MTH2302D : Devoir

Amira Tamakloe

```r
source('charger.R')
mondata<-charger(2131198)
```

Phase 1 : Analyse statistique descriptive et inférence

```r
# Matrice de corrélation des 3 variables (mpg, displacement et weight)
mondataSubset  <- mondata[ , c("mpg", "displacement", "weight")]
cor(mondataSubset)
```

```
##                     mpg displacement     weight
## mpg           1.0000000   -0.7901391 -0.8264434
## displacement -0.7901391    1.0000000  0.9329057
## weight       -0.8264434    0.9329057  1.0000000
```

Un test de corrélation comme effectué avec la formule cor() détermine la dépendance entre plusieurs variables. Dans notre cas, nous nous intéressons à la dépendance entre trois variables soient mpg, displacement et weight. Dans notre cas, le test de corrélation effectué pour les 3 variables est celui de pearson puisqu'il est le test effectué par défaut. Une valeur de 1 dans la matrice signifie que la corrélation entre deux variables est parfaite et plus celle-ci se rapproche de 1 plus elle est dépandante. Ainsi, la valeur du test détermine la force de leur dépendance. Le signe par contre obtenu, c'est-à-dire qu'il soit positif ou négatif détermine le sens de variation des deux variables auxquels. Si la valeur est négative, les deux variables évoluent dans deux sens différent, c'est-à-dire que lorsqu'une d'entre elle diminue l'autre augment. Si le signe est positif, les deux augmentent ou diminuent ensemble. On en conclut ainsi que la variable mpg à une corrélation quand même forte avec les deux autres variables mais elles évoluent dans des sens différents. La variable de displacement évolue dans le même sens que le poid et leur corrélation est très forte.

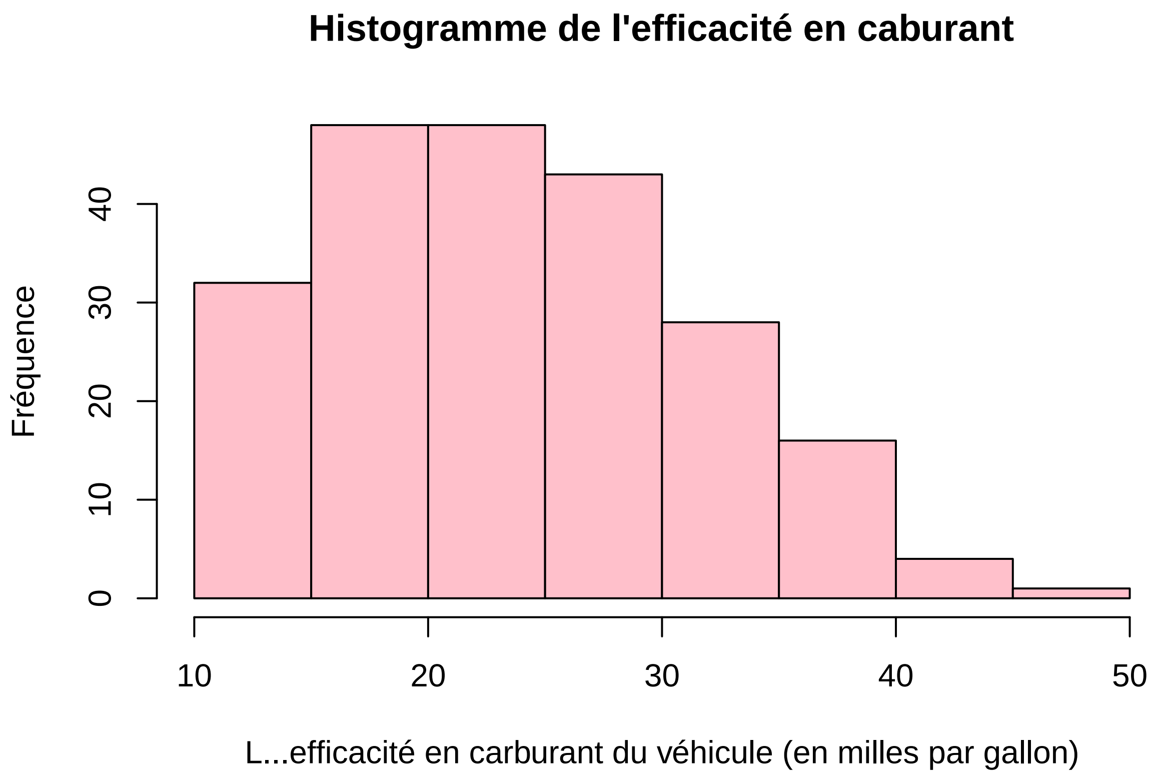
Phase 1.b Graphique et tableaux associés à la variable d'efficacité du carburant

```r
#Histogramme
hist(mondata$mpg, col="pink",border="black", main=paste("Histogramme de l'efficacité en caburant"),
     xlab="L’efficacité en carburant du véhicule (en milles par gallon)",ylab="Fréquence")
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'L’efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <e2>
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'L’efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <80>
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'L’efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <99>
```



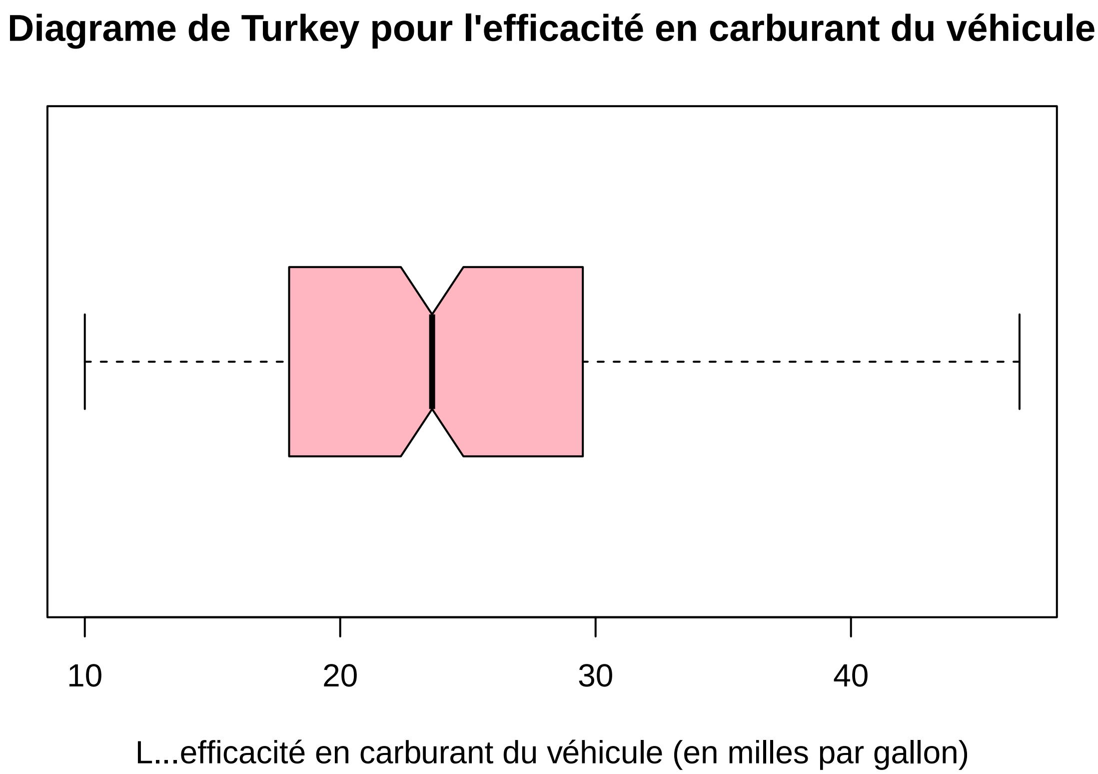
L'histogramme montre que l'étalement des données de l'efficacité en carburant du véhicule est étale de 10 milles/galon à 50 milles/gallon. On s'aperçoit aussi la fréquence des données est plus grande lorsque l'on se situe vers la gauche c'est à dire, une efficacité en carburant du véhicule en milles/gallon plus faible. Ainsi, la majorité des données se situent à l'extrémité gauche. On peut estimer que la moyenne sera un peu plus que 20 milles/gallon.

```
#Diagramme de Tukey
boxplot(mondata$mpg,col="lightpink", horizontal=TRUE,notch=TRUE, main=paste("Diagrame de Turkey pour l'e
```

```
## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'L'efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <e2>
```

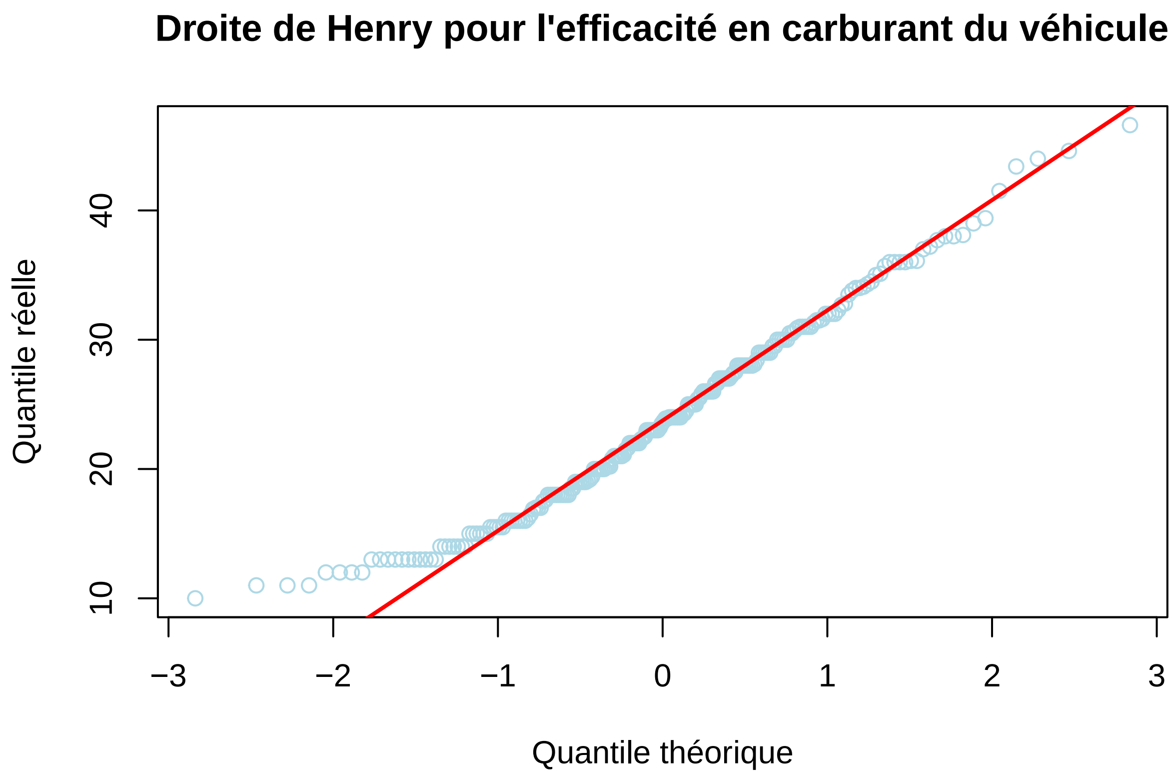
```
## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'L'efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <80>
```

```
## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'L'efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <99>
```



Le diagramme de Tukey montre que la valeur minimale de l'efficacité en carburant du véhicule en milles/gallon est de 10 et que sa valeur maximale est un peu en dessous de 50 milles/gallon. La médiane de l'efficacité avoisine 23 milles/gallon. Tandis que celle du premier quartile est en dessous de 20, à peu près 18 milles/gallon et celle du troisième quartile est très près de 30 milles/gallon. La distribution est à peu près symétrique, car la portion droite de la boîte et la moustache droite sont à peu près égale à celle de gauche.

```
#Droite de Henry
qqnorm(mondata$mpg, col="lightblue", main=paste("Droite de Henry pour l'efficacité en carburant du véhi
qqline(mondata$mpg, col="red", lwd=2)
```



La droite de Henry nous permet d'évaluer la normalité de la distribution. On s'aperçoit que les données sont majoritairement sur la line rouge, ce qui signifie que la majorité des valeurs suivent une loi normal. Cependant, on peut aperçevoir quelques valeurs à l'extrémité gauche qui ne touchent pas la droite. Pour confirmer mon hypothèse que les données suivent une loi normal malgré les quelques données aberrantes à l'extrémité gauche, il est possible de faire d'autres tests comme celui de shapiro pour en avoir la certitude.

```r
#Test de Shapiro
shapiro.test(mondata$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mondata$mpg
## W = 0.97491, p-value = 0.0005895
```

