POLYTECHNIQUE MONTRÉAL

LE GÉNIE EN PREMIÈRE CLASSE

DÉPARTEMENT DE MATHÉMATIQUES
ET 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 : Bakashov	**Prénom : Marsel**
Groupe : 03	**Matricule : 2147174**

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 `R` pour 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 des matières

Phase 1 : Analyse statistique descriptive et inférence

A) (2 points)

Examinez les liens entre les variables quantitatives de l'étude. Pour cela, produisez une matrice des corrélations pour l'ensemble des trois variables quantitatives et commentez brièvement.

```
mcor <- cor(mondata[,1:3])
round(mcor,2)
```

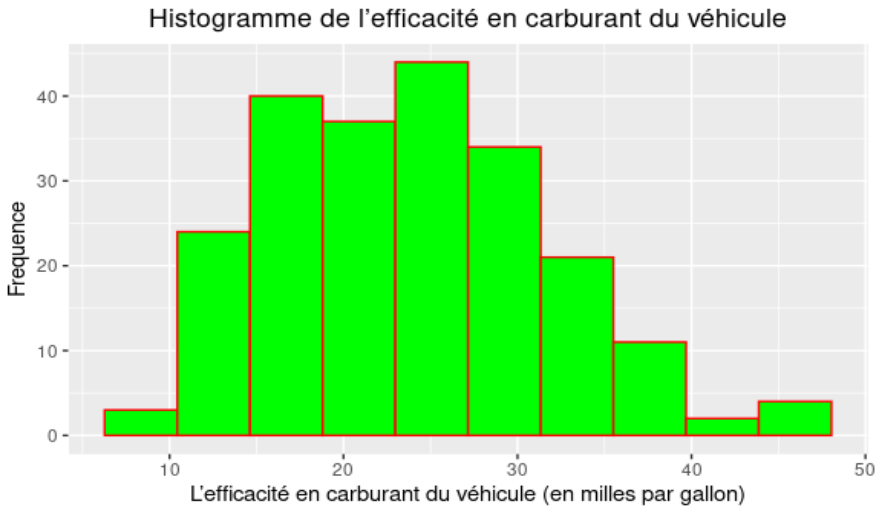
```
##                mpg displacement weight
## mpg           1.00        -0.79  -0.82
## displacement -0.79         1.00   0.93
## weight       -0.82         0.93   1.00
```

Tout d'abord, on peut observer une symétrie par rapport à la diagonale. Une relation linéaire positive existe entre la variable weight et la variable displacement. Ainsi, plus la cylindrée du moteur du véhicule augmente, plus son poids augmente, ce qui semble logique. Une relation linéaire négative existe entre la variable displacement et la variable mpg. En d'autres mots, plus une voiture aura une cylindrée élevée, moins elle aura une bonne efficacité en carburant. Cette même relation est observée entre mpg et weight. Ainsi, plus une voiture sera lourde, moins elle sera efficace au niveau de la consommation de carburant. Les relations suivent la logique des choses. En effet, si le poids du véhicule augmente, il est logique que l'efficacité diminue, cela est prouvé par le -0.82. De plus, si la cylindrée augmente, il est logique que l'efficacité diminue également, ce qui est prouvé par le -0.79. De plus, si la cylindrée augmente, la taille de la voiture doit augmenter en conséquent, pour permettre d'avoir l'espace et les composantes pour cette cylindrée, cela implique donc une augmentation du poids, prouvé par le 0.93. En suivant les chiffres de la matrice et la logique des choses, il est possible d'émettre l'hypothèse que le poids du véhicule à la plus grande influence sur l'efficacité en carburant du véhicule, étant donné que leur corrélation est plus grande que la corrélation entre la cylindrée et l'efficacité. Cela est logique étant donné qu'un véhicule plus lourd nécessite une plus grande consommation de carburant afin de le transporter. De plus, si la cylindrée augmente, cela augmente nécessairement le poids du véhicule.

B) (8 points)

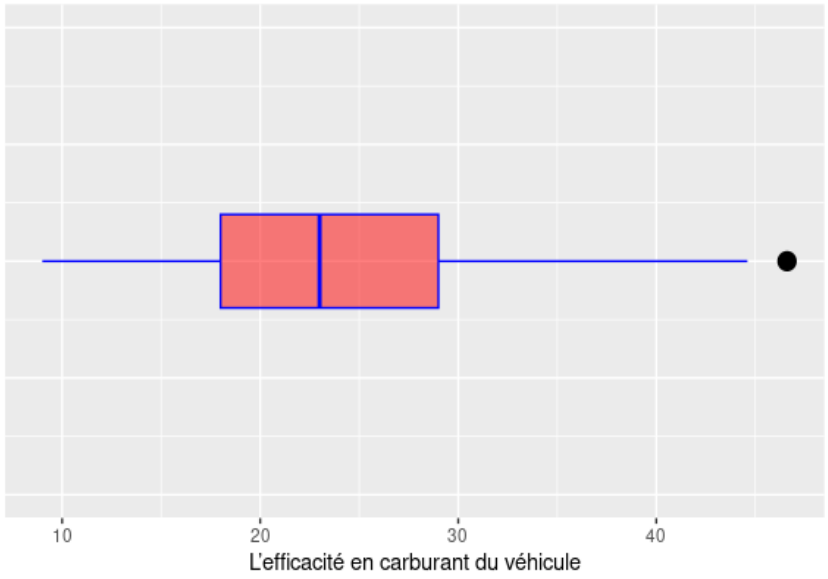
Pour chacune des trois variables Y (l'efficacité en carburant), X1 (la cylindrée) et X2 (le poids), produisez les graphiques et les tableaux demandés ci-dessous et interprétez brièvement les résultats dans chaque cas :

• un histogramme et un diagramme de Tukey (ou «Box Plot») ;

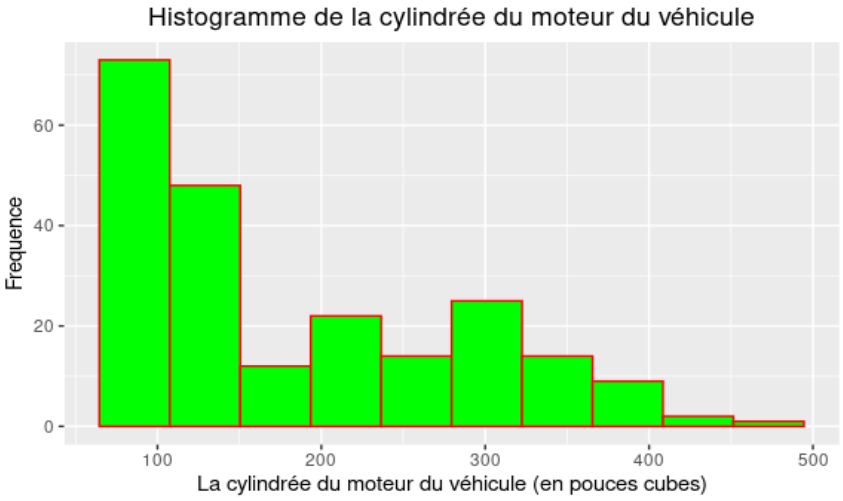


Le premier facteur qui saute aux yeux en regardant cet histogramme est la répartition des véhicules par rapport à leur efficacité en carburant. En effet, on remarque rapidement qu'il y a une forte concentration de véhicules dans l'intervalle [15,25[. Ainsi, on peut en déduire qu'une grande partie des véhicules ont une efficacité se situant entre 15 mpg et 25 mpg.

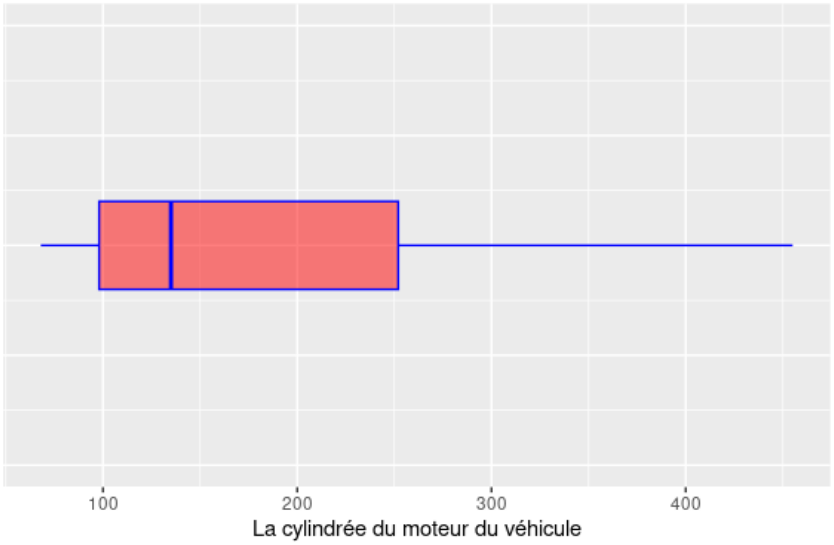
De plus, cet histogramme ne semble pas posséder une distribution normale. En effet, la distribution n'est pas symétrique, elle est asymétrique vers la gauche. Cela nous fait instinctivement penser à une distribution en cloche (gaussienne).



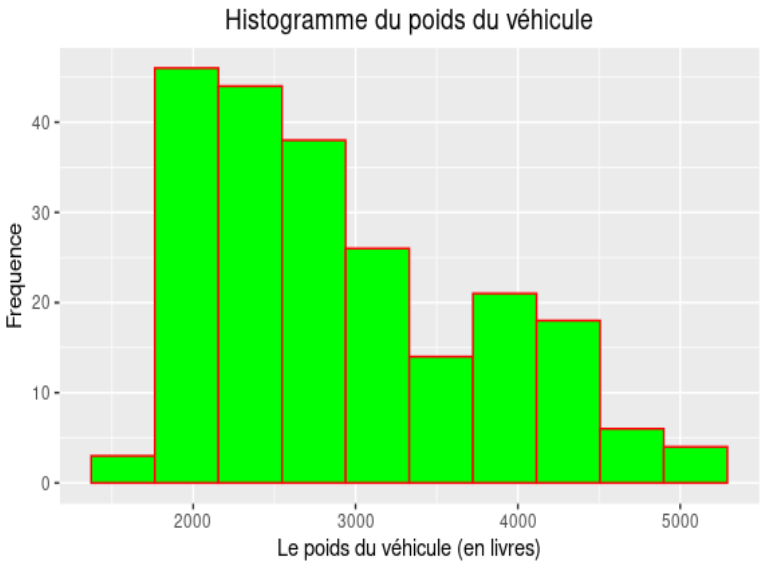
Avec ce diagramme de Tukey, on remarque encore une fois que la distribution tend vers la gauche. On remarque également que la majorité des véhicules ont une efficacité se situant entre 18 mpg et 29 mpg. En d'autres mots 18 est le premier quartile et 29 est le troisième quartile. La valeur minimale est de 9 et la valeur maximale est de 44. On remarque également qu'il y a un point aberrant qui a une valeur beaucoup plus grande que les autres. La médiane est d'environ 23



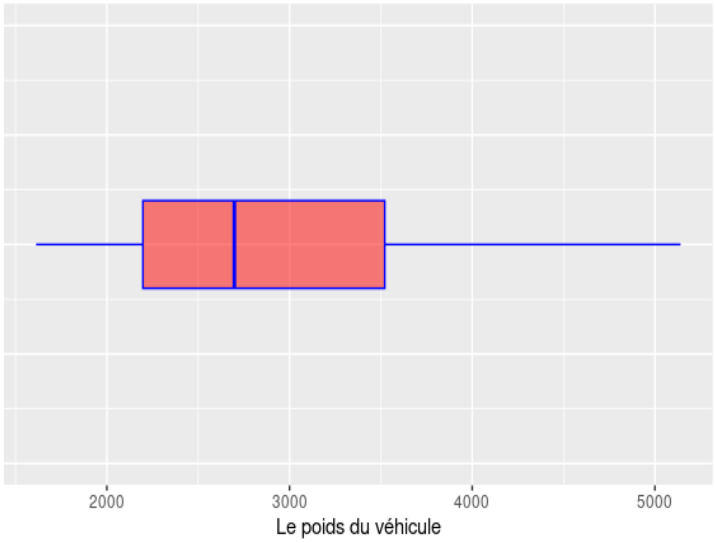
On remarque rapidement que la grande majorité des véhicules ont une cylindrée se situant dans l'intervalle [0,150[. En d'autres mots, la majorité des véhicules ont une cylindrée se situant entre 0 pouces cubes et 150 pouces cubes. Cet histogramme est asymétrique vers la gauche, mais ne possède pas une forme de cloche (distribution gaussienne).



