SEP0832

Méthode d'échantillonnage

Rapport

Etude du prix d'une voiture

Élèves :
Garance Gabaut
Alexandre Brunet
Dylan Meyer
Viktorria Kabakova

Enseignant :
Amor Keziou

20 mai 2023

Table des matières

1 Introduction

La détermination du prix des véhicules est un sujet d'intérêt général pour les consommateurs, les constructeurs et les acteurs de l'industrie automobile. Le prix d'une voiture est souvent considéré comme un indicateur de la qualité et de la valeur d'un véhicule, influençant les décisions d'achat des consommateurs. Les prix dépendent de facteurs tels que le modèle, les fonctionnalités et options disponibles, ainsi que la demande du marché.

Les caractéristiques et options disponibles pour une voiture ont un impact majeur sur son prix. Des options telles que la connectivité Bluetooth ou la commande vocale peuvent considérablement augmenter le prix d'une voiture. De même, des caractéristiques telles que le modèle, le type de carburant utilisé ou le nombre de sièges peuvent avoir une incidence sur le prix d'un véhicule.

Étudier le prix des voitures en fonction de leurs caractéristiques est important pour les constructeurs automobiles et les consommateurs. Les constructeurs automobiles doivent comprendre les préférences des consommateurs et adapter leurs produits en conséquence pour rester compétitifs. D'autre part, les consommateurs peuvent utiliser ces informations pour comparer les caractéristiques et les prix de différentes voitures sur le marché et prendre des décisions d'achat éclairées.

En fin de compte, la tarification des voitures est un processus complexe impliquant de nombreux facteurs. En comprenant les caractéristiques et les options qui affectent le prix d'une voiture, les constructeurs automobiles et les consommateurs peuvent prendre des décisions éclairées pour assurer leur succès à long terme. Nous allons donc voir dans quelle mesure les caractéristiques et les options des voitures peuvent influencer leur prix.

Pour étudier la question, nous allons d'abord sélectionner les variables que nous allons utiliser pour le modèle de régression. Ensuite, nous choisirons le modèle de régression le plus approprié pour expliquer au mieux les prix des voitures, en utilisant les variables que nous avons préalablement sélectionnées. Ensuite, nous vérifierons la validité des hypothèses associées au modèle de régression choisi. Dans le cas où toutes les hypothèses sont respectées, nous établirons des prédictions grâce au modèle choisi.

Pour cela, nous allons utiliser une base de données comportant 2602 voitures différentes avec plus de 70 caractéristiques et options différentes qui serviront à expliquer leur prix.

Contributions

Alexandre BRUNET s'est occupé de l'introduction ainsi que de la programmation, de l'interprétation et de la rédaction de la validation des hypothèses. Il s'est également chargé de l'importation et du nettoyage de la base de données.

Viktoriia KABAKOVA s'est occupée de la programmation pour la présélection des variables. Cependant, étant donné qu'une mauvaise méthode a été utilisée, Alexandre BRUNET s'est chargé de la corriger et d'utiliser une autre méthode plus appropriée pour effectuer le travail. De plus, il a entièrement rédigé cette partie.

Dylan MEYER s'est occupé de la programmation, de l'interprétation et de la rédaction du choix des variables selon l'algorithme pas à pas. Il a également réalisé le code pour le modèle LASSO ainsi que le RLM pour le critère AIC, et a rédigé cette partie.

Garance GABAUT a appliqué l'algorithme génétique pour sélectionner les variables, a interprété et rédigé cette partie. Elle a construit le modèle RLM selon AIC et le modèle ridge. Elle a effectué l'interprétation et la rédaction des modèles ridge, LASSO et RLM selon AIC.

Alexandre BRUNET et Garance GABAUT se sont réparti le travail de mise en forme du code final et du rapport.

Nous avons rencontré quelques problèmes avec la base de données. Dylan MEYER, Alexandre BRUNET et Garance GABAUT ont travaillé ensemble pour résoudre tous les problèmes liés à celle-ci.

2 Choix des variables pour l'intégration du modèle de régression

Dans un modèle de régression, lorsque la taille du modèle est grande, un problème de sur ajustement peut survenir, d'où la nécessité de procéder à une sélection de variables que l'on introduira par la suite dans les modèles que l'on testera.

Dans un premier temps, nous allons classer les variables selon leur significativité pour apporter une vue d'ensemble sur le nombre de variables que l'on pourrait *apriori* utilisé. Dans un deuxième temps, nous sélectionnerons les variables grâce à un algorithme adapté.

