POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

DÉPARTEMENT DE MATHÉMATIQUESET DE GÉNIE INDUSTRIEL

MTH2302D - PROBABILITÉS ET STATISTIQUE

# Devoir - Automne 2022

**Date de remise : 7 décembre avant 23h59 (dans Moodle)**

---

**Veuillez remplir le tableau suivant et joindre cette page à votre rapport.**

| **Identification de l'étudiant** | |
|---|---|
| **Nom : Chowdhury** | **Prénom : Rasel** |
| **Groupe : 03** | **Matricule : XXXXXXX** |

Placer les deux fichiers DevoirD_A22.csv et charger.R dans le répertoire de travail de R. En utilisant votre **matricule**, exécuter ensuite (dans cet ordre) les deux commandes suivantes dans Rpour générer votre ensemble de données personnalisées 'mondata' :

source('**charger.R**')
mondata<-charger(**matricule**)

---

| **Question** | **Note** |
|---|---|
| **a)** | **/2** |
| **b)** | **/8** |
| **c)** | **/8** |
| **d)** | **/14** |
| **e)** | **/6** |
| **Présentation** | **/2** |
| **TOTAL** | **/40** |

Mercredi le 7 décembre 2022

# Table de matières

**Chargement des données en fonction de ma matricule à partir du fichier <<charger.r>> :**

```{r}
source('charger.R');
mondata <- charger(XXXXXXX);
```

**Affichage des données obtenus suite au chargement du fichier <<charger>> :**

| | mpg | displacement | weight | Origin |
|---|---|---|---|---|
| **156** | 16.0 | 400.0 | 4220 | 1 |
| **215** | 32.0 | 91.0 | 1965 | 0 |
| **94** | 30.5 | 97.0 | 2190 | 0 |
| **78** | 16.0 | 225.0 | 3439 | 1 |
| **12** | 21.1 | 134.0 | 2515 | 0 |
| **207** | 29.0 | 135.0 | 2525 | 1 |
| **138** | 20.0 | 114.0 | 2582 | 0 |
| **103** | 18.0 | 199.0 | 2774 | 1 |
| **110** | 26.6 | 151.0 | 2635 | 1 |
| **234** | 31.0 | 79.0 | 2000 | 0 |
| **144** | 25.4 | 183.0 | 3530 | 0 |
| **36** | 33.5 | 98.0 | 2075 | 1 |

| | mpg | displacement | weight | Origin |
|---|---|---|---|---|
| **243** | 19.0 | 156.0 | 2930 | 0 |
| **183** | 30.0 | 98.0 | 2155 | 1 |
| **70** | 12.0 | 350.0 | 4499 | 1 |
| **44** | 25.0 | 97.5 | 2126 | 1 |
| **125** | 19.1 | 225.0 | 3381 | 1 |
| **154** | 16.0 | 250.0 | 3278 | 1 |
| **66** | 11.0 | 350.0 | 3664 | 1 |
| **182** | 37.0 | 119.0 | 2434 | 0 |
| **91** | 14.0 | 340.0 | 3609 | 1 |
| **228** | 39.0 | 86.0 | 1875 | 1 |
| **146** | 22.0 | 122.0 | 2395 | 1 |
| **237** | 20.2 | 200.0 | 2965 | 1 |
| **93** | 31.0 | 76.0 | 1649 | 0 |
| **63** | 27.5 | 134.0 | 2560 | 0 |
| **198** | 23.2 | 156.0 | 2745 | 1 |
| **118** | 35.0 | 72.0 | 1613 | 0 |

| | mpg | displacement | weight | Origin |
|---|---|---|---|---|
| **124** | 14.0 | 318.0 | 4457 | 1 |
| **90** | 44.6 | 91.0 | 1850 | 0 |
| **80** | 15.5 | 304.0 | 3962 | 1 |
| **158** | 20.2 | 302.0 | 3570 | 1 |
| **75** | 36.0 | 107.0 | 2205 | 0 |
| **64** | 26.0 | 97.0 | 2265 | 0 |
| **160** | 13.0 | 360.0 | 3821 | 1 |
| **42** | 25.0 | 140.0 | 2572 | 1 |
| **224** | 19.0 | 250.0 | 3302 | 1 |
| **35** | 31.5 | 89.0 | 1990 | 0 |
| **240** | 11.0 | 429.0 | 4633 | 1 |
| **230** | 18.0 | 232.0 | 2945 | 1 |
| **200** | 28.0 | 120.0 | 2625 | 1 |
| **29** | 32.3 | 97.0 | 2065 | 0 |
| **164** | 26.0 | 98.0 | 2265 | 0 |
| **87** | 14.0 | 318.0 | 4077 | 1 |

| | mpg | displacement | weight | Origin |
|---|---|---|---|---|
| **92** | 21.5 | 231.0 | 3245 | 1 |
| **194** | 30.0 | 88.0 | 2065 | 0 |
| **22** | 15.0 | 390.0 | 3850 | 1 |
| **246** | 34.2 | 105.0 | 2200 | 1 |
| **109** | 20.0 | 225.0 | 3651 | 1 |
| **121** | 34.0 | 112.0 | 2395 | 1 |
| **208** | 11.0 | 400.0 | 4997 | 1 |
| **15** | 13.0 | 307.0 | 4098 | 1 |
| **132** | 25.0 | 90.0 | 2223 | 0 |
| **175** | 18.5 | 360.0 | 3940 | 1 |
| **233** | 23.0 | 120.0 | 2957 | 0 |
| **47** | 29.0 | 97.0 | 1940 | 0 |
| **169** | 27.2 | 141.0 | 3190 | 0 |
| **7** | 18.5 | 250.0 | 3525 | 1 |
| **213** | 21.0 | 140.0 | 2401 | 1 |
| **199** | 36.0 | 120.0 | 2160 | 0 |

| | mpg | displacement | weight | Origin |
|---|---|---|---|---|
| **216** | 25.8 | 156.0 | 2620 | 1 |
| **141** | 28.0 | 107.0 | 2464 | 0 |
| **26** | 20.0 | 262.0 | 3221 | 1 |
| **96** | 34.4 | 98.0 | 2045 | 1 |
| **133** | 33.8 | 97.0 | 2145 | 0 |
| **192** | 14.0 | 302.0 | 4042 | 1 |
| **19** | 32.0 | 83.0 | 2003 | 0 |
| **250** | 17.0 | 260.0 | 4060 | 1 |
| **127** | 21.0 | 120.0 | 2979 | 0 |
| **43** | 15.0 | 383.0 | 3563 | 1 |
| **9** | 13.0 | 350.0 | 4274 | 1 |
| **46** | 29.5 | 98.0 | 2135 | 0 |

# Partie 1 : Analyse statistique descriptive et inférence

## 1.a) Matrice de corrélation

**Production de la matrice de corrélation des 3 variables quantitatives, soit l'efficacité en carburant du véhicule (en milles par gallon) *(représenté dans ce rapport sous la forme de la variable Y),* la cylindrée du moteur du véhicule (en pouces cubes) (*représenté dans ce rapport sous la forme de la variable X1)* ainsi que le poids du véhicule (en livres) (*représenté dans ce rapport sous la forme de la variable X2).***

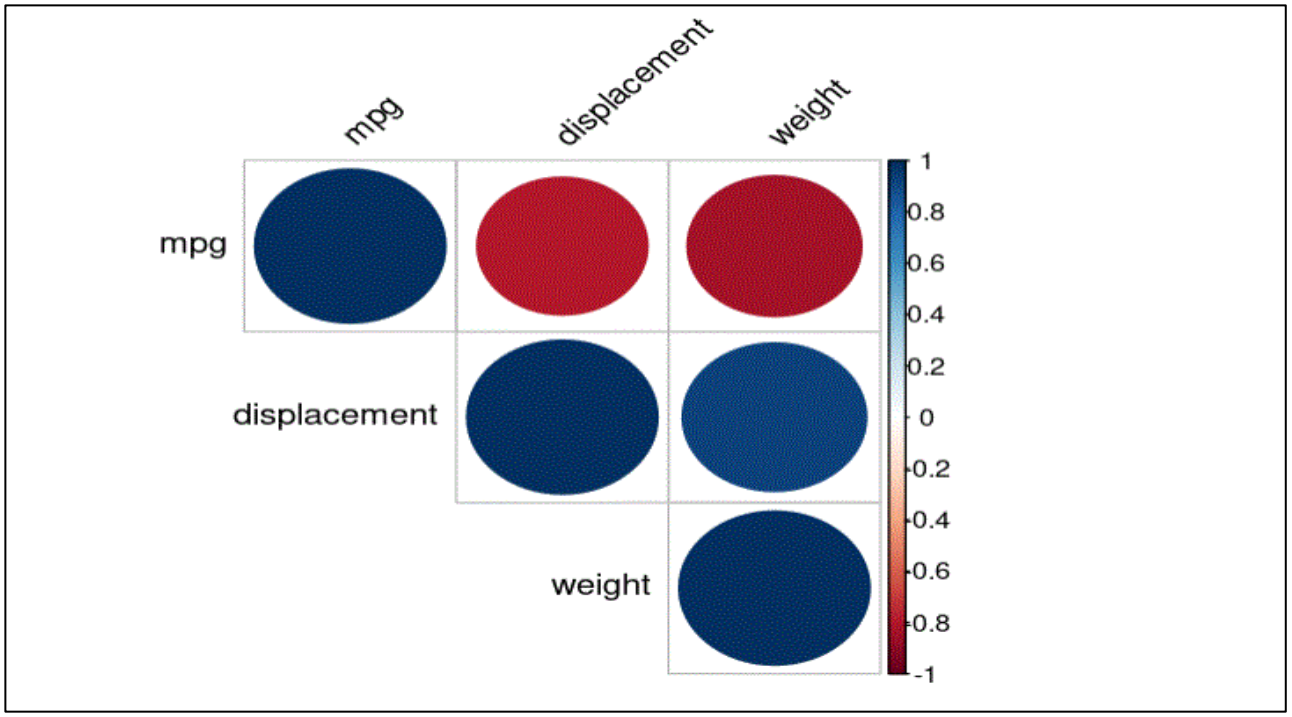
```{r}
mondata_Correle <- mondata[, c(1,2,3)]
head(mondata_Correle,3)

Affichage_mondata_Correle <- cor(mondata_Correle)
round(Affichage_mondata_Correle,2)
```

```
#                mpg displacement weight
# mpg           1.00        -0.80  -0.83
# displacement -0.80         1.00   0.93
# weight       -0.83         0.93   1.00
```

D'apres la matrice de corrélation et les coefficients générés, il peut être déduit que les 3 valeurs quantitatives sont fortement corrélées entre eux, et ce qu'ils soient positifs ou négatifs. Deux graphiques seront reproduits ci-dessous pour montrer cet aspect d'un point de vue visuel des données matricielles **(même si le point de vue n'est pas une tache demandée, il sera tout de même reproduit pour supporter les données observées).**

**Production d'un corrélogramme :**

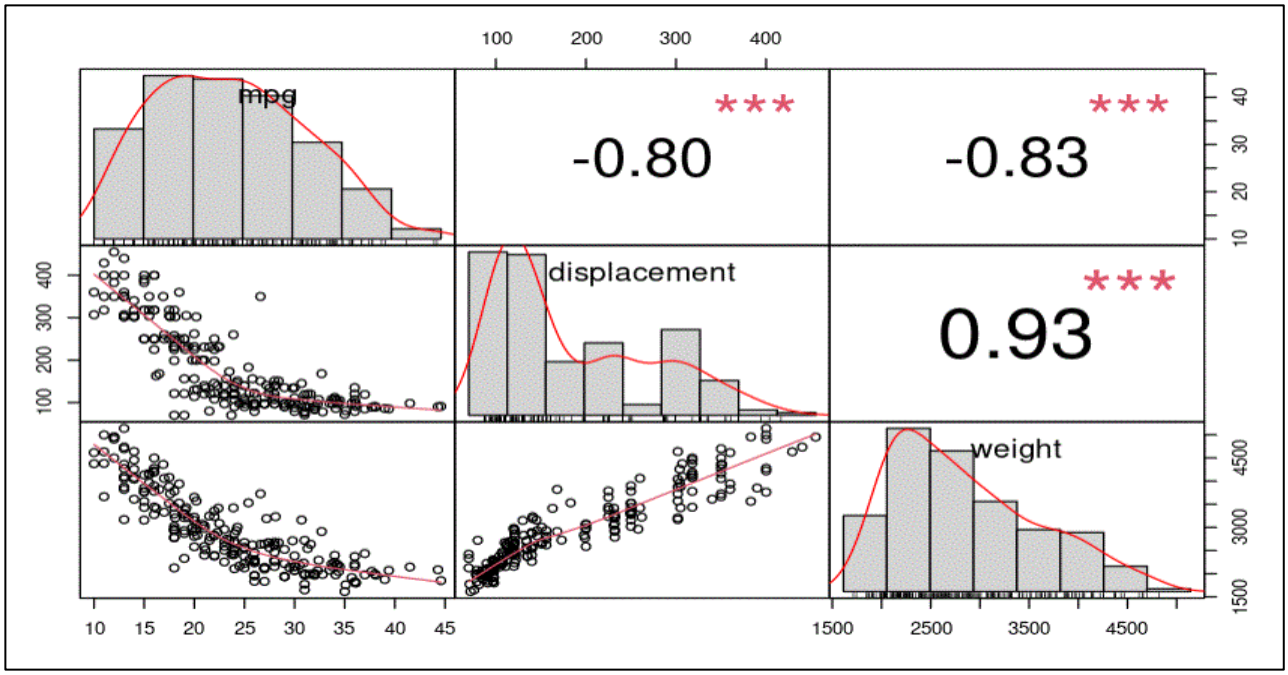
```{r}
install.packages("corrplot")
library(corrplot)
mondata_Correle <- mondata[, c(1,2,3)]

corrplot(Affichage_mondata_Correle, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```




Ici, les corrélations positives sont représentées en bleus alors que les corrélations négatives sont représentées en rouges. L'intensité de leur couleur représente la présence de fortes corrélations entres les valeurs respectives. La légende de droite fait référence aux différents coefficients et les couleurs qu'ils adoptent.

**Production d'un graphique :**

```{r}
install.packages("PerformanceAnalytics")
library("PerformanceAnalytics")