Dans ce diagramme de Tukey, on peut observer que l'intervalle [100,250[est le plus populaire. Tout comme l'histogramme, ce diagramme est asymétrique et tend vers la gauche. On peut également observer les valeurs suivantes :
Q1 : 100
Q3 : 250
Médiane : 125
Max : 450
Min : 50
Une grande majorité des véhicules tournent autour de la médiane allant du premier quartile au deuxième quartile. Ensuite, le reste est majoritaire supérieur au premier quartile, mais il y en a quelques un qui y sont inférieur, mais en minorité.

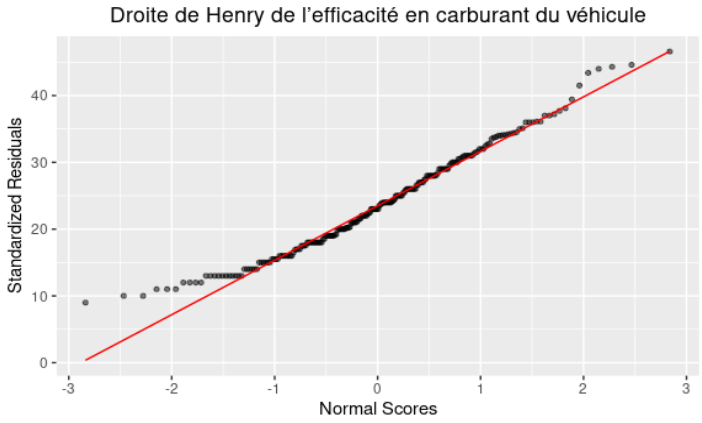


La répartition la plus fréquente est dans l'intervalle [1900, 2900[. En d'autres mots, la majorité des véhicules ont un poids si situant entre 1900 livres et 2900 livres. Cet histogramme a une forme de cloche, donc il ressemble à une distribution gaussienne asymétrique vers la gauche encore une fois.



Dans ce diagramme de Tukey, on peut observer que l'intervalle [2200 ,3500[est le plus populaire. Tout comme l'histogramme, ce diagramme est asymétrique et tend vers la gauche. On peut également observer les valeurs suivantes :
Q1 : 2200
Q3 : 3500
Médiane : 2650
Max : 5200
Min : 1600
Une grande majorité des véhicules tournent autour de la médiane allant du premier quartile au deuxième quartile. Ensuite, le reste est majoritaire supérieur au premier quartile, mais il y en a quelques un qui y sont inférieur, mais en minorité.

• une droite de Henry (ou «Normal Probability Plot») et un test de normalité (Shapiro-Wilk) ;

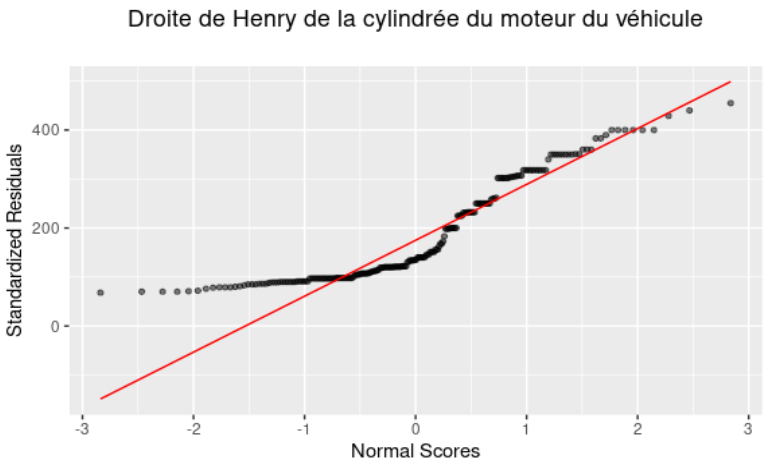



```
##  Shapiro-Wilk normality test
##
## data:  mondata$mpg
## W = 0.97628, p-value = 0.0009237
```

La droite de Henry permet d'ajuster et de visualiser si un échantillon suit une distribution gaussienne ou non. Selon l'histogramme, l'efficacité en carburant du véhicules suivait une distribution gaussienne. En observant notre droite, il est possible d'affirmer que les points qui sont alignés avec la droite rouge suivent une distribution normale (donc une distribution gaussienne). Les points qui sont situés dans l'intervalle [-1,2 ; 1,9] sont relativement alignés avec la droite et ils suivent donc une distribution normale en conséquent. Les valeurs qui sont hors de cet intervalle convergent vers la droite de normalité.

En ce qui concerne le test de Shapiro-Wilk, il a pour but de vérifier si la distribution de l'échantillon est normale. Pour ce faire, ce test calcul la p-value et la compare avec la valeur de alpha, qui est de 0,05 dans notre cas. Si p-value > alpha, alors on peut accepter l'hypothèse qui affirme que la distribution de l'échantillon est normale. Si p-value < alpha, alors on rejette l'hypothèse qui affirme que la distribution de l'échantillon est normale. Dans notre cas, 0.0009237 < 0,05, donc on rejette l'hypothèse affirmant la normalité de la distribution de l'échantillon.

Il est donc pas possible d'affirmer que l'efficacité en carburant du véhicule suit une distribution normale.

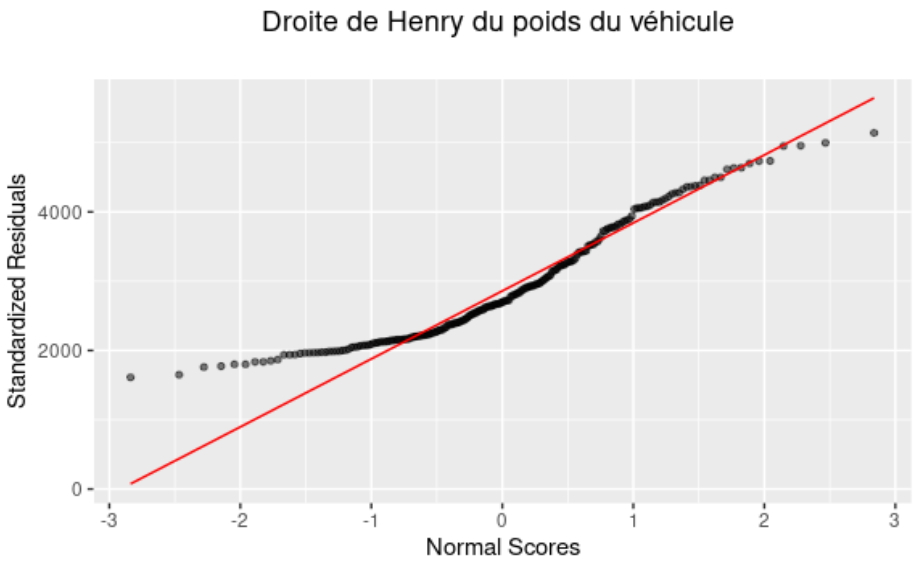



```
##  Shapiro-Wilk normality test
##
## data:  mondata$displacement
## W = 0.86054, p-value = 2.808e-13
```

La droite de Henry permet d'ajuster et de visualiser si un échantillon suit une distribution gaussienne ou non. Selon l'histogramme, l'efficacité en carburant du véhicules suivait une distribution gaussienne. En observant notre droite, il est possible d'affirmer que les points qui sont alignés avec la droite rouge suivent une distribution normale (donc une distribution gaussienne). Les points qui sont situés dans l'intervalle [0,5 ; 0,8] sont relativement alignés avec la droite et ils suivent donc une distribution normale en conséquent. Les valeurs qui sont hors de cet intervalle convergent vers la droite de normalité. Peu de points suivent suite une distribution normale, donc on aura tendance à affirmer que ce n'est pas le cas pour tous les points en regardant la droite de Henry. Toutefois, il faut effectuer le test de Shapiro-Wilk

En ce qui concerne le test de Shapiro-Wilk, il a pour but de vérifier si la distribution de l'échantillon est normale. Pour ce faire, ce test calcul la p-value et la compare avec la valeur de alpha, qui est de 0,05 dans notre cas. Si p-value > alpha, alors on peut accepter l'hypothèse qui affirme que la distribution de l'échantillon est normale. Si p-value < alpha, alors on rejette l'hypothèse qui affirme que la distribution de l'échantillon est normale. Dans notre cas, 2.808e-13< 0,05, donc on rejette l'hypothèse affirmant la normalité de la distribution de l'échantillon. Ce qui suit la conclusion que nous avons eu avec la droite de Henry.

Il est donc pas possible d'affirmer que l'efficacité en carburant du véhicule suit une distribution normale.




```
##  Shapiro-Wilk normality test
##
## data:  mondata$weight
## W = 0.92796, p-value = 6.651e-09
```

La droite de Henry permet d'ajuster et de visualiser si un échantillon suit une distribution gaussienne ou non. Selon l'histogramme, l'efficacité en carburant du véhicules suivait une distribution gaussienne. En observant notre droite, il est possible d'affirmer que les points qui sont alignés avec la droite rouge suivent une distribution normale (donc une distribution gaussienne). Les points qui sont situés dans l'intervalle [-0,8 ; 1] sont relativement alignés avec la droite et ils suivent donc une distribution normale en conséquent. Les valeurs qui sont hors de cet intervalle convergent vers la droite de normalité. Quelques points suivent la droite de normalité, on pourrait donc hésiter sur le fait que cet échantillon suit une distribution normale. Pour ne pas avoir de doutes, effectuons le test de Shapiro-Wilk.

En ce qui concerne le test de Shapiro-Wilk, il a pour but de vérifier si la distribution de l'échantillon est normale. Pour ce faire, ce test calcul la p-value et la compare avec la valeur de alpha, qui est de 0,05 dans notre cas. Si p-value > alpha, alors on peut accepter l'hypothèse qui affirme que la distribution de l'échantillon est normale. Si p-value < alpha, alors on rejette l'hypothèse qui affirme que la distribution de l'échantillon est normale. Dans notre cas, 6.651e-09 < 0,05, donc on rejette l'hypothèse affirmant la normalité de la distribution de l'échantillon.

Il est donc pas possible d'affirmer que l'efficacité en carburant du véhicule suit une distribution normale.

• un tableau de statistiques descriptives comprenant : moyenne, quartiles, écart type, erreur type, intervalle de confiance pour la moyenne ;

Nous pouvons utiliser ces tableaux de statistiques descriptives pour confirmer les résultats obtenus dans les observations précédentes, notamment dans les diagrammes de Tukey.

