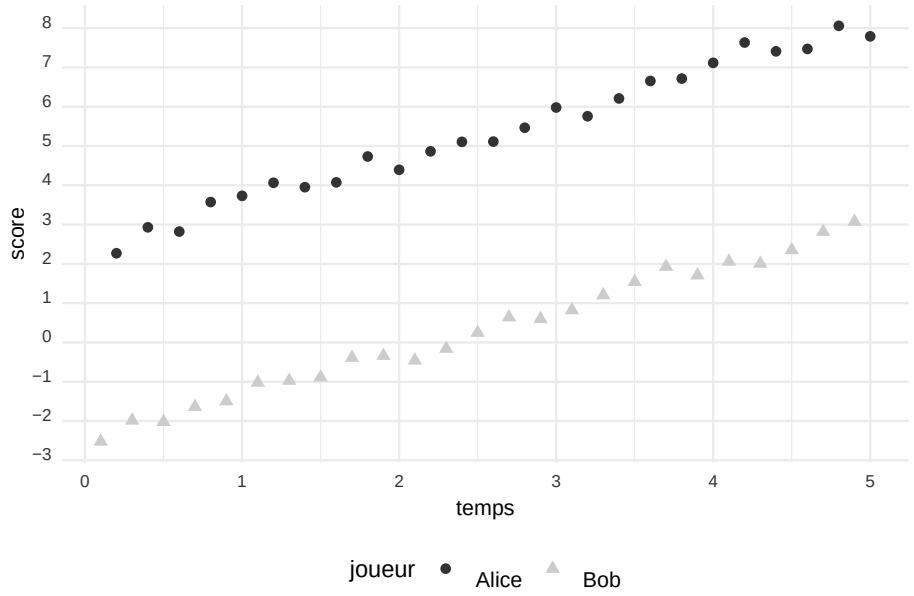

Figure 1: Scores d'Alice et de Bob en fonction du nombre d'heures de jeu.

2.1 Bob et Alice remarquent que leurs scores à un jeu de société peuvent être modélisés en fonction du nombre d'heures de jeu à l'aide du modèle linéaire

$$\texttt{score}_i = \beta_0 + \beta_1 \texttt{temps}_i + \beta_2 \texttt{joueur}_i + \beta_3 \texttt{temps}_i \texttt{joueur}_i + \varepsilon_i,$$

où ε_i est un terme d'erreur de moyenne nulle et $\texttt{joueur}_i$ est une variable binaire égale à un si le ie score est celui d'Alice et zéro pour celui de Bob.
En vous basant sur la Figure 1, que peut-on dire quant aux signes des coefficients $\widehat{\beta}_1, \ldots, \widehat{\beta}_3$?

Solution
On peut déduire l'ordonnée à l'origine et la pente à partir de la Figure 1 et reparamétriser le modèle. L'équation de la pente pour Alice est $2.5 + 1.1\texttt{temps}$ et celle de Bob est $-2.5 + 1.1\texttt{temps}$. Le paramètre $\widehat{\beta}_0$ correspond à l'ordonnée à l'origine de la catégorie de référence, -2.5, et la pente $\widehat{\beta}_1$ est celle de la personne de référence, 1.1. Les autres paramètres sont la différence moyenne entre l'ordonnée à l'origine/la pente du score d'Alice moins ceux de Bob, à savoir $(\widehat{\beta}_2 = 5, \widehat{\beta}_3 = 0)$. Il suffit maintenant de considérer le signe des coefficients et on déduit que $\widehat{\beta}_0 < 0, \widehat{\beta}_1 > 0, \widehat{\beta}_2 > 0$ et $\widehat{\beta}_3 = 0$.

2.2 On considère un modèle de régression pour expliquer l'impact de l'éducation et du nombre d'enfants sur le salaire des femmes, à savoir:

$$\texttt{logsalaire}_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \beta_2 \mathrm{X}_{i2} + \varepsilon_i,$$

Compilé le 18/09/2020 à 10:17

où

$$X_1 = \begin{cases} 0, & \text{si la femme n'a pas de diplôme secondaire,} \\ 1, & \text{si la femme a un diplôme secondaire mais pas de diplôme collégial,} \\ -1, & \text{si la femme a un diplôme collégial.} \end{cases}$$

$$X_2 = \begin{cases} 0, & \text{si la femme n'a pas d'enfant,} \\ 1, & \text{si la femme a 1 ou 2 enfants,} \\ -1, & \text{si la femme a 3 enfants ou plus.} \end{cases}$$

Selon ce modèle, quelle serait la **différence** moyenne en log-salaire entre (i) une femme qui possède un diplôme collégial et qui a trois enfants et (ii) la moyenne de toutes les femmes dans l'échantillon, en supposant que la taille de chacun des neuf groupes est la même (plan équilibré)?