mondata_Correle <- mondata[, c(1,2,3)]
chart.Correlation(mondata_Correle <- mondata[, c(1,2,3)]
, histogram=TRUE, pch=19)
```




Sur la diagonale, on trouve la distribution de chaque variable. Ce point sera d'ailleurs représenté plus en détail dans le rapport. La partie qui se trouvent en-dessous de la diagonale principale, soit la partie qui nous intéresse montrent sous forme de graphique la relation entre deux variables respectives. La distance très rapprochée des multiples points représente l'intensité de corrélation.

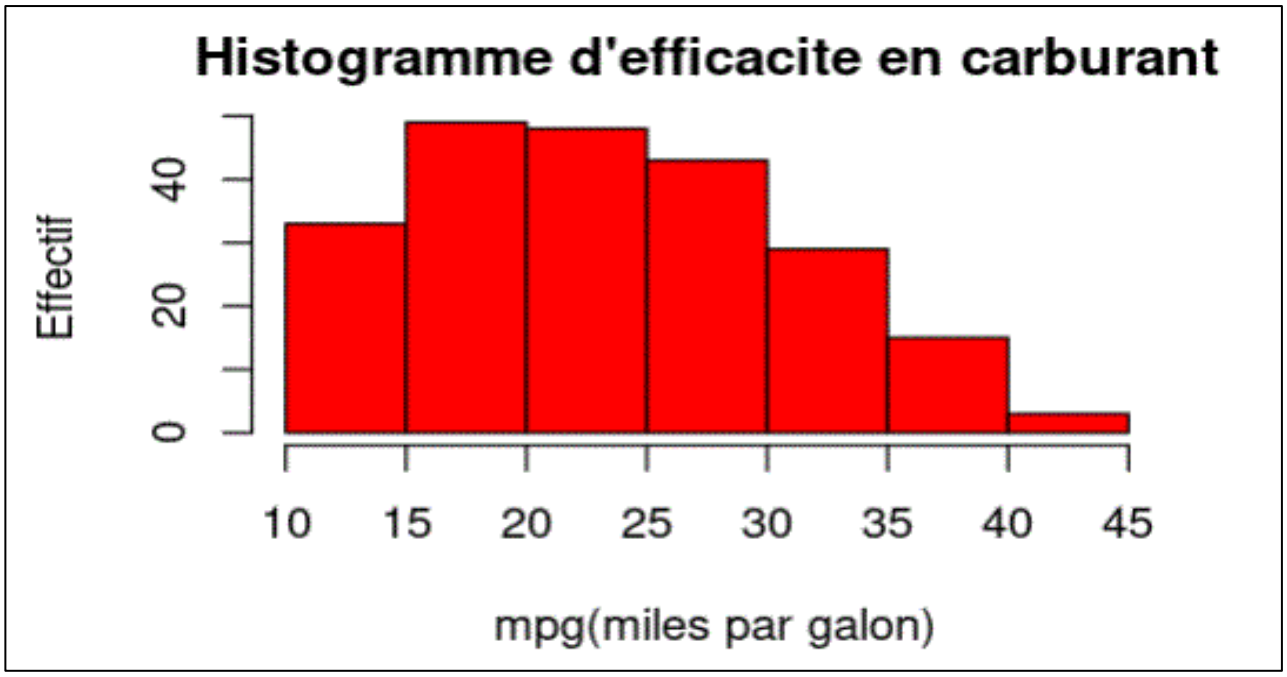
## 1.b) Création des différents diagrammes

**Efficacité en carburant**

***(Représenter sous la variable Y dans le code)***

**Traçage de l'histogramme d'efficacité en carburant :**

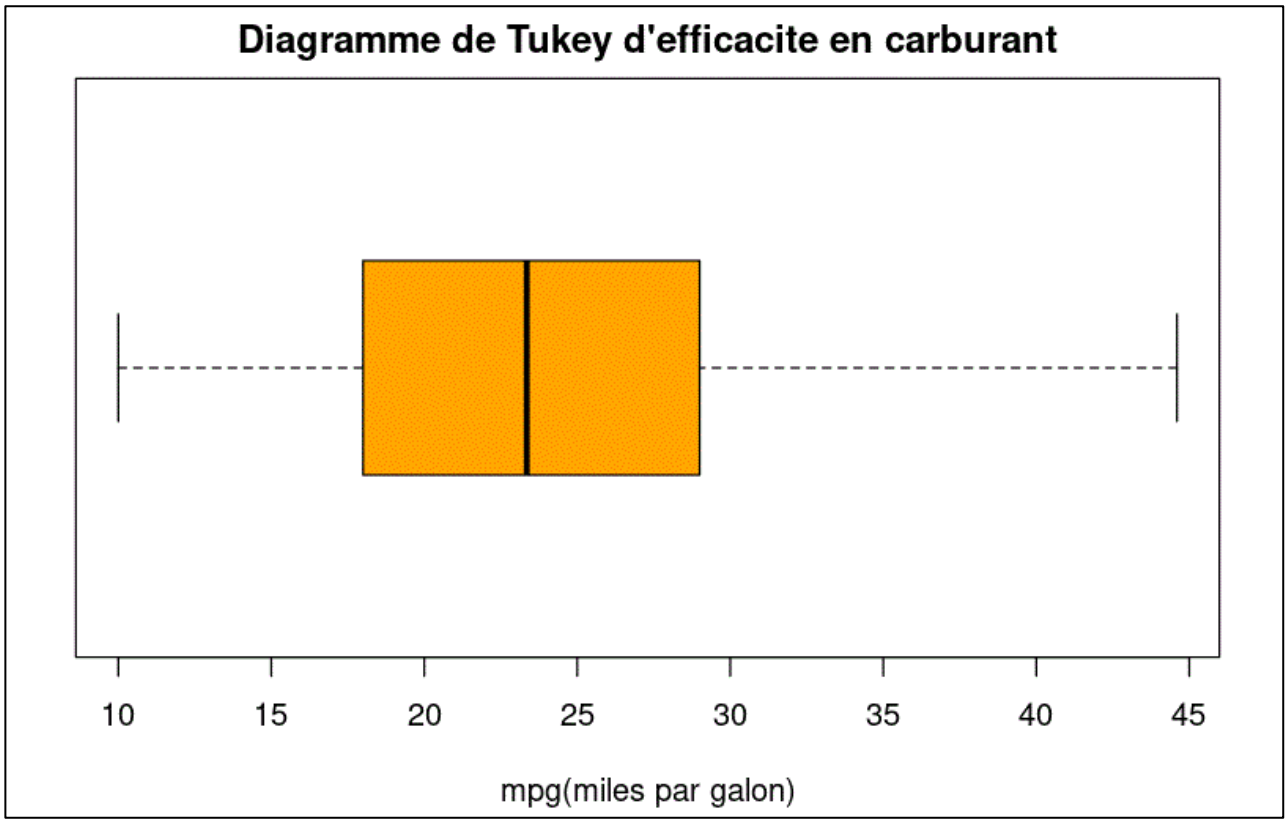
```{r}
Y <- mondata$mpg
hist(Y, col = "red", border = "black",
    main = "Histogramme d'efficacite en carburant",
    xlab = " mpg(miles par galon)",
    ylab = "Effectif")
```




La répartition la plus fréquente se fait autour des intervalles [15,20[, suivi de l'intervalle [20,25[. Il peut donc être déduit que l'efficacité en carburant en mpg se fait plus fréquemment autour de ces intervalles. Mise à part ça, ce diagramme a une distribution similaire à une distribution gaussienne. Cependant, puisque la distribution n'est pas centrée, mais plutôt asymétrique vers la gauche (la grande majorité des valeurs se trouvent vers cette direction), quiconque peut penser que la distribution n'est pas normale. Pour confirmer l'aspect visuel avec certitude, d'autres test seront effectué.

**Tracage du diagramme de Tukey d'efficacité en carburant :**

```{r}
boxplot(Y, main = "Diagramme de Tukey d'efficacite en carburant",
        horizontal = TRUE,
        col = "orange", xlab = "mpg(miles par galon)")
```



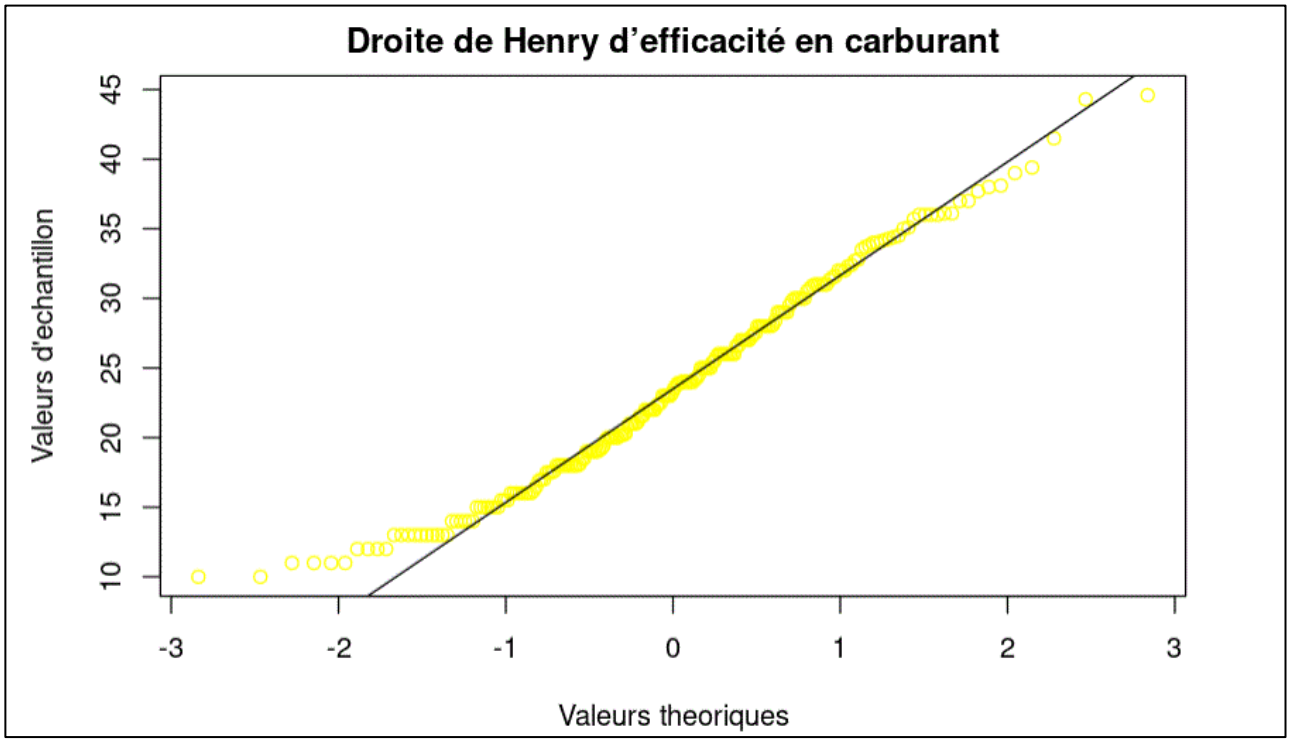

Le diagramme de Tukey apporte plus de clarté sur certaines statistiques et sur la répartition de la distribution. Mise a part le fait que la grande majorité des données se trouvent approximativement dans l'intervalle [18,29[, le diagramme de Tukey amène une clarification quant aux différents quartiles, à la médiane ainsi que les valeurs maximales et minimales.

- Q1 : ~ 18
- Médiane : ~ 23
- Q3 : ~ 29
- Valeur minimum : ~ 10
- Valeur maximum : ~ 45

Le diagramme de Tukey confirme le fait que la distribution se fait vers la gauche.

**Traçage de la droite de Henry d'efficacité en carburant:**

```{r}
qqnorm(Y, col = "yellow", main = "Droite de Henry d'efficacité en carburant", xlab = "Valeurs theoriques", ylab = "Valeurs d'echantillon")

qqline(Y, col = "Black", main = "Droite de Henry", xlab = "Valeurs theoriques", ylab = "Valeurs d'echantillon")

```




**Test de normalite (Shapiro-Milk) d'efficacite en carburant :**

```{r}
shapiro.test(Y)
```

# Shapiro-Wilk normality test
#
# data:  Y
# W = 0.97824, p-value = 0.001781

La droite de Henry et le test de Shapiro-Wilk permettent tous les deux de tester la normalité de la distribution de l'échantillon. En commençant par l'interprétation de la droite de Henry, suivi du test de Shapiro-Wilk, on en déduit que :

- Le nuage de points qui est aligné avec la droite indique qu'une grande majorité des points suit une distribution normale. Quiconque peut donc avoir une certaine idée sur la normalité de la distribution de l'échantillon en analysant cette droite. Quant au graphique, il est possible de voir que les points qui sont alignés avec la droite se trouvent dans l'intervalle [-1.5, 1.5]. Pour les points se trouvant sur la borne inférieure, soit les points qui ne respectent pas nécessairement la tendance de la courbe (ceux se trouvant en-dessus de celle-ci) convergent vers ce derrière, et pour les points se trouvant sur la borne supérieur (en-dessous de la courbe), une manifestation pareille est visible. Cependant, pour tester la normalité de la distribution de manière rigoureuse, il faut à tout prix passer par le test de Shapiro-Wilk.

- Le test de Shapiro-Wilk a pour but de tester la distribution normale de l'échantillon avec le calcul de la P-Value (PV). Pour tester la normalité de l'échantillon, une comparaison entre cette dernière et la valeur de alpha (0,05) est nécessaire.

  Si PV < alpha : Rejet de l'hypothèse selon laquelle la distribution de l'échantillon est normale

  Si PV > alpha : Acceptation de l'hypothèse. Cependant, la valeur de PV n'indique rien sur la distribution de l'échantillon.

- Dans notre cas, puisque PV = 0.001781 < 0.05, on rejette l'hypothèse que la distribution suit une loi normale. Certes, a premier vu, que ce soit par le biais de la distribution à bande ou encore par le fait qu'une grande majorité des points sont sur la droite de Henry mène à conclure que la distribution est normale. Cependant, puisque ce dernier a échoué le test de Shapiro-Wilk, soit l'étape décisive, on conclut que la distribution n'est pas normale.

**Table de statistiques descriptives d'efficacité en carburant :**

Pour ce qui suit ci-dessous, IC correspond à l'intervalle de confiance a 95% et n correspond à la taille de l'échantillon, soit 220. Pour calculer l'intervalle de confiance, le calcul des bornes, soit L et U est nécessaire et leur formule sera décrit sous bas :

**L = X(bar) – (Zα/2*σ) / √ n**
**U = X(bar) + (Zα/2*σ) / √ n**

```{r}
Tab1 <- data.frame (

#Moyenne
  Moyenne = mean(Y),

#Premier Quartile
  Q1 = quantile(Y, 0.25),

#Troisieme Quartile
  Q3 = quantile(Y, 0.75),

#Ecart-Type
  ET = sd(Y),

#Erreur-type
  ERT = (sd(Y) / sqrt(220)),

  IC = paste("[", toString(c((mean(Y) - qnorm(0.975)*sd(Y)/sqrt(220)))),
             (mean(Y) + qnorm(0.975)*sd(Y)/sqrt(220)), "]")
  )
  Tab1


```

| # | Moyenne | Q1 | Q3 | ET | ERT | IC |
|---|---|---|---|---|---|---|
| # | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <chr> |
| # 25% | 23.70545 | 18 | 29 | 7.583915 | 0.5113075 | [22.703310346708 24.7075987442011] |

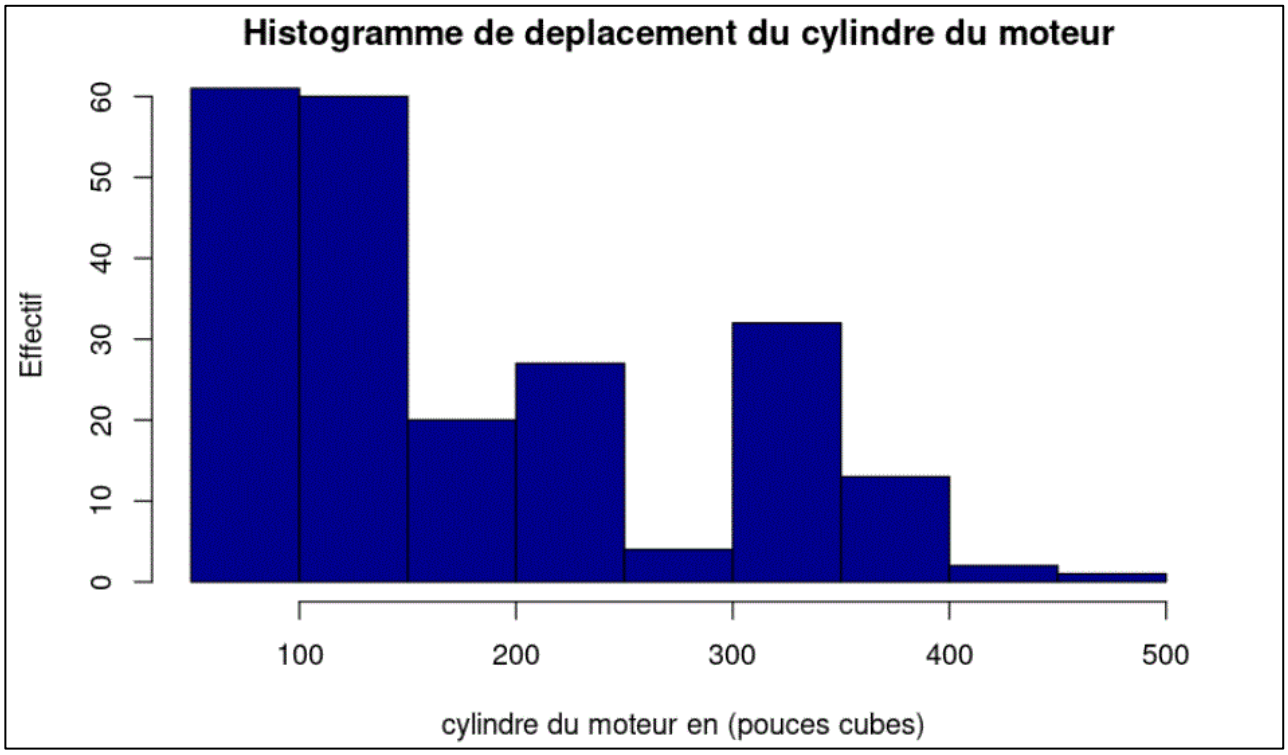
Ou << Moyenne >> représente la moyenne, << Q1 >>, le premier quartile, << Q3 >>, le troisième quartile, << ET >>, l'erreur-type, << ERT >>, l'erreur type et << IC>>, l'intervalle de confiance

## La cylindrée du moteur

## (*Représenter sous la variable X1 dans le code*)

**Traçage de l'histogramme de déplacement du cylindre du moteur en pouces cubes :**

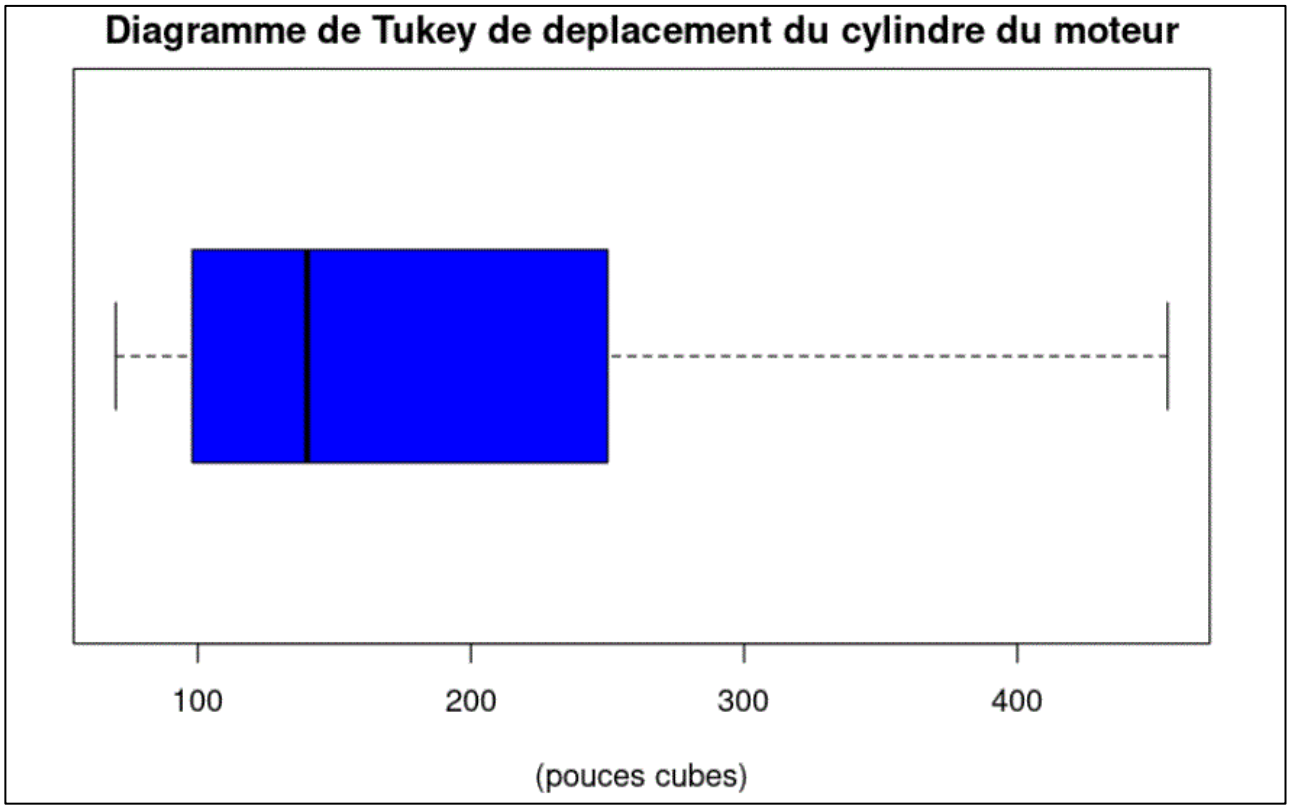
```r
X1 <- mondata$displacement
hist(X1, col = "dark blue", border = "black",
     main = "Histogramme de deplacement du cylindre du moteur",
     xlab = "cylindre du moteur en (pouces cubes)",
     ylab = "Effectif"
)

```




La répartition la plus fréquente se fait autour des intervalles [0, 100[, suivi de l'intervalle [100,150[. [. Il peut donc être déduit que le déplacement du cylindre du moteur se fait plus fréquemment autour de ces intervalles. Mise à part ça, ce diagramme a une distribution différente de celle d'une distribution gaussienne. En effet, la distribution n'est pas centrée, mais plutôt asymétrique vers la gauche (la grande majorité des valeurs se trouvent vers cette direction) et de plus, elle n'a pas la forme d'une cloche. Pour confirmer l'aspect visuel avec certitude, d'autres test seront effectué.

**Traçage du diagramme de Turkey d'efficacité de déplacement du cylindre du moteur :**