Statistiques descriptives		95% CI
mpg		23, 25
Moyenne	24	
Écart-type	8	
Erreur type	1	
Minimum	9	
Maximum	47	
Premier quartile	18	
Médiane	23	
Troisième quartile	29	

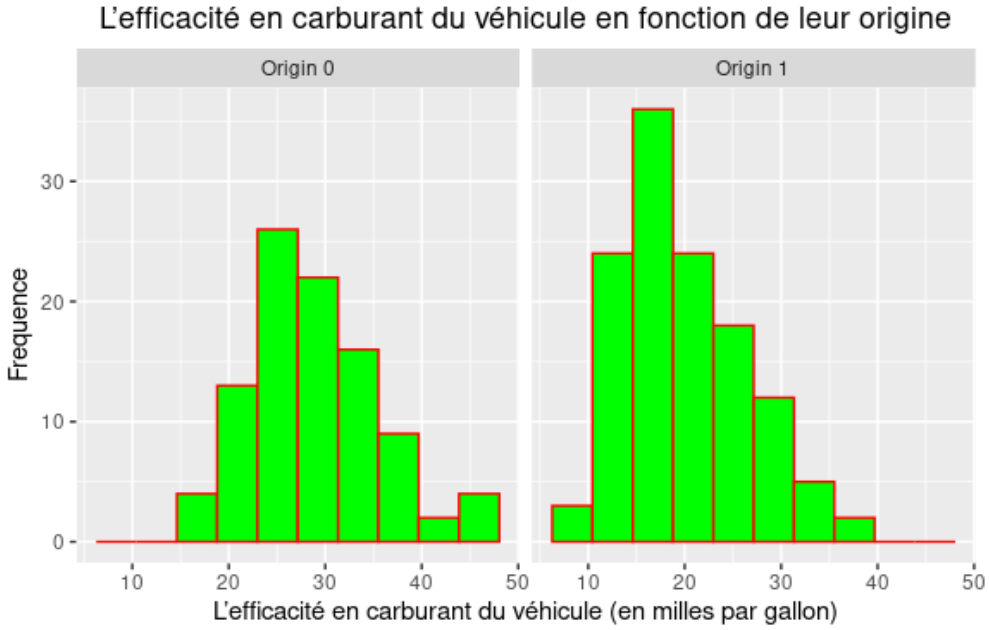
Statistiques descriptives		95% CI
displacement		171, 198
Moyenne	184	
Écart-type	102	
Erreur type	7	
Minimum	68	
Maximum	455	
Premier quartile	98	
Médiane	135	
Troisième quartile	252	

Statistiques descriptives		95% CI
weight		2,806, 3,032
Moyenne	2,919	
Écart-type	851	
Erreur type	57	
Minimum	1,613	
Maximum	5,140	
Premier quartile	2,198	
Médiane	2,698	
Troisième quartile	3,521	

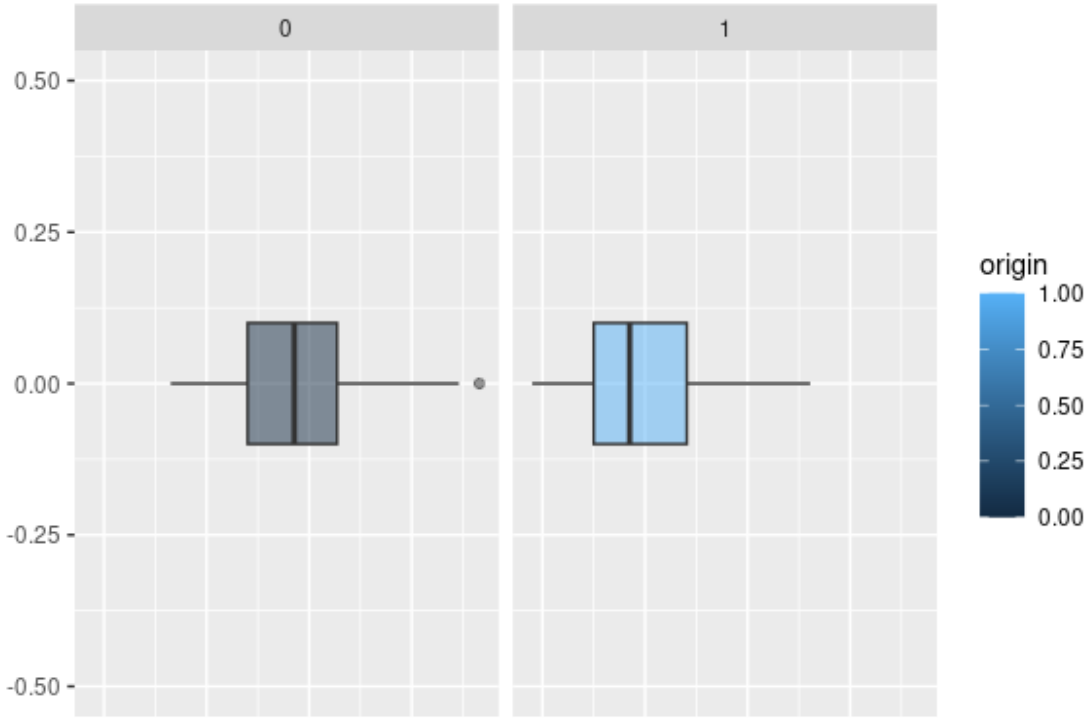
C) (8 points)

Afin de vérifier si l'efficacité en carburant d'un véhicule dépend de l'origine de celui-ci, on peut considérer deux groupes de données selon la variable origin et effectuer une comparaison des deux groupes en termes de moyenne, symétrie et variabilité. Pour ce faire, effectuez les analyses suivantes et donnez une brève conclusion :

- deux histogrammes juxtaposés, et deux diagrammes de Tukey (ou «Box Plot») juxtaposés ;



L'histogramme des véhicules de l'origine 0 semble suivre une distribution normale, alors que l'histogramme des véhicules d'origine 1 semble suivre une distribution gaussienne asymétrique vers la gauche. Les véhicules de l'origine 0 se trouvent en majorité dans l'intervalle [22, 27[et ceux de l'origine 1 se trouvent en majorité dans l'intervalle [15, 18[. Étant donné qu'il y a quand même beaucoup de différences entre les deux histogrammes, il est possible d'affirmer que l'origine du véhicule pourrait avoir une influence sur son efficacité.



On peut confirmer que l'échantillon des véhicules de l'origine 1 ont une distribution asymétrique vers la gauche. En observant les données du tableau de statistiques descriptives ci-dessous, on peut apercevoir qu'il y a des différences entre les moyennes, les variances, les écarts-types, les minimums, les maximums, les quartiles et les intervalles de confiance. On peut donc supposer que l'origine a bel et bien une incidence sur l'efficacité en carburant du véhicule.

• un tableau des statistiques descriptives par groupe : moyenne, quartiles, variance, écart type, intervalle de confiance pour la moyenne ;

Statistiques descriptives	*0*	**95% CI**	*1*	**95% CI**
mpg		27, 30		19, 21
Moyenne	29		20	
Variance	46		41	
Écart-type	7		6	
Minimum	16		9	
Maximum	47		36	
Premier quartile	24		15	
Médiane	29		18	
Troisième quartile	33		24	

• un test d'hypothèses sur l'égalité des variances des deux groupes ;

```
##  F test to compare two variances
##
## data:  mpg by origin
## F = 1.1153, num df = 95, denom df = 123, p-value = 0.5667
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7657915 1.6395772
## sample estimates:
## ratio of variances
##            1.11526
```

$$H0 : \sigma_1^2 = \sigma_2^2 \mid H1 : \sigma_1^2 \neq \sigma_2^2$$

Le test de Fisher ou le test de d'hypothèses sur l'égalité des variances des deux groupes permet de tester si les deux lois normales ont une différence significative entre les deux variances. On pose l'hypothèse H0 : les variances sont égales et H1 : les variances ne sont pas égales. On obtiendra une p-value à la suite de ce test et si cette p-value est inférieur à alpha (0,05 dans notre cas), on rejette l'hypothèse H0. Dans notre cas, 0.5667 > 0.05, donc on garde l'hypothèse H0 et on peut affirmer qu'il est possible que les deux efficacités de carburant aient des variances similaires.

• un test d'hypothèses sur l'égalité des moyennes des deux groupes.

```
##  Welch Two Sample t-test
##
## data:  mpg by origin
## t = 9.7929, df = 198.71, 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:
##   7.02920 10.57389
## sample estimates:
## mean in group 0 mean in group 1
##        28.73542        19.93387
```

$$H0 : \mu_1 = \mu_2 \mid H1 : \mu_1 \neq \mu_2$$

En ce concerne le test d'hypothèses sur l'égalité des moyennes des deux groupes, on rejettera H0 si notre p-value est inférieur à alpha (0,05 dans notre cas). On obtient 2.2e-16 < 0,05, donc on rejette H0. Ainsi, les deux efficacités n'ont pas une moyenne similaire, donc l'origine du véhicule pourrait exercer une influence sur l'efficacité du véhicule.

Phase 2 : Recherche d'un modèle.

D) (15 points)

Pour chacun des huit modèles ci-dessus :

• (5 points) Effectuez l'ajustement (i.e. obtenir le tableau des coefficients de régression, le tableau d'analyse de la variance).

• (5 points) Tester la signification du modèle et effectuez une analyse des résidus (normalité, homoscédasticité, points atypiques, etc.)

Modèle 1

```
## Call:
## lm(formula = Y ~ X1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6978  -3.0598  -0.6526   2.3310  16.8728
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.944473   0.674347   51.82   <2e-16 ***
## X1          -0.060666   0.003202  -18.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.853 on 218 degrees of freedom
## Multiple R-squared:  0.6221, Adjusted R-squared:  0.6204
## F-statistic: 358.9 on 1 and 218 DF,  p-value: < 2.2e-16

## (Intercept)          X1
## 34.94447336 -0.06066641

## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X1          1 8453.2  8453.2  358.86 < 2.2e-16 ***
## Residuals 218 5135.1    23.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

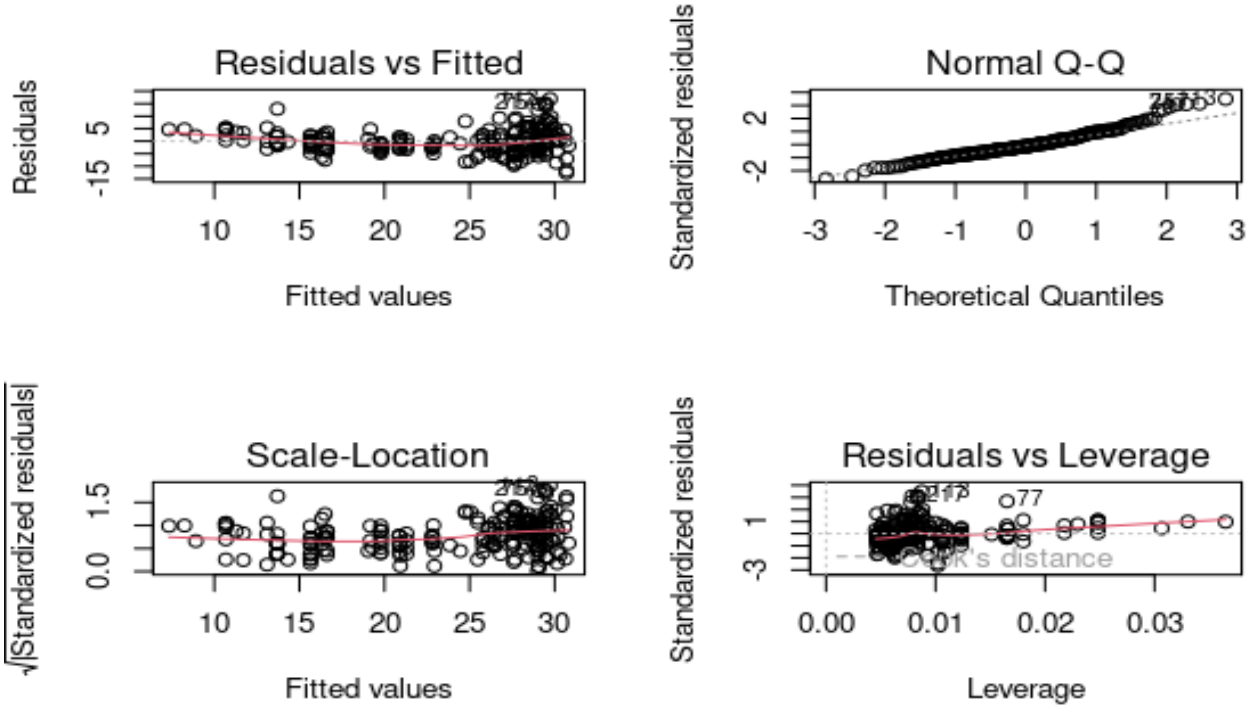
Tableau des coefficients de régression

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept) (β0)	34.944473	0.674347	51.82	<2e-16
X1 (β1)	-0.060666	0.003202	-18.94	<2e-16

Tableau d'analyse de la variance

	Df	Sum sq	Mean sq	F value	Pr(>F)
X1	1	8453.2	8453.2	358.86	< 2.2e-16
Residuals	218	5135.1	23.6		

Selon les tableaux, on remarque que la valeur ajustée de R2 est de 0.6204, ce qui est peu significatif, donc il est probable que les données de l'échantillon ne correspondent pas à ce modèle. De plus, le coefficient de B1 est relativement faible, ce dernier exprime le lien entre l'efficacité du carburant et la cylindrée du moteur.




