Projet de statistiques

Travail sur la régression linéaire

(TP2 Trinôme 8)

CHENG Wenxin
ZHAO Wenbo
BORDIER Benjamin

L3 MIAGE
Année 2022/2023

Introduction :

1. Introduction de la population concernée

Nous examinons un échantillon de données sur la population carcérale purgeant des peines. L'objectif de cette étude est d'identifier les variables qui ont un impact sur la durée de la détention. Nous avons ici un total de 800 échantillons de recherche, et 8 variables de référence.

2. les variables étudiées

Sur nos 8 variables, il y a 3 variables qualitatives et 5 variables quantitatives. Puisque dans les 3 variables quantitatives, les données sont identifiées par 0 et 1, nous pensons que sous R, cela n'est pas propice à la bonne lecture et à la génération adaptée des données. Le fait de le changer en fonction du type de codage et de ce qui est nécessaire rendra la sortie bien plus agréable à lire.

•`marie` : État civil des individus.

- Il s'agit d'une **variable qualitative nominale** à 2 modalités (non-marié /marié).

```
class(prison$marie)
prison$marie=factor(prison$marie,labels = c("non","oui"))
```

•`coupable` : si l'individu est coupable.

- Il s'agit d'une **variable qualitative nominale** à 2 modalités (non-coupable /coupable).

```
class(prison$coupable)
prison$coupable=factor(prison$coupable,labels = c("non","oui"))
```

•`motif` : Le motif de l'individu pour commettre l'infraction.

- Il s'agit d'une **variable qualitative nominale** à 3 catégories (autres, braquage et meurtre)

```
class(prison$motif)
prison$motif=factor(prison$motif,labels = c("autres","braquage","meurtre"))
```

•`nivetud` : nombre d'années d' études.

- Il s'agit d'une **variable quantitative discrète**.

•`nbcondamn` : nombre de condamnations antérieures.

- Il s'agit d'une **variable quantitative discrète**.

•`regle` : nombre de règles enfreintes pendant le séjour en prison.

- Il s'agit d'une **variable quantitative discrète**.

•`age` : age (en années) au début du séjour en prison.

- Il s'agit d'une **variable quantitative continue.**

•`durprison` : durée de la détention (en mois).
- Il s'agit d'une **variable quantitative continue.**

```
>   summary(prison)
   durprison       marie      coupable     nivetud
 Min.   :  1.00   non:593   non:557   Min.   : 1.000
 1st Qu.:  6.00   oui:207   oui:243   1st Qu.: 8.000
 Median : 12.00                       Median :10.000
 Mean   : 19.59                       Mean   : 9.751
 3rd Qu.: 25.00                       3rd Qu.:12.000
 Max.   :219.00                       Max.   :19.000
     regle             motif       nbcondamn           age
 Min.   : 0.000   autres  :558   Min.   : 0.000   Min.   :17.00
 1st Qu.: 0.000   braquage:201   1st Qu.: 0.000   1st Qu.:22.00
 Median : 0.000   meurtre : 41   Median : 0.000   Median :26.00
 Mean   : 1.199                  Mean   : 1.491   Mean   :29.12
 3rd Qu.: 1.000                  3rd Qu.: 2.000   3rd Qu.:34.00
 Max.   :27.000                  Max.   :28.000   Max.   :78.00
```

3. le but de l'étude

Comme indiqué dans l'énoncé, l'objectif de cette étude est d'identifier les variables qui ont un effet sur la durée de la détention.
En d'autres termes, nous voulons savoir si **l'état civil, le fait d'être coupable ou non, le motif, la durée des études, l'âge, le nombre d'infractions antérieures et le nombre de règles enfreintes en prison** ont un effet sur **le durprison**, et si c'est le cas, nous voulons étudier et mesurer cet effet de manière plus approfondie.

