Devoir 2

Partie théorique

T-1

#1

On veut montrer que $\bar{e} = \sum_{i=1}^{n} \frac{e_i}{n} = 0$. On peut donc développer la formule de $\bar{e}$:

$$\begin{aligned}\bar{e} &= \sum_{i=1}^{n} \frac{e_i}{n} = \sum_{i=1}^{n} \frac{Y_i - \hat{Y}_i}{n} \\ &= \bar{Y} - \frac{\sum_{i=1}^{n} x_i^{'}\hat{\beta}}{n}\end{aligned}$$

Nous pouvons écrire la somme de droite de cette façon:

$$\begin{aligned}\sum_{i=1}^{n} x_i^{'}\hat{\beta} &= \hat{\beta}_0 \sum_{i=1}^{n} 1 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1} + ... + \hat{\beta}_{p'} \sum_{i=1}^{n} x_{i,p'} \\ \frac{\sum_{i=1}^{n} x_i^{'}\hat{\beta}}{n} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + ... + \hat{\beta}_{p'} \bar{x}_{p'}\end{aligned}$$

Tel que donnée dans l'énoncé, on peut prendre pour acquis que:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}_1 - ... - \hat{\beta_{p'}} \bar{x}_{p'}$$

Ainsi, en remplaçant les valeurs obtenues avec l'équation plus haut, on obtient:

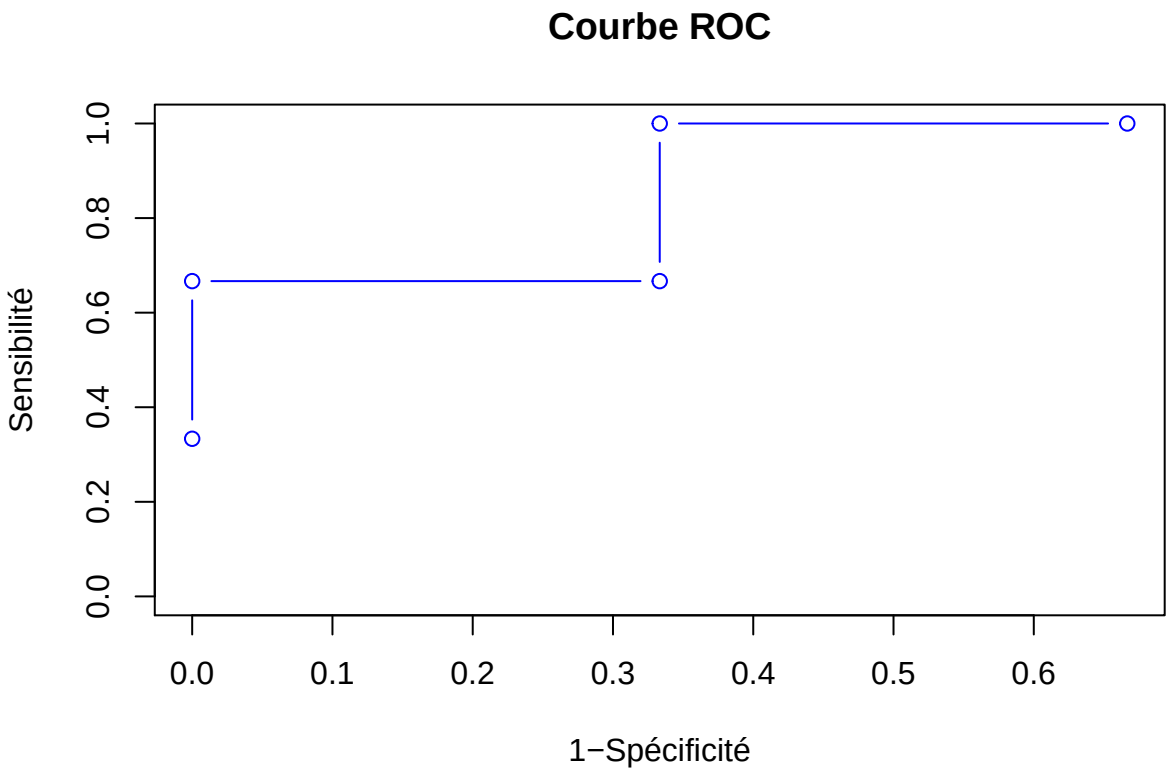
$$\begin{aligned}\bar{e} &= \bar{Y} - \frac{\sum_{i=1}^{n} x_i^{'}\hat{\beta}}{n} \\ &= \bar{Y} - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + ... \hat{\beta}_{p'} \bar{x}_{p'}\right) \\ &= \bar{Y} - \left(\left[\bar{Y} - \hat{\beta}_1 \bar{x}_1 - ... - \hat{\beta_{p'}} \bar{x}_{p'}\right] + \hat{\beta}_1 \bar{x}_1 + ... + \hat{\beta}_{p'} \bar{x}_{p'}\right) \\ &= \bar{Y} - \bar{Y} \\ \bar{e} &= 0\end{aligned}$$

#4

Pour faire la courbe ROC, il faut calculer la valeur de $\hat{\pi}_i = P(Y_i = 1, x_i) = \frac{exp(-4.2+1.2x_i)}{1+exp(-4.2+1.2x_i)}$. Par la suite on calcule une prédiction de $\hat{Y}_i$ selon un certain seuil u_k : si $\hat{\pi}_i > u_k$, alors $\hat{Y}_i = 1$, sinon $\hat{Y}_i = 0$. On calule alors la sensibilié et la spécificité pour les différents seuils. Voici un tableau des résulats ($\hat{Y}_i$ est représenté comme étant $\hat{Y}_u$ selon la valeur de u_k) pour les valeurs de u_k =0.1, 0.2, 0.5, 0.8 et 0.9 ainsi que le graphique de la courbe ROC:

Observation	x_i	$\hat{\pi}_i$	Y_i	$\hat{Y}_{u=0.1}$	$\hat{Y}_{u=0.2}$	$\hat{Y}_{u=0.5}$	$\hat{Y}_{u=0.8}$	$\hat{Y}_{u=0.9}$
obs 1	1	0.0474259	0	0	0	0	0	0
obs 2	2	0.1418511	0	1	0	0	0	0
obs 3	3	0.3543437	1	1	1	0	0	0
obs 4	4	0.6456563	0	1	1	1	0	0
obs 5	5	0.8581489	1	1	1	1	1	0
obs 6	6	0.9525741	1	1	1	1	1	1

Métriques	$\hat{Y}_{u=0.1}$	$\hat{Y}_{u=0.2}$	$\hat{Y}_{u=0.5}$	$\hat{Y}_{u=0.8}$	$\hat{Y}_{u=0.9}$
VP	3.0000000	3.0000000	2.0000000	2.0000000	1.0000000
FN	0.0000000	0.0000000	1.0000000	1.0000000	2.0000000
VN	1.0000000	2.0000000	2.0000000	3.0000000	3.0000000
FP	2.0000000	1.0000000	1.0000000	0.0000000	0.0000000
Sensibilité	1.0000000	1.0000000	0.6666667	0.6666667	0.3333333
Spécificité	0.3333333	0.6666667	0.6666667	1.0000000	1.0000000



T-2

#4

Le modèle est défini de la façon suivante:

$$Y_{ij} = \beta_0 + \gamma_{i0} + (\beta_1 + \gamma_{i1})x_{ij} + \epsilon_{ij}$$

On peut définir la matrice de betas et de gammas pour nos tests d'hypothèse qui prendront tous la même forme:

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \gamma_{10} \\ \gamma_{11} \\ \gamma_{20} \\ \gamma_{21} \end{bmatrix}$$

$$H_0 : \boldsymbol{L} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \boldsymbol{d}$$

$$H_1 : \boldsymbol{L} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} \neq \boldsymbol{d}$$

a)

Pour tester si l'effet moyen du médicament sur la pression sanguine moyenne dans cette population de souris est nul, on doit regarder l'espérance de Y lorsque les effets aléatoires sont nuls. Alors,

$$E[Y; x_0 = x + 1] - E[Y; x_0 = x] = 0$$

Ce qui revient à tester si $\beta_1 = 0$.

