Code

Analyse de l’impact de quelques facteurs sur le poids de bébés à la naissance

Florent VALBON, Julien MASSIP (TP B1)

23 décembre 2018

Le but de ce rapport est d’établir des liens entre le poids à la naissance de bébés et plusieurs facteurs :

- le poids de naissance du bébé (en grammes) (BWT = birth weight),
- tabagisme durant la grossesse (Y=oui; N=non) (SMOKE)
- race de la mère (1=blanche; 2=noire; 3=autre) (RACE)
- antécédents d’hypertension (Y=oui; N=non) (HT)
- présence d’irritabilité utérine (Y=oui; N=non) (UI)

Voici une partie du tableau de données utilisé pour l’étude :

	BWT <int>	SMOKE <fctr>	RACE <int>	HT <fctr>	UI <fctr>
1	2523	N	2	N	Y
2	2551	N	3	N	N
3	2557	Y	1	N	N
4	2594	Y	1	N	Y
5	2600	Y	1	N	Y
6	2622	N	3	N	N

6 rows

1 Plan d’expérience

Observons dans un premier temps la répartition des individus dans les différents groupes :

```
, , HT = N, UI = N

     RACE
SMOKE  1  2  3
    N 39 11 43
    Y 39  9  8

, , HT = Y, UI = N

     RACE
SMOKE  1  2  3
    N  1  2  4
    Y  4  1  0

, , HT = N, UI = Y

     RACE
SMOKE  1  2  3
    N  4  3  8
    Y  9  0  4

, , HT = Y, UI = Y

     RACE
SMOKE  1  2  3
    N  0  0  0
    Y  0  0  0
```

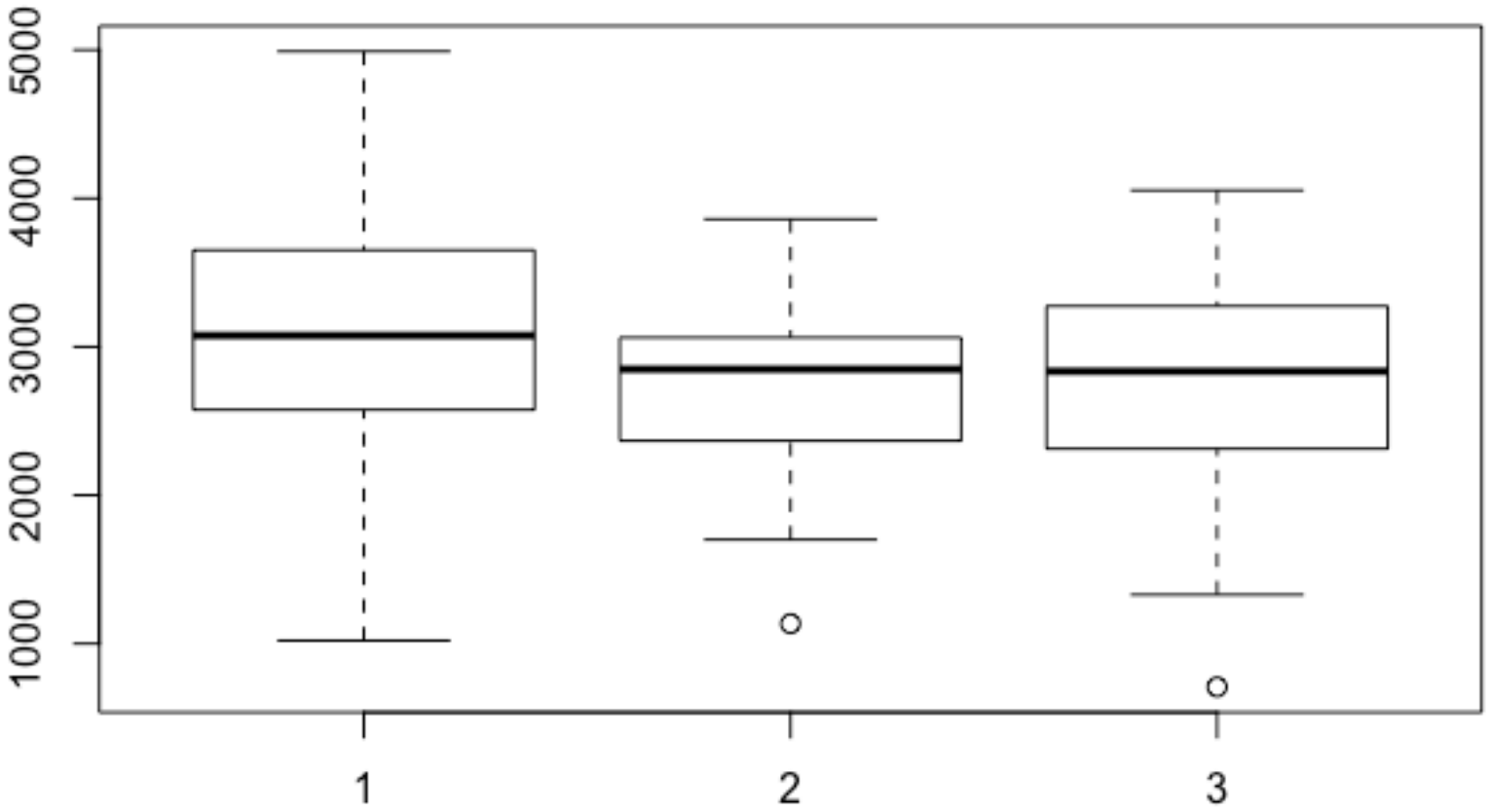
On voit que nous n’avons aucun individu ayant à la fois fait de l’hypertension et eu des irritations utérines, ce qui empêchera certaines analyses. De plus, le relativement faible nombre d’individus ayant des antécédents d’hypertension ou d’irritabilité utérine diminuera la fiabilité des tests sur l’impact de ces facteurs sur le poids des bébés. On dit ainsi que le plan d’expérience n’est ni complet, ni équilibré.

2 Analyse à un facteur

Nous voulons commencer par analyser la pertinence de chaque facteur pris individuellement à l’aide d’une analyse de la variance (ANOVA).

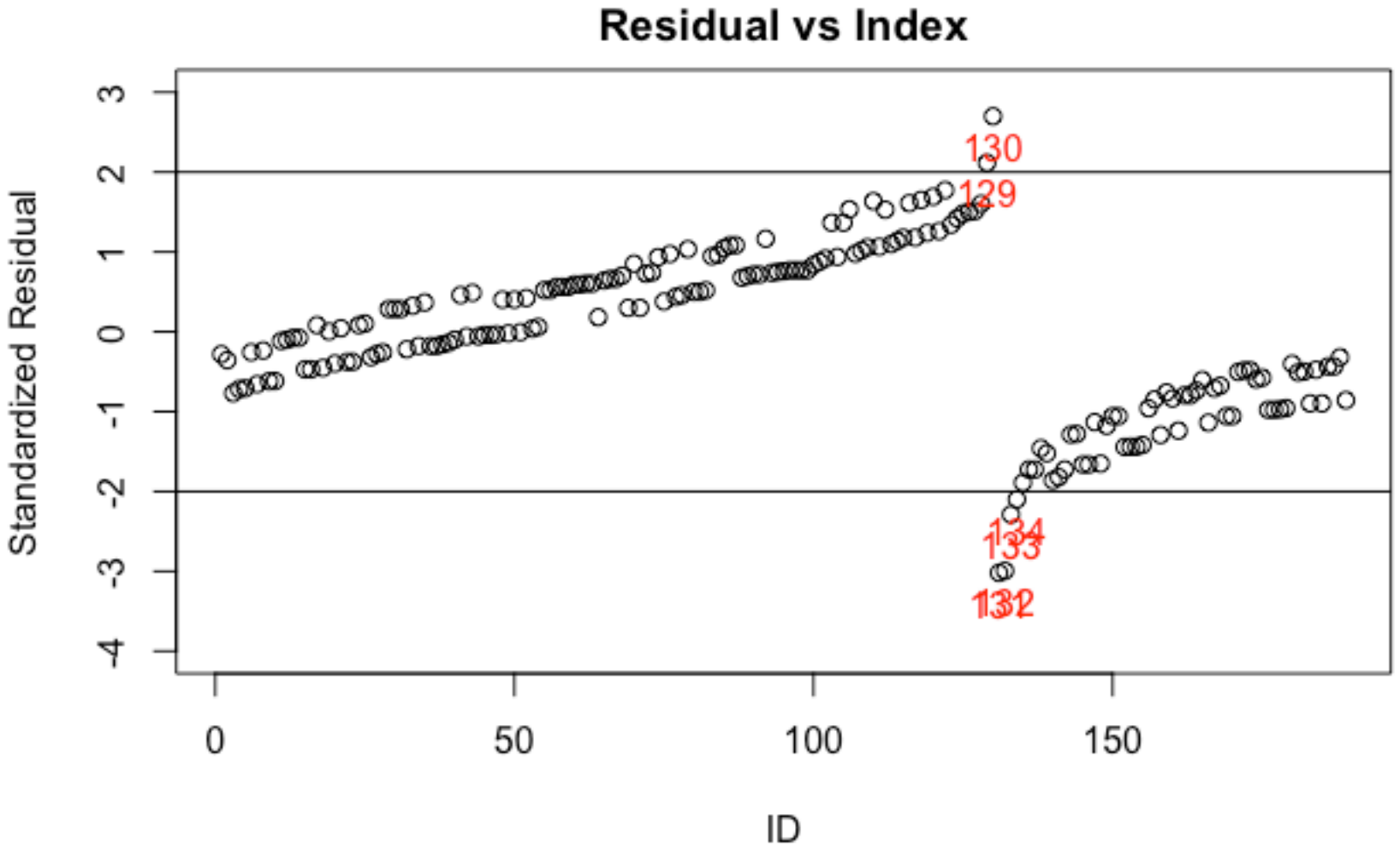
2.1 Race

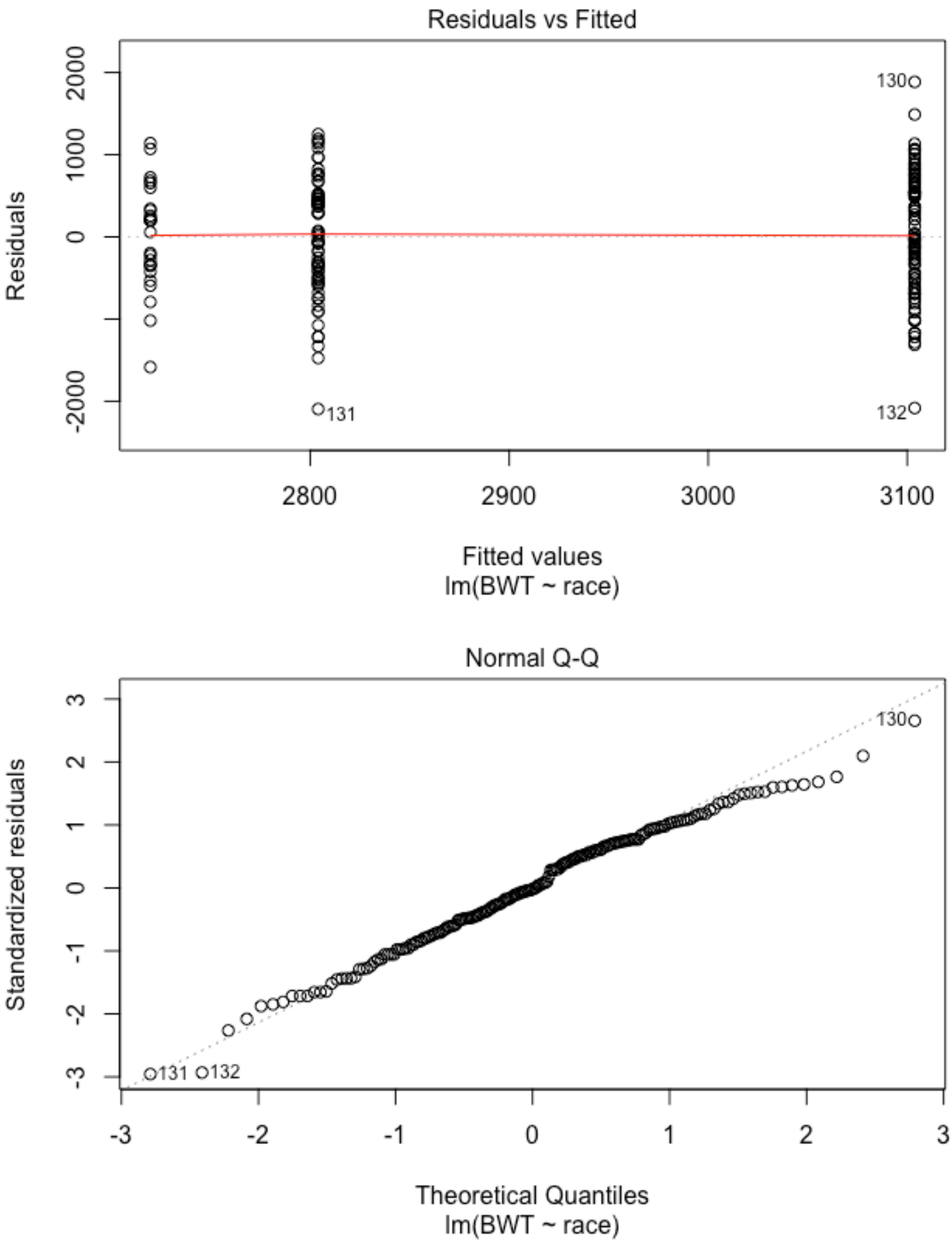
Analysons d’abord à l’aide d’un boxplot l’impact de la race de la mère sur le poids du bébé :

