Evaluation 1 - Rapport

Technologie de l'e-commerce et mobiles - Big Data

Khalladi Mohamed - B32-DA

2023-10-19

Contents

1 ANOVA 1

1.1 Les civilisations précolombiennes

Le Ministère de la Culture de Batracie a étudié le nombre de récipients contenant de la bière fermentée sur divers sites archéologiques correspondants chacun à l'un des 4 types de civilisations précolombiennes suivantes: Cuacuacomeqiqi, Oxomatl, Tlaloc et Tenochtitlan.

Observe-ton des différences significatives entre les quatre traitements et quels sont ceux qui sont, s'ils existent, à résultats similaires ?

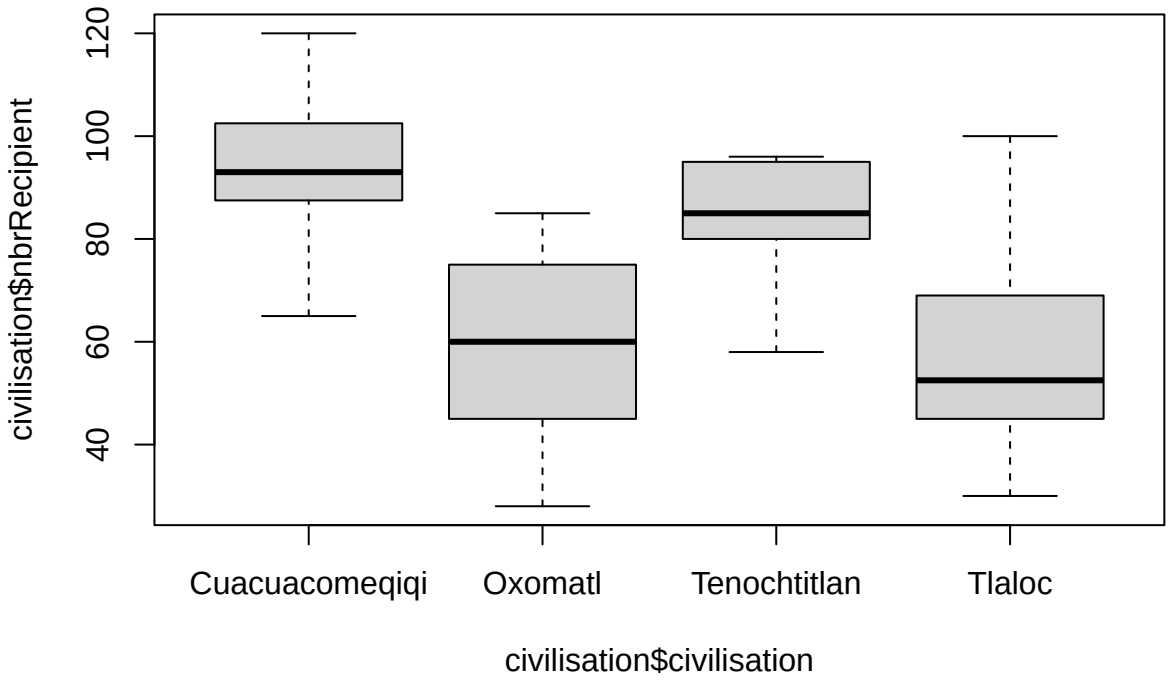
```
setwd("C:\\Users\\amine\\OneDrive\\Bureau\\EcomStat\\Labo\\datasets")

civilisation <- read.table("civilisation.csv", header=TRUE, sep=";", dec=".")

summary(civilisation)
```

```
##   nbrRecipient    civilisation
##  Min.   : 28.00   Length:46
##  1st Qu.: 55.75   Class :character
##  Median : 79.00   Mode  :character
##  Mean   : 75.04
##  3rd Qu.: 93.00
##  Max.   :120.00
```

```
boxplot(civilisation$nbrRecipient~civilisation$civilisation)
```



Avec notre boxplot, on peut remarquer que les variances sont assez différentes, mais il y a un rapprochement entre le groupe Oxomatl et Tenochtitlan. Nous allons donc effectuer différents tests pour vérifier de manière objective si le traitement a une influence sur le nombre de récipients.

```
model<-lm(civilisation$nbrRecipient~civilisation$civilisation)
model
```

```
##
## Call:
## lm(formula = civilisation$nbrRecipient ~ civilisation$civilisation)
##
## Coefficients:
##                               (Intercept)       civilisation$civilisationOxomatl
##                                     94.67                                  -35.37
## civilisation$civilisationTenochtitlan        civilisation$civilisationTlaloc
##                                   -11.78                                  -36.92
```

Ici on a donc créer notre modéle linéaire, qui nous donne donc l'intercept qui est l'estimation de mu du premier groupe et la différence (ecart à la moyenne) des autres par rapport au premier intercept.
Maintenant on peut comparer les moyennes entres-elles et vérifier si le traitement a bien une influence.
H0 : mu1 = mu2 = mu3 . . .
H1 : Au moins une différence.

