Devoir 4 - Solutions

Patrick Boily

2023-02-25

Préliminaires 1

Nous importons l'ensemble `Autos.xlsx` se retrouvant sur Brightspace. Nous ne nous intéressons qu'aux véhicules de type VPAS, avec prédicteurs `VKM.q` (X_1, distance quotidienne moyenne, en km) et `Age` (X_2, age du véhicule, en années), et réponse `CC.q` (Y, consommation de carburant quotidienne moyenne, en L).

```
library(tidyverse)  # pour avoir acces a select() et |>
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
Autos <- readxl::read_excel("Data/Autos.xlsx") |>
  filter(Type == "VPAS") |> select(VKM.q,Age,CC.q)
str(Autos)
```

```
## tibble [494 x 3] (S3: tbl_df/tbl/data.frame)
##  $ VKM.q: num [1:494] 208 196 173 169 165 161 154 153 151 147 ...
##  $ Age  : num [1:494] 6 9 7 5 0 20 18 11 4 1 ...
##  $ CC.q : num [1:494] 19 19 14 15 18 14 16 13 14 13 ...
```

```
x1 = Autos$VKM.q
x2 = Autos$Age
y = Autos$CC.q
```

Q31

Calculez directement le coefficient de détermination multiple et le coefficient de détermination multiple ajusté (sans utiliser `lm()`). Qu'est-ce que ces valeurs vous disent au sujet de la qualité de l'ajustement multiple dans l'ensemble de données?

Solution: on commence par trouver le vecteur des valeurs ajustées $\hat{\mathbf{Y}}$ (sans utiliser `lm()`, comme le demande la question).

```
n = nrow(Autos)
p = 2
X = cbind(rep(1,n), x1, x2)
(b = solve(t(X)%*%X) %*% t(X) %*% y)
```

```
##            [,1]
##    -0.014050253
## x1  0.095157626
## x2  0.007384133
```

```
y.hat = X %*% b
```

Si $\beta_0 \neq 0$, le coefficient de détermination R^2 est $r^2_{Y,\hat{Y}}$, le carré de la corrélation de Pearson entre les valeurs réelles de la réponse et les valeurs ajustées de cette dernière.

L'estimateur b_0 n'est pas nul, mais cela ne revient pas nécessairement à dire que $\beta_0 \neq 0$. L'erreur-type $s\{b_0\}$ est donnée par:

```
e = y - y.hat
(SSE = sum(e^2))
```

```
## [1] 2120.459
```

```
MSE = SSE/(n-p)
sigma.b = MSE * solve(t(X) %*% X)
```

L'intervalle de confiance de β_0 à environ 95% est ainsi

```
c(b[1] - qt(1-0.05/2,n-p)*sqrt(sigma.b[1,1]), b[1] + qt(1-0.05/2,n-p)*sqrt(sigma.b[1,1]))
```

```
## [1] -0.4207085  0.3926080
```

Alors on ne sait pas vraiment si $\beta_0 \neq 0$ et il faut utiliser une autre approche; pourquoi ne pas se servir de la définition: $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$?

Nous avons déjà calculé SSE; voici SST:

```
(SST = sum((y-mean(y))^2))
```

```
## [1] 8632.024
```

d'où

```
(R.2 = 1 - SSE/SST)
```

```
## [1] 0.7543498
```

Le coefficient de détermination multiple, quant à lui, est:

```
(R.2.a = 1 - (n-1)/(n-p)*SSE/SST)
```

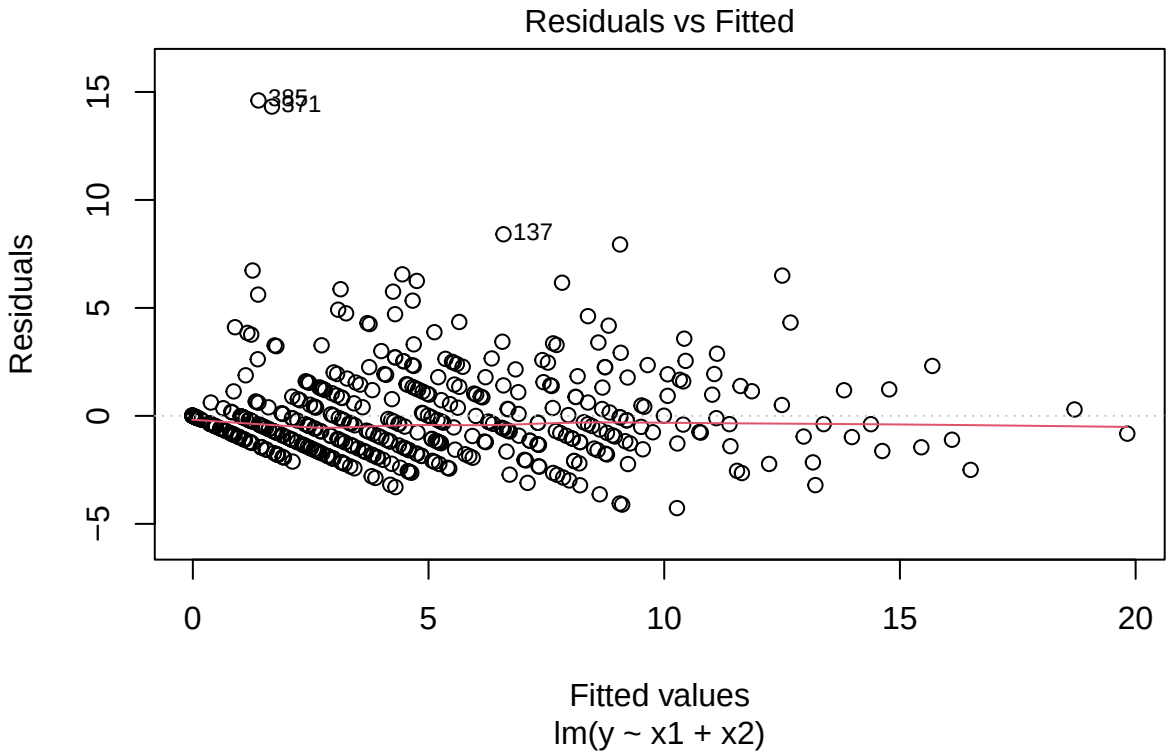
```
## [1] 0.7538505
```

Q32

Est-ce que l'hypothèse de linéarité est raisonable? Justifiez votre réponse.

Solution: on peut bien commencer par visualiser les résidus et les valeurs ajustées.

```
mod = lm(y ~ x1 + x2)
plot(mod, which=1)
```



Ouais, à vue d'oeil, la linéarité ne semble pas garantie (il y a définitivement une tendance dans les résidus). Nous allons utiliser le test RESET de Ramsey afin de tester la spécification linéaire de la fonction de la moyenne.

```
library(lmtest)
resettest(mod, powers=c(2,3))
```

```
##
##  RESET test
##
## data:  mod
## RESET = 0.19966, df1 = 2, df2 = 489, p-value = 0.8191
```

Comme la valeur p du test est assez élevée, on ne peut donc pas rejeter l'hypothèse nulle que le modèle est mal spécifiée selon le test RESET de Ramsey (!).

Mais ce n'est pas la seule façon de s'y prendre; par exemple, avec le test d'inadéquation vu dans les notes, on viendrait à en conclure que la réponse moyenne n'est pas une combinaison linéaire des prédicteurs (mais à peine) – en pratique, c'est plus ou moins toujours comme cela que ça se déroule... les résultats sont assez rarement catégoriques. Tant que vous justifiez votre réponse.

