



Compte-rendu d'analyse

valant évaluation dans le cadre de :

Diplôme : Master de mathématiques appliquées, 1^{ère} année

Année universitaire : 2022-2023

Module d'enseignement : SEP0832

Responsable : Amor Keziou

Comptant pour : 50 %

Projet sur les méthodes d'échantillonnage

Elif ERTAS

elif.ertas@etudiant.univ-reims.fr

Fatima AAGOUR

fatima.aagour@etudiant.univ-reims.fr

Ibrahima TANDIA

ibrahima.tandia@etudiant.univ-reims.fr

Aldjia IBEGHOUCHENE

aldjia.ibeghouchene@etudiant.univ-reims.fr

Métriques

Finalisé le : 18 mai 2023

Page(s) : 17

Références(s) : 0

Figure(s) : 0

Table(s) : 1

Théorème(s) : 0

Résumé : Ce projet présente les différentes méthodes de rééchantillonage, de validation croisée et de bootstrap pour l'estimation de l'erreur en régression.

Mots-clés : Regression, Prediction, Erreur.

Matériel supplémentaire :

git: <https://github.com/pregnault/urcadown>

bookdown: https://bookdown.org/yih_huynh/Guide-to-R-Book/diamonds.html

Table des matières

1	Introduction	3
2	Explication de la base de donnée	3
3	Construction du modèle RLM complet	4
3.1	Non-corrélation des erreurs	5
3.2	Homoscédasticité	6
3.3	Linéarité	6
3.4	Normalité	7
4	Classement des variables explicatives	8
5	Influence des outliers	9
6	Sélection du meilleur modèle	9
6.1	Méthode exhaustive	9
6.2	Méthode génétique	10
7	Estimation de l'erreur de prévision des modèles sélectionnés	10
7.1	Méthode de l'ensemble de validation	10
7.2	Méthode LOOCV	11
7.3	Méthode K-fold cross validation	12
7.4	Bootstrap	12
8	Conclusion	17

1. Introduction

Dans le cadre de ce projet, nous allons étudier la variation de prix sur des diamants en fonction de leurs nombreuses caractéristiques telles que la taille, le poids ou leur couleur.

Nous cherchons donc à trouver une solution à la problématique suivante :

Comment peut-on expliquer le prix d'un diamant en fonction de ses caractéristiques ?

Afin de répondre à notre problématique, nous allons mettre en place des modèles de régressions linéaires multiples puis nous chercherons le meilleur modèle qui décrit au mieux la variable expliquée **price**.

Notre travail se divisera en 3 étapes : la construction du modèle de régression linéaire multiple complet et la vérification de ses hypothèses, la sélection des meilleurs modèles et l'estimation de l'erreur de prévision de ces derniers afin de sélectionner le meilleur.

En ce qui concerne la répartition des tâches, voici ce que chacun d'entre nous à apporter :

- Fatima : Recherche de la base de donnée, construction du modèle RLM et vérification des hypothèses d'homoscédasticité et d'auto-corrélation, analyse des résidus, sélection du meilleur modèle avec la méthode exhaustive.
- Elif : Rédaction et mise en page, explication de la base de donnée, vérification des hypothèses de linéarité et de normalité, influence des outliers, sélection du meilleur modèle avec la méthode génétique, conclusion.
- Aldjia : Estimation de l'erreur de prévision par l'approche de l'ensemble de validation, leave-one-out VC et K-fold CV pour chacun des modèles, mise en place de l'outil bootstrap.
- Ibrahima : Interprétation du bootstrap.

2. Explication de la base de donnée

La base de données **diamonds** est une collection de données réelles sur les caractéristiques et les prix de plus de 50 000 diamants. Elle contient notamment leur poids, leur qualité de coupe, leur couleur, leur pureté et leur prix. Les données ont été recueillies auprès de détaillants de diamants aux États-Unis par la société d'analyse des données Tipper et reprises par ggplot2, une bibliothèque de visualisation de données de R.

Les diamants inclus dans la base de données ont été sélectionnés de manière aléatoire à partir d'un échantillon représentatif de diamants disponibles sur le marché. Les caractéristiques des diamants ont été évaluées et mesurées par des gemmologues professionnels, et les prix ont été enregistrés par les détaillants.

Les informations contenues dans cette base de données sont utilisées pour prédire le prix des diamants en fonction de leurs caractéristiques. L'analyse des données de cette base peut aider à comprendre les facteurs qui influencent le prix des diamants et peut être utile pour les acheteurs et les vendeurs de diamants.