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: civilisation$nbrRecipient
##                            Df Sum Sq Mean Sq F value    Pr(>F)
## civilisation$civilisation   3  12397  4132.4  15.139 7.991e-07 ***
## Residuals                  42  11465   273.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le test de l'anova nous renvoie une p-value miniscule (7.991e-07), on peut donc rejeter l'hypothése H0, et dire que la différence est bien significative.

```
summary(model)
```

```
##
## Call:
## lm(formula = civilisation$nbrRecipient ~ civilisation$civilisation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.300 -11.825  -1.278  11.758  42.250
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           94.667      4.266  22.192  < 2e-16 ***
## civilisation$civilisationOxomatl     -35.367      6.745  -5.243 4.81e-06 ***
## civilisation$civilisationTenochtitlan -11.778      6.966  -1.691   0.0983 .
```

```
## civilisation$civilisationTlaloc        -36.917      6.399  -5.769 8.52e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.52 on 42 degrees of freedom
## Multiple R-squared:  0.5195, Adjusted R-squared:  0.4852
## F-statistic: 15.14 on 3 and 42 DF,  p-value: 7.991e-07
```

Le summary du modéle linéaire repond a cette question grâce au test de fisher (p-value) On peut donc rejeter H0 , La différence est bien significative entre les populations.
On utilise un test de conformité de moyenne.
H0 : mu1 = 0 si la p-value est trop petite on rejette le H0 et on peut faire confiance a la valeur de mu1 donner par l'intercept. Après on unitilise un test d'homogénéité de moyenne :
H0 : sigma2 = 0 ou mu1 = mu2

Considérons maintenant que les variances soient inégales :

```
oneway.test(civilisation$nbrRecipient~civilisation$civilisation, var.equal=FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  civilisation$nbrRecipient and civilisation$civilisation
## F = 14.03, num df = 3.000, denom df = 21.431, p-value = 2.791e-05
```

Ici le oneway.test va passer par welch, et on obtient une p-value miniscule, cela nous raméne a conclure qu'il y'a bien une différence significative.
On peut également aller voir de plus prés en comparant par paire, pour vérifier quels populations varient fortement.

```
pairwise.t.test(civilisation$nbrRecipient, civilisation$civilisation,
p.adjust.method ="none", pool.sd=TRUE)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  civilisation$nbrRecipient and civilisation$civilisation
##
##              Cuacuacomeqiqi Oxomatl Tenochtitlan
## Oxomatl      4.8e-06        -       -
## Tenochtitlan 0.0983         0.0034  -
## Tlaloc       8.5e-07        0.8276  0.0013
##
## P value adjustment method: none
```

Le test pairwise.t.test effectue un test d'homogénéité des moyennes pour chaque paire de groupes
On peut remarquer qu'avec un seuil de tolérance de 5 %, il y a une grande similarité entre les groupes Oxomatl et Tenochtitlan, ce qui confirme notre conclusion basée sur l'observation du boxplot.

2 Régression et corrélation multiple

2.1 Les accidents sur les routes du Minnesota

Le fichier accidents2.csv contient des données sur le taux d'accidents de voiture sur les autoroutes du Minnesota (1973). Ces données, collectées sur 39 grands segments d'autoroute, ont évidemment été collectées pour essayer de déterminer les raisons de ces accidents.

On demande d'étudier l'éventuelle relation entre ce taux d'accidents et nombre de signaux routiers par mile associé à la largeur de la bande d'urgence latérale. Dans un second temps, on demande d'ajouter comme 3ème variable explicative le nombre d'entrées par mile d'autoroute.

```
setwd("C:\\Users\\amine\\OneDrive\\Bureau\\EcomStat\\Labo\\datasets")

accidents <- read.table("accidents2.csv", h=TRUE, sep=";", dec=",", row.names=1)
accidents
```