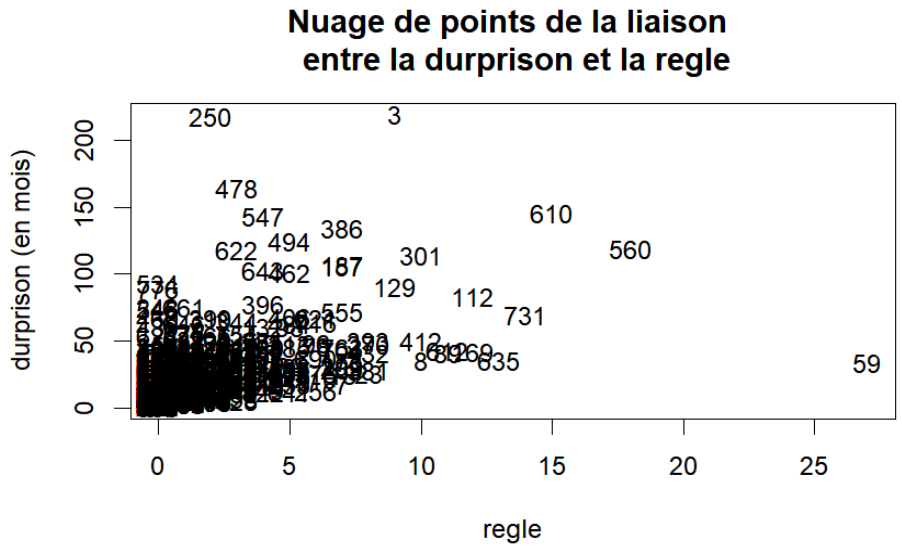
Partie 1 : régression linéaire simple

Problématique : la durée de détention peut-elle être modélisée (linéairement) en fonction du nombre de règles enfreintes pendant le séjour en prison?

1. Réaliser le nuage de points de la durée de détention en fonction du nombre de règles enfreintes en prison et calculer le coefficient de corrélation linéaire entre les variables durprison et règle. Commenter.

a) Nuage de points de la durée de détention en fonction du nombre de règles enfreintes en prison :

```
plot(regle, durprison,
     type = "n",
     main = "Nuage de points de la liaison \n entre la durprison et la regle",
     xlab = "regle",
     ylab = "durprison (en mois)")
text(regle,durprison, 1:800)
```



Commentaire :

Il semble y avoir une corrélation linéaire positive entre règle et durprison.

b) Coefficient de corrélation linéaire entre règle et durprison:

```
>     cor(durprison, regle)
[1] 0.5234419
>     cor.test(durprison, regle)

        Pearson's product-moment correlation

data:  durprison and regle
t = 17.354, df = 798, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4712247 0.5720027
sample estimates:
      cor
0.5234419
```

Commentaire :

P-value = 2.2e-16 < 5% rejet de H0 c'est-à-dire rejet de "pas de corrélation linéaire".

Il existe une association linéaire positive et significative relativement forte entre durprison et règle, (coeff de corr linéaire empirique ≈ 0,52).

Donc la régression linéaire de durprison sur règle est raisonnable.

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

Modèle 1 : durprison_i = a regle_i + b + e_i pour i=1,...,n (n=800)

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 règle et interprétation du paramètre estimé). Représenter la droite de régression de la durée de détention sur le nombre de règles enfreintes en prison sur le nuage de points.

a) Estimer le modèle et commenter les résultats

Régression linéaire de durprison sur regle:

```
>     regression1=lm(durprison~regle)
>     regression1

Call:
lm(formula = durprison ~ regle)

Coefficients:
(Intercept)        regle
     13.487        5.095
```

Commentaire :

Coefficients de régression estimés : a_chapeau = 5.095 et b_chapeau=13.487

Équation de la ligne de régression : y = 5.095 x + 13.487

```
>     summary(regression1)

Call:
lm(formula = durprison ~ regle)

Residuals:
     Min       1Q   Median       3Q      Max
-117.060   -8.880   -5.487    4.513  194.322

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.4870     0.7639   17.66   <2e-16 ***
regle         5.0953     0.2936   17.35   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.18 on 798 degrees of freedom
Multiple R-squared:  0.274,     Adjusted R-squared:  0.2731
F-statistic: 301.2 on 1 and 798 DF,  p-value: < 2.2e-16
```

Commentaire :

1. Le coefficient déterministe R^2 ≈ 0.27, 27% de la variation totale de la durée de détention est expliquée par le modèle 1, c'est-à-dire par la variation de la durée de détention. Ainsi, il y a toujours une grande partie de la variation qui n'est pas expliquée par la variable `regle`. R^2 est relativement proche de 0 et le modèle ne s'ajuste pas bien aux données.