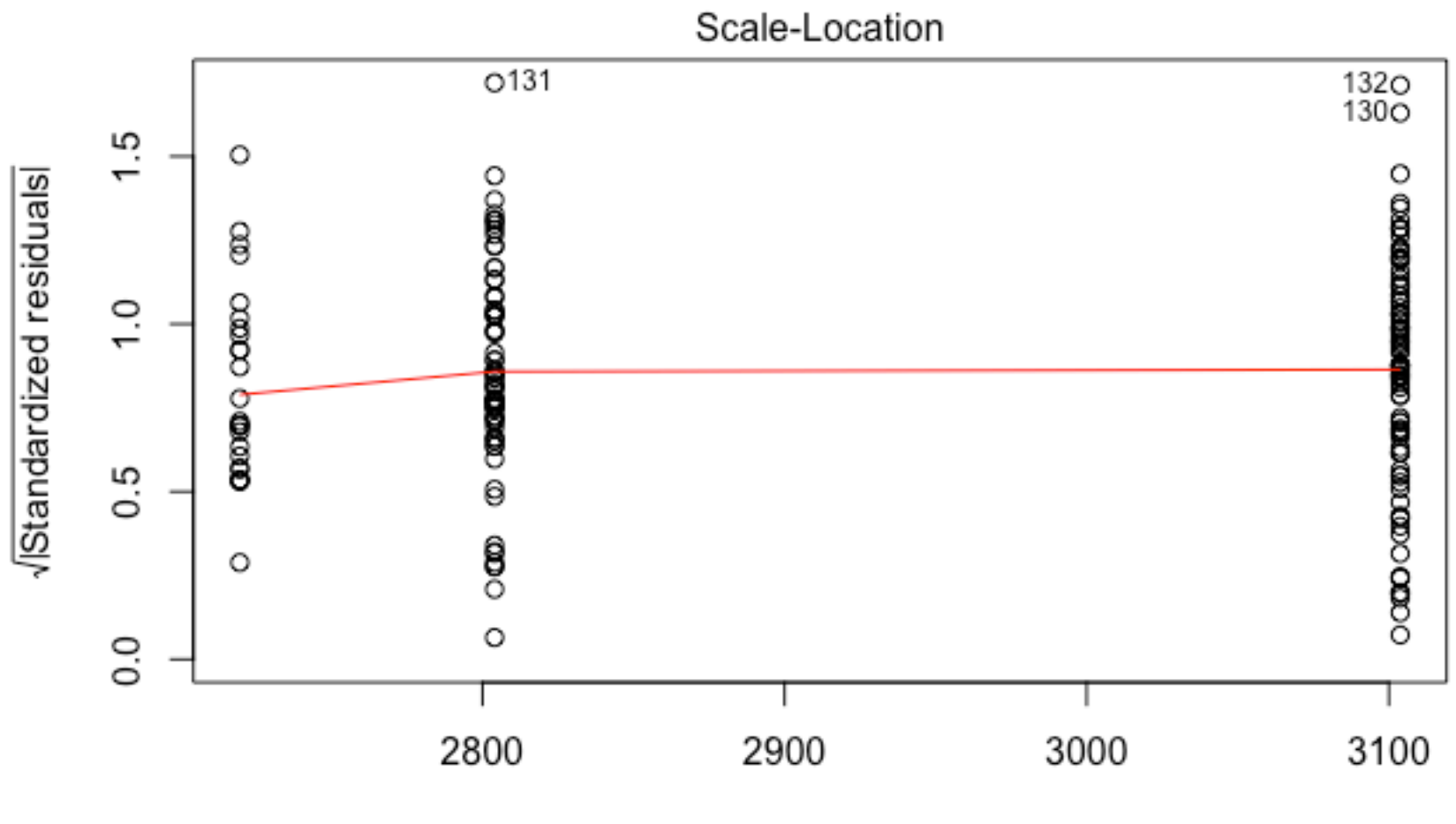


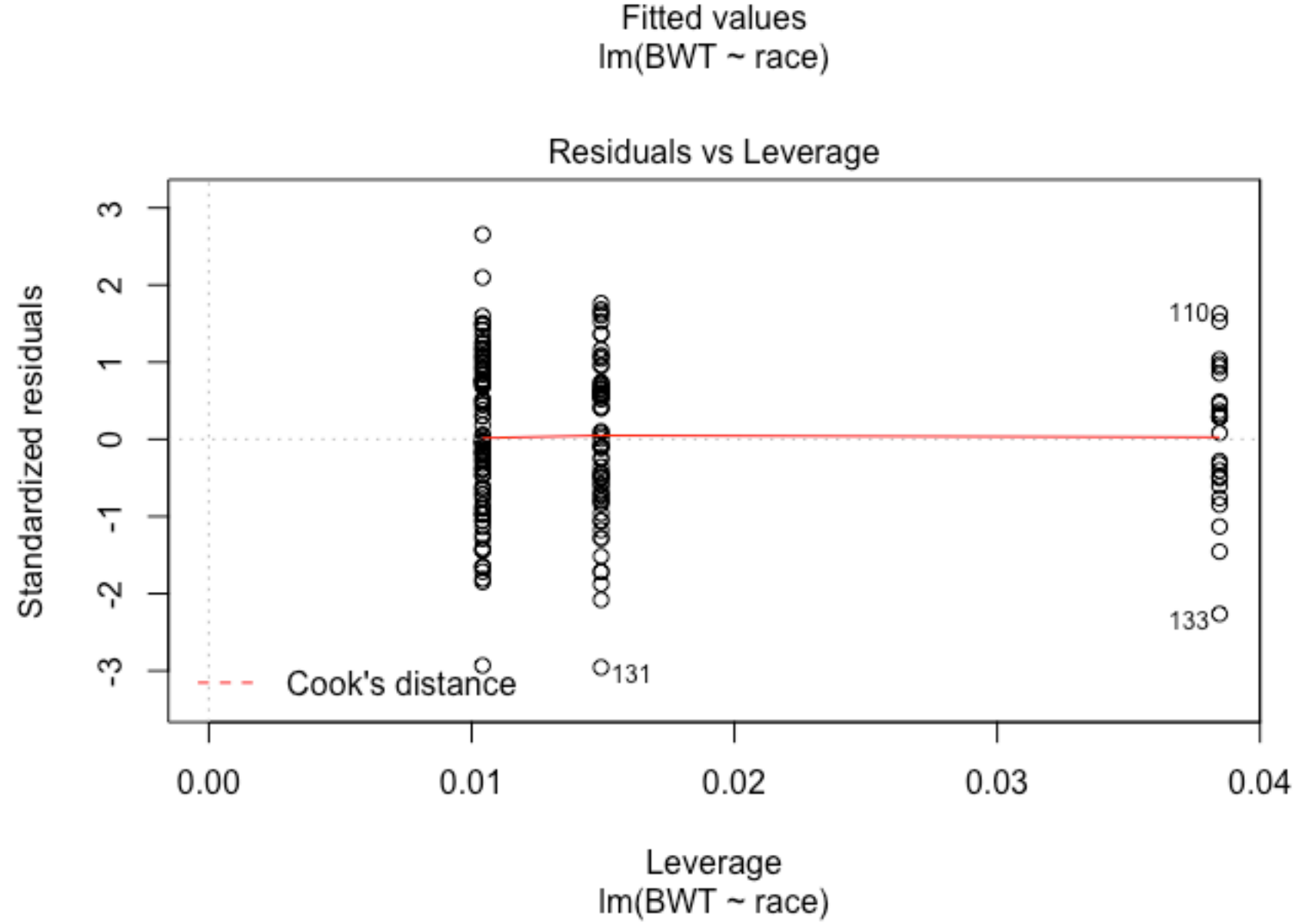
Il semblerait à première vue que la race de la mère ait un impact sur le poids du bébé. Cependant, avant de procéder à l'ANOVA pour le confirmer, nous devons vérifier que le modèle linéaire gaussien associé est valide.

Nous devons pour cela vérifier plusieurs hypothèses : les résidus sont gaussiens, de même loi et indépendants, nous voulons aussi savoir si nous avons des valeurs aberrantes ou isolés qui fausseraient le modèle.




Residuals vs Fitted
Residuals
2000
1000
0
-2000
130
131
132
2800
2900
3000
3100
Fitted values
lm(BWT ~ race)
Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
130
131
132
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(BWT ~ race)





On peut voir sur le Normal Q-Q que les points sont suffisamment alignés sur la première bissectrice pour accepter l'hypothèse de normalité des résidus, ainsi que l'hypothèse que les résidus suivent la même loi.

On observe qu'il n'y a aucun point dans le dernier graphique dont la valeur dépasse le seuil $\frac{3p}{n} = 0.048$: on en déduit qu'il n'y a pas de points leviers.

De plus, le premier graphique indique que quelques valeurs sont aberrantes, nous enlèverons les données d'identifiants $131, 132$ qui sont suffisamment éloignés de la bande $[-2, 2]$.

Cependant, ce graphique montre clairement une structure pour les résidus studentisés, ce qui compromet l'hypothèse d'indépendance.

Nous effectuerons malgré cela une ANOVA en prenant compte que notre modèle est peut-être faux.

L'ANOVA sur le modèle avec deux valeurs supprimés donne :

