Projet 2
sur la régression linéaire

(À l'attention de Mme Casanova Sandrine et Mme Lecomte Evelyne)

Zou Yongkang

Hassan Yusuf Zakaria

(TP2 Binôme 6)

L3 MIAGE

Annee 2021/2022

Introduction :

1. la population concernée par l'étude

Dans notre étude, nous disposons de données relevées sur un échantillon de 500 médecins spécialistes de Midi-Pyrénées en 1999. (Taille de l'échantillon : 500) Puis, notre échantillon représente une partie de la population étudiée, les 2082 médecins spécialistes en Midi-Pyrenee en 1999.

2. les variables étudiées

`prasex` : sexe du médecin (Femme pour feminin et Homme pour masculin)
- Il s'agit d'une **variable qualitative nominale** à 2 modalité (`Femme/Homme`).

`praspe3` : spécialité du médecin (en 3 catégories : chirurgicale, médicale et mixte)
- Il s'agit d'une **variable qualitative nominale** à 3 modalités (`Chirurgicale/Médicale/Mixte`).

`secteur` : secteur tarifaire du médecin (en 2 catégories : secteur 1 et secteur 2)
- Il s'agit d'une **variable qualitative nominale** à 2 modalités (`1/2`).

`cltota` : nombre total de patients vus par le médecin en 1999
- Il s'agit d'une **variable quantitative discrète**.

`honormke` : honoraires totaux du médecin pour l'année 1999 (en millier d'euros)
- Il s'agit d'une **variable quantitative continue**.

`rless16` : proportion de patients âgés de moins de 16 ans dans la patientèle du médecin
- Il s'agit d'une **variable quantitative continue**.

`r60to69` : proportion de patients âgés de 60 à 69 ans dans la patientèle du médecin
- Il s'agit d'une **variable quantitative continue**.

`rmore70` : proportion de patients âgés de plus de 70 ans dans la patientèle du médecin
- Il s'agit d'une **variable quantitative continue**.

3. le but de l'étude

Comme indiqué dans l'énoncé, le but est d'identifier les variables qui ont un effet sur les honoraires.

Autrement dit, nous voulons savoir si **le sexe**, **la specisalité**, **le secteur**, **le nombre total de patients**, **la proportion de patients âgés de moins de 16 ans, de 60 à 69 ou de plus de 70 ans dans la patientèle** ont un impact sur **les honoraires** du medecin et si oui, on voulons encore mesurer cet effet.

Partie 1 : régression linéaire simple

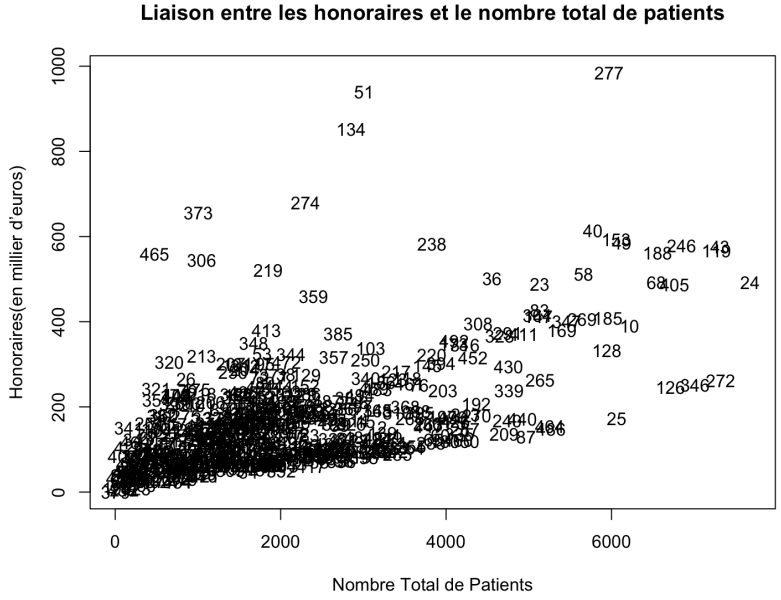
Problématique : les honoraires totaux d'un médecin peuvent-ils être modélisés (linéairement) en fonction du nombre de ses patients ?