2.1 Classement des variables explicatives signficatives

Pour juger de la significativité des variables, nous avons à notre disposition le test de Student et le test de Fisher. Etant donné que nous sommes dans une régression linéaire, nous avons décidé d'utiliser le test de Fisher (figure (1)). En effet, ce test examine si le modèle dans son ensemble est significatif, c'est-à-dire s'il y a au moins une variable indépendante qui a un effet significatif sur la variable dépendante. Il compare la variance expliquée par le modèle (variance due à la relation entre les variables indépendantes et la variable dépendante) avec la variance non expliquée (variance résiduelle) pour déterminer si le modèle est statistiquement significatif.

```
> values
Analysis of Variance Table

Response: Price
                Df     Sum Sq    Mean Sq   F value Pr(>F)
Power.hp         1 4.9382e+13 4.9382e+13 2089.7905 <2e-16 ***
Displacement.l   1 2.9252e+12 2.9252e+12  123.7890 <2e-16 ***
Torque.lbft      1 1.4260e+13 1.4260e+13  603.4544 <2e-16 ***
MPG.City         1 2.1067e+12 2.1067e+12   89.1510 <2e-16 ***
MPG.Highway      1 4.7772e+12 4.7772e+12  202.1667 <2e-16 ***
Length.in        1 5.9890e+11 5.9890e+11   25.3445 <2e-16 ***
Width.in         1 4.4038e+11 4.4038e+11   18.6365 <2e-16 ***
Brand           41 2.4682e+13 6.0199e+11   25.4756 <2e-16 ***
Wheelbase.in     1 2.0568e+11 2.0568e+11    8.7040 0.0033 **
Clearance.in     1 1.0887e+11 1.0887e+11    4.6071 0.0322 *
Height.in        1 1.0685e+11 1.0685e+11    4.5216 0.0338 *
Cylinders        1 8.9348e+10 8.9348e+10    3.7811 0.0523 .
Doors            1 5.1515e+10 5.1515e+10    2.1801 0.1403
Seats            1 3.7863e+10 3.7863e+10    1.6023 0.2060
Gearbox.Type     2 1.9941e+10 9.9703e+09    0.4219 0.6560
Fuel.Type        3 1.6563e+10 5.5209e+09    0.2336 0.8730
Drivetrain       2 2.4131e+09 1.2066e+09    0.0511 0.9502
Residuals      655 1.5478e+13 2.3630e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 1 – Anova de la table de données pour le classement des varriables significatives

Parmi les 18 variables étudiées, seules 6 ne présentent pas de signification statistique, à savoir : *Cylinders*, *Doors*, *Seats*, *Gearbox.Type*, *Fuel.type* et *Drivetrain*. Nous pouvons dire

cela car ces variables sont associées à une p-value inférieure au seuil de 5 %. En revanche, les autres variables sont statistiquement significatives. De plus, il est intéressant de noter que toutes les variables sont significatives à un niveau similaire (très proche de 0), à l'exception de trois d'entre elles : *Wheelbade.in*, *Clearance.in* et *Height.in*. Ces variables sont moins significatives que les précendentes, mais le sont tout de même, car elles sont associées à une p-value maximum de 3,38%, ce qui reste inférieur à notre seuil de 5%. Nous disposons donc de 11 variables. Nous allons maintenant approfondir la sélection des variables en utilisant des méthodes appropriées.

2.2 Sélection définitive des variables

La sélection des variables peut être effectuée à l'aide de multiples méthodes, chacune présentant ses avantages et ses inconvénients. Dans cette étude, nous allons comparer deux approches différentes : tout d'abord en appliquant un algorithme génétique, puis en utilisant une méthode de recherche pas-à-pas. Plus précisément, pour cette dernière méthode, nous appliquerons une approche descendante.

2.2.1 L'alogrithme génétique

Notre première idée était d'appliquer une recherche exhaustive, mais nous avons rapidement dû l'abandonner, car le nombre de modèles à construire était beaucoup trop élevé. Nous avons alors décidé d'appliquer un algorithme génétique, qui est mieux adapté aux situations où le nombre de variables est important. Toutefois, cette méthode peut être coûteuse en temps de calcul. Dans notre cas, elle a nécessité un grand nombre d'itérations avant de converger. Nous l'avons appliqué pour deux critères différents : selon le critère de sélection AIC [1] et selon le BIC [2]. Les résultats diffèrent légèrement d'un critère à l'autre. Le meilleur modèle pour prédire le prix d'une voiture selon la méthode génétique se compose de 9 variables avec le critère AIC : *Power.hp*, *Displacement.l*, *Torque.lbft*, *MPG.City*, *MPG.Highway*, *Height.in*,*length.in*, *Width.in* et *Wheelbase.in*, et de seulement 7 variables selon le critère BIC *Power.hp*, *Displacement.l*, *Torque.lbft*, *MPG.City*, *MPG.Highway*, *Width.in* et *Clearance.in*.

2.2.2 L'algoritme pas-à-pas

Nous avons finalement souhaité appliquer l'algorithme de recherche pas-à-pas (méthode descendante), également selon les critères AIC et BIC. Cet algorithme a pour particularité de construire les modèles explicatifs de manière récursive, c'est-à-dire que des variables explicatives sont supprimées ou ajoutées à chaque étape. En simplifiant les modèles à chaque étape, cet algorithme gagne en rapidité, mais perd en précision. Les résultats obtenus diffèrent légèrement selon le critère : lorsqu'on se base sur le critère AIC, le meilleur modèle se compose des variables *Power.hp*, *Displacement.l*, *Torque.lbft*, *MPG.City*, *MPG.Highway*, *Height.in*,*length.in*, *Width.in* et *Wheelbase.in*, tandis que le critère BIC renvoie *Power.hp*, *Displacement.l*, *Torque.lbft*, *MPG.City*, *MPG.Highway*, *Height.in* et *Width.in*.

1. signifie *Akaike information criterion*
2. signifie *bayesian information criterion*

Le second modèle contient seulement 7 variables, contrairement au premièr modèle qui en contient donc 9. Le résultat renvoyé par R selon le critère AIC est visible sur la figure (2).

```
Call:
lm(formula = Price ~ Power.hp + Displacement.l + Torque.lbft +
    MPG.City + MPG.Highway + Height.in + Length.in + Width.in +
    Wheelbase.in, data = df)

Coefficients:
   (Intercept)        Power.hp  Displacement.l     Torque.lbft        MPG.City     MPG.Highway
Height.in
         97898            4149          -45921           -2679           32716          -25498
6974
     Length.in        Width.in    Wheelbase.in
         -4404           -4432            3414

> TdiffPasaPasAIC
Time difference of 0.04039407 secs
```

FIGURE 2 – Résultat retourné par R lors de l'application de l'algorithme pas-à-pas appliqué selon le critère AIC

Power.hp, *Displacement.l*, *Torque.lbft*, *MPG.City*, *MPG.Highway*, *Height.in*,*length.in*, *Width.in* et *Wheelbase.in*

2.2.3 Comparaison des résultats

Le premier point que l'on peut aborder est le temps d'exécution de chaque algorithme. Dans le cas de l'algorithme génétique, la durée d'exécution est d'environ 20 secondes, tandis que l'algorithme pas-à-pas ne prend que quelques millièmes de secondes pour s'exécuter. Il est donc évident que l'algorithme pas à pas est nettement plus rapide.

En ce qui concerne les résultats obtenus, on constate que pour le critère AIC, les résultats sont les mêmes. Pour le critère BIC, une variable diffère : avec l'algorithme génétique, nous obtenions la variable *Clearance.in* alors qu'avec l'algorithme pas à pas, nous avions *Height.in*. Cette variable se trouve justement dans les variables sélectionnées par les deux méthodes avec le critère AIC. Nous décidons de conserver les résultats obtenus selon le critère BIC.

Désormais, nous allons regarder quelle modèle de régression est le plus adapté et donne le meilleur résultat avec les variables conservées.

3 Choix de la régression

Nous allons effectuer plusieurs régressions : une régression linéaire multiple (RLM), une Ridge et une Lasso. chacune d'entre elles seront décomposées dans une partie.

3.1 Régression linéaire multiple

La régression linéaire multiple est une technique permettant d'établir une relation entre une variable cible *(le prix dans notre cas)* et plusieurs variables explicatives (*Power.hp*, *Displacement.l*, *Torque.lbft*, *MPG.City*, *MPG.Highway*, *Height.in*,*length.in*, *Width.in* et *Wheelbase.in*). Elle permet de prédire la valeur de la variable réponse en fonction des valeurs des variables explicatives. Nous avons donc souhaité voir si ce modèle pouvait fonctionner avec nos variables. Sur la figure (3) se trouve les résultats de la régression linéaire. Les résultats de cette régression linéaire fournissent des estimations des coefficients de régression pour chaque variable explicative. Cependant, il convient de noter que certaines hypothèses statistiques n'ont pas encore été vérifiées, telles que l'homoscédasticité des résidus et nous n'avons pas encore estimé l'erreur de prévision du modèle.

La valeur de R-carré multiple est d'environ 0.64, ce qui signifie que les variables explicatives incluses dans le modèle expliquent environ 64% de la variation observée dans le prix.

Cependant, ces résultats sont préliminaires et leur interprétation n'est pas encore exploitable puisqu'on n'a pas vérifié les hypothèses qui garantissent la validité des résultats et de la fiabilité des conclusions faites à partir du modèle.