$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$
$$\boldsymbol{d} = \begin{bmatrix} 0 \end{bmatrix}$$

b)

Pour tester si l'effet du médicament sur la pression sanguine moyenne dans la famille 1 est nul, ceci revient à tester si la valeur devant x_{1j} est nulle pour la famille 1. Ceci revient alors à tester si

$$\beta_1 + \gamma_{11} = 0$$
$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$
$$\boldsymbol{d} = \begin{bmatrix} 0 \end{bmatrix}$$

c)

Afin de tester si l'effet du médicament sur la pression sanguine moyenne est nul chez chacune des deux familles de souris, alors on souhaite tester si les valeurs devant x_{ij} pour les 2 familles sont nulles. Alors, on peut tester simulatément ceci:

$$\beta_1 + \gamma_{11} = 0 \text{ et } \beta_1 + \gamma_{21} = 0$$
$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$
$$\boldsymbol{d} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

d)

Afin de tester si une hausse d'une unité de la dose de médicament dans la famille 1 est équivalente à une hausse d'une unité de la dose de médicament dans la famille 2 en termes d'effet sur la pression sanguine moyenne, on peut réécrire le tout de la manière suivante:

$$E[Y_1; x_{10}^* = x_{10} + 1] - E[Y_1; x_{10}^* = x_{10}] = E[Y_2; x_{20}^* = x_{20} + 1] - E[Y_2; x_{20}^* = x_{20}]$$

Alors ceci revient à tester si

$$\beta_1 + \gamma_{11} = \beta_1 + \gamma_{21}$$
$$\gamma_{11} - \gamma_{21} = 0$$

$$\boldsymbol{L} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$
$$\boldsymbol{d} = \begin{bmatrix} 0 \end{bmatrix}$$

#5

a)

On peut définir le modèle de cette façon: $Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_{ij} + \epsilon_{ij}$. Voici la notation matricielle:

$$\mathbf{Y}' = \begin{bmatrix} Y_{11} & Y_{12} & Y_{21} & Y_{22} & Y_{31} & Y_{32} \end{bmatrix}$$
$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{21} \\ 1 & x_{22} \\ 1 & x_{31} \\ 1 & x_{32} \end{bmatrix}$$
$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$
$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \gamma_{30} \end{bmatrix}$$
$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$
$$\epsilon' = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{21} & \epsilon_{22} & \epsilon_{31} & \epsilon_{32} \end{bmatrix}$$

On peut également remplacer les valeurs symboliques de $\mathbf{Y}$ et $\mathbf{X}$ par leurs valeurs numériques:

$$\mathbf{Y}' = \begin{bmatrix} 70 & 80 & 50 & 60 & 100 & 70 \end{bmatrix}$$
$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

b)

On trouve les matrices de variance:

$$\mathbf{D} = Var(\boldsymbol{\gamma}) = \begin{bmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_0^2 & 0 \\ 0 & 0 & \sigma_0^2 \end{bmatrix}$$

$$\mathbf{V} = Var(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = Var(\mathbf{Y}) = \mathbf{ZDZ}^{'} + \mathbf{V}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 + \sigma_0^2 & \sigma_0^2 & 0 & 0 & 0 & 0 \\ \sigma_0^2 & \sigma^2 + \sigma_0^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 + \sigma_0^2 & \sigma_0^2 & 0 & 0 \\ 0 & 0 & \sigma_0^2 & \sigma^2 + \sigma_0^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 + \sigma_0^2 & \sigma_0^2 \\ 0 & 0 & 0 & 0 & \sigma_0^2 & \sigma^2 + \sigma_0^2 \end{bmatrix}$$

c)

Ces deux valeurs peuvent s'estimer avec les formules suivantes:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{'}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}^{'}{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}$$

$$\hat{\boldsymbol{\gamma}} = \mathbf{DZ}^{'}{\boldsymbol{\Sigma}}^{-1}(\mathbf{Y} - \boldsymbol{X\beta})$$

On peut utiliser R pour calculer les valeurs numériques de ces estimés:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 80 \\ -16.67 \end{bmatrix}$$

$$\hat{\boldsymbol{\gamma}} = \begin{bmatrix} 2.67 \\ -13.33 \\ 10.67 \end{bmatrix}$$

d)

On peut calculer l'estimé à partir de cette équation:

$$\hat{\boldsymbol{Y_i}} = \boldsymbol{V_i}{\boldsymbol{\Sigma_i}}^{-1}\boldsymbol{X_i^{'}}\hat{\boldsymbol{\beta}} + (\boldsymbol{I}_{n_i * n_i} - \boldsymbol{V_i}{\boldsymbol{\Sigma_i}}^{-1})\boldsymbol{Y_i}$$

Où:

$$\boldsymbol{V_i}{\boldsymbol{\Sigma_i}}^{-1} = \begin{bmatrix} 0.6 & -0.4 \\ -0.4 & 0.6 \end{bmatrix}$$

De cette façon, on obtient la valeur moyenne de $\hat{\boldsymbol{Y_i}} = 56.67$ pour notre estimé de la note moyenne obtenue par l'individu 2 dans les cours où il utilise un manuel de langue anglaise.

#6

a)

Dans cette situation, le paramètre d'intérêt est β_3 puisque celui-ci estimera l'effet sur la valeur de Y_{ij} au fil du temps seulement lorsque $x_i = 1$.

b)

Il serait raisonnable de choisir les strutures AR(1) et UN(1) puisque cette paire a la plus petite valeur d'AIC pour la méthode ML.

c)

On procède à un test d'hypothèse formel en testant un modèle complet (ligne 1) et un modèle réduit sans la pente aléatoire (ligne 2):

$$H_0 : Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_i + \beta_2 t_j + \beta_3 x_i t_j + \epsilon_{ij}$$
$$H_1 : Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_i + (\beta_2 + \gamma_{i2}) t_j + \beta_3 x_i t_j + \epsilon_{ij}$$

On pose : $\epsilon = 2(\ell_1 - \ell_0) = 2(-88 + 89.5) = 3$. (On pourrait également utiliser les mesures REML pour des résultats semblables).

Nous rejeterons l'hypothèse si la p-value du test est trop élevée:

$$pvalue = 0.5P[\chi^2_{m_1 - m_0 - 1} > \epsilon] + 0.5P[\chi^2_{m_1 - m_0} > \epsilon]$$

$m_0 = 1$(1 variance) et $m_1 = 2$(2 variances). Par conséquent:

$$p = 0.5P[\chi^2_0 > \epsilon] + 0.5P[\chi^2_1 > \epsilon] = 0 + 0.5 * 0.08326 = 0.04163226$$