Le test de Shapiro nous renseigne sur la normalité de la variable en question c'est à dire l'efficacité en carburant du véhicule. La valeur de p obtenue est 0.0005895 est inférieure à 0.05 ce qui signifie que nous ne pouvons pas supposer la normalité. Ainsi, on contredit l'hypothèse posé précèdement. L'hypothèse sur la normalité est rejetté, ainsi la variable ne suit pas une loi normale.

```r
#Statistiques descriptives
Quartile <- quantile(mondata$mpg , c(0.25,0.5,0.75))
Moyenne <- mean(mondata$mpg)
ÉcartType <- sd(mondata$mpg)
ErreurType <- ÉcartType/ sqrt(length(mondata$mpg))
quantile(mondata$mpg,c(.05,.95))
```

```
##     5%    95%
## 13.000 37.225
```

```r
InterValleDeConfiance <- t.test(mondata$mpg, conf.level = 0.95)
InterValleDeConfiance.min = c(13.000)
```

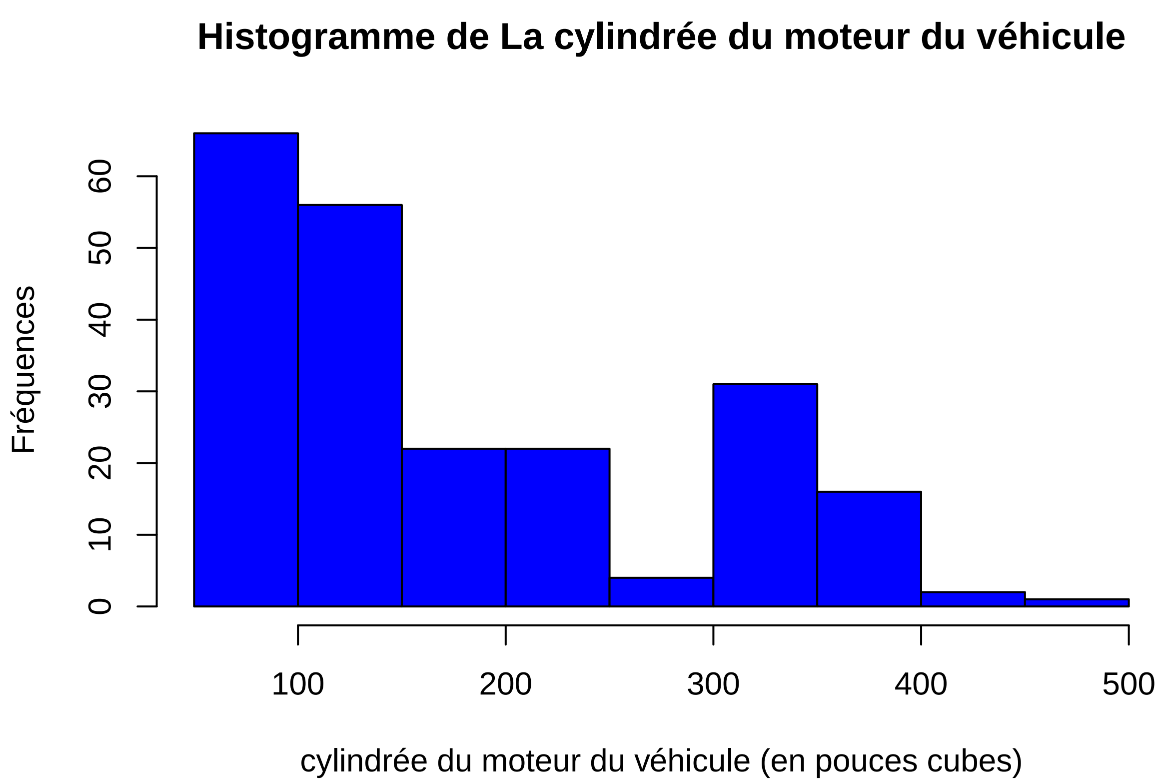
```r
InterValleDeConfiance.max = c(37.225)
table <- data.frame( Moyenne=Moyenne,
      "Écart Type"= ÉcartType ,
      "Erreur Type"= ErreurType,
      "min IDC" = InterValleDeConfiance.min,
      "max IDC"= InterValleDeConfiance.max,
      "p25"= Quartile[1],
      "Mediane"= Quartile[2],
      "p75"= Quartile[3],
      row.names = "mpg")
table
```

```
##      Moyenne Écart.Type Erreur.Type min.IDC max.IDC p25 Mediane  p75
## mpg 23.97455   7.802657    0.526055      13  37.225  18    23.6 29.5
```

La moyenne est un indicateur de position d'un échantillon car c'est la valeur unique que tous devrait avoir pour que leur total soit inchangé. Dans notre cas elle est de 23.97455. L'écart type quant à lui représente la dispersion des données autour de la moyenne et dans notre cas celui-ci est de 7.8 milles/gallon. L'écart type est élevés donc les données sont très dispersés. L'erreur type représente l'écart approximatif d'un échantillon et puisque 0.526055 est faible alors l'estimation de la moyenne est précise. L'intervalle de confiance à un niveau de confiance de 95% ce qui veut dire que la moyenne de l'efficacité en carburant se situe à 95% entre les deux valeurs de min.IDC et max.IDC. Ainsi, cette valeur se situe entre 13 et 37.225. Les quartiles sont respectivement p25 qui représentent 25% de l'échantillon est à 18 milles/gallon, la médiane qui représente le point milieu d'un échantillon est de 23.6, le p75 représentent 75% de la population et est fixé à 29.5 mille/gallon.

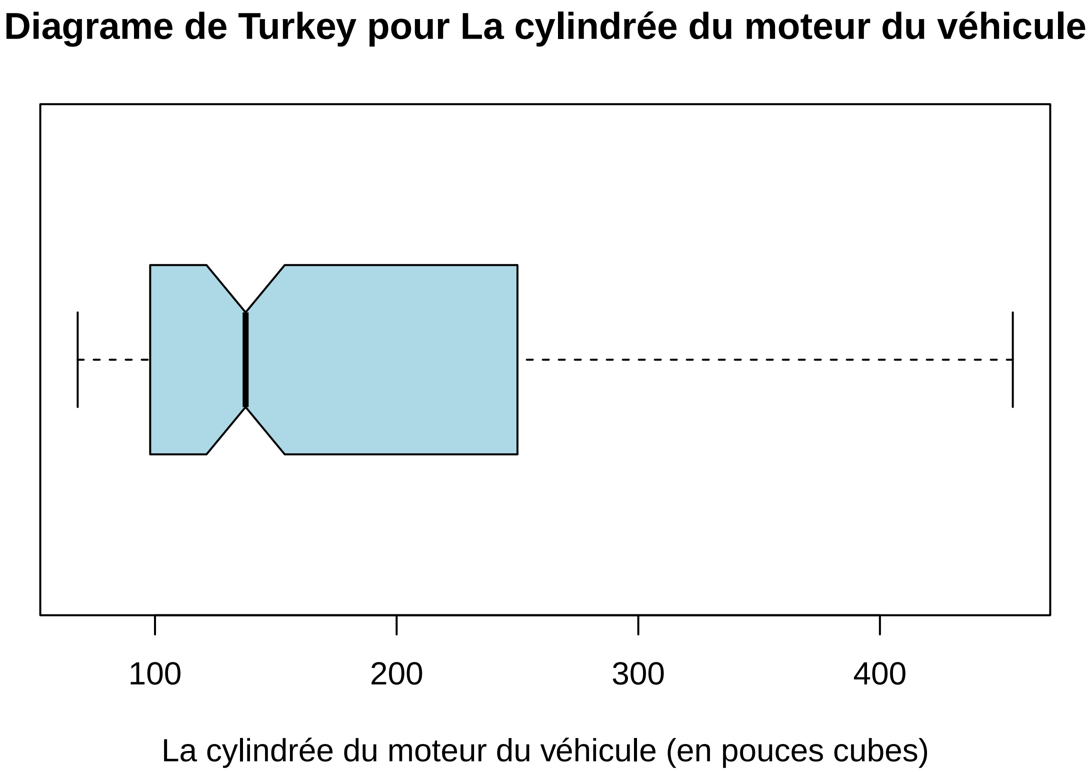
Graphique et tableaux associés à la cylindrée du moteur du véhicule :

```r
#histogramme
hist(mondata$displacement, col="blue",border="black", main=paste("Histogramme de La cylindrée du moteur
    xlab="cylindrée du moteur du véhicule (en pouces cubes)",ylab="Fréquences")
```



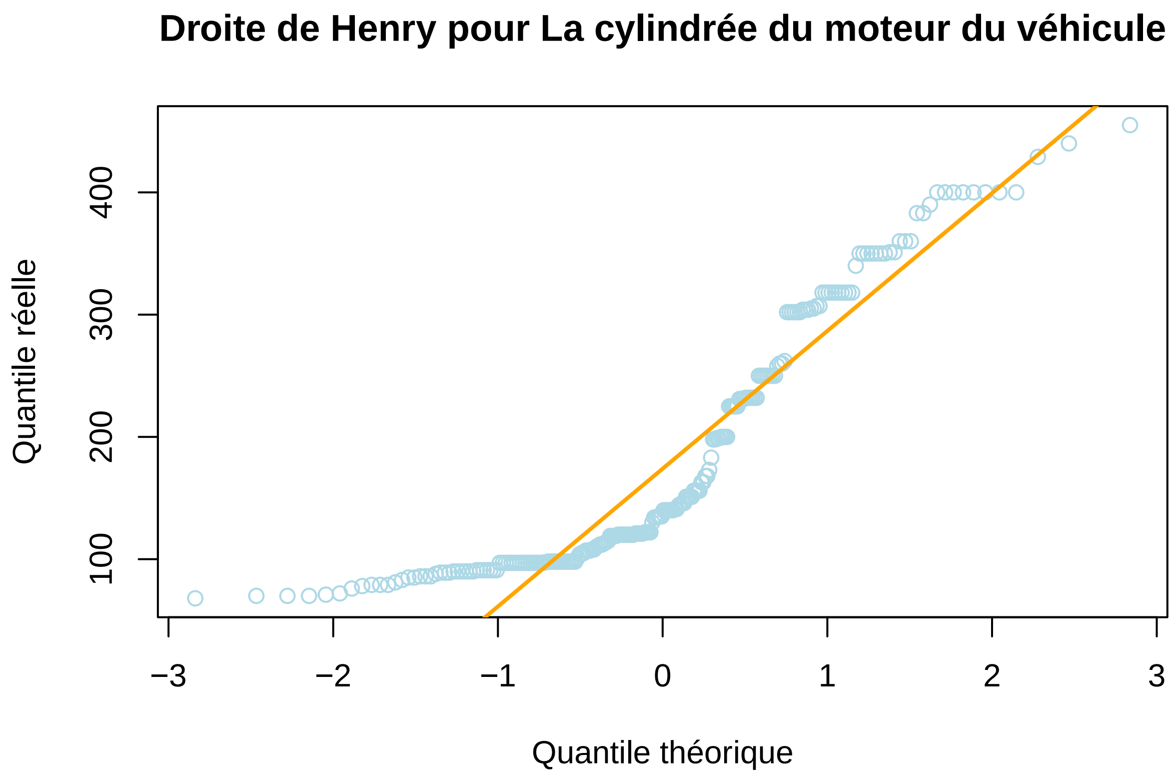
L'histogramme montre que l'étalement des données de cylindrée du moteur du véhicule est étalé de 100 pouces cubes à 500 pouces cubes. On s'aperçoit aussi la fréquence des données est plus grande lorsque l'on se situe vers la gauche. Ainsi, la majorité des données se situent à l'extrémité gauche.

```
#diagramme de tukey
boxplot(mondata$displacement,col="lightblue", horizontal=TRUE, notch = TRUE, main=paste("Diagrame de Tu
```



Le diagramme de Tukey montre que le minimum de la cylindrée du moteur du véhicule en pouces cube est inférieure à 100 pouces cubes et que sa valeur maximale est à peu près 450 pouces cubes. La médiane de la cylindrée du moteur avoisine 140 pouces/cubes. Tandis que celle du premier quartile est à peu près à 100 pouces cubes et celle du troisième quartile est à peu près 250 pouces cubes. La distribution est positivement asymétrique, car la portion droite de la boîte et la moustache droite sont plus longues qu'à gauche de la médiane.

```
#Droite de henry

qqnorm(mondata$displacement, col="lightblue", main=paste("Droite de Henry pour La cylindrée du moteur du
qqline(mondata$displacement, col="orange", lwd=2)
```