```{r}
boxplot(X1, main = "Diagramme de Tukey de deplacement du cylindre du moteur",
        horizontal = TRUE,
        col = "blue", xlab = "(pouces cubes)")

```



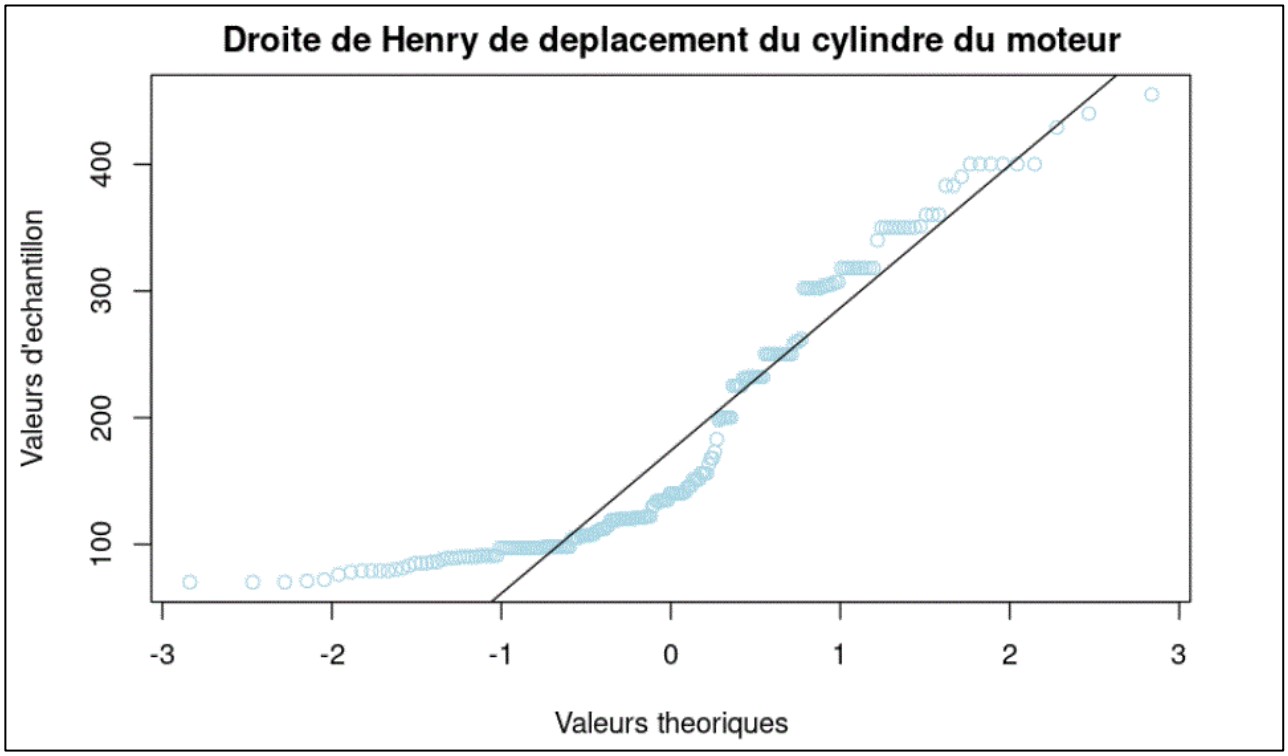

Le diagramme de Turkey apporte plus de clarté sur certaines statistiques et sur la répartition de la distribution. Mise à part le fait que la grande majorité des données se trouvent approximativement dans l'intervalle [100, 250[, le diagramme de Tukey amène une clarification quant aux différents quartiles, a la médiane ainsi que les valeurs maximales et minimales.

- Q1 ~ 100
- Médiane : ~ 125
- Q3 ~ 250
- Valeur minimum : ~ 75
- Valeur maximum : ~ 450

Le diagramme de Tukey confirme le fait que la distribution se fait vers la gauche.

**Traçage de la droite de Henry de déplacement du cylindre du moteur :**

```{r}
qqnorm(X1, col = "light blue", main = "Droite de Henry de deplacement du cylindre du moteur", xlab = "Valeurs theoriques", ylab = "Valeurs d'echantillon")

qqline(X1, col = "Black", main = "Droite de Henry", xlab = "Valeurs theoriques", ylab = "Valeurs d'echantillon")
```




**Test de normalité (Shapiro-Milk) de déplacement du cylindre du moteur :**

```{r}
shapiro.test(X1)
```

# Shapiro-Wilk normality test
#
# data:  X1
# W = 0.86362, p-value = 4.1e-13

La droite de Henry et le test de Shapiro-Wilk permettent tous les deux de tester la normalité de la distribution de l'échantillon. En commençant par l'interprétation de la droite de Henry, suivi du test de Shapiro-Wilk, on en déduit que :

- Le nuage de points qui est aligné avec la droite indique que seuls un nombre faible de points suit tous une distribution normale. Quiconque peut donc avoir une certaine idée sur la normalité de la distribution de l'échantillon en analysant cette droite. Quant au graphique, il est possible de voir que les points qui sont alignés avec la droite se trouvent dans l'intervalle [0.5, 1.0]. Pour les points se trouvant sur la borne inférieure, soit les points qui ne respectent pas nécessairement la tendance de la courbe (ceux se trouvant en-dessus de celle-ci) convergent vers ce derrière, et pour les points se trouvant sur la borne supérieur (en-dessous de la courbe), une manifestation pareille est visible. Cependant, pour tester la normalité de la distribution de manière rigoureuse, il faut à tout prix passer par le test de Shapiro-Wilk.

- Le test de Shapiro-Wilk a pour but de tester la distribution normale de l'échantillon avec le calcul de la P-Value (PV). Pour tester la normalité de l'échantillon, une comparaison entre cette dernière et la valeur de alpha (0,05) est nécessaire.

  Si PV < alpha : Rejet de l'hypothèse selon laquelle la distribution de l'échantillon est normale

  Si > alpha : Acceptation de l'hypothèse. Cependant, la valeur de PV n'indique rien sur la distribution de l'échantillon.

- Dans notre cas, puisque PV = 4.1e-13 < 0.05, on rejette l'hypothèse. Il peut donc être conclut avec certitude que la distribution n'est pas normale. Cela vient d'ailleurs confirmer l'aspect visuel du diagramme a bande (qui ne ressemble en rien à une distribution normale) et celle de la droite de Henry ou il est possible de voir que seul un faible nombre de point respecte la tendance normale de la droite.

**Table de statistiques descriptives de déplacement du cylindre du moteur :**

Pour ce qui suit ci-dessous, IC correspond à l'intervalle de confiance a 95% et n correspond à la taille de l'échantillon, soit 220. Pour calculer l'intervalle de confiance, le calcul des bornes, soit L et U est nécessaire et leur formule sera décrit sous bas :

**L = X(bar) – (Zα/2*σ) / √ n**
**U = X(bar) + (Zα/2*σ) / √ n**

```{r}
Tab2 <- data.frame (

#Moyenne
  Moyenne = mean(X1),

#Premier Quartile
  Q1 = quantile(X1, 0.25),

#Troisieme Quartile
  Q3 = quantile(X1, 0.75),

#Ecart-Type
  ET = sd(X1),

#Erreur-type
  ERT = (sd(X1) / sqrt(220)),

  IC = paste("[", toString(c((mean(X1) - qnorm(0.975)*sd(X1)/sqrt(220)))),
             (mean(X1) + qnorm(0.975)*sd(X1)/sqrt(220)), "]")
  )
  Tab2


```

| # | Moyenne | Q1 | Q3 | ET | ERT | IC |
|---|---|---|---|---|---|---|
| # | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <chr> |
| # 25% | 183.6341 | 98 | 250 | 100.8309 | 6.798015 | [170.310226128024 196.957955690158 ] |

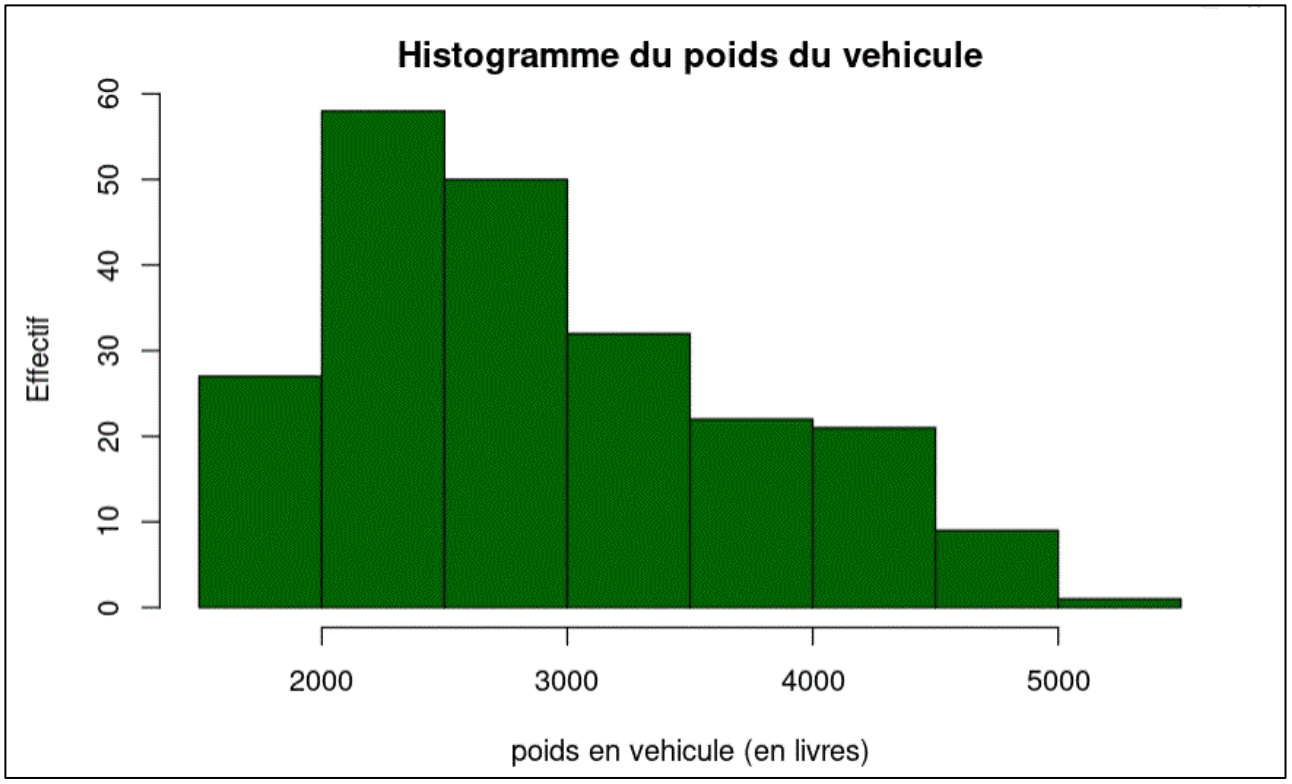
Ou << Moyenne >> représente la moyenne, << Q1 >>, le premier quartile, << Q3 >>, le troisième quartile, << ET >>, l'erreur-type, << ERT >>, l'erreur type et << IC>>, l'intervalle de confiance

## **Le poids du véhicule**

## ***(Représenter sous la variable X2 dans le code)***

**Traçage de l'histogramme du poids du véhicules :**

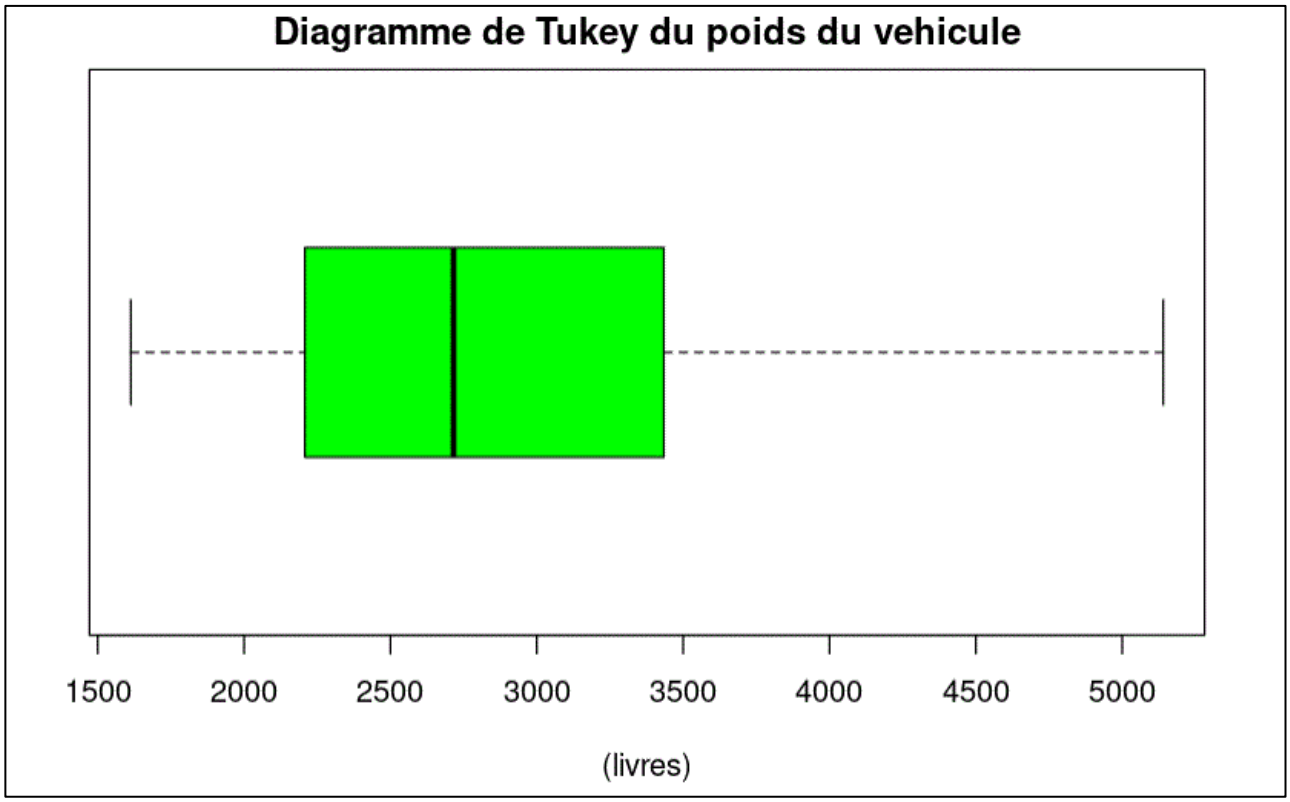
```r
X2 <- mondata$weight
hist(X2, col = "dark green", border = "black",
    main = "Histogramme du poids du vehicule",
    xlab = "poids en vehicule (en livres)",
    ylab = "Effectif"
)
```




La répartition la plus fréquente se fait autour des intervalles [2000, 2500[, suivi de l'intervalle [2500,3000[. Il peut donc être déduit que le poids du véhicule se fait plus fréquemment autour de ces intervalles. Mise à part ça, ce diagramme a une distribution similaire à une distribution gaussienne. Cependant, puisque la distribution n'est pas centrée, mais plutôt asymétrique vers la gauche (la grande majorité des valeurs se trouvent vers cette direction), quiconque peut penser que la distribution n'est pas normale. Pour confirmer l'aspect visuel avec certitude, d'autres test seront effectué.

**Traçage du diagramme de Turkey du poids du véhicule :**

