LSTAT2120 - Modèles linéaires - Projet
Prédiction de la consommation de carburant en fonction de différents paramètres

Axel Struys - Alexis Buckens

9 janvier 2017

1 Introduction

Pour réaliser ce projet, nous avons choisi une base de données en provenance de l'UCI machine learning repository. Afin d'analyser les données, nous avons utilisé SAS university edition et R. Ce dataset décrit plusieurs modèles de voitures en fonction de différentes caractéristiques. Plusieurs variables dépendantes pourraient être choisies pour ce dataset. Afin de restreindre notre analyse à une seule variable dépendante, nous avons choisi une question de recherche : "Quelles caractéristiques prédisent la consommation de carburant sur autoroute d'une voiture ?". Le dataset contenait à l'origine 26 variables : 16 continues et 10 nominales. Pour l'analyse, nous avons choisi 13 variables continues et 2 variables nominales. Nous avons exclu 4 variables continues car elles étaient des variables dépendantes. Nous avons exclu 8 variables nominales afin de simplifier le modèle. De plus, la variable highwaympg exprimait la consommation de carburant en miles par gallons. Afin d'exprimer cette consommation en unité européenne, nous l'avons transformée en L/100Km. Aussi, nous avons converti les unités de longueur en centimètres, les unités de poids en kilogrammes et les unités de volume en litres (dm^3). Il y avait 205 observations à l'origine, mais pour 6 d'entre-elles, il y avait des valeurs manquantes. Nous avons supprimé ces observations. Le dataset ainsi nettoyé contient 199 observations.

Description des variables Dans la table 1 nous présentons les statistiques descriptives pour les 13 variables continues, ainsi que les unités correspondantes. La table 2 présente les fréquences observées des deux variables discrètes : Aspiration et Engine type. Ces deux variables ont deux niveaux. On peut remarquer que les fréquences observées ne correspondent pas aux fréquences attendues : 40 observations par cellule. Pour l'analyse de la variance, nous sommes donc dans une situation où les cellules sont de tailles inégales. Brève description des variables :

1. Wheelbase : c'est la distance entre les roues avant et arrière de la voiture
2. Length/width/height : longueur/largeur/hauteur de la voiture.
3. Curbweigth : masse à vide du véhicule, en ordre de marche (sans les passagers et autres marchandises).
4. Enginesize : c'est le volume du moteur, exprimé en litres.
5. Bore : c'est le diamètre des pistons du moteur.
6. Stroke : c'est la distance totale parcourue par le piston lors d'un cycle.
7. Compression ratio : c'est le taux de compression du piston, c'est un ratio entre le volume maximal en début de cycle, et le volume minimal que le piston comprime, juste avant l'explosion.
8. Horsepower : Cheval-vapeur en français. C'est une unité qui exprime la puissance du moteur.
9. Peakrpm : nombre maximal de tour/minutes du moteur.
10. Fueltype : type de carburant (essence ou diesel)
11. Aspiration : admission d'air classique (atmosphérique) ou turbocomprimée.
12. Highwaylkm100 : consommation de carburant sur l'autoroute, en litres par 100 km.

Enfin, la table 3 présente la moyenne de consommation sur autoroute en fonction du type de carburant et du type d'aspiration d'air. Ce tableau est représenté graphiquement sur la figure 1. De façon purement descriptive, on peut constater une consommation de carburation supérieure pour une aspiration de type turbo, ainsi qu'une consommation supérieure pour l'essence par rapport au diesel. Néanmoins, nous devrons confirmer ces affirmation en testant les effets principaux de ces variables, ce que nous feront plus loin dans ce travail.

2 Sélection du modèle de régression

2.1 Multicolinéarité

Afin de sélectionner un bon modèle de régression, il faut en premier lieu éviter la multicolinéarité. Celle-ci est mesurée par le VIF (variance inflation factor). Celui-ci indique à quel point une variable augmente artificiellement la somme des carrés du modèle. Aussi, c'est une mesure de la corrélation de la variable avec les autres variables du modèle. Une variable est donc redondante si la proportion de la variance qu'elle explique est aussi expliquée par une autre variable du modèle. Si cette proportion est trop grande, alors il est utile de supprimer cette variable indépendante afin de simplifier le modèle. En termes d'algèbre linéaire, c'est une mesure de la dépendance

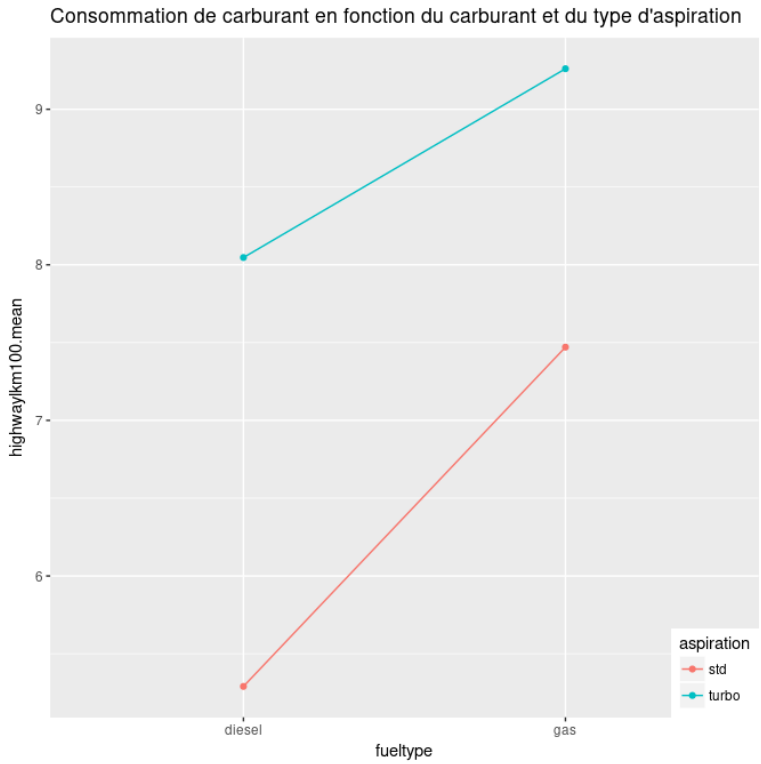


Figure 1 – Moyenne de la consommation (L/100km) en fonction du type de carburant et de l'aspiration

