TP4: régression linéaire simple
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. On trouve
$$\texttt{scores} = 698.93 - 2.28 \times \texttt{studprof}$$
On ne peut pas interpréter $\widehat{\beta}_0 = 698.93$ car le rapport entre élèves et professeurs ne sera jamais proche de zéro. L'estimation $\widehat{\beta}_1$ peut être interprétée de la façon suivante : en moyenne, par élève supplémentaire par professeur, le score diminue de 2.28 points. C'est-à-dire, si on compare deux écoles, dont la première a x élèves par professeur et la deuxième $(x+1)$ élèves par professeur, les performances moyennes des élèves de la première école seront de 2.28 points plus élevées que les performances moyennes des élèves de la deuxième école.
3. En regardant le résumé de la régression, on trouve $\widehat{\sigma}_{\widehat{\beta}_1} = 0.4798$. On calcule l'intervalle de confiance par
$$\begin{aligned}(l, u) &= \left(\widehat{\beta}_1 - \widehat{\sigma}_{\widehat{\beta}_1} t_{0.975}, \widehat{\beta}_1 + \widehat{\sigma}_{\widehat{\beta}_1} t_{0.975}\right) \\ &= (-2.2798 - 0.4798 \times 1.9657,\ -2.2798 + 0.4798 \times 1.9657) \\ &= (-3.22, -1.34)\end{aligned}$$
On est "confiant à 95 %" que l'intervalle $(-3.22, -1.34)$ contient la vraie valeur β_1. Pour un élève supplémentaire par professeur, la performance moyenne des élèves diminuera de 1.34 à 3.22 points. On peut également calculer que l'on est "confiant à 99.99 %" que l'intervalle $(-4.16, -0.39)$ contient la vraie valeur β_1. La valeur $\beta_1 = 0$ n'étant toujours pas dans l'intervalle, même pour un niveau de confiance très élevé, il est très peu probable que β_1 soit zéro. Autrement dit, il est très peu probable que le rapport entre élèves et professeurs n'influence pas les performances moyennes des élèves.
4. Soit $H_0 : \beta_1 = 0$ et $H_1 : \beta_1 \neq 0$. La p-valeur est la probabilité d'obtenir la même valeur de la statistique du test ou une valeur plus extrême (c'est-à-dire plus éloignée de l'hypothèse nulle), sous l'hypothèse nulle. Ici, on a observé $\widehat{\beta}_1 = -2.28$. Étant donné que l'on s'intéresse à une test bilatéral, il faut regarder des deux côtés de la distribution ; la p-valeur est donc la probabilité d'avoir observé un coefficient inférieur à -2.28 ou un coefficient supérieur à 2.28. On trouve
$$\mathbb{P}[\widehat{\beta}_1 \leq -2.28 \text{ ou } \widehat{\beta}_1 \geq 2.28 \mid H_0] = 2\,\mathbb{P}[\widehat{\beta}_1 \leq -2.28 \mid H_0]$$

$$= 2\,\mathbb{P}\left[T \leq \frac{-2.2798}{0.4798}\right]$$
$$= 2\,\mathbb{P}\left[T \leq -4.7516\right] = 0.00000278$$

On peut rejeter l'hypothèse nulle $H_0 : \beta_1 = 110$ avec un "niveau de confiance" de maximum 99.99972%. Interprétation : on est "confiant à 99.99972%" que le rapport entre élèves et professeurs est un bon prédicteur de la performance moyenne des élèves.

5. Oui, un test unilatéral a plus de sens, car nous avons bien trouvé une association négative entre le rapport élèves / professeurs et les performances des élèves. L'hypothèse alternative serait donc $H_1 : \beta_1 < 0$. La p-valeur est deux fois plus petite que la p-valeur du test bilatéral ; on trouve

$$0.5 \times 0.00000278 = 0.00000139$$

Interprétation : on est "confiant à 99.99986%" que le rapport entre élèves et professeurs est un bon prédicteur de la performance moyenne des élèves et que cette performance moyenne s'améliore quand le rapport entre élèves et professeurs diminue.

6. Si on note X = rapport entre élèves et professeurs et Y = performances moyennes des élèves, on trouve une corrélation de $r_{xy} = -0.2264$ et donc un coefficient de détermination de $r_{xy}^2 = 0.0512$. On pourrait également calculer

$$\text{SSE} = 7794.109, \qquad \text{SSR} = 144315.5, \qquad \text{SST} = 152109.6$$

et SSE / SST = 1− SSR / SST = 0.0512. Interprétation : 5.12 % de la variabilité des performances moyennes des élèves peut être expliquée par le rapport entre élèves et professeurs de l'école.

7. On voit que la tendance linéaire est assez faible et qu'il reste une grande variabilité autour de la droite de régression estimée. Ceci est confirmé par le coefficient de détermination, qui est également faible. Or, nous avons vu grâce aux tests d'hypothèse précédents que le rapport entre élèves et professeurs est un bon prédicteur des performances. A première vue, cela peut sembler contradictoire ; mais il est tout à fait possible que ce rapport ait un impact sur les performances sans pour autant pouvoir expliquer une grande partie de la variabilité de ces performances. En effet, on voit sur le nuage de points que la variabilité des performances est grande (c'est-à-dire, le SST est grand par rapport au SSE). Pour améliorer le modèle, on devrait chercher d'autres facteurs qui peuvent expliquer une partie de cette variabilité. Dans le TP suivant, nous allons regarder comment combiner plusieurs prédicteurs pour expliquer un maximum de variabilité de la performance des élèves.

8. Si on ajuste un modèle de régression simple avec le revenu moyen du quartier comme variable explicative et les performances des élèves comme variable à expliquer, on trouve un $\widehat{\beta}_1$ hautement significatif (c'est-à-dire, l'hypothèse nulle $H_0 : \beta_1 = 0$ est rejeté avec quasi certitude) et un coefficient de détermination égal à 0.5076. On pourrait donc dire que le revenu moyen du quartier est un meilleur prédicteur des performances moyennes des élèves que le rapport entre élèves et professeurs, car il explique 50.76% de la variabilité des performances moyennes des élèves.

9. En comparant les coefficients de détermination, on voit que le pourcentage d'élèves qui ont droit aux repas à prix réduits est le meilleur prédicteur des performances moyennes des élèves, expliquant 75.48% de sa variabilité.

10. On voit que les points rouges, représentant des quartiers avec revenus faibles, sont associés à un grand nombre d'élèves ayant droit aux repas à prix réduits, tandis que les points bleues, représentant des quartiers avec revenus élevés, sont associés à un petit nombre d'élèves ayant droit aux repas à prix réduits, ce qui est intuitif. Les performances moyennes des élèves sont meilleures pour les écoles des quartiers avec revenus élevés que pour les écoles des quartiers avec revenus faibles. En ajoutant les deux droites de régression, on voit que, pour les écoles des quartiers avec revenus élevés, les performances moyennes diminuent plus vite quand le pourcentage d'élèves qui ont droit au repas à prix réduits augmente que pour les écoles des quartiers avec revenus faibles.