Course Project: Part 2

Due: **March 16, 2018, 11:59 PM**, via Blackboard

This part of the mini-project is all about inference and understanding relationships in your data.

Important note:

- Keep your report as concise as you can.

1 Inference on the outcome variables

(a) Estimate the mean of your continuous outcome variable. Estimate the mean of your binary outcome variable (i.e., the fraction of rows for which this variable is 1).

(b) For both your continuous outcome variable and your binary outcome variable, construct a confidence interval around the mean. Explain any assumptions you made in constructing these intervals.

(c) Do you believe your confidence intervals are valid? Why or why not?

(d) For your continuous outcome variable, test for normality in two ways: (1) QQ plots, and (2) the K-S test. What can you conclude?

(e) Choose a reasonable binary treatment variable. We want to study its effect on the continuous outcome variable. What is the mean outcome when the binary treatment variable is 0 and 1, respectively? Use the Student's t-test to give a 95% confidence interval for the difference in these mean outcomes. Choose a test statistic and use a permutation test to report a p-value for the hypothesis that this binary treatment variable has a non-zero effect on the outcome. Explain how you chose the test statistic. What can you conclude based on the obtained p-value?

Note: If you do not have a suitable choice of a binary treatment variable in your dataset, create your own from a continuous treatment variable (for example, 1 if your continuous treatment variable exceeds a fixed threshold, and 0 otherwise).

(f) Can you interpret your answer to part (e) causally? Why or why not?

2 Investigating relationships

(a) For each treatment variable, run a simple regression against your outcome variables and report p-values for the slopes. What do you observe? Compare your results to question 2(g) from mini-project: part 1. Do higher estimated correlations always correspond to smaller p-values?

(b) Fit a linear regression model with all treatment variables against your continuous outcome variable. Fit a logistic regression model with all treatment variables against your binary outcome variable.

(c) For your linear regression model, look at which coefficients are significant in the output according to `StatsModels`. Describe what statistical significance means for these coefficients. Do you believe the results? Why or why not? Compare your results to part (a). Are there variables that now appear more or less relevant?

(d) Comment on potential problems with your analysis on multiple hypothesis testing. Suggest how applying the Bonferroni correction would change your interpretation of significant coefficients.

(e) For the relationships that you found significant in part (c), would you be willing to interpret them as causal relationships? If not, why not? What other covariates do you think might be confounding your ability to infer causal relationships? Are there variables you would like to remove from the regression, e.g., post-treatment variables? Do your best in including or excluding variables to arrive at the most causally sound conclusions. Explain your thought process – what did you do and why?

(f) Repeat parts (c) and (e) for your logistic regression model.