Solution
On modélise la moyenne de chaque groupe (analyse de variance à deux facteurs, sans interaction). Puisqu'on a le même nombre de femmes dans chaque catégorie, la moyenne globale est la somme de chacune des catégories, soit $\widehat{\beta}_0$. L'équation de la moyenne ajustée pour la catégorie de référence en (i) est $\widehat{\beta}_0 - \widehat{\beta}_1 - \widehat{\beta}_2$ et la différence est donc $-\widehat{\beta}_1 - \widehat{\beta}_2$.

2.3 On considère le log du prix de vente de maisons en fonction de leur localisation (urbain ou rural), de la surface du garage (en pieds carrés), et d'un indicateur dénotant la présence ou l'absence de garage. Le modèle linéaire postulé est le suivant

$$\texttt{logprix} = \beta_0 + \beta_1 \texttt{garage} + \beta_2 \texttt{surface} + \beta_3 \mathbf{1}_{\texttt{loc=urbain}} + \varepsilon,$$

où ε est un terme d'erreur de moyenne nulle et `garage` une variable indicatrice

$$\texttt{garage} = \begin{cases} 0, & \text{si la maison a un garage (}\texttt{surface} > 0\text{);} \\ 1, & \text{si la maison n'a pas de garage (}\texttt{surface} = 0\text{).} \end{cases}$$

On suppose que le modèle a été ajusté par la méthode des moindres carrés et qu'on obtient $\widehat{\beta}_1 > 0$ et $\widehat{\beta}_2 > 0$. Laquelle des affirmations suivantes est **toujours** correcte?

(a) Toutes choses étant égales par ailleurs, les maisons avec garage sont en moyenne plus chères que celles sans garage.
(b) Toutes choses étant égales par ailleurs, les maisons sans garage sont toujours moins chères que celles avec garage.
(c) Toutes choses étant égales par ailleurs, les maisons avec garage sont en moyenne moins chères que celles sans garage.
(d) La localisation (urbain versus rural) est négativement corrélée avec la surface du garage.
(e) Aucune de ces réponses.

Solution
Le toutes choses étant égales par ailleurs ne veut rien dire parce qu'on ne peut fixer garage sans affecter superficie! Si la variable `garage` = 0 en l'absence de garage et que $\beta_1 < 0$, $\beta_2 > 0$, on ne peut rien conclure.

2.4 On considère un modèle de régression simple pour le prix d'une voiture électrique en fonction de son autonomie

Compilé le 18/09/2020 à 10:17

(distance); le modèle est

$$\texttt{prix}^{\texttt{USD}} = \beta_0^i + \beta_1^i \texttt{distance}^{\texttt{mi.}} + \varepsilon^i,$$

où ε est un terme d'erreur de moyenne nulle. Vos amis ont collecté des données où la variable prix est mesurée en dollars américains (`USD`) et la variable distance est mesurée en miles (`mi.`), et ont ajusté le modèle de régression afin d'obtenir les estimés $(\widehat{\beta_0^i}, \widehat{\beta_1^i})$.
Vous aimeriez connaître les estimés du modèle où la variable prix est exprimée en dollars canadiens (`CAD`) et la variable distance est exprimée en kilomètres (`km`), c'est-à-dire

$$\texttt{prix}^{\texttt{CAD}} = \beta_0^m + \beta_1^m \texttt{distance}^{\texttt{km}} + \varepsilon^m.$$

Sachant que 1 USD vaut 1.39 CAD et que 1 mile égale 1.61 km, quelle est la valeur de C dans l'équation $\widehat{\beta_1^i} = C\widehat{\beta_1^m}$?

Solution
En substituant les nouvelles variables dans l'équation, on obtient

$$\texttt{prix}^{\texttt{USD}} = 1.39\texttt{prix}^{\texttt{CAD}} \beta_0^i + \beta_1^i 1.61 \texttt{distance}^{\texttt{km}} + \varepsilon^i.$$

On divise chaque côté de l'équation par 1.39 pour obtenir l'équation en fonction de $\widehat{\beta_1^m}$ et on déduit $C = 0.621 = 1.61/1.39$.

Compilé le 18/09/2020 à 10:17