TP4: régression linéaire simple
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

Commandes utiles pour le TP4

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`
`qt()`	calcule un quantile de la distribution student t
`pt()`	calcule la fonction de répartition de la distribution student t
`confint()`	permet de calculer des intervalles de confiance pour les coefficients estimées d'une régression
`mean()`	donne la moyenne d'une variable quantitative
`cor()`	donne la corrélation entre deux variables quantitatives
`sum()`	effectue la somme d'un vecteur d'éléments
`fitted()`	donne les valeurs ajustées $\widehat{y}_1, \ldots, \widehat{y}_n$ à partir d'un objet crée par `lm`
`subset()`	sélectionne une sous-collection de variables et/ou observations d'un tableau
`plot()`	crée un graphique (linéaire ou en nuage de points)
`abline()`	trace une ligne sur un graphique
`points()`	ajoute des points à un graphique déjà réalisé
`legend()`	ajoute une légende à un graphique déjà réalisé

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 "TP4". 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
— `math` : performance moyenne des élèves dans les mathématiques

Explication supplémentaire de la variable `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.

Dans ce TP, on se concentrera surtout sur les performances moyennes des élèves dans deux domaines (mathématiques et lecture) et sur les rapports entre élèves et professeur. On va également vérifier quelle variable explicative serait la meilleure pour prédire les performances moyennes des élèves. Le TP suivant, nous allons utiliser le même jeu de données pour faire une analyse plus approfondie qui prend en compte toutes les variables.

Exercices

1. On voudrait savoir si on peut prédire les performances des élèves en fonction du rapport élèves/professeurs. Commencez par ajouter deux colonnes à ce tableau de données : la première, appelée `score`, représente la moyenne des scores en mathématiques et lecture ; la deuxième, `studprof`, représente le rapport entre élèves (`students`) et professeurs (`teachers`).
2. Ajustez ensuite le modèle de régression linéaire et donnez la droite de régression estimée. Interprétez soigneusement les coefficients estimés $\widehat{\beta}_0$, $\widehat{\beta}_1$.
3. Donnez un intervalle de confiance de niveau $\alpha = 0.05$ pour β_1 et interprétez-le. Vérifiez votre réponse avec la fonction `confint`. Utilisez cette fonction pour obtenir un intervalle de confiance de niveau $\alpha = 0.0001$. Est-ce qu'il est probable que $\beta_1 = 0$?
4. Écrivez la formule avec laquelle on calcule la p-valeur d'un test d'hypothèse bilatéral pour vérifier si le rapport entre élèves et professeurs influence les scores moyens des élèves. Calculez la p-valeur associée et vérifiez votre réponse en regardant le résumé du modèle linéaire dans `R`. Interprétez la p-valeur.
5. Est-ce qu'un test unilatéral a plus de sens qu'un test bilatéral dans ce contexte ? Donnez l'hypothèse alternative de ce test unilatéral. Déduisez la p-valeur du test unilatéral de la p-valeur du test bilatéral.
6. Calculez le coefficient de détermination de deux manières différentes : en utilisant le fait qu'il est égal au coefficient de corrélation linéaire au carré, et en utilisant la formule SSE/SST. Vérifiez votre réponse en regardant le résumé du modèle linéaire ajusté précédemment et interprétez ce coefficient.
7. Faites un nuage de points et ajoutez-y la droite de régression estimée. Le rapport entre élèves et professeurs est-il un bon prédicteur de la performance des élèves si on regarde le nuage de points et le coefficient de détermination ? Est-ce en contradiction avec votre conclusion aux questions 3 à 5 ?

8. Est-ce que le revenu moyen du quartier de l'école est un bon prédicteur des performances moyennes des élèves ? Peut-on dire que c'est un meilleur prédicteur que le rapport entre élèves et professeurs ?
9. Parmi les variables `lunch`, `expenditure`, `income`, et `studprof`, laquelle est le meilleur prédicteur des performances moyennes des élèves ?
10. Faites un nuage de points entre la variable sélectionnée à la question 9 et les performances moyennes des élèves. Colorez en rouge les points correspondants aux écoles avec un revenu du quartier inférieur ou égal à la moyenne et en bleu les points correspondants aux écoles avec un revenu du quartier supérieur à la moyenne (et ajoutez une légende au graphique). Que voyez-vous ? Ajustez encore deux modèles de régression, un pour les points rouges, et l'autre pour les points bleues, et ajoutez les deux droites de régression estimées au graphique.