Assignment 02 - Data Viz, Wrangling and Advanced Regression

Due Date

Thursday February 12, 2022 by midnight.

For this assignment, we will be analyzing data from the Southern California Children's Health Study. The learning objectives are to conduct data wrangling and visualize the data with key questions in mind.

Data Wrangling

You will need to download two CHS datasets: the individual and regional. The individual data includes personal and health characteristics of children in 12 communities across Southern California. The regional data include air quality measurements at the community level. A data dictionary is available here Once downloaded, you can merge these datasets using townname. Once combined, you will need to do the following:

1. After merging the data, make sure you don't have any duplicates by counting the number of rows. Make sure it matches.

   In the case of missing values, impute data using the average within the variables "male" and "hispanic." If you are interested (and feel adventurous) in the theme of Data Imputation, take a look at this paper on "Multiple Imputation" using the Amelia R package here.

2. Create a new categorical variable named "obesity_level" using the BMI measurement (underweight BMI<14; normal BMI 14-22; overweight BMI 22-24; obese BMI>24). To make sure the variable is rightly coded, create a summary table that contains the minimum BMI, maximum BMI, and the total number of observations per category.

3. Create another categorical variable named "smoke_gas_exposure" that summarizes "Second Hand Smoke" and "Gas Stove." The variable should have four categories in total.

4. Create four summary tables showing the average (or proportion, if binary) and sd of "Forced expiratory volume in 1 second (ml)" and asthma indicator by town, sex, obesity level, and "smoke_gas_exposure."

Looking at the Data

The primary questions of interest are: 1. What is the association between BMI and FEV (forced expiratory volume)? 2. What is the association between smoke and gas exposure and FEV? 3. What is the association between PM2.5 exposure and FEV?

Follow the EDA checklist from week 3 and the previous assignment. Be sure to focus on the key variables.

Visualization Create the following figures and interpret them. Be sure to include easily understandable axes, titles, and legends.

1. Facet plot showing scatterplots with regression lines of BMI vs FEV by "townname".
2. Stacked histograms of FEV by BMI category and FEV by smoke/gas exposure. Use different color schemes than the ggplot default.
3. Barchart of BMI by smoke/gas exposure.
4. Statistical summary graphs of FEV by BMI and FEV by smoke/gas exposure category.
5. A leaflet map showing the concentrations of PM2.5 mass in each of the CHS communities.
6. Choose a visualization to examine whether PM2.5 mass is associated with FEV.

Advanced Regression

Construct a multiple linear regression model to examine the association between weight and FEV, adjusted for age, sex, and race. First use weight as a linear predictor variable, and then fit another model where you put a cubic regression spline on weight. Provide summaries of your models, plots of the linear and non-linear associations, and interpretation of the linear and non-linear associations.