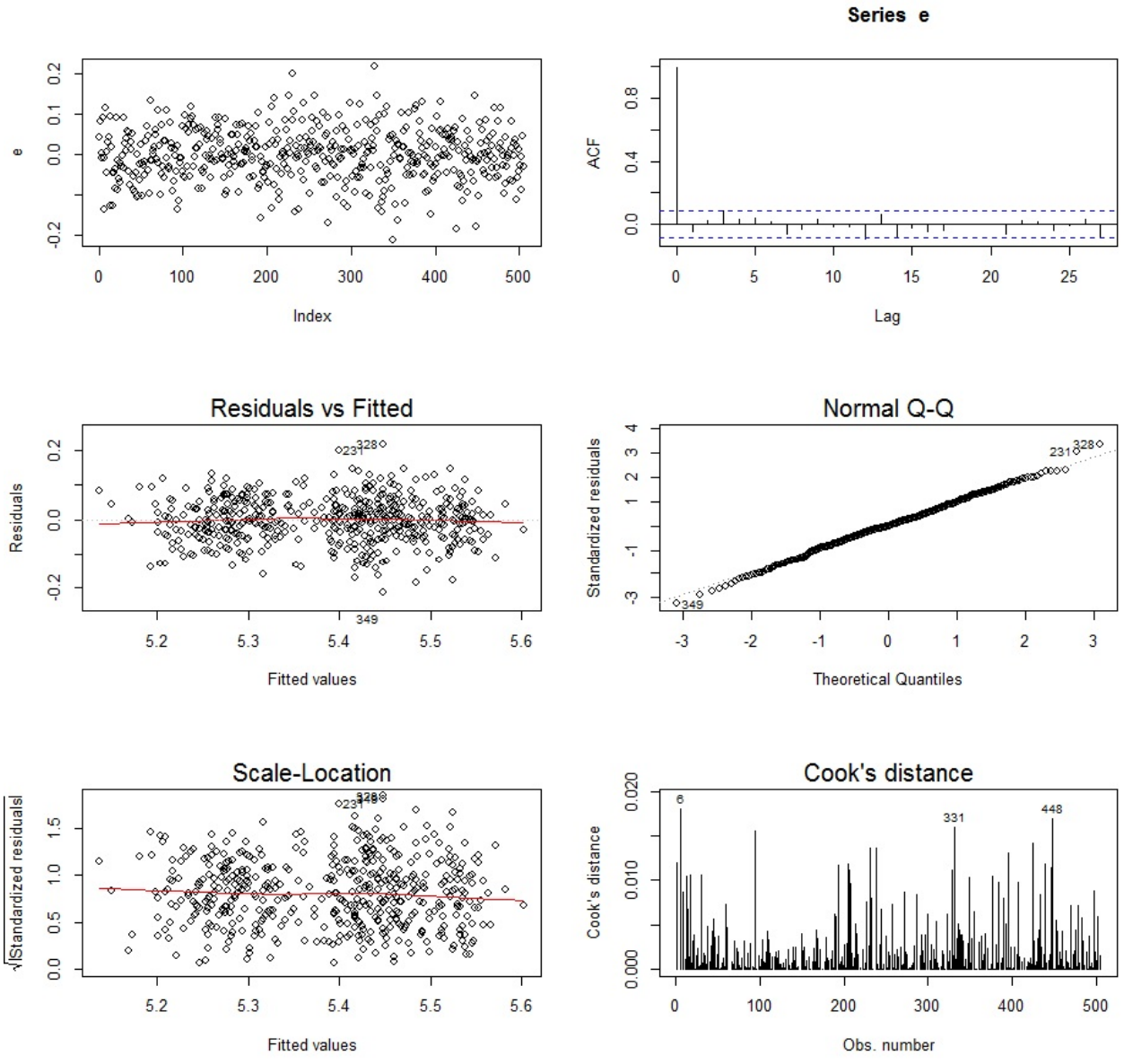
Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.7482	0.1399	26.79	0.0000	***
X1	0.0206	0.0016	12.49	0.0000	***
X2F	-0.0336	0.0092	-3.67	0.0003	***
X2G	-0.0900	0.0149	-6.03	0.0000	***
X3	0.0025	0.0007	3.71	0.0002	***

```
Residual standard error:  0.06484 on 500 degrees of freedom
Multiple R-squared:  0.7189, Adjusted R-squared:  0.7167
F-statistic:  319.7 on 4 and 500 DF, p-value:  < 2.2e-16
```

Puis :

```
e = residuals(reg)
par(mfrow = c(3, 2))
plot(e)
acf(e)
plot(reg, 1:4)
```



1. Est-ce que le modèle considéré est performant ? Est-ce que les hypothèses standards semblent vérifiées ? Justifier vos réponses.

2. Proposer des idées pour éventuellement améliorer le modèle.

Exercice 12. On dispose d'un jeu de données issues d'expériences chirurgicales. On souhaite expliquer une variable quantitative Y à partir de 5 variables quantitatives $X1, X2, X3, X4, X5$ et 3 variables qualitatives binaires $X6, X7, X8$.

1. Décrire brièvement l'enjeu des commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/chirurgie.txt",
header = F, col.names = c("X1", "X2", "X3", "X4", "X5", "X6", "X7",
"X8", "Y"))
attach(w)
head(w)
```

Cela renvoie :

	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	6.70	62	81	2.59	50	0	1	0	695
2	5.10	59	66	1.70	39	0	0	0	403
3	7.40	57	83	2.16	55	0	0	0	710
4	6.50	73	41	2.01	48	0	0	0	349
5	7.80	65	115	4.30	45	0	0	1	2343
6	5.80	38	72	1.42	65	1	1	0	348

2. Décrire brièvement l'enjeu des commandes R suivantes :

```
X6 = as.factor(X6)
X7 = as.factor(X7)
X8 = as.factor(X8)
```

Est-ce que ces commandes sont vraiment indispensables ?

3. On exécute les commandes R suivantes :

```
reg = lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8)
summary(reg)
```

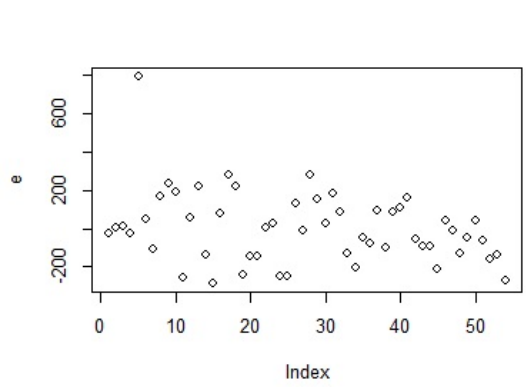
Cela renvoie :

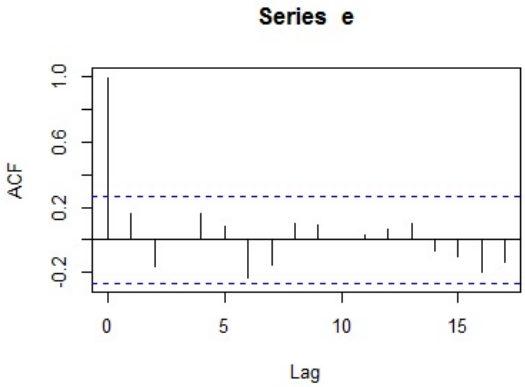
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1148.8230	242.3279	-4.74	0.0000	***
X1	62.3904	24.4697	2.55	0.0143	*
X2	8.9731	1.8741	4.79	0.0000	***
X3	9.8881	1.7417	5.68	0.0000	***
X4	50.4127	44.9593	1.12	0.2681	
X5	-0.9510	2.6486	-0.36	0.7212	
X61	15.8743	58.4746	0.27	0.7873	
X71	7.7131	64.9564	0.12	0.9060	
X81	320.6969	85.0701	3.77	0.0005	***

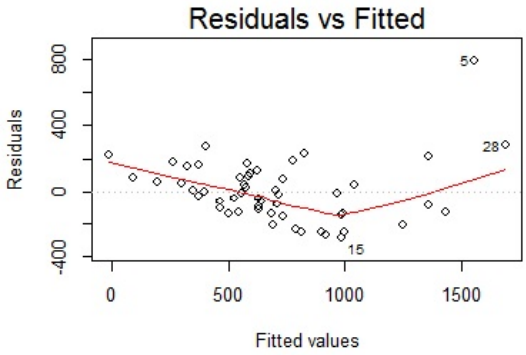
```
Residual standard error:  201.4 on 45 degrees of freedom
Multiple R-squared:  0.7818, Adjusted R-squared:  0.7431
F-statistic:  20.16 on 8 and 45 DF, p-value:  1.607e-12
```

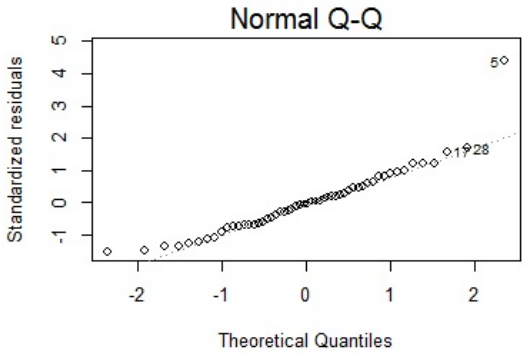
Puis :

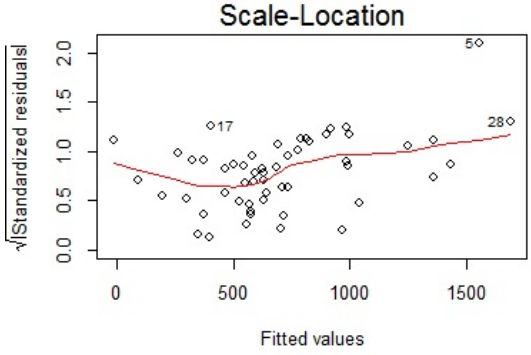
```
e = residuals(reg)
par(mfrow = c(3, 2))
plot(e)
acf(e)
plot(reg, 1:4)
```

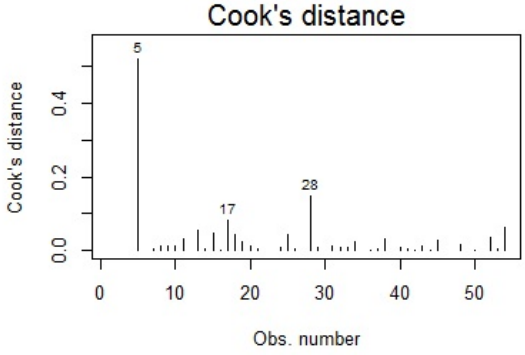












Est-ce que tout semble satisfaisant ?

4. Pour améliorer les hypothèses standards, on propose un nouveau modèle considérant une transformation logarithmique de Y :

```
reg2 = lm(log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8)
summary(reg2)
```

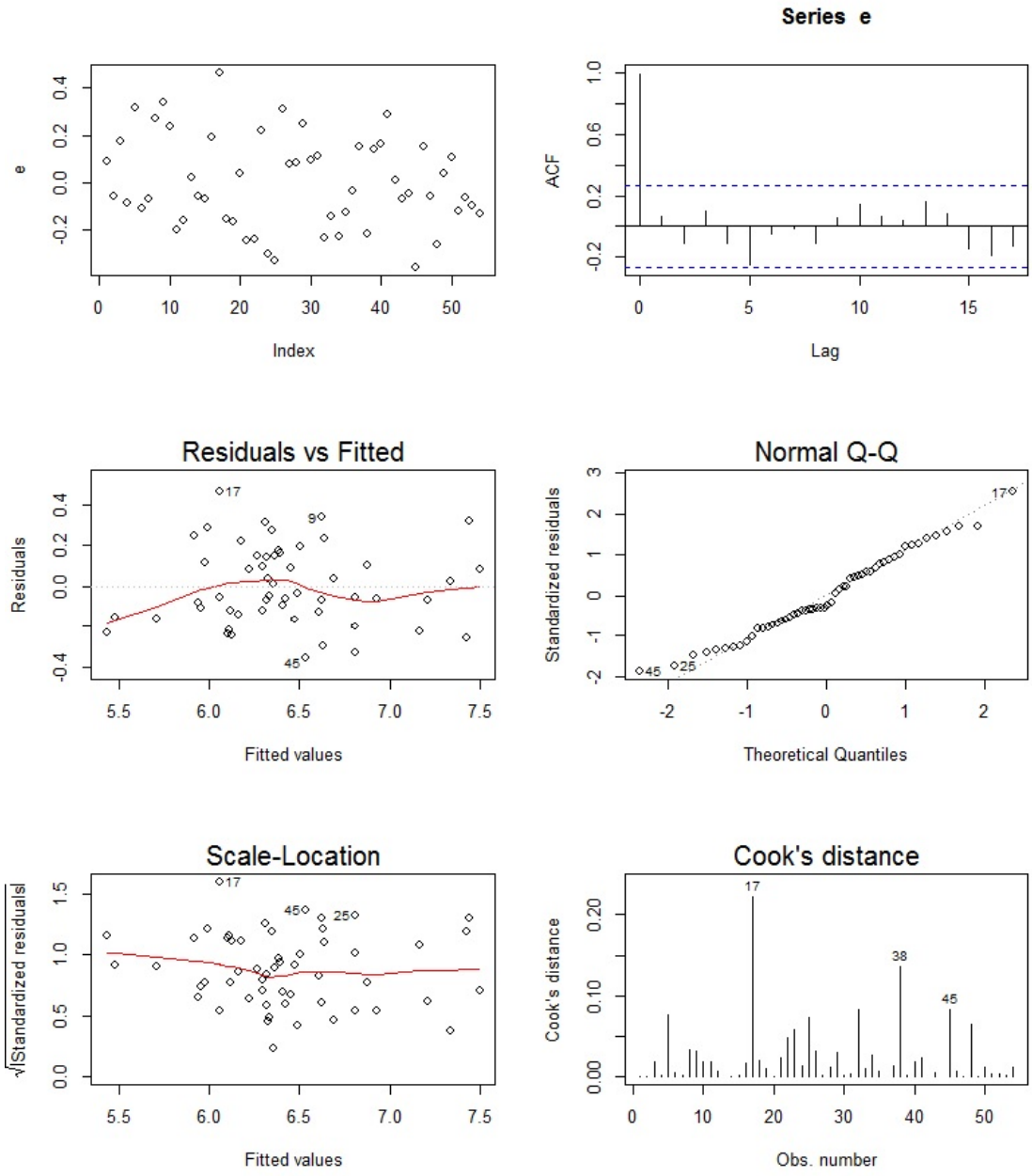
Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.0509	0.2517	16.09	0.0000	***
X1	0.0686	0.0254	2.70	0.0098	**
X2	0.0135	0.0019	6.91	0.0000	***
X3	0.0149	0.0018	8.26	0.0000	***
X4	0.0079	0.0467	0.17	0.8659	
X5	-0.0036	0.0028	-1.30	0.2014	
X61	0.0842	0.0607	1.39	0.1728	
X71	0.0573	0.0675	0.85	0.4002	
X81	0.3882	0.0884	4.39	0.0001	***

```
Residual standard error:  0.2093 on 45 degrees of freedom
Multiple R-squared:  0.8461, Adjusted R-squared:  0.8187
F-statistic:  30.93 on 8 and 45 DF, p-value:  7.823e-16
```

Puis :

```
e = residuals(reg2)
par(mfrow = c(3, 2))
plot(e)
acf(e)
plot(reg2, 1:4)
```



Est-ce que les points problématiques du modèle initial se sont améliorés ? Est-ce que le nouveau modèle est meilleur que l'ancien ?

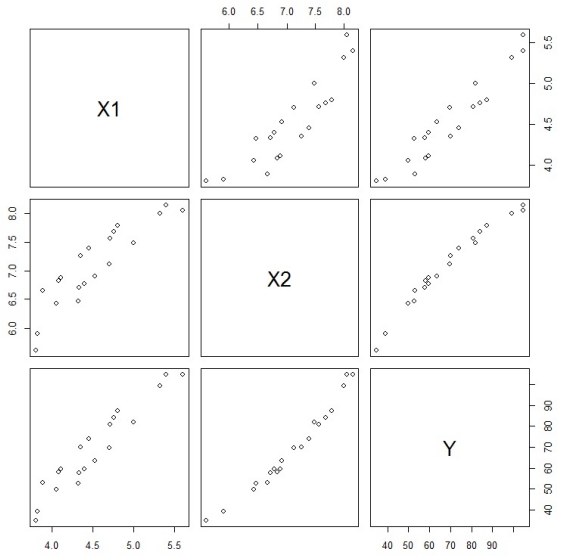
Exercice 13. On cherche à expliquer le poids d'un œuf en fonction de sa hauteur et de sa largeur. Pour 20 œufs, on dispose :

- du poids (variable Y),
- de la largeur (variable $X1$),
- de la hauteur (variable $X2$).

On exécute les commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/oeufs.txt", header =
T)
attach(w)
pairs(w)
```

Cela renvoie :



Puis on considère une modélisation simple avec le modèle de *rlm* :

```
reg = lm(Y ~ X1 + X2)
summary(reg)
```

Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-129.5199	4.6078	-28.11	0.0000	***
X1	14.8336	1.9684	7.54	0.0000	***
X2	18.5823	1.4684	12.66	0.0000	***

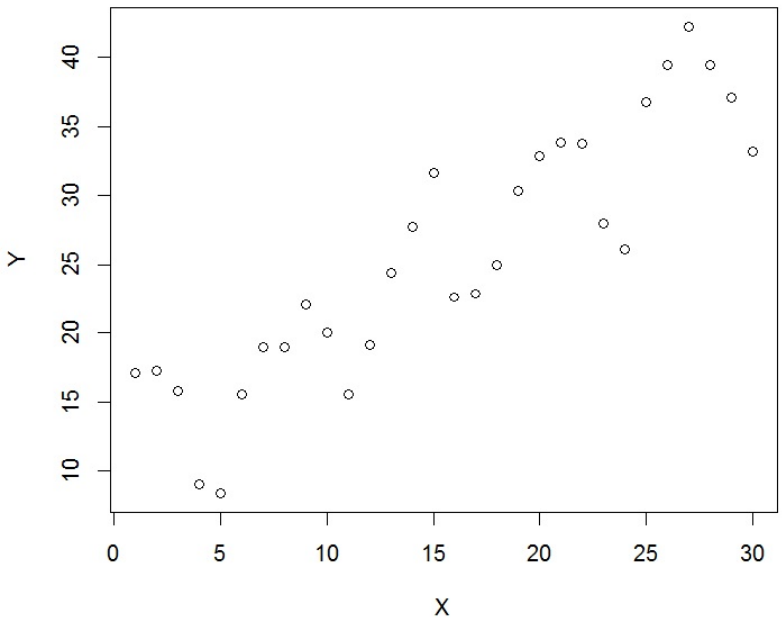
```
Residual standard error:  1.956 on 17 degrees of freedom
Multiple R-squared:  0.9916, Adjusted R-squared:  0.9907
F-statistic:  1009 on 2 and 17 DF, p-value:  < 2.2e-16
```

Que pensez-vous de la modélisation proposée ?

Exercice 14. On dispose d'un jeu de données à partir duquel on souhaite expliquer une variable quantitative Y à partir d'une variable quantitative X.

1. Décrire brièvement l'enjeu des commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/xp1.txt",
header = T)
attach(w)
plot(X, Y)
```



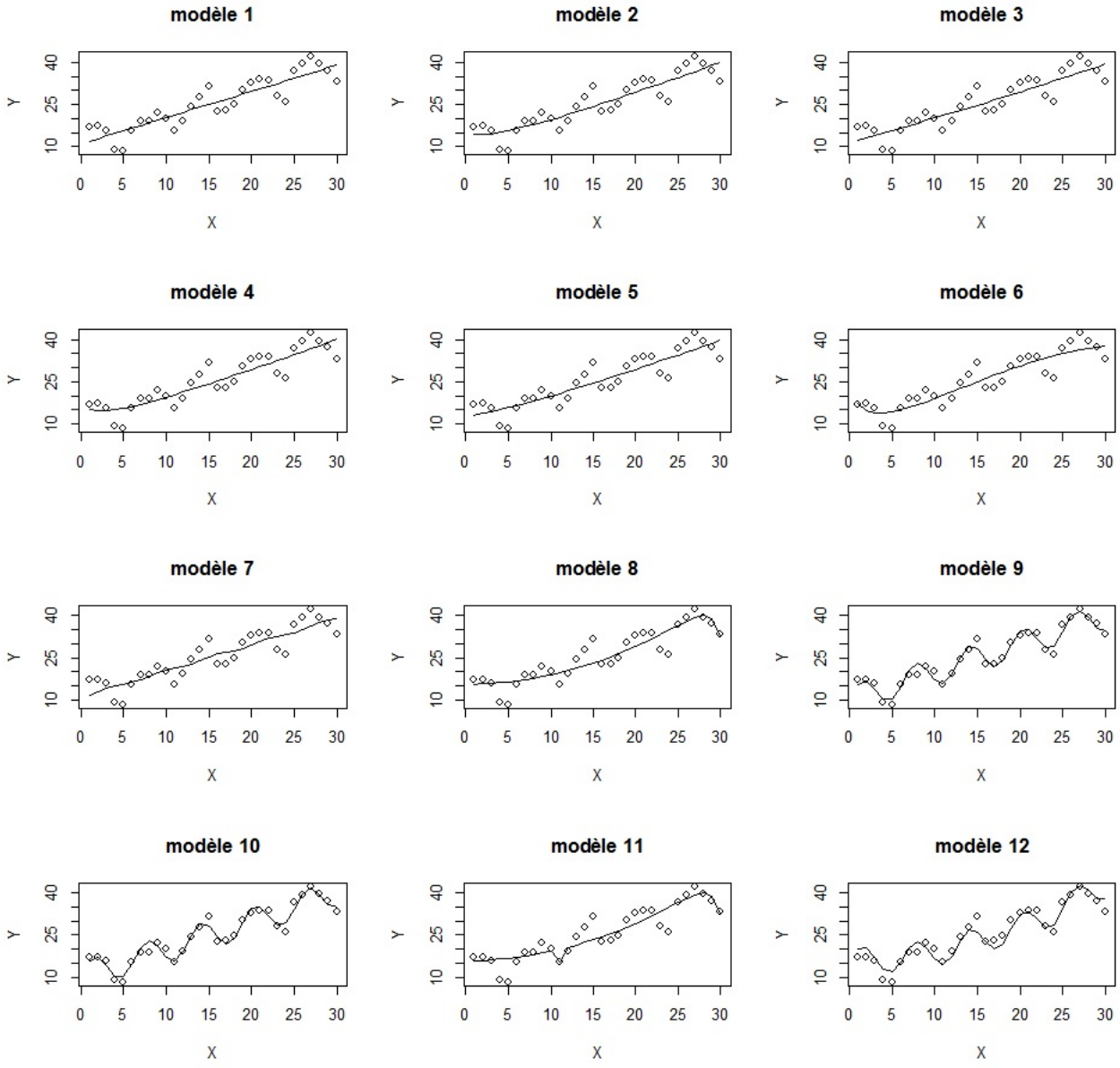
2. On étudie les 12 modèles de *rlm* suivants :

Modèle 1 : $Y = \beta_0 + \beta_1 X + \epsilon$,
Modèle 2 : $Y = \beta_0 + \beta_1 X + \beta_2 \sqrt{X} + \epsilon$,
Modèle 3 : $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$,
Modèle 4 : $Y = \beta_0 + \beta_1 X + \beta_2 \log(X) + \epsilon$,
Modèle 5 : $Y = \beta_0 + \beta_1 X + \beta_2 \sqrt{X + 10} + \epsilon$,
Modèle 6 : $Y = \beta_0 + \beta_1 X + \beta_2 \log(X) + \beta_3 X^3 + \epsilon$,
Modèle 7 : $Y = \beta_0 + \beta_1 X + \beta_2 \cos(X) + \epsilon$,
Modèle 8 : $Y = \beta_0 + \beta_1 X^2 + \beta_2 \exp(X) + \epsilon$,
Modèle 9 : $Y = \beta_0 + \beta_1 X + \beta_2 \sin(X) + \epsilon$,
Modèle 10 : $Y = \beta_0 + \beta_1 X + \beta_2 \sin(X) + \beta_3 X^2 + \epsilon$,
Modèle 11 : $Y = \beta_0 + \beta_1 X^2 + \beta_2 \tan(X) + \beta_3 \exp(X) + \epsilon$,

Modèle 12 : $Y = \beta_0 + \beta_1 X^2 + \beta_2 \cos(X) + \beta_3 \sin(X) + \epsilon$.

Pour chacun des modèles, proposer des commandes R pour estimer les coefficients de régression.

3. Pour chacun des modèles, on trace la fonction de régression estimée. Les résultats sont :



Visuellement, quels sont les trois meilleurs ajustements du nuage de points ?

4. Commenter les résultats suivants :

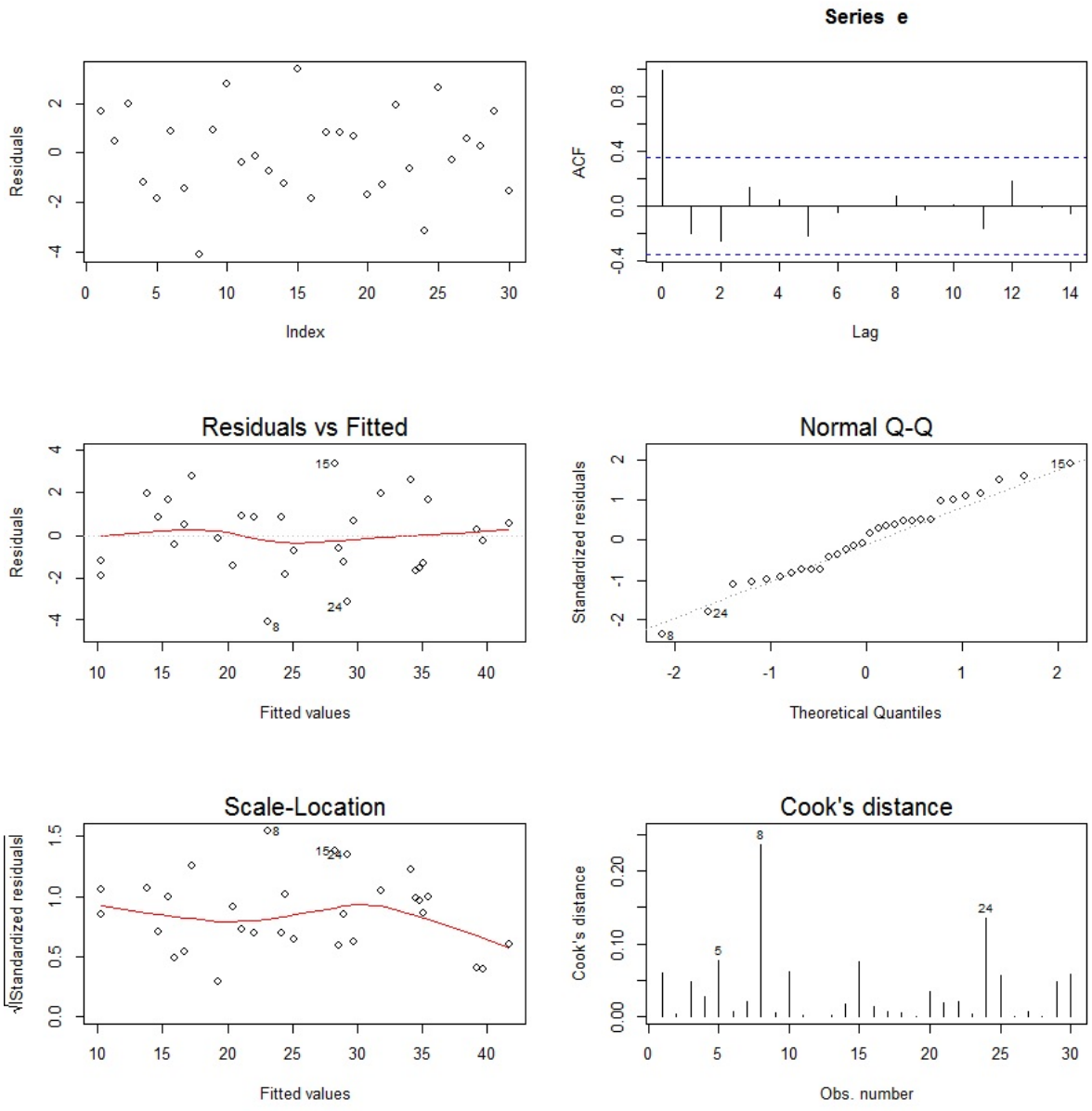
	AIC	BIC	R2 ajusté
mod1	174.4600	178.6636	0.7948047
mod2	175.4747	181.0795	0.7940801
mod3	176.3576	181.9624	0.7879303
mod4	174.8504	180.4552	0.7983214
mod5	176.1074	181.7121	0.7896918
mod6	174.4505	181.4565	0.8066660
mod7	176.3029	181.9077	0.7883165
mod8	173.7106	179.3154	0.8058402
mod9	126.2329	131.8376	0.9601114
mod10	128.0018	135.0077	0.9588951
mod11	174.3432	181.3492	0.8073560
mod12	151.9644	158.9704	0.9086332

Est-ce que cela conforte votre réponse de la question précédente ? Quel est alors le meilleur modèle selon les critères considérés ?

5. Décrire brièvement l'enjeu des commandes R suivantes :

