# Etudes ; modèles de régression

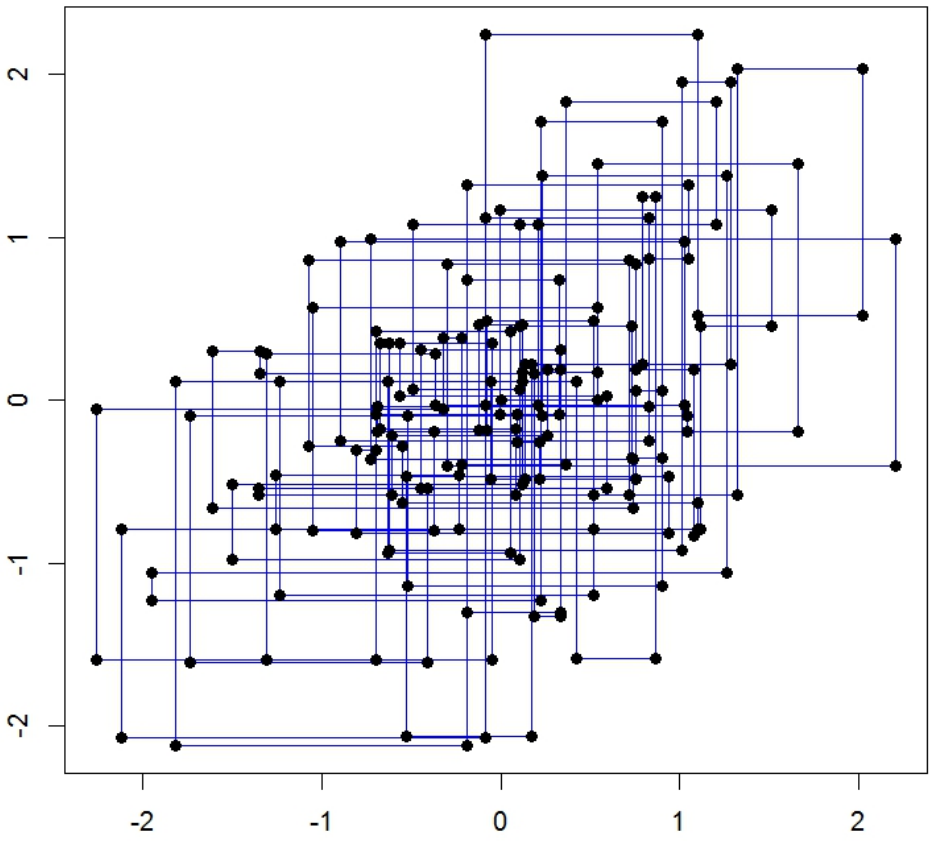
Christophe Chesneau

Université de Caen Basse-Normandie

# Études ; modèles de régression

Christophe Chesneau

http://www.math.unicaen.fr/~chesneau/




Caen, le 10 Novembre 2016

# Table des matières



**~ Note ~**

Ce document collecte les études statistiques du cours *Modèles de Régression* du M2 orienté statistique de l'université de Caen. Un des objectifs est de donner des pistes de réflexion à la construction de modèles prédictifs à partir de données de nature différente. Le logiciel utilisé est R.

*L'écriture des modèles de régression a été simplifiée et les commandes sont aussi directes que possible. Les interprétations d'analyse doivent être perçues comme des propositions.*

Je vous invite à me contacter pour tout commentaire :

`christophe.chesneau@gmail.com`

Bonne lecture !

# 0 Étude n° 0 : Pression artérielle systolique

## Contexte

La pression artérielle systolique est la pression maximale du sang dans les artères au moment de la contraction du cœur. Celle-ci a été mesurée pour 29 individus de différents âges. Ainsi, pour chacun d'entre eux, on dispose :

– de leur pression systolique en mmHg (variable $Y$),

– de leur âge en années (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$.

Dans un premier temps, ouvrir une fenêtre R et taper les commandes :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude0.txt",
header = T)
```

Une partie des données s'affiche en faisant :

```
head(w)
```

| | X1 | Y |
|---|---|---|
| 1 | 39 | 144 |
| 2 | 45 | 138 |
| 3 | 47 | 145 |
| 4 | 65 | 162 |
| 5 | 46 | 142 |
| 6 | 67 | 170 |

On obtient une brève description des données en faisant :

```
str(w)
```

On associe les variables $Y$ et $X1$ aux valeurs associées en faisant :

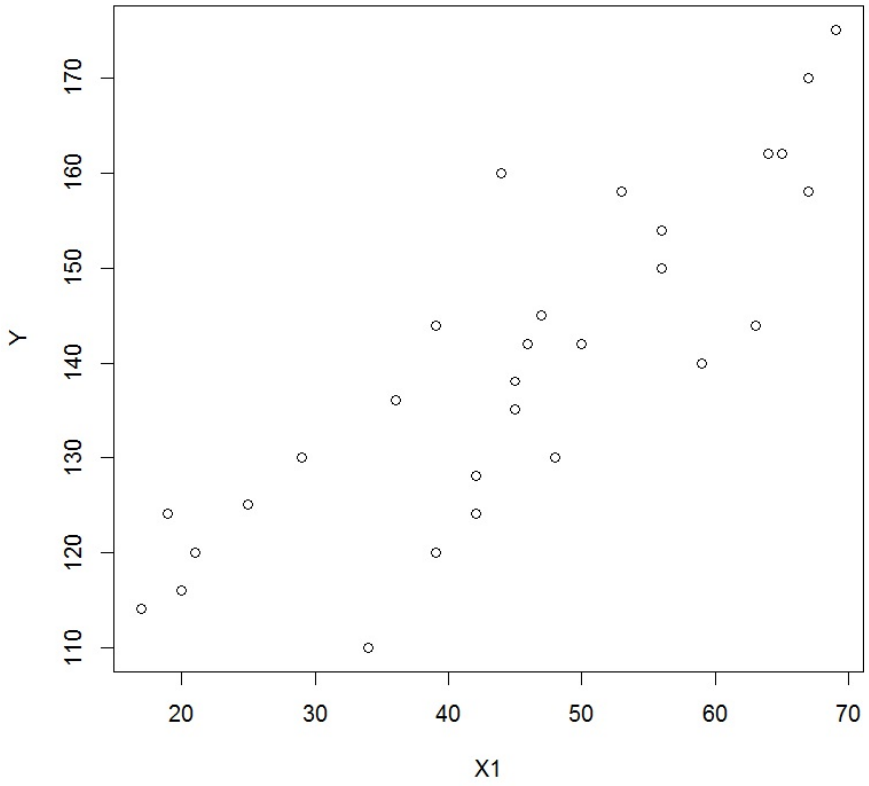
```
attach(w)
```

# Régression linéaire simple

## Analyse du nuage de points

On trace le nuage de points $\{(x_{1,i}, y_i),\ i \in \{1, \ldots, n\}\}$ :

```
plot(X1, Y)
```




On constate qu'une liaison linéaire entre $Y$ et $X1$ est envisageable.

## Modélisation

Une première approche est de considérer le modèle de *rls* :

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

où $\beta_0$ et $\beta_1$ sont 2 coefficients inconnus, et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma$ inconnu.

La présence de $\beta_0$ est justifiée car même un très jeune individu peut avoir une pression artérielle systolique élevée.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

## Estimations

La modélisation de la *rls* et les estimations des paramètres par la méthode des *mco* s'obtiennent par les commandes :

```
reg = lm(Y ~ X1)

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 97.0771 | 5.5276 | 17.56 | 0.0000 | *** |
| X1 | 0.9493 | 0.1161 | 8.17 | 0.0000 | *** |

```
Residual standard error: 9.563 on 27 degrees of freedom

Multiple R-squared: 0.7122, Adjusted R-squared: 0.7015

F-statistic: 66.81 on 1 and 27 DF, p-value: 8.876e-09
```

– Estimations ponctuelles de $\beta_0$ et $\beta_1$ :

| $\widehat{\beta}_0$ | $\widehat{\beta}_1$ |
|---|---|
| 97.0771 | 0.9493 |

– Estimations ponctuelles des écart-types des estimateurs de $\beta_0$ et $\beta_1$ :

| $\widehat{\sigma}(\widehat{\beta}_0)$ | $\widehat{\sigma}(\widehat{\beta}_1)$ |
|---|---|
| 5.5276 | 0.1161 |

– $t_{obs}$ :

| $H_1$ | $\beta_0 \neq 0$ | $\beta_1 \neq 0$ |
|---|---|---|
| $t_{obs}$ | 17.56 | 8.17 |

– Test de Student pour $\beta_1$ : influence de $X1$ sur $Y$ : p-valeur $< 0.001$, $***$ : hautement significative,

– $R^2 = 0.7122$ et $\overline{R}^2 = 0.7015$ : cela est satisfaisant,

– Test de Fisher : p-valeur $= 8.876e{-}09 < 0.001$, $***$ : l'utilisation du modèle de *rls* est pertinente.

La valeur prédite de $Y$ quand $X1 = 8$ (par exemple) est donnée par les commandes :

```
predict(reg, data.frame(X1 = 8))
```

Cela renvoie 104.6717.

Ainsi, la pression artérielle systolique moyenne d'un enfant de 8 ans est de 104.6717 mmHg.

On peut aussi s'intéresser :

– aux intervalles de confiance pour $\beta_0$ et $\beta_1$ au niveau 95% (par exemple).

Les commandes sont :

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

Cela renvoie :

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 85.7354850 | 108.418684 |
| X1 | 0.7110137 | 1.187631 |

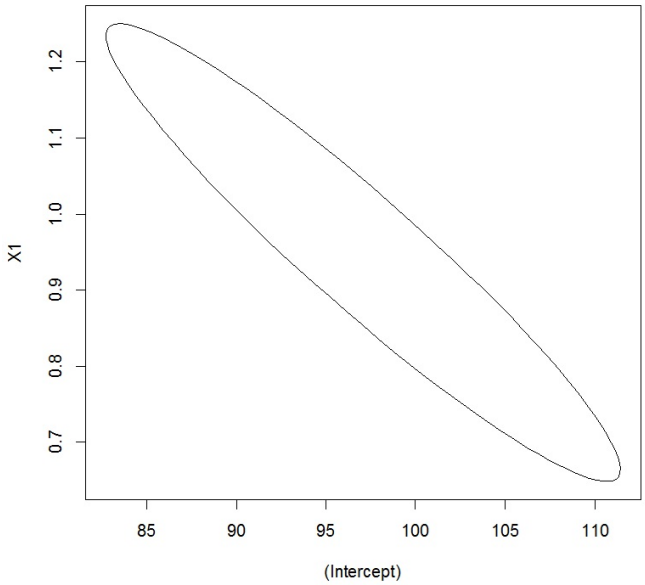
Le tableau donne les bornes inférieures et supérieures des intervalles de confiance de $\beta_0$ et $\beta_1$ :

| $i_{\beta_0}$ | $i_{\beta_1}$ |
|---|---|
| [85.7354850, 108.418684] | [0.7110137, 1.187631] |

– à l'ellipsoïde de confiance de $(\beta_0, \beta_1)$ au niveau 95% (par exemple).

Les commandes sont :

```
library(ellipse)

plot(ellipse(reg, c(1, 2), level = 0.95, type = "l"))
```




– à l'intervalle de confiance pour la valeur moyenne de $Y$ quand $X1 = 8$ (par exemple).

Les commandes sont :

```
predict(reg, data.frame(X1 = 8), interval = "confidence")
```

Cela renvoie :

| fit | lwr | upr |
|---|---|---|
| 104.6717 | 95.11582 | 114.2275 |

La première valeur est celle de la valeur prédite de $Y$ quand $X1 = 8$ (déjà vue), les deux autres correspondent aux bornes inférieures et supérieures des intervalles de confiance de cette valeur. Ainsi, pour $X1 = 8$, on a

$$i_{y_x} = [95.11582,\ 114.2275].$$

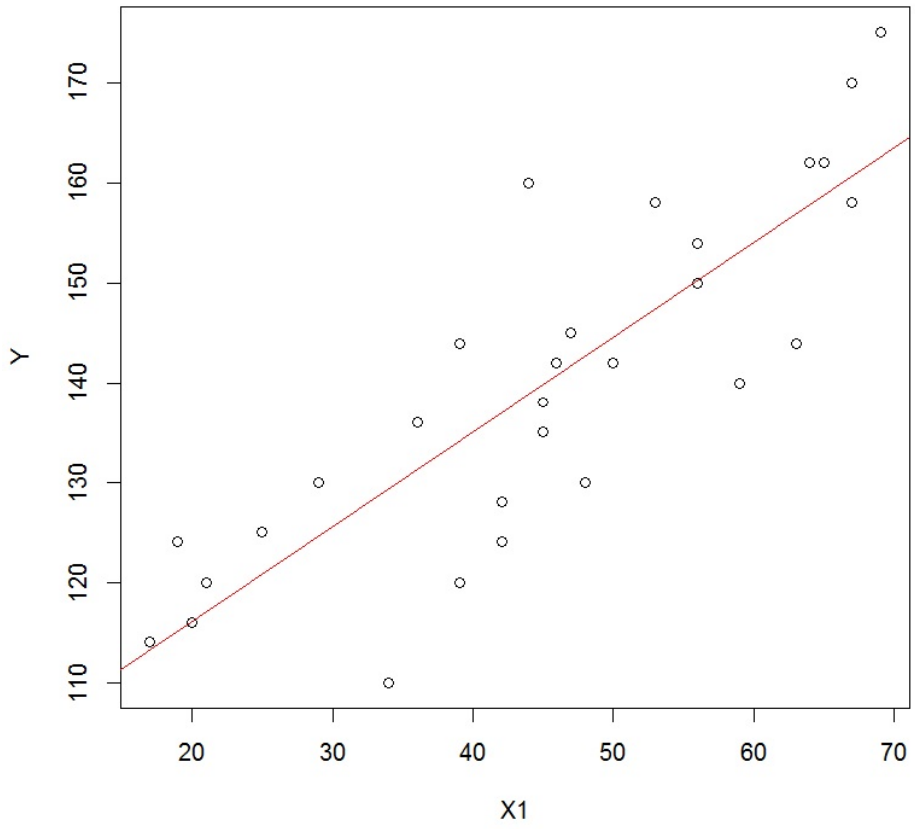
### Droite de régression

En utilisant les estimations ponctuelles de $\beta_0$ et $\beta_1$, l'équation de la droite de régression est :

$$y = 97.0771 + 0.9493x.$$

On la visualise avec les commandes :

```
abline(reg, col = "red")
```

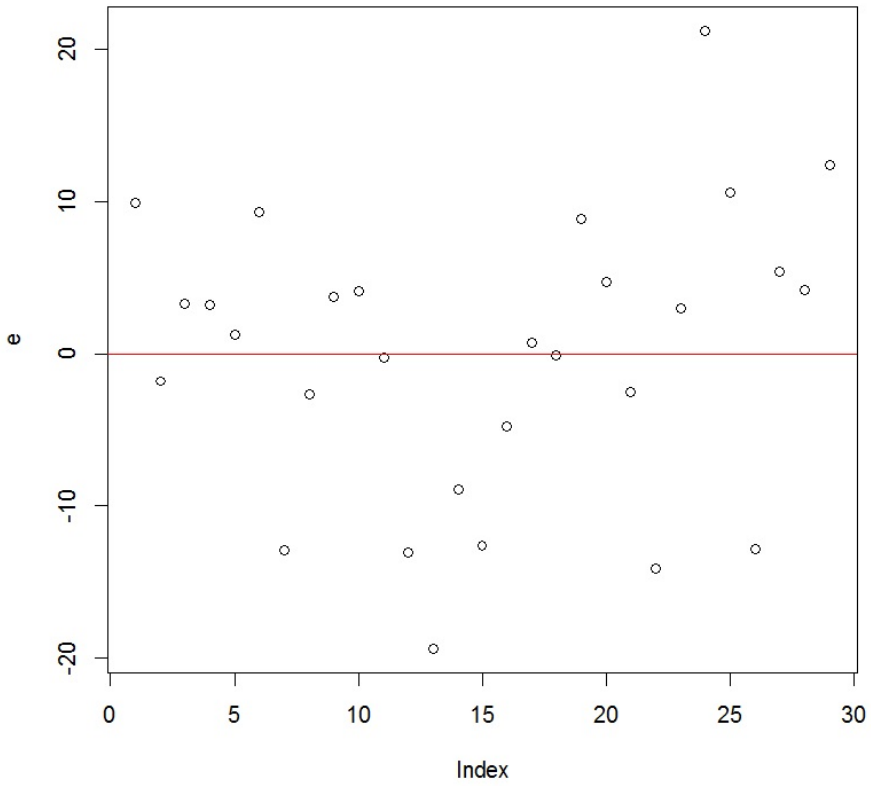


# Validation des hypothèses

## Analyse graphique des résidus

On examine les résidus en faisant :

```
e = residuals(reg)

plot(e)

abline(h = 0, col = "red")
```

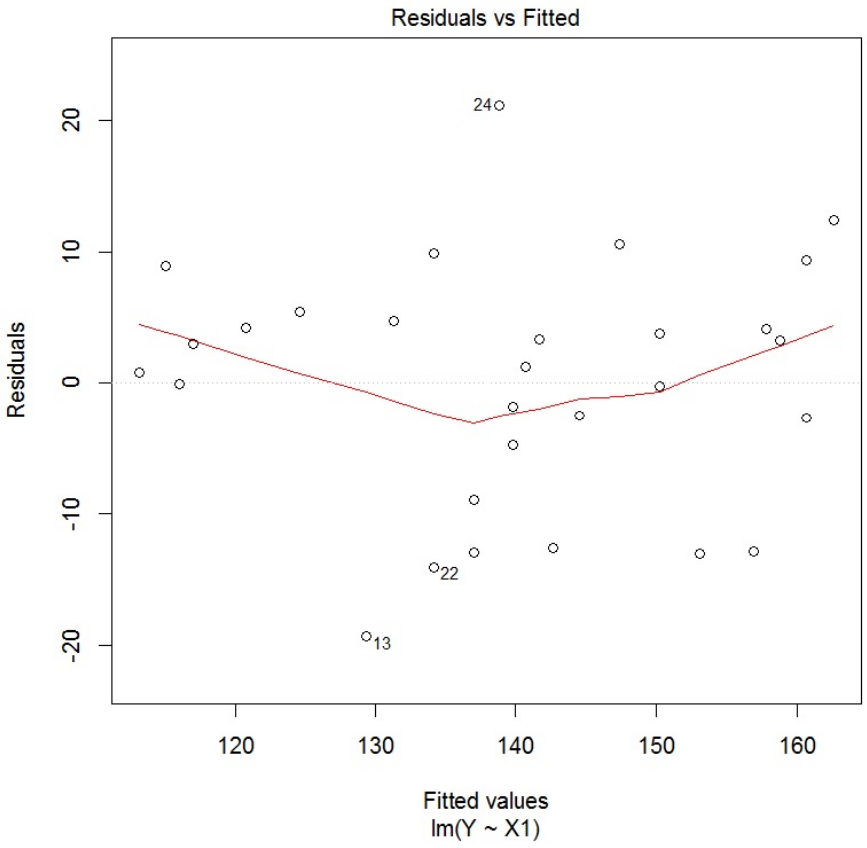



On constate une relative symétrie des résidus par rapport à l'axe des abscisses et pas de structure évidente. Cela est encourageant pour la validation des hypothèses standards du modèle de *rlm*.

### Indépendance de $\epsilon$ et $X1$

On trace le nuage de points {(résidus$_i$, prédictions en $x_{1,i}$)} :

```
plot(reg, 1)
```



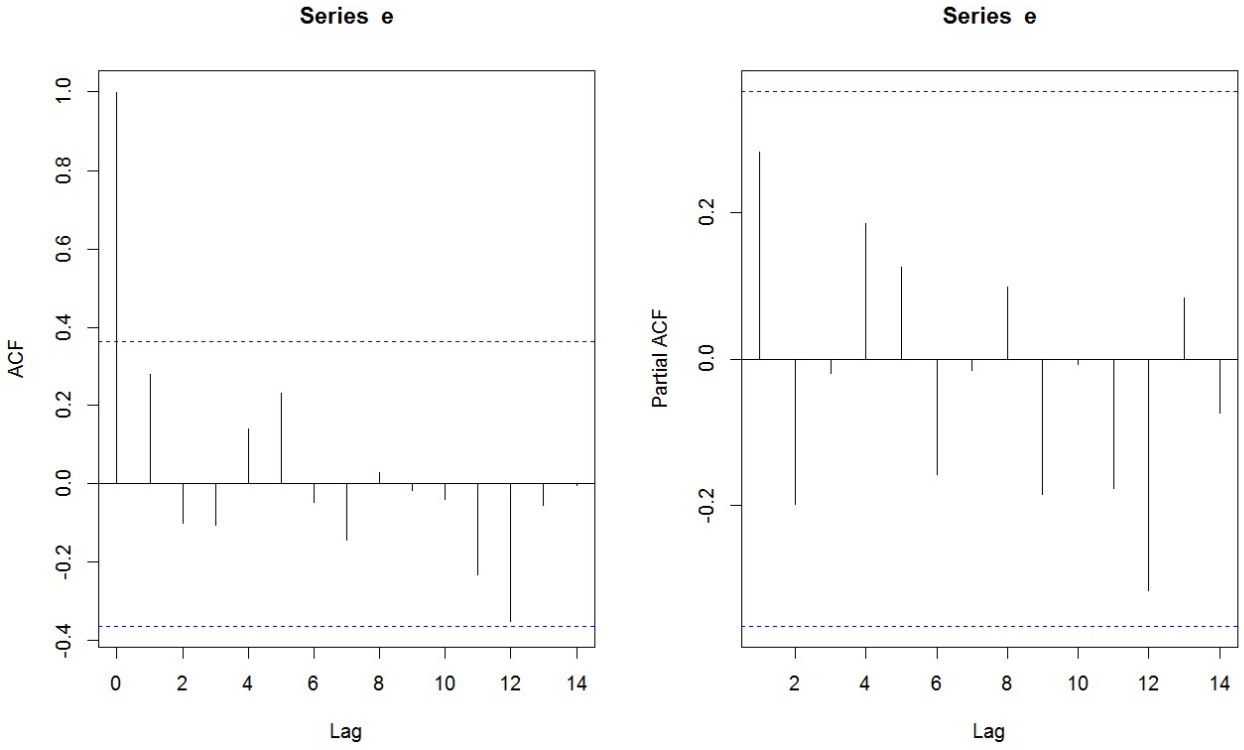

On constate que le nuage de points obtenu n'est pas ajustable par une "ligne" et la moyenne des valeurs de la ligne rouge est quasi nulle ; on admet que $\epsilon$ et $X1$ sont indépendantes.

### Indépendance de $\epsilon_1, \ldots, \epsilon_n$

Les observations de $(Y, X1)$ portent sur des individus tous différents, il doit donc y avoir indépendance de $\epsilon_1, \ldots, \epsilon_n$.

On vérifie cela avec les graphiques *acf* et *pacf* :

```
par(mfrow = c(1, 2))
acf(e)
pacf(e)
```

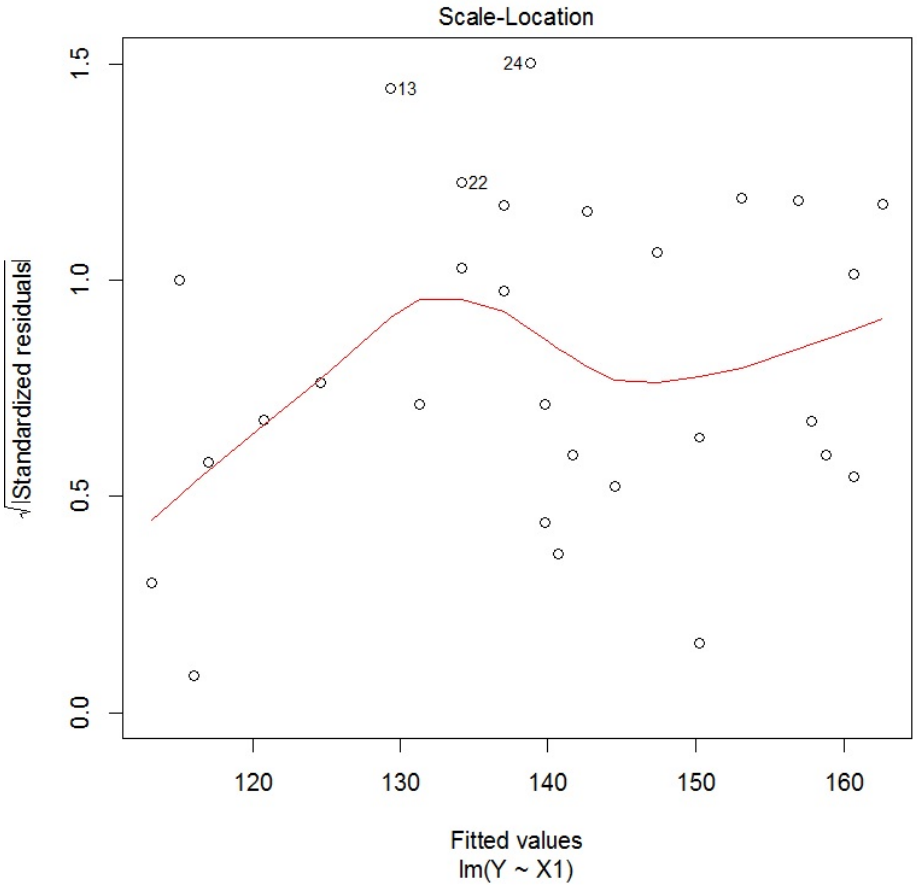



On ne constate aucune structure particulière et peu de bâtons dépassent les bornes limites ; on admet l'indépendance de $\epsilon_1, \ldots, \epsilon_n$.

### Égalité des variances de $\epsilon_1, \ldots, \epsilon_n$

Une indication graphique sur l'égalité des variances de $\epsilon_1, \ldots, \epsilon_n$ est donnée par :

```
plot(reg, 3)
```




On ne constate pas de structure particulière, ce qui traduit une égalité des variances.

On étudie celle-ci avec le test de Breusch-Pagan :

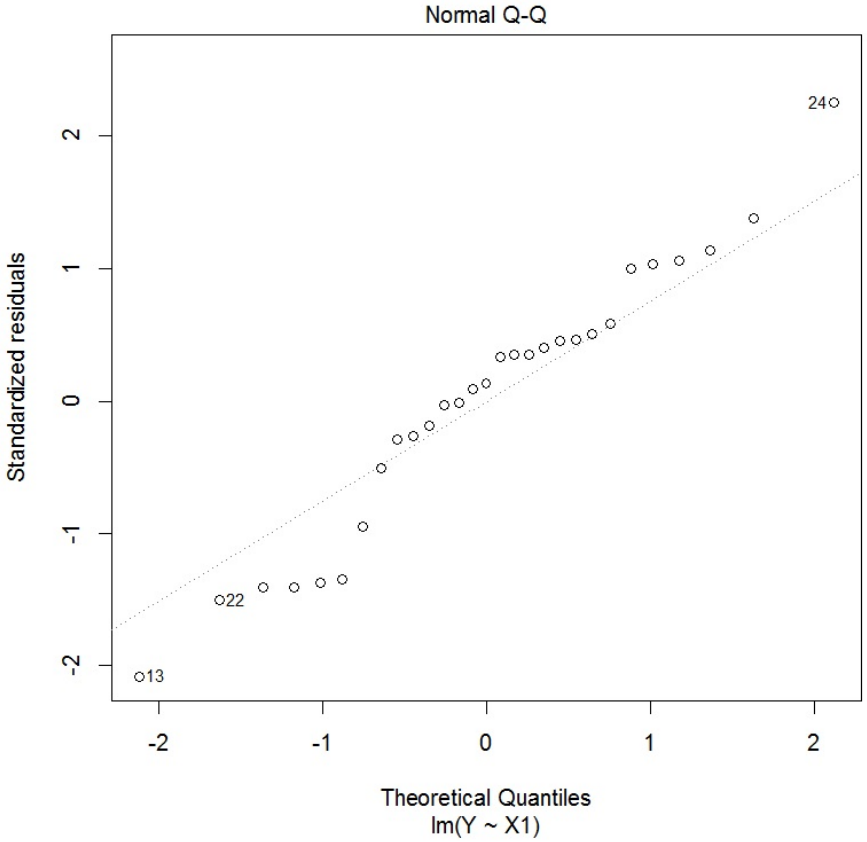
```
library(lmtest)

bptest(reg)
```

Cela renvoie : p-valeur $= 0.7636$. Comme p-valeur $> 0.05$, on admet l'égalité des variances.

### Normalité de $\epsilon_1, \ldots, \epsilon_n$

On trace le QQ plot associé :

```
plot(reg, 2)
```




On constate que les points sont à peu près alignés, ce qui traduit la normalité de $\epsilon_1, \ldots, \epsilon_n$.

On peut vérifier cela avec le test de Shapiro-Wilk :

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

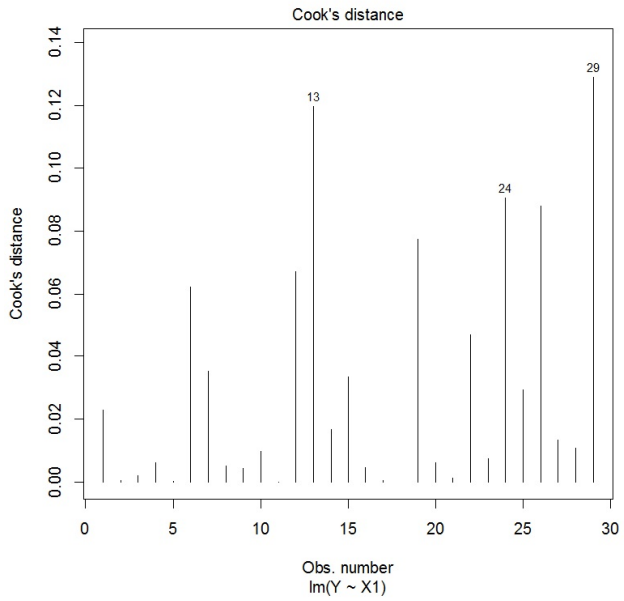
Cela renvoie : p-valeur $= 0.38$. Comme p-valeur $> 0.05$, on admet la normalité de $\epsilon_1, \ldots, \epsilon_n$.

# Compléments

## Détection des valeurs anormales

On étudie les distances de Cook des observations :

```
plot(reg, 4)
```




Aucune d'entre elles ne dépasse 1, il n'y a pas de valeur anormale a priori.

## AIC et BIC

En complément du $\overline{R}^2$, calculons le AIC et le BIC du modèle.

```
AIC(reg)
```

Cela renvoie 217.1864.

```
BIC(reg)
```

Cela renvoie 221.2883.

# Conclusion et études similaires

## Conclusion

L'étude statistique mise en œuvre montre que le modèle de *rlm* est adapté au problème ; les hypothèses permettant la validation des principaux résultats d'estimation sont vérifiées.

## Étude similaire 1 : Scores

On peut considérer le jeu de données "scores" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/scores.txt",
header = T)
```

Une étude a été menée auprès de 19 étudiants afin de mettre en évidence une relation entre le score (note) final à un examen de mathématiques et le temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose :

- du temps de révision en heures (variable $X1$),
- du score obtenu sur 800 points (variable $Y$).

## Étude similaire 2 : Fibres

On peut considérer le jeu de données "fibres" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/fibres.txt",
header = T)
```

Une étude s'intéresse à la vitesse de propagation de l'influx nerveux dans une fibre nerveuse. Pour 16 fibres nerveuses différentes, on considère :

- le diamètre en microns (variable $X1$),
- la vitesse de l'influx nerveux en m/s (variable $Y$).

On souhaite expliquer $Y$ à partir de $X1$.

### Étude similaire 3 : Toluca

On peut considérer le jeu de données "toluca" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/toluca.txt",
header = T)
```

L'entreprise Toluca fabrique des pièces de rechange pour l'équipement de réfrigération. Pour une pièce particulière, le processus de production prend un certain temps. Dans le cadre d'un programme d'amélioration des coûts, l'entreprise souhaite mieux comprendre la relation entre :

- la taille du lot (variable $X1$),
- nombre total d'heures de travail (variable $Y$).

Les données ont été rapportées pour 25 lots représentatifs de taille variable.

### Étude similaire 4 : Eaux usées

On peut considérer le jeu de données "eaux usées" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/eauxusées.txt",
header = T)
```

Une nouvelle machine pour le traitement des eaux usées est à l'étude. En particulier, les ingénieurs s'intéressent à :

- la vitesse de filtration mesurée en pour cent (variable $X1$),
- l'humidité des granulés en kg-DS/m/h (variable $Y$).

Les données ont été rapportées pour 20 expériences indépendantes.

On souhaite expliquer $Y$ à partir de $X1$.

## Étude similaire 5 : Blé

On peut considérer le jeu de données "blé" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/blé.txt", header = T)
```

L'étude porte sur le rendement d'une culture de blé en fonction de la hauteur de pluie printanière. Pour 54 parcelles, on dispose :

- du rendement de blé (variable $Y$),
- de la hauteur de pluie en mètres (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$.

## Étude similaire 6 : Oxygène

On peut considérer le jeu de données "oxygène" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/oxygène.txt",
header = T)
```

Une étude a été menée auprès de 31 individus afin de mettre en évidence une relation entre la consommation d'oxygène et le temps de l'effort lors de séances d'aérobie.

Pour chaque individu, on dispose :

- du temps de l'effort en minutes (variable $X1$),
- de la consommation d'oxygène (variable $Y$).

## Étude similaire 7 : Loyers

On peut considérer le jeu de données "loyers" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/loyers.txt",
header = T)
```

Dans un quartier parisien, une étude a été menée afin de mettre en évidence une relation entre le loyer mensuel et la surface des appartements ayant exactement 3 pièces.

Pour 30 appartements de ce type, on dispose :

– de la surface en mètres carrés (variable $X1$),

– du loyer mensuel en francs (variable $Y$).

## Étude complémentaire : ANOVA à 1 facteur

L'étude porte sur l'effet d'un médicament $Y$ sur des malades classés en 3 groupes en fonction de leur âge. Les données sont :

Groupe 1 : 14.7, 15.2, 16.3, 17.1, 19.1

Groupe 2 : 17.6, 19.9, 23.1, 22.5, 21.4

Groupe 3 : 22.4, 25.1, 27.0, 28.1, 30.3

On peut alors introduire une variable qualitative $X1$ avec les 3 modalités : Groupe 1, Groupe 2 et Groupe 3.

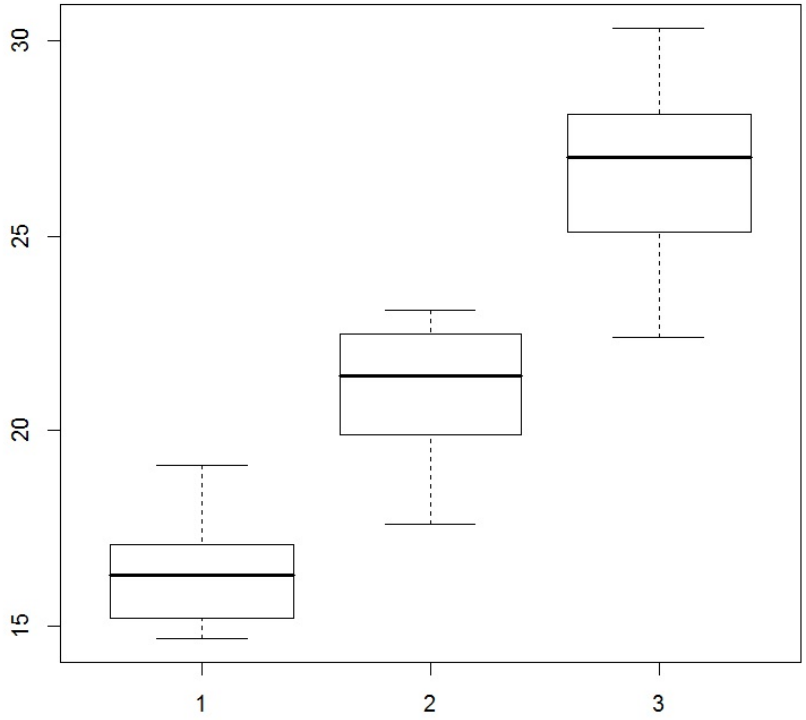
On souhaite étudier l'influence de $X1$ sur $Y$.

Comme $X1$ est qualitative, on ne peut pas modéliser directement ce problème par une *rls* standard.

Pour étudier l'influence de $X1$ sur $Y$, on se place dans le cadre ANOVA à 1 facteur.

Les commandes clés sont :

```
w = data.frame(matrix(c(14.7, 15.2, 16.3, 17.1, 19.1, 17.6, 19.9, 23.1,
22.5, 21.4, 22.4, 25.1, 27.0, 28.1, 30.3, rep(1 : 3, each = 5)), ncol = 2,
byrow = FALSE))
colnames(w) = c("Y", "X1")
w$X1 = as.factor(w$X1)
attach(w)
boxplot(Y ~ X1)
```




Les 3 boîtes sont à des niveaux très différents ce qui traduit une influence de $X1$ sur $Y$. On confirme cela avec le test statistique approprié :

```
mod = aov(Y ~ X1)

summary(mod)
```

Cela renvoie :

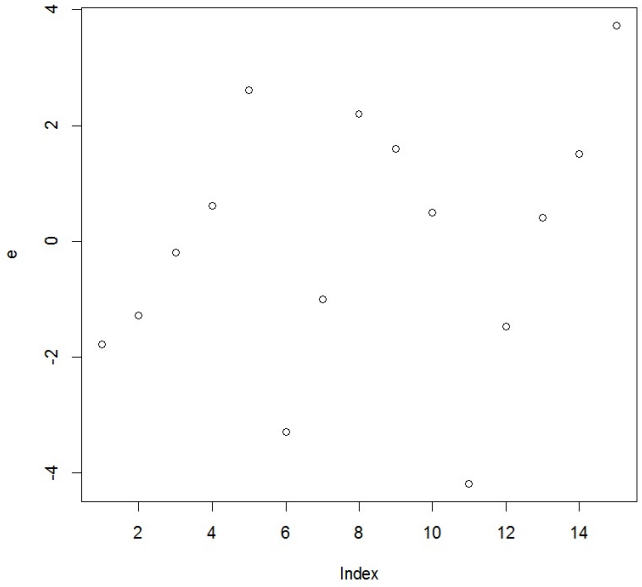
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|---|---|---|---|---|---|---|
| X1 | 2 | 256.35 | 128.17 | 22.75 | 0.0001 | $***$ |
| Residuals | 12 | 67.62 | 5.63 | | | |

On a : p-valeur $< 0.0001$ : $***$. Il y a une influence hautement significative de $X1$ sur $Y$.

La validation des hypothèses de l'ANOVA à 1 facteur repose sur les mêmes bases que la *rls*.

Les résidus s'obtiennent par les commandes :

```
e = residual(mod)
plot(e)
```




Il n'y a rien de suspect à signaler ; on peut vérifier l'indépendance des résidus avec les corrélogrammes : `acf(e)` et `pacf(e)`, la normalité des résidus avec le test de Shapiro-Wilk : `shapiro.test(e)` et l'homoscédasticité avec le test de Bartlett : `bartlett.test(e, X1)`.

Pour compléter l'étude, on peut aussi comparer deux à deux les moyennes de $Y$ sous les modalités : Groupe 1, Groupe 2 et Groupe 3, par le test de TukeyHDS :

```
TukeyHSD(mod)
```

Celui-ci montre une différence significative pour les groupes 3 et 1, et les groupes 3 et 2 quant à $Y$.

*Note : l'étude n° 2 "Cigarettes" proposera une approche alternative en modélisant les modalités de variables qualitatives par des variables binaires et en les intégrant dans une rlm.*

# 1 Étude n° 1 : Horloges

## Contexte

Dans une vente aux enchères, 32 horloges anciennes différentes ont trouvé preneur. Pour chacune d'entre elles, on dispose :

- du prix de vente en pounds sterling (variable $Y$), (1 pound = 1.2553 euros),
- de l'âge de l'horloge en années (variable $X1$),
- du nombre de personnes qui ont fait une offre sur celle-ci (variable $X2$).

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

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude1.txt",
header = T)
head(w)
```

| | X1 | X2 | Y |
|---|---|---|---|
| 1 | 127 | 13 | 1235 |
| 2 | 115 | 12 | 1080 |
| 3 | 127 | 7 | 845 |
| 4 | 150 | 9 | 1522 |
| 5 | 156 | 6 | 1047 |
| 6 | 182 | 11 | 1979 |

On associe les variables $Y$, $X1$ et $X2$ aux valeurs associées :

```
attach(w)
```

# Régression linéaire multiple

## Modélisation

Une première approche est de considérer le modèle de *rlm* :

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

où $\beta_0$, $\beta_1$ et $\beta_2$ sont 3 coefficients inconnus et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma$ inconnu.