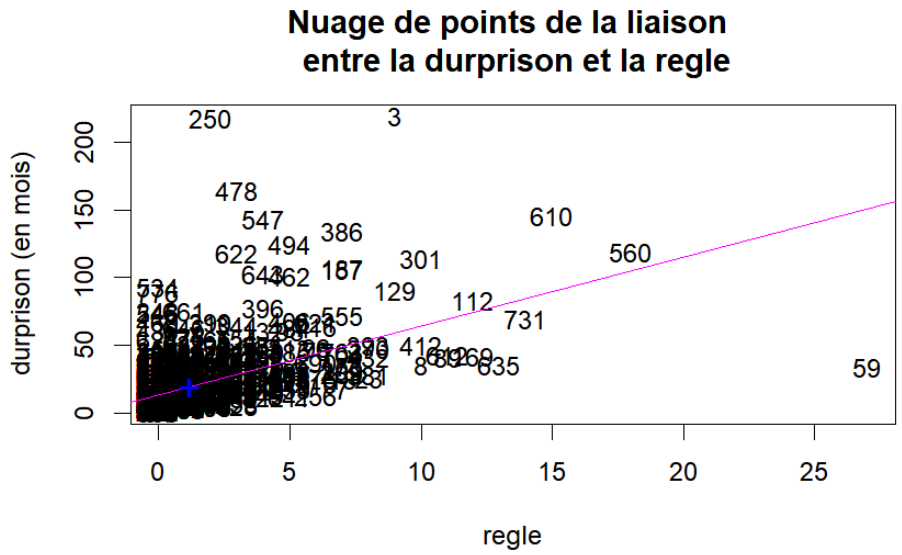
2. Tester la validité globale du modèle :
   on teste H0 et H1 :
   H0 : Tous les paramètres sont nuls sauf la constante, c'est-à-dire a = 0 dans une régression linéaire simple (une seule variable explicative).
   H1 : au moins un paramètre sauf la constante est différent de 0, c'est-à-dire que a est différent de 0 (dans une régression linéaire simple)
   P-value = 2.2e-16 < 5% Nous rejetons H0, alors le modèle est globalement valide.

3. Test pour la signification de la variable `regle` :
   H0 : a=0 contre H1 : a≠0
   P-value=2e-16 < 5% Nous rejetons H0, alors la variable `regle` est significative.

4. Le paramètre estimé â=5.0953 peut être interprété comme suit : pour chaque violation supplémentaire de `regle`, la durée d'emprisonnement augmente de 5.0953 mois.

b) Ajouter une droite de régression de la durée de détention sur le nombre de règles enfreintes en prison sur le nuage de points.

```
abline(regression1, col="magenta")
points(mean(regle), mean(durprison), pch="+", col="blue", cex=1.5)
```

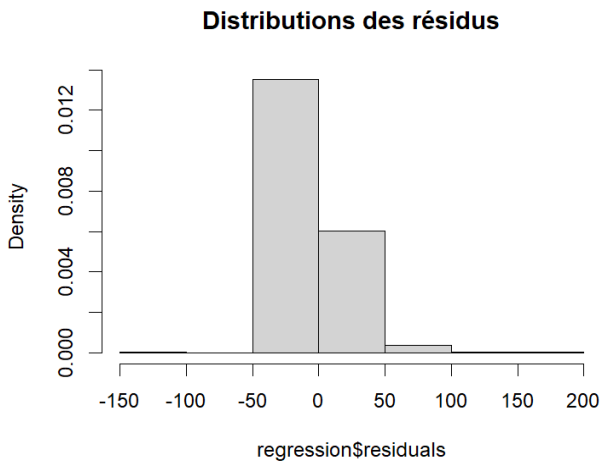


Commentaire :
Nous vérifions que la droite de régression passe par le point moyen (point représentant la moyenne) du nuage.

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

a) Histogramme:

```
> #normalité des résidus
> hist(regression1$residuals,freq=FALSE,
+      main="Distributions des résidus")
```

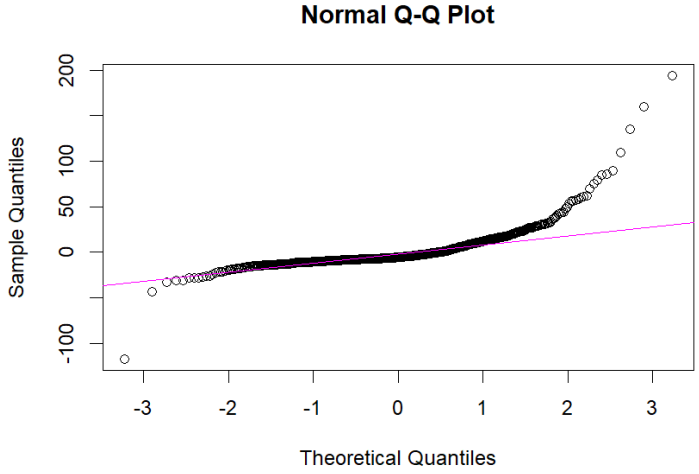


Commentaires :