	mean	std.dev	median	min	max
wheelbase (cm)	251.09	15.46	246.38	219.96	307.09
length (cm)	442.21	31.72	439.93	358.39	528.57
width (cm)	167.40	5.53	166.12	153.16	183.64
height (cm)	136.69	6.08	137.41	121.41	151.89
curbweight (kg)	1160.66	239.50	1094.97	674.94	1844.31
enginesize (L)	2.10	0.68	1.97	1.00	5.34
bore (cm)	8.45	0.70	8.41	6.45	10.01
stroke (cm)	8.25	0.79	8.36	5.26	10.59
compressionratio	10.17	4.03	9.00	7.00	23.00
horsepower (ch)	104.15	40.05	95.00	48.00	288.00
peakrpm (tr/min)	5107.79	467.59	5200.00	4150.00	6600.00
highwaylkm100 (L/100km)	8.00	1.85	7.84	4.36	14.70

Table 1 – Statistiques descriptives des 13 variables continues

Fueltype	Aspiration	Count
Diesel	Standard	7
Diesel	Turbo	13
Gas	Standard	155
Gas	Turbo	24

Table 2 – Fréquences observées pour les variables Aspiration & Engine type

linéaire entre les variables. Si une variable est linéairement dépendante d'une autre, l'inversion de la matrice devient impossible. Dans ce cas, il n'est pas possible de calculer les termes de la régression. Plus une variable se rapproche de la dépendance linéaire, et plus le modèle devient instable. Un critère pour la grandeur VIF est qu'aucune variable ne peut avoir un VIF supérieur à 10. De plus, la moyenne des VIF ne doit pas être grandement supérieure à 1.

Nous avons lancé un premier modèle via SAS en incluant toutes les variables afin de vérifier la multicolinéarité. La table 4 montre que curbweight, enginesize et horsepower ont un VIF supérieur à / proche de 10. En réalisant un corrélation sur toutes les variables numériques (voir table 9 en annexe) on remarque que curbweight, enginesize et horsepower sont fortement corrélées entre elles. De plus, elles sont corrélées avec les autres variables. Ces corrélations font sens d'un point de vue théorique, étant donné que la puissance d'un moteur est directement liée à son volume. Néanmoins, horsepower est en moyenne moins corrélées avec les autres variables que les deux premières variables, et son VIF est plus petit. Nous allons donc supprimer ces deux variables, et garder horsepower. En réalisant à nouveau une régression, on constate que tous les VIF sont inférieurs à 10, et que le VIF moyen est de $3,363$. On peut considérer ceci comme acceptable.

Fueltype	Aspiration	highwaylkm100.mean
Diesel	Standard	5.44
Diesel	Turbo	8.11
Gas	Standard	7.90
Gas	Turbo	9.37

Table 3 – Moyenne de la consommation (L/100km) en fonction du type de carburant et de l'aspiration

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr > \|t\|$	Variance Inflation
Intercept	1	−2.68801	2.81974	−0.95	0.3417	0
wheelbase	1	−0.00454	0.00765	−0.59	0.5537	7.51130
length	1	0.00685	0.00407	1.68	0.0938	8.93723
width	1	0.00656	0.01855	0.35	0.7241	5.64192
height	1	0.01087	0.01079	1.01	0.3151	2.31344
curbweight	1	0.00381	0.00067915	5.61	<.0001	14.19486
enginesize	1	1.55782	0.20856	7.47	<.0001	10.80793
bore	1	−0.19462	0.09008	−2.16	0.0320	2.11806
stroke	1	−0.23587	0.06268	−3.76	0.0002	1.32045
horsepower	1	−0.00822	0.00337	−2.44	0.0158	9.79074
peakrpm	1	0.00048201	0.00013387	3.60	0.0004	2.10236
fueldummy	1	−1.09864	0.10096	−10.88	<.0001	1.98744
aspiration-dummy	1	0.58015	0.08188	7.09	<.0001	2.18891

Table 4 – Premier modèle incluant tous les facteurs : on peut voir que curbweight, enginesize et horsepower ont un vif supérieur à / proche de 10.

2.2 Méthodes de sélections du modèle de régression

Après avoir vérifié la multicolinéarité, plusieurs méthodes sont disponibles afin de sélectionner les variables à inclure dans le modèle final.

Méthodes de type 1 Il est possible de choisir le meilleur modèle parmi tous les modèles possibles, sur base d'un critère ; Nous comparerons ici le critère de Mallows et le critère du coefficient de détermination ajusté. Grâce à SAS, nous pouvons rapidement calculer le meilleur modèles. Un inconvénient majeur de ces méthodes est que plus le nombre de variables dans le modèle augmente, plus cela demande de ressources de calcul étant donné que le logiciel doit calculer toutes les permutations possibles. Dans notre cas, le nombre de variable de bases est raisonnable, ce qui permet l'emploi de ces méthodes. Les tables 10 et 11 en annexe permettent de comparer les 5 meilleurs modèles en fonction de ces deux critères. En fonction du critère de Mallows, dont l'estimateur est C(p), le meilleur modèle ne contient que

6 variables, et explique 81% de la variance. En utilisant le critère du R^2 ajusté, le meilleur modèle contient 8 variables, et explique aussi 81% de la variance. Néanmoins, il nous semble plus pertinent d'être parcimonieux dans le nombre de variables explicatives à inclure. Dans ce cas, le modèle incluant 6 variables est plus intéressant. On pourrait même considérer le quatrième meilleur modèle selon le critère de Mallows, qui ne contient que 5 variables et pourtant explique toujours 81% de la variance.