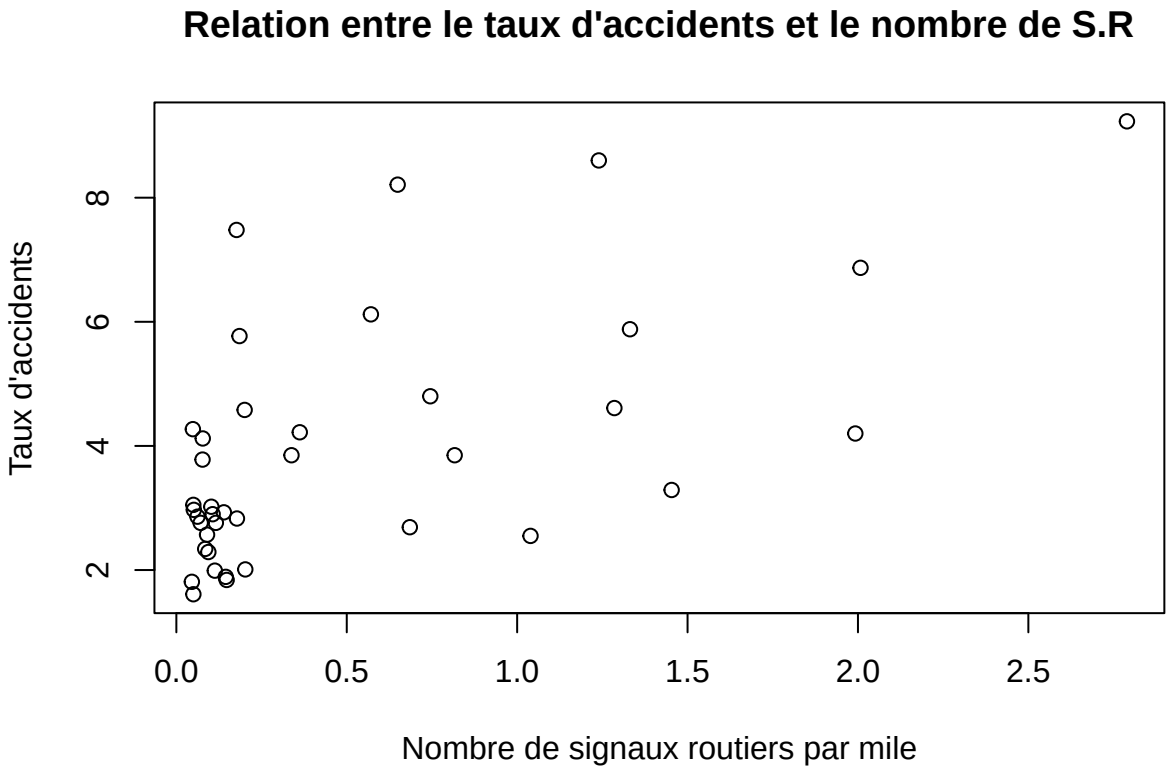
```
##    rate   len adt trks      sigs1 slim shld lane acpt  itg lwid htype
## 1  4.58  4.99  69    8 0.20040080   55   10    8  4.6 1.20   12   FAI
## 2  2.86 16.11  73    8 0.06207325   60   10    4  4.4 1.43   12   FAI
## 3  3.02  9.75  49   10 0.10256410   60   10    4  4.7 1.54   12   FAI
## 4  2.29 10.65  61   13 0.09389671   65   10    6  3.8 0.94   12   FAI
## 5  1.61 20.01  28   12 0.04997501   70   10    4  2.2 0.65   12   FAI
## 6  6.87  5.97  30    6 2.00750419   55   10    4 24.8 0.34   12    PA
## 7  3.85  8.57  46    8 0.81668611   55    8    4 11.0 0.47   12    PA
## 8  6.12  5.24  25    9 0.57083969   55   10    4 18.5 0.38   12    PA
## 9  3.29 15.79  43   12 1.45333122   50    4    4  7.5 0.95   12    PA
## 10 5.88  8.26  23    7 1.33106538   50    5    4  8.2 0.12   12    PA
## 11 4.20  7.03  23    6 1.99224751   60   10    4  5.4 0.29   12    PA
## 12 4.61 13.28  20    9 1.28530120   50    2    4 11.2 0.15   12    PA
## 13 4.80  5.40  18   14 0.74518518   50    8    2 15.2 0.00   12    PA
## 14 3.85  2.96  21    8 0.33783784   60   10    4  5.4 0.34   12    PA
## 15 2.69 11.75  27    7 0.68510638   55   10    4  7.9 0.26   12    PA
## 16 1.99  8.86  22    9 0.11286682   60   10    4  3.2 0.68   12    PA
## 17 2.01  9.78  19    9 0.20224949   60   10    4 11.0 0.20   12    PA
## 18 4.22  5.49   9   11 0.36214936   50    6    2  8.9 0.18   12    PA
## 19 2.76  8.63  12    8 0.11587485   55    6    2 12.4 0.14   13    PA
## 20 2.55 20.31  12    7 1.03923683   60   10    4  7.8 0.05   12    PA
## 21 1.89 40.09  15   13 0.14494388   55    8    4  9.6 0.05   12    PA
## 22 2.34 11.81   8    8 0.08467400   60   10    2  4.3 0.00   12    PA
## 23 2.83 11.39   5    9 0.17779631   50    8    2 11.1 0.00   12    PA
## 24 1.81 22.00   5   15 0.04545454   60    7    2  6.8 0.00   12    PA
## 25 9.23  3.58  23    6 2.78932961   40    2    4 53.0 0.56   12    MA
## 26 8.60  3.23  13    6 1.23959752   45    2    2 17.3 0.31   12    MA
## 27 8.21  7.73   7    8 0.64936611   55    8    2 27.3 0.13   12    MA
## 28 2.93 14.41  10   10 0.13939625   55    6    2 18.0 0.00   12    MA
## 29 7.48 11.54  12    7 0.17665511   45    3    2 30.2 0.09   12    MA
## 30 2.57 11.10   9    8 0.09009009   60    7    2 10.3 0.00   12    MA
## 31 5.77 22.09   4    8 0.18526935   45    3    2 18.2 0.00   11    MA
## 32 2.90  9.39   5   10 0.10649627   55    1    2 12.3 0.00   13    MA
## 33 2.97 19.49   4   13 0.05130836   55    4    2  7.1 0.00   12    MA
## 34 1.84 21.01   5   12 0.14759638   55    8    2 14.0 0.00   10    MA
## 35 3.78 27.16   2   10 0.07681885   55    3    2 11.3 0.04   12    MA
```

```
## 36 2.76 14.03   3    8 0.07127584   50    4    2 16.3 0.07   12    MA
## 37 4.27 20.63   1   11 0.04847310   55    4    2  9.6 0.00   11    MA
## 38 3.05 20.06   3   11 0.04985045   60    8    2  9.0 0.00   12    MC
## 39 4.12 12.91   1   10 0.07745933   55    3    2 10.4 0.00   12    MC
```