```
> summary(modeleRLM)

Call:
lm(formula = Price ~ ., data = DataFrameModele)

Residuals:
    Min      1Q  Median      3Q     Max
-824991  -67674   13163   89160 1980118

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    97898.4   177211.5   0.552  0.58082
Power.hp        4149.4      203.3  20.409  < 2e-16 ***
Displacement.l -45920.8    10589.9  -4.336 1.66e-05 ***
Torque.lbft     -2679.0      209.2 -12.804  < 2e-16 ***
MPG.City        32715.9     3665.5   8.925  < 2e-16 ***
MPG.Highway    -25497.6     4100.5  -6.218 8.59e-10 ***
Width.in        -4431.6     1550.7  -2.858  0.00439 **
Wheelbase.in     3413.9     1715.0   1.991  0.04690 *
Height.in        6974.0     1771.2   3.937 9.05e-05 ***
Length.in       -4403.7     1444.8  -3.048  0.00239 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 239800 on 709 degrees of freedom
Multiple R-squared:  0.6469,    Adjusted R-squared:  0.6424
F-statistic: 144.3 on 9 and 709 DF,  p-value: < 2.2e-16
```

FIGURE 3 – Résultats obtenus lors de l'exécution du modèle de régression linéaire multiple

Désormais, il est question d'effectuer une nouvelle sélection des variables afin d'améliorer le modèle construit. Nous allons effectuer cette sélection selon 2 critères : l'AIC et le BIC.

3.1.1 Selon le critère AIC

Nous commençons par le critère AIC. Sur la figure (4) se trouve la sortie *R* lorsqu'on exécute le code permettant la nouvelle sélection des variables. On constate que toutes les variables incluses dans la formule du modèle sont conservées, puisqu'elles ont toutes un p-value inférieure au seuil de 5%. On voit également que la p-value du modèle est inférieure à notre seuil de significativité. De ce fait, le modèle dans son ensemble est statistiquement significatif pour expliquer la variation de la variable dépendante à partir des variables indépendantes.

```
> summary(modele_aic)

Call:
lm(formula = Price ~ Power.hp + Displacement.l + Torque.lbft +
    MPG.City + MPG.Highway + Width.in + Wheelbase.in + Height.in +
    Length.in, data = DataFrameModele)

Residuals:
    Min      1Q  Median      3Q     Max
-824991  -67674   13163   89160 1980118

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     97898.4   177211.5   0.552  0.58082
Power.hp         4149.4      203.3  20.409  < 2e-16 ***
Displacement.l -45920.8    10589.9  -4.336 1.66e-05 ***
Torque.lbft     -2679.0      209.2 -12.804  < 2e-16 ***
MPG.City        32715.9     3665.5   8.925  < 2e-16 ***
MPG.Highway    -25497.6     4100.5  -6.218 8.59e-10 ***
Width.in        -4431.6     1550.7  -2.858  0.00439 **
Wheelbase.in     3413.9     1715.0   1.991  0.04690 *
Height.in        6974.0     1771.2   3.937 9.05e-05 ***
Length.in       -4403.7     1444.8  -3.048  0.00239 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 239800 on 709 degrees of freedom
Multiple R-squared:  0.6469,    Adjusted R-squared:  0.6424
F-statistic: 144.3 on 9 and 709 DF,  p-value: < 2.2e-16
```

FIGURE 4 – Sortie R suite à une sélection des variables sur un modèle RLM selon le critère AIC.

Nous avons souhaité estimer l'erreur de ce modèle à partir de la méthode d'apprentissage/validation, qui consiste à diviser la base de données en deux parties de façon aléatoire. Les 2/3 des données servent à l'apprentissage, contre 1/3 pour la validation. Ensuite, nous regardons la moyenne des différences au carré entre les valeurs de l'ensemble de validation et celles prédites. Pour affiner les résultats, nous répétons $m = 1\ 000$ fois le processus. L'erreur finale est la moyenne de chaque erreur de prévision calculée.

On a calculé la moyenne de l'erreur pour i itérations, avec i allant de 1 à m. La figure (5) représente justement la moyenne estimée de l'erreur en fonction du nombre d'itérations. On constate qu'arrivé à un certain nombre d'itérations, la moyenne de l'erreur estimée stagne. Il n'est donc sans doute pas pertinent de prendre autant d'itérations. La seconde remarque que l'on peut faire se porte sur la valeur de l'erreur. En effet, celle-ci est supérieure à 10^{10}, ce qui est énorme. En particulier, elle vaut 58 386 264 547. Regardons désormais les résultats pour le critère BIC.

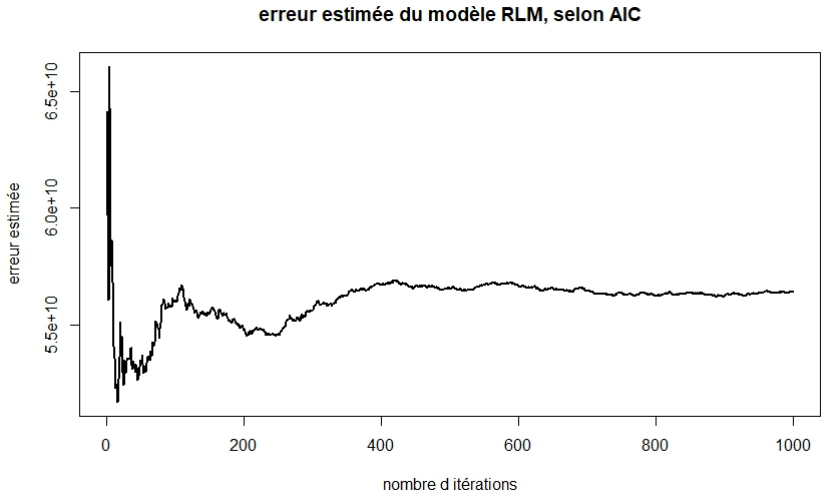

FIGURE 5 – Moyenne de l'erreur estimée en fonction du nombre d'itérations, selon le critère AIC.

3.1.2 Selon le critère BIC

À nouveau, et comme en témoigne la sortie R sur la figure (6), on constate que toutes les variables sont conservés. De plus, la p-value du modèle est inférieure au seuil de 5%, ainsi le modèle est statistiquement significatif pour expliquer la variable *Price* avec les variables explicatives.