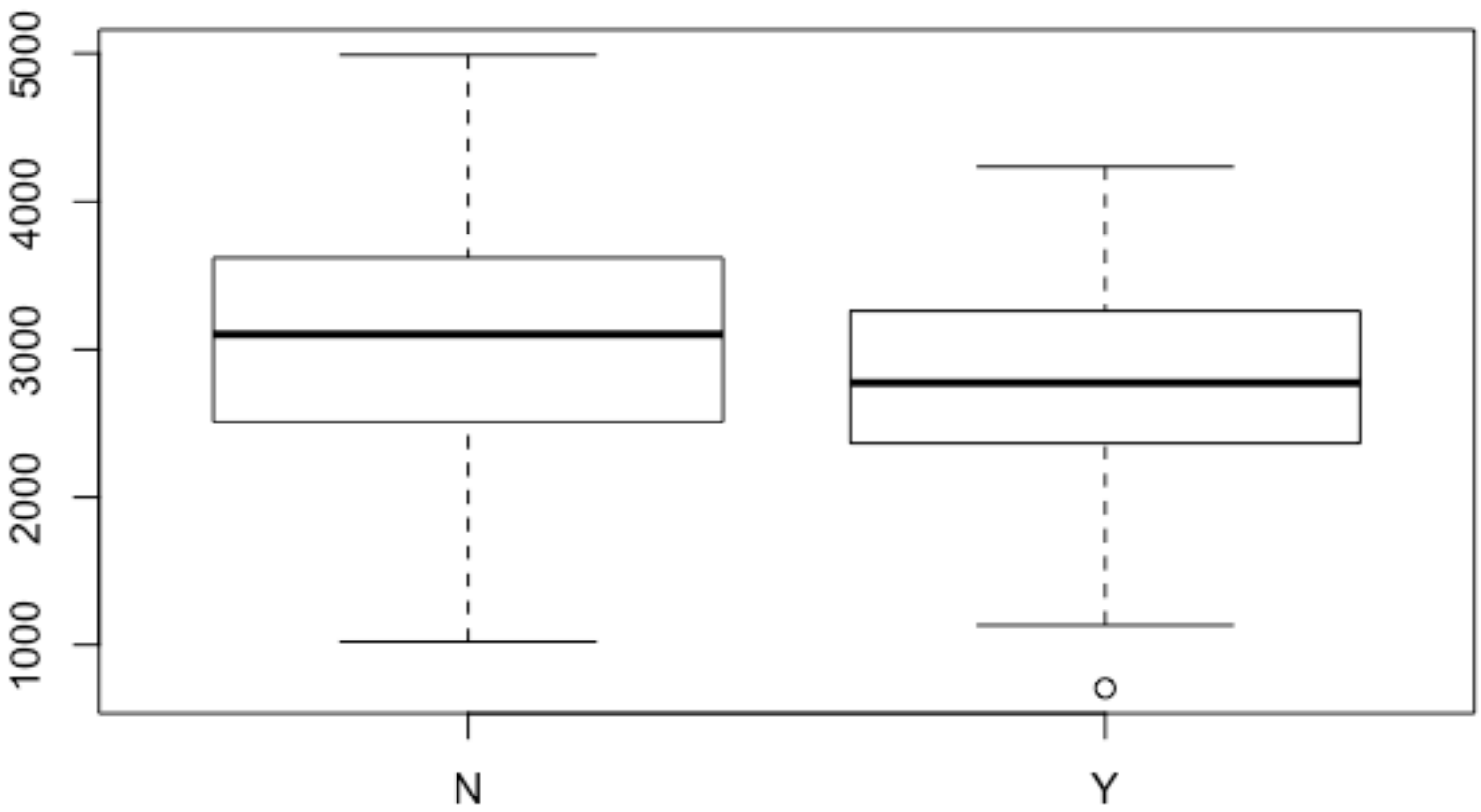
```
             Df   Sum Sq Mean Sq F value Pr(>F)
race.modif    2  5118538 2559269   5.475 0.0049 **
Residuals   184 86007391  467431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On conclus par l'ANOVA avec une p-valeur de 0.0049, que la race a un impact sur le poids du bébé à la naissance.

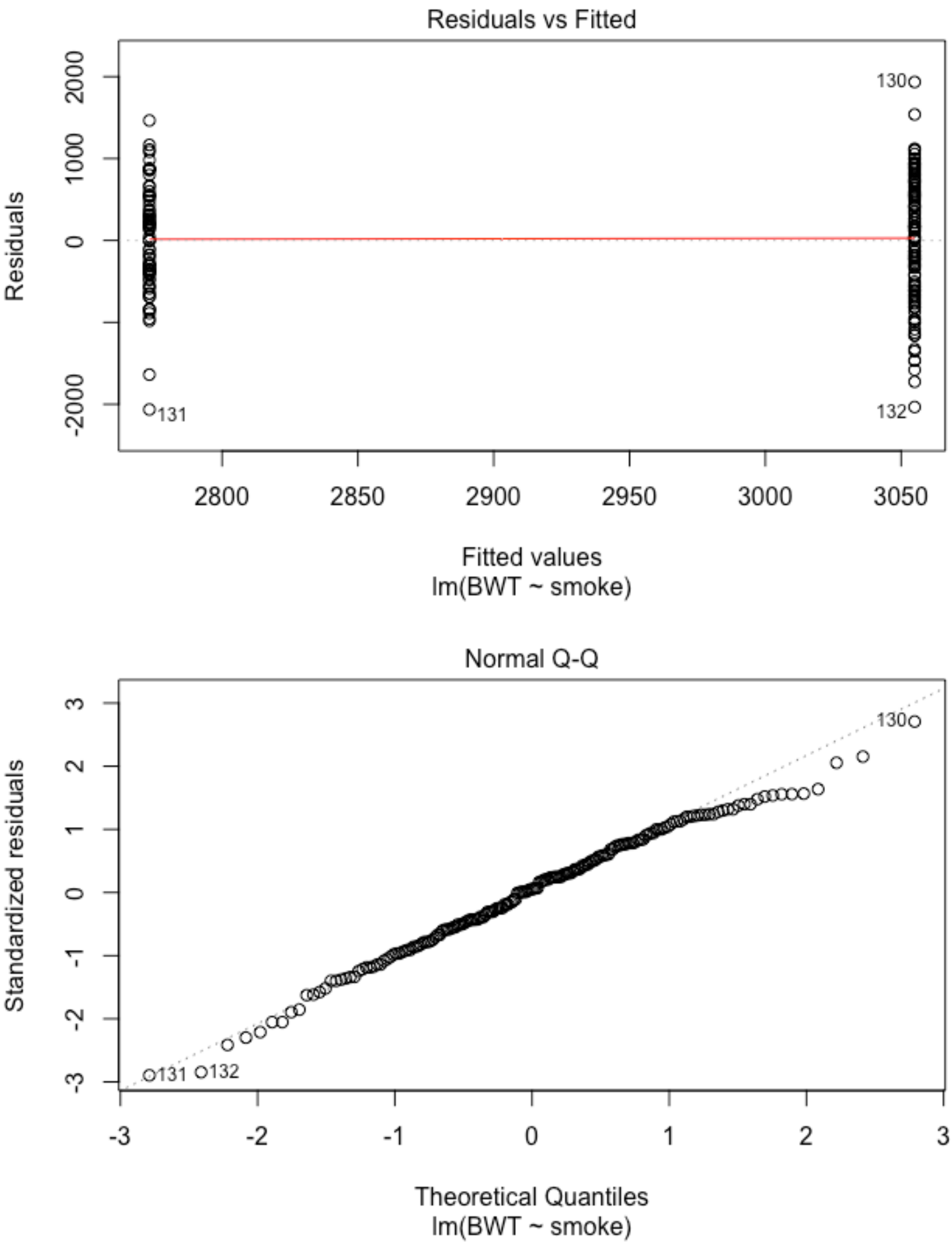
2.2 Fumeur

Analysons si le fait que la mère fume a un effet sur le poids du bébé à la naissance.

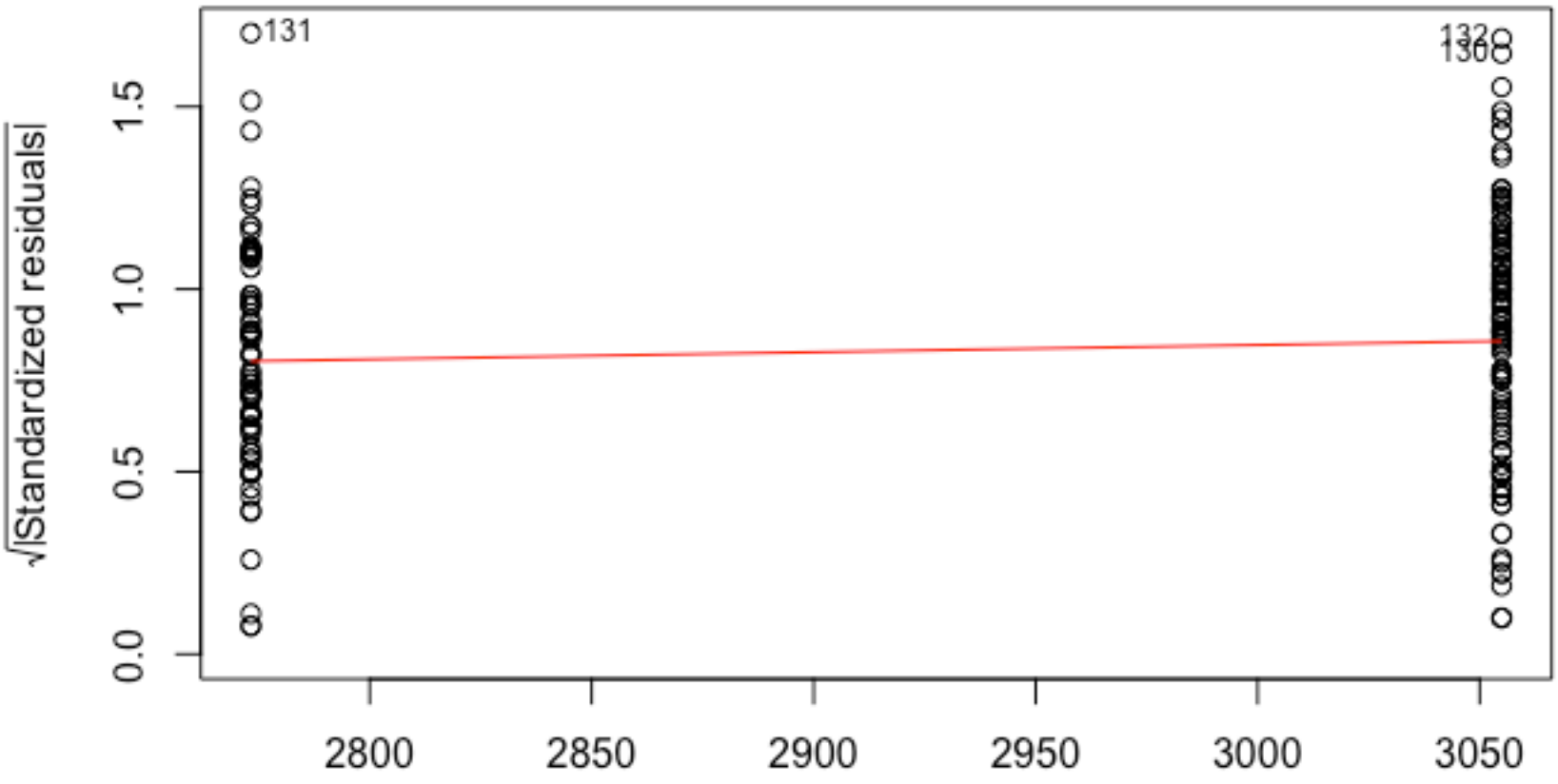
À priori il semblerait que ce soit le cas d'après le boxplot. Nous allons le confirmer avec l'ANOVA.



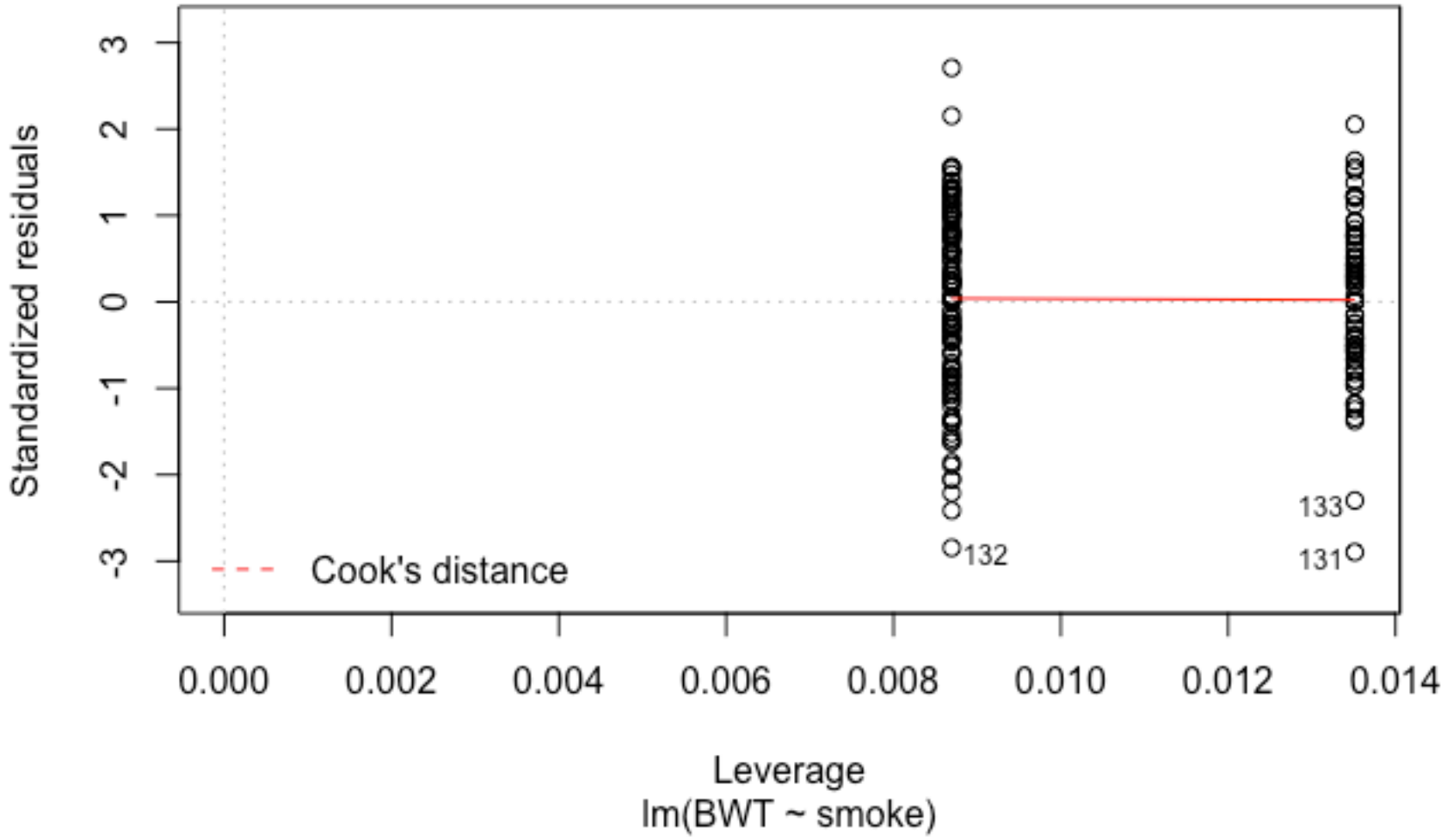
Vérifions tout d'abord si le modèle linéaire associé est valide. Les graphiques suivant montrent clairement qu'ils y a trois valeurs aberrantes (d'indices $130, 131, 132$). Là encore, les valeurs d'indices 131 et 132 sont encore aberrantes. Il n'y a pas de points isolés en revanche et les résidus sont gaussiens de même variance. Cependant on observe ici aussi une structure dans les résidus studentisés contredisant l'hypothèse d'indépendance des résidus. Nous effectuons donc une ANOVA en enlevant ces trois points.

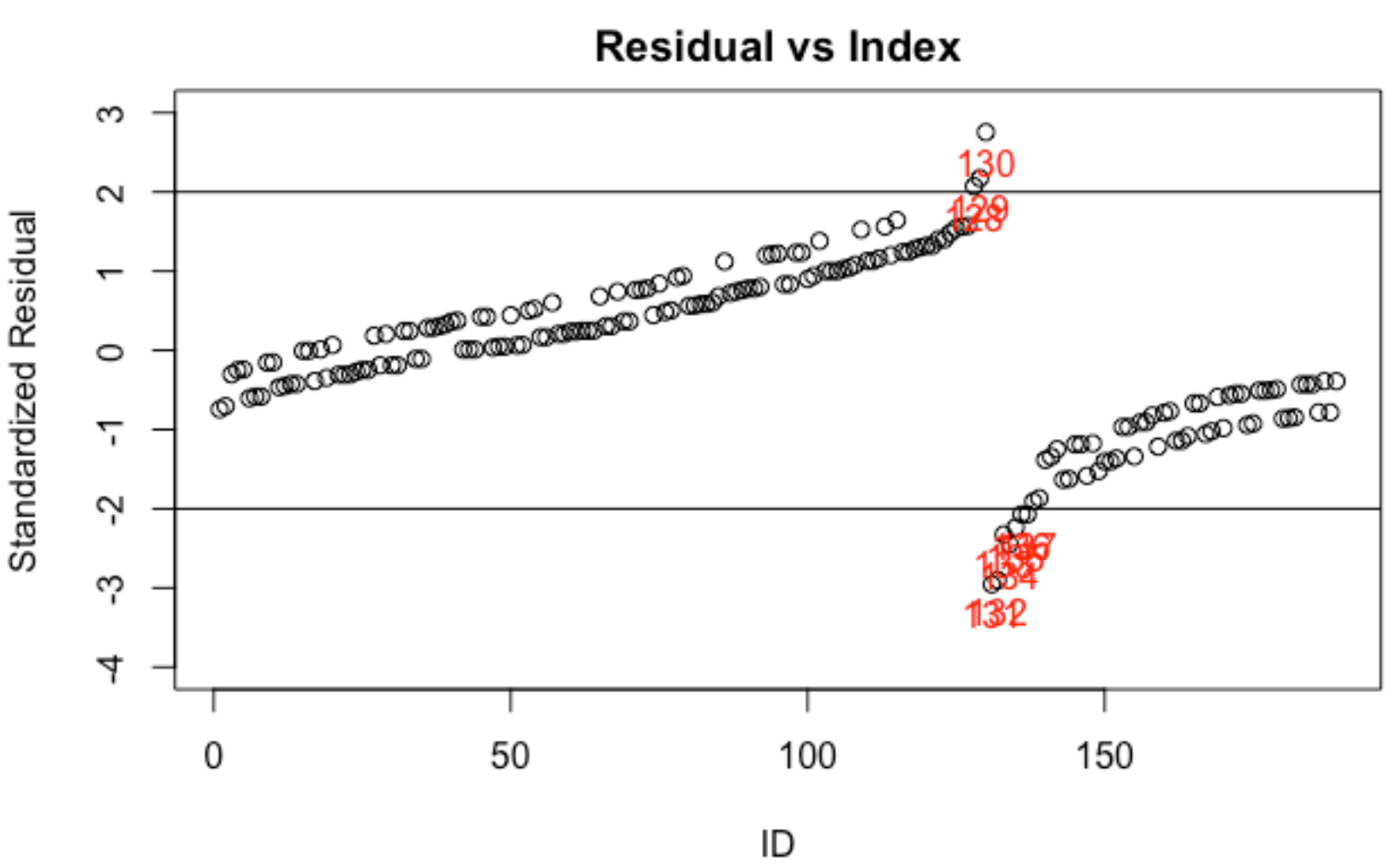

Residuals vs Fitted
Residuals
2000
1000
0
-2000
130
131
132
2800
2850
2900
2950
3000
3050
Fitted values
lm(BWT ~ smoke)
Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
130
131
132
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(BWT ~ smoke)

Scale-Location



Residuals vs Leverage





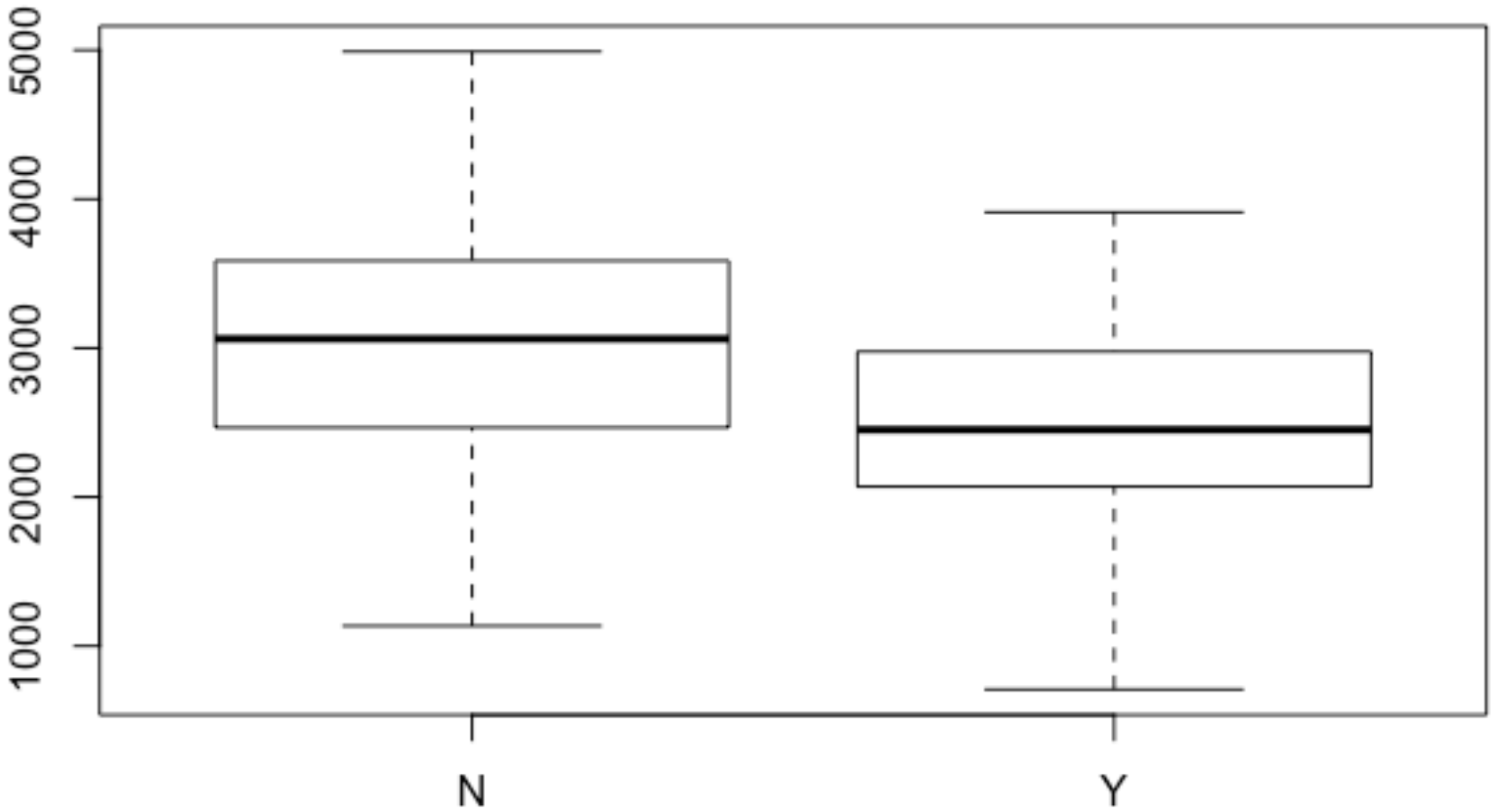
L’ANOVA sur le modèle modifié est résumé ici :

```
              Df   Sum Sq Mean Sq F value Pr(>F)
SMOKE.modif    1  2868268 2868268   6.272 0.0131 *
Residuals    184 84142716  457297
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

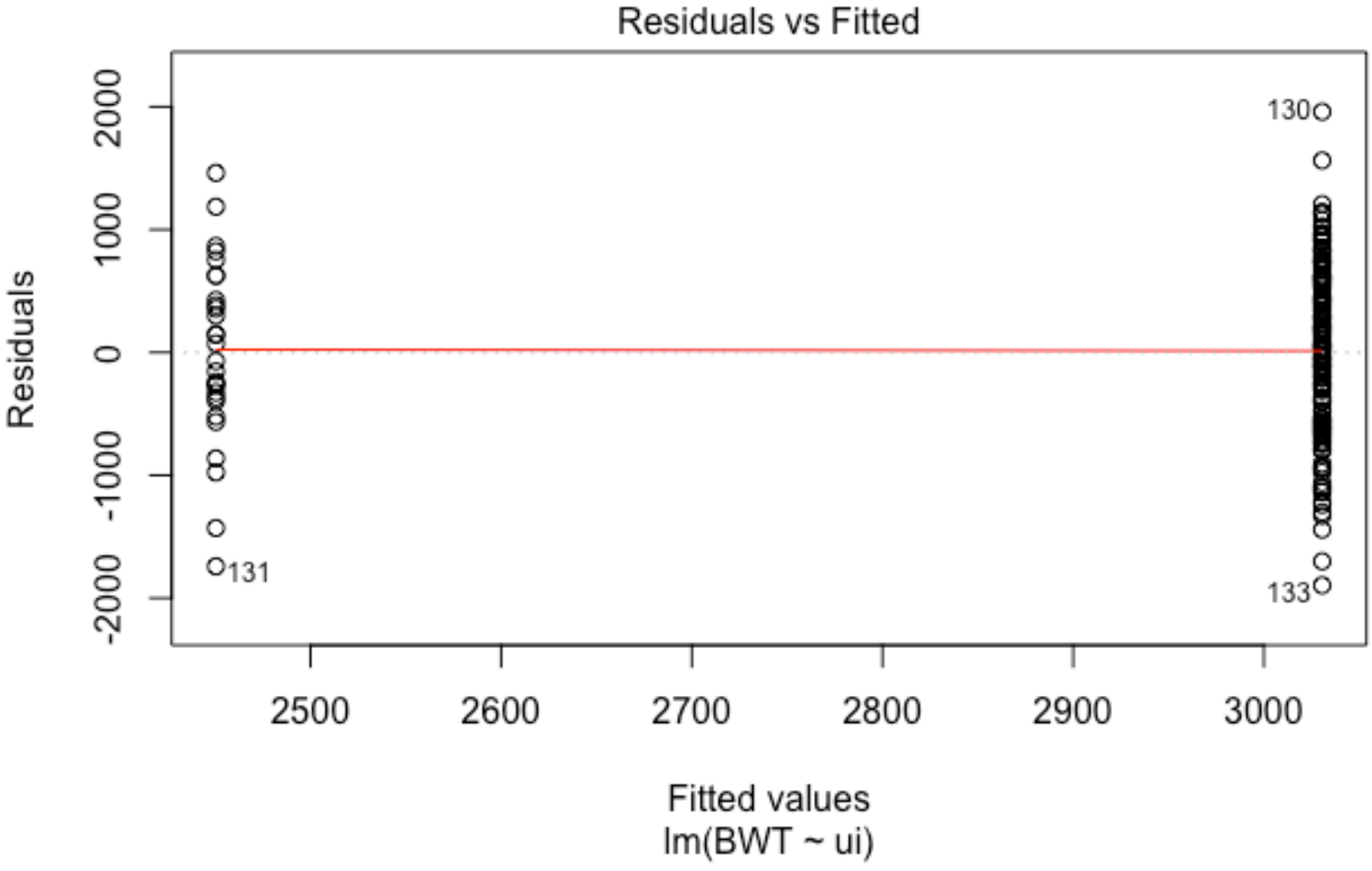
On conclut que le fait que la mère fume a un impact sur le poids du bébé à la naissance.