```
reg1 = lm(Y ~ X)
reg2 = lm(Y ~ X + sqrt(X))
reg3 = lm(Y ~ X + I(X^2))
reg4 = lm(Y ~ X + log(X))
reg5 = lm(Y ~ X + sqrt(X+10))
reg6 = lm(Y ~ X + log(X) + I(X^3))
reg7 = lm(Y ~ X + cos(X))
reg8 = lm(Y ~ I(X^2) + exp(X))
reg9 = lm(Y ~ X + sin(X))
reg10 = lm(Y ~ X + sin(X) + I(X^2))
reg11 = lm(Y ~ I(X^2) + tan(X) + exp(X))
reg12 = lm(Y ~ I(X^2) + cos(X) + sin(X))
mod1=c(AIC(reg1), BIC(reg1), summary(reg1)$adj.r.squared)
mod2=c(AIC(reg2), BIC(reg2), summary(reg2)$adj.r.squared)
mod3=c(AIC(reg3), BIC(reg3), summary(reg3)$adj.r.squared)
mod4=c(AIC(reg4), BIC(reg4), summary(reg4)$adj.r.squared)
mod5=c(AIC(reg5), BIC(reg5), summary(reg5)$adj.r.squared)
mod6=c(AIC(reg6), BIC(reg6), summary(reg6)$adj.r.squared)
mod7=c(AIC(reg7), BIC(reg7), summary(reg7)$adj.r.squared)
mod8=c(AIC(reg8), BIC(reg8), summary(reg8)$adj.r.squared)
mod9=c(AIC(reg9), BIC(reg9), summary(reg9)$adj.r.squared)
mod10=c(AIC(reg10), BIC(reg10), summary(reg10)$adj.r.squared)
mod11=c(AIC(reg11), BIC(reg11), summary(reg11)$adj.r.squared)
mod12=c(AIC(reg12), BIC(reg12), summary(reg12)$adj.r.squared)
r = rbind(mod1, mod2, mod3, mod4, mod5, mod6, mod7, mod8, mod9, mod10,
mod11, mod12)
d = data.frame(r)
colnames(d) = c("AIC", "BIC", "R2 ajusté")
d
```

6. On considère le meilleur modèle. Commenter les graphiques associés suivants :



Est-ce que tout semble satisfaisant ?

Exercice 15. On dispose d'un jeu de données avec une variable à expliquer Y et 5 variables explicatives $X1$, $X2$, $X3$, $X4$ et $X5$.

1. On considère les commandes R suivantes :

```
reg1 = lm(Y ~ X1 + X2 + X3 + X4 + X5)
reg2 = lm(Y ~ X1 + X2 + X3)
anova(reg1, reg2)
```

Cela renvoie :

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	8	1.30				
2	10	1.38	-2	-0.08	0.23	0.7980

   (a) Expliciter les modèles de *rlm* que considèrent les commandes `reg1` et `reg2`.

   (b) Expliciter l'hypothèse nulle H_0 associée au test statistique considéré.

   (c) Comment interpréter le résultat du test statistique considéré ?

   (d) À partir des informations dont vous disposez, quels modèles choisiriez-vous.

2. On considère les commandes R suivantes :

```
reg1 = lm(Y ~ X1 + X2 + X3 + X4 + X5)
reg3 = lm(Y ~ X1 + X4 + X5)
anova(reg1, reg3)
```

Cela renvoie :

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	8	1.30					
2	10	5.49	-2	-4.19	12.89	0.0031	**

   (a) Expliciter les modèles de *rlm* que considèrent les commandes `reg1` et `reg3`.

   (b) Expliciter l'hypothèse nulle H_0 associée au test statistique considéré.

   (c) Comment interpréter le résultat du test statistique considéré ?

3. On considère les commandes R suivantes :

```
reg1 = lm(Y ~ X1 + X2 + X3 + X4 + X5)
drop1(reg1)
```

Cela renvoie :

	Df	Sum of Sq	RSS	AIC
<none>			1.30	-21.27
X1	1	0.12	1.42	-22.04
X2	1	1.29	2.59	-13.65
X3	1	0.59	1.89	-18.04
X4	1	0.05	1.35	-22.73
X5	1	0.00	1.30	-23.27

Comment interpréter les résultats de ce tableau ? À partir de celui-ci, expliciter le meilleur des modèles considérés suivant le critère AIC.

Exercice 16. On dispose d'un jeu de données `w`, dont les variables ont été attachées. La commande `str(w)` de R renvoie :

'data.frame': 150 obs. of 26 variables:
$ X1 : num 5.92 4.57 5.44 4.91 5.7 ...
$ X2 : num 6.51 3.64 5.5 5.85 6.33 ...
$ X3 : num 6.88 4.38 6.6 6.54 5.28 ...
$ X4 : num 5.16 5.06 4.48 5.71 5.02 ...
$ X5 : num 6.42 3.16 4.99 4.67 4.23 ...
$ X6 : num 5.83 4.6 5.52 5.26 5.77 ...
$ X7 : num 5.36 4.58 6.24 5.79 4.47 ...
$ X8 : num 6.63 4.17 5.4 5.07 4.77 ...
$ X9 : num 5.73 4.26 5.46 5.5 6.06 ...
$ X10: num 6.11 4.44 5.57 4.64 5.5 ...
$ X11: num 4.82 5.58 5.04 6.33 3.37 ...
$ X12: num 6.11 5.71 5.87 4.85 5.45 ...
$ X13: num 6.61 5.07 5.02 6.13 6.35 ...
$ X14: num 4.82 4.25 4.91 6.41 5.38 ...
$ X15: num 6.3 4.55 4.83 4.22 5.75 ...
$ X16: num 5.45 4.12 5.23 5.33 4.84 ...
$ X17: num 6.42 4.2 4.85 6.06 5.41 ...
$ X18: num 6.71 4.8 5.31 5.68 4.56 ...
$ X19: num 5.56 4.16 5.03 5.29 5.54 ...
$ X20: num 7.15 5.64 5.15 5.29 6.38 ...
$ X21: num 6.47 3.93 4.93 5.04 5.45 ...
$ X22: num 6.31 4.61 5.33 4.55 5.44 ...
$ X23: num 5.41 4.1 4.58 5.4 4.65 ...
$ X24: num 5.92 5.64 4.97 6.02 5.5 ...
$ X25: num 5.12 5.62 5.3 5.87 5.12 ...
$ Y : num 42.8 30.4 36.3 36.6 38.3 ...

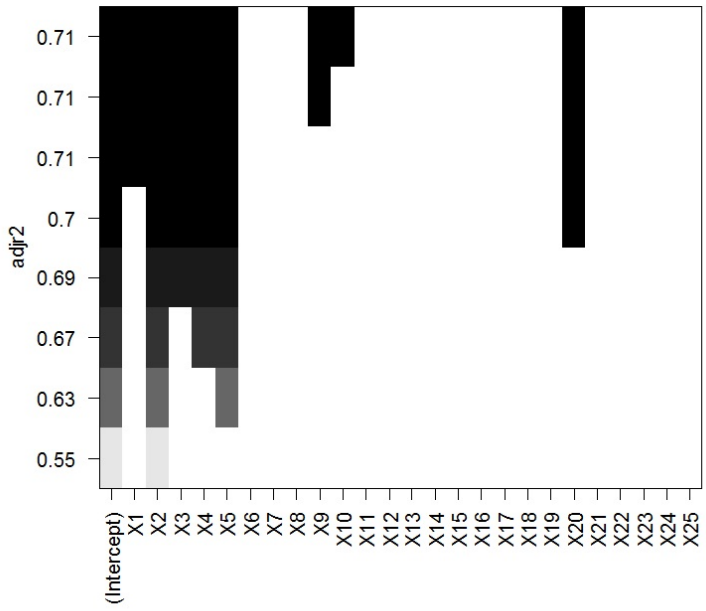
On désire expliquer la variable Y à partir des autres variables. Pour ce faire, on considère le modèle de *rlm*.

1. Combien y-a t'il de variables explicatives ? De quelle natures sont-elles ? Combien y-a t'il de données en tout ?

2. On exécute les commandes R suivantes :

```
library(leaps)
v = regsubsets(Y ~ ., w, method = "exhaustive")
plot(v, scale = "adjr2")
```

On obtient le graphique :



Expliquer l'enjeu des commandes R effectuées et interpréter le graphique obtenu. A-t-on oublié une option importante dans la commande `regsubsets` ?

3. On exécute les commandes R suivantes :

```
reg2 = lm(Y ~ X1 + X2 + X3 + X4 + X5 + X9 + X10 + X20)
summary(reg2)
```

Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	11.3475	1.3459	8.43	0.0000	***
X1	0.6858	0.2618	2.62	0.0098	**
X2	1.7217	0.2665	6.46	0.0000	***
X3	0.9497	0.2643	3.59	0.0005	***
X4	1.0511	0.2566	4.10	0.0001	***
X5	1.0510	0.2506	4.19	0.0000	***
X9	-0.4248	0.2693	-1.58	0.1169	
X10	0.4074	0.2818	1.45	0.1504	
X20	-0.7378	0.2831	-2.61	0.0101	*

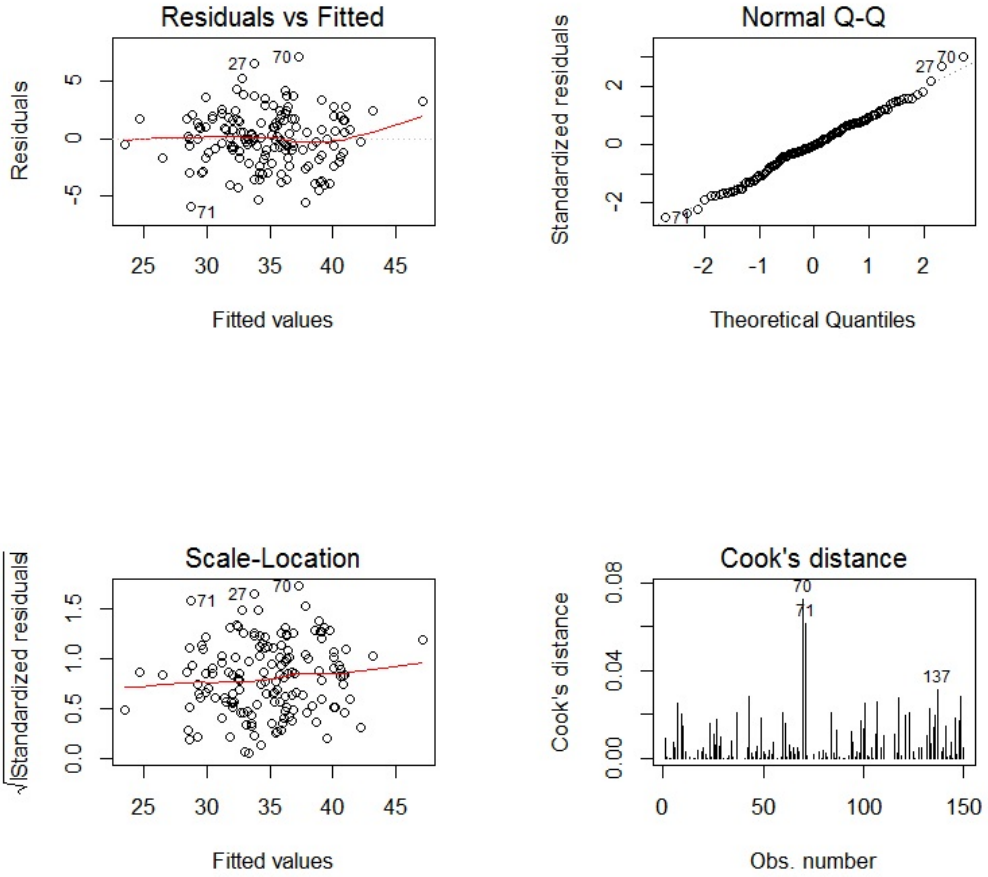
```
Residual standard error:  2.428 on 141 degrees of freedom
Multiple R-squared:  0.7281, Adjusted R-squared:  0.7127
F-statistic:  47.2 on 8 and 141 DF, p-value:  < 2.2e-16
```

Que pensez-vous de ces résultats ?

4. On exécute les commandes R suivantes :

```
par(mfrow = c(2, 2))
plot(reg2, 1:4)
```

Cela renvoie :



Est-ce que vous déceler un problème ? Expliquer votre réponse.

5. On exécute la commande R suivantes :

```
predict(reg2, data.frame(X1 = 5.1, X2 = 5.2, X3 = 4.8, X4 = 5.2,
X5 = 6.1, X9 = 5.0, X10 = 4.9, X20 = 5.6))
```

Retrouver le résultat numérique de cette commande R en utilisant votre calculette.

Exercice 17. L'étude porte sur le taux de criminalité dans différents états américains en 1960. Ainsi, pour 47 états, on dispose :

- du taux de criminalité (variable Y),
- de la durée moyenne du niveau de scolarité (fois 10) (variable $X1$),
- du budget de la police par habitant de l'état (variable $X2$),
- du nombre d'actifs pour 1000 hommes âgés de 14 à 24 ans (variable $X3$),
- du nombre d'hommes pour 1000 femmes (variable $X4$),
- du nombre d'habitants de l'état (unité 100000) (variable $X5$),
- du nombre de chômeurs pour 1000 habitants de 14 à 24 ans (variable $X6$),
- du nombre de chômeurs pour 1000 habitants de 35 à 39 ans (variable $X7$),
- du revenu médian des familles en dizaines de dollars (variable $X8$),
- du taux (sur 1000) de familles en dessous du niveau de pauvreté (variable $X9$).

On exécute les commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/crimes.txt", header
= T)
attach(w)
reg = lm(Y ~ ., w)
summary(reg)
```

Cela renvoie :

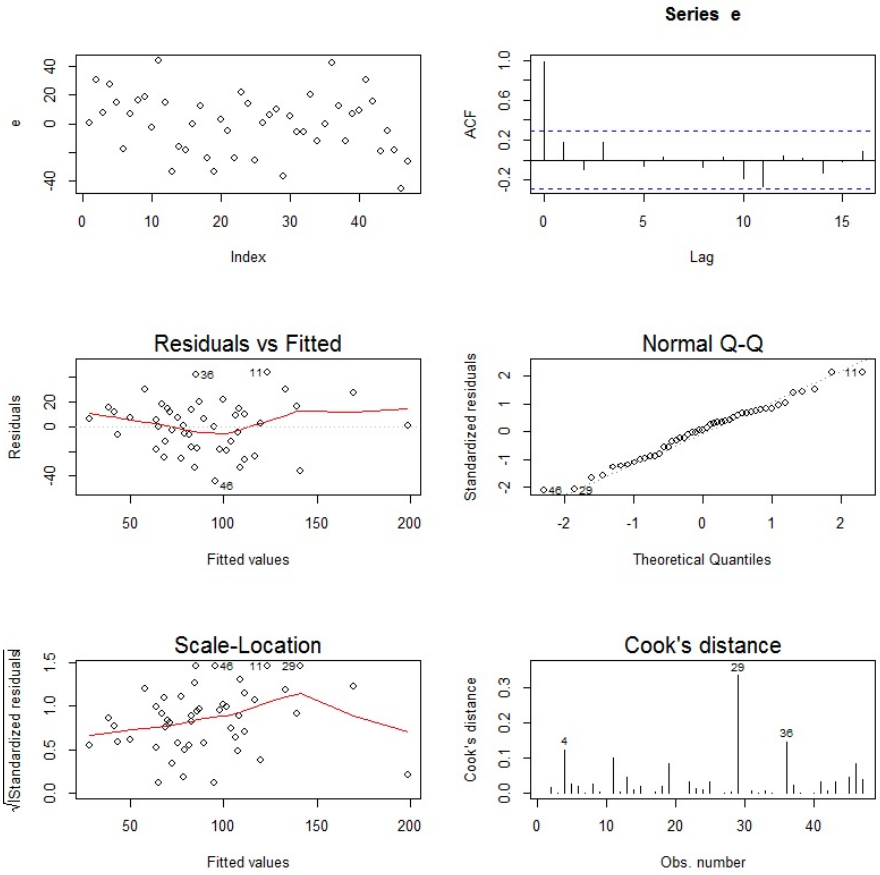
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-594.1726	154.4946	-3.85	0.0005	***
X1	1.4370	0.6483	2.22	0.0329	*
X2	1.0217	0.2334	4.38	0.0001	***
X3	-0.0323	0.1311	-0.25	0.8070	
X4	0.2869	0.2041	1.41	0.1683	
X5	-0.0416	0.1348	-0.31	0.7596	
X6	-0.6525	0.4198	-1.55	0.1286	
X7	1.4554	0.8666	1.68	0.1015	
X8	0.0823	0.1058	0.78	0.4416	
X9	0.7900	0.2207	3.58	0.0010	***

```
Residual standard error:  22.94 on 37 degrees of freedom
Multiple R-squared:  0.7171, Adjusted R-squared:  0.6482
F-statistic:  10.42 on 9 and 37 DF, p-value:  8.393e-08
```

Puis :

```
e = residuals(reg)
par(mfrow = c(3, 2))
plot(e)
acf(e)
plot(reg, 1:4)
```

Cela renvoie :

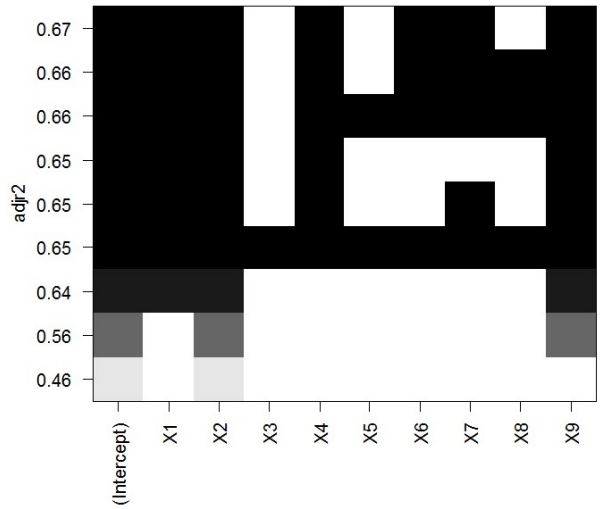


1. Est-ce que le modèle considéré est performant ? Est-ce que les hypothèses standards semblent vérifiées ? Justifier vos réponses.

2. On exécute les commandes R suivantes :

```
v = regsubsets(Y ~ ., w, method = "backward", nvmax = 9)
plot(v, scale = "adjr2")
```

On obtient :



Comment interprêter ce graphique ? Quel modèle de *rlm* s'en dégage ? Donner les commandes R associées. On le notera désormais `reg2`.

3. On considère les commandes :

```
anova(reg, reg2)[6]
```

Cela renvoie :

	Pr(>F)
1	
2	0.8734

Que peut-on en conclure ? Entre `reg` et `reg2`, quel modèle privilégeriez-vous ?

Exercice 18. On considère le jeu de données `p9.10` de la librairie `MPV`. L'étude porte sur la profondeur de l'ornière des chaussées en asphalte préparées dans des conditions différentes. On exécute les commandes R suivantes :

```
library(MPV)
w = p9.10
head(w)
```

Cela renvoie :

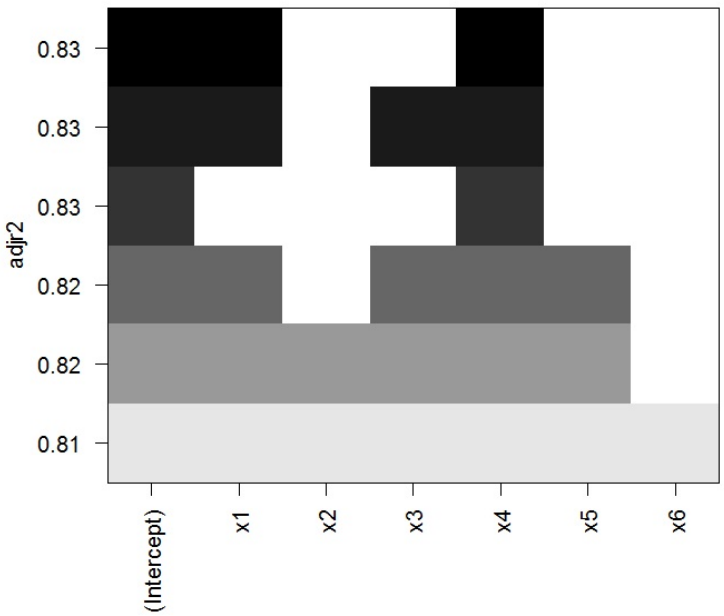
	y	x1	x2	x3	x4	x5	x6
1	0.83	0.45	4.68	4.87	-1.00	8.40	4.92
2	1.11	0.15	5.19	4.50	-1.00	6.50	4.56
3	1.17	0.15	4.82	4.73	-1.00	7.90	5.32
4	1.10	0.52	4.85	4.76	-1.00	8.30	4.87
5	0.92	0.23	4.86	4.95	-1.00	8.40	3.78
6	1.03	0.46	5.16	4.45	-1.00	7.40	4.40

La variable à expliquer est `y` et les variables explicatives sont `x1`, `x2`, `x3`, `x4`, `x5` et `x6`. Le modèle de *rlm* incluant toutes ces variables est envisageable.

1. Expliciter ce modèle. Détailler toutes les hypothèses standards permettant la validation des résultats statistiques usuels.

2. Dans les commandes qui suivent, que cherche-t-on à faire ? En quoi consiste la méthode inhérente à `method = "backward"` ? Quelle conclusion tirer du graphique ?

```
library(leaps)
v = regsubsets(y ~ ., w, method = "backward")
plot(v, scale = "adjr2")
```



Exercice 19.

1. Décrire brièvement l'enjeu des commandes R suivantes :

```
X = 1:100 / 10
Y = 2 + X + 1.2 * rnorm(100, 0, (1 + 0.7 * X^2) / 9)
reg = lm(Y ~ X)
summary(reg)
```

Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.8540	0.9168	3.11	0.0024	∗∗∗
X	0.8482	0.1576	5.38	0.0000	∗∗∗

```
Residual standard error:  2.029 on 98 degrees of freedom
Multiple R-squared:  0.5972, Adjusted R-squared:  0.5931
F-statistic:  145.3 on 1 and 98 DF, p-value:  < 2.2e-16
```

2. On exécute les commandes R suivantes :

```
library(lmtest)
bptest(reg)
```

Cela renvoie : p-valeur $= 4.63e-06$. Quel problème est ainsi mis en évidence ? Est-ce normal vu la construction du modèle ?

3. Décrire brièvement l'enjeu des commandes R suivantes :

```
reg2 = lm(Y ~ X, weights = 1 / sample(1:10, 100, T))
summary(reg2)
```

Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.4483	0.8785	3.93	0.0002	∗∗∗
X	0.5809	0.1586	3.66	0.0004	∗∗∗

```
Residual standard error:  2.341 on 98 degrees of freedom
Multiple R-squared:  0.1204, Adjusted R-squared:  0.1115
F-statistic:  13.42 on 1 and 98 DF, p-value:  0.0004043
```

Est-ce que le modèle `reg2` est meilleur que `reg1` ?

4. Décrire brièvement l'enjeu des commandes R suivantes :

```
u = (1 + 0.7 * X^2) / 9
reg3 = lm(Y ~ X, weights = 1 / u^2)
summary(reg3)
```

Cela renvoie :

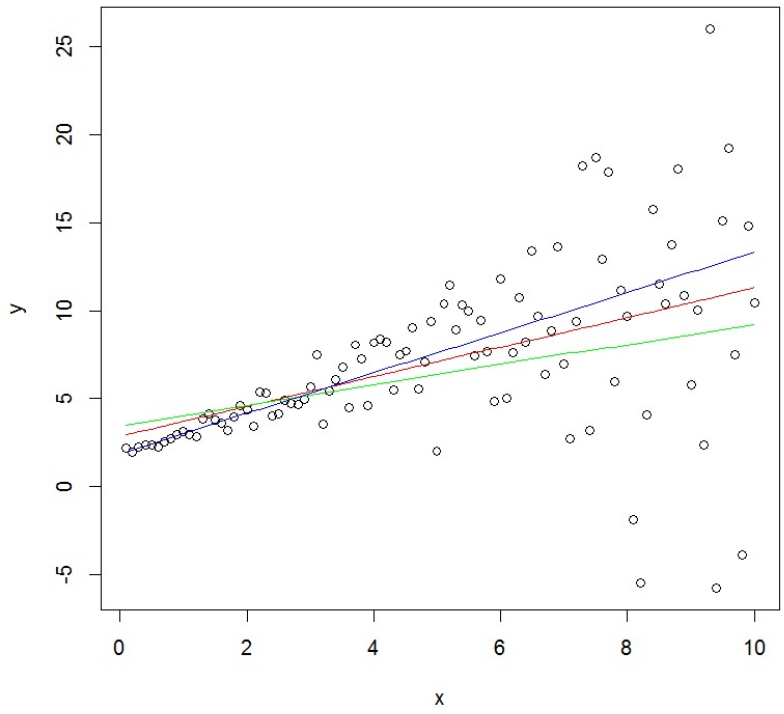
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.8654	0.0650	28.72	0.0000	***
X	1.1503	0.0573	20.06	0.0000	***

```
Residual standard error:  1.245 on 98 degrees of freedom
Multiple R-squared:  0.8042, Adjusted R-squared:  0.8022
F-statistic:  402.5 on 1 and 98 DF, p-value:  < 2.2e-16
```

Est-ce que le modèle `reg3` est meilleur que `reg1` ?