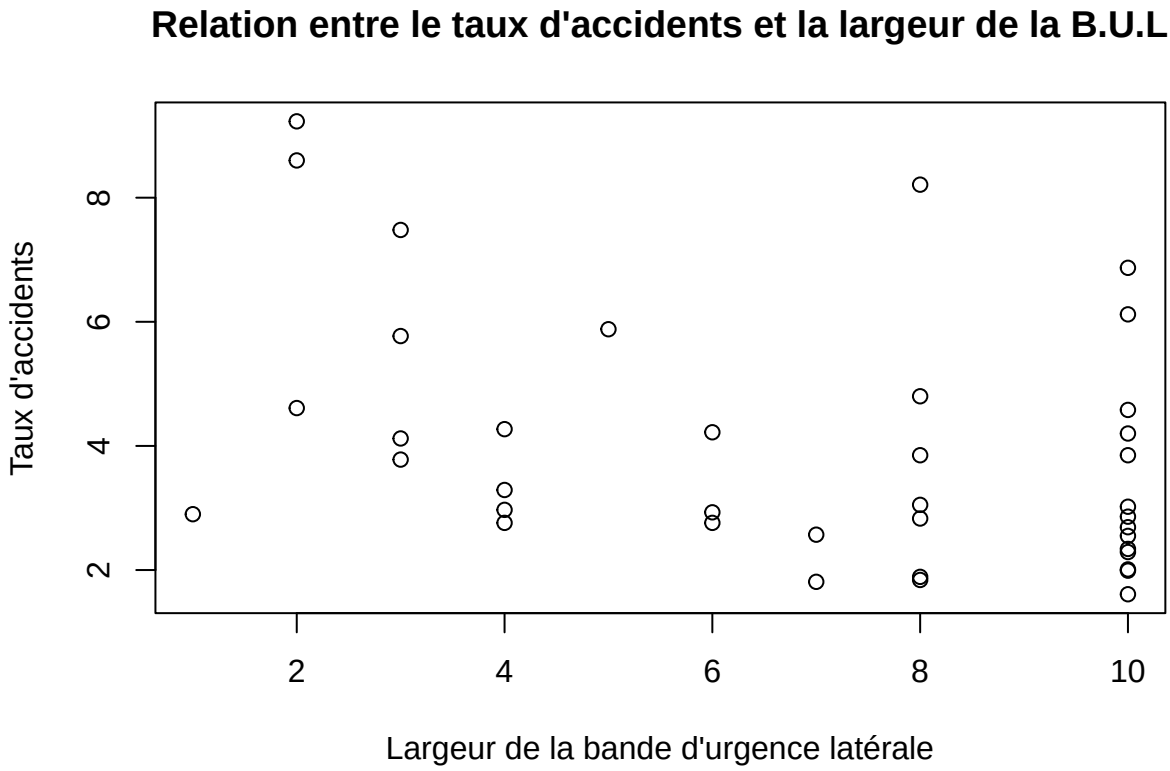
```
summary(accidents)
```

```
##       rate            len              adt             trks      
##  Min.   :1.610   Min.   : 2.960   Min.   : 1.00   Min.   : 6.000  
##  1st Qu.:2.630   1st Qu.: 7.995   1st Qu.: 5.00   1st Qu.: 8.000  
##  Median :3.050   Median :11.390   Median :13.00   Median : 9.000  
##  Mean   :3.933   Mean   :12.884   Mean   :19.62   Mean   : 9.333  
##  3rd Qu.:4.595   3rd Qu.:17.800   3rd Qu.:24.00   3rd Qu.:11.000  
##  Max.   :9.230   Max.   :40.090   Max.   :73.00   Max.   :15.000  
##      sigs1              slim          shld            lane      
##  Min.   :0.04545   Min.   :40   Min.   : 1.000   Min.   :2.000  
##  1st Qu.:0.08738   1st Qu.:50   1st Qu.: 4.000   1st Qu.:2.000  
##  Median :0.17666   Median :55   Median : 8.000   Median :2.000  
##  Mean   :0.51072   Mean   :55   Mean   : 6.872   Mean   :3.128  
##  3rd Qu.:0.71515   3rd Qu.:60   3rd Qu.:10.000   3rd Qu.:4.000  
##  Max.   :2.78933   Max.   :70   Max.   :10.000   Max.   :8.000  
##       acpt             itg              lwid          htype          
##  Min.   : 2.20   Min.   :0.0000   Min.   :10.00   Length:39          
##  1st Qu.: 6.95   1st Qu.:0.0000   1st Qu.:12.00   Class :character   
##  Median :10.30   Median :0.1300   Median :12.00   Mode  :character   
##  Mean   :12.16   Mean   :0.2964   Mean   :11.95                      
##  3rd Qu.:14.60   3rd Qu.:0.3600   3rd Qu.:12.00                      
##  Max.   :53.00   Max.   :1.5400   Max.   :13.00
```

```
# Tracer le nuage de points entre rate et sigs1
plot(accidents$sigs1, accidents$rate, xlab = "Nombre de signaux routiers par mile",
ylab = "Taux d'accidents",
main ="Relation entre le taux d'accidents et le nombre de S.R")
```




```
# Tracer le nuage de points entre rate et shld
plot(accidents$shld, accidents$rate, xlab = "Largeur de la bande d'urgence latérale",
ylab = "Taux d'accidents",
main = "Relation entre le taux d'accidents et la largeur de la B.U.L")
```




```
# Créer le modèle de régression multiple
modele1 <- lm(rate ~ sigs1 + shld, data = accidents)
modele1
```

```
##
## Call:
## lm(formula = rate ~ sigs1 + shld, data = accidents)
##
## Coefficients:
## (Intercept)        sigs1         shld
##      4.4974       1.6848      -0.2073
```