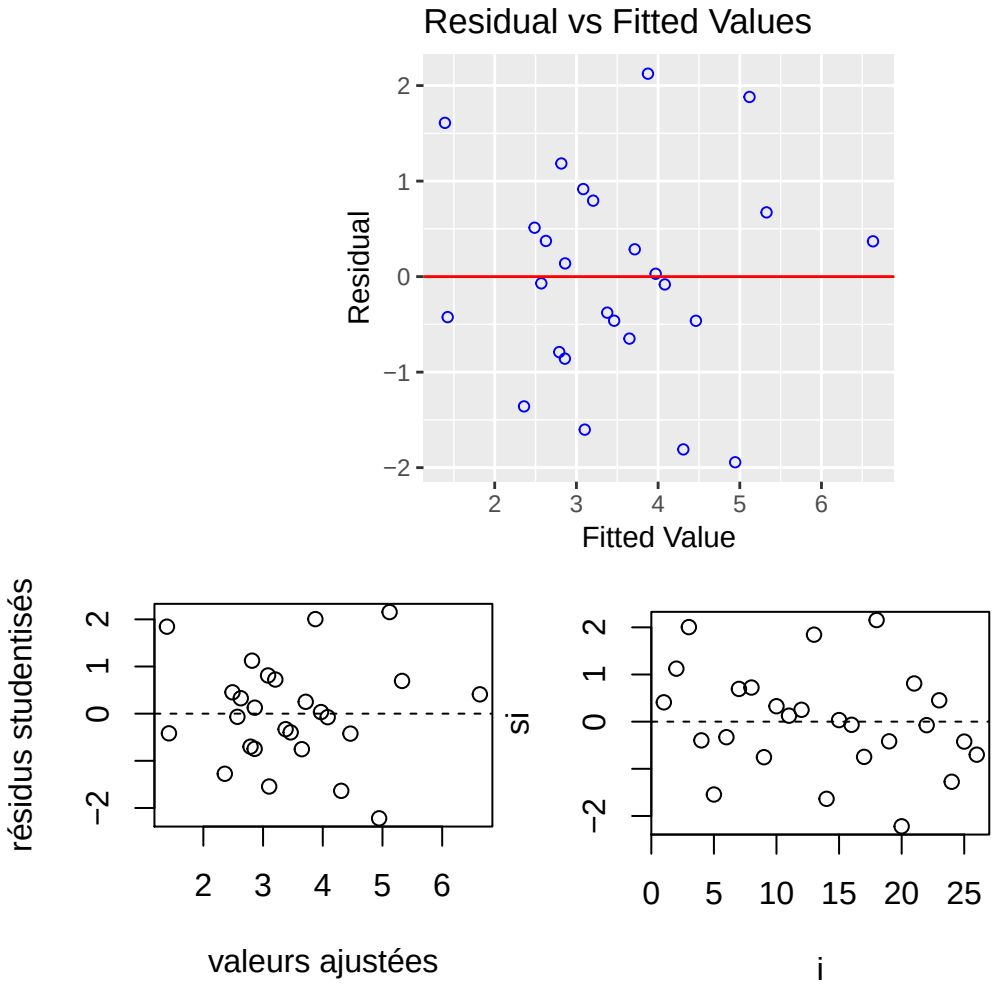
La p-value est inférieur à 0.05, par conséquent on rejète l'hypothèse du modèle réduit avec ce seuil de 5% et on conserve l'effet aléatoire γ_{i2}. Il est toutefois à noter que la p-value est assez proche de 0.05 avec une valeur de 0.0416.

Partie pratique

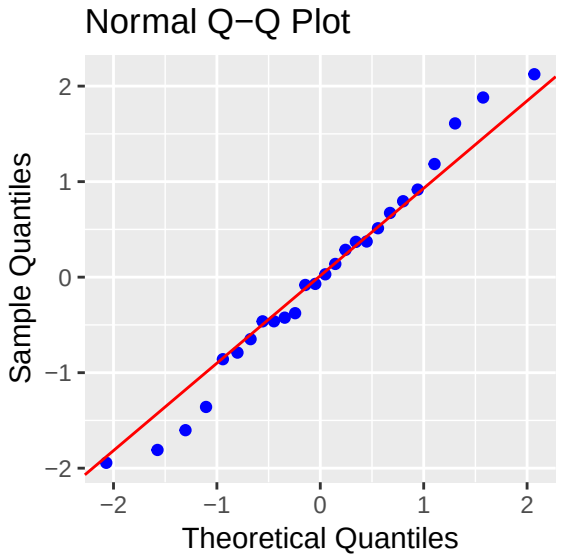
P-1

On reprend le modèle complet du TP1, sans la variable *LG9*, car il y avait de la multicolinéarité en sa présence.

Commençons par regarder les graphiques de résidus du modèle complet.



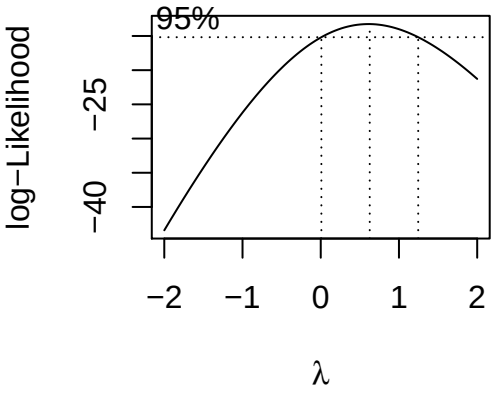
L'allure des résidus semble correct. Il y a peut-être un petit manque de linéarité. Le test de *Box-cox* va être un bon indicateur si on doit faire une transformation sur la variable endogène. De plus on pourrait tester quelques intéractions. Comme on l'indiquait lors de TP1, plus un enfant de 2ans est pesant, mois son indice d'obésité à 18ans sera élevé. Cela peut sembler contre intuitif. Peut-être que la grandeur de l'enfant est vraiment importante à 2ans pour avoir une bonne mesure du poids. On va donc tester d'ajouter un intéraction entre le poids et la grandeur de l'enfant à 2ans.




```
## ------------------------------------------------
##        Test             Statistic       pvalue
## ------------------------------------------------
## Shapiro-Wilk              0.9817         0.9083
## Kolmogorov-Smirnov        0.0635         0.9997
## Cramer-von Mises          1.8419         0.0000
## Anderson-Darling          0.1428         0.9663
## ------------------------------------------------
```

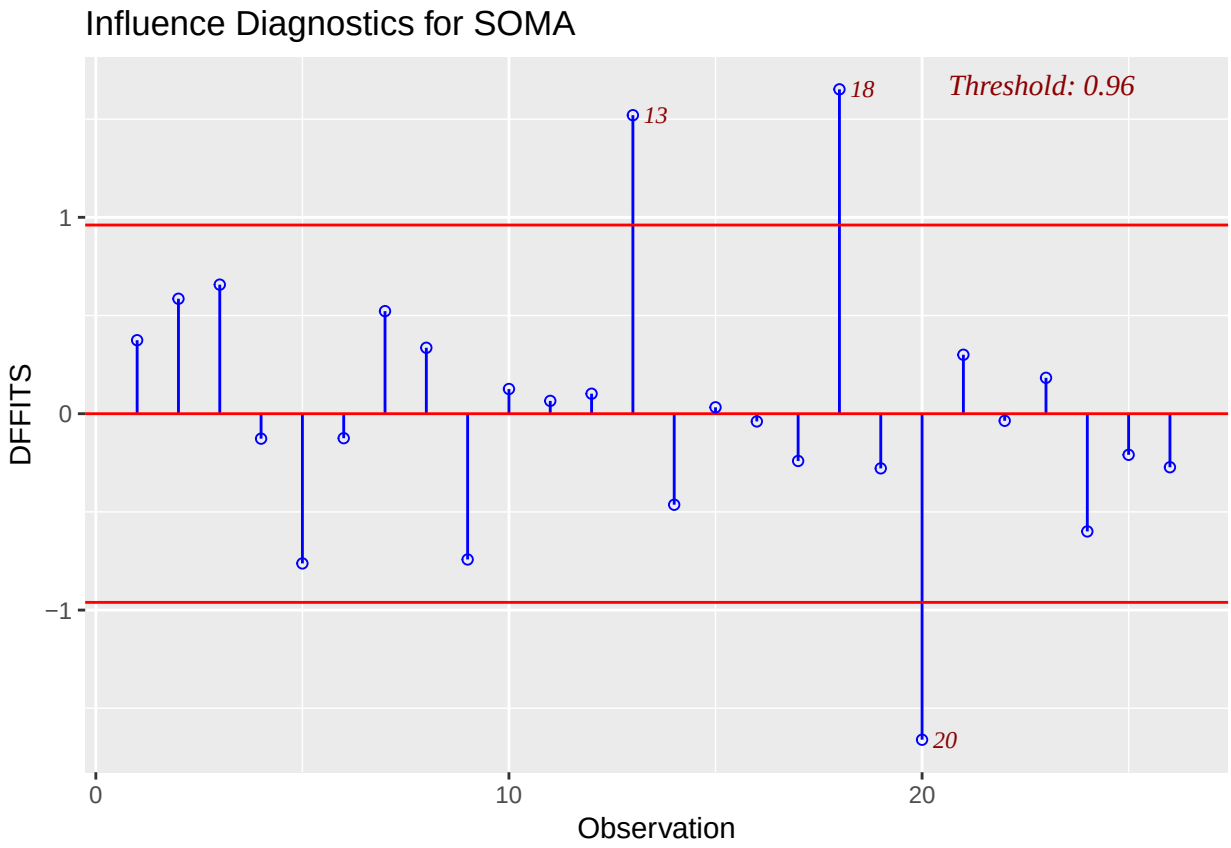
Le graphique de *QQ-plot* montre que les résidus semblent être distribué de façon normal. De plus le test de *Shapiro-Wilk* fait en sorte qu'on ne rejette pas l'hypothèse de normalité des résidus.