L'histogramme des résidus est asymétrique, s'étalant vers la droite. Cela confirme que l'hypothèse de normalité des erreurs n'est pas entièrement vérifiée sur ces données.

b) QQ-Plot :

```
> #quantile-quantile plot
> qqnorm(regression1$residuals)
> qqline(regression1$residuals, col="magenta")
```



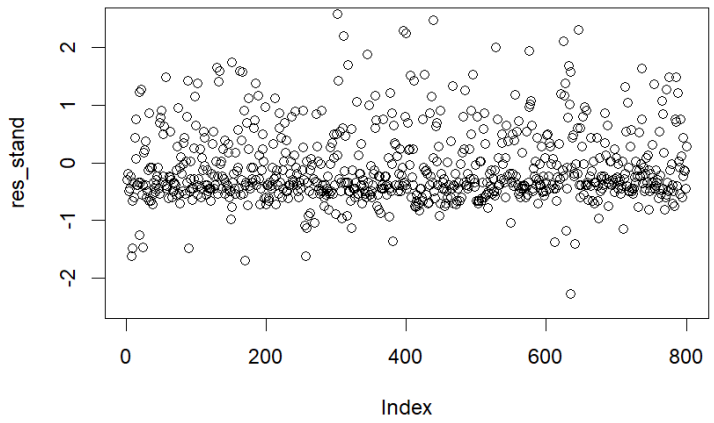
Commentaires :

Pour le Q-Q plot, les points doivent être alignés si les erreurs sont normalement distribuées. Ici les points ne sont pas alignés donc pas gaussien. D'après le graphique, nous pouvons conclure que l'hypothèse de normalité des résidus n'est pas vraiment vérifiée.
Ensuite, nous devons trouver les valeurs aberrantes.

c) Repérer d'éventuels points aberrants

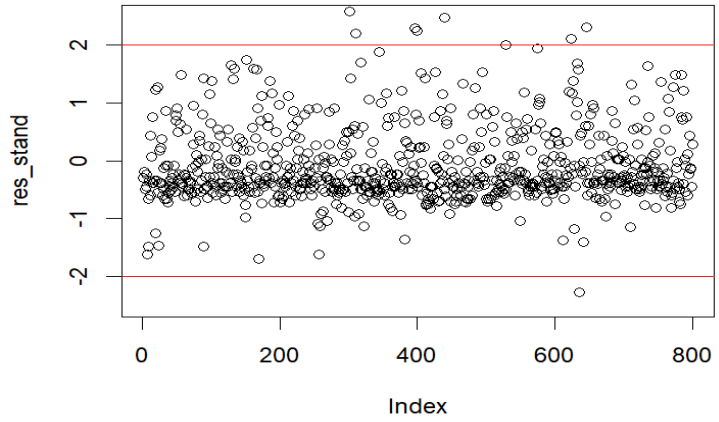
résidus standardisés : res_stand = regression1$residuals / sd(regression1$residuals)
Donc on a *(res_stand = regression1$residuals/19.18)*
19.18 est l'écart-type estimé des résidues, donné dans la sortie de la régression.

```
plot(res_stand,ylim=c(-2.5,2.5))
```



Nous utilisons la fonction abline() pour identifier les individus dont les valeurs absolues des résidus normalisés sont supérieures à 2. Ensuite, nous pouvons également utiliser identify() afin d'identifier les anomalies.

```
abline(h=-2, col="red")
abline(h=2, col="red")
identify(res_stand)
```



Un individu est dit atypique (ou outlier ou aberrant) si son résidu standardisé a une valeur en dehors de l'intervalle[-2, 2] (pour n - k > 30) ici n=800,k=8, n-k=792>30.

Enfin, nous utilisons la fonction which pour filtrer les points d'échantillonnage dont nous avons besoin.

```
which(res_stand < -2)
which(res_stand > 2)
```

```
> which(res_stand < -2)
 59 635
 59 635
>     which(res_stand > 2)
  3 157 187 248 250 301 310 386 396 399 439 460 462
  3 157 187 248 250 301 310 386 396 399 439 460 462
478 494 528 534 546 547 610 622 624 643 646 661 776
478 494 528 534 546 547 610 622 624 643 646 661 776
```

Commentaires :
Les prisonniers qui apparaissent sur `res_stand > 2` avaient des mois de détention beaucoup plus élevés que les autres, ce qui explique pourquoi leurs mois de détention étaient plus mal prédits dans le modèle.

