ECOLE NATIONALE

DES SCIENCES APPLIQUÉES OUJDA

Rapport de Projet
Modélisation Statistique

Sujet de Projet:

La Régression linéaire simple est multiple pour la prédiction des prix des voitures

Préparé par : *YOUSSEF BOUJYDAH*
HICHAM GHOUCH
YOUSSEF OULADDEHOU

Encadrant : *Mme ELMEHDI Rachida*

Résumé/Abstract

La modélisation dans les mathématiques consiste à définir une expression mathématique qui décrit un phénomène réel en faisant introduire des variables (vecteurs au sens algébrique). C'est donc chercher pour ce phénomène un modèle adéquat, noté mathématiquement par une fonction qui est souvent une fonction inconnue et à déterminer. Par ailleurs, il existe des phénomènes dont lesquels il faut intervenir un certain ensemble de contraintes entre les variables qui doivent être prises en considération dans la détermination de la fonction.
Les modèles cherchés ont trois caractéristiques principales. Ils sont hypothétiques, modifiables et adéquats pour certains problèmes dans certaines situations ? Il n'y a donc pas le bon modèle d'une manière absolue.
La modélisation est un des objets de la statistique. De ce fait, la modélisation statistique consiste principalement dans la représentation des bases de données observées par des modèles théoriques qui décrivent au mieux ces données en utilisant des variables explicatives et des variables à expliquer et en tenant compte surtout de leur nature aléatoire. Elle sert entre autres à faire de l'inférence et à établir des prédictions afin d'en tirer des conclusions et d'en prendre des décisions.
Les méthodes de modélisation statistique sont diverses et variées mais les principales parmi elles sont le modèle linéaire (gaussien), le modèle linéaire généralisé, les modèles non linéaires, les modèles mixtes, les modèles pour données répétées, les modèles pour séries chronologiques, l'analyse discriminante et la classification, …

Dans ce projet, on va se focaliser sur le modèle linéaire (gaussien) : régression linéaire simple, régression linéaire multiple

Listes des figures

Table des matières

Introduction

1- **Problématique** :

Le problème qu'on traite dans ce projet c'est la prédiction des prix des voitures utilisées en utilisant la régression linéaire simple , multiple .

le jeu de données utilisé est "CarPriceDataset.csv"

La première partie de ce projet consiste à établir un modèle de la régression linéaire simple pour prédire le prix des voitures en fonction de la Puissance .

La deuxième partie de ce projet consiste à établir un modèle de la régression linéaire multiple pour prédire le prix des voitures en fonction de plusieurs critères .

2- présentation de dataset :

notre dataset est comme suit

```
> data<-read.csv("CarPriceDataset.csv")
> head(data)
  Year Kilometers_Driven Mileage Engine  Power Seats Price
1 2010             72000   26.60    998  58.16     5  1.75
2 2015             41000   19.67   1582 126.20     5 12.50
3 2011             46000   18.20   1199  88.70     5  4.50
4 2012             87000   20.77   1248  88.76     7  6.00
5 2013             40670   15.20   1968 140.80     5 17.74
6 2012             75000   21.10    814  55.20     5  2.35
>
```

Régression linéaire Simple

Introduction :

La régression linéaire simple est un modèle de régression linéaire avec une seule variable explicative . C'est-à-dire qu'il concerne des points d'échantillonnage à deux dimensions avec une variable indépendante et une variable dépendante (conventionnellement, les coordonnées *x* et *y* dans un système de coordonnées cartésiennes) et trouve un fonction linéaire (une ligne droite non verticale) qui, aussi précisément que possible, prédit les valeurs des variables dépendantes en fonction de la variable indépendante.

1-présentation de données :

le dataset utilisé pour la régression linéaire simple , avec y la variable à expliquer est le prix "Price" , et x la variable explicative est la puissance "Power", les nombres des variables d'entraînement n= 6019

```
> simpleData<-data[c("Power","Price")]
> head(simpleData)
   Power Price
1  58.16  1.75
2 126.20 12.50
3  88.70  4.50
4  88.76  6.00
5 140.80 17.74
6  55.20  2.35
> summary(simpleData)
     Power            Price
 Min.   : 34.2   Min.   :  0.440
 1st Qu.: 78.0   1st Qu.:  3.500
 Median : 98.6   Median :  5.640
 Mean   :113.3   Mean   :  9.479
 3rd Qu.:138.0   3rd Qu.:  9.950
 Max.   :560.0   Max.   :160.000
> |
```

Figure 1 : data summary

La commande sammray nous permet d'avoir une vision générale sur notre base de données

Et d'afficher les différents indices statistiques (la moyenne, ….)