La droite de Henry nous permet d'évaluer la normalité de la distribution. On s'aperçoit que la majorité des données ne touchent pas la ligne orange, ce qui signifie que la majorité des valeurs ne suivent pas une loi normal. Pour confirmer mon hypothèse, il est possible de faire d'autres tests comme celui de shapiro pour en avoir la certitude.

```
#Test de Shapiro
shapiro.test(mondata$displacement)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mondata$displacement
## W = 0.85343, p-value = 1.2e-13
```

Le test de Shapiro nous renseigne sur la normalité de la variable en question c'est à dire l'efficacité en carburant du véhicule. La valeur de p obtenue est 1.2e-13, elle est extrêmement inférieure à 0.05 ce qui signifie que nous ne pouvons pas supposer la normalité.

```
Quartile <- quantile(mondata$displacement)
Moyenne <- mean(mondata$displacement)
ÉcartType <- sd(mondata$displacement)
ErreurType <- ÉcartType/ sqrt(length(mondata$displacement))
quantile(mondata$displacement)
```

```
##    0%   25%   50%   75%  100%
##  68.0  98.0 137.5 250.0 455.0
```

```
InterValleDeConfiance <- t.test(mondata$displacement, conf.level = 0.95)
InterValleDeConfiance.min = c(80.9)
InterValleDeConfiance.max = c(390.5)
table <- data.frame( Moyenne=Moyenne,
      "Écart Type"= ÉcartType ,
```

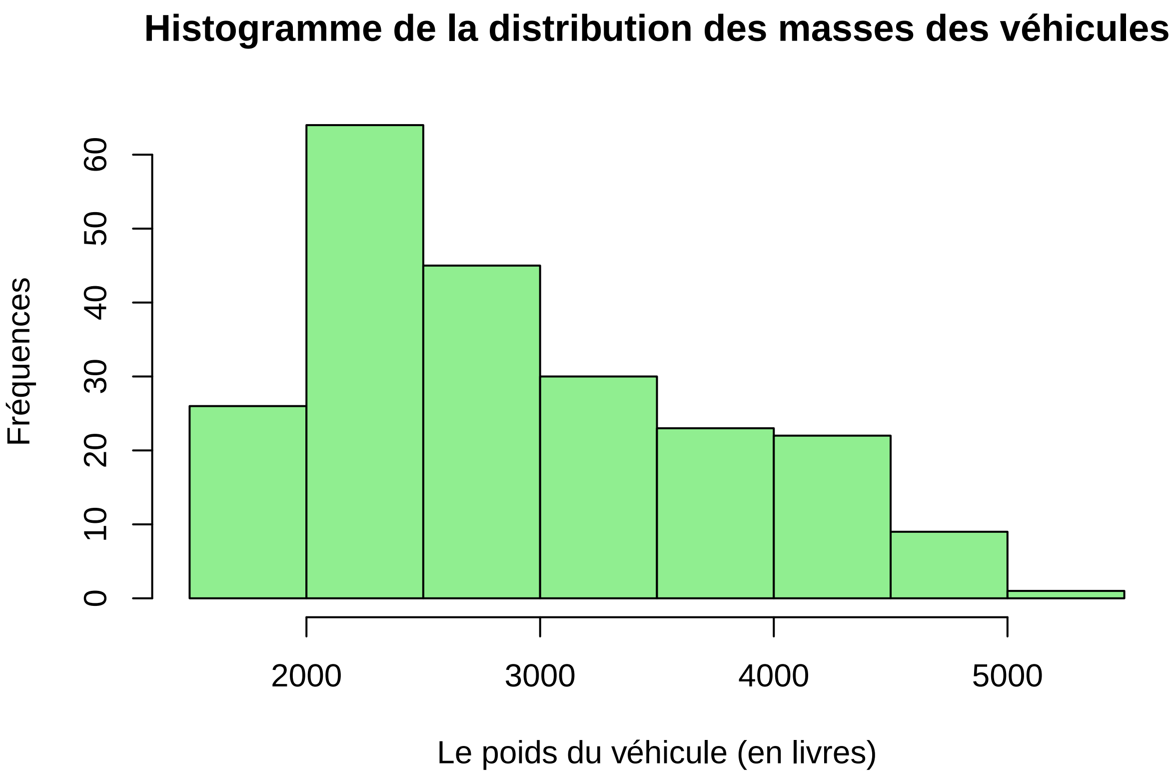
```
      "Erreur Type"= ErreurType,
      "min IDC" = InterValleDeConfiance.min,
      "max IDC"= InterValleDeConfiance.max,
      "min"=Quartile[1],
      "p25"= Quartile[2],
      "Mediane"= Quartile[3],
      "p75"= Quartile[4],
      "max"= Quartile[5],
      row.names = "displacement")
table
```

```
##              Moyenne Écart.Type Erreur.Type min.IDC max.IDC min p25 Mediane p75
## displacement 183.275   103.0308    6.946336    80.9   390.5  68  98   137.5 250
##              max
## displacement 455
```

La moyenne est un indicateur de position d'un échantillon car c'est la valeur unique que tous devrait avoir pour que leur total soit inchangé. Dans notre cas elle est de 183.275 L'écart type quant à lui représente la dispersion des données autour de la moyenne et dans notre cas celui-ci est de 103.0308. L'écart type est élevés donc les données sont très dispersés. L'erreur type représente l'écart approximatif d'un échantillon et puisque 6.946336 est élevé alors l'estimation de la moyenne n'est pas précise. L'intervalle de confiance à un niveau de confiance de 95% ce qui veut dire que la moyenne de l'efficacité en carburant se situe à 95% entre les deux valeurs de min.IDC et max.IDC. Ainsi, cette valeur se situe entre 80.9 et 390.5. Les quartiles sont respectivement p25 qui représentent 25% de l'échantillon est à 98 pouces cubes, la médiane qui représente le point milieu d'un échantillon est de 137.5, le p75 représentent 75% de la population et est fixé à 250 pouces cubes.

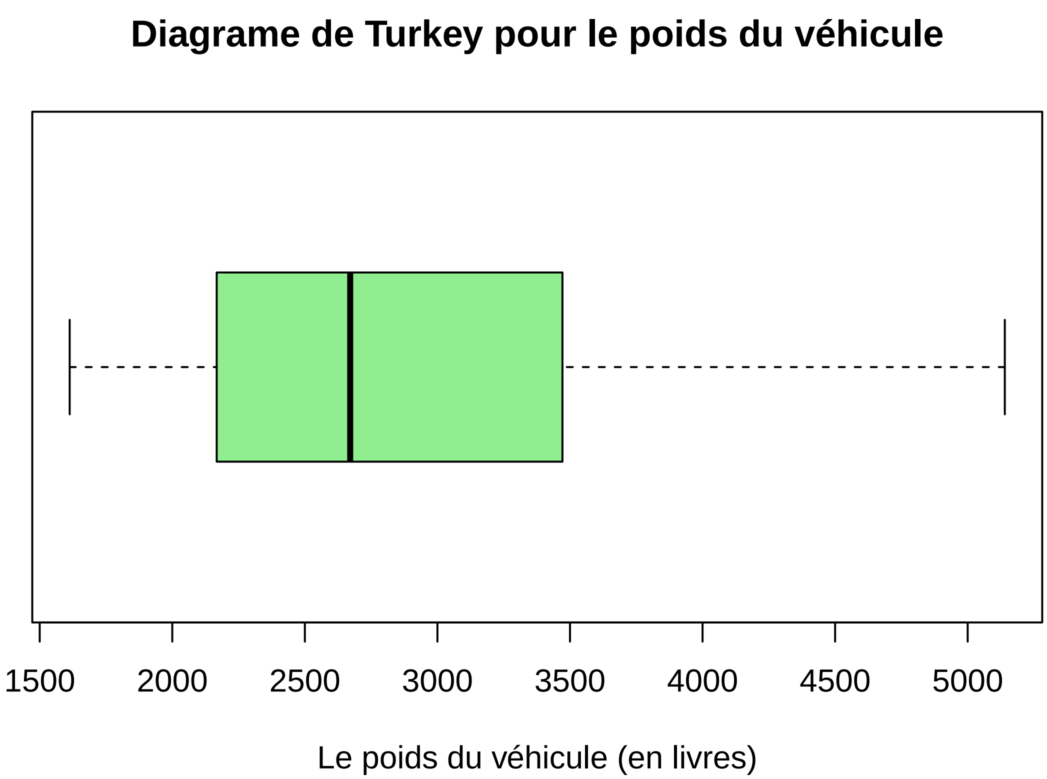
Graphique et tableaux associés associé au poids du véhicule :

```
#histogramme
hist(mondata$weight, col="lightgreen",border="black", main=paste("Histogramme de la distribution des ma
     xlab="Le poids du véhicule (en livres)",ylab="Fréquences")
```



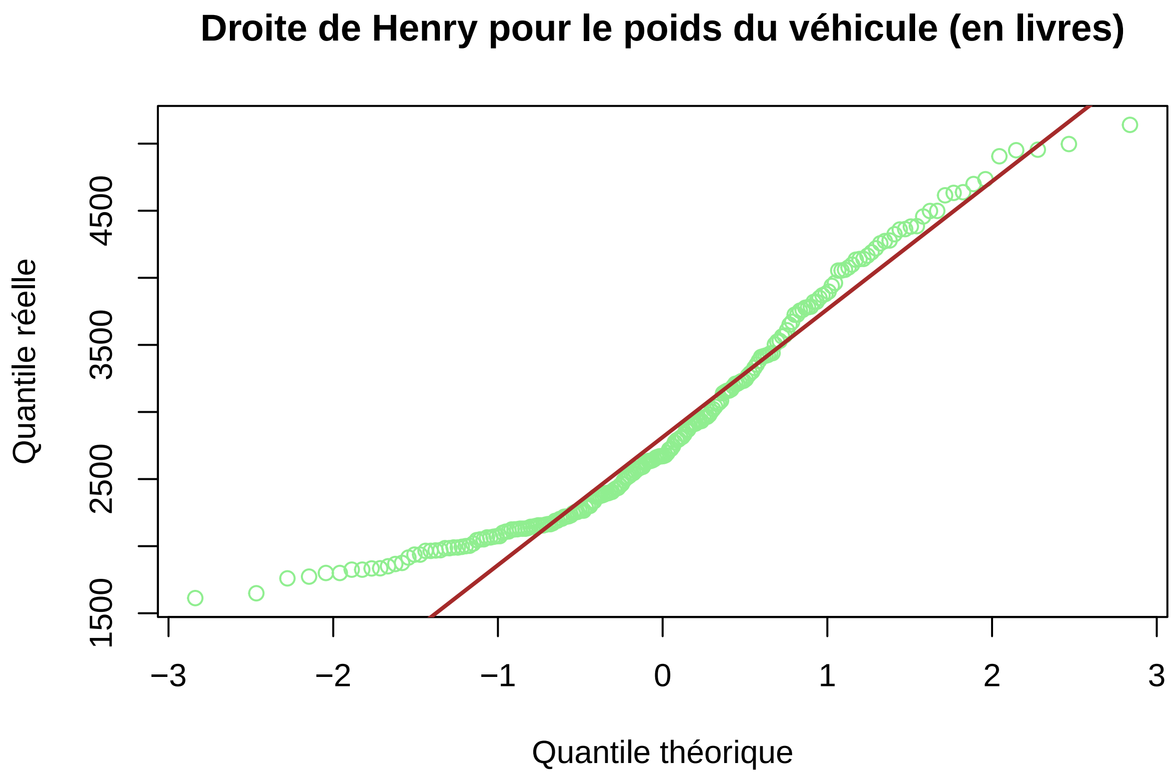
L'histogramme montre que l'étalement des données Le poids du véhicule (en livres) est étalé de 2000 livres à 5000 livres. On peut estimer que la moyenne sera très près de 3000.

```
#Diagramme de Tukey
boxplot(mondata$weight,col="lightgreen", horizontal=TRUE, main=paste("Diagrame de Turkey pour le poids d
```



Le diagramme de Tukey montre que le poid minimum dvéhicule en livre est un peu plus de 1500 livres et que sa valeur maximale est à peu près 5000 livres. La médiane poid du véhicule avoisine 2650 livres. Tandis que celle du premier quartile est à peu près 2200 livres et celle du troisième quartile est un peu inférieure à 3500 livres. La distribution est positivement asymétrique, car la portion droite de la boîte et la moustache droite sont plus longues qu'à gauche de la médiane.

```
#Droite de Henry
qqnorm(mondata$weight, col="lightgreen", main=paste("Droite de Henry pour le poids du véhicule (en livre
qqline(mondata$weight, col="brown", lwd=2)
```