Méthodes de type 2 Une deuxième approche est d'inclure ou exclure séquentiellement des variables du modèle. Plusieurs méthodes sont disponibles : forward selection, backward elimination, forward stepwise, etc... Un avantage de ces méthodes est qu'elles sont plus économiques en termes de temps de calcul pour l'ordinateur. Dans le cas d'une inclusion séquentielle, on commence avec la variable indépendante qui explique le plus la variable dépendante (Le plus grand R^2). Ensuite, on va ajouter séquentiellement des variables, toujours en utilisant ce critère, jusqu'à atteindre le significance level to stop : la p valeur de la variable ajoutée est plus grande qu'une p valeur déterminée. Dans le cas d'une backward élimination, on commence avec toutes les variables, et on supprime séquentiellement. Pour une selection stepwise, on recalcule à chaque étape si on peut supprimer une variable précédemment ajoutée. Nous allons comparer ces trois techniques. Pour la méthode forward selection, nous avons choisi un seuil d'entrée de 0.1. Pour la méthode backward élimination, nous avons choisi un seuil d'élimination de 0.15. Enfin, pour la méthode forward stepwise, les seuils d'entrée et d'élimination étaient de 0.1 et 0.15 respectivement. Dans les tables 12, 13 et 14 en annexe, on peut constater que les 3 méthodes convergent vers la même solution : les variables heigth, stroke, bore et peakrpm sont éliminées, tandis que horsepower, length, fuel, wheelbase, aspiration et width sont conservées. Remarquons que ces méthodes proposent les mêmes 6 variables que le meilleur modèle selon le critère de Mallows.

Méthodes de type 3 Ce type de méthodes est utile lorsque le nombre de variables dans le modèle est très grand. Nous allons utiliser la méthode LASSO vue au cours, via SAS. Ici aussi, l'étape optimale retient les même 6 variables que précédemment (voir table 15 en annexe).

Modèle sélectionné Les différentes méthodes que nous avons présenté convergent vers le même modèle à sélectionner, contenant 6 variables.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	$Pr > F$
Model	6	553.85484	92.30914	141.16	<.0001
Error	192	125.55930	0.65395		
Corrected Total	198	679.41413			

Root MSE	0.80867	R-Square	0.8152
Dependent Mean	8.00271	Adj R-Sq	0.8094
Coeff Var	10.10501		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > $\|t\|$
Intercept	1	−13.19524	2.75453	−4.79	<.0001
wheelbase	1	0.01934	0.00913	2.12	0.0354
length	1	0.01492	0.00466	3.20	0.0016
width	1	0.04188	0.02374	1.76	0.0792
horsepower	1	0.02182	0.00239	9.14	<.0001
fueldummy	1	−0.72761	0.11985	−6.07	<.0001
aspirationdummy	1	0.19410	0.08603	2.26	0.0252

Table 5 – Résumé du modèle final sélectionné

3 Diagnostic du modèle linéaires et mesures correctrices

3.1 Hypothèses sous-jascentes au modèle linéaire

3.1.1 Vérification

Nous allons maintenant nous intéresser à la vérification des hypothèses sous-jascentes à la regression. Il y a trois principales hypothèses : l'homoscédasticité, la normalité des résidus et l'indépendance des observations. Une autre question pertinente est aussi de savoir s'il y a bien une linéarité de la fonction de régression. Les figures 2 et 3 présentent les différents graphes nous permettant de répondre à ces questions.

Homoscédasticité Nous pouvons répondre à cette question en observant le graphe des résidus par rapport aux valeurs prédites. Nous pouvons voir que les erreurs se répartissent aléatoirement autour de 0, indiquant une variance homogène. Il n'y a pas de structure claire se dégageant du graphe. Néanmoins, sur la figure 3, nous constatons que la variance des deux variables catégorielles est différente pour chaque niveau. Il y a un problème d'héteroskédasticité pour ces deux variables catégorielles.

Normalité des résidus Le Q-Q plot nous permet de comparer les quantiles de notre distribution aux quantiles de la distribution normale. Si tous les points sont sur la diagonale, alors notre distribution est normale. Nous pouvons constater que les résidus sont en moyenne situés sur la droite.