2-Régression linéaire simple et *l'estimation des paramètres*:

```
>  SimpleRegressionModel<-lm(Price~Power,data = simpleData)
>
> summary(SimpleRegressionModel)

Call:
lm(formula = Price ~ Power, data = simpleData)

Residuals:
    Min      1Q  Median      3Q     Max
-57.352  -2.959   0.080   2.197 127.660

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.785420   0.217425  -40.41   <2e-16 ***
Power        0.161275   0.001737   92.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.175 on 6017 degrees of freedom
Multiple R-squared:  0.5888,    Adjusted R-squared:  0.5887
F-statistic:  8616 on 1 and 6017 DF,  p-value: < 2.2e-16
```

Figure 2: summary régression linéaire simple

La régression linéaire R utilise la fonction lm() pour créer un modèle de régression à partir de la base de données train, et la stocke dans la variable SimpleRegressionModel

La commande summary permet de donner un aperçu sur la régression sur la base de données.

L'estimation des paramètres :

L'intercepté (β_0) = -8.785420

La pente ($\beta 1$) = 0.161275

Donc la droite de régression de notre modèle est : **Y = -8.785420 + 0.161275 *X**

3.Coefficient de détermination et coefficient de corrélation linéaire :

Le coefficient de détermination exprime le rapport entre la variance expliquée par le modèle de régression et la variance totale. Il sert à évaluer la qualité de l'ajustement

On a ici le R^2 =0.5888

Donc le modèle n'ajuste pas bien le nuage de pointes

4- Tests d'hypothèse au niveau de risque 5 %:

Significativité des paramètres (du modèle) :

Test de student :

Pour β_0:

l'hypothèse :

H0 : $\beta_0 = 0$ VS H1 : $\beta_0 ! = 0$

Nous avons Pr <2e-16 est très inférieure à 0,05 donc on rejette l'hypothèse H0, alors β_0 est très significative dans notre modèle.

Pour β_1:

l'hypothèse : H0 : $\beta_1 = 0$ VS H1 : $\beta_1 ! = 0$

Nous avons Pr <2e-16 est largement inférieure à 0,05 donc on rejette l'hypothèse H0, alors β_1 est très significative dans notre modèle.

Alors on va tracer les données et la droit de régression linéaire simple

```
> plot(simpleData$Power,simpleData$Price,xlab = "power",ylab = "price",lwd=1)
> abline(-8.785420,0.161275,lwd=2)
>
```

Figure 3: commande plot régression linéaire simple

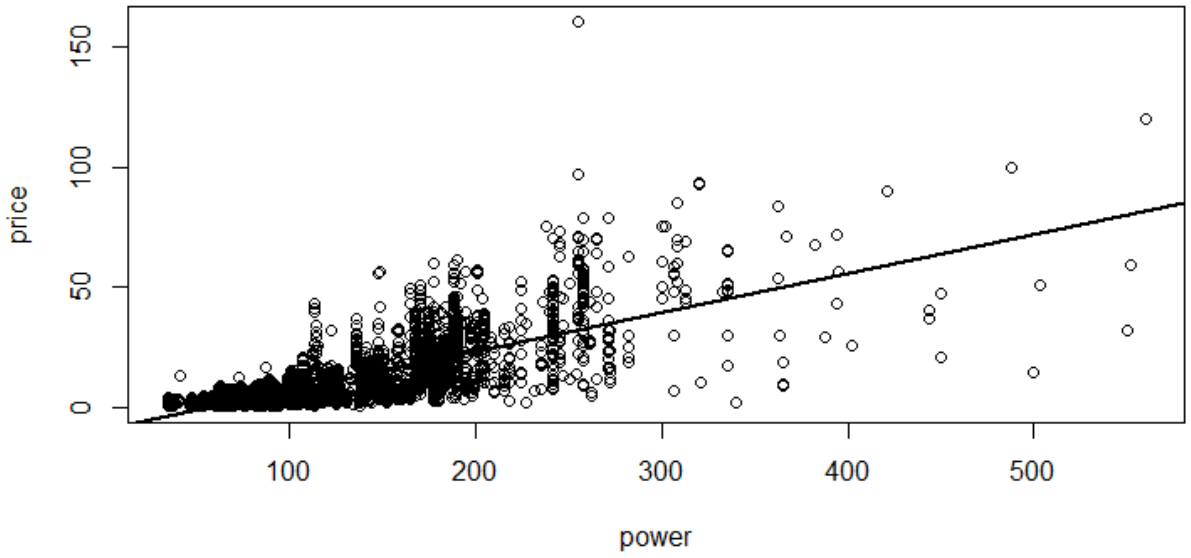

Figure 4: droite de régression linéaire simple

5-Intervalle de confiance :

L'intervalle de confiance est un indicateur mathématique qui permet de chiffrer la zone d'incertitude, lors d'une enquête ou d'un sondage portant sur un échantillon de population.