```{r}
boxplot(X2, main = "Diagramme de Tukey du poids du vehicule",
        horizontal = TRUE,
        col = "green", xlab = "(livres)")

```



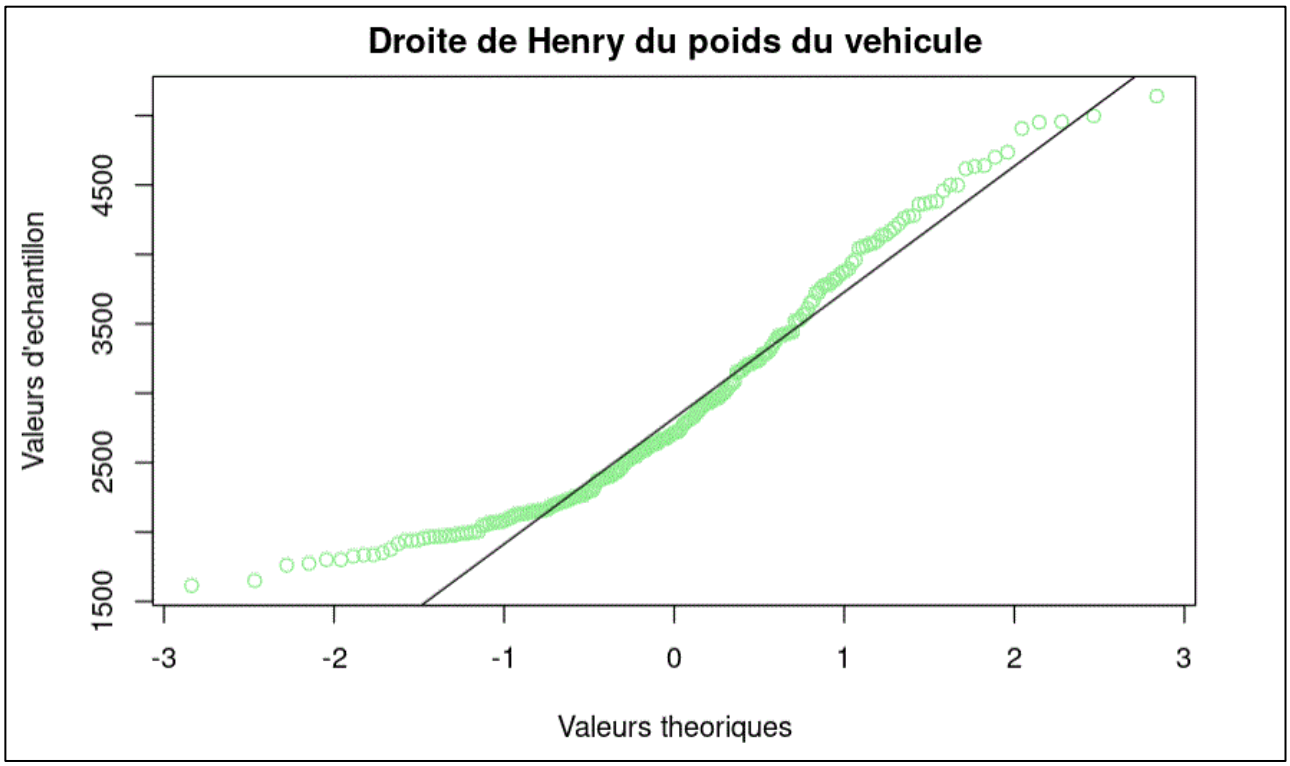

Le diagramme de Turkey apporte plus de clarté sur certaines statistiques et sur la répartition de la distribution. Mise à part le fait que la grande majorité des données se trouvent approximativement dans l'intervalle [2250, 3500[, le diagramme de Tukey amène une clarification quant aux différents quartiles, a la médiane ainsi que les valeurs maximales et minimales.

- Q1 ~ 2250
- Médiane : ~ 2700
- Q3 ~ 3400
- Valeur minimum : ~ 1600
- Valeur maximum : ~ 5250

Le diagramme de Tukey confirme le fait que la distribution se fait vers la gauche.

**Traçage de la droite de Henry du poids du véhicule :**

```{r}
qqnorm(X2, col = "light green", main = "Droite de Henry du poids du vehicule", xlab = "Valeurs theoriques", ylab = "Valeurs d'echantillon")

qqline(X2, col = "Black", main = "Droite de Henry", xlab = "Valeurs theoriques", ylab = "Valeurs d'echantillon")
```




Test de normalité (Shapiro-Milk) du poids du véhicule

```{r}
shapiro.test(X2)
```

# Shapiro-Wilk normality test

# data:  X2
# W = 0.93645, p-value = 3.479e-08

La droite de Henry et le test de Shapiro-Wilk permettent tous les deux de tester la normalité de la distribution de l'échantillon. En commençant par l'interprétation de la droite de Henry, suivi du test de Shapiro-Wilk, on en déduit que :

- Le nuage de points qui est aligné avec la droite indique que plusieurs points suivent une distribution normale. Quiconque peut donc avoir une certaine idée sur la normalité de la distribution de l'échantillon en analysant cette droite. Quant au graphique, il est possible de voir que les points qui sont alignés avec la droite se trouvent dans l'intervalle [-1.0, 1.0]. Pour les points se trouvant sur la borne inférieure, soit les points qui ne respectent pas nécessairement la tendance de la courbe (ceux se trouvant en-dessus de celle-ci) convergent vers ce derrière, et pour les points se trouvant sur la borne supérieur (en-dessous de la courbe), une manifestation pareille est visible. Cependant, pour tester la normalité de la distribution de manière rigoureuse, il faut à tout prix passer par le test de Shapiro-Wilk.

- Le test de Shapiro-Wilk a pour but de tester la distribution normale de l'échantillon avec le calcul de la P-Value (PV). Pour tester la normalité de l'échantillon, une comparaison entre cette dernière et la valeur de alpha (0,05) est nécessaire.

  Si PV < alpha : Rejet de l'hypothèse selon laquelle la distribution de l'échantillon est normale

  Si > alpha : Acceptation de l'hypothèse. Cependant, la valeur de PV n'indique rien sur la distribution de l'échantillon.

- Dans notre cas, puisque PV = 3.479e-08 < 0.05, on rejette l'hypothèse que la distribution suit une loi normale. Comme pour la première distribution, le diagramme à bande mène à croire que la distribution est normale, alors que cette hypothèse ne se manifeste pas pour autant une fois avoir réalisé la droite de Henry. Cependant, l'étape brise-glace, soit le test de Shapiro-Milk conclut que la distribution n'est pas normale.

**Table de statistiques descriptives du poids du véhicule :**

Pour ce qui suit ci-dessous, IC correspond à l'intervalle de confiance a 95% et n correspond à la taille de l'échantillon, soit 220. Pour calculer l'intervalle de confiance, le calcul des bornes, soit L et U est nécessaire et leur formule sera décrit sous bas :

**L = X(bar) – (Zα/2*σ) / √ n**
**U = X(bar) + (Zα/2*σ) / √ n**

```{r}
Tab3 <- data.frame (

#Moyenne
  Moyenne = mean(X2),

#Premier Quartile
  Q1 = quantile(X2, 0.25),

#Troisieme Quartile
  Q3 = quantile(X2, 0.75),

#Ecart-Type
  ET = sd(X2),

#Erreur-type
  ERT = (sd(X2) / sqrt(220)),

  IC = paste("[", toString(c((mean(X2) - qnorm(0.975)*sd(X2)/sqrt(220)))),
             (mean(X2) + qnorm(0.975)*sd(X2)/sqrt(220)), "]")
  )
  Tab3


```

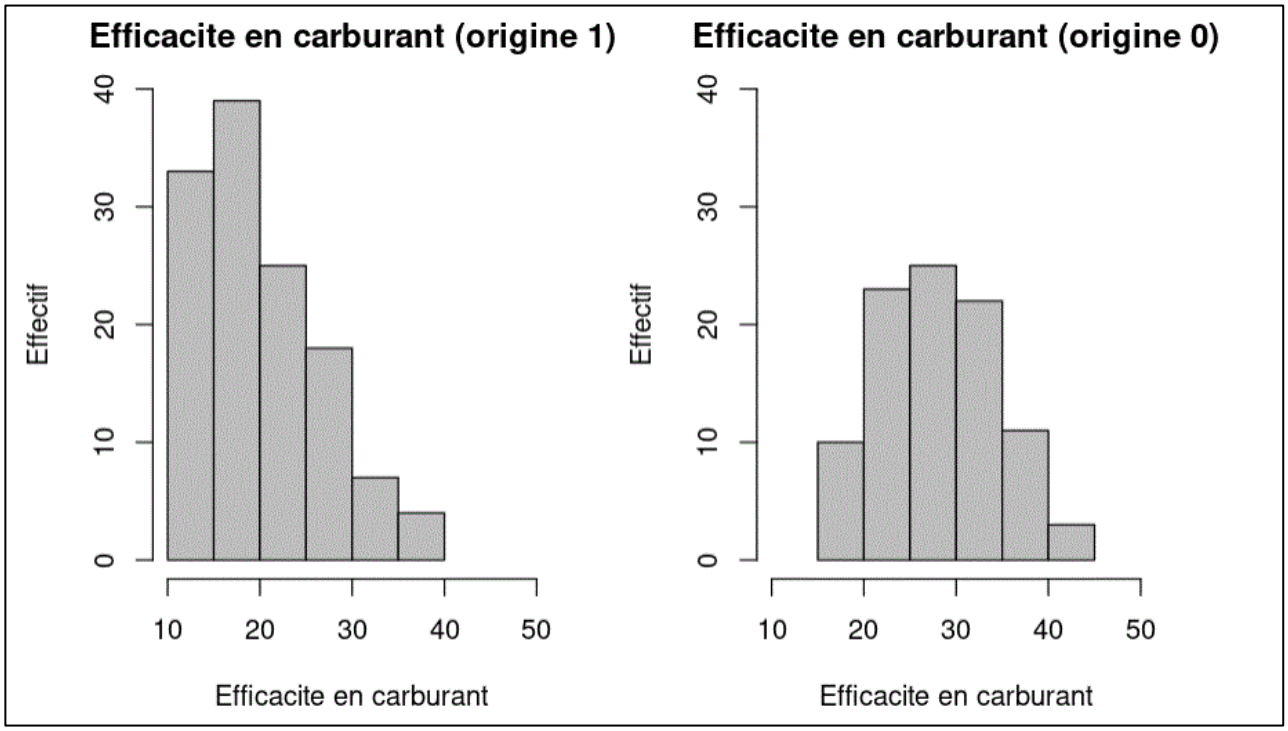
| # | Moyenne | Q1 | Q3 | ET | ERT | IC |
|---|---|---|---|---|---|---|
| # | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <chr> |
| # 25% | 2912.1 | 2208.75 | 3433 | 835.2034 | 56.3094 | [2801.73560050381 3022.46439949619] |

Ou << Moyenne >> représente la moyenne, << Q1 >>, le premier quartile, << Q3 >>, le troisième quartile, << ET >>, l'erreur-type, << ERT >>, l'erreur type et << IC>>, l'intervalle de confiance

## 1.c) L’efficacité en carburant d’un véhicule selon le code d’origine

**Histogramme juxtapose d'efficacité en carburant selon le code d'origine :**

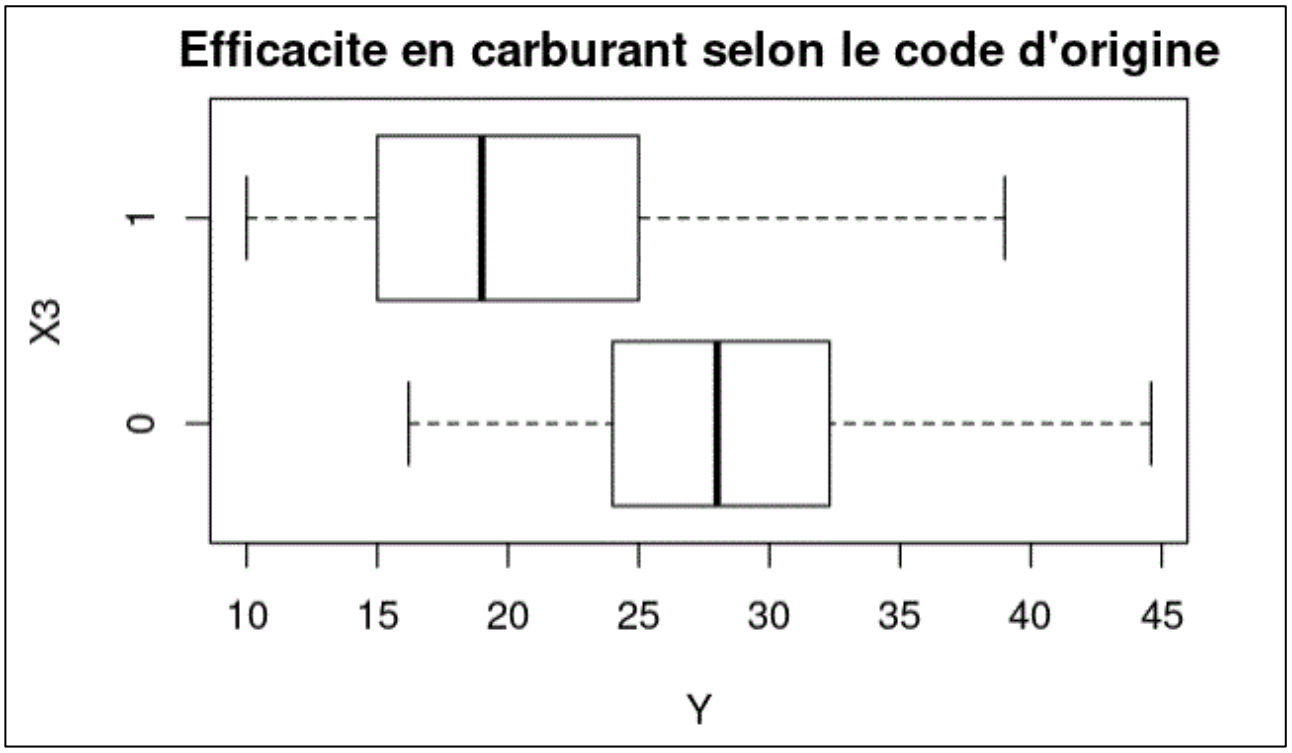
```{r}
Y <- mondata$mpg
X3 <- mondata$origin
split.screen(1:2)

screen(2); hist(Y[X3==0], col = "grey", main = "Efficacite en carburant (origine 0)", border = "black", xlab = "Efficacite en carburant", ylab = "Effectif", xlim=c(10,50), ylim=c(0,40))

screen(1); hist(Y[X3==1], col = "grey", main = "Efficacite en carburant (origine 1)", border = "black", xlab = "Efficacite en carburant", ylab = "Effectif", xlim=c(10,50), ylim=c(0,40))

```




En comparant ces deux histogrammes cote à cote, il peut être dit que l'efficacité en carburant pour le code d'origine 0 a tendance à suivre une distribution en cloche plus uniforme que pour la région 1. Pour le premier graphique, soit << Efficacité en carburant (origine 1) >>, les valeurs les plus fréquentes se trouvent dans l'intervalle [15,20[suivi de l'intervalle [10,15[. Pour ce qui est du deuxième graphique, soit << Efficacité en carburant (origine 0) >>, la répartition est plus uniforme et les valeurs les plus fréquentes se trouvent dans l'intervalle [25,30[suivi de l'intervalle [20,25[. La répartition est beaucoup plus écartée pour le graphique 1 que pour le graphique 0. Par ailleurs, la distribution pour le graphique reflétant l'origine 0 se fait un peu plus vers la gauche tandis que celle reflétant l'origine 1 se fait légèrement plus vers la droite.

**Diagrammes de Tukey juxtapose reflétant l'efficacité en carburant selon le code d'origine :**

```{r}
boxplot(Y~X3, main = "Efficacite en carburant selon le code d'origine", horizontal= TRUE, col = "white")
```




*Dans le graphique ci-dessus, << Y >> représente l'efficacité en carburant du véhicule (en milles par gallon) et << X3 >> représente le code d'origine du véhicule, et peut prendre deux valeurs, soit 0 ou 1.*

Les diagrammes de Tukey amènent des informations plus détaillées sur les statistiques descriptives et il est possible de voir des différences majeures. Ces différences entre les deux seront listés en bas.

<u>Pour ce qui est du code d'origine 1 :</u>
- Q1 ~ 14
- Médiane : ~ 18
- Q3 ~ 25
- Valeur minimum : ~ 10
- Valeur maximum : ~ 39

<u>Pour ce qui est du code d'origine 0 :</u>
- Q1 ~ 24
- Médiane : ~ 28
- Q3 ~ 33
- Valeur minimum : ~ 16
- Valeur maximum : ~ 45

**Tableau des statistiques descriptives par groupe :**

**Origine 1**

```{r}
TabOrig1 <- data.frame(
MoyenneOrig1 = mean(Y[X3==1]),
EcartTypeOrig1 = sd(Y[X3==1]),
VarianceOrig1 = (sd(Y[X3==1]))^2,
Q1Orig1 = quantile(Y[X3==1], .25),
Q3Orig1 = quantile(Y[X3==1], .75),
IntervalleConfianceOrig1 = paste("[",toString(c(mean(Y[X3==1]) -
qnorm(.975)*(sd(Y[X3==1])/sqrt(length(Y[X3==1]))) , mean(Y[X3==1]) +
qnorm(.975)*(sd(Y[X3==1])/sqrt(length(Y[X3==1])))),"]")
)

TabOrig1
```

| # | MoyenneOrig1 | EcartTypeOrig1 | VarianceOrig1 | Q1Orig1 |
|---|---|---|---|---|
| # | <dbl> | <dbl> | <dbl> | <dbl> |
| # 25% | 20.30317 | 6.682119 | 44.65071 | 15 |

| # | Q3Orig1 | IntervalleConfianceOrig1 |
|---|---|---|
| # | <dbl> | <dbl> |
| # 25% | 24.875 | [19.1364267720863, 21.4699224342629] |

**Origine 0**

```{r}
TabOrig0 <- data.frame(
MoyenneOrig0 = mean(Y[X3==0]),
EcartTypeOrig0 = sd(Y[X3==0]),
VarianceOrig0 = (sd(Y[X3==0]))^2,
Q1Orig0 = quantile(Y[X3==0], .25),
Q3Orig0 = quantile(Y[X3==0], .75),
IntervalleConfianceOrig0 = paste("[",toString(c(mean(Y[X3==0]) -
qnorm(.975)*(sd(Y[X3==0])/sqrt(length(Y[X3==0]))) , mean(Y[X3==0]) +
qnorm(.975)*(sd(Y[X3==0])/sqrt(length(Y[X3==0])))),"]")
)

TabOrig0
```

| # | MoyenneOrig0 | EcartTypeOrig0 | VarianceOrig0 | Q1Orig0 |
|---|---|---|---|---|
| # | <dbl> | <dbl> | <dbl> | <dbl> |
| # 25% | 28.26596 | 6.222636 | 38.72119 | 24 |

| # | Q3Orig0 | IntervalleConfianceOrig0 |
|---|---|---|
| # | <dbl> | <dbl> |
| # 25% | 32.225 | [27.008021476358, 29.523893417259] |

**Test d'hypothèse de Student sur l'égalité des variances de l'efficacité en carburant du véhicule lorsque le code d'origine du véhicule est égal a 1 et lorsqu'il est égal a 0 :**

On pose H0 : les variances sont égales et on pose H1: les variances ne sont pas égales. La valeur de alpha est égale à 0.05. Ce test permet entres autres de tester H0 tel que H0 : les variances suivent une loi normale et sont égales. Pour ce test, l'hypothèse sera rejetée si la P-value obtenue est inférieur à 0.05.