5. Décrire brièvement l'enjeu des commandes R suivantes :

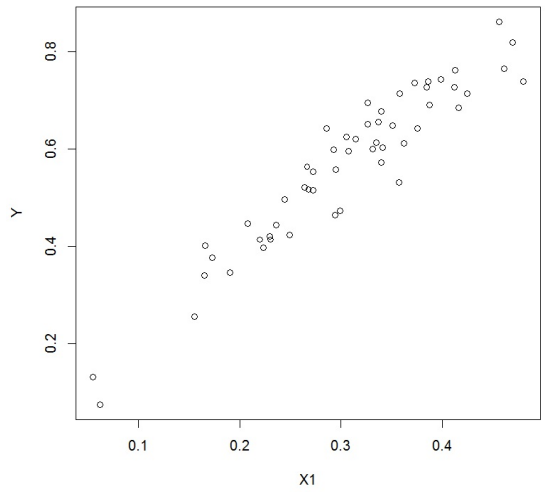
```
plot(x, y)
lines(X, fitted(reg), col = "red")
lines(X, fitted(reg2), col = "green")
lines(X, fitted(reg3), col = "blue")
```



Exercice 20. L'étude porte sur le rendement d'une culture de blé (variable Y) à partir de la quantité/hauteur de pluie printanière (variable $X1$). Les données sont disponibles ici :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/blé.txt",
header = T)
attach(w)
str(w)
```

On souhaite expliquer Y à partir de $X1$. Le nuage de points est :



1. Deux modèles de régression sont considérés : `reg1` et `reg2`, lesquels sont donnés par les commandes R suivantes :

```
reg1 = lm(Y ~ X1)
reg2 = lm(Y ~ poly(X1, 2))
```

Écrire les expressions mathématiques de ces deux modèles de régression.

2. On exécute la commande R suivantes :

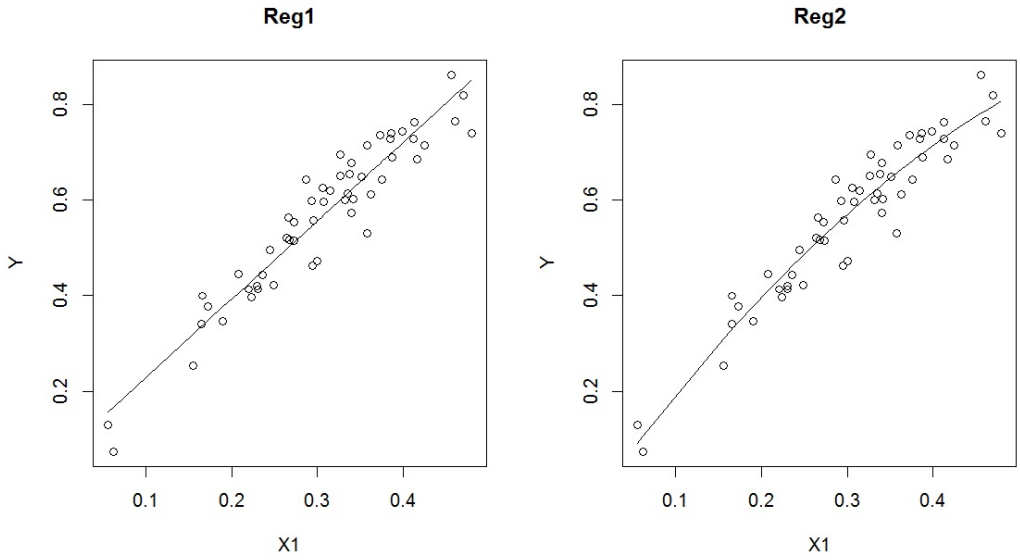
```
anova(reg1, reg2)
```

Cela renvoie :

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	52	0.14				
2	51	0.12	1	0.02	7.75	0.0075

Est-ce que l'on peut affirmer que les deux modèles diffèrent avec un risque faible de se tromper ? Si oui, quel est le degré de significativité ?

3. On trace les droites de régression associées. Laquelle vous semble ajuster le mieux le nuage de points ?



4. Les critères AIC et BIC donnent :

	`reg1`	`reg2`
AIC	−162.0788	−167.7176
BIC	−156.1119	−159.7617

   Quelle modélisation est la meilleure a priori ?

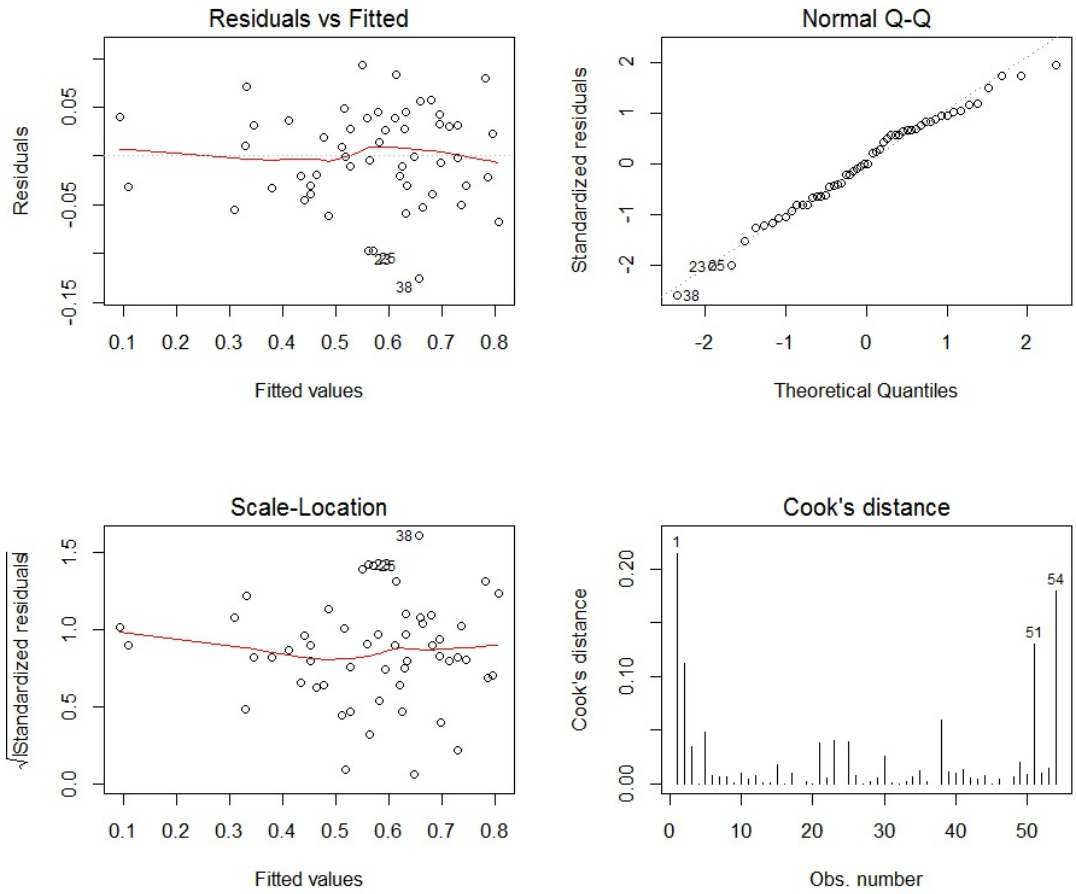
5. Est-ce que vous validez la qualité et la pertinence du modèle `reg2` au vu du tableau et des graphiques suivants :

   - Tableau :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.5662	0.0067	85.03	0.0000	***
poly(X1, 2)1	1.1310	0.0489	23.12	0.0000	***
poly(X1, 2)2	-0.1362	0.0489	-2.78	0.0075	**

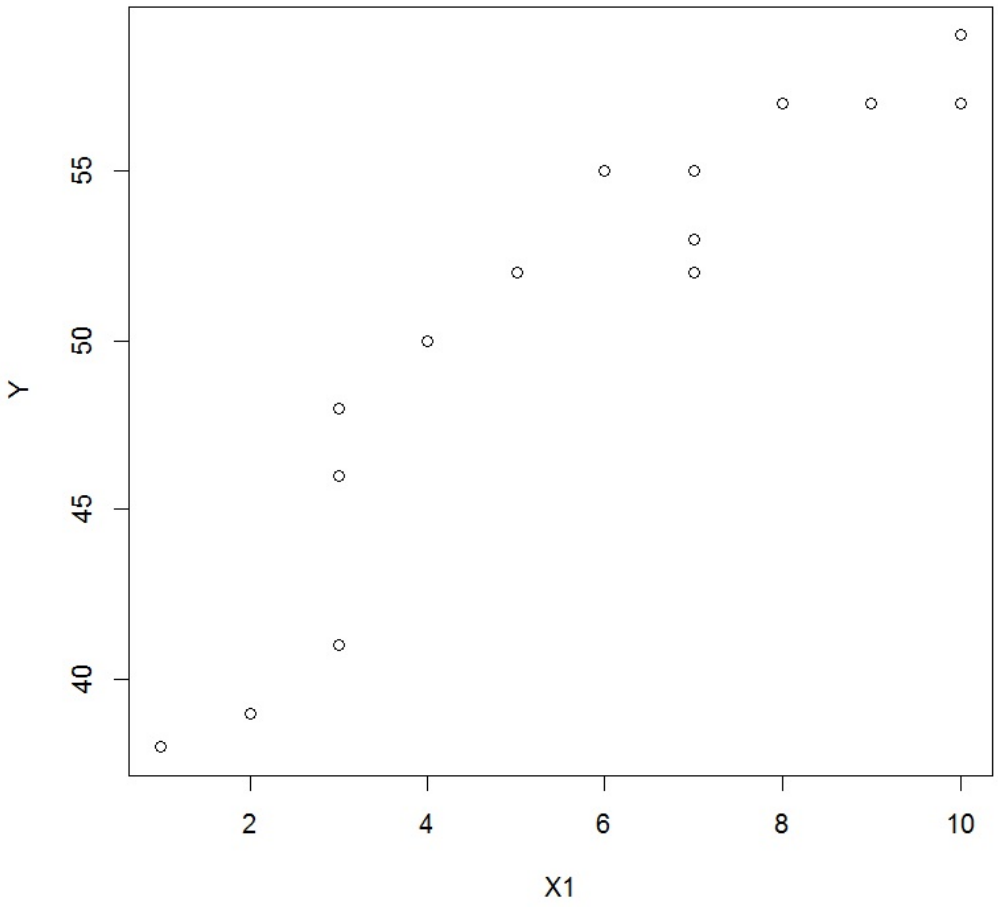
```
Residual standard error:  0.04893 on 51 degrees of freedom
Multiple R-squared:  0.914, Adjusted R-squared:  0.9106
F-statistic:  271 on 2 and 51 DF, p-value:  < 2.2e-16
```

- Graphiques :


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. number

Exercice 21. Un limnologue s'intéresse à la relation existante entre la clarté d'un lac (variable Y) et de sa taille (variable $X1$). On souhaite expliquer Y à partir de $X1$. Les données sont :

```
Y = c(46, 52, 55, 57, 41, 57, 59, 52, 55, 38, 48, 39, 53, 50, 57)
X1 = c(3, 5, 6, 10, 3, 9, 10, 7, 7, 1, 3, 2, 7, 4, 8)
plot(X1, Y)
```

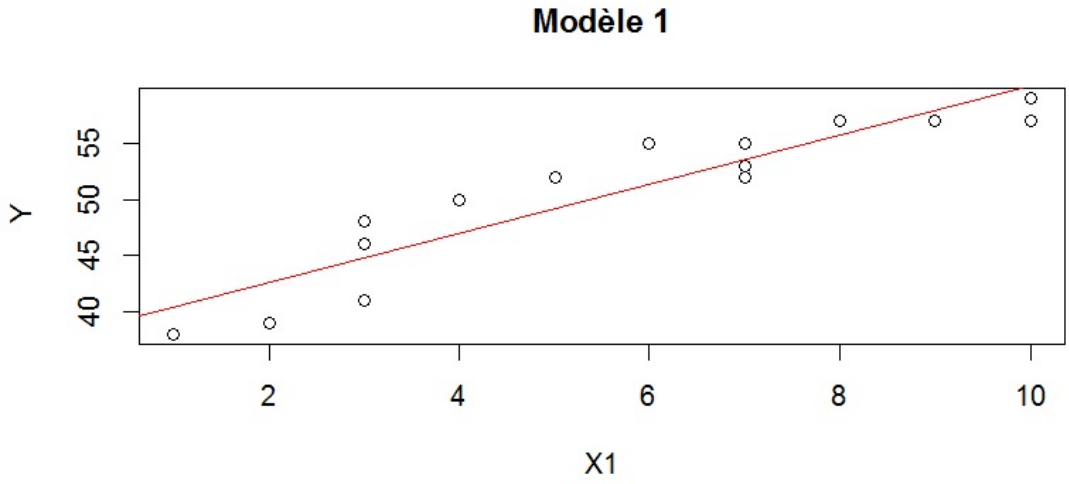


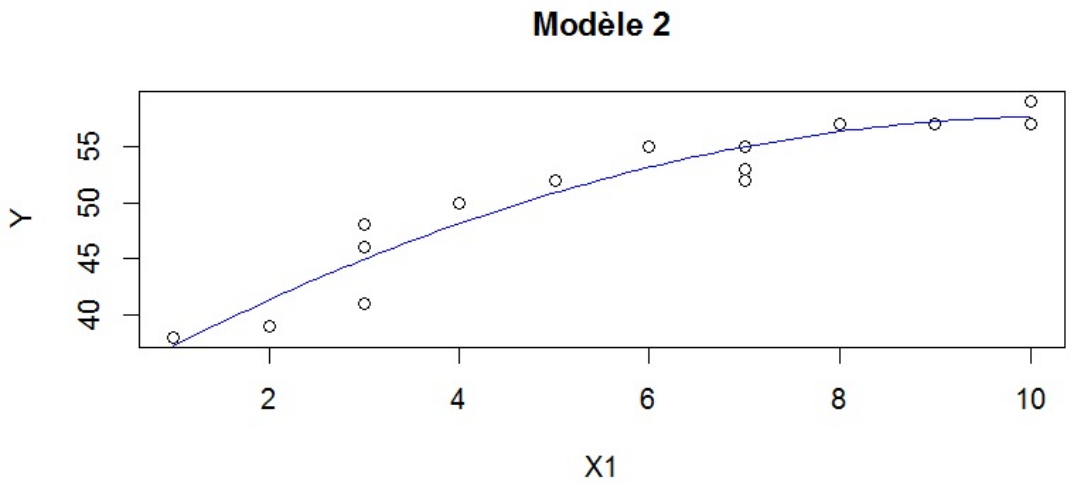
1. Deux modèles de régression sont considérés : `reg1` et `reg2`, lesquels sont donnés par les commandes R suivantes :

```
reg1 = lm(Y ~ X1)
reg2 = lm(Y ~ poly(X1, 2))
```

Une fois les estimations faites pour les 2 modèles, on trace les droites/lignes de regression.

On obtient :





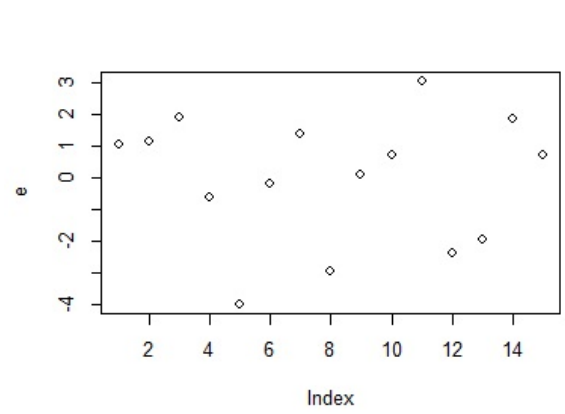
Quelle droite/ligne semble mieux ajuster le nuage de points ?

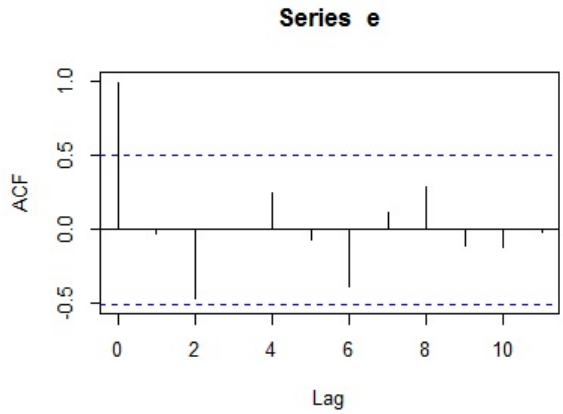
2. On donne les résultats suivantes :

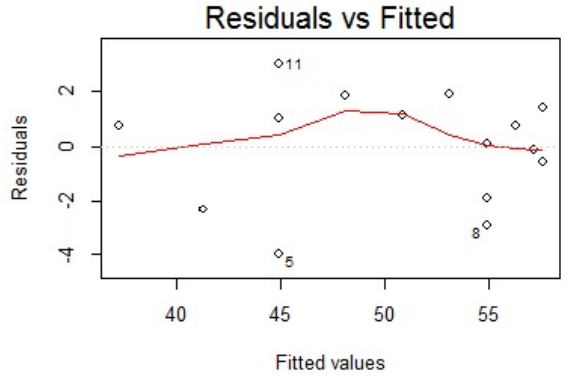
	AIC	BIC	R2 ajusté
modèle 1	76.01	78.14	0.85
modèle 2	70.18	73.01	0.90

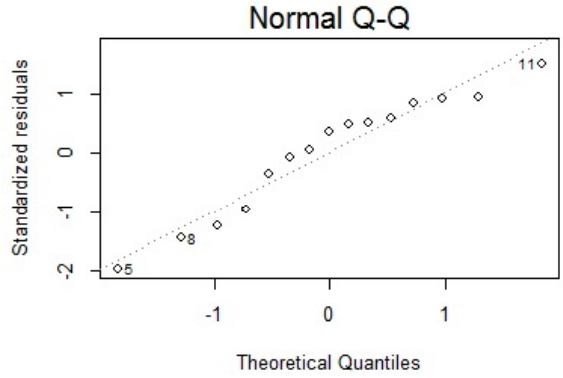
Quel modèle est le meilleur ?

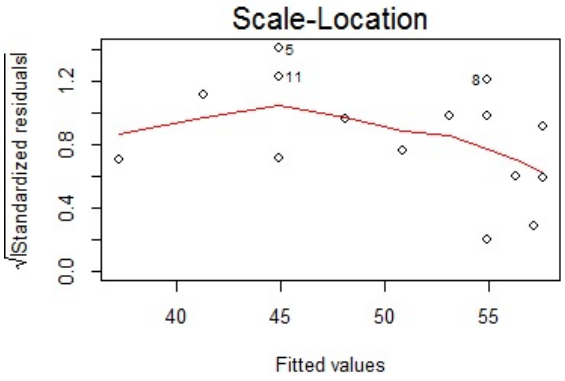
3. On considère le meilleur modèle. Commenter les graphiques associés suivants :

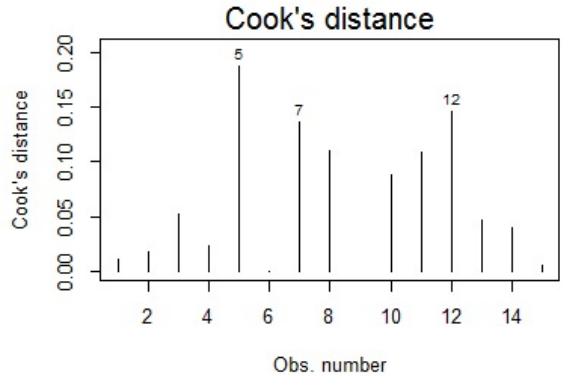












Est-ce que tout semble satisfaisant ?

Exercice 22. On considère le jeu de données `FEV`, le nom signifiant Forced Expiratory Volume. L'étude porte sur la relation entre le volume expiratoire forcé (FEV) et le fait de fumer. Pour ce faire, on considère un échantillon de 654 jeunes âgés de 3 à 19 ans dans la région de l'est de Boston entre le milieu et la fin des années 1970. Pour chacun d'entre eux, on relève les valeurs des variables suivantes:

- `age` : variable quantitative (discrète) (en années)
- `fev` : variable quantitative (en litres)
- `ht` : (pour heigth) taille (en inches)
- `sex` : variable binaire (Femme codé par 0, Homme codé par 1)
- `smoke` : variable binaire (Non fumeur codé par 0, Fumeur codé par 1)

On souhaite donc expliquer `fev` en fonction des autres variables.
On fait les commandes R suivantes :

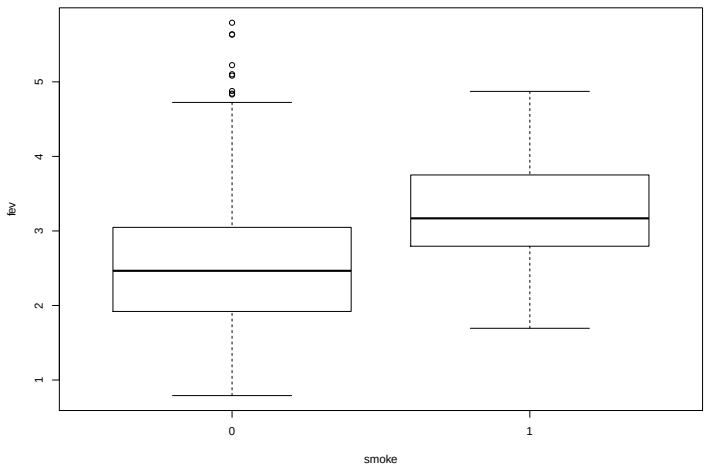
```
w = read.table(file = "https://chesneau.users.lmno.cnrs.fr/fev.dat.txt",
col.names = c("age", "fev", "ht", "sex", "smoke"),
colClasses = c(rep("numeric", 3), "factor", "factor")) attach(w)
str(w)
```

Cela renvoie :

```
'data.frame': 654 obs. of  5 variables:
 $ age  : num  9 8 7 9 9 8 6 6 8 9 ...
 $ fev  : num  1.71 1.72 1.72 1.56 1.9 ...
 $ ht   : num  57 67.5 54.5 53 57 61 58 56 58.5 60 ...
 $ sex  : Factor w/ 2 levels "0","1": 1 1 1 2 2 1 1 1 1 1 ...
 $ smoke: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

1. Dans cette question, on cherche à expliquer `fev` en fonction de `age` et `smoke` uniquement.

   (a) On exécute des commandes qui donnent la sortie suivante :



Quelles sont ces commandes ? Commenter le graphique obtenu.

(b) Quelles sont les enjeux des commandes suivantes :

```
plot(age[smoke == "0"], fev[smoke == "0"], cex = 0.1, xlab = "age", ylab
= "fev")
points(age[smoke == "1"], fev[smoke == "1"], cex = 0.1, col = "red")
```

(c) On décide d'aller un peu plus loin. On exécute les commandes suivantes :

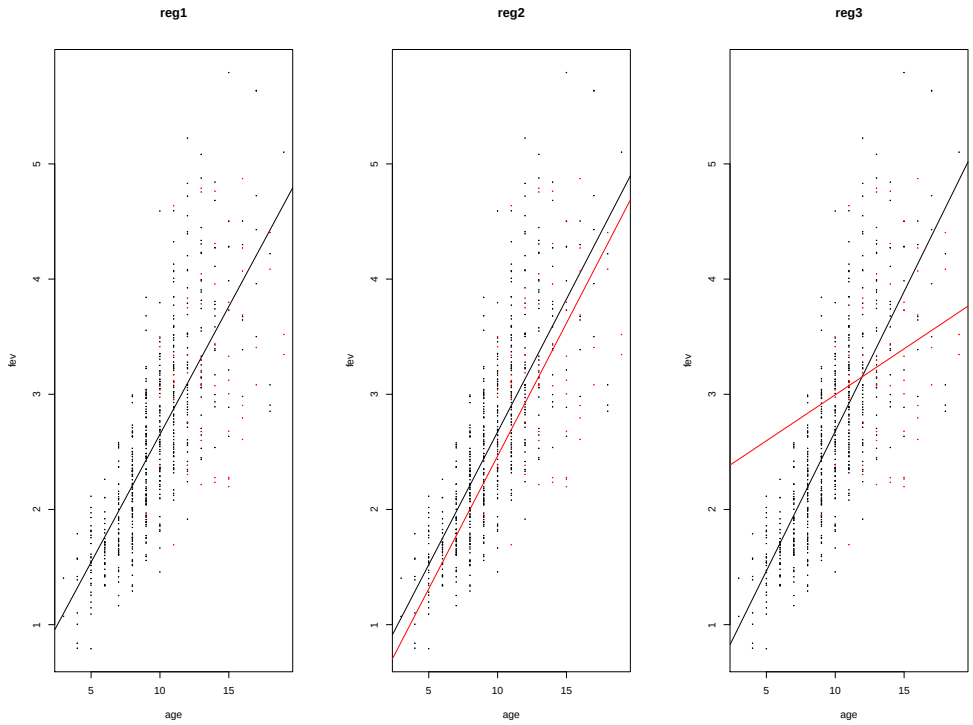
```
reg1 = lm(fev ~ age)
reg2 = lm(fev ~ age + smoke)
reg3 = lm(fev ~ age * smoke)
```

Écrire les formules mathématiques (formes génériques) des modèles considérés.

(d) Puis on fait :

```
par(mfrow = c(1, 3))
plot(age[smoke == "0"], fev[smoke == "0"], cex = 0.1, xlab = "age", ylab
= "fev", main = "reg1")
points(age[smoke == "1"], fev[smoke == "1"], cex = 0.1, col = "red")
abline(reg1)
plot(age[smoke == "0"], fev[smoke == "0"], cex = 0.1, xlab = "age", ylab
= "fev", main = "reg2")
abline(reg2$coefficients[1], reg2$coefficients[2])
points(age[smoke == "1"], fev[smoke == "1"], cex = 0.1, col = "red")
abline(reg2$coefficients[1] + reg2$coefficients[3],
reg2$coefficients[2], col = "red")
plot(age[smoke == "0"], fev[smoke == "0"], cex = 0.1, xlab = "age", ylab
= "fev", main = "reg3")
abline(reg3$coefficients[1], reg3$coefficients[2])
points(age[smoke == "1"], fev[smoke == "1"], cex = 0.1, col = "red")
abline(reg3$coefficients[1] + reg3$coefficients[3], reg3$coefficients[2]
+ reg3$coefficients[4], col = "red")
```

Cela renvoie :



Commenter les graphiques obtenus et tirer des conclusions intéressantes sur les modèles `reg1`, `reg2` et `reg3` (notamment, en commentant les droites affichées).

2. Dorénavant, on prend en compte toutes les variables. On considère un quatrième modèle `reg4`. On fait les commandes :

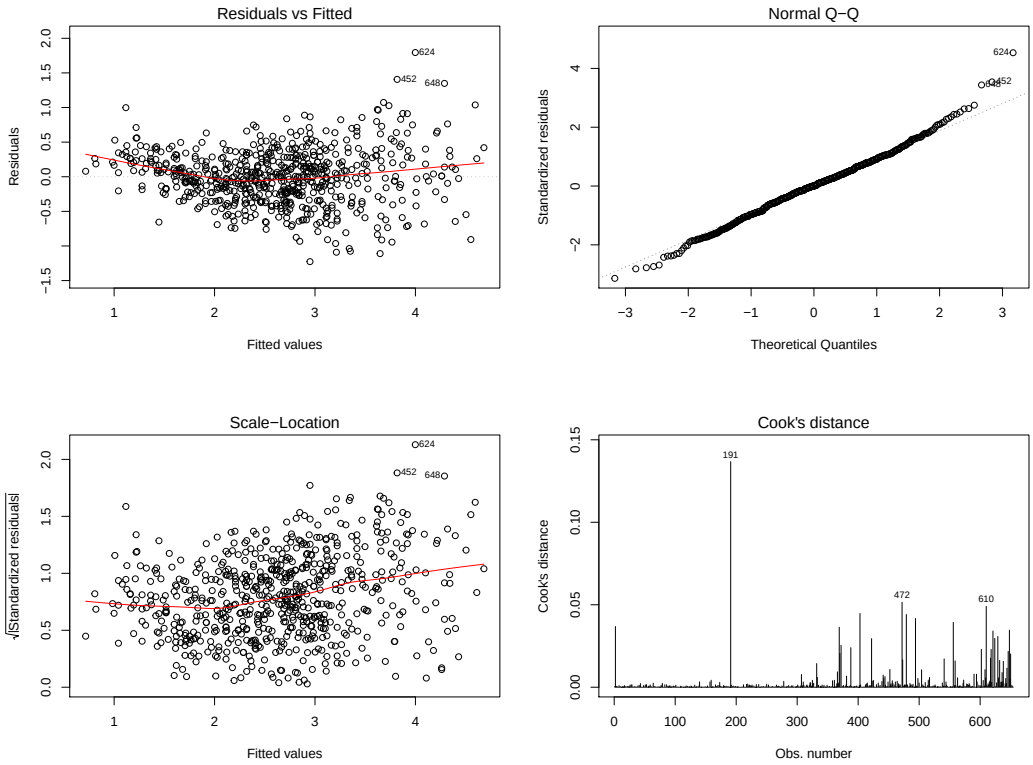
```
reg4 = lm(fev ~ (age + ht) * sex * smoke)
```

   (a) Combien de variables explicatives vont être considérées dans le modèle `reg4` ? Écrire la formule mathématique (forme générique) du modèle `reg4`.

   (b) Partant de `reg4`, proposer des commandes pour :

   - donner un résumé statistique complet des estimations du modèle,
   - tester la normalité des résidus,
   - tester l'égalité des variances des résidus par rapport à `sex` d'une part, et par rapport à `smoke` d'autre part,

- obtenir en une seule figure les 4 graphiques permettant l'analyse fine des hypothèses standards :



- Au vu des graphiques précédents, est-ce que les hypothèses standards semblent être vérifiées ? Argumenter.