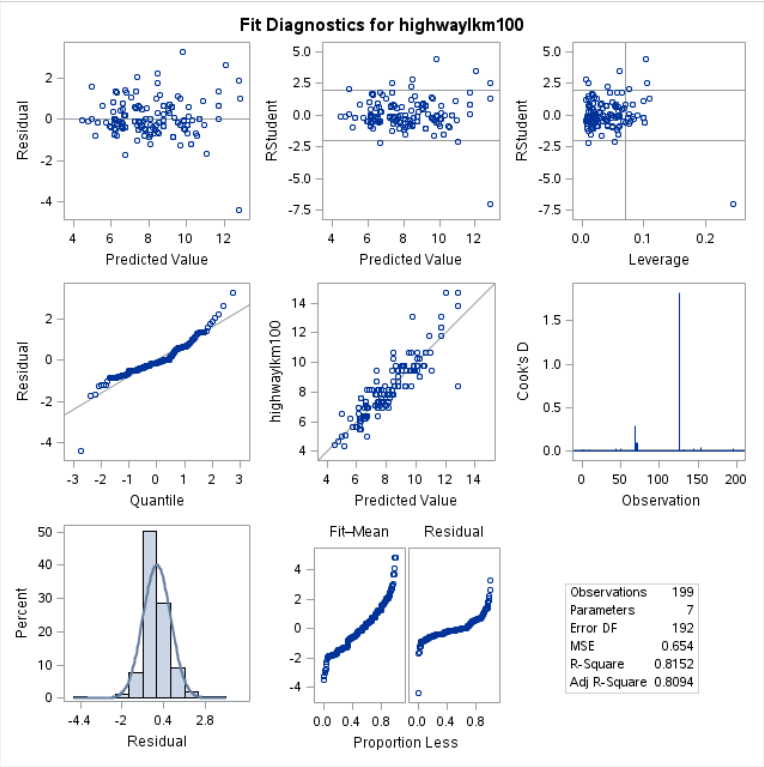


Figure 2 – Résumé des différents diagnostiques du modèle de régression

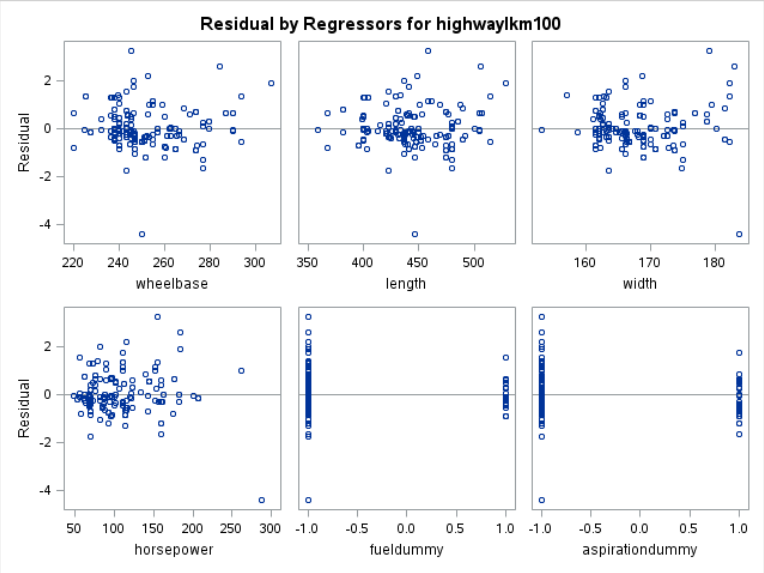


Figure 3 – Graphes des résidus par régresseurs

Durbin-Watson D	1.406
$Pr < DW$	<.0001
$Pr > DW$	1.0000
Number of Observations	199
1st Order Autocorrelation	0.288

Table 6 – Test de Durbin-Watson pour l'autocorrélation

Néanmoins, les valeurs extrêmes de la distribution s'écartent de la distribution normale. Sur l'histogramme des résidus, nous pouvons voir que leur densité se rapproche d'une normale. Nous pouvons donc faire l'hypothèse, prudente, que les résidus suivent une distribution normale.

Indépendance des observation D'un point de vue théorique, chaque modèle de voiture de ce dataset est différent. Néanmoins, il faut être prudent avec cette affirmation. En effet, il est raisonnable de penser qu'un constructeur produisant plusieurs modèle de voiture réutilise certains composants dans différents modèles, offrant des performances similaires. Dans ce cas, il pourrait exister des observations qui ne sont pas indépendantes les unes des autres. Via SAS, nous avons réalisé un test de Durbin-Watson pour l'autocorrélation. Celui-ci indique une autocorrélation positive signification de premier ordre de 0.288.

Linéarité de la fonction de régression Afin de répondre à cette question, il faut regarder les graphes des régresseurs en fonction des résidus. Sur la figure 3, on constate que chaque régresseur à une relation linéaire avec les résidus. Si un régresseur avait une relation quadratique, on aurait pu ajouter un terme au carré afin de corriger ce problème.

3.1.2 Correction

Correction de l'heteroskedasticité et de l'autocorrélation Nous avons vu que la variance des variables catégorielles n'était pas homogène. En outre, il y avait une autocorrélation de premier ordre significative. Elle était certes faible, mais nous avons voulu la corriger. Comme suggéré précédemment, la non-indépendance provient peut-être du fait que les différents modèles de voiture provenant d'un même constructeur ont des caractéristiques similaires. Comme suggéré par Neter(2004), une solution possible à cette autocorrélation est d'inclure des variables indépendantes susceptibles d'être la source de cette non indépendance. Ainsi, nous avons relancé un modèle en ajoutant les modèles de voitures comme variables indépendantes. Etant donné qu'il y avait 21 modèles de voitures, cela nous rajoute 20 variables catégorielles dans le modèle, le complexifiant énormément. De plus, afin de corriger l'hétéroscédasticité, nous avons réalisé une transformation de box-cox sur la variable dépendantes. La transformation de Box-Cox est une

transformation sur la variable dépendante, selon un paramètre λ. La formule est donnée par : $Y' = \frac{y^\lambda - 1}{\lambda}$ pour $\lambda \neq 0$. Ce paramètre λ est estimé par le logiciel selon la méthode du maximum de vraisemblance. Dans notre cas, SAS a choisi $\lambda = 0.5$.

3.2 Observations aberrantes et influentes

De plus, des observation aberrantes et influentes peuvent influencer la construction du modèle de régression. Nous avons calculés ces observations après correction du modèle par rapport aux hypothèses sous-jascentes. Les graphes relatifs aux observations aberrantes et influentes sont en annexe.

Observations aberrantes Sur le graphe 5a nous pouvons observer les leviers par rapport au numéro d'observation. C'est une mesure des observations aberrantes par rapport aux régresseurs. On peut utiliser comme limite d'identification des leviers $L > \frac{2p}{n}$. Dans notre cas, cette limite est de 0.27.

