TP5: régression linéaire multiple
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

Commandes utiles pour le TP5

Commande	Description de la commande
`subset()`	sélectionne une sous-collection de variables et/ou observations d'un tableau
`cor()`	donne la/les corrélation(s) d'une collection des variables quantitatives
`lm()`	ajuste un modèle de régression linéaire
`summary()`	donne le résumé d'une variable ou d'un objet créé par `lm`
`predict()`	prédit une valeur à partir d'un objet créé par `lm`

Rappel : pour l'utilisation de la fonction `subset()` en plus de détail, voir l'énoncé du TP3.

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

Récupérez le fichier `schools.RDS`, disponible sur webcampus, et enregistrez-le sur votre ordinateur dans un dossier "TP5". Créez un nouveau script `R` et enregistrez-le dans le même dossier. Ouvrez le fichier sur Rstudio en utilisant la commande suivante :

```
data <- readRDS("schools.RDS")
```

La base de données `schools.RDS` contient des données sur $n = 420$ écoles primaires en Californie pour l'année 1998/1999. Les variables reprises dans ce jeu de données sont :

- `school` : le nom de l'école.
- `county` : la commune dans laquelle se trouve l'école.
- `type` : le type d'école, avec 6 (`6year`) ou 8 (`8year`) années consécutives.
- `students` : le nombre d'élèves inscrits dans l'école.
- `teachers` : le nombre de professeurs qui enseignent dans l'école.
- `lunch` : le pourcentage d'élèves qui ont droit aux repas à prix réduits.
- `computer` : nombre d'ordinateurs à l'école.
- `expenditure` : le montant que l'école dépense par élève.
- `income` : le revenu moyen du quartier de l'école (en milliers de dollars).
- `read` : performance moyenne des élèves dans la lecture (en points).
- `math` : performance moyenne des élèves dans les mathématiques (en points).

Explication supplémentaire de `type` : les écoles avec 8 années consécutives regroupent l'école primaire et les premières deux années de l'école secondaire (le "middle school" en anglais), tandis que les écoles avec 6 années consécutives sont des écoles primaires traditionnelles.

Le but de ce TP est de trouver un modèle de régression multiple simple (c'est-à-dire, avec le moins de variables explicatives) mais avec un bon pouvoir prédictif des performances des élèves (`scores`). Les exercices ci-dessous vous guident vers un modèle particulier, mais n'oubliez pas qu'il n'y a jamais un seul "bon" modèle ; le choix de variables explicatives à garder peut dépendre d'un statisticien à l'autre !

Exercices

1. Comme au TP4, commencez par ajouter des colonnes à ce tableau de données : la première, appelée `scores`, représente la moyenne des scores en mathématiques et lecture ; la deuxième, `studprof`, représente le rapport entre élèves (`students`) et professeurs (`teachers`) ; la troisième, `compstud`, représente le nombre d'ordinateurs par élève. Ensuite, vous pouvez supprimer les colonnes suivantes (que l'on n'utilisera plus dans la suite) de ce tableau de données : `school`, `students`, `teachers`, `computer`, `read`, et `math`.
2. Calculez la corrélation entre les régresseurs quantitatifs et interprétez les corrélations relativement élevées (≥ 0.5 ou ≤ -0.5). Quels régresseurs semblent pertinents à inclure dans une régression multiple si le but est de prédire la variable `scores` ? Est-ce que vous pensez que le modèle va souffrir de multicolinéarité ?
3. Ajustez un modèle de régression multiple, appelé `m1`, pour prédire `scores` en fonction de `type`, `lunch`, `expenditure`, `income`, `studprof` et `compstud`. Est-ce que ce modèle a un bon pouvoir prédictif ? Est-ce qu'au moins une des variables explicatives est significative ?
4. Intuitivement, avoir un plus grand nombre d'ordinateurs par élève à l'école devrait avoir un impact positif sur les performances des élèves. Formalisez cette intuition en écrivant les hypothèses du test correspondant, donnez la p-valeur et interprétez-là. Est-ce que la variable `compstud` est significative (utilisez un niveau de $\alpha = 0.01$) ?
5. On considère d'enlever les variables `studprof` et/ou `expenditure` du modèle `m1` s'ils n'améliorent pas le pouvoir prédictif du modèle. Quelle est votre conclusion ?
6. Ajustez un modèle de régression multiple, appelé `m2`, pour prédire les performances des élèves en fonction de `type`, `lunch, income,` et `compstud`. Écrivez l'équation de la surface de régression estimée et interprétez soigneusement les coefficients estimés.
7. On considère d'ajouter la variable explicative `county` au modèle `m2`. Est-ce une bonne idée ? Pourquoi (pas) ?
8. Prédisez les scores de deux élèves d'une école à 8 années (`type = 8year`) avec les caractéristiques suivantes :
   (a) 50% des élèves ont droit aux repas à prix réduits, le revenu moyen est de 10 mille dollars, et il y a 0.2 ordinateurs par élève.
   (b) 80% des élèves ont droit aux repas à prix réduits, le revenu moyen est de 40 mille dollar, et il y a 0.3 ordinateurs par élève.

   S'agit-il de bonnes prédictions ? *Indice : regarder des nuages de points entre les variables* `lunch, income, compstud`.