```
> summary(modele_bic)

Call:
lm(formula = Price ~ Power.hp + Displacement.l + Torque.lbft +
    MPG.City + MPG.Highway + Width.in + Height.in, data = DataFrameModele)

Residuals:
    Min      1Q  Median      3Q     Max
-805129  -65902   17104   89825 2024627

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -16393.5   160791.3  -0.102 0.918821
Power.hp         4142.0      198.2  20.898  < 2e-16 ***
Displacement.l -53315.1    10087.0  -5.286 1.67e-07 ***
Torque.lbft     -2738.2      202.1 -13.547  < 2e-16 ***
MPG.City        34873.7     3606.7   9.669  < 2e-16 ***
MPG.Highway    -28653.6     3986.6  -7.187 1.67e-12 ***
Width.in        -5537.7     1477.7  -3.747 0.000193 ***
Height.in        4274.5     1473.1   2.902 0.003827 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 241100 on 711 degrees of freedom
Multiple R-squared:  0.6422,    Adjusted R-squared:  0.6387
F-statistic: 182.3 on 7 and 711 DF,  p-value: < 2.2e-16
```

FIGURE 6 – Sortie R suite à une sélection des variables du modèle RLM selon le critère BIC.

Nous avons ensuite procédé à l'estimation de l'erreur de la même manière que pour le critère AIC. Malheureusement, nous avons constaté que l'erreur obtenue était à nouveau très élevée, s'élevant à 57 462 549 234 . La figure (7) illustre l'évolution de l'erreur estimée pour le modèle RLM. Comme précédemment avec le critère AIC, l'erreur estimée converge rapidement. Nous allons maintenant examiner les résultats obtenus avec la régression Lasso.

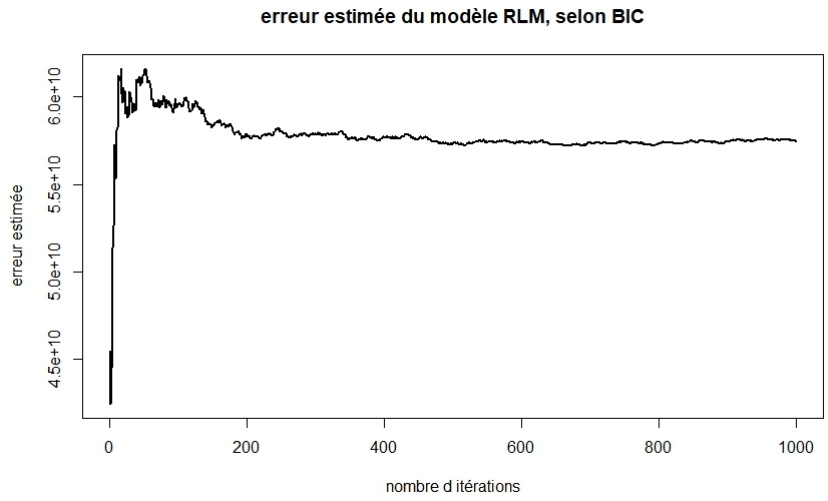

FIGURE 7 – Évolution de l'erreur estimée du modèle RLM en fonction du nombre d'itérations, selon le critère BIC.

3.2 Régression LASSO

La régression LASSO [3] a pour but de minimiser le critère des moindres carrés pénalisé par la norme L_1 des estimateurs des coefficients. Nous avons souhaité voir les résultats avec nos variables. Pour cela, nous avons déterminé la valeur la plus adaptée de λ, le paramètre de pénalisation.

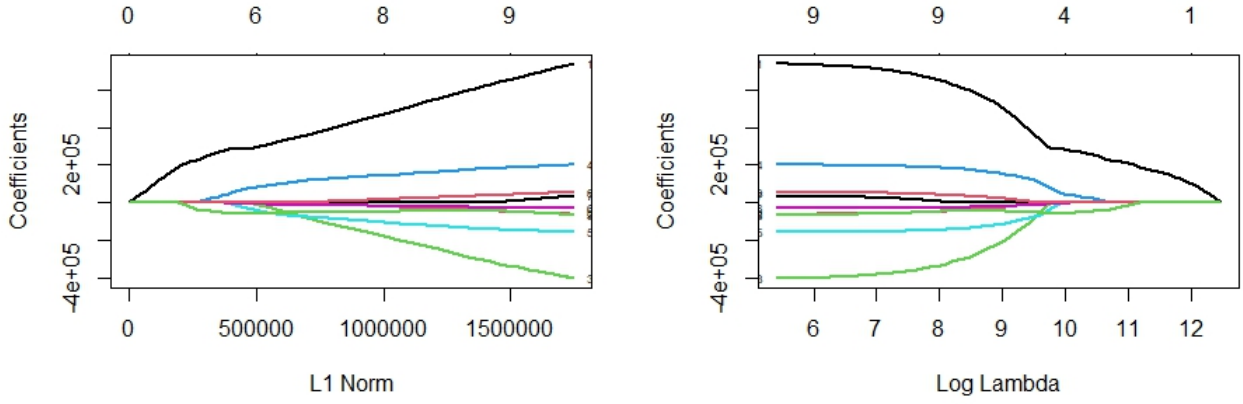

FIGURE 8 – Résultats de la régression LASSO

La figure (8) permet de visualiser les résultats de la régression LASSO et à faciliter l'interprétation des coefficients estimés en fonction de la valeur du paramètre de pénalisation. La figure à gauche représente les valeurs des coefficients estimés en fonction de la valeur du logarithme du paramètre de pénalisation λ. On constate que la courbe 1 croît particulièrement vite : la variable qui lui est associé, *Power.hp*, joue un rôle important dans la prédiction de la variable dépendante. D'autres courbes, comme la 6, 7 ou 8 (associées aux variables *Wheelbase.in, Height.in et Width.in*) ne semblent pas évoluer avec