```{r}
var.test(Y[X3==0], Y[X3==1])
```

```
# F test to compare two variances
#
# data:  Y[X3 == 0] and Y[X3 == 1]
# F = 0.8672, num df = 93, denom df = 125, p-value = 0.4705
# alternative hypothesis: true ratio of variances is not equal to 1
# 95 percent confidence interval:
# 0.5952851 1.2770684
# sample estimates:
# ratio of variances
#         0.8672022
```

PV = 0.4705 ce qui est supérieur à alpha = 0.05. Il peut donc être conclu que les deux codes origines du véhicule suivent une loi normale de variances similaires.

**Test d'hypothèse de Student sur l'égalité des moyennes de l'efficacité en carburant du véhicule lorsque le code d'origine du véhicule est égal a 1 et lorsqu'il est égal a 0 :**

On pose H0 : les moyennes sont égales et on pose H1: les moyennes ne sont pas égales. La valeur de alpha est égale à 0.05.

```{r}
t.test(Y[X3==0], Y[X3==1])

```

```
# Welch Two Sample t-test
#
# data:  Y[X3 == 0] and Y[X3 == 1]
# t = 9.0963, df = 207.56, p-value < 2.2e-16
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  6.236998 9.688568
# sample estimates:
# mean of x mean of y
     28.26596 20.30317
```

PV = < 2.2e-16 ce qui est inférieur à alpha = 0.05. Il peut donc être conclu que les deux codes d'origines du véhicule ne suivent pas une loi normale de moyennes similaires.

Puisque le les deux variances et les moyennes ne sont pas les mêmes, il ne peut pas être conclut que le code d'origine du véhicule a une incidence sur l'efficacité en carburant.

# Phase 2 : Recherche d'un modèle

## 1.d) Analyse de 8 modèles

L'objectif de la partie 2 est de trouver les variables qui expliqueraient le mieux la performance d'un véhicule en fonction de de ces derniers. Pour ce faire, il nous est donné 8 modèles, dont 4 sont en fonction de la cylindrée du moteur du véhicule en pouces cubes (X1), et 4 sont en fonction du poids du véhicule en livres (X2). Dans les pages qui suivent, une analyse plus détaillée des modèles suivants seront effectués, soit :

- **Modèle 1 : $Y = \beta0 + \beta1*X1 + \varepsilon$**
- **Modèle 2 : $Y = \beta0 + \beta1*X1^2 + \varepsilon$**
- **Modèle 3 : $Y = \beta0*X1^{(\beta1)} * e^{(\varepsilon)}$**
- **Modèle 4 : $Y = \beta0*e^{(\beta1*X1+\varepsilon)}$**
- **Modèle 5 : $Y = \beta0 + \beta1*X2 + \varepsilon$**
- **Modèle 6 : $Y = \beta0 + \beta1*X2^2 + \varepsilon$**
- **Modèle 7 : $Y = \beta0*X2^{(\beta1)} * e^{(\varepsilon)}$**
- **Modèle 8 : $Y = \beta0*e^{(\beta1*X2+\varepsilon)}$**

## **Modèle 1 : Y = β0 + β1*X1 + ε**

```{r}
X1 <- mondata$displacement
linModel1 <- lm(Y~X1)
summary(linModel1)
```

```
# Call:
# lm(formula = Y ~ X1)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -12.5337  -2.9113  -0.6399   2.3964  15.3282
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 34.740048   0.641517   54.15   <2e-16 ***
# X1          -0.060090   0.003064  -19.61   <2e-16 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 4.572 on 218 degrees of freedom
# Multiple R-squared:  0.6383,            Adjusted R-squared:  0.6366
# F-statistic: 384.7 on 1 and 218 DF,  p-value: < 2.2e-16
```

```{r}
coef(linModel1)
```

```
# (Intercept)          X1
# 34.74004805 -0.06009011
```

```{r}
anaVariance1 <- anova(linModel1)
anaVariance1
```

```
# Analysis of Variance Table
#
# Response: Y
#           Df Sum Sq Mean Sq F value    Pr(>F)
# X1         1 8039.6  8039.6  384.66 < 2.2e-16 ***
# Residuals 218 4556.3    20.9
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Regardant l'équation du modèle 1, celui-ci correspond à une équation simple de la régression linéaire. Il n'y a donc aucune modification à apporter à cette dernière, mise à part son ajustement. On obtient ainsi le tableau de coefficient de la régression suivante :

**Tableau de coefficient de la régression du Modèle 1**

| | Estimation | Erreur Standard | Probabilité sur le test | Valeur de PV |
|---|---|---|---|---|
| B0 | 34.740048 | 0.641517 | 54.15 | < 2e-16 |
| B1 | -0.060090 | 0.003064 | -19.61 | < 2e-16 |

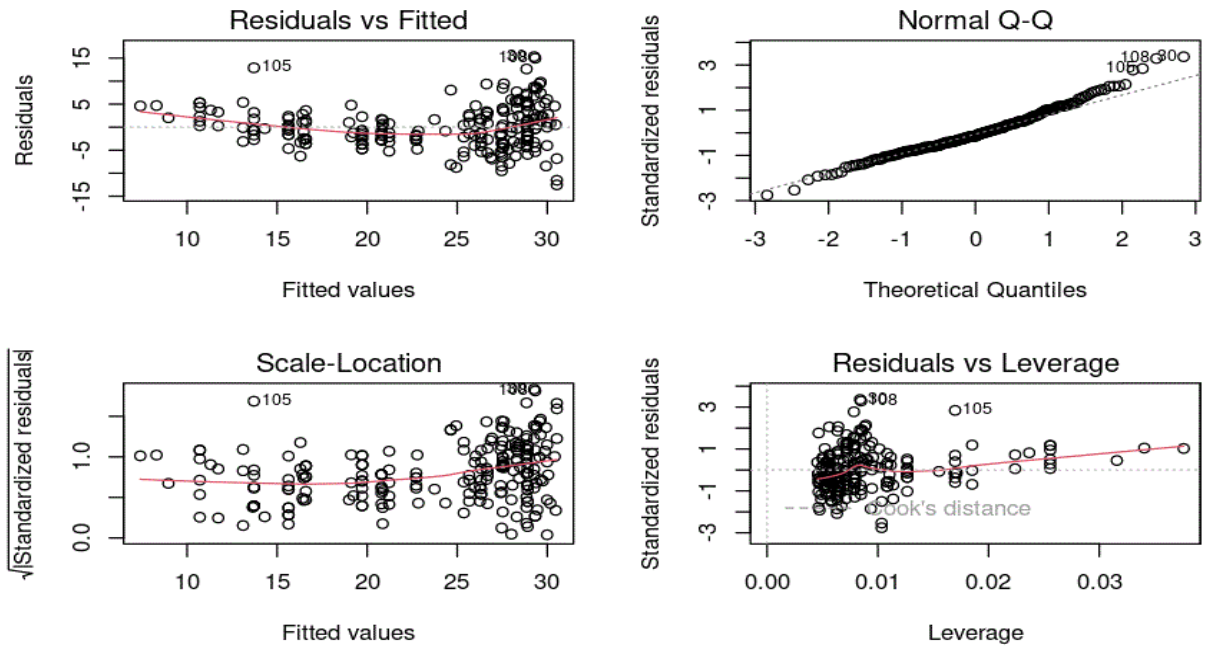
On obtient également le tableau d'analyse de la variance suivante :

**Tableau de d'analyse de la variance du Modèle 1**

| | Degrés de liberté (Df) | Somme des carrés (Sum Sq) | Moyenne des carrés (Mean Sq) | Valeur de F0 (F Value) |
|---|---|---|---|---|
| Régression X1 | 1 | 8039.6 | 8039.6 | 384.66 |
| Erreurs (Résiduals) | 218 | 4556.3 | 20.9 | N/A |

Test de signification du modèle :

```r
par(mfrow = c(2,2))
plot(linModel1)
```



```{r}
residus1 <- rstudent(linModel1)
shapiro.test(residus1)
```

```
# Shapiro-Wilk normality test
#
# data:  residus1
# W = 0.98007, p-value = 0.003352
```

Le test de signification du modèle consiste à comparer la valeur du PV obtenue lors de l'analyse de la variance à celle de alpha, soit 0.05. La valeur obtenue est de 0.003352 soit un nombre inférieur à 0,05. Il peut donc être affirmer que la variable du modèle est significative.

Pour ce qui est des graphiques, le schéma du haut en présente 4, soit **Residuals vs Fitted, Normal Q-Q, Scale-Location et Residuals vs Leverage** et ces derniers seront élaborés en détails ici-bas.

**Residuals vs Fitted**

*Ce graphique permet d'attester la linéarité des résidus. S'il s'agit d'une tendance linéaire, la courbe en rouge du graphique devrait être droite. Dans la mesure ou la courbe en rouge est droite, il serait possible d'affirmer que les résidus suivent une tendance linéaire. Dans le cas contraire, les résidus ne suivent pas de tendance linéaire.* Dans notre cas, il est possible de voir qu'il ne s'agit pas d'une courbe. En effet, les points ont tendances à former une espèce de courbe vers la fin de la régression éliminant ainsi toute possibilité à avoir une tendance linéaire droite. Il peut donc être conclut que la tendance n'est pas linéaire.

**Normal Q-Q**

*Ce graphique est similaire à celle de la droite de Henry, dans le sens ou les résidus constituent un nuage de points. Si ce nuage de point est aligné avec la droite se trouvant dans le graphique, il est possible de conclure que les résidus suivent une loi normale.* Dans notre cas, il est possible de constater que le nuage de points est assez bien aligné à la droite, ce qui nous permet de de juger de la normalité de la distribution. En utilisant le test de Shapiro-Wilk, 0.003353, soit la p-value trouvé, est inférieur à alpha, 0,05. Il peut donc être dit que la distribution n'est pas normale.

**Scale-Location**

*Ce graphique permet de montrer l'égalité de variances des résidus, communément appelé homoscédasticité. Si la répartition des résidus se fait de manière homogène sur le graphique, on parle alors d'homoscédasticité. Dans le cas contraire, on parle d'hétéroscédasticité.* Pour ce qui de la répartition des éléments dans le graphique, il est possible de voir que les valeurs sont vraiment étalées, ce qui nous suggère une hétéroscédasticité. Certes, les points ne sont pas placés de manière homogène autour de celle-ci.

**Résiduals vs Leverage**

*Ce graphique permet d'identifier les points atypiques qui ont une incidence plus importante que d'autres éléments de la population sur les données. Il sera possible d'identifier les points atypiques en analysant la distance de Cook. Dans le cas où les valeurs se trouvent en dehors de cet encadrement, ces derniers peuvent être considéré comme ayant une incidence importante sur les données.* Dans notre cas, il est possible de voir que peu de points atypiques sont placés à l'extérieur de la droite de la distance de Cook ce qui permet de déduire qu'un nombre faible de points ont une incidence importante sur les données.

## **Modèle 2 : Y = β0 + β1*X1^2 + ε**

```{r}
B02 <- rep(X1)
B12 <- B02^2

linModel2 <- lm(Y~B12)
summary(linModel2)
```

```
# Call:
# lm(formula = Y ~ B12)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -10.5173  -3.4723  -0.9601   2.9676  16.5005
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  2.912e+01  4.649e-01   62.64   <2e-16 ***
# B12         -1.236e-04  7.306e-06  -16.91   <2e-16 ***
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 4.999 on 218 degrees of freedom
# Multiple R-squared:  0.5675,            Adjusted R-squared:  0.5655
# F-statistic: 286.1 on 1 and 218 DF,  p-value: < 2.2e-16
```

```{r}
coef(linModel2)
```

```
#   (Intercept)          B12
# 29.1227403682 -0.0001235635
```

```{r}
anaVariance2 <- anova(linModel2)
anaVariance2
```

```
# Analysis of Variance Table
#
# Response: Y
#           Df Sum Sq Mean Sq F value    Pr(>F)
# B12        1 7148.4  7148.4  286.07 < 2.2e-16 ***
# Residuals 218 5447.5    25.0
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Tableau de coefficient de la régression du Modèle 2**

| | Estimation | Erreur Standard | Probabilité sur le test | Valeur de PV |
|---|---|---|---|---|
| B0 | 2.912e+01 | 4.649e-01 | 62.64 | < 2e-16 |
| B1 | -1.236e-04 | 7.306e-06 | -16.91 | < 2e-16 |

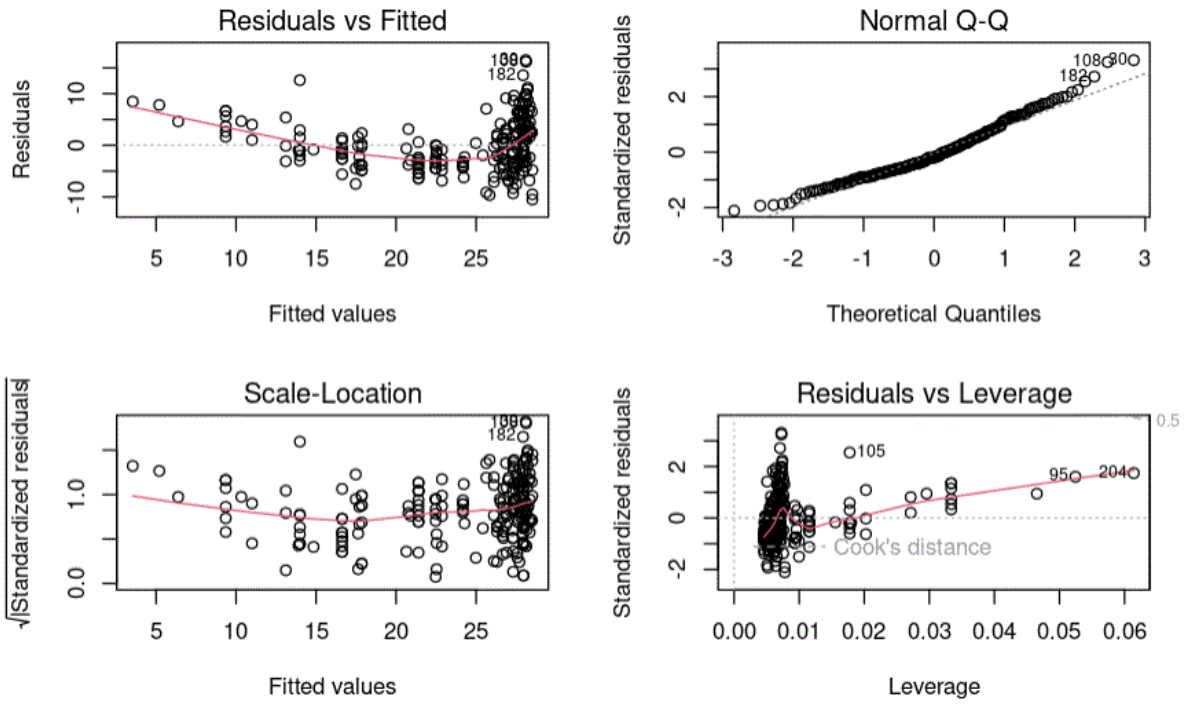
On obtient également le tableau d'analyse de la variance suivante :

**Tableau de d'analyse de la variance du Modèle 2**

| | Degrés de liberté (Df) | Somme des carrés (Sum Sq) | Moyenne des carrés (Mean Sq) | Valeur de F0 (F Value) |
|---|---|---|---|---|
| Régression X1 | 1 | 7148.4 | 7148.4 | 286.07 |
| Erreurs (Résiduals) | 218 | 5447.5 | 25.0 | N/A |

Test de signification du modèle :