La droite de Henry nous permet d'évaluer la normalité de la distribution. On s'aperçoit que beaucoup de données ne touchent pas la ligne rouge, ce qui signifie que la distribution pour le poid ne suit pas une loi normal. Pour confirmer mon hypothèse, il est possible de faire d'autres tests comme celui de shapiro pour en avoir la certitude.

```
#test de shapiro
shapiro.test(mondata$weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mondata$weight
## W = 0.92963, p-value = 9.109e-09
```

Le test de Shapiro nous renseigne sur la normalité de la variable en question c'est à dire l'efficacité en carburant du véhicule. La valeur de p obtenue est 9.109e-09 est inférieure à 0.05 ce qui signifie que nous ne pouvons pas supposer la normalité. Ainsi, on confirme l'hypothèse posé précèdement.

```
#tableau statistiques
Quartile <- quantile(mondata$weight)
Moyenne <- mean(mondata$weight)
ÉcartType <- sd(mondata$weight)
ErreurType <- ÉcartType/ sqrt(length(mondata$weight))
```

```
quantile(mondata$weight,c(.05,.95))
```

```
##      5%     95%
## 1866.15 4498.05
```

```
InterValleDeConfiance <- t.test(mondata$weight, conf.level = 0.95)
InterValleDeConfiance.min = c(1866.15)
InterValleDeConfiance.max = c(4498.05)
table <- data.frame( Moyenne=Moyenne,
      "Écart Type"= ÉcartType ,
      "Erreur Type"= ErreurType,
      "min IDC" = InterValleDeConfiance.min,
      "max IDC"= InterValleDeConfiance.max,
      "min"=Quartile[1],
      "p25"= Quartile[2],
      "Mediane"= Quartile[3],
      "p75"= Quartile[4],
      "max"= Quartile[5],
      row.names = "weight")
table
```

```
##         Moyenne Écart.Type Erreur.Type min.IDC max.IDC  min     p25 Mediane
## weight 2904.159   849.6087     57.2806 1866.15 4498.05 1613 2169.25    2671
##            p75  max
## weight 3455.25 5140
```

La moyenne est un indicateur de position d'un échantillon car c'est la valeur unique que tous devrait avoir pour que leur total soit inchangé. Dans notre cas elle est de 2904.159 livres. L'écart type quant à lui représente la dispersion des données autour de la moyenne et dans notre cas celui-ci est de 849.6087. L'écart type est élevés donc les données sont très dispersés. L'erreur type représente l'écart approximatif d'un échantillon et puisque 57.2806 est élevé alors l'estimation de la moyenne n'est pas précise. L'intervalle de confiance à un niveau de confiance de 95% ce qui veut dire que la moyenne de poid d'un véhicule se situe à 95% entre les deux valeurs de min.IDC et max.IDC. Ainsi, cette valeur se situe entre 1866.15 et 4498.05 Les quartiles sont respectivement p25 qui représentent 25% de l'échantillon est à 2169.25 livres, la médiane qui représente le point milieu d'un échantillon est de 2671 livres, le p75 représentent 75% de la population et est fixé à 3455.25 livres.

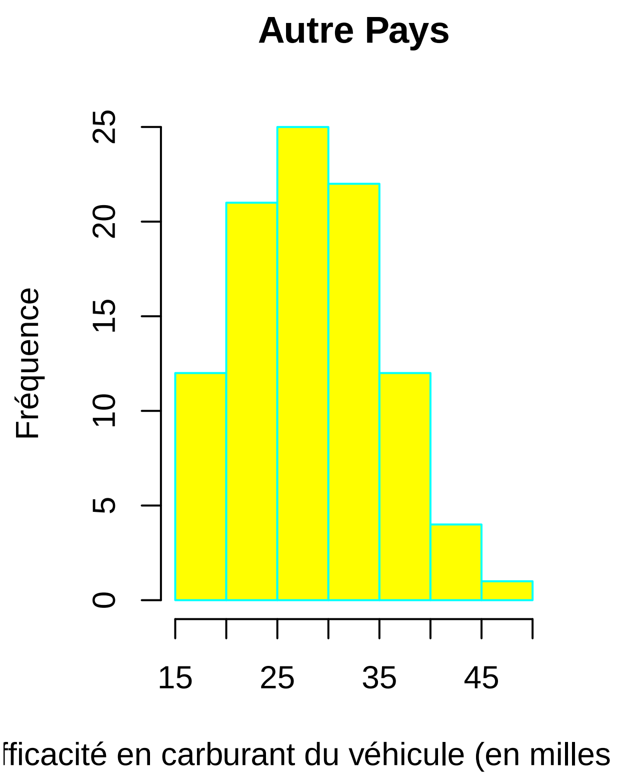
Analyses pour vérifier si l'efficacité en carburant d'un véhicule dépend de l'origine de celui-ci et donnez une brève conclusion :

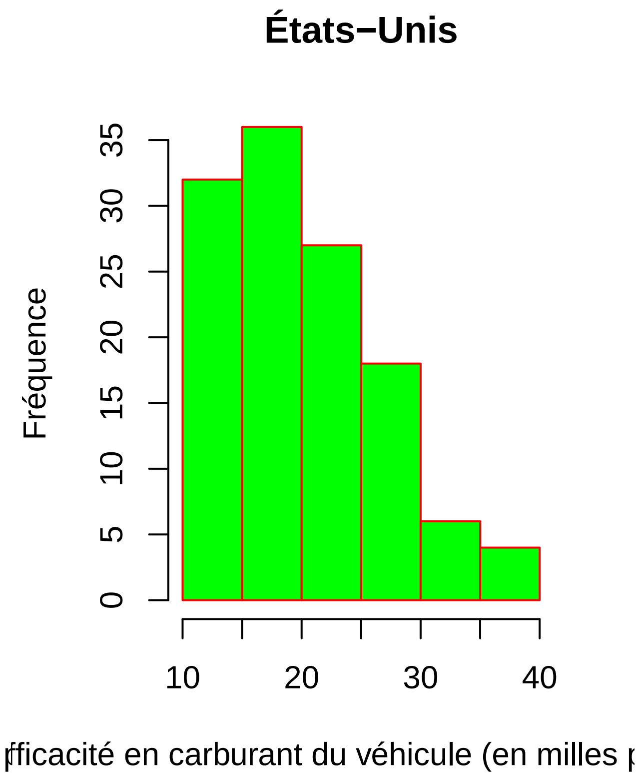
```
#deux histogrammes juxtaposés
layout(matrix(1:2,1,2))
hist(mondata$mpg[mondata$origin=="0"], col="yellow",border="cyan",
     main=paste("Autre Pays"),xlab="L'efficacité en carburant du véhicule (en milles par gallon)",ylab=
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'L'efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'L'efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'L'efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <99>
```

```
hist(mondata$mpg[mondata$origin=="1"], col="green",border="red",
     main=paste("États-Unis"),xlab="L'efficacité en carburant du véhicule (en milles par gallon)",ylab=
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'L'efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'L'efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'L'efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <99>
```



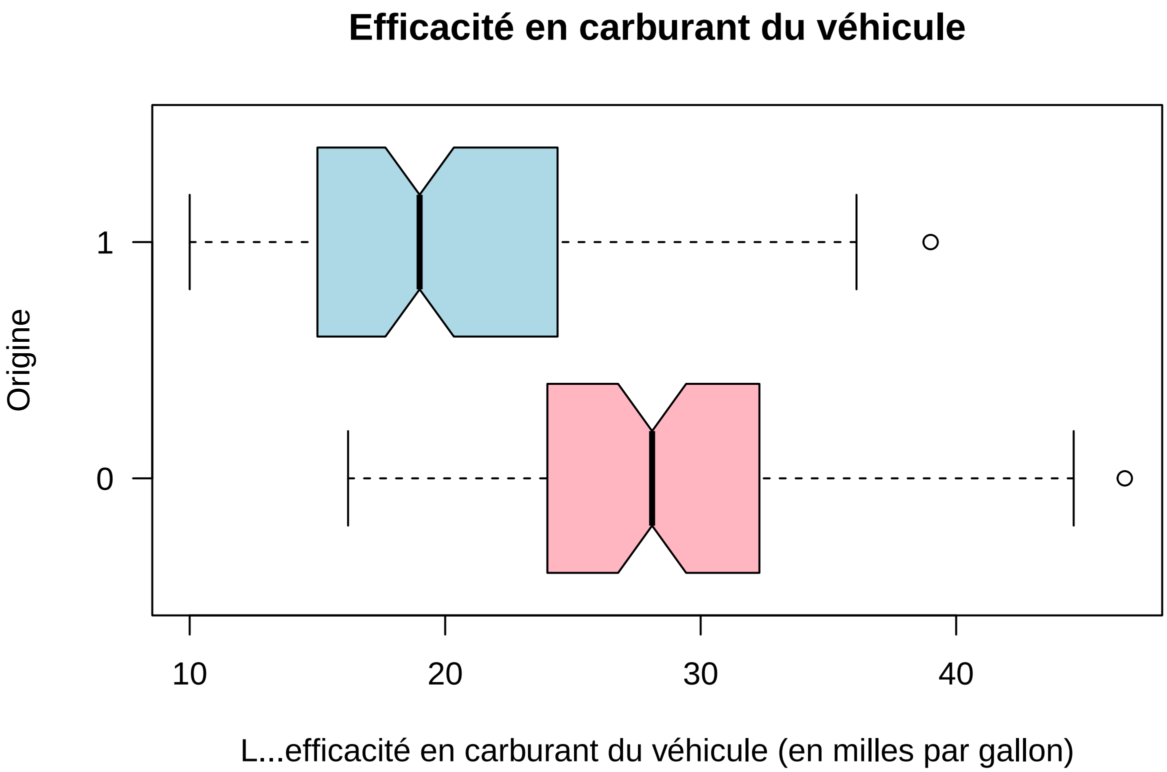


Les deux histogrammes juxtaposés nous indiquent que la distribution pour l'efficacité en carburant du véhicule pour des véhicules provenant d'autres pays montrent un peu une forme de cloche, ainsi on peut considérer que la distribution est approximativement normal. La fréquence pour l'efficacité en carburant des véhicules provenant des états unis sont majoritairement situé à l'extrémité gauche. Ainsi, la majorité des données se situent à l'extrémité gauche. La moyenne de l'efficacité en carburant du véhicule pour les véhicules provenant d'autres pays se situent à peu près à 30 milles/gallons alors que ceux provenant des états-unis se situe à peu près à 20 milles/gallons.

```
#deux diagrammes de Tukey (ou «Box Plot») juxtaposés
boxplot(mondata$mpg~mondata$origin,
col=c("lightpink","lightblue"),
horizontal=TRUE,
notch=TRUE,
main=paste("Efficacité en carburant du véhicule"),
ylab="Origine",
xlab="L'efficacité en carburant du véhicule (en milles par gallon)",
```

```
las=1)
```

```
## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'L’efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'L’efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'L’efficacité en carburant du véhicule (en milles par
## gallon)' in 'mbcsToSbcs': dot substituted for <99>
```