Partie 2 : régression linéaire multiple

Problématique : quels sont les déterminants de la durée de détention?

1. Effectuer la régression linéaire multiple de la durée de détention en fonction de toutes les variables explicatives disponibles (écrire 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 la durée de détention? On ne demande pas, dans cette question, de commenter ces effets.

Modèle 2 : durprison_i = a_0 + a_1*marie_i + a_2 + *coupable_i + a_3*nivetud_i
+ a_4*regle_i + a_5*motifbraquage_i + a_6*motifmeurtre_i
+ a_7*nbcondamn_i + a_8*age_i + e_i pour i=1,...,n (n=800)

```
> regression2=lm(durprison~marie+coupable+nivetud+regle+motif+nbcondamn+age)
> summary(regression2)

Call:
lm(formula = durprison ~ marie + coupable + nivetud + regle +
    motif + nbcondamn + age)

Residuals:
    Min      1Q  Median      3Q     Max
-90.700  -7.077  -1.878   3.046 152.052

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.34113    3.45444   2.994 0.002843 **
marieoui     -2.13754    1.39347  -1.534 0.125435
coupableoui  13.18265    2.13695   6.169 1.10e-09 ***
nivetud      -0.70161    0.24990  -2.808 0.005114 **
regle         4.18193    0.26812  15.597  < 2e-16 ***
motifbraquage  4.11942    2.20263   1.870 0.061822 .
motifmeurtre  17.22226    3.27174   5.264 1.82e-07 ***
nbcondamn     0.73845    0.22186   3.328 0.000914 ***
age           0.15837    0.06765   2.341 0.019477 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.55 on 791 degrees of freedom
Multiple R-squared:  0.4638,    Adjusted R-squared:  0.4583
F-statistic: 85.51 on 8 and 791 DF,  p-value: < 2.2e-16
```

Commentaires :

1. Le coefficient déterministe R^2 ≈ 0.46, 46% de la variation totale de la durée de détention est expliquée par le modèle 2.
   On voit que le R^2 a augmenté par rapport au modèle précédent, mais cela était prévisible car le R^2 augmente automatiquement dès lors qu'on rajoute des variables ; il ne permet donc pas de comparer deux modèles entre eux.
   Pour cela, il faut utiliser le R2 ajusté qui pénalise les modèles en fonction de leur nombre de paramètres. Ici le R2 ajusté vaut 0.4583 contre 0.2731 dans le modèle 1 donc le modèle 2 est meilleur que le modèle 1.

2. Tester la validité globale du modèle :
   On teste H0 contre H1 :
   H0 : Tous les paramètres sont nuls sauf la constante, c'est-à-dire a = 0 dans une régression linéaire simple (une seule variable explicative).
   H1 : au moins un paramètre (sauf la constante) est différent de 0, c'est-à-dire que a est différent de 0 (dans une régression linéaire simple)
   P-value = 2.2e-16 < 5% Nous rejetons H0, alors le modèle est globalement valide.

3. Test pour la signification du paramètre de règle :
   a) Pour `marieoui`, on voit que P-value = 0.125435 < 5% , donc la variable `maire` n'est pas significative.
   b) Pour `motif`, une variable qualitative nominale à 3 catégorie (autres braquage et meurtre), Il faut une analyse plus approfondie.

```
> regression2=lm(durprison~marie+coupable+nivetud+regle+motif+nbcondamn+age)
> regression2bis=lm(durprison~marie+coupable+nivetud+regle+nbcondamn+age)
> anova(regression2bis,regression2)
Analysis of Variance Table

Model 1: durprison ~ marie + coupable + nivetud + regle + nbcondamn +
    age
Model 2: durprison ~ marie + coupable + nivetud + regle + motif + nbcondamn +
    age
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    793 224522
2    791 216731  2    7791.2 14.218 8.586e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On voit que P-value de `motif` est 8.586e-07 (<5%), donc, `motif` a un effet significatif.

Par conséquent, nous pouvons maintenant conclure que `coupable, nivetud, regle, motif, nbcondamn, age` ont un effet significatif sur la durée de la détention.

2. Certains coefficients de la régression précédente n'étant pas significatifs, recommencer un ajustement de régression linéaire multiple par la méthode de pas à pas vue en TP de façon à avoir à la fin un modèle ne contenant que des coefficients significatifs à s %. Pour ce pas à pas, vous pouvez être amenés 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:Écrire le modèle de régression linéaire théorique correspondant,Commenter le R et le test de validité globale du modèle,Interpréter les paramètres estimés.

1. Ici, on veut un pas à pas avec un critère qui permette d'avoir à la fin toutes les variables significatives.

a) On enlève d'abord la variable `marie` qui est la moins significative du modèle (celle avec la plus grande p-valeur de la sortie de régression).