1. **Réaliser le nuage de points des honoraires en fonction du nombre total de patients et calculer le coefficient de corrélation linéaire entre les variables** `honormke` **et** `cltota`**. Commenter.**
    a. Nuage de points de points des honoraires en fonction du nombre total de patients

    Code (Nuage de Points):

```
plot(cltota,honormke, type="n",main="Liaison entre les honoraires et le nombre total de patients", xlab="Nombre Total de Patients",ylab="Honoraires(en millier d'euros)")
text(cltota,honormke,1:500)
```

    Sorties :



    Commentaire :

    Il semble exister une liaison linéaire positive entre les honoraires et le nombre total de patients.

    b. **le coefficient de corrélation linéaire entre les variables** `honormke` **et** `cltota`

    Code (Coefficient de coefficient de corrélation linéaire et test associé) :

```
cor(honormke,cltota)
cor.test(honormke,cltota)
```

    Sorties :

```
> cor(honormke,cltota)
[1] 0.5627235
> cor.test(honormke,cltota)

        Pearson's product-moment correlation

data:  honormke and cltota
t = 15.191, df = 498, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4996904 0.6198283
sample estimates:
      cor
0.5627235
```

Commentaire :

Il existe une liaison linéaire significative (p<5%) positive (r=0,56) entre les honoraires et le nombre total de patients des médecins.

2. **le modèle de régression linéaire simple théorique correspondant**

   Modèle 1 : $Honormke_i = \alpha_0 + \alpha_1 Cltota_i + e_i$ avec $i = 1,\ldots,n = 500$

3. **Estimer le modèle et commenter les résultats (coefficient de détermination R^2, test de validité globale du modèle, test de significativité du paramètre associé à la variable `cltota` et interprétation du paramètre estimé). Représenter la droite de régression des honoraires sur le nombre total de patients sur le nuage de points.**

   a. **Régression linéaire des honoraires sur le nombre total de patients**

Code (regression) :

```
regression=lm(honormke~cltota)
regression
```

Sorties :

```
> regression=lm(honormke~cltota)
> regression

Call:
lm(formula = honormke ~ cltota)

Coefficients:
(Intercept)       cltota
   63.36484      0.04929
```

Commentaire :

$\widehat{\alpha_1} = 0,05$ et $\widehat{\alpha_0} = 63,36$

L'équation de la droite de régression est : y = 0,05x - 63,36

Code (summary de regression) :

```
summary(regression)
```

Sorties :

```
> summary(regression)

Call:
lm(formula = honormke ~ cltota)

Residuals:
    Min      1Q  Median      3Q     Max
-189.91  -69.32  -16.97   41.05  728.09

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.364835   8.286916   7.646 1.07e-13 ***
cltota       0.049293   0.003245  15.191  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 108.1 on 498 degrees of freedom
Multiple R-squared:  0.3167,    Adjusted R-squared:  0.3153
F-statistic: 230.8 on 1 and 498 DF,  p-value: < 2.2e-16
```

Commentaire :

1. R^2 = 0,31 : 31% de la variation totale des honoraires est expliquée par ce modèle linéaire, c'est-à-dire, par la variation du nombre total de patients. Donc, il reste encore une grande partie de la variabilité des honoraires n'est pas expliquée par la variation du nombre total de patients.
2. Test de validité globale du modèle
    a. H0 : tous les paramètres sont nuls sauf la constante
    b. p-value < 2.2e-16 < 5%, donc on rejette H0
    c. donc le modèle 1 est globalement valide
3. Test de significativité du paramètre de `Cltota`
    a. H0 : $\alpha_1 = 0$ contre H1 : $\alpha_1 \neq 0$
    b. p-value < 2e-16 < 5% donc on rejette H0
    c. donc la variable `Cltota` est significative
4. Donc, nous pouvons commenter son paramètre estimé $\widehat{\alpha_1} = 0,05$.