Les deux diagrammes de Tukey juxtaposée corroborent mes propos précédents. l’efficacité en carburant du véhicule provenant d’autre pays est approximativement symétrique, car les deux moitiés de la boîte sont de longueurs sensiblement égales. Tandis que l’efficacité en carburant du véhicule provenant des États-Unis à une distribution positivement asymétrique, car la portion droite de la boîte et la moustache droite sont plus longues qu’à gauche de la médiane. De plus, la médiane pour l’efficacité en carburant du véhicule provenant d’autre pays est à peu près de 29 contrairement à celle des véhicule provenant des états unis qui est de 20.

```
#un tableau des statistiques descriptives pour groupe 0
Quartile <- quantile(mondata$mpg[mondata$origin=="0"])
Variance <- var(mondata$mpg[mondata$origin=="0"])
Moyenne <- mean(mondata$mpg[mondata$origin=="0"])
ÉcartType <- sd(mondata$mpg[mondata$origin=="0"])
ErreurType <- ÉcartType/ sqrt(length(mondata$mpg[mondata$origin=="0"]))
quantile(mondata$mpg[mondata$origin=="0"],c(.05,.95))
```

```
##    5%   95%
## 18.80 39.82
```

```r
InterValleDeConfiance <- t.test(mondata$mpg[mondata$origin=="0"], conf.level = 0.95)
InterValleDeConfiance.min = c(18.80)
InterValleDeConfiance.max = c(39.82)
table <- data.frame( Moyenne=Moyenne,
      "Écart Type"= ÉcartType ,
      "Erreur Type"= ErreurType,
      "Variance"= Variance,
      "min IDC" = InterValleDeConfiance.min,
      "max IDC"= InterValleDeConfiance.max,
      "min"=Quartile[1],
      "p25"= Quartile[2],
      "Mediane"= Quartile[3],
      "p75"= Quartile[4],
      "max"= Quartile[5],
      row.names = "Autres pays")
table
```

```
##              Moyenne Écart.Type Erreur.Type Variance min.IDC max.IDC  min p25
## Autres pays 28.66186   6.702121   0.6804973 44.91843    18.8   39.82 16.2  24
##             Mediane  p75  max
## Autres pays    28.1 32.3 46.6
```

```r
#un tableau des statistiques descriptives pour groupe 1
Quartile <- quantile(mondata$mpg[mondata$origin=="1"])
Moyenne <- mean(mondata$mpg[mondata$origin=="1"])
Variance <- var(mondata$mpg[mondata$origin=="1"])
ÉcartType <- sd(mondata$mpg[mondata$origin=="1"])
ErreurType <- ÉcartType/ sqrt(length(mondata$mpg[mondata$origin=="1"]))
InterValleDeConfiance <- t.test(mondata$mpg[mondata$origin=="1"], conf.level = 0.95)
InterValleDeConfiance.min = c(12.00)
InterValleDeConfiance.max = c(33.25)
table <- data.frame( "Moyenne" =Moyenne,
      "Écart Type"= ÉcartType ,
      "Erreur Type"= ErreurType,
      "Variance"= Variance,
      "min IDC" = InterValleDeConfiance.min,
      "max IDC"= InterValleDeConfiance.max,
      "min"=Quartile[1],
      "p25"= Quartile[2],
      "Mediane"= Quartile[3],
      "p75"= Quartile[4],
      "max"= Quartile[5],
      row.names = "États-Unis")
table
```

```
##             Moyenne Écart.Type Erreur.Type Variance min.IDC max.IDC min p25
## États-Unis 20.27805    6.53428   0.5891761 42.69681      12   33.25  10  15
##            Mediane  p75 max
## États-Unis      19 24.4  39
```

```r
Quartile0 <- quantile(mondata$mpg[mondata$origin=="0"])
Moyenne0 <- mean(mondata$mpg[mondata$origin=="0"])
ÉcartType0 <- sd(mondata$mpg[mondata$origin=="0"])
ErreurType0 <- ÉcartType/ sqrt(length(mondata$mpg[mondata$origin=="0"]))
InterValleDeConfiance0 <- t.test(mondata$mpg[mondata$origin=="0"], conf.level = 0.95)
```

```r
InterValleDeConfiance0.min = c(18.80)
InterValleDeConfiance0.max = c(39.82)

Quartile1 <- quantile(mondata$mpg[mondata$origin=="1"])
Moyenne1 <- mean(mondata$mpg[mondata$origin=="1"])
ÉcartType1 <- sd(mondata$mpg[mondata$origin=="1"])
ErreurType1 <- ÉcartType/ sqrt(length(mondata$mpg[mondata$origin=="1"]))
InterValleDeConfiance1 <- t.test(mondata$mpg[mondata$origin=="1"], conf.level = 0.95)
InterValleDeConfiance1.min = c(12.00)
InterValleDeConfiance1.max = c(33.25)
varianceDesOrigines <- by(mondata$mpg, mondata$origin, function(x) var(x))
varianceDesOrigines
```

```
## mondata$origin: 0
## [1] 44.91843
## ------------------------------------------------------------
## mondata$origin: 1
## [1] 42.69681
```

```r
table <- data.frame( Moyenne=c(Moyenne0,Moyenne1),
      "Écart Type"= c(ÉcartType0, ÉcartType1) ,
      "Erreur Type"= c(ErreurType0, ErreurType1) ,
      "min IDC" = c(InterValleDeConfiance0.min,InterValleDeConfiance1.min) ,
      "max IDC"= c(InterValleDeConfiance0.max,InterValleDeConfiance1.max),
      "Variance" = c(varianceDesOrigines[1], varianceDesOrigines[2]),
      "min"=c(Quartile0[1], Quartile1[1]),
      "p25"= c(Quartile0[2], Quartile1[2]),
      "Mediane"= c(Quartile0[3], Quartile1[3]),
      "p75"= c(Quartile0[4], Quartile1[4]),
      "max"= c(Quartile0[5], Quartile1[5]),
      row.names = c("Autres Pays", "États-Unis"))
table
```

```
##              Moyenne Écart.Type Erreur.Type min.IDC max.IDC Variance  min p25
## Autres Pays 28.66186   6.702121   0.6634556    18.8   39.82 44.91843 16.2  24
## États-Unis  20.27805   6.534280   0.5891761    12.0   33.25 42.69681 10.0  15
##             Mediane  p75  max
## Autres Pays    28.1 32.3 46.6
## États-Unis     19.0 24.4 39.0
```

la moyenne et la moyenne pour L'efficacité en carburant du véhicule (en milles par gallon) est supérieure lorsque la provenance du pays est les états-unis est inférieure à lorqu'elle provient d'autres pays. Dans les deux cas, la moyenne et la médiane sont quasiment pareille dans les deux cas et particulièrement lorsque la provenance du véhicule est d'autres pays, on peut ainsi en déuire que la distribution est symétrique. L'écart type, l'erreur type est sensiblement pareille pour les deux groupes. On peut en conclure que l'efficacité en carburant est meilleur lorsque l'origine est d'autres pays. La moyenne plus grande et les valeurs min et max de l'intervalle de confiance sont supérieure à celle des États-Unis. Ainsi, les résultats sont préférables lorsque la provenance est d'autres pays.

```r
var.test(mpg~origin, data=mondata)
```

```
##
##  F test to compare two variances
##
## data:  mpg by origin
## F = 1.052, num df = 96, denom df = 122, p-value = 0.787
```

```
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7224487 1.5453969
## sample estimates:
## ratio of variances
##           1.052032
```

```
varianceDesOrigines <- by(mondata$mpg, mondata$origin, function(x) var(x))
varianceDesOrigines
```

```
## mondata$origin: 0
## [1] 44.91843
## ------------------------------------------------------------
## mondata$origin: 1
## [1] 42.69681
```

Le test d'hypothèses sur l'égalité des variances des deux groupes. hypothèses Ho : Les deux variances sont égales Hypothèse H1 : Les deux variances ne sont pas égales. La p-value des 2 variances est de 0.787 et puisqu'elle est plus grande que 0.05, l'hypothèse sur l'égalité des deux variances est accepté.

```
t.test(mpg~origin, data=mondata)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by origin
## t = 9.3142, df = 203.77, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   6.609076 10.158538
## sample estimates:
## mean in group 0 mean in group 1
##        28.66186        20.27805
```

```
moyenneDesOrigines <- by(mondata$mpg, mondata$origin, function(x) mean(x))
moyenneDesOrigines
```

```
## mondata$origin: 0
## [1] 28.66186
## ------------------------------------------------------------
## mondata$origin: 1
## [1] 20.27805
```

Le test d'hypothèse sur l'égalité des moyennes des deux groupes. Hypothèses: Ho : Les deux moyennes sont égales. H1 : Les deux moyennes ne sont pas égales. La p-value des 2 moyennes est inférieure à 2.2e-16 et puisqu'elle est inférieure à 0.05, l'hypothèse sur l'égalité des deux moyennes est rejetté.

Recherche d'un modèle : Modèle 1

```
#Modèle 1
modele1 <- lm(mondata$mpg~mondata$displacement)
summary(modele1)
```

```
##
## Call:
## lm(formula = mondata$mpg ~ mondata$displacement)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -12.753  -3.063  -0.556   2.441  16.805
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          34.941400   0.660625   52.89   <2e-16 ***
## mondata$displacement -0.059838   0.003144  -19.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.793 on 218 degrees of freedom
## Multiple R-squared:  0.6243, Adjusted R-squared:  0.6226
## F-statistic: 362.3 on 1 and 218 DF,  p-value: < 2.2e-16
```

Test de signification du modèle 1 :

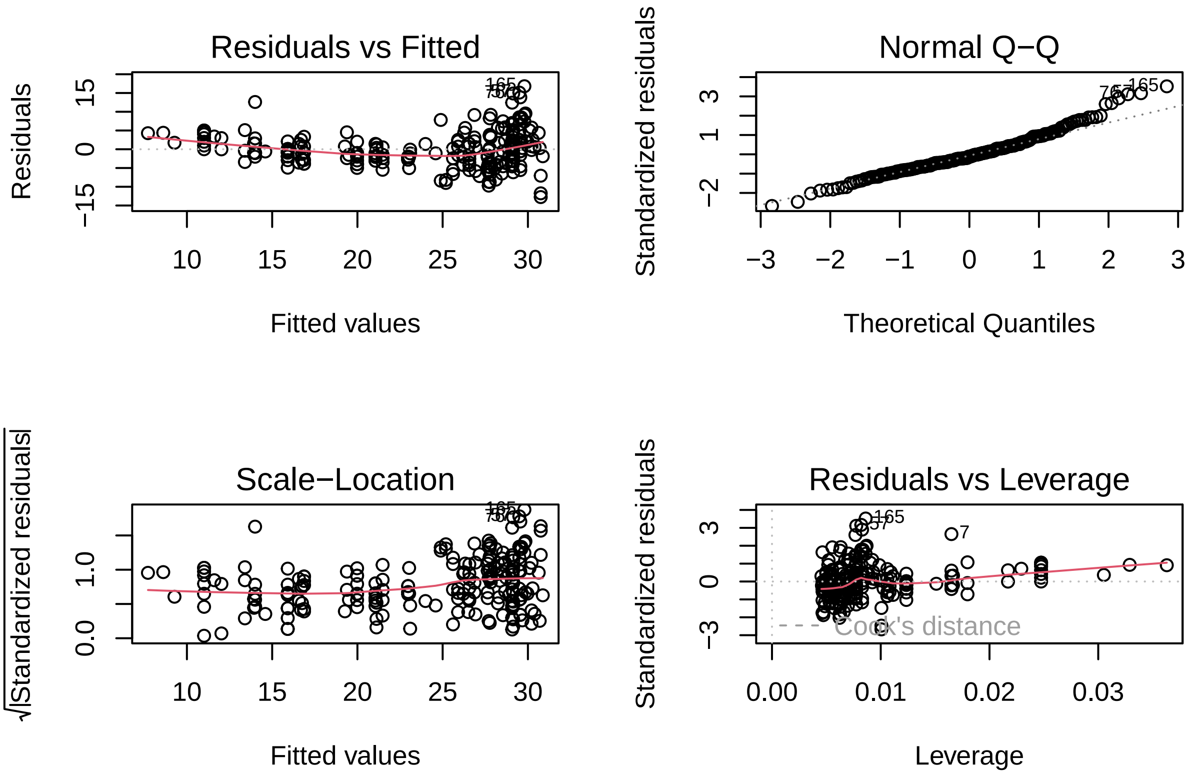
```
#modele 1
anova(modele1)
```

```
## Analysis of Variance Table
##
## Response: mondata$mpg
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## mondata$displacement   1 8324.1  8324.1  362.28 < 2.2e-16 ***
## Residuals            218 5009.0    23.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
shapiro.test(residuals(modele1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modele1)
## W = 0.97182, p-value = 0.0002219
```

```
par(mfrow=c(2,2))
plot(modele1)
```




```
table <- data.frame("B0"=c(34.941400),
                    "B1"=c(-0.059838),
                    row.names = "valeur")
table
```

```
##              B0        B1
## valeur 34.9414 -0.059838
```

Intervalle de confiance du modèle 1 :

```
#modele 1
confint(modele1, level = 0.95)
```

```
##                           2.5 %      97.5 %
## (Intercept)          33.6393701 36.24343031
## mondata$displacement -0.0660344 -0.05364209
```

```
table <- data.frame("min"=c(33.6393701,-0.0660344),
                    "max"=c(36.24343031, -0.05364209),
                    row.names = c("B0", "B1"))
table
```

```
##           min         max
## B0 33.6393701 36.24343031
## B1 -0.0660344 -0.05364209
```

On voit dans le graphique Normal Q-Q que la distribution suit une loi normal car les points se situent sur la droite. De plus, la valeur de R squared est 0.6243 ce qui n'est pas si loin de 1 donc le modèle est valable. La valeur de p dans le tableau d'analyse de la variance est 2.2e-16 et puisque la p value est inférieure à 0.05, on rejette l'hypothès qui dit que $B1 = 0$ et on accepte celle qui dit que B1 n'est pas égale à 0. La cylindré du moteur du véhicule à un grand impact sur l'efficacité en carburant du véhicule. La valeur de p dans le test de shapiro est de 0.0002219 ce qui inférieur à 0.05 ce qui veut dire qu'on rejette H0 et on peut en conclure que les résidus ne suivent pas une loi normale. On infirme ainsi l'information du graphique normal Q-Q. De plus, la dispersion des résidus dans les 3 autres graphiques permettent de confirmer que le modèle n'est pas homogène donc de ce fait même invalide.