La base de donnée contient 10 variables :

- **price** : le prix du diamant en dollars américains
- **carat** : le poids du diamant en carats
- **cut** : la qualité de la coupe du diamant (Fair, Good, Very Good, Premium, Ideal)

- **color** : la couleur du diamant (de J, la plus jaune, à D, la plus blanche)
- **clarity** : la pureté du diamant (I1 (inclusions visibles à l'oeil nu), SI1 (petites inclusions), SI2 (inclusions visibles à l'oeil nu), VS1 (inclusions très petites), VS2 (inclusions petites), VVS1 (inclusions très très petites), VVS2 (inclusions très petites))
- **depth** : la profondeur totale du diamant en pourcentage de la largeur moyenne
- **table** : la largeur du plateau du diamant en pourcentage de la largeur moyenne
- **x** : la longueur en millimètres
- **y** : la largeur en millimètres
- **z** : la profondeur en millimètres

TABLE 1 – Tableau des 6 premières observations de la base de données diamonds

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

3. Construction du modèle RLM complet

Nous cherchons à trouver parmi les 9 caractéristiques que nous avons dans la base de données, celles qui expliquent le mieux le prix des diamants.

Cela se traduit par la recherche des variables explicatives qui décrivent le mieux notre variable cible.

Dans un premier temps, nous allons étudier le modèle complet, celui contenant toutes les variables explicatives.

Call:

```
lm(formula = price ~ ., data = diamonds)
```

Residuals:

Min	1Q	Median	3Q	Max
-21376.0	-592.4	-183.5	376.4	10694.2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5753.762	396.630	14.507	< 2e-16 ***
carat	11256.978	48.628	231.494	< 2e-16 ***
cut.L	584.457	22.478	26.001	< 2e-16 ***
cut.Q	-301.908	17.994	-16.778	< 2e-16 ***
cut.C	148.035	15.483	9.561	< 2e-16 ***
cut^4	-20.794	12.377	-1.680	0.09294 .
color.L	-1952.160	17.342	-112.570	< 2e-16 ***
color.Q	-672.054	15.777	-42.597	< 2e-16 ***
color.C	-165.283	14.725	-11.225	< 2e-16 ***
color^4	38.195	13.527	2.824	0.00475 **

```

color^5      -95.793    12.776   -7.498 6.59e-14 ***
color^6      -48.466    11.614   -4.173 3.01e-05 ***
clarity.L    4097.431   30.259   135.414 < 2e-16 ***
clarity.Q    -1925.004   28.227   -68.197 < 2e-16 ***
clarity.C    982.205    24.152   40.668 < 2e-16 ***
clarity^4    -364.918    19.285   -18.922 < 2e-16 ***
clarity^5    233.563    15.752   14.828 < 2e-16 ***
clarity^6     6.883     13.715    0.502  0.61575
clarity^7    90.640     12.103    7.489  7.06e-14 ***
depth        -63.806    4.535   -14.071 < 2e-16 ***
table        -26.474     2.912   -9.092 < 2e-16 ***
x            -1008.261   32.898   -30.648 < 2e-16 ***
y              9.609     19.333    0.497  0.61918
z            -50.119     33.486   -1.497  0.13448
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 1130 on 53916 degrees of freedom
 Multiple R-squared: 0.9198, Adjusted R-squared: 0.9198
 F-statistic: 2.688e+04 on 23 and 53916 DF, p-value: < 2.2e-16

On observe que toutes les variables sont significatives sauf clarity^6, y et z.

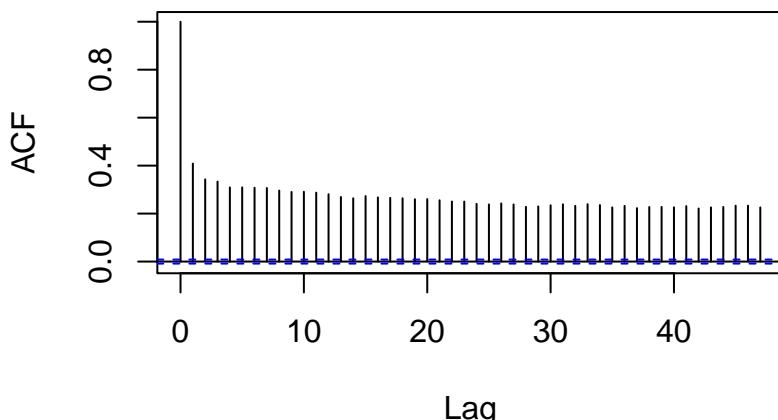
3.1. Non-corrélation des erreurs

Nous voulons tester si les erreurs sont corrélées ou non. Pour cela on utilise le test de Durbin-Watson .

H_0 : les erreurs sont non-correlées

H_1 : les erreurs sont corrélées

Auto-corrélations des erreurs



Durbin-Watson test

