Final Project Proposal - Data Science 608 Spring, 2022

Eric Hirsch

**Investigating and Discovering Interaction Terms in Selected Data Sets**

Data visualization is an effective tool for displaying the results of an analysis. However, I would like to explore using it to discover insights into datasets prior to analysis. In particular, my project will develop a visualization framework for investigating, discovering, and evaluating interaction terms.

1. Background

As part of the data exploration process, a data analyst needs to consider the possibility that observations aren’t independent. There may be groupings or connections between observations that aren’t automatically evident or obvious. One such example of this are interactions – where the slope of an independent variable against the dependent variable may vary depending upon another variable.

R and python provide numerous ways to model and display interactions, but the sheer number of possibilities makes an open-ended exploration a daunting task. An interactive framework which can provide researchers insight into where interactions might be occurring, and a quick visual look at how interactions manifest throughout the database, could be very beneficial to this process.

1. Details

This exploration tool will facilitate the exploration of interactions through three steps:

* 1. Display histograms and box plots, as well as scatterplots against the dependent variable, for all variables in the data set. (If the dependent variable is binary, the algorithm will choose boxplots instead of scatterplots.) This will allow the researcher to use distributions or loess fits in scatterplots to develop hunches about good candidates for interaction terms.
  2. Show two-slope scatterplots for interactions between any chosen independent variable against all of the other variables on the dependent variable.
  3. Allow the researcher to investigate more closely chosen interactions, including getting more details about the interaction such as its effect on adjusted r-