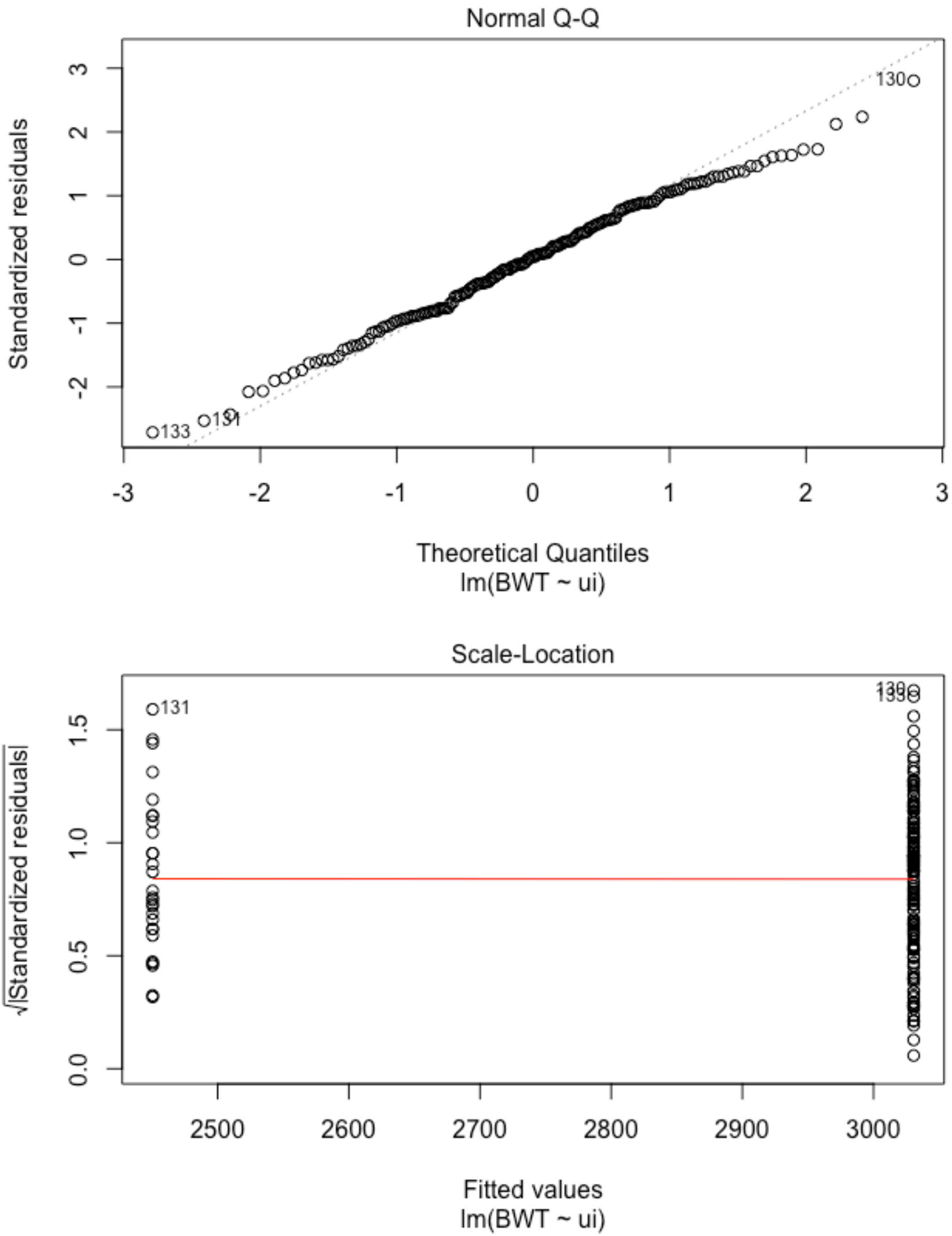
2.3 Irritabilité Utérine

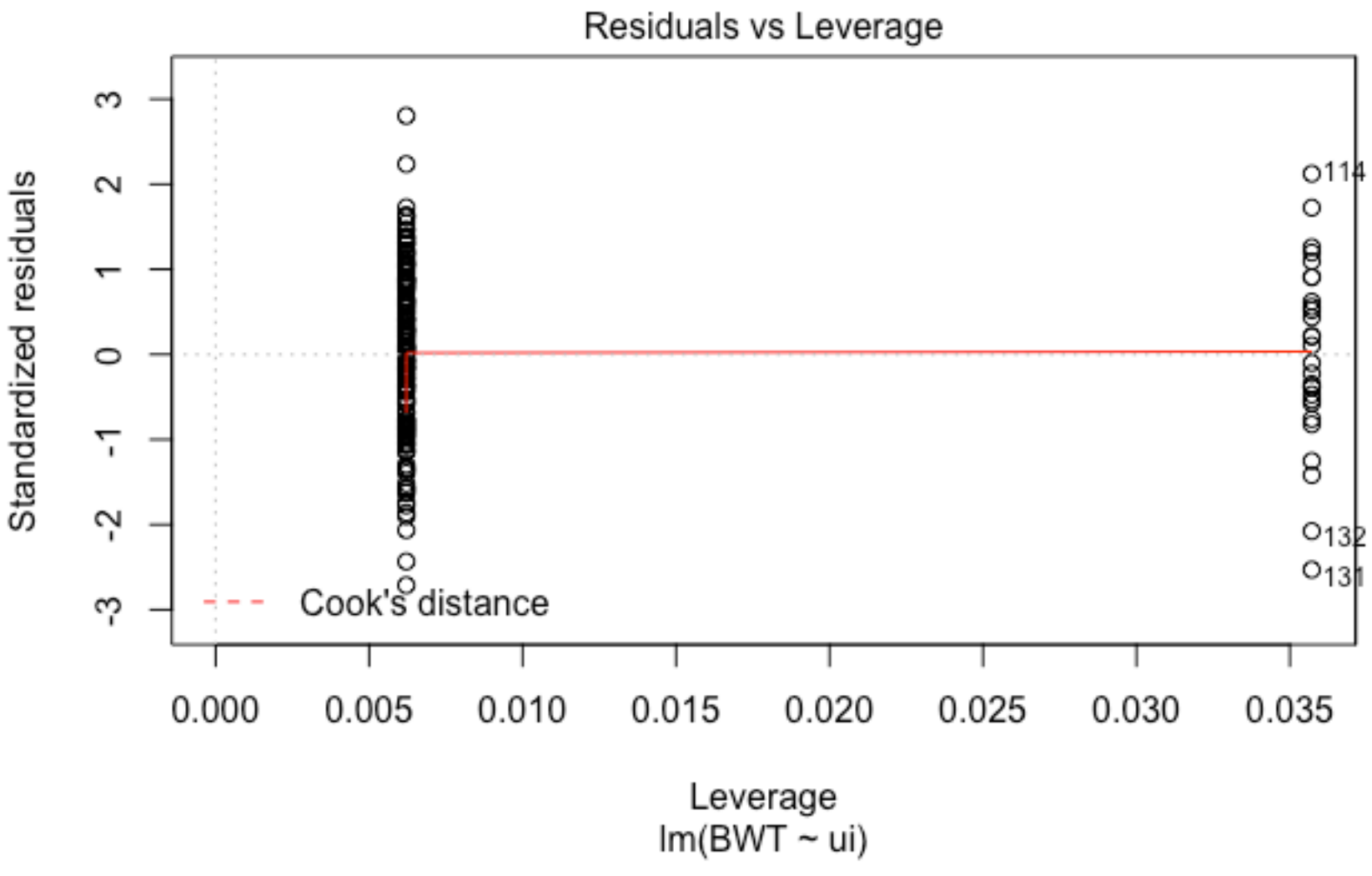
Il semblerait que le facteur UI ait lui aussi un effet sur le poids.



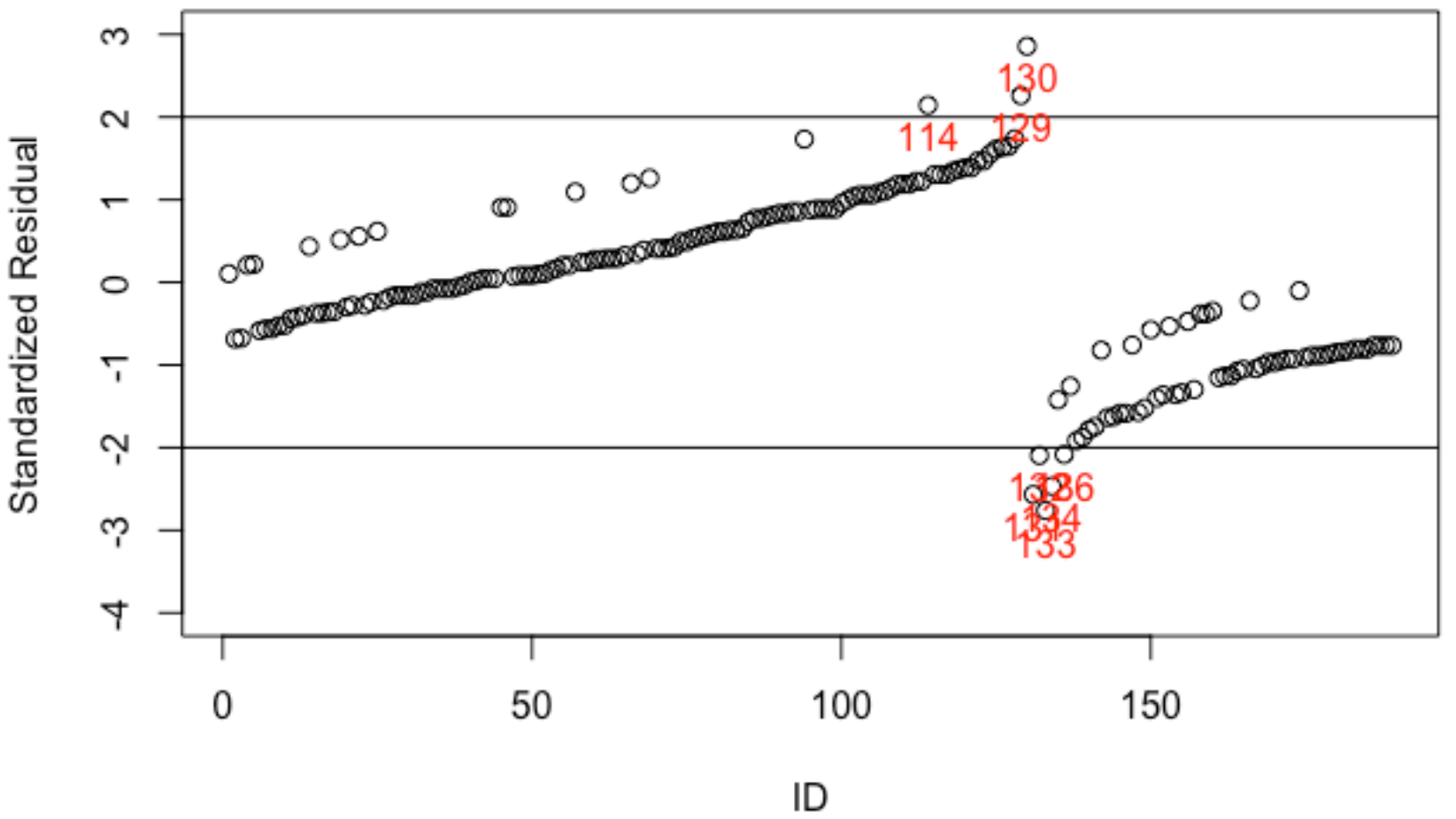
Ici encore, les graphiques montrent qu'il y a trois valeurs aberrantes : nous les supprimons donc du modèle linéaire avant de faire l'ANOVA. Cependant on observe ici aussi une structure dans les résidus studentisés contredisant l'hypothèse d'indépendance des résidus.