```
data: RLM
DW = 1.1831, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is not 0
```

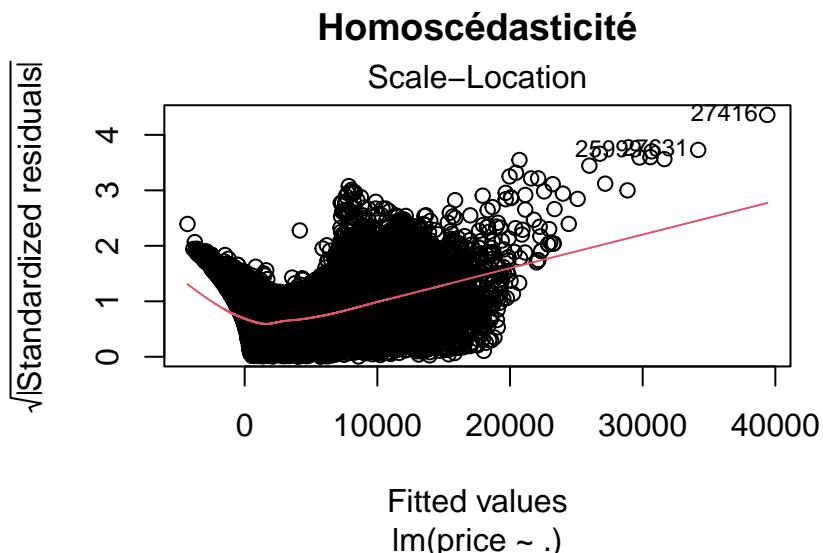
La p-value est significative donc on rejette l'hypothèse nulle, les erreurs sont donc corrélées.

3.2. Homoscédasticité

Nous voulons vérifier l'hypothèse d'homoscédasticité. Pour cela on utilise le test de **Breusch-Pagan**.

H_0 : l'erreur est homoscédastique

H_1 : l'erreur est hétéroscléastique



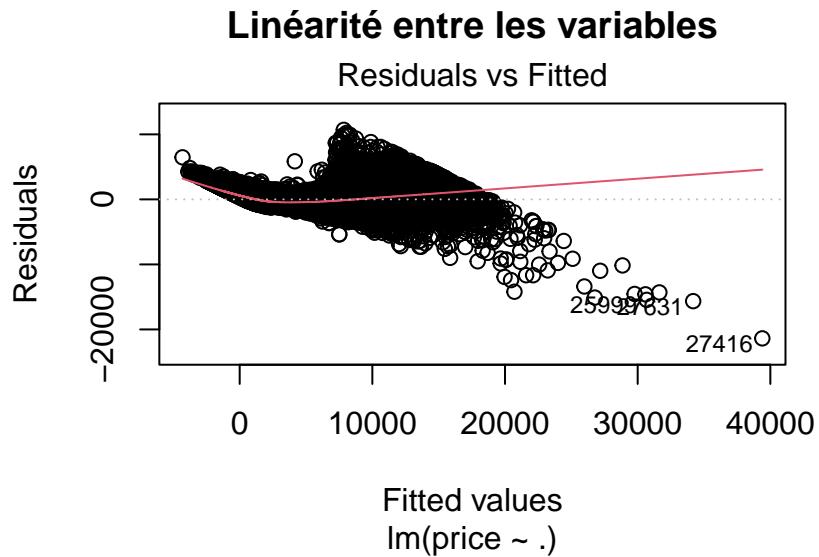
Breusch-Pagan test

```
data: RLM
BP = 102326, df = 23, p-value < 2.2e-16
```

On obtient une p-value significative donc on rejette l'hypothèse nulle. L'erreur est donc hétéroscléastique.

3.3. Linéarité

Nous voulons tester l'hypothèse de linéarité entre la variable réponse **price** et les variables explicatives. Pour cela nous allons appliquer une régression linéaire local des résidus en fonction des valeurs ajustées.



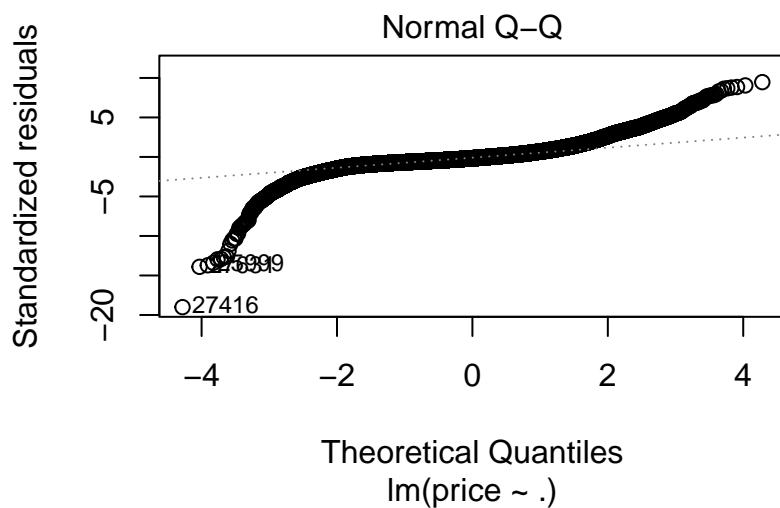
La courbe rouge représentant la droite d'ajustement, elle approche légèrement la courbe horizontale en pointillée. Cela nous permet de valider la linéarité entre les variables. On estime donc que la linéarité de ce modèle est vérifiée.

3.4. Normalité

On veut tester la normalité de l'erreur du modèle. Pour cela nous allons faire des visualisation graphique puis utiliser le test de **Kolmogorov-Smirnov**.

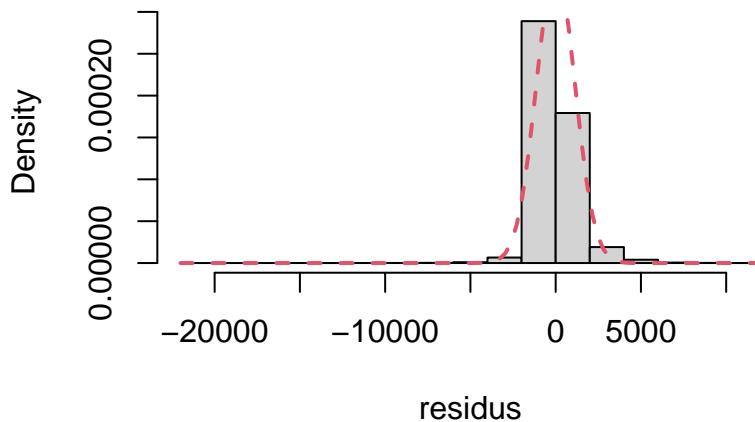
$$H_0 : \text{l'erreur est normale}$$

$$H_1 : \text{l'erreur n'est pas normale}$$



On observe que les points ne se situent pas autour de la droite en pointillée, cela veut dire que l'erreur ne suit pas une loi normale.

Histogramme des résidus



La courbe rouge représentant la densité de la loi normale, on observe que l'histogramme ne suit pas la même tendance donc cela confirme bien que l'erreur ne suit pas une loi normale.

Nous allons effectuer le test de **Kolmogorov-Smirnov** pour vérifier l'hypothèse.