```
> regression3=lm(durprison~coupable+nivetud+regle+motif+nbcondamn+age)
> summary(regression3)

Call:
lm(formula = durprison ~ coupable + nivetud + regle + motif +
    nbcondamn + age)

Residuals:
    Min      1Q  Median      3Q     Max
-91.081  -6.903  -2.166   2.895 153.102

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.34557    3.45739   2.992 0.002855 **
coupableoui   13.03131    2.13649   6.099 1.66e-09 ***
nivetud       -0.69022    0.25000  -2.761 0.005899 **
regle          4.20949    0.26774  15.722  < 2e-16 ***
motifbraquage  4.22879    2.20336   1.919 0.055312 .
motifmeurtre  17.30645    3.27407   5.286 1.62e-07 ***
nbcondamn      0.77012    0.22109   3.483 0.000522 ***
age            0.13314    0.06567   2.027 0.042970 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.57 on 792 degrees of freedom
Multiple R-squared:  0.4622,    Adjusted R-squared:  0.4574
F-statistic: 97.22 on 7 and 792 DF,  p-value: < 2.2e-16
```

b) Nous pouvons voir qu'actuellement le P-value de motifbraquage est 0.055312 (>0.05), alors pour enlever le motifbraquage, nous allons créer des indicatrices.

```
> indiM=as.numeric(motif=="meurtre")
> regression4=lm(durprison~coupable+nivetud+regle+indiM+nbcondamn+age)
> summary(regression4)
```

```
Call:
lm(formula = durprison ~ coupable + nivetud + regle + indiM +
    nbcondamn + age)

Residuals:
    Min      1Q  Median      3Q     Max
-88.792  -6.913  -2.131   3.093 152.535

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.75270    3.45671   3.111 0.001933 **
coupableoui 16.17264    1.37551  11.758  < 2e-16 ***
nivetud     -0.69883    0.25039  -2.791 0.005380 **
regle        4.27432    0.26605  16.066  < 2e-16 ***
indiM       14.19187    2.84841   4.982 7.71e-07 ***
nbcondamn    0.80002    0.22091   3.621 0.000312 ***
age          0.12704    0.06571   1.933 0.053537 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.59 on 793 degrees of freedom
Multiple R-squared:  0.4597,    Adjusted R-squared:  0.4556
F-statistic: 112.4 on 6 and 793 DF,  p-value: < 2.2e-16
```

c) Nous pouvons voir qu'actuellement le P-value d'âge est 0.053537 (>0.05), alors on enlève l'âge.

```
> regression5=lm(durprison~coupable+nivetud+regle+indiM+nbcondamn)
> summary(regression5)

Call:
lm(formula = durprison ~ coupable + nivetud + regle + indiM +
    nbcondamn)

Residuals:
    Min      1Q  Median      3Q     Max
-86.952  -6.855  -2.090   3.028 153.227