La présence de $\beta_0$ est justifiée car même une horloge récente avec une offre peut avoir un prix non négligeable.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

## Estimations

La modélisation de la *rlm* et les estimations des paramètres par la méthode des *mco* s'obtiennent par les commandes :

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

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -1336.7221 | 173.3561 | -7.71 | 0.0000 | *** |
| X1 | 12.7362 | 0.9024 | 14.11 | 0.0000 | *** |
| X2 | 85.8151 | 8.7058 | 9.86 | 0.0000 | *** |

```
Residual standard error: 133.1 on 29 degrees of freedom

Multiple R-squared: 0.8927, Adjusted R-squared: 0.8853

F-statistic: 120.7 on 2 and 29 DF, p-value: 8.769e-15
```

– Estimations ponctuelles de $\beta_0$, $\beta_1$ et $\beta_2$ :

| $\widehat{\beta}_0$ | $\widehat{\beta}_1$ | $\widehat{\beta}_2$ |
|---|---|---|
| $-1336.7221$ | 12.7362 | 85.8151 |

– Estimations ponctuelles des écart-types des estimateurs de $\beta_0$, $\beta_1$ et $\beta_2$ :

| $\widehat{\sigma}(\widehat{\beta}_0)$ | $\widehat{\sigma}(\widehat{\beta}_1)$ | $\widehat{\sigma}(\widehat{\beta}_2)$ |
|---|---|---|
| 173.3561 | 0.9024 | 8.7058 |

– $t_{obs}$ :

| $H_1$ | $\beta_0 \neq 0$ | $\beta_1 \neq 0$ | $\beta_2 \neq 0$ |
|---|---|---|---|
| $t_{obs}$ | $-7.71$ | 14.11 | 9.86 |

– Degrés de significativité :

| $H_1$ | $\beta_0 \neq 0$ | $\beta_1 \neq 0$ | $\beta_2 \neq 0$ |
|---|---|---|---|
| degré | $***$ | $***$ | $***$ |

– $R^2 = 0.8927$ et $\overline{R}^2 = 0.8853$ : cela est tout à fait correct,

– Test de Fisher : p-valeur $< 0.001$, $***$ : l'utilisation du modèle de *rlm* est pertinente.

La valeur prédite de $Y$ quand $X1 = 157$ et $X2 = 11$ (par exemple) est donnée par

```
predict(reg, data.frame(X1 = 157, X2 = 11))
```

Cela renvoie 1606.828.

Par conséquent, une horloge qui a 157 ans et sur laquelle 11 personnes ont fait une offre sera vendue, en moyenne, 1606.828 pounds.

On peut aussi s'intéresser :

– aux intervalles de confiance pour $\beta_0$, $\beta_1$ et $\beta_2$ au niveau 99% (par exemple) :

```
confint(reg, level = 0.99)
```

Cela renvoie :

| | 0.5 % | 99.5 % |
|---|---|---|
| (Intercept) | -1814.55843 | -858.88567 |
| X1 | 10.24889 | 15.22351 |
| X2 | 61.81871 | 109.81156 |

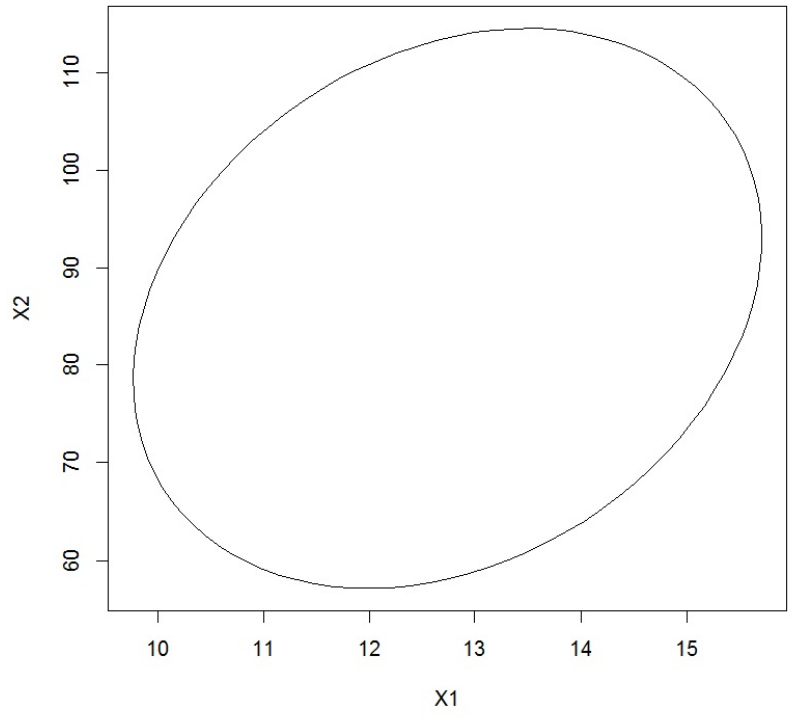
Le tableau donne les bornes inférieures et supérieures des intervalles de confiance de $\beta_0$, $\beta_1$ et $\beta_2$ :

| $i_{\beta_0}$ | $i_{\beta_1}$ | $i_{\beta_2}$ |
|---|---|---|
| $[-1814.55843,\ -858.88567]$ | $[10.24889,\ 15.22351]$ | $[61.81871,\ 109.81156]$ |

– à l'ellipsoïde de confiance de $(\beta_1, \beta_2)$ au niveau 99% (par exemple).

Les commandes sont :

```
library(ellipse)

plot(ellipse(reg, c(2, 3), level = 0.99), type = "l")
```



– à l'intervalle de confiance pour la valeur moyenne de $Y$ quand $X1 = 157$ et $X2 = 11$ (par exemple) :

```
predict(reg, data.frame(X1 = 157, X2 = 11), interval = "confidence")
```

Cela renvoie :

| fit | lwr | upr |
|---|---|---|
| 1606.828 | 1545.249 | 1668.407 |

La première valeur du tableau est celle de la valeur prédite de $Y$ quand $X1 = 157$ et $X2 = 11$ (déjà vue), les deux autres correspondent aux bornes inférieures et supérieures des intervalles de confiance de cette valeur.
Ainsi, pour $(X1, X2) = (157, 11) = x$, on a

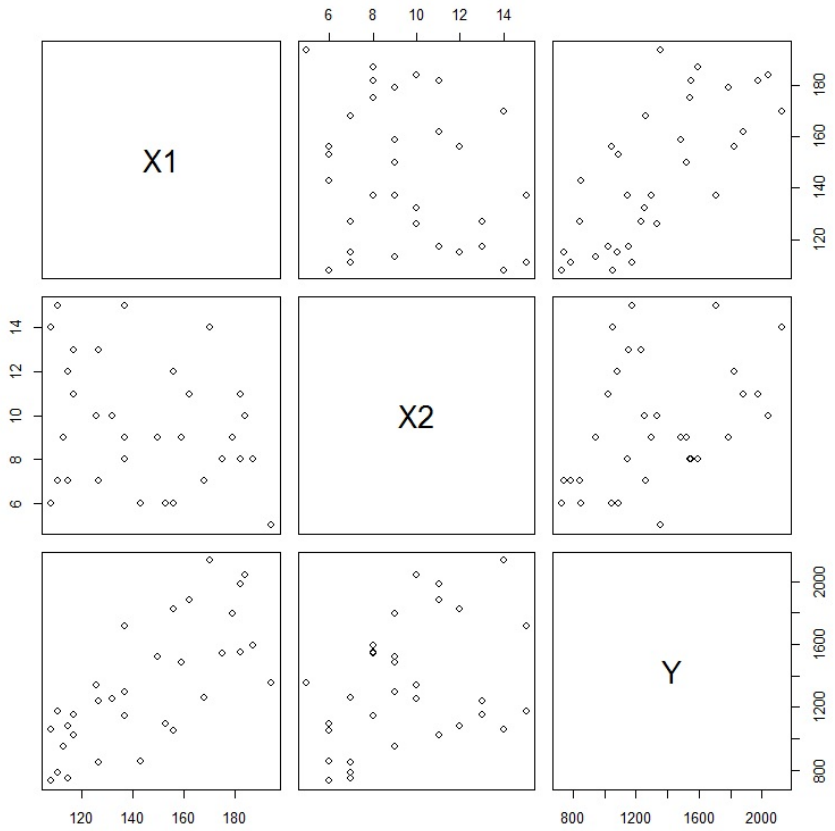
$$i_{y_x} = [1545.249,\ 1668.407].$$

# Validation des hypothèses

## Analyse des nuages de points

On trace les nuages de points des variables par pairs :

```
plot(w)
```



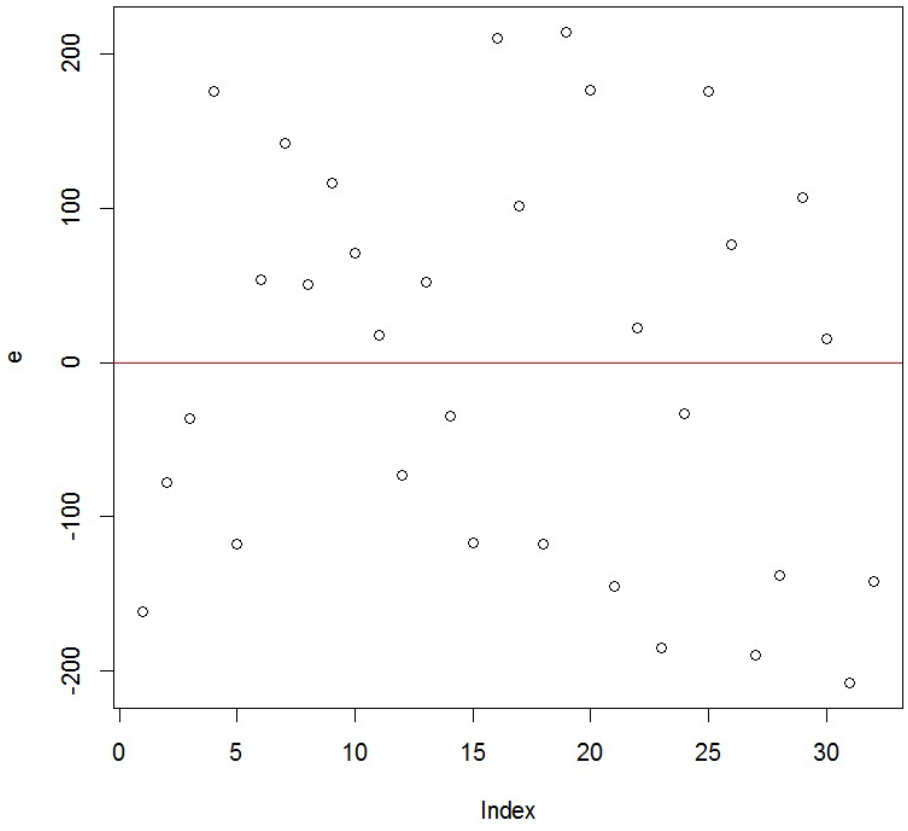

On remarque qu'une liaison linéaire entre $Y$ et $X1$ est effectivement envisageable. C'est un peu moins clair entre $Y$ et $X2$.

Cette analyse amène une vague idée de modélisation ; des méthodes plus rigoureuses seront présentées dans des études futures.

### Analyse graphique des résidus

On examine les résidus en faisant :

```
e = residuals(reg)
plot(e)
abline(h = 0, col = "red")
```

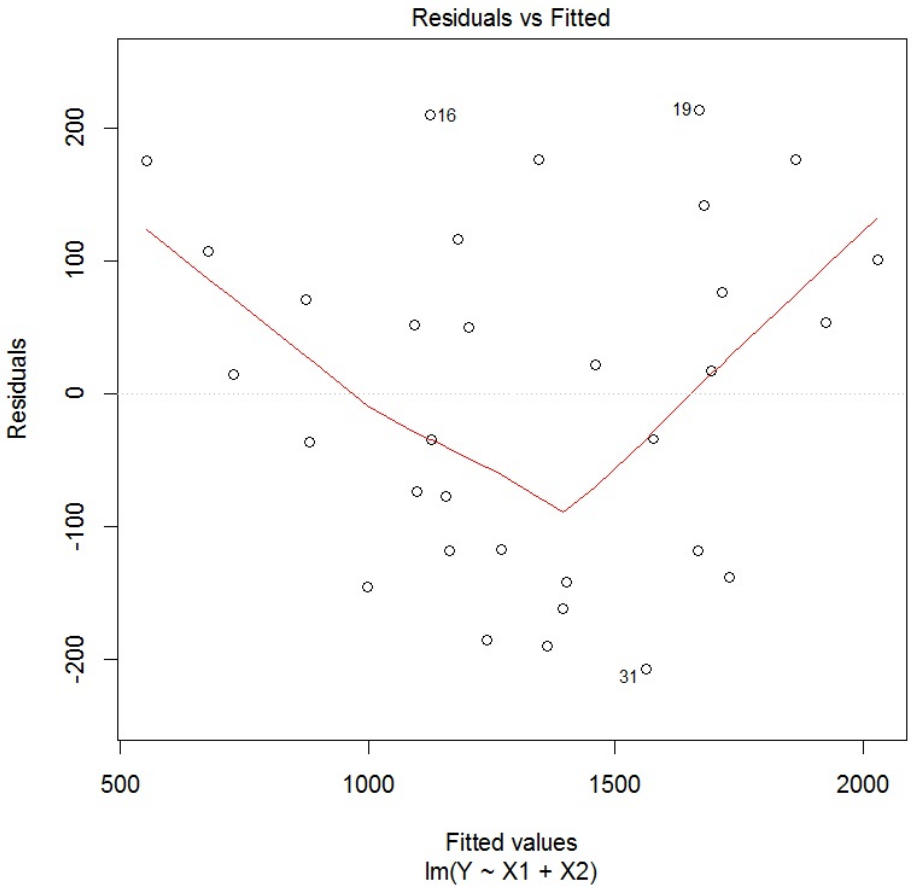



On constate une relative symétrie des résidus par rapport à l'axe des abscisses et pas de structure apparente. Cela est encourageant pour la validation des hypothèses standards.

### Indépendance de $\epsilon$ et $X1, X2$

On trace le nuage de points $\{(\text{résidus}_i, \text{prédictions en } (x_{1,i}, x_{2,i}))\}$ :

```
plot(reg, 1)
```



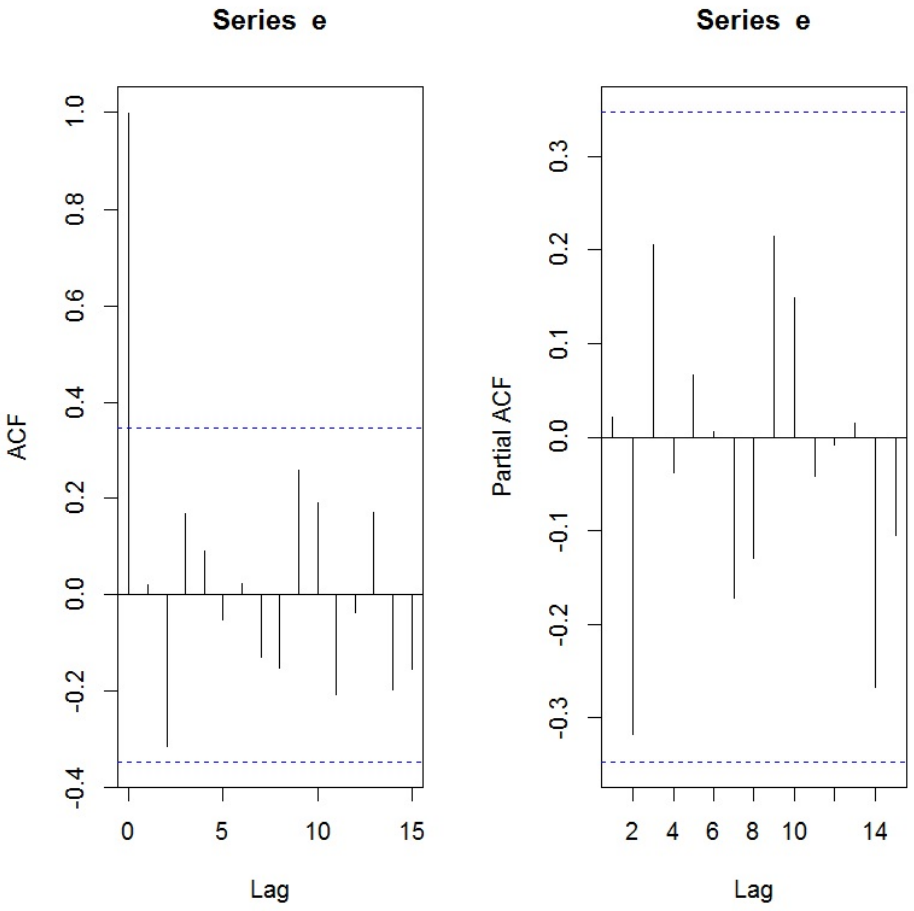

On constate que le nuage de points obtenu est difficilement ajustable par une "ligne" et la moyenne des valeurs de la ligne rouge est quasi nulle ; on admet que $\epsilon$ et $X1, X2$ sont indépendantes.

### Indépendance de $\epsilon_1, \ldots, \epsilon_n$

Les observations de $(Y, X1, X2)$ portent sur des horloges toutes différentes, il doit donc y avoir indépendance de $\epsilon_1, \ldots, \epsilon_n$.

On examine cela avec les graphiques *acf* et *pacf* :

```
par(mfrow = c(1, 2))
acf(e)
pacf(e)
```

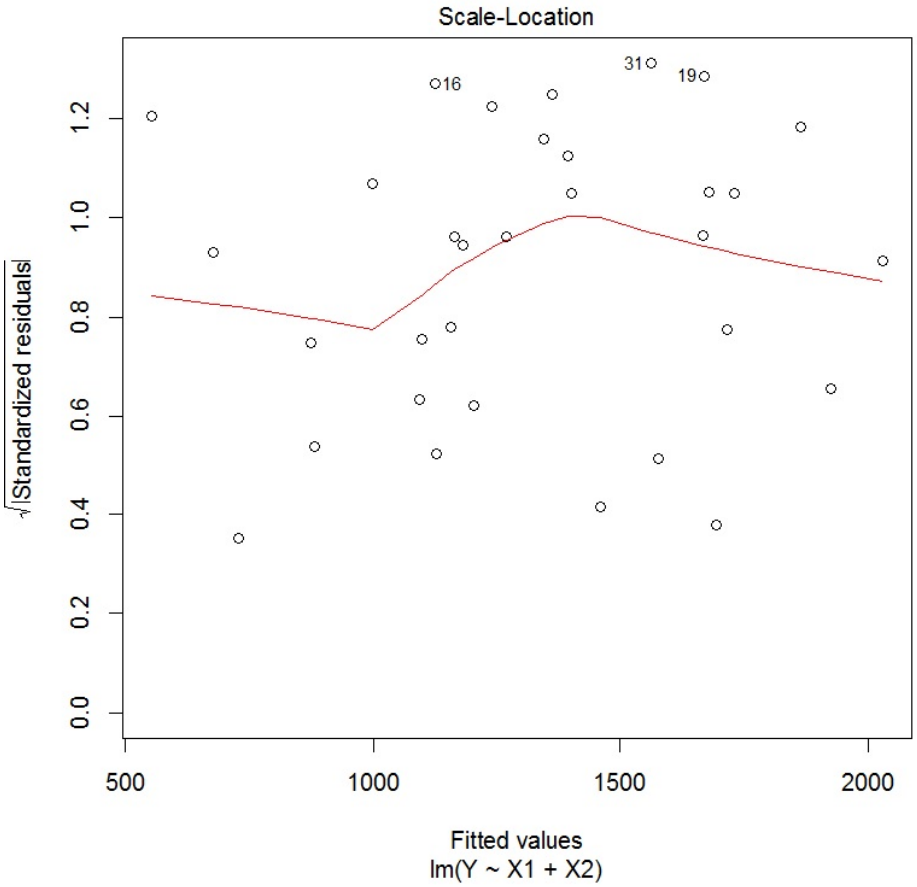



On ne constate aucune structure particulière (et pas de bâtons dépassent les bornes limites, à part le premier, ce qui est normal) ; on admet l'indépendance de $\epsilon_1, \ldots, \epsilon_n$.

### Égalité des variances de $\epsilon_1, \ldots, \epsilon_n$

Une indication graphique sur l'égalité des variances de $\epsilon_1, \ldots, \epsilon_n$ est donnée par :

```
plot(reg, 3)
```




On ne constate pas de structure particulière, ce qui traduit une égalité des variances.

On étudie l'égalité des variances avec le test de Breusch-Pagan :

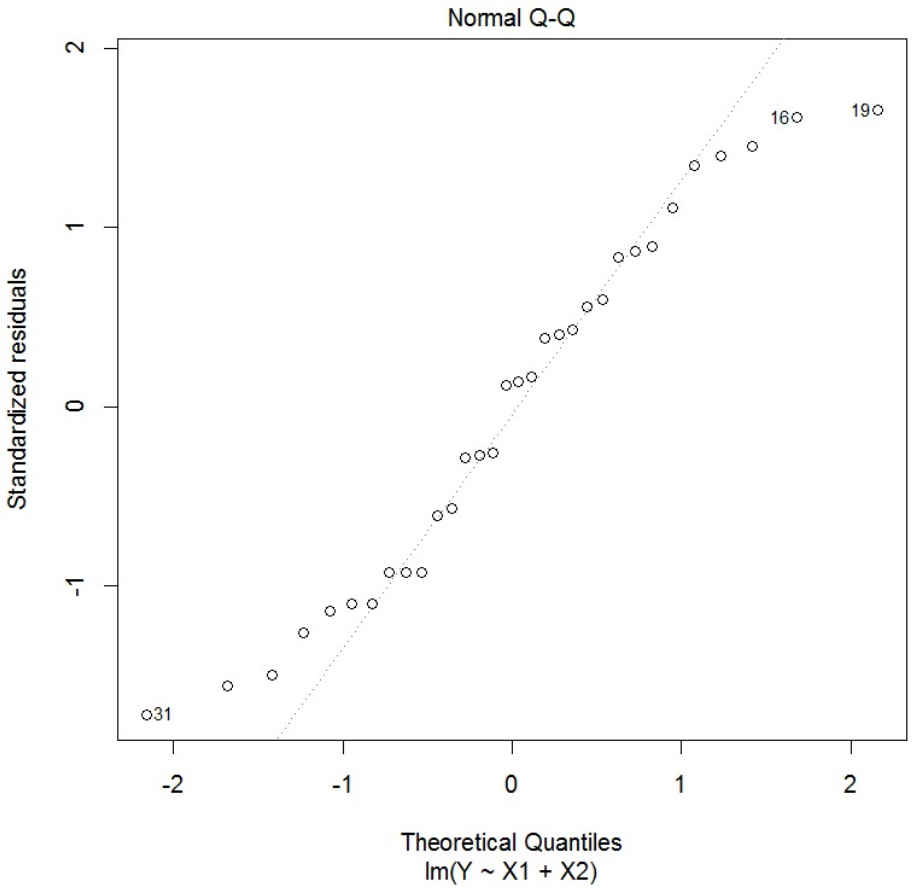
```
library(lmtest)

bptest(reg)
```

Cela renvoie : p-valeur = 0.8038. Comme p-valeur $> 0.05$, on admet l'égalité des variances.

### Normalité de $\epsilon_1, \ldots, \epsilon_n$

On trace le QQ plot associé :

```
plot(reg, 2)
```




On constate que les points sont à peu près alignés, ce qui traduit la normalité de $\epsilon_1, \ldots, \epsilon_n$.

On peut vérifier cela avec le test de Shapiro-Wilk :

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

Cela renvoie : p-valeur $= 0.1215$. Comme p-valeur $> 0.05$, on admet la normalité de $\epsilon_1, \ldots, \epsilon_n$.

# Compléments

## Étude de la multicolinéarité

On calcule le carré du coefficient de corrélation entre $X1$ et $X2$ :

```
cor(X1, X2)^2
```

Cela renvoie 0.06438861, lequel est éloigné de $R^2 = 0.8811$. Donc, par la règle de Klein, il n'y a pas de lien linéaire entre $X1$ et $X2$.

On peut obtenir la même conclusion avec les *vif* :

```
library(car)

vif(reg)
```

Cela renvoie :

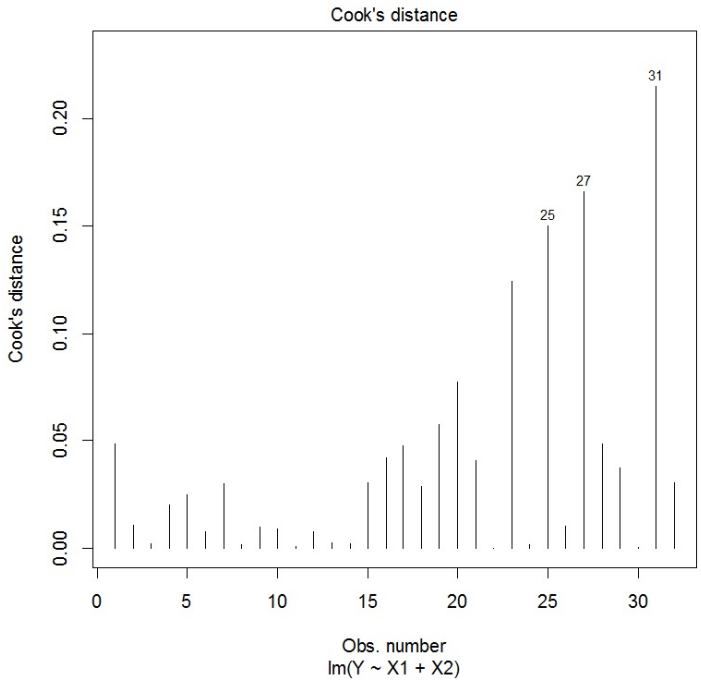
| $V_1$ | $V_2$ |
|---|---|
| 1.06882 | 1.06882 |

Comme les *vif* sont inférieurs à 5, il n'y a pas de lien linéaire entre $X1$ et $X2$.

## Détection des valeurs anormales

On étudie les distances de Cook des observations :

```
plot(reg, 4)
```




Aucune d'entre elles ne dépasse 1, il n'y a pas de valeur anormale a priori.

## AIC et BIC

En complément du $\overline{R}^2$, calculons

– le AIC :

```
AIC(reg)
```

Cela renvoie 408.71.

– le BIC :

```
BIC(reg)
```

Cela renvoie 414.5729.

# Conclusion et études similaires

## Conclusion

L'étude statistique mise en œuvre montre que le modèle de *rlm* est adapté au problème ; les hypothèses permettant la validation des principaux résultats d'estimation sont vérifiées.

## Étude similaire 1 : Fromages

On peut considérer le jeu de données "fromages" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/fromages.txt",
header = T)
```

Le goût d'un fromage dépend de la concentration de plusieurs composés chimiques, dont :

- la concentration d'acide acétique (variable $X1$),
- la concentration d'hydrogène sulfuré (variable $X2$),
- la concentration d'acide lactique (variable $X3$).

Pour 30 types de fromage, on dispose du score moyen attribué par des consommateurs (variable $Y$).

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

## Étude similaire 2 : Analyse des ventes

On peut considérer le jeu de données "dwayne" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/dwayne.txt",
header = T)
```

Une entreprise "Dwayne Portrait Studio" a fait une analyse des ventes sur la base de données à partir de 21 villes. Pour chacune d'entre elles, on dispose :

- des ventes en milliers de dollars (variable $Y$),
- du nombre de personnes de moins de 16 ans (variable $X1$),
- du revenu disponible par habitant en milliers de dollars (variable $X2$).

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

# 2 Étude n° 2 : Cigarettes

## Contexte

Les données considérées sont issues d'une étude sur le taux de tabagisme au travail en Angleterre. Pour 25 groupes professionnels, on dispose :

- d'un indice de cigarettes fumées (variable $X1$),
- d'un indice de mortalité par le cancer du poumon (variable $Y$),
- de l'occupation principale du groupe (variable $X2$, de modalités Outdoor, Factory, Office).

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude2.txt",
header = T)
head(w)
```

| | X1 | Y | X2 |
|---|---|---|---|
| 1 | 77 | 84 | Outdoor |
| 2 | 137 | 116 | Outdoor |
| 3 | 117 | 123 | Factory |
| 4 | 94 | 128 | Factory |
| 5 | 116 | 155 | Factory |
| 6 | 102 | 101 | Factory |

On associe les variables $Y$, $X1$ et $X2$ aux valeurs associées :

```
attach(w)
```

# Régression linéaire multiple

## Modélisation

Notons que $Y$ et $X1$ sont des variables quantitatives, alors que $X2$ est qualitative nominale de modalités : Office, Outdoor et Factory. Pour modéliser $X2$, on considère les 3 variables :

$$X2Office = \mathbf{1}_{\{X2=Office\}}, \qquad X2Outdoor = \mathbf{1}_{\{X2=Outdoor\}}, \qquad X2Factory = \mathbf{1}_{\{X2=Factory\}},$$

chacune valant 1 si $X2$ est égale à la modalité associée, et 0 sinon. Comme ces variables sont liées (leur somme fait 1), nous allons considérer uniquement 2 variables.

Une convention consiste à éliminer celle qui correspond à la situation la plus courante :

```
summary(X2)
```

Cela renvoie :

| Factory | Office | Outdoor |
|---|---|---|
| 11 | 7 | 7 |

Donc $X2Factory$ est la plus courante, on l'élimine et on considère $X2Office$ et $X2Outdoor$.

Une autre convention consiste à éliminer la modalité arrivant premier dans l'ordre alphabétique ; c'est ce que fait le logiciel R. On élimine alors encore $X2Factory$.

On modélise alors le problème général comme une *rlm* en prenant en compte la possible interaction de $X2$ avec $X1$ :

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2Office + \beta_3 X2Outdoor + \beta_4 X1 : X2Office + \beta_5 X1 : X2Outdoor + \epsilon,$$

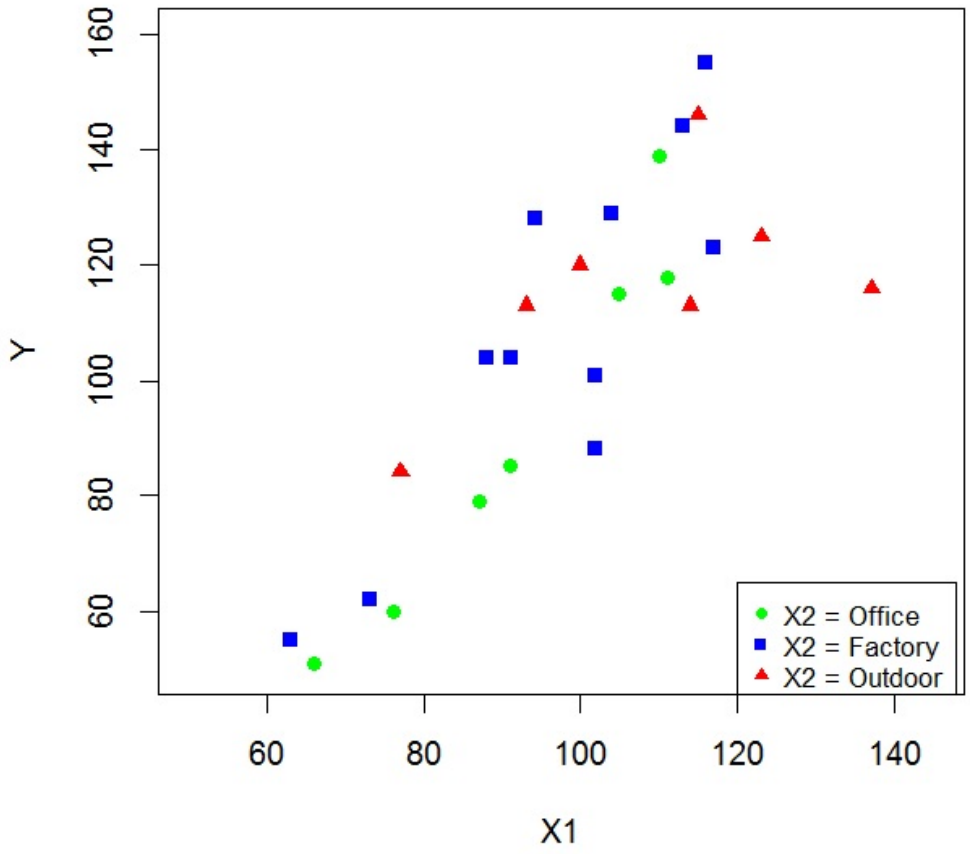
où $\beta_0, \ldots, \beta_5$ sont 6 coefficients inconnus et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma$ inconnu.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

### Nuage et sous-nuages de points

Le nuage de points de $(X1, Y)$ en fonction des modalités de $X2$ est donné par :

```
plot(X1[X2 == "Office"], Y[X2 == "Office"], pch = 16, ylab = "Y",
xlab = "X1", xlim = c(50, 145), ylim = c(50, 160), col = "green")
points(X1[X2 == "Factory"], Y[X2 == "Factory"], pch = 15, col = "blue")
points(X1[X2 == "Outdoor"], Y[X2 == "Outdoor"], pch = 17, col = "red")
legend(x = 120, y = 65, c("X2 = Office", "X2 = Factory", "X2 = Outdoor"),
cex = 0.8, col = c("green", "blue", "red"), pch = c(16, 15, 17))
```



Si on considère les sous-nuages de points correspondants aux 3 modalités de $X2$, il est envisageable de les ajuster par une droite. Le fait que les points correspondants aux différentes modalités soient mélangés traduit la présence d'une interaction entre $X1$ et $X2$ sur $Y$.

## Estimations

La modélisation de la *rlm* avec les variables explicatives $X1$, $X2Office$, $X2Outdoor$, $X1 : X2Office$ et $X1 : X2Outdoor$, et les estimations des paramètres par la méthode des *mco* s'obtiennent par les commandes :

```
reg = lm(Y ~ X1 * X2)

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -43.9772 | 26.8714 | -1.64 | 0.1182 | |
| X1 | 1.5774 | 0.2741 | 5.75 | 0.0000 | $***$ |
| X2Office | -31.1132 | 42.4761 | -0.73 | 0.4728 | |
| X2Outdoor | 100.6033 | 42.9926 | 2.34 | 0.0304 | $*$ |
| X1:X2Office | 0.2378 | 0.4455 | 0.53 | 0.5996 | |
| X1:X2Outdoor | -1.0232 | 0.4102 | -2.49 | 0.0220 | $*$ |

```
Residual standard error: 14.95 on 19 degrees of freedom

Multiple R-squared: 0.793, Adjusted R-squared: 0.7385

F-statistic: 14.56 on 5 and 19 DF, p-value: 6.165e-06
```

– Estimations ponctuelles de $\beta_0, \ldots, \beta_5$ :

| $\widehat{\beta}_0$ | $\widehat{\beta}_1$ | $\widehat{\beta}_2$ | $\widehat{\beta}_3$ | $\widehat{\beta}_4$ | $\widehat{\beta}_5$ |
|---|---|---|---|---|---|
| $-43.9772$ | 1.5774 | $-31.1132$ | 100.6033 | 0.2378 | $-1.0232$ |

– Estimations ponctuelles des écart-types des estimateurs de $\beta_0, \ldots, \beta_5$ :

| $\widehat{\sigma}(\widehat{\beta}_0)$ | $\widehat{\sigma}(\widehat{\beta}_1)$ | $\widehat{\sigma}(\widehat{\beta}_2)$ | $\widehat{\sigma}(\widehat{\beta}_3)$ | $\widehat{\sigma}(\widehat{\beta}_4)$ | $\widehat{\sigma}(\widehat{\beta}_5)$ |
|---|---|---|---|---|---|
| 26.8714 | 0.2741 | 42.4761 | 42.9926 | 0.4455 | 0.4102 |

– $t_{obs}$ :

| $H_1$ | $\beta_0 \neq 0$ | $\beta_1 \neq 0$ | $\beta_2 \neq 0$ | $\beta_3 \neq 0$ | $\beta_4 \neq 0$ | $\beta_5 \neq 0$ |
|---|---|---|---|---|---|---|
| $t_{obs}$ | $-1.637$ | $5.754$ | $-0.732$ | $2.340$ | $0.534$ | $-2.495$ |

– Degrés de significativité :

| $H_1$ | $\beta_0 \neq 0$ | $\beta_1 \neq 0$ | $\beta_2 \neq 0$ | $\beta_3 \neq 0$ | $\beta_4 \neq 0$ | $\beta_5 \neq 0$ |
|---|---|---|---|---|---|---|
| degré | | $***$ | | $*$ | | $*$ |

– $R^2 = 0.793$ et $\overline{R}^2 = 0.7385$ : cela est correct,

– Test de Fisher : p-valeur $= 6.165e{-}06 < 0.001$, $***$ : l'utilisation du modèle de *rlm* est pertinente.

