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

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

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

Réponses (performances des élèves)

1. Le modèle est significatif, c'est-à-dire, `income` aide à prédire `scores` et explique 50.76% de la variabilité de ce dernier. Or, le nuage de points montre une relation non-linéaire entre `income` et `scores` ; on voit une forte augmentation des performances pour une augmentation du revenu de 5 à 25 mille dollars mais une faible augmentation des performances pour une augmentation du revenu de 25 à 45 mille dollars.
2. Le deuxième modèle, avec $\log x$, explique un peu plus de variabilité de Y que le premier modèle (56.25% contre 54.72%). En plus, la courbe de régression estimée du deuxième modèle à l'air de mieux suivre le nuage de points que la courbe de régression estimée du premier modèle.
3. Voir fichier `R` ; les performances des élèves augmentent respectivement de (a) 3.47 points, (b) 1.78 points et (c) 3.47 points. Une augmentation du revenu moyen de mille dollars ne mène pas à une augmentation constante des performances des élèves car le modèle n'est pas linéaire ; c'est-à-dire,

$$\mu_2(x+1) - \mu_2(x) = \beta_1 \left[\log(x+1) - \log(x)\right]$$

dépend de la valeur de x. En effet, on a $\widehat{\beta}_1 = 36.42$, et donc

$$36.42 \times [\log(11) - \log(10)] = 3.47$$
$$36.42 \times [\log(21) - \log(20)] = 1.78$$

Or, une augmentation de 10% du revenu moyen implique une augmentation constante des performances des élèves, car

$$\begin{aligned}\mu_2(1.1x) - \mu_2(x) &= \beta_1 \left[\log(1.1x) - \log(x)\right] \\ &= \beta_1 \left[\log(1.1) + \log(x) - \log(x)\right] = \beta_1 \log(1.1)\end{aligned}$$

ne dépend pas de la valeur de x. En effet, les performances augmentent de $36.42 \times \log(1.1) = 3.47$ points quand le revenu moyen augmente de 10% (et avec $36.42 \times \log(1.2) = 6.64$ points quand le revenu moyen augmente de 20%, etc).
4. Voir fichier `R` ; les deux graphiques montrent que les résidus sont centrés autour de zéro, que leur variance est approximativement constante, et que les quantiles des résidus ressemblent fort à ceux d'une distribution normale. L'hypothèse $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ semble donc raisonnable.

Réponses (faucons de Iowa City)

1. La droite de régression estimée est

   ```
   Wing = 88.97 + 0.348 Feather
   ```

   La variable `Feather` est significative et 81.01% de la variabilité de la largeur de l'aile est expliquée par la longueur de sa plume principale. A priori, on dirait que le modèle est bon, mais il faut évidemment investiguer le nuage de points entre les deux variables avant de conclure.

2. Voir fichier `R`. On voit que le nuage de points est très hétérogène, c'est-à-dire, on a l'impression de voir deux ou trois nuages de points avec des comportements différents. La droite de régression estimée capte bien la tendance générale mais on a l'impression de pouvoir mieux faire.

3. Voir fichier `R`. On pourrait avoir un meilleur modèle en divisant les faucons en différentes variétés. Nous allons donc ajuster un modèle de régression multiple avec `Feather` et `Species` comme variables explicatives.

4. `RT` et `SS` sont deux variables indicatrices (égales à 1 si le faucon est de cette variété et 0 si non). Le modèle estimé est

   ```
   Wing = 142.36  - 12.21 RT - 40.02 SS + 0.24 Feather
   ```

   On peut interpréter les coefficients estimés :

   — On n'interprète pas l'intercept car `Feather = 0` n'a pas de sens.

   — En moyenne, si la plume principale de l'aile est de 1 mm plus longue, alors l'aile sera de 0.24 mm plus large, peu importe la variété de faucon.

   — En moyenne, pour une même longueur de la plume principale de l'aile, un faucon "red-tailed" aura une aile qui est 12.21 mm moins large que celle d'un faucon "Cooper's hawks".

   — En moyenne, pour une même longueur de la plume principale de l'aile, un faucon "sharp-shinned" aura une aile qui est 40.02 mm moins large que celle d'un faucon "Cooper's hawks".

   On peut également écrire trois droites de régression, une pour chaque variété,

   $$\begin{aligned} \texttt{Wing} &= 130.15 + 0.24\ \texttt{Feather}, && \text{pour les faucons RT} \\ \texttt{Wing} &= 102.33 + 0.24\ \texttt{Feather}, && \text{pour les faucons SS} \\ \texttt{Wing} &= 142.36 + 0.24\ \texttt{Feather}, && \text{pour les faucons CH} \end{aligned}$$

   Le nuage de points et le coefficient de détermination suggère que ce modèle est supérieur au modèle de régression linéaire simple. Le graphique montre que l'on pourrait encore améliorer le modèle en permettant des pentes différentes pour les trois droites.