On effectue la détermination de l'intervalle de confiance de notre modèle par la commande 'confint(SimpleRegressionModel)'

Et on obtient :

```
> confint(SimpleRegressionModel)
                  2.5 %     97.5 %
(Intercept) -9.2116505 -8.3591890
Power        0.1578689  0.1646811
>
```

Figure5: intervalle de confiance des paramètres du model

l'intervalle de confiance de β_0 c'est [-9.2116505 ;-8.3591890] avec 95% de confiance

l'intervalle de confiance de β_1 c'est [0.1578689 ; 0.1646811] avec 95% de confiance

6-intervalle de Prédiction :

soit $X_{i^*} = 95$ donc $\widetilde{Y_{i^*}} = -8.785420 + 0.161275 * 95 = 6.535705$

donc l'intervalle de prédiction est : $Y_i \in [\widetilde{Y_{i^*}} - \sigma_1 * t_{tab} ; \widetilde{Y_{i^*}} + \sigma_1 * t_{tab}]$

soit

h_{i*}=0.6533 e

$\sigma_1^2 = \sigma_2^2 * (1 + h_{i*})$

t_{tab}=2.228

resudial standard errer =7.175 σ_1 =(7.175*7.175*(1+0.6533))^(1/2)

σ_1=9.22

intervalle=[*6.535705*- 9.22*2.228; *6.535705*+9.22*2.228]

intervalle de prédiction de Y^* est [-14.006455 ; 27.077865]

plus l'intervalle de prédiction est large et plus la prédiction perd de précision

Régression Linéaire Multiple

Introduction :

L'analyse de régression multiple est utilisée lorsque nous voulons prédire la valeur d'une variable en fonction de la valeur de deux ou plusieurs autres variables. La variable que nous voulons prédire est appelée variable dépendante et la variable que nous utilisons pour prédire la variable dépendante est appelée variable indépendante. Dans ce cas, la variable dépendante est le prix et la variable indépendante est toute autre variable de l'ensemble de données.

1-Présentation des données :

Dans cette partie de projet on va construire un modèle de régression multiple ou on va prédire le prix des voiture en fonction de Année , Kilométrage parcouru , Puissance , moteur , les places.

On utilise la commande summary(multipledata) pour donner une aperçu sur la base de données

```
> multipledata<-read.csv("CarPriceDataset.csv")
> summary(multipledata)
      Year      Kilometers_Driven    Mileage          Engine         Power
 Min.   :1998   Min.   :    171   Min.   : 0.00   Min.   :  72   Min.   : 34.2
 1st Qu.:2011   1st Qu.:  34000   1st Qu.:15.17   1st Qu.:1198   1st Qu.: 78.0
 Median :2014   Median :  53000   Median :18.15   Median :1493   Median : 98.6
 Mean   :2013   Mean   :  58738   Mean   :18.13   Mean   :1621   Mean   :113.3
 3rd Qu.:2016   3rd Qu.:  73000   3rd Qu.:21.10   3rd Qu.:1969   3rd Qu.:138.0
 Max.   :2019   Max.   :6500000   Max.   :33.54   Max.   :5998   Max.   :560.0
     Seats            Price
 Min.   : 0.000   Min.   :  0.440
 1st Qu.: 5.000   1st Qu.:  3.500
 Median : 5.000   Median :  5.640
 Mean   : 5.279   Mean   :  9.479
 3rd Qu.: 5.000   3rd Qu.:  9.950
 Max.   :10.000   Max.   :160.000
> .
```

Figure 6 : summary data

La commande summray nous permet d'avoir une vision générale sur notre base de données
Et d'afficher les différents indices statistiques (la moyenne, ...)