La valeur prédite de $Y$ quand $X1 = 106$ et $X2 = Office$ (par exemple) est donnée par :

```
predict(reg, data.frame(X1 = 106, X2 = "Office"))
```

Cela renvoie 117.323.

Ainsi, un individu qui a un indice de cigarettes de 106 et dont la principale occupation est $Office$ aura, en moyenne, un indice de mortalité du cancer du poumon de 117.323.

On peut aussi s'intéresser :

– aux intervalles de confiance pour $\beta_0, \ldots, \beta_5$ au niveau 95% (par exemple) :

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

Cela renvoie :

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -100.22 | 12.27 |
| X1 | 1.00 | 2.15 |
| X2Office | -120.02 | 57.79 |
| X2Outdoor | 10.62 | 190.59 |
| X1:X2Office | -0.69 | 1.17 |
| X1:X2Outdoor | -1.88 | -0.16 |

Ainsi, on a les intervalles de confiance suivants :

| $i_{\beta_0}$ | $i_{\beta_1}$ | $i_{\beta_2}$ |
|---|---|---|
| $[-100.22,\ 12.27]$ | $[1.00,\ 2.15]$ | $[-120.02,\ 57.79]$ |

| $i_{\beta_3}$ | $i_{\beta_4}$ | $i_{\beta_5}$ |
|---|---|---|
| $[10.62,\ 190.59]$ | $[-0.69,\ 1.17]$ | $[-1.88,\ -0.16]$ |

– à l'intervalle de confiance pour la valeur moyenne de $Y$ quand $X1 = 106$ et $X2 = Office$ (par exemple) :

```
predict(reg, data.frame(X1 = 106, X2 = "Office"), interval = "confidence")
```

Cela renvoie :

| fit | lwr | upr |
|---|---|---|
| 117.323 | 101.786 | 132.86 |

Ainsi, pour $(X1, X2) = (106, Office) = x$, on a

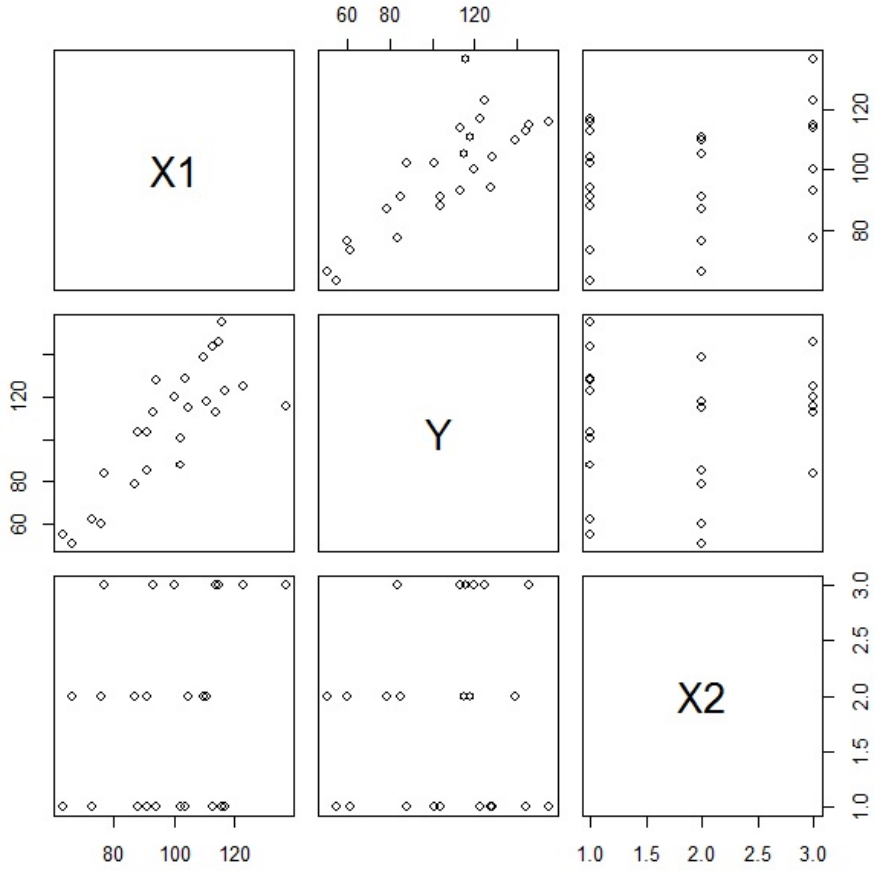
$$i_{y_x} = [101.786,\ 132.86].$$

# Validation des hypothèses

## Analyse des nuages de points

On trace les nuages de points des variables par pairs :

```
plot(w)
```

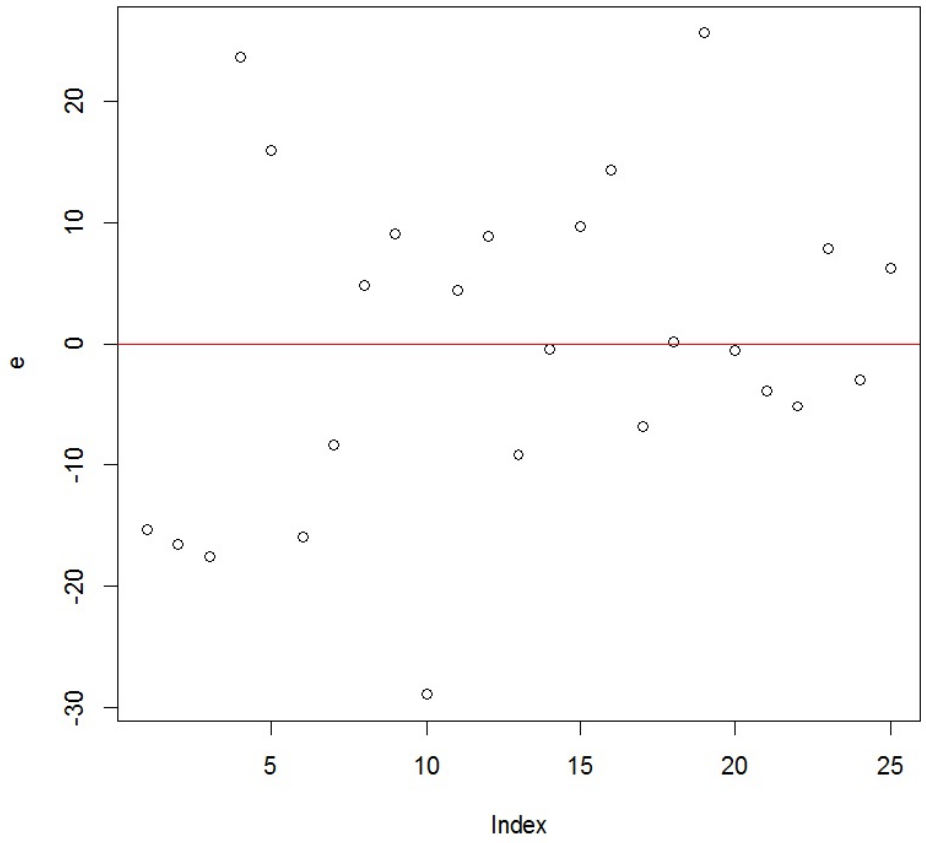



On remarque que la liaison linéaire entre $Y$ et $X1$ est envisageable, avec une légère structure en forme de mégaphone présageant une inégalité des variances d'erreurs. Cette analyse est juste pour se donner une vague idée ; on ne peut rien en conclure.

### Analyse graphique des résidus

On examine les résidus en faisant :

```
e = residuals(reg)

plot(e)

abline(h = 0, col = "red")
```

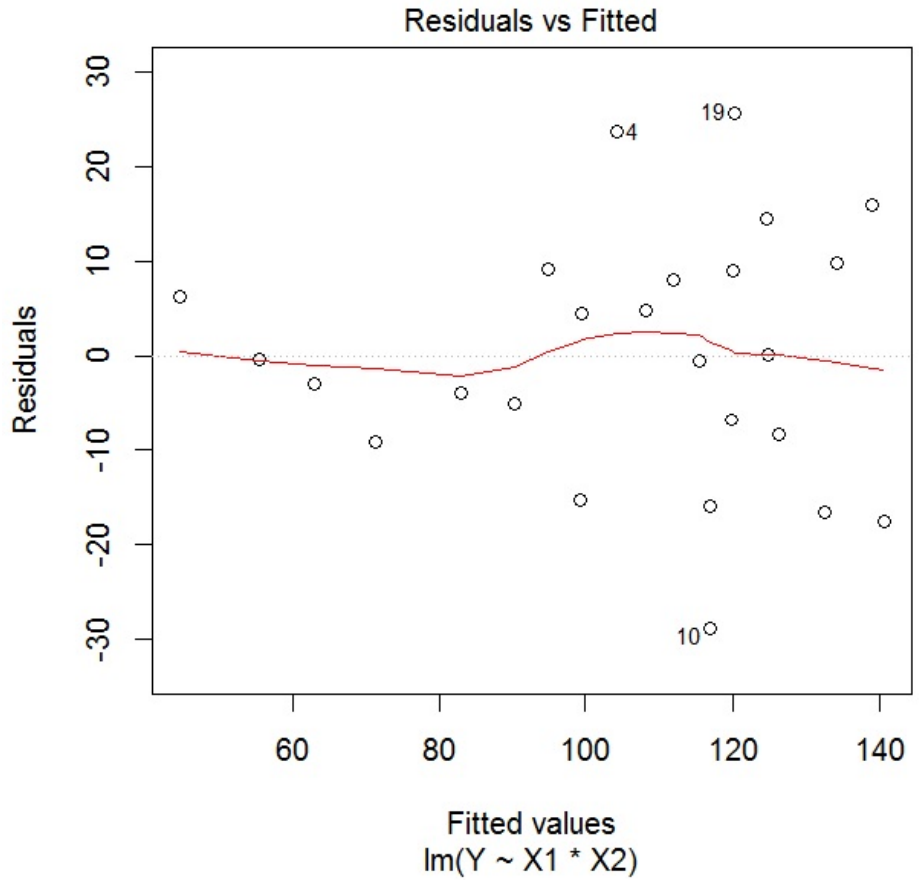



On ne distingue pas de structure. Toutefois, l'uniformité de la répartition des points est discutable.

### Indépendance de $\epsilon$ et $X1, X2$

On trace le nuage de points {(résidus$_i$, prédictions en $(x_{1,i}, x_{2,i})$} :

```
plot(reg, 1)
```



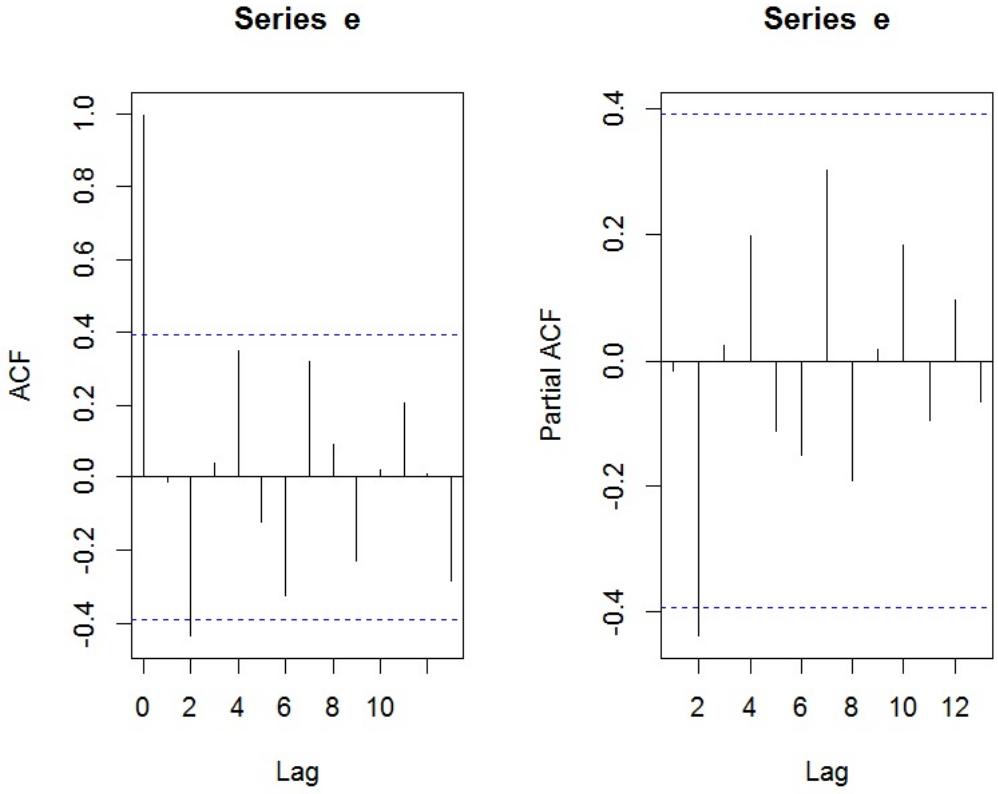

On constate que le nuage de points obtenu n'est pas ajustable par une "ligne" et la moyenne des valeurs de la ligne rouge est quasi nulle ; on admet que $\epsilon$ et $X1, X2$ sont indépendantes.

### Indépendance de $\epsilon_1, \ldots, \epsilon_n$

Les observations de $(Y, X1, X2)$ portent sur des groupes tous différents, il doit donc y avoir indépendance de $\epsilon_1, \ldots, \epsilon_n$.

On examine cela avec les graphiques *acf* et *pacf* :

```
par(mfrow = c(1, 2))

acf(e)

pacf(e)
```




On ne constate aucune structure particulière et peu de bâtons dépassent les bornes limites ; on admet l'indépendance de $\epsilon_1, \ldots, \epsilon_n$.

### Égalité des variances de $\epsilon_1, \ldots, \epsilon_n$

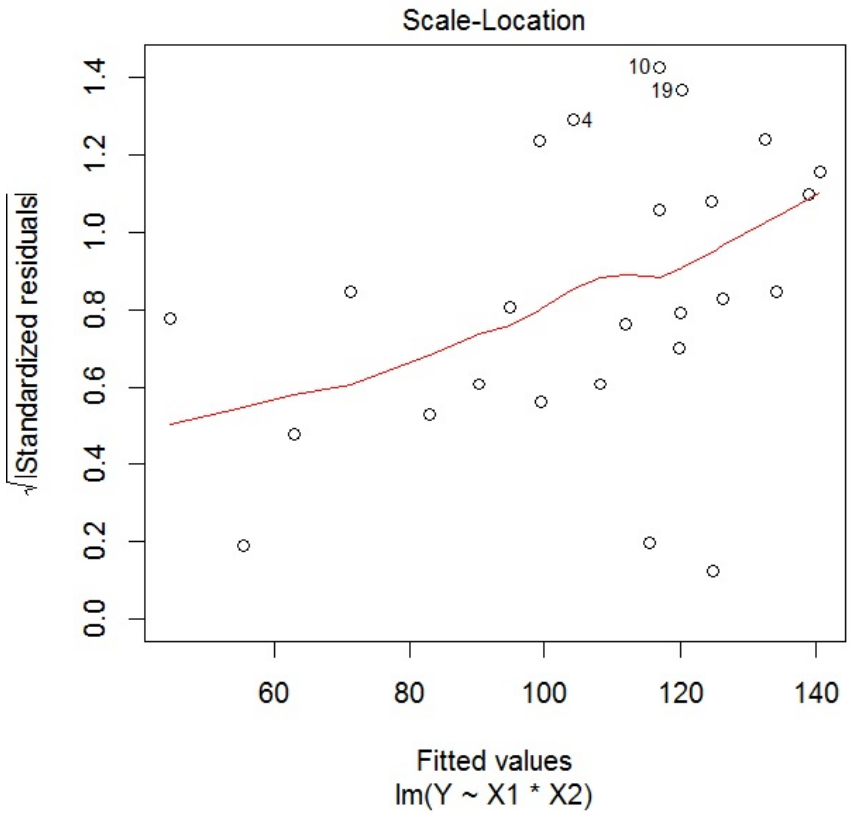
Comme il y a une variable qualitative dans le modèle, on vérifie l'hypothèse $\mathbb{V}(\epsilon_1) = \ldots = \mathbb{V}(\epsilon_n)$ en faisant :

- une première analyse graphique (nuage de points adapté et boites à moustaches pour chacune des modalités),
- un test statistique adapté pour confirmer/infirmer.

Analyse graphique :

```
plot(reg, 3)
```

Cela renvoie :




On constate une légère structure en forme de mégaphone ; l'inégalité des variances d'erreurs est à étudier.

Analysons les boites à moustaches :

```
boxplot(e ~ X2)
```


Les boites ont des étendues différentes mais il est difficile de conclure avec certitude.

Il faut alors faire un test statistique adapté pour conclure de manière rigoureuse. On utilise le test de Bartlett :

```
library(stats)

bartlett.test(e, X2)
```

Cela renvoie : p-valeur $= 0.2123 > 0.05$, donc on admet l'égalité des variances.

On aurait aussi pu faire le test de Levene :

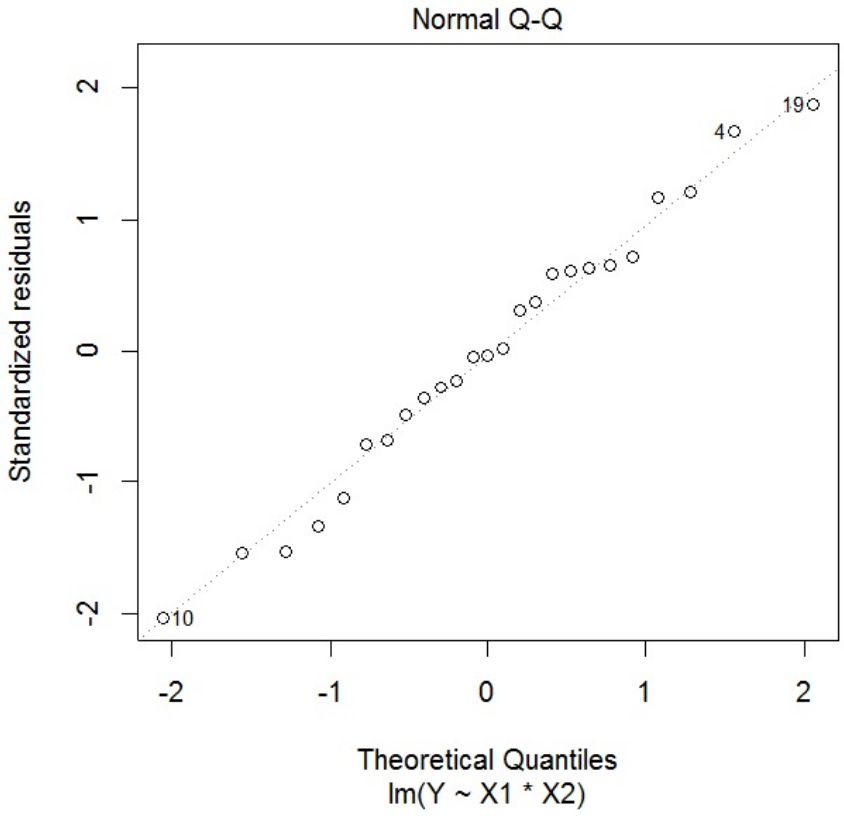
```
library(lawstat)

levene.test(e, X2)
```

Cela renvoie : p-valeur $= 0.2387 > 0.05$, on retrouve la conclusion précédente.

### Normalité de $\epsilon_1, \ldots, \epsilon_n$

On trace le QQ plot associé :

```
plot(reg, 2)
```




On constate que les points sont à peu près alignés, ce qui traduit la normalité de $\epsilon_1, \ldots, \epsilon_n$.

On peut vérifier cela avec le test de Shapiro-Wilk :

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

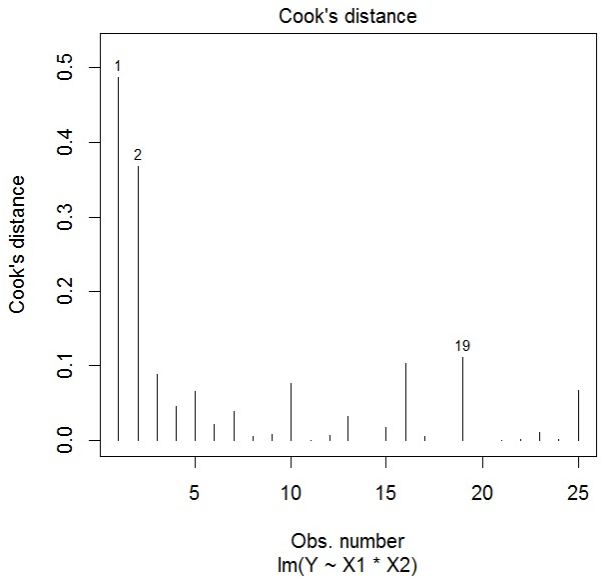
Cela renvoie : p-valeur $= 0.9737$. Comme p-valeur $> 0.05$, on admet la normalité de $\epsilon_1, \ldots, \epsilon_n$.

# Compléments

## Détection des valeurs anormales

On étudie les distances de Cook des observations :

```
plot(reg, 4)
```




Aucune d'entre elles ne dépasse 1, il n'y a pas de valeur anormale a priori.

## AIC et BIC

En complément du $\overline{R}^2$, calculons le AIC et le BIC du modèle :

```
AIC(reg)
```

Cela renvoie 213.3076.

```
BIC(reg)
```

Cela renvoie 221.8397.

# Conclusion et études similaires

## Conclusion

L'étude statistique mise en œuvre montre que le modèle de *rlm* usuel n'est pas très bien adapté au problème (le $\overline{R}^2$ est éloigné de 1). On arrive quand même à améliorer (un peu) le modèle en transformant les variables.

## Étude similaire 1 : Biscuits

On peut considérer le jeu de données "biscuits" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/biscuits.txt",
header = T)
```

Une entreprise a étudié les effets de trois types de promotions différentes sur les ventes d'une marque spécifique de biscuits :

- Configuration 1 : Les biscuits étaient sur leur plateau d'origine, mais des échantillons gratuits ont été donnés dans le magasin,
- Configuration 2 : Les biscuits étaient sur leur plateau d'origine, mais avec plus d'espace d'étalage.
- Configuration 3 : Les biscuits étaient sur des étagères spéciales à la fin de l'allée, en plus de leur espace d'étalage habituel.

La société a sélectionné 15 magasins pour participer à l'étude. Pour chaque magasin, l'un des trois types de promotion a été réparti de façon aléatoire, avec 5 magasins affectés à chaque promotion. Ainsi, pour chacun d'entre eux, on dispose :

- du nombre de boîtes de biscuits vendues au cours de la période de promotion (variable $X1$),
- du nombre de boîtes de biscuits vendues durant la période antérieure de même durée (variable $Y$),
- de la configuration adoptée (variable $X2$).

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

Quelques commandes préliminaires sont données ci-dessous :

```
X2 = as.factor(X2)
plot(X1[X2 == "2"], Y[X2 == "2"], pch = 15, ylab = "Y", xlab = "X1",
xlim = c(15, 32), ylim = c(20, 50), col = "green")
points(X1[X2 == "3"], Y[X2 == "3"], pch = 16, ylab = "Y", xlab = "X1",
xlim = c(15, 32), ylim = c(20, 50), col = "blue")
points(X1[X2 == "1"], Y[X2 == "1"], pch = 16, ylab = "Y", xlab = "X1",
xlim = c(15, 32), ylim = c(20, 50), col = "red")
reg = lm(Y ~ X1 + X2)
```

On a considéré le modèle sans interaction car les points associés aux différentes modalités de $X2$ ne sont pas vraiment mélangés.

## Étude similaire 2 : Graines

On peut considérer le jeu de données "graines" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/graines.txt",
header = T)
```

Une étude mesure la capacité d'une plante à produire des graines selon la nature brouté ou non du champ dans lequel elle se trouve. Pour 40 plantes, on dispose :

- du poids total de graines produites après expérimentation en mg (variable $Y$),
- de la nature du champ (variable $X1$, avec $X1$ ="brouté" si le champ a été brouté avant l'expérience, $X1$ = "nonbrouté" sinon),
- du diamètre au collet de la racine avant expérimentation en mm (variable $X2$).

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

### Étude similaire 3 : Salaires

On peut considérer le jeu de données "salaires" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/salaires.txt",
header = T)
```

Dans une grande entreprise, pour 50 individus, on dispose :

- du salaire mensuel en euros (variable $Y$),
- de l'ancienneté en année (variable $X1$),
- du sexe (variable $X2$, avec $X2 = 0$ pour femme, $X2 = 1$ pour homme).

L'objectif est d'expliquer $Y$ à partir de $X1$ et $X2$.

### Étude similaire 4 : Samara

On peut considérer le jeu de données "samara" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/samara.txt",
header = T)
```

Pour 36 samaras (petits fruits ailés) tombant de 3 érables aux caractéristiques différentes, on dispose :

- de leur vitesse maximale de chute (variable $Y$),
- de leur "disque de chargement" (une quantité calculée en fonction de leur taille et le poids) (variable $X1$),
- de l'érable d'où il tombe (variable $X2$, avec $X2 = 1$ s'il tombe de l'érable 1, $X2 = 2$ pour l'érable 2 et $X2 = 3$ pour l'érable 3).

L'objectif est d'expliquer $Y$ à partir de $X1$ et $X2$.

## Étude similaire 5 : Profs

On peut considérer le jeu de données "profs" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/profs.txt", header = T)
```

Dans une étude statistique, 23 professeurs sont évalués quant à la qualité de leur enseignement. Pour chacun d'entre eux, on dispose :

- d'un indice de performance globale donné par les étudiants (variable $Y$),
- des résultats de 4 tests écrits donnés à chaque professeur (variables $X1$, $X2$, $X3$ et $X4$),
- du sexe (variable $X5$, avec $X5 = 0$ pour femme, $X5 = 1$ pour homme).

L'objectif est d'expliquer $Y$ à partir de $X1$, $X2$, $X3$, $X4$ et $X5$.

## Étude similaire 6 : Évaluations

On peut considérer le jeu de données "evaluations" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/evaluations.txt",
header = T)
```

Dans une étude statistique, 200 étudiants sont évalués sur plusieurs matières. Pour chacun d'entre eux, on dispose :

- de leur note pour l'examen d'orthographe (variable $Y$),
- de 3 autres notes pour les matières suivantes : mathématiques, sciences physiques et sciences sociales (variables $X1$, $X2$ et $X3$),
- du sexe (variable $X4$).

L'objectif est d'expliquer $Y$ à partir de $X1$, $X2$, $X3$ et $X4$.

Une fois l'analyse terminée, on pourra étudier le modèle de *rlm* avec transformation de $Y$ :

```
reg = lm(log(100 - Y) ~ X1 + X2 + X3 + X4)
```

### Étude similaire 7 : NBA

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 $X3$).

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

# 3 Étude n° 3 : Mensurations

## Contexte

On s'intéresse au lien éventuel entre le poids d'un homme et divers caractéristiques physiques. Pour 22 hommes en bonne santé âgés de 16 à 30 ans, on dispose :

- du poids en kg (variable $Y$),
- de la circonférence maximale de l'avant-bras en cm (variable $X1$),
- de la circonférence maximale du biceps en cm (variable $X2$),
- de la distance autour de la poitrine directement sous les aisselles en cm (variable $X3$),
- de la distance autour du cou, à peu près à mi-hauteur, en cm (variable $X4$),
- de la distance autour des épaules, mesurées autour de la pointe des omoplates, en cm (variable $X5$),
- de la distance autour de la taille au niveau de la ligne de pantalon, en cm (variable $X6$),
- de la hauteur de la tête aux pieds en cm (variable $X7$),
- de la circonférence maximum du mollet en cm (variable $X8$),
- de la circonférence de la cuisse, mesurée à mi-chemin entre le genou et le haut de la jambe, en cm (variable $X9$),
- de la circonférence de la tête en cm (variable $X10$).

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude3.txt",
header = T)
head(w)
```

| | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 77.00 | 28.50 | 33.50 | 100.00 | 38.50 | 114.00 | 85.00 | 178.00 | 37.50 | 53.00 | 58.00 |
| 2 | 85.50 | 29.50 | 36.50 | 107.00 | 39.00 | 119.00 | 90.50 | 187.00 | 40.00 | 52.00 | 59.00 |
| 3 | 63.00 | 25.00 | 31.00 | 94.00 | 36.50 | 102.00 | 80.50 | 175.00 | 33.00 | 49.00 | 57.00 |
| 4 | 80.50 | 28.50 | 34.00 | 104.00 | 39.00 | 114.00 | 91.50 | 183.00 | 38.00 | 50.00 | 60.00 |
| 5 | 79.50 | 28.50 | 36.50 | 107.00 | 39.00 | 114.00 | 92.00 | 174.00 | 40.00 | 53.00 | 59.00 |
| 6 | 94.00 | 30.50 | 38.00 | 112.00 | 39.00 | 121.00 | 101.00 | 180.00 | 39.50 | 57.50 | 59.00 |

On associe les variables $Y$, $X1, \ldots, X10$ aux valeurs associées en faisant :

```
attach(w)
```

## Régression linéaire multiple

### Modélisation

On modélise le problème comme une *rlm* :

$$\begin{aligned} Y &= \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \beta_5 X5 + \beta_6 X6 + \beta_7 X7 \\ &+ \beta_8 X8 + \beta_9 X9 + \beta_{10} X10 + \epsilon, \end{aligned}$$

où $\beta_0, \ldots, \beta_{10}$ sont 11 coefficients inconnus et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma$ inconnu.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

### Estimations

La modélisation de la *rlm* avec les variables explicatives $X1, \ldots, X10$, et les estimations des paramètres par la méthode des *mco* s'obtiennent par les commandes :

```
reg = lm(Y ~ ., w)

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -69.5171 | 29.0374 | -2.39 | 0.0356 | $*$ |
| X1 | 1.7818 | 0.8547 | 2.08 | 0.0612 | . |
| X2 | 0.1551 | 0.4853 | 0.32 | 0.7553 | |
| X3 | 0.1891 | 0.2258 | 0.84 | 0.4201 | |
| X4 | -0.4818 | 0.7207 | -0.67 | 0.5175 | |
| X5 | -0.0293 | 0.2394 | -0.12 | 0.9048 | |
| X6 | 0.6614 | 0.1165 | 5.68 | 0.0001 | $***$ |
| X7 | 0.3178 | 0.1304 | 2.44 | 0.0329 | $*$ |
| X8 | 0.4459 | 0.4125 | 1.08 | 0.3029 | |
| X9 | 0.2972 | 0.3051 | 0.97 | 0.3509 | |
| X10 | -0.9196 | 0.5201 | -1.77 | 0.1047 | |

```
Residual standard error: 2.287 on 11 degrees of freedom
Multiple R-squared: 0.9772, Adjusted R-squared: 0.9565
F-statistic: 47.17 on 10 and 11 DF, p-value: 1.408e-07
```

– $R^2 = 0.9772$ et $\overline{R}^2 = 0.9565$ : cela est très correct,

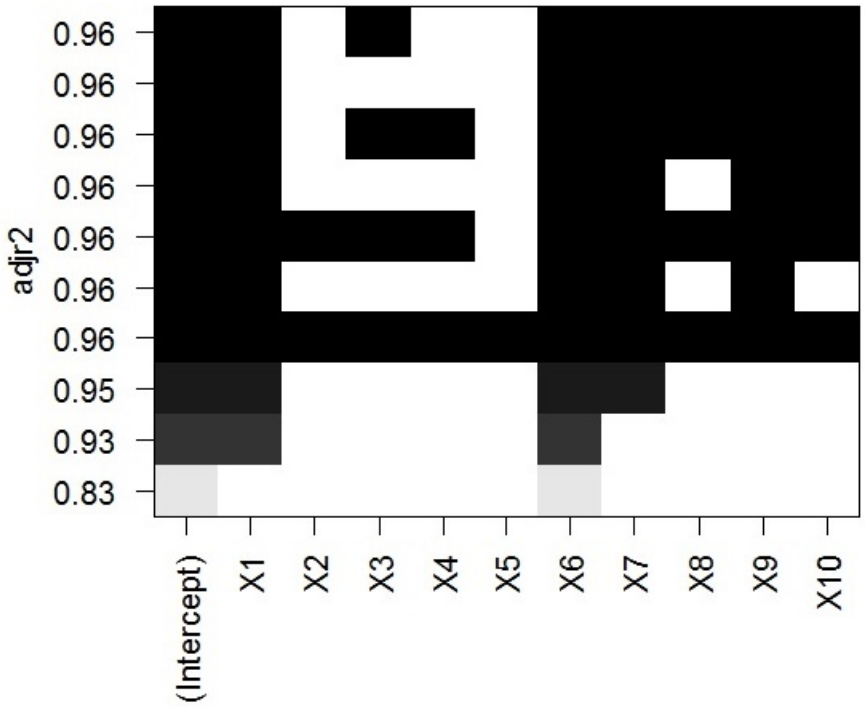
– Test de Fisher : p-valeur $< 0.001$, $***$ : l'utilisation du modèle de *rlm* est pertinente.

Comme beaucoup de variables explicatives ne sont pas significatives, on peut éventuellement faire une sélection de variables avant toute chose.

## Sélection de variables

On propose de faire une sélection de variables via l'approche exhaustive :

```
library(leaps)
v = regsubsets(Y ~ ., w, method = "exhaustive", nvmax = 10)
plot(v, scale = "adjr2")
```




On a précisé `nvmax = 10` car il y a plus de 8 variables explicatives dans le modèle, R considérant une combinaison de 8 variables explicatives par défaut.

On constate que le plus grand $\overline{R}^2$ est obtenu avec la présence de toutes les variables sauf $X2$, $X4$ et $X5$.

On refait une *rlm* avec les variables restantes :

```
reg2 = update(reg, .~. - X2 - X4 - X5)

summary(reg2)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -76.0501 | 24.0581 | -3.16 | 0.0069 | ** |
| X1 | 1.6259 | 0.5096 | 3.19 | 0.0065 | ** |
| X3 | 0.1380 | 0.1310 | 1.05 | 0.3103 | |
| X6 | 0.6365 | 0.0987 | 6.45 | 0.0000 | *** |
| X7 | 0.2687 | 0.0815 | 3.30 | 0.0053 | ** |
| X8 | 0.5468 | 0.3475 | 1.57 | 0.1379 | |
| X9 | 0.3212 | 0.2508 | 1.28 | 0.2211 | |
| X10 | -0.8221 | 0.4116 | -2.00 | 0.0656 | . |

```
Residual standard error: 2.07 on 14 degrees of freedom

Multiple R-squared: 0.9762, Adjusted R-squared: 0.9644

F-statistic: 82.18 on 7 and 14 DF, p-value: 2.744e-10
```

– $R^2 = 0.9762$ et $\overline{R}^2 = 0.9644$ : cela est un peu mieux que dans le modèle précédent,

– Test de Fisher : p-valeur $= 2.744e - 10$ : cela est un peu mieux que dans le modèle précédent.

Ainsi, le deuxième nouveau modèle est meilleur que le premier en termes de $\overline{R}^2$.

Il est aussi intéressant de voir si les modèles sont significativement différents. Pour ce faire, on utilise les commandes :

```
anova(reg, reg2)
```

Cela renvoie :

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| 1 | 11 | 57.52 | | | | |
| 2 | 14 | 59.97 | -3 | -2.45 | 0.16 | 0.9235 |

Comme p-valeur$> 0.05$, les 2 modèles ne sont pas significativement différents.

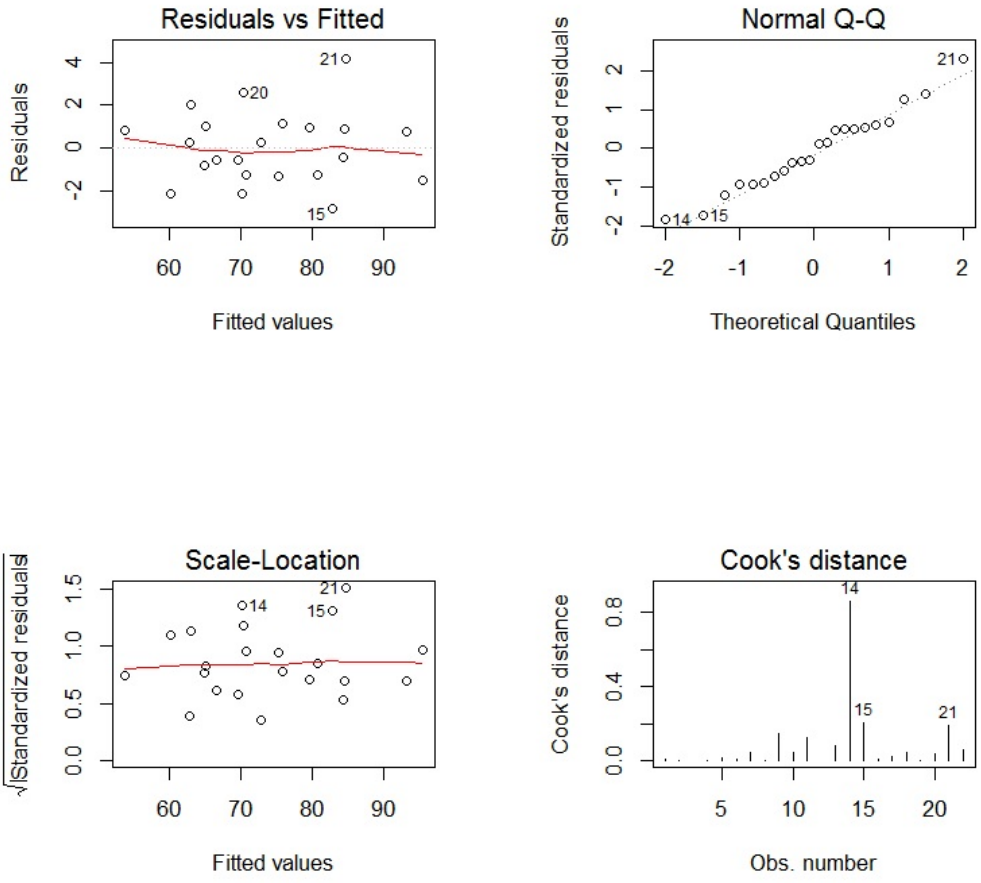
Dans le suite, on privilégie toutefois le deuxième vu qu'il présente moins de variables explicatives et un meilleur $\overline{R}^2$.

# Validation des hypothèses

## Analyse graphique groupée

On trace les graphiques nous permettant de conclure à la validation ou non des hypothèses de base :

```
par(mfrow = c(2, 2))

plot(reg2, 1:4)
```



On obtient des résultats graphiques corrects. En particulier :

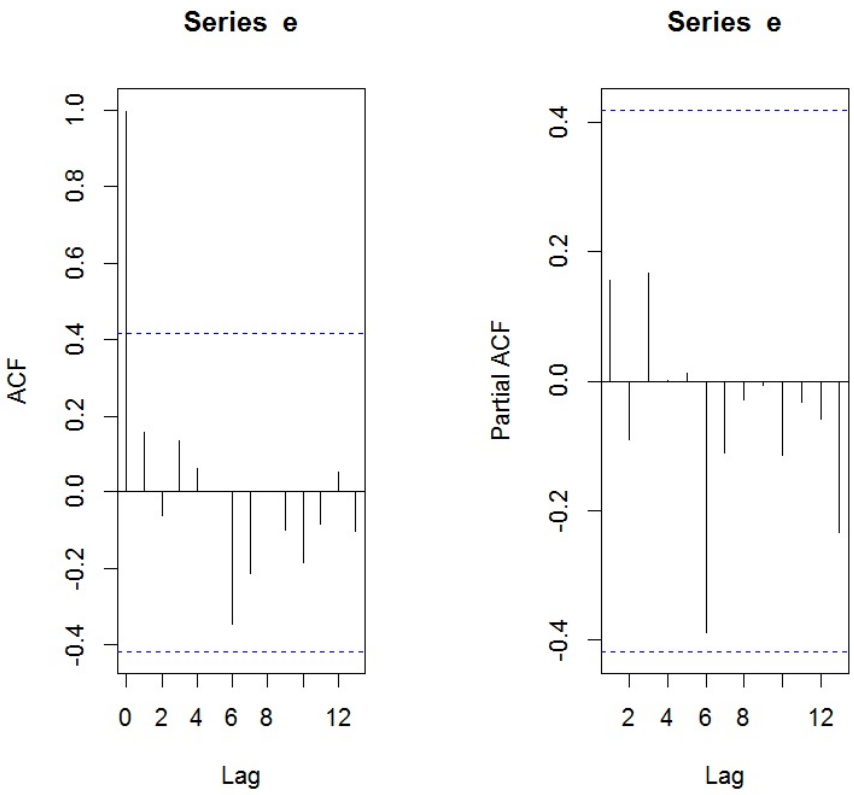
- Celui en haut à gauche présente un nuage de points difficilement ajustable par une "ligne" : on admet que $\epsilon$ et $X1, \ldots, X10$ sont indépendantes,
- Celui en haut à droite montre un alignement des points sur la diagonale : on admet la normalité de $\epsilon_1, \ldots, \epsilon_n$,
- Celui en bas à gauche ne montre pas de structure particulière : on admet l'égalité des variances de $\epsilon_1, \ldots, \epsilon_n$,
- Celui en bas à droite montre les distances de Cook : aucune ne dépasse 1, il n'y a pas de valeurs anormales (toutefois, la distance de Cook associée à l'individu 14 étant élevée, on peut envisager de l'enlever).

### Indépendance de $\epsilon_1, \ldots, \epsilon_n$

Les observations de $(Y, X1, \ldots, X10)$ portent sur des individus tous différents, il est donc normal que $\epsilon_1, \ldots, \epsilon_n$ soient indépendantes.

On examine cela avec les graphiques *acf* et *pacf* :

```
e = residuals(reg2)
par(mfrow = c(1, 2))
acf(e)
pacf(e)
```




On ne constate aucune structure particulière, confirmant l'indépendance des erreurs.

# Compléments

## AIC et BIC

En complément du $\overline{R}^2$, calculons le AIC et le BIC du modèle.