```
Warning in ks.test.default(residuals(RLM), "pnorm"): aucun ex-aequo ne devrait
être présent pour le test de Kolmogorov-Smirnov
```

Asymptotic one-sample Kolmogorov-Smirnov test

```
data: residuals(RLM)
D = 0.59223, p-value < 2.2e-16
alternative hypothesis: two-sided
```

La p-value est significative donc on rejette l'hypothèse nulle, l'erreur ne suit pas une loi normale.

4. Classement des variables explicatives

Dans cette partie, nous allons classer les variables explicatives de la plus significative à la moins significative selon les valeurs des p-values du test de Fisher.

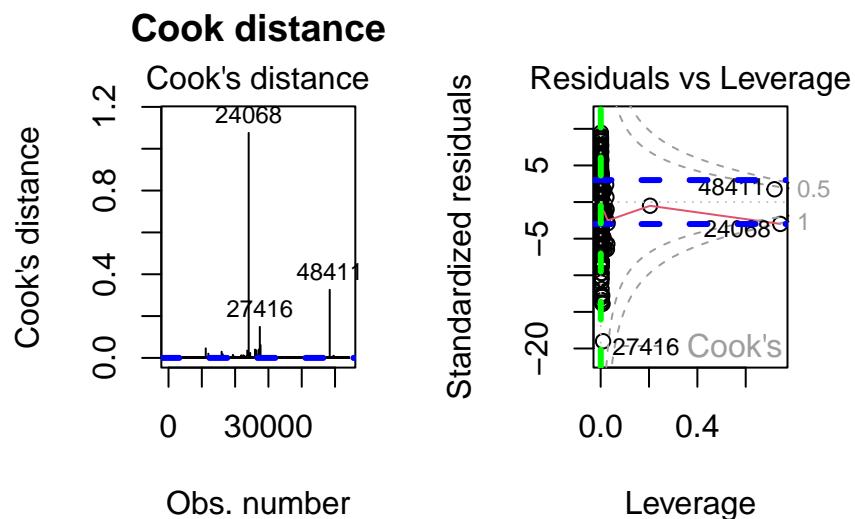
carat	cut	color	clarity	x	table
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.622873e-17
depth	z	y			
4.920523e-02	1.344776e-01	7.623472e-01			

On observe que les variables les plus significatives sont **carat** , **cut** , **color** , **clarity** , et **x** . Celle qui est la moins significative est **y** .

5. Influence des outliers

Dans cette partie nous essayerons de détecter les variables qui ont des outliers et des points leviers extrêmes. Un petit changement à leur niveau peut engendrer de grands changements.

Afin de mesurer l'influence d'une observation dans le modèle, nous pouvons se servir de la distance de Cook.



Une observation est considérée comme excessivement influente si sa distance de Cook est supérieure à $4/(53940 - 9 - 1) = 7.417022e - 05$.

On observe ici que les diamants n°24 068, 27 416 et 48 411 sont excessivement influente.

Tout les points ayant un résidus studentisés supérieur à 3 en valeur absolue sont considérés comme des outliers. Les diamants n°24 068 et 48 411 sont donc des outliers.

Tout les points ayant un point levier supérieure à $2 * (9 + 1)/53940 = 0.0003$ sont considérés comme des points leviers extrêmes.

Ici, on remarque que les diamants n°24 068 et 48 411 ont une influence supérieure à celle du reste des diamants.

6. Sélection du meilleur modèle

Afin de trouver le meilleur modèle, une multitude de méthodes de sélection de modèle existe. Nous allons donc tester les deux méthodes qui nous semblent être les plus pertinentes, il s'agit de la méthode exhaustive et génétique.

6.1. Méthode exhaustive

Une méthode de sélection est l'algorithme exhaustive. Nous allons nous baser sur les critères de l'AIC et du BIC.

Selon le critère de l'AIC, le modèle sélectionner est :

"price ~ 1 + cut + color + clarity + carat + depth + table + x + z"

Selon le critère du BIC, le modèle sélectionné est :

"price ~ 1 + cut + color + clarity + carat + depth + table + x"

6.2. Méthode génétique

Une autre méthode de sélection est l'algorithme génétique. Nous allons nous baser sur les critères de l'AIC et du BIC.

Selon le critère de l'AIC, le modèle sélectionné est :

"price ~ 1 + cut + color + clarity + carat + depth + table + x + z"