On obtient donc l'hyperplan d'ajustement suivant :
rate = 4.497 + 1.685*sigs1 - 0.207*shld -> y = 4.497 + 1.685*x1 – 0.207*x2

```
# Afficher un résumé du modèle
summary(modele1)
```

```
##
## Call:
## lm(formula = rate ~ sigs1 + shld, data = accidents)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8267 -0.8936 -0.2926  0.4840  4.2770
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.49738    0.65166   6.901 4.42e-08 ***
## sigs1        1.68475    0.36786   4.580 5.38e-05 ***
## shld        -0.20729    0.08053  -2.574   0.0143 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.496 on 36 degrees of freedom
## Multiple R-squared:  0.4627, Adjusted R-squared:  0.4329
## F-statistic:  15.5 on 2 and 36 DF,  p-value: 1.391e-05
```

```
sqrt(0.4329)
```

```
## [1] 0.6579514
```

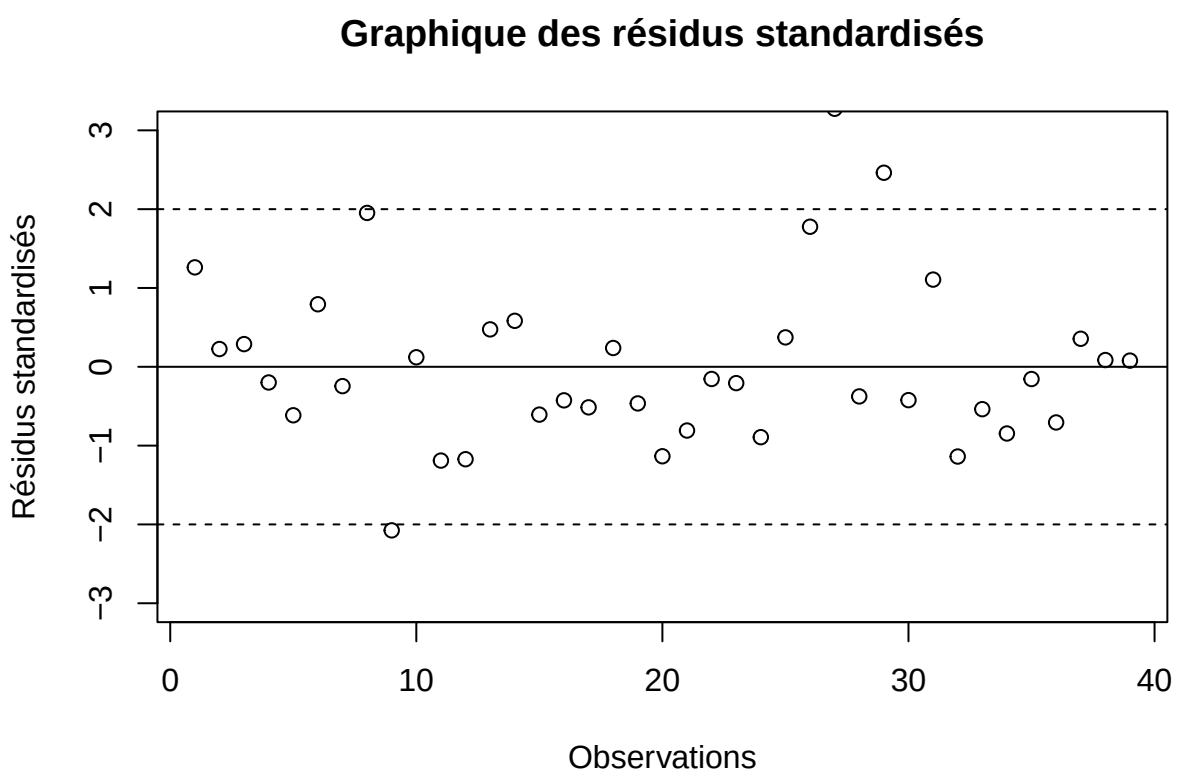
Avec un seuil de tolérance fixé à 0.05, on peut donc en conclure :
- Rejet de H0 pour Beta 0
- Rejet de H0 pour Beta 1
- Rejet de H0 pour Beta 2