```
AIC(reg2)
```

Cela renvoie 102.4963.

```
BIC(reg2)
```

Cela renvoie 112.3157.

## Étude de la multicolinéarité

On étudie la multicolinéarité éventuelle des variables explicatives à l'aide des *vif* :

```
library(car)

vif(reg2)
```

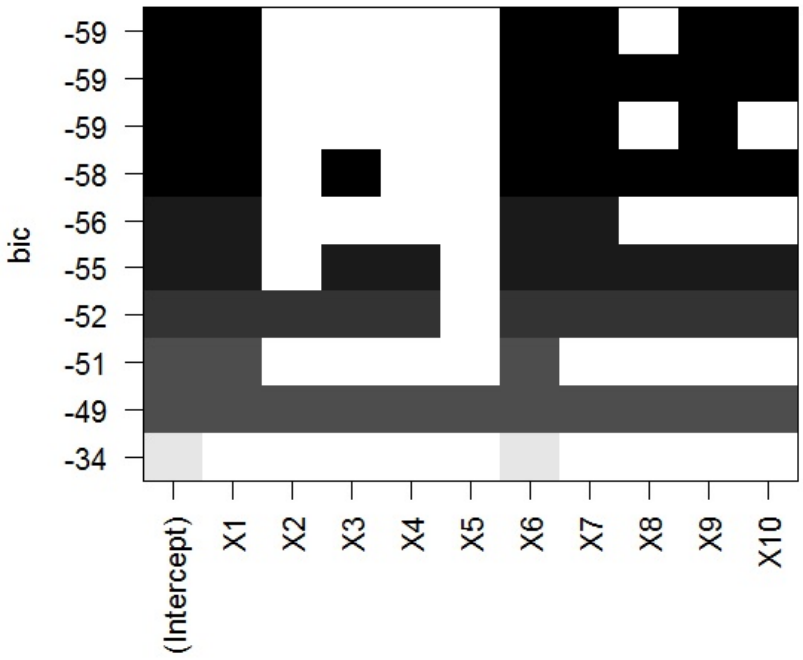
Aucune des valeur ne dépasse 5, il n'y a pas de (multi) colinéarité.

# Alternative possible

Des méthodes de sélection de variables autre que l'approche exhaustive sont possibles, ainsi que d'autre critère que le $\overline{R}^2$.

Par exemple :

```
library(leaps)
v = regsubsets(Y ~ ., w, method = "backward", nvmax = 10)
plot(v, scale = "bic")
```




Précisons que l'axe des ordonnées n'affiche pas exactement le BIC de chaque modèle, mais une quantité qui en dépendant. On constate que la plus petite de ces quantités est obtenue avec la présence de toutes les variables sauf $X2$, $X3$, $X4$, $X5$ et $X8$.

On refait une *rlm* avec les variables restantes.

```
reg3 = update(reg, . ~ . - X2 - X3 - X4 - X5 - X8)
```

Pour avoir directement le modèle de *rlm* avec les variables sélectionnées (sans les écrire) :

```
reg3 = step(reg, direction = "backward", k = log(length(Y)))
```

On obtient le BIC :

```
BIC(reg3)
```

Cela renvoie 110.7438.

Comme ce BIC est plus petit que le modèle précédent, i.e., $BIC = 112.3157$, on peut le considérer comme meilleur.

Il est aussi intéressant de voir si ils sont significativement différents :

```
anova(reg3, reg)
```

Cela renvoie :

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| 1 | 16 | 73.96 | | | | |
| 2 | 11 | 57.52 | 5 | 16.43 | 0.63 | 0.6822 |

Comme p-valeur$> 0.05$, les 2 modèles ne sont pas significativement différents.

On peut aussi regarder les caractéristiques de la nouvelle *rlm* et étudier la validation des hypothèses.

```
summary(reg3)
```

Cela renvoie :

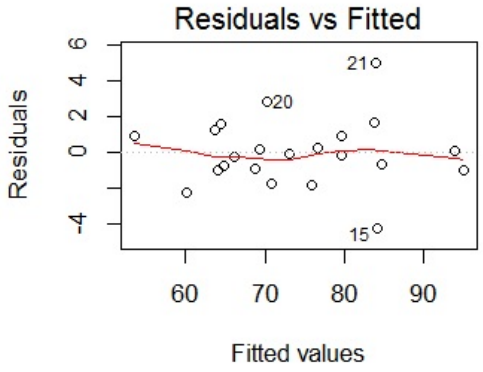
| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -80.4533 | 24.7202 | -3.25 | 0.0050 | ** |
| X1 | 2.1232 | 0.4454 | 4.77 | 0.0002 | *** |
| X6 | 0.6656 | 0.1004 | 6.63 | 0.0000 | *** |
| X7 | 0.2770 | 0.0818 | 3.39 | 0.0038 | ** |
| X9 | 0.5232 | 0.2272 | 2.30 | 0.0351 | * |
| X10 | -0.6371 | 0.3951 | -1.61 | 0.1264 | |

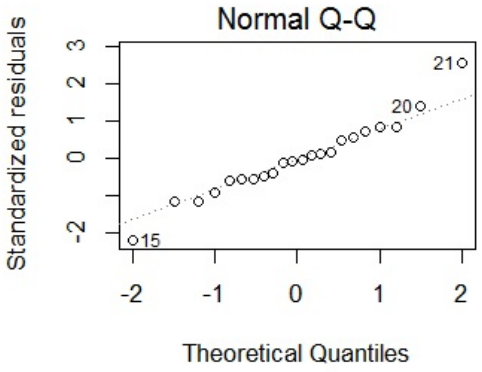
```
Residual standard error: 2.15 on 16 degrees of freedom
Multiple R-squared: 0.9707, Adjusted R-squared: 0.9615
F-statistic: 106 on 5 and 16 DF, p-value: 1.10e-11
```

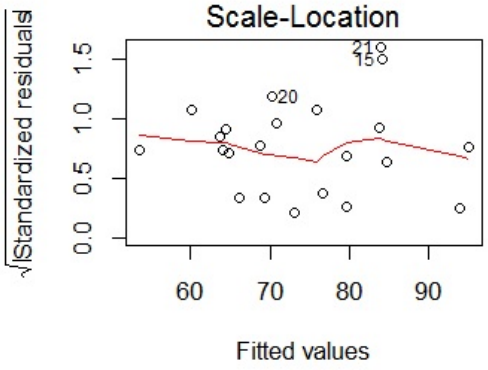
Ainsi, le $R^2$ est logiquement moins bon que dans le modèle précédent, mais comme cela n'est plus notre critère de référence, on n'y fait pas attention.

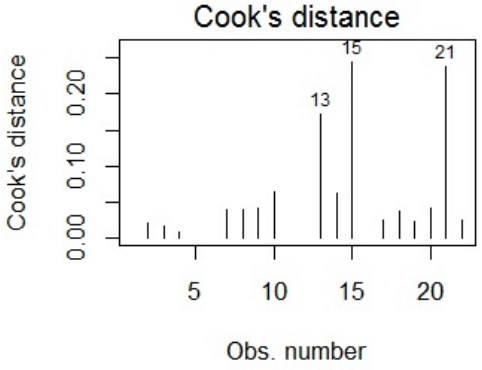
Validation des hypothèses :

```
par(mfrow = c(2, 2))

plot(reg3, 1:4)
```




Il n'y a pas de problème à signaler.

# Conclusion et études similaires

## Conclusion

L'étude statistique mise en œuvre montre plusieurs méthodes pour sélectionner des variables explicatives de manière pertinente (sur le plan mathématique).

## Étude similaire 1 : Épines

On peut considérer le jeu de données "épines" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/épines.txt",
header = T)
```

On s'intéresse au lien éventuel entre la composition minérale des épines d'un mélèze (arbres des régions tempérées de l'hémisphère nord) et sa taille. Pour 26 mélèzes, on dispose :

- de la taille de l'arbre (variable $Y$),
- du pourcentage d'azote dans les épines (variable $X1$),
- du pourcentage de phosphore dans les épines (variable $X2$),
- du pourcentage de potassium dans les épines (variable $X3$),
- du pourcentage de cendre résiduelle dans les épines (variable $X4$).

L'objectif est de proposer le meilleur modèle de *rlm* suivant le critère de votre choix.

## Étude similaire 2 : Selection test

On peut considérer le jeu de données "selection-test" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/selection-test.txt",
header = T)
```

Ce jeu de données est artificiel. Utiliser le modèle de *rlm* pour expliquer $Y$ à partir des 25 variables : $X1, \ldots, X25$. Ensuite, proposer le meilleur modèle suivant le critère de votre choix.

## Étude similaire 3 : Bébé

On peut considérer le jeu de données "bébé" :

```
w = read.table("http://math.agrocampus-ouest.fr/infoglueDeliverLive/
digitalAssets/19623_bebe.txt, header = T, sep = ";")
```

On cherche à expliquer le poids d'un bébé en fonction de plusieurs variables liées au père et à la mère. Considérer le modèle de *rlm* donné par les commandes suivantes :

```
w = na.omit(w)
attach(w)
reg = lm(PoidsBB ~ Nbsem + TailleBB + PoidsPlacenta + AgedelaMère + TailMere
+ PoidsMere + Agedupère + TailPere + PoidsPere + NbGrossess + NbEnfants)
```

Ensuite, proposer le meilleur modèle, utilisant une partie des variables du modèle initiale, suivant le critère de votre choix.

Pour aller plus loin : Étudier la pertinence du modèle de *rlm* décrit par les commandes suivantes :

```
reg2 = lm(PoidsBB ~ Nbsem + TailleBB + PoidsPlacenta + PoidsMere +
PoidsPere)
```

Comprendre l'enjeu des commandes suivantes :

```
par(mfrow = c(2, 2))
plot(reg2, 1:4)
ww = w[-56, ]
attach(ww)
reg3 = lm(PoidsBB   Nbsem + TailleBB + PoidsPlacenta + PoidsMere +
PoidsPere, data = ww)
summary(reg3)
par(mfrow = c(2, 2))
plot(reg3, 1:4)
```

# 4 Étude n° 4 : Pression artérielle diastolique

## Contexte

La pression artérielle diastolique est la pression minimale du sang au moment du relâchement du cœur. Celle-ci a été mesurée pour 54 individus de différents âges. Ainsi, pour chacun d'entre eux, on dispose :

- de leur pression diastolique en mmHg (variable $Y$),
- de leur âge en années (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$.

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude4.txt",
header = T)
head(w)
```

| | X1 | Y |
|---|---|---|
| 1 | 27 | 73 |
| 2 | 21 | 66 |
| 3 | 22 | 63 |
| 4 | 24 | 75 |
| 5 | 25 | 71 |
| 6 | 23 | 70 |

On associe les variables $Y$ et $X1$ aux valeurs associées en faisant :

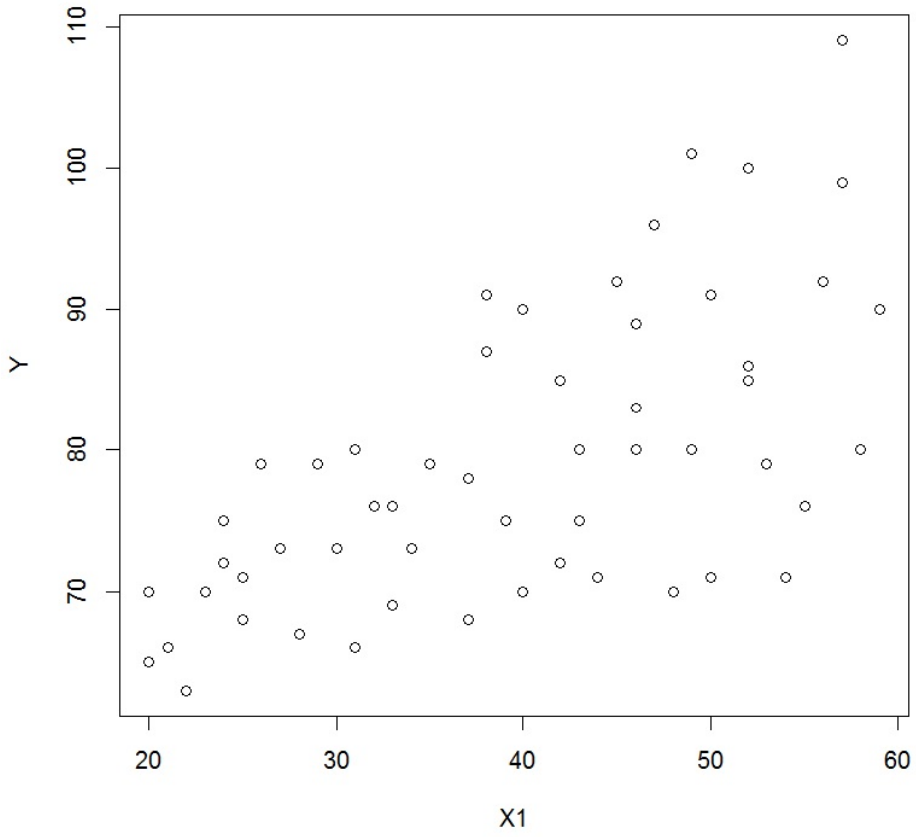
```
attach(w)
```

# Régression linéaire simple

## Analyse du nuage de points

On trace le nuage de points $\{(x_{1,i}, y_i),\ i \in \{1, \ldots, n\}\}$ :

```
plot(X1, Y)
```




On constate qu'une liaison linéaire entre $Y$ et $X1$ est difficile, mais envisageable dans une première étude.

## Modélisation

Une première approche est de considérer le modèle de *rls* :

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

où $\beta_0$ et $\beta_1$ sont 2 coefficients inconnus, et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma$ inconnu.

La présence de $\beta_0$ est justifiée car même un très jeune individu peut avoir une pression artérielle diastolique élevée.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

## Estimations

La modélisation de la *rls* et les estimations des paramètres par la méthode des *mco* s'obtiennent par les commandes :

```
reg = lm(Y ~ X1)

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 56.1569 | 3.9937 | 14.06 | 0.0000 | *** |
| X1 | 0.5800 | 0.0970 | 5.98 | 0.0000 | *** |

```
Residual standard error: 8.146 on 52 degrees of freedom

Multiple R-squared: 0.4077, Adjusted R-squared: 0.3963

F-statistic: 35.79 on 1 and 52 DF, p-value: 2.05e-07
```

– Estimations ponctuelles de $\beta_0$ et $\beta_1$ :

| $\widehat{\beta}_0$ | $\widehat{\beta}_1$ |
|---|---|
| 56.1569 | 0.5800 |

– Estimations ponctuelles des écart-types des estimateurs de $\beta_0$ et $\beta_1$ :

| $\widehat{\sigma}(\widehat{\beta}_0)$ | $\widehat{\sigma}(\widehat{\beta}_1)$ |
|---|---|
| 3.9937 | 0.0970 |

– $t_{obs}$ :

| $H_1$ | $\beta_0 \neq 0$ | $\beta_1 \neq 0$ |
|---|---|---|
| $t_{obs}$ | 14.06 | 5.98 |

– Test de Student pour $\beta_1$ : influence de $X1$ sur $Y$ : p-valeur $< 0.001$, $***$ : hautement significative,

– $R^2 = 0.4077$ et $\overline{R}^2 = 0.3963$ : cela n'est pas très convaincant,

– Test de Fisher : p-valeur $= 2.05e-07 < 0.001$, $***$ : l'utilisation du modèle de *rlm* est pertinente.

La valeur prédite de $Y$ quand $X1 = 90.2$ (par exemple) est donnée par les commandes :

```
predict(reg, data.frame(X1 = 90.2))
```

Cela renvoie 108.4757.

Ainsi, la pression artérielle diastolique moyenne d'un individu de 90.2 ans est de 108.4757 mmHg.

On peut aussi s'intéresser :

– aux intervalles de confiance pour $\beta_0$ et $\beta_1$ au niveau 95% (par exemple). Les commandes sont :

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

Cela renvoie :

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 48.1430367 | 64.1708221 |
| X1 | 0.3854841 | 0.7745775 |

Le tableau donne les bornes inférieures et supérieures des intervalles de confiance de $\beta_0$ et $\beta_1$ :

$$i_{\beta_0} = [48.1430367,\ 64.1708221], \qquad i_{\beta_1} = [0.3854841,\ 0.7745775].$$

– à l'intervalle de confiance pour la valeur moyenne de $Y$ quand $X1 = 90.2$ (par exemple).

Les commandes sont :

```
predict(reg, data.frame(X1 = 90.2), interval = "confidence")
```

Cela renvoie :

| fit | lwr | upr |
|---|---|---|
| 108.4757 | 98.37854 | 118.5729 |

La première valeur est celle de la valeur prédite de $Y$ quand $X1 = 90.2$ (déjà vue), les deux autres correspondent aux bornes inférieures et supérieures des intervalles de confiance.
Ainsi, pour $X1 = 90.2$, on a

$$i_{y_x} = [98.37854,\ 118.5729].$$

On étudie les critères AIC et BIC :

```
AIC(reg)
```

Cela renvoie 383.7369.

```
BIC(reg)
```

Cela renvoie 389.7039.

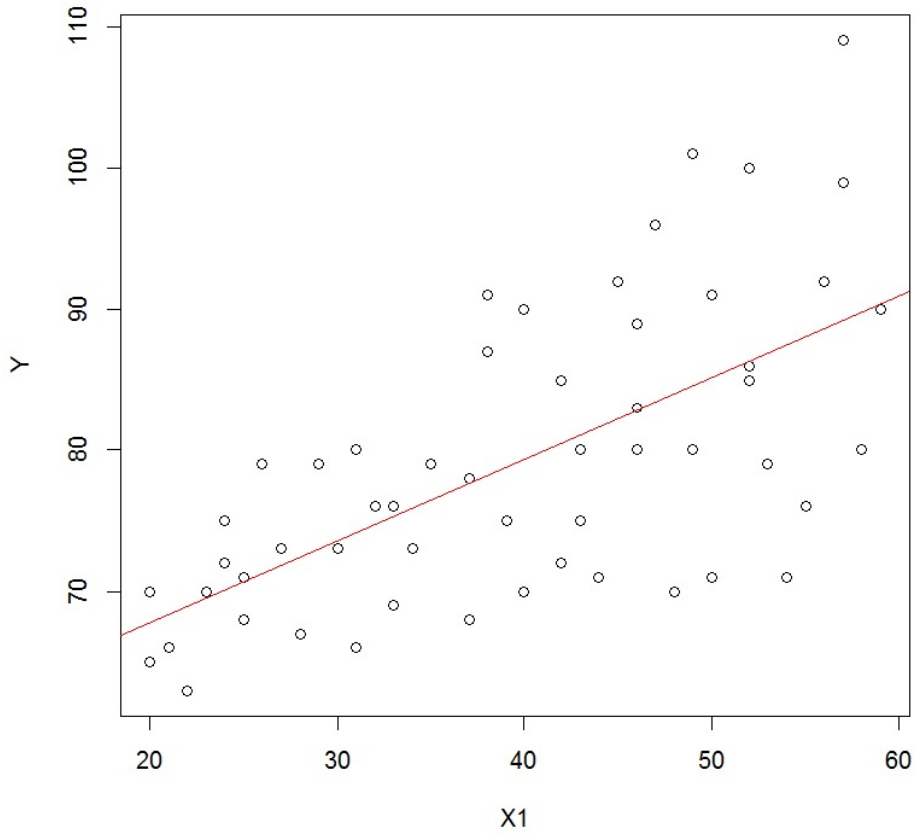
### Droite de régression

En utilisant les estimations ponctuelles de $\beta_0$ et $\beta_1$, l'équation de la droite de régression est :

$$y = 56.1569 + 0.58x.$$

On la visualise avec les commandes :

```
abline(reg, col = "red")
```

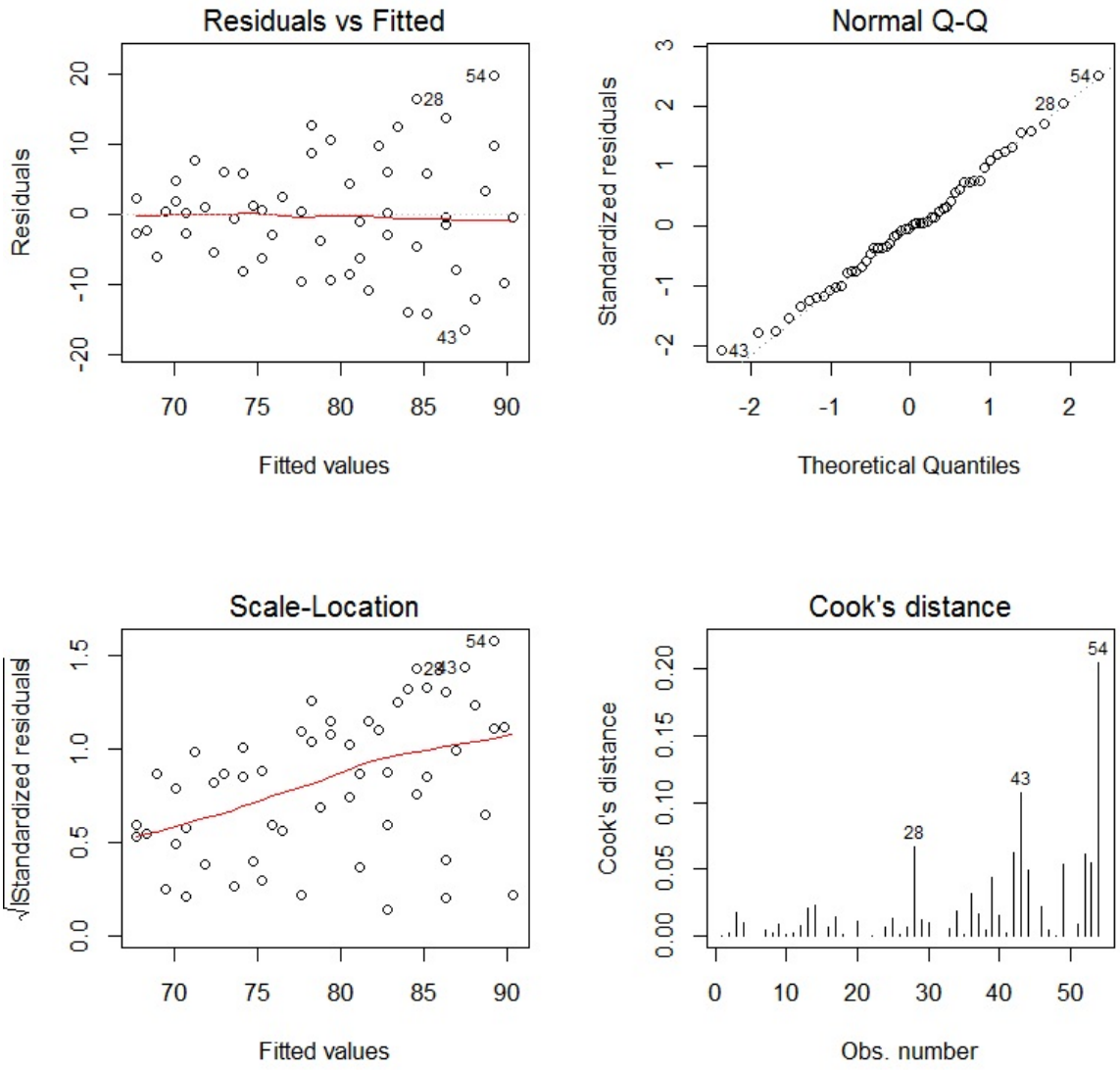


# Validation des hypothèses

## Analyse graphique groupée

On trace les graphiques nous permettant de conclure à la validation ou non des hypothèses de base :

```
par(mfrow = c(2, 2))

plot(reg, 1:4)
```



On obtient des résultats graphiques corrects. En particulier :

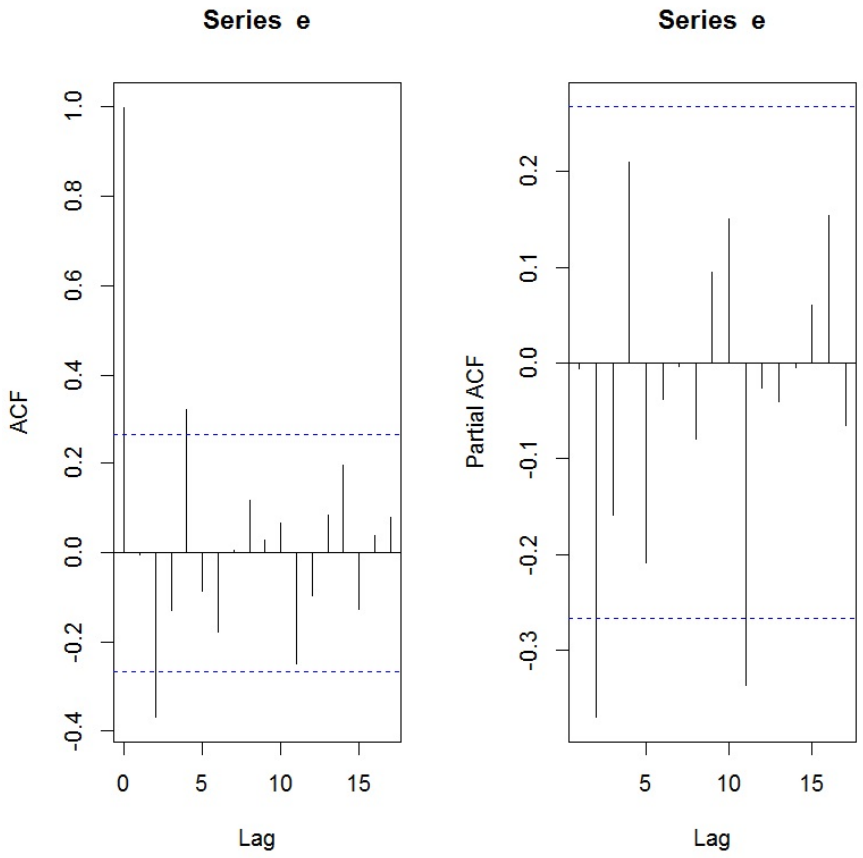
- Celui en haut à gauche présente un nuage de points difficilement ajustable par une "ligne" et la moyenne des valeurs de la ligne rouge est quasi nulle : on admet que $\epsilon$ et $X1$ sont indépendantes,
- Celui en haut à droite montre un alignement des points sur la diagonale : on admet la normalité de $\epsilon_1, \ldots, \epsilon_n$,
- Celui en bas à gauche montre une structure en forme de mégaphone : l'égalité des variances de $\epsilon_1, \ldots, \epsilon_n$ est a étudier plus finement,
- Celui en bas à droite montre les distances de Cook : aucune ne dépasse 1, il n'y a pas de valeurs anormales.

### Indépendance de $\epsilon_1, \ldots, \epsilon_n$

Les observations de $(Y, X1)$ portent sur des individus tous différents, il est donc normal $\epsilon_1, \ldots, \epsilon_n$ soient indépendantes.

On examine cela avec les graphiques *acf* et *pacf* :

```
e = residuals(reg)
par(mfrow = c(1, 2))
acf(e)
pacf(e)
```




On ne constate aucune structure particulière et peu de bâtons dépassent les bornes limites, confirmant l'indépendance des erreurs.

En complément, on peut aussi s'intéresser au test de Durbin-Watson :

```
library(lmtest)

dwtest(reg)
```

Cela renvoie : p-valeur = $0.3087 > 0.05$. Il n'y a donc pas de dépendance AR(1) dans les erreurs.

### Égalité des variances de $\epsilon_1, \ldots, \epsilon_n$

On a constaté une structure en forme de mégaphone dans le graphique Scale-Location.

On étudie plus finement celle-ci avec le test de Breusch-Pagan :

```
library(lmtest)

bptest(reg)
```

Cela renvoie : p-valeur = $0.0003981 < 0.05$. L'égalité des variances de $\epsilon_1, \ldots, \epsilon_n$ est donc rejetée ; il y a de l'hétéroscédasticité.

Nous allons traiter ce problème avec deux méthodes :

- en transformant judicieusement $Y$,
- avec la méthode des *mcqg*.

# Transformation de variables

## Transformation de Box-Cox

On peut aussi considérer une transformation de $Y$ dépendante d'un paramètre inconnu, comme la transformation de Box-Cox :

$$bc_\lambda(u) = \begin{cases} \dfrac{u^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0, \\ \ln(u) & \text{sinon.} \end{cases}$$

On effectue alors une *rlm* sur $bc_\lambda(Y)$ et on estime le réel $\lambda$ tel que la loi des résidus du modèle soit aussi proche que possible d'une loi normale (cette estimation se fait avec l'*emv* $\widehat{\lambda}$ de $\lambda$) :

```
libray(car)
reg = lm(Y ~ X1)
reg2 = powerTransform(reg)
reg2
```

Cela renvoie :

```
Estimated transformation parameters
Y1
-2.123196
```

Ainsi, l'estimation ponctuelle $\widehat{\lambda}$ de $\lambda$ par la méthode du *mv* est $-2.123196$.

On va alors faire une *rlm* avec les variables $bc_{-2.123196}(Y)$ et $X1$ :

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

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 0.4709 | 0.0000 | 107405.69 | 0.0000 | $***$ |
| X1 | 0.0000 | 0.0000 | 6.27 | 0.0000 | $***$ |

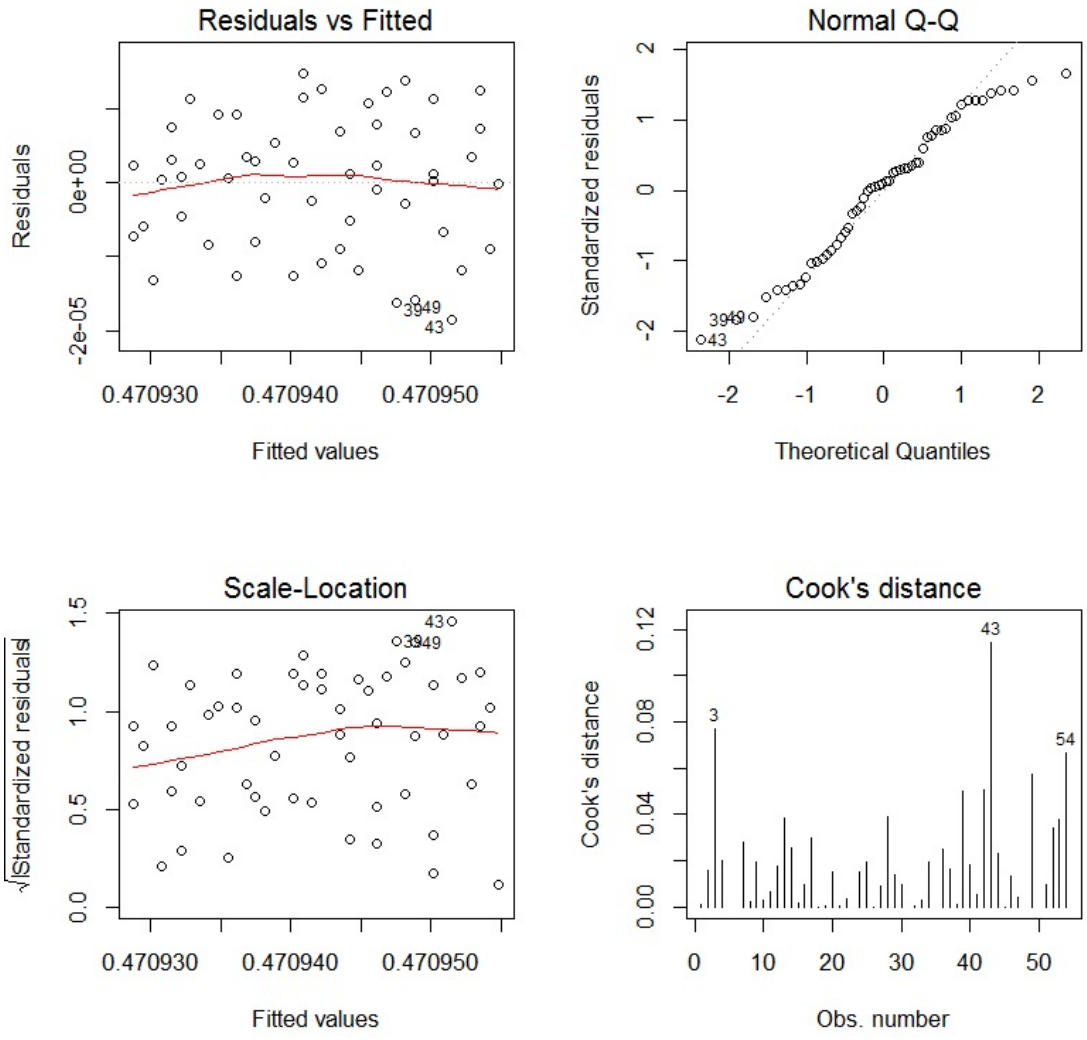
```
Residual standard error: 8.943e-06 on 52 degrees of freedom
Multiple R-squared: 0.4306, Adjusted R-squared: 0.4197
F-statistic: 39.33 on 1 and 52 DF, p-value: 7.157e-08
```

On constate que le $\overline{R}^2$ est sensiblement meilleur que celui du modèle de *rlm* initial, i.e., $\overline{R}^2 = 0.4197$ contre $\overline{R}^2 = 0.3963$. Il reste cependant faible. Avec ce critère, le nouveau modèle est sensiblement meilleur que le modèle initial.

### Analyse graphique groupée

On trace les graphiques nous permettant de conclure à la validation ou non des hypothèses de base :

```
par(mfrow = c(2, 2))
plot(reg3, 1:4)
```




Tout semble satisfaisant. L'hétéroscédasticité est bien traitée ; cela se voit sur le graphique Scale-Location. Peut-être que la normalité des erreurs est à étudier :

```
e3 = residuals(reg3)

shapiro.test(e3)
```

Cela renvoie : p-valeur = 0.1109. Comme p-valeur $> 0.05$, on admet la normalité de $\epsilon_1, \ldots, \epsilon_n$.

## Méthode des *mcqg*

### Estimations

On peut aussi corriger l'hétéroscédasticité en utilisant la méthode des *mcqg* :

$$\widehat{\beta} = (X^t \widetilde{\Omega}^{-1} X)^{-1} X^t \widetilde{\Omega}^{-1} Y,$$

avec une matrice $\widetilde{\Omega}$ bien choisie (nous allons utiliser la "méthode II" du cours).

Partant du modèle de *rlm* initial et ses résidus, on propose les commandes :

```
rege = lm(log(e^2) ~ X1)

regmcqg = lm(Y ~ X1, weights = exp(-fitted(rege)))

summary(regmcqg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 55.9586 | 2.8812 | 19.42 | 0.0000 | *** |
| X1 | 0.5858 | 0.0846 | 6.93 | 0.0000 | *** |

```
Residual standard error: 1.865 on 52 degrees of freedom

Multiple R-squared: 0.4799, Adjusted R-squared: 0.4699

F-statistic: 47.97 on 1 and 52 DF, p-value: 6.488e-09
```

On constate un meilleur $\overline{R}^2$ que celui du modèle initial et que le modèle utilisant la transformation de Box-Cox, même s'il reste faible.