Normal Q-Q
Standardized residuals
-3
-2
-1
0
1
2
3
130
133
131
Theoretical Quantiles
lm(BWT ~ ui)
Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
131
2500
2600
2700
2800
2900
3000
Fitted values
lm(BWT ~ ui)



Residual vs Index



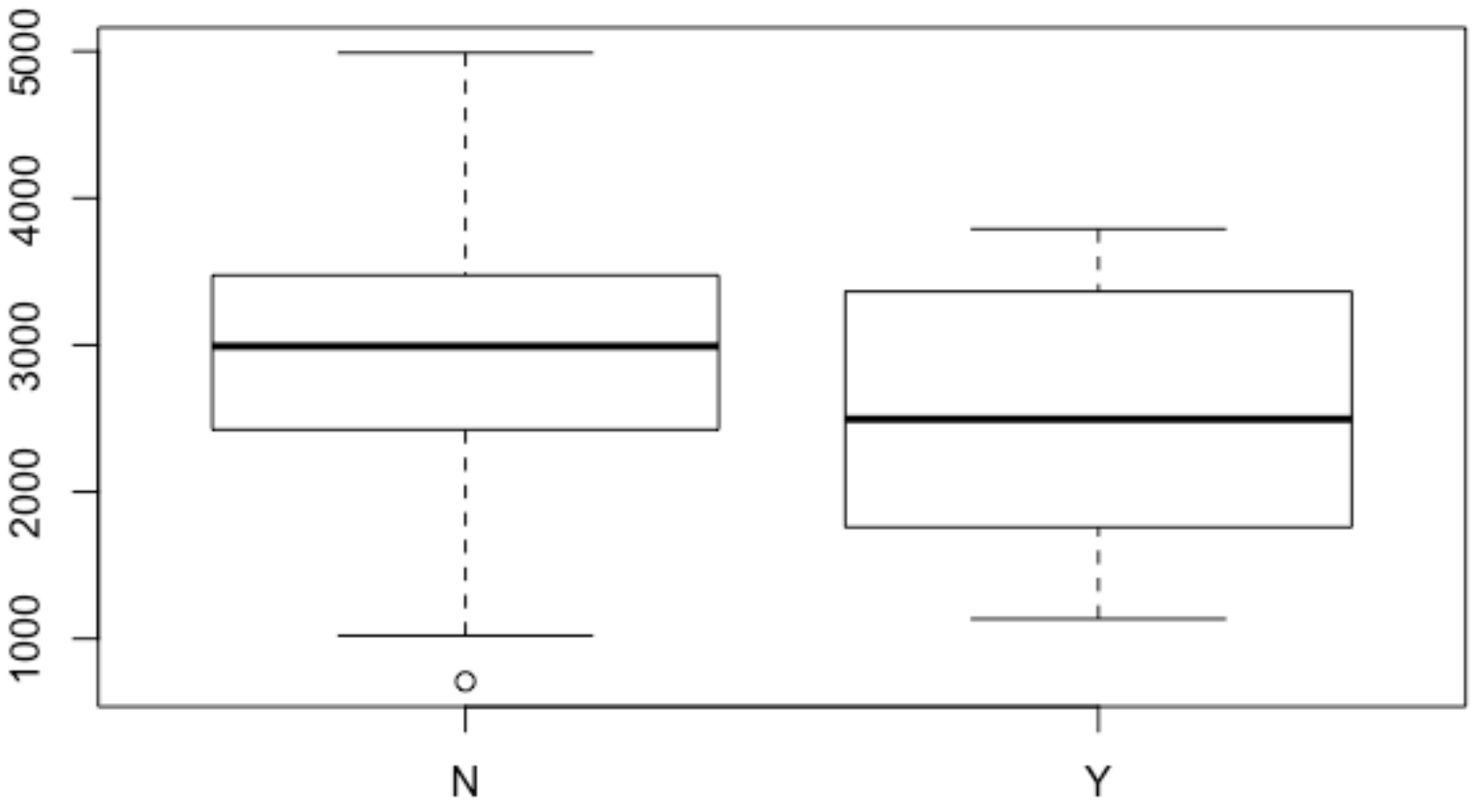
L’ANOVA donne :

```
             Df   Sum Sq Mean Sq F value   Pr(>F)
ui            1  8028747 8028747   16.34 7.73e-05 ***
Residuals   187 91888305  491381
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On conclut par l'ANOVA que l'irritabilité utérine a un effet sur le poids du bébé à la naissance.

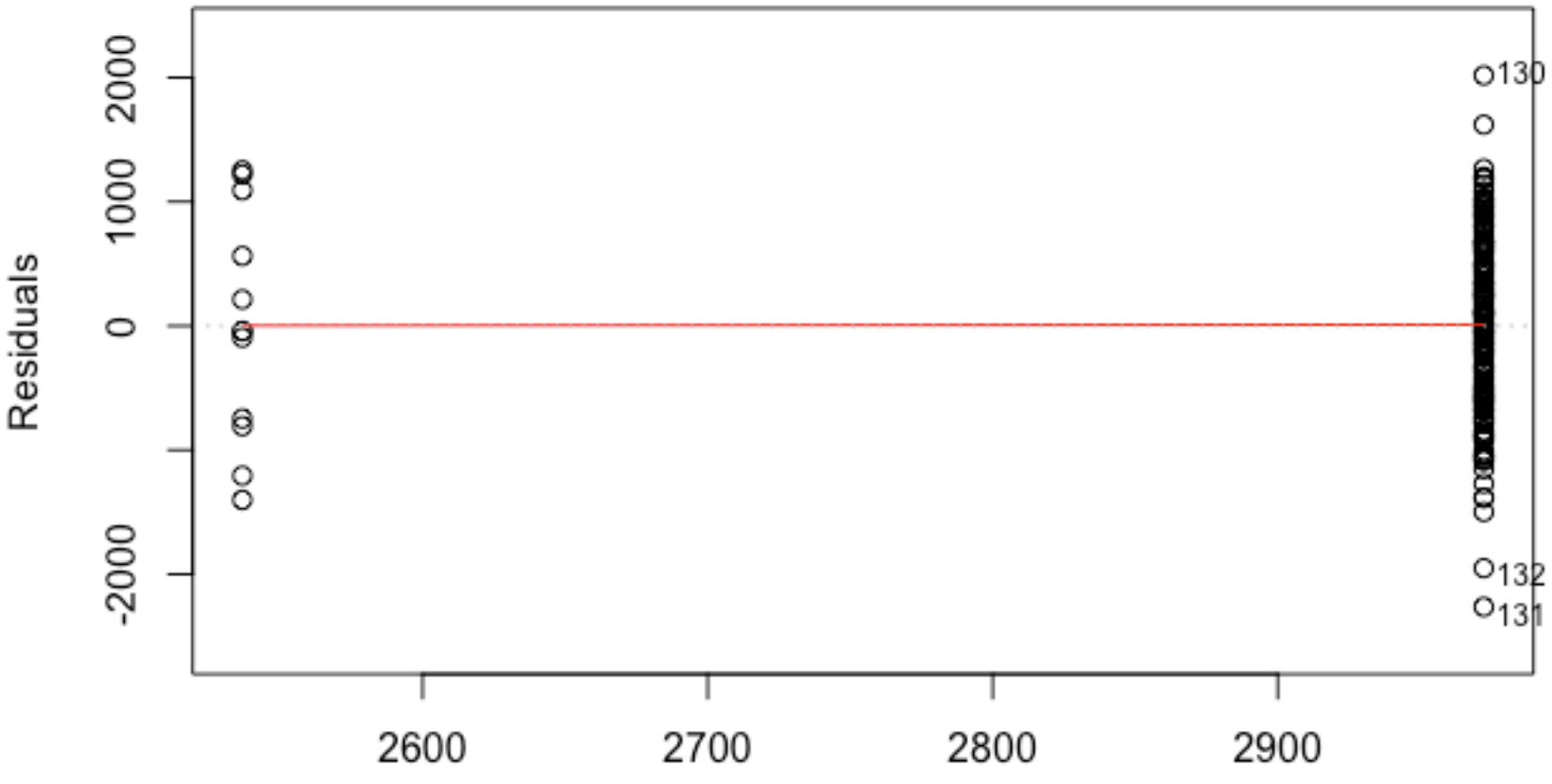
2.4 Haute tension

Il s'emblerait que le facteur HT ait un effet sur le poids.

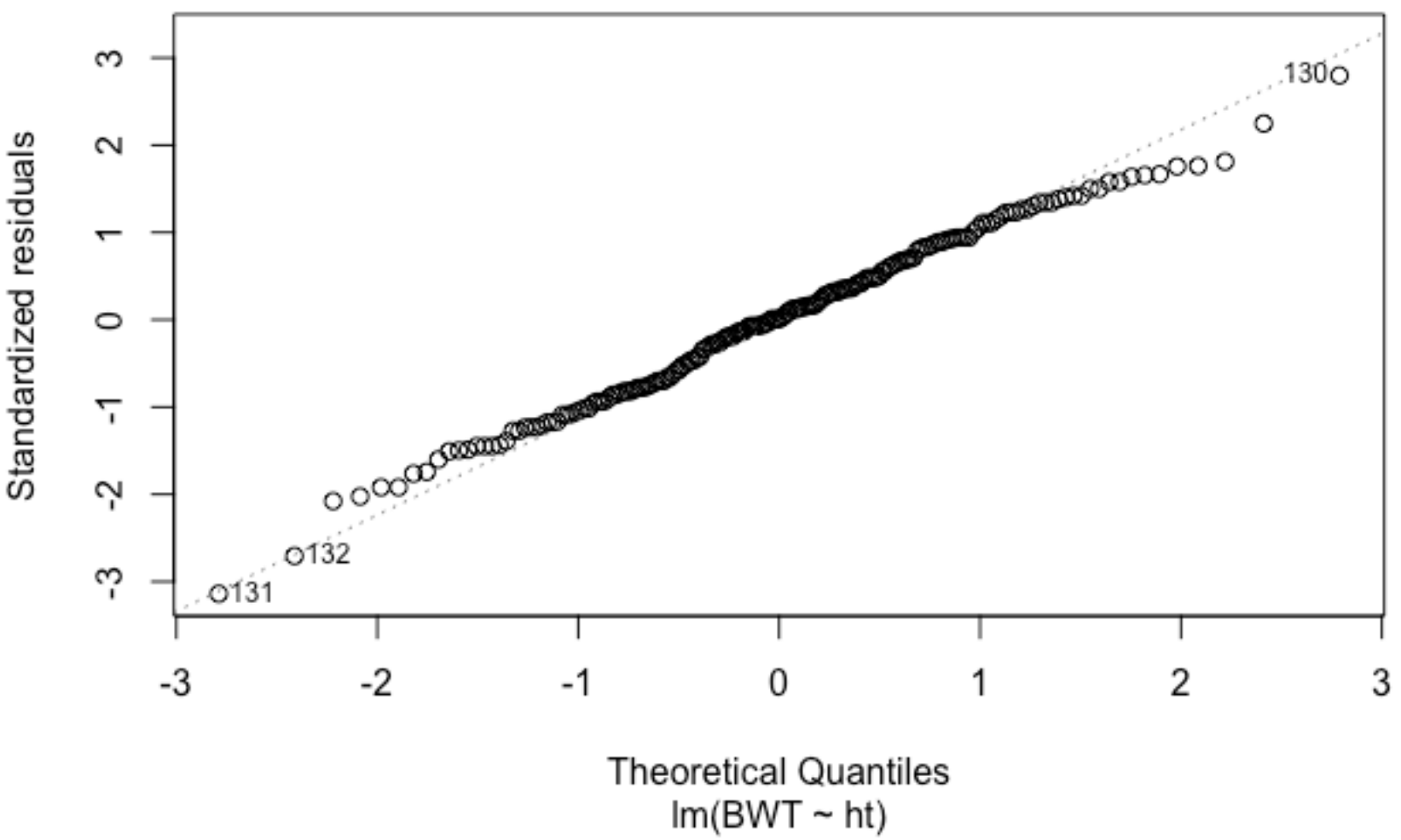


On retire ici trois valeurs aberrantes du modèle avant de faire l'ANOVA : celles d'indices $130, 131$ et 132. On note que les résidus ne sont là encore pas indépendants et que le modèle ne sera de toutes façons pas parfait.

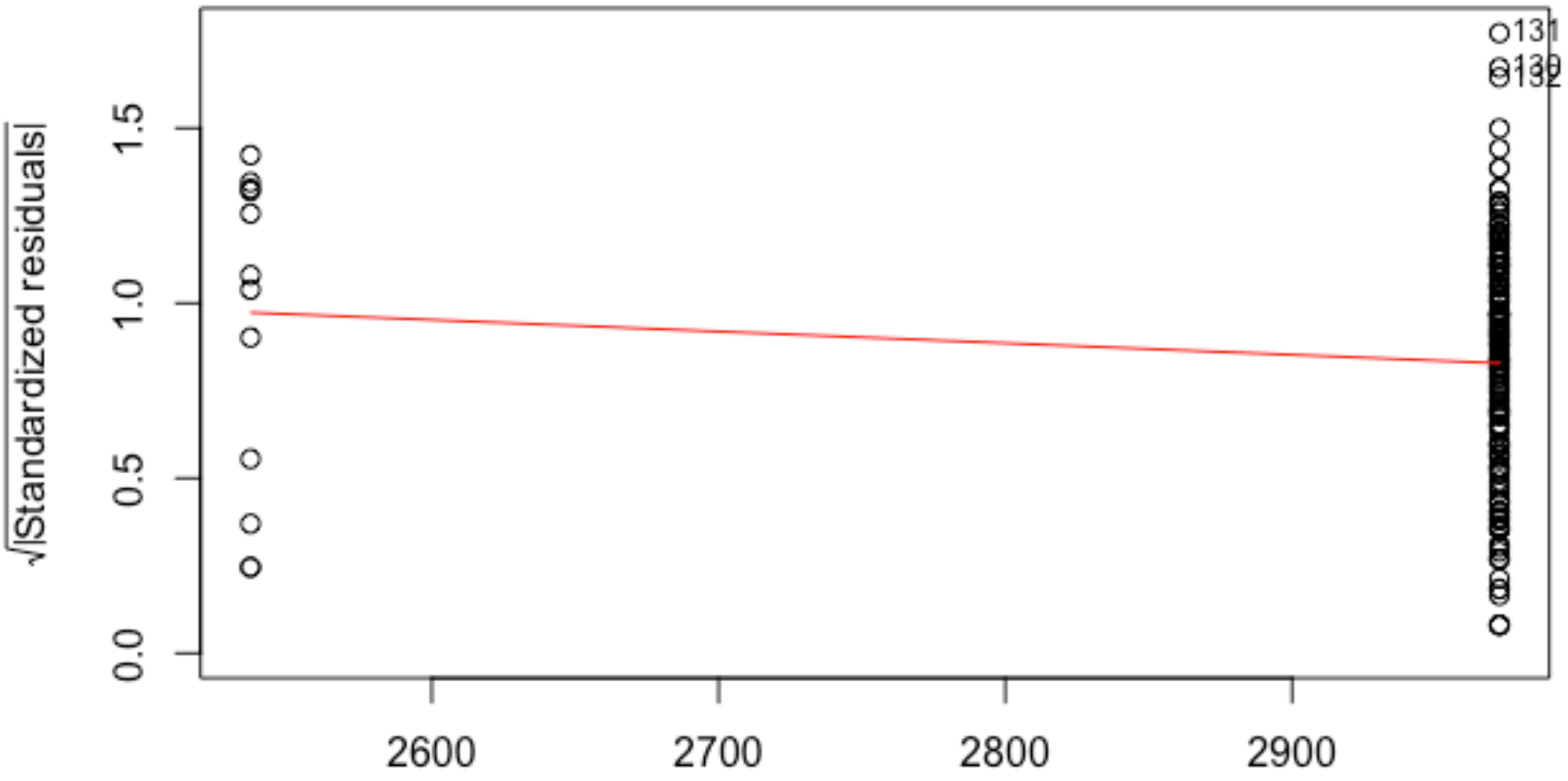
Residuals vs Fitted



Normal Q-Q

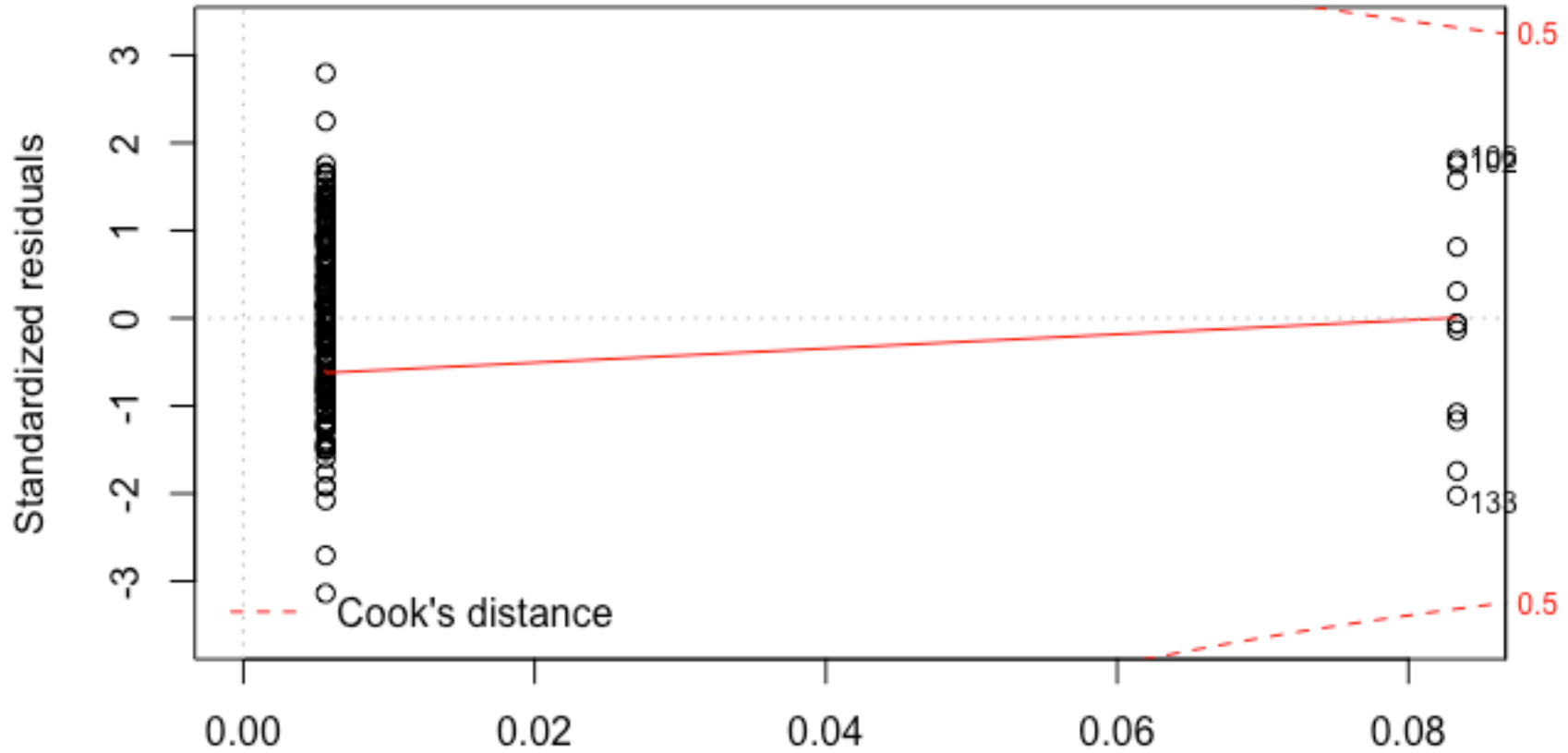


Scale-Location

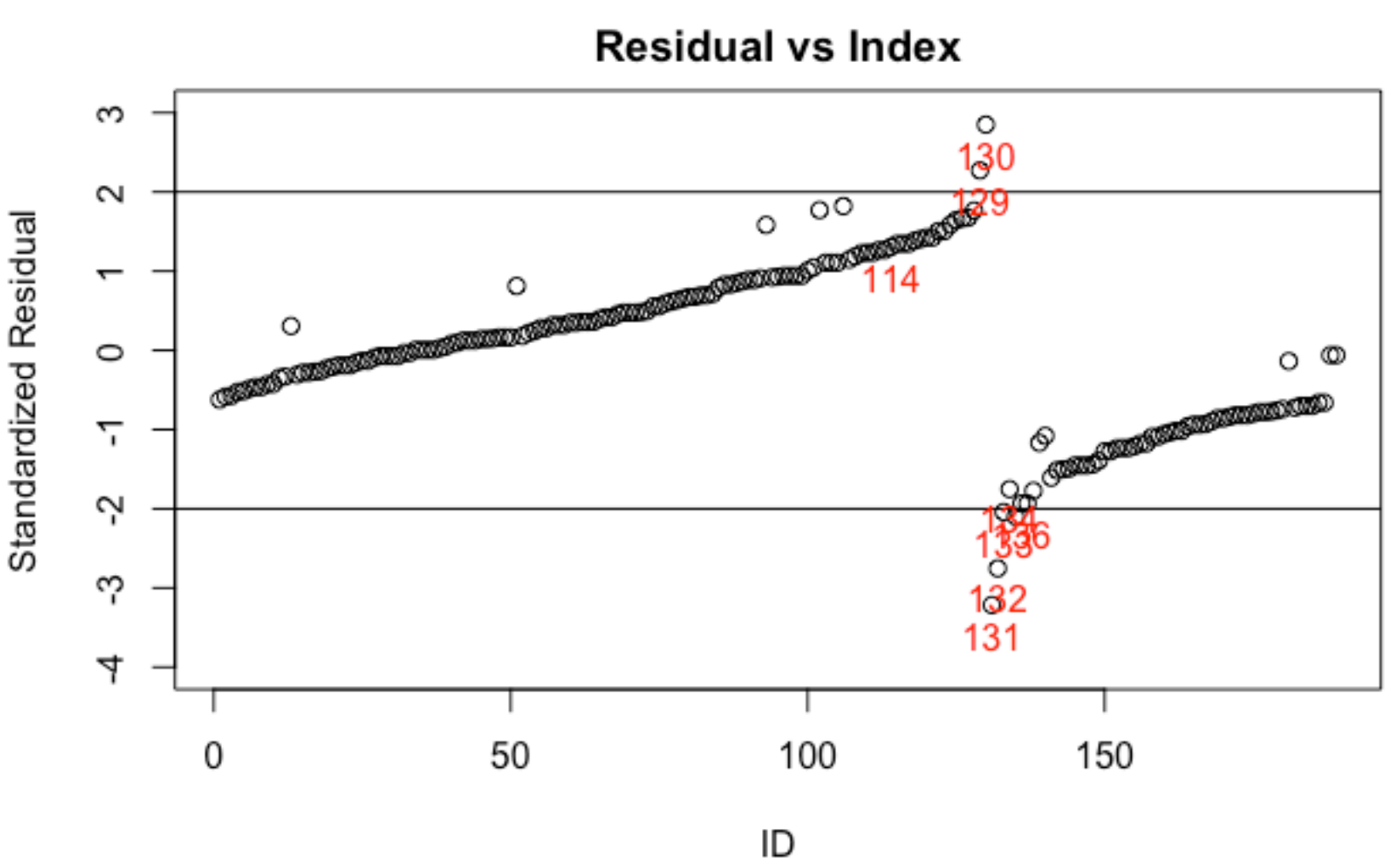


Fitted values
lm(BWT ~ ht)

Residuals vs Leverage



Leverage
lm(BWT ~ ht)



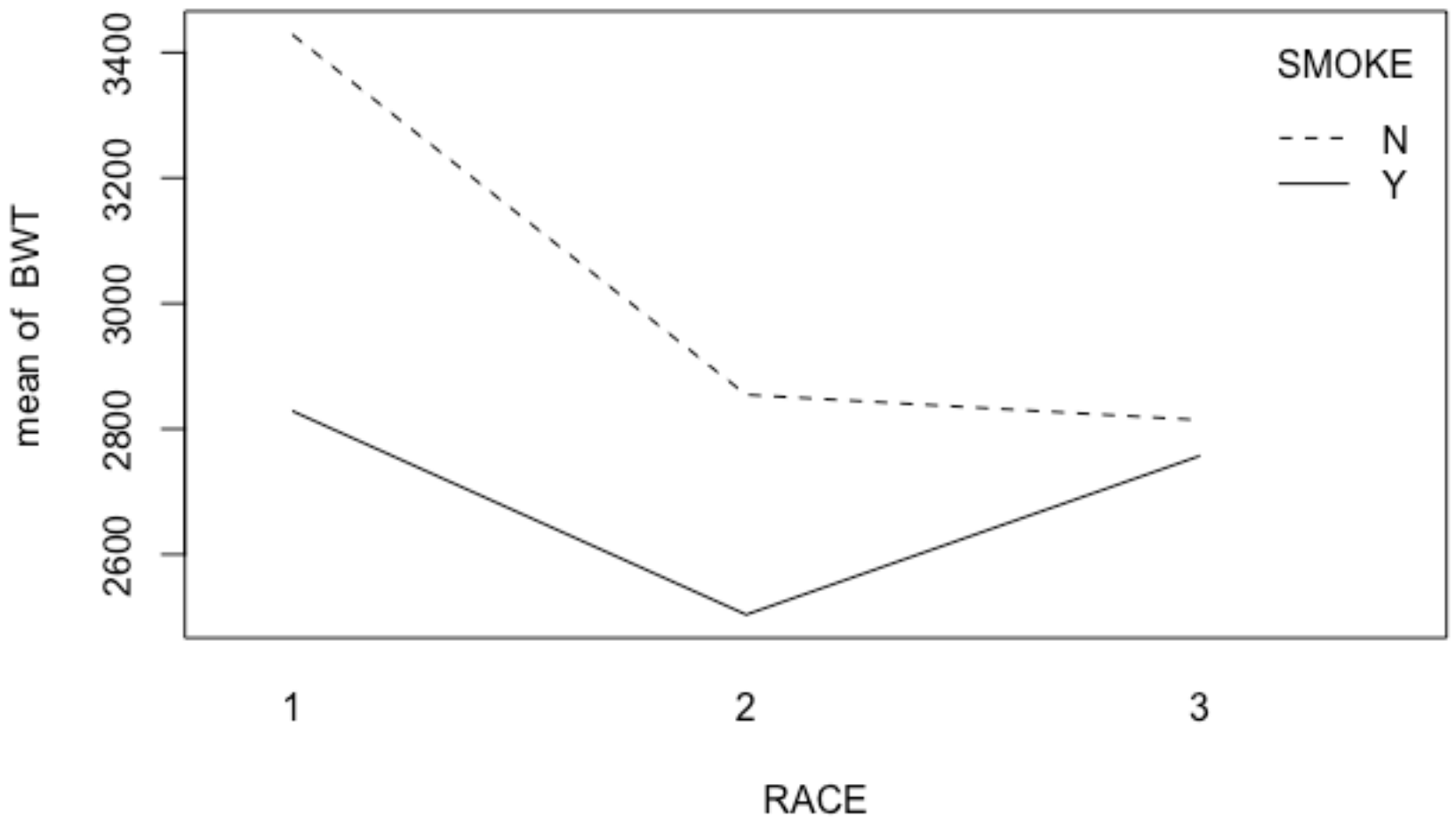
L’ANOVA donne :