Sur le graphe 5b nous pouvons observer les résidus effacés studentisés par rapport au numéro d'observation. C'est une mesure des observations aberrantes par rapport à la réponse Y. On peut utiliser comme règle de décision $|d_i^*| > t_{171;0.975;=1.960}$ afin de considérer les observation comme aberrantes par rapport à Y.

Observations influentes Le graphe des DFFITS (figure 5c) permet d'observer les observations influentes par rapport aux valeurs prédites. On peut calculer la valeur limite, dans notre cas, par $|DFITT| = 2 \times \sqrt{27/199} = 0.7366$. Les observations en dehors de cette valeur seront considérées comme influentes. Sur le graphe, on peut voir que quelques valeurs peuvent être considérées comme influentes. Le graphe des distances de Cook (figure 5d) permet de détecter la présence d'observations influentes sur les valeurs prédites, ainsi que sur les coefficients de la régression. On peut décider qu'une observation est influente via la règle, dans notre cas, $D_i > F_{27,172;0.95} = 1.55$. On peut constater qu'aucune valeur ne dépasse ce seuil.

Correction des observations aberrantes Il y a plusieurs mesures correctrices possibles. La première est d'ajouter des variables indépendantes qui auraient pu être oubliées. Il est aussi possible d'ajouter des termes d'interaction et quadratiques. Une mesure simple, mais manquant d'élégance – et pourtant systématiquement employée dans le milieu de la recherche en psychologie – est de supprimer ces observation aberrantes. Dans notre cas, on pourrait essayer d'ajouter des variables ou des interactions. Néanmoins, notre première mesure correctrice à déjà ajouté beaucoup de variables dans le modèle, le complexifiant. En utilisant les critères plus haut, nous avons calculé qu'il y avait 25 observation influentes dans le dataset. Si nous décidions

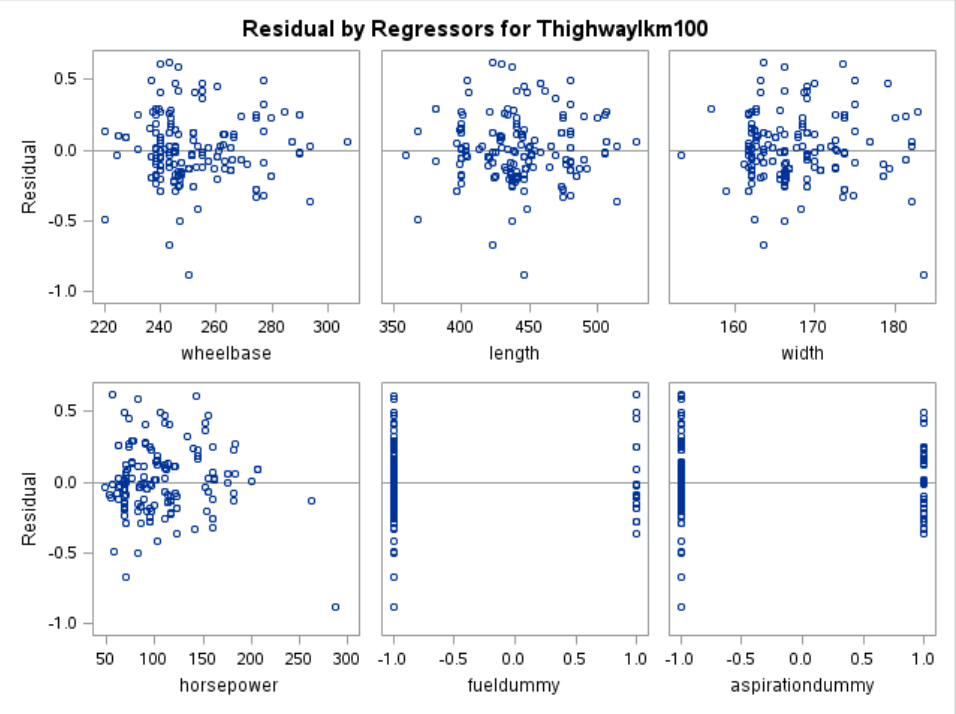


Figure 4 – Graphes des résidus par régresseurs, modèle corrigé

de supprimer ces observation, cela ferait beaucoup à enlever, risquant de biaiser l'échantillon. On pourrait se retrouver avec des modèles de voitures n'étant plus représentées dans cet échantillon. En effet, en analysant plus finement les observations considérées comme aberrantes, nous avons remarquées que celles-ci étaient exclusivement des observations en provenance de marques de voitures sous-représentées dans l'échantillon. Dans ce cas, il n'est pas convenable de supprimer ces observations. Nous allons donc les garder.

Correction des observation influentes Etant donné qu'il n'y a pas ou peu d'observation influentes, nous avons décidé de ne pas les enlever.

3.3 Modèle final après corrections

Sur le graphe 4 nous pouvons constater que les variances des prédicteurs catégoriels est maintenant beaucoup plus homogène. De plus, le test de Durbin-Watson(table 8) pour l'autocorrélation nous montre que celle-ci à été fortement réduite, passant a 0.054. Néanmoins, elle est toujours significative, indiquant qu'il existe toujours une petite source d'autocorrélation, non-identifiée.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	$Pr > F$
Model	26	71.45596	2.74831	50.93	<.0001
Error	172	9.28210	0.05397		
Corrected Total	198	80.73806			

