BIOST 509: In-Class Exercise 8

Instructor: Amy Willis, Biostatistics, UW

Due date: 6:30pm on November 22, 2019 via Canvas

Instructions

Submit your answers to the below questions in a R Script (.R), Word (.doc or .docx) or pdf file to Canvas. Provide the code that you used to get the results, the output of the code as comments, and your answers to the questions as comments.

Optional but encouraged: Use R Markdown to create the pdf file with your results.

The style of this week's homework is choose-your-own-adventure!

- If you are looking for more challenge, there is one question with no subparts. You will need to figure out the necessary steps by yourself.
- If you are looking for less challenge, the question is broken down into subparts that guide you through the steps.

Dataset

All questions for this homework are related to an airquality dataset with daily air quality measurements from New York between May and September 1973. The following variables are included:

- `Ozone`: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
- `Solar.R`: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
- `Wind`: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
- `Temp`: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.
- `Month`: Number of month
- `Day`: Number of day during month

The `airquality` is available from Canvas in `Pages/Module 8 materials` or `Files/datasets`.

Questions: unguided version

Create a tibble with 3 columns: 'month', 'slopes', 'slope_ses'. The 'slope_estimate' column will contain the fitted slope for a linear regression model with ozone as the response and solar radiation as the predictor **based on data from only one 'month'**. The 'slope_se' will contain the standard error of the slope estimate. Also, make a single plot with a subplot for each month showing solar radiation on the x-axis and ozone on the y-axis for that month.

Questions: guided version

1. How many months are there in the `airquality` dataset? Store this number in the variable `num_months`.
2. Fit a linear regression model with ozone as the response (dependent) variable and solar radiation as the predictor (independent) variable. What is the slope coefficient and standard error of the slope? Store these numbers in the variables `est_full` and `se_full`

3. Create two empty vectors of length `num_months`, called `slopes` and `slope_ses`
4. Write a loop that loops through the months in `airquality` and calculates the slope coefficient and standard error of the slope for a linear regression model with ozone as the response and solar radiation as the predictor *using only the data in a single month*. Store these values in the vectors `slopes` and `slope_ses`.
5. Create a tibble or data frame with three columns:
    - 'month': This column should contain the months in the `airquality` dataset.
    - `slopes`: You created this column in Q4
    - `slope_ses`: You created this column in Q4

There should be one row per month, and therefore the number of rows in the dataset should be the same as your answer to Question 1.

5. *Optional:* Make a plot *for each month* showing solar radiation on the x-axis and ozone on the y-axis. You should experiment with `facet_wrap` and `facet_grid` (shown briefly in lecture) and see which you prefer!