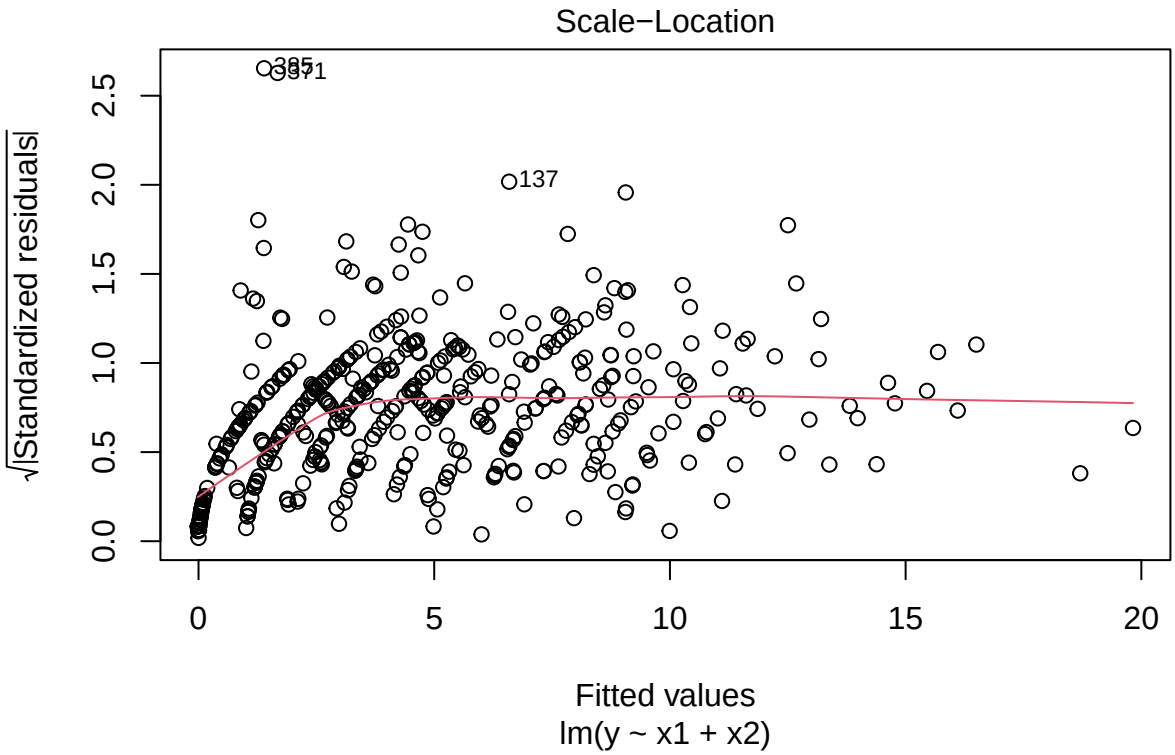
Q33

Est-ce que l'hypothèse de variance constante est raisonable? Justifiez votre réponse.

Solution: comme le modèle linéaire est présumément bien spécifié, on peut vérifier l'homoscédasticité du modèle (la variance de l'erreur ne dépend pas des prédicteurs).

On commence avec le diagramme des racines des résidus standardisés en fonction des valeurs ajustées. Si le modèle est homoscédastique, on devrait s'attendre à ce qu'il y ait une tendance horizontale (près de 1) dans le diagramme. Mais, s'il y a une tendance prononcée dans le diagramme, ceci suggère que la variance n'est pas constante.

```
plot(mod, which=3)
```



Mmmhhh... ce n'est pas diable. Utilisons le test Breusch-Pagan Studentizé, mettons (ce n'est pas la seule option).

```
e <- mod$residuals
# le test de Breusch-Pagan
mod.BP <- lm(e^2 ~ x1 + x2)
R.2 <- summary(mod.BP)$r.squared
# valeur observée de la statistique de test de Breush-Pagan Studentizé
n*R.2
```

```
## [1] 0.748989
```

Mais la valeur p du test BP est:

```
1 - pchisq(n*R.2,p-1)
```

```
## [1] 0.3867965
```

La valeur p est assez élevée pour que nous ne puissions pas rejeter l'hypothèse d'homoscédasticité (valeur constante... en venons nous à la même conclusion avec le test de Brown-Forsythe ou le test de White?)

Q34

Est-ce que l'hypothèse de l'indépendance des termes d'erreur est raisonable? Justifiez votre réponse.

Solution: à peu près le seul truc que l'on peut s'imaginer utiliser ici, c'est la corrélation entre les résidus e_i et les valeurs ajustées $\hat{y}_i$.

Nous avons:

```
cor(e,y.hat)
```

```
## [1] -8.953371e-17
```

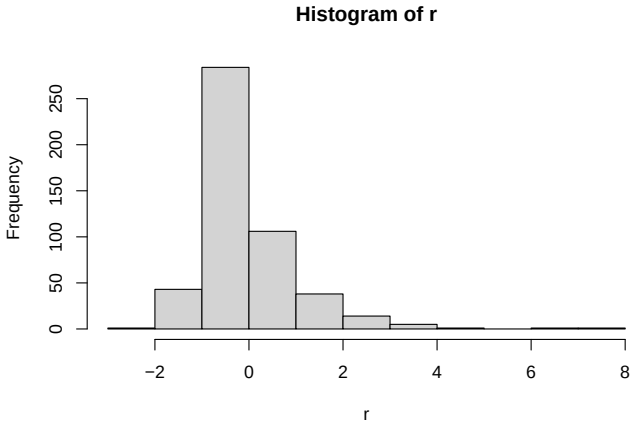
La corrélation est vraiment faible, alors nous n'avons pas à nous inquiéter outre mesure: les termes d'erreurs sont sans doute indépendents.

Q35

Est-ce que l'hypothèse de la normalité des termes d'erreur est raisonable? Justifiez votre réponse.

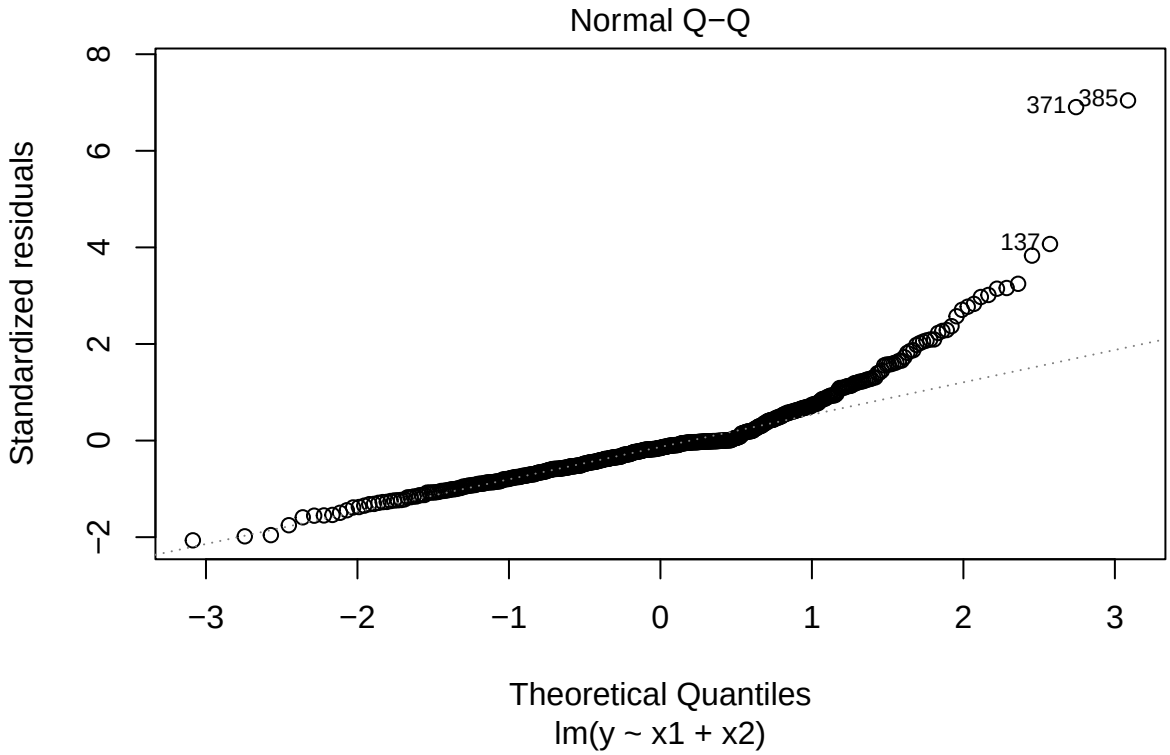
Solution: on commence par un tracé de l'histogramme des résidus studentisés:

```
H = X %*% solve(t(X) %*% X) %*% t(X)
h = diag(H)
r = e/sqrt(MSE*(1-h))
hist(r)
```