```
             Df   Sum Sq Mean Sq F value Pr(>F)
HT.modif      1  2254943 2254943   4.895 0.0282 *
Residuals   184 84756041  460631
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On en conclut par qu’une haute tension a un effet sur le poids du bébé à la naissance.

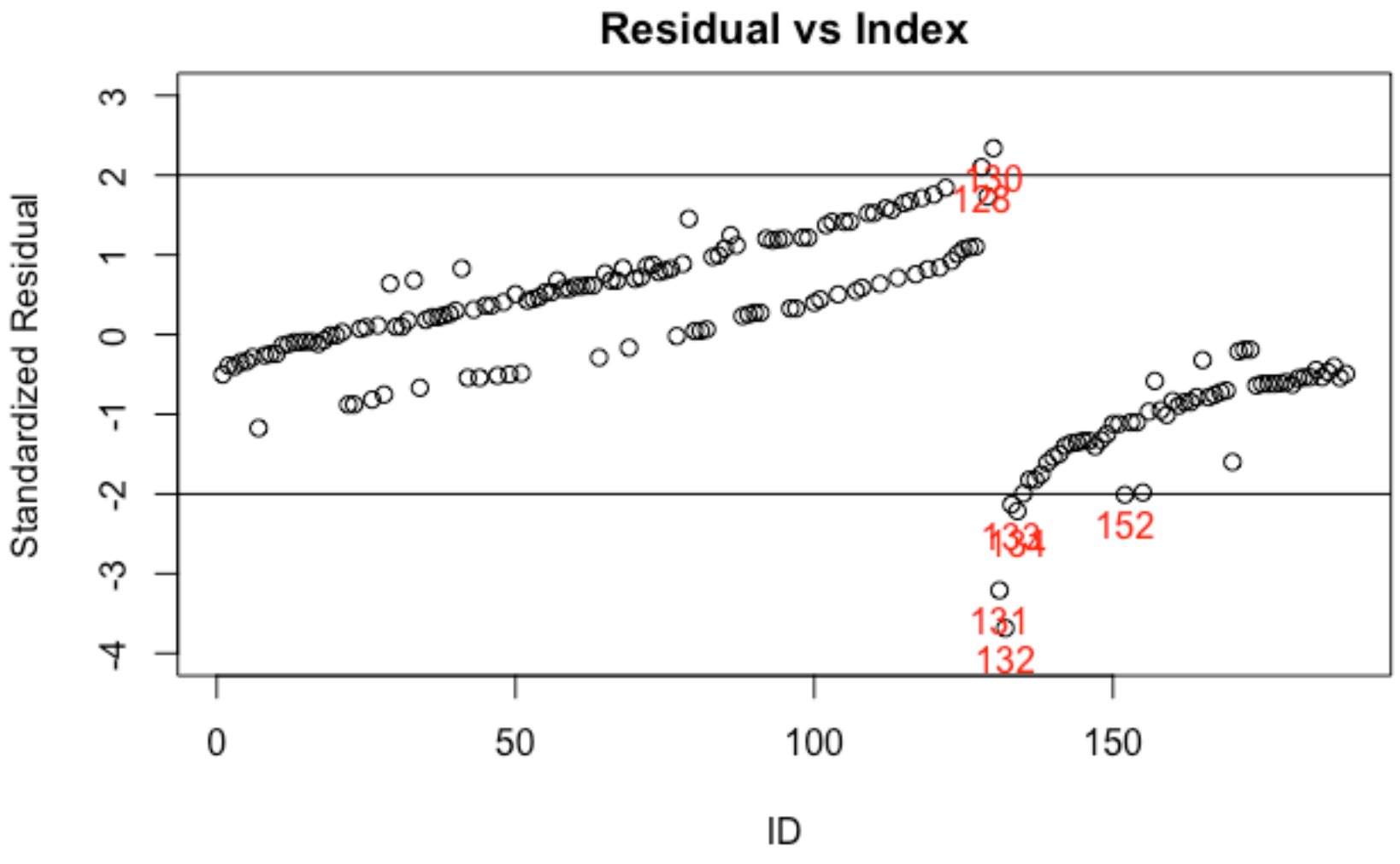
3 Analyse à deux facteurs

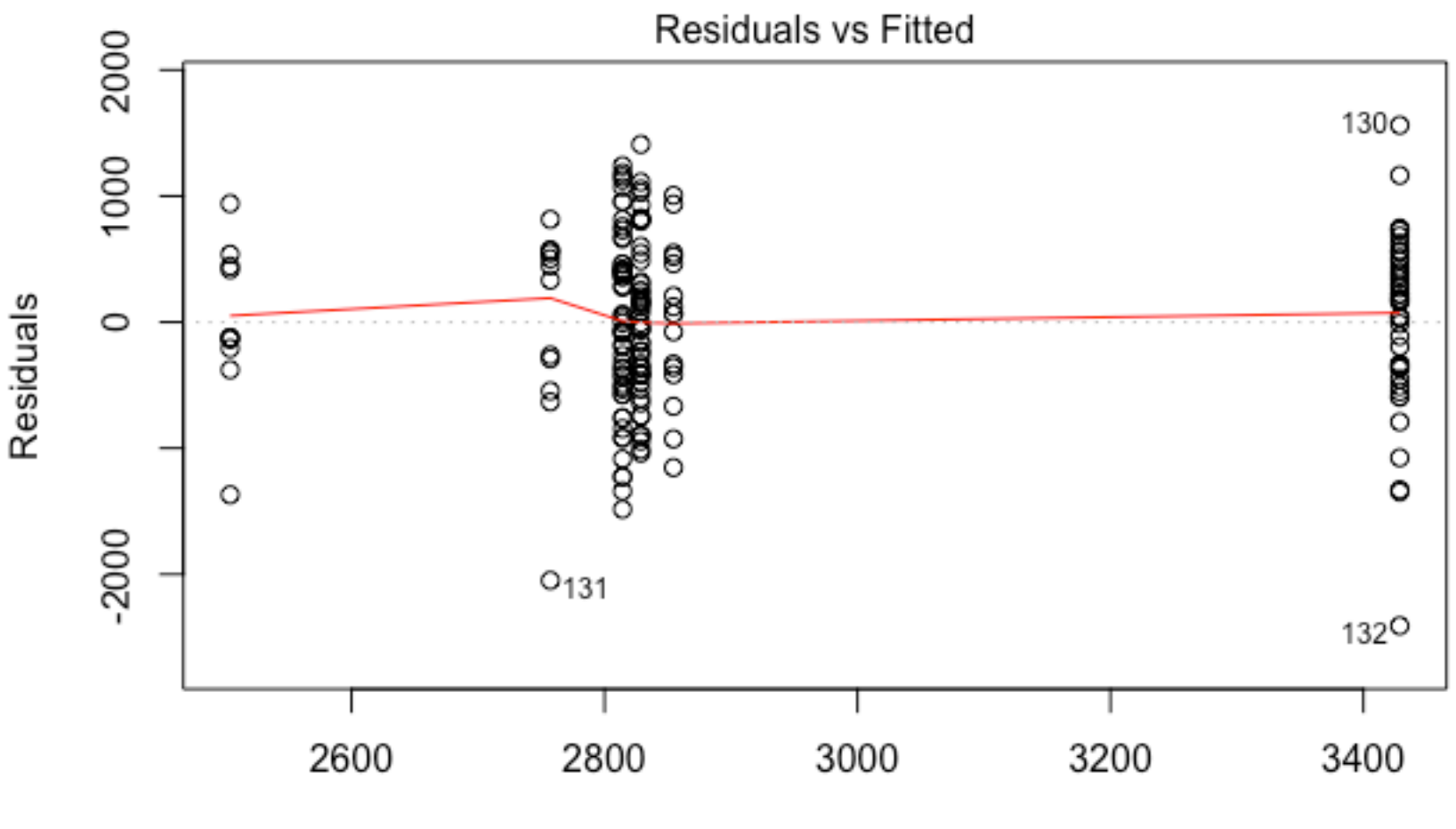
Intéressons-nous maintenant à l’analyse du poids en fonction de deux paramètres : le tabagisme et la race. Pour cela, commençons par analyser un graphique du poids du bébé en fonction de la race de la mère et de son tabagisme durant la grossesse.