```
shapiro.test(rstudent(linModel1))

##
##  Shapiro-Wilk normality test
##
## data:  rstudent(linModel1)
## W = 0.96332, p-value = 1.878e-05
```

Le test de significativité

On doit comparer la valeur de notre p-value avec notre alpha (0,05). Si la p-value est supérieure à alpha, alors il est possible de conclure que le modèle n'est pas significatif. On obtient 1.878e-05 < 0,05, donc notre modèle peut être significatif. De plus, on peut affirmer que notre échantillon ne suit pas une loi normale.

Residuals vs Fitted

Dans ce graphique, si les résidus (points) suivent une droite, il est possible d'affirmer que ces derniers suivent une tendance linéaire.

Dans notre cas, il est rapidement apercevable que les résidus ne suivent pas une tendance linéaire. En effet, ceux-ci ont plutôt l'air d'être dispersés aléatoirement et ne suivent pas en conséquent une droite.

Normal Q-Q

Ce graphique permet de vérifier si les résidus suivent une distribution normale, en vérifiant si les points sont tous autour de la droite principale.

Une grande majorité des points semblent suivre la droite de normalité. On aurait tendance à dire que la distribution est normale. Toutefois, le test Shapiro-Wilk vient nous prouver le contraire.

Scale-Location

Ce graphique nous montre si les résidus sont étalés de façon équilibrée, il nous donnera des informations sur l'homoscédasticité des résidus. Si les valeurs sont réparties de façon aléatoire, cela est signe d'une mauvaise homoscédasticité.

On aperçoit que les valeurs sont reparties de façon aléatoire, on pourrait donc affirmer que les résidus ne font pas preuve d'homoscédasticité.

Residuals vs Leverage

Ce graphique permet de vérifier si des valeurs ont une plus grande influence sur les autres données, ces valeurs seront alors dans la distance de Cook.

On ne retrouve pas de résidus dans la distance de Cook. En conséquent, on pourrait affirmer que peu de points ont une grande influence sur les données.

Modèle 2

```
##
## Call:
## lm(formula = Y ~ X1^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6978  -3.0598  -0.6526   2.3310  16.8728
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.944473   0.674347   51.82   <2e-16 ***
## X1          -0.060666   0.003202  -18.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.853 on 218 degrees of freedom
## Multiple R-squared:  0.6221, Adjusted R-squared:  0.6204
## F-statistic: 358.9 on 1 and 218 DF,  p-value: < 2.2e-16

linModel2$coefficients

## (Intercept)          X1
## 34.94447336 -0.06066641

anova(linModel2)

## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X1          1 8453.2  8453.2  358.86 < 2.2e-16 ***
## Residuals 218 5135.1    23.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

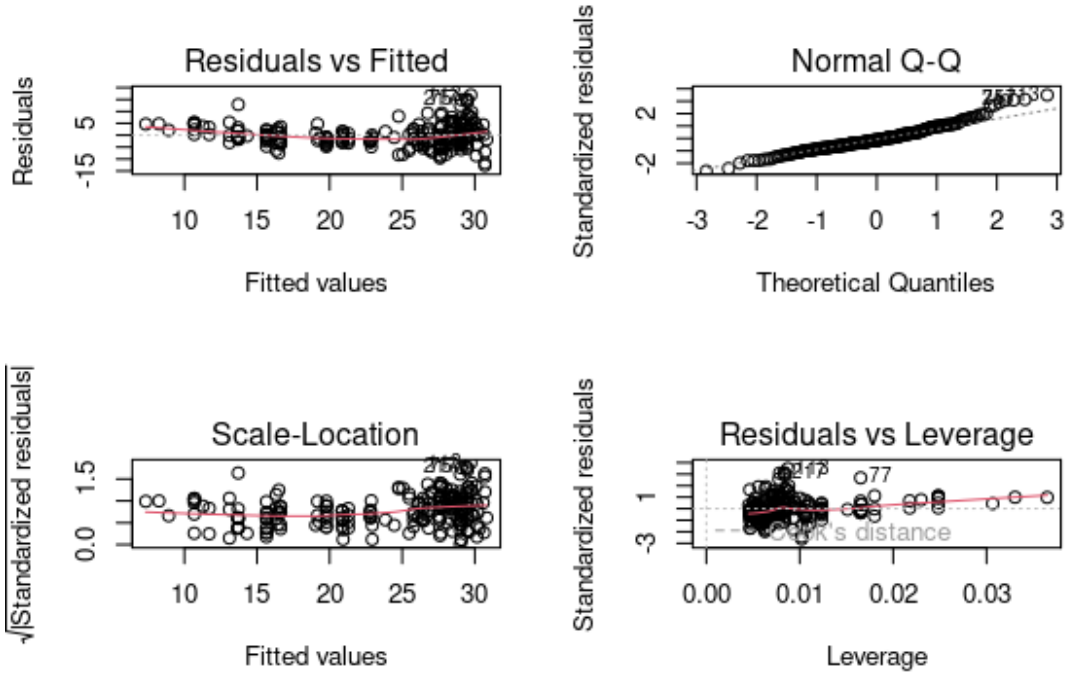
Tableau des coefficients de régression

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept) (β0)	34.944473	0.674347	51.82	<2e-16
X1 (β1)	-0.060666	0.003202	-18.94	<2e-16

Tableau d'analyse de la variance

	Df	Sum sq	Mean sq	F value	Pr(>F)
X1	1	8453.2	8453.2	358.86	< 2.2e-16
Residuals	218	5135.1	23.6		

Le modèle 2, à la différence du modèle 1, vient mettre la variable X1 au carré, afin de vérifier si on peut observer une tendance quadratique. Étant donné que les données des deux tableaux ne changent pas, on pourrait dire qu'il n'y a pas de tendance quadratique dans cet échantillon.




```
shapiro.test(rstudent(linModel2))

##
##  Shapiro-Wilk normality test
##
## data:  rstudent(linModel2)
## W = 0.96332, p-value = 1.878e-05
```

Le test de significativité

On doit comparer la valeur de notre p-value avec notre alpha (0,05). Si la p-value est supérieure à alpha, alors il est possible de conclure que le modèle n'est pas significatif. On obtient 1.878e-05 < 0,05, donc notre modèle peut être significatif. De plus, on peut affirmer que notre échantillon ne suit pas une loi normale.

Residuals vs Fitted

Dans ce graphique, si les résidus (points) suivent une droite, il est possible d'affirmer que ces derniers suivent une tendance linéaire.

Dans notre cas, il est rapidement apercevable que les résidus ne suivent pas une tendance linéaire. En effet, ceux-ci ont plutôt l'air d'être dispersés aléatoirement et ne suivent pas en conséquent une droite.

Normal Q-Q

Ce graphique permet de vérifier si les résidus suivent une distribution normale, en vérifiant si les points sont tous autour de la droite principale.

Une grande majorité des points semblent suivre la droite de normalité. On aurait tendance à dire que la distribution est normale. Toutefois, le test Shapiro-Wilk vient nous prouver le contraire.

Scale-Location

Ce graphique nous montre si les résidus sont étalés de façon équilibrée, il nous donnera des informations sur l'homoscédasticité des résidus. Si les valeurs sont réparties de façon aléatoire, cela est signe d'une mauvaise homoscédasticité.

On aperçoit que les valeurs sont reparties de façon aléatoire, on pourrait donc affirmer que les résidus ne font pas preuve d'homoscédasticité.

Residuals vs Leverage

Ce graphique permet de vérifier si des valeurs ont une plus grande influence sur les autres données, ces valeurs seront alors dans la distance de Cook.

On ne retrouve pas de résidus dans la distance de Cook. En conséquent, on pourrait affirmer que peu de points ont une grande influence sur les données.

Modèle 3

```
##
## Call:
## lm(formula = log(Y) ~ log(X1))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66556 -0.12026  0.00221  0.13383  0.59515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.85295    0.11992   48.81   <2e-16 ***
## log(X1)     -0.54067    0.02352  -22.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1861 on 218 degrees of freedom
## Multiple R-squared:  0.7079, Adjusted R-squared:  0.7065
## F-statistic: 528.2 on 1 and 218 DF,  p-value: < 2.2e-16

linModel3$coefficients

## (Intercept)     log(X1)
##   5.8529463  -0.5406658

anova(linModel3)

## Analysis of Variance Table
##
## Response: log(Y)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## log(X1)     1 18.2947 18.2947  528.23 < 2.2e-16 ***
## Residuals 218  7.5502  0.0346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

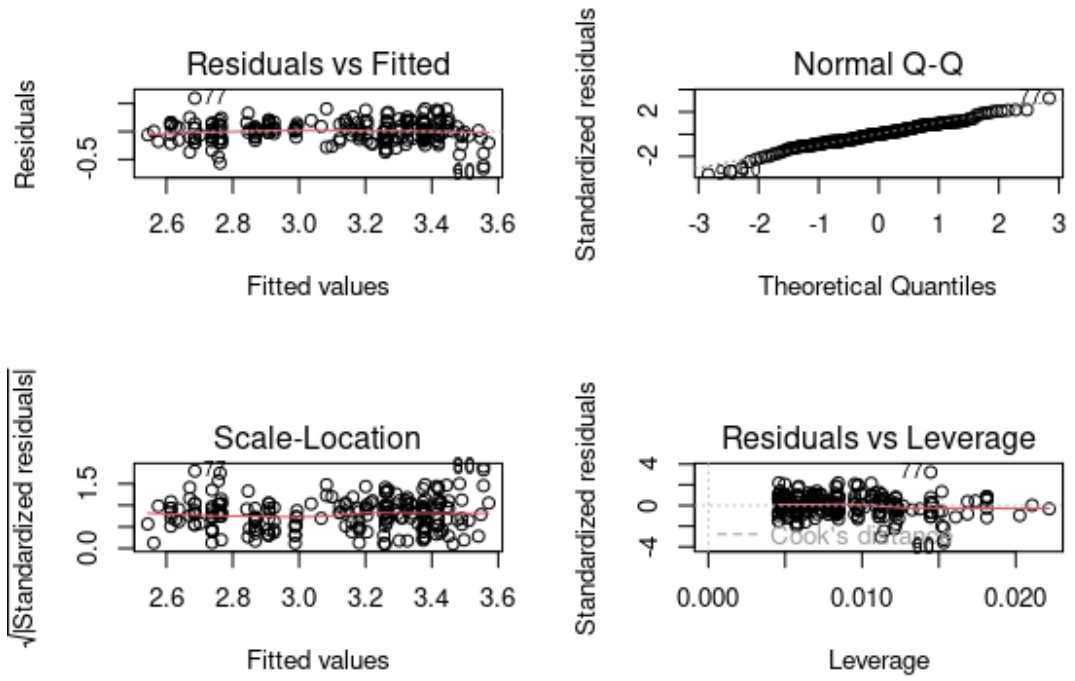
Tableau des coefficients de régression

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept) (β0)	5.85295	0.11992	48.81	<2e-16
X1 (β1)	-0.54067	0.02352	-22.98	<2e-16

Tableau d'analyse de la variance

	Df	Sum sq	Mean sq	F value	Pr(>F)
X1	1	18.2947	18.2947	528.23	< 2.2e-16
Residuals	218	7.5502	0.0346		

La F value est plus grande que dans les modèles précédents suggérant que X1 exercerait une influence significative.