```
```{r}
par(mfrow = c(2,2))
plot(linModel2)
```
```



```{r}
residus2 <- rstudent(linModel2)
shapiro.test(residus2)
```

```
#             Shapiro-Wilk normality test
#
# data:  residus2
# W = 0.96897, p-value = 9.385e-05
```

Le test de signification du modèle consiste à comparer la valeur du PV obtenue lors de l'analyse de la variance à celle de alpha, soit 0.05. La valeur obtenue est de 9.385e-05 soit un nombre inférieur à 0,05. Il peut donc être affirmer que la variable du modèle est significative.

Pour ce qui est des graphiques, le schéma du haut en présente 4, soit **Residuals vs Fitted, Normal Q-Q, Scale-Location et Residuals vs Leverage** et ces derniers seront élaborés en détails ici-bas.

**Residuals vs Fitted**

*Ce graphique permet d'attester la linéarité des résidus. S'il s'agit d'une tendance linéaire, la courbe en rouge du graphique devrait être droite. Dans la mesure ou la courbe en rouge est droite, il serait possible d'affirmer que les résidus suivent une tendance linéaire. Dans le cas contraire, les résidus ne suivent pas de tendance linéaire.* Dans notre cas, il est possible de voir qu'il ne s'agit pas d'une courbe comme dans la première situation. En effet, les points ont tendances à former une espèce de courbe vers la fin de la régression éliminant ainsi toute possibilité à avoir une tendance linéaire droite. Il peut donc être conclut que la tendance n'est pas linéaire.

**Normal Q-Q**

*Ce graphique est similaire à celle de la droite de Henry, dans le sens ou les résidus constituent un nuage de points. Si ce nuage de point est aligné avec la droite se trouvant dans le graphique, il est possible de conclure que les résidus suivent une loi normale.* Dans notre cas, il est possible de constater que le nuage de points est assez bien aligné à la droite, ce qui nous permet de de juger de la normalité de la distribution. En utilisant le test de Shapiro-Wilk, 9.385e-05, soit la p-value trouvé, est inférieur à alpha, 0,05. Il peut donc être dit que la distribution n'est pas normale.

**Scale-Location**

*Ce graphique permet de montrer l'égalité de variances des résidus, communément appelé homoscédasticité. Si la répartition des résidus se fait de manière homogène sur le graphique, on parle alors d'homoscédasticité. Dans le cas contraire, on parle d'hétéroscédasticité.* Pour ce qui de la répartition des éléments dans le graphique, il est possible de voir que les valeurs sont vraiment étalées, ce qui nous suggère une hétéroscédasticité. Certes, les points ne sont pas placés de manière homogène autour de celle-ci.

**Résiduals vs Leverage**

*Ce graphique permet d'identifier les points atypiques qui ont une incidence plus importante que d'autres éléments de la population sur les données. Il sera possible d'identifier les points atypiques en analysant la distance de Cook. Dans le cas où les valeurs se trouvent en dehors de cet encadrement, ces derniers peuvent être considéré comme ayant une incidence importante sur les données.* Dans notre cas, il est possible de voir que de multiples points atypiques sont placés à l'extérieur de la droite de la distance de Cook ce qui permet de déduire que plusieurs points ont une incidence importante sur les données. La majorité de ces points atypiques se trouvent au tout début du graphique.

## **Modèle 3 : Y = β0*X1^(β1) * e^(ε)**

```{r}
B03 <- rep(Y)
B13 <- rep(X1)
lnB03 <- log(B03)
lnB13 <- log(B13)
linModel3 <- lm(lnB03~lnB13)

summary(linModel3)
```

```
# Call:
# lm(formula = lnB03 ~ lnB13)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -0.6649 -0.1216  0.0011  0.1333  0.5920
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  5.84249    0.11742   49.76   <2e-16 ***
# lnB13       -0.53835    0.02303  -23.38   <2e-16 ***
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.1791 on 218 degrees of freedom
# Multiple R-squared:  0.7148,            Adjusted R-squared:  0.7135
# F-statistic: 546.4 on 1 and 218 DF,  p-value: < 2.2e-16
```

```{r}
coef(linModel3)
```

```
# (Intercept)       lnB13
#    5.842488   -0.538350
```

```{r}
anaVariance3 <- anova(linModel3)
anaVariance3
```

```
# Analysis of Variance Table
#
# Response: lnB03
#           Df  Sum Sq Mean Sq F value    Pr(>F)
# lnB13      1 17.5270 17.5270  546.45 < 2.2e-16 ***
# Residuals 218  6.9922  0.0321
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Apres transformation de l'équation, on obtient :

**Tableau de coefficient de la régression du Modèle 3**

| | Estimation | Erreur Standard | Probabilité sur le test | Valeur de PV |
|---|---|---|---|---|
| B0 | 5.84249 | 0.11742 | 49.76 | < 2e-16 |
| B1 | -0.53835 | 0.02303 | -23.38 | < 2e-16 |

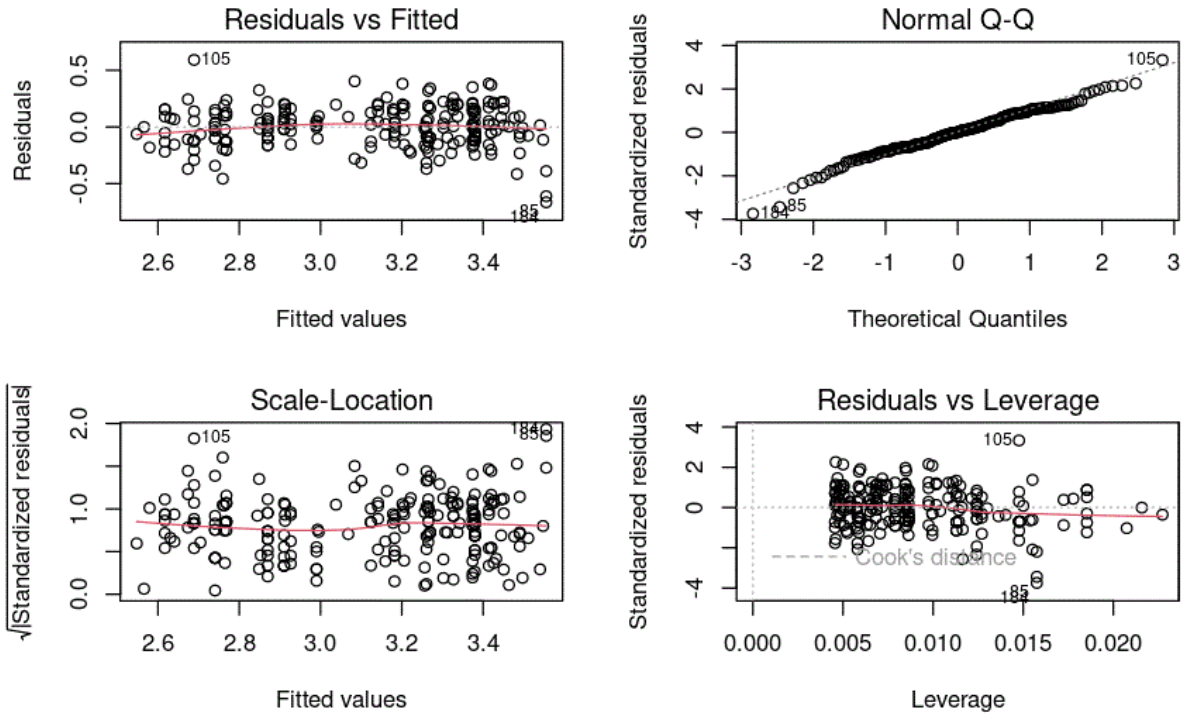
On obtient également le tableau d'analyse de la variance suivante :

**Tableau de d'analyse de la variance du Modèle 3**

| | Degrés de liberté (Df) | Somme des carrés (Sum Sq) | Moyenne des carrés (Mean Sq) | Valeur de F0 (F Value) |
|---|---|---|---|---|
| Régression X1 | 1 | 17.5270 | 17.5270 | 546.45 |
| Erreurs (Résiduals) | 218 | 6.9922 | 0.0321 | N/A |

Test de signification du modèle :

```{r}
par(mfrow = c(2,2))
plot(linModel3)
```



```{r}
residus3 <- rstudent(linModel3)
shapiro.test(residus3)
```

```
#             Shapiro-Wilk normality test
#
# data:  residus3
# W = 0.98329, p-value = 0.01059
```

Le test de signification du modèle consiste à comparer la valeur du PV obtenue lors de l'analyse de la variance à celle de alpha, soit 0.05. La valeur obtenue est de 0.01059 soit un nombre inférieur à 0,05. Il peut donc être affirmer que la variable du modèle est significative.

Pour ce qui est des graphiques, le schéma du haut en présente 4, soit **Residuals vs Fitted, Normal Q-Q, Scale-Location et Residuals vs Leverage** et ces derniers seront élaborés en détails ici-bas.

**Residuals vs Fitted**

*Ce graphique permet d'attester la linéarité des résidus. S'il s'agit d'une tendance linéaire, la courbe en rouge du graphique devrait être droite. Dans la mesure ou la courbe en rouge est droite, il serait possible d'affirmer que les résidus suivent une tendance linéaire. Dans le cas contraire, les résidus ne suivent pas de tendance linéaire.* Dans notre cas, il est possible de voir que la courbe en rouge suit assez bien la trajectoire de la droite du graphique. Il peut donc être conclu que la tendance est linéaire.

**Normal Q-Q**

*Ce graphique est similaire à celle de la droite de Henry, dans le sens ou les résidus constituent un nuage de points. Si ce nuage de point est aligné avec la droite se trouvant dans le graphique, il est possible de conclure que les résidus suivent une loi normale*. Dans notre cas, il est possible de constater que le nuage de points est assez bien aligné à la droite, ce qui nous permet de de juger de la normalité de la distribution. En utilisant le test de Shapiro-Wilk, 0.01059, soit la p-value trouvé, est inférieur à alpha, 0,05. Il peut donc être dit que la distribution n'est pas normale.

**Scale-Location**

*Ce graphique permet de montrer l'égalité de variances des résidus, communément appelé homoscédasticité. Si la répartition des résidus se fait de manière homogène sur le graphique, on parle alors d'homoscédasticité. Dans le cas contraire, on parle d'hétéroscédasticité.* Pour ce qui de la répartition des éléments dans le graphique, il est possible de voir que les valeurs sont éparpillées un peu partout, ce qui nous suggère une hétéroscédasticité.

**Résiduals vs Leverage**

*Ce graphique permet d'identifier les points atypiques qui ont une incidence plus importante que d'autres éléments de la population sur les données. Il sera possible d'identifier les points atypiques en analysant la distance de Cook. Dans le cas où les valeurs se trouvent en dehors de cet encadrement, ces derniers peuvent être considéré comme ayant une incidence importante sur les données.* Dans notre cas, il est possible de voir que peu de points atypiques sont placés à l'extérieur de la droite de la distance de Cook ce qui permet de déduire qu'un nombre faible de points ont une incidence importante sur les données.

**Modèle 4 : Y = β0*e^(β1*X1+ε)**

```{r}
B04 <- rep(Y)
lnB04 <- log(B04)
linModel4 <- lm(lnB04~X1)

summary(linModel4)
```

```
# Call:
# lm(formula = lnB04 ~ X1)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.54136 -0.10678 -0.01168  0.12854  0.63675
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  3.6286313  0.0249708  145.31   <2e-16 ***
# X1          -0.0028128  0.0001193  -23.59   <2e-16 ***
# ---
# Signif. codes:
# 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.178 on 218 degrees of freedom
# Multiple R-squared:  0.7185,            Adjusted R-squared:  0.7172
# F-statistic: 556.3 on 1 and 218 DF,  p-value: < 2.2e-16
```

```{r}
coef(linModel4)
```

```
# (Intercept)          X1
#  3.628631291 -0.002812781
```

```{r}
anaVariance4 <- anova(linModel4)
anaVariance4
```

```
# Analysis of Variance Table
#
# Response: lnB04
#           Df  Sum Sq Mean Sq F value    Pr(>F)
# X1         1 17.6158 17.6158  556.29 < 2.2e-16 ***
# Residuals 218  6.9033  0.0317
# ---
# Signif. codes:
# 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Apres transformation de l'équation, on obtient :

**Tableau de coefficient de la régression du Modèle 4**

| | Estimation | Erreur Standard | Probabilité sur le test | Valeur de PV |
|---|---|---|---|---|
| B0 | 3.6286313 | 0.0249708 | 145.31 | < 2e-16 |
| B1 | -0.0028128 | 0.0001193 | -23.59 | < 2e-16 |

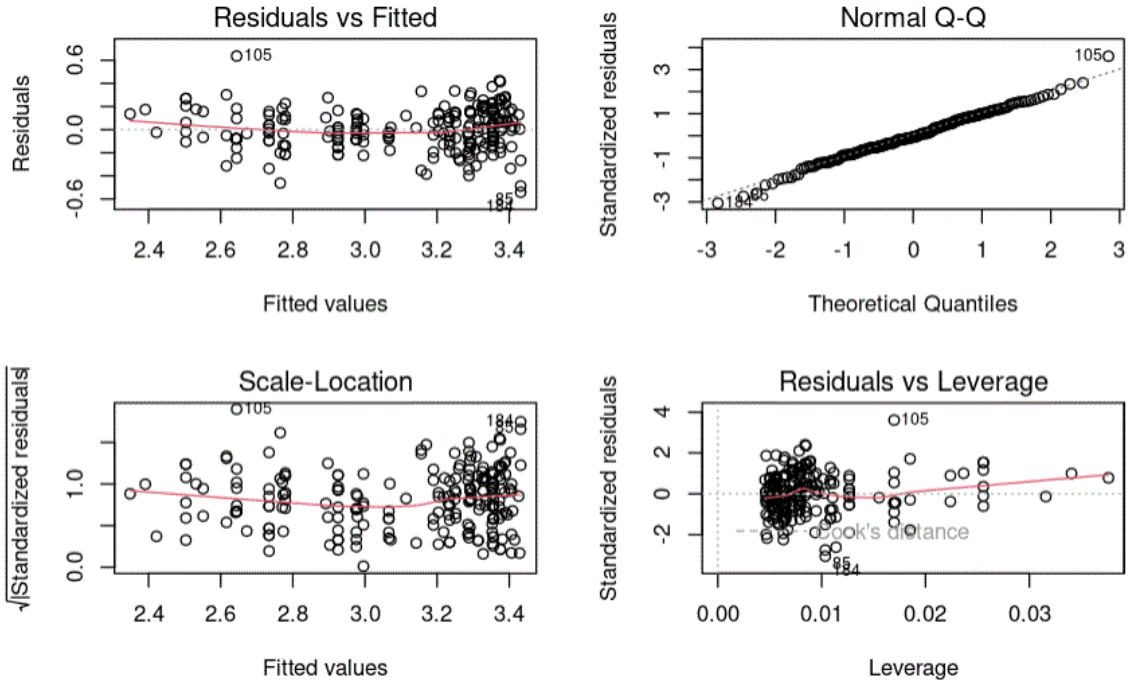
On obtient également le tableau d'analyse de la variance suivante :

**Tableau de d'analyse de la variance du Modèle 4**

| | Degrés de liberté (Df) | Somme des carrés (Sum Sq) | Moyenne des carrés (Mean Sq) | Valeur de F0 (F Value) |
|---|---|---|---|---|
| Régression X1 | 1 | 17.6158 | 17.6158 | 556.29 |
| Erreurs (Résiduals) | 218 | 6.9033 | 0.0317 | N/A |

Test de signification du modèle :

```r
par(mfrow = c(2,2))
plot(linModel4)
```



```{r}
residus4 <- rstudent(linModel4)
shapiro.test(residus4)
```

```
#            Shapiro-Wilk normality test
#
# data:  residus4
# W = 0.99343, p-value = 0.4416
```

Le test de signification du modèle consiste à comparer la valeur du PV obtenue lors de l'analyse de la variance à celle de alpha, soit 0.05. La valeur obtenue est de 0.4416 soit un nombre supérieur à 0,05. Il peut donc être affirmer que la variable du modèle est non-significative.

Pour ce qui est des graphiques, le schéma du haut en présente 4, soit **Residuals vs Fitted, Normal Q-Q, Scale-Location et Residuals vs Leverage** et ces derniers seront élaborés en détails ici-bas.

**Residuals vs Fitted**

*Ce graphique permet d'attester la linéarité des résidus. S'il s'agit d'une tendance linéaire, la courbe en rouge du graphique devrait être droite. Dans la mesure ou la courbe en rouge est droite, il serait possible d'affirmer que les résidus suivent une tendance linéaire. Dans le cas contraire, les résidus ne suivent pas de tendance linéaire.* Dans notre cas, il est possible de voir que la courbe en rouge suit assez bien la trajectoire de la droite du graphique. Il peut donc être conclu que la tendance est linéaire.

**Normal Q-Q**

*Ce graphique est similaire à celle de la droite de Henry, dans le sens ou les résidus constituent un nuage de points. Si ce nuage de point est aligné avec la droite se trouvant dans le graphique, il est possible de conclure que les résidus suivent une loi normale.* Dans notre cas, il est possible de constater que le nuage de points est assez bien aligné à la droite, ce qui nous permet de de juger de la normalité de la distribution. En utilisant le test de Shapiro-Wilk, 0.4416, soit la p-value trouvé est supérieur à alpha, 0,05. Il peut donc être dit que la distribution est normale.

**Scale-Location**

*Ce graphique permet de montrer l'égalité de variances des résidus, communément appelé homoscédasticité. Si la répartition des résidus se fait de manière homogène sur le graphique, on parle alors d'homoscédasticité. Dans le cas contraire, on parle d'hétéroscédasticité.* Pour ce qui de la répartition des éléments dans le graphique, il est possible de voir que les valeurs sont vraiment étalées, ce qui nous suggère une hétéroscédasticité. Certes, les points ne sont pas placés de manière homogène autour de celle-ci.

**Résiduals vs Leverage**

*Ce graphique permettra d'identifier les points atypiques qui ont une incidence plus importante que d'autres éléments de la population sur les données. Il sera possible d'identifier les points atypiques en analysant la distance de Cook. Dans le cas où les valeurs se trouvent en dehors de cet encadrement, ces derniers peuvent être considéré comme ayant une incidence importante sur les données.* Dans notre cas, il est possible de voir que peu de points atypiques sont placés à l'extérieur de la droite de la distance de Cook ce qui permet de déduire qu'un nombre faible de points ont une incidence importante sur les données.

**Modèle 5 : Y = β0 + β1*X2 + ε**