Selon le critère du BIC, le modèle sélectionné est :

"price ~ 1 + cut + color + clarity + carat + depth + table + x"

Nous avons trouvé plusieurs modèles selon les méthodes et les critères. Pour la suite de l'étude, nous allons estimer les erreurs de prévision de ces modèles afin de sélectionner celui qui est le plus optimal.

7. Estimation de l'erreur de prévision des modèles sélectionnés

Nous estimerons correctement l'erreur théorique de prévision de chaque modèle, afin de choisir le modèle ayant l'erreur estimée la plus faible. Les méthodes de validation croisée permettent d'évaluer efficacement cette erreur. On présente ici trois méthodes différentes : méthode de l'ensemble de validation, méthode K-fold CV et méthode LOOCV.

Modèle 1 : price ~ 1 + cut + color + clarity + carat + depth + table + x

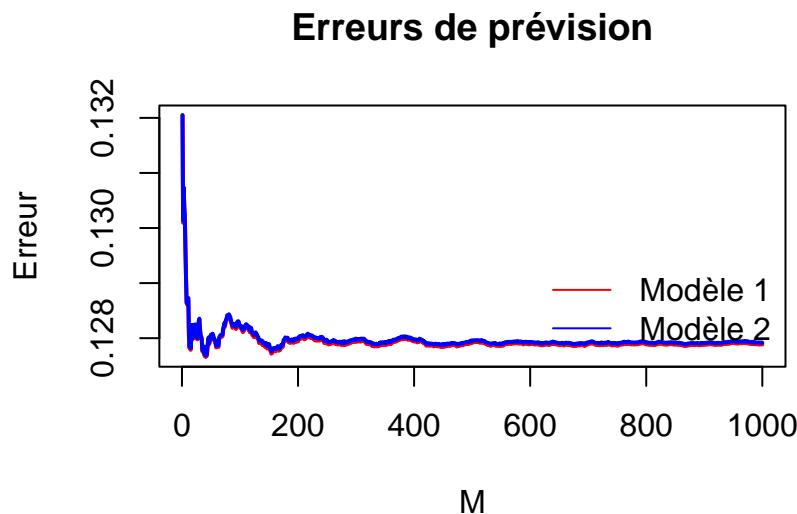
Modèle 2 : price ~ 1 + cut + color + clarity + carat + depth + table + x + z

7.1. Méthode de l'ensemble de validation

La méthode de l'ensemble de validation est une méthode d'apprentissage et de validation. C'est pourquoi, on a commencé par séparer la base de données en 2 parties, la première contient les 2/3 de la base, ces données seront utilisées pour l'apprentissage et le restant pour tester les prédictions en évaluant la valeur de l'erreur.

Puisque la base de données a été séparé en deux, on peut supposer que l'estimation de l'erreur n'est pas totalement fiable. Donc, pour avoir une meilleure estimation de l'erreur, on a estimé 1000 fois l'erreur en prenant à chaque fois de nouvelles valeurs dans la base de données d'apprentissage, donc la régression linéaire change et donc les valeurs prédites également.

On a donc cherché la meilleure estimation de l'erreur pour chacun des 2 modèles qu'on a sélectionné précédemment. Pour avoir une idée visuelle, pour voir comment varie l'estimation de l'erreur en fonction des itérations on a tracer les graphiques pour chaque modèle :



```
[1] "Résultats des estimations par la méthode de l'ensemble de validation : "
[2] "Estimation de l'erreur du modele1 =  0.127885106426773"
[3] "Estimation de l'erreur du modele2 =  0.127916765647575"
```

On observe donc que plus le nombre d'itérations augmente, plus l'erreur se stabilise et donc approche au mieux l'erreur réelle. Le modèle 1 possède une plus faible erreur que le deuxième modèle selon la méthode d'apprentissage/validation.

Nous allons effectuer une autre méthode pour vérifier ce résultat.

7.2. Méthode LOOCV

La méthode leave-one-out Cross-Validation est l'une des méthodes de re-échantillonnage qui va nous servir à estimer l'erreur théorique de notre modèle RLM. Elle permet également de sélectionner le meilleur modèle tout en évitant le problème de sur-ajustement. Le meilleur modèle est donc celui qui possède la plus faible erreur estimée.

Algorithmiquement cela se calcule de cette façon :