Root MSE	0.23231	R-Square	0.8850
Dependent Mean	3.62184	Adj R-Sq	0.8677
Coeff Var	6.41400		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr > \|t\|$
Intercept	1	1.51831	1.16128	1.31	0.1928
wheelbase	1	0.00149	0.00350	0.43	0.6703
length	1	0.00757	0.00169	4.47	<.0001
width	1	−0.01703	0.00848	−2.01	0.0462
horsepower	1	0.00970	0.00099419	9.76	<.0001
fueldummy	1	−0.25512	0.04049	−6.30	<.0001
aspirationdummy	1	0.05965	0.03127	1.91	0.0581
makealfa_rom	1	0.19949	0.18123	1.10	0.2725
makeaudi	1	0.40580	0.12018	3.38	0.0009
makebmw	1	0.00791	0.11774	0.07	0.9465
makechevrole	1	−0.36691	0.17746	−2.07	0.0402
makedodge	1	0.03518	0.12475	0.28	0.7783
makehonda	1	0.04394	0.11959	0.37	0.7137
makeisuzu	1	−0.03369	0.15311	−0.22	0.8261
makejaguar	1	0.29561	0.16727	1.77	0.0790
makemazda	1	0.01375	0.10973	0.13	0.9004
makemercedes	1	0.75409	0.12097	6.23	<.0001
makemercury	1	−0.14793	0.24914	−0.59	0.5535
makemitsubis	1	0.00513	0.11195	0.05	0.9635
makenissan	1	−0.07021	0.10892	−0.64	0.5201
makepeugot	1	0.33728	0.10528	3.20	0.0016
makeplymouth	1	−0.02969	0.12848	−0.23	0.8175
makeporsche	1	−0.45421	0.18292	−2.48	0.0140
makesaab	1	−0.19943	0.13100	−1.52	0.1298
makesubaru	1	0.19383	0.11209	1.73	0.0856
maketoyota	1	−0.03065	0.09664	−0.32	0.7515
makevolkswag	1	0.06063	0.11148	0.54	0.5872

Table 7 – Modèle final après corrections

Durbin-Watson D	1.888
$Pr < DW$	0.0122
$Pr > DW$	0.9878
Number of Observations	199
1st Order Autocorrelation	0.054

Table 8 – Test de Durbin-Watson après correction via ajout des constructeurs de voitures comme variables de contrôle.

4 Interprétation du modèle

Etant donné que nous avons introduit les marques de voiture comme contrôle de l'autocorrélation, nous ne les interpréterons pas ici. Une première remarque quant à ce modèle est le R^2 de 0.88, indiquant un fort ajustement du modèle, captant la quasi intégralité de la variance, ne laissant que 12% d'erreur. Néanmoins, remarquons les problèmes d'overfitting que ce modèle implique. En effet, en utilisant 20 degrés de libertés pour corriger les problèmes d'autocorrélation, nous introduisons beaucoup de complexité dans le modèles. Il est important de savoir que trop de variables dans un modèle linéaire peut conduire à une surestimation de la significativité des coefficients. Dans le milieu de la recherche scientifique, il est souvent mal vu d'introduire trop de variables contrôles dans un modèle de régression. Mais ce n'est pas l'objet de ce travail. Notons aussi que la transformation de Box-Cox rend l'interprétation des coefficients plus difficile. En effet, ceux-ci ne reflètent plus la variable dépendante d'origine, mais une variable transformée. L'interprétation ne peut donc se faire qu'en terme de signes des coefficients (A moins de réaliser la transformation inverse). Globalement, la variable horsepower explique le plus la consommation de carburant. Ceci est logique étant donné que plus un moteur est puissant, plus on s'attend à ce qu'il consomme du carburant afin de délivrer cette puissance. Vient ensuite le type de carburant, mais nous interpréterons cette variable catégorielle plus bas. Ensuite, la longueur influence aussi la consommation de carburant ; plus une voiture est longue, plus elle consomme de carburant. Une voiture plus longue implique un plus grand volume, et donc plus de poids, nécessitant plus d'énergie pour la faire avancer. De façon marginale, la largeur influence négativement la consommation de carburant. Notons que ce coefficient est difficile à interpréter, étant donné que dans le premier modèle, celui-ci était positif. Etant donné la significativité marginale, il est raisonnable de penser que nous nous trouvons dans une situation ou l'hypothèse H0 est rejetée à tort. Dans ce cas, il est plus prudent de ne pas interpréter ce coefficient. Wheelbase n'est pas significatif, mais étant corrélé à length, il est possible que la variance que celui-ci explique a été captée par length. Notons aussi l'intercept, qui représente la valeur de Y lorsque tous les prédicteurs sont à 0. Néanmoins, nous pensons qu'il est hasardeux d'interpréter un intercept, surtout lorsque la variable dépendante a été transformée.

5 Test des prédicteurs catégoriels

Nous avions codé, manuellement, des codes de contrastes pour les deux variables catégorielles. Ainsi, pour la variable fuel nous avons attribué la valeur "-1" si le carburant était l'essence, et "1" si le carburant était le diesel. Pour la variable aspiration, nous avons attribué la valeur "-1" si l'ar-

rivée d'air était standard, et "1" s'il y avait une aspiration turbo. Sur le tableau 7 on peut constater que seul la variable catégorielle correspondant au type de carburant est significative. Comme nous avons utilisé des codes de contrastes, il est facile d'interpréter les coefficients β estimés. Comme ce coefficient est négatif, cela indique que lorsque le carburant est diesel, le moteur consomme moins de carburant. Inversement pour l'essence. L'aspiration est à la limite de la significativité. Nous voyons qu'une aspiration turbo augmentera la consommation, et une aspiration standard la diminuera. En ajoutant un terme d'interaction au modèle (tableau non inclu pour des raison d'économie de pages) nous remarquons que le terme d'interaction n'est pas significatif (p=0.32). De plus, en ajoutant ce terme d'interaction, la variable aspiration n'est plus significative (p=0.24). La variable fuel reste significative (p¡0.000). Comme ce terme d'interaction n'est pas significatif, nous allons le retirer du modèle. Ceci est logique, en se référant au graphe 1 : les droites sont quasiment parallèles, indiquant, de manière graphique, qu'il n'y a pas d'interaction entre les deux variables.