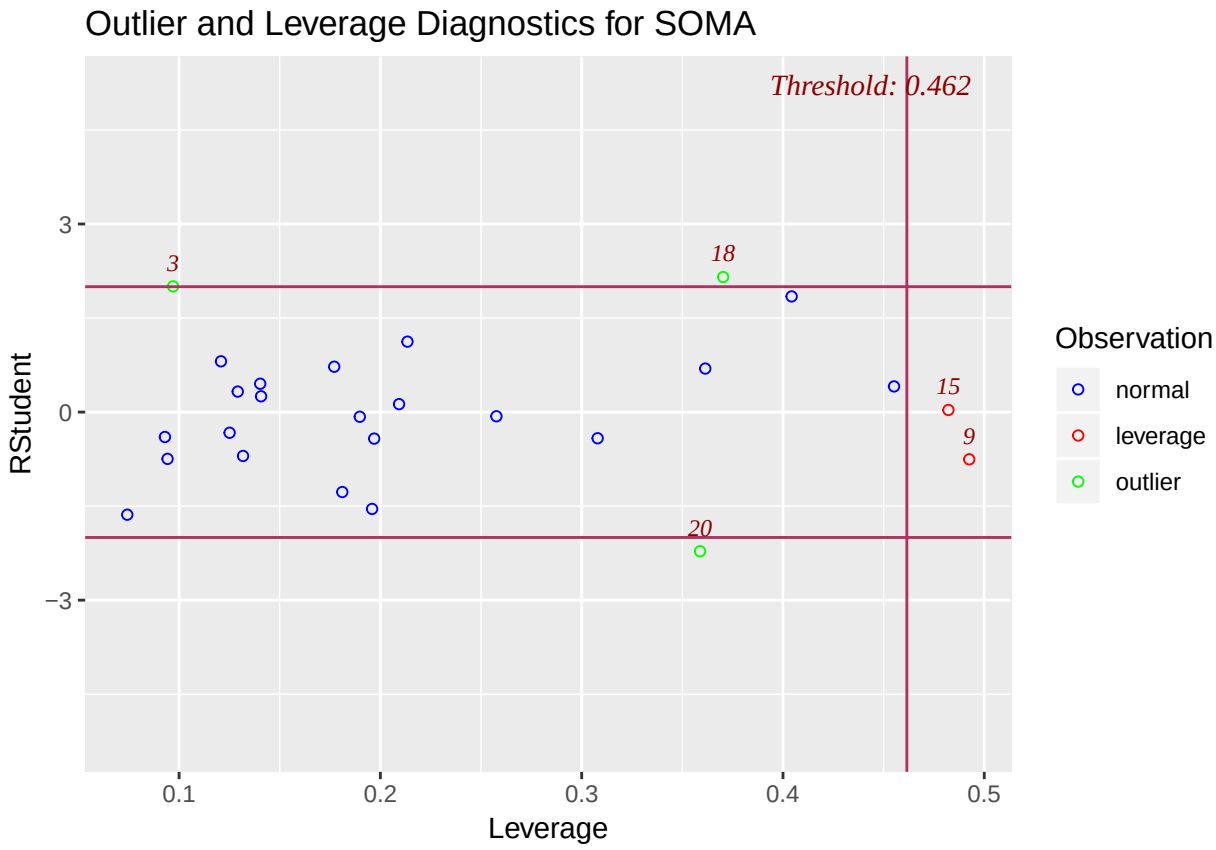
Regardons le graphique des valeurs de lambda avec la transformation de *Box-Cox*



Il serait approprié de faire notre modèle de régression linéaire sur $\sqrt{Y}$.

Regardons maintenant certaines statistiques sur l'influence des données.


Influence Diagnostics for SOMA
Threshold: 0.96
18
13
20
1
0
−1
DFFITS
0
10
20
Observation


Outlier and Leverage Diagnostics for SOMA
Threshold: 0.462
3
18
15
9
20
3
0
−3
RStudent
0.1
0.2
0.3
0.4
0.5
Leverage
Observation
normal
leverage
outlier

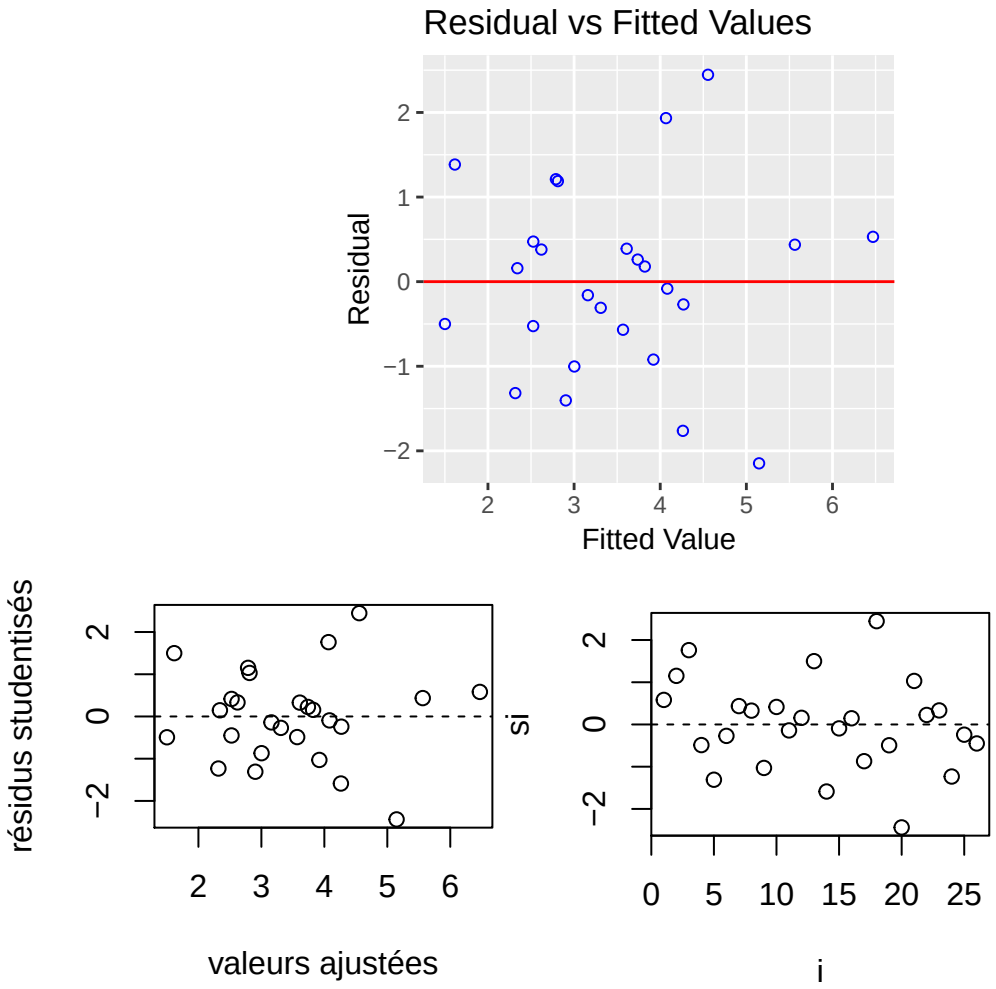
Les données 18 et 20 peuvent potentiellement causer des problèmes. Ils ont une certaines influences sur les valeurs prédites. Cependant, avec le graphique des résidus en fonction des h_{ii} on remarque que ces données sont des outliers, mais cependant il n'ont pas un grand effet de levier. Donc il est correct de les conserver pour faire le modèle.

```
## Influence measures of
##   lm(formula = SOMA ~ WT2 + HT2 + WT9 + HT9 + ST9, data = data_tp1) :
##
##      dfb.1_  dfb.WT2  dfb.HT2   dfb.WT9  dfb.HT9  dfb.ST9   dffit cov.r
## 1   0.04425 -0.18433  0.07681  0.349617 -0.13496 -0.11998  0.3743 2.369
## 2  -0.39929 -0.16795  0.02902 -0.220200  0.34352  0.05570  0.5854 1.176
## 3   0.16078 -0.07779 -0.34725  0.002016  0.33029 -0.21208  0.6572 0.476
## 4  -0.05484 -0.05434  0.03816 -0.037054  0.01912 -0.00702 -0.1271 1.427
## 5  -0.13170 -0.06071 -0.29496  0.321007  0.34528 -0.01053 -0.7624 0.832
## 6   0.00757  0.07639 -0.02067  0.016556  0.00431 -0.06391 -0.1247 1.503
## 7   0.30913 -0.02873 -0.17624  0.098046  0.01841 -0.33692  0.5223 1.833
## 8   0.21946  0.08754 -0.27888 -0.023209  0.13021  0.03967  0.3357 1.404
## 9   0.13875 -0.39245  0.24789 -0.095989 -0.33213  0.43727 -0.7424 2.247
## 10  0.00896 -0.00840  0.03031 -0.065439 -0.03105  0.06500  0.1259 1.510
## 11  0.01999 -0.00842 -0.03664 -0.028349  0.03263  0.01709  0.0652 1.712
## 12 -0.00750  0.03857 -0.02502 -0.006189  0.03843 -0.07036  0.1018 1.552
## 13  0.38604  1.15313 -0.46370 -0.280101 -0.07569  0.62402  1.5202 0.849
## 14  0.05803  0.14674 -0.06693 -0.267915  0.02820  0.15995 -0.4635 0.666
## 15 -0.00363 -0.01163 -0.00452  0.015736  0.00432  0.01298  0.0325 2.626
## 16 -0.00225 -0.01639 -0.01528  0.001660  0.02379 -0.01275 -0.0393 1.830
## 17 -0.12164 -0.13408  0.09727  0.046846  0.00732 -0.05621 -0.2407 1.263
## 18  0.10178 -0.52295  1.32112  0.649441 -1.49317 -0.28097  1.6518 0.582
## 19  0.13407  0.01695  0.03299  0.120774 -0.13260 -0.20091 -0.2783 1.861
## 20 -1.45139  0.52312  0.48744 -0.659767  0.56448 -0.22455 -1.6603 0.532
## 21 -0.10549  0.05531  0.20255 -0.078981 -0.13391 -0.01540  0.3002 1.262
## 22  0.02184  0.00905 -0.02445  0.000114  0.00573  0.01667 -0.0360 1.676
## 23 -0.01427 -0.02469 -0.05581 -0.110167  0.09298  0.07897  0.1828 1.484
## 24  0.27548 -0.41443 -0.05045  0.233820 -0.12899  0.13793 -0.5995 1.015
## 25  0.07858  0.12946 -0.07860 -0.001309 -0.01637  0.07674 -0.2093 1.601
## 26  0.19511 -0.00542 -0.17152  0.096971  0.02724 -0.02501 -0.2726 1.345
##      cook.d    hat inf
## 1  0.024359 0.4552   *
## 2  0.056375 0.2134
## 3  0.062543 0.0970
## 4  0.002810 0.0930
## 5  0.090599 0.1959
## 6  0.002711 0.1251
## 7  0.046668 0.3614
## 8  0.019242 0.1771
## 9  0.093879 0.4925   *
## 10 0.002764 0.1291
## 11 0.000746 0.2093
## 12 0.001813 0.1407
## 13 0.343804 0.4043   *
## 14 0.033036 0.0743
## 15 0.000185 0.4822   *
## 16 0.000271 0.2576
```