```
shapiro.test(rstudent(linModel3))

##
##  Shapiro-Wilk normality test
##
## data:  rstudent(linModel3)
## W = 0.98211, p-value = 0.006909
```

Le test de significativité

On doit comparer la valeur de notre p-value avec notre alpha (0,05). Si la p-value est supérieure à alpha, alors il est possible de conclure que le modèle n'est pas significatif. On obtient 0.006909< 0,05, donc notre modèle peut être significatif. De plus, on peut affirmer que notre échantillon ne suit pas une loi normale.

Residuals vs Fitted

Dans ce graphique, si les résidus (points) suivent droite, il est possible d'affirmer que ces derniers suivent une tendance linéaire.

Les résidus forment une courbe de trajectoire qui est assez droite. Il serait donc possible d'affirmer que les résidus suivent une tendance linéaire.

Normal Q-Q

Ce graphique permet de vérifier si les résidus suivent une distribution normale, en vérifiant si les points sont tous autour de la droite principale.

On peut apercevoir que les résidus sont relativement bien alignés avec la droite sauf aux extrémités. On pourrait donc penser que c'est une distribution normale, mais le test de Shapiro-Wilk nous confirme le contraire.

Scale-Location

Ce graphique nous montre si les résidus sont étalés de façon équilibrée, il nous donnera des informations sur l'homoscédasticité des résidus. Si les valeurs sont réparties de façon aléatoire, cela est signe d'une mauvaise homoscédasticité.

On aperçoit que les valeurs sont reparties de façon aléatoire, on pourrait donc affirmer que les résidus ne font pas preuve d'homoscédasticité.

Residuals vs Leverage

Ce graphique permet de vérifier si des valeurs ont une plus grande influence sur les autres données, ces valeurs seront alors dans la distance de Cook.

On ne retrouve pas de résidus dans la distance de Cook. En conséquent, on pourrait affirmer que peu de points ont une grande influence sur les données.

Modèle 4

```
##
## Call:
## lm(formula = log(Y) ~ X1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57650 -0.11262 -0.01162  0.12735  0.63696
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.6313215  0.0258848  140.29   <2e-16 ***
## X1          -0.0028211  0.0001229  -22.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1863 on 218 degrees of freedom
## Multiple R-squared:  0.7073, Adjusted R-squared:  0.7059
## F-statistic: 526.7 on 1 and 218 DF,  p-value: < 2.2e-16
```

```
linModel4$coefficients
```

```
##  (Intercept)           X1
##  3.631321505 -0.002821058
```

```
anova(linModel4)
```

```
## Analysis of Variance Table
##
## Response: log(Y)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X1          1 18.2788 18.2788  526.66 < 2.2e-16 ***
## Residuals 218  7.5661  0.0347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

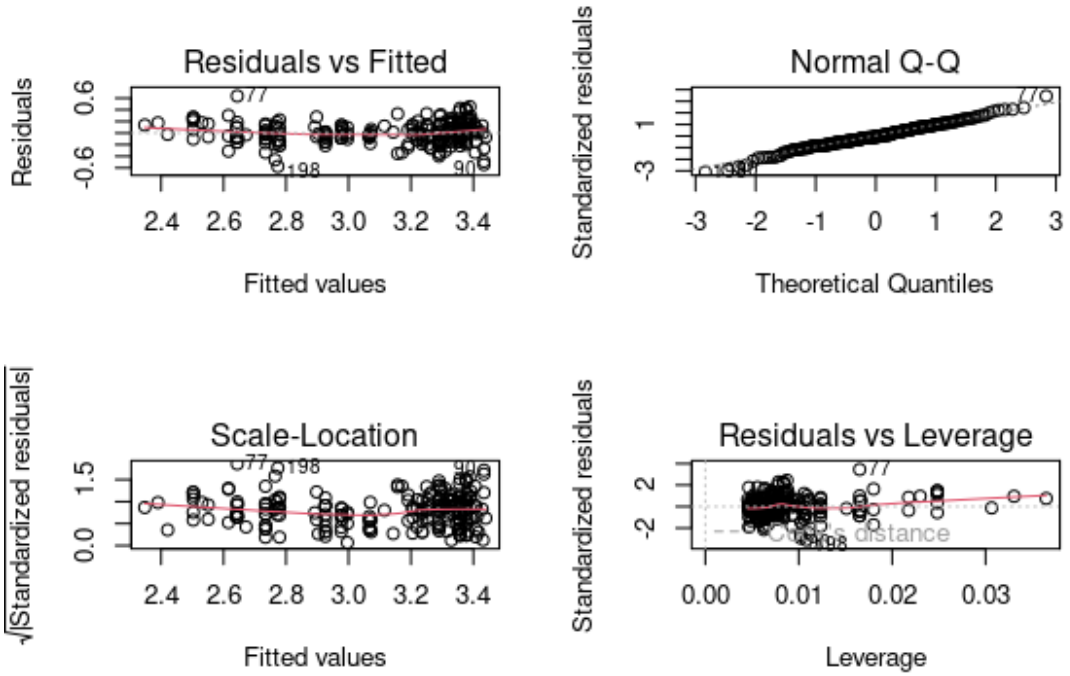
Tableau des coefficients de régression

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept) (β0)	3.6313215	0.0258848	140.29	<2e-16
X1 (β1)	-0.0028211	0.0001229	-22.95	<2e-16

Tableau d'analyse de la variance

	Df	Sum sq	Mean sq	F value	Pr(>F)
X1	1	18.2788	18.2788	526.66	< 2.2e-16
Residuals	218	7.5661	0.0347		

La F value est plus grande que dans les modèles précédents suggérant que X1 exercerait une influence significative. Toutefois, la valeur de B1 est très faible.




```
shapiro.test(rstudent(linModel4))

##
##  Shapiro-Wilk normality test
##
## data:  rstudent(linModel4)
## W = 0.99109, p-value = 0.1977
```

Le test de significativité

On doit comparer la valeur de notre p-value avec notre alpha (0,05). Si la p-value est supérieure à alpha, alors il est possible de conclure que le modèle n'est pas significatif. On obtient 0.1977 > 0,05, donc notre modèle peut être non significatif. De plus, on peut affirmer que notre échantillon suit une loi normale.

Residuals vs Fitted

Dans ce graphique, si les résidus (points) suivent droite, il est possible d'affirmer que ces derniers suivent une tendance linéaire.

Les résidus forment une courbe de trajectoire qui n'est pas droite. Il serait donc possible d'affirmer que les résidus ne suivent pas une tendance linéaire.

Normal Q-Q

Ce graphique permet de vérifier si les résidus suivent une distribution normale, en vérifiant si les points sont tous autour de la droite principale.

On peut apercevoir que les résidus sont relativement bien alignés avec la droite sauf aux extrémités. On pourrait donc penser que c'est une distribution normale et le test de Shapiro-Wilk nousle confirme.

Scale-Location

Ce graphique nous montre si les résidus sont étalés de façon équilibrée, il nous donnera des informations sur l'homoscédasticité des résidus. Si les valeurs sont réparties de façon aléatoire, cela est signe d'une mauvaise homoscédasticité.

On aperçoit que les valeurs sont reparties de façon aléatoire, on pourrait donc affirmer que les résidus ne font pas preuve d'homoscédasticité.

Residuals vs Leverage

Ce graphique permet de vérifier si des valeurs ont une plus grande influence sur les autres données, ces valeurs seront alors dans la distance de Cook.

On ne retrouve pas de résidus dans la distance de Cook. En conséquent, on pourrait affirmer que peu de points ont une grande influence sur les données.

Modèle 5

```
##
## Call:
## lm(formula = Y ~ X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8378  -2.9317  -0.4811   2.3942  16.6554
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.0413411  1.0796593   42.64   <2e-16 ***
## X2          -0.0076288  0.0003552  -21.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.472 on 218 degrees of freedom
## Multiple R-squared:  0.6791, Adjusted R-squared:  0.6776
## F-statistic: 461.3 on 1 and 218 DF,  p-value: < 2.2e-16

linModel5$coefficients

##  (Intercept)           X2
## 46.041341096 -0.007628798

anova(linModel5)

## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X2          1 9227.7  9227.7  461.33 < 2.2e-16 ***
## Residuals 218 4360.5    20.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

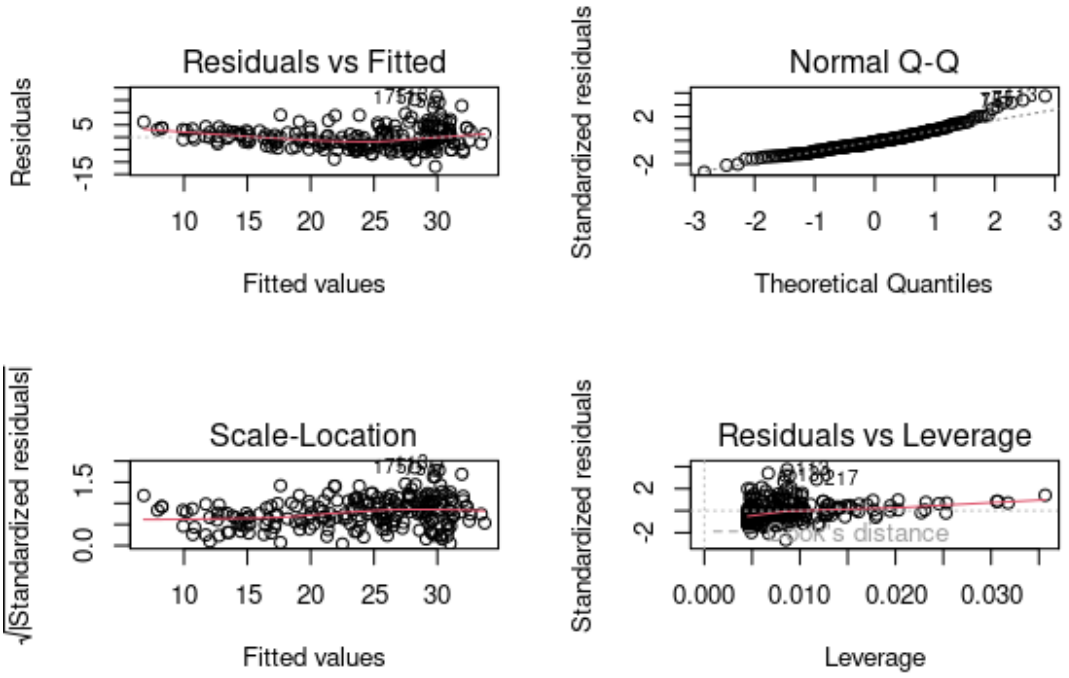
Tableau des coefficients de régression

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept) (β0)	46.0413411	1.0796593	42.64	<2e-16
X1 (β1)	-0.0076288	0.0003552	-21.48	<2e-16

Tableau d'analyse de la variance

	Df	Sum sq	Mean sq	F value	Pr(>F)
X1	1	9227.7	9227.7	461.33	< 2.2e-16
Residuals	218	4360.5	20.0		

Les résultats obtenus dans ce modèle 5 sont très similaires aux résultats du modèle 1, toutefois dans ce cas, notre X1 est le poids du véhicule. Encore une autre fois, le B1 est très faible. La F value est relativement grande et suggère que le poids est une variable significative.