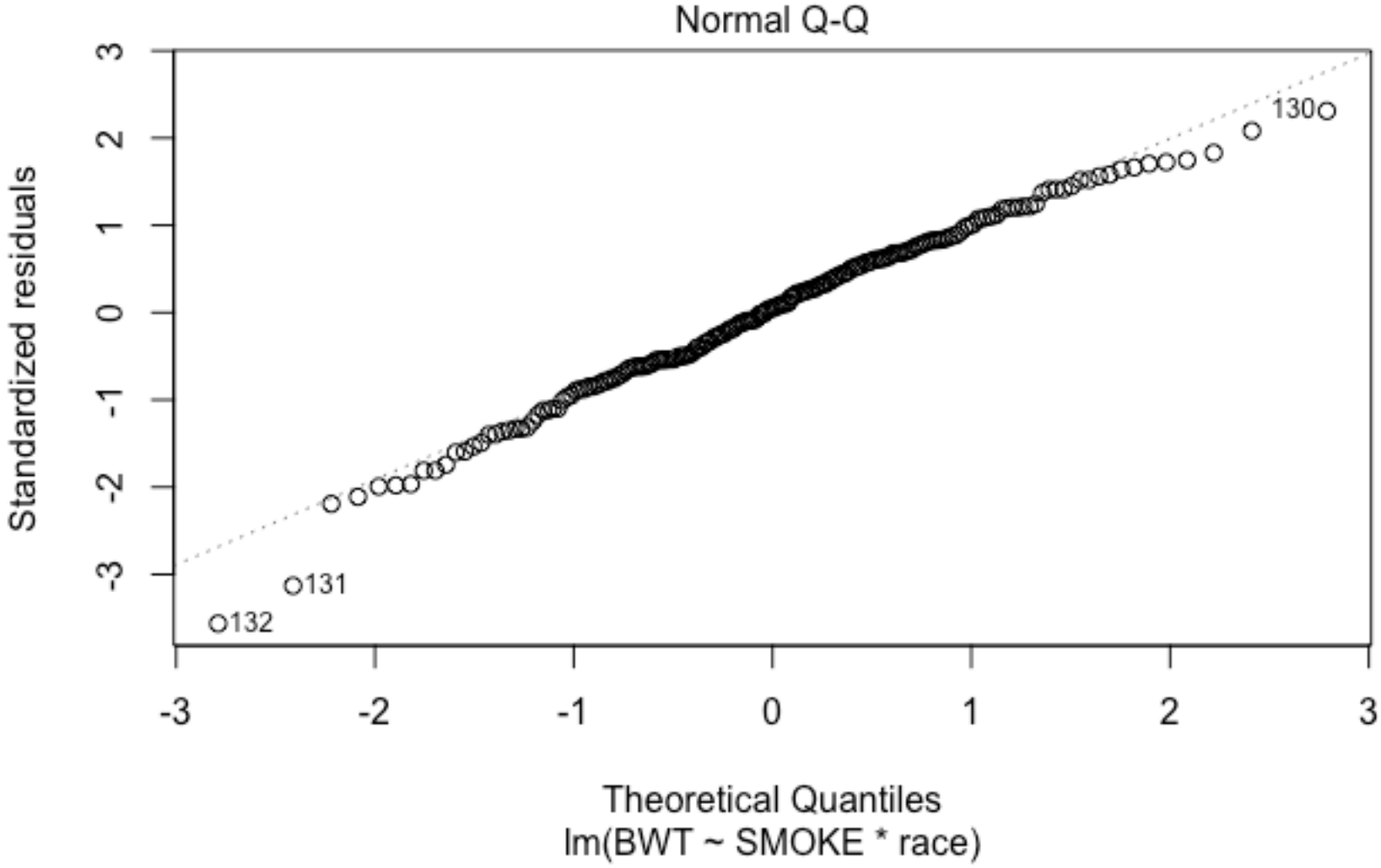


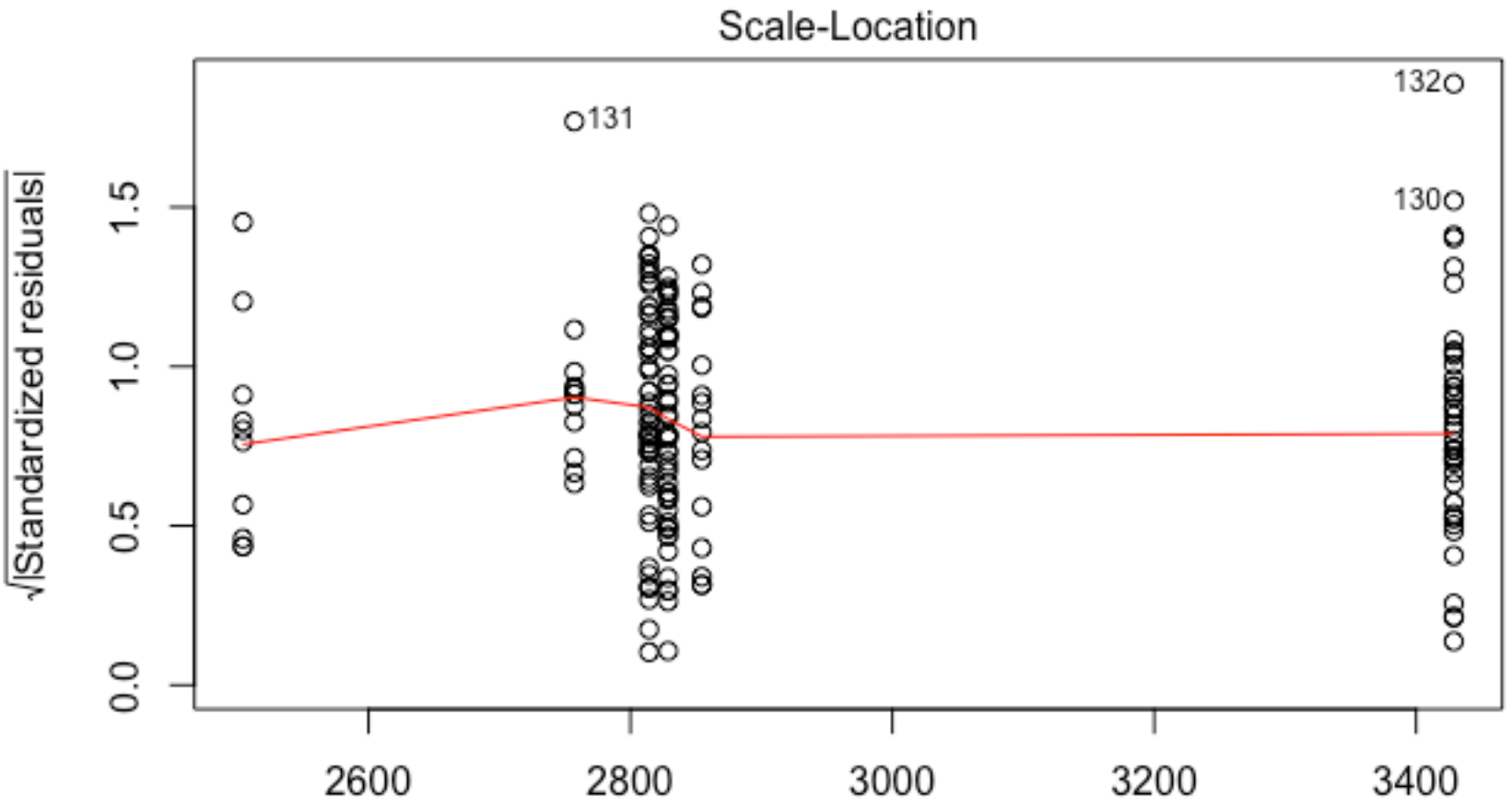
Le fait de fumer diminue considérablement le poids du bébé pour les races 1 et 2. On peut aussi remarquer que si pour les non-fumeurs, la moyenne des poids est la même pour les race 2 et 3, pour les fumeurs, la moyenne de la race 2 est bien plus faible que celle de la race 3.

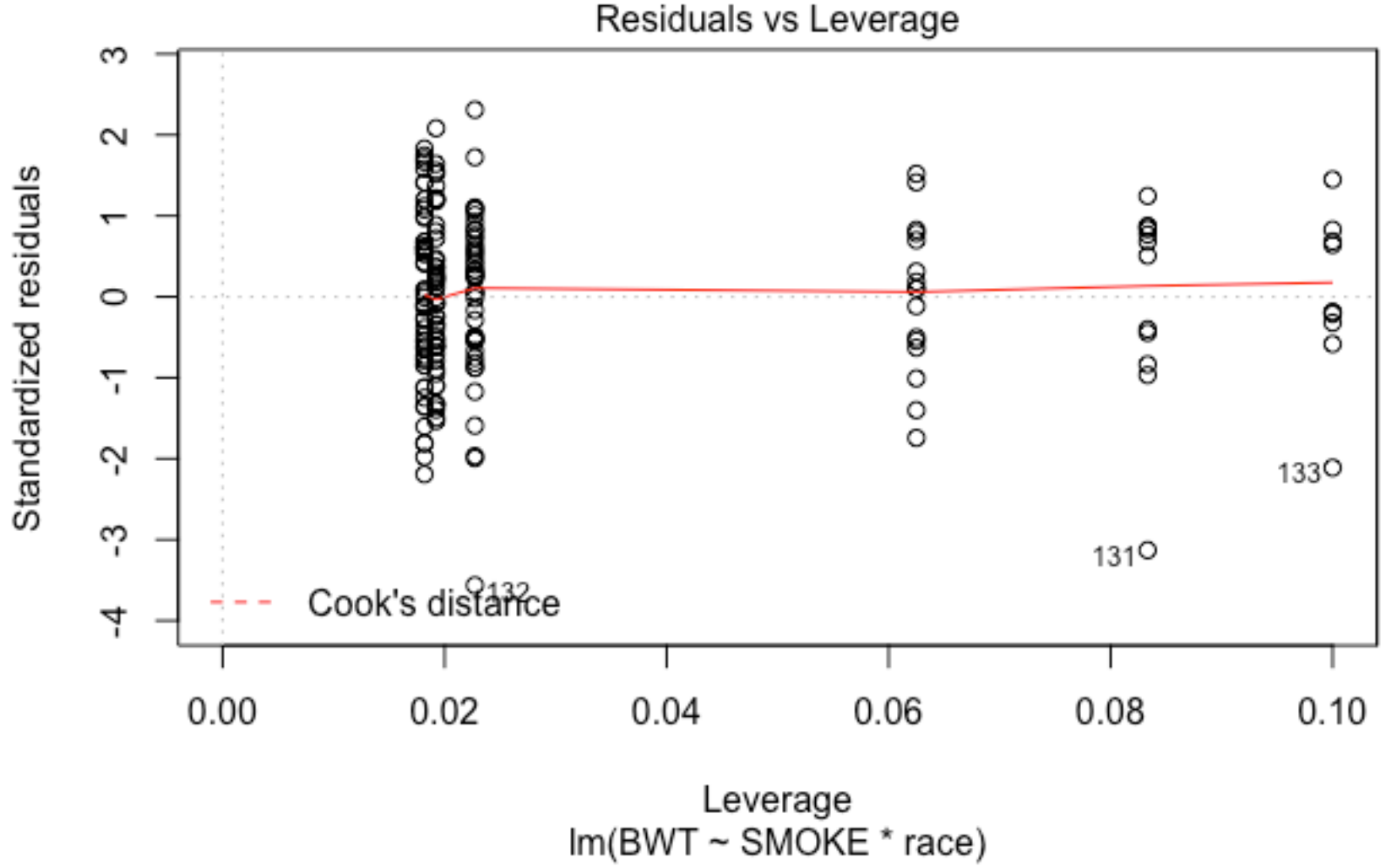
Il semblerait donc à première vue que l'interaction de ces deux facteurs aurait un effet. Vérifions cela à l'aide d'une ANOVA, après avoir validé le modèle linéaire.




Residuals vs Fitted
Residuals
2000
1000
0
-2000
130
131
132
2600
2800
3000
3200
3400
Fitted values
lm(BWT ~ SMOKE * race)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
130
131
132
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(BWT ~ SMOKE * race)





On conclut de ces graphiques que : les résidus sont normaux de même loi, mais pas indépendants car une structure se dégage des résidus studentisés ; nous avons quelques valeurs aberrantes, mais pas suffisamment fortes compte tenus de la taille de nos données pour être enlevés ; que pour le seuil $\frac{3p}{n} = 0.095$, nous avons une dizaine de points leviers et que nous n'avons pas de points suspects pour la distance de Cook.

Nous effectuerons donc notre ANOVA en prenant compte que notre modèle n'est pas parfait.

```
Analysis of Variance Table

Response: BWT
            Df   Sum Sq Mean Sq F value    Pr(>F)
SMOKE        1  3573406 3573406  7.6503 0.0062584 **
race         2  8768299 4384149  9.3861 0.0001317 ***
SMOKE:race   2  2097537 1048769  2.2453 0.1088037
Residuals  183 85477810  467092
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Au vus des résultats de l’ANOVA, puisque nous avons une p-valeur non significative de 0.11, l’interaction fumeur et race n’a pas vraiment d’impact significatif sur le poids du bébé : on peut considérer que seul le tabagisme et la race influent sur le poids dans ce cas.

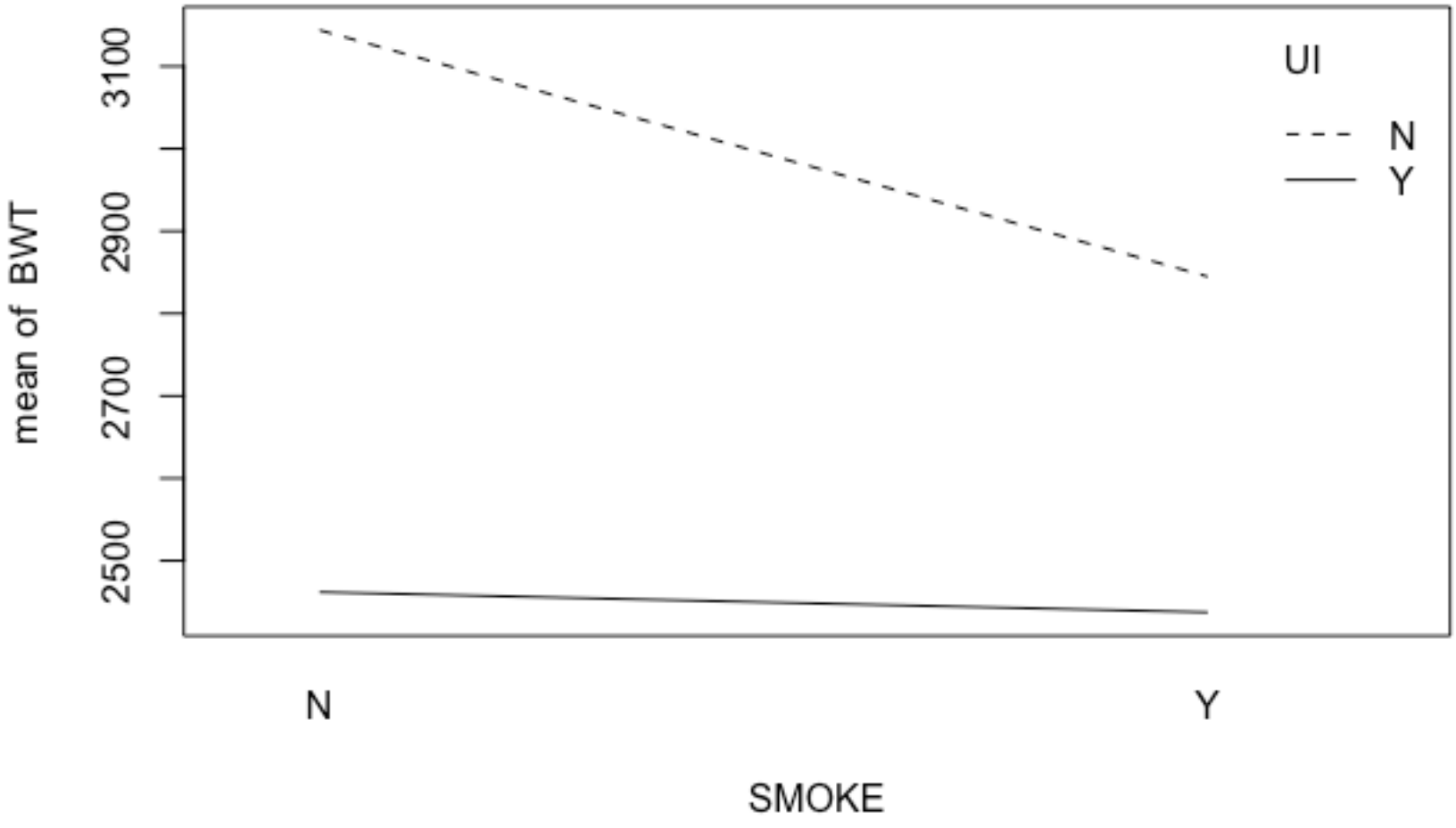
```
Analysis of Variance Table

Response: BWT
           Df   Sum Sq Mean Sq F value    Pr(>F)
SMOKE       1  3573406 3573406  7.5487 0.0065995 **
race        2  8768299 4384149  9.2614 0.0001468 ***
Residuals 185 87575348  473380
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On confirme ainsi que le tabagisme et la race ont bien un effet significatif sur le poids du bébé dans notre modèle.

