Master 1 Data Science — Année universitaire 2022-2023
Gilles Stupfler — gilles.stupfler@univ-angers.fr

Modèles de régression
Epreuve de contrôle continu (TP), durée 1h30

L'utilisation des notes de cours et des feuilles de TD est autorisée. Celle de tout autre document est interdite.

Vous répondrez à l'aide d'un fichier R Markdown dont le gabarit vous est fourni et que vous rendrez. Votre code sera commenté et vos réponses soigneusement justifiées.

Les routines standard pour l'estimation du modèle linéaire sont `lm`, `aov`, `anova` et `Anova`. Leur utilisation, dûment justifiée, est bien évidemment vivement conseillée. L'utilisation de toute autre routine pré-implémentée pour l'estimation, la construction d'intervalles de confiance... pourra être pénalisée, en particulier si l'interprétation qui en est faite ne correspond pas avec sa programmation.

Exercice 1 (Données `pollution.csv`)**.** Le jeu de données `pollution.csv` contient $n = 111$ observations liées à la qualité de l'air à New York (USA), collectées quotidiennement de mai à septembre 1973. Ces données contiennent quatre variables :

- `rad`, une mesure de la radiation solaire,
- `temp`, une mesure de la température exprimée en degrés Fahrenheit,
- `wind`, une mesure de la vitesse de vent exprimée en *miles* par heure,
- `ozone`, une mesure de la concentration atmosphérique en ozone exprimée en parties par milliard.

On s'intéresse au lien entre la variable à expliquer `ozone` et les trois autres variables. On commence par explorer le lien existant entre `ozone` et `wind`.

1. Expliquer pourquoi exprimer `ozone` comme un polynôme du second degré en `wind` semble approprié.
2. Estimer le modèle de régression exprimant `ozone` comme un polynôme du second degré en `wind`.

3. Construire un intervalle de confiance à 95% pour le terme quadratique en `wind`. Ce terme est-il significatif ?

4. Le modèle est-il significatif ?

On intègre maintenant la variable `temp`.

5. Estimer le modèle de régression exprimant `ozone` comme un polynôme du second degré en `wind` et en `temp`, ne comportant pas de terme d'interaction entre `wind` et `temp`.

6. Expliquer pourquoi, dans ce modèle, le nombre de degrés de liberté résiduels est égal à 106.

7. Le modèle est-il significatif ?

8. Que pensez-vous de l'hypothèse d'homoscédasticité des résidus dans ce contexte ? De l'hypothèse de normalité ?

9. Tester la significativité des deux termes quadratiques simultanément dans un cadre de modèles emboîtés. Comment décririez-vous le modèle alternatif ?

10. Commenter la pertinence d'intégrer la variable `rad` au modèle.

Exercice 2 (Données `agriculture.csv`)**.** Le jeu de données `agriculture.csv`, de taille $n = 96$, contient les quatre variables suivantes :

- `yield`, une mesure du rendement par hectare d'une certaine culture agricole,
- `density`, une variable catégorielle à deux modalités concernant la densité de la plantation,
- `fertilizer`, une variable catégorielle à trois modalités concernant le type d'engrais utilisé,
- `block`, une variable catégorielle à quatre modalités concernant le type de traitement (pesticide...) utilisé.

On souhaite comprendre le rendement moyen, et on va donc modéliser `yield`. On commence par différencier selon la variable `fertilizer`.

1. Ecrire en toutes lettres un modèle ANOVA à un facteur dans ce cadre. On précisera le nombre de groupes associé et les effectifs de chaque groupe, et l'objectif de l'utilisation du modèle ANOVA.

2. Tester la présence d'un effet dû à `fertilizer`.

On rajoute maintenant l'information portée par `density`.

3. Ecrire en toutes lettres un modèle ANOVA à deux facteurs dans ce cadre. On précisera le nombre de groupes associé et les effectifs de chaque groupe.

4. Les groupes sont-ils équilibrés ou déséquilibrés ?

5. En supposant l'absence d'un effet d'interaction, tester la présence d'un effet marginal dû à `fertilizer`, et la présence d'un effet marginal dû à `density`.

6. Tester la présence d'un effet d'interaction entre `fertilizer` et `density`.

7. Retrouver les résultats des deux questions précédentes en utilisant seulement la fonction `lm` sur des modèles emboîtés.

On souhaite enfin comprendre l'influence de `block`.

8. Dans ce but, on exécute la commande

   ```
   summary(aov(yield~fertilizer+density+block, data=agriculture))
   ```

   et on obtient la sortie suivante :

   ```
               Df Sum Sq Mean Sq F value   Pr(>F)
   fertilizer   1  5.743   5.743  17.265 7.27e-05 ***
   density      1  5.122   5.122  15.397 0.000168 ***
   block        1  0.486   0.486   1.461 0.229823
   Residuals   92 30.603   0.333
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   ```

   Dans l'exécution de cette commande, une grosse erreur a été commise. Laquelle ?

9. Tester la présence d'un effet dû à `block` sur `yield`.