Projet - Modèles linéaires

Axel Struys - Alexis Buckens

30 décembre 2016

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.

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. Il faudra donc adapter notre manière de calculer les effets principaux et les interactions pour ces deux variables discrètes. 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.

	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

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

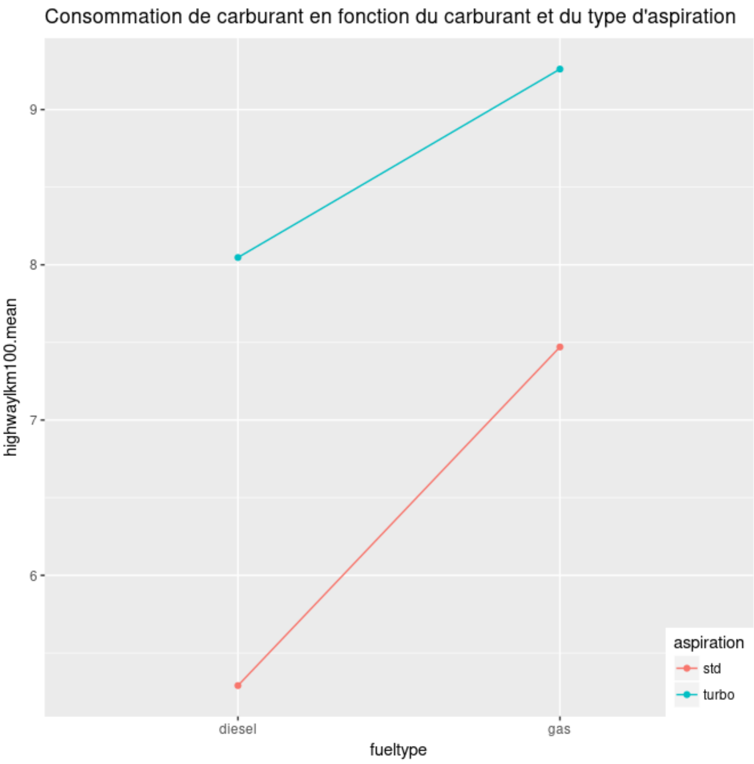

Figure 1 – Moyenne de la consommation (L/100km) en fonction du type de carburant et de l'aspiration

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 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 12 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.

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

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.

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 5 et 6 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

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 5 – 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 6 – 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 7 – 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 8 – Résumé de la méthode de Backward elimination : 4 variables sont supprimées au seuil $p > 0.15$

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 7, 8 et 9 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.

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

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	fuel-dummy		3	0.0234	0.7967	21.12	22.45	<.0001
4	wheel-base		4	0.0105	0.8073	12.12	10.62	0.0013
5	aspiration-dummy		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 9 – 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 10 – Sélection lasso. Le step optimal est le 6[e].

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 11 – Résumé du modèle final sélectionné

	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 12 – Corrélations entre les variables numériques : on remarque que curbweigth, enginesize et horsepower sont fortement corrélées entre elles

3 Vérification des hypothèses sous-jascentes au modèle linéaire

4 Annexes

A Vérification des VIF