2-Matrice de corrélation :

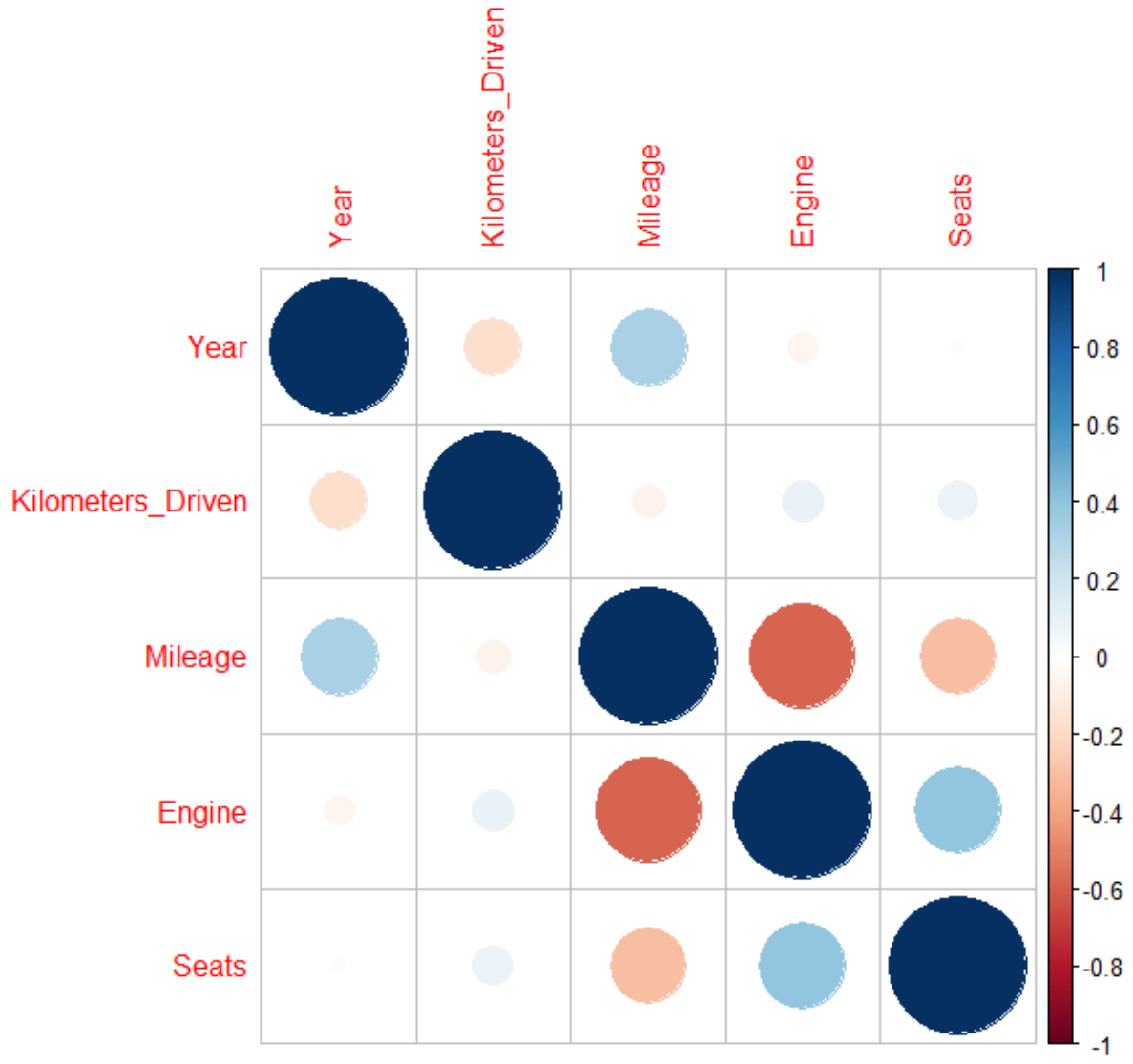

Figure 7: plot de la matrice de corrélation.

Remarque :

D'aprés le plot de la matrice de corrélation on remarque que toutes les variables sont indépendantes