```
n <- n <- nrow(diamonds) # nombre d'observations
modele1 <- glm(formula = price ~ 1 + cut + color + clarity + carat + depth + table +
  x, data = bdd_diamonds)
estimation_erreur_modele1 <- cv.glm(data = bdd_diamonds, glmfit = modele1, K = n)$delta[1] #erreur

modele2 <- glm(formula = price ~ 1 + cut + color + clarity + carat + depth + table +
  x + z, data = bdd_diamonds)
estimation_erreur_modele2 <- cv.glm(data = bdd_diamonds, glmfit = modele2, K = n)$delta[1] #erreur

print(c("Résultats des estimations par LOOCV : ", paste("Estimation de l'erreur du modele1 = ",
  as.character(estimation_erreur_modele1)), paste("Estimation de l'erreur du modele2 = ",
  as.character(estimation_erreur_modele2))))
```

Dans notre cas, n étant le nombre d'observations qui correspond à 53940 l'algorithme prend beaucoup de temps à s'exécuter. Ce n'est donc pas une méthode à privilégier dans notre étude.

Essayons alors une autre méthode.

7.3. Méthode K-fold cross validation

La méthode K-fold cross validation consiste à diviser de manière aléatoire les données en K (ici K=10) groupes et de répéter le calcul de l'erreur K fois en prenant le premier groupe comme ensemble de validation et les K-1 pour ajuster le modèle. On remarque de plus que si on prend le même K que dans la méthode LOOCV alors on trouve le même résultat.

Nous appliquons donc cette méthode aux deux modèles sélectionnés dans l'objectif de trouver le meilleur modèle pour notre jeu de données.

```
[1] "Résultats des estimations par 10-fold CV : "
[2] "Estimation de l'erreur du modele1 =  0.127882243485329"
[3] "Estimation de l'erreur du modele2 =  0.127883027268616"
```

On observe une erreur semblable l'une à l'autre.

7.4. Bootstrap

Le Bootstrap est un outil performant pour évaluer l'incertitude d'un estimateur ou d'une méthode d'apprentissage. Nous allons l'utiliser pour estimer l'erreur théorique de notre modèle, c'est à dire évaluer l'efficacité de notre estimateur.

7.4.1 Modèle 1

```
(Intercept)      cut.L      cut.Q      cut.C      cut^4      color.L
5935.106750    584.717366 -302.036924  148.065481 -21.252994 -1952.127536
      color.Q      color.C      color^4      color^5      color^6      clarity.L
-672.207017   -165.451426   38.260893  -95.815562 -48.440803  4096.912291
      clarity.Q     clarity.C     clarity^4     clarity^5     clarity^6     clarity^7
-1924.681238    982.003925 -364.870240  233.449467   6.973269   90.738422
      carat        depth       table         x
11256.968051   -66.769313  -26.457328 -1029.477901
```

```
(Intercept)      cut.L      cut.Q      cut.C      cut^4      color.L
5600.02559     631.55149  -341.92341   180.39065 -34.81660 -1940.56898
      color.Q      color.C      color^4      color^5      color^6      clarity.L
-663.05079    -157.77326   56.06089  -103.09030 -42.27702  4159.60466
      clarity.Q     clarity.C     clarity^4     clarity^5     clarity^6     clarity^7
-2038.64382    1051.21799 -404.21067   262.81235 -15.67683  102.40942
      carat        depth       table         x
11243.90836   -63.65931  -24.69928 -1028.00092
```

```
(Intercept)      cut.L      cut.Q      cut.C      cut^4      color.L
4702.698556    550.542103 -284.430970  144.191074 -1.851404 -1946.325268
      color.Q      color.C      color^4      color^5      color^6      clarity.L
-652.882246   -164.661946   60.992975  -79.459289 -31.143230  4282.770316
      clarity.Q     clarity.C     clarity^4     clarity^5     clarity^6     clarity^7
```

```
-2074.605684  1131.408177  -455.654733   287.545128   -31.305828    84.676114
      carat        depth       table         x
10851.434192  -60.211464  -21.758809  -880.626436
```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

```
boot(data = diamonds, statistic = f_estimateurs_w, R = 2000)
```

Bootstrap Statistics :

	original	bias	std. error
t1*	5935.106750	38.60703168	624.074192
t2*	584.717366	-0.59352647	28.786562
t3*	-302.036924	0.51718311	23.275493
t4*	148.065481	-0.15272172	17.705944
t5*	-21.252994	-0.48091291	12.397842
t6*	-1952.127536	-1.08586377	22.788568
t7*	-672.207017	-0.14569777	18.699148
t8*	-165.451426	0.20244010	16.429221
t9*	38.260893	-0.05895468	14.018605
t10*	-95.815562	-0.14252777	13.207568
t11*	-48.440803	-0.08191026	11.359107
t12*	4096.912291	-2.03783424	68.671080
t13*	-1924.681238	2.42963107	65.023783
t14*	982.003925	-1.32887813	52.557966
t15*	-364.870240	1.59743682	38.276712
t16*	233.449467	-0.02500584	23.594266
t17*	6.973269	0.58299468	15.199919
t18*	90.738422	-0.31894858	11.758674
t19*	11256.968051	12.47619578	186.803649
t20*	-66.769313	-0.22886812	5.441078
t21*	-26.457328	-0.08163992	3.034744
t22*	-1029.477901	-5.07119394	74.002421

Ensuite, on fait une comparaison des valeurs de notre estimateur et celles de la régression linéaire.

Call:

```
lm(formula = price ~ 1 + cut + color + clarity + carat + depth +
  table + x, data = diamonds)
```

Residuals:

Min	1Q	Median	3Q	Max
-21385.0	-592.4	-183.7	376.5	10694.6

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5935.107	378.328	15.688	< 2e-16 ***

```

cut.L      584.717   22.476   26.015 < 2e-16 ***
cut.Q     -302.037   17.983  -16.795 < 2e-16 ***
cut.C      148.065   15.459   9.578 < 2e-16 ***
cut^4     -21.253   12.364  -1.719  0.08562 .
color.L    -1952.128  17.342 -112.568 < 2e-16 ***
color.Q    -672.207   15.777  -42.608 < 2e-16 ***
color.C    -165.451   14.724  -11.236 < 2e-16 ***
color^4    38.261    13.526   2.829  0.00468 **
color^5    -95.816   12.776  -7.500  6.50e-14 ***
color^6    -48.441   11.614  -4.171  3.04e-05 ***
clarity.L  4096.912  30.253  135.423 < 2e-16 ***
clarity.Q  -1924.681  28.224  -68.192 < 2e-16 ***
clarity.C  982.004   24.149   40.664 < 2e-16 ***
clarity^4  -364.870   19.285  -18.920 < 2e-16 ***
clarity^5  233.449   15.751   14.822 < 2e-16 ***
clarity^6  6.973    13.715   0.508  0.61114
clarity^7  90.738    12.103   7.497  6.63e-14 ***
carat      11256.968  48.600  231.626 < 2e-16 ***
depth      -66.769    4.091  -16.322 < 2e-16 ***
table      -26.457    2.911  -9.089 < 2e-16 ***
x          -1029.478   20.549  -50.098 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 1130 on 53918 degrees of freedom
 Multiple R-squared: 0.9198, Adjusted R-squared: 0.9198
 F-statistic: 2.944e+04 on 21 and 53918 DF, p-value: < 2.2e-16

7.4.2 Modèle 2

Même chose pour le second modèle :

(Intercept)	cut.L	cut.Q	cut.C	cut^4	color.L
5768.781806	584.599569	-302.211146	148.445883	-20.619129	-1952.178617
color.Q	color.C	color^4	color^5	color^6	clarity.L
-672.074909	-165.277290	38.193376	-95.779638	-48.452123	4097.612686
clarity.Q	clarity.C	clarity^4	clarity^5	clarity^6	clarity^7
-1925.132656	982.321523	-364.975671	233.634885	6.870585	90.622159
carat	depth	table	x	z	
11257.752405	-64.002583	-26.501176	-1000.354147	-47.925349	
(Intercept)	cut.L	cut.Q	cut.C	cut^4	color.L
5633.597166	576.562450	-276.940909	145.850942	-13.870252	-1960.535078
color.Q	color.C	color^4	color^5	color^6	clarity.L
-674.442704	-153.657211	4.032955	-79.997424	-45.292563	4131.825182
clarity.Q	clarity.C	clarity^4	clarity^5	clarity^6	clarity^7
-1985.055765	1012.385205	-382.249570	224.739075	3.781080	80.721153
carat	depth	table	x	z	
11491.165242	-57.629070	-25.604896	-957.829135	-259.493568	
(Intercept)	cut.L	cut.Q	cut.C	cut^4	color.L

```
5291.54103  581.08563 -295.69576   159.90764  -51.42187 -1938.65222
  color.Q     color.C    color^4     color^5    color^6 clarity.L
-639.90007 -150.97068   38.60274  -102.01933  -61.01016 4120.02161
  clarity.Q   clarity.C  clarity^4  clarity^5  clarity^6 clarity^7
-1905.80416  993.22859 -357.50612   224.33528  -13.68487  75.36256
  carat       depth      table        x         z
11270.82954 -57.79911 -23.60212 -951.01119 -147.56007
```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

```
boot(data = diamonds, statistic = f_estimateurs_w, R = 2000)
```