3. On considère un cinquième modèle `reg5`. On fait les commandes :

```
reg5 = lm(sqrt(fev) ~ (age + ht) * sex * smoke)
```

(a) Quelle est la différence majeure entre `reg4` et `reg5` ? Est-ce que `reg5` est un modèle de régression linéaire multiple par rapport à la variable `fev` ?

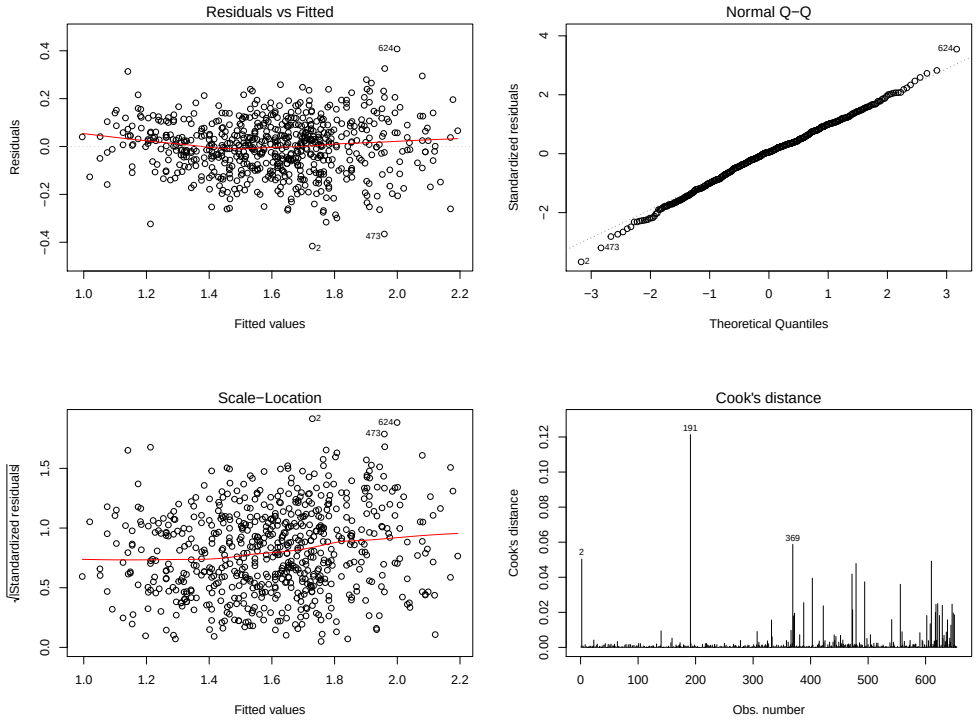
(b) Donner les commandes permettant d'obtenir :

- la sortie suivante :

```
Call:
lm(formula = sqrt(fev) ~ (age + ht) * sex * smoke)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41605 -0.07214  0.00583  0.07629  0.40760

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.411329   0.111759  -3.680 0.000252 ***
age              0.021251   0.004093   5.192 2.79e-07 ***
```

```
ht                0.029191   0.002326  12.548  < 2e-16 ***
sex1             -0.196287   0.141885  -1.383 0.167015
smoke1            1.303848   0.564767   2.309 0.021280 *
age:sex1          0.002710   0.006010   0.451 0.652154
ht:sex1           0.003381   0.003036   1.114 0.265795
age:smoke1       -0.022988   0.009351  -2.458 0.014218 *
ht:smoke1        -0.016049   0.008560  -1.875 0.061258 .
sex1:smoke1      -2.745050   0.752979  -3.646 0.000288 ***
age:sex1:smoke1   0.015007   0.014411   1.041 0.298069
ht:sex1:smoke1    0.038515   0.011650   3.306 0.000999 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1158 on 642 degrees of freedom
Multiple R-squared:  0.8106,Adjusted R-squared:  0.8073
F-statistic: 249.7 on 11 and 642 DF,  p-value: < 2.2e-16
```

- obtenir en une seule figure les 4 graphiques suivants :



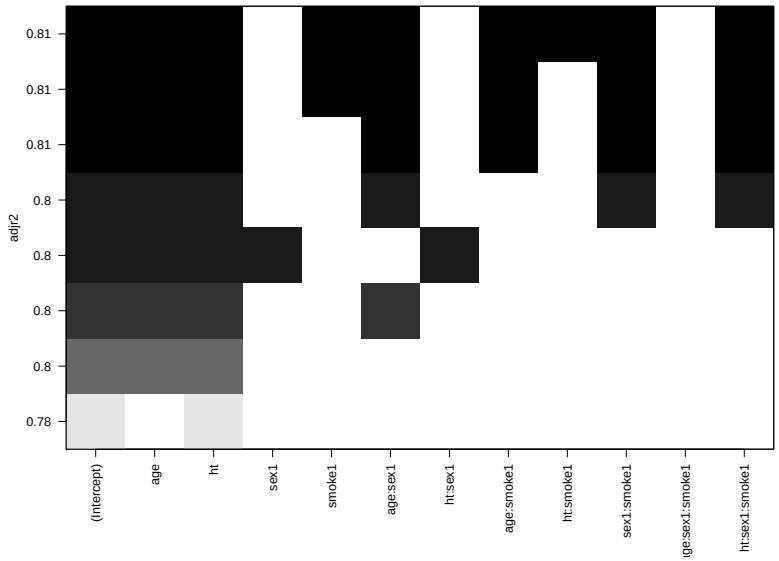
- Au vu des graphiques précédents, est-ce que les hypothèses standards semblent être vérifiées ? Argumenter et comparer avec le graphique analogue obtenu pour le modèle `reg4`.
- Que pensez-vous de la qualité prédictive du modèle `reg5` ?

(c) En utilisant `reg5`, donner une prédiction de la valeur moyenne de `fev` pour une fille non-fumeuse de 17 ans mesurant 53 inches.

4. Finalement, on fait les commandes :

```
library(leaps)
v = regsubsets(sqrt(fev)   (age + ht) * sex * smoke, w, method =
"exhaustive")
plot(v, scale = "adjr2")
```

Cela renvoie :



Expliquer l'enjeu de ce graphique. Que peut-on en conclure ?

5. Pour finir, proposer un modèle de régression et les commandes associées pour déterminer les chances qu'un jeune âgé entre 3 et 19 ans choisi au hasard fume en fonction de son âge, de son sexe et de son volume expiratoire forcé.

Exercice 23. On s'intéresse au jeu de données `longley`. On exécute les commandes R suivantes :

```
data(longley)
attach(longley)
str(longley)
```

Cela renvoie :

```
'data.frame':   16 obs. of  7 variables:
 $ GNP.deflator: num  83 88.5 88.2 89.5 96.2 ...
 $ GNP         : num  234 259 258 285 329 ...
 $ Unemployed  : num  236 232 368 335 210 ...
 $ Armed.Forces: num  159 146 162 165 310 ...
 $ Population  : num  108 109 110 111 112 ...
 $ Year        : int  1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 ...
 $ Employed    : num  60.3 61.1 60.2 61.2 63.2 ...
```

Plus précisément, pour 16 années d'observations allant de 1947 à 1962, on dispose

- du déflateur de prix implicite du Produit National Brut (avec 1954 = 100) (variable `GNP.deflator`),
- du produit national brut (variable `GNP`),
- du nombre de chômeurs (variable `Unemployed`),
- du nombre de personnes dans les forces armées (variable `Armed.Forces`),
- de la population "non institutionnalisée" âgée de plus de 14 ans (variable `Population`),
- de l'année (variable `Year`),
- du nombre de personnes employées (variable `Employed`).

1. On exécute les commandes R suivantes :

```
reg = lm(Employed ~ ., data = longley)
reg2 = step(reg, direction = "both", k = log(length(Employed)))
summary(reg2)
```

Cela renvoie :

```
Call:
lm(formula = Employed ~ GNP + Unemployed + Armed.Forces + Year,
    data = longley)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42165 -0.12457 -0.02416  0.08369  0.45268
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.599e+03  7.406e+02  -4.859 0.000503 ***
GNP          -4.019e-02  1.647e-02  -2.440 0.032833 *
Unemployed   -2.088e-02  2.900e-03  -7.202 1.75e-05 ***
Armed.Forces -1.015e-02  1.837e-03  -5.522 0.000180 ***
Year          1.887e+00  3.828e-01   4.931 0.000449 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2794 on 11 degrees of freedom
Multiple R-squared:  0.9954,    Adjusted R-squared:  0.9937
F-statistic: 589.8 on 4 and 11 DF,  p-value: 9.5e-13
```

Décrire avec précision l'enjeu des commandes précédentes, ainsi que les sorties obtenues. Est-ce que le modèle semble performant ?

2. On exécute les commandes R suivantes :

```
shapiro.test(resid(reg2))$p.value
```

Cela renvoie :

```
[1] 0.600544
```

Puis :

```
library(lmtest)
raintest(reg2)$p.value
```

Cela renvoie :

```
[1] 0.08005809
```

Puis :

```
bptest(reg2)$p.value
```

Cela renvoie :

```
       BP
0.9700813
```

Puis :

```
dwtest(long.lm)$p.value
```

Cela renvoie :

```
[1] 0.6885454
```

Et enfin :

```
library(car)
vif(reg2)
```

Cela renvoie :

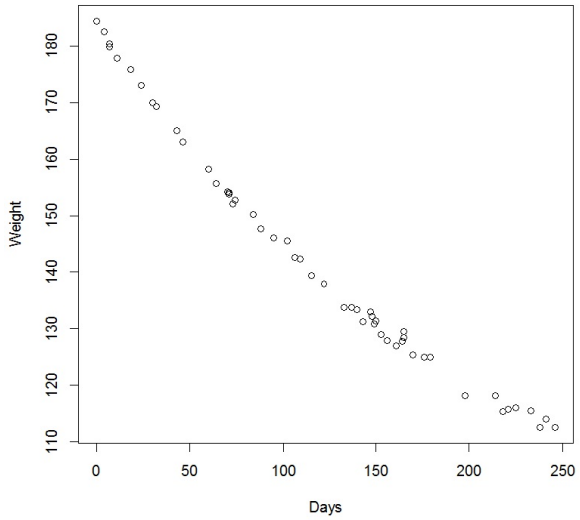
```
        GNP   Unemployed Armed.Forces          Year
515.123851    14.108642     3.141581   638.128041
```

Décrire avec précision l'enjeu des commandes précédentes, ainsi que les sorties obtenues. Est-ce que le modèle est dénué de problème ?

Exercice 24. On considère le jeu de données `wtloss` de la librairie `MASS`. Celui-ci donne le poids en kilogrammes (variable `Weight`) d'un patient obèse à 52 points de temps (variable `Days`) sur une période d'un programme de perte de poids de 8 mois. On souhaite expliquer la variable `Weight` en fonction de la variable `Days`. Pour avoir une idée plus précise, on exécute les commandes R suivantes :

```
library(MASS)
attach(wtloss)
plot(Days, Weight)
```

Cela renvoie :



1. On considère le modèle de *rlm* standard :

$$\text{Weight} = \beta_0 + \beta_1 \text{Days} + \epsilon.$$

   Étant donné le contexte, il y a de la dépendance temporelle dans les variables d'erreurs. Quel repère graphique permet de mettre en évidence cette dépendance ? Donner les commandes R associées.

2. Quel test statistique permet de mettre en évidence une structure de dépendance AR(1) des erreurs ? Préciser les hypothèses associées à ce test statistique.

3. En admettant une dépendance AR(1) des erreurs, proposer un modèle de régression adapté. Donner les commandes R associées.

4. On s'interesse maintenant à l'ajustement simple du nuage de points, en mettant de côté l'aspect temporel des données. On considère alors les commandes R :

```
s =c(b0 = 92, b1 = 93, b2 = 120)
reg = nls(Weight ~ b0 + b1 * 2^(-Days / b2), start = s)
summary(reg)
```

Cela renvoie :

```
Formula: Weight ~ b0 + b1 * 2^(-Days / b2)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
b0   81.374      2.269   35.86   <2e-16 ***
b1  102.684      2.083   49.30   <2e-16 ***
b2  141.910      5.295   26.80   <2e-16 ***

Residual standard error: 0.8949 on 49 degrees of freedom

Number of iterations to convergence: 3
Achieved convergence tolerance: 2.98e-06
```

Quelle méthode est utilisée dans ces commandes ? Quel algorithme utilise-t-elle ? Donner une estimation ponctuelle du poids du patient à 265 jours de traitement.

5. Comprendre l'enjeu des commandes suivantes :

```
plot(Days, Weight)
x = seq(0, 255, 1)
lines(x, predict(reg, new = list(Days = x)))
```

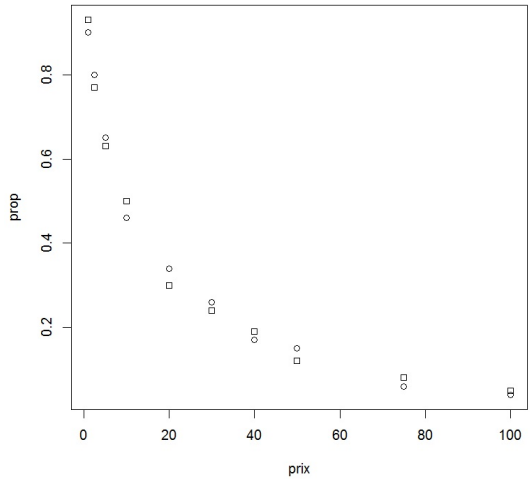
6. Proposer d'autres méthodes possibles pour l'ajustement de ce nuage de points.

Exercice 25. Bertrand souhaite expliquer la proportion d'individus dans une ville qui serait susceptible d'acheter un ordinateur (variable `prop`) avec le prix de vente de l'ordinateur (en centaines de dollars) (variable `prix`) et la ville de résidence (variable `ville` à deux modalités : A et B). Pour se faire, un échantillon de 20 individus est considéré. Les données sont consultables dans les commandes suivantes :

```
prop = c(.9, .8, .65, .46, .34, .26, .17, .15, .06, .04, .93, .77, .63,
.5, .3, .24, .19, .12, .08, .05)
prix = c(1, 2.5, 5, 10, 20, 30, 40, 50, 75, 100, 1, 2.5, 5, 10, 20, 30,
40, 50, 75, 100)
ville = c("A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "B", "B",
"B", "B", "B", "B", "B", "B", "B", "B")
```

1. Bertrand décide de recoder la variable `ville` en variable binaire : il pose `ville` $= 0$ si `ville = A` et `ville` $= 1$ si `ville = B`. Écrire toutes les commandes que Bertrand doit faire.

2. À l'aide d'une seule ligne de commande, assez courte, Bertrand obtient le graphique suivant :



   (On a des cercles si `ville` $= 0$ et des carrés si `ville` $= 1$). Quelle peut être cette ligne de commande ?

3. Bertrand pense à essayer le modèle, nommé `nlm`, écris sous forme générique :

$$\texttt{prop} = \beta_0 + \beta_1 e^{\beta_2 \texttt{prix}} + \beta_3 \texttt{ville} + \epsilon.$$

   Écrire les commandes exactes correspondant à ce modèle. Comment obtient-on un résumé des estimations liées à ce modèle ?

4. Bertrand exécute les commandes suivantes :

```
xgrid = seq(min(prix), max(prix), length = 200)
lines(xgrid, coef(nlm)[1] + coef(nlm)[2]*exp(coef(nlm)[3] * xgrid), lty
= 1 )
lines(xgrid, coef(nlm)[1] + coef(nlm)[2]*exp(coef(nlm)[3] * xgrid) +
coef(nlm)[4], lty = 2 )
```

   Qu'obtient-il alors ?

Exercice 26. On s'intéresse au jeu de données `sbp` (pour systolic blood pressure) de la librairie `multcomp`. Pour 69 individus, on dispose

- de la pression systolique en mmHg (variable `sbp`),
- de leur âge en année (variable `age`),
- de leur sexe (variable `gender` à 2 modalités : `female` pour femme et `male` pour homme).

On souhaite expliquer du mieux possible la variable `sbp` en fonction des autres variables.

1. On exécute les commandes R suivantes :

```
library(multcomp)
data(sbp)
summary(sbp)
```

Cela renvoie :

```
     gender         sbp             age
 male  :40   Min.   :110.0   Min.   :17.00
 female:29   1st Qu.:135.0   1st Qu.:36.00
             Median :149.0   Median :47.00
             Mean   :148.7   Mean   :46.14
             3rd Qu.:162.0   3rd Qu.:59.00
             Max.   :185.0   Max.   :70.00
```

Dès lors, répondre aux questions suivantes relatives aux données `sbp`:

(a) Si l'on choisi un individu au hasard, combien a t-on de chances que ce soit une femme ?

(b) Si l'on choisi un individu au hasard, combien a t-on de chances que sa pression systolique soit supérieure à 149 ?

(c) À quelle valeur la pression systolique des trois quarts des individus est-elle inférieure ?

2. On exécute les commandes R suivantes :

```
reg = lm(sbp ~ age * gender, data = sbp)
summary(reg)
```

Cela renvoie :

```
Call:
lm(formula = sbp ~ age * gender, data = sbp)

Residuals:
    Min      1Q  Median      3Q     Max
-20.647  -3.410   1.254   4.314  21.153
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      110.03853    4.73610  23.234  < 2e-16 ***
age                0.96135    0.09632   9.980 9.63e-15 ***
genderfemale     -12.96144    7.01172  -1.849   0.0691 .
age:genderfemale  -0.01203    0.14519  -0.083   0.9342
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.946 on 65 degrees of freedom
Multiple R-squared:  0.7759,    Adjusted R-squared:  0.7656
F-statistic: 75.02 on 3 and 65 DF,  p-value: < 2.2e-16
```

(a) Quel est le modèle statistique considéré ? Donner la formule mathématique (forme générique) du modèle, avec les estimations associées.

Pour les 3 *questions ci-dessous, on suppose que le modèle* `reg` *est bon, ce qui sera remis en cause par la suite.*

(b) Déterminer un intervalle de confiance pour le coefficient de régression associé à la variable `age` au niveau 95% (on pourra utiliser l'approximation normale pour la valeur du quantile qui intervient dans la définition de l'intervalle de confiance au niveau $100(1-\alpha)\%$, à savoir: $t_\alpha(\nu) \approx z_\alpha = 1.96$).

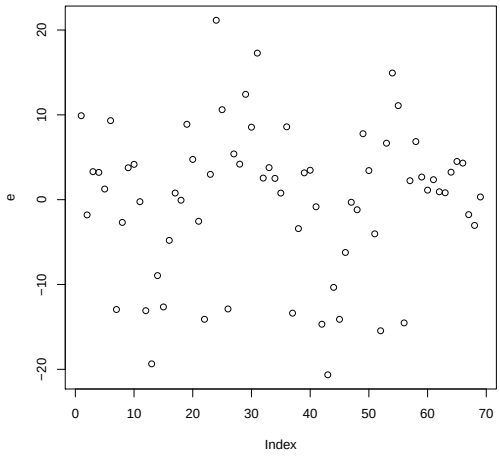
(c) Est-ce que les influences combinées de `age` et `gender` génèrent un effet supplémentaire dans l'explication de `sbp` ?

(d) Donner une prédiction moyenne de la valeur de `sbp` pour une femme vérifiant `age = 22`.

3. En vue de valider le modèle, on fait une étude grossière des résidus du modèle. On exécute les commandes R suivantes :

```
e = residuals(reg)
plot(e)
```

Cela renvoie :

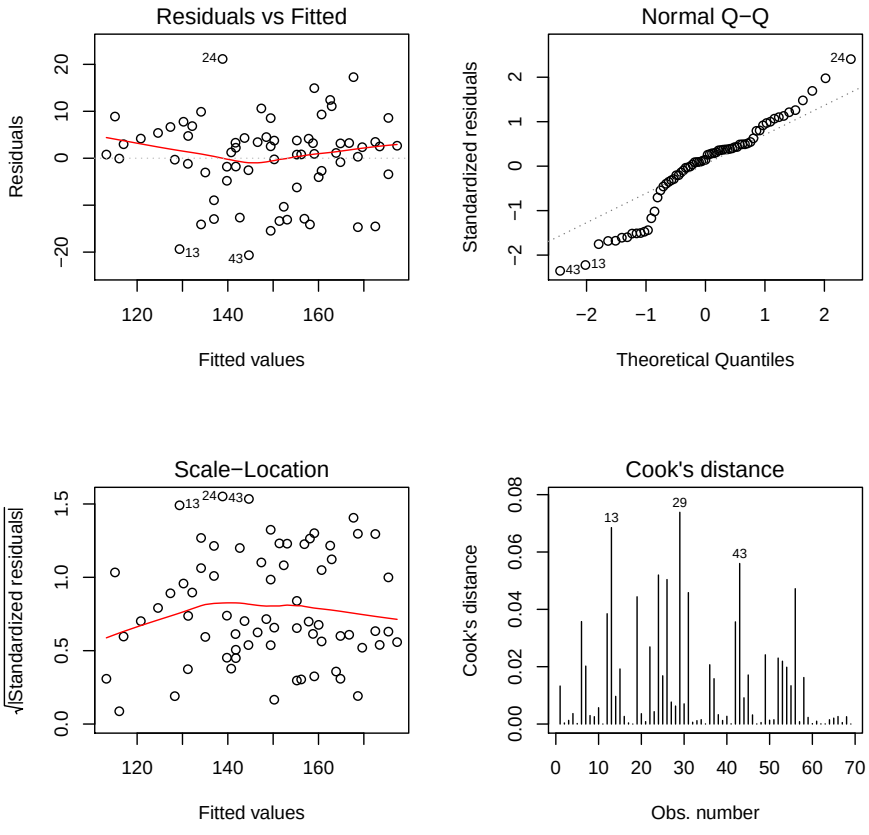


Que pouvez-vous dire sur ce graphique des résidus ? Est-ce que tout semble bon ?

4. On décide d'analyser plus finement les hypothèses. On exécute les commandes R suivantes :

```
par(mfrow = c(2, 2))
plot(reg, 1:4)
```

Cela renvoie :



Un problème particulier se profile. Lequel est-ce ? Commenter.

5. On exécute les commandes R suivantes :

```
shapiro.test(e)$p.value; bartlett.test(e, gender)$p.value
```

Cela renvoie :

```
[1] 0.01620704
[1] 0.5120409
```

(a) Quels sont les tests statistiques considérés ?

(b) Que peut-on dire des résultats numériques obtenus pour ces tests ?

(c) Dès lors, que peut-on dire des degrés de significativé (étoiles) obtenus dans le `summary(reg)` ?

(d) À la place du test statistique correspondant aux commandes : `bartlett.test(e, gender)`, quel autre test statistique aurait-on pu utiliser pour analyser le même phénomène ?

6. Par curiosité, on fait :

```
library(lmtest)
dwtest(reg)$p.value
```

Cela renvoie :

```
[1] 0.0004296367
```

Quel test statistique a t-on mis en œuvre ici ? Bien que la p-valeur vérifie p-valeur < 0.05, on prétend qu'il n'y a absolument aucun problème de dépendance inquiétante dans les données. D'après-vous, qu'est-ce qui pourrait expliquer cette affirmation, dans la structure même du jeu de données ?

7. On propose d'étudier un nouveau modèle noté `reg2`. Pour ce faire, on exécute les commandes R suivantes :

```
reg2 = lm(sbp ~ age + gender, data = sbp)
summary(reg2)
```

Cela renvoie :

```
Call:
lm(formula = sbp ~ age + gender, data = sbp)

Residuals:
    Min      1Q  Median      3Q     Max
-20.705  -3.299   1.248   4.325  21.160
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  110.28698    3.63824  30.313  < 2e-16 ***
age            0.95606    0.07153  13.366  < 2e-16 ***
genderfemale -13.51345    2.16932  -6.229  3.7e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.878 on 66 degrees of freedom
Multiple R-squared:  0.7759,    Adjusted R-squared:  0.7691
F-statistic: 114.2 on 2 and 66 DF,  p-value: < 2.2e-16
```

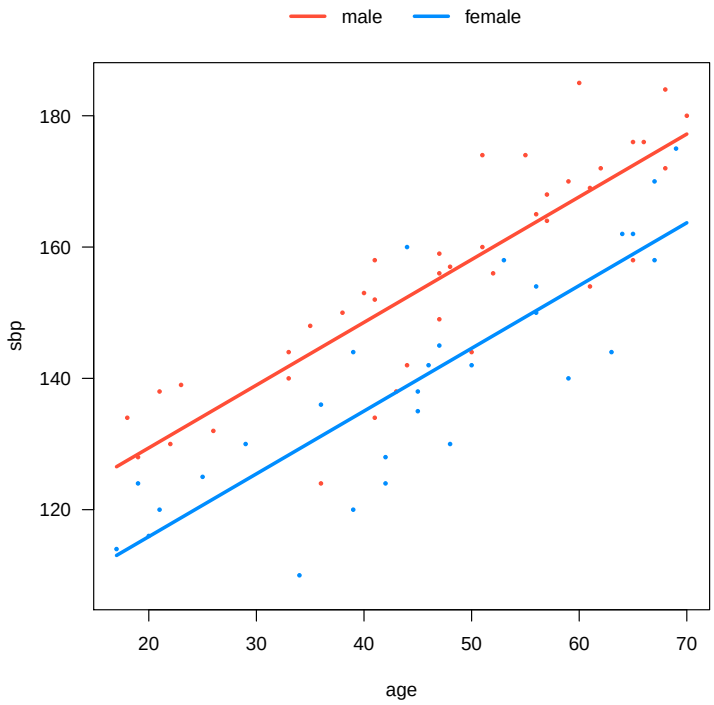
(a) Qu'a t-on fait de différent dans le nouveau modèle par rapport au modèle `reg` ? Quels seraient les arguments justifiant l'étude de ce modèle ?

(b) Quel est le principal changement statistique constaté dans le `summary(reg)`?

8. On fait les commandes R suivantes (non vues en cours) :

```
library(visreg)
visreg(reg2, "age", by = "gender", overlay = TRUE, band = FALSE)
```

Cela renvoie :



(a) Extrapoler l'enjeu de ces commandes en analysant bien la figure et sa légende.

(b) Donner les équations exactes des 2 droites tracées.

9. On exécute les commandes R suivantes :

```
e2 = residuals(reg2)
shapiro.test(e2)$p.value
```

Cela renvoie :

```
[1] 0.015673
```

Que peut-on en conclure ? Comparer avec le résultat analogue du modèle `reg`.

10. N'étant pas satisfait de ce qui précède, on propose d'étudier un nouveau modèle noté `reg3`. Pour ce faire, on exécute les commandes R suivantes :

```
reg3 = lm(sbp ~ poly(age,2, raw = T) + gender, data = sbp)
summary(reg3)
```

Cela renvoie :

```
Call:
lm(formula = sbp ~ poly(age, 2, raw = T) + gender, data = sbp)

Residuals:
    Min      1Q  Median      3Q     Max
-19.326  -3.948   1.561   5.106  23.153

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             124.695190   8.603200  14.494  < 2e-16 ***
poly(age, 2, raw = T)1    0.212495   0.409925   0.518   0.6060
poly(age, 2, raw = T)2    0.008473   0.004602   1.841   0.0702 .
genderfemale            -13.600805   2.131606  -6.381 2.12e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.721 on 65 degrees of freedom
Multiple R-squared:  0.787,     Adjusted R-squared:  0.7772
F-statistic: 80.05 on 3 and 65 DF,  p-value: < 2.2e-16
```

(a) Donner la formule mathématique (forme générique) du modèle `reg3`. Quel type de modèle est-ce ? (plusieurs réponses étant possibles, justifier la votre).

(b) Commenter la qualité prédictive du modèle. Comparer avec celle des modèles `reg` et `reg2`.

(c) De plus, on fait :

```
e3 = residuals(reg3)
shapiro.test(e3)$p.value
```

Cela renvoie :

```
[1] 0.06867984
```

Que peut-on en conclure par rapport aux modèles `reg` et `reg2` ?