L'histogramme a une longue queue vers la droite, mais ce ne sont peut-être que des valeurs aberrantes. Cela vaut la peine de tracer le diagrame quantile-quantile.

```
plot(mod, which=2)
```



La queue vers la droite semble effectivement être problématique. Suffisamment, en fait, pour en conclure que les termes d'erreurs ne suivent pas une loi normale.

Q36

Dans son ensemble, est-ce que vous croyez que le modèle de régression linéaire multiple est approprié? Justifiez votre réponse.

Solution: nous avons vu que nous ne pouvons ni rejeter l'hypothèse de spécification linéaire pour la moyenne de la réponse, ni celle de l'homoscédasticité (variance constante), ni celle de l'indépendence des termes d'erreur, mais que l'hypothèse de la normalité peut fort probablement être rejetée, surtout dans le régime de la longue queue à droite. Dépendammentdu test utilisé, on pourrait aussi finir par en conclure que la moyenne n'est pas une combinaison linéaire des prédicteurs.

Visuellement, les tracés diagnostiques ne sont pas hyper-appétissants; de toute évidence, il y a de la structure dans les données (p-ê puisque `Age` est une variable ordinale?), mais cette structure n'est pas capturée par les tests formels.

Et même lorsque l'on capture un problème avec la normalité, le problème ne semble pas si énorme que cela. En réalité, les données ne sont jamais normales, mais on veut savoir si le fait qu'elles ne le sont pas viendra occasionner des problèmes. Mais il y a peut-être quand même quelque chose à corriger ici.

Nous semblons nous trouver dans un cas limite: le modèle n'est pas idéal, bien sûr, mais de là à dire qu'il n'est pas approprié… j'imagine que cela dépend des applications que nous avons en tête, mais avoir à me prononcer, je lui donnerais un "C" – note de passage, certes, mais je ne lui fournirais pas de lettre de recommendation, disons.

S'il y avait un truc à mettre en doute, cela pourrait peut-être être qu'il semble y avoir une composante non-aléatoire aux données – j'ai rarement vu de telles tendances dans des données. C'est l'effet que cela me donne, tout du moins.

Et vous, qu'en pensez-vous?

Q37

Utilisez les mesures correctives appropriées afin d'améliorer les résultats d'ajustement multiple.

Solution: vu la solution à la question 36, on peut essayer de jouer avec la normalité. Il n'y a pas de transformations suggérées par la tendance des réponses, des résidus, et des valeurs ajustées; dans de tels cas, on peut toujours s'essayer avec Box-Cox (on utilise une variable y transformée $\mapsto y + \eta$, η petit, afin d'éviter d'avoir à diviser par 0 dans la transformation de Box-Cox ou encore de prendre le log de 0 ... on pourait peut-être aussi se débarasser des observations pour lesquelles $y = 0$ et voir ce dont il en découle, ou encore n'utiliser que des valeurs positives de λ?).

Avec $\eta = 0.01$, nous avons:

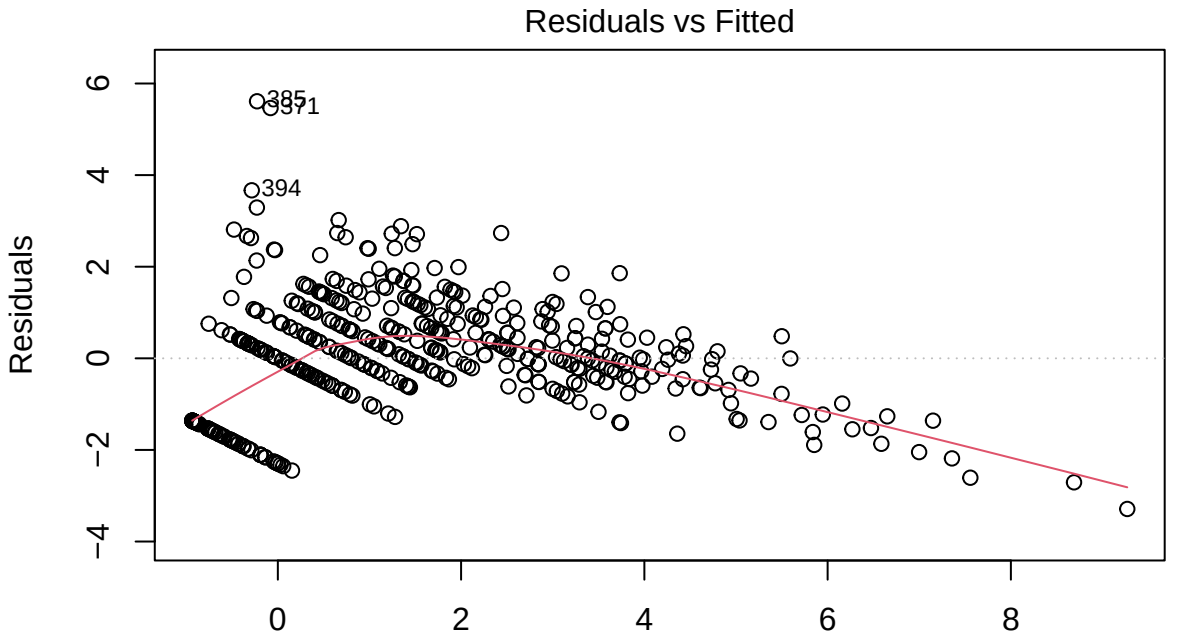
```
k = -1 + (0:100)*0.0261
SSE=array()
i = 0
for(lambda in k){
  i = i + 1
  y.lambda = ((y+0.01)^(lambda)-1)/lambda
  mod.tmp = lm(y.lambda ~ x1 + x2)
  SSE[i] = sum(mod.tmp$residuals^2)
}

k[which(SSE == min(SSE))]
```