Modèle2

```
#Modèle 2
modele2 <- lm(mondata$mpg~((mondata$displacement)^2))
summary(modele2)
```

```
##
## Call:
## lm(formula = mondata$mpg ~ ((mondata$displacement)^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.753  -3.063  -0.556   2.441  16.805
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          34.941400   0.660625   52.89   <2e-16 ***
## mondata$displacement -0.059838   0.003144  -19.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.793 on 218 degrees of freedom
## Multiple R-squared:  0.6243, Adjusted R-squared:  0.6226
## F-statistic: 362.3 on 1 and 218 DF,  p-value: < 2.2e-16
```

Test de signification du modèle 2 :

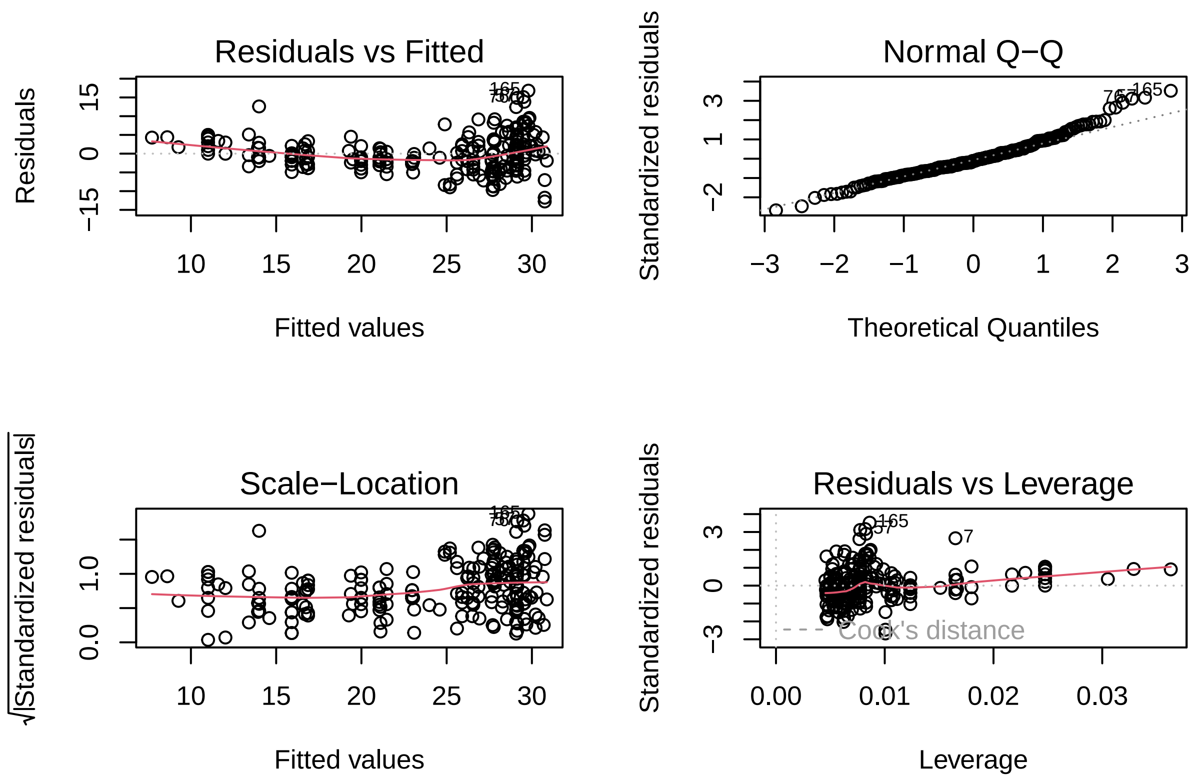
```
#modèle 2
anova(modele2)
```

```
## Analysis of Variance Table
##
## Response: mondata$mpg
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## mondata$displacement   1 8324.1  8324.1  362.28 < 2.2e-16 ***
## Residuals            218 5009.0    23.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
shapiro.test(residuals(modele2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modele2)
## W = 0.97182, p-value = 0.0002219
```

```
par(mfrow=c(2,2))
plot(modele2)
```




```r
table <- data.frame("B0"=c(34.941400),
                    "B1"=c(-0.059838),
                    row.names = "valeur")
table
```

```
##              B0        B1
## valeur 34.9414 -0.059838
```

On voit dans le graphique Normal Q-Q que la distribution suit une loi normal car les points se situent sur la droite. De plus, la valeur de R squared est 0.6243 ce qui n'est pas si loin de 1 donc le modèle est valable. La valeur de p dans le tableau d'analyse de la variance est 2.2e-16 et puisque la p value est inférieure à 0.05, on rejette l'hypothès qui dit que $B1 = 0$ et on accepte celle qui dit que B1 n'est pas égale à 0. La cylindré du moteur du véhicule à un grand impact sur l'efficacité en carburant du véhicule. La valeur de p dans le test de shapiro est de 0.0002219 ce qui inférieur à 0.05 ce qui veut dire qu'on rejette H0 et on peut en conclure que les résidus ne suivent pas une loi normale. On infirme ainsi l'information du graphique normal Q-Q. De plus, la dispersion des résidus dans les 3 autres graphiques permettent de confirmer que le modèle n'est pas homogène donc de ce fait même invalide.

Modèle 3

```r
#Modèle 3
modele3 <- lm(log(mondata$mpg)~log(mondata$displacement))
summary(modele3)
```

```
##
## Call:
## lm(formula = log(mondata$mpg) ~ log(mondata$displacement))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66894 -0.12200 -0.00008  0.12582  0.58338
```

```
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.83420    0.11536   50.57   <2e-16 ***
## log(mondata$displacement) -0.53546    0.02265  -23.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1788 on 218 degrees of freedom
## Multiple R-squared:  0.7193, Adjusted R-squared:  0.718
## F-statistic: 558.7 on 1 and 218 DF,  p-value: < 2.2e-16
```

Test de signification du modèle 3 :

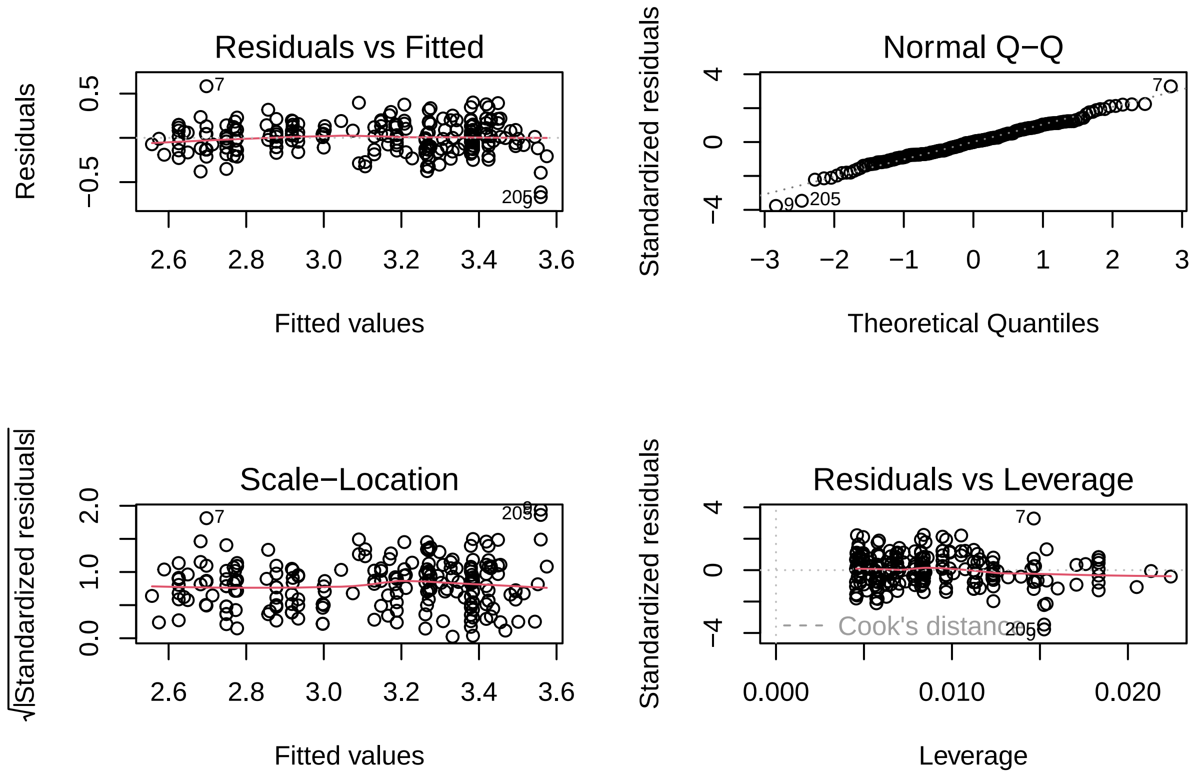
```r
anova(modele3)
```

```
## Analysis of Variance Table
##
## Response: log(mondata$mpg)
##                             Df  Sum Sq Mean Sq F value    Pr(>F)
## log(mondata$displacement)   1 17.8614  17.861  558.72 < 2.2e-16 ***
## Residuals                 218  6.9692   0.032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
shapiro.test(residuals(modele3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modele3)
## W = 0.98723, p-value = 0.04617
```

```r
par(mfrow=c(2,2))
plot(modele3)
```




```
table <- data.frame("B0"=c(5.83420),
                    "B1"=c(-0.059838),
                    row.names = "valeur")
table
```

```
##            B0        B1
## valeur 5.8342 -0.059838
```

On voit dans le graphique Normal Q-Q que la distribution suit une loi normal car les points se situent sur la droite. De plus, la valeur de R squared est 0.7193 ce qui n'est pas si loin de 1 donc le modèle est valable. La valeur de p dans le tableau d'analyse de la variance est 2.2e-16 et puisque la p value est inférieure à 0.05, on rejette l'hypothèse qui dit que $B1 = 0$ et on accepte celle qui dit que B1 n'est pas égale à 0. La cylindré du moteur du véhicule à un grand impact sur l'efficacité en carburant du véhicule. La valeur de p dans le test de shapiro est de 0.04617 ce qui est légèrement inférieur à 0.05 ce qui veut dire qu'on rejette H0 et on peut en conclure que les résidus ne suivent pas une loi normale. On infirme ainsi l'information du graphique normal Q-Q car la majorité des points se trouvaient sur la droite sauf ceux à l'extrémité gauche. De plus, la dispersion des résidus dans les 3 autres graphiques permettent de confirmer que le modèle n'est pas homogène donc de ce fait même invalide.

Modèle 4

```
#Modèle 4
modele4 <- lm(log(mondata$mpg)~mondata$displacement)
summary(modele4)
```

```
##
## Call:
## lm(formula = log(mondata$mpg) ~ mondata$displacement)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.54517 -0.11075 -0.01295  0.11666  0.61958
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.629096   0.024796  146.36   <2e-16 ***
## mondata$displacement -0.002765   0.000118  -23.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1799 on 218 degrees of freedom
## Multiple R-squared:  0.7158, Adjusted R-squared:  0.7145
## F-statistic: 549.1 on 1 and 218 DF,  p-value: < 2.2e-16
```

test de signification du modèle 4:

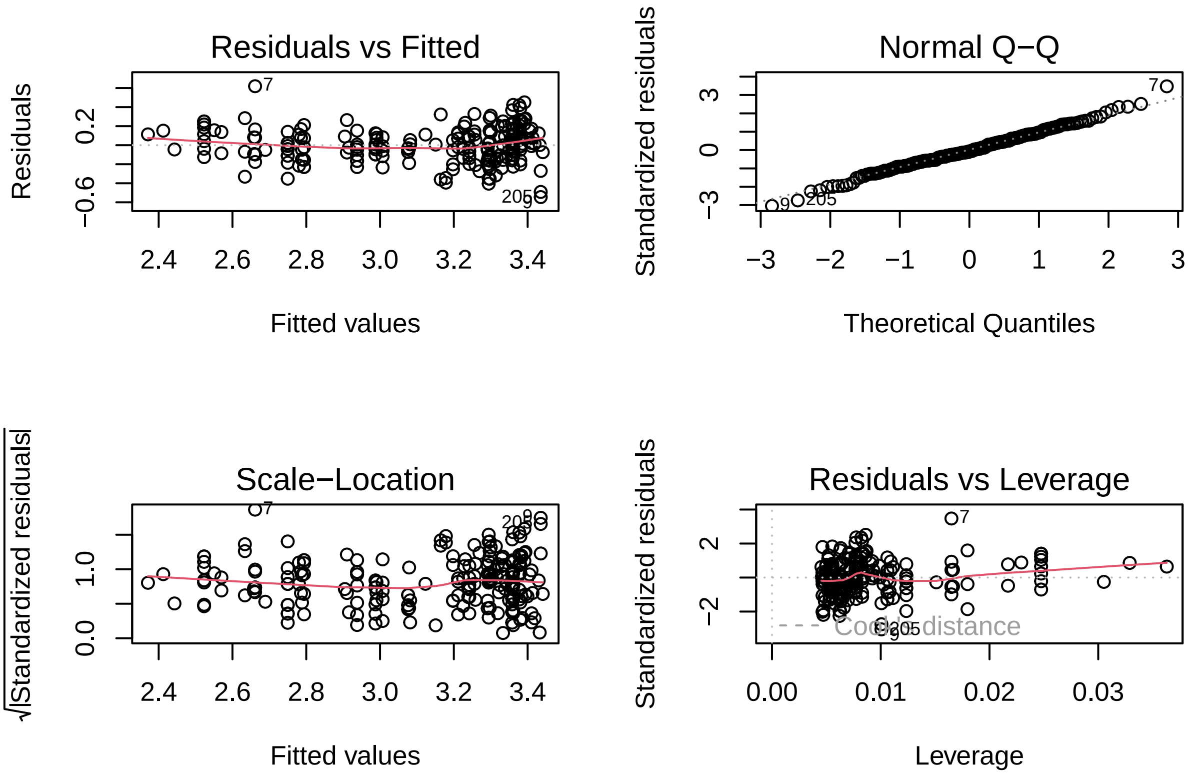
```
#modèle 4
anova(modele4)
```

```
## Analysis of Variance Table
##
## Response: log(mondata$mpg)
##                       Df  Sum Sq Mean Sq F value    Pr(>F)
## mondata$displacement   1 17.7740 17.7740   549.1 < 2.2e-16 ***
## Residuals            218  7.0566  0.0324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
shapiro.test(residuals(modele4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modele4)
## W = 0.9954, p-value = 0.7515
```

```
par(mfrow=c(2,2))
plot(modele4)
```




```
table <- data.frame("B0"=c(3.629096),
                    "B1"=c(-0.002765),
                    row.names = "valeur")
table
```

```
##               B0        B1
## valeur 3.629096 -0.002765
```

On voit dans le graphique Normal Q-Q que la distribution suit une loi normal car les points se situent sur la droite. De plus, la valeur de R squared est 0.7158 ce qui n'est pas si loin de 1 donc le modèle est valable. La valeur de p dans le tableau d'analyse de la variance est 2.2e-16 et puisque la p value est inférieure à 0.05, on rejette l'hypothès qui dit que $B1 = 0$ et on accepte celle qui dit que B1 n'est pas égale à 0. La cylindré du moteur du véhicule à un grand impact sur l'efficacité en carburant du véhicule. La valeur de p dans le test de shapiro est de 0.7515 ce qui supérieure à 0.05 ce qui veut dire qu'on accepte H0 et on peut en conclure que les résidus suivent une loi normale. On confirme ainsi l'information du graphique normal Q-Q.

Modèle 5

```
#Modèle 5
modele5 <- lm(mondata$mpg~mondata$weight)
summary(modele5)
```

```
##
## Call:
## lm(formula = mondata$mpg ~ mondata$weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8959  -2.8309  -0.4423   2.1864  16.5979
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    46.0168548  1.0594816   43.43   <2e-16 ***
## mondata$weight -0.0075899  0.0003502  -21.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.403 on 218 degrees of freedom
## Multiple R-squared:  0.683,  Adjusted R-squared:  0.6816
## F-statistic: 469.7 on 1 and 218 DF,  p-value: < 2.2e-16
```

Test de signification du modèle 5:

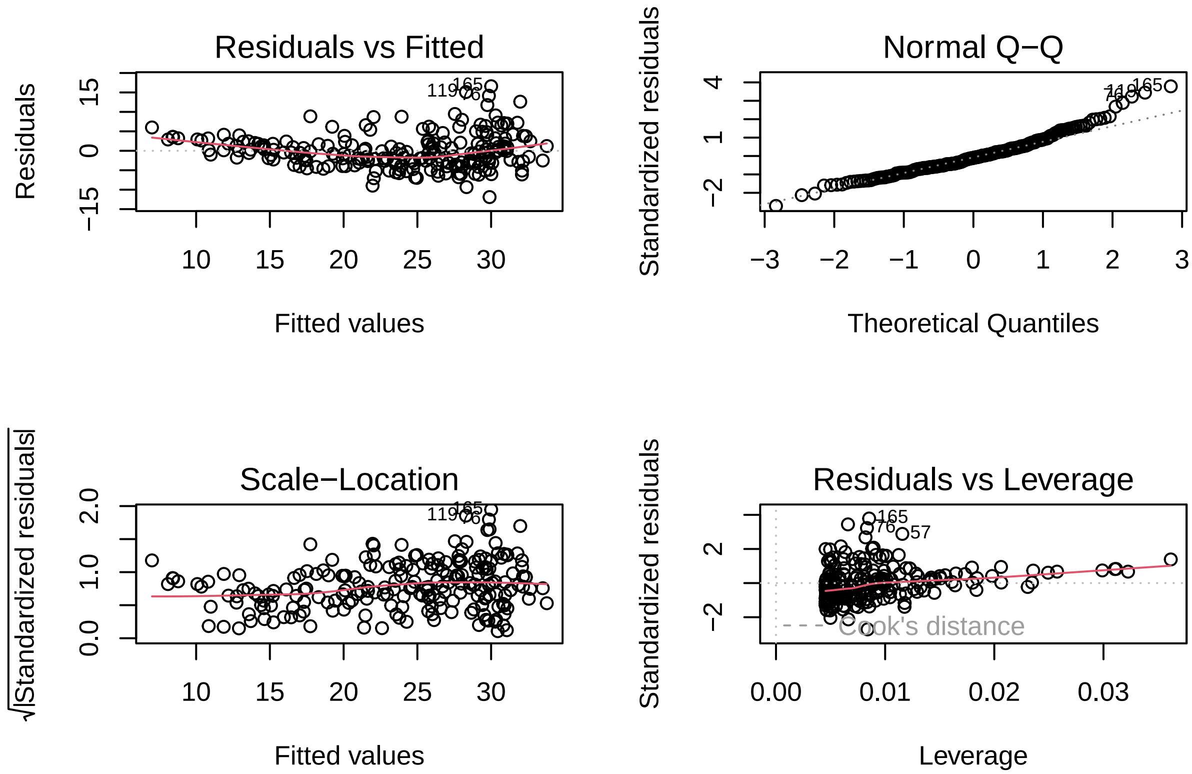
```
#Modèle 5
anova(modele5)
```

```
## Analysis of Variance Table
##
## Response: mondata$mpg
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## mondata$weight   1 9106.6  9106.6  469.72 < 2.2e-16 ***
## Residuals      218 4226.5    19.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
shapiro.test(residuals(modele5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modele5)
## W = 0.96367, p-value = 2.07e-05
```

```
par(mfrow=c(2,2))
plot(modele5)
```




```r
table <- data.frame("B0"=c(46.0168548),
                    "B1"=c(-0.0075899),
                    row.names = "valeur")
table
```

```
##                B0         B1
## valeur 46.01685 -0.0075899
```

Intervalle de confiance du modèle 5 :

```r
confint(modele5, level = 0.95)
```

```
##                        2.5 %       97.5 %
## (Intercept)    43.928716550 48.104993046
## mondata$weight -0.008280127 -0.006899695
```

```r
table <- data.frame("min"=c(43.928716550,-0.008280127),
                    "max"=c(48.104993046, -0.006899695),
                    row.names = c("B0", "B1"))
table
```

```
##             min          max
## B0 43.928716550 48.104993046
## B1 -0.008280127 -0.006899695
```

On voit dans le graphique Normal Q-Q que la distribution ne suit pas une loi normal car les points situés à l'extrémité droite ne touchent pas la droite. De plus, la valeur de R squared est 0.683 ce qui n'est pas si loin de 1 donc le modèle est valable. La valeur de p dans le tableau d'analyse de la variance est 2e-16 et puisque la p value est inférieure à 0.05, on rejette l'hypothès qui dit que $B1 = 0$ et on accepte celle qui dit que B1 n'est pas égale à 0. Le poids en livre du véhicule à un grand impact sur l'efficacité en carburant du véhicule. La valeur de p dans le test de shapiro est de 2.07e-05 ce qui inférieur à 0.05 ce qui veut dire qu'on rejette H0 et on peut en conclure que les résidus ne suivent pas une loi normale. On confirme ainsi l'information du graphique normal Q-Q.

Modèle 6

```
#Modèle 6
modele6 <- lm(mondata$mpg~(mondata$weight)^2)
summary(modele6)
```

```
##
## Call:
## lm(formula = mondata$mpg ~ (mondata$weight)^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8959  -2.8309  -0.4423   2.1864  16.5979
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    46.0168548  1.0594816   43.43   <2e-16 ***
## mondata$weight -0.0075899  0.0003502  -21.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.403 on 218 degrees of freedom
## Multiple R-squared:  0.683,  Adjusted R-squared:  0.6816
## F-statistic: 469.7 on 1 and 218 DF,  p-value: < 2.2e-16
```

Tet de signification du modèle 6:

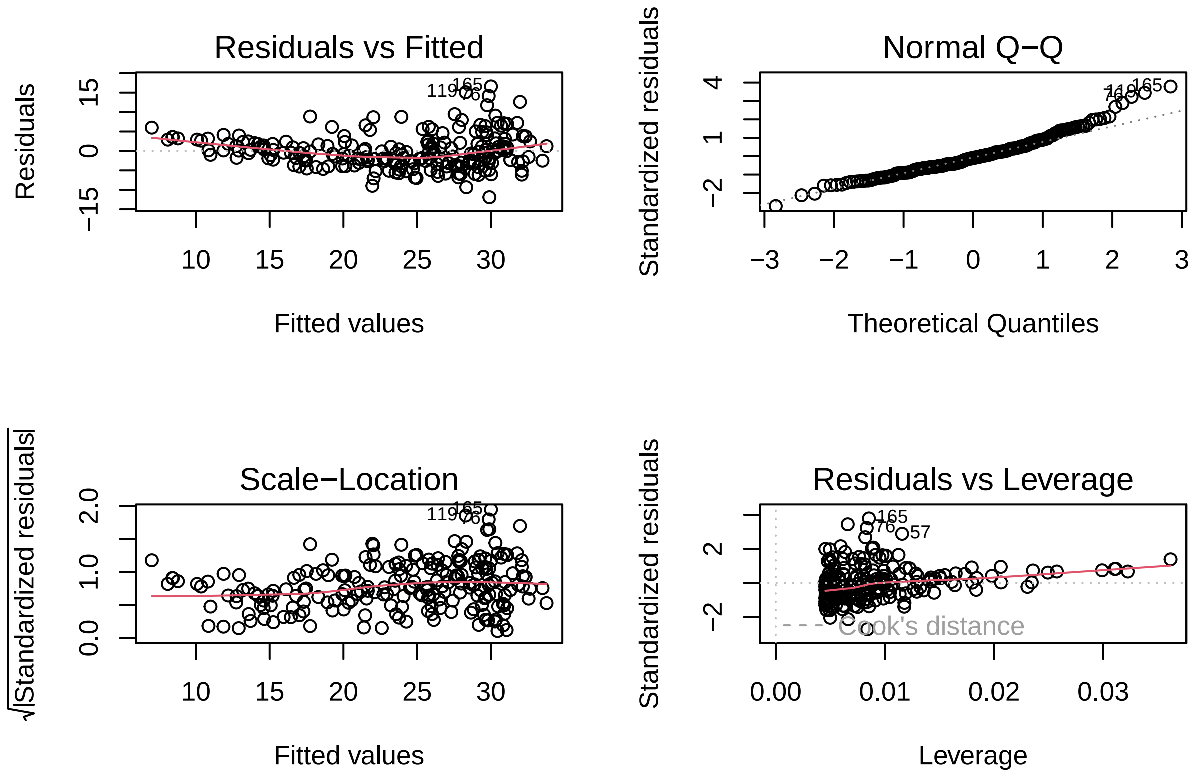
```
#Modèle 6
anova(modele6)
```

```
## Analysis of Variance Table
##
## Response: mondata$mpg
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## mondata$weight   1 9106.6  9106.6  469.72 < 2.2e-16 ***
## Residuals      218 4226.5    19.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
shapiro.test(residuals(modele6))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modele6)
## W = 0.96367, p-value = 2.07e-05
```

```
par(mfrow=c(2,2))
plot(modele6)
```




```r
table <- data.frame("B0"=c(46.0168548),
                    "B1"=c(-0.0075899),
                    row.names = "valeur")
table
```

```
##              B0         B1
## valeur 46.01685 -0.0075899
```

On voit dans le graphique Normal Q-Q que la distribution ne suit pas une loi normal car les points se situant à l'extrémité droite ne sont pas sur la droite. De plus, la valeur de R squared est 0.683 ce qui n'est pas si loin de 1 donc le modèle est valable. La valeur de p dans le tableau d'analyse de la variance est 2e-16 et puisque la p value est inférieure à 0.05, on rejette l'hypothès qui dit que $B1 = 0$ et on accepte celle qui dit que B1 n'est pas égale à 0. Le poids en livre du véhicule à un grand impact sur l'efficacité en carburant du véhicule. La valeur de p dans le test de shapiro est de 2.07e-05 ce qui inférieur à 0.05 ce qui veut dire qu'on rejette H0 et on peut en conclure que les résidus ne suivent pas une loi normale. On confirme ainsi l'information du graphique normal Q-Q.