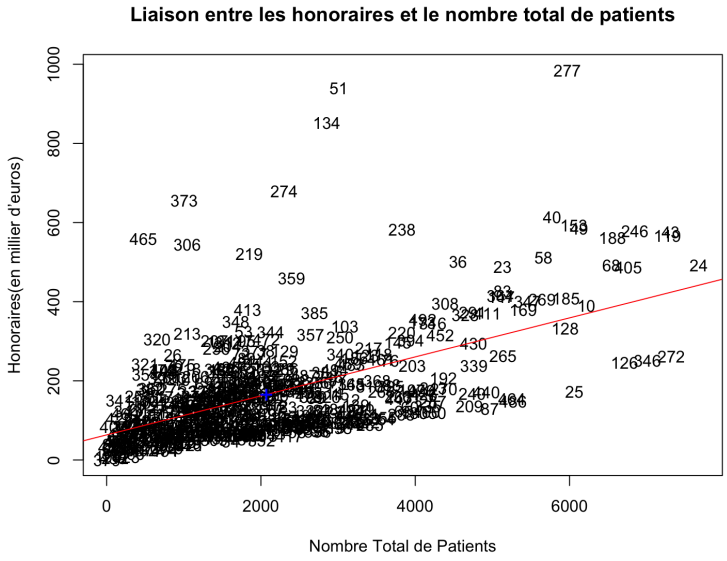
C'est-à-dire, 1 patient supplémentaire vu par le médecin fait augmenter ses honoraires totaux de 50 euros (0,05 millier d'euros).

b. la droite de régression des honoraires sur le nombre total de patients

Code :

```
abline(regression,col="red")
points(mean(cltota),mean(honormke),pch="+",col="blue",cex=1.5)
```

Sorties :



Commentaire :

Nous vérifions que la droite de régression passe par le point moyen de nuage.

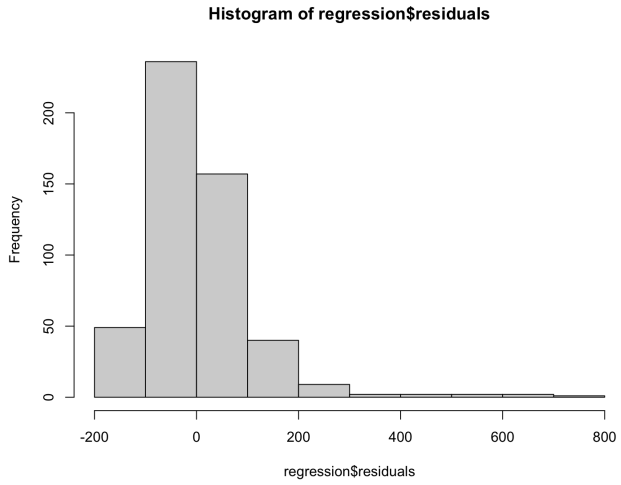
4. **Vérifier la normalité des résidus (histogramme + QQ-plot) et donner un graphique permettant de repérer d'éventuels points aberrants.**

   a. **Histogramme**

Code :

```
hist(regression$residuals)
```

Sorties :



Commentaire :

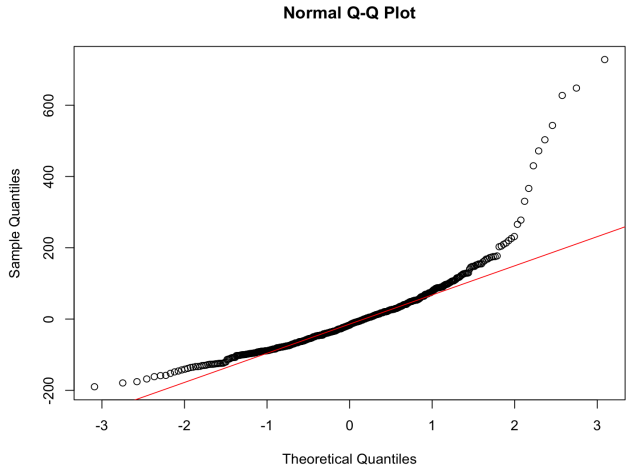
L'histogramme des résidus est très asymétrique avec un étalement à droite ce qui confirme que l'hypothèse de normalité des erreurs n'est pas vérifiée tout à fait sur ces données.

   b. **Quantile-quantile plot (Q-Q plot)**

Code :

```
qqnorm(regression$residuals)
qqline(regression$residuals,col="red")
```

Sortie :



Commentaire :