Bootstrap Statistics :

	original	bias	std. error
t1*	5768.781806	-96.91961132	692.846316
t2*	584.599569	-0.14803635	29.811182
t3*	-302.211146	0.04309746	24.346484
t4*	148.445883	0.35664769	18.217764
t5*	-20.619129	0.08613758	12.795388
t6*	-1952.178617	-0.22844255	22.624144
t7*	-672.074909	0.16117734	18.332258
t8*	-165.277290	0.28505562	16.568851
t9*	38.193376	-0.07943798	13.897913
t10*	-95.779638	-0.05176485	12.757707
t11*	-48.452123	-0.26281318	10.938440
t12*	4097.612686	-1.34987585	70.803062
t13*	-1925.132656	2.29910060	65.543498
t14*	982.321523	-1.11428596	54.866707
t15*	-364.975671	1.53629419	38.252907
t16*	233.634885	-0.43529327	23.527397
t17*	6.870585	0.35089103	15.462934
t18*	90.622159	-0.09019363	12.090014
t19*	11257.752405	6.46690123	185.879895
t20*	-64.002583	1.68221900	7.324008
t21*	-26.501176	0.04879901	3.115977
t22*	-1000.354147	16.63764092	91.512653
t23*	-47.925349	-31.05215335	86.615010

Call:

```
lm(formula = price ~ 1 + cut + color + clarity + carat + depth +
  table + x + z, data = diamonds)
```

Residuals:

Min	1Q	Median	3Q	Max
-21378.8	-592.5	-183.5	376.3	10694.1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5768.782	395.474	14.587	< 2e-16 ***
cut.L	584.600	22.476	26.010	< 2e-16 ***
cut.Q	-302.211	17.983	-16.805	< 2e-16 ***
cut.C	148.446	15.461	9.601	< 2e-16 ***
cut^4	-20.619	12.371	-1.667	0.09559 .
color.L	-1952.179	17.342	-112.572	< 2e-16 ***
color.Q	-672.075	15.777	-42.599	< 2e-16 ***
color.C	-165.277	14.725	-11.224	< 2e-16 ***
color^4	38.193	13.526	2.824	0.00475 **
color^5	-95.780	12.776	-7.497	6.64e-14 ***
color^6	-48.452	11.614	-4.172	3.02e-05 ***
clarity.L	4097.613	30.256	135.431	< 2e-16 ***
clarity.Q	-1925.133	28.226	-68.205	< 2e-16 ***
clarity.C	982.322	24.150	40.676	< 2e-16 ***
clarity^4	-364.976	19.285	-18.926	< 2e-16 ***
clarity^5	233.635	15.751	14.833	< 2e-16 ***
clarity^6	6.871	13.715	0.501	0.61640
clarity^7	90.622	12.103	7.487	7.13e-14 ***
carat	11257.752	48.602	231.630	< 2e-16 ***
depth	-64.003	4.517	-14.168	< 2e-16 ***
table	-26.501	2.911	-9.103	< 2e-16 ***
x	-1000.354	28.795	-34.740	< 2e-16 ***
z	-47.925	33.194	-1.444	0.14880

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 '	1		

Residual standard error: 1130 on 53917 degrees of freedom
Multiple R-squared: 0.9198, Adjusted R-squared: 0.9198
F-statistic: 2.81e+04 on 22 and 53917 DF, p-value: < 2.2e-16

On trouve de légère différence au niveau de certaines valeurs, l'erreur données par notre estimateur est plus élevée que celle donné par la fonction lm. De plus, pour chacun des deux modèles l'estimateur de la variance est élevé et celui du biais est faible : c'est le résultat lorsqu'on est grande dimension. Par le résultat de la régression, on choisit le premier modèle. Ce modèle semble être le meilleur, de plus il a été sélectionné par d'autres méthodes utilisées précédemment et il possède un biais faible : le but étant en général de diminuer le biais.

8. Conclusion

En analysant plusieurs méthodes de sélection de modèles et en estimant les erreurs de prévision, nous avons conclu que le prix d'un diamant peut être expliqué par différents facteurs clés. Ces facteurs comprennent la qualité de la coupe, la couleur, la pureté, le poids, la profondeur, la largeur et la longueur du diamant.

La qualité de la coupe joue un rôle important dans la détermination du prix d'un diamant. Une coupe précise et de haute qualité peut augmenter sa brillance et sa valeur. De même, la couleur du diamant est un autre aspect essentiel. Les diamants incolores ou légèrement colorés sont généralement considérés comme plus précieux que ceux qui présentent une teinte plus prononcée.

La pureté du diamant est également un facteur significatif. Les diamants sans imperfections visibles sont considérés comme plus purs et plus précieux. De plus, le poids du diamant, exprimé en carats, joue un rôle majeur dans sa valorisation. Les diamants plus lourds ont tendance à être plus coûteux, tout en tenant compte des autres critères de qualité.

La profondeur, la largeur et la longueur du diamant sont des mesures qui influencent également son prix. Ces dimensions affectent la manière dont la lumière interagit avec la pierre, ce qui peut avoir un impact sur son éclat et sa valeur.

Pour conclure, à travers notre étude, nous avons constaté que le prix d'un diamant est fortement influencé par la qualité de la coupe, la couleur, la pureté, le poids, la profondeur, la largeur et la longueur du diamant. Comprendre ces facteurs peut nous aider à mieux évaluer et apprécier la valeur d'un diamant lors de son estimation ou de son achat.