3. LASSO Signifiant *Least Absolute Shrinkage and Selection Operator.*

le coefficient de pénalisation. Cela signifie que le coefficient correspondant à ces variables ne change pas de manière significative à mesure que la valeur de λ change.

La figure de droite est une représentation de l'erreur en fonction du paramètre de pénalisation. On constate que les courbes numérotées par 2, 9, 8, 7, et 6 stagnent. Cela indique que l'erreur du modèle ne change pas de manière significative à mesure que le paramètre de pénalisation évolue. Au contraire, les courbes 1 et 3 (*variables Power.hp et Torque.lbft*) croient ou décroient considérablement. elles semblent donc davantage être influencées lorsque le paramètre de pénalisation augmente.

Nous avons fini par représenter les résultats de la validation croisée pour le modèle créé. Ils sont visibles sur la figure (9).

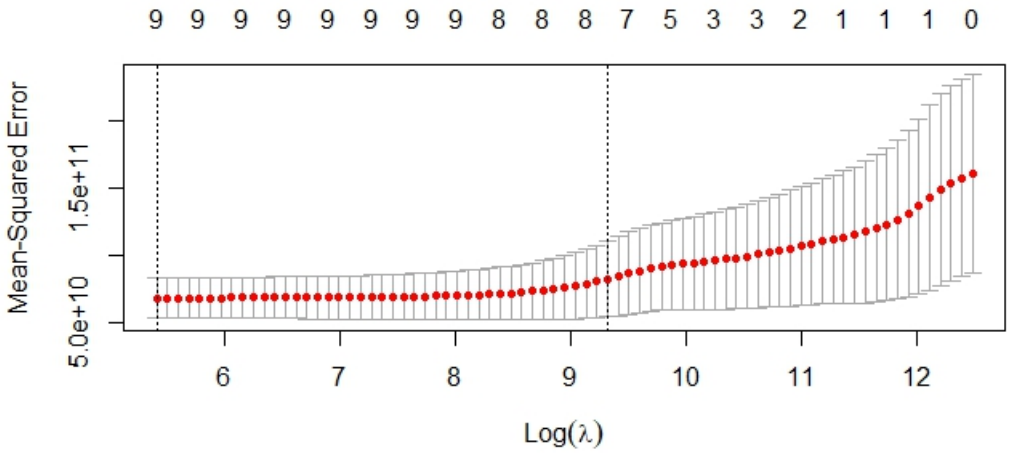

FIGURE 9 – Résultats de la validation croisée pour le modèle LASSO.

La courbe rouge représente l'erreur de validation croisée en fonction de la valeur de λ. On constate que la courbe rouge croit, cela signifie que l'erreur de validation croisée augmente à mesure que la valeur de λ augmente : le modèle devient moins performant à mesure que la pénalisation augmente.

Ensuite, les intervalles gris sont les intervalles de confiance, qui sont ici très grands, ce qui témoigne d'une incertitude sur la performance du modèle.

Pour finir, les droites verticales, lorsque $log(\lambda) \approx 9.3$ et $log(\lambda) \approx 5.5$, indiquent les meilleures valeurs de λ afin d'optimiser les résultats. R nous retourne en particulier $\lambda = 294.5327$, et donc $log(\lambda) = 5.68$. Cela correspond à la première droite.

A nouveau, l'erreur de prévision obtenue est très grande : 66 025 435 691. Nous avons finalement regardé les coefficients retourné par la régression, et les résultats se trouve sur la figure (10).

On constate que 2 variables sont considérées comme non pertinentes : les variables *Displacement.l* et *Wheelbase.in*. On remarque que ces deux variables font partie de celle dont les courbes sur la figure (8) ne semblaient pas évoluer lorsque le paramètre de pénalisation évoluait. Les deux interprétations concordent donc. On note aussi que le coefficient le plus élevé est celui associé à la variable *Power.hp* : cela signifie qu'il existe une relation plus forte entre la variable *Price* et celle-ci. On remarque aussi que la courbe associée à cette variable faisait partie de celles qui évoluaient le plus lorsque le paramètre λ évoluait. C'est à nouveau cohérent. Nous allons finir par regarder le modèle Ridge.