6 Prédiction

On peut également tenter de calculer l'intervalle de prédiction sur une nouvelle valeur. La première étape consiste à choisir cette nouvelle valeur sur laquelle estimer l'intervalle. Pour ce faire, on peut demander à l'ordinateur de générer une nouvelle valeur avec une distribution de moyenne et variance égale à celle de chacune des variables. On peut alors ensuite appliquer notre modèle à cette valeur, et calculer l'intervalle de prédiction et l'intervale de confiance autour de cette valeur. On fixera la valeur de alpha à 0.95 pour le calcul des intervalles.

Si on calcule l'intervalle de prédiction sur une nouvelle donnée construite par cette méthode, on obtient une valeur prédite de 6.577 et un intervalle compris entre 4.86 et 8.29. Si les valeurs semblent à priori raisonnables (elles ne sont en effet pas très éloignées de la moyenne de la variable dépendante), il peut néanmoins être frustrant de constater la largeur de l'intervalle de prédiction. La borne supérieure est en effet près du double de la borne inférieure (1.7 fois celle-ci exactement). La largeur de cet intervalle est peut-être normal, mais il est intéressant de se poser la question de ce qui a pu aller de travers et provoquer cet intervalle qui semble excessivement large.

Si on considère que cet intervalle large n'est pas naturel et provient d'une erreur dans le modèle, on peut rechercher d'où provient ce résultat. On pourrait penser que la transformation, en modifiant les unité pourrait être à l'origine de cet écart démesuré. Néanmoins, la même technique appliquée aux unités d'origine donne un résultat similaire au résultat obtenu précédemment. Il est donc plus probable que l'intervalle, aussi large qu'il soit est tout simplement l'intervalle de prediction correct pour le modèle

sélectionné.

7 Intervalle de confiance de la variance de l'erreur

Pour cette question, la réponse n'est pas explicite. Néanmoins, en sachant que la somme des carrés de l'erreur peut être considéré comme la variance de l'erreur du modèle, il est possible de construire un intervalle de confiance autour de cette variance, en utilisant les formules classiques d'inférence :

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}\right]$$

Dans notre cas :

$$\left[\frac{198 \times 9.28}{\chi^2_{198;0.025}}, \frac{198 \times 9.28}{\chi^2_{198;0.975}}\right]$$

L'intervalle de confiance ainsi créé est : $[7.72, 11.48]$

8 Conclusion

Ce travail nous a permis d'explorer en profondeur les différentes possibilités permettant de raffiner un modèle linéaire. Nous avons montré comment sélectionner les variables un modèle, vérifier les hypothèses sous-jascentes à la régression, et comment détecter les valeurs aberrantes et influentes. Une limite importante à ce travail est probablement la présence d'une non indépendance des observation, étant donné que certaines voitures ont des caractéristiques en commun. Néanmoins, celle-ci a pu être contrôlée via l'ajout des constructeurs de voiture comme prédicteurs. De plus, une autre limite est la transformation de Box-Cox, qui rend difficile l'interprétation des coefficients. Nous aurions pu aussi utiliser la méthode des moindres carrés pondérés afin de contrôler l'hétéroscédasticité.

9 Annexes

A Vérification des VIF

	wheel-base	length	width	height	curb-weight	engi-nesize	bore	stroke	horse-power	pea-krpm
wheel-base	1.00	0.88	0.80	0.59	0.78	0.57	0.49	0.17	0.36	−0.35
length	0.88	1.00	0.84	0.50	0.88	0.69	0.61	0.12	0.56	−0.28
width	0.80	0.84	1.00	0.29	0.87	0.75	0.56	0.18	0.64	−0.22
height	0.59	0.50	0.29	1.00	0.30	0.03	0.18	−0.05	−0.11	−0.28
curb-weight	0.78	0.88	0.87	0.30	1.00	0.86	0.65	0.17	0.75	−0.27
engine-size	0.57	0.69	0.75	0.03	0.86	1.00	0.59	0.21	0.83	−0.21
bore	0.49	0.61	0.56	0.18	0.65	0.59	1.00	−0.07	0.58	−0.26
stroke	0.17	0.12	0.18	−0.05	0.17	0.21	−0.07	1.00	0.09	−0.07
horse-power	0.36	0.56	0.64	−0.11	0.75	0.83	0.58	0.09	1.00	0.13
pea-krpm	−0.35	−0.28	−0.22	−0.28	−0.27	−0.21	−0.26	−0.07	0.13	1.00

Table 9 – Corrélations entre les variables numériques : on remarque que curb-weigth, enginesize et horsepower sont fortement corrélées entre elles

B Sorties de la sélection du modèle

Number in Model	C(p)	R-Square	Variables in Model
6	7.8733	0.8152	wheelbase length width horsepower fueldummy aspirationdummy
7	8.2218	0.8168	wheelbase length width horsepower peakrpm fueldummy aspirationdummy
8	8.5198	0.8184	wheelbase length width bore horsepower peakrpm fueldummy aspirationdummy
5	9.0007	0.8122	wheelbase length horsepower fueldummy aspirationdummy
7	9.1320	0.8159	wheelbase length width bore horsepower fueldummy aspirationdummy

Table 10 – Sélection des 5 meilleurs modèles en fonction du critère de Mallows (C(p)).