```{r}
X2 <- mondata$weight

linModel5 <- lm(Y~X2)
summary(linModel5)
```

```
# Call:
# lm(formula = Y ~ X2)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -11.6801  -2.7944  -0.3729   2.5103  14.3242
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 45.7823928  1.0253126   44.65   <2e-16 ***
# X2          -0.0075811  0.0003385  -22.40   <2e-16 ***
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 4.184 on 218 degrees of freedom
# Multiple R-squared:  0.6970,            Adjusted R-squared:  0.6957
# F-statistic: 501.6 on 1 and 218 DF,  p-value: < 2.2e-16
```

```{r}
coef(linModel5)
```

```
# (Intercept)          X2
# 45.782392807 -0.007581106
```

```{r}
anaVariance5 <- anova(linModel5)
anaVariance5
```

```
# Analysis of Variance Table
#
# Response: Y
#           Df Sum Sq Mean Sq F value    Pr(>F)
# X2         1   8780  8780.0  501.59 < 2.2e-16 ***
# Residuals 218   3816    17.5
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Regardant l'équation du modèle 5, celui-ci correspond à une équation simple de la régression linéaire. Il n'y a donc aucune modification à apporter à cette dernière, mise à part son ajustement. On obtient ainsi le tableau de coefficient de la régression suivante :

**Tableau de coefficient de la régression du Modèle 5**

| | Estimation | Erreur Standard | Probabilité sur le test | Valeur de PV |
|---|---|---|---|---|
| B0 | 45.7823928 | 1.0253126 | 44.65 | < 2e-16 |
| B1 | -0.0075811 | 0.0003385 | -22.40 | < 2e-16 |

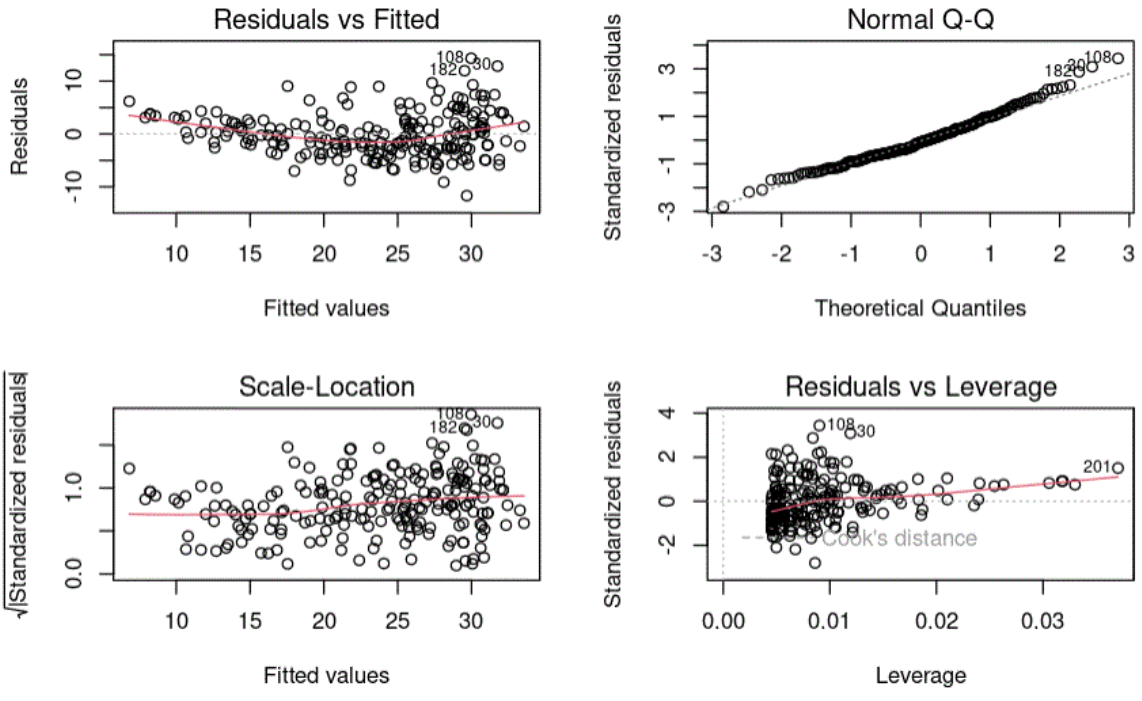
On obtient également le tableau d'analyse de la variance suivante :

**Tableau de d'analyse de la variance du Modèle 5**

| | Degrés de liberté (Df) | Somme des carrés (Sum Sq) | Moyenne des carrés (Mean Sq) | Valeur de F0 (F Value) |
|---|---|---|---|---|
| Régression X1 | 1 | 8780.0 | 8780.0 | 501.59 |
| Erreurs (Résiduals) | 218 | 3816 | 17.5 | N/A |

Test de signification du modèle :

```r
par(mfrow = c(2,2))
plot(linModel5)
```



```{r}
residus5 <- rstudent(linModel5)
shapiro.test(residus5)
```

```
#            Shapiro-Wilk normality test
#
# data:  residus5
# W = 0.981, p-value = 0.004645
```

Le test de signification du modèle consiste à comparer la valeur du PV obtenue lors de l'analyse de la variance à celle de alpha, soit 0.05. La valeur obtenue est de 0.004645 soit un nombre inférieur à 0,05. Il peut donc être affirmer que la variable du modèle est significative.

Pour ce qui est des graphiques, le schéma du haut en présente 4, soit **Residuals vs Fitted, Normal Q-Q, Scale-Location et Residuals vs Leverage** et ces derniers seront élaborés en détails ici-bas.

**Residuals vs Fitted**

*Ce graphique permet d'attester la linéarité des résidus. S'il s'agit d'une tendance linéaire, la courbe en rouge du graphique devrait être droite. Dans la mesure ou la courbe en rouge est droite, il serait possible d'affirmer que les résidus suivent une tendance linéaire. Dans le cas contraire, les résidus ne suivent pas de tendance linéaire.* Dans notre cas, il est possible de voir qu'il ne s'agit pas d'une courbe. En effet, les points ont tendances à former une espèce de courbe éliminant ainsi toute possibilité à avoir une tendance linéaire droite. Il peut donc être conclut que la tendance n'est pas linéaire.

**Normal Q-Q**

*Ce graphique est similaire à celle de la droite de Henry, dans le sens ou les résidus constituent un nuage de points. Si ce nuage de point est aligné avec la droite se trouvant dans le graphique, il est possible de conclure que les résidus suivent une loi normale.* Dans notre cas, il est possible de constater que le nuage de points est assez bien aligné a la droite, ce qui nous permet de de juger de la normalité de la distribution. En utilisant le test de Shapiro-Wilk, 0.004645, soit la p-value trouvé, est inférieur à alpha, 0,05. Il peut donc être dit que la distribution n'est pas normale.

**Scale-Location**

*Ce graphique permet de montrer l'égalité de variances des résidus, communément appelé homoscédasticité. Si la répartition des résidus se fait de manière homogène sur le graphique, on parle alors d'homoscédasticité. Dans le cas contraire, on parle d'hétéroscédasticité.* Pour ce qui de la répartition des éléments dans le graphique, il est possible de voir que les valeurs sont vraiment étalées partout de manière hétérogène, ce qui suggère une hétéroscédasticité.

**Résiduals vs Leverage**

*Ce graphique permettra d'identifier les points atypiques qui ont une incidence plus importante que d'autres éléments de la population sur les données. Il sera possible d'identifier les points atypiques en analysant la distance de Cook. Dans le cas où les valeurs se trouvent en dehors de cet encadrement, ces derniers peuvent être considéré comme ayant une incidence importante sur les données.* Dans notre cas, il est possible de voir que peu de points atypiques sont placés à l'extérieur de la droite de la distance de Cook ce qui permet de déduire qu'un nombre faible de points ont une incidence importante sur les données.

## **Modèle 6 : Y = β0 + β1*X2^2 + ε**

```{r}
B06 <- rep(X2)
B16 <- B06^2

linModel6 <- lm(Y~B16)
summary(linModel6)
```

```
# Call:
# lm(formula = Y ~ B16)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -10.9860  -3.0778  -0.7911   2.4396  15.1281
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  3.409e+01  6.015e-01   56.69   <2e-16 ***
# B16         -1.132e-06  5.656e-08  -20.02   <2e-16 ***
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 4.511 on 218 degrees of freedom
# Multiple R-squared:  0.6477,            Adjusted R-squared:  0.6461
# F-statistic: 400.9 on 1 and 218 DF,  p-value: < 2.2e-16
```

```{r}
coef(linModel6)
```

```
#   (Intercept)          B16
#  3.409449e+01 -1.132355e-06
```

```{r}
anaVariance6 <- anova(linModel6)
anaVariance6
```

```
# Analysis of Variance Table
#
# Response: Y
#           Df Sum Sq Mean Sq F value    Pr(>F)
# B16        1 8159.0  8159.0  400.88 < 2.2e-16 ***
# Residuals 218 4436.9    20.4
#
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Tableau de coefficient de la régression du Modèle 6**

| | Estimation | Erreur Standard | Probabilité sur le test | Valeur de PV |
|---|---|---|---|---|
| B0 | 3.409e+01 | 6.015e-01 | 56.69 | < 2e-16 |
| B1 | -1.132e-06 | 5.656e-08 | -20.02 | < 2e-16 |

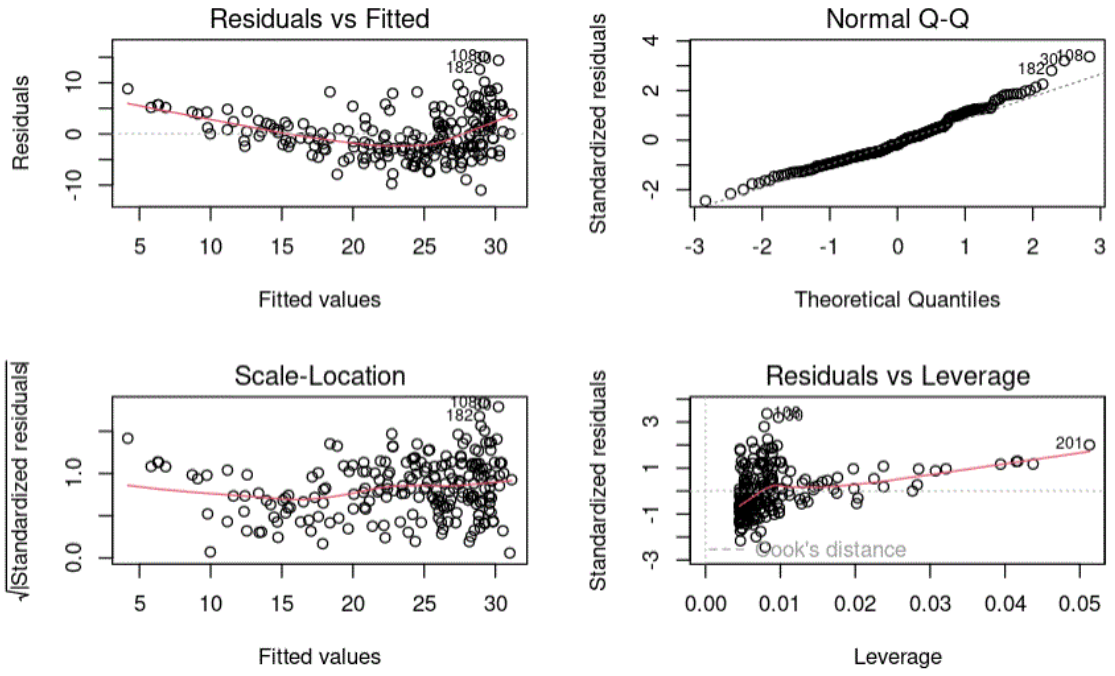
On obtient également le tableau d'analyse de la variance suivante :

**Tableau de d'analyse de la variance du Modèle 6**

| | Degrés de liberté (Df) | Somme des carrés (Sum Sq) | Moyenne des carrés (Mean Sq) | Valeur de F0 (F Value) |
|---|---|---|---|---|
| Régression X1 | 1 | 8159.0 | 8159.0 | 400.88 |
| Erreurs (Résiduals) | 218 | 4436.9 | 20.4 | N/A |

Test de signification du modèle :

```
```{r}
par(mfrow = c(2,2))
plot(linModel6)
```
```



```{r}
residus6 <- rstudent(linModel6)
shapiro.test(residus6)
```

```
#             Shapiro-Wilk normality test
#
# data:  residus6
# W = 0.97648, p-value = 0.0009865
```

Le test de signification du modèle consiste à comparer la valeur du PV obtenue lors de l'analyse de la variance à celle de alpha, soit 0.05. La valeur obtenue est de 0.0009865 soit un nombre inférieur à 0,05. Il peut donc être affirmer que la variable du modèle est significative.

Pour ce qui est des graphiques, le schéma du haut en présente 4, soit **Residuals vs Fitted, Normal Q-Q, Scale-Location et Residuals vs Leverage** et ces derniers seront élaborés en détails ici-bas.

**Residuals vs Fitted**

*Ce graphique permet d'attester la linéarité des résidus. S'il s'agit d'une tendance linéaire, la courbe en rouge du graphique devrait être droite. Dans la mesure ou la courbe en rouge est droite, il serait possible d'affirmer que les résidus suivent une tendance linéaire. Dans le cas contraire, les résidus ne suivent pas de tendance linéaire.* Dans notre cas, il est possible de voir qu'il ne s'agit pas d'une courbe. En effet, les points ont tendances à former une espèce de courbe éliminant ainsi toute possibilité à avoir une tendance linéaire droite. Il peut donc être conclut que la tendance n'est pas linéaire.

**Normal Q-Q**

*Ce graphique est similaire à celle de la droite de Henry, dans le sens ou les résidus constituent un nuage de points. Si ce nuage de point est aligné avec la droite se trouvant dans le graphique, il est possible de conclure que les résidus suivent une loi normale.* Dans notre cas, il est possible de constater que le nuage de points est assez bien aligné à la droite, ce qui nous permet de juger de la normalité de la distribution. En utilisant le test de Shapiro-Wilk, 0.0009865, soit la p-value trouvé, est inférieur à alpha, 0,05. Il peut donc être dit que la distribution n'est pas normale.

**Scale-Location**

*Ce graphique permet de montrer l'égalité de variances des résidus, communément appelé homoscédasticité. Si la répartition des résidus se fait de manière homogène sur le graphique, on parle alors d'homoscédasticité. Dans le cas contraire, on parle d'hétéroscédasticité.* Pour ce qui de la répartition des éléments dans le graphique, il est possible de voir que les valeurs sont vraiment étalées, ce qui nous suggère une hétéroscédasticité.

**Résiduals vs Leverage**

*Ce graphique permettra d'identifier les points atypiques qui ont une incidence plus importante que d'autres éléments de la population sur les données. Il sera possible d'identifier les points atypiques en analysant la distance de Cook. Dans le cas où les valeurs se trouvent en dehors de cet encadrement, ces derniers peuvent être considéré comme ayant une incidence importante sur les données.* Dans notre cas, il est possible de voir que peu de points atypiques sont placés à l'extérieur de la droite de la distance de Cook ce qui permet de déduire qu'un nombre faible de points ont une incidence importante sur les données.

## Modèle 7 : Y = β0*X2^(β1) * e^(ε)

```{r}
B07 <- rep(Y)
B17 <- rep(X2)
lnB07 <- log(B07)
lnB17 <- log(B17)
linModel7 <- lm(lnB07~lnB17)

summary(linModel7)
```

```
# Call:
# lm(formula = lnB07 ~ lnB17)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -0.5134 -0.1009 -0.0019  0.1052  0.4696
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 11.48434    0.31424   36.55   <2e-16 ***
# lnB17       -1.05475    0.03957  -26.66   <2e-16 ***
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.1625 on 218 degrees of freedom
# Multiple R-squared:  0.7653,            Adjusted R-squared:  0.7642
# F-statistic: 710.7 on 1 and 218 DF,  p-value: < 2.2e-1
```

```{r}
coef(linModel7)
```

```
# (Intercept)       lnB17
#   11.484338   -1.054754
```

```{r}
anaVariance7 <- anova(linModel7)
anaVariance7
```

```
# Analysis of Variance Table
#
# Response: lnB07
#           Df  Sum Sq Mean Sq F value    Pr(>F)
# lnB17      1 18.7635 18.7635  710.68 < 2.2e-16 ***
# Residuals 218  5.7557  0.0264
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Apres transformation de l'équation, on obtient :