On étudie les critères AIC et BIC :

```
AIC(regmcqg)
```

Cela renvoie 369.7255.

```
BIC(regmcqg)
```

Cela renvoie 375.6925.

Ces résultats sont un peu meilleurs que ceux du modèle initial.

## Analyse du modèle transformé $\widetilde{\Omega}^{-1/2}Y$

On peut vérifier les hypothèses standards du modèle transformé avec, pour écriture matricielle :

$$\widetilde{\Omega}^{-1/2}Y = \widetilde{\Omega}^{-1/2}X + \widetilde{\Omega}^{-1/2}\epsilon.$$

```
n = length(Y)
omega = exp(fitted(rege)) * diag(n)
```

On calcule $\Omega^{-1/2}$ :

```
omegasqrtinv = exp(-fitted(rege) / 2) * diag(n)
```

On calcule les transformations $\Omega^{-1/2}Y$ et $\Omega^{-1/2}X$ :

```
Yo = omegasqrtinv %*% Y
X = cbind(1, X1)
Xo = omegasqrtinv %*% X
```

On fait une *rlm* sur $Yo$ et les colonnes de $Xo$ :

```
regmcqg2 = lm(Yo ~ Xo[ ,1] + Xo[ ,2] - 1)
summary(regmcqg2)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 55.9586 | 2.8812 | 19.42 | 0.0000 | $***$ |
| X1 | 0.5858 | 0.0846 | 6.93 | 0.0000 | $***$ |

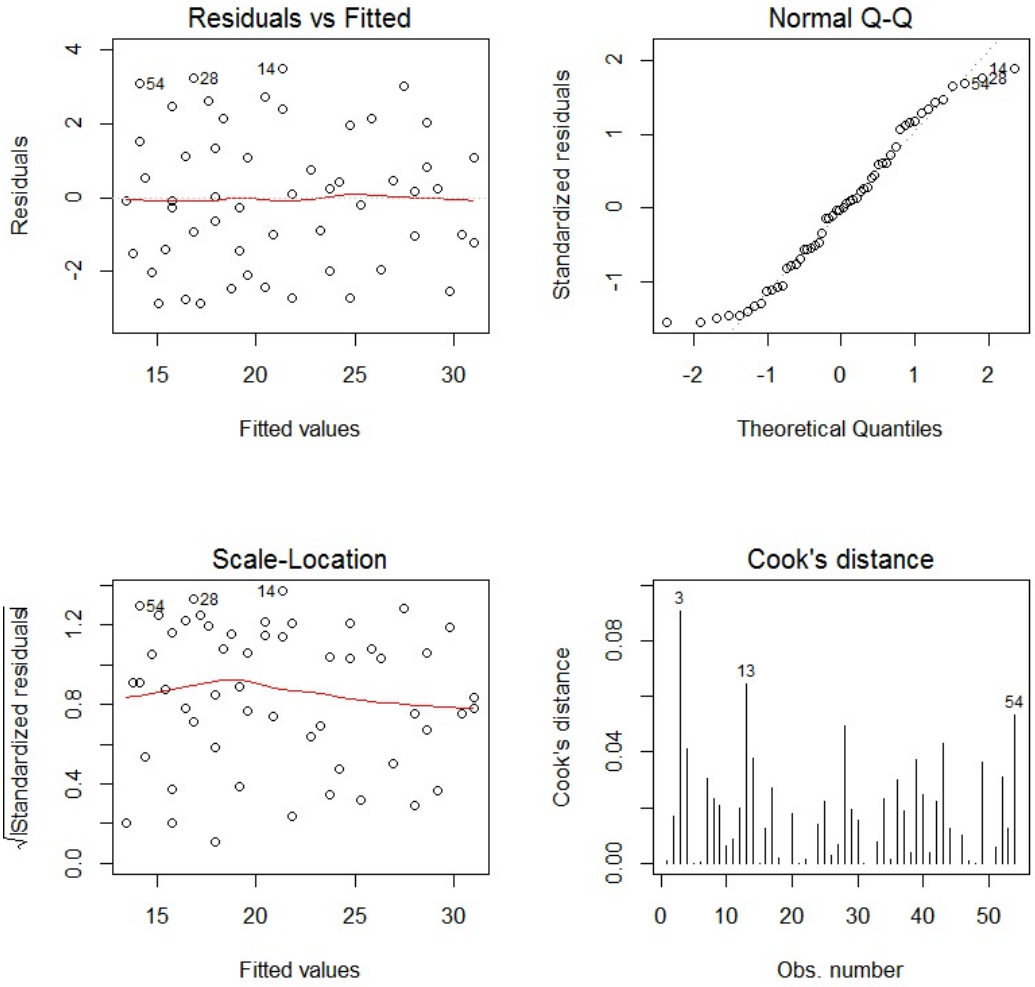
```
Residual standard error: 1.865 on 52 degrees of freedom
Multiple R-squared: 0.993, Adjusted R-squared: 0.9927
F-statistic: 47.97 on 1 and 52 DF, p-value: 6.488e-09
```

Notons que $\overline{R}^2$ est très bon. Il est bien meilleur que celui du modèle initial mais ceux-ci ne sont pas comparables ; ce ne sont pas les mêmes variables qui sont considérées dans le modèle.

Le vrai $\overline{R}^2$ est 0.4699.

Les hypothèses sont validées via une analyse graphique :

```
par(mfrow = c(2,2))

plot(regmcqg2, 1:4)
```




Tout semble satisfaisant. L'hétéroscédasticité est bien traitée ; cela se voit sur le graphique Scale-Location. Peut-être que la normalité des erreurs est à étudier :

```
e2 = residuals(regmcqg2)

shapiro.test(e2)
```

Cela renvoie : p-valeur $= 0.05842 > 0.05$. On admet de justesse la normalité des erreurs.

Ainsi, pour le nouveau modèle :

– Estimations ponctuelles de $\beta_0$ et $\beta_1$ :

| $\widehat{\beta}_0$ | $\widehat{\beta}_1$ |
|---|---|
| 55.9586 | 0.5858 |

– Estimations ponctuelles des écart-types des estimateurs de $\beta_0$ et $\beta_1$ :

| $\widehat{\sigma}(\widehat{\beta}_0)$ | $\widehat{\sigma}(\widehat{\beta}_1)$ |
|---|---|
| 2.8812 | 0.0846 |

– $t_{obs}$ :

| $H_1$ | $\beta_0 \neq 0$ | $\beta_1 \neq 0$ |
|---|---|---|
| $t_{obs}$ | 19.42 | 6.93 |

La valeur prédite de $Y$ quand $X1 = 90.2$ est donnée par :

```
predict(regmcqg, data.frame(X1 = 90.2))
```

ou

```
cbind(1, 90.2) %*% regmcqg2$coeff
```

Cela renvoie 108.7962.

Avec ce nouveau modèle, la pression artérielle diastolique moyenne d'un individu de 90.2 ans est de 108.7962 mmHg.

## Conclusion

L'étude statistique mise en œuvre montre que le modèle de *rlm* usuel n'est pas très bien adapté au problème. Cela est en partie dû à de l'hétéroscédasticité des erreurs.

On arrive toutefois à une modélisation acceptable en faisant une correction de l'hétéroscédasticité avec :

- la transformation de Box-Cox pour $Y$,
- la méthode des *mcqg*.

# 5 Étude n° 5 : Super et Ultra

## Contexte

Un produit charcuterie est vendu dans deux supermarchés concurrents : Super et Ultra. Sa particularité est que son prix (en euros) peut fluctuer significativement d'une semaine à l'autre. Pour 20 semaines consécutives, on dispose :

- du nombre de ventes de ce produit dans Super (variable $Y$),
- de son prix dans Super (variable $X1$),
- de son prix dans Ultra (variable $X2$).

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude5.txt",
header = T)
head(w)
```

| | Y | X1 | X2 |
|---|---|---|---|
| 1 | 1201 | 1.99 | 1.96 |
| 2 | 1286 | 1.79 | 1.90 |
| 3 | 1432 | 1.89 | 1.97 |
| 4 | 1058 | 2.09 | 1.95 |
| 5 | 1234 | 1.99 | 1.95 |
| 6 | 999 | 1.99 | 1.86 |

On associe les variables $Y$, $X1$ et $X2$ aux valeurs associées en faisant :

```
attach(w)
```

# Régression linéaire multiple

## Modélisation

Une première approche est de considérer le modèle de *rlm* :

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

où $\beta_0$, $\beta_1$ et $\beta_2$ sont 3 coefficients inconnus et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma$ inconnu.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

## Estimations

La modélisation de la *rlm* et les estimations des paramètres par la méthode des *mco* s'obtiennent par les commandes :

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

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 220.3120 | 953.6647 | 0.23 | 0.8201 | |
| X1 | -976.0475 | 163.1165 | -5.98 | 0.0000 | *** |
| X2 | 1456.1069 | 443.3859 | 3.28 | 0.0044 | ** |

```
Residual standard error: 115.5 on 17 degrees of freedom

Multiple R-squared: 0.758, Adjusted R-squared: 0.7296

F-statistic: 26.63 on 2 and 17 DF, p-value: 5.778e-06
```

– $R^2 = 0.758$ et $\overline{R}^2 = 0.7296$ : cela est correct,

– Test de Fisher : p-valeur $< 0.001$, $**$ : l'utilisation du modèle de *rlm* est pertinente.

Si on veut la valeur prédite de $Y$ quand $X1 = 1.98$ et $X2 = 1.92$ (par exemple), on exécute :

```
predict(reg, data.frame(X1 = 1.98, X2 = 1.92))
```

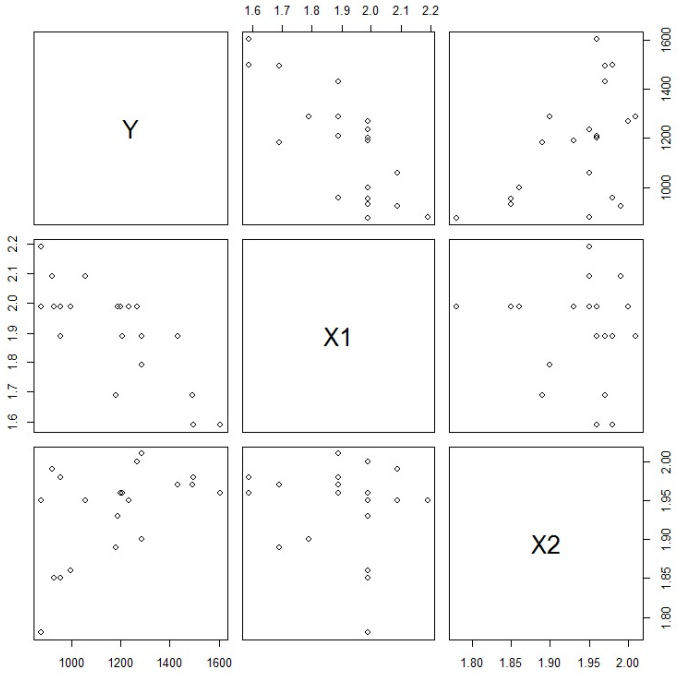
Cela renvoie 1083.463.

# Améliorations et validation des hypothèses

## Analyse des nuages de points

On trace les nuages de points des variables par pairs :

```
plot(w)
```

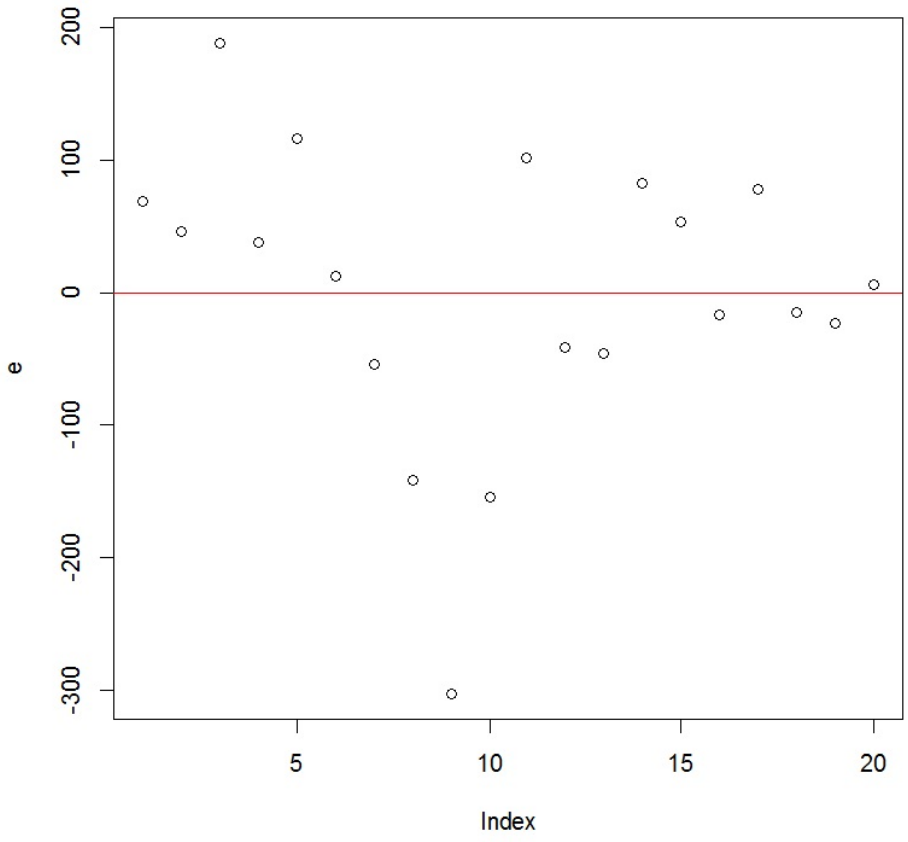



On remarque que la liaison linéaire entre $Y$ et $X1$ est envisageable. Cela est un peu moins clair entre $Y$ et $X2$.

### Analyse graphique des résidus

On examine les résidus en faisant :

```
e = residuals(reg)

plot(e)

abline(h = 0, col = "red")
```

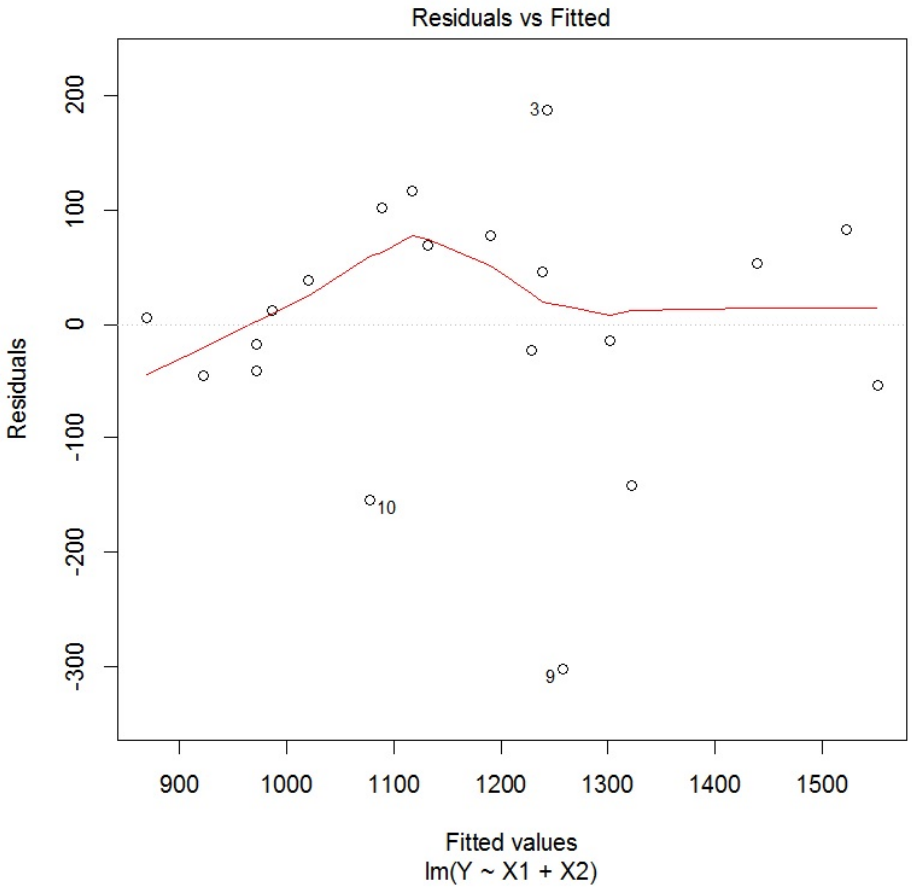



On constate une structure, les hypothèses standards ne semblent pas être vérifiées.

### Indépendance de $\epsilon$ et $X1, X2$

On trace le nuage de points {(résidus$_i$, prédictions en $(x_{1,i}, x_{2,i})$)} :

```
plot(reg, 1)
```



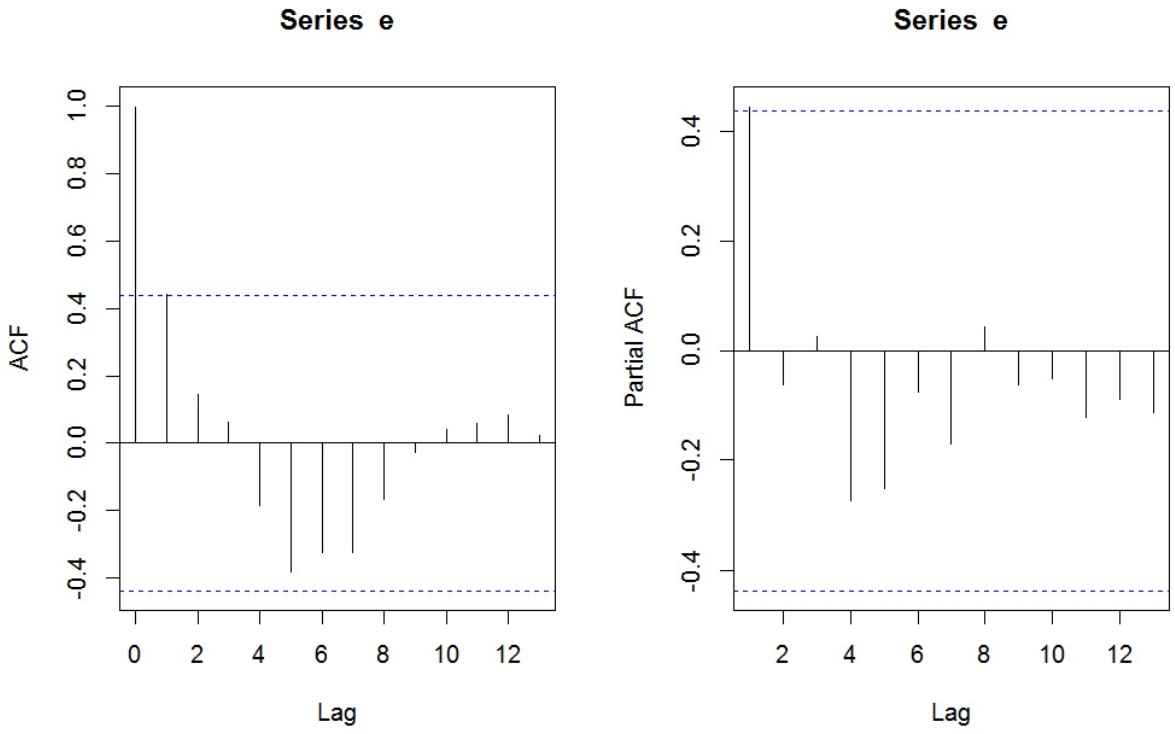

On constate que le nuage de points obtenu est ajustable par une "ligne". Un problème se profile.

### Indépendance de $\epsilon_1, \ldots, \epsilon_n$

Les observations de $(Y, X1, X2)$ dépendent du temps, il est donc probable que $\epsilon_1, \ldots, \epsilon_n$ soient dépendants.

On examine cela avec les graphiques *acf* et *pacf* :

```
par(mfrow = c(1, 2))
acf(e)
pacf(e)
```




On constate une structure dans la taille et le signe des bâtons, confirmant la dépendance des erreurs.

On peut arrêter l'analyse de ce modèle ici ; il faut traiter cette dépendance. Cela va nous mener à d'autres modèles plus adaptés.

# Traitement de la dépendance

## Test de Durbin-Watson

On confirme la dépendance en faisant le test de Durbin-Watson :

```
lmtest

dwtest(reg)
```

Cela renvoie : p-valeur = $0.01649 < 0.05$, la dépendance des erreurs est confirmée. La structure AR(1) de $\epsilon_1, \ldots, \epsilon_n$ est significative.

Autrement dit, on admet l'existence d'un $\rho \in ]-1, 1[$ et $n$ *var iid* $u_1, \ldots, u_n$ suivant chacune la loi normale $\mathcal{N}(0, v^2)$ tels que

$$\epsilon_i = \rho\epsilon_{i-1} + u_i.$$

Cette structure AR(1) va nous guider pour trouver un modèle bien adapté au problème.

## Méthode des *mcg*

Dans un premier temps, on estime le $\rho$ en le considérant comme un coefficient de régression :

```
n = length(Y)

e = residuals(reg)

rho = lm(e[-1] ~ e[-n] - 1)$coeff[1]

rho
```

Cela renvoie :

```
e[-n]

0.4462422
```

Cette estimation ponctuelle est donc éloignée de 0.

On termine le calcule en considérant l'*emcg* :

$$\widehat{\beta} = (X^t \widetilde{\Omega}^{-1} X)^{-1} X^t \widetilde{\Omega}^{-1} Y.$$

On propose les commandes :

```
omega = matrix(rep(0, n^2), n, n)
for (i in 1:n){
for (j in 1:n){
omega[i, j] = (1 / (1-rho^2)) * rho^(abs(i - j))
}}
fnMatSqrtInverse = function(mA) {
ei = eigen(mA)
d = ei$values
d = (d + abs(d)) / 2
d2 = 1 / sqrt(d)
d2[d == 0] = 0
return(ei$vectors %*% diag(d2) %*% t(ei$vectors))
}
Yo = fnMatSqrtInverse(omega) %*% Y
X = cbind(1, X1, X2)
Xo = fnMatSqrtInverse(omega) %*% X
reg2 = lm(Yo ~ Xo[ ,1] + Xo[ ,2] + Xo[ ,3] - 1)
summary(reg2)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| Xo[, 1] | 220.8456 | 753.5125 | 0.29 | 0.7730 | |
| Xo[, 2] | -1064.1358 | 138.3313 | -7.69 | 0.0000 | $***$ |
| Xo[, 3] | 1544.7454 | 343.6588 | 4.49 | 0.0003 | $***$ |

```
Residual standard error: 101.5 on 17 degrees of freedom
Multiple R-squared: 0.9826, Adjusted R-squared: 0.9796
F-statistic: 320.8 on 3 and 17 DF, p-value: 3.693e-15
```

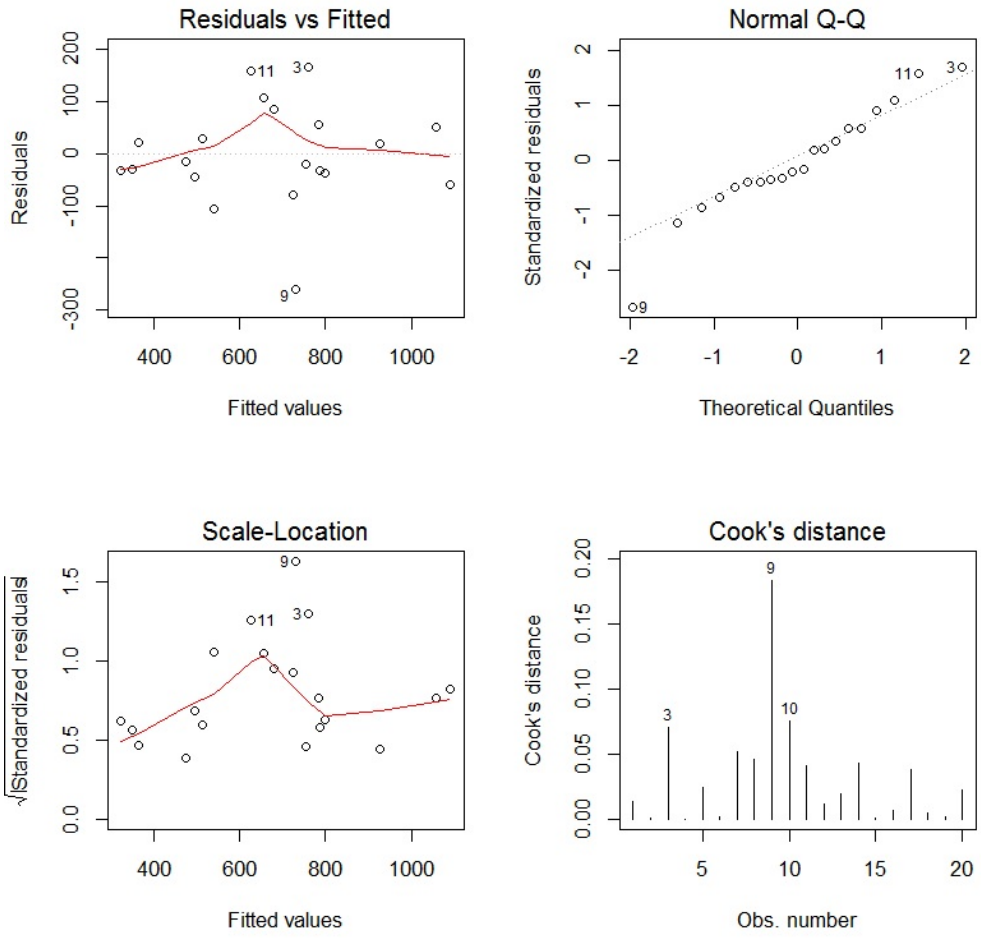
– une estimation ponctuelle de $\beta_0$ est 220.8456,

– une estimation ponctuelle de $\beta_1$ est $-1064.1358$,

– une estimation ponctuelle de $\beta_2$ est 1544.7454.

On vérifie les hypothèses standards via une analyse groupée :

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



Tout semble satisfaisant. Peut-être que la normalité des erreurs est à confirmer :

```
e2 = residuals(reg2)

shapiro.test(e2)
```

Cela renvoie : p-valeur $= 0.267 > 0.05$. On admet donc la normalité des erreurs.

Aussi, étudions l'hypothèse d'homoscédasticité (égalité des variances des erreurs du nouveau modèle) :

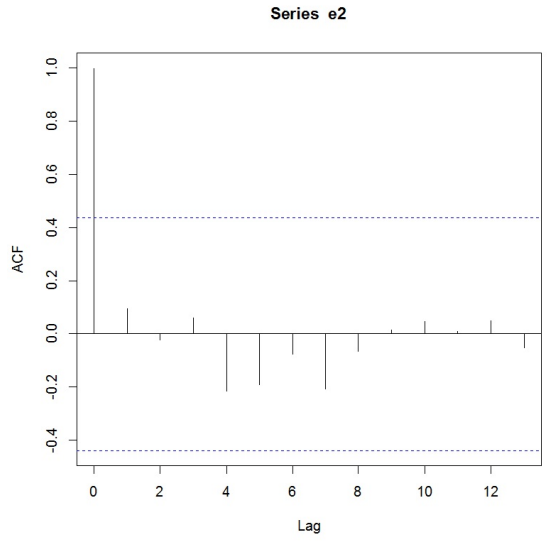
```
library(lmtest)

bptest(reg2)
```

Cela renvoie p-value $= 0.4971 > 0.05$. On admet donc l'égalité des variances des erreurs.

On peut aussi s'interroger sur la présence de l'individu 9 qui semble avoir une forte influence dans tous les domaines de l'analyse.

Notons que la dépendance ne ressort plus de l'*acf* des nouvelles erreurs :

```
acf(reg2)
```




On peut alors faire, entre autre, de l'estimation ponctuelle : pour une valeur de $(X1, X2) = (x_1, x_2) = x$ donnée, la valeur prédite de $Y$ est

$$d_x = 220.8456 - 1064.1358x_1 + 1544.7454x_2.$$

Avec $x_1 = 1.98$ et $x_2 = 1.92$, le nombre de vente moyen du produit est donnée par les commandes :

```
cbind(1, 1.98, 1.92) %*% reg2$coeff
```

Cela renvoie 1079.768, soit 1079.

# Conclusion et étude similaire

## Conclusion

Le modèle de *rlm* standard n'est pas bien adapté au problème car ses erreurs sont clairement corrélées (suivant une structure AR(1)). Pour traiter cela, on a utilisé la méthode des *mcg* avec une estimation de $\rho$ par la méthode des *mco*. Le modèle obtenu est plus convaincant.

Une extension possible de cette étude serait d'estimer $\rho$ par la méthode du maximum de vraisemblance et d'analyser les performance de l'estimateur des *mcg* associé :

```
library(nlme)

reg = gls(Y ~ X1 + X2, correlation = corARMA(p = 1, q = 0), method = "ML")
```

## Étude similaire : Traffic

On peut considérer le jeu de données "traffic" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/traffic.txt",

header = T)
```

L'étude porte sur la concentration en CO prise à 8 mètres d'une autoroute à Los Angeles. Sur une semaine d'été, on dispose :

- de l'heure (de minuit à minuit) (variable $X3$),
- CO : de la concentration moyenne de CO (variable $Y$),
- d'un indicateur de la densité moyenne du trafic (variable $X1$),
- un indicateur de la vitesse moyenne perpendiculaire du vent (variable $X2$).

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

# 6 Étude n° 6 : Dugongs

## Contexte

La taille a été mesurée pour 27 dugongs de différents âges. Ainsi, pour chacun d'entre eux, on dispose :

– de leur taille en mètres (variable $Y$),

– de leur âge en années (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$.

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude6.txt",
header = T)
head(w)
```

| | X1 | Y |
|---|---|---|
| 1 | 1.00 | 1.80 |
| 2 | 1.50 | 1.85 |
| 3 | 1.50 | 1.87 |
| 4 | 1.50 | 1.77 |
| 5 | 2.50 | 2.02 |
| 6 | 4.00 | 2.27 |

On associe les variables $Y$ et $X1$ aux valeurs associées en faisant :

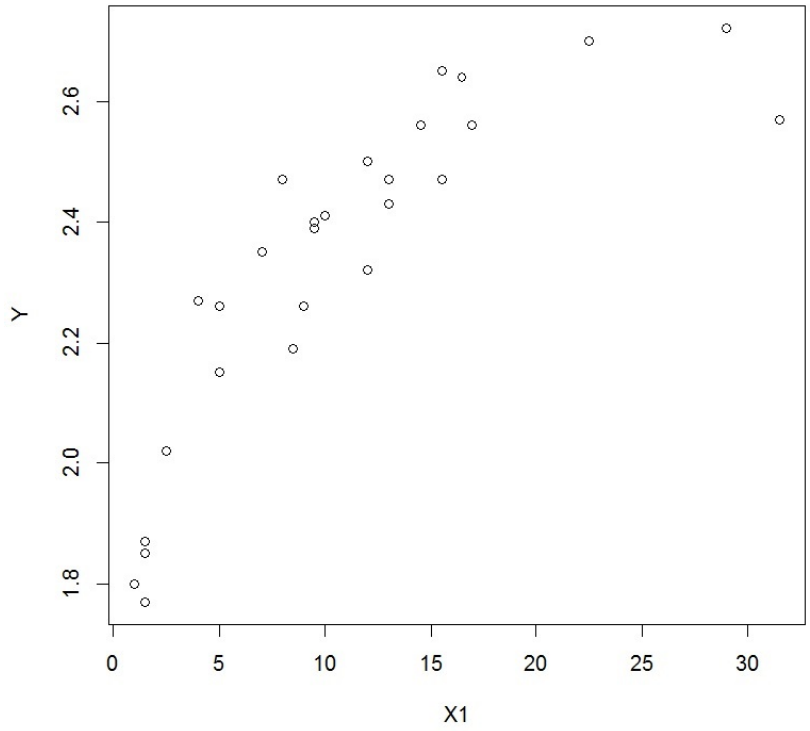
```
attach(w)
```

# Régression linéaire simple

## Analyse du nuage de points

On trace le nuage de points $\{(x_{1,i}, y_i),\ i \in \{1, \ldots, n\}\}$ :

```
plot(X1, Y)
```




On constate une allure courbe du nuage de points. Une liaison linéaire entre $Y$ et $X1$ est étudiable. Toutefois, cela peut clairement être amélioré avec une régression non-linéaire. Nous étudierons cela par la suite.

## Modélisation

Une première approche est de considérer le modèle de *rls* :

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

où $\beta_0$ et $\beta_1$ sont 2 coefficients inconnus, et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma$ inconnu.

La présence de $\beta_0$ est justifiée car même un dugong très jeune peut avoir une taille non négligeable.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

## Estimations

La modélisation de la *rls* et les estimations des paramètres par la méthode des *mco* s'obtiennent par les commandes :

```
reg = lm(Y ~ X1)

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 2.0183 | 0.0522 | 38.67 | 0.0000 | $***$ |
| X1 | 0.0290 | 0.0039 | 7.43 | 0.0000 | $***$ |

```
Residual standard error: 0.1564 on 25 degrees of freedom

Multiple R-squared: 0.6883, Adjusted R-squared: 0.6758

F-statistic: 55.21 on 1 and 25 DF, p-value: 8.794e-08
```

– Test de Student pour $\beta_1$ : influence de $X1$ sur $Y$ : $***$ : hautement significative,

– $R^2 = 0.6883$ et $\overline{R}^2 = 0.6758$ : cela est satisfaisant,

– Test de Fisher : p-valeur $< 0.001$, $***$ : l'utilisation du modèle de *rls* est pertinente.

La valeur prédite de $Y$ quand $X1 = 2$ (par exemple) est donnée par les commandes :

```
predict(reg, data.frame(X1 = 2))
```

Cela renvoie 2.076197.

Ainsi, la taille moyenne d'un dugong de 2 ans est de 2.076197 m.

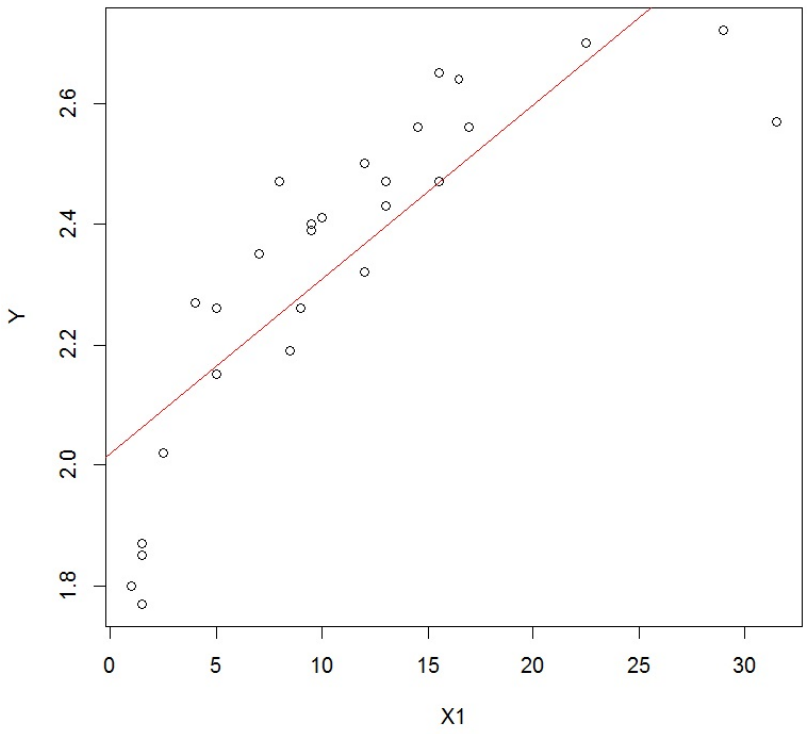
## Droite de régression

En utilisant les estimations de ponctuelles $\beta_0$ et $\beta_1$, l'équation de la droite de régression est :

$$y = 97.0771 + 0.9493x.$$

On la visualise avec les commandes :

```
abline(reg, col = "red")
```



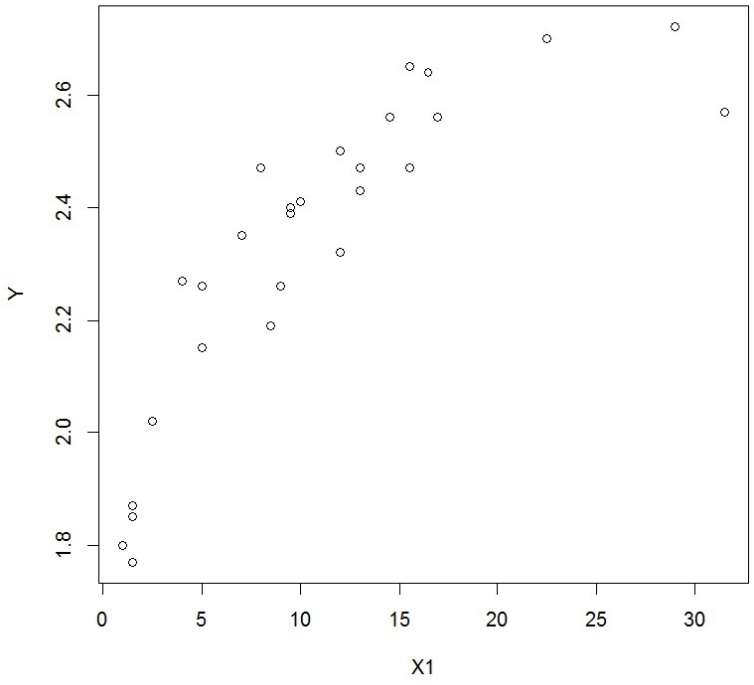

Nous voyons, sans surprise, que cette droite ajuste mal le nuage de points.

# Régression non-linéaire

## Modélisation

Une autre possibilité de modélisation est une régression non-linéaire, justifiée vu l'allure courbe du nuage de points $\{(x_{1,i}, y_i),\ i \in \{1, \ldots, n\}\}$ :

```
plot(X1, Y)
```




On peut vérifier la pertinence d'un tel modèle avec le test de Rainbow :

```
library(lmtest)

raintest(reg)
```

Cela renvoie : p-valeur = 0.004197. Comme p-valeur $< 0.05$, le modèle de régression non-linéaire est adapté aux données.