On obtient un r-squared de 0.4627 mais pour une régression multiple, il est préférable d'utiliser le Adjusted R-squared car le R-Squared ne va cesser d'augmenter en ajoutant plus de régresseurs au modèle.
On obtient 0.657 qui est un taux correct donc on va considérer que les 2 variables ici présentes exercent une influence sur le taux d'accidents.
Enfin, nous comparons le modèle simple et complet pour vérifier à nouveau si nos régresseurs sont fiables ou non.
H0 : y = beta0 + epsilon
H0 : beta1 = 0 et beta2 = 0
Nous obtenons une p-value très faible (1.391e-05), ce qui suggère que nos régresseurs jouent un rôle significatif dans notre modèle.

```
# Calcul des résidus standardisés
residus.studentises <- rstudent(modele1)

# Tracer le graphique des résidus standardisés
plot(residus.studentises, ylim = c(-3, 3),
xlab = "Observations",
ylab = "Résidus standardisés",
main = "Graphique des résidus standardisés")

# Tracer les lignes horizontales à -2, 0 et 2
abline(h = c(-2, 0, 2), lty = c(2, 1, 2))
```

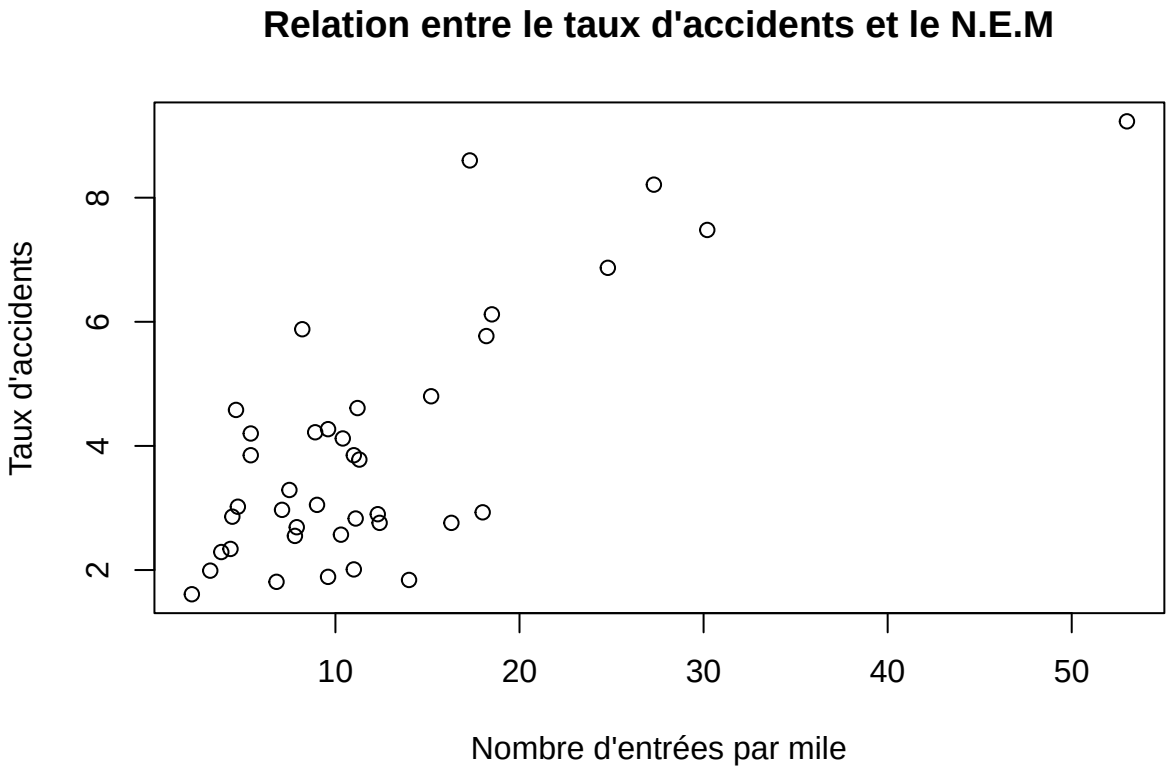


L'étude des résidus montre que la corrélation est faible. Les points ne sont pas suffisamment dispersés. Certaines valeur se trouvent en dehors de l'intervalle et il y a plus de points négatifs que de points positifs.

2.2 Ajout de la variable : nombre d'entrées par mile d'autoroute

Testons avec le nombre d'entrée par mile d'autoroute en plus :

```
# Tracer le nuage de points entre rate et acpt
plot(accidents$acpt, accidents$rate,
xlab = "Nombre d'entrées par mile",
ylab = "Taux d'accidents",
main = "Relation entre le taux d'accidents et le N.E.M")
```




```
# Créer le modèle de régression multiple
modele2 <- lm(rate ~ sigs1 + shld + acpt, data = accidents)
modele2
```

```
##
## Call:
## lm(formula = rate ~ sigs1 + shld + acpt, data = accidents)
##
## Coefficients:
## (Intercept)        sigs1         shld         acpt
##     2.58285      0.92569     -0.07702      0.11571
```

Voici l'hyperplan d'ajustement formé :
y = 2.583 + 0.926*x1 - 0.077*x2 + 0.116*x3
Analysons maintenant le summary :

```
# Afficher un résumé du modèle
summary(modele2)
```

```
##
## Call:
## lm(formula = rate ~ sigs1 + shld + acpt, data = accidents)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.91363 -0.98276 -0.08176  0.65134  3.02187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.58285    0.70958   3.640 0.000872 ***
## sigs1        0.92569    0.35540   2.605 0.013409 *
## shld        -0.07702    0.07375  -1.044 0.303523
## acpt         0.11571    0.02779   4.164 0.000194 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.24 on 35 degrees of freedom
## Multiple R-squared:  0.6407, Adjusted R-squared:  0.6099
## F-statistic: 20.81 on 3 and 35 DF,  p-value: 6.502e-08
```

```
sqrt(0.609)
```

```
## [1] 0.7803845
```