(d) On informe que toutes les hypothèses statistiques validant le modèle sont vérifiées. Dès lors, utiliser `reg3` pour (re)donner une prédiction moyenne de la valeur de `sbp` pour une femme vérifiant `age = 22`.

11. On pousse l'étude plus loin en introduisant 2 nouveaux modèles, que l'on cherche à comparer avec les précédents. On exécute les commandes R suivantes :

```
reg4 = lm(sbp ~ log(age) + gender, data = sbp)
reg5 = lm(sbp ~ I(age^2) + gender, data = sbp)
```

Puis ont fait :

```
AIC(reg); AIC(reg2); AIC(reg3); AIC(reg4); AIC(reg5); BIC(reg);
BIC(reg2); BIC(reg3); BIC(reg4); BIC(reg5)
```

Cela renvoie :

```
[1] 504.0718
[1] 502.0791
[1] 500.5714
[1] 517.0452
[1] 498.856
[1] 515.2424
[1] 511.0155
[1] 511.7419
[1] 525.9817
[1] 507.7924
```

Que nous suggère ces résultats numériques ? Trouver une cohérence entre le modèle qui ressort et les analyses statistiques associées au modèle `reg3`.

12. Dès lors, on exécute les commandes R suivantes :

```
e5 = residuals(reg5)
shapiro.test(e5)$p.value
```

Cela renvoie :

```
[1] 0.1340853
```

Plus fort, on informe que toutes les hypothèses statistiques validant le modèle concerné sont vérifiées. Aussi, on fait :

```
which.max(c(summary(reg1)$adj.r.squared, summary(reg2)$adj.r.squared,
summary(reg3)$adj.r.squared, summary(reg4)$adj.r.squared,
summary(reg5)$adj.r.squared))
```

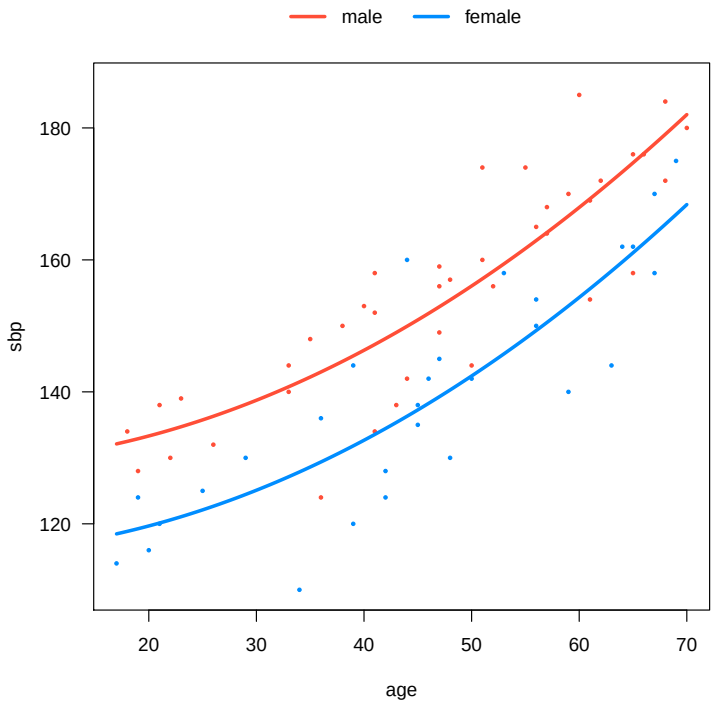
Cela renvoie :

```
[1] 5
```

En guise d'illustration, on fait :

```
library(visreg)
visreg(reg5, "age", by = "gender", overlay = TRUE, band = FALSE)
```

Cela renvoie :



Au vu de ce qui précède, si l'on ne devait garder qu'un seul modèle, lequel serait-ce ?

Exercice 27. Une étude porte sur la capacité d'accueil idéale d'un centre de loisirs en fonction du nombre d'individus habitants de sa commune de ratachement. Pour 8 communes, on dispose des données suivantes :

commune	population	accueil
1	9780	102
2	20130	123
3	29670	168
4	40210	259
5	49890	354
6	61040	480
7	70120	679
8	79870	997

1. On execute les commandes suivantes :

```
population = c(9780, 20130, 29670, 40210, 49890, 61040, 70120, 79870)
accueil = c(102, 123, 168, 259, 354, 480, 679, 997)
reg = nls(accueil ~ exp(a * population + b),
start = list(a = 0.00003, b = 4))
summary(reg)
```

   Cela renvoie :

```
Formula: accueil ~ exp(a * population + b)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 3.504e-05  1.010e-06   34.71 3.81e-08 ***
b 4.089e+00  7.228e-02   56.58 2.05e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.13 on 6 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 4.631e-06
```

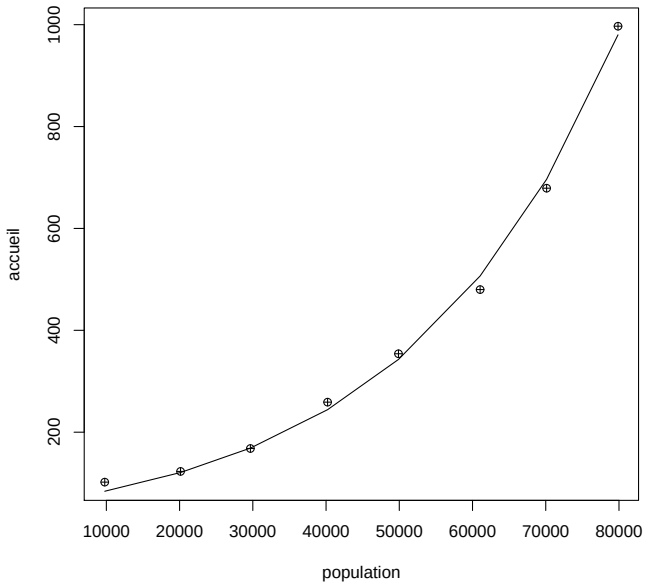
   Donner la formule mathématique (forme générique) du modèle considéré. Quelle est la nature de ce modèle ? Dès lors, quel est le degré de significativité du lien entre `accueil` et `population` ?

2. En admettant que le modèle considéré est correct, prédire la capacité d'accueil idéale moyenne d'un centre de loisirs dans une commune de 82768 habitants.

3. On considère les commandes suivantes :

```
plot(population, accueil, pch = 10)
lines(population, predict(reg))
```

Cela renvoie :



Puis on execute les commandes suivantes :

```
plot(population, summary(reg)$residuals, type = "h", col = "red")
abline(h = 0)
```

Cela renvoie un graphique affichant plusieurs bâtons. Utiliser le graphique précédent pour tracer, approximativement, ces bâtons (on ne se préoccupera pas des unités de mesures, juste le positionnement de ces bâtons par rapport à l'axe $y = 0$ (plutot dirigés vers le bas ou vers le haut ?) et leur longueur (grand ou petit, en fonction de ceux qui les précèdent et/ou succèdent)).

4. Proposer un modèle équivalent à `reg` en utilisant la fonction `lm`. Est-ce que les estimations ponctuelles des coefficients inconnus sont calculées de la même façon ? Justifier.

Exercice 28. On considère le jeu de données `superviseur`. Pour 27 entreprises, on dispose

- du nombre d'encadrants (variable `Y`),
- du nombre d'apprentis (variable `X`).

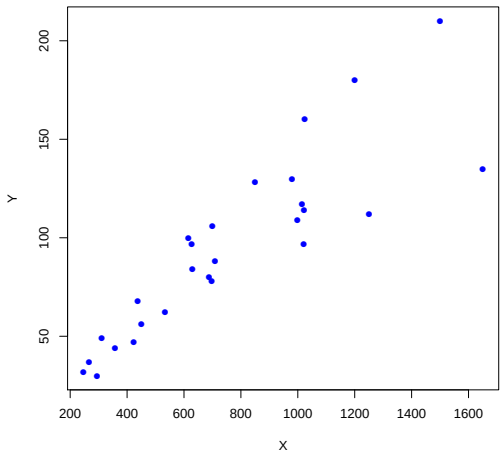
On souhaite expliquer du mieux possible `Y` en fonction de `X`.

1. Un premier modèle est considéré.

   (a) On exécute les commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/superviseur.txt",
header = T)
attach(w)
plot(X, Y, pch = 16, col = 4)
```

   Cela renvoie :



   Puis on enchaine avec :

```
reg1 = lm(Y ~ X)
summary(reg1)
```

   Cela renvoie :

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-53.294  -9.298  -5.579  14.394  39.119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.44806    9.56201   1.511    0.143
X            0.10536    0.01133   9.303 1.35e-09 ***
```

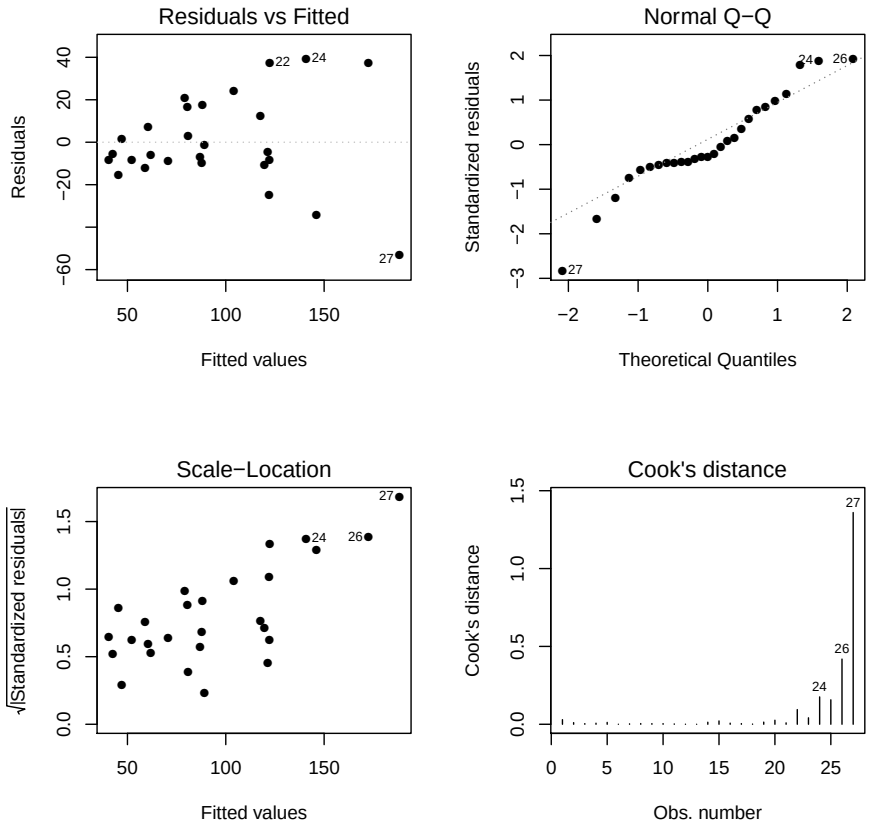
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

Residual standard error: 21.73 on 25 degrees of freedom
Multiple R-squared:  0.7759,    Adjusted R-squared:  0.7669
F-statistic: 86.54 on 1 and 25 DF,  p-value: 1.35e-09
```

Écrire la formule générique du modèle `reg1`. Avec les informations dont vous disposez, est-ce que ce modèle peut être considéré comme satisfaisant ?

(b) On fait :

```
par(mfrow = c(2, 2))
plot(reg1, which = c(1:4), add.smooth = FALSE, pch = 16)
par(mfrow = c(1, 1))
```

Cela renvoie :



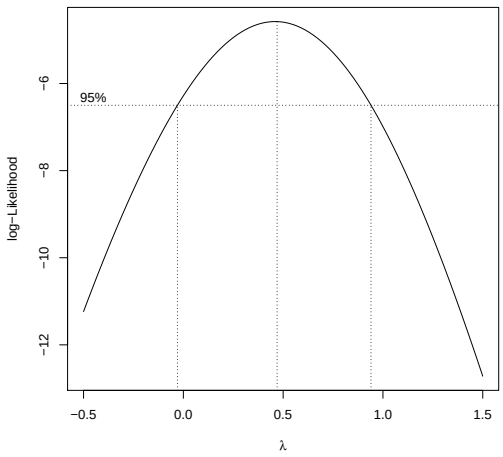
Analyser ces graphiques un à un.

2. On envisage un nouveau modèle.

(a) On exécute les commandes :

```
library(MASS)
boxcox(Y ~ X, data = w, lam = seq(-0.5, 1.5, 1/10))
```

Cela renvoie :



A quoi correspond ce graphique ? Que cherche-t-on à faire ?

(b) On exécute les commandes :

```
reg2 = lm(sqrt(Y) ~ X)
summary(reg2)
```

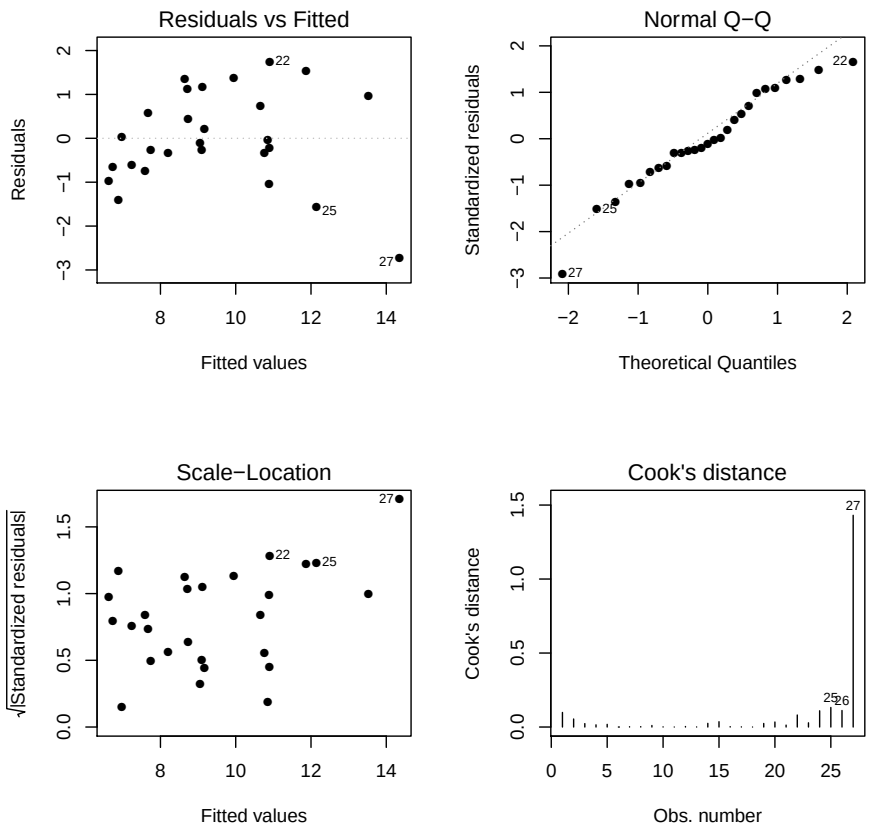
```
Call:
lm(formula = sqrt(Y) ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7309 -0.6279 -0.1091  0.8540  1.7403

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.2653598  0.4776900  11.023 4.34e-11 ***
X           0.0055058  0.0005658   9.731 5.54e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.086 on 25 degrees of freedom
Multiple R-squared:  0.7911,    Adjusted R-squared:  0.7828
F-statistic: 94.69 on 1 and 25 DF,  p-value: 5.543e-10
```

Écrire la formule générique du modèle `reg2`. Pourquoi ce modèle est justifié ? Avec les informations dont vous disposez, est-ce que ce modèle peut être considéré comme satisfaisant ?

(c) On enchaîne en faisant :

```
par(mfrow = c(2, 2))
plot(reg2, which = c(1:4), add.smooth = FALSE, pch = 16)
par(mfrow = c(1, 1))
```

Cela renvoie :



Analyser ces graphiques un à un.

3. Dès lors, on envisage encore un nouveau modèle.

   (a) On exécute les commandes :

```
wi = 1 / (X ^2)
reg3 = lm(Y ~ X, weights = wi)
summary(reg3)
```

Cela renvoie :

```
Call:
lm(formula = Y ~ X, weights = wi)

Weighted Residuals:
      Min        1Q    Median        3Q       Max
-0.041477 -0.013852 -0.004998  0.024671  0.035427

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.803296   4.569745   0.832    0.413
X           0.120990   0.008999  13.445 6.04e-13 ***
```

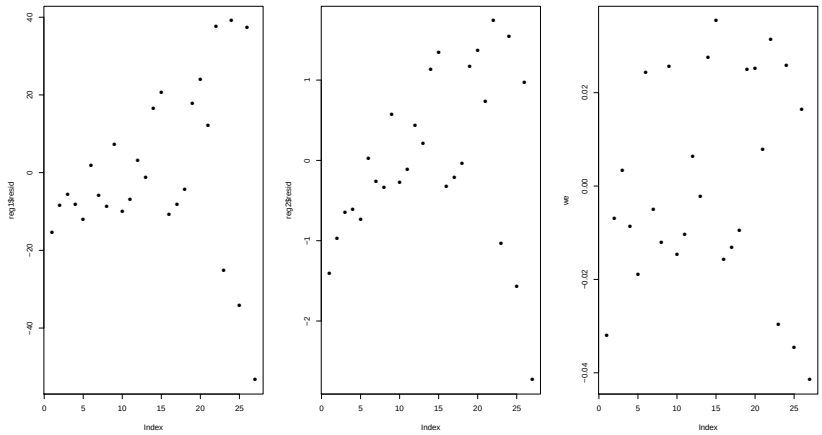
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02266 on 25 degrees of freedom
Multiple R-squared:  0.8785,    Adjusted R-squared:  0.8737
F-statistic: 180.8 on 1 and 25 DF,  p-value: 6.044e-13
```

Écrire la formule générique du modèle `reg3`. Que cherche-t-on à résoudre comme problème en introduisant le vecteur `wi` ? Avec les informations dont vous disposez, est-ce que ce modèle peut être considéré comme satisfaisant ?

(b) On enchaîne en faisant :

```
par(mfrow = c(1, 3))
plot(reg1$resid, pch = 16)
plot(reg2$resid, pch = 16)
rqwi = sqrt(wi)
we = rqwi * reg3$resid
plot(we, pch = 16)
par(mfrow = c(1, 1))
```

Cela renvoie :



Analyser ces graphiques un à un. Que peut-on en conclure à ce stade de l'analyse ?

4. Un nouveau modèle est proposé.

(a) On exécute les commandes :

```
X2 = X * X
reg4 = lm(log(Y) ~ X + X2)
summary(reg4)
```

Cela renvoie :

```
Call:
lm(formula = log(Y) ~ X + X2)
```

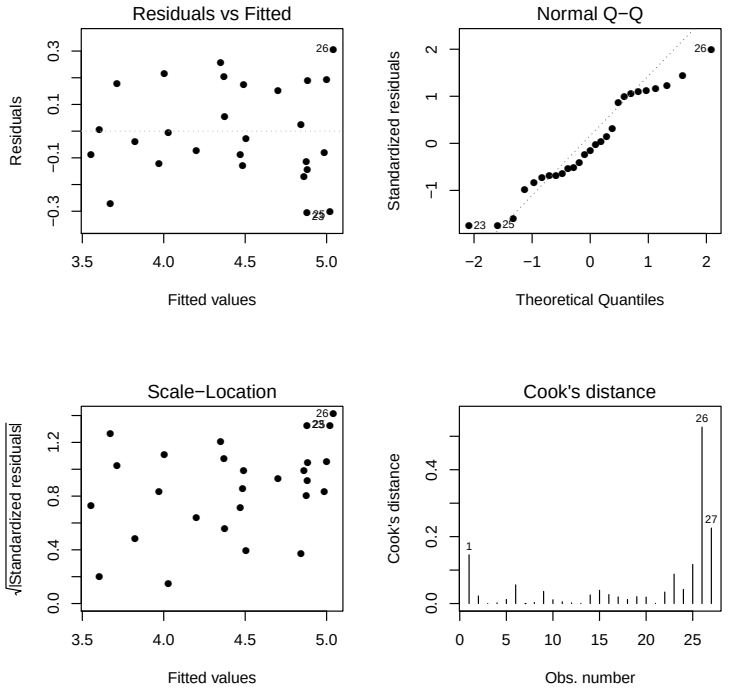
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.30589 -0.11705 -0.02707  0.17593  0.30657

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.852e+00  1.566e-01  18.205 1.50e-15 ***
X            3.113e-03  3.989e-04   7.803 4.90e-08 ***
X2          -1.102e-06  2.238e-07  -4.925 5.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1817 on 24 degrees of freedom
Multiple R-squared:  0.8857,    Adjusted R-squared:  0.8762
F-statistic: 92.98 on 2 and 24 DF,  p-value: 4.976e-12
```

Écrire la formule générique du modèle `reg4`. Proposer des commandes R alternatives qui aurait permis de définir un modèle identique à `reg4` (on s'attend à deux façons de faire différentes). Avec les informations dont vous disposez, est-ce que ce modèle peut être considéré comme satisfaisant ?

(b) On enchaîne en faisant :

```
par(mfrow = c(2, 2))
plot(reg4, which = c(1:4), add.smooth = FALSE, pch = 16)
par(mfrow = c(1, 1))
```

Cela renvoie :



Analyser ces graphiques un à un.

(c) On fait :

```
library(car)
vif(reg4)
```

Cela renvoie :

```
       X       X2
17.75033 17.75033
```

Que peut-on en conclure ?

5. Après l'analyse de `reg4`, un nouveau modèle est proposé.

(a) On exécute les commandes :

```
Xb = X - mean(X)
Xb2 = Xb * Xb
reg5 = lm(log(Y) ~ Xb + Xb2)
vif(reg5)
```

Cela renvoie :

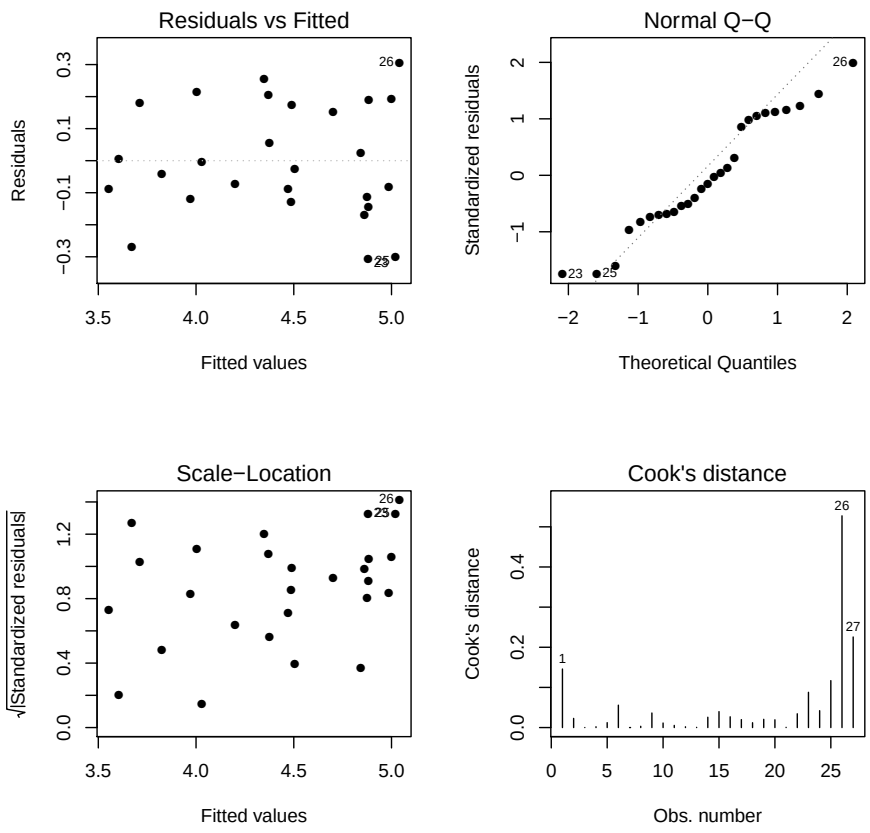
```
      Xb      Xb2
1.253509 1.253509
```

Écrire la formule générique du modèle `reg5`. Comparer les constructions des modèles `reg5` et `reg4`, et expliquer pourquoi `reg5` est préférable à `reg4`.

(b) On enchaîne en faisant :

```
par(mfrow = c(2, 2))
plot(reg5, which = c(1:4), add.smooth = FALSE, pch = 16)
par(mfrow = c(1, 1))
```

Cela renvoie :



Puis on fait :

```
shapiro.test(reg5$resid)$p.value
```

Cela renvoie :

```
[1] 0.2244218
```

Que peut-on dire de ces sorties ?

(c) On fait :

```
summary(reg5)
```

```
Call:
lm(formula = log(Y) ~ Xb + Xb2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30589 -0.11705 -0.02707  0.17593  0.30657

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.580e+00  4.640e-02  98.688  < 2e-16 ***
```

```
Xb           1.439e-03  1.060e-04  13.573 9.38e-13 ***
Xb2         -1.102e-06  2.238e-07  -4.925 5.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1817 on 24 degrees of freedom
Multiple R-squared:  0.8857,    Adjusted R-squared:  0.8762
F-statistic: 92.98 on 2 and 24 DF,  p-value: 4.976e-12
```

Que peut-on dire de cette sortie ?

(d) On exécute les commandes suivantes :

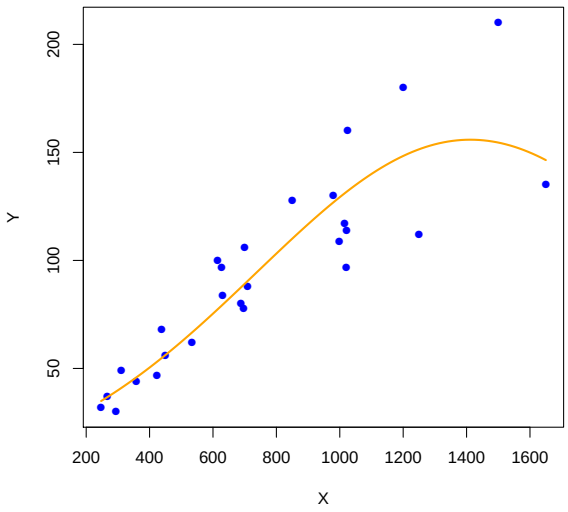
```
a = predict(reg5, data.frame(Xb = 800 - mean(X), Xb2 = (800 -
mean(X))^2))
a
```

Cela renvoie :

```
         1
4.636296
```

Avec l'aide de `a`, prédire le nombre moyen d'encadrants nécessaires dans une entreprise pour superviser 800 apprentis.

(e) Sans écrire de valeurs numériques (en utilisant, par exemple, les commandes `reg5$coef` ou `mean(X)`), proposer des commandes R permettant d'obtenir l'ajustement du nuage de points par le modèle `reg5` décrit dans le graphique ci-dessous.



(f) Commenter le graphique précédent. Est-ce que l'utilisation du modèle `reg5` pour avoir une prédiction moyenne de `Y` avec `X` supérieure à 1700 est fiable ?

Exercice 29. On s'intéresse au jeu de données `cats` de la librairie `MASS`. Pour 144 individus, on dispose

- du poids du cœur (variable `Hwt`),
- du poids du corps (variable `Bwt`),
- de leur sexe (variable `Sex` à 2 modalités : `F` pour femme et `M` pour homme).

On souhaite expliquer du mieux possible la variable `Hwt` en fonction des autres variables.

1. On exécute les commandes R suivantes :

```
library(MASS)
attach(cats)
reg = lm(Hwt ~ Bwt * Sex)
summary(reg)
```

Cela renvoie :

```
Call:
lm(formula = Hwt ~ Bwt * Sex)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7728 -1.0118 -0.1196  0.9272  4.8646