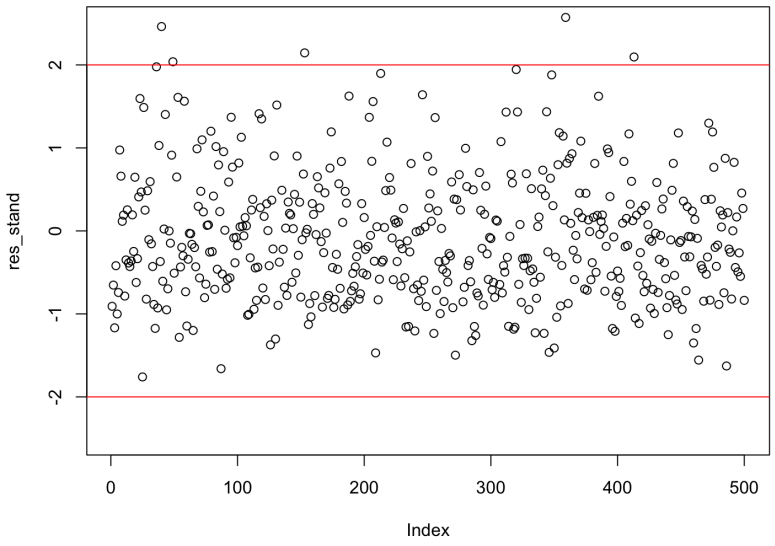
Nous voyons que les points ne sont pas vraiment alignés ce qui remet en cause l'hypothèse de normalité des erreurs

c. Repérer d'éventuels points aberrants

Code :

```
res_stand=regression$residuals/sd(regression$residuals)
plot(res_stand,ylim=c(-2.5,2.5))
abline(h=-2,col="red")
abline(h=2,col="red")
which(res_stand < -2)
which(res_stand > 2)
```

Sorties :



```
> which(res_stand < -2)
named integer(0)
> which(res_stand > 2)
 40  49  51 134 153 219 238 274 277 306 359 373 413 465
 40  49  51 134 153 219 238 274 277 306 359 373 413 465
```

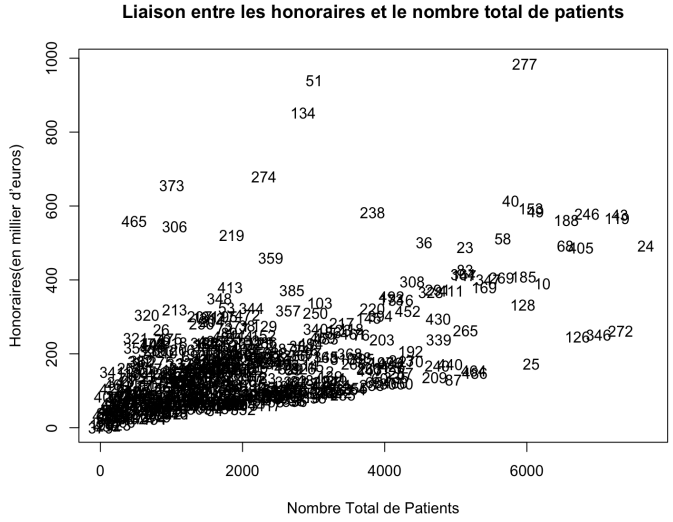
Commentaire :

Nous pouvons voir pourquoi il est aberrant en le repérant sur le nuage de points des honoraires en fonction du nombre total de patients.

Code (Nuage de Points):

```
plot(cltota,honormke, type="n",main="Liaison entre les honoraires et le nombre total de patients", xlab="Nombre Total de Patients",ylab="Honoraires(en millier d'euros)")
text(cltota,honormke,1:500)
```

Sorties :



Cmmentaire :

Les médecins dont l'indice est présent sur la sortie de la fonction " which(res_stand > 2)" ont des honoraires beaucoup plus hauts que les autres par rapport à leur nombre total de patients vus par le médecin, ce qui explique que ses honoraires soient mal prédit par la droite.

Partie 2 : régression linéaire multiple

Problématique : Quels sont les déterminants des honoraires totaux d'un médecin ?

1. **Effectuer la régression linéaire multiple du salaire en fonction de toutes les variables explicatives disponibles (ecrire le modèle théorique correspondant) et commenter les résultats obtenus (R^2, test de validité globale du modèle). Quelles sont les variables qui ont un effet significatif sur les honoraires ? On ne demande pas, dans cette question, de commenter ces effets.**

   <u>modèle 2</u> :

   $$Honormke_i = \alpha_0 + \alpha_1 Cltota_i + \alpha_2 Rless16_i + \alpha_3 R60to69_i + \alpha_4 Rmore70_i + \alpha_5 Secteur_i + \alpha_6 PrasexHomme_i + \alpha_7 Praspe3Médicale_i + \alpha_8 Praspe3Mixte_i + e_i$$

   avec $i = 1,\ldots, n = 500$

   Code :

```
regression1=lm(honormke~cltota+rless16+r60to69+rmore70+secteur+
prasex+praspe3)
summary(regression1)
```