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.2222     2.5744   5.913 4.99e-09 ***
coupableoui  16.0638     1.3767  11.668  < 2e-16 ***
nivetud      -0.7925     0.2461  -3.221  0.00133 **
regle         4.1801     0.2620  15.955  < 2e-16 ***
indiM        14.8856     2.8306   5.259 1.87e-07 ***
nbcondamn     0.9707     0.2029   4.785 2.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.62 on 794 degrees of freedom
Multiple R-squared:  0.4571,    Adjusted R-squared:  0.4537
F-statistic: 133.7 on 5 and 794 DF,  p-value: < 2.2e-16
```

Commentaire:

Toutes les variables sont significatives donc on a fini le pas à pas et on retient comme modèle final le Modèle 5.

d) Modèle final

Par conséquent, nous pouvons conclure que `coupable, nivetud, regle, motif_meutre, nbcondamn` a un effet plus significatif sur `durprison`.

Modèle final: durprison_i = a_0 + a_1*coupable_i + a_2 + *nivetud_i + a_3*regle_i + a_4*motifbraquage_i + a_5*nbcondamn_i + e_i pour i=1,...,n (n=800)

2. Coefficient R^2

Le coefficient déterministe R^2 ≈ 0.46, 46% de la variation totale de la durée de détention est expliquée par le modèle final.

Il faut utiliser le R^2 ajusté qui pénalise les modèles en fonction de leur nombre de paramètres pour comparer deux modèles.

R^2=0.4537 .Dans le modèle 2: R^2=0.4583. Il est quasiment identique et comme on préfère garder un modèle avec moins de paramètres et où toutes les variables sont significatives, on retient bien le modèle final.

3. Tester la validité globale du modèle

on teste H0 contre H1 :
H0 : Tous les paramètres sont nuls sauf la constante, c'est-à-dire a = 0 dans une régression linéaire simple (une seule variable explicative).
H1 : au moins un paramètre sauf la constante est différent de 0, c'est-à-dire que a est différent de 0 (dans une régression linéaire simple).
P-value = 2.2e-16 < 5% Nous rejetons H0, alors le modèle final est globalement valide.

4. Interpréter les paramètres estimés

c1_chapeau = 16.06
Toutes choses égales par ailleurs, une personne `coupable` passe en moyenne 16.06 mois de plus en `durprison` que la personne non coupable.

c2_chapeau = -0.79
Toutes choses égales par ailleurs, une `nivetud` supplémentaire fait diminuer la `durprison` de 0.82 mois en moyenne.

c3_chapeau = 4.18
Toutes choses égales par ailleurs, une `règle` enfreinte pendant le séjour en prison de plus fait augmenter le `durprison` de 4.18 mois de moyenne.

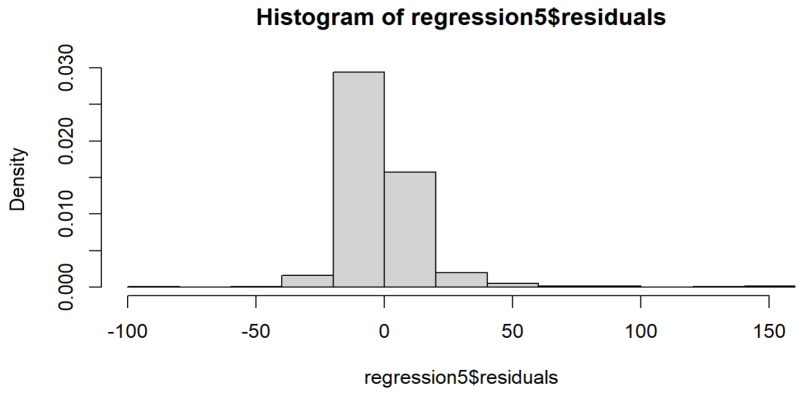
c4_chapeau = 14.89
Toutes choses égales par ailleurs, une incarcération pour `motif_meurtre` augmente en moyenne la `durprison` de 14.89 mois de plus que les autres motifs.

c5_chapeau = 0.97
Toutes choses égales par ailleurs, une `condamnation` antérieure de plus fait augmenter la `durprison` de 0.97 mois en moyenne.

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

a) Histogramme

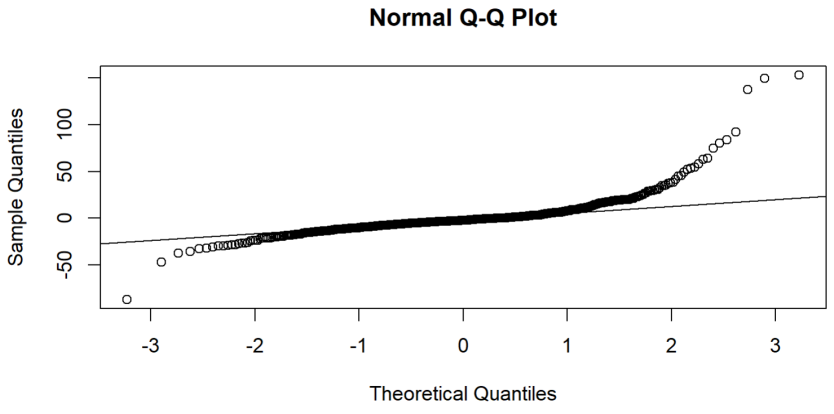
```
> hist(regression5$residuals,freq=FALSE)
```



Commentaires :
L'histogramme des résidus est asymétrique, s'étalant vers la droite. Cela confirme que l'hypothèse de normalité des erreurs n'est pas entièrement vérifiée sur ces données.

b) QQ-plot

