Project — Part 3

Data:

The raw data consist of Bixi usage records from the 2023 season, wherein each observation consists of an individual trip and includes details on the start time and place, as well as the end time and place. Note that, as in the previous parts of the project, only trips under 2 hours in the months extending from May to October, inclusively, are considered. In this third part of the project we will explore the **average trip duration** at the **station level** from the 2023 season. The modified dataset includes the following variables:

`station`	departure station name
`arrondissement`	borough where the station is located
`mm`	departure month, ranging from 5 (May) to 10 (October)
`dd`	departure date, ranging from 1 to 31
`wday`	weekday, taking on values `Sunday` through `Saturday`
`dur`	average trip duration (in minutes)

Mandate:

The goal of this third part of the project is to explore the factors which affect BIXI trip durations. Throughout, be sure that your analyses allow you to answer the specific questions in an appropriate and adequate manner. Whenever a statistical model is used, be sure to

- report estimated coefficients, along with uncertainty measures,
- provide appropriate parameter interpretations, as pertaining to the question,
- provide relevant conclusions that reflect the context, as pertaining to the question,
- discuss the validity of the analysis carried out,
- discuss any shortcomings or limitations of the analysis carried out.

Any time a statistical test is carried out, be sure to clearly state the underlying hypotheses in terms of the model parameters, provide the value of the test statistic and resulting p-value, and provide a relevant conclusion which reflects the context.

Business questions:

The goal here is to explore whether the average trip duration varies across boroughs, and whether the average trip duration differs on weekends in comparison to weekdays. To this end, consider a linear regression model with `dur` as the response variable, and as covariates include the borough variable (`arrondissement`) and a binary variable indicating whether the observed response corresponds to a weekend day (i.e., Saturday or Sunday). Consider only the main effects of these two covariates.

0. Before beginning, carry out a **brief** exploratory data analysis by answering the following questions:
   a) Verify whether there are any missing values in the data.
   b) How many distinct stations are there in the data?
   c) How many observations are there in the data?
   d) Provide descriptive statistics for the number of observations at the different stations.
1. **Model 1**: Fit a linear regression model assuming independent observations.
   a) Interpret the **intercept** and the regression parameter for the **weekend** covariate in the fitted model.
   b) What happens if a fixed effect for `station` is added to the model? Explain.
2. **Model 2:** Fit a linear regression model with a **random intercept** at the **station level** (assuming independent random errors).
   a) Based on the fitted model, what is the estimated within-station correlation?
   b) Is the assumption of independence (as in **Model 1**) reasonable here? Consider an appropriate statistical test, using $\alpha = 1\%$.
   c) Based on the fitted model, does the average trip duration vary significantly across the boroughs? Consider an appropriate statistical test, using $\alpha = 1\%$.

3. **Model 3:** Fit a linear regression model with a **random intercept** and a **random effect** for the weekend variable, both at the **station level**, assuming **independent random effects** (and assuming independent random errors).

   a) Based on the station-level predictions, which station has the greatest increase in average trip duration on weekends? Explain your answer.

   b) Based on the fitted model, is there significant variation in the effect of weekend on the average trip duration across the stations? Consider an appropriate statistical test, using $\alpha = 1\%$.

   c) Based on the fitted model, what is the correlation between two weekday observations leaving from the same station? What about two weekend observations leaving from the same station? What about two weekday observations leaving from different stations?