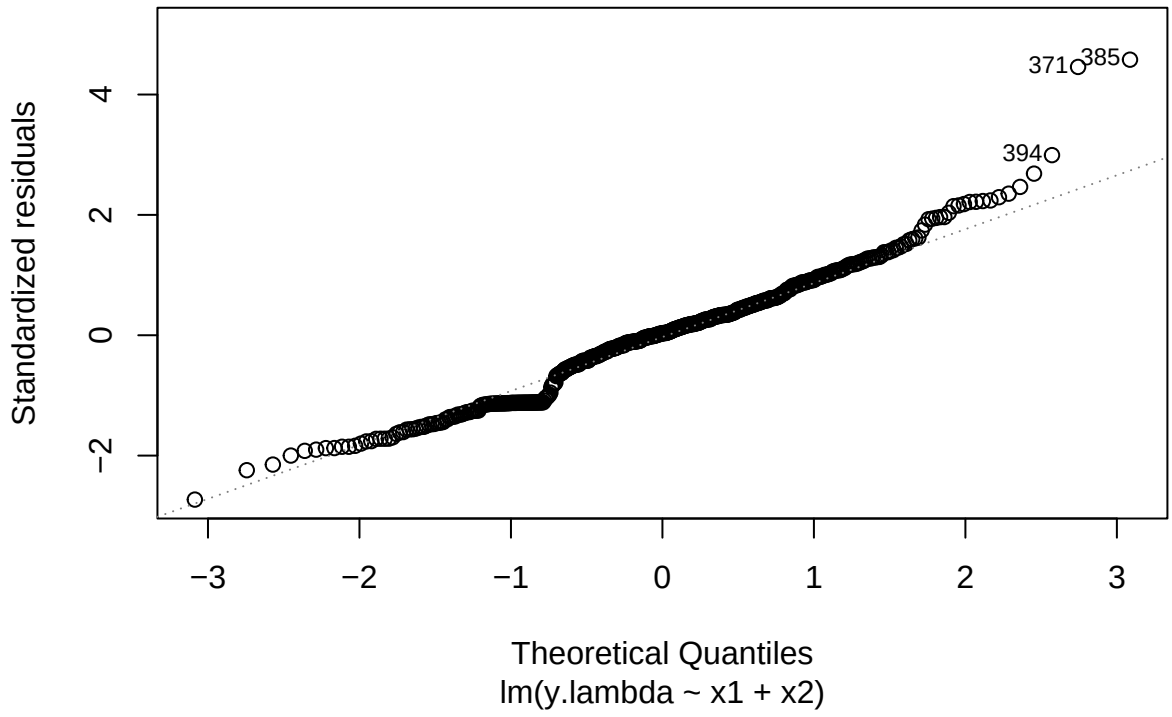
```
## [1] 0.4355
```

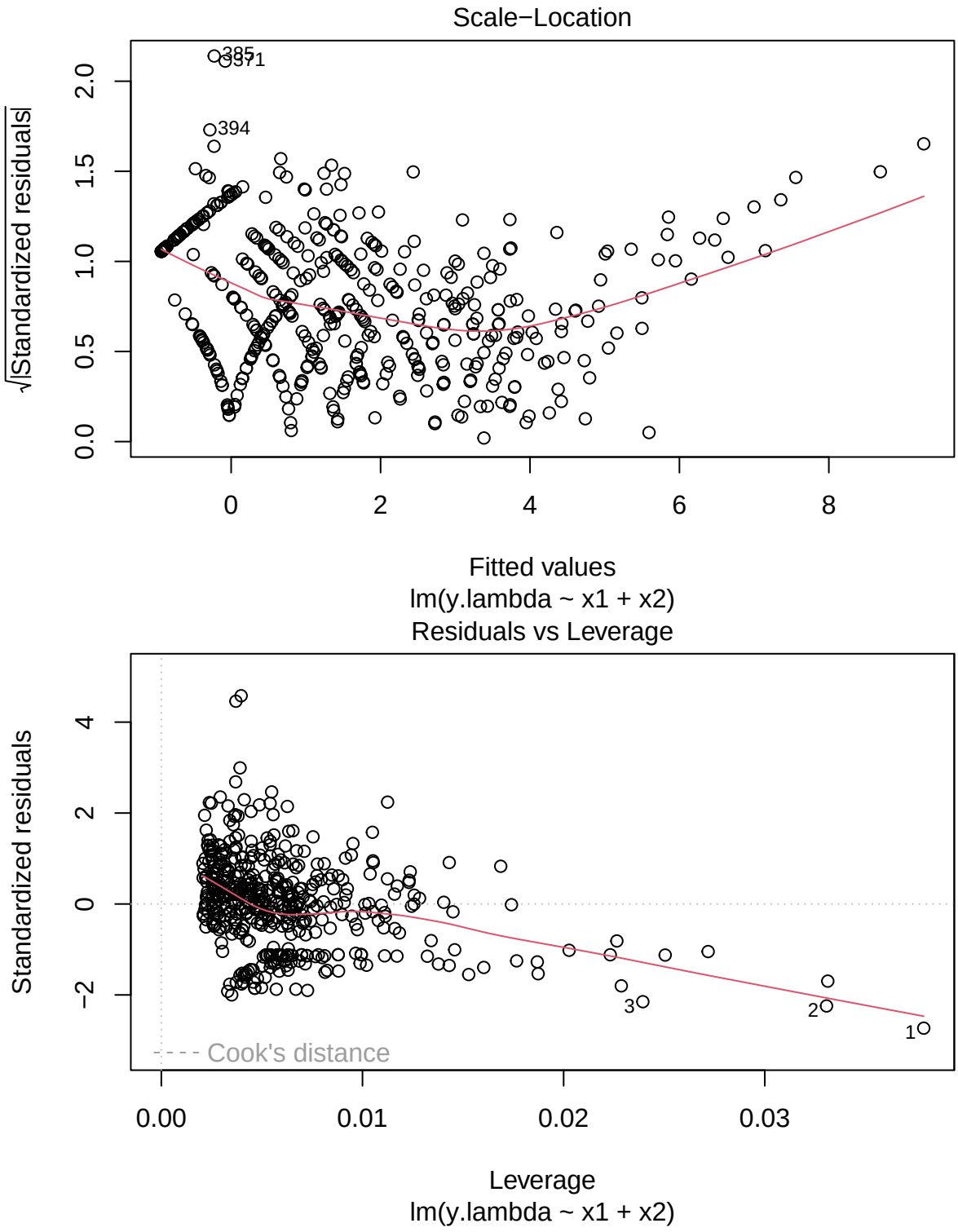
```
y.lambda = (y^(k[which(SSE == min(SSE))])-1)/k[which(SSE == min(SSE))]
mod.lambda = lm(y.lambda ~ x1 + x2)
summary(mod.lambda)
```

```
##
## Call:
## lm(formula = y.lambda ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2889 -0.7754  0.0399  0.7043  5.6113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.934875   0.122405  -7.638 1.17e-13 ***
## x1           0.048999   0.001448  33.830  < 2e-16 ***
## x2           0.002220   0.010850   0.205    0.838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.228 on 491 degrees of freedom
## Multiple R-squared:    0.7,  Adjusted R-squared:  0.6988
## F-statistic: 572.8 on 2 and 491 DF,  p-value: < 2.2e-16
```

```
plot(mod.lambda)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y.lambda ~ x1 + x2)
385
371
394


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y.lambda ~ x1 + x2)
371
385
394



La transformation optimale résoud le problème de la normalité des termes d'erreur, mais au détrimant des autres suppositions; étant donné que nous n'étions pas si "fâché'' de la qualité du modèle au départ, nous pourrions le laisser comme tel.

Vaut-il la peine d'utiliser les moindres carrés pondérés? On utilise $|e_i| \approx \sigma_i$, et $w_i = \frac{1}{\hat{s}_i}$, où les $\hat{s}_i$ sont les valeurs ajustées des $|e_i|$ en fonction

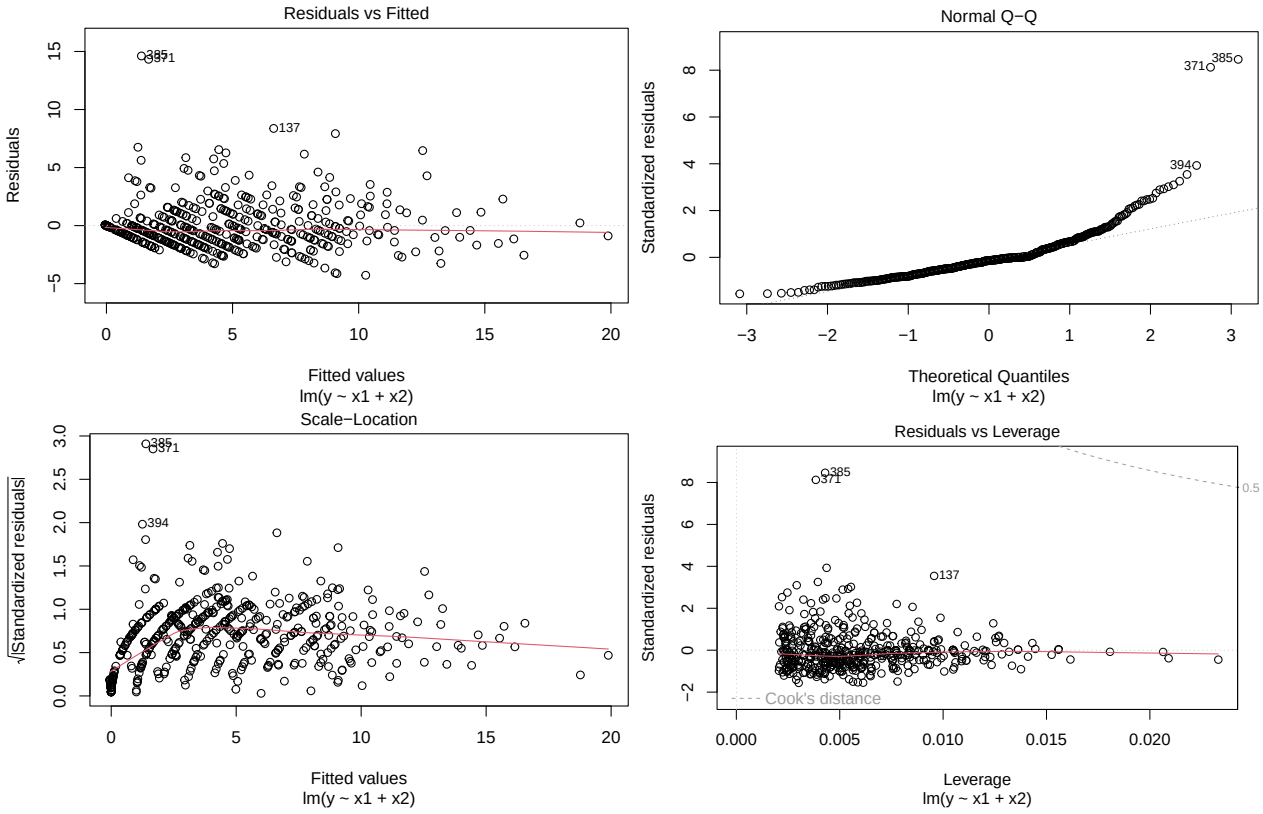
```r
poids <- 1 / lm(abs(mod$residuals) ~ x1 + x2)$fitted.values^2
mod.wls <- lm(y ~ x1 + x2, weights=poids)
(MSE.w = sum(mod.wls$residuals^2)/(n-p))
```