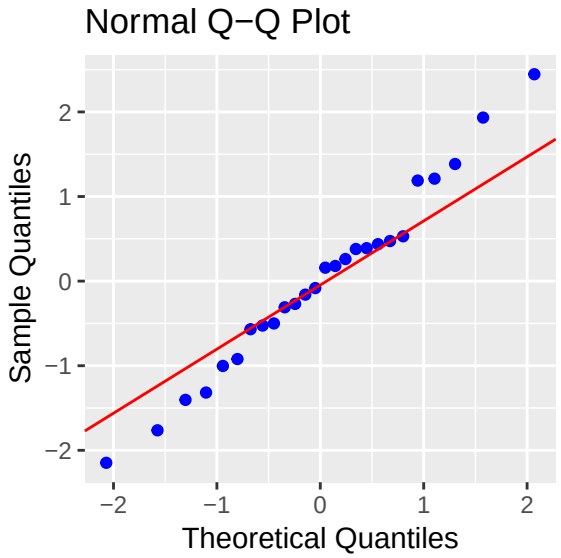
```
## 17 0.009877 0.0943
## 18 0.384723 0.3703   *
## 19 0.013469 0.3079
## 20 0.384029 0.3588   *
## 21 0.015284 0.1208
## 22 0.000227 0.1898
## 23 0.005799 0.1403
## 24 0.058079 0.1811
## 25 0.007614 0.1969
## 26 0.012712 0.1318
```

En regardant la distance de cook ainsi que le *covratio*, on remarque que certaine données influence la valeur des *beta* et la variance des estimateurs. Comme nous avons que 26 données, nous allons tout de même les conserver dans le modèle.

On reprend maintenant le modèle final du TP1 avec les variables *WT2, WT9, HT9 & ST9* Commençons par regarder les graphiques de résidus du modèle choisi.



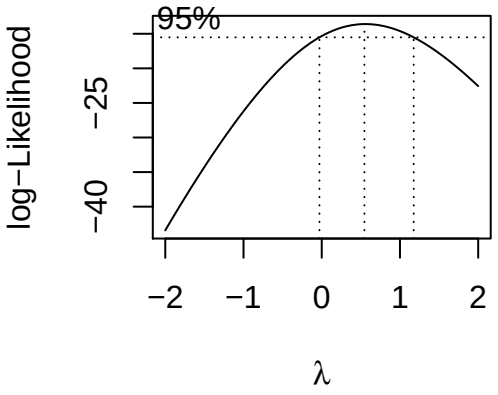
L'allure des résidus semble correct. Les graphiques ont pratiquement la même allure que pour le modèle complet. Les mêmes conclusions sont tirées.




```
## ------------------------------------------------
##         Test              Statistic       pvalue
## ------------------------------------------------
## Shapiro-Wilk               0.9856         0.9654
## Kolmogorov-Smirnov         0.1219         0.7912
## Cramer-von Mises           1.896          0.0000
## Anderson-Darling           0.1937         0.8844
## ------------------------------------------------
```

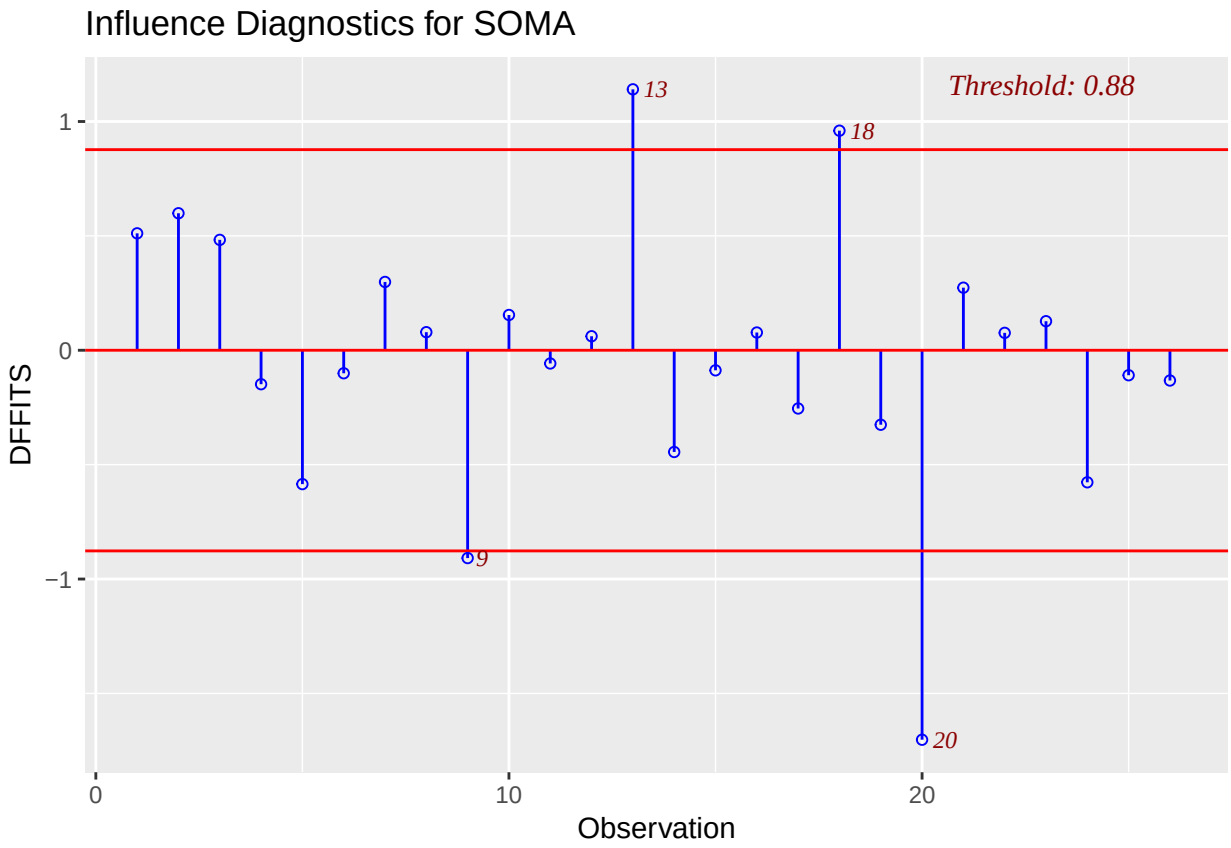
Le graphique de *QQ-plot* montre que les résidus semblent être distribué de façon normal. De plus le test de *Shapiro-Wilk* fait en sorte qu'on ne rejette pas l'hypothèse de normalité des résidus.