   Sorties :

```
> regression1=lm(honormke~cltota+rless16+r60to69+rmore70+secteur+prasex+praspe3)
> summary(regression1)

Call:
lm(formula = honormke ~ cltota + rless16 + r60to69 + rmore70 +
    secteur + prasex + praspe3)

Residuals:
    Min      1Q  Median      3Q     Max
-185.05  -50.29  -10.67   32.35  658.73

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -32.308555  28.063247  -1.151 0.250179
cltota           0.051906   0.002994  17.336  < 2e-16 ***
rless16         -0.621828   0.287393  -2.164 0.030970 *
r60to69          2.672099   1.201393   2.224 0.026591 *
rmore70          0.374484   0.575629   0.651 0.515631
secteur         44.430071  12.733274   3.489 0.000528 ***
prasexHomme     49.186256   9.458922   5.200 2.93e-07 ***
praspe3Médicale -6.118731  15.635775  -0.391 0.695724
praspe3Mixte    -72.402969  16.842576  -4.299 2.07e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 91.17 on 491 degrees of freedom
Multiple R-squared:  0.5203,    Adjusted R-squared:  0.5125
F-statistic: 66.57 on 8 and 491 DF,  p-value: < 2.2e-16
```

   Commentaire :

   1. $R^2 = 0,52$ ⇒ 52% de la variation des honoraires est expliquée par le modèle linéaire, c'est-à-dire par la variation des variables explicatives du modèle.
      a. On voit que le R^2 a augmenté par rapport au modele precedent, mais cela était prévisible car le R^2 augmente automatiquement dès lors qu'on ajoute de variables. Il ne permet donc pas de comparer deux modèles entre eux.

b. On utilise le R^2 ajusté qui pénalise les modèles en fonction de leur nombre de paramètres : ici, le R^2 ajusté vaut 0,51 contre 0,31 dans le modèle 1.
c. Donc, le modèle 2 est meilleur que le modèle 1.

2. Test de validité globale du modèle
   a. H0 : tous les paramètres sont nuls sauf la constante
      ⇒ H0 : $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = 0$
   b. p-value < 2.2e-16 < 5% donc on rejette H0
   c. donc le modèle 2 est globalement valide
3. Test de significativité des paramètres des variables explicatives :
   a. Pour l'effet de `praspe3` (une variable qualitative explicative à 3 modalités)

Code :

```
regression2=lm(honormke~cltota+rless16+r60to69+rmore70+secteur+
prasex)
anova(regression2,regression1)
```

Sorties :

```
> anova(regression2,regression1)
Analysis of Variance Table

Model 1: honormke ~ cltota + rless16 + r60to69 + rmore70 + secteur + prasex
Model 2: honormke ~ cltota + rless16 + r60to69 + rmore70 + secteur + prasex +
    praspe3
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    493 4434965
2    491 4081620  2    353345 21.253 1.406e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Commentaire :

p-value=1.406e-09<5%, donc `praspe3` a un effet significatif sur les honoraires.

b. Toutes les variables ont un effet significatif sur les honoraires sauf `rmore70` (p-value=0,52>5%)

Les variables ayant un impact significatif sur les honoraires sont cltota, rless16, r60to69, secteur, prasexHomme et praspe3Mixte.

2. **Certains coefficients de la régression précédentes n'étant pas significatifs, recommencer un ajustement de régression linéaire multiple par la modèle de pas à pas vue en TP de façon à avoir à la fin un modèle ne contenant que des coefficients significatifs à 5%. Pour ce pas à pas, vous pouvez être amenées à créer une (des) variables(s) indicatrice(s). Donner toutes les sorties R de ce pas à pas. Pour le modèle final obtenu par le pas à pas :**
   Ecrire le modèle de régression linéaire théorique correspondant
   Commenter le R^2 et le test de validité globale du modèle
   Interpreter les parametres estimes

a. On enlève d'abord la variable la moins significative du modèle :

praspe3Médicale

Code :