**Tableau de coefficient de la régression du Modèle 7**

| | Estimation | Erreur Standard | Probabilité sur le test | Valeur de PV |
|---|---|---|---|---|
| B0 | 11.48434 | 0.31424 | 36.55 | < 2e-16 |
| B1 | -1.05475 | 0.03957 | -26.66 | < 2e-16 |

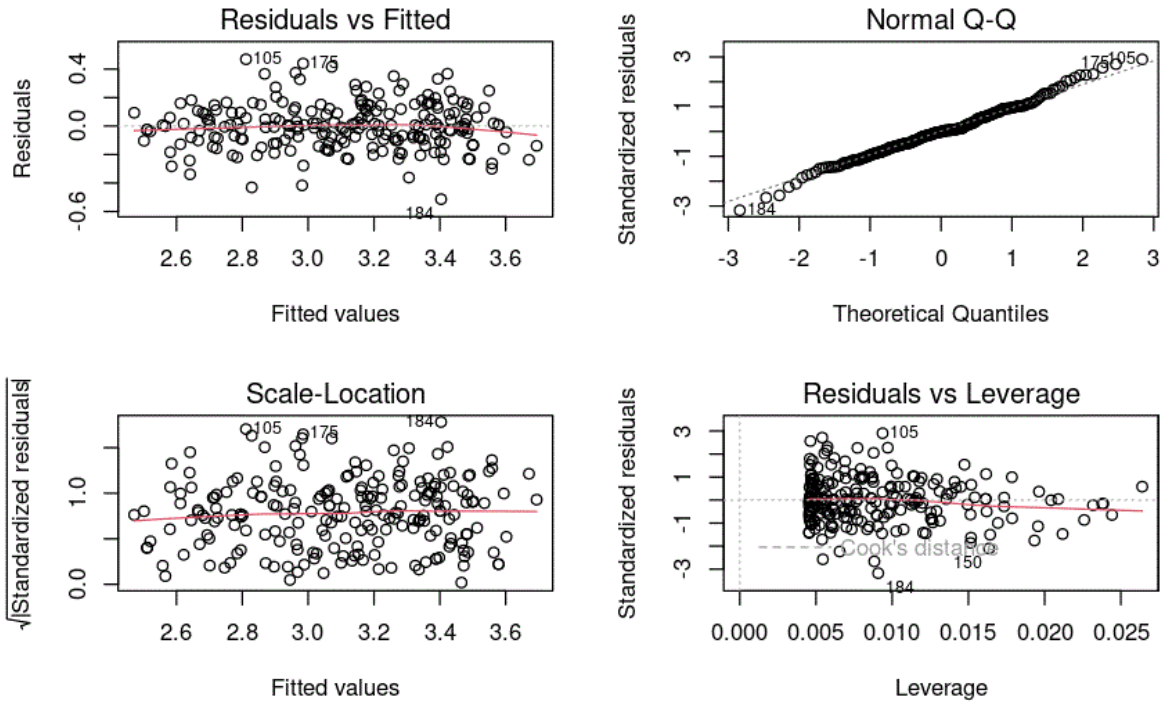
On obtient également le tableau d'analyse de la variance suivante :

**Tableau de d'analyse de la variance du Modèle 7**

| | Degrés de liberté (Df) | Somme des carrés (Sum Sq) | Moyenne des carrés (Mean Sq) | Valeur de F0 (F Value) |
|---|---|---|---|---|
| Régression X1 | 1 | 18.7635 | 18.7635 | 710.68 |
| Erreurs (Résiduals) | 218 | 5.7557 | 0.0264 | N/A |

Test de signification du modèle :

```{r}
par(mfrow = c(2,2))
plot(linModel7)

```



```{r}
residus7 <- rstudent(linModel7)
shapiro.test(residus7)

```

```
#            Shapiro-Wilk normality test
#
# data:  residus7
# W = 0.99146, p-value = 0.2256
```

Le test de signification du modèle consiste à comparer la valeur du PV obtenue lors de l'analyse de la variance à celle de alpha, soit 0.05. La valeur obtenue est de 0.2256 soit un nombre supérieur à 0,05. Il peut donc être affirmer que la variable du modèle est non-significative.

Pour ce qui est des graphiques, le schéma du haut en présente 4, soit **Residuals vs Fitted, Normal Q-Q, Scale-Location et Residuals vs Leverage** et ces derniers seront élaborés en détails ci-dessous.

**Residuals vs Fitted**

*Ce graphique permet d'attester la linéarité des résidus. S'il s'agit d'une tendance linéaire, la courbe en rouge du graphique devrait être droite. Dans la mesure ou la courbe en rouge est droite, il serait possible d'affirmer que les résidus suivent une tendance linéaire. Dans le cas contraire, les résidus ne suivent pas de tendance linéaire.* Dans notre cas, il est possible de voir qu'il s'agit d'une courbe. En effet, l'allure générale de la courbe suit celle d'une droite ayant une tendance linéaire droite. Il peut donc être conclut que la tendance est linéaire.

**Normal Q-Q**

*Ce graphique est similaire à celle de la droite de Henry, dans le sens ou les résidus constituent un nuage de points. Si ce nuage de point est aligné avec la droite se trouvant dans le graphique, il est possible de conclure que les résidus suivent une loi normale.* Dans notre cas, il est possible de constater que le nuage de points est assez bien aligné a la droite, ce qui nous permet de de juger de la normalité de la distribution. En utilisant le test de Shapiro-Wilk, 0. 2256, soit la p-value trouvé, est supérieur à alpha, 0,05. Il peut donc être dit que la distribution est normale.

**Scale-Location**

*Ce graphique permet de montrer l'égalité de variances des résidus, communément appelé homoscédasticité. Si la répartition des résidus se fait de manière homogène sur le graphique, on parle alors d'homoscédasticité. Dans le cas contraire, on parle d'hétéroscédasticité.* Pour ce qui de la répartition des éléments dans le graphique, il est possible de voir que les valeurs sont assez étalées, ce qui nous suggère une hétéroscédasticité. Certes, les points ne sont pas placés de manière très homogène autour de celle-ci.

**Résiduals vs Leverage**

*Ce graphique permettra d'identifier les points atypiques qui ont une incidence plus importante que d'autres éléments de la population sur les données. Il sera possible d'identifier les points atypiques en analysant la distance de Cook. Dans le cas où les valeurs se trouvent en dehors de cet encadrement, ces derniers peuvent être considéré comme ayant une incidence importante sur les données.* Dans notre cas, il est possible de voir que peu de points atypiques sont placés à l'extérieur de la droite de la distance de Cook ce qui permet de déduire qu'un nombre faible de points ont une incidence importante sur les données.

## Modèle 8 : Y = β0*e^(β1*X2+ε)

```{r}
B08 <- rep(Y)
lnB08 <- log(B08)
linModel8 <- lm(lnB08~X2)

summary(linModel8)
```

```
# Call:
# lm(formula = lnB08 ~ X2)
#
# Residuals:
#     Min       1Q   Median       3Q      Max
# -0.49804 -0.10087 -0.00742  0.10659  0.45380
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  4.133e+00  3.977e-02   103.9   <2e-16 ***
# X2          -3.506e-04  1.313e-05   -26.7   <2e-16 ***
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.1623 on 218 degrees of freedom
# Multiple R-squared:  0.7658,            Adjusted R-squared:  0.7648
# F-statistic:   713 on 1 and 218 DF,  p-value: < 2.2e-16
```

```{r}
coef(linModel8)
```

```
#   (Intercept)           X2
#  4.1330824376 -0.0003505971
```

```{r}
anaVariance8 <- anova(linModel8)
anaVariance8
```

```
# Analysis of Variance Table
#
# Response: lnB08
#           Df  Sum Sq Mean Sq F value    Pr(>F)
# X2         1 18.7778 18.7778     713 < 2.2e-16 ***
# Residuals 218  5.7413  0.0263
# ---
# Signif. codes:
# 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Apres transformation de l'équation, on obtient :

**Tableau de coefficient de la régression du Modèle 8**

| | Estimation | Erreur Standard | Probabilité sur le test | Valeur de PV |
|---|---|---|---|---|
| B0 | 4.133e+00 | 3.977e-02 | 103.9 | < 2e-16 |
| B1 | -3.506e-04 | 1.313e-05 | -26.7 | < 2e-16 |

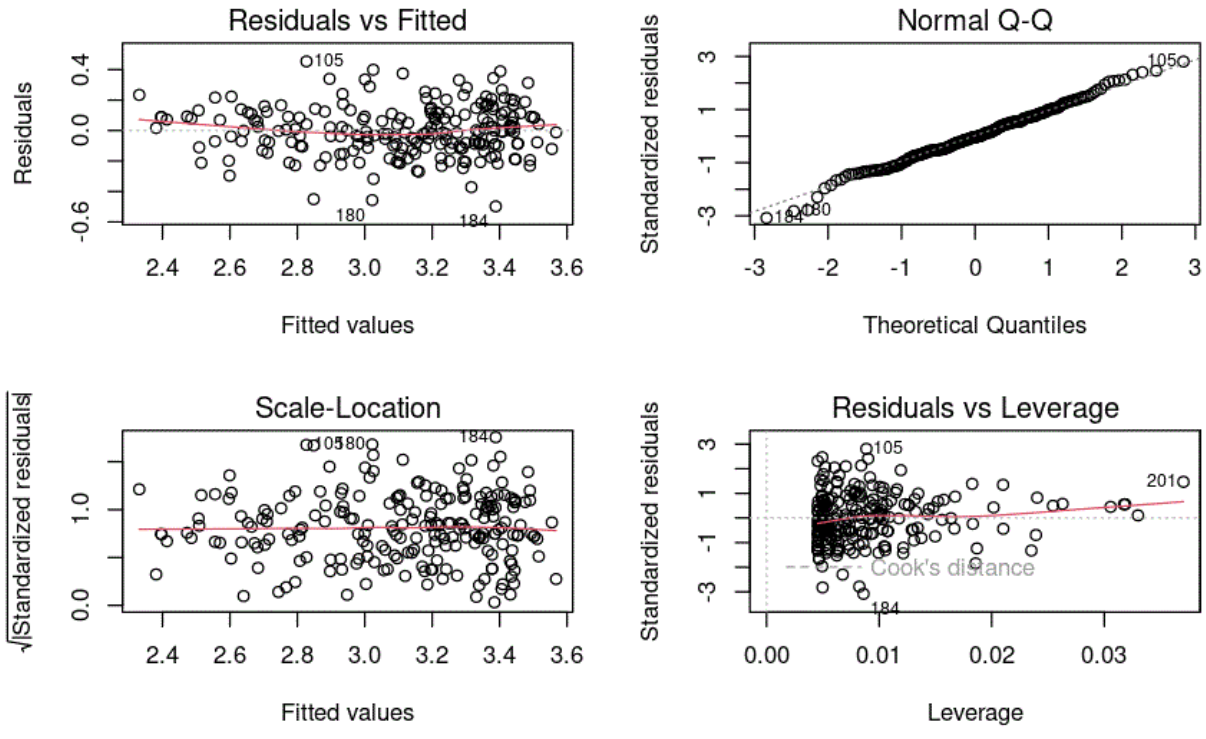
On obtient également le tableau d'analyse de la variance suivante :

**Tableau de d'analyse de la variance du Modèle 8**

| | Degrés de liberté (Df) | Somme des carrés (Sum Sq) | Moyenne des carrés (Mean Sq) | Valeur de F0 (F Value) |
|---|---|---|---|---|
| Régression X1 | 1 | 18.7778 | 18.7778 | 713 |
| Erreurs (Résiduals) | 218 | 5.7413 | 0.0263 | N/A |

Test de signification du modèle :

```{r}
par(mfrow = c(2,2))
plot(linModel8)

```



```{r}
residus8 <- rstudent(linModel8)
shapiro.test(residus8)

```

```
#            Shapiro-Wilk normality test
#
# data:  residus8
# W = 0.99313, p-value = 0.4016
```

Le test de signification du modèle consiste à comparer la valeur du PV obtenue lors de l'analyse de la variance à celle de alpha, soit 0.05. La valeur obtenue est de 0.4016 soit un nombre supérieur à 0,05. Il peut donc être affirmer que la variable du modèle est non-significative.

Pour ce qui est des graphiques, le schéma du haut en présente 4, soit **Residuals vs Fitted, Normal Q-Q, Scale-Location et Residuals vs Leverage** et ces derniers seront élaborés en détails ci-dessous.

**Residuals vs Fitted**

*Ce graphique permet d'attester la linéarité des résidus. S'il s'agit d'une tendance linéaire, la courbe en rouge du graphique devrait être droite. Dans la mesure ou la courbe en rouge est droite, il serait possible d'affirmer que les résidus suivent une tendance linéaire. Dans le cas contraire, les résidus ne suivent pas de tendance linéaire.* Dans notre cas, il est possible de voir qu'il s'agit d'une courbe. En effet, l'allure générale de la courbe suit celle d'une droite ayant une tendance linéaire droite. Il peut donc être conclut que la tendance est linéaire.

**Normal Q-Q**

*Ce graphique est similaire à celle de la droite de Henry, dans le sens ou les résidus constituent un nuage de points. Si ce nuage de point est aligné avec la droite se trouvant dans le graphique, il est possible de conclure que les résidus suivent une loi normale.* Dans notre cas, il est possible de constater que le nuage de points est assez bien aligné a la droite, ce qui nous permet de de juger de la normalité de la distribution. En utilisant le test de Shapiro-Wilk, 0.4016, soit la p-value trouvé, est supérieur à alpha, 0,05. Il peut donc être dit que la distribution est normale.

**Scale-Location**

*Ce graphique permet de montrer l'égalité de variances des résidus, communément appelé homoscédasticité. Si la répartition des résidus se fait de manière homogène sur le graphique, on parle alors d'homoscédasticité. Dans le cas contraire, on parle d'hétéroscédasticité.* Pour ce qui de la répartition des éléments dans le graphique, il est possible de voir que les valeurs sont assez étalées, ce qui nous suggère une hétéroscédasticité. Certes, les points ne sont pas placés de manière très homogène autour de celle-ci.

**Résiduals vs Leverage**

*Ce graphique permettra d'identifier les points atypiques qui ont une incidence plus importante que d'autres éléments de la population sur les données. Il sera possible d'identifier les points atypiques en analysant la distance de Cook. Dans le cas où les valeurs se trouvent en dehors de cet encadrement, ces derniers peuvent être considéré comme ayant une incidence importante sur les données.* Dans notre cas, il est possible de voir que peu de points atypiques sont placés à l'extérieur de la droite de la distance de Cook ce qui permet de déduire qu'un nombre faible de points ont une incidence importante sur les données.

## Intervalle de confiance β0 et β1 pour les modèles 1 et 5

```{r}
intconfiance1 <- confint(linModel1, level = 0.95)
intconfiance1
```

```
#                 2.5 %      97.5 %
# (Intercept) 33.47567794 36.00441816
# X1          -0.06612861 -0.05405162
```

IC de β0 a 95% du modèle 1 : [33.47567794 36.00441816]
IC de β1 a 95% du modèle 1 : [ -0.06612861 -0.05405162]

```{r}
intconfiance5 <- confint(linModel5, level = 0.95)
intconfiance5
```

```
#                 2.5 %       97.5 %
# (Intercept) 43.761598496 47.803187118
# X2          -0.008248259 -0.006913952
```

IC de β0 a 95% du modèle 5 : [43.761598496 47.803187118]
IC de β1 a 95% du modèle 5 : [ -0.008248259 -0.006913952]

Puisque 0 n'est pas comprise dans l'intervalle de confiance de β0 et de β1 autant pour le modèle 1 que pour modèle 5, il peut être affirmer que les deux modèles sont significatifs, comme le suggère le test de signification du modèle.

## Comparaison des 8 modèles et choix du meilleur

Afin de choisir le meilleur modèle, la première étape sera de subdiviser le tout en deux catégories :

- Les modèles dont la variable est significative (p-value > 0,05)
- Les modèles dont la variable est non-significative (p-value < 0,05)

Dans la première catégorie, on trouve les modèles suivants :
- 4 – 7 -8

Dans la deuxième catégorie, on trouve les modèles suivants :
- 1 – 2 – 3 – 5 – 6

En se basant sur le critère de linéarité, il est pas mal évident que le graphique 7 et 8 ressortent des autres graphiques. En jetant un coup d'œil assez critique sur ces derniers, ces deux graphiques sont ceux qui, de loin, ont l'allure de linéarité la plus attendue. En se basant ensuite sur la répartition des points de ces deux graphiques dans le but de déduire celui qui est le plus uniforme, il est possible de voir que le graphique 8 est de loin, le plus homogène des deux. Pour pousser les choses un peu plus loin, il est aussi possible de comparer les valeurs de R^2 entres eux, et de voir qu'effectivement que le modèle 8 possède la meilleure valeur de R^2 de tous ses autres compétiteurs (0.7658). Il est ainsi possible de conclure que le graphique 8 est le meilleur des 8 modèles.

## Calcul de l’intervalle de prévision pour l’efficacité en carburant d’un véhicule

## 1.e) Intervalle de prévision

```{r}
predict(linModel8, newdata = data.frame(X1 = 190, X2 = 2500), interval = "predict", level = 0.95)
```

```
#       fit      lwr      upr
1 3.25659 2.935838 3.577342
```

L’intervalle du modèle 8 se situe entre [2.935838 ; 3.577342]. Cet intervalle correspond à une estimation sur lequel les valeurs futures peuvent se situer dépendamment des observations faites.

## Références

Les plus utiles étaient sans aucun doute :

- https://odr.inrae.fr/
- https://pmarchand1.github.io/
- https://delladata.fr/