```
> coef(reg.cvlasso)
10 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)    113715.401
Power.hp       362963.788
Displacement.l      .
Torque.lbft    -79102.829
MPG.City       119682.560
MPG.Highway    -74645.059
Width.in       -12065.180
Wheelbase.in        .
Height.in        1341.517
Length.in      -51203.248
```

FIGURE 10 – Coefficients obtenus lors de la régression LASSO.

3.3 Le modèle Ridge

La régression Ridge a pour but de minimiser le critère des moindres carrés, cette fois pénalisé par la norme L_2, des estimateurs des coefficients. Nous avons souhaité l'appliquer aussi à nos données afin de comprendre la variable *Price*. La figure (11) montre les résultats obtenus.

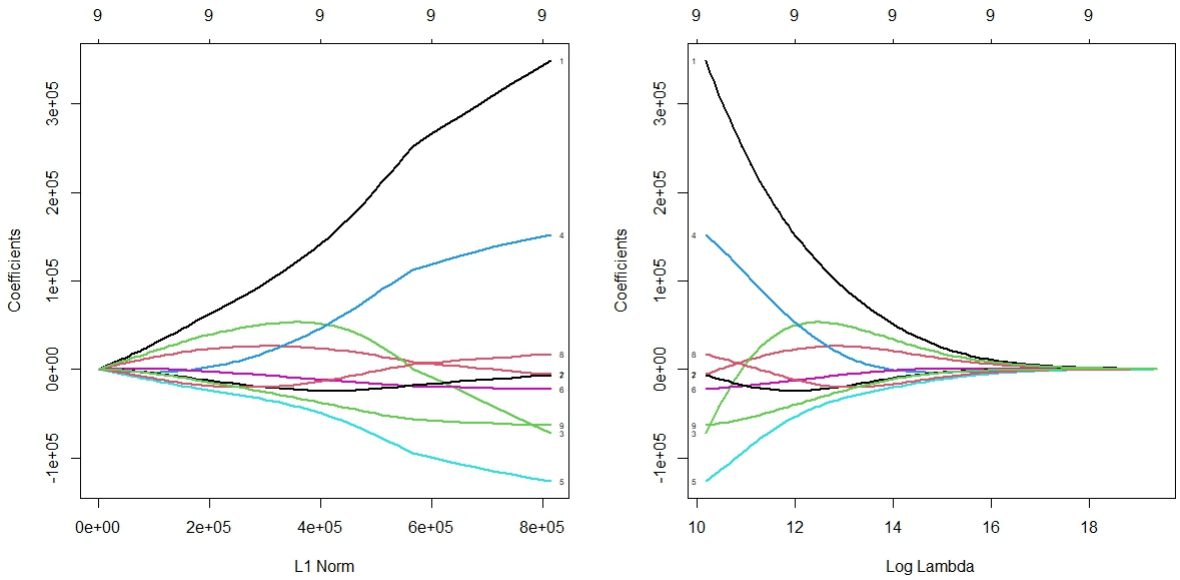

FIGURE 11 – Résultats de la régression Ridge

A nouveau, la courbe associée à la variable *Power.hp*, montrant l'évolution des coefficients en fonction de la valeur du paramètre de pénalisation, évolue particulièrement en fonction du paramètre de pénalisation. De même lorsqu'on regarde l'évolution de l'erreur en fonction du paramètre de pénalisation. A *contrario*, les variables *Displacement.l* et *Wheelbase.in* sont aussi représentées dans les deux cas par des courbes qui stagnent énormément. Nous avons ensuite représenté sur la figure (12) le processus de validation croisée pour sélectionner la valeur optimal de λ.

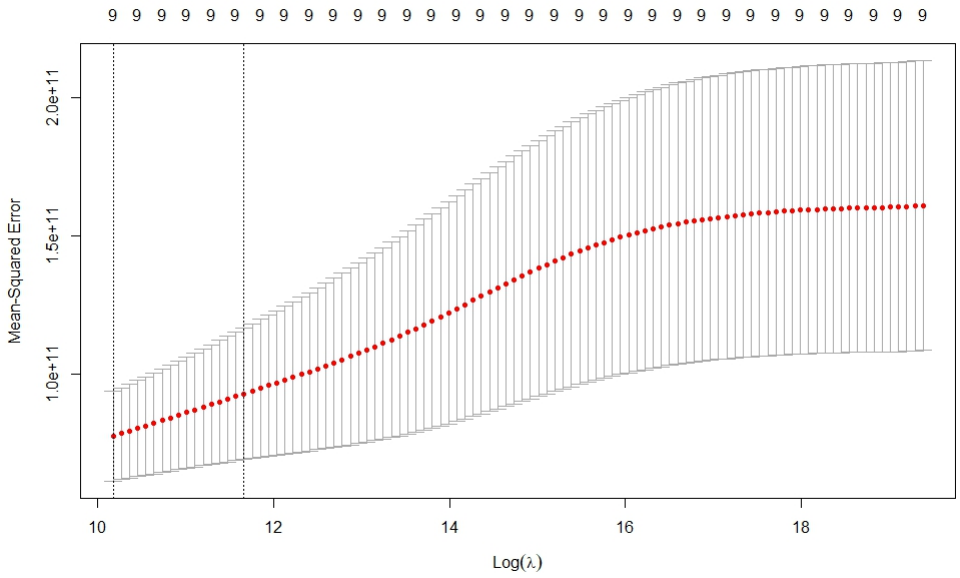

FIGURE 12 – Résultats de la validation croisée pour le modèle Ridge.

L'erreur croît à mesure que la valeur de λ augmente. On constate qu'il existe 2 valeurs de λ optimales : environ 10.2 et 11.8. R nous retourne en particulier que la valeur de λ optimal était $log(26219.74) = 10.17427$, soit la valeur la plus à gauche. La valeur de l'erreur retenue est 67 714 363 312. À nouveau, cette valeur est très élevée.

La conclusion concernant le choix du modèle est difficile à établir, car les 5 modèles présentent des erreurs très élevées. Cela peut indiquer que les variables incluses dans les modèles ne suffisent pas à expliquer pleinement la variation de la variable dépendante (Price). Nous allons tout de même poursuivre cette étude avec le modèle RLM selon le critère BIC puisque c est ce modèle qui a l'erreur la moins élevée.

4 Validation des hypothèses

Après avoir identifié le meilleur modèle (ou le moins mauvais), il est essentiel de valider les hypothèses associées à ce modèle afin d'évaluer sa capacité à réaliser des prédictions fiables.

4.1 La multicolinéarité des variables

La première chose que nous devons regarder, c'est bien la colinéarité des variables. Pour cela, nous avons utilisé le VIF [4]. Le VIF est utilisé pour évaluer la multicolinéarité entre les variables indépendantes. La multicolinéarité se produit lorsqu'il y a une corrélation élevée entre les variables indépendantes, ce qui peut causer des problèmes dans l'interprétation des coefficients de régression. Selon les recommandations, le VIF d'une variable ne doit pas dépasser la valeur 4. Cependant, dans notre cas, la plupart des variables dépassent cette valeur, comme on peut le voir sur la figure (13). À partir de ce test, nous pouvons déjà conclure que le modèle est inadapté pour établir des prévisions fiables.