3-Régression linéaire multiple :

```
> multipleModel<-lm(Price~Year+Kilometers_Driven+Mileage+Engine+Power+Seats,data = multipledata)
> summary(multipleModel)

Call:
lm(formula = Price ~ Year + Kilometers_Driven + Mileage + Engine +
    Power + Seats, data = multipledata)

Residuals:
    Min      1Q  Median      3Q     Max
-50.203  -3.049  -0.681   1.961 123.456

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.178e+03  5.554e+01 -39.211  < 2e-16 ***
Year               1.080e+00  2.773e-02  38.949  < 2e-16 ***
Kilometers_Driven  1.692e-06  9.137e-07   1.852  0.06405 .
Mileage           -6.225e-02  2.398e-02  -2.596  0.00945 **
Engine             2.725e-03  3.434e-04   7.936 2.48e-15 ***
Power              1.329e-01  3.536e-03  37.575  < 2e-16 ***
Seats             -1.129e+00  1.300e-01  -8.689  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.323 on 6012 degrees of freedom
Multiple R-squared:  0.6809,    Adjusted R-squared:  0.6806
F-statistic:  2138 on 6 and 6012 DF,  p-value: < 2.2e-16
```

Figure 8: summary régression linéaire multiple

La régression linéaire R utilise la fonction lm() pour créer un modèle de régression à partir de la base de données train, et la stocke dans la variable MultipleModel

La commande summary permet de donner un aperçu sur la régression sur la base de données.

L'estimation des paramètres

β_0 = -2.178e+03

β_1 = 1.08

β_2 = 1.692e-06

β_3 = -6.225e-02

β_4 = 2.725e-03

β_5 = 1.329e-01

β_6 = -1.129e+00

4- Tests d'hypothèse au niveau de risque 5 %*:*

Significativité des paramètres :

Test de student :

$$H_0 : \beta_0 = 0 \quad VS \quad H_1 : \beta_0 \neq 0$$

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_0 est très significative dans notre modèle.

$$H_0 : \beta_1 = 0 \quad VS \quad H_1 : \beta_1 \neq 0$$

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_1 est très significative dans notre modèle.

$$H_0 : \beta_2 = 0 \quad VS \quad H_1 : \beta_2 \neq 0$$

Nous avons 0.06405 est superieur à 0,05 donc on accepte l'hypothèse H_0, alors on élimine β_2 et on ré-estime le modèle

$$H_0 : \beta_3 = 0 \quad VS \quad H_1 : \beta_3 \neq 0$$

Nous avons Pr 0.00945 est inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_3 est très significative dans notre modèle.

$$H_0 : \beta_4 = 0 \quad VS \quad H_1 : \beta_4 \neq 0$$

Nous avons 2.48e-15 est très inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_4 est très significative dans notre modèle.

$$H_0 : \beta_5 = 0 \quad VS \quad H_1 : \beta_5 \neq 0$$

Nous avons < 2e-16 est inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_5 est significative dans notre modèle.

$$H_0 : \beta_6 = 0 \quad VS \quad H_1 : \beta_6 \neq 0$$

Nous avons < 2e-16 est très inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_6 est très significative dans notre modèle.

Ré-estimation du modèle :

Voici les résultats après l'élimination de β_2

```
> multipleModel<-lm(Price~Year+Mileage+Engine+Power+Seats,data = multipledata)
> summary(multipleModel)

Call:
lm(formula = Price ~ Year + Mileage + Engine + Power + Seats,
    data = multipledata)

Residuals:
    Min      1Q  Median      3Q     Max
-50.250  -3.040  -0.683   1.967 123.416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.160e+03  5.473e+01 -39.467  < 2e-16 ***
Year         1.071e+00  2.733e-02  39.198  < 2e-16 ***
Mileage     -5.991e-02  2.395e-02  -2.502   0.0124 *
Engine       2.772e-03  3.426e-04   8.092 7.02e-16 ***
Power        1.326e-01  3.534e-03  37.524  < 2e-16 ***
Seats       -1.121e+00  1.299e-01  -8.628  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.324 on 6013 degrees of freedom
Multiple R-squared:  0.6807,    Adjusted R-squared:  0.6805
F-statistic:  2564 on 5 and 6013 DF,  p-value: < 2.2e-16
```

Figure 9: modéle après ré-estimation.

Test de student (après ré-estimation) :

$$H_0 : \beta_0 = 0 \quad VS \quad H_1 : \beta_0 \neq 0$$

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_0 est très significative dans notre modèle.

$$H_0 : \beta_1 = 0 \quad VS \quad H_1 : \beta_2 \neq 0$$

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_2 est très significative dans notre modèle

$$H_0 : \beta_3 = 0 \quad VS \quad H_1 : \beta_3 \neq 0$$

Nous avons Pr 0.0124 est inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_3 est très significative dans notre modèle.

$$H_0 : \beta_4 = 0 \quad VS \quad H_1 : \beta_4 \neq 0$$

Nous avons 7.02e−16 est très inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_4 est très significative dans notre modèle.

$$H_0 : \beta_5 = 0 \quad VS \quad H_1 : \beta_5 \neq 0$$

Nous avons < 2e-16 est très inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_5 est très significative dans notre modèle.

$$H_0 : \beta_6 = 0 \quad VS \quad H_1 : \beta_6 \neq 0$$

Nous avons < 2e-16 est très inférieure à 0,05 donc on rejette l'hypothèse H_0, alors β_6 est très significative dans notre modèle.