```
shapiro.test(rstudent(linModel5))

##
##  Shapiro-Wilk normality test
##
## data:  rstudent(linModel5)
## W = 0.9526, p-value = 1.213e-06
```

Le test de significativité

On doit comparer la valeur de notre p-value avec notre alpha (0,05). Si la p-value est supérieure à alpha, alors il est possible de conclure que le modèle n'est pas significatif. On obtient 1.213e-06< 0,05, donc notre modèle peut être significatif. De plus, on peut affirmer que notre échantillon ne suit pas une loi normale.

Residuals vs Fitted

Dans ce graphique, si les résidus (points) suivent droite, il est possible d'affirmer que ces derniers suivent une tendance linéaire.

Les résidus forment une courbe de trajectoire qui n'est pas droite. Il serait donc possible d'affirmer que les résidus ne suivent pas une tendance linéaire.

Normal Q-Q

Ce graphique permet de vérifier si les résidus suivent une distribution normale, en vérifiant si les points sont tous autour de la droite principale.

On peut apercevoir que les résidus sont relativement bien alignés avec la droite sauf aux extrémités. On pourrait donc penser que c'est une distribution normale, mais le test de Shapiro-Wilk nous confirme le contraire.

Scale-Location

Ce graphique nous montre si les résidus sont étalés de façon équilibrée, il nous donnera des informations sur l'homoscédasticité des résidus. Si les valeurs sont réparties de façon aléatoire, cela est signe d'une mauvaise homoscédasticité.

On aperçoit que les valeurs sont reparties de façon aléatoire, on pourrait donc affirmer que les résidus ne font pas preuve d'homoscédasticité.

Residuals vs Leverage

Ce graphique permet de vérifier si des valeurs ont une plus grande influence sur les autres données, ces valeurs seront alors dans la distance de Cook.

On ne retrouve pas de résidus dans la distance de Cook. En conséquent, on pourrait affirmer que peu de points ont une grande influence sur les données.

Modèle 6

```
##
## Call:
## lm(formula = Y ~ X2^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8378  -2.9317  -0.4811   2.3942  16.6554
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.0413411  1.0796593   42.64   <2e-16 ***
## X2          -0.0076288  0.0003552  -21.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.472 on 218 degrees of freedom
## Multiple R-squared:  0.6791, Adjusted R-squared:  0.6776
## F-statistic: 461.3 on 1 and 218 DF,  p-value: < 2.2e-16

linModel6$coefficients

##  (Intercept)           X2
## 46.041341096 -0.007628798

anova(linModel6)

## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X2          1 9227.7  9227.7  461.33 < 2.2e-16 ***
## Residuals 218 4360.5    20.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

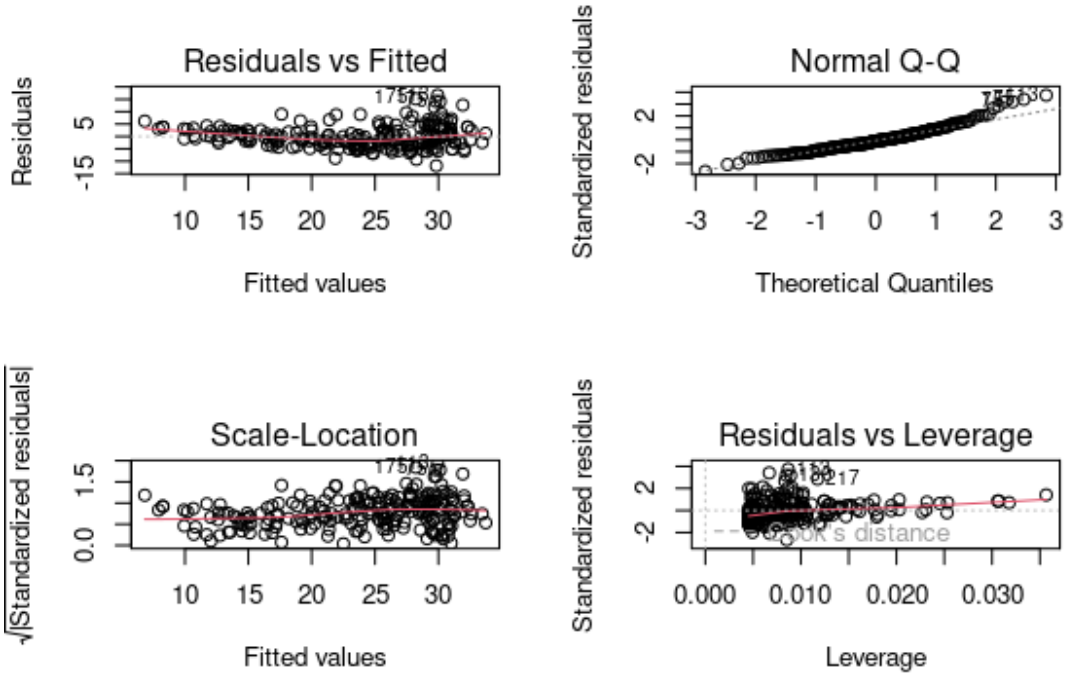
Tableau des coefficients de régression

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept) (β0)	46.0413411	1.0796593	42.64	<2e-16
X1 (β1)	-0.0076288	0.0003552	-21.48	<2e-16

Tableau d'analyse de la variance

	Df	Sum sq	Mean sq	F value	Pr(>F)
X1	1	9227.7	9227.7	461.33	< 2.2e-16
Residuals	218	4360.5	20.0		

Tout comme le modèle 2, le modèle 6 vérifie si on suit une tendance quadratique, mais avec X1 étant le poids du véhicule. Les valeurs sont en conséquent identiques au modèle 5, car il n'y a pas de tendance quadratique.




```
shapiro.test(rstudent(linModel6))

##
##  Shapiro-Wilk normality test
##
## data:  rstudent(linModel6)
## W = 0.9526, p-value = 1.213e-06
```

Le test de significativité

On doit comparer la valeur de notre p-value avec notre alpha (0,05). Si la p-value est supérieure à alpha, alors il est possible de conclure que le modèle n'est pas significatif. On obtient 1.213e-06< 0,05, donc notre modèle peut être significatif. De plus, on peut affirmer que notre échantillon ne suit pas une loi normale.

Residuals vs Fitted

Dans ce graphique, si les résidus (points) suivent droite, il est possible d'affirmer que ces derniers suivent une tendance linéaire.

Les résidus forment une courbe de trajectoire qui n'est pas droite. Il serait donc possible d'affirmer que les résidus ne suivent pas une tendance linéaire.

Normal Q-Q

Ce graphique permet de vérifier si les résidus suivent une distribution normale, en vérifiant si les points sont tous autour de la droite principale.

On peut apercevoir que les résidus sont relativement bien alignés avec la droite sauf aux extrémités. On pourrait donc penser que c'est une distribution normale, mais le test de Shapiro-Wilk nous confirme le contraire.

Scale-Location

Ce graphique nous montre si les résidus sont étalés de façon équilibrée, il nous donnera des informations sur l'homoscédasticité des résidus. Si les valeurs sont réparties de façon aléatoire, cela est signe d'une mauvaise homoscédasticité.

On aperçoit que les valeurs sont reparties de façon aléatoire, on pourrait donc affirmer que les résidus ne font pas preuve d'homoscédasticité.

Residuals vs Leverage

Ce graphique permet de vérifier si des valeurs ont une plus grande influence sur les autres données, ces valeurs seront alors dans la distance de Cook.

On ne retrouve pas de résidus dans la distance de Cook. En conséquent, on pourrait affirmer que peu de points ont une grande influence sur les données.

Modèle 7

```
##
## Call:
## lm(formula = log(Y) ~ log(X2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51720 -0.10534 -0.00395  0.09967  0.47163
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.56674    0.32020   36.12   <2e-16 ***
## log(X2)     -1.06502    0.04031  -26.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.168 on 218 degrees of freedom
## Multiple R-squared:  0.762,  Adjusted R-squared:  0.7609
## F-statistic: 698.1 on 1 and 218 DF,  p-value: < 2.2e-16
```

```
linModel7$coefficients
```

```
## (Intercept)     log(X2)
##   11.566735   -1.065018
```

```
anova(linModel7)
```

```
## Analysis of Variance Table
##
## Response: log(Y)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## log(X2)     1 19.6947 19.6947   698.1 < 2.2e-16 ***
## Residuals 218  6.1502  0.0282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

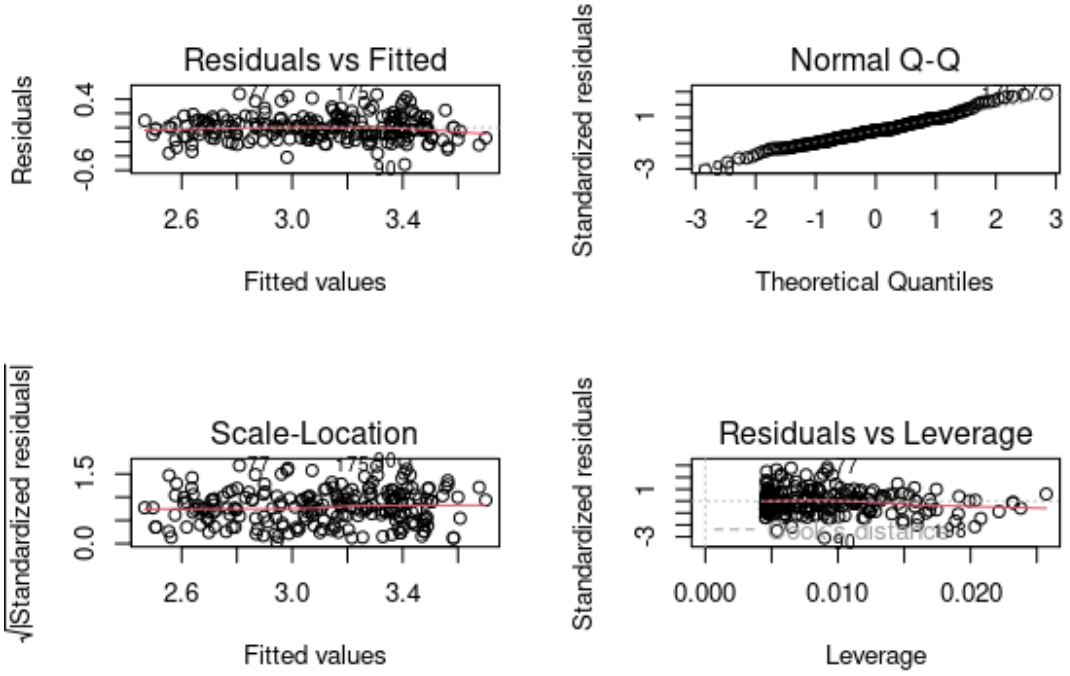
Tableau des coefficients de régression

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept) (β0)	11.56674	0.32020	36.12	<2e-16
X1 (β1)	-1.06502	0.04031	-26.42	<2e-16

Tableau d'analyse de la variance

	Df	Sum sq	Mean sq	F value	Pr(>F)
X1	1	19.6947	19.6947	698.1	< 2.2e-16
Residuals	218	6.1502	0.0282		

On a une grande F value suggérant que la variable est significative.




