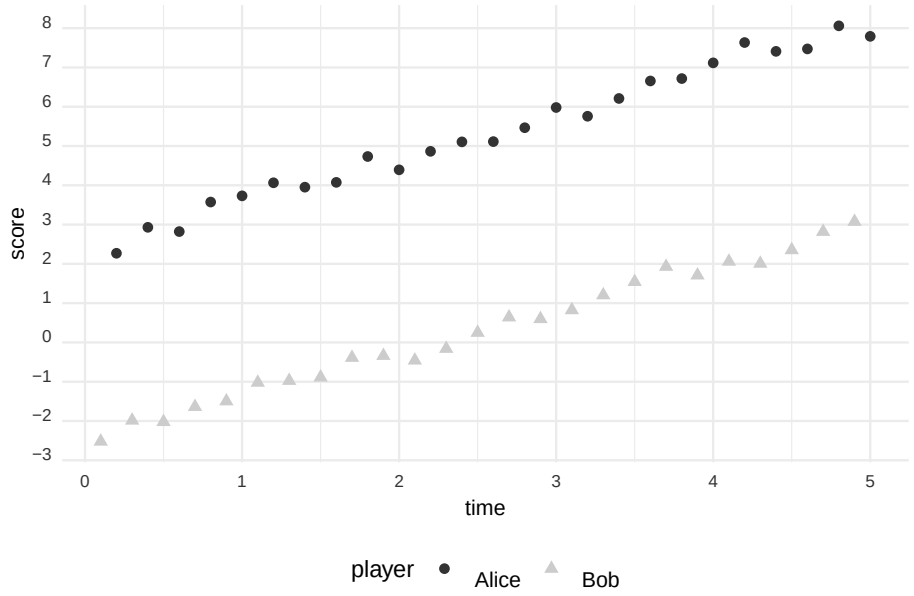


Figure 1: Scores of Alice and Bob as a function of the number of hours of play.

2.1 Bob and Alice decide to play board games during the Pandemic. They notice their scores (as a function of the number of hours they play) could be modelled using a linear model of the form

$$\texttt{score}_i = \beta_0 + \beta_1 \texttt{time}_i + \beta_2 \texttt{player}_i + \beta_3 \texttt{time}_i \texttt{player}_i + \varepsilon_i,$$

where $\varepsilon_i$ is a mean-zero error term and $\texttt{player}_i$ is a binary indicator equal to 1 if the $i$th score belongs to Alice and 0 if it belongs to Bob.
In view of Figure 1, what can we say about the sign of $\widehat{\beta}_1, \ldots, \widehat{\beta}_3$?

**Solution**
It suffices to check the respective intercept and slope and reparametrize the model. The equation of the slope for Alice is $2.5 + 1.1\texttt{time}$ and that of Bob is $-2.5 + 1.1\texttt{time}$. The parameter $\widehat{\beta}_0$ corresponds to the intercept of the baseline, namely -2.5 and the slope $\widehat{\beta}_1$ to the slope of the baseline, 1.1. The other parameters are mean difference between the intercept/slope of Alice minus that of Bob, meaning ($\widehat{\beta}_2 = 5, \widehat{\beta}_3 = 0$). It remains to consider the sign of the coefficients.

We consider a regression model to explain the impact of education and the number of children on the salary of women, viz.

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

where

$$X_1 = \begin{cases} 0, & \text{if the woman did not complete high school,} \\ 1, & \text{if the woman completed high school, but not college,} \\ -1, & \text{if the woman completed college.} \end{cases}$$

$$X_2 = \begin{cases} 0, & \text{if the woman has no children,} \\ 1, & \text{if the woman has 1 or 2 children,} \\ -1, & \text{if the woman has 3 or more children.} \end{cases}$$

According to the model, what would be the mean **difference** in log-salary between (i) a woman who completed college and has three children and (ii) the average log-salary of all women in the sample, assuming the sub-sample size in each of the nine group is the same (balanced design)?

**Solution**

2.2 We model a different mean for each of the nine categories (two-way ANOVA additive model. Since we have the same number of women in each category (balanced design), the overall mean is the sum of each fitted value, namely $\widehat{\beta}_0$. The equation for the fitted mean of the reference in (i) is $\widehat{\beta}_0 - \widehat{\beta}_1 - \widehat{\beta}_2$ and thus the difference is $-\widehat{\beta}_1 - \widehat{\beta}_2$.

2.3 We consider log of housing price as a function of location (urban or rural), whether or not the house includes a garage and the surface of the latter (in square feet). The postulated linear model is

$$\texttt{logprice} = \beta_0 + \beta_1 \texttt{garage} + \beta_2 \texttt{area} + \beta_3 \mathbf{1}_{\texttt{loc=urban}} + \varepsilon,$$

where $\varepsilon$ is a mean-zero error term and `garage` is an indicator variable,

$$\texttt{garage} = \begin{cases} 0, & \text{if the house has a garage } (\texttt{area} > 0); \\ 1, & \text{if the house doesn't have a garage } (\texttt{area} = 0). \end{cases}$$

Suppose we fit the model via least squares and find $\widehat{\beta}_1 > 0$ and $\widehat{\beta}_2 > 0$. Which of the following statement is **always** correct?

(a) Everything else being equal, houses with garages are on average more expensive than ones without a garage.
(b) Everything else being equal, houses with garages are always less expensive than ones without a garage.
(c) Everything else being equal, houses with garages are on average cheaper than ones without a garage.
(d) Location (urban versus rural) is negatively correlated with garage surface.
(e) None of the above

**Solution**

None of the above. Everything else being constant is meaningless: you cannot fix the dummy for garage without impacting area! If `garage` = 0 when there is no garage and $\beta_1 < 0$, $\beta_2 > 0$, we cannot conclude.

2.4 We consider a simple linear regression model for the price of an electric car as a function of its autonomy (distance); the model is

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

where $\varepsilon$ is a zero-mean error term. Your friends collects some data in which the price is expressed in American dollars (`USD`) and distance is measured in miles (`mi.`) and run the regression to get estimates $(\widehat{\beta_0^i}, \widehat{\beta_1^i})$.
You would like to know the estimates for the model with the price expressed in Canadian dollars (`CAD`) and distance

expressed in kilometers (km), i.e.,

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

Knowing that 1 USD is 1.39 CAD and that 1 mile is 1.61km, what is the value of $C$ in the equation $\widehat{\beta_1^i} = C\widehat{\beta_1^m}$?

**Solution**
Substituting the metric/Canadian measures in the first equation, we get

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

Divide throughout by 1.39 to obtain the equation in terms of $\widehat{\beta_1^m}$; we deduce $C = 0.621 = 1.61/1.39$.



Compiled 18/09/2020 at 10:15