```
indiF1=as.numeric(praspe3 == "Médicale")
indiF2=as.numeric(praspe3 == "Mixte")
regression3=lm(honormke~cltota+rless16+r60to69+rmore70+secteur+
prasex+indiF2)
summary(regression3)
```

Sorties :

```
> indiF1=as.numeric(praspe3 == "Médicale")
> indiF2=as.numeric(praspe3 == "Mixte")
>
> regression3=lm(honormke~cltota+rless16+r60to69+rmore70+secteur+prasex+indiF2)
> summary(regression3)

Call:
lm(formula = honormke ~ cltota + rless16 + r60to69 + rmore70 +
    secteur + prasex + indiF2)

Residuals:
    Min      1Q  Median      3Q     Max
-185.02  -51.06   -9.42   31.53  657.91

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -40.489529  18.705547  -2.165   0.0309 *
cltota        0.051790   0.002977  17.396  < 2e-16 ***
rless16      -0.627758   0.286747  -2.189   0.0290 *
r60to69       2.695921   1.198816   2.249   0.0250 *
rmore70       0.372295   0.575106   0.647   0.5177
secteur      46.851471  11.119184   4.214 2.99e-05 ***
prasexHomme  49.533183   9.409177   5.264 2.11e-07 ***
indiF2      -67.195780  10.316391  -6.513 1.82e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.1 on 492 degrees of freedom
Multiple R-squared:  0.5201,    Adjusted R-squared:  0.5133
F-statistic: 76.18 on 7 and 492 DF,  p-value: < 2.2e-16
```

Commentaire :

Le p-value de `rmore70` est 0.52 (>5%), donc, `rmore70` a un effet NS(non significative).

b. On enlève la variable la moins significative du modèle : `rmore70`

Code :

```
regression4=lm(honormke~cltota+rless16+r60to69+secteur+prasex+i
ndiF2)
summary(regression4)
```

Sorties :

```
> regression4=lm(honormke~cltota+rless16+r60to69+secteur+prasex+indiF2)
> summary(regression4)

Call:
lm(formula = honormke ~ cltota + rless16 + r60to69 + secteur +
    prasex + indiF2)

Residuals:
    Min      1Q  Median      3Q     Max
-185.03  -50.58   -8.88   31.27  655.96

