Data Analysis Assignment 1

IDS 702

DUE 11:55 PM Sept 16, 2022

Instructions

This assignment involves linear regression. The data can be found on Sakai: go to Resources → Data Analysis Assignment Datasets → Assignment 1. Please type your solutions using R Markdown. The final output file should be ".pdf". Submissions should be made on gradescope: go to Assignments → Data Analysis Assignment 1.

DO NOT INCLUDE R CODE OR OUTPUT IN YOUR SOLUTIONS/REPORTS *All R code must be included in an appendix, and R outputs should be converted to nicely formatted tables. Feel free to use R packages such as* `kable`*,* `xtable`*,* `stargazer`*, etc.*

Also, you can round up ALL numbers/estimates to 2 decimal places (4 decimal places at the most to avoid exact zeros when possible).

Reminder: You are allowed and even encouraged to talk to each other about general concepts, or to the instructor/TAs. However, the write-ups, solutions, and code MUST be entirely your own work.

Questions

Question 1 below is taken and adapted from Chapter 7 of Ramsey, F.L. and Schafer, D.W. (2013), "The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed).".

Side Note: We will use textbook datasets on some of the data analysis assignments. This is intentional as a way to start with clean and small datasets. Projects will focus a bit more on "messy" datasets.

1. RESPIRATORY RATES FOR CHILDREN. A high respiratory rate is a potential diagnostic indicator of respiratory infection in children. To judge whether a respiratory rate is truly "high," however, a physician must have a clear picture of the distribution of "normal" respiratory rates. To this end, Italian researchers measured the respiratory rates of 618 children that are at most 3 years old.
   The data for this question can be found in the file "Respiratory.csv" on Sakai.
   a. Do exploratory analysis on the data and include a useful plot that a physician could use to assess a "normal" range of respiratory rates for children of any age between 0 and 3.
   b. Write down a regression model for predicting respiratory rates from age. Make sure to use the right mathematical notation.
   c. Fit the model to the data. Include a table showing the output from the regression model including the estimated intercept, slope, residual standard error, and proportion of variation explained by the model.
   d. Interpret your results. In the context of the problem, what do you conclude? Your interpretation should mention an appropriate p-value, 95% confidence interval, and R^2 value.
   e. Is there enough evidence that the model assumptions are reasonable for this data? Include appropriate plots in your answer.

2. AIRBNB LISTINGS FOR SEATTLE, WA. AirBnB is a rental online marketplace. The company itself is based in San Francisco CA, and there are millions of listings in cities across the world. In this problem, you will only focus on data for AirBnB listings in Queen Anne, Seattle, WA. Specifically, you will try to understand how certain factors influence the price of a listing. The data we will use here is a very small subset of the overall available data. For more on the data, or if you are interested in using AirBnB data, see http://insideairbnb.com/get-the-data.html.
The data for this question can be found in the file "Listings_QueenAnne.txt" on Sakai.

   a. Analyze the data using `host_is_superhost`,`host_identity_verified`, `room_type`, `accommodates`, `bathrooms` and `bedrooms` as predictors. You should start by doing EDA, then model fitting, and model assessment. You should consider transformations if needed.
   b. Include the output from the final regression model that you used, as well as evidence that the model fits the assumptions reasonably well. Your regression output should includes a table with coefficients and SEs, p-values, and confidence intervals.
   c. Interpret the results of your fitted model in the context of the data.
   d. Are there any (potential) outliers, leverage points or influential points? Provide evidence to support your response. Also, if there are influential points and/or outliers, exclude the points, fit your model without them, and report the changes in your overall conclusions.
   e. Overall, are there any potential limitations of this analysis? If yes, what are two potential limitations?

Code Book

Variable	Description
id	Unique identifier for listings
host_is_superhost	Whether or not the host is a "superhost," meaning they satisfy AirBnB's criteria for high-quality listings, high response rate, and reliability
host_identity_verified	Whether or not the host has verified their identity with AirBnB
room_type	Entire home/apt, private room, or shared room
accommodates	Number of people that the listing can accommodate
bathrooms	Number of bathrooms in the listing
bedrooms	Number of bedrooms in the listing
price	Price of the listing for one night *(Use this as the response variable)*

Grading

40 points: 20 points for each question