Number in Model	Adjusted R-Square	R-Square	Variables in Model
8	0.8108	0.8184	wheelbase length width bore horsepower peakrpm fueldummy aspirationdummy
9	0.8108	0.8194	wheelbase length width bore stroke horsepower peakrpm fueldummy aspirationdummy
10	0.8103	0.8199	wheelbase length width height bore stroke horsepower peakrpm fueldummy aspirationdummy
7	0.8101	0.8168	wheelbase length width horsepower peakrpm fueldummy aspirationdummy
9	0.8100	0.8187	wheelbase length width height bore horsepower peakrpm fueldummy aspirationdummy

Table 11 – Sélection des 5 meilleurs modèles en fonction du critère du R^2

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Sq	Model R-Sq	C(p)	F Value	$Pr > F$
1	horsepower	1	0.6579	0.6579	162.071	378.80	<.0001
2	length	2	0.1155	0.7734	43.5431	99.87	<.0001
3	fueldummy	3	0.0234	0.7967	21.1259	22.45	<.0001
4	wheelbase	4	0.0105	0.8073	12.1210	10.62	0.0013
5	aspiration-dummy	5	0.0049	0.8122	9.0007	5.04	0.0259
6	width	6	0.0030	0.8152	7.8733	3.11	0.0792

Table 12 – Résumé de la méthode de Forward selection : 6 variables sont retenues au seuil $p < 0.1$.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Sq	Model R-Sq	C(p)	F Value	$Pr > F$
1	height	9	0.0005	0.8194	9.5281	0.53	0.4683
2	stroke	8	0.0010	0.8184	8.5198	0.99	0.3200
3	bore	7	0.0016	0.8168	8.2218	1.71	0.1930
4	peakrpm	6	0.0016	0.8152	7.8733	1.65	0.2006

Table 13 – Résumé de la méthode de Backward elimination : 4 variables sont supprimées au seuil $p > 0.15$

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Sq	Model R-Sq	C(p)	F Value	$Pr > F$
1	horsepower		1	0.6579	0.6579	162.07	378.80	<.0001
2	length		2	0.1155	0.7734	43.54	99.87	<.0001
3	fueldummy		3	0.0234	0.7967	21.12	22.45	<.0001
4	wheelbase		4	0.0105	0.8073	12.12	10.62	0.0013
5	aspirationdummy		5	0.0049	0.81	9.0007	5.04	0.0259
6	width		6	0.0030	0.8152	7.87	3.11	0.0792

Table 14 – Résumé de la méthode stepwise forward selection : 6 variables sont retenues au seuil de sélection $p < 0.1$, et aucune variable n'est retirée au seuil $p > 0.15$

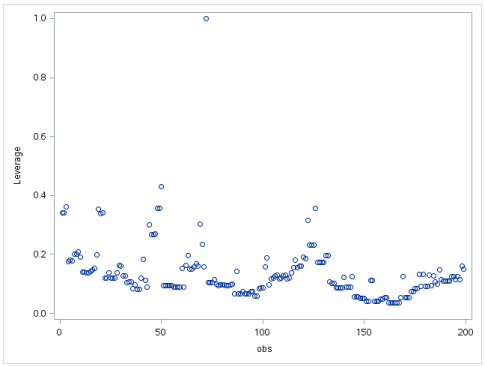
LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	BIC
0	Intercept		1	245.4286
1	horsepower		2	188.7603
2	length		3	117.5676
3	width		4	−22.4925
4	fueldummy		5	−22.8460
5	wheelbase		6	−64.1505
6	aspirationdummy		7	−71.8419*
7	peakrpm		8	−71.8178
8	stroke		9	−71.5047
9	bore		10	−70.8762
10	height		11	−71.4571

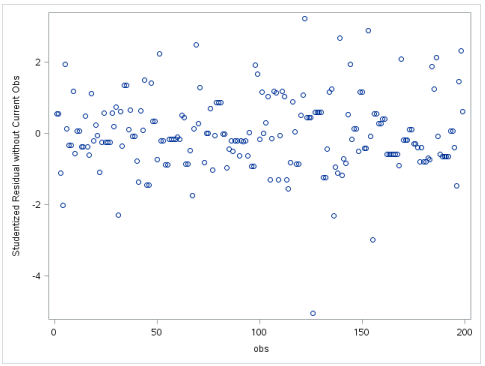
* Optimal Value of Criterion

Table 15 – Sélection lasso. Le step optimal est le 6^{e}.

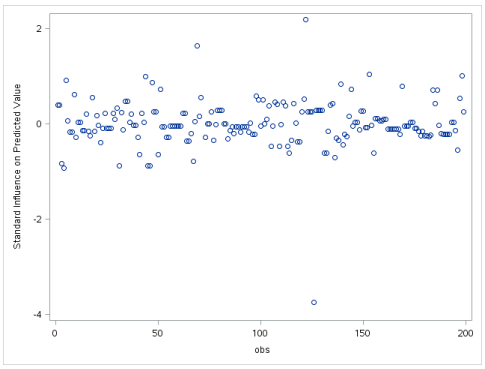
C Graphes des valeurs aberrantes et influentes



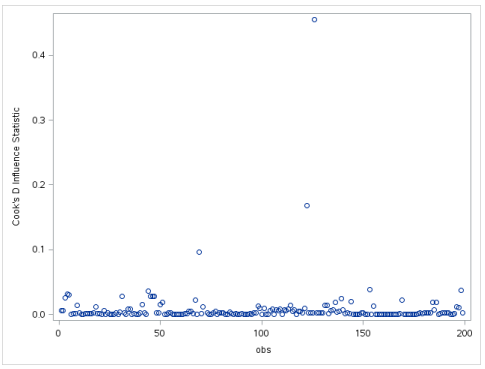
(a) Graphes des leviers par numéro d'observation



(b) Graphes des résidus effacés studentisés par numéro d'observation



(c) Graphes des DFITTS par numéro d'observation



(d) Graphes des distances de cooks par numéro d'observation

Figure 5 – Graphes des différentes mesures de valeurs aberrantes et influentes, par observation