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.868257  18.572950  -2.254   0.0246 *
cltota        0.051689   0.002971  17.397  < 2e-16 ***
rless16      -0.609659   0.285212  -2.138   0.0330 *
r60to69       3.289805   0.771247   4.266 2.39e-05 ***
secteur      47.126127  11.104538   4.244 2.63e-05 ***
prasexHomme  49.925886   9.384067   5.320 1.58e-07 ***
indiF2      -67.487170  10.300492  -6.552 1.43e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 91.04 on 493 degrees of freedom
Multiple R-squared:  0.5197,    Adjusted R-squared:  0.5139
F-statistic: 88.92 on 6 and 493 DF,  p-value: < 2.2e-16
```

Commentaire :

Toutes les variables sont significatives (toutes les p-values sont inférieures à 5%), donc on a fini le pas à pas. On retient ce modèle comme le modèle final.

c. **<u>Modele final</u>** :

$$Honormke_i = \alpha_0 + \alpha_1 Cltota_i + \alpha_2 Rless16_i + \alpha_3 r60to69_i + \alpha_4 secteur_i + \alpha_5 prasexHomme_i + \alpha_6 indiF2_i + e_i$$

avec $i = 1,\dots, n = 500$

d. le R^2

i. $R^2 = 0,52$ ⇒ 52% de la variation des honoraires est expliquée par le modèle linéaire, c'est-à-dire par la variation des variables explicatives du modèle.

ii. Ici, le R^2 ajusté vaut 0,51 contre 0,51 dans le modèle 1.

iii. Le modèle final est un modèle bien formé sortie du modèle 1.

e. Test de validité globale du modèle

i. H0 : tous les paramètres sont nuls sauf la constante
⇒ H0 : $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$

ii. p-value < 2.2e-16 < 5% donc on rejette H0

iii. donc le modèle final est globalement valide

f. Interprétation les paramètres estimés

i. $\widehat{\alpha_1}$ = 0,05 : Toutes choses égales par ailleurs, lorsque le nombre de patients vus par le médecin augmente de 1, les honoraires totaux augmentent de 50 euros (0,05 millier d'euros).

ii. $\widehat{\alpha_2}$ = -0,61 : Toutes choses égales par ailleurs, lorsque la proportion des patients âgés de moins de 16 ans dans la patientèle augmente de 1%, les honoraires totaux du médecin diminuent de 610 euros (0,61 millier d'euros).

iii. $\widehat{\alpha_3}$ = 3,29 : Toutes choses égales par ailleurs, lorsque la proportion des patients âgés de 60 à 69 ans dans la patientèle augmente de 1%, les honoraires totaux du médecin augmentent de 3290 euros (3,92 millier d'euros).

iv. $\widehat{\alpha_4}$ = 47,13 : Toutes choses égales par ailleurs, les médecins dans le secteur 2 gagnent en moyenne 47130 euros (47,13 millier d'euros) de plus par an que les médecins dans le secteur 1.

v. $\widehat{\alpha_5}$ = 49,93 : Toutes choses égales par ailleurs, les hommes médecins gagnent en moyenne 49930 euros (49,93 millier d'euros) de plus par an que les femmes médecins.

vi. $\widehat{\alpha_6}$ = -67,49 : Toutes choses égales par ailleurs, les médecins mixtes gagnent en moyenne 67490 euros (67,49 millier d'euros) de moins par an que les autres médecins.

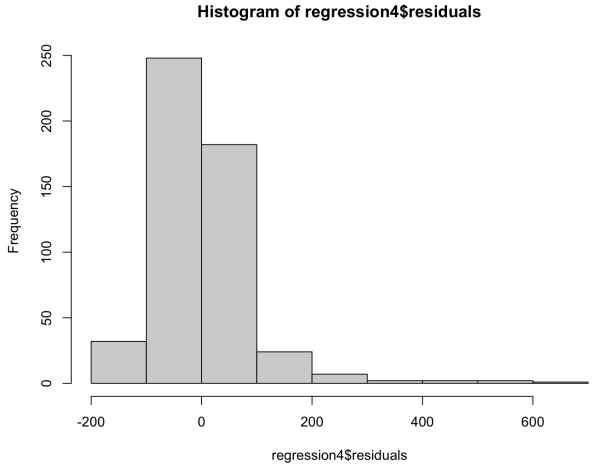
3. **Vérifier la normalité des résidus (histogramme + QQ-plot) et donner un graphique permettant de repérer d'éventuels points aberrants.**

   a. **Histogramme**

Code :

```
hist(regression4$residuals)
```

Sorties :



Commentaire :

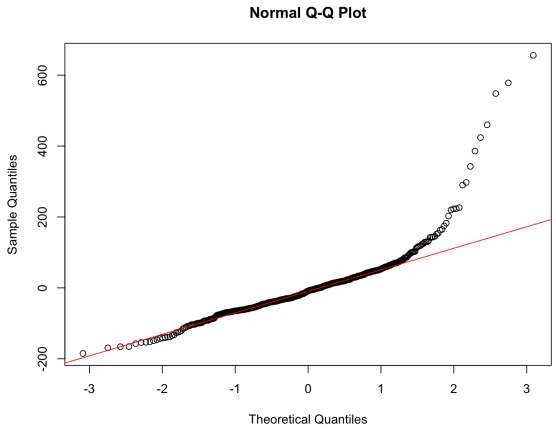
L'histogramme des résidus est très asymétrique avec un étalement à droite ce qui confirme que l'hypothèse de normalité des erreurs n'est pas vérifiée tout à fait sur ces données.

   b. **Quantile-quantile plot (Q-Q plot)**

Code :

```
qqnorm(regression4$residuals)
qqline(regression4$residuals,col="red")
```

Sortie :



Commentaire :

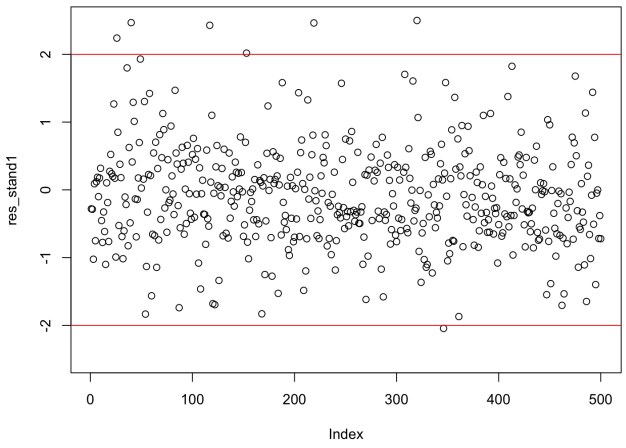
Les points ne sont pas alignés donc on peut remettre en cause l'hypothèse de normalité des erreurs. On peut penser qu'il existe donc des valeurs aberrantes.

c. Repérer d'éventuels points aberrants

Code :