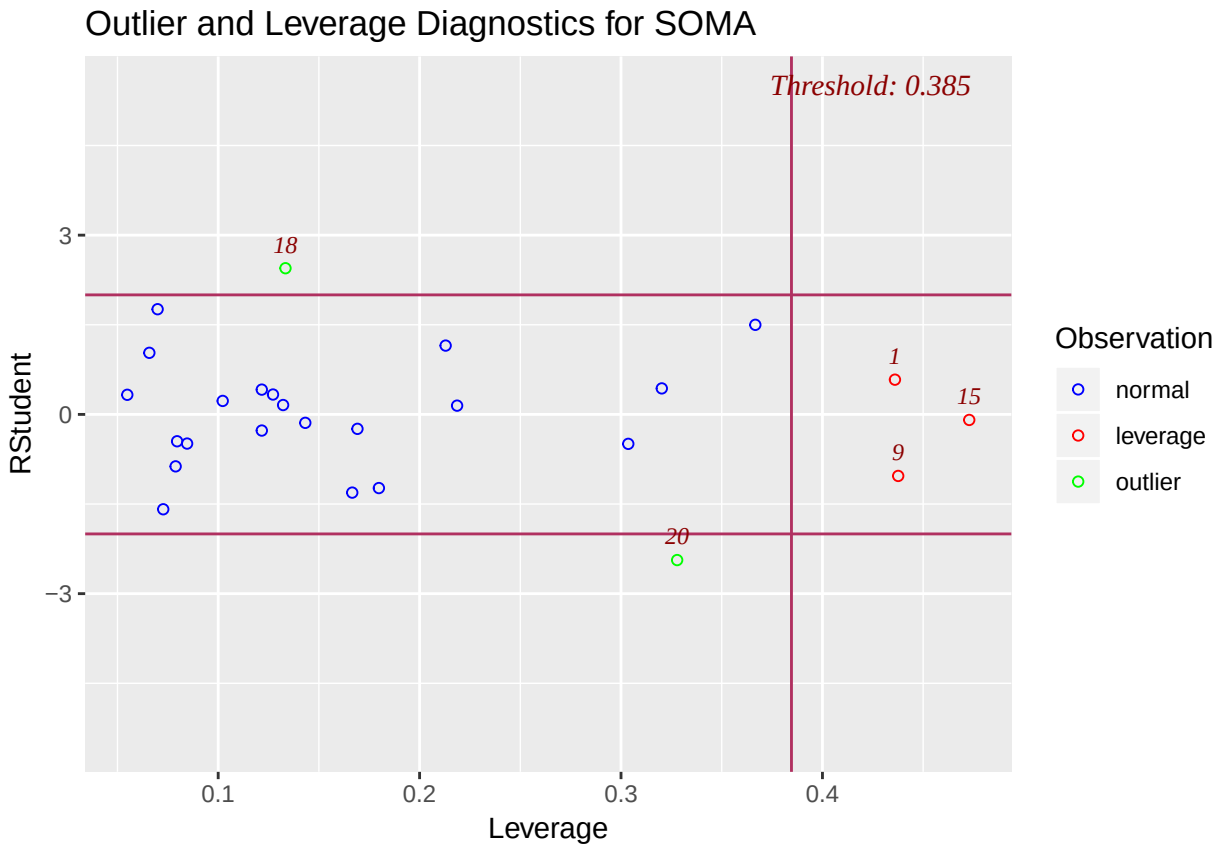
Regardons le graphique des valeurs de lambda avec la transformation de *Box-Cox*



Il serait aussi approprié de faire notre modèle de régression linéaire sur $\sqrt{Y}$.

Regardons maintenant certaines statistiques sur l'influence des données pour le modèle final.


Influence Diagnostics for SOMA
Threshold: 0.88
13
18
9
20
1
0
−1
0
10
20
DFFITS
Observation


Outlier and Leverage Diagnostics for SOMA
Threshold: 0.385
18
1
15
9
20
3
0
−3
0.1
0.2
0.3
0.4
RStudent
Leverage
Observation
normal
leverage
outlier

Les données 18 et 20 peuvent potentiellement causer des problèmes. Ils ont une certaines influences sur les valeurs prédites. Cependant, avec le graphique des résidus en fonction des h_{ii} on remarque que ces données sont des outliers, mais cependant il n'ont pas un grand effet de levier. Donc il est correct de les conserver pour faire le modèle.

```
## Influence measures of
##   lm(formula = SOMA ~ WT2 + WT9 + HT9 + ST9, data = data_tp1) :
##
##      dfb.1_  dfb.WT2  dfb.WT9  dfb.HT9  dfb.ST9   dffit cov.r   cook.d
## 1   0.12709 -0.23487  0.47943 -0.15795 -0.15854  0.5112 2.081 0.053958
## 2  -0.44765 -0.17207 -0.22983  0.51358  0.05993  0.5989 1.176 0.070651
## 3  -0.00248 -0.17859  0.03267  0.10874 -0.21093  0.4827 0.668 0.042359
## 4  -0.05117 -0.05396 -0.05041  0.07666 -0.00446 -0.1484 1.314 0.004569
## 5  -0.25504 -0.14122  0.29385  0.16317 -0.03067 -0.5851 1.015 0.066226
## 6  -0.00199  0.06030  0.01535 -0.01119 -0.05398 -0.1005 1.427 0.002115
## 7   0.15573 -0.05672  0.07089 -0.08626 -0.21501  0.2986 1.791 0.018553
## 8   0.04231 -0.00277  0.00227 -0.03607  0.00634  0.0790 1.315 0.001305
## 9   0.37558 -0.42608 -0.15838 -0.28897  0.59851 -0.9083 1.753 0.164535
## 10  0.03324  0.00240 -0.08705 -0.01773  0.08585  0.1543 1.392 0.004960
## 11 -0.00334  0.02354  0.02644 -0.01091 -0.01486 -0.0577 1.482 0.000699
## 12 -0.01355  0.01988 -0.00226  0.01815 -0.04523  0.0613 1.462 0.000788
## 13  0.14996  0.83432 -0.18425 -0.42927  0.46107  1.1407 1.183 0.245635
## 14  0.02916  0.12787 -0.25439 -0.02393  0.14988 -0.4446 0.760 0.036862
## 15  0.01777  0.03803 -0.04435 -0.00455 -0.03441 -0.0877 2.416 0.001613
## 16  0.02296  0.04905 -0.00696 -0.03928  0.03039  0.0776 1.625 0.001265
## 17 -0.09941 -0.12444  0.04286  0.11847 -0.05521 -0.2545 1.151 0.013108
## 18  0.79325 -0.08126  0.50038 -0.77887 -0.15913  0.9599 0.398 0.148930
## 19  0.19959  0.03505  0.13896 -0.17844 -0.23405 -0.3254 1.725 0.021974
## 20 -1.48555  0.78232 -0.76530  1.33176 -0.19524 -1.7025 0.517 0.469262
## 21 -0.01432  0.16173 -0.12379  0.00950  0.00319  0.2736 1.055 0.014930
## 22 -0.03375 -0.00254 -0.00758  0.04412 -0.04192  0.0761 1.404 0.001212
## 23 -0.03351 -0.03366 -0.07664  0.05491  0.05424  0.1270 1.422 0.003371
## 24  0.27563 -0.44253  0.23213 -0.21819  0.12946 -0.5772 1.079 0.065015
## 25  0.02652  0.06152  0.00381 -0.05468  0.03939 -0.1091 1.514 0.002491
## 26  0.08089 -0.04176  0.07182 -0.07819 -0.02519 -0.1322 1.319 0.003633
##       hat inf
## 1  0.4360   *
## 2  0.2129
## 3  0.0699
## 4  0.0846
## 5  0.1666
## 6  0.1217
## 7  0.3202   *
## 8  0.0549
## 9  0.4376   *
## 10 0.1217
## 11 0.1432
## 12 0.1322
## 13 0.3667
## 14 0.0727
## 15 0.4729   *
## 16 0.2186
```

```
## 17 0.0789
## 18 0.1334
## 19 0.3036   *
## 20 0.3279   *
## 21 0.0658
## 22 0.1022
## 23 0.1272
## 24 0.1798
## 25 0.1692
## 26 0.0796
```

