TP6: variables qualitatives et vérification des hypothèses du modèle
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

Commandes utiles pour le TP6

Commande	Description de la commande
`lm()`	ajuste un modèle de régression linéaire
`summary()`	donne le résumé d'une variable ou d'un objet créé par `lm`
`plot()`	crée un graphique (linéaire ou en nuage de points)
`abline()`	trace une ligne droite sur un graphique
`points()`	ajoute des points à un graphique déjà réalisé
`lines()`	ajoute une ligne (pas nécessairement une droite) à un graphique déjà réalisé
`legend()`	ajoute une légende à un graphique déjà réalisé
`seq()`	génère un vecteur avec valeurs de `from` jusqu'au `to`, de longueur `length`
`min()`	donne la valeur minimale de son argument
`max()`	donne la valeur maximale de son argument
`predict()`	prédit une valeur à partir d'un objet créé par `lm`
`fitted()`	donne les valeurs ajustées $\widehat{y}_1, \ldots, \widehat{y}_n$ à partir d'un objet créé par `lm`
`resid()`	donne les résidus $e_1, \ldots, e_n$ à partir d'un objet créé par `lm`
`qqnorm()`	crée un graphe quantiles-quantiles (QQ-plot)
`qqline()`	ajoute une droite à un QQ-plot déjà réalisé
`subset()`	sélectionne une sous-collection de variables et/ou observations d'un tableau

Rappel : pour l'utilisation de la fonction `subset()` en plus de détail, voir l'énoncé du TP3. Pour les fonctions concernant les graphiques, voir l'énoncé du TP1.

Données "schools" : performances des élèves à l'école primaire

La base de données `schools.RDS` contient des données sur $n = 420$ écoles primaires en Californie pour l'année 1998/1999. Pour une description des variables reprises dans ce jeu de données, voir le TP4 ou le TP5. Comme au TP4, commencez par ajouter une colonne à ce tableau de données appelée `scores`, qui représente la moyenne des scores en mathématiques et lecture.

Exercices

1. Ajustez un modèle de régression linéaire simple pour prédire les performances des élèves en fonction du revenu moyen du quartier de l'école. Faites un nuage de points et ajoutez-y

la droite de régression estimée. S'agit-il d'un bon modèle ? Pourquoi (pas) ?

2. On considère les deux modèles suivants, où $Y =$ `scores` et $x =$ `income` :

$$\mu_1(x) = \beta_0 + \beta_1\sqrt{x} \tag{1}$$
$$\mu_2(x) = \beta_0 + \beta_1 \log x \tag{2}$$

   Ajustez ces deux modèles, et ajoutez les deux courbes de régression au nuage de points de la question précédente. Lequel des deux modèles préfériez-vous ?

3. Supposons que l'on garde le deuxième modèle ; comment est-ce que les performances des élèves augmentent quand le revenu moyen du quartier augmente de
   (a) 10 à 11 mille dollar ?
   (b) 20 à 21 mille dollar ?
   (c) 20 à 22 mille dollar ?

   On remarque que (a) et (b) représentent une augmentation de mille dollars, tandis que (a) et (c) représentent une augmentation de 10%. Pourquoi obtient-on la même réponse dans (a) et (c) ?

4. Faites un nuage de points entre les valeurs ajustées du modèle et les résidus. Faites également un QQ-plot des résidus. Les hypothèses sur les résidus semblent-elles vérifiées ?

Données "hawks" : faucons de Iowa City

La base de données `hawks` contient des données sur $n = 907$ faucons observés autour de Iowa City. On s'intéresse aux variables suivantes :

— `Feather` : la longueur (en mm) de la plume principale de l'aile.
— `Wing` : la largeur (en mm) de l'aile.
— `Species` : la variété du faucon, red-tailed (`RT`), sharp-shinned (`SS`), ou Cooper's hawks (`CH`).

Des mesures sur la longueur de la plume principale sont plus faciles à obtenir que des mesures sur la largeur de l'aile, d'où l'intérêt de prédire `Wing` en fonction de `Feather`.

Exercices

1. Ajustez une droite de régression pour prédire `Wing` en fonction de `Feather` et écrivez l'équation de la droite estimée. S'agit il d'un bon modèle ?
2. Faites un nuage de points entre `Feather` et `Wing`. Ajoutez la droite de régression au nuage de points. S'agit il d'un bon modèle ?
3. En utilisant différentes couleurs, mettez les différentes variétés de faucons en évidence sur un nuage de points. Que voyez-vous ?
4. Ajustez un modèle de régression multiple pour prédire `Wing`. Interprétez soigneusement les coefficients estimés, écrivez la droite estimée pour chaque variété, et ajoutez ces droites au nuage de points. S'agit-il d'un meilleur modèle que le modèle de régression simple ? Voyez-vous encore un moyen d'améliorer ce modèle ?