```
shapiro.test(rstudent(linModel7))

##
##  Shapiro-Wilk normality test
##
## data:  rstudent(linModel7)
## W = 0.98668, p-value = 0.04696
```

Le test de significativité

On doit comparer la valeur de notre p-value avec notre alpha (0,05). Si la p-value est supérieure à alpha, alors il est possible de conclure que le modèle n'est pas significatif. On obtient 0.04696 < 0,05, donc notre modèle peut être significatif. De plus, on peut affirmer que notre échantillon ne suit pas une loi normale.

Residuals vs Fitted

Dans ce graphique, si les résidus (points) suivent droite, il est possible d'affirmer que ces derniers suivent une tendance linéaire.

Les résidus forment une courbe de trajectoire qui est droite. Il serait donc possible d'affirmer que les résidus suivent une tendance linéaire.

Normal Q-Q

Ce graphique permet de vérifier si les résidus suivent une distribution normale, en vérifiant si les points sont tous autour de la droite principale.

On peut apercevoir que les résidus sont relativement bien alignés avec la droite sauf aux extrémités. On pourrait donc penser que c'est une distribution normale, mais le test de Shapiro-Wilk nous confirme le contraire.

Scale-Location

Ce graphique nous montre si les résidus sont étalés de façon équilibrée, il nous donnera des informations sur l'homoscédasticité des résidus. Si les valeurs sont réparties de façon aléatoire, cela est signe d'une mauvaise homoscédasticité.

On aperçoit que les valeurs sont reparties de façon aléatoire, on pourrait donc affirmer que les résidus ne font pas preuve d'homoscédasticité.

Residuals vs Leverage

Ce graphique permet de vérifier si des valeurs ont une plus grande influence sur les autres données, ces valeurs seront alors dans la distance de Cook.

On ne retrouve pas de résidus dans la distance de Cook. En conséquent, on pourrait affirmer que peu de points ont une grande influence sur les données.

Modèle 8

```
##
## Call:
## lm(formula = log(Y) ~ X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50237 -0.10633 -0.00992  0.09648  0.45388
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.143e+00  4.021e-02  103.05   <2e-16 ***
## X2          -3.533e-04  1.323e-05  -26.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1666 on 218 degrees of freedom
## Multiple R-squared:  0.766,  Adjusted R-squared:  0.7649
## F-statistic: 713.6 on 1 and 218 DF,  p-value: < 2.2e-16

linModel8$coefficients

##   (Intercept)            X2
##  4.1432714372 -0.0003533544

anova(linModel8)

## Analysis of Variance Table
##
## Response: log(Y)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X2          1 19.7972 19.7972  713.63 < 2.2e-16 ***
## Residuals 218  6.0477  0.0277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

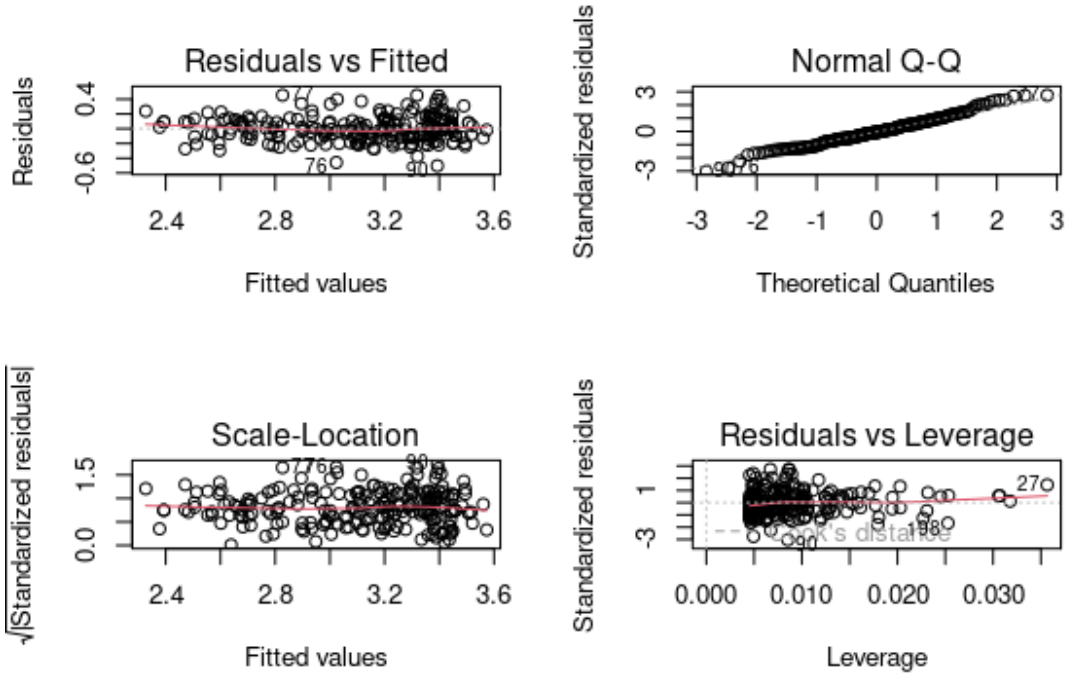
Tableau des coefficients de régression

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept) (β0)	4.143e+00	4.021e-02	103.05	<2e-16
X1 (β1)	-3.533e-04	1.323e-05	-26.71	<2e-16

Tableau d'analyse de la variance

	Df	Sum sq	Mean sq	F value	Pr(>F)
X1	1	19.7972	19.7972	713.63	< 2.2e-16
Residuals	218	6.0477	0.0277		

On a une grande F value suggérant que la variable est significative.




```
shapiro.test(rstudent(linModel8))

##
##  Shapiro-Wilk normality test
##
## data:  rstudent(linModel8)
## W = 0.98725, p-value = 0.05519
```

Le test de significativité

On doit comparer la valeur de notre p-value avec notre alpha (0,05). Si la p-value est supérieure à alpha, alors il est possible de conclure que le modèle n'est pas significatif. On obtient 0. 0.05519 > 0,05, donc notre modèle peut être non significatif. De plus, on peut affirmer que notre échantillon suit une loi normale.

Residuals vs Fitted

Dans ce graphique, si les résidus (points) suivent droite, il est possible d'affirmer que ces derniers suivent une tendance linéaire.

Les résidus forment une courbe de trajectoire qui est droite. Il serait donc possible d'affirmer que les résidus suivent une tendance linéaire.

Normal Q-Q

Ce graphique permet de vérifier si les résidus suivent une distribution normale, en vérifiant si les points sont tous autour de la droite principale.

On peut apercevoir que les résidus sont relativement bien alignés avec la droite sauf aux extrémités. On pourrait donc penser que c'est une distribution normale et le test de Shapiro-Wilk nous le confirme.

Scale-Location

Ce graphique nous montre si les résidus sont étalés de façon équilibrée, il nous donnera des informations sur l'homoscédasticité des résidus. Si les valeurs sont réparties de façon aléatoire, cela est signe d'une mauvaise homoscédasticité.

On aperçoit que les valeurs sont reparties de façon aléatoire, on pourrait donc affirmer que les résidus ne font pas preuve d'homoscédasticité.

Residuals vs Leverage

Ce graphique permet de vérifier si des valeurs ont une plus grande influence sur les autres données, ces valeurs seront alors dans la distance de Cook.

On ne retrouve pas de résidus dans la distance de Cook. En conséquent, on pourrait affirmer que peu de points ont une grande influence sur les données.

• (3 points) Donner un intervalle de confiance pour chacun des paramètres β0 et β1 des modèles 1 et 5.

```
confint(linModel1)

##                    2.5 %      97.5 %
## (Intercept) 33.61539919 36.27354753
## X1          -0.06697816 -0.05435467

confint(linModel5)

##                     2.5 %       97.5 %
## (Intercept) 43.913434591 48.169247600
## X2          -0.008328828 -0.006928768
```

Modèle	Intervalle de confiance de B0	Intervalle de confiance de B1
Modèle 1	[33.61539919; 36.27354753]	[-0.06697816 ; -0.05435467]
Modèle 2	[43.913434591 ; 48.169247600}]	[-0.008328828 ; -0.006928768]

Étant donné qu'on ne retrouve pas 0 dans les intervalles de confiance, on peut affirmer que les deux modèles sont significatifs.

• (2 points) En conclusion : effectuez une comparaison et dire lequel des huit modèles est préférable aux autres. Justifiez votre choix en précisant les critères utilisés.

p-value > 0.05 : 4, 7 (sa p-value étant très proche de 0,05) , 8
p-value < 0.05 : 1, 2, 3, 5, 6

R^2 modèle 1 : 0.6204
R^2 modèle 2 : 0.6204
R^2 modèle 3 : 0.7065
R^2 modèle 4 : 0.7059
R^2 modèle 5 : 0.6776
R^2 modèle 6 : 0.6776
R^2 modèle 7 : 0.7609
R^2 modèle 8 : 0.7649

En se fiant qu'au R2, on pourrait croire que les modèles 8,7,4,3 seraient les meilleurs, car ce coefficient de corrélation est un indicateur de la qualité du modèle. Toutefois, il faudrait considérer d'autres critères afin de faire notre choix. Prenons les graphiques que nous avons générés. Les modèles 3,4 et 8 ont des graphiques Q-Q qui confirment qu'ils suivent une distribution normale. De plus, le test de Shapiro-Wilk nous dit que les modèles 4,7 et 8 suivent une distribution normale. Par élimination, il nous reste donc les modèles 4 et 8. **Le modèle 8** ayant le plus grand coefficient de corrélation sera retenu.

Maintenant, penchons-nous sur l'aspect logique. La cylindrée du moteur du véhicule exerce une influence significative sur l'efficacité du véhicule ou bien le poids du véhicule a une influence significative sur l'efficacité du véhicule. Selon l'aspect logique, plus un véhicule est lourd, moins il sera efficace, donc le poids exerce une influence directe sur l'efficacité du véhicule. La cylindrée du moteur peut également exercer une influence sur l'efficacité du véhicule, mais elle sera moindre.

E) (5 points)

Sur la base du meilleur modèle que vous avez obtenu en d), calculez un intervalle de prévision pour l'efficacité en carburant d'un véhicule ayant les caractéristiques suivantes : X1 = 190 ; X2 = 2500. Commentez brièvement votre résultat.
Remarque. Notez que le modèle que vous avez obtenu en d) n'utilise pas nécessairement toutes les valeurs ci-dessus.

```
      ##        fit      lwr      upr
## 1 3.259885 2.930689 3.589082
```

L'intervalle de prévision du modèle 8 se situe entre [2.930689; 3.589082]. Y a une probabilité de 0.95 de se situer dans cet intervalle pour des valeurs futures.

Annexe

Code

````
---
title: "Devoir A2022"
author: "Marsel Bakashov"
date: "`r format(Sys.time(), '%d %b %Y')`"
output:
  word_document: default
  html_notebook: default
  pdf_document: default
lang: fr
---
```{r setup, message = FALSE}
library(dplyr)
library(knitr)
library(ggplot2)
library(gtsummary)

opts_chunk$set(
  fig.width = 6,
  fig.height = 4,
  fig.align = 'center'
)

reset_gtsummary_theme()
```

```{r}
# Fonction pour calculer le coefficient de variation.
cv <- function(x) {
  return(sd(x) / mean(x))
}

# Fonction pour calculer l'écart interquartile.
iqr <- function(x) {
  return(quantile(x, .75) - quantile(x, .25))
}
```

```{r}
source('charger.R')
mondata<-charger(2147174)
mondata
```
#Phase 1 : Analyse statistique descriptive et inference

a) (2 points) Examinez les liens entre les variables quantitatives de l'étude. Pour cela, produisez une matrice des corrélations pour l'ensemble des trois variables quantitatives et commentez brièvement
```{r}
mcor <- cor(mondata[,1:3])
round(mcor,2)
```

b) (8 points) Pour chacune des trois variables Y (l'efficacité en carburant), X1 (la cylindrée) et X2 (le poids), produisez
les graphiques et les tableaux demandés ci-dessous et interprétez brièvement les résultats dans chaque cas :
````

• un histogramme et un diagramme de Tukey (ou «Box Plot») ;

```{r}
mondata%>%
  ggplot(aes(x=mpg)) +
  geom_histogram(
    col = "red",
    fill = "green",
    bins = 10
  ) +
  labs(x='L’efficacité en carburant du véhicule (en milles par gallon)
', y ='Frequence') +
  ggtitle('Histogramme de l’efficacité en carburant du véhicule') +
  theme(plot.title = element_text(hjust = 0.5))

mondata%>%
  ggplot(aes(x=displacement)) +
  geom_histogram(
    col = "red",
    fill = "green",
    bins = 10
  ) +
  labs(x='La cylindrée du moteur du véhicule (en pouces cubes)
', y ='Frequence') +
  ggtitle('Histogramme de la cylindrée du moteur du véhicule') +
  theme(plot.title = element_text(hjust = 0.5))

mondata%>%
  ggplot(aes(x=weight)) +
  geom_histogram(
    col = "red",
    fill = "green",
    bins = 10
  ) +
  labs(x='Le poids du véhicule (en livres)
', y ='Frequence') +
  ggtitle('Histogramme du poids du véhicule') +
  theme(plot.title = element_text(hjust = 0.5))

mondata%>%
  ggplot(aes(x=mpg)) +
  geom_boxplot(
    color = 'blue',
    fill = 'red',
    alpha = 0.5,
    width = 0.2,
    outlier.color = 'black',
    outlier.fill = 'black',
    outlier.alpha = 1,
    outlier.size = 4
  ) +
ylim(-0.5, 0.5) +
labs(x='L’efficacité en carburant du véhicule') +
theme(
  axis.ticks.y = element_blank(),
  axis.text.y = element_blank()
)
```

```
mondata%>%
  ggplot(aes(x=displacement)) +
  geom_boxplot(
    color = 'blue',
    fill = 'red',
    alpha = 0.5,
    width = 0.2,
    outlier.color = 'black',
    outlier.fill = 'black',
    outlier.alpha = 1,
    outlier.size = 4
  ) +
ylim(-0.5, 0.5) +
labs(x='La cylindrée du moteur du véhicule') +
theme(
  axis.ticks.y = element_blank(),
  axis.text.y = element_blank()
)

mondata%>%
  ggplot(aes(x=weight)) +
  geom_boxplot(
    color = 'blue',
    fill = 'red',
    alpha = 0.5,
    width = 0.2,
    outlier.color = 'black',
    outlier.fill = 'black',
    outlier.alpha = 1,
    outlier.size = 4
  ) +
ylim(-0.5, 0.5) +
labs(x='Le poids du véhicule') +
theme(
  axis.ticks.y = element_blank(),
  axis.text.y = element_blank()
)

