TP5: régression linéaire multiple
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

NB : toutes les réponses ont été arrondis à deux chiffres après la virgule.

Réponses

1. Voir fichier `R`.
2. Interprétation des corrélations :
   - `expenditure` et `studprof` sont négativement corrélées ($r = -0.62$) : un grand nombre d'élèves par professeur est associé à peu de dépenses par élève. En effet, il est moins cher d'avoir des grandes classes.
   - `lunch` et `income` sont négativement corrélées ($r = -0.68$) : un petit nombre d'élèves qualifiant pour un repas à prix réduit est associé à un revenu élevé dans le quartier de l'école. En effet, pour qualifier pour un repas à prix réduit, il faut que les parents aient un revenu peu élevé.

   La variable à expliquer, `scores`, est fort corrélée avec `lunch` ($r = -0.87$) et `income` ($r = 0.71$). C'est-à-dire, les écoles dans lesquelles beaucoup d'élèves ont droit aux repas à prix réduits ont des moins bonnes performances qu'une école moyenne, tandis que les écoles se situant dans des quartiers avec des revenus élevés ont de meilleurs performances qu'une école moyenne. Les variables `lunch` et `income` semblent donc pertinentes à inclure dans la régression multiple ; or, étant donné qu'elles sont fort corrélées entre elles, le modèle risque de souffrir de multicolinéarité.
3. Ce modèle a un bon pouvoir prédictif — en effet, son r^2 indique que 79.42% de la variabilité des performances des élèves peut être expliquée par ces $p = 6$ régresseurs. L'hypothèse nulle du test de Fisher,

   $$H_0 : \beta_1 = \ldots = \beta_6 = 0$$

   est rejetée avec une p-valeur qui est ≈ 0, donc on est virtuellement certain qu'au moins un des régresseurs est significatif.
4. On s'intéresse au test d'hypothèse $H_0 : \beta_{\texttt{compstud}} = 0$ contre $H_1 : \beta_{\texttt{compstud}} > 0$. La p-valeur de ce test est donnée par $0.00802/2 \approx 0.004$, c'est-à-dire, sous l'hypothèse nulle, la probabilité d'observer un coefficient de régression supérieur ou égal à $\widehat{\beta}_{\texttt{studprof}} = 18.8$ est de 0.004. Alternativement, on peut dire que l'on est confiant à 99.6% que le nombre d'ordinateurs par élève est un bon prédicteur des performances des élèves. Le régresseur `compstud` est donc significatif pour un niveau $\alpha = 0.01$ car la p-valeur du test est inférieure à α.

5. Voir fichier `R`. Enlever les variables `studprof` et `expenditure` ne diminue quasiment pas le pouvoir prédictif du modèle (comme mesuré par le r^2) et on décide d'enlever les deux variables.
6. Le modèle ajusté est

   ```
   scores = 666.86 - 2.61 type - 0.50 lunch + 0.54 income + 25.14 compstud
   ```

   où `type = 1` pour une école à 8 années et `type = 0` pour une école à 6 années. On se rappelle que `scores`, la variable à expliquer, représente la performance moyenne des élèves en math et en lecture, exprimée en points. On interprète les coefficients "ceteris paribus" (c'est-à-dire, on suppose que l'on compare deux écoles qui ont exactement les mêmes valeurs pour toutes les autres variables) :
   - On n'interprète pas l'intercept de 666.86 car cela n'a pas de sens de poser

     `type = lunch = income = compstud = 0`
   - Un élève d'une école à 6 années peut s'attendre à 2.61 points supplémentaires par rapport à un élève d'une école à 8 années.
   - Une augmentation de 1% du nombre d'élèves qualifiant pour un repas à prix réduit est associée, en moyenne, avec une diminution des performances de 0.5 points.
   - Une augmentation du revenu moyen du quartier de l'école de 1000 dollars est associée, en moyenne, avec 0.54 points supplémentaires.
   - Une augmentation d'un ordinateur supplémentaire par élève est associée, en moyenne, avec une augmentation des performances de 25.14 points.
7. La variable `county` (la commune dans laquelle se trouve l'école) peut prendre 45 valeurs différentes. En l'incluant dans la régression, on ajoute donc 44 coefficients de régression supplémentaires à estimer, ce qui semble excessif. En regardant le résumé d'un tel modèle, on voit effectivement que la grande majorité de ces coefficients ne sont pas significatifs. Or, le coefficient de détermination (r^2) s'améliore un peu. Si on connaissait l'emplacement géographique de ces communes, on pourrait essayer de les diviser en quelques régions (par exemple, "sud", "ouest", "nord" et "est") et inclure ces régions dans la régression. Ceci nous permettrait d'ajouter des informations concernant la localisation de l'école sans pour autant faire exploser le nombre de coefficients à estimer.
8. On pose `type = 1` et
   (a) `lunch = 50, income = 10, compstud = 0.2`. On obtient

   $$\texttt{scores} \approx 666.86 - 2.61 \times 1 - 0.50 \times 50 + 0.54 \times 10 + 25.14 \times 0.2 \approx 649.89$$

   (b) `lunch = 80, income = 40, compstud = 0.3`. On obtient

   $$\texttt{scores} \approx 666.86 - 2.61 \times 1 - 0.50 \times 80 + 0.54 \times 40 + 25.14 \times 0.3 \approx 653.74$$

   C'est-à-dire, un élève d'une école à 8 années avec les caractéristiques décrites en (a) peut s'attendre à un score de 649.89, tandis qu'un élève d'une école à 8 années avec les caractéristiques décrites en (b) peut s'attendre à un score de 653.74. Les nuages de points entre les régresseurs (voir fichier `R`) montrent que l'on ne devrait pas faire une prédiction avec les valeurs de (b). En effet, les écoles dans les quartiers si riches (`income = 40`) dans lesquelles 80% des élèves ont droit aux repas à prix réduits n'éxistent pas. Il s'agit donc d'une extrapolation irréaliste.