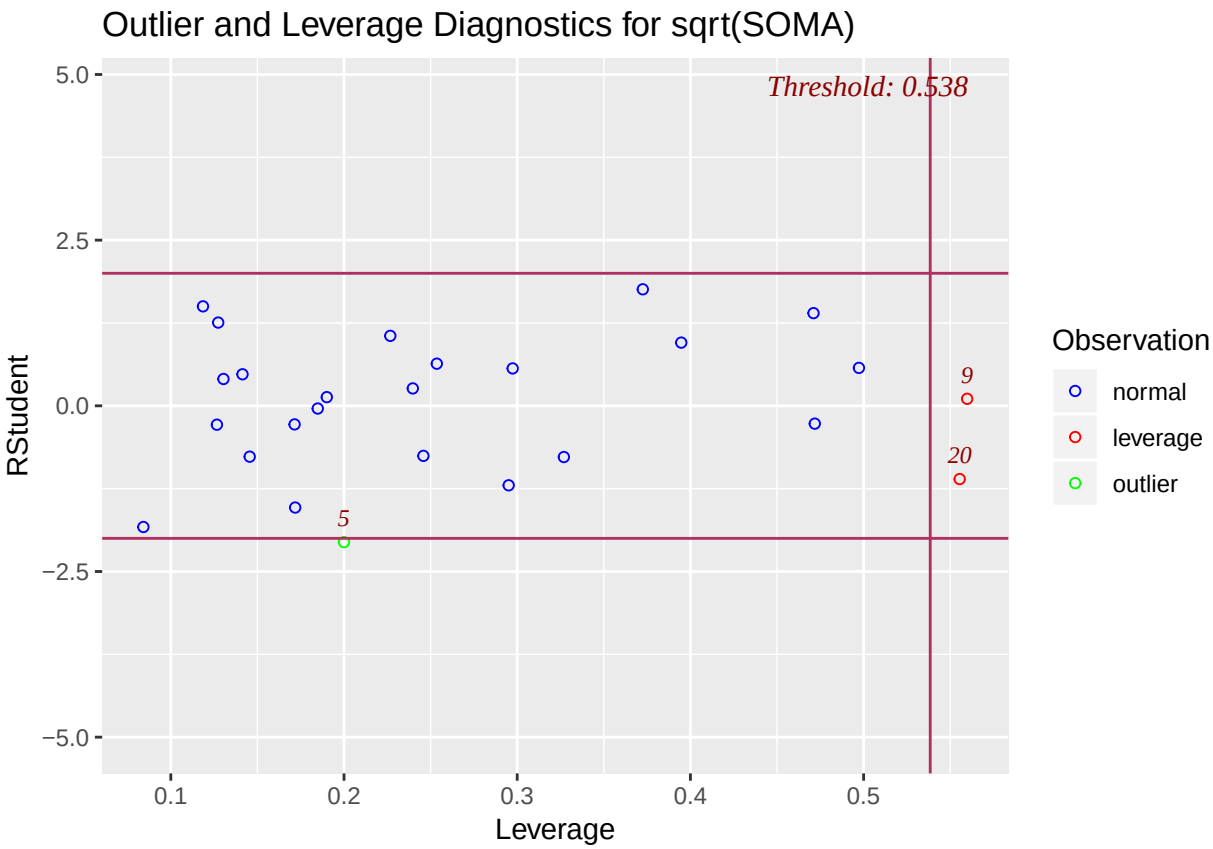
En regardant la distance de cook ainsi que le *covratio*, on remarque que certaines données influencent la valeur des *beta* et la variance des estimateurs. Comme nous avons que 26 données, nous allons tout de même les conserver dans le modèle.

Le modèle final est celui où l'on effectue la régression sur $\sqrt{Y}$. De plus on ajoute la variable *HT2* ainsi que l'intéraction entre *WT2 et HT2* dans le modèle. Nous avons maintenant une relation positive entre le poids à 2ans et l'indice d'obésité. Voici maintenant la prédiction du modèle.

```
## [1] "Prédiction pour Y"
```

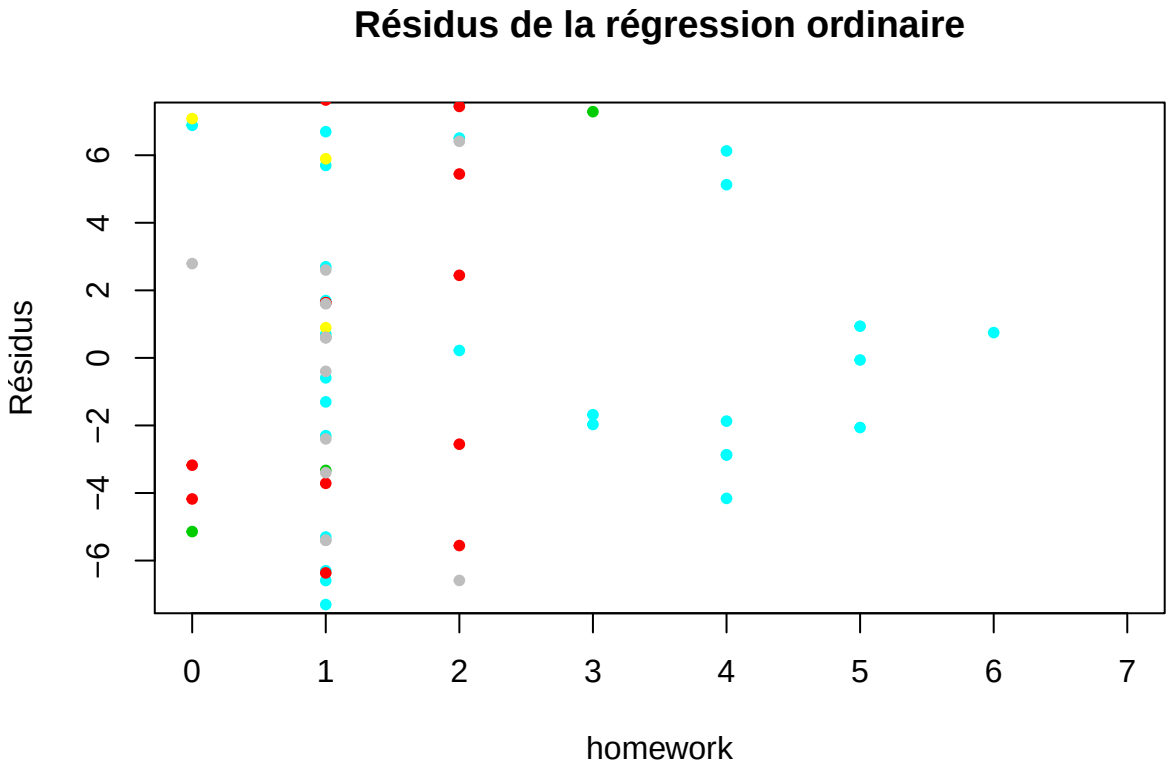
```
##        fit      lwr      upr
## 1 7.544987 3.842645 12.48464
```

Un graphique intéressant à regarder est les résidus en fonction de h_{ii}. Il n'y a maintenant plus d'observation qui sont vraiment outlier.



