TP II: Regression linéaire multiple

Nous utiliserons le langage `R` pour ce TP. Le corrigé sera à faire sous forme de `.Rmd`.

Les données `Cars.txt` décrivent les caractéristiques de 38 véhicules des Etats-Unis. L'objectif est de prédire la consommation des véhicules (`MPG`: plus le chiffre est élevé, moins la voiture consomme) à partir de leurs caractéristiques (poids, rapport de pont, puissance...). Descriptif des variables du jeu de données:

- `Country`: Pays du constructeur
- `Car`: Marque et modèle
- `MPG`: Miles par gallon, mesure du kilométrage gazeux
- `Weight`: Poids
- `DriveRatio`: Rapport de pont du véhicule
- `Horsepower`: Puissance
- `Displacement`: Déplacement du véhicule en pouces cubes
- `Cylinders`: Nombre de cylindrées du véhicule

Exercice 1. Etude exploratoire

1. Créer dans `R` un objet nommé `cars` à partir du fichier de données `Cars.txt`. En tirer un objet nommé `cars.quantitative.vars` ne contenant que les variables quantitatives de `cars`.
2. Faire une étude exploratoire des variables de `cars.quantitative.vars`. Calculer les corrélations entre les variables continues. Que peut-on en conclure?
3. Créer un sous-objet de `cars.quantitative.vars` nommé `carsntyp` dans lequel le véhicule `Buick Estate Wagon` est supprimé. Refaire la question précédente. Que se passe t-il? Pourquoi a-t-on enlevé ce véhicule?

Exercice 2. Influence d'un point atypique sur la modélisation

1. Estimer le modèle complet expliquant la variable `MPG` en fonction de toutes les autres variables de l'objet `cars.quantitative.vars`. Interpréter les résultats obtenus. Ce modèle vous semble-il satisfaisant? Quelles sont, à 5%, les variables significatives?
2. Estimer le modèle complet expliquant la variable `MPG` en fonction de toutes les autres variables de l'objet `carsntyp`. Interpréter les résultats obtenus. Ce modèle vous semble-il satisfaisant? Quelles sont, à 5%, les variables significatives?

Exercice 3. Sélection de variables

1. Voici une procédure de choix de modèle à la main par élimination. On partira du modèle complet:
   - Identifier la variable explicative pour laquelle le test de Student est le moins significatif (i.e. plus grande p-value).
   - La retirer du modèle et relancer l'estimation.

   On itérera ces 2 étapes jusqu'à ce que tous les coefficients soient significatifs à 5%. Attention! La variable constante est généralement conservée dans tous les modèles.
2. La procédure `step` permet également de chercher le meilleur modèle de régression. Tester les différentes options et comparer avec le modèle obtenu "à la main".
3. Ici, le nombre de variables explicatives n'étant pas trop important, on peut également faire une recherche exhaustive parmi tous les modèles possibles. Tester cela en choisissant différents critères de sélection.

Exercice 4. Prédiction A partir du modèle complet, puis du meilleur modèle obtenu selon le critère BIC, prédire le `MPG` des véhicules dont les caractéristiques sont les suivantes :

Country	Car	Weight	DriveRatio	Horsepower	Displacement	Cylinders	MPG
US	Pontiac	3.654	3.044	95	120	8	?
France	CitroenC3	2.99	3.101	102	192	5	?
Germany	AudiA3	3.22	2.885	65	136	8	?
Japan	ToyotaCorona	4.001	3.965	128	145	7	?