On considère alors la modélisation :

$$Y = f(X1) + \epsilon,$$

où $f$ désigne une fonction inconnue, et $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Pour ce faire, nous allons utiliser 4 méthodes :

- la considération de $\log(X_1)$ dans un modèle de *rls*,
- la régression polynomiale,
- la méthode itérative,
- la méthode à noyau.

## Transformation logarithmique de $X1$

En remarquant l'allure logarithmique du nuage de points, on peut étudier le modèle :

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

Les estimations des paramètres par la méthode des *mco* s'obtiennent par les commandes :

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

summary(reg)
```

Cela renvoie :

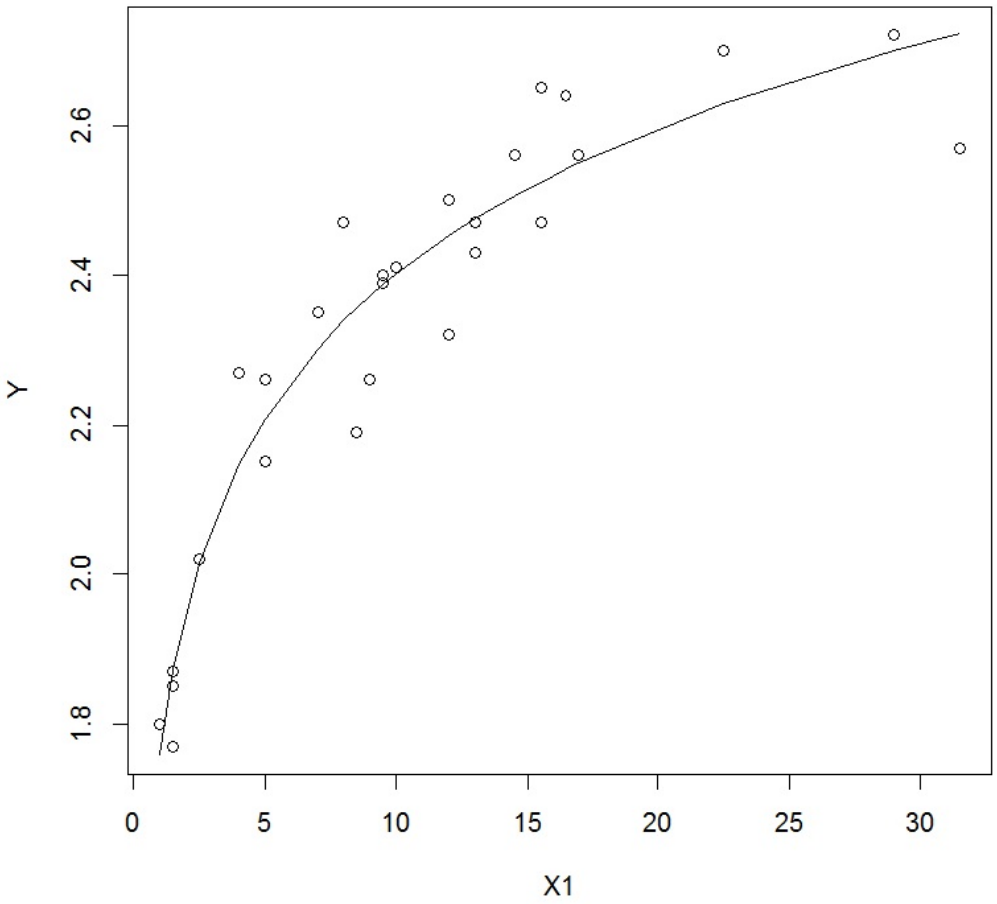
| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 1.7585 | 0.0397 | 44.34 | 0.0000 | *** |
| log(X1) | 0.2794 | 0.0176 | 15.92 | 0.0000 | *** |

```
Residual standard error: 0.08393 on 25 degrees of freedom

Multiple R-squared: 0.9102, Adjusted R-squared: 0.9066

F-statistic: 253.4 on 1 and 25 DF, p-value: 1.36e-14
```

Les résultats de significativité des coefficients sont bons, avec un $\overline{R}^2 = 0.9066$.

On représente l'estimation obtenue sur un graphique :

```
plot(X1, Y)

lines(X1, fitted(reg))
```




Valeurs de AIC et BIC :

```
AIC(reg)
```

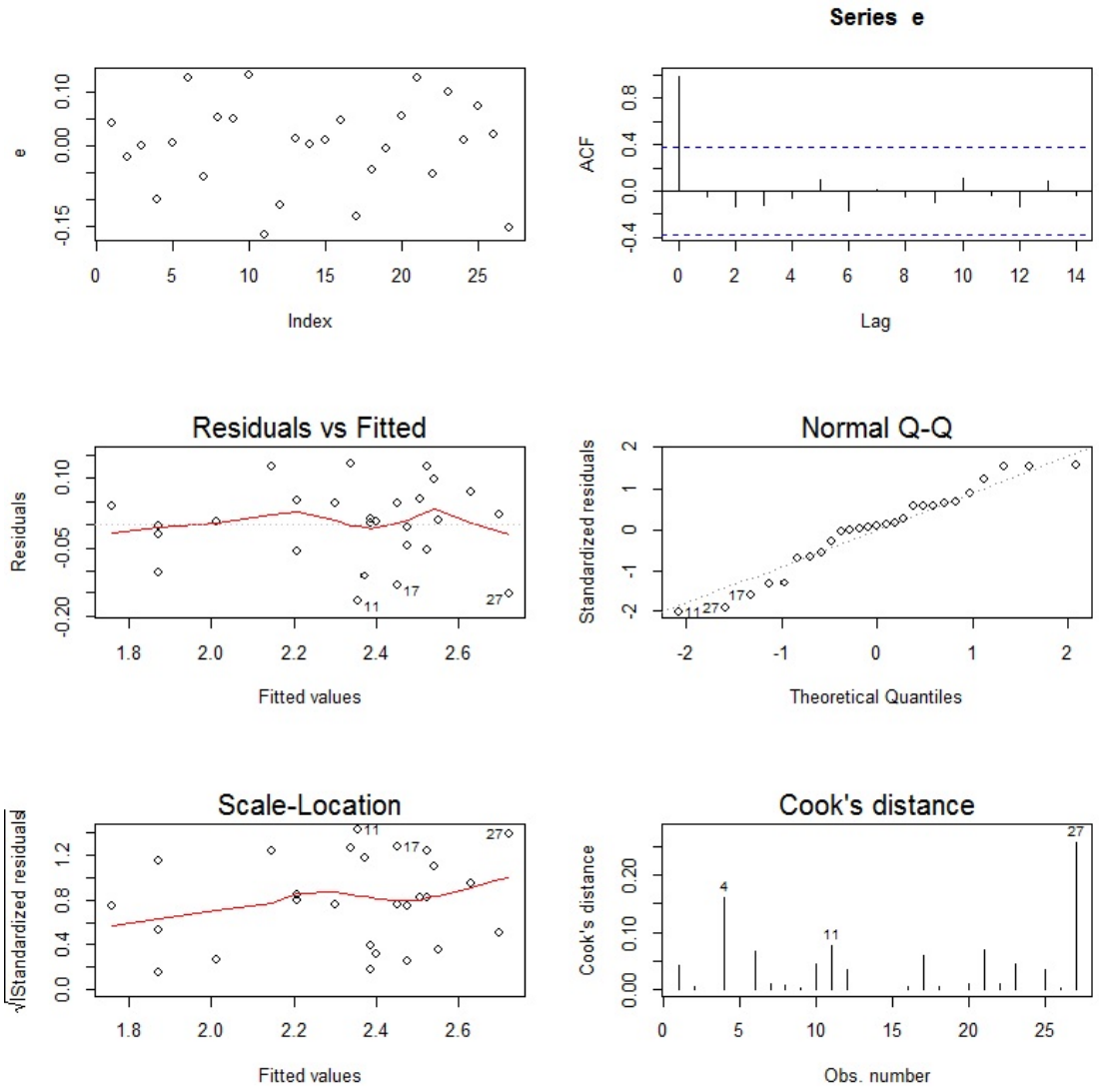
Cela renvoie $-53.25219$.

```
BIC(reg)
```

Cela renvoie $-49.36468$.

On trace les graphiques nous permettant de conclure à la validation ou non des hypothèses de base :

```
e = residuals(reg)

par(mfrow = c(3, 2))

plot(e)

acf(e)

plot(reg, 1:4)
```




Tout semble satisfaisant.

### Régression polynomiale

Les commandes relatives au modèle de régression polynomiale (avec orthonormalisation) sont :

```
reg = lm(Y ~ poly(X1, 8))

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 2.3352 | 0.0158 | 148.14 | 0.0000 | $***$ |
| poly(X1, 8)1 | 1.1619 | 0.0819 | 14.19 | 0.0000 | $***$ |
| poly(X1, 8)2 | -0.6325 | 0.0819 | -7.72 | 0.0000 | $***$ |
| poly(X1, 8)3 | 0.1043 | 0.0819 | 1.27 | 0.2191 | |
| poly(X1, 8)4 | -0.2295 | 0.0819 | -2.80 | 0.0118 | $*$ |
| poly(X1, 8)5 | 0.1193 | 0.0819 | 1.46 | 0.1626 | |
| poly(X1, 8)6 | -0.0931 | 0.0819 | -1.14 | 0.2704 | |
| poly(X1, 8)7 | -0.0479 | 0.0819 | -0.58 | 0.5663 | |
| poly(X1, 8)8 | 0.0418 | 0.0819 | 0.51 | 0.6162 | |

À partir du degré 4, les coefficients associés aux puissances de $X1$ ne sont plus significatifs. On propose de les retirer de la modélisation.

```
reg = lm(Y ~ poly(X1, 4))

summary(reg)
```

Cela renvoie :

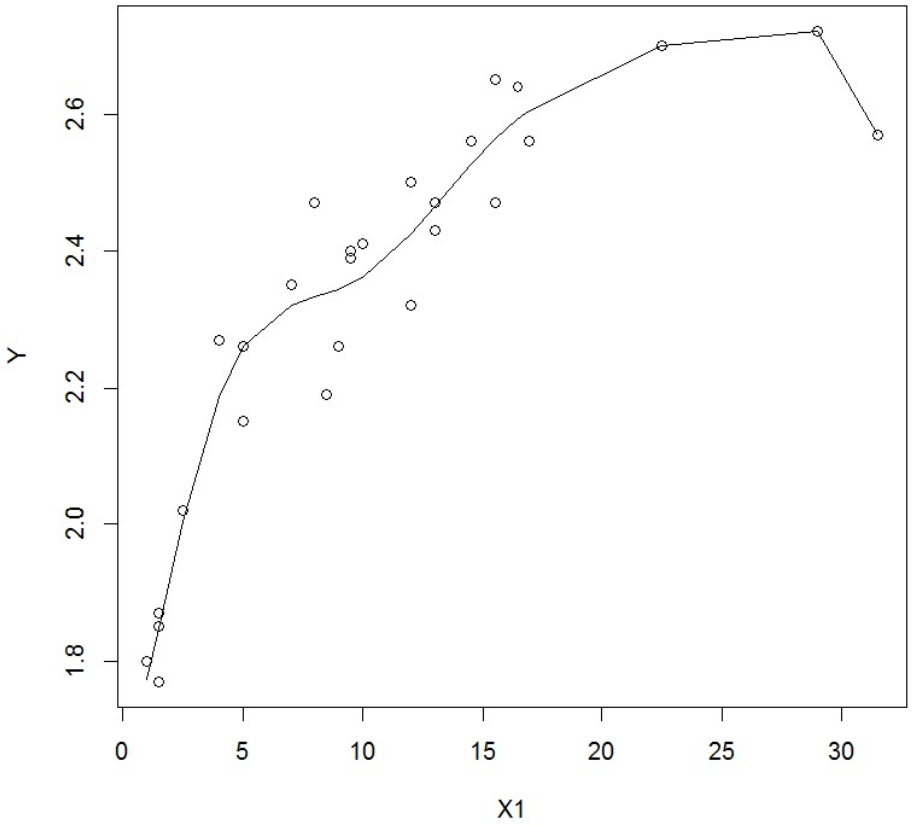
| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 2.3352 | 0.0158 | 148.09 | 0.0000 | $***$ |
| poly(X1, 4)1 | 1.1619 | 0.0819 | 14.18 | 0.0000 | $***$ |
| poly(X1, 4)2 | -0.6325 | 0.0819 | -7.72 | 0.0000 | $***$ |
| poly(X1, 4)3 | 0.1043 | 0.0819 | 1.27 | 0.2163 | |
| poly(X1, 4)4 | -0.2295 | 0.0819 | -2.80 | 0.0104 | $*$ |

```
Residual standard error: 0.08193 on 22 degrees of freedom

Multiple R-squared: 0.9247, Adjusted R-squared: 0.911

F-statistic: 67.54 on 4 and 22 DF, p-value: 4.933e-12
```

Les résultats de significativité des coefficients sont bons, avec un $\overline{R}^2 = 0.911$.

On représente l'estimation obtenue sur un graphique :

```
plot(X1, Y)

lines(X1, fitted(reg))
```




Valeurs de AIC et BIC :

```
AIC(reg)
```

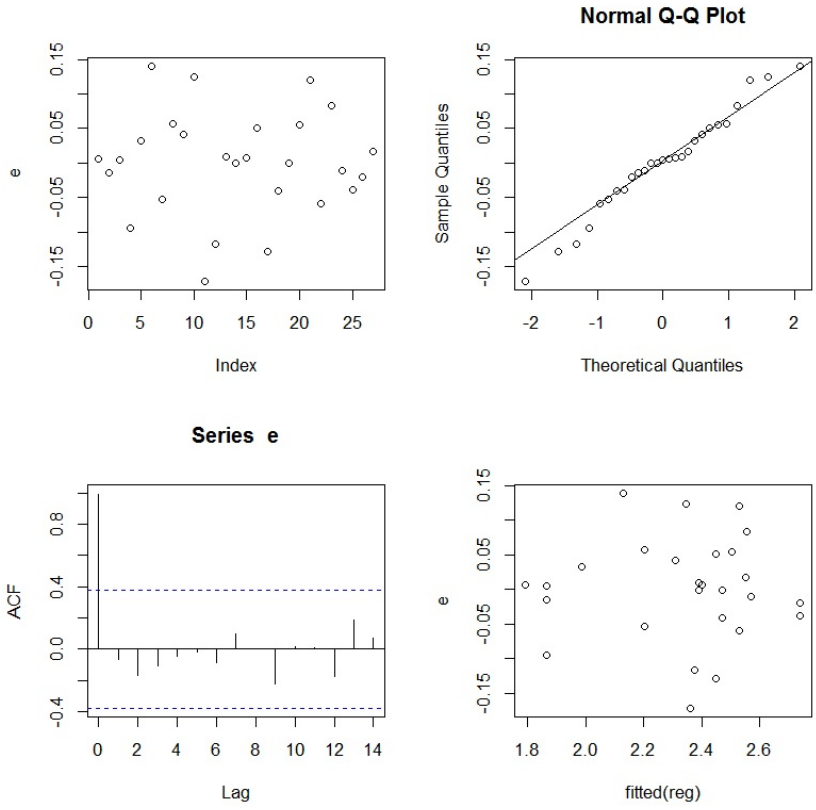
Cela renvoie $-52.00602$.

```
BIC(reg)
```

Cela renvoie $-44.23099$.

Pour valider les hypothèses standards, voici une proposition d'analyse :

```
e = residuals(reg)
par(mfrow = c(2, 2))
plot(e)
qqnorm(e)
qqline(e)
acf(e)
plot(fitted(reg), e)
```




Tout semble satisfaisant. En particulier :

– le graphique des résidus est satisfaisant : uniformité dans la répartition et symétrie,

– le QQ plot montre des points à peu près alignés,

– il n'y a pas de structure dans l'*acf*,

– il n'y a pas de structure dans le nuage de points $\{(e_i, y_i - e_i);\ i \in \{1, \ldots, n\}\}$.

### Méthode itérative

Vu l'allure du nuage de points, on peut envisager une modélisation de $f$ sous la forme :

$$f(x) = \beta_0 - \beta_1 \beta_2^x,$$

où $\beta_0$, $\beta_1$ et $\beta_2$ sont 3 coefficients inconnus.

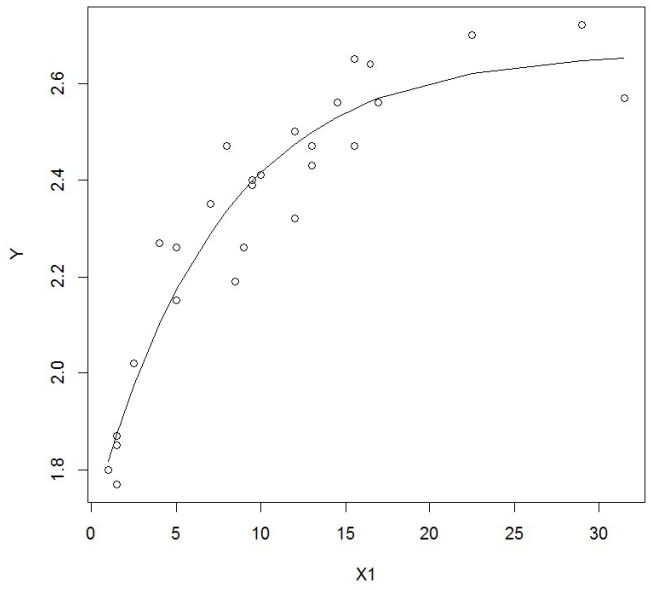
```
library(MASS)
s = c(b0 = 2, b1 = 9, b2 = 0.1)
reg = nls(Y ~ b0 - b1 * b2 ^X1, start = s)
reg
```

Cela renvoie :

```
Nonlinear regression model
  model: Y ~ b0 - b1 * b2 ^X1
  data: parent.frame()
  b0 b1 b2
2.6666   0.9725   0.8735
residual sum-of-squares: 0.1868
Number of iterations to convergence: 8
Achieved convergence tolerance: 7.309e-07
```

On représente l'estimation obtenue sur un graphique :

```
plot(X1, Y)
lines(X1, fitted(reg))
```




Valeurs de AIC et BIC :

```
AIC(reg)
```

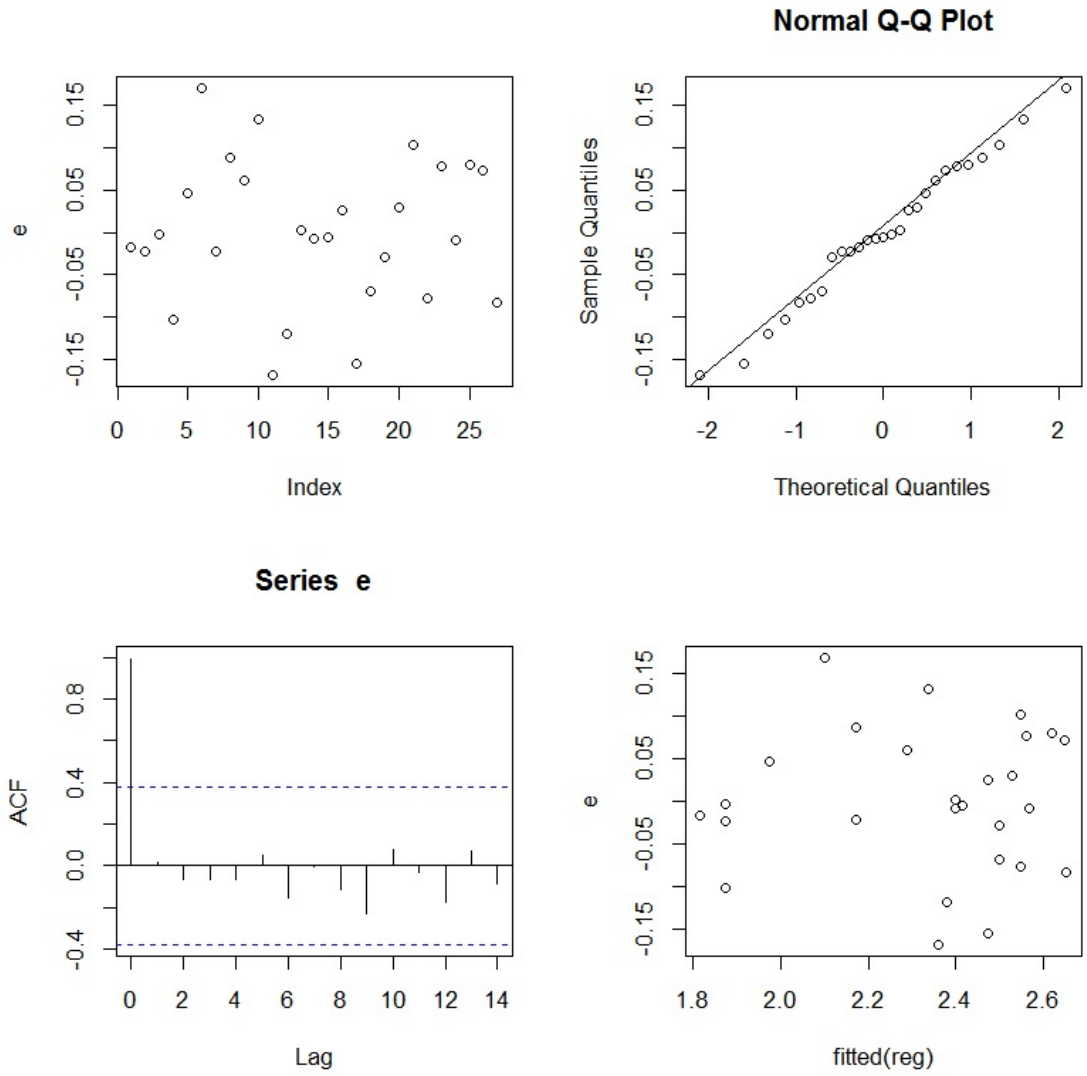
Cela renvoie −49.66917.

```
BIC(reg)
```

Cela renvoie −44.48582.

Pour valider les hypothèses standards, voici une proposition d'analyse :

```
e = residuals(reg)
par(mfrow = c(2, 2))
plot(e)
qqnorm(e)
qqline(e)
acf(e)
plot(fitted(reg), e)
```




Tout semble satisfaisant.

### Méthode à noyau

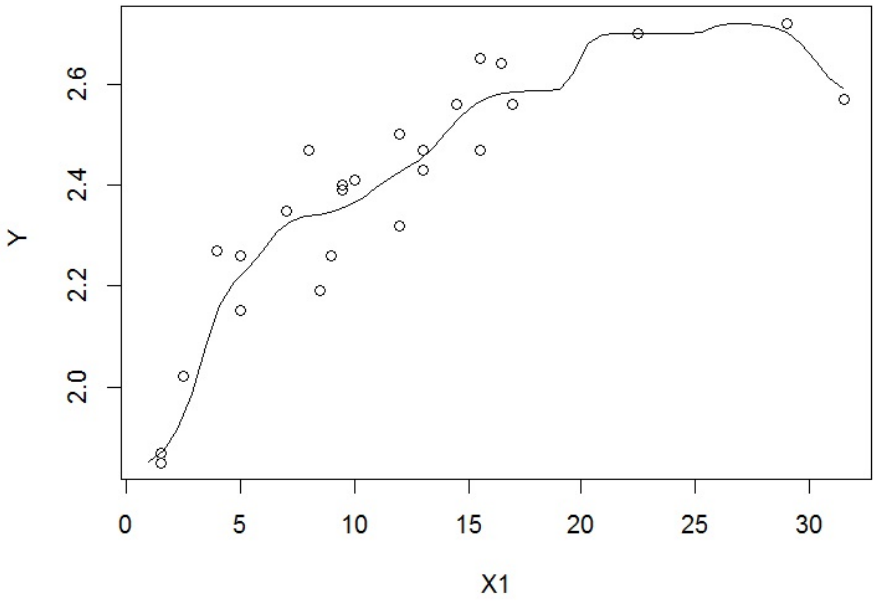
Les commandes suivantes appellent l'estimateur de Nadaraya-Watson :

```
library(np)
reg = npreg(Y ~ X1)
summary(reg)
```

Cela renvoie $R^2 = 0.9322363$. La définition de ce coefficient diffère de celle utilisée pour le modèle de *rlm*. Toutefois, le fait qu'il soit proche de 1 traduit une bonne estimation de $f$.

On représente graphiquement l'estimateur obtenu :

```
plot(reg)
points(X1, Y)
```



On constate un très bon ajustement du nuage de points.

La valeur prédite de $Y$ quand $X1 = 2$ (par exemple) est donnée par les commandes :

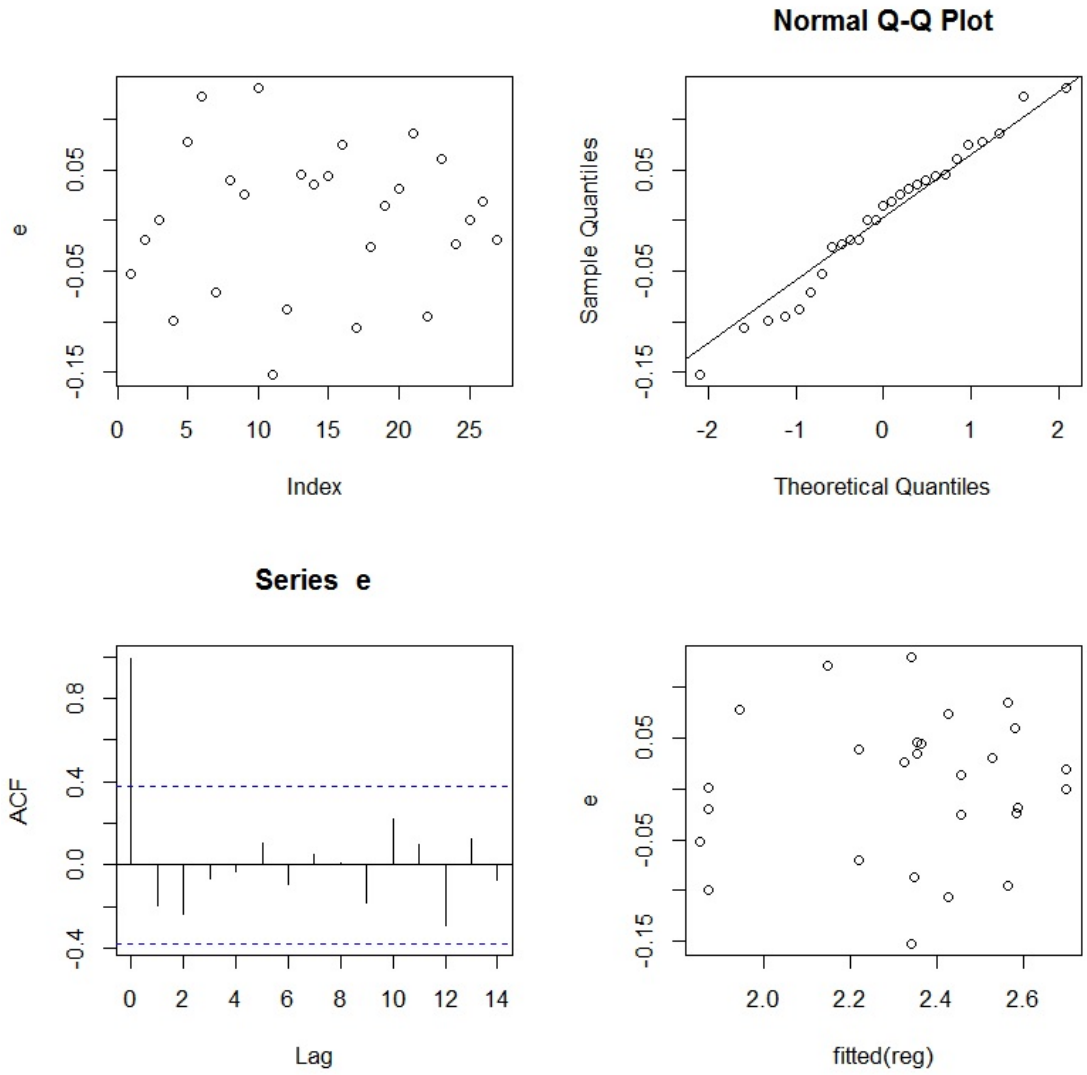
```
predict(reg, newdata = data.frame(X1 = 2))
```

Cela renvoie 1.897279, contre 2.076197 pour le premier modèle, la première valeur étant plus fiable.

Ainsi, la taille moyenne d'un dugong de 2 ans est de 1.897279 m.

Pour valider les hypothèses standards, voici une proposition d'analyse :

```
e = residuals(reg)
par(mfrow = c(2, 2))
plot(e)
qqnorm(e)
qqline(e)
acf(e)
plot(fitted(reg), e)
```




Tout semble satisfaisant.

# Conclusion et études similaires

## Conclusion

L'étude statistique mise en œuvre montre que le modèle de régression non-linéaire est mieux adapté au problème que le modèle de *rlm*.

## Étude similaire 1 : Lapins

On peut considérer le jeu de données "lapins" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/lapins.txt",
header = T)
```

En Australie, la quantité d'humidité dans un œil a été mesurée pour 71 lapins de différents âges. Ainsi, pour chacun d'entre eux, on dispose :

- de leur quantité d'humidité en millilitres (variable $Y$),
- de leur âge en jours (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$.

## Étude similaire 2 : Blé

On peut considérer le jeu de données "blé" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/blé.txt", header = T)
```

L'étude porte sur le rendement d'une culture de blé en fonction de la hauteur de pluie printanière. Pour 54 parcelles, on dispose :

- du rendement de blé (variable $Y$),
- de la hauteur de pluie en mètres (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$.

### Étude similaire 3 : Caries

On peut considérer le jeu de données "caries" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/caries.txt",
header = T)
```

L'étude porte sur le nombre de caries chez l'enfant en fonction du taux de fluorure dans l'eau du robinet. Pour 21 villes différentes, on dispose :

– du nombre de caries pour 100 enfants examinés (variable $Y$),

– de la teneur en fluorure dans l'eau du robinet (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$. On traitera une valeur aberrante.

### Étude similaire 4 : Canon

On peut considérer le jeu de données "canon" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/canon.txt", header = T)
```

On s'intéresse à la portée d'un canon suivant l'angle de tir. Pour 17 angles différents, on dispose :

– de la portée du tir en mètres (variable $Y$),

– des degrés de l'angle (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$.

### Étude similaire 5 : Neige

On peut considérer le jeu de données "neige" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/neige.txt", header = T)
```

On s'intéresse à la relation entre la teneur en eau de la neige tombée et le rendement en eau en pouces dans une rivière des États-Unis.

On dispose :

– du rendement en eau de la rivière d'Avril à Juillet (variable $Y$),

– de la teneur en eau de la neige (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$.

# 7 Étude n° 7 : Savoir-faire

## Contexte

On considère un jeu de données "fabriqué pour l'occasion" dans lequel on souhaite expliquer une variable quantitative $Y$ à partir de 4 variables quantitatives : $X1, \ldots, X4$.

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude7.txt",
header = T)
head(w)
```

| | Y | X1 | X2 | X3 | X4 |
|---|---|---|---|---|---|
| 1 | -145.76 | -71.80 | -115.34 | 91.47 | 3.12 |
| 2 | -15.30 | 86.10 | -195.29 | -95.93 | 3.46 |
| 3 | 1558.48 | -63.60 | -35.75 | -282.05 | 355.90 |
| 4 | -20.77 | -201.71 | -35.34 | 114.66 | 5.10 |
| 5 | 541.84 | -2.65 | 52.28 | -164.89 | 3.28 |
| 6 | 135.84 | 8.59 | -144.82 | -8.52 | 6.30 |

On associe les variables $Y$ et $X1, \ldots, X4$ aux valeurs associées en faisant :

```
attach(w)
```

# Régression linéaire multiple

## Modélisation

Une première approche est de considérer le modèle de *rls* :

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \epsilon,$$

où $\beta_0, \ldots, \beta_4$ sont 5 coefficients inconnus, et $\epsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma$ inconnu.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

## Estimations

La modélisation de la *rlm* avec les variables explicatives $X1, \ldots, X4$, et les estimations des paramètres par la méthode des *mco* s'obtiennent par les commandes :

```
reg = lm(Y ~ ., w)

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 570.5813 | 11.6173 | 49.11 | 0.0000 | $***$ |
| X1 | 2.1013 | 0.0992 | 21.18 | 0.0000 | $***$ |
| X2 | 3.1131 | 0.0970 | 32.09 | 0.0000 | $***$ |
| X3 | 0.2300 | 0.0966 | 2.38 | 0.0176 | $*$ |
| X4 | 9.0417 | 0.3966 | 22.80 | 0.0000 | $***$ |

```
Residual standard error: 215 on 495 degrees of freedom

Multiple R-squared: 0.7972, Adjusted R-squared: 0.7956

F-statistic: 486.5 on 4 and 495 DF, p-value: < 2.2e-16
```

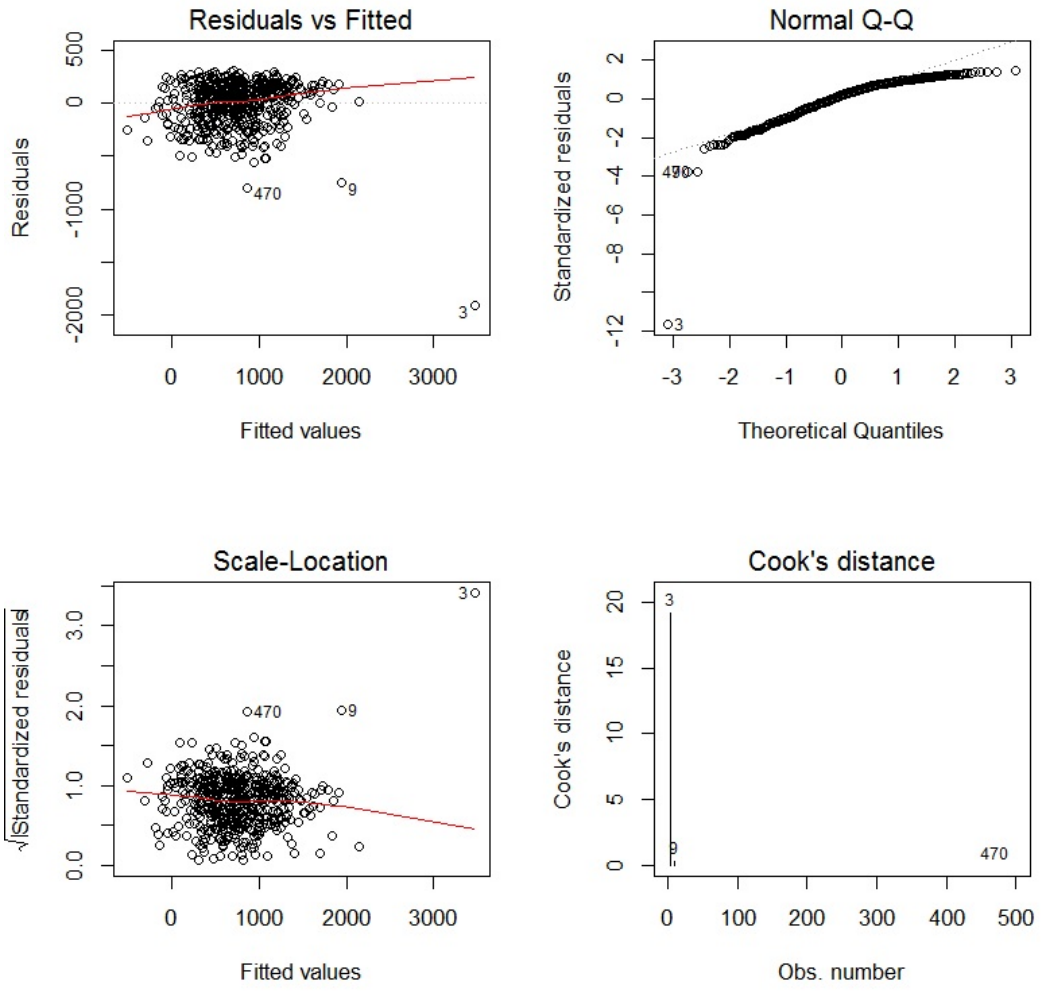
- $R^2 = 0.7972$ et $\overline{R}^2 = 0.7956$ : cela est correct,
- Test de Fisher : p-valeur $< 0.001$, $***$ : l'utilisation du modèle de *rlm* est pertinente.

# Validation des hypothèses

## Analyse groupée des hypothèses

On trace les graphiques nous permettant de conclure à la validation ou non des hypothèses de base :

```
par(mfrow = c(2, 2))

plot(reg, 1:4)
```



Parmi les points problématiques, il y a présence de 2 valeurs anormales associées aux individus 3 et 9. En effet, leurs distances de Cook dépassent 1.

### Traitement des valeurs anormales

On propose d'exclure les individus 3 et 9 du jeu de donnés et de recommencer l'analyse :

```
reg2 = lm(Y ~ ., w, subset = -c(3, 9))

summary(reg2)
```

Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 493.0292 | 10.0539 | 49.04 | 0.0000 | *** |
| X1 | 1.9940 | 0.0777 | 25.66 | 0.0000 | *** |
| X2 | 3.0627 | 0.0758 | 40.40 | 0.0000 | *** |
| X3 | 0.0024 | 0.0766 | 0.03 | 0.9747 | |
| X4 | 14.7980 | 0.4474 | 33.07 | 0.0000 | *** |