Significativité de modèle :

Test de Fisher:

Le test de significativité global du modèle. Il consiste à tester si tous les coefficients du modèle, à l'exception de l'intercepte, sont nuls. Les hypothèses sont :

$$H_0 : \beta_1 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \text{ VS } H_1 : \exists j \text{ tel que } \beta_j \neq 0$$

```
Residual standard error: 6.324 on 6013 degrees of freedom
Multiple R-squared:  0.6807,    Adjusted R-squared:  0.6805
F-statistic:  2564 on 5 and 6013 DF,  p-value: < 2.2e-16
```

Figure 11: test de Fisher

On a la p-value est largement inferieur a 0.05 donc on peut rejette fortement H_0

5.Coefficient de détermination:

Le coefficient de détermination exprime le rapport entre la variance expliquée par le modèle de régression et la variance totale. Il sert à évaluer la qualité de l'ajustement
On ici le $R^2 = 0.68 = 68\%$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.160e+03  5.473e+01 -39.467  < 2e-16 ***
Year         1.071e+00  2.733e-02  39.198  < 2e-16 ***
Mileage     -5.991e-02  2.395e-02  -2.502   0.0124 *
Engine       2.772e-03  3.426e-04   8.092 7.02e-16 ***
Power        1.326e-01  3.534e-03  37.524  < 2e-16 ***
Seats       -1.121e+00  1.299e-01  -8.628  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.324 on 6013 degrees of freedom
Multiple R-squared:  0.6807,    Adjusted R-squared:  0.6805
F-statistic:  2564 on 5 and 6013 DF,  p-value: < 2.2e-16
```

Figure 12: coefficient de détermination R^2

On ici le $R^2 = 0.68$

Donc le modèle n'ajuste pas bien le nuage de pointes

Car la variance totale est très élevés

6-Intervalle de confiance :

L'intervalle de confiance est un indicateur mathématique qui permet de chiffrer la zone

d'incertitude, lors d'une enquête ou d'un sondage portant sur un échantillon de population.

On effectue la détermination de l'intervalle de confiance de notre modèle par la commande '*confint(modele3)*'

Et on obtient :

```
> confint(multipleModel)
                    2.5 %        97.5 %
(Intercept) -2.267385e+03 -2.052800e+03
Year         1.017794e+00  1.124956e+00
Mileage     -1.068623e-01 -1.296567e-02
Engine       2.100737e-03  3.443916e-03
Power        1.256961e-01  1.395535e-01
Seats       -1.375832e+00 -8.663715e-01
>
```

Figure 13: intervalle de confiance des paramètres du model

Donc

$\beta_0 \in$ [-2.267385e+03 ; -2.052800e+03]

$\beta_1 \in$ [1.017794e+00 ; 1.124956e+00]

$\beta_3 \in$ [-1.068623e-01 ; -1.296567e-02]

$\beta_4 \in$ [2.100737e-03 ; 3.443916e-03]

$\beta_5 \in$ [1.256961e-01 ; 1.395535e-01]

$\beta_6 \in$ [-1.375832e+00 ; -8.663715e-01]

7- Prédiction :

Dans cette partie on va faire un prédiction ponctuelle et une prédiction par intervalle

Voici les données pour lesquels on va prédire le nouveau résultat

	Year	Mileage	Engine	Power	Seats
1	2014	20	1200	90	5

Figure 14: les données de test (test).

Prédiction ponctuelle :

On a obtenue le résultat suivant avec la commande ' predict() '

```
> predict(multipleModel,newdata = test)
       1
6.115979
```

Figure 15: le résultat de la prédiction ponctuelle .

Donc $\hat{Y}_{i^*}$=6,115979

Prédiction par intervalle :

De même avec la commande ' predict() ' on obtient le résultat suivant

```
> predict(multipleModel,newdata = test,interval = "prediction")
       fit      lwr      upr
1 6.115979 -6.28329 18.51525
>
```

Figure 16: le résultat de la prédiction par intervalle .

Donc :

$Y_{i^*} \in$ [-6.28329 ; 18.51525]