```
## [1] 4.310329
```

```r
summary(mod.wls)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, weights = poids)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3891 -0.8744 -0.2108  0.5062 12.9508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.054328   0.196208  -0.277    0.782
## x1           0.095581   0.002849  33.544   <2e-16 ***
## x2           0.010207   0.018019   0.566    0.571
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.534 on 491 degrees of freedom
## Multiple R-squared:  0.6965, Adjusted R-squared:  0.6953
## F-statistic: 563.5 on 2 and 491 DF,  p-value: < 2.2e-16
```

```r
plot(mod.wls)
```

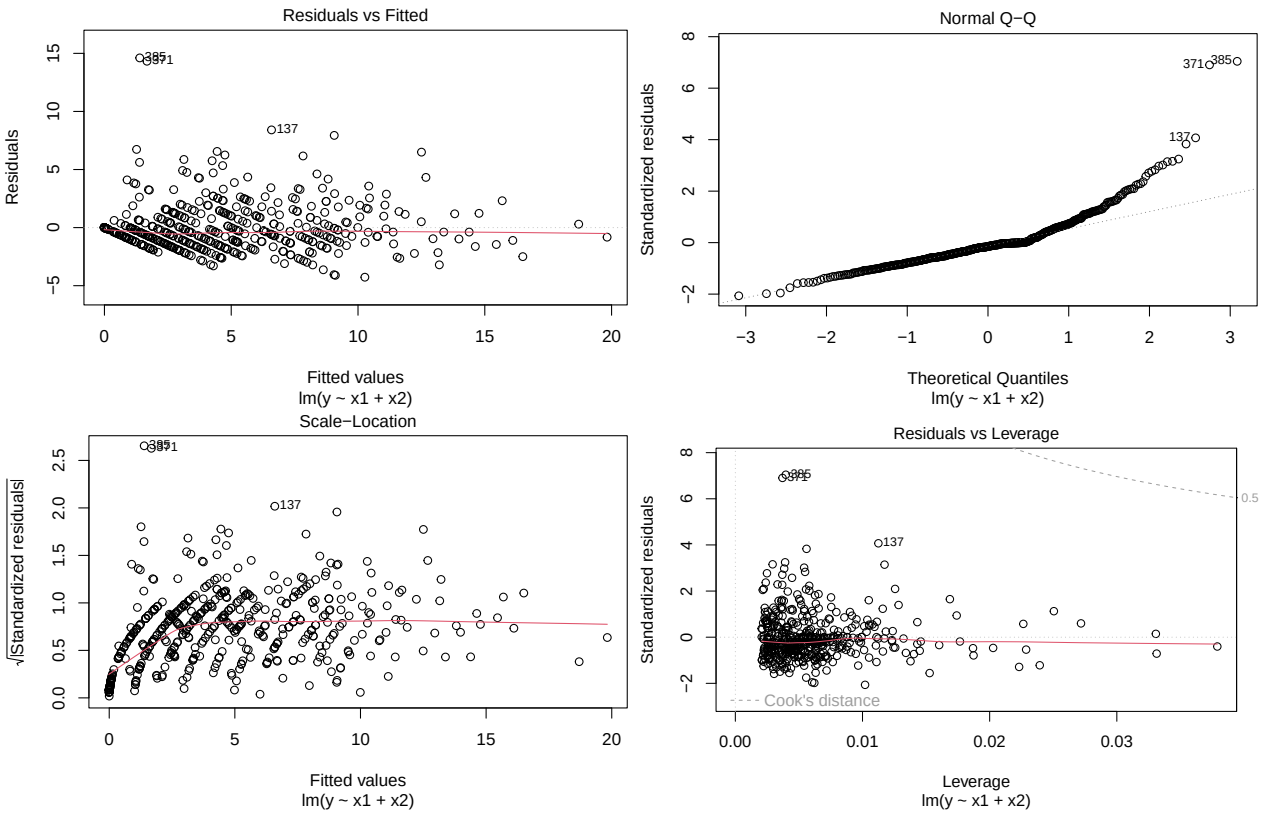


En comparant avec le modèle initial:

```
mod <- lm(y ~ x1 + x2)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2704 -1.2115 -0.3180  0.6609 14.6080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.014050   0.207183  -0.068    0.946
## x1           0.095158   0.002452  38.815   <2e-16 ***
## x2           0.007384   0.018365   0.402    0.688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.078 on 491 degrees of freedom
## Multiple R-squared:  0.7543, Adjusted R-squared:  0.7533
## F-statistic: 753.9 on 2 and 491 DF,  p-value: < 2.2e-16
```

```
plot(mod)
```



Encore une fois, pas super excitant. Ni la transformation de Box-Cox et le modèle WLS n'améliore vraiment la situation (faible réduction de SSE), mais ce n'est vraiment pas exceptionnel dans les 2 cas.

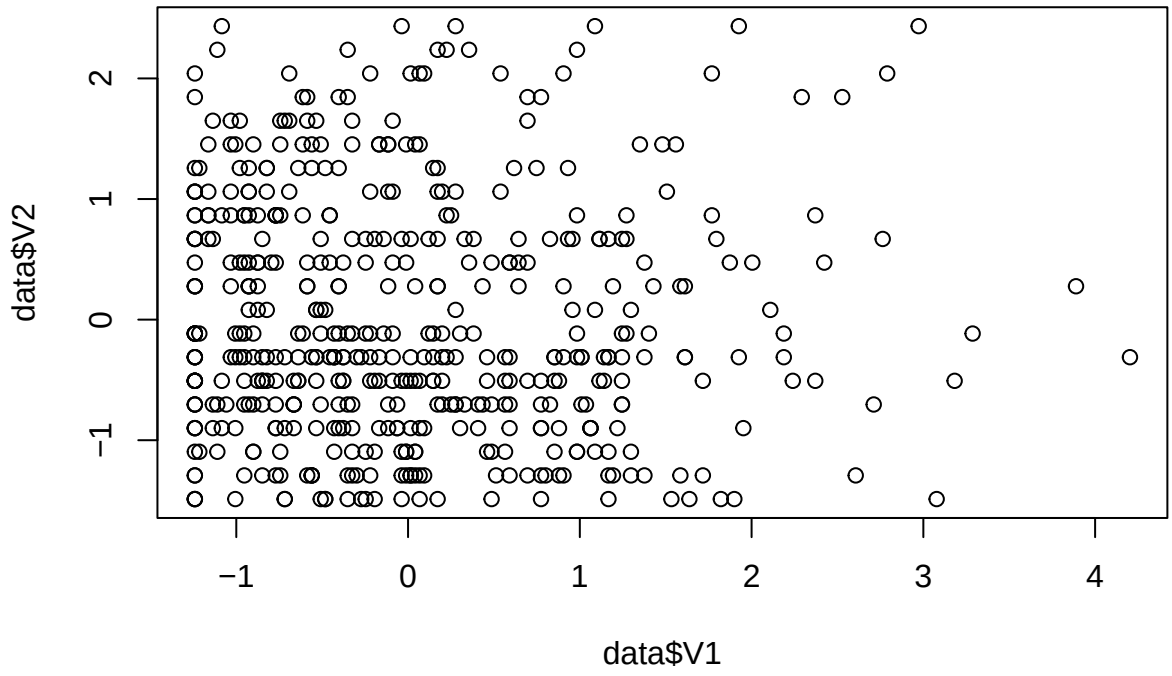
Q38

Les prédicteurs de l'ensemble de données sont-ils multicollinéaires? Justifiez votre réponse.

Solution: il n'y a que deux prédicteurs, $x1$ et $x2$. Il suffit d'aller calculer le coefficient de détermination R^2 entre les deux.

Commençons par centrer et standardiser les données (pas essentiel pour la multi-collinéarité, mais je veux vous montrer comment on le ferait).

```
data = X[,2:3]
data = data.frame(scale(data, center=TRUE, scale=TRUE))
colnames(data) = c("V1","V2")
plot(data$V1, data$V2)
```




```
summary(lm(V1 ~ V2, data=data))
```

```
##
## Call:
## lm(formula = V1 ~ V2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2983 -0.7962 -0.1951  0.6388  4.1911
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.023e-16  4.500e-02    0.00    1.000
## V2          -3.830e-02  4.505e-02   -0.85    0.396
##
## Residual standard error: 1 on 492 degrees of freedom
## Multiple R-squared:  0.001467,   Adjusted R-squared:  -0.0005627
## F-statistic: 0.7227 on 1 and 492 DF,  p-value: 0.3957
```

Le coefficient de détermination est minuscule: les prédicteurs ne sont pas collinéaires.

Q39

Pour cette question, nous allons laisser tomber la variable `Age`. Ajustez la réponse à une régression cubique centrée sur le prédicteur $x_1 = X_1 - \overline{X}_1$ en ajoutant une variable à la fois, afin d'obtenir

$$E\{Y \mid x_1\} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3.$$

En utilisant $\alpha = 0.05$, testez $H_0 : \beta_2 = \beta_3 = 0$ vs. $H_1 : \beta_2 \neq 0$ ou $\beta_3 \neq 0$.

Solution: nous centrons la variable x_1.

```
x.c = x1 - mean(x1)
x.c.2 = x.c^2
x.c.3 = x.c^3
```

```
mod.3 = lm(y ~ x.c + x.c.2 + x.c.3)
```

```
anova(mod.3)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq   F value Pr(>F)
## x.c         1 6510.9  6510.9 1505.1763 <2e-16 ***
## x.c.2       1    0.9     0.9    0.2195 0.6396
## x.c.3       1    0.6     0.6    0.1475 0.7011
## Residuals 490 2119.6     4.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si H_0 est valide, la statistique

$$F^* = \frac{\text{SSR}(R)/(4-2)}{\text{SSR}(F)/(n-4)} = \frac{\text{SSR}(x^2, x^3|x)/(4-2)}{\text{SSE}(x, x^2, x^3)/(n-4)}$$

suit une loi $F(4-2, n-4)$, où $q = 2$ est le nombre de paramètres dans le modèle réduit (R), $v = 4$ est le nombre de paramètres dans le modèle complet (F) et $n - 4 = 490$ est le # de degr'es de libert'e de l'erreur.

Si $\alpha = 0.05$, la valeur critique est $F(0.95; 2, 490)$:

```
qf(0.95,2,490)
```

```
## [1] 3.014122
```

Puisque

$$F^* = \frac{\left[\text{SSR}(x^2|x) + \text{SSR}(x^3|x, x^2)\right]/2}{\text{SSE}(x, x^2, x^3)/490}$$

```
((0.9+0.6)/2)/(2119.6/490)
```

```
## [1] 0.1733818
```

alors $F^* < F(0.95; 2, 490)$ et on ne rejette pas H_0 à ce niveau de confiance.

Q40

Pour cette question, nous ré-introduisons la variable `Age`. Préparez un modèle polynomial de degré 2 en X_1 et X_2 qui inclu un terme d'interaction (le modèle complet) et un modèle n'étant que de degré 1 en X_1 et X_2, mais qui contient quand même un terme d'interaction (modèle réduit). Déterminez les coefficients dans les deux cas. Lequel des deux modèles est préférable?