```
> #Test de multicolinéarité
> vif(regb)
    Power.hp  Torque.lbft     MPG.City  MPG.Highway     Width.in Wheelbase.in    Height.in
    9.688113     7.632764     6.033417     7.686354     1.621784     5.673829     2.581309
   Length.in
    6.760775
```

FIGURE 13 – Facteurs d'inflation de la variance pour le modèle RLM au sens du BIC.

4.2 La linéarité du modèle

La régression linéaire suppose que le modèle soit linéaire en les paramètres du modèle. Cette hypothèse est extrêmement importante pour évaluer si la relation entre la variable dépendante et les variables indépendantes est véritablement linéaire. Elle permet de s'assurer que l'utilisation d'un modèle de régression linéaire est appropriée et que les résultats de l'analyse sont fiables. Ainsi, cette hypothèse doit être systématiquement validée pour effectuer des prévisions fiables. La figure (14) présente le test de Rainbow, qui vise précisément à tester la linéarité du modèle. Dans notre cas, le test confirme que le modèle est linéaire aux paramètres.

```
> #Test de linéarité
> raintest(regb)

        Rainbow test

data:  regb
Rain = 0.3004, df1 = 771, df2 = 761, p-value = 1
```

FIGURE 14 – Test de Rainbow pour le modèle RLM au sens du BIC.

4. Facteur d'Inflation de la Variance

4.3 La normalité des résidus

Une autre hypothèse que nous devons vérifier concerne les résidus du modèle, en s'assurant qu'ils suivent une distribution normale. Cette hypothèse revêt une importance significative pour plusieurs raisons. Tout d'abord, elle garantit l'impartialité des estimations des paramètres. Ensuite, elle assure la précision des prévisions, car une distribution normale des résidus implique que les erreurs de prédiction sont aléatoires et symétriques autour de zéro. Enfin, cette hypothèse permet le diagnostic du modèle, car si elle n'est pas vérifiée, cela indique des problèmes potentiels dans le modèle. Dans notre cas, nous pouvons visualiser si les résidus suivent effectivement une distribution normale à l'aide de la figure (15).

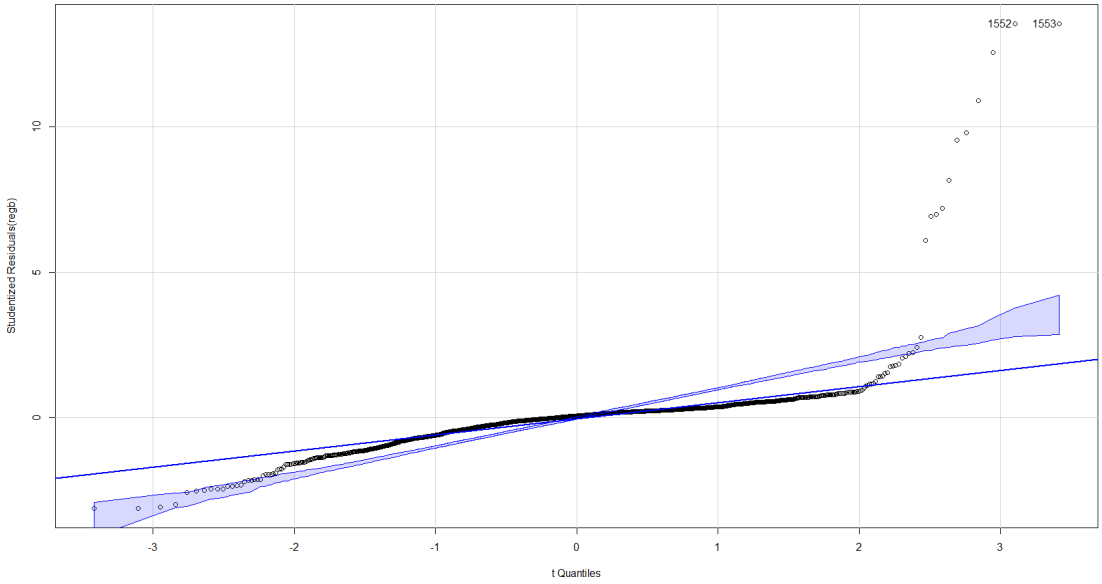

FIGURE 15 – Graphe Quantile-Quantile pour le modèle RLM au sens du BIC.

Apparemment, les points ne semblent pas suivre la droite bleue, ce qui suggère que les résidus ne suivent pas une distribution normale. Nous pouvons confirmer cette observation en consultant la figure (16), qui teste la normalité des résidus de quatre manières différentes.

```
> ols_test_normality(residuals(regb))
-----------------------------------------------
       Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.5632         0.0000
Kolmogorov-Smirnov        0.2037         0.0000
Cramer-von Mises        133.7316         0.0000
Anderson-Darling        120.8028         0.0000
-----------------------------------------------
```

FIGURE 16 – Tests de normalité pour les résidus du modèle RLM au sens du BIC.

Les différents tests de normalité nous confirme bien que les résidus ne suivent pas une loi normale. Nous concluons donc que cette hypothèse du modèle n'est pas validée.

4.4 La stabilité de la variance : l'homoscédasticité

Une autre hypothèse concernant les résidus est celle de l'homoscédasticité, qui stipule que la variance des résidus est constante dans le temps. Nous pouvons évaluer si le modèle présente une homoscédasticité des résidus en observant la figure (17). Nous remarquons que le nuage de points forme une structure, ce qui suggère que les variances des résidus ne sont pas égales (il semble y avoir une hétéroscédasticité).

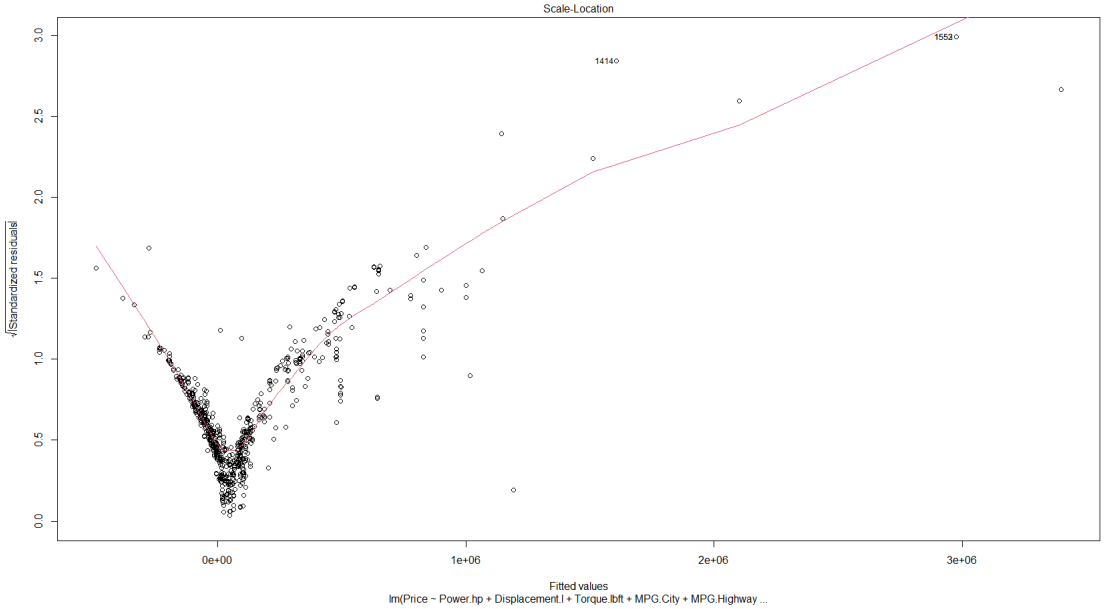

FIGURE 17 – Visualisation graphique de l'homoscédasticité pour les résidus du modèle RLM au sens du BIC.

Pour confirmer le refus de l'hypothèse d'égalité des variances, nous avons procédé à un test (figure (18)). Ce test rejette l'hypothèse d'homogénéité des variances, elle n'est donc pas validée pour notre modèle.