```
> qqnorm(regression5$residuals)
> qqline(regression5$residuals)
```



Commentaire :

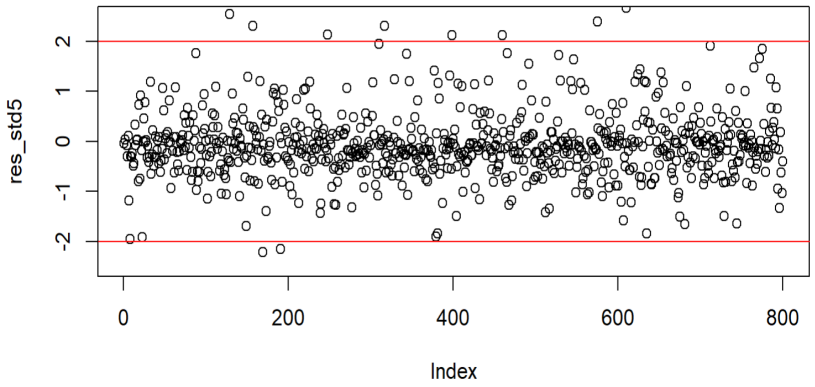
Pour le Q-Q plot, les points doivent être alignés si les erreurs sont normalement distribuées. Ici les points ne sont pas alignés donc pas gaussien. D'après le graphique, nous pouvons conclure que l'hypothèse de normalité des résidus n'est pas vraiment vérifiée.

Ensuite, nous devons trouver les valeurs aberrantes.

c) Repérer d'éventuels points aberrants

Nous pouvons également voir par l'histogramme que la gauche et la droite sont symétriques et donc non gaussiennes.

```
> res_std5=regression4$residuals/16.62
> plot(res_std5,ylim=c(-2.5,2.5))
> abline(h=-2,col="red")
> abline(h=2,col="red")
```



Un individu est dit atypique (ou outlier ou aberrant) si son résidu standardisé a une valeur en dehors de l'intervalle[-2, 2] (pour n - k > 30) ici n=800,k=8, n-k=792>30.

```
> which(abs(res_std5)> 2) # beaucoup de points aberrants ici
  3  59 129 157 169 187 191 248 250 301 317 386 399 439 460 462 478 494 534 547 575 610 612
  3  59 129 157 169 187 191 248 250 301 317 386 399 439 460 462 478 494 534 547 575 610 612
622 643 661 776
622 643 661 776
```

Commentaire:

Il y a beaucoup de points aberrants ici, la normalité n'est pas vérifiée, on ne peut pas interpréter les paramètres estimés, le test de significativité n'est donc plus valide.

Conclusion de l'étude

Faire une courte synthèse de votre étude en précisant les limites éventuelles de cette étude quant à la fiabilité de vos résultats.

Nous pouvons conclure que `coupable`, `nivetud`, `regle`, `motif_meutre`, `nbcondamn` ont un effet plus significatif sur `durprison`.

Le coefficient déterministe R^2 ≈ 0.46, 46% de la variation totale de la durée de détention est expliqué par le modèle final. Ainsi, il y a encore d'autres parties de la variation qui ne sont pas expliquées, elles sont représentées par les 54% restants.

Limites possibles de cette étude

- La taille de notre échantillon doit être aussi grande que possible. Trouver des associations significatives dans les données sera un défi si l'échantillon est petit, car les tests statistiques nécessitent souvent des échantillons de plus grande taille pour assurer une représentation équitable et fiable, ce qui peut être une limite.

- L'opportunité de l'étude. Les données trop anciennes ne sont souvent pas représentatives du passé récent, car les variables que nous étudions évoluent au fil du temps. Nous devons donc actualiser les données en permanence pour garantir leur exactitude et leur fiabilité.

- Problèmes liés aux échantillons d'étude et à la sélection des échantillons. Les inexactitudes d'échantillonnage se produisent lorsque des méthodes d'échantillonnage probabiliste sont utilisées pour sélectionner un échantillon qui ne représente pas fidèlement l'ensemble de la population ou du groupe d'intérêt. Par conséquent, notre étude souffrira peut-être de "biais d'échantillonnage" ou de "biais de sélection". Par exemple, si nous sélectionnons un échantillon de prisonniers dans un seul quartier d'une ville, il ne représente pas fidèlement l'ensemble de la population carcérale du pays. Nous avons donc élargi notre recherche lors de la sélection de l'échantillon afin de garantir un échantillon diversifié.