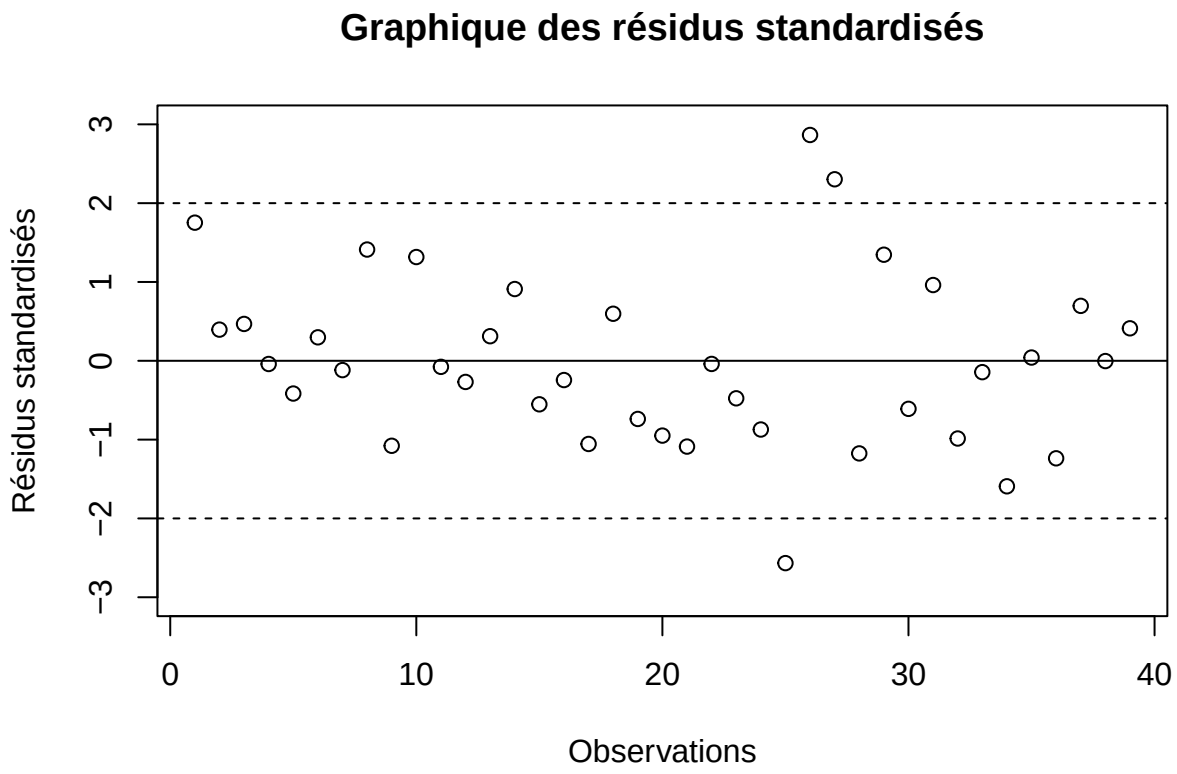
Avec un seuil de tolérance fixé à 0.05, on peut donc en conclure :
Rejet de H0 pour Beta 0
Rejet de H0 pour Beta 1
Acceptation de H0 pour Beta 2 (Beta 2 = 0). Donc perte de l'influence de la largeur de la bande d'urgence latérale.
Rejet de H0 pour Beta 3
On obtient un r-squared de 0.6407 mais pour une régression multiple, il est préférable d'utiliser le Adjusted R-squared car le R-Squared ne va cesser d'augmenter en ajoutant plus de régresseurs au modèle.
On obtient 0.780 qui est un taux élevé et vu la p-value 6.502e-08 qui est très faible, on peut rejeter H0 et on peut dire qu'il y a une haute corrélation mais elle peut mener à des erreurs vue que la p-value « shld » est trop élevée et donc pas fiable.

Enfin, nous comparons le modèle simple et complet pour vérifier à nouveau si nos régresseurs sont fiables ou non.
H0 : y = beta0 + epsilon
H0 : beta1 = 0, beta2 = 0 et beta3 = 0
Nous obtenons une p-value très faible (6.502e-08), ce qui suggère que nos régresseurs jouent un rôle significatif dans notre modèle.

```
# Calcul des résidus standardisés
residus.studentises <- rstudent(modele2)

# Tracer le graphique des résidus standardisés
plot(residus.studentises,
ylim = c(-3, 3),
xlab = "Observations",
ylab = "Résidus standardisés",
main = "Graphique des résidus standardisés")

# Tracer les lignes horizontales à -2, 0 et 2
abline(h = c(-2, 0, 2), lty = c(2, 1, 2))
```