```
> #Tester l'homoscédascité (égalité des variances)
> ols_test_score(regb)

 Score Test for Heteroskedasticity
 ---------------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: fitted values of Price

          Test Summary
 -------------------------------
 DF             =    1
 Chi2           =    396.717
 Prob > Chi2    =    2.854996e-88
```

FIGURE 18 – Test d'homoscédasticité pour les résidus du modèle RLM au sens du BIC.

4.5 L'indépendance des résidus

En plus de la normalité et de l'homoscédasticité des résidus, nous devons vérifier s'ils sont indépendants. La vérification de cette hypothèse dans une régression linéaire est essentielle pour garantir la validité des résultats et l'efficacité des tests statistiques associés. En d'autres termes, la validité de cette hypothèse permettrait d'estimer les paramètres du modèle sans biais, d'améliorer la précision des prévisions et de valider le modèle dans son ensemble (ces raisons sont les mêmes que celles de l'hypothèse précédente). Dans la figure (19), le test de Durbin-Watson démontre que, dans le modèle choisi, les résidus présentent une dépendance.

```
> #Tester si les résidus sont indépendants entre eux
> durbinWatsonTest(regb)
 lag Autocorrelation D-W Statistic p-value
   1       0.3685508      1.261812       0
 Alternative hypothesis: rho != 0
```

FIGURE 19 – Test d'indépendance des résidus du modèle RLM au sens du BIC.

4.6 L'orthogonalité du modèle

Une dernière hypothèse concerne l'orthogonalité du modèle. En réalité, cette hypothèse consiste à démontrer l'absence de corrélation entre les variables indépendantes et les résidus. Nous avons précédemment constaté que les variables étaient corrélées entre elles et les résidus également. Maintenant, nous devons vérifier si nous observons la même chose entre les variables et les résidus, en nous référant à la figure (20). Nous remarquons clairement que les variables et les résidus ne sont pas orthogonaux entre eux. Cela est évident car le nuage de points suit la courbe rouge, ce qui ne devrait pas être le cas. Par conséquent, l'hypothèse d'orthogonalité n'est pas validée.

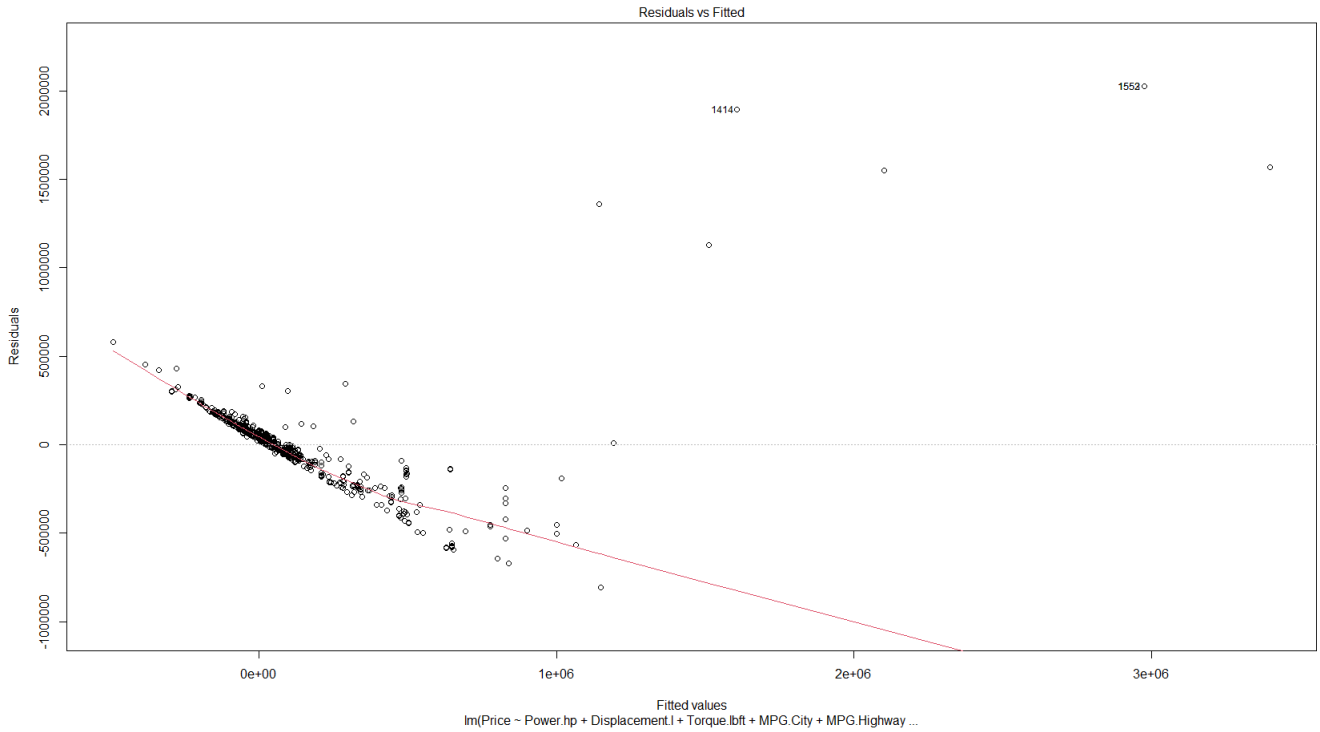

FIGURE 20 – Visualisation de l'orthogonalité du modèle RLM au sens du BIC.

En conclusion, il est impossible d'utiliser ce modèle pour établir des prévisions précises et fiables, car la plupart des hypothèses n'ont pas été validées. Par conséquent, il est inutile de formuler des prévisions dans ce contexte.

5 Conclusion

Parmi les 76 variables que nous avions à notre disposition, seuls 7 sont à la fois exploitables et significatifs dans le modèle, au sens du BIC. Plus précisément, le prix d'une voiture est significativement expliqué par sa puissance (en chevaux), sa taille de la chambre de combustion de chaque piston (en litre), son couple du moteur (en livre/pieds), sa consommation en Miles par Galon en ville, sa consommation en Miles par Galon sur une autoroute et sa taille (largeur et hauteur, en pouces).

Parmi les modèles Lasso, Ridge, RLM au sens de l'AIC et RLM au sens du BIC, c'est ce dernier qui présente une erreur de prévision plus faible, laissant supposer qu'il pourrait être plus précis pour les prévisions. Toutefois, il convient de noter que ce modèle ne peut pas être utilisé de manière efficace pour effectuer des prévisions en raison du non-respect de nombreuses hypothèses (presque toute, sauf une seule). Par conséquent, toute prévision basée sur ce modèle serait à la fois peu précise et peu fiable.

Table des figures