w203_Lab3_feedback_to_Group1

Joanna Wang, Douglas Xu

November 27, 2018

1. Introduction the introduction is very clear with background information + research question + identified parameters. However, maybe the research question should be asked in a more open format? So that you leave enough room for other parameters that you may find strongly correlated in the analysis

2. The initial data loading and cleaning: the data cleaning is very well executed on the variables of key interest. Also, anomalies are identified, but you are being very careful about removing anything that you do not believe is error.
   one thing that might be worth thinking about is that, should identifying key variables come before data cleaning or after. It might make more logical sense to make decision on the key explanatory variables after cleaning the dataset initially

3. Model building process: very good univariate analysis on each explanatory variable and outcome variable. **a few suggestions**:
   - in the EDA, analyze several more variables
   - more explanation on why are those explanatory variables chosen over the rest of dataset
   - use potential transformations to variables that are not close to normal distribution

4. Regression model: nice detailed reporting on model coefficients, and plotting of standard errors **a few suggestions**: - more detailed discussion of the coefficients from each model, and what conclusions can be drawn from the model
   - more discussion on how the coefficients are complying with the 6 CLM assumptions
   - more detailed discussion between each model, especially in how the third model helps in verifying the validity of previous two models
   - compare the model coefficient in a table

5. Omitted Variable Discussion: according to the question and answer in piazza, omitted variable seems to be variable that is outside of the dataset. Therefore, there should be some discussion on some important variable that are important for the model, but not present in the dataset

Several small things

1. the units in $avgsen$ is actually days
2. maybe add a few more comments after each graph to show how it helps you in analyzing the data and the model
3. have a comparison table between the three different models