Modèle 7

```r
#Modèle 7
modele7 <- lm(log(mondata$mpg)~log(mondata$weight))
summary(modele7)
```

```
##
## Call:
## lm(formula = log(mondata$mpg) ~ log(mondata$weight))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51608 -0.10557 -0.00818  0.08908  0.46281
##
```

```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          11.39849    0.31047   36.71   <2e-16 ***
## log(mondata$weight) -1.04320    0.03911  -26.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1634 on 218 degrees of freedom
## Multiple R-squared:  0.7655, Adjusted R-squared:  0.7644
## F-statistic: 711.5 on 1 and 218 DF,  p-value: < 2.2e-16
```

test de signification du modèle 7:

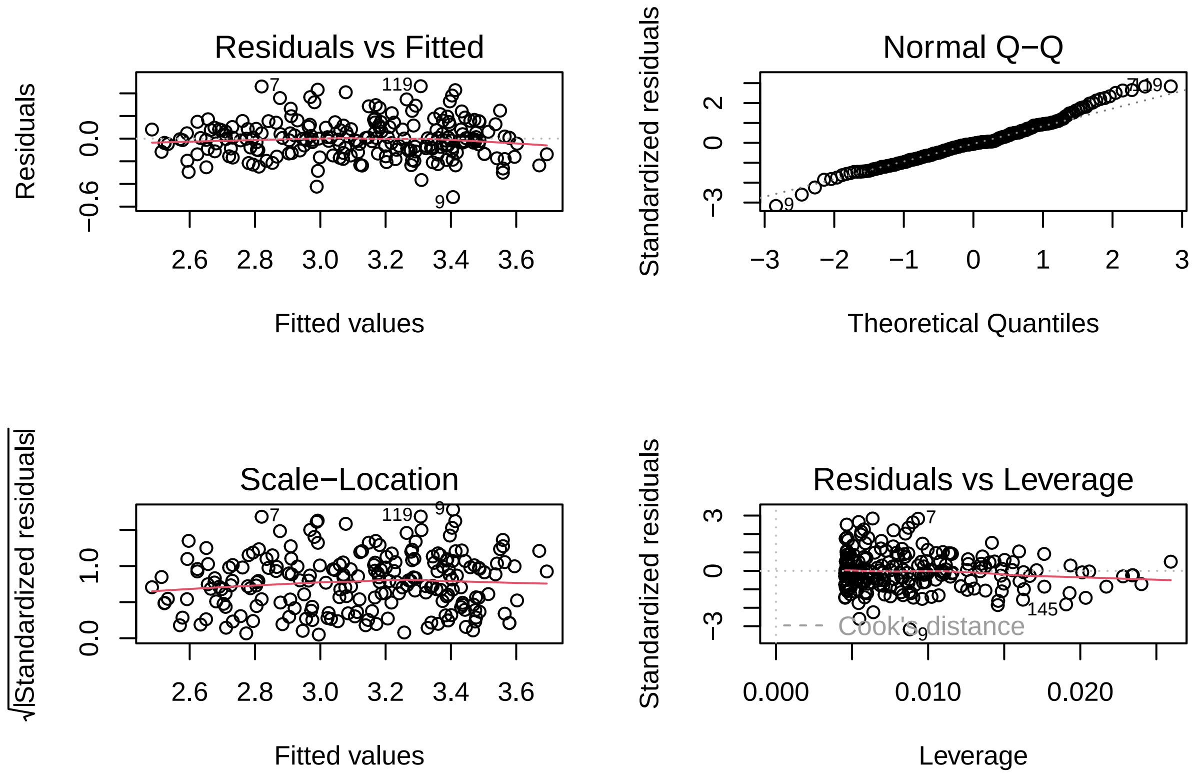
```
#modèle 7
anova(modele7)
```

```
## Analysis of Variance Table
##
## Response: log(mondata$mpg)
##                       Df  Sum Sq Mean Sq F value    Pr(>F)
## log(mondata$weight)   1 19.0068 19.0068  711.48 < 2.2e-16 ***
## Residuals           218  5.8238  0.0267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
shapiro.test(residuals(modele7))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modele7)
## W = 0.9832, p-value = 0.01028
```

```
par(mfrow=c(2,2))
plot(modele7)
```




```
table <- data.frame("B0"=c(11.39849),
                    "B1"=c(-1.04320),
                    row.names = "valeur")
table
```

```
##               B0      B1
## valeur 11.39849 -1.0432
```

On voit dans le graphique Normal Q-Q que la distribution ne suit pas une loi normal car les points se situant aux extrémités ne sont pas sur la droite. De plus, la valeur de R squared est 0.7655 ce qui n'est pas si loin de 1 donc le modèle est valable. La valeur de p dans le tableau d'analyse de la variance est 2e-16 et puisque la p value est inférieure à 0.05, on rejette l'hypothès qui dit que B1 = 0 et on accepte celle qui dit que B1 n'est pas égale à 0. Le poids en livre du véhicule à un grand impact sur l'efficacité en carburant du véhicule. La valeur de p dans le test de shapiro est de 0.01028 ce qui inférieur à 0.05 ce qui veut dire qu'on rejette H0 et on peut en conclure que les résidus ne suivent pas une loi normale. On confirme ainsi l'information du graphique normal Q-Q.

Modèle 8

```
#Modèle 8
X2= mondata$weight
modele8 <- lm(log(mondata$mpg)~X2)
summary(modele8)
```

```
##
## Call:
## lm(formula = log(mondata$mpg) ~ X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50260 -0.09636 -0.00817  0.09016  0.45069
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.130e+00  3.927e-02  105.16   <2e-16 ***
## X2          -3.469e-04  1.298e-05  -26.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1632 on 218 degrees of freedom
## Multiple R-squared:  0.7661, Adjusted R-squared:  0.7651
## F-statistic: 714.2 on 1 and 218 DF,  p-value: < 2.2e-16
```

Test de signification du modèle 8:

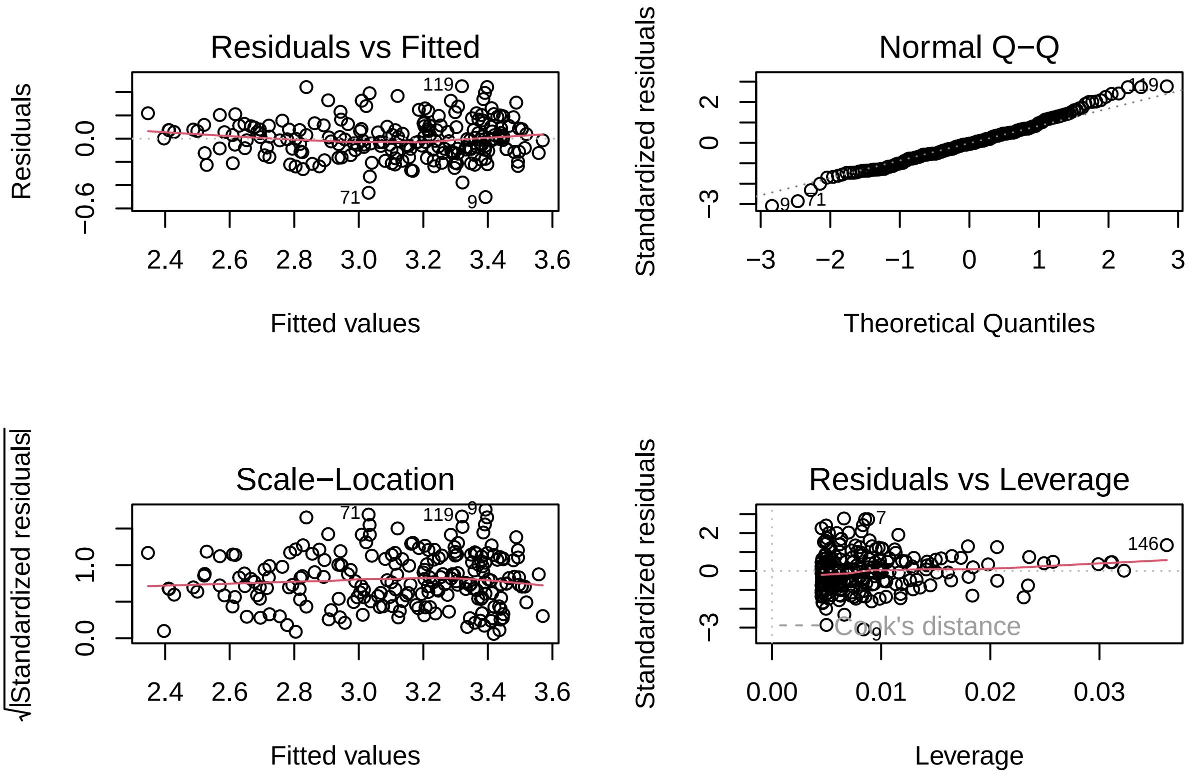
```
anova(modele8)
```

```
## Analysis of Variance Table
##
## Response: log(mondata$mpg)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X2          1 19.0236 19.0236  714.18 < 2.2e-16 ***
## Residuals 218  5.8069  0.0266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
shapiro.test(residuals(modele8))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modele8)
## W = 0.98879, p-value = 0.08367
```

```
par(mfrow=c(2,2))
plot(modele8)
```




```
table <- data.frame("B0"=c(4.130e+00),
                    "B1"=c(-3.469e-04),
                    row.names = "valeur")
table
```

```
##            B0         B1
## valeur 4.13 -0.0003469
```

On voit dans le graphique Normal Q-Q que la distribution ne suit pas une loi normal car les points se situant aux deux extrémités ne sont pas sur la droite. De plus, la valeur de R squared est 0.7661 ce qui n'est pas si loin de 1 donc le modèle est valable. La valeur de p dans le tableau d'analyse de la variance est 2e-16 et puisque la p value est inférieure à 0.05, on rejette l'hypothès qui dit que $B1 = 0$ et on accepte celle qui dit que B1 n'est pas égale à 0. Le poids en livre du véhicule à un grand impact sur l'efficacité en carburant du véhicule. La valeur de p dans le test de shapiro est de 0.08367 ce qui est supérieur à 0.05 ce qui veut dire qu'on accepte H0 et on peut en conclure que les résidus suivent une loi normale. On infirme ainsi l'information du graphique normal Q-Q.

La comparaison des 8 modèles nous indiquent que le modèle à préconiser est ainsi le modèle 8. Il a la valeur de R carrée la plus proche de 1 ce qui veut dire que c'est le modèle le plus valable. De plus, le poids en livre à un impact important sur la l'efficacité en carburant du véhicule et les résidus suivent une loi normale.

Calculez un intervalle de prévision pour l'efficacité en carburant d'un véhicule ayant les caractéristiques suivantes : $X1 = 190$; $X2 = 2500$.

```
#le meilleur modèle est le modèle 8

predict(modele8, data.frame(X2=2500), level = 0.95, interval="prediction")
```

```
##        fit      lwr      upr
## 1 3.262535 2.939969 3.585101
```

```r
exp(3.262535)
```

```
## [1] 26.11566
```

```r
exp(2.939969)
```

```
## [1] 18.91526
```

```r
exp(3.585101)
```

```
## [1] 36.057
```

Le modèle 8 prédit que l'intervalle de prévision est de [18.91526; 36.057]. Cet intervalle pour L'efficacité en carburant du véhicule (en milles par gallon) est relativement grand mais les données coîncide avec les valeurs de l'échantillon. J'en conclus que le modèle 8 est valables et mais qu'il faudrait considérer plus de facteurs pour avoir un intervalle réduit.