```
res_stand1=regression4$residuals/sd(regression4$residuals)
plot(res_stand1,ylim=c(-2.5,2.5))
abline(h=-2,col="red")
abline(h=2,col="red")
which(res_stand1 < -2)
which(res_stand1 > 2)
```

Sorties :




```
> which(res_stand1 < -2)
346
346
> which(res_stand1 > 2)
 26  40  51 117 134 153 219 238 274 277 306 320 359 373 465
 26  40  51 117 134 153 219 238 274 277 306 320 359 373 465
```

Commentaire :

Nous pouvons voir qu'il y a beaucoup de points aberrants ici.

Conclusion : Synthèse d'étude

pour le modèle final :

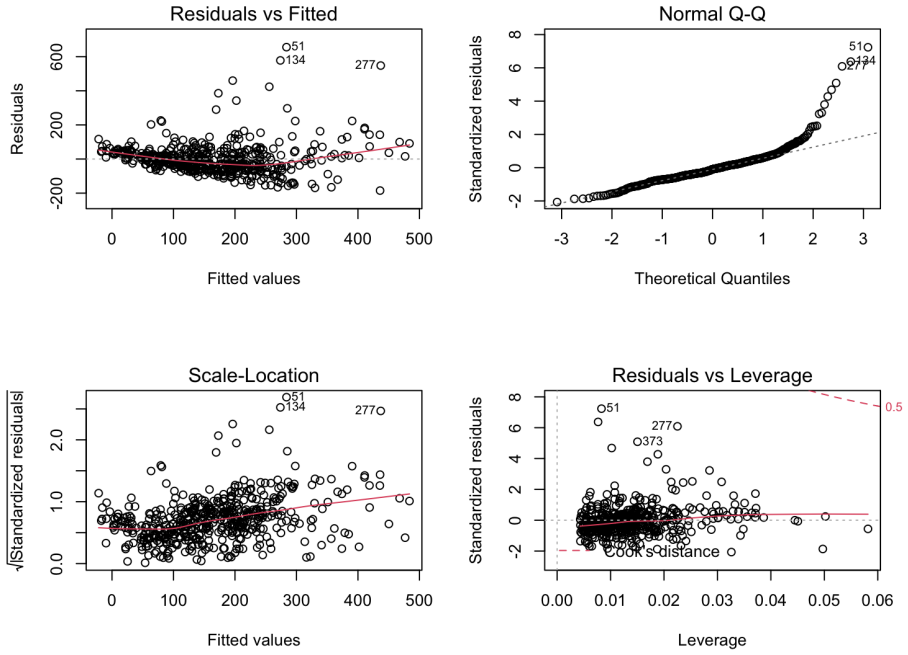
$VIF = \frac{1}{1-R_j^2} = 2.06 > 2$, donc on a un problème de colinéarité sévère. Il y a peut-être une linéarité totale dont l'une des variables est une combinaison linéaire des autres, ou une colinéarité approchée dont l'une forte corrélation entre les variables explicatives rend l'estimation des coefficients de régression imprécise.

$R^2 = 0.52$, c'est-à-dire, il y a encore d'autres variables explicatives qui sont significatives et ne sont pas prises en compte dans notre étude.

Code :

```
par(mfrow=c(2,2))
plot(regression4)
```

Sorties :



Commentaire :

Le 4ème graphique, Residuals vs. Leverage, permet d'identifier les points avec une forte influence. Les lignes pointillées marquent les seuils de 0.5 pour la distance de Cook. Ici, aucun des trois points extrêmes (`51`, `277` et `373`) ne dépasse 0,5, mais leur influence est supérieure de beaucoup à celle du reste des points. Une solution possible serait d'ignorer ces trois points extrêmes.

Donc, on peut essayer d'augmenter la taille d'échantillon, prendre plus des variables explicatives possibles en compte et réaliser une ACP sur les variables explicatives pour supprimer les corrélations et obtenir une estimation plus précise.