```
• une droite de Henry (ou «Normal Probability Plot») et un test de normalité (Shapiro-Wilk) ;
```{r}
mondata%>%
  ggplot(aes(sample=mpg)) +
  stat_qq(
    alpha = 0.5,
    size = 1.2
  ) +
  stat_qq_line(
    color = 'red'
  )+
  labs(x='Normal Scores
', y ='Standardized Residuals') +
  ggtitle('Droite de Henry de l’efficacité en carburant du véhicule') +
  theme(plot.title = element_text(hjust = 0.5))

mondata%>%
  ggplot(aes(sample=displacement)) +
```

```
  stat_qq(
    alpha = 0.5,
    size = 1.2
  ) +
  stat_qq_line(
    color = 'red'
  )+
  labs(x='Normal Scores
', y ='Standardized Residuals') +
  ggtitle('Droite de Henry de la cylindrée du moteur du véhicule
') +
  theme(plot.title = element_text(hjust = 0.5))

mondata%>%
  ggplot(aes(sample=weight)) +
  stat_qq(
    alpha = 0.5,
    size = 1.2
  ) +
  stat_qq_line(
    color = 'red'
  )+
  labs(x='Normal Scores
', y ='Standardized Residuals') +
  ggtitle('Droite de Henry du poids du véhicule
') +
  theme(plot.title = element_text(hjust = 0.5))


shapiro.test(mondata$mpg)
shapiro.test(mondata$displacement)
shapiro.test(mondata$weight)
```
• un tableau de statistiques descriptives comprenant : moyenne, quartiles, écart type, erreur type, intervalle de confiance pour la moyenne ;
```{r}
mondata%>%
  tbl_summary(
    include = mpg,
    label = mpg ~ "mpg",
    type = mpg ~ "continuous2",
    statistic = mpg ~ c(
      "{mean}",
      "{sd}",
      "{se}",
      "{min}",
      "{max}",
      "{p25}",
      "{median}",
      "{p75}"
      )
    )-> mpg.stats.table

mpg.stats.table %>%
  add_stat_label(
    label = mpg ~ c(
      "Moyenne",
```

```
      "Écart-type",
      "Erreur type",
      "Minimum",
      "Maximum",
      "Premier quartile",
      "Médiane",
      "Troisième quartile"
    )
  ) -> mpg.stats.table

mpg.stats.table %>%
  modify_header(
    label ~ "**Statistiques descriptives**",
    all_stat_cols() ~ ""
  ) -> mpg.stats.table

mpg.stats.table %>%
  bold_labels() %>%
  italicize_levels() -> mpg.stats.table

mpg.stats.table %>%
  add_ci()

mondata%>%
  tbl_summary(
    include = displacement,
    label = displacement ~ "displacement",
    type = displacement ~ "continuous2",
    statistic = displacement ~ c(
      "{mean}",
      "{sd}",
      "{se}",
      "{min}",
      "{max}",
      "{p25}",
      "{median}",
      "{p75}"
      )
    )-> displacement.stats.table

displacement.stats.table %>%
  add_stat_label(
    label = displacement ~ c(
      "Moyenne",
      "Écart-type",
      "Erreur type",
      "Minimum",
      "Maximum",
      "Premier quartile",
      "Médiane",
      "Troisième quartile"
    )
  ) -> displacement.stats.table

displacement.stats.table %>%
  modify_header(
    label ~ "**Statistiques descriptives**",
```

```
    all_stat_cols() ~ ""
  ) -> displacement.stats.table

displacement.stats.table %>%
  bold_labels() %>%
  italicize_levels() -> displacement.stats.table

displacement.stats.table %>%
  add_ci()

mondata%>%
  tbl_summary(
    include = weight,
    label = weight ~ "weight",
    type = weight ~ "continuous2",
    statistic = weight ~ c(
      "{mean}",
      "{sd}",
      "{se}",
      "{min}",
      "{max}",
      "{p25}",
      "{median}",
      "{p75}"
      )
    )-> weight.stats.table

weight.stats.table %>%
  add_stat_label(
    label = weight ~ c(
      "Moyenne",
      "Écart-type",
      "Erreur type",
      "Minimum",
      "Maximum",
      "Premier quartile",
      "Médiane",
      "Troisième quartile"
    )
  ) -> weight.stats.table

weight.stats.table %>%
  modify_header(
    label ~ "**Statistiques descriptives**",
    all_stat_cols() ~ ""
  ) -> weight.stats.table

weight.stats.table %>%
  bold_labels() %>%
  italicize_levels() -> weight.stats.table

weight.stats.table %>%
  add_ci()

```

c) (8 points) Afin de vérifier si l'efficacité en carburant d'un véhicule dépend de l'origine de celui-ci, on peut considérer deux groupes de données selon la variable origin et effectuer une comparaison des deux groupes en termes de moyenne,symétrie et variabilité. Pour ce faire, effectuez les analyses suivantes et donnez une brève conclusion :

• deux histogrammes juxtaposés, et deux diagrammes de Tukey (ou «Box Plot») juxtaposés ;

```{r}
origin.labeller <- function(n) {
  return(paste("Origin", n))
}

mondata%>%
  filter(origin %in% c(0,1))%>%
  ggplot(aes(x=mpg)) +
  geom_histogram(
    col = "red",
    fill = "green",
    bins = 10
  ) +
  facet_wrap(
    ~factor(origin),
    labeller = as_labeller(origin.labeller)
  )+
  labs(x='L'efficacité en carburant du véhicule (en milles par gallon)', y ='Frequence') +
  ggtitle('L'efficacité en carburant du véhicule en fonction de leur origine') +
  theme(plot.title = element_text(hjust = 0.5))

mondata%>%
  ggplot(aes(x=mpg, fill=origin)) +
  geom_boxplot(
    alpha = 0.5,
    width = 0.2
  ) +
  facet_wrap(
    ~origin, nrow=1
  )+
  ylim(-0.5, 0.5) +
  labs(x='L'efficacité en carburant du véhicule') +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank()
)
```

• un tableau des statistiques descriptives par groupe : moyenne, quartiles, variance, écart type, intervalle de confiance pour la moyenne ;

```{r}
mondata%>%
  tbl_summary(
    include = mpg,
    by = origin,
    label = mpg ~ "mpg",
    type = mpg ~ "continuous2",
    statistic = mpg ~ c(
      "{mean}",
      "{var}",
      "{sd}",
```

```
      "{min}",
      "{max}",
      "{p25}",
      "{median}",
      "{p75}"
      )
    )-> mpg.stats.table

mpg.stats.table %>%
  add_stat_label(
    label = mpg ~ c(
      "Moyenne",
      "Variance",
      "Écart-type",
      "Minimum",
      "Maximum",
      "Premier quartile",
      "Médiane",
      "Troisième quartile"
    )
  ) -> mpg.stats.table

mpg.stats.table %>%
  modify_header(label ~ "**Statistiques descriptives**") %>%
  modify_header(all_stat_cols() ~ "*{level}*"
)-> mpg.stats.table

mpg.stats.table %>%
  bold_labels() %>%
  italicize_levels() -> mpg.stats.table

mpg.stats.table %>%
  add_ci()
```
• un test d'hypothèses sur l'égalité des variances des deux groupes ;
```{r}
var.test(mpg ~ origin, data=mondata, alternative = "two.sided")
```
• un test d'hypothèses sur l'égalité des moyennes des deux groupes.
```{r}
t.test(mpg ~ origin, data=mondata, alternative = "two.sided")
```
#Phase 2 : Recherche d'un modèle
On s'intéresse dans cette phase à la détermination d'un modèle permettant d'expliquer la performance d'un véhicule en
fonction des différentes variables de l'étude. Pour ce faire, on envisage des modèles de régression en considérant l'efficacité en carburant comme variable dépendante, Y .
d) (15 points) On considère les huit modèles suivants :
Pour chacun des huit modèles ci-dessus :

• (5 points) Effectuez l'ajustement (i.e. obtenir le tableau des coefficients de régression, le tableau d'analyse de la variance).
```{r}
Y <- mondata$mpg
X1 <- mondata$displacement
X2 <- mondata$weight
X3 <- mondata$origin
```

```
linModel1 <- lm(Y ~ X1)
summary(linModel1)
linModel1$coefficients
anova(linModel1)
```

```{r}
linModel2 <- lm(Y ~ X1^2)
summary(linModel2)
linModel2$coefficients
anova(linModel2)
```

```{r}
linModel3 <- lm(log(Y) ~log(X1) )
summary(linModel3)
linModel3$coefficients
anova(linModel3)
```

```{r}
linModel4 <- lm(log(Y) ~X1)
summary(linModel4)
linModel4$coefficients
anova(linModel4)
```

```{r}
linModel5 <- lm(Y ~ X2)
summary(linModel5)
linModel5$coefficients
anova(linModel5)
```

```{r}
linModel6 <- lm(Y ~ X2^2)
summary(linModel6)
linModel6$coefficients
anova(linModel6)
```

```{r}
linModel7 <- lm(log(mondata$mpg) ~ log(mondata$weight) )
summary(linModel7)
linModel7$coefficients
anova(linModel7)
```

```{r}
linModel8 <- lm(log(Y) ~ X2)
summary(linModel8)
linModel8$coefficients
anova(linModel8)
```

• (5 points) Tester la signification du modèle et effectuez une analyse des résidus (normalité, homoscédasticité, points atypiques, etc.)

```{r}
par(mfrow = c(2,2))
plot(linModel1)
shapiro.test(rstudent(linModel1))
```

```{r}
par(mfrow = c(2,2))
plot(linModel2)
shapiro.test(rstudent(linModel2))
```

```{r}
par(mfrow = c(2,2))
plot(linModel3)
shapiro.test(rstudent(linModel3))
```

```{r}
par(mfrow = c(2,2))
plot(linModel4)
shapiro.test(rstudent(linModel4))
```

```{r}
par(mfrow = c(2,2))
plot(linModel5)
shapiro.test(rstudent(linModel5))
```

```{r}
par(mfrow = c(2,2))
plot(linModel6)
shapiro.test(rstudent(linModel6))
```

```{r}
par(mfrow = c(2,2))
plot(linModel7)
shapiro.test(residuals(linModel7))
```

```{r}
par(mfrow = c(2,2))
plot(linModel8)
shapiro.test(residuals(linModel8))
```

• (3 points) Donner un intervalle de confiance pour chacun des paramètres β0 et β1 des modèles 1 et 5.

```{r}
confint(linModel1)
```

```{r}
confint(linModel5)
```

• (2 points) En conclusion : effectuez une comparaison
```{r}

```
e) (5 points) Sur la base du meilleur modèle que vous avez obtenu en d), calculez un intervalle de prévision pour l'efficacité en carburant d'un véhicule ayant les caractéristiques suivantes : X1 = 190; X2 = 2500.
Commentez brièvement votre résultat.
Remarque. Notez que le modèle que vous avez obtenu en d) n'utilise pas nécessairement toutes les valeurs ci-dessus
```{r}
predict(linModel8, newdata = data.frame(X1 = 190, X2 = 2500), interval = "predict", level = 0.95)
```