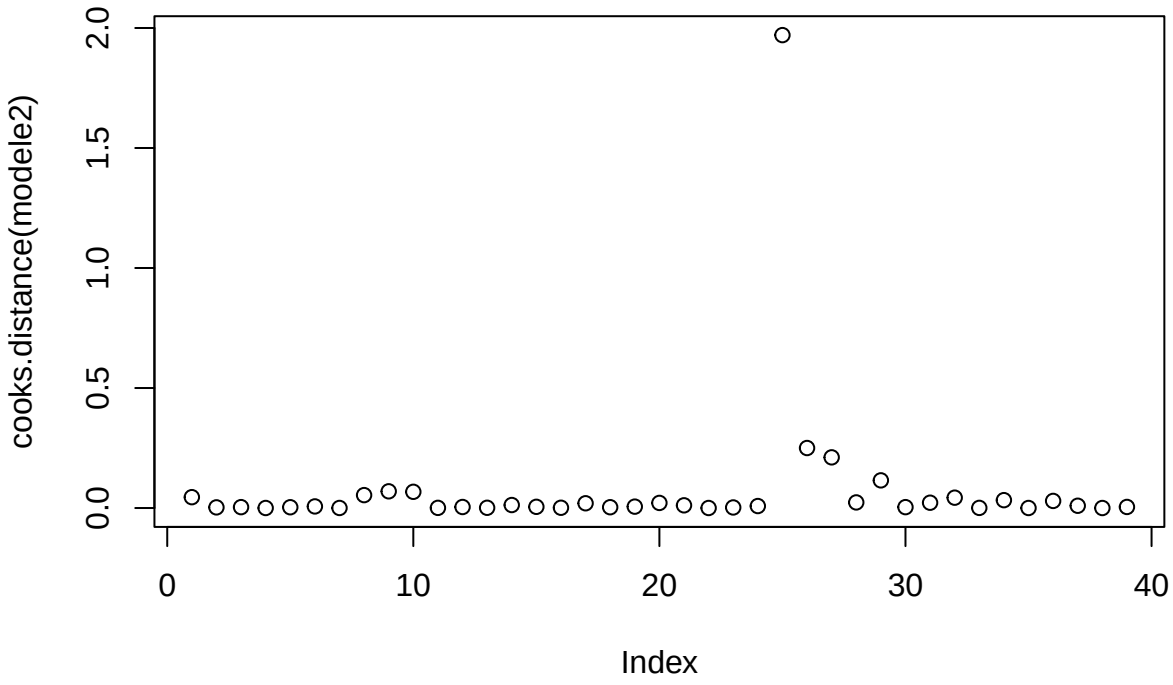
L'etude des résidus nous montre que la corrélation n'est pas hyper bonne, les points ne sont pas assez dispersés et certaines valeur se trouvent en dehors de l'intervalle, cela nous confirme qu'il existe un meilleur modéle.

2.3 Complément : La distance de Cook

```
cooks.distance(modele2)
```

```
##            1            2            3            4            5            6
## 4.517589e-02 2.665735e-03 3.612047e-03 2.808641e-05 3.223111e-03 6.606566e-03
##            7            8            9           10           11           12
## 1.410971e-04 5.373214e-02 6.955170e-02 6.751307e-02 6.102989e-04 4.198708e-03
##           13           14           15           16           17           18
## 9.698514e-04 1.241240e-02 4.913588e-03 1.041009e-03 1.968698e-02 3.287452e-03
##           19           20           21           22           23           24
## 5.767956e-03 2.114908e-02 1.122239e-02 2.699416e-05 2.268157e-03 8.165894e-03
##           25           26           27           28           29           30
## 1.970309e+00 2.500101e-01 2.110625e-01 2.333586e-02 1.152286e-01 3.611473e-03
##           31           32           33           34           35           36
## 2.205497e-02 4.312689e-02 4.684723e-04 3.271217e-02 4.266306e-05 2.950739e-02
##           37           38           39
## 9.368889e-03 1.333928e-07 4.371100e-03
```

```
plot(cooks.distance(modele2))
```



La distance de Cook pour chacun des points du nuage est la distance entre les paramètres estimés par la régression avec et sans ce point.

Avec notre modèle, on observe qu'à l'indice 25, on a une valeur supérieure à 1, ce qui pourrait biaiser notre estimation des coefficients de régression (point aberrant).

2.4 Complément : Le critère AIC

```
step(modele2)
```

```
## Start:  AIC=20.58
## rate ~ sigs1 + shld + acpt
##
##         Df Sum of Sq    RSS    AIC
## - shld   1    1.6778 55.529 19.781
## <none>               53.851 20.584
## - sigs1  1   10.4384 64.290 25.494
## - acpt   1   26.6768 80.528 34.277
##
## Step:  AIC=19.78
## rate ~ sigs1 + acpt
##
```

```
##         Df Sum of Sq    RSS    AIC
## <none>               55.529 19.781
## - sigs1  1     9.590 65.119 23.994
## - acpt   1    39.822 95.352 38.866
```

```
##
## Call:
## lm(formula = rate ~ sigs1 + acpt, data = accidents)
##
## Coefficients:
## (Intercept)        sigs1         acpt
##      1.9269       0.8807       0.1280
```

est un critère de comparaison de modèles:
pour un sous-modèle donné, il propose une estimation de la perte d'information lorsqu'on utilise ce modèle pour (prédire) les données.

Avec notre modèle, on observe que si on enlève la variable "shld", on aura le moins d'information perdue (AIC = 19.781). Donc, R nous propose un modèle simplifié qui peut décrire les données avec le plus petit nombre de paramètres possible.