##P-2

Tout d'abord, il est important de commencer l'analyse en considérant les effets aléatoires possibles dans le modèle. Puisque les variables `ratio` ainsi que `meanses` ne varient pas en fonction de l'école, ces variables ne sont pas utilisable comme pente aléatoire dans le modèle. Par ailleurs, il serait probable que le nombre d'heures de travail de l'étudiant à la maison par semaine varient selon l'école. Afin de constaté ceci, il est possible de faire un modèle complet de régression standard et visualiser sur un graphique les résidus en fonction de la variable `homework` pour ces différentes écoles. Afin d'avoir une vue allégée du graphique, seulement les 5 premières écoles sont dans le graphique des résidus plus bas.



Bien qu'il soit difficile à analyser, on semble voir une certaine dépendance avec l'école.

(a)

Commençons en créant un modèle linéaire mixte sans considérer la variable `meanses`.

Alors, avec le modèle suivant:

$$Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_{ij1} + \beta_2 x_{i2} + (\beta_3 + \gamma_{i3}) x_{ij3} + \epsilon_{ij}$$

où
Y_{ij} = résultat de l'étudiant j dans l'école i en mathématique.
x_{ij1} = l'indicateur de 1 pour blanc pour l'étudiant j dans l'école i.
x_{i2} = est le ratio du nomde d'étudiants par enseignant dans les classes de l'école i.
x_{ij3} = le nombre d'heures de travail à la maison de l'étudiant j dans l'école i.

on peut tester si on obtient un AIC minimal avec les blocs de la matrice D non-structuré (UN) ou diagonale principale (UN(1)). Par simplicité et limitation du package `lme4`, la matrice V sera de composante de variance (VC).

```
##     VC et UN      VC et UN(1)
```

```
## AIC 3643.262          3666.629
```

Alors, on choisit le modèle avec les blocs de la matrice D non-structuré (UN).

On peut poursuivre en effectuant un test de la nécessité des effets aléatoires avec le test suivant:

$$H_0 : Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_{ij1} + \beta_2 x_{i2} + \beta_3 x_{ij3} + \epsilon_{ij}$$
$$H_1 : Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_{ij1} + \beta_2 x_{i2} + (\beta_3 + \gamma_{i3}) x_{ij3} + \epsilon_{ij}$$

Et nous pourrons obtenir la p-value avec cette équation:

$$p = 0.5P[\chi_1^2 > \epsilon] + 0.5P[\chi_2^2 > \epsilon]$$

puisque nous avons que $m_0 = 1$(1 variance) et $m_1 = 3$(2 variances et 1 covariance). On obtient une valeur de $\epsilon = 2(\ell_1 - \ell_0) = 89.74556$ et on obtient une p-value très près de 0. Alors, on rejette le l'hypothèse nulle et nous ne pouvons pas simplifier le modèle.

Il est possible d'effectuer une sélection des effets fixes avec la méthodes backward, où nous rejetons la variable `ratio`.

```
## [1] "Résumé des coefficients des effets fixes"
```

```
## (Intercept)       white    homework
##   44.019758    3.300014    1.903103
```

```
## [1] "Résumé des effets aléatoires"
```

```
## $schid
##       (Intercept)   homework
## 6053    4.4055629  0.2448821
## 6327   13.8592245 -7.8453317
## 6467   -5.0779757  4.1820446
## 7194    5.1343844 -4.2449442
## 7472    3.1258114 -4.8279724
## 7474   10.6523351 -4.4787070
## 7801    7.6905919 -4.6275450
## 7829    4.8144131 -4.5816031
## 7930   -7.2576202  5.4227929
## 24371  -6.2371857  2.4271454
## 24725 -11.3814602  3.3947014
## 25456   5.5535448 -4.9940643
## 25642   4.1265536 -3.3944956
## 26537  -0.8034290  1.8455423
## 46417   0.5194043  2.2395233
## 47583  -4.2400423  1.1974074
## 54344  -7.4448939  1.3461689
## 62821  10.7107189 -0.4949450
## 68448  -7.4166165  3.7705810
## 68493  -6.2215976  3.2229537
## 72080  -1.9702841  2.3861839
## 72292  -8.7018273  4.1025471
## 72991  -3.8396126  3.7071342
```

Alors, pour répondre aux questions de l'énoncé (a), l'effet du nombre d'heure est positif, alors plus un étudiant passe d'heure d'étude par semaine, meilleur sont ses résultats en mathématiques. De plus, l'augmentation moyenne sur le résultat en mathématique d'une heure d'étude supplémentaire par semaine sur l'ensemble des étudiant est de 1.903103($\hat{\beta}_3$). Par ailleurs, puisque nous avons rejeté le modèle sans la pente aléatoire devant la variable x_{ij3}, on peut conclure qu'il est raisonnable de croire que l'effet du nombre d'heures varient d'une école à une autre.

(b)

En procédant de la même façon que les étapes en (a) et en y rajoutant la variable `meanses`, on conserve la structure des variances VC et UN puisqu'on obtient un AIC minimal. De plus, en effectuant le test de la nécessité des effets aléatoires, on obtient un résultat identique, c'est-à-dire qu'on conserve la pente aléatoire devant la variable x_{ij3} (nombre d'heures de travail par semaine tel que défini plus haut).

En utilisant la méthode backward pour la sélection des effets fixes, on rejette, en 2 étapes, la variable `ratio` ainsi que l'interaction `meanses*homework`

```
## [1] "Résumé des coefficients des effets fixes"
```

```
## (Intercept)       white    homework     meanses
##   44.702203    3.114922    1.925085    4.892483
```

```
## [1] "Résumé des effets aléatoires"
```

```
## $schid
##       (Intercept)   homework
## 6053    0.7955302  0.1671122
## 6327   10.6841274 -6.9633524
## 6467   -5.5008579  4.2295462
## 7194    6.1971941 -4.2164797
## 7472    4.9610280 -4.7656103
## 7474   10.8284267 -4.4721868
## 7801    8.5316493 -4.6276834
## 7829    7.3630183 -4.3946052
## 7930   -8.3637902  5.3596980
## 24371  -5.0440909  2.4523099
## 24725  -8.6148618  3.3969895
## 25456   5.9455044 -4.7524984
## 25642   8.3953235 -3.8543474
## 26537  -4.6855665  1.4033056
## 46417   0.7681519  2.2498679
## 47583  -4.3887822  1.0435040
## 54344  -8.2255629  1.3886616
## 62821   5.4262853 -0.5889649
## 68448  -7.1480214  3.7467238
## 68493  -4.7023652  3.1041034
## 72080  -1.4669293  2.3654021
## 72292  -7.0099571  4.0740467
## 72991  -4.7454535  3.6544576
```

Alors, pour répondre aux questions du numéro, l'ajout de la variable `meanses` au modèle n'a pas diminué le besoin d'inclure des effets aléatoires. De plus, l'effet moyen sur le résultat en mathématique d'une heure d'étude supplémentaire par semaine sur l'ensemble des étudiants est environ le même qu'obtenu précédemment, soit $\hat{\beta}_3 = 1.925085$. Par ailleurs, puisque nous avons rejeté le modèle sans la pente aléatoire devant la variable x_{ij3}, on peut conclure qu'il est raisonnable de croire que l'effet du nombre d'heures varient d'une école à une autre.
Alors, le fait d'inclure ou non une variable unique à chaque école dans le modèle n'a pas réduit l'importance d'avoir des effets aléatoires dans le modèle et n'a non plus pas changé les conclusions.