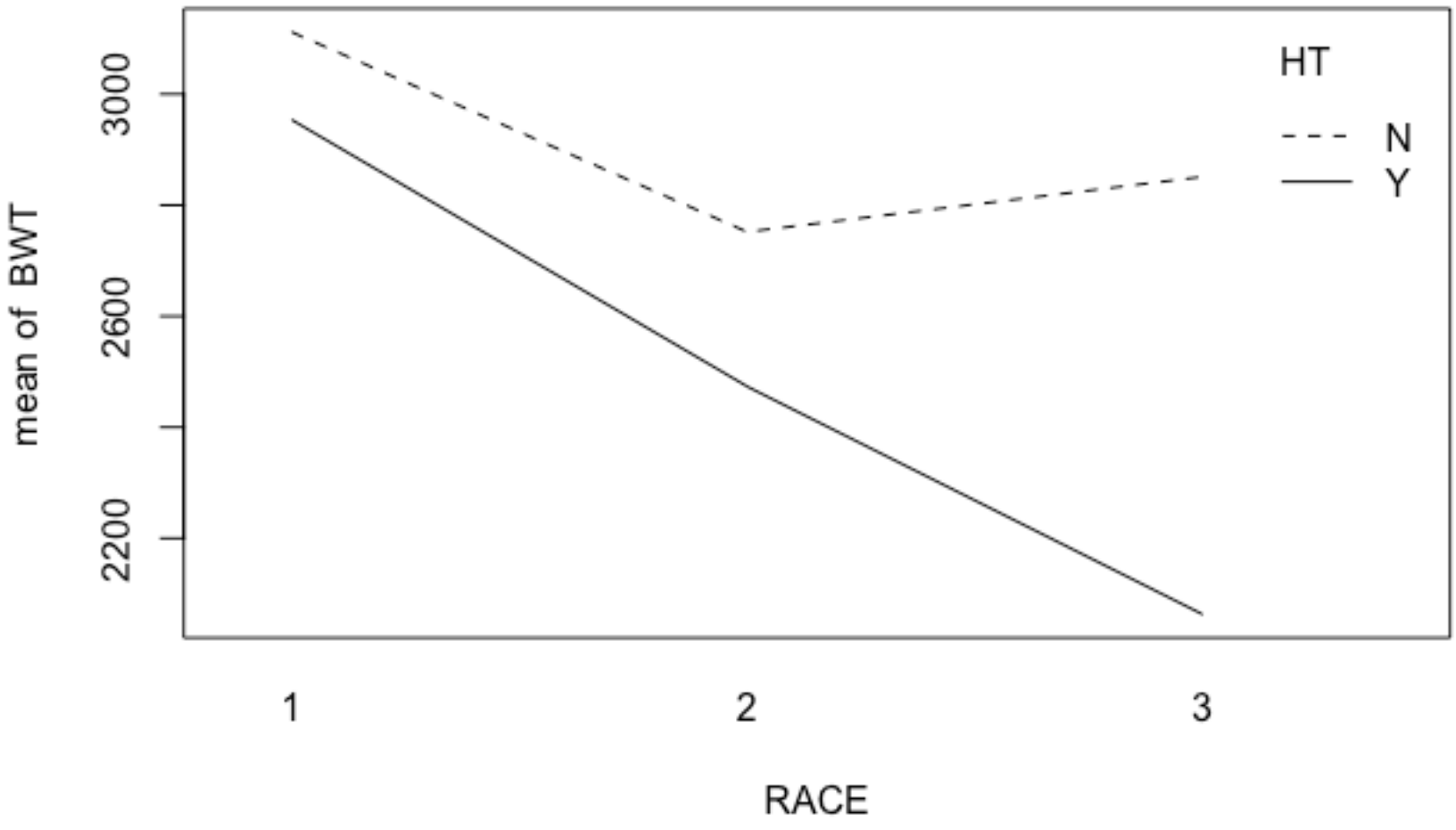
4 Autre modèle à deux facteurs

4.1 Première observation

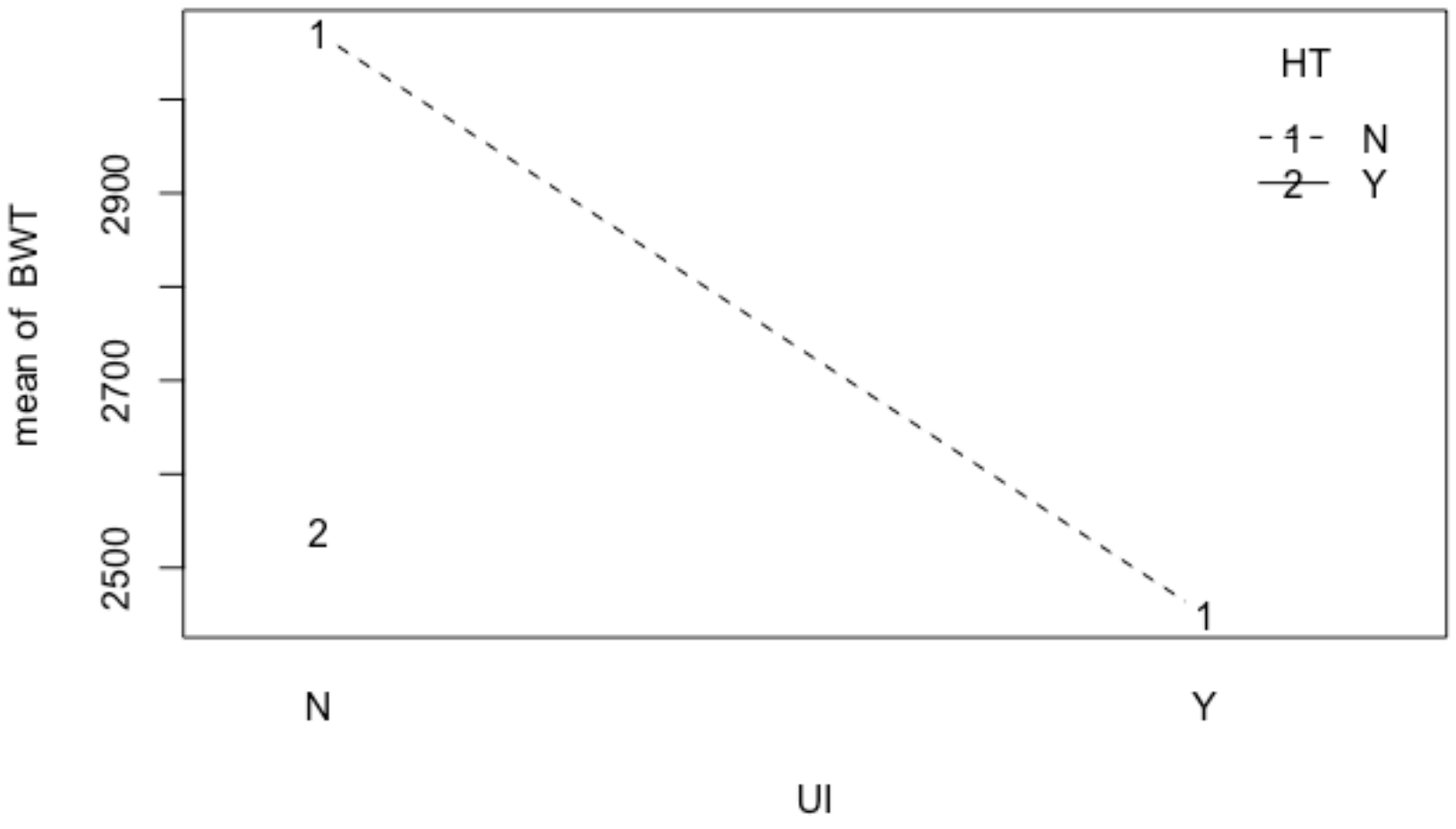
Intéressons-nous maintenant à d’autres modèles à 3 facteurs : UI-SMOKE, HT-RACE et UI-HT. Pour cela, commençons par analyser les graphiques suivants :



On peut voir sur ce graphique que la tendance du poids des bébés en fonction du tabagisme de la mère est modifié par la présence d'irritation utérine, ce qui pourrait suggérer un impact de l'interaction.



On voit un phénomène similaire sur ce graphique, qui suggèrerait de même l'impact de l'interaction sur le poids du bébé.



Malheureusement, ce graphique ne peut pas être interprété car, comme le montre le tableau d'expérience en début de rapport, nous n'avons aucun individu avec des antécédents d'hypertension et d'irritabilité utérine, ce qui empêche l'observation d'une différence de tendance.

4.2 ANOVA

Faisons maintenant des ANOVA pour vérifier si les facteurs dans les deux premiers modèles sont indépendants, sachant que nous n'avons pas les données nécessaires pour faire ce travail sur le troisième modèle. Pour chaque modèle, nous ferons une ANOVA du modèle initial, et d'un modèle supposé sans interaction.

ANOVA pour le modèle SMOKE-UI

```
Analysis of Variance Table

Response: BWT
           Df   Sum Sq Mean Sq F value    Pr(>F)
SMOKE       1  3573406 3573406  7.4702 0.0068817 **
UI          1  7405281 7405281 15.4807 0.0001178 ***
SMOKE:UI    1   442511  442511  0.9251 0.3374036
Residuals 185 88495854  478356
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Response: BWT
           Df   Sum Sq Mean Sq F value    Pr(>F)
SMOKE       1  3573406 3573406  7.4732 0.0068673 **
UI          1  7405281 7405281 15.4869 0.0001173 ***
Residuals 186 88938365  478163
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Avec une p-valeur non significatif de 0.34, il est raisonnable de considérer l'effet de ces deux facteurs comme étant nul. D'ailleurs la suppression de cette interaction ne modifie pas significativement les p-value des autres tests. De plus, l'ANOVA nous indique que, dans un modèle où l'on considère uniquement les facteurs UI et SMOKE, le premier aurait un effet plus important sur le poids du bébé.

Anova pour le modèle RACE-HT

```
Analysis of Variance Table

Response: BWT
           Df   Sum Sq Mean Sq F value   Pr(>F)
race        2  5070608 2535304  5.0329 0.007451 **
HT          1  1777535 1777535  3.5287 0.061907 .
race:HT     2   884006  442003  0.8774 0.417589
Residuals 183 92184904  503743
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Response: BWT
           Df   Sum Sq Mean Sq F value   Pr(>F)
race        2  5070608 2535304  5.0396 0.007394 **
HT          1  1777535 1777535  3.5333 0.061719 .
Residuals 185 93068910  503075
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

De même, nous pouvons considérer qu'il n'y pas d'interaction entre les facteurs RACE et HT. De plus, dans un modèle où l'on considère uniquement les facteurs RACE et HT, le premier facteur est plus significatif que le second pour modéliser le poids du bébé.

ANOVA pour le modèle HT-UI

```
Analysis of Variance Table

Response: BWT
           Df   Sum Sq Mean Sq F value    Pr(>F)
HT          1  2132014 2132014  4.4694   0.03584 *
UI          1  9059202 9059202 18.9912 2.169e-05 ***
Residuals 186 88725836  477021
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Si on ne peut pas se prononcer quand à un éventuel effet de l'interaction de ces deux facteurs, nous pouvons cependant conclure que, dans un modèle utilisant les facteurs UI et HT, le premier est le plus significatif.

4.3 Critère de sélection de modèle

Nous allons à présent sélectionner le modèle à privilégier selon le critère R_a^2 : nous sélectionnerons le modèle avec la valeur R_a^2 la plus élevé, modèle que nous supposons sans interactions, ce qui est en accord avec les ANOVA précédents.

```

Call:
lm(formula = BWT ~ SMOKE + race)

Residuals:
     Min       1Q   Median       3Q      Max
-2314.15  -441.93    35.99   491.85  1654.85

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3335.15      91.75  36.352  < 2e-16 ***
SMOKEY       -427.22     109.01  -3.919 0.000125 ***
race2        -451.14     153.07  -2.947 0.003619 **
race3        -454.62     116.44  -3.904 0.000132 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 688 on 185 degrees of freedom
Multiple R-squared:  0.1235,    Adjusted R-squared:  0.1093
F-statistic:  8.69 on 3 and 185 DF,  p-value: 2.007e-05


Call:
lm(formula = BWT ~ SMOKE + UI)

Residuals:
     Min       1Q   Median       3Q      Max
-1797.78  -461.32    47.22   501.22  1862.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3127.78      67.08  46.624  < 2e-16 ***
```

```
SMOKEY       -256.46     103.25  -2.484 0.013883 *
UIY          -558.28     141.86  -3.935 0.000117 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 691.5 on 186 degrees of freedom
Multiple R-squared:  0.1099,    Adjusted R-squared:  0.1003
F-statistic: 11.48 on 2 and 186 DF,  p-value: 1.99e-05


Call:
lm(formula = BWT ~ race + HT)

Residuals:
     Min       1Q   Median       3Q      Max
-2118.85  -502.85     7.47   512.47  1865.47

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3124.53      73.23  42.667  < 2e-16 ***
race2        -358.78     157.39  -2.280  0.02377 *
race3        -296.68     112.92  -2.627  0.00933 **
HTY          -399.19     212.37  -1.880  0.06172 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 709.3 on 185 degrees of freedom
Multiple R-squared:  0.06854,   Adjusted R-squared:  0.05343
F-statistic: 4.538 on 3 and 185 DF,  p-value: 0.004283


Call:
lm(formula = BWT ~ HT + UI)

Residuals:
     Min       1Q   Median       3Q      Max
-1741.43  -522.43    -8.38   529.62  1919.62

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3070.38      56.58  54.265  < 2e-16 ***
HTY          -533.63     207.25  -2.575   0.0108 *
UIY          -619.95     142.26  -4.358 2.17e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 690.7 on 186 degrees of freedom
Multiple R-squared:  0.112, Adjusted R-squared:  0.1025
F-statistic: 11.73 on 2 and 186 DF,  p-value: 1.593e-05
```

Ainsi, si nous prenons comme critère le R^2_a ajusté, nous gardons le modèle SMOKE-RACE, considéré sans interaction. Le choix serait le même avec le critère R^2. Il faut cependant se souvenir que le modèle retenu n’est pas parfait, en particulier que les résidus ne sont pas indépendants. Il pourrait donc être intéressant

de modéliser le poids des bébés à l'aide de ces deux facteurs dans un modèle plus général que celui des régressions linéaires gaussiennes. De plus, il faudrait faire des tests supplémentaires avec de nouvelles données pour vérifier qu'un modèle à seulement deux facteurs est suffisant pour pouvoir prédire le poids des bébés.