Solution: préparons les variables séparément (nous allons centrer de nouveau).

```
x1.c = x1 - mean(x1) # pred centre x1
x2.c = x2 - mean(x2) # pred centre x2
x1.c.2 = x1.c^2      # pred centre x1^2
x1.c.x2.c = x1.c * x2.c # terme d'interaction
x2.c.2 = x2.c^2      # pred centre x2^2
```

Le modèle complet est:

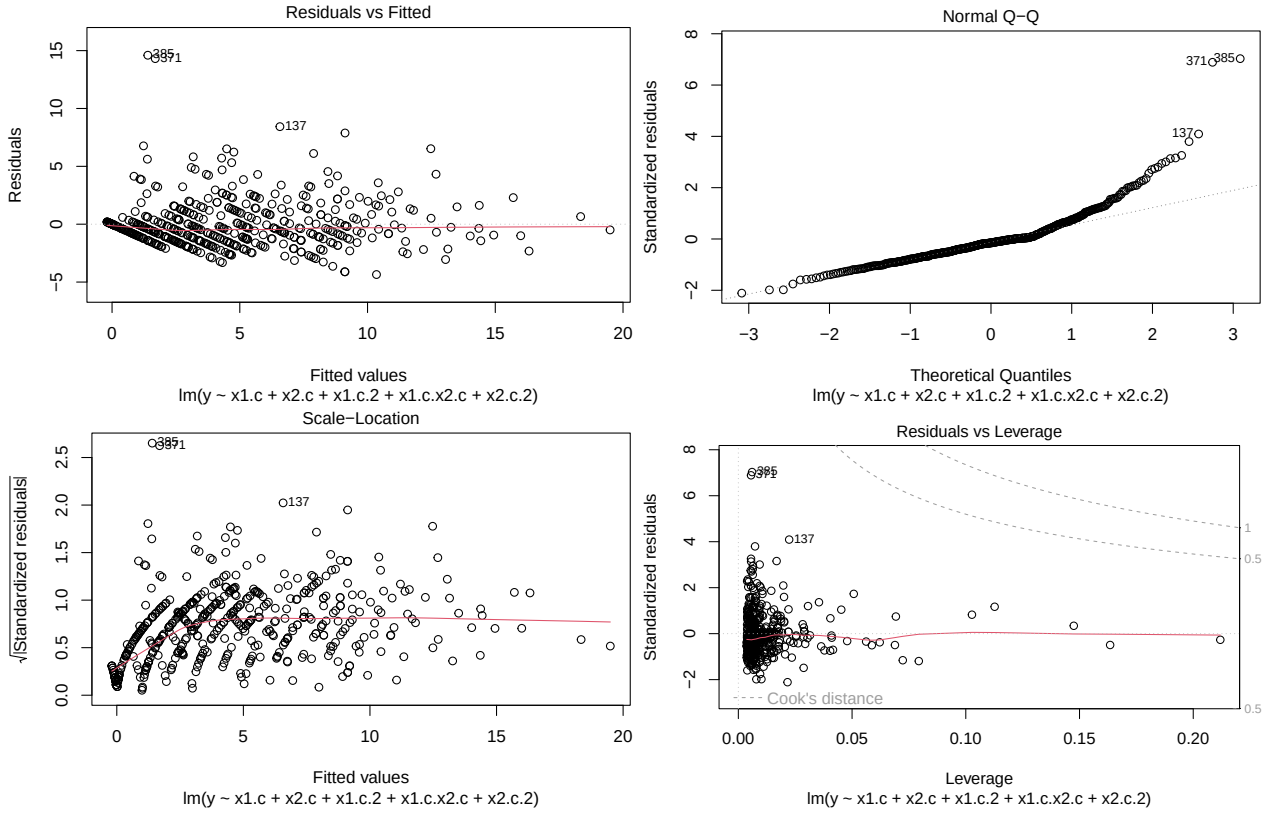
```
mod.c <- lm(y ~ x1.c + x2.c + x1.c.2 + x1.c.x2.c + x2.c.2)
summary(mod.c)
```

```
##
## Call:
## lm(formula = y ~ x1.c + x2.c + x1.c.2 + x1.c.x2.c + x2.c.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3504 -1.2037 -0.3125  0.6755 14.6008
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.596e+00  1.485e-01  30.943   <2e-16 ***
## x1.c         9.610e-02  3.030e-03  31.721   <2e-16 ***
## x2.c         9.678e-03  2.052e-02   0.472    0.637
## x1.c.2      -2.242e-05  4.652e-05  -0.482    0.630
## x1.c.x2.c   -2.415e-04  4.754e-04  -0.508    0.612
## x2.c.2      -3.931e-04  3.441e-03  -0.114    0.909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.083 on 488 degrees of freedom
## Multiple R-squared:  0.7546, Adjusted R-squared:  0.7521
## F-statistic: 300.1 on 5 and 488 DF,  p-value: < 2.2e-16
```

```
anova(mod.c)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq   F value Pr(>F)
## x1.c        1 6510.9  6510.9 1500.0292 <2e-16 ***
## x2.c        1    0.7     0.7    0.1609 0.6885
## x1.c.2      1    1.1     1.1    0.2465 0.6198
## x1.c.x2.c   1    1.2     1.2    0.2698 0.6037
## x2.c.2      1    0.1     0.1    0.0131 0.9091
## Residuals 488 2118.2     4.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(mod.c)
```



Le modèle réduit est:

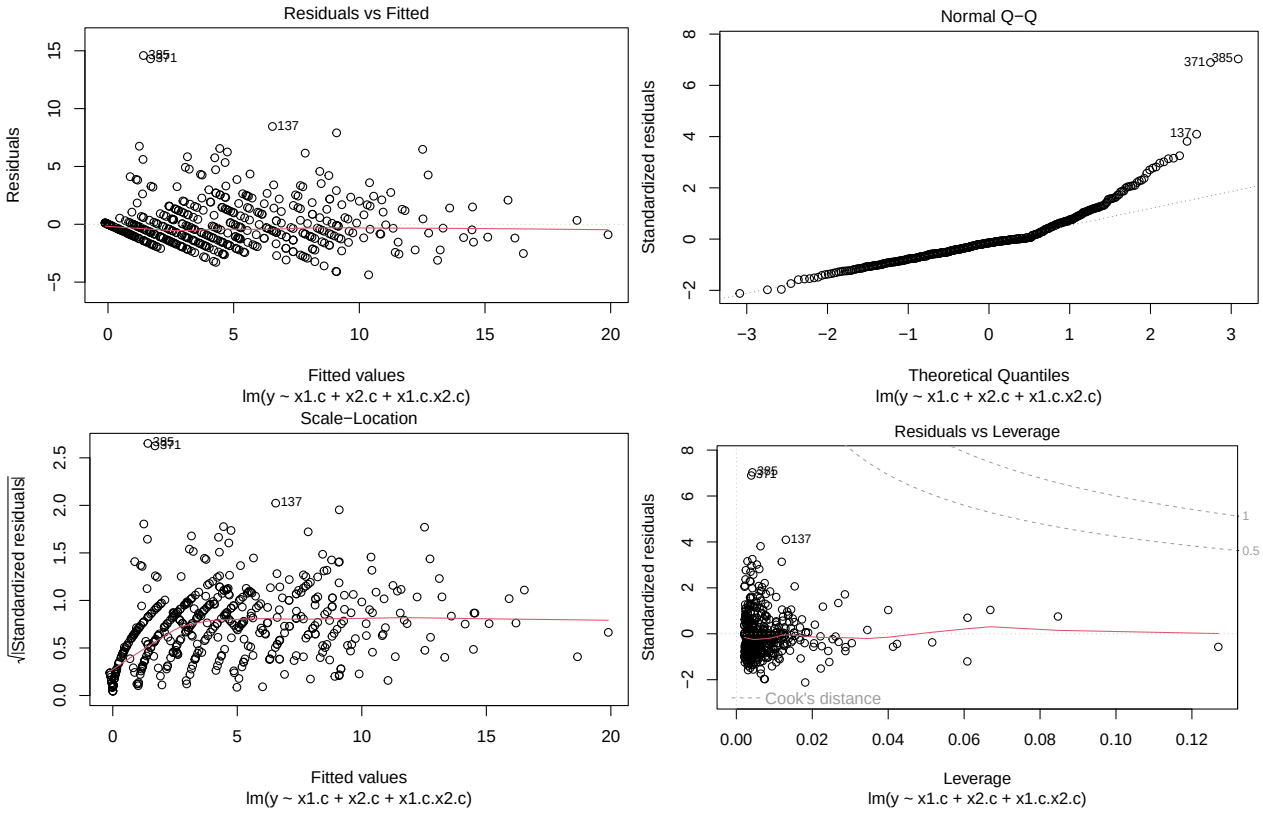
```
mod.r <- lm(y ~ x1.c + x2.c + x1.c.x2.c)
summary(mod.r)
```

```
##
## Call:
## lm(formula = y ~ x1.c + x2.c + x1.c.x2.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3702 -1.1903 -0.3072  0.6568 14.5903
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.5527608  0.0936336  48.623   <2e-16 ***
## x1.c         0.0952362  0.0024577  38.750   <2e-16 ***
## x2.c         0.0080798  0.0184235   0.439    0.661
## x1.c.x2.c   -0.0002545  0.0004725  -0.538    0.590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.08 on 490 degrees of freedom
## Multiple R-squared:  0.7545, Adjusted R-squared:  0.753
## F-statistic:   502 on 3 and 490 DF,  p-value: < 2.2e-16
```

```
anova(mod.r)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq   F value Pr(>F)
## x1.c        1 6510.9  6510.9 1505.4352 <2e-16 ***
## x2.c        1    0.7     0.7    0.1614 0.6880
## x1.c.x2.c   1    1.3     1.3    0.2900 0.5905
## Residuals 490 2119.2     4.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(mod.r)
```



En comparant avec le modèle qui ne contient qu'un seul prédicteur, on se rend compte que ni l'utilisation de termes d'interaction ou de termes d'ordre supérieurs ne peuvent vraiment venir aider le modèle.

```
mod.0 <- lm(y ~ x1.c)
summary(mod.0)
```

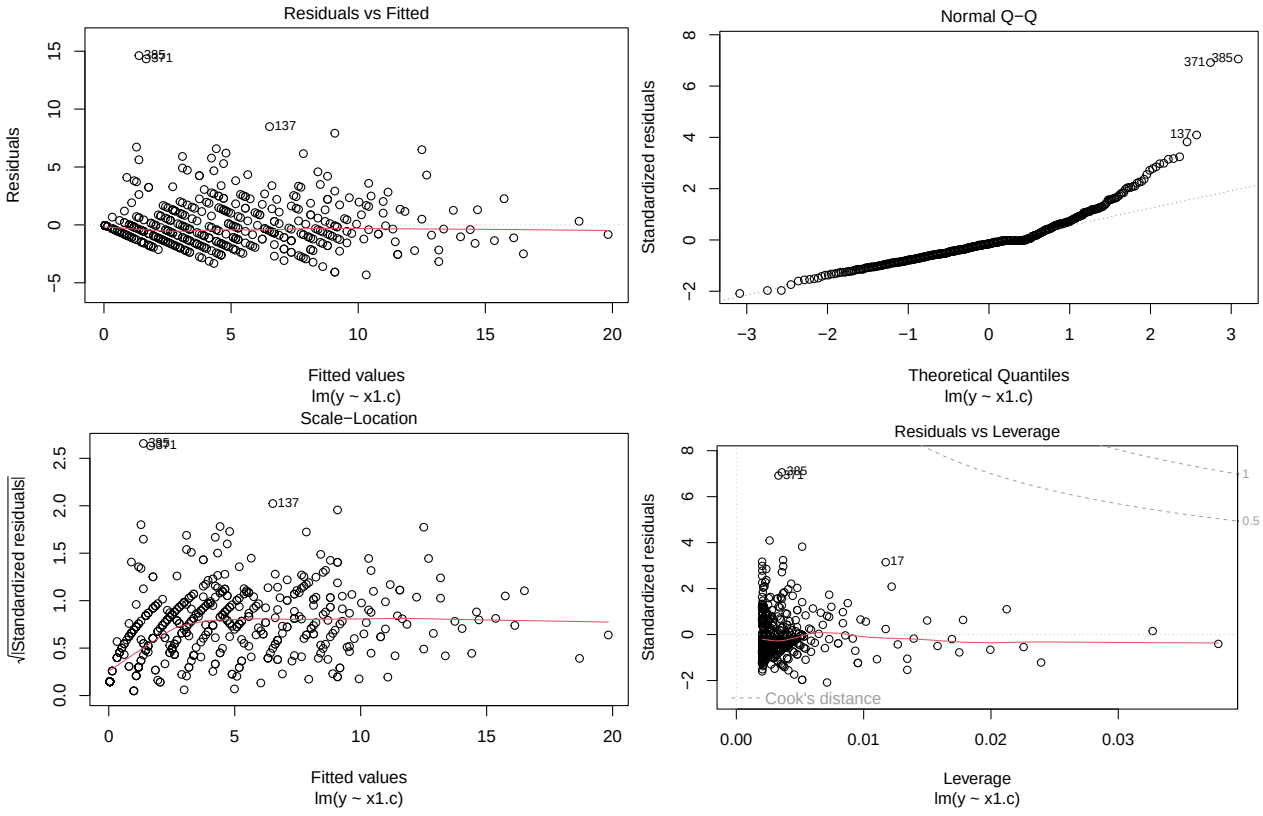
```
##
## Call:
## lm(formula = y ~ x1.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3167 -1.1852 -0.3217  0.7075 14.6245
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.554656   0.093420   48.76   <2e-16 ***
## x1.c        0.095120   0.002448   38.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.076 on 492 degrees of freedom
## Multiple R-squared:  0.7543, Adjusted R-squared:  0.7538
## F-statistic:  1510 on 1 and 492 DF,  p-value: < 2.2e-16
```

```
anova(mod.0)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## x1.c        1 6510.9  6510.9  1510.2 < 2.2e-16 ***
## Residuals 492 2121.2     4.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(mod.0)
```



Alors pourquoi avoir fait tout ce travail (Q31 à Q40)? Quand un modèle ne nous laisse ni chaud, ni froid, on a souvent tendance à s'essayer avec des truc de plus en plus sophistiqués (transformations, ajout de prédicteurs et de termes d'interaction, etc.) Cela fonctionne souvent. Mais il ne faut pas non plus aller chercher midi à 14 heures et reconnaître que le meilleur modèle sera parfois quand même assez médiocre. Et c'est acceptable. Les données réelles ne se portent pas toujours à nos attentes/demandes.