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9813     1.8428   1.618 0.107960
Bwt           2.6364     0.7759   3.398 0.000885 ***
SexM         -4.1654     2.0618  -2.020 0.045258 *
Bwt:SexM      1.6763     0.8373   2.002 0.047225 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.442 on 140 degrees of freedom
Multiple R-squared:  0.6566,    Adjusted R-squared:  0.6493
F-statistic: 89.24 on 3 and 140 DF,  p-value: < 2.2e-16
```

   (a) Quel est le modèle statistique considéré ? Donner une écriture mathématique du modèle (écriture générique), avec les estimations associées.

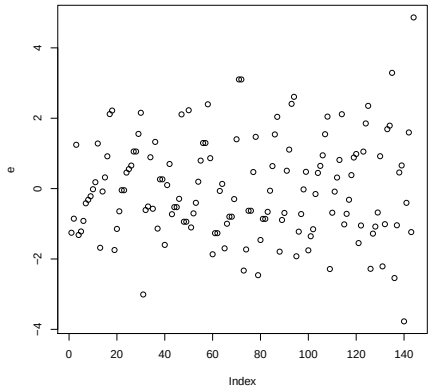
   (b) Est-ce que les influences combinées de `Bwt` et `Sex` génère un effet supplémentaire dans l'explication de `Hwt` ?

   (c) En supposant que le modèle est bon, donner une prédiction moyenne de la valeur de `Hwt` pour une femme vérifiant `Bwt = 2.55`.

2. En vue de valider le modèle, on fait une étude grossière des résidus du modèle. On exécute les commandes R suivantes :

```
e = residuals(reg)
plot(e)
```

Cela renvoie :

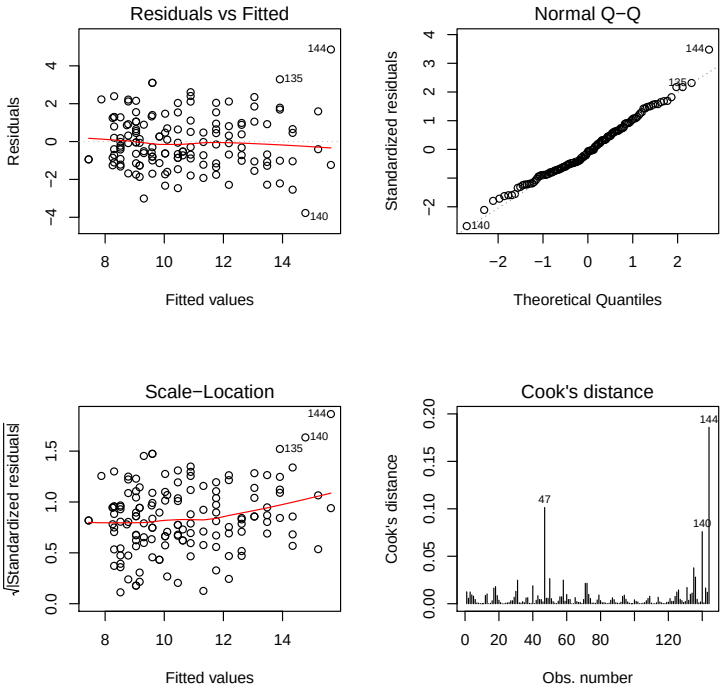


Que pouvez-vous dire sur ce graphique des résidus ? Est-ce que tout semble bon ?

3. On décide d'analyser plus finement les hypothèses. On exécute les commandes R suivantes :

```
par(mfrow = c(2, 2))
plot(reg, 1:4)
```

Cela renvoie :



Faire une analyse de chacun de ces graphiques, avec identification des problèmes possibles.

4. On exécute les commandes R suivantes :

```
shapiro.test(e)$p.value; bartlett.test(e, Sex)$p.value
```

Cela renvoie :

```
[1] 0.1806282
[1] 0.02453587
```

(a) Quels sont les tests statistiques considérés ?

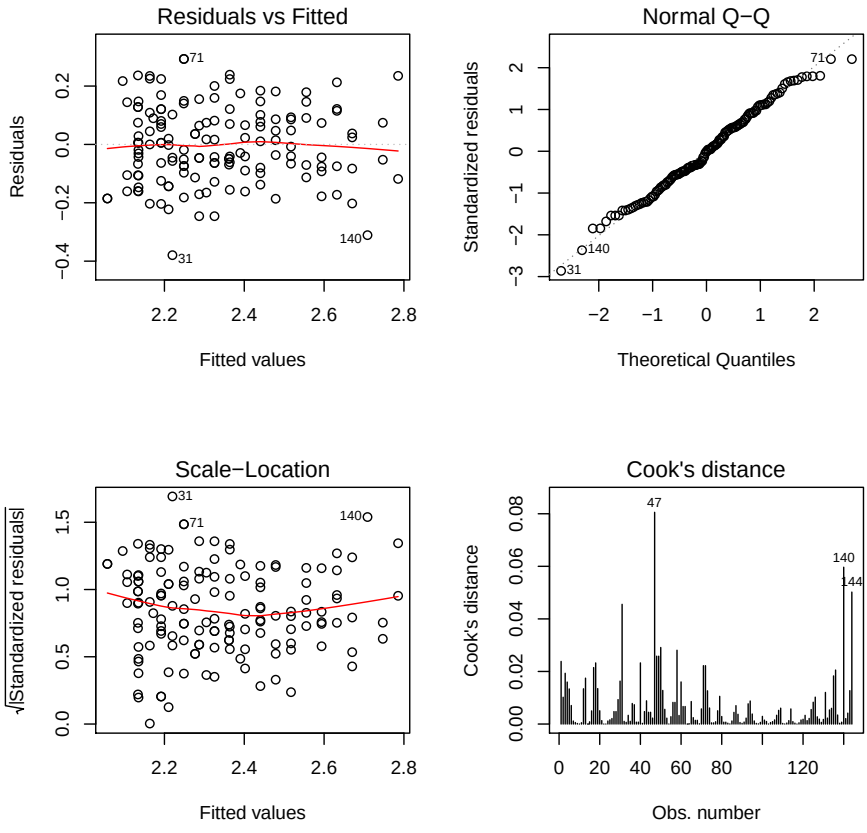
(b) Que peut-on dire des résultats numériques obtenus ?

(c) À la place du test statistique correspondant aux commandes : `bartlett.test(e, Sex)`, quel autre test statistique aurait-on pu utiliser pour analyser le même phénomène ?

5. On propose de résoudre d'un coup les problèmes potentiels. Pour ce faire, on exécute les commandes R suivantes :

```
reg2 = lm(log(Hwt) ~ Bwt * Sex)
par(mfrow = c(2, 2))
plot(reg2, 1:4)
```

Cela renvoie :



Pour compléter cette analyse visuelle, on exécute les commandes R suivantes :

```
e2 = residuals(reg2)
shapiro.test(e2)$p.value; bartlett.test(e2, Sex)$p.value
```

Cela renvoie :

```
[1] 0.4159872
[1] 0.6894161
```

(a) Quel est le modèle statistique considéré ? Donner une écriture mathématique du modèle (écriture générique).

(b) Est-ce que les principaux problèmes semblent être résolus ?

(c) De plus, on fait :

```
library(lmtest)
dwtest(reg2)$p.value
```

Cela renvoie :

```
[1] 0.008049465
```

Quel test statistique a t-on mis en œuvre ici ? Bien que la p-valeur vérifie p-valeur$<$ 0.05, on prétend qu'il n'y a absolument aucun problème de dépendance inquiétante dans les données. D'après-vous, qu'est-ce qui pourrait expliquer cette affirmation, dans la structure même du jeu de données ?

6. Le modèle `reg2` étant solide, on s'intéresse maintenant aux estimations des paramètres et à la qualité du modèle. On exécute les commandes R suivantes :

```
summary(reg2)
```

Cela renvoie :

```
Call:
lm(formula = log(Hwt) ~ Bwt * Sex)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37948 -0.09021 -0.00104  0.09133  0.29291

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.53962    0.17131   8.987 1.56e-15 ***
Bwt          0.28350    0.07213   3.931 0.000133 ***
SexM        -0.24939    0.19166  -1.301 0.195312
Bwt:SexM     0.09989    0.07784   1.283 0.201515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.134 on 140 degrees of freedom
Multiple R-squared:  0.6438,    Adjusted R-squared:  0.6362
F-statistic: 84.36 on 3 and 140 DF,  p-value: < 2.2e-16
```

Que remarque t-on par rapport au modèle défni par `reg` ?

7. On exécute les commandes R suivantes :

```
reg3 = lm(log(Hwt) ~ Bwt)
summary(reg3)
```

Cela renvoie :

```
Call:
lm(formula = log(Hwt) ~ Bwt)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37985 -0.09095 -0.00508  0.09634  0.28459

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.34156    0.06382   21.02   <2e-16 ***
Bwt          0.36618    0.02307   15.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1339 on 142 degrees of freedom
Multiple R-squared:  0.6395,    Adjusted R-squared:  0.637
F-statistic: 251.9 on 1 and 142 DF,  p-value: < 2.2e-16
```

(a) Quel est le modèle statistique considéré ? Donner une écriture mathématique du modèle (écriture générique), avec les estimations associées.

(b) On exécute les commandes R suivantes :

```
summary(reg2)$adj.r.squared <= summary(reg3)$adj.r.squared ;
AIC(reg3) <= AIC(reg2); BIC(reg3) <= BIC(reg2)
```

Cela renvoie :

```
[1] TRUE
[1] TRUE
[1] TRUE
```

Que peut-on en conclure ?

(c) Après vérification, le modèle `reg3` est bon. Dès lors,

i. donner un intervalle de confiance pour le coefficient de régression associé à `Bwt` au niveau 95% (pour le $t_\alpha(\nu)$, on utilisera l'approximation normale : $t_\alpha(\nu) \approx z_\alpha \approx 1.96$),

ii. donner une prédiction moyenne de la valeur de `Hwt` pour une femme vérifiant `Bwt = 2.55`. Comparer avec le résultat obtenu à la question 1− c).

Exercice 30. On considère le jeu de données `state`. Pour 50 états américains, on dispose

- du nombre d'habitants (unité 1000) (1er juillet 1975) (variable `Population`),
- du revenu par habitant (1974) (variable `Income`),
- du pourcentage d'analphabétisme de la population (1970) (variable `Illiteracy`),
- de l'espérance de vie en années (1969-1971) (variable `Life.Exp`),
- du taux de meurtre et d'homicide involontaire coupable pour 100 000 habitants (1976) (variable `Murder`),
- du pourcentage de diplômés du secondaire (1970) (variable `HS.Grad`),
- du nombre moyen de jours avec la température min 32 degrés (1931-1960) dans la capitale ou la grande ville (variable `Frost`),
- de la superficie en miles carrés (variable `Area`).

On souhaite expliquer du mieux possible la variable `Life.Exp` en fonction des autres variables.

1. On exécute les commandes R suivantes :

```
data(state)
w = data.frame(state.x77, row.names = state.abb, check.names = T)
reg = lm(Life.Exp ~ ., data = w)
summary(reg)
```

Cela renvoie :

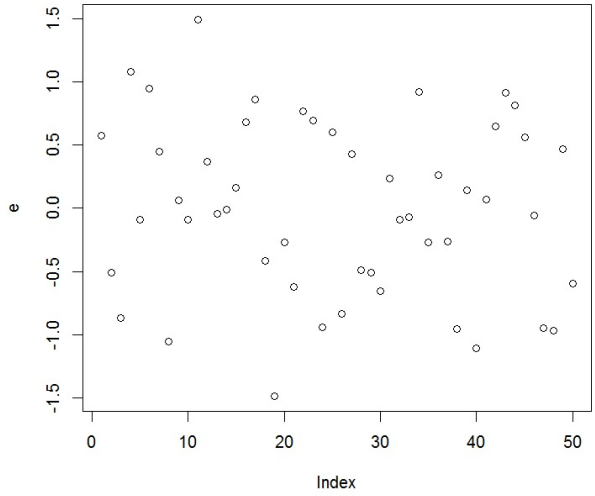
```
Call:
lm(formula = Life.Exp ~ ., data = w)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48895 -0.51232 -0.02747  0.57002  1.49447

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .
Income      -2.180e-05  2.444e-04  -0.089   0.9293
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
Area        -7.383e-08  1.668e-06  -0.044   0.9649
---

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared:  0.7362,    Adjusted R-squared:  0.6922
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

Quel est le modèle statistique considéré ? En admettant que les hypothèses de base soient satisfaites, est-ce que le modèle `reg` est satisfaisant ?

2. On exécute les commandes :

```
e = residuals(reg)
plot(e)
```

Cela renvoie :



Qu'étudie t-on ici ? Est-ce que vous décelez un problème ?

3. On execute les commandes :

```
library(car)
vif(reg)
```

Cela renvoie :

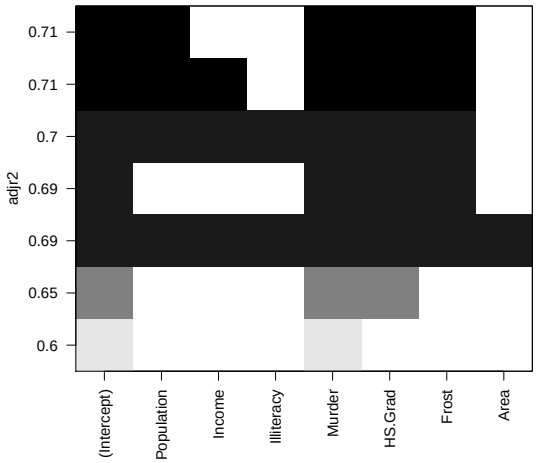
```
Population     Income Illiteracy     Murder    HS.Grad      Frost       Area
  1.499915   1.992680   4.403151   2.616472   3.134887   2.358206   1.789764
```

Qu'étudie t-on ici ? Est-ce que vous décelez un problème ?

4. On exécute les commandes :

```
library(leaps)
v = regsubsets(Life.Exp ~ ., w, method = "exhaustive")
plot(v, scale = "adjr2")
```

Cela renvoie :



Qu'étudie t-on ici ? Quel modèle ressort de ce graphique et pourquoi ? Proposer des commandes R associées. On notera ce modèle `reg2`.

5. On exécute les commandes :

```
summary(reg2)
```

Cela renvoie :

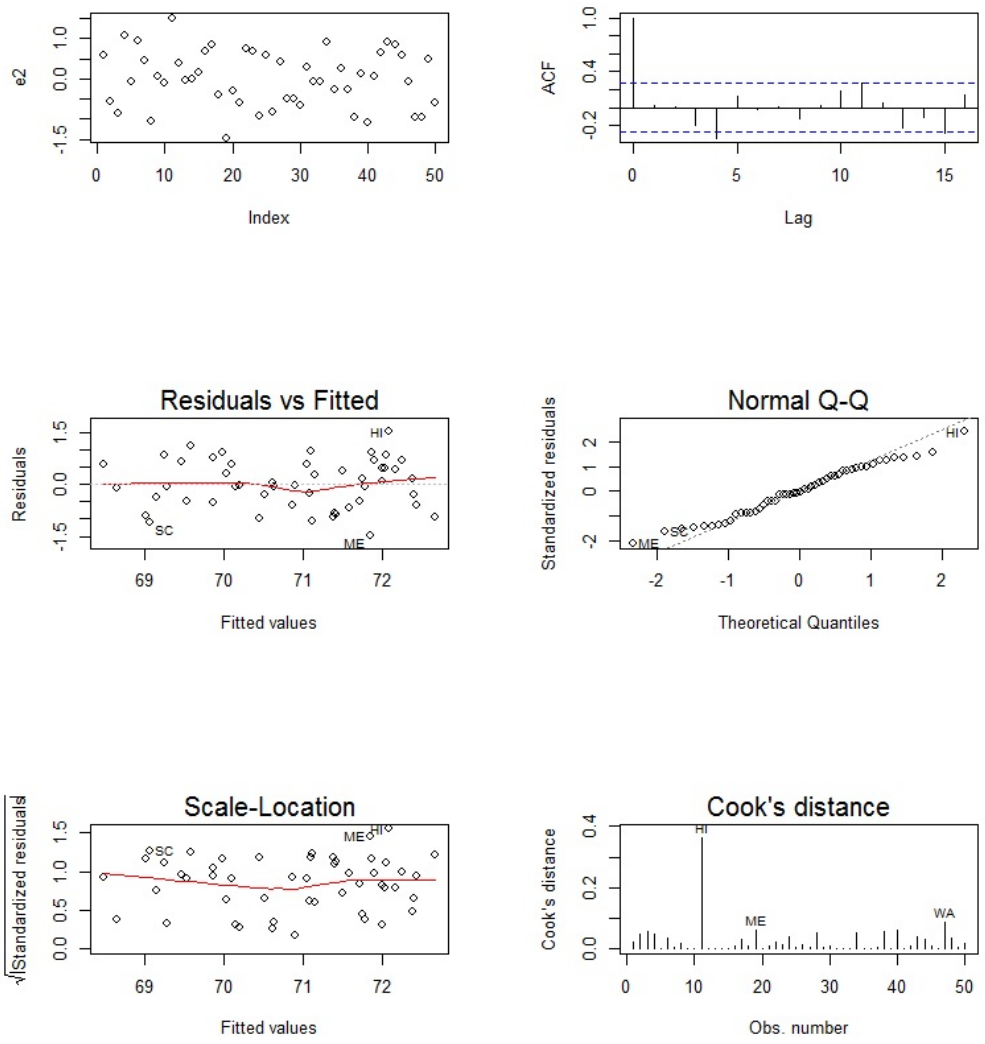
```
Call:
lm(formula = Life.Exp ~ . - Income - Illiteracy - Area, data = w)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47095 -0.53464 -0.03701  0.57621  1.50683

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
---

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736,     Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

En admettant que les hypothèses de base soient satisfaites, est-ce que le modèle `reg2` est satisfaisant ?

6. On exécute les commandes :

```
e2 = residuals(reg2)
par(mfrow = c(3, 2))
plot(e2)
acf(e2)
plot(reg2, 1:4)
```

Cela renvoie :



Qu'étudie t-on ici, graphique par graphique ? Est-ce que vous décelez un problème ?

7. On exécute les commandes :

```
shapiro.test(e2)
```

Cela renvoie :

```
        Shapiro-Wilk normality test

data:  e2
W = 0.97935, p-value = 0.525
```

Puis :

```
library(lmtest)
bptest(reg2)
```

Cela renvoie :

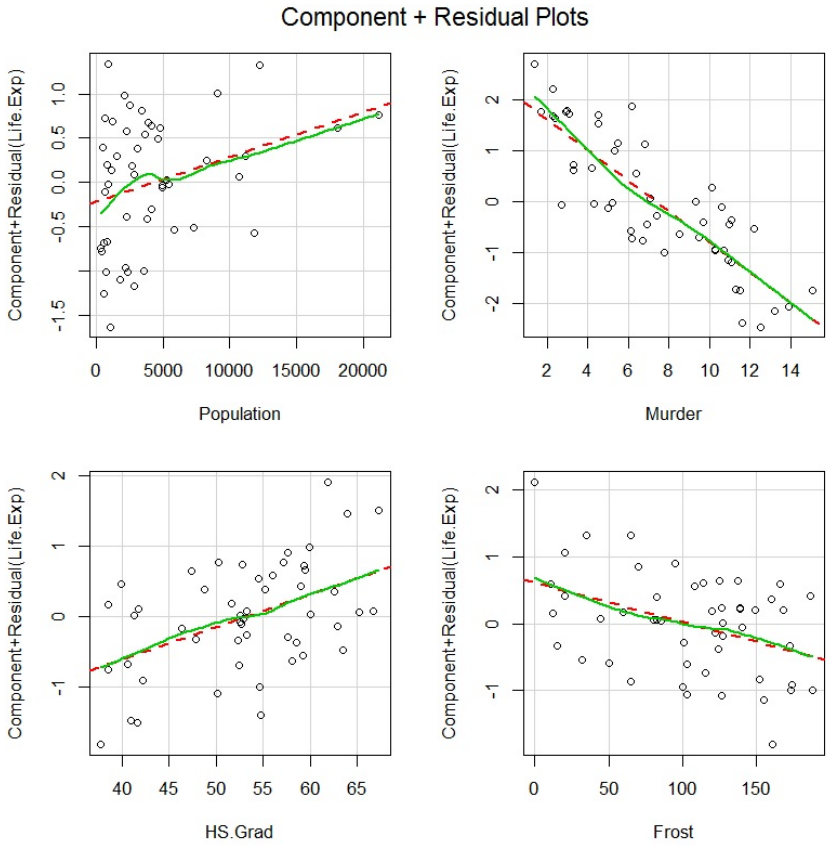
```
        studentized Breusch-Pagan test

data:  reg2
BP = 6.2721, df = 4, p-value = 0.1797
```

Qu'étudie t-on ici ? Est-ce que vous décelez un problème ?

8. On exécute les commandes :

```
library(car)
crPlots(reg2)
```

Cela renvoie :



Qu'étudie t-on ici ? Qu'en déduisez-vous ?

9. Faire une conclusion de ce qui précède.

Exercice 31. On considère le jeu de données `w` donné par les commandes R suivantes :

```
X1 = c(32, 64, 96, 118, 126, 144, 152.5, 158)
X2 = c(574.6, 468.6, 538.0, 385.9, 279.8, 91.3, 141.7, 127.1)
Y = c(99.5, 104.8, 108.5, 100, 86, 64, 35.3, 15)
w = data.frame(Y, X1, X2)
```

On souhaite expliquer `Y` en fonction de `X1` et `X2`.

1. On exécute les commandes :

   ```
   reg = lm(Y ~ X1 + X2)
   summary(reg)
   ```

   Cela renvoie :

   ```
   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   (Intercept) 21.80500   79.27293   0.275    0.794
   X1           0.02587    0.43296   0.060    0.955
   X2           0.15943    0.10008   1.593    0.172

   Residual standard error: 22.04 on 5 degrees of freedom
   Multiple R-squared:  0.7187,    Adjusted R-squared:  0.6062
   F-statistic: 6.387 on 2 and 5 DF,  p-value: 0.04197
   ```

   Est-ce que le modèle `reg` est satisfaisant ?

2. On exécute les commandes :

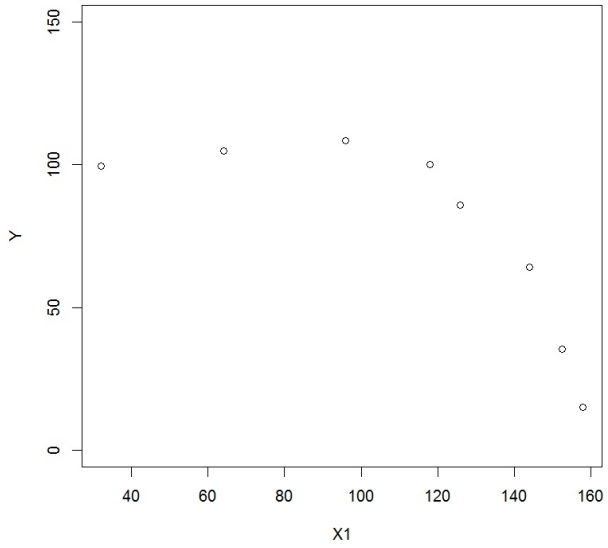
   ```
   plot(X1, X2)
   ```

   Cela affiche un graphique, dans lequel on constate immédiatement un phénomène. On a finalement pas intérêt à mettre la variable `X2` dans le modèle. Quel peut-être ce phénomène ?

3. On exécute les commandes suivantes :

   ```
   plot(X1, Y, ylim = c(0, 150))
   ```

Cela renvoie :



Manifestement, le lien entre `Y` et `X1` est plus nonlinéaire que linéaire. On considère alors 3 modèles de régression, notés `reg1`, `reg2` et `reg3`, donnés par les commandes suivantes :

```
reg1 = lm(Y ~ X1)
reg2 = lm(Y ~ poly(X1, 2, raw = TRUE))
reg3 = lm(Y ~ poly(X1, 3, raw = TRUE))
```

Expliciter les formules mathématiques associées à `reg2` et `reg3`. Proposer des commandes R donnant également `reg2` et `reg3` sans utiliser la commande `poly`.

4. On exécute les commandes suivantes :

```
AIC(reg1) ; AIC(reg2) ; AIC(reg3); BIC(reg1) ; BIC(reg2) ; BIC(reg3)
```

Cela renvoie :

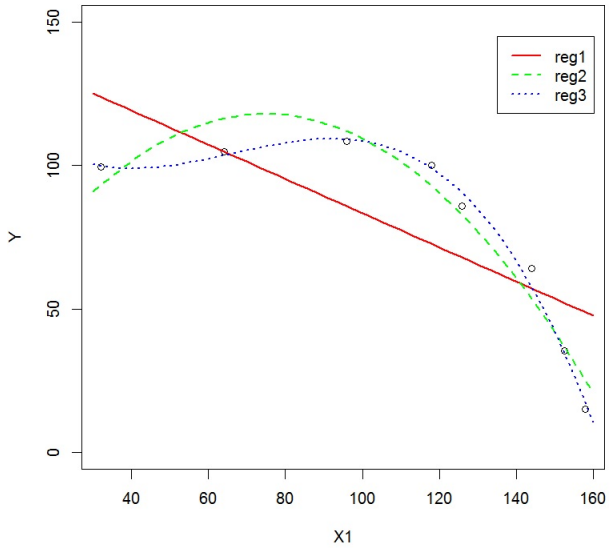
```
[1] 77.71092
[1] 63.01831
[1] 49.56936
[1] 77.94924
[1] 63.33607
[1] 49.96657
```

Quel modèle ressort de ces résultats numériques ? Pourquoi ?

5. On exécute les commandes suivantes :

```
xx = seq(30, 160, length = 50)
plot(X1, Y, ylim = c(0, 150))
lines(xx, predict(reg1, data.frame(X1 = xx)), col = "red", lty = 1, lwd
= 2)
lines(xx, predict(reg2, data.frame(X1 = xx)), col = "green", lty = 2,
lwd = 2)
lines(xx, predict(reg3, data.frame(X1 = xx)), col = "blue", lty = 3, lwd
= 2)
legend(135, 145, legend=c("reg1", "reg2", "reg3"), col=c("red", "green",
"blue"), lty=1:3, lwd = rep(2,3))
```

Cela renvoie :



Est-ce que ce graphique confirme le résultat de la question précédente ? Pourquoi ?

Exercice 32. On considère le jeu de données "anesthésie" :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/anesthésie.txt",
header = T)
```

Trente patients ont reçu un certain niveau de dosage d'agent anesthésique pendant 15 minutes. Puis une incision leur est faite. Il est ensuite noté si le patient a bougé ou pas lors de l'incision. Ainsi, pour chaque patient, on dispose :

- du dosage de l'agent anesthésique pendant 15 minutes (variable $X1$),
- du fait qu'il ait bougé ou pas (variable Y, avec $Y = 1$ pour bougé).

On souhaite expliquer Y à partir de $X1$.

1. On exécute les commandes R suivantes :

```
attach(w)
library(stats)
reg = glm(Y ~ X1, family = binomial)
summary(reg)
predict.glm(reg, data.frame(X1 = 1.25), type = "response")
confint.default(reg, level = 0.95)
```

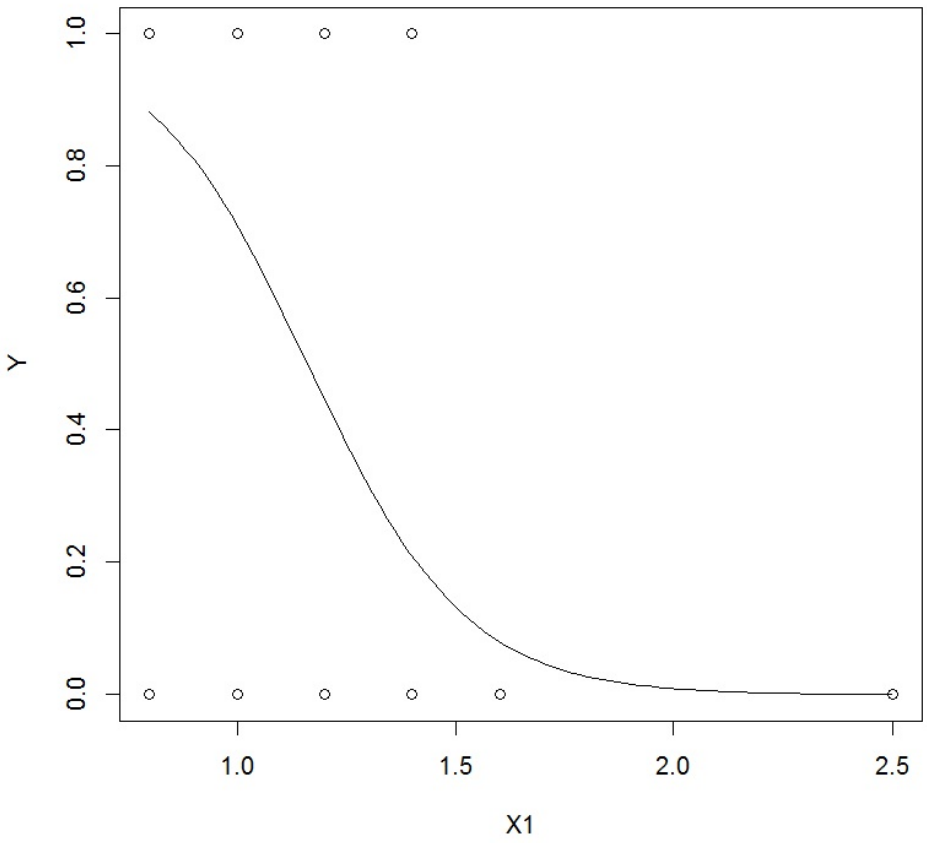
Cela renvoie :

```
> summary(reg)

Call:
glm(formula = Y ~ X1, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.06900  -0.68666  -0.03413   0.74407   1.76666

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    6.469      2.418   2.675  0.00748 **
X1            -5.567      2.044  -2.724  0.00645 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 27.754  on 28  degrees of freedom
AIC: 31.754

Number of Fisher Scoring iterations: 5

> predict.glm(reg, data.frame(X1 = 1.25), type="response")
       1
0.379946
> confint.default(reg, level=0.95)
                2.5 %    97.5 %
(Intercept)  1.728560 11.208790
X1          -9.572126 -1.561398
```

Quel est le modèle considéré ? Est-ce qu'un patient ayant eu pour dosage `X1 = 1.25` a plus de chance de bouger que de ne pas bouger ?

2. On exécute les commandes R suivantes :

```
plot(X1, Y)
curve(predict(reg, data.frame(X1 = x), type = "response"), add = T)
```



Que représente ce graphique ?

3. On exécute les commandes R suivantes :

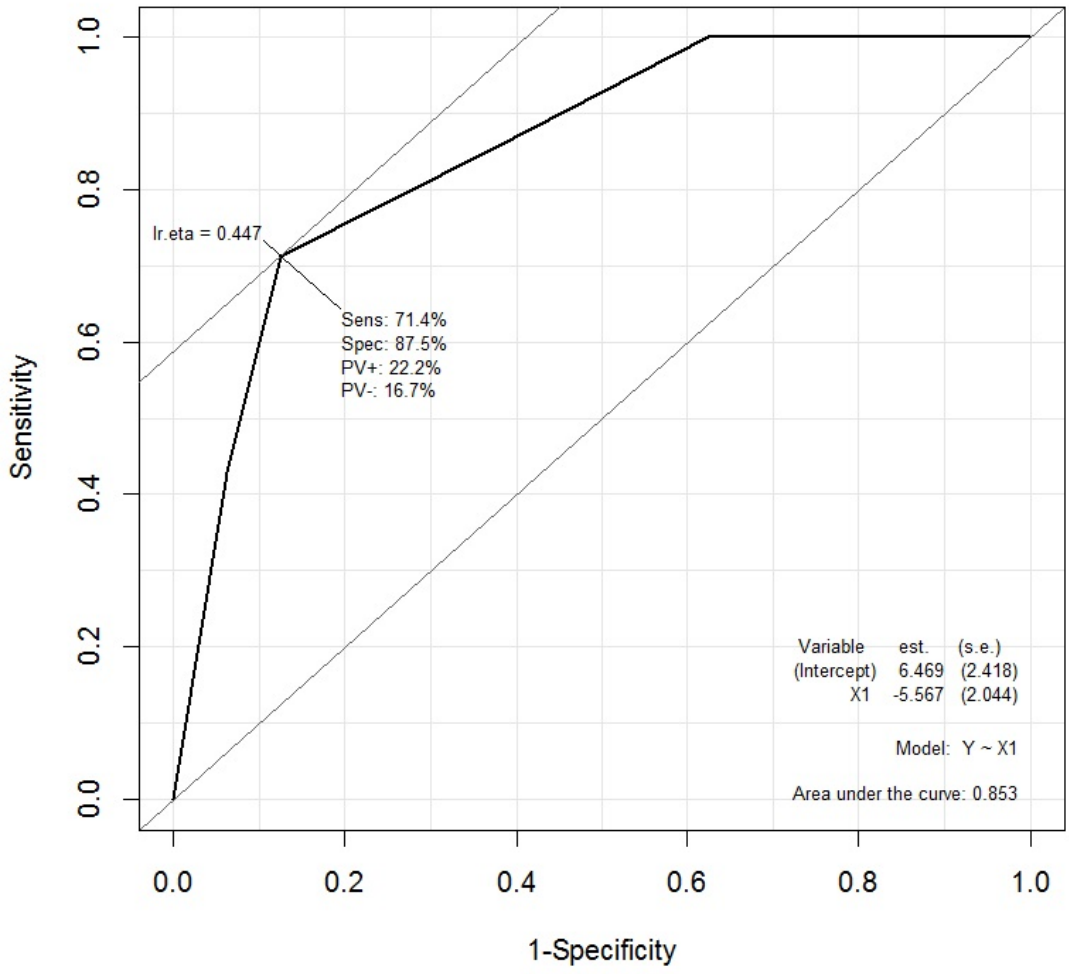
```
pred.prob = predict(reg, type = "response")
pred.mod = factor(ifelse(pred.prob >0.5, "1", "0"))
mc = table(Y, pred.mod)
t = (mc[1, 2] + mc[2, 1]) / sum(mc)
t
```

Cela renvoie $t = 0.2$.

Que représente cette quantité ? Est-ce que le résultat est satisfaisant ?

4. On exécute les commandes R suivantes :

```
library(Epi)
ROC(form = Y ~ X1, plot = "ROC")
```



Quelle est l'aire sous la courbe ROC ? Est-ce que cela est satisfaisant ?

Exercice 33. Une entreprise de vente par correspondance souhaite étudier le lien possible entre le montant total annuel des dépenses multimédia en euros d'un consommateur (variable `DEP`) et le fait qu'il réponse favorablement à une offre de promotion en fin d'année. Le responsable marketing prélève au hasard 12 consommateurs. Pour chacun d'eux, il note `PROMO` $= 1$ s'il a répondu favorablement à l'offre de promotion et 0 sinon (créant ainsi une variable `PROMO`). Il note le montant annuel de ses commandes en multi-média `DEP`. Les résultats, transcrits dans une `data.frame` nommée `MULTIMED`, sont les suivants :

PROMO	DEP
1	1000
1	990
0	55
1	1100
0	70
0	90
0	100
1	420
1	890
1	840
0	700
0	350

1. L'observation 1 ou 0 est la réalisation d'une variable aléatoire réelle Y. Quelle est la loi suivie par cette variable aléatoire réelle ? Sachant que `Dep` $= x$, une modélisation possible consiste à relier la probabilité $p(x) = \mathbb{P}(\{Y = 1\} \mid \{\texttt{Dep} = x\})$ à x par une expression où apparaissent deux coefficients inconnus β_0 et β_1. Ecrire cette expression.

2. On execute des commandes qui renvoient, entre autre :

```
Call:
glm(formula = Y ~ X, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.62555  -0.26147  -0.01377   0.32829   1.62214

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.024489   2.283132  -1.763   0.0780 .
X            0.007193   0.003559   2.021   0.0432 *
---
```

   (a) Écrire toutes les commandes faites pour arriver à cette sortie.

   (b) Le montant annuel des commandes influe-t-il significativement sur la probabilité de répondre favorablement à l'offre de promotion ?

   (c) Sachant qu'un client a passé dans l'année un montant de commandes égal à 500 euros, quelle est la probabilité que ce client réponde favorablement à l'offre de promotion ?

Exercice 34. Une banque s'intéresse au comportement de remboursement ou de non-remboursement d'un futur emprunteur pour un prêt immobilier. Les variables suivantes sont considérées :

- Y : comportement de remboursement ou de non-remboursement d'un emprunteur,
- Y_* : capacité de remboursement de l'emprunteur en euros,
- X_1 : montant du prêt demandé en euros,
- X_2 : revenu annuel de l'emprunteur en euros,
- X_3 : âge de l'emprunteur en années,
- X_4 : nombre de mois que l'emprunteur est client dans la banque.

Seules des informations sur (Y, X_1, X_2, X_3, X_4) sont disponibles ; Y_* est une variable dite latente.
On dispose de $n = 567$ observations de (Y, X_1, X_2, X_3, X_4) constituant les données.
En considérant ces variables comme aléatoires, une modélisation possible pour Y est :

$$Y = \begin{cases} 1 & \text{si } Y_* \geq 900, \\ 0 & \text{sinon.} \end{cases}$$

De plus, une liaison linéaire entre Y_* et $X_1, \ldots, X_4$ est envisageable : on peut modéliser Y_* par une *rlm* :

$$Y_* = (900 + \beta_0) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \gamma,$$

où $\beta_0, \ldots, \beta_4$ sont 5 coefficients inconnus et γ est une *var*

- symétrique (γ et $-\gamma$ suivent la même loi) de densité f_γ et de fonction de répartition F_γ,
- indépendante de $X_1, \ldots, X_4$.

Pour tout $x = (x_1, \ldots, x_4)$, on souhaite estimer la probabilité :

$$p(x) = \mathbb{P}(\{Y = 1\} | \{(X_1, \ldots, X_p) = x\}).$$

1. Montrer que
$$p(x) = F_\gamma \left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4\right).$$

2. On suppose maintenant que γ suit la loi logistique $\mathcal{L}(1)$, ce qui nous amène au modèle de régression logistique standard :
$$\begin{aligned} p(x) &= \text{logit}^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4) \\ &= \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}. \end{aligned}$$
Remplacer le symbole □ dans les commandes R ci-dessous pour considérer le modèle souhaité :

```
reg = glm(Y ~ X1 + X2 + X3 + X4, family = □)
```

3. Le test de Hosmer-Lemeshow renvoie une p-valeur $= 0.78$. Que peut-on en conclure ?

4. Quel objet mathématique résulte des commandes R ci-dessous ?

```
pred.prob = predict(reg, type = "response")
pred.mod = factor(ifelse(pred.prob > 0.5, "1", "0"))
mystere = table(Y, pred.mod)
mystere
```

5. La matrice de confusion associée au modèle de régression logistique est :

$$MC = \begin{pmatrix} 275 & 40 \\ 28 & 224 \end{pmatrix}$$

Quel est le taux d'erreur ? Est-il satisfaisant ?

6. On dispose des valeurs de $X_1, \ldots, X_4$ pour un client ayant fait un dossier d'emprunt dans la banque, à savoir : $X_1 = 20000$, $X_2 = 36000$, $X_3 = 39$ et $X_4 = 65$.

Que peut-on conclure des commandes R suivantes et du résultat ?

```
predict.glm(reg, data.frame(X1 = 20000, X2 = 36000, X3 = 39, X4 = 65),
type = "response")
```

Cela renvoie 0.79.

Exercice 35. On considère le jeu de données `mtcars` de la librairie `datasets`. Les données ont été extraites de la revue Motor Trend US 1974 et comprend la consommation de carburant ainsi que 10 aspects de la conception automobile et de performance pour 32 voitures (de modèles 1973-74). On exécute les commandes R suivantes :

```
library(datasets)
w = subset(mtcars, select = c(mpg, am, vs))
head(w)
```

Cela renvoie :

	mpg	am	vs
Mazda RX4	21.00	1.00	0.00
Mazda RX4 Wag	21.00	1.00	0.00
Datsun 710	22.80	1.00	1.00
Hornet 4 Drive	21.40	0.00	1.00
Hornet Sportabout	18.70	0.00	0.00
Valiant	18.10	0.00	1.00

La variable à expliquer est `vs` et les variables explicatives sont `mpg` et `am`. La variable `vs` est binaire, la variable `mpg` est quantitative et la variable `am` est binaire.

1. Expliciter le modèle considéré dans les commandes suivantes :

```
reg = glm(vs ~ mpg * am, w, family = binomial)
```

2. On exécute les commandes suivantes :

```
summary(reg)
```

Cela renvoie :

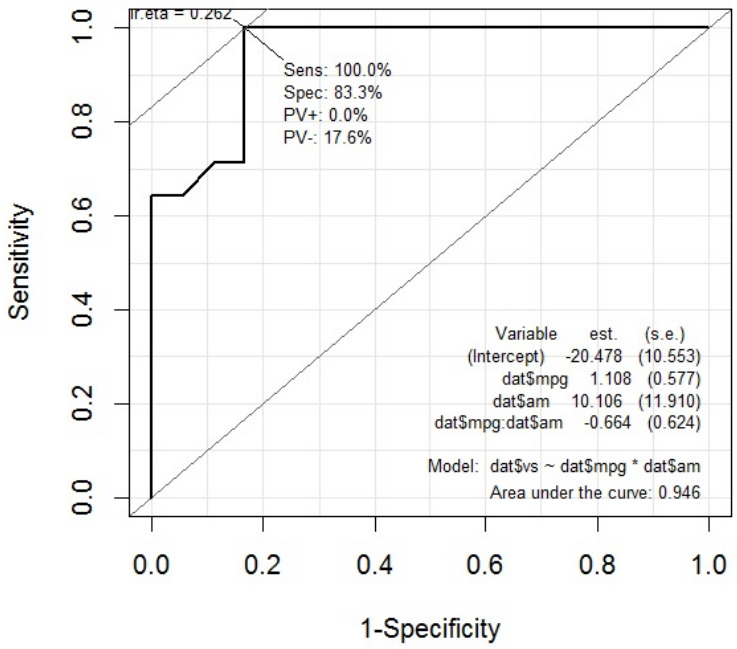
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-20.4784	10.5525	-1.94	0.0523	.
mpg	1.1084	0.5770	1.92	0.0547	.
am	10.1055	11.9104	0.85	0.3962	
mpg:am	-0.6637	0.6242	-1.06	0.2877	

À partir de ce tableau, calculer la valeur renvoyée par la commande suivante :

```
predict(reg, data.frame(mpg = 19.5, am = 1), type = "response")
```

3. On exécute les commandes suivantes :

```
library(Epi)
ROC(form = w$vs ~ w$mpg * w$am, plot = "ROC")
```



En utilisant le graphique ci-dessus, expliquer pourquoi le modèle considéré est bon.

Exercice 36. On a relevé les informations des caractères sexe (caractère binaire de modalités "homme" (h) et "femme" (f)), poids (en kilogrammes) et taille (en centimètres) sur un échantillon d'hommes et de femmes. Ces informations sont collectées dans le jeu de données `Quetelet` qui contient trois colonnes : `sexe`, `poids` et `taille`. Pour en savoir plus sur les données, on fait les commandes R suivantes :

```
Quetelet = read.csv("https://chesneau.users.lmno.cnrs.fr/quetelet.csv",
header = T)
head(Quetelet)
```

Cela renvoie :

```
  sexe poids taille
1    h    60    170
2    f    57    169
3    f    51    172
4    f    55    174
5    f    50    168
6    f    50    161
```

Puis on fait :

```
str(Quetelet)
```

Cela renvoie :

```
'data.frame':   66 obs. of  3 variables:
$ sexe  : chr  "h" "f" "f" "f" ...
$ poids : int  60 57 51 55 50 50 48 72 52 64 ...
$ taille: int  170 169 172 174 168 161 162 189 160 175 ...
```

On s'intéresse à l'indice de masse corporel (imc ou indice de Quetelet) d'un individu à partir des données. La formule de cet indice est la suivante :

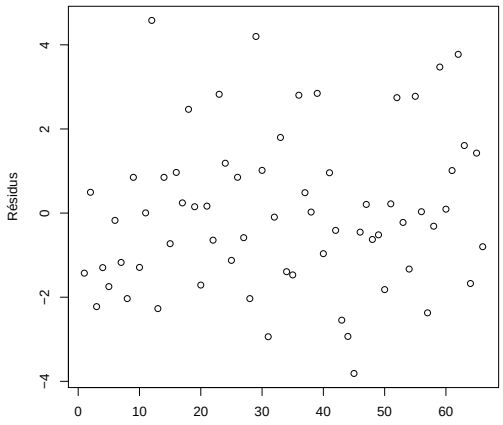
$$\text{imc} = \frac{\text{poids en kilogrammes}}{(\text{taille en mètres})^2}.$$

Cet indice permet de mesurer la corpulence de l'homme adulte. On adopte la classification suivante : maigreur (imc inférieur à 18.5), normal (imc de 18.5 à 25), risque de surpoids (imc de 25 à 30), et obésité (imc supérieur à 30).

1. Dans un premier temps, on souhaite expliquer l'imc d'un individu en fonction de son sexe.

   (a) À partir des informations ci-dessus, proposer toutes les commandes R pour créer un modèle adapté.

(b) Donner les commandes permettant d'obtenir le graphique des résidus associé suivant :

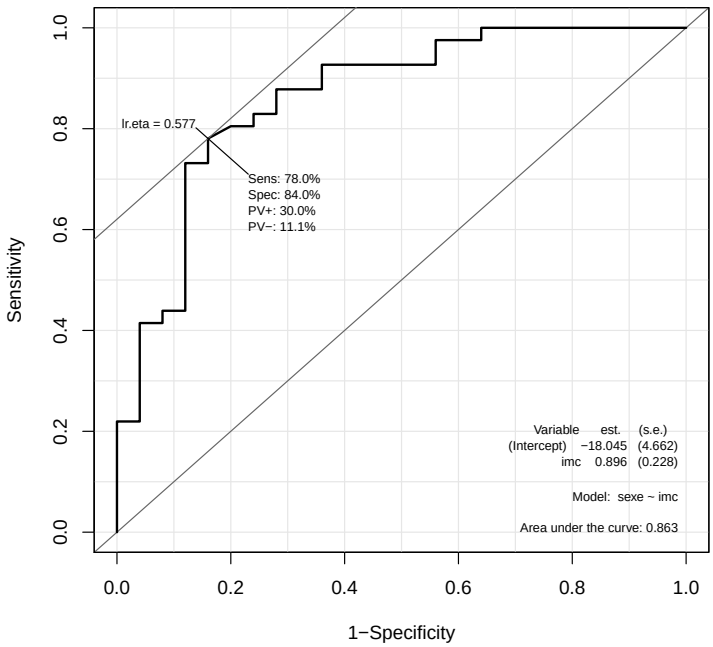


Que peut-on dire de ce graphique ?

2. Dorénavant, on souhaite expliquer le sexe d'un individu en fonction de son imc.

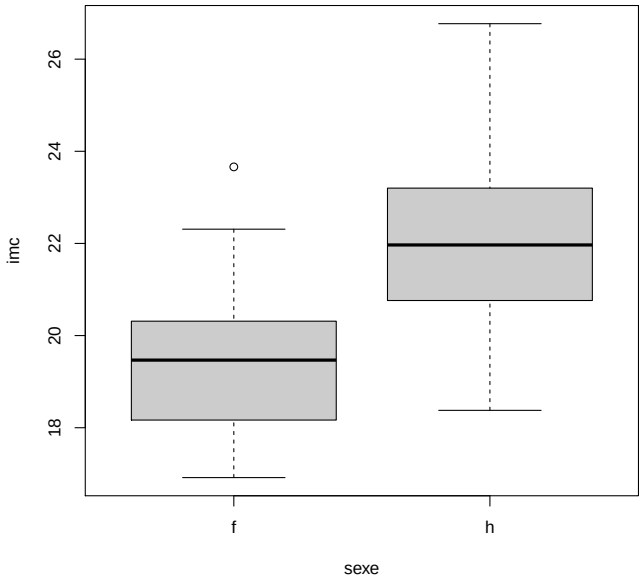
(a) À partir des informations ci-dessus, proposer toutes les commandes R pour créer un modèle adapté.

(b) Proposer les commandes R amenant au graphique suivant :



Que peut-on dire de ce graphique ?

3. Pour finir, indépendamment des modèles précédemment construits, proposer les commandes R amenant au graphique suivant :



Que présente ce graphique ? Donner toutes les informations que l'on peut en déduire.

Exercice 37. On considère le jeu de données `hdv2003` de la librairie `questionr`.

1. On exécute les commandes R suivantes :

```
library(questionr)
data(hdv2003)
w = hdv2003
str(w)
```

Cela renvoie :

```
'data.frame':   2000 obs. of  20 variables:
 $ id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ age          : int  28 23 59 34 71 35 60 47 20 28 ...
 $ sexe         : Factor w/ 2 levels "Homme","Femme": 2 2 1 1 2 2 2  ...
 $ nivetud      : Factor w/ 8 levels "N'a jamais fait d'etudes",..: 8 ...
 $ poids        : num  2634 9738 3994 5732 4329 ...
 $ occup        : Factor w/ 7 levels "Exerce une profession",..: 1 3 1  ...
 $ qualif       : Factor w/ 7 levels "Ouvrier specialise",..: 6 NA 3 3 ...
 $ freres.soeurs: int  8 2 2 1 0 5 1 5 4 2 ...
 $ clso         : Factor w/ 3 levels "Oui","Non","Ne sait pas": 1 1 2  ...
 $ relig        : Factor w/ 6 levels "Pratiquant regulier",..: 4 4 4 3 ...
 $ trav.imp     : Factor w/ 4 levels "Le plus important",..: 4 NA 2  ...
 $ trav.satisf  : Factor w/ 3 levels "Satisfaction",..: 2 NA 3 1 NA  ...
 $ hard.rock    : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 1 1 1 1 ...
 $ lecture.bd   : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 1 1 1 1 ...
 $ peche.chasse : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 2 2 1 1 ...
 $ cuisine      : Factor w/ 2 levels "Non","Oui": 2 1 1 2 1 1 2 2 1 1 ...
 $ bricol       : Factor w/ 2 levels "Non","Oui": 1 1 1 2 1 1 1 2 1 1 ...
 $ cinema       : Factor w/ 2 levels "Non","Oui": 1 2 1 2 1 2 1 1 2 2 ...
 $ sport        : Factor w/ 2 levels "Non","Oui": 1 2 2 2 1 2 1 1 1 2 ...
 $ heures.tv    : num  0 1 0 2 3 2 2.9 1 2 2 ...
```

Quelles sont les natures des variables `sport`, `freres.soeurs`, `qualif` et `age` ?

2. On exécute les commandes R suivantes :

```
w = subset(w, select = c(sport, freres.soeurs, qualif, age))
w = na.omit(w)
attach(w)
str(w)
```

Cela renvoie :

```
'data.frame': 1653 obs. of  4 variables:
 $ sport        : Factor w/ 2 levels "Non","Oui": 1 2 2 1 2 1 1 2 1 2 ...
 $ freres.soeurs: int  8 2 1 0 5 1 5 2 3 4 ...
 $ qualif       : Factor w/ 7 levels "Ouvrier specialise",..: 6 3 3 6 6 2 2 7 6 2 ...
 $ age          : int  28 59 34 71 35 60 47 28 65 47 ...
```

Décrire l'enjeu des commandes précédentes.

3. On exécute les commandes R suivantes :

```
sport = as.numeric(sport == "Oui")
table(sport)
```

Cela renvoie :

```
sport
   0    1
1064  589
```

Décrire l'enjeu des commandes précédentes, ainsi que la sortie.

4. Désormais, on souhaite expliquer la variable `sport` à l'aide des variables `freres.soeurs`, `qualif` et `age`, sans considérer d'éventuelle interaction entre ces variables.

   (a) Proposer un modèle adapté au contexte. Donner son écriture générique compète.

   (b) La partie principale du résumé statistique du modèle adapté est la suivante :

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.763839   0.262374   2.911 0.003600 **
freres.soeurs                 -0.057035   0.022480  -2.537 0.011177 *
age                           -0.045185   0.003989 -11.329  < 2e-16 ***
qualifOuvrier qualifie         0.561974   0.237471   2.366 0.017957 *
qualifTechnicien               1.403301   0.299731   4.682 2.84e-06 ***
qualifProfession intermediaire 1.516603   0.257659   5.886 3.95e-09 ***
qualifCadre                    1.841658   0.238597   7.719 1.18e-14 ***
qualifEmploye                  0.712462   0.213179   3.342 0.000832 ***
qualifAutre                    0.770811   0.351109   2.195 0.028138 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   Que peut-on dire de l'influence de `age` sur la variable `sport` ? Que peut-on dire de l'influence de `qualifOuvrier qualifie` sur la variable `sport` ?

   (c) On exécute les commandes R suivantes :

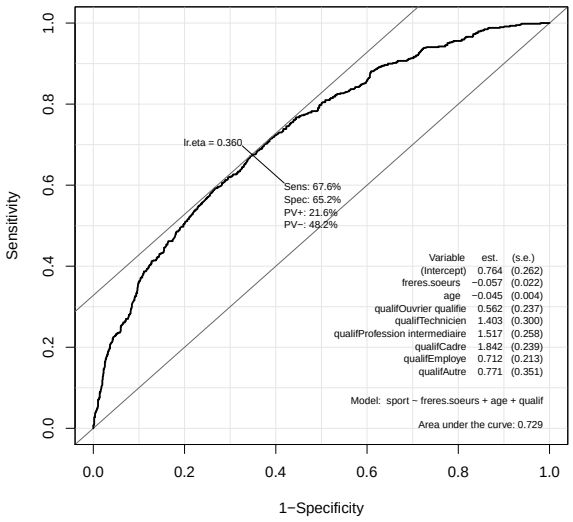
```
predict(reg, data.frame(freres.soeurs = 7, qualif = "Employe",
age = 41), type = "response")
```

   Donner la valeur numérique attendue. Est-ce qu'un individu ayant 7 frères et sœurs, employé et de 41 ans a plus de chances de faire du sport que non ?

   (d) On exécute les commandes R suivantes :

```
library(Epi)
ROC(form = sport ~ freres.soeurs + age + qualif, plot = "ROC")
```

Cela renvoie :



Que nous apprend ce graphique ?

(e) On exécute les commandes R suivantes :

```
pred.prob = predict(reg, type = "response")
pred.mod = factor(ifelse(pred.prob > 0.5, "1", "0"))
mc = table(sport, pred.mod)
(mc[1, 2] + mc[2, 1])/length(sport)
```

Cela renvoie :

```
[1] 0.2970357
```

A quel objet mathématique correspond `mc` ? A quel objet mathématique correspond la valeur numérique de la sortie ? Que peut-on en dire ? Est-ce que cela est en contradiction avec ce que vous avez pu observer à la question précédente ?

(f) Quel autre outil aurait-on pu utiliser pour évaluer la qualité prédictive du modèle ?

Exercice 38. On s'intéresse au lien éventuel entre la consommation d'alcool et la présence de malformations sur le fœtus durant la grossesse. On dispose du tableau suivant :

Consommation d'alcool	Présence	Absence
0	48	17066
< 1	38	14464
$1-2$	5	788
$3-5$	1	126
$6-8$	1	37

1. Décrire brièvement l'enjeu des commandes R suivantes :

```
Alcool = factor(c ("0", "<1", "1-2", "3-5", "6-8"),
levels = c ("0", "<1", "1-2", "3-5", "6-8"))
malformations = c(48, 38, 5, 1, 1)
n = c(17066, 14464, 788, 126, 37) + malformations
```

   Que représente exactement le vecteur `n` ?

2. Décrire brièvement l'enjeu des commandes R suivantes :

```
reg = glm(malformations / n ~ Alcool, family = binomial, weights = n)
summary(reg)
```

   Cela renvoie :

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.8736	0.1445	-40.64	0.0000	***
Alcool<1	-0.0682	0.2174	-0.31	0.7538	
Alcool1-2	0.8136	0.4713	1.73	0.0843	.
Alcool3-5	1.0374	1.0143	1.02	0.3064	
Alcool6-8	2.2627	1.0237	2.21	0.0271	*

```
Null deviance:  6.2020e+00 on 4 degrees of freedom
Residual deviance:  1.8163e-13 on 0 degrees of freedom
AIC: 28.627
```

3. On exécute les commandes R suivantes :

```
cbind(logit = predict(reg), fitted.prop = predict(reg,
type = "response"))
```

   Cela renvoie :

   Que représentent ces valeurs ?

	logit	fitted.prop
1	-5.87	0.00
2	-5.94	0.00
3	-5.06	0.01
4	-4.84	0.01
5	-3.61	0.03

4. Dans cette question, on propose une autre modélisation reposant sur la régression logistique. Décrire brièvement l'enjeu des commandes R suivantes :

```
Centres = c(0, 0.5, 1.5, 4, 7)
reg2 = glm(malformations / n ~ Centres, family = binomial, weights = n)
summary(reg2)
```

Cela renvoie :

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.9605	0.1154	-51.64	0.0000	***
Centres	0.3166	0.1254	2.52	0.0116	*

Puis :

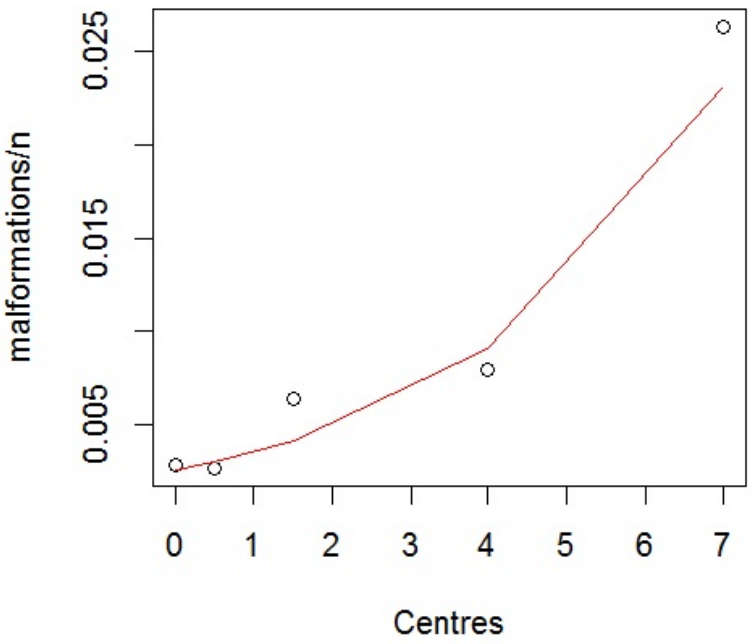
```
cbind(logit = predict(reg2), fitted.prop = predict(reg2,
type = "response"))
```

Cela renvoie :

	logit	fitted.prop
1	-5.96	0.00
2	-5.80	0.00
3	-5.49	0.00
4	-4.69	0.01
5	-3.74	0.02

De plus :

```
plot(Centres, malformations / n)
lines(Centres, fitted.values(reg2), col = "red")
```



Est-ce que le modèle vous semble bon ?

Exercice 39. On souhaite expliquer une variable qualitative Y à 3 modalités, codées par : 1, 2 et 3, à partir de 2 variables quantitatives $X1$ et $X2$. Précisons que les modalités sont sans lien hiérarchique/ordre. Les données sont disponibles ici :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/multinom.txt",
header = T)
attach(w)
head(w)
```

	X1	X2	Y
1	62.94	42.91	1
2	48.09	53.13	2
3	30.50	81.08	1
4	38.07	37.30	3
5	37.26	78.70	1
6	83.14	24.58	3

L'objectif principal est d'estimer la probabilité (ou proportion) inconnue

$$p_k(x) = \mathbb{P}(\{Y = u_k\}|\{(X1, X2) = x\}), \qquad x = (x_1, x_2),$$

avec $u_1 = 1$, $u_2 = 2$ et $u_3 = 3$, à l'aide des données.

1. On adopte le modèle de régression multinomiale :

   ```
   library(nnet)
   reg = multinom(Y ~ X1 + X2)
   ```

   Expliciter ce modèle. Combien de coefficients sont à estimer ?

2. On exécute les commandes R suivantes :

   ```
   summary(reg)
   ```

   Cela renvoie :

   ```
   multinom(formula = Y ~ X1 + X2)
   Coefficients:
   (Intercept) X1           X2
   2 14.53405 -0.1902811 -0.1316235
   3 22.84242 -0.2193699 -0.2779681
   Std.  Errors:
   (Intercept) X1          X2
   2 2.214416 0.02366630 0.02241691
   3 2.431125 0.02437192 0.02827328
   ```

   Que représente ces valeurs ?

3. On exécute les commandes R suivantes :

```
predict(reg, data.frame(X1 = 58, X2 = 32), type = "probs")
```

Les valeurs renvoyées peuvent se mettre sous la forme :

$\widehat{p}_1(x)$	$\widehat{p}_2(x)$	$\widehat{p}_3(x)$
$\circ$	$\star$	$*$

Remplacer $\circ$, $\star$ et $*$ par les valeurs attendues.

4. Quel est l'enjeu de la commande R suivante ?

```
predict(reg, data.frame(X1 = 58, X2 = 32), type = "class")
```

Cela renvoie 3.

5. Pour tester la significativité des variables explicatives, on exécute :

```
z = summary(reg)$coeff / summary(reg)$standard.errors
pvaleur = 2 * (1 - pnorm(abs(z), 0, 1))
pvaleur
```

Cela renvoie :

```
(Intercept)      X1             X2
2 5.260326e-11 8.881784e-16 4.315606e-09
3 0.000000e+00 0.000000e+00 0.000000e+00
```

Que peut-on en conclure ?

6. On détermine la matrice de confusion du modèle :

```
pr = predict(reg)
mc = table(Y, pr)
mc
```

Cela renvoie la matrice :

$$MC = \begin{pmatrix} 238 & 10 & 12 \\ 13 & 43 & 7 \\ 13 & 1 & 163 \end{pmatrix}$$

Calculer le taux d'erreur. Est-ce que le modèle a une bonne qualité prédictive ?

Exercice 40. Stéphanie considère le jeu de données `cpus` de la librarie `MASS`. Sur un échantillon de 209 processeurs, on dispose, entre autre, de sa performance : `perf`, ainsi que six variables techniques : `syct`, `mmin`, `mmax`, `cach`, `chmin` et `chmax`. Stéphanie désire expliquer `perf` en fonction de `syct`, `mmin`, `mmax`, `cach`, `chmin` et `chmax`.

1. Pour atteindre son objectif, Stéphanie a une première idée de modélisation ; elle pense à utiliser le modèle de *rlm*. D'abord, elle fait les commandes :

```
library(MASS)
w = data.frame(cpus)
attach(w)
str(w)
```

Cela renvoie :

```
'data.frame':   209 obs. of  9 variables:
 $ name   : Factor w/ 209 levels "ADVISOR 32/60",..: 1 3 2 4 5 6 8 9 10 7 ...
 $ syct   : int  125 29 29 29 29 26 23 23 23 23 ...
 $ mmin   : int  256 8000 8000 8000 8000 8000 16000 16000 16000 32000 ...
 $ mmax   : int  6000 32000 32000 32000 16000 32000 32000 32000 64000 64000 ...
 $ cach   : int  256 32 32 32 32 64 64 64 64 128 ...
 $ chmin  : int  16 8 8 8 8 8 16 16 16 32 ...
 $ chmax  : int  128 32 32 32 16 32 32 32 32 64 ...
 $ perf   : int  198 269 220 172 132 318 367 489 636 1144 ...
 $ estperf: int  199 253 253 253 132 290 381 381 749 1238 ...
```

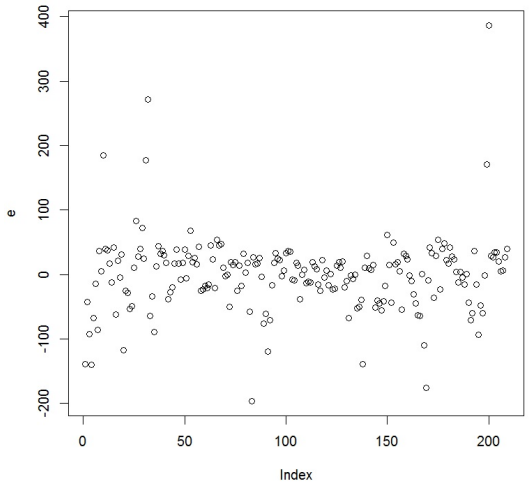
Elle créée alors une data frame `ww` contenant uniquement les variables explicatives `syct`, `mmin`, `mmax`, `cach`, `chmin` et `chmax`. Elle utilise ensuite la commande `lm` pour mettre en œuvre le modèle de *rlm*, nommé `reg`, sans y écrire le noms des variables explicatives dedans. Puis elle demande un résumé des estimations liées à ce modèle. Donner toutes les commandes que Stéphanie a faites.

2. L'exécution des commandes de la question précédente renvoie, entre autre :

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.590e+01  8.045e+00  -6.948 4.99e-11 ***
syct         4.886e-02  1.752e-02   2.789  0.00579 **
mmin         1.529e-02  1.827e-03   8.371 9.42e-15 ***
mmax         5.571e-03  6.418e-04   8.680 1.33e-15 ***
cach         6.412e-01  1.396e-01   4.594 7.64e-06 ***
chmin       -2.701e-01  8.557e-01  -0.316  0.75263
chmax        1.483e+00  2.201e-01   6.738 1.64e-10 ***
---

Residual standard error: 59.99 on 202 degrees of freedom
Multiple R-squared:  0.8649,    Adjusted R-squared:  0.8609
F-statistic: 215.5 on 6 and 202 DF,  p-value: < 2.2e-16
```

Stéphanie trouve le modèle performant. Expliquer cette impression.

Dès lors, Stéphanie souhaite le valider mathématiquement. Elle décide alors de tracer le graphique des résidus pour analyse. Elle obtient le graphique suivant :



Que pensez-vous de ce graphique ? Partant des commandes de la question précédente, écrire toutes les commandes que Stéphanie a faites pour arriver à cette sortie.

3. Stéphanie décide de construire un autre modèle simplifié en transformant `perf` en variable qualitative `C` à trois modalités : "mauvaise", "moyenne" et "bonne". Ainsi, elle crée un vecteur colonne à 209 éléments dont le i-ème élément vaut

   - "mauvaise" si sa performance est < 32,
   - "bonne" si sa performance est > 72,
   - "moyenne" si sa performance est $\in [32, 72]$.

   Une fois créée, elle fait d'autres commandes qui renvoient :

```
C
  bonne mauvaise  moyenne
     69       62       78
```

   Partant des commandes de la question précédente, écrire toutes les commandes que Stéphanie a faites pour arriver à cette sortie.

4. Ensuite, Stéphanie fait les commandes :

```
library(nnet)
regnew = multinom(C ~ .  , data = ww)
pr = predict(regnew)
mc = table(C, pr)
mc
```

Cela renvoie :

```
          pr
C          bonne mauvaise moyenne
  bonne       59        1       9
  mauvaise     0       49      13
  moyenne      3       15      60
```

Stéphanie est contente de ce résultat.

Quel est le modèle considéré ? Quel est le taux d'erreur ? Est-ce que Stéphanie a raison d'être contente de son analyse ?

Exercice 41. On souhaite expliquer une variable qualitative Y à 4 modalités codées par : $--$, $-$, $+$ et $++$, à partir de 2 variables quantitatives $X1$ et $X2$. Précisons que les modalités ont un lien hiérarchique : $-- < - < + < ++$. Les données sont disponibles ici :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/polr.txt",
header = T)
attach(w)
head(w)
```

	X1	X2	Y
1	171.08	24.32	++
2	173.39	32.06	++
3	185.91	28.03	+
4	175.49	27.22	+
5	175.91	22.39	-
6	187.01	29.64	+

L'objectif principal est d'estimer la probabilité (ou proportion) inconnue

$$p_k(x) = \mathbb{P}(\{Y = u_k\}|\{(X1, X2) = x\}), \qquad x = (x_1, x_2),$$

avec $k \in \{1, 2, 3, 4\}$, $u_1 = --$, $u_2 = -$, $u_3 = +$ et $u_4 = ++$, à l'aide des données.

1. Expliciter un modèle de régression adapté aux données.
2. Décrire brièvement l'enjeu des commandes R suivantes :

```
library(MASS)
reg = polr(Y ~ X1 + X2, method = "logistic")
summary(reg)
```

Cela renvoie :

```
Coefficients:
Value Std.  Error t value
X1 0.06467 0.02887 2.240
X2 -0.03743 0.02193 -1.707
Intercepts:
Value Std.  Error t value
-|- 9.0328 5.1009 1.7708
-|+ 10.2082 5.1187 1.9943
+|++ 11.4268 5.1565 2.2160
```

3. Décrire brièvement l'enjeu des commandes R suivantes :

```
predict(reg, data.frame(X1 = 187, X2 = 25), type = "class")
```

Cela renvoie : $++$.

Exercice 42. On considère le jeu de données `espèces`. On s'intéresse au nombre d'espèces animales présentent sur un site écologique à partir du diamètre, de la profondeur et de l'oxygène dissous du site. Ainsi, pour 13 sites différents, on dispose :

- du nombre d'espèces animales (variable Y),
- du diamètre (variable $X1$),
- de la profondeur (variable $X2$),
- de l'oxygène dissous (variable $X3$).

On souhaite expliquer Y à partir de $X1$, $X2$ et $X3$.

Pour se donner une idée plus précise des données, on exécute les commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/espèces.txt", header
= T)
attach(w)
head(w)
```

Cela renvoie :

	X1	X2	X3	Y
1	40	10.00	8.40	21
2	100	25.00	8.10	21
3	50	12.50	7.90	17
4	5	7.50	6.15	17
5	5	30.00	7.90	8
6	10	15.00	9.60	29

1. On exécute les commandes R suivantes :

```
reg1 = glm(Y ~ X1 + X2 + X3, family = poisson)
```

Quel modèle de régression est considéré par ces commandes ? Pourquoi celui-ci est approprié ? Expliciter ce modèle.

2. Décrire brièvement l'enjeu des commandes R suivantes :

```
lamb = predict.glm(reg1, data.frame(X1 = 19, X2 = 12, X3 = 8),
type = "response")
probs = dpois(0:100, lamb)
which.max(probs)
```

Cela renvoie 20.

Quelle est la valeur la plus probable pour Y si $X1 = 19$, $X2 = 12$ et $X3 = 8$?

3. Quel objet mathématique renvoie les commandes R suivantes ?

```
loglamb = predict.glm(reg, data.frame(X1 = 20, X2 = 13, X3 = 7),
se.fit = TRUE)
icloglamb = c(loglamb$fit - 1.96 * loglamb$se.fit,
loglamb$fit + 1.96 * loglamb$se.fit)
ic = exp(icloglamb)
ic
```

4. On exécute les commandes R suivantes :

```
reg1 = glm(Y ~ X1 + X2 + X3, family = poisson)
reg2 = glm(Y ~ log(X1) + X2 + X3, family = poisson)
reg3 = glm(Y ~ X1 + log(X2) + X3, family = poisson)
reg4 = glm(Y ~ X1 + X2 + log(X3), family = poisson)
reg5 = glm(Y ~ log(X1) + log(X2) + X3, family = poisson)
reg6 = glm(Y ~ log(X1) + X2 + log(X3), family = poisson)
reg7 = glm(Y ~ X1 + log(X2) + log(X3), family = poisson)
reg8 = glm(Y ~ log(X1) + log(X2) + log(X3), family = poisson)
```

Puis :

```
a = cbind(AIC(reg1), AIC(reg2), AIC(reg3), AIC(reg4), AIC(reg5),
AIC(reg6), AIC(reg7), AIC(reg8))
b = cbind(BIC(reg1), BIC(reg2), BIC(reg3), BIC(reg4), BIC(reg5),
BIC(reg6), BIC(reg7), BIC(reg8))
rbind(a, b)
```

Cela renvoie :

	1	2	3	4	5	6	7	8
1	87.96	89.21	85.02	87.41	84.85	89.13	84.13	84.73
2	90.22	91.47	87.28	89.67	87.11	91.39	86.39	86.99

D'après les critères considérés, quel modèle est le meilleur ? Expliciter ce modèle.

5. On exécute les commandes R suivantes :

```
library(AER)
dispersiontest(reg7)
```

Cela renvoie : p-valeur = 0.4341. Que peut-on en déduire ?

Exercice 43. Chacun des 300 employés d'une grande compagnie d'assurance sont assignés au hasard à l'un des 3 nouveaux logiciels de formation notés A, B et C. Leur nombre d'appels au support informatique au cours du mois suivant est comptabilisé. Ainsi, pour chacun des 300 employés, on dispose :

- du logiciel utilisé (variable $X1$),
- du nombre d'années d'expérience (variable $X2$),
- du score d'un test de connaissances en informatique (sur 100) (variable $X3$),
- du nombre d'appels au support informatique au cours du mois suivant (variable Y).

On souhaite expliquer Y à partir de $X1$, $X2$ et $X3$.

1. Décrire brièvement l'enjeu des commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/formation.txt",
header = T)
attach(w)
head(w)
```

Cela renvoie :

	X1	X2	X3	Y
1	A	3.92	60	6
2	A	5.83	64	3
3	A	0.92	51	8
4	A	8.50	58	2
5	A	7.83	59	1
6	A	1.17	49	3

Puis :

```
table(Y)
```

Cela renvoie :

	0	1	2	3	4	5	6	7	8	9	10	11	12
Y	6	27	42	61	70	39	23	17	9	2	2	1	1

2. On adopte le modèle de régression de Poisson :

```
reg = glm(Y ~ X1 + X2 + X3, family = poisson)
summary(reg)
```

Cela renvoie :

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.9927	0.1592	12.52	0.0000	$***$
X1B	-0.1705	0.0732	-2.33	0.0198	$*$
X1C	-0.0078	0.0702	-0.11	0.9112	
X2	-0.0280	0.0103	-2.72	0.0066	$**$
X3	-0.0092	0.0030	-3.05	0.0023	$**$

Expliciter ce modèle et préciser les estimations ponctuelles des coefficients principaux. Est-ce que l'on peut affirmer que $X3$ influe très significativement sur Y ?

3. Décrire brièvement l'enjeu des commandes R suivantes :

```
lamb = predict.glm(reg, data.frame(X1 = "A", X2 = 9.4, X3 = 51),
type = "response")
probs = dpois(0:100, lamb)
which.max(probs)
```

Cela renvoie 4.

(a) Quelle est la valeur la plus probable pour Y si $X1 = A$, $X2 = 9.4$ et $X3 = 51$?

(b) Retrouver, par le calcul, la valeur de `lamb`

4. Décrire brièvement l'enjeu des commandes R suivantes :

```
loglamb = predict.glm(reg, data.frame(X1 = "A", X2 = 9.4, X3 = 51),
se.fit = TRUE)
icloglamb = c(loglamb$fit - 1.96 * loglamb$se.fit,
loglamb$fit + 1.96 * loglamb$se.fit)
ic = exp(icloglamb)
ic
```

Cela renvoie : 3.076122, 4.040153

5. On exécute les commandes R suivantes :

```
deviance(reg) / df.residual(reg)
```

Cela renvoie 1.036946.

Quelle règle met-on en œuvre et que peut-on en conclure ?

6. On utilise le test de Cameron et Trivedi :

```
library(AER)
dispersiontest(reg)
```

Cela renvoie : p-valeur = 0.6282.

Que peut-on en conclure ?

Exercice 44. Une entreprise de petites livraisons possède deux sites d'où partent les véhicules. Sur un échantillon de 20 véhicules (10 dans chaque site), on relève le nom du site (variable `site`), le nombre de kilomètres parcourus depuis la première mise en circulation (variable `km`) et le nombre de sinistres déclarés depuis la première mise en circulation (variable `nb`). Les résultats sont les suivants :

site	km	nb
S1	81838	1
S1	8730	0
S1	55088	0
S1	54309	0
S1	19780	0
S1	45591	0
S1	20767	0
S1	86343	0
S1	104371	4
S1	2084	0
S2	5192	0
S2	71643	0
S2	87022	0
S2	125544	2
S2	122785	1
S2	104632	0
S2	129189	3
S2	97319	0
S2	116152	2
S2	45931	0

L'entreprise se pose la question suivante : le nombre moyen de sinistres est-il dépendant du nombre de kilomètres parcourus et si oui, cette dépendance est-elle la même quelque soit le site ?

1. Proposer un modèle susceptible d'apporter une réponse fiable à l'entreprise. Écrire sa formule mathématique. Proposer des commandes R adaptées.

2. Une sortie partielle des commandes précédentes renvoient :

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.448e+00  4.748e+00  -1.990   0.0466 *
km           1.037e-04  4.734e-05   2.191   0.0285 *
siteS2      -2.905e+00  7.925e+00  -0.367   0.7140
km:siteS2    6.838e-07  6.958e-05   0.010   0.9922
```

   Que peut-on en conclure ?

3. Un nouveau véhicule vérifie `km` = 107671 et `site` = S1, mais on ne sait rien de la valeur de `nb`. En utilisant le modèle précédent, donner une estimation ponctuelle de la probabilité que `nb` soit égale à 5.

Exercice 45. On considère le jeu de données `Sitka` de la librairie `MASS`. Il contient des mesures répétées sur la taille (logarithmique) de 79 épinettes de Sitka, dont 54 ont été cultivées dans des serres enrichies en ozone et 25 ont été sous contrôle. La taille a été mesurée cinq fois en 1988, à intervalles plus ou moins mensuels. Ainsi, pour chacun des 79 épinettes, on dispose de 5 mesures et :

- de le taille (logarithmique) d'une épinette (variable `size`),
- du moment de la mesure en jours depuis le 1er Janvier de 1988 (variable `Time`),
- du numéro de l'épinette (variable `tree`),
- du traitement "ozone" ou "contrôle" de l'épinette (variable `treat`).

On souhaite expliquer `size` à partir de `Time` et `treat`.

1. Décrire brièvement l'enjeu des commandes R suivantes :

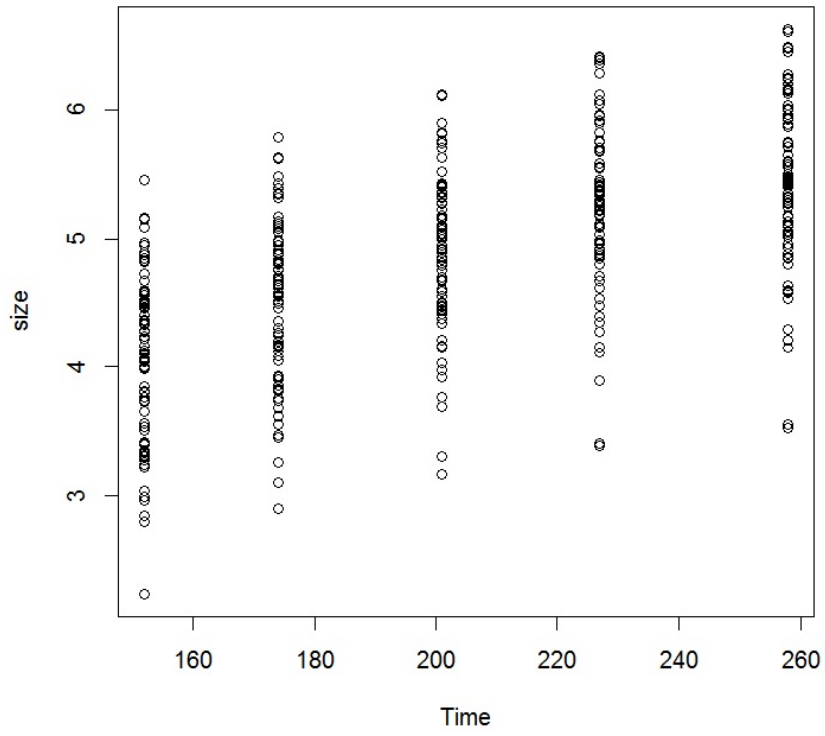
```
library(MASS)
data(Sitka)
attach(Sitka)
head(Sitka, 20)
```

Cela renvoie :

	size	Time	tree	treat
1	4.51	152.00	1	ozone
2	4.98	174.00	1	ozone
3	5.41	201.00	1	ozone
4	5.90	227.00	1	ozone
5	6.15	258.00	1	ozone
6	4.24	152.00	2	ozone
7	4.20	174.00	2	ozone
8	4.68	201.00	2	ozone
9	4.92	227.00	2	ozone
10	4.96	258.00	2	ozone
11	3.98	152.00	3	ozone
12	4.36	174.00	3	ozone
13	4.79	201.00	3	ozone
14	4.99	227.00	3	ozone
15	5.03	258.00	3	ozone
16	4.36	152.00	4	ozone
17	4.77	174.00	4	ozone
18	5.10	201.00	4	ozone
19	5.30	227.00	4	ozone
20	5.36	258.00	4	ozone

Puis :

```
plot(Time, size)
```



2. Est-ce que les modèles de régression `reg1` et `reg2` décris dans les commandes suivantes sont adaptés au problème ? Expliquer votre réponse.

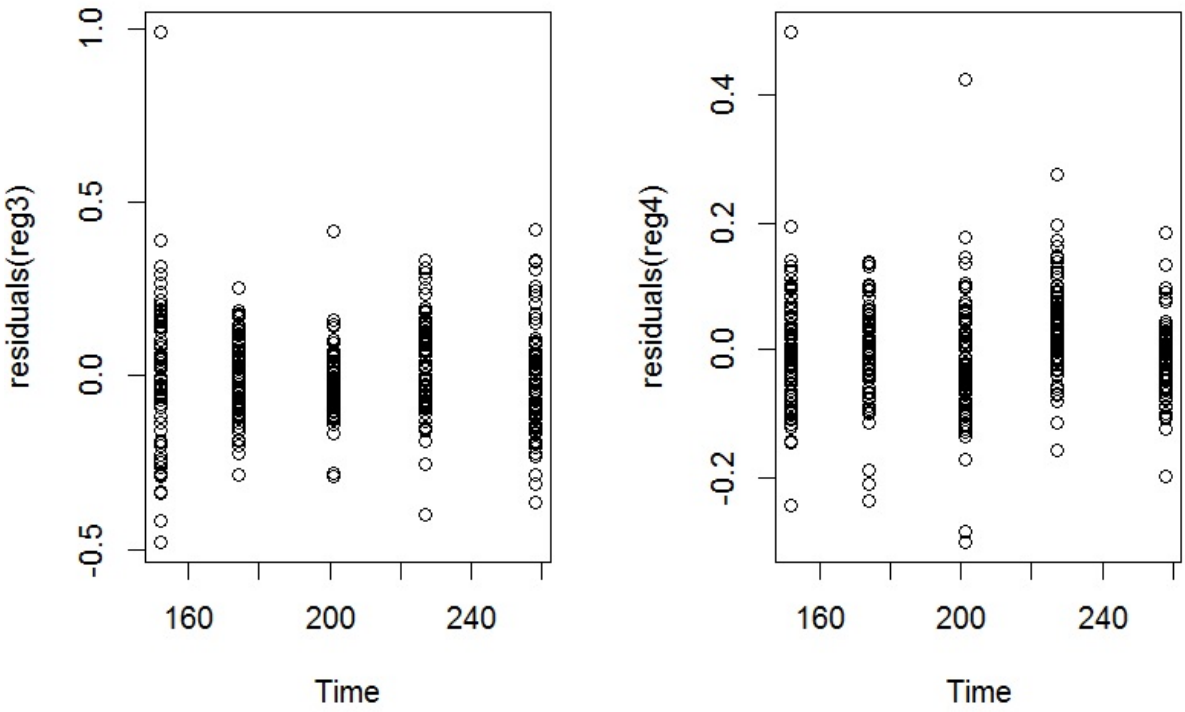
```
reg1 = lm(size ~ poly(Time, 2) * treat)
tree = as.factor(tree)
reg2 = lm(size ~ poly(Time, 2) * treat + tree)
```

3. Est-ce que les modèles de régression `reg3` et `reg4` décris dans les commandes suivantes sont adaptés au problème ? Expliquer votre réponse.

```
library(lme4)
library(lmerTest)
reg3 = lmer(size ~ poly(Time, 2) + treat + (1 | tree))
reg4 = lmer(size ~ poly(Time, 2) + treat + (Time | tree))
```

4. On exécute les commandes R suivantes :

```
par(mfrow = c(1, 2))
plot(Time, residuals(reg3))
plot(Time, residuals(reg4))
```



Au vu de ces graphiques, quel modèle semble être le plus adapté ?