Statistique Avancée (Cours électif MA2822)
Centrale 1A et 2A, 2016-2017
christine.keribin@math.u-psud.fr

Devoir Maison

Le fichier `Centrale-DM.data` contient le prix de vente de voitures d'occasion ainsi que différentes variables les caractérisant, notamment: le prix en milliers d'euros (`price`), l'âge en mois (`age`), le kilométrage en milliers de kilomètres (`km`), le nombre de mois jusqu'à la révision suivante (`TIA`), la présence de l'ABS (oui, `extra1`=1), la présence de toit ouvrant (oui, `extra2`=1).

1. Importer les données dans R, en nommant les variables comme dans le fichier de données. Vérifiez que le data-frame obtenu contient 172 observations et 12 variables. Renommer la variable `extra1` en `ABS`.
2. Représenter un scatter plot des quatre premières variables, puis, sur une fenêtre en deux parties, les boxplots du prix en fonction de l'ABS puis en fonction de l'option de toit ouvrant. Mettre un titre pour chacun des boxplots. Commenter les graphiques obtenus.
3. Étude du prix en fonction de la présence ou non de l'ABS:
   (a) Y a-t-il une différence à considérer la variable ABS comme qualitative ou quantitative?
   (b) Interpréter les résultats de la régression simple du prix en fonction de l'ABS.
4. Étude du prix en fonction du kilométrage:
   (a) Ajuster une régression linéaire simple et interpréter.
   (b) Déterminer un intervalle de confiance du prix d'une voiture de 50 000 km, puis d'une voiture de 135 000 km. Justifier le fait que les intervalles sont de longueur différente.
   (c) Vérifier que la variable `kop1` du fichier de données est obtenue en centrant et réduisant la variable `km`. L'ajustement obtenu avec la variable centrée réduite est-il le même qu'avec la variable initiale? Compléter votre réponse par une justification théorique.
   (d) On considère la régression polynomiale de degré 3 sur la variable `km`: rappeler la définition du modèle. Est-ce toujours un modèle linéaire? L'ajuster dans R en l'écrivant sous la forme `price~km + I(km^2) + I(km^3)` (Modèle M3). Commenter les résultats.
   (e) La fonction `anova` appliquée au modèle M3 donne-t-elle les mêmes résultats qu'en l'appliquant au modèle M3b défini par `price~ I(km^3) + I(km^2)+ km`? Commenter.
   (f) La fonction `vif` du package `car` calcule le facteur d'inflation de variance. Rappeler sa définition et retrouver la sortie de cette fonction par calcul direct. Commenter.
   (g) Il est possible d'effectuer la régression polynomiale d'ordre 3 en utilisant les variables `kop1`, `kop2` et `kop3` qui sont orthogonales et engendrent le même espace que `km`, `km^2` et `km^3`.

- Vérifier que ces variables sont orthogonales, centrées, réduites.
- Effectuer la régression et commenter.
- Construire directement les variables `kop` à partir de `km`.

5. On prend maintenant en compte toutes les variables (avec les polynômes d'ordre 3 pour `km` et `age`).
   (a) Sélectionner les variables par des méthodes différentes
   (b) Valider le modèle retenu

Consignes:
Le devoir maison donne lieu à un compte-rendu *rédigé* à effectuer en binôme:

- Ne pas oublier de définir un titre, une introduction pour préciser la problématique étudiée et le plan du travail, et une conclusion.
- Commenter les résultats obtenus, inclure les graphiques pertinents dans le corps du texte.
- Les résultats doivent être justifiés. La notation prendra en compte la clarté et le soin de la rédaction.
- A remettre sous forme d'un `pdf` contenant le compte-rendu et d'un fichier texte `.R` contenant les commandes. Les fichiers seront nommés avec votre `NOM`, soit `NOM1-NOM2.pdf` et `NOM1-NOM2.R`.
- A envoyer pour le **2 mars** à mon adresse électronique