```
Residual standard error: 167.8 on 493 degrees of freedom

Multiple R-squared: 0.8761, Adjusted R-squared: 0.8751

F-statistic: 871.2 on 4 and 493 DF, p-value: < 2.2e-16
```

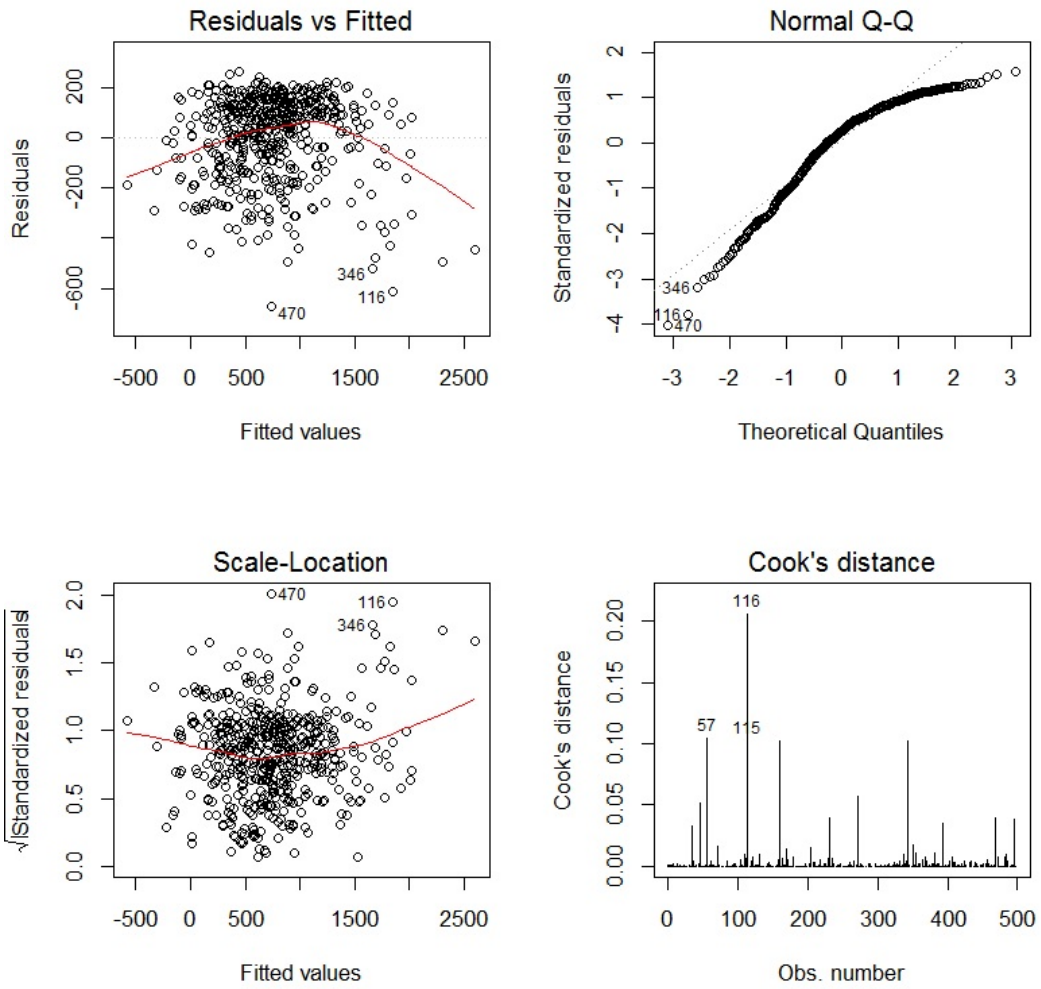
On constate :

– un meilleur $\overline{R}^2$ : on a $\overline{R}^2 = 0.8751$, contre $\overline{R}^2 = 0.7956$ pour la première analyse,

– une variable $X3$ non significative (sa p-valeur est même proche de 1).

Étudions la validation des hypothèses :

```
par(mfrow = c(2, 2))

plot(reg2, 1:4)
```




Parmi les points problématiques, celui de la non normalité des erreurs est à étudier.

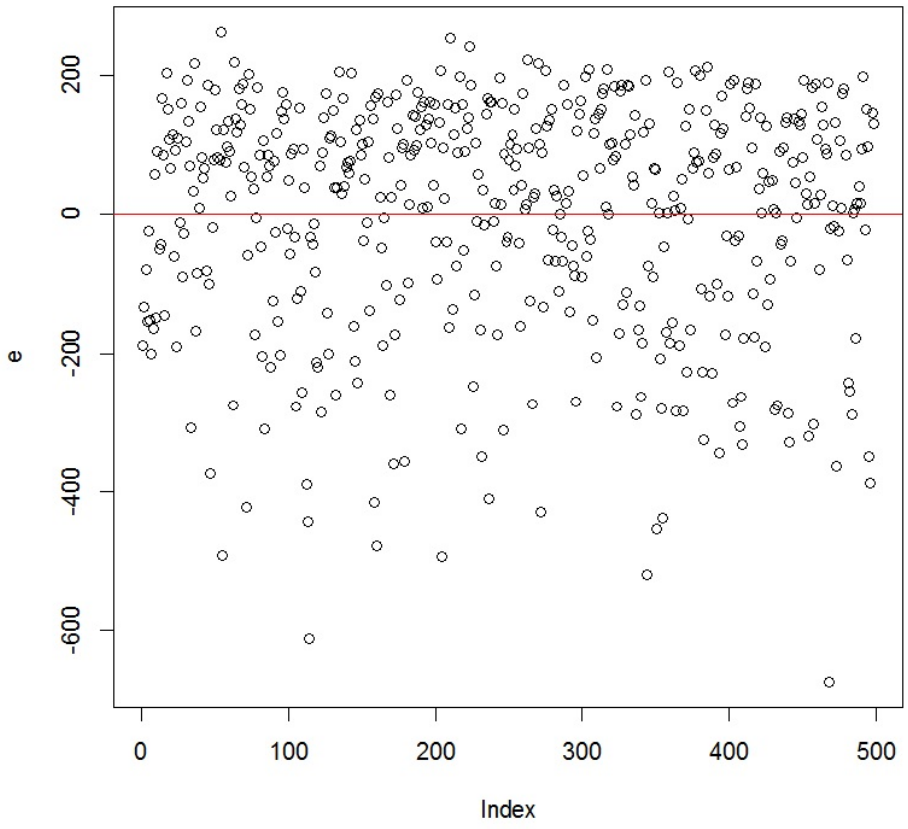
Complétons cette analyse graphique par d'autres analyses ciblées et des tests statistiques adaptés.

## Analyse graphique des résidus

On examine les résidus en faisant :

```
e = residuals(reg2)

plot(e)

abline(h = 0, col = "red")
```

Cela renvoie :

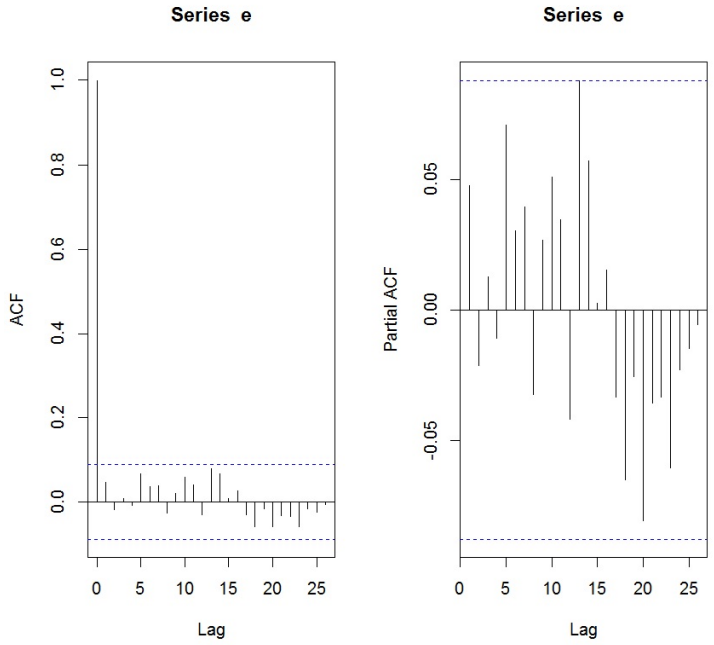



Les points ne sont pas du tout uniformément répartis et il n'y a pas de symétrie. On ne constate aucune structure toutefois.

### Indépendance de $\epsilon_1, \ldots, \epsilon_n$

On étudie l'indépendance de $\epsilon_1, \ldots, \epsilon_n$ avec les graphiques *acf* et *pacf* :

```
par(mfrow = c(1, 2))

acf(e)

pacf(e)
```




On ne constate aucune structure particulière et peu de bâtons dépassent les bornes limites. Cela traduit l'indépendance de $\epsilon_1, \ldots, \epsilon_n$.

### Égalité des variances de $\epsilon_1, \ldots, \epsilon_n$

Étudions l'égalité des variances avec le test de Breusch-Pagan :

```
library(lmtest)

bptest(reg2)
```

Cela renvoie : p-valeur $= 1.323e-07 < 0.001 ***$. Il n'y a pas donc pas égalité des variances.

### Normalité de $\epsilon_1, \ldots, \epsilon_n$

On peut vérifier cela avec le test de Shapiro-Wilk :

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

Cela renvoie : p-valeur $= 5.089e-16$. Comme p-valeur $< 0.05 * * *$, on rejette la normalité de $\epsilon_1, \ldots, \epsilon_n$.

### Étude de la multicolinéarité

On étudie la valeur des *vif* :

```
library(car)

vif(reg2)
```

Cela renvoie :

| $V_1$ | $V_2$ | $V_3$ | $V_4$ |
|---|---|---|---|
| 1.009763 | 1.010481 | 1.026481 | 1.017855 |

Comme ils sont tous inférieurs à 5, il n'y a pas de multicolinéarité entre $X1$, $X2$, $X3$ et $X4$.

## Tentatives d'amélioration

### Idées

Pour tenter d'améliorer le modèle `reg2`, au vu des analyses précédentes, on propose :

- le retrait de la variable $X3$ dans le modèle (laquelle est loin d'être significative ; sa p-valeur est proche de 1),
- l'étude des liaisons linéaires de $X1$, $X2$ et $X4$ sur $Y$ en vue de la construction d'un modèle non-linéaire.

### Suppression de $X3$

On retire $X3$ du modèle ce qui donne :

```
reg3 = lm(Y ~ X1 + X2 + X4, w, subset = -c(3, 9))

summary(reg3)
```

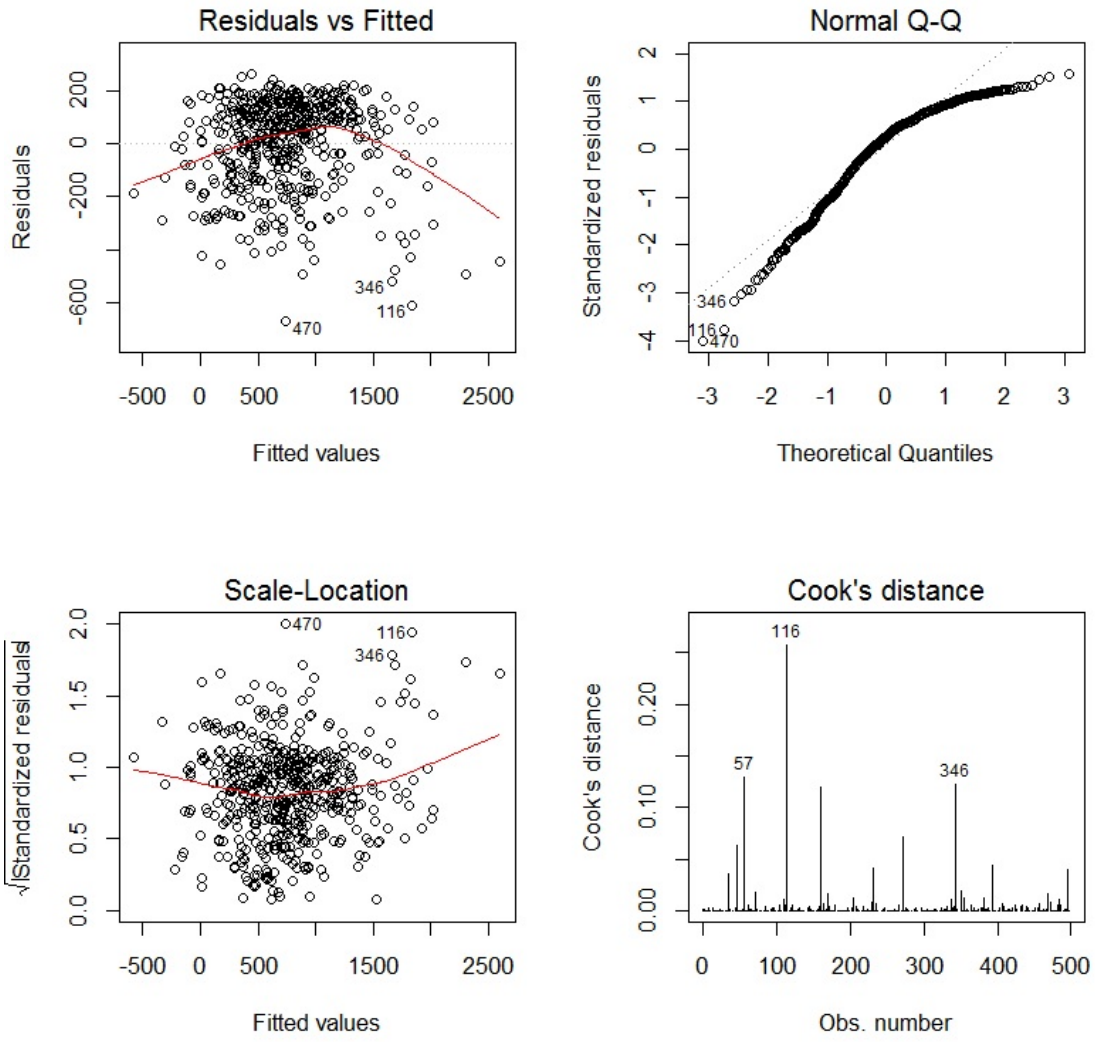
Cela renvoie :

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 493.0297 | 10.0437 | 49.09 | 0.0000 | *** |
| X1 | 1.9939 | 0.0774 | 25.75 | 0.0000 | *** |
| X2 | 3.0629 | 0.0755 | 40.57 | 0.0000 | *** |
| X4 | 14.7997 | 0.4437 | 33.35 | 0.0000 | *** |

```
Residual standard error: 167.6 on 494 degrees of freedom

Multiple R-squared: 0.8761, Adjusted R-squared: 0.8753

F-statistic: 1164 on 3 and 494 DF, p-value: < 2.2e-16
```

Étudions la validation des hypothèses :

```
par(mfrow = c(2, 2))

plot(reg3, 1:4)
```

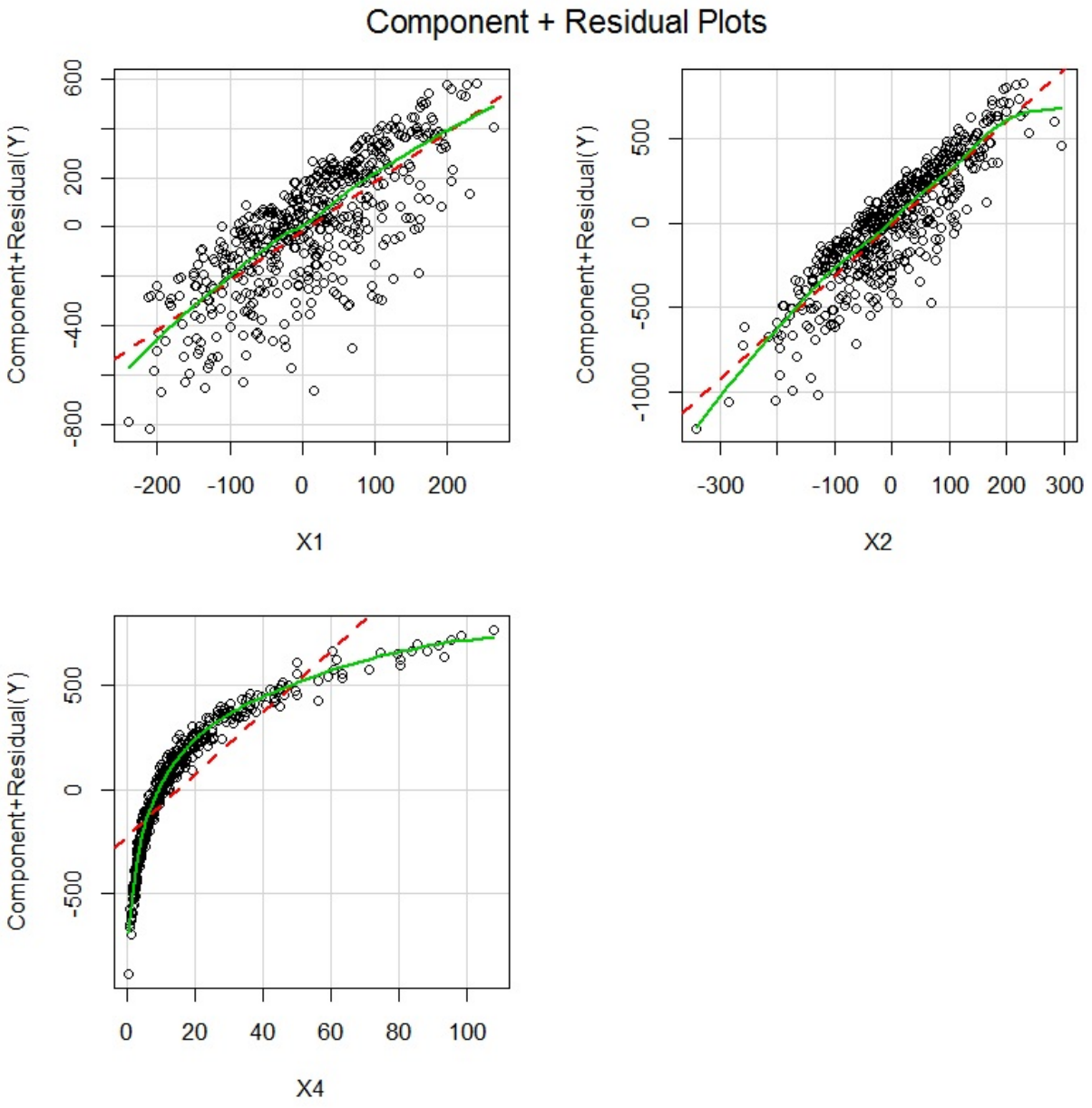



Encore une fois, plusieurs nombreux problèmes se profilent indiquant que le modèle de régression linéaire standard n'est pas adapté.

## Analyse des liens linéaires

On étudie la pertinence des liens linéaires de $X1$, $X2$ et $X4$ sur $Y$ en faisant :

```
library(car)

crPlots(reg3)
```



On constate que :

– la liaison linéaire entre $Y$ et $X1$ est claire,

– la liaison linéaire entre $Y$ et $X2$ est claire,

– la liaison linéaire entre $Y$ et $X4$ est inexistante ; on peut ajuster le nuage de points avec la fonction $y = \ln(x)$.

Par conséquent, il est judicieux de considérer le modèle de régression non-linéaire :

```
reg4 = lm(Y ~ X1 + X2 + log(X4), w, subset = -c(3, 9))

summary(reg4)
```

Cela renvoie :

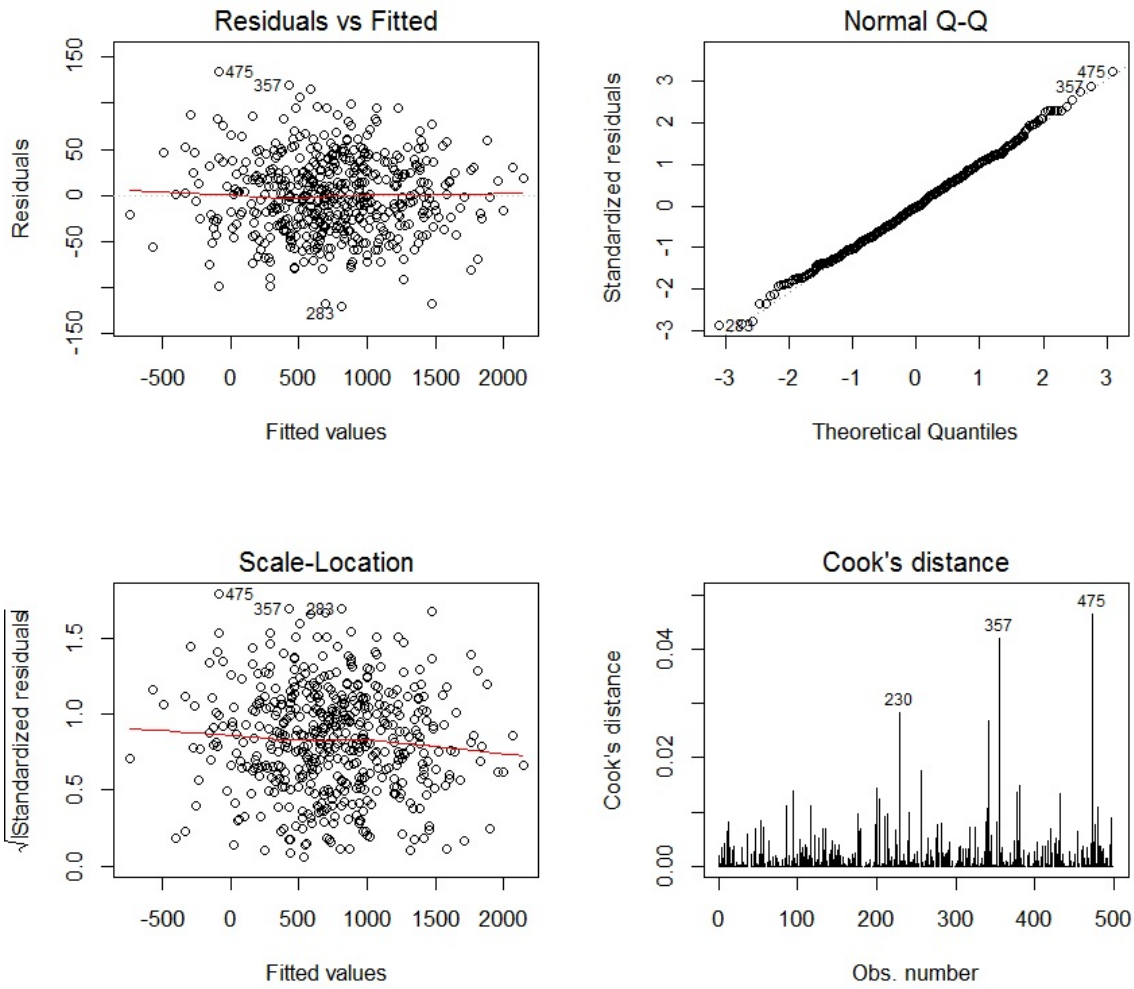
| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 50.9889 | 4.6007 | 11.08 | 0.0000 | $***$ |
| X1 | 2.0135 | 0.0194 | 103.67 | 0.0000 | $***$ |
| X2 | 3.0314 | 0.0190 | 159.95 | 0.0000 | $***$ |
| log(X4) | 298.8710 | 1.8900 | 158.13 | 0.0000 | $***$ |

```
Residual standard error: 42.08 on 494 degrees of freedom

Multiple R-squared: 0.9922, Adjusted R-squared: 0.9921

F-statistic: 2.093e+04 on 3 and 494 DF, p-value: < 2.2e-16
```

Tous les tests statistiques sont $***$ et $\overline{R}^2 = 0.9921$, ce qui est très bon.

Étudions la validation des hypothèses :

```
par(mfrow = c(2, 2))

plot(reg4, 1:4)
```




Tout est satisfaisant. Cela se confirme avec les tests statistiques usuels.

Ce dernier modèle semble tout à fait correct pour faire de la prédiction entre autre. On a :

– Estimations ponctuelles de $\beta_0$, $\beta_1$, $\beta_2$ et $\beta_3$ :

| $\widehat{\beta}_0$ | $\widehat{\beta}_1$ | $\widehat{\beta}_2$ | $\widehat{\beta}_3$ |
|---|---|---|---|
| 50.9889 | 2.0135 | 3.0314 | 298.8710 |

– Estimations ponctuelles des écart-types des estimateurs de $\beta_0$, $\beta_1$, $\beta_2$ et $\beta_3$ :

| $\widehat{\sigma}(\widehat{\beta}_0)$ | $\widehat{\sigma}(\widehat{\beta}_1)$ | $\widehat{\sigma}(\widehat{\beta}_2)$ | $\widehat{\sigma}(\widehat{\beta}_3)$ |
|---|---|---|---|
| 4.6007 | 0.0194 | 0.0190 | 1.8900 |

– $t_{obs}$ :

| $H_1$ | $\beta_0 \neq 0$ | $\beta_1 \neq 0$ | $\beta_2 \neq 0$ | $\beta_3 \neq 0$ |
|---|---|---|---|---|
| $t_{obs}$ | 11.08 | 103.67 | 159.95 | 158.13 |

– Les tests de Student nous disent que les coefficients $\beta_0, \ldots, \beta_5$ sont différents de 0 avec le degré de significativité $***$,

– $R^2 = 0.9922$ et $\overline{R}^2 = 0.9921$ : cela est très correct,

– Test de Fisher : p-valeur $< 0.001$, $***$ : le modèle est très bien expliqué par le modèle de *rlm*.

La valeur prédite de $Y$ quand $X1 = 106$, $X2 = 9$ et $X4 = 109$ (par exemple) est donnée par :

```
predict(reg4, data.frame(X1 = 106, X2 = 9, X4 = 109))
```

Cela renvoie 1693.814.

# Conclusion et étude similaire

## Conclusion

Le modèle de *rlm* à soulever plusieurs problèmes, dont le traitement de 2 valeurs anormales (la suppression de $X3$, variable non significative) et un lien non-linéaire entre $Y$ et $X4$. On a alors construit un modèle de régression non-linéaire plus adapté.

## Étude similaire : Calmars

On peut considérer le jeu de données "calmars" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/calmars.txt",
header = T)
```

Une étude a été menée dans le but d'étudier le poids des calmars dévorés par les requins. Les variables explicatives sont relatives au bec du calmar. Pour 22 calmars, on dispose :

– de la longueur rostrale en pouces (variable $X1$),

– de la longueur de l'aile en pouces (variable $X2$),

– de la longueur du rostre au cran (variable $X3$),

– de la longueur du cran à l'aile (variable $X4$),

– de la largeur en pouces (variable $X5$),

– du poids du calmar en livres (variable $Y$).

On souhaite expliquer $Y$ à partir de $X1$.

Après l'étude de nombreux aspects, on doit aboutir au fait que le modèle le plus adapté est multiplicatif : $Y = X1 \times X5 \times \xi$ :

```
reg = lm(log(Y) ~ log(X1) + log(X5))
```

# 8 Étude n° 8 : Anisophyllea

## Contexte

On s'intéresse à la présence d'Anisophyllea (arbustes des zones humides originaires des régions tropicales) en fonction de l'altitude. On dispose de 15 mesures avec :

- un indicateur de présence (variable $Y \in \{0, 1\}$, avec $Y = 1$ pour présence),
- l'altitude en mètres (variable $X1$).

On souhaite expliquer $Y$ à partir de $X1$.

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude8.txt",
header = T)
head(w)
```

| | X1 | Y |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 4 | 1 |
| 3 | 6 | 1 |
| 4 | 8 | 1 |
| 5 | 10 | 1 |
| 6 | 12 | 0 |

On associe les variables $Y$ et $X1$ aux valeurs associées en faisant :

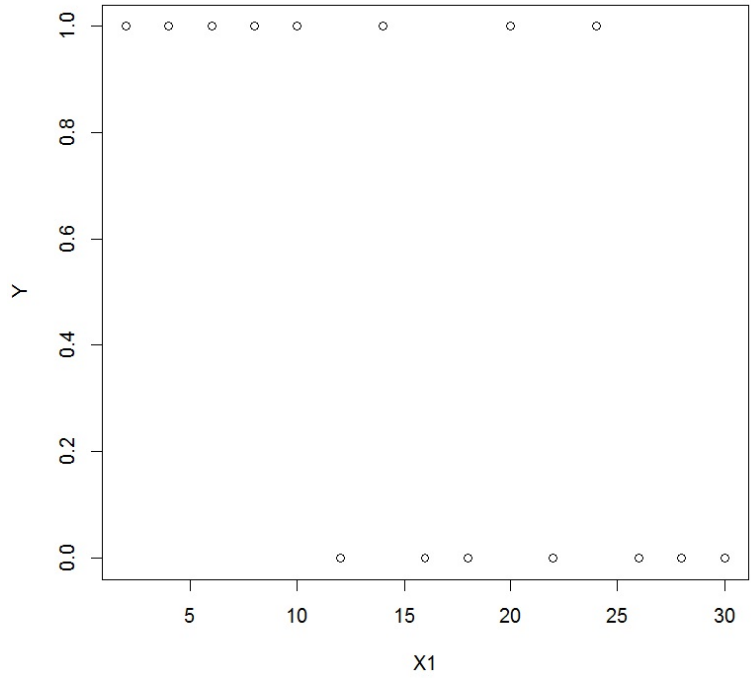
```
attach(w)
```

# Régression logistique simple

## Nuage de points

On trace le nuage de points $\{(x_{1,i}, y_i),\ i \in \{1, \ldots, n\}\}$ :

```
plot(X1, Y)
```




## Modélisation

On veut estimer la probabilité (ou proportion) inconnue

$$p(x) = \mathbb{P}(\{Y = 1\}|\{X1 = x\}),$$

à partir des données.

On adopte le modèle de régression logistique :

$$\text{logit}(p(x)) = \beta_0 + \beta_1 x, \qquad \text{logit}(y) = \ln\left(\frac{y}{1-y}\right),$$

où $\beta_0$ et $\beta_1$ sont des réels inconnus.

Soit encore,

$$p(x) = \text{logit}^{-1}(\beta_0 + \beta_1 x), \qquad \text{logit}^{-1}(y) = \frac{e^y}{1+e^y}.$$

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

## Estimations

La modélisation de la régression logistique simple et les estimations des paramètres par la méthode du *mv* s'obtiennent par les commandes :

```
library(stats)

reg = glm(Y ~ X1, family = binomial)

summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 3.3774 | 1.7643 | 1.91 | 0.0556 | . |
| X1 | -0.1974 | 0.0975 | -2.02 | 0.0430 | * |

```
Null deviance: 20.728 on 14 degrees of freedom

Residual deviance: 14.092 on 13 degrees of freedom

AIC: 18.092

Number of Fisher Scoring iterations: 4
```

– Estimations ponctuelles de $\beta_0$ et $\beta_1$ :

| $\widehat{\beta}_0$ | $\widehat{\beta}_1$ |
|---|---|
| 3.3774 | $-0.1974$ |

– Estimations ponctuelles des écart-types des estimateurs de $\beta_0$ et $\beta_1$ :

| $\widehat{\sigma}(\widehat{\beta}_0)$ | $\widehat{\sigma}(\widehat{\beta}_1)$ |
|---|---|
| 1.7643 | 0.0975 |

– $t_{obs}$ :

| $H_1$ | $\beta_0 \neq 0$ | $\beta_1 \neq 0$ |
|---|---|---|
| $z_{obs}$ | 1.91 | $-2.02$ |

– Test de Wald pour $\beta_1$ : influence de $X1$ sur $Y$ : p-valeur $= 0.0430 \in ]0.01, 0.05[$, $*$ : significative,

– $AIC = 18.092$.

La valeur prédite de $Y$ quand $X1 = 25$ (par exemple) est donnée par les commandes :

```
predict.glm(reg, data.frame(X1 = 25), type = "response")
```

Cela renvoie 0.1739867.

Ainsi, la probabilité de trouver de l'Anisophyllea à 25 mètres d'altitude est de 0.1739867. Il y a donc de fortes chances qu'il n'y ait pas d'Anisophyllea à cette altitude.

On peut aussi s'intéresser :

– au test de la déviance afin de confirmer l'influence de $X1$ sur $Y$ :

```
anova(reg, test = "Chisq")
```

Cela renvoie : p-valeur $= 0.009993 \in ]0.001, 0.01[$ $**$, soit une influence très significative de $X1$ sur $Y$. Ce test utilisant des outils autres que celui de Wald, on ne s'étonnera pas que le résultat diffère un peu.

– à l'estimateur du rapport des côtes associé à $X1$ :

```
exp(coef(reg))
```

Cela renvoie : 0.8208606.

Ainsi, l'augmentation d'une unité de $X_1$ entraîne une augmentation des chances que $\{Y = 0\}$ se réalise.

– aux intervalles de confiance pour $\beta_0$ et $\beta_1$ au niveau 95% (par exemple).

Les commandes sont :

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

Cela renvoie :

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -0.08056611 | 6.835399063 |
| X1 | -0.38855154 | -0.006252325 |

Ainsi,

$$i_{\beta_0} = [-0.08056611,\ 6.835399063], \qquad i_{\beta_1} = [-0.38855154, 0.006252325].$$

– à l'intervalle de confiance pour $p(x)$ au niveau 95% avec $x = 25$ :

```
logitp = predict.glm(reg, data.frame(X1 = 25), se.fit = TRUE)
iclogit = c(logitp$fit - 1.96 * logitp$se.fit,
logitp$fit + 1.96 * logitp$se.fit)
ic = exp(iclogit) / (1 + exp(iclogit))
ic
```

Cela renvoie :

| 1 | 1 |
|---|---|
| 0.02666024 | 0.61828999 |

Ainsi, pour $X1 = 25$, on a

$$i_{p(x)} = [0.02666024,\ 0.61828999].$$

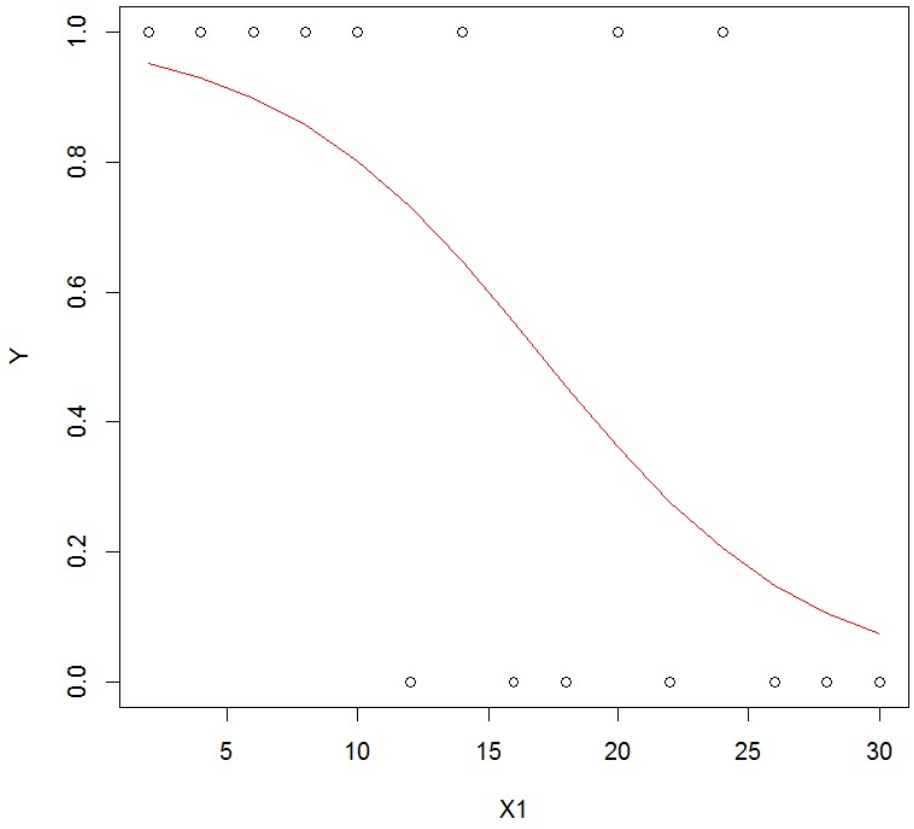
### Ligne logistique

En utilisant les estimations ponctuelles de $\beta_0$ et $\beta_1$, l'estimation ponctuelle de $p(x)$ est

$$\widetilde{p}(x) = \text{logit}^{-1}(3.3774 - 0.1974x).$$

On peut représenter graphiquement la fonction $p(x)$, $x \in [0, 30]$ avec les commandes :

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



## Pertinence du modèle

On peut étudier la pertinence du modèle de régression logistique avec :

– le test de Hosmer-Lemeshow :

```
library(ResourceSelection)

hoslem.test(Y, fitted(reg), g = 10)
```

Cela renvoie : p-valeur= $0.3551 > 0.05$, donc le modèle est bien adapté au problème.

– l'utilisation des résidus de Pearson :

```
s2 = sum(residuals(reg, type = "pearson")^2)

ddl = df.residual(reg)

pvaleur = 1 - pchisq(s2, ddl)

pvaleur
```

Cela renvoie : p-valeur= $0.4926908 > 0.05$, donc le modèle est bien adapté au problème.

Tout indique que notre modèle est compatible avec les données.

## Détection des valeurs anormales

On peut étudier la présence de valeurs anormales en calculant les résidus de Pearson et identifier ceux dont la magnitude dépasse 2 :

```
e = rstandard(reg, type = "pearson")

e[abs(e) > 2]
```

Cela renvoie :

```
12

2.143909
```

Ainsi, d'après ce critère mathématique, les valeurs associées à l'individu 12 sont anormales.

Voyons si cela ressort avec les distances de Cook :

```
cooks.distance(reg)[cooks.distance(reg) > 1]
```

L'individu 12 ressort effectivement mais la distance reste inférieure à 1. Son retrait du jeu de donnée est toutefois envisageable.

## Qualité du modèle

### Taux d'erreur

Afin d'examiner le taux d'erreur du modèle logistique considéré, on calculer les prédictions de groupe :

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

Cela renvoie :

```
1  2  3  4  5  6  7  8  9  10  11  12  13  14  15
1  1  1  1  1  1  1  1  0  0   0   0   0   0   0
```

Pour avoir la matrice de confusion, on exécute :

```
mc = table(Y, pred.mod)
mc
```

Cela renvoie :

$$MC = \begin{pmatrix} 5 & 2 \\ 2 & 6 \end{pmatrix}$$

Le taux d'erreur est donné par :

```
t = (sum(mc) - sum(diag(mc))) / sum(mc)
t
```

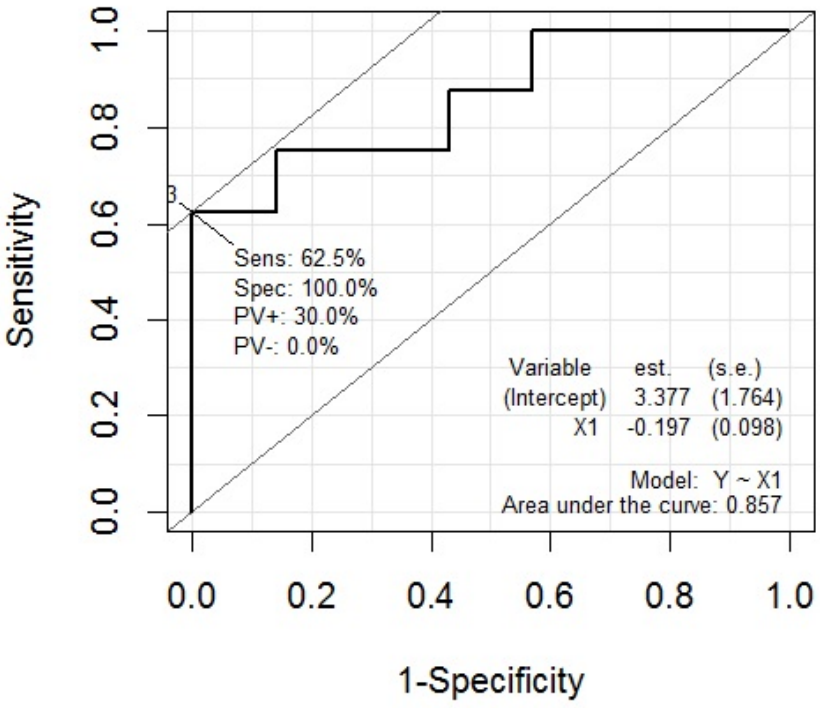
Cela renvoie 0.2666667, soit environ $26,66\%$

La petitesse du taux d'erreur traduit la bonne qualité prédictive du modèle.

## Courbe ROC

On obtient la courbe ROC en faisant :

```
library(Epi)

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




On constate que la courbe est proche des axes $x = 0$ et $y = 1$ ; elle est éloignée de l'axe $x = 1/2$. Cela confirme que le modèle est correct.

L'aide de la courbe ROC est de 0.857.

## Modélisation probit

On peut aussi étudier la modélisation probit :

$$p(x) = \Phi\left(\beta_0 + \beta_1 x\right), \qquad \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

La modélisation de la régression probit et les estimations des paramètres par la méthode du *mv* s'obtiennent par les commandes :

```
library(stats)

reg2 = glm(Y ~ X1, family = binomial(link = "probit"))

summary(reg2)
```

Cela renvoie :

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 2.0959 | 0.9744 | 2.151 | 0.0315 | * |
| X1 | -0.1217 | 0.0534 | -2.279 | 0.0227 | * |

```
Null deviance: 20.728 on 14 degrees of freedom

Residual deviance: 13.927 on 13 degrees of freedom

AIC: 17.927

Number of Fisher Scoring iterations: 6
```

- Estimations ponctuelles de $\beta_0$ et $\beta_1$ : pour $\beta_0$ (Intercept) : 2.0959, ppour $\beta_1$ : $-0.1217$,
- Estimations ponctuelles des écart-types des estimateurs de $\beta_0$ et $\beta_1$ : pour $\beta_0$ : 0.9744, pour $\beta_1$ : 0.0534,
- $z_{obs}$ : pour $\beta_0$ : 2.151, pour $\beta_1$ : $-2.279$,
- Test de Wald pour $\beta_1$ : influence de $X1$ sur $Y$ : p-valeur $= 0.0227 \in ]0.01, 0.05[$, $*$ : significative,
- $AIC = 17.927$.

On remarque que le AIC du nouveau modèle est plus petit que celui du premier, mais c'est tellement infime qu'il est difficile d'en conclure quelque chose.

On peut comparer la significativité des 2 modèles avec les commandes :

```
anova(reg, reg2)
```

Cela renvoie : p-valeur $= 0.16492 > 0.05$.

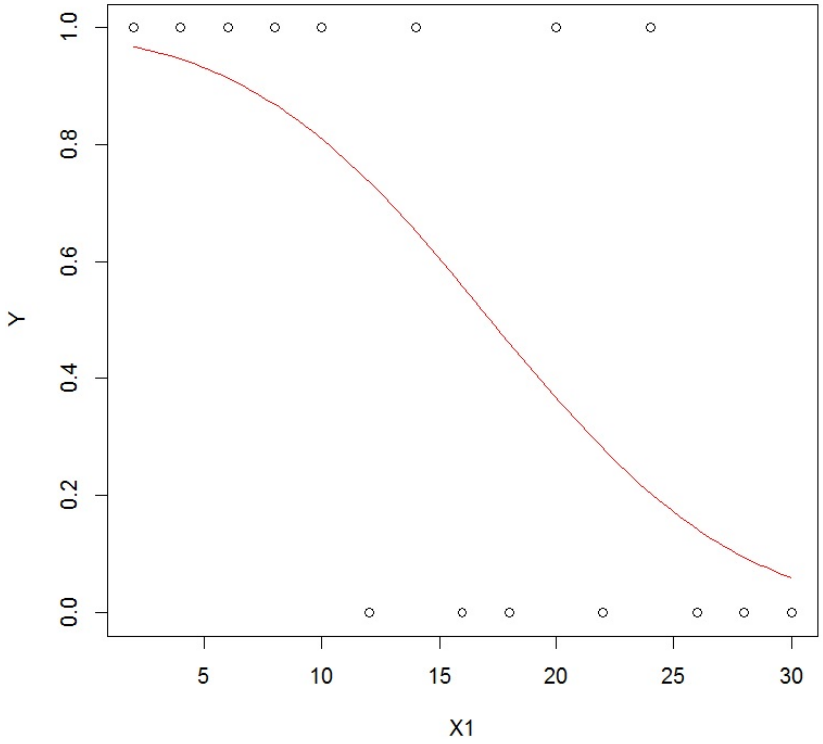
Ainsi, les modèles ne sont pas significativement différents.

L'estimation ponctuelle de $p(x)$ avec cette nouvelle modélisation est

$$\widetilde{p}(x) = \Phi\left(2.0959 - 0.1217x\right)$$

que l'on peut représenter graphiquement avec les commandes :

```
plot(X1, Y)
curve(predict(reg2, data.frame(X1 = x), type = "response"), col = "red",
add = T)
```



# Conclusion et études similaires

## Conclusion

On a construit un modèle de régression logistique satisfaisant pour expliquer $Y$ à partir de $X1$. Une valeur anormale est éventuellement à traiter.

## Étude similaire 1 : Blattes

On peut considérer le jeu de données "blattes" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/blattes.txt",
header = T)
```

Sur 8 groupes contenant un nombre de cafards différent, on administre un insecticide, puis on compte le nombre de cafards morts dans chaque groupe après 5 heures. Ainsi, pour chaque groupe, on dispose :

– du dosage de l'insecticide (variable $X1$),

– du nombre total de cafards morts (variable $Y^*$, laquelle ne prend pas que les valeurs $\{0, 1\}$).

L'objectif est de prédire la proportion moyenne de cafards morts après 5 heures pour une dose d'insecticide fixée.

On remarquera que les données sont groupées. Par conséquent, on utilisera les commandes :

```
reg = glm(cbind(morts, total - morts) ~ X1, family = binomial)
```

## Étude similaire 2 : Puits

On peut considérer le jeu de données "puits" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/puits.txt", header = T)
```

Dans une région du Bangladesh, les puits ont été étiquetés avec le taux en arsenic de l'eau et une indication "safe" ou "unsafe". Les habitants utilisant de l'eau non-potable ont été encouragés à changer de puits. Quelques années plus tard il a été observé que 57.5% des 3020 familles qui consommaient de l'eau non potable avaient changé de puits. Ainsi, pour chaque famille, on dispose :

- de son départ ou non d'un puits "unsafe" vers un puits "safe" (variable $Y \in \{0, 1\}$, avec $Y = 1$ pour départ),
- du niveau de contamination par l'arsenic dans le puits d'origine de la maison, en centaines de microgrammes par litre ; tous ceux au-dessus de 0.5 ont été identifiés comme le niveau "safe" (variable $X1$),
- de la distance en mètres du puits connu "safe" le plus proche (variable $X2$).

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

## Étude similaire 3 : Anesthésie

On peut considérer le jeu de données "anesthésie" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/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 $Y \in \{0, 1\}$, avec $Y = 1$ pour bougé).

On souhaite expliquer $Y$ à partir de $X1$.

## Étude similaire 4 : Prostate

On peut considérer le jeu de données "anesthésie" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/prostate.txt",
header = T)
```

L'étude porte sur 53 patients atteints par un cancer de la prostate. On souhaite expliquer l'atteinte du système lymphatique par ce cancer à l'aide de certaines variables. Ainsi, pour chaque patient, on dispose :

- de l'âge en années (variable $X1$),
- du niveau d'acide phosphatase (variable $X2$),
- du résultat positif ou négatif des rayons (variable $X3 \in \{0,1\}$, avec $X3 = 1$ pour positif),
- de la taille (grande ou pas) de la tumeur (variable $X4 \in \{0,1\}$, avec $X4 = 1$ pour grande),
- du grade tumoral, sérieux ou non, d'après la biopsie (variabe $X5 \in \{0,1\}$, avec $Y = 1$ pour sérieux),
- du fait que le cancer ait atteint le système lymphatique ou pas (variable $Y \in \{0,1\}$, avec $Y = 1$ pour atteint).

On souhaite expliquer $Y$ à partir de $X1$, $X2$, $X3$, $X4$ et $X5$ (on pourra également introduire la variable $X6 = \log(X2)$ et faire une sélection de variables pour optimiser le modèle).

# 9 Étude n° 9 : Marques

## Contexte

On s'intéresse au choix d'une marque de produits parmi 3 proposées en fonction de l'âge et du sexe du client. On dispose de 735 données avec :

- un indicateur de la marque choisie (variable $Y \in \{1, 2, 3\}$),
- un indicateur de sexe du client (variable $X1 \in \{0, 1\}$ avec $X1 = 1$ pour homme et $X1 = 0$ pour femme),
- l'âge du client (variable $X2$).

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

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude9.txt",
header = T)
head(w)
```

| | Y | X1 | X2 |
|---|---|---|---|
| 1 | 1 | 0 | 24 |
| 2 | 1 | 0 | 26 |
| 3 | 1 | 0 | 26 |
| 4 | 1 | 1 | 27 |
| 5 | 1 | 1 | 27 |
| 6 | 3 | 1 | 27 |

On associe les variables $Y$, $X1$ et $X2$ aux valeurs associées en faisant :

```
attach(w)
```

On précise que $X1$ est une variable qualitative (binaire) en faisant :

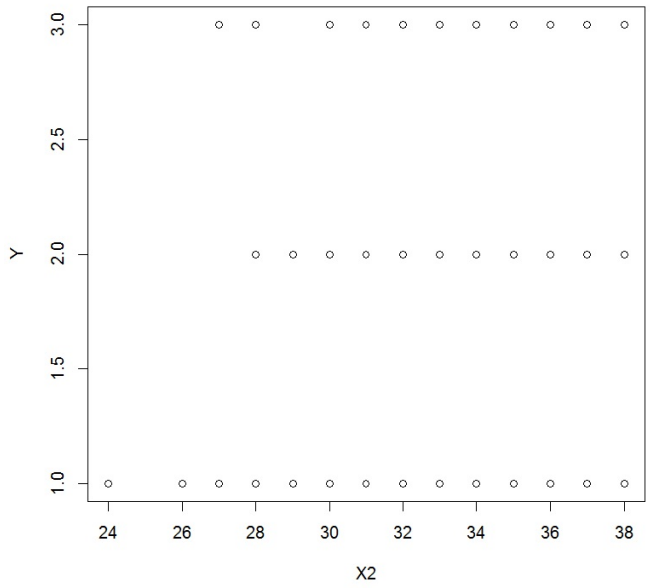
```
X1 = as.factor(X1)
```

# Régression multinomiale (polytomique non-ordonnée)

## Nuage de points

On trace le nuage de points $\{(x_{2,i}, y_i),\ i \in \{1, \ldots, n\}\}$ :

```
plot(X2, Y)
```




## Modélisation

On veut 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 partir des données.

Notons que les modalités sont sans lien hiérarchique/ordre.

On adopte le modèle de régression multinomiale : pour tout $k \in \{2, 3\}$,

$$\ln\left(\frac{p_k(x)}{p_1(x)}\right) = \beta_0^{(k)} + \beta_1^{(k)} x_1 + \beta_2^{(k)} x_2,$$

soit encore,

$$p_k(x) = \frac{\exp(\beta_0^{(k)} + \beta_1^{(k)} x_1 + \beta_2^{(k)} x_2)}{1 + \sum_{k=2}^{3} \exp(\beta_0^{(k)} + \beta_1^{(k)} x_1 + \beta_2^{(k)} x_2)},$$

où $\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)}$ désigne 3 réels inconnus.

Notons que $p_1(x) = 1 - (p_2(x) + p_3(x))$.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

## Estimations

La modélisation de la régression multinomiale se fait par les commandes :

```
library(nnet)

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

On estime les coefficients de régression inconnus en faisant :

```
summary(reg)
```

Cela renvoie :

```
multinom(formula = Y ~ X1 + X2)

Coefficients:

(Intercept) X1           X2

2 -11.77469 0.5238197 0.3682075

3 -22.72141 0.4659488 0.6859087

Std. Errors:

(Intercept) X1          X2

2 1.774614 0.1942467 0.05500320

3 2.058030 0.2260895 0.06262657
```

```
Residual Deviance: 1405.941
```

```
AIC: 1417.941
```

– Estimations ponctuelles de $\beta_0^{(2)}$, $\beta_1^{(2)}$, $\beta_2^{(2)}$, $\beta_0^{(3)}$, $\beta_1^{(3)}$ et $\beta_2^{(3)}$ :

| $\widehat{\beta}_0^{(2)}$ | $\widehat{\beta}_1^{(2)}$ | $\widehat{\beta}_2^{(2)}$ |
|---|---|---|
| $-11.77469$ | 0.5238197 | 0.3682075 |

| $\widehat{\beta}_0^{(3)}$ | $\widehat{\beta}_1^{(3)}$ | $\widehat{\beta}_2^{(3)}$ |
|---|---|---|
| $-22.72141$ | 0.4659488 | 0.6859087 |

– Estimations ponctuelles des écart-types des estimateurs de $\beta_0^{(2)}$, $\beta_1^{(2)}$, $\beta_2^{(2)}$, $\beta_0^{(3)}$, $\beta_1^{(3)}$ et $\beta_2^{(3)}$ :

| $\widehat{\sigma}(\widehat{\beta}_0^{(2)})$ | $\widehat{\sigma}(\widehat{\beta}_1^{(2)})$ | $\widehat{\sigma}(\widehat{\beta}_2^{(2)})$ |
|---|---|---|
| 1.774614 | 0.1942467 | 0.05500320 |

| $\widehat{\sigma}(\widehat{\beta}_0^{(3)})$ | $\widehat{\sigma}(\widehat{\beta}_1^{(3)})$ | $\widehat{\sigma}(\widehat{\beta}_2^{(3)})$ |
|---|---|---|
| 2.058030 | 0.2260895 | 0.06262657 |

– Déviance : 1405.941,

– AIC : 1417.941.

On a donc une estimation de $p_k(x)$ :

$$\widehat{p}_k(x) = \frac{\exp(\widehat{\beta}_0^{(k)} + \widehat{\beta}_1^{(k)} x_1 + \widehat{\beta}_2^{(k)} x_2)}{1 + \sum_{k=2}^{3} \exp(\widehat{\beta}_0^{(k)} + \widehat{\beta}_1^{(k)} x_1 + \widehat{\beta}_2^{(k)} x_2)}.$$

La probabilité qu'un client vérifiant $(X1, X2) = (1, 41) = x$ satisfait $Y = u_k$ pour $k \in \{1, 2, 3\}$ est :

```
predict(reg, data.frame(X1 = 1, X2 = 41), type = "probs")
```

Cela renvoie :

| $\widehat{p}_1(x)$ | $\widehat{p}_2(x)$ | $\widehat{p}_3(x)$ |
|---|---|---|
| 0.002494372 | 0.116707528 | 0.880798101 |

On constate que la probabilité $p_3(x)$ est la plus forte. La modalité la plus probable pour le client considéré est donc $u_3 = 3$.

Cela se confirme avec les commandes :

```
predict(reg, data.frame(X1 = 1, X2 = 41), type = "class")
```

Cela renvoie 3.

On peut aussi déterminer les intervalles de confiance pour $\beta_0^{(2)}$, $\beta_1^{(2)}$, $\beta_2^{(2)}$, $\beta_0^{(3)}$, $\beta_1^{(3)}$ et $\beta_2^{(3)}$, au niveau 95% (par exemple).

Les commandes sont :

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

Cela renvoie :

```
, , 2

2.5 % 97.5 %

(Intercept) -15.2528706 -8.2965119

X1 0.1431032 0.9045363

X2 0.2604032 0.4760118

, , 3

2.5 % 97.5 %

(Intercept) -26.75507739 -18.6877495

X1 0.02282138 0.9090761

X2 0.56316287 0.8086545
```

Ainsi :

| $i_{\beta_0^{(2)}}$ | $i_{\beta_1^{(2)}}$ | $i_{\beta_2^{(2)}}$ |
|---|---|---|
| $[-15.2528706,\ -8.2965119]$ | $[0.1431032,\ 0.9045363]$ | $[0.2604032,\ 0.4760118]$ |

| $i_{\beta_0^{(3)}}$ | $i_{\beta_1^{(3)}}$ | $i_{\beta_2^{(3)}}$ |
|---|---|---|
| $[-26.75507739,\ -18.6877495]$ | $[0.02282138,\ 0.9090761]$ | $[0.56316287,\ 0.8086545]$ |

### Influence des variables explicatives

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 3.243428e-11 0.007003615 2.1672e-11
3 0.000000e+00 0.039312243 0.0000e+00
```

Ainsi :

| $H_1$ | $\beta_0^{(2)} \neq 0$ | $\beta_1^{(2)} \neq 0$ | $\beta_2^{(2)} \neq 0$ |
|---|---|---|---|
| p-valeur | $3.243428e-11$ | $0.007003615$ | $2.1672e-11$ |
| degré | $***$ | $**$ | $***$ |

| $H_1$ | $\beta_0^{(3)} \neq 0$ | $\beta_1^{(3)} \neq 0$ | $\beta_2^{(3)} \neq 0$ |
|---|---|---|---|
| p-valeur | $0.000000e+00$ | $0.039312243$ | $0.0000e+00$ |
| degré | $***$ | $*$ | $***$ |

On peut aussi faire le test statistique global. On considère les hypothèses :

$$H_0 : \beta_1^{(k)} = \beta_2^{(k)} = 0 \text{ pour tout } k \in \{2, 3\} \qquad \text{contre}$$

$$H_1 : \text{il y a au moins un coefficient non nul.}$$

On le met en oeuvre en faisant :

```
reg0 = multinom(Y ~ 1)
rv = reg0$deviance - reg$deviance
ddl = reg$edf - reg0$edf
pvaleur = 1 - pchisq(rv, ddl)
pvaleur
```

Cela renvoie : p-valeur $\simeq 0$. Ainsi, la régression multinomiale est hautement significative.

## Qualité du modèle

On détermine la matrice de confusion du modèle :

```
pr = predict(reg)

mc = table(Y, pr)

mc
```

Cela renvoie la matrice :

$$MC = \begin{pmatrix} 58 & 136 & 13 \\ 18 & 238 & 51 \\ 10 & 101 & 110 \end{pmatrix}$$

On détermine le taux d'erreur :

```
t = (sum(mc) - sum(diag(mc))) / sum(mc)

t
```

Cela renvoie 0.447619.

Même si ce taux est inférieur à 0.5, il est quand même assez élevé.

Disons que le modèle est d'une qualité prédictive moyenne.

# Conclusion et étude similaire

## Conclusion

On a construit un modèle de régression multinomiale pour expliquer $Y$ à partir de $X1$ et $X2$. Toutefois, celui-ci est clairement améliorable, peut-être en prenant en compte des variables supplémentaires comme la catégorie socio-professionnelle des clients par exemple.

## Étude similaire : Iris

On peut considérer le jeu de données "iris".

Pour 3 variétés d'iris :

Setosa, Versicolor , Virginica,

et pour 150 iris par variété, on considère les variables :

- la longueur d'un pétale (variable $X1$),
- la largeur d'un pétale (variable $X2$),
- la longueur d'un sépale (variable $X3$),
- la largeur d'un sépale (variable $X4$).

On cherche à expliquer l'espèce d'iris : $\{setosa, versicolor, virginica\}$ (variable $Y$ = Species) à partir de $X1$, $X2$, $X3$ et $X4$.

Ainsi, la problématique est : à partir des données recueillies, peut-on prévoir la variété d'un iris choisit au hasard uniquement à partir des mesures effectuées sur $X1$, $X2$, $X3$ et $X4$ ?

Voici les commandes utiles pour appréhender ce jeu de données :

On affiche le haut du jeu de données :

```
head(iris)
```

Un petit résumer de plusieurs aspects :

```
str(iris)
```

Nous allons mettre en œuvre une régression multinomiale :

```
library(nnet)
```

```
reg = multinom(Species ~ ., data = iris)
```

On affiche les estimations ponctuelles :

```
summary(reg)
```

On détermine la matrice de confusion :

```
pr = predict(reg, iris)
```

```
mc = table(iris$Species, pr)
```

```
mc
```

On détermine le taux d'erreur :

```
t = (sum(mc) - sum(diag(mc))) / sum(mc)
```

```
t
```

# 10 Étude n° 10 : Nageurs

## Contexte

On s'intéresse au nombre d'infections de l'oreille de jeunes nageurs. Pour 187 nageurs, on dispose de :

- un indicateur d'assiduité à la nage (variable $X1 \in \{Freq, Occas\}$),
- un indicateur sur le lieu habituel de nage (variable $X2 \in \{Beach, NonBeach\}$),
- un indicateur sur la tranche d'âge (variable $X3 \in \{[15, 19], [20, 24], [25, 29]\}$)
- un indicateur sur le sexe (variable $X4 \in \{Female, Male\}$),
- le nombre d'infections de l'oreille (variable $Y \in \mathbb{N}$).

On souhaite expliquer $Y$ à partir de $X1$, $X2$, $X3$ et $X4$.

Les données sont disponibles ici :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/Etude10.txt",
header = T)
head(w)
```

| | X1 | X2 | X3 | X4 | Y |
|---|---|---|---|---|---|
| 1 | Occas | NonBeach | 15-19 | Male | 0 |
| 2 | Occas | NonBeach | 15-19 | Male | 0 |
| 3 | Occas | NonBeach | 15-19 | Male | 0 |
| 4 | Occas | NonBeach | 15-19 | Male | 0 |
| 5 | Occas | NonBeach | 15-19 | Male | 0 |
| 6 | Occas | NonBeach | 15-19 | Male | 0 |

On associe les variables $Y$ et $X1$, $X2$, $X3$ et $X4$ aux valeurs associées en faisant :

```
attach(w)
```

# Régression de Poisson

## Modélisation

Les variables $X1$, $X2$, $X3$ et $X4$ sont qualitatives. On définit donc de nouvelles variables correspondants à chacune des modalités :

- pour $X1$ : $X1Occas$ et $X1Freq$,
- pour $X2$ : $X2Beach$ et $X2NonBeach$,
- pour $X3$ : $X315-19$, $X320-24$ et $X325-29$,
- pour $X4$ : $X4Female$ et $X4Male$,

avec, par exemple,

$$X1Occas = \mathbf{1}_{\{X1=Occas\}}$$

qui vaut 1 si $X1$ est égale à la modalité $Occas$, et 0 sinon.

On en ignore une pour chaque $Xi$, $i \in \{1,2,3,4\}$, pour ne pas créer de la dépendance inutile. Il reste donc 5 variables au total :

$$X1Occas = \mathbf{1}_{\{X1=Occas\}}, \qquad X2NonBeach = \mathbf{1}_{\{X2=NonBeach\}}, \qquad X320-24 = \mathbf{1}_{\{X3=20-24\}}$$

$$X325-29 = \mathbf{1}_{\{X3=25-29\}}, \qquad X4Male = \mathbf{1}_{\{X4=Male\}}.$$

Sachant que

$$\{(X1Occas, X2NonBeach, X320-24, X325-29, X4Male) = (x_1, x_2, x_3, x_4, x_5) = x\}$$

comme $Y$ est une variable de comptage à valeurs entières, on peut supposer que $Y \sim \mathcal{P}(\lambda(x))$, où $\lambda(x)$ désigne le nombre moyen inconnu d'infections.

On veut estimer $\lambda(x)$ à partir des données.

Posons

| $u_0$ | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $u_6$ | $u_7$ | $u_8$ | $u_9$ | $u_{10}$ | $u_{11}$ | $u_{12}$ | $u_{13}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_1x_2$ | $x_1x_3$ | $x_1x_4$ | $x_1x_5$ | $x_2x_3$ | $x_2x_4$ | $x_2x_5$ | $x_3x_5$ |

| $u_{14}$ | $u_{15}$ | $u_{16}$ | $u_{17}$ | $u_{18}$ | $u_{19}$ | $u_{20}$ | $u_{21}$ | $u_{22}$ | $u_{23}$ |
|---|---|---|---|---|---|---|---|---|---|
| $x_4x_5$ | $x_1x_2x_3$ | $x_1x_2x_4$ | $x_1x_2x_5$ | $x_1x_3x_5$ | $x_1x_4x_5$ | $x_2x_3x_5$ | $x_2x_4x_5$ | $x_1x_2x_3x_5$ | $x_1x_2x_4x_5$ |

On adopte le modèle de régression de Poisson (avec interactions car on travaille avec des variables qualitatives) :

$$\ln(\lambda(x)) = \sum_{i=0}^{23} \beta_i u_i$$

où et $\beta_0, \ldots, \beta_{23}$ sont des réels inconnus.

Objectifs : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

## Estimations

La modélisation de la régression de Poisson (avec interactions) et les estimations des paramètres par la méthode du *mv* s'obtiennent par les commandes :

```
library(stats)
reg = glm(Y ~ X1 * X2 * X3 * X4, family = poisson)
summary(reg)
```

Cela renvoie :

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 0.3567 | 0.2236 | 1.60 | 0.1107 | |
| X1Occas | 0.1133 | 0.3028 | 0.37 | 0.7082 | |
| X2NonBeach | 0.0488 | 0.4655 | 0.10 | 0.9165 | |
| X320-24 | -1.0498 | 0.5000 | -2.10 | 0.0358 | * |
| X325-29 | -0.3567 | 0.4655 | -0.77 | 0.4435 | |
| X4Male | -0.5798 | 0.3354 | -1.73 | 0.0839 | . |
| X1Occas:X2NonBeach | 0.7340 | 0.5742 | 1.28 | 0.2012 | |
| X1Occas:X320-24 | 0.3285 | 0.6592 | 0.50 | 0.6182 | |
| X1Occas:X325-29 | 0.7517 | 0.5576 | 1.35 | 0.1777 | |
| X2NonBeach:X320-24 | 0.5390 | 0.7265 | 0.74 | 0.4581 | |
| X2NonBeach:X325-29 | -1.4351 | 1.1762 | -1.22 | 0.2224 | |
| X1Occas:X4Male | 0.5153 | 0.4330 | 1.19 | 0.2341 | |
| X2NonBeach:X4Male | 0.5153 | 0.5490 | 0.94 | 0.3479 | |
| X320-24:X4Male | 1.2730 | 0.6748 | 1.89 | 0.0592 | . |
| X325-29:X4Male | -0.5188 | 0.6922 | -0.75 | 0.4536 | |
| X1Occas:X2NonBeach:X320-24 | -0.6957 | 0.9199 | -0.76 | 0.4495 | |
| X1Occas:X2NonBeach:X325-29 | -0.2127 | 1.3413 | -0.16 | 0.8740 | |
| X1Occas:X2NonBeach:X4Male | -0.9180 | 0.6801 | -1.35 | 0.1771 | |
| X1Occas:X320-24:X4Male | -2.9030 | 1.2935 | -2.24 | 0.0248 | * |
| X1Occas:X325-29:X4Male | -0.3462 | 0.8206 | -0.42 | 0.6731 | |
| X2NonBeach:X320-24:X4Male | -1.1831 | 0.9158 | -1.29 | 0.1964 | |
| X2NonBeach:X325-29:X4Male | 1.9696 | 1.3390 | 1.47 | 0.1413 | |
| X1Occas:X2NonBeach:X320-24:X4Male | 3.9762 | 1.4946 | 2.66 | 0.0078 | ** |
| X1Occas:X2NonBeach:X325-29:X4Male | 0.6356 | 1.5374 | 0.41 | 0.6793 | |

```
Null deviance: 824.51 on 286 degrees of freedom

Residual deviance: 703.72 on 263 degrees of freedom

AIC: 1124.1

Number of Fisher Scoring iterations: 6
```

On a les estimations ponctuelles des coefficients $\beta_0, \ldots, \beta_{23}$, ainsi que les degrés de significativité des variables considérées sur $Y$.

La valeur prédite de $Y$ quand $X1 = Occas$, $X2 =$ "$NonBeach$", $X3 =$ "$25-29$",

$X4 =$ "$Male$" (par exemple) est donnée par :

```
predict.glm(reg, data.frame(X1 = "Occas", X2 = "NonBeach", X3 = "25-29",

X4 = "Male"), type = "response")
```

Cela renvoie 3.571429.

Ainsi, le nombre moyen d'infections pour un individu vérifiant les modalités précédentes est de 3.571429.

La probabilité que cet individu satisfait $Y = 5$, i.e.,

$$\widehat{p}_5(x) = \exp(-\widehat{\lambda}(x))\frac{(\widehat{\lambda}(x))^5}{5!},$$

est donnée par :

```
probs = dpois(5, 3.571429)
probs
```

Cela renvoie 0.1361372.

On peut aussi s'intéresser à la significativité de la régression : On considère les hypothèses :

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{23} = 0 \qquad \text{contre} \qquad H_1 : \text{il y a au moins un coefficient non nul.}$$

On le met en œuvre avec les commandes :

```
reg = glm(Y ~ X1 * X2 * X3 * X4, family = poisson)
reg0 = glm(Y ~ 1, family = poisson)
anova(reg0, reg, test = "Chisq")
```

Cela renvoie : p-valeur $= 3e-15 * ** < 0.001$. La régression est donc hautement significative.

On peut s'intéresser :

– aux intervalles de confiance pour $\beta_0, \ldots, \beta_{23}$ au niveau 95% (par exemple) :

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

– à l'intervalle de confiance pour $\lambda(x)$ au niveau 95% avec la valeur de $x$ précédente :

```
loglamb = predict.glm(reg, data.frame(X1 = "Occas", X2 = "NonBeach",
X3 = "25-29", X4 = "Male"), se.fit = TRUE)
icloglamb = c(loglamb$fit - 1.96 * loglamb$se.fit,
loglamb$fit + 1.96 * loglamb$se.fit)
ic = exp(icloglamb)
ic
```

## Pertinence du modèle

On peut étudier la pertinence du modèle de régression de Poisson avec :

– le test de Hosmer-Lemeshow :

```
library(ResourceSelection)

hoslem.test(Y, fitted(reg), g = 10)
```

Cela renvoie : p-valeur= $1 > 0.05$, donc le modèle est bien adapté au problème.

– l'utilisation des résidus de Pearson :

```
s2 = sum(residuals(reg, type = "pearson")^2)

ddl = df.residual(reg)

pvaleur = 1 - pchisq(s2, ddl)

pvaleur
```

Cela renvoie : p-valeur$\simeq 0 < 0.05$, ce qui contredit la conclusion du test de Hosmer-Lemeshow.

– La règle du pouce donne :

$$\frac{D}{n-(p+1)} = \frac{703.72}{263} = 2.675741,$$

lequel est bien plus grand que 1.

Il y a manifestement un problème de modélisation.

## Détection des valeurs anormales

On peut étudier la présence de valeurs anormales en calculant les résidus de Pearson et identifier ceux dont la magnitude dépasse 2 :

```
e = rstandard(reg, type = "pearson")

sum(abs(e) > 2)
```

Cela renvoie :

```
35
```

Ainsi, d'après ce critère mathématique, il y a 35 valeurs anormales.

Voyons si cela ressort avec les distances de Cook :

```
cooks.distance(reg)[cooks.distance(reg) > 1]
```

Cela renvoie 0.

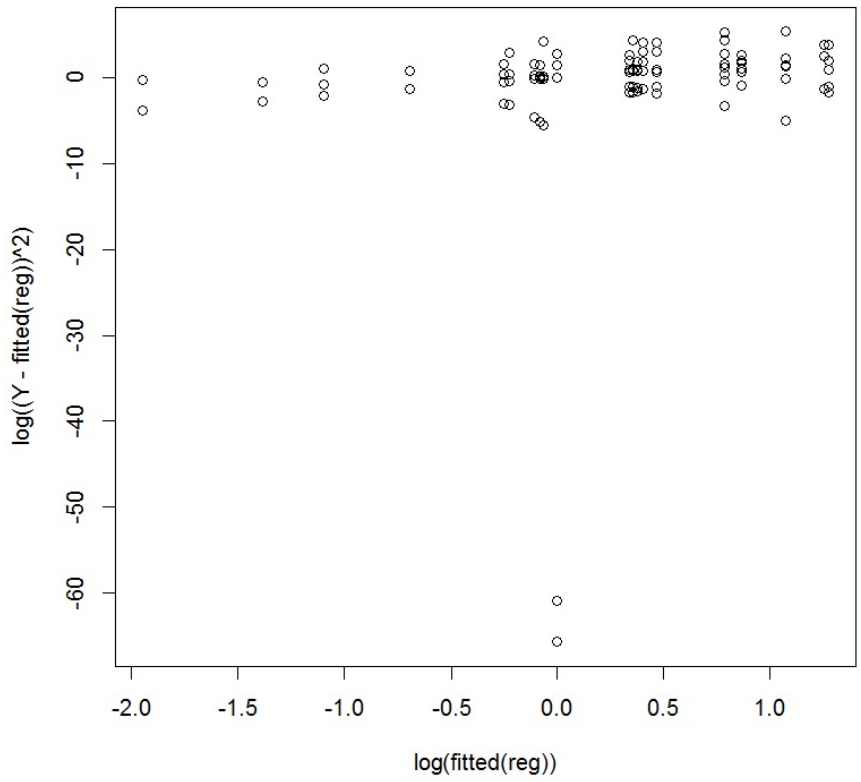
Les résultats semblent contradictoires, peut-être à cause d'une dispersion anormale des valeurs.

## Étude de la dispersion

On peut peut visualiser l'éventuelle anormalité de la dispersion en faisant :

```
plot(log(fitted(reg)), log((Y - fitted(reg))^2))
```

Cela renvoie :



Le nuage de points est difficilement ajustable par une droite, notamment à cause de plusieurs points excentrés.

On peut également mettre en relief cette dispersion anormale avec le test de Cameron et Trivedi :

```
library(AER)

dispersiontest(reg)
```

Cela renvoie : p-valeur $= 3.376e-05 < 0.05$. L'anormalité de la dispersion est confirmée.

On peut corriger cela en injectant son estimation dans le modèle :

```
phi = sum(residuals(reg, type = "pearson")^2) / df.residual(reg)

summary(reg, dispersion = phi)
```

Cela renvoie :

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.3567 | 0.3976 | 0.90 | 0.3697 |
| X1Occas | 0.1133 | 0.5384 | 0.21 | 0.8333 |
| X2NonBeach | 0.0488 | 0.8277 | 0.06 | 0.9530 |
| X320-24 | -1.0498 | 0.8891 | -1.18 | 0.2377 |
| X325-29 | -0.3567 | 0.8277 | -0.43 | 0.6665 |
| X4Male | -0.5798 | 0.5964 | -0.97 | 0.3310 |
| X1Occas:X2NonBeach | 0.7340 | 1.0211 | 0.72 | 0.4723 |
| X1Occas:X320-24 | 0.3285 | 1.1722 | 0.28 | 0.7793 |
| X1Occas:X325-29 | 0.7517 | 0.9916 | 0.76 | 0.4484 |
| X2NonBeach:X320-24 | 0.5390 | 1.2918 | 0.42 | 0.6765 |
| X2NonBeach:X325-29 | -1.4351 | 2.0914 | -0.69 | 0.4926 |
| X1Occas:X4Male | 0.5153 | 0.7700 | 0.67 | 0.5034 |
| X2NonBeach:X4Male | 0.5153 | 0.9762 | 0.53 | 0.5976 |
| X320-24:X4Male | 1.2730 | 1.1999 | 1.06 | 0.2888 |
| X325-29:X4Male | -0.5188 | 1.2309 | -0.42 | 0.6734 |
| X1Occas:X2NonBeach:X320-24 | -0.6957 | 1.6358 | -0.43 | 0.6706 |
| X1Occas:X2NonBeach:X325-29 | -0.2127 | 2.3851 | -0.09 | 0.9289 |
| X1Occas:X2NonBeach:X4Male | -0.9180 | 1.2093 | -0.76 | 0.4478 |
| X1Occas:X320-24:X4Male | -2.9030 | 2.3002 | -1.26 | 0.2069 |
| X1Occas:X325-29:X4Male | -0.3462 | 1.4593 | -0.24 | 0.8125 |
| X2NonBeach:X320-24:X4Male | -1.1831 | 1.6285 | -0.73 | 0.4675 |
| X2NonBeach:X325-29:X4Male | 1.9696 | 2.3811 | 0.83 | 0.4081 |
| X1Occas:X2NonBeach:X320-24:X4Male | 3.9762 | 2.6577 | 1.50 | 0.1346 |
| X1Occas:X2NonBeach:X325-29:X4Male | 0.6356 | 2.7338 | 0.23 | 0.8162 |

Les estimations ponctuelles ne changent pas.

En revanche, la significativité des variables explicatives sur $Y$ change.

Après cette correction, on constate qu'aucune variable n'est significative.

# Conclusion et études similaires

## Conclusion

On a construit un modèle de régression de Poisson pour traiter le problème. Toutefois, plusieurs indicateurs montrent qu'il n'est pas très bien adapté au problème, même si une correction de dispersion améliore certains aspects.

## Étude similaire 1

On peut considérer le jeu de données "clients" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/clients.txt",
header = T)
```

Pendant 2 semaines, une étude de marché compte les clients d'un magasin en fonction de leur provenance de 110 secteurs de recensement (ces secteurs sont des régions métropolitaines ayant une population d'environ 4000 habitants chacune). Diverses caractéristiques démographiques ont également été obtenues. Pour chaque secteur, on dispose :

- du nombre de clients visitant le magasin (variable $Y$),
- du nombre de logement (variable $X1$),
- du revenu personnel moyen annuel (variable $X2$, en dollars),
- de l'âge moyen (variable $X3$),
- de la distance du magasin concurrent le plus proche du secteur (variable $X4$, en miles),
- de la distance du secteur au magasin (variable $X5$, en miles),

On souhaite expliquer $Y$ à partir de $X1$, $X2$ $X3$, $X4$ et $X5$.

## Étude similaire 2

On peut considérer le jeu de données "circuits" :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/circuits.txt",
header = T)
```

L'étude s'intéresse aux nombre de défauts dans la fabrication de circuits imprimés par 2 procédés différents : $A$ et $B$. Pour 20 circuits, on dispose :

– du nombre de défauts (variable $Y$),

– du procédé utilisé (variable $X1$ avec $X1 \in \{A, B\}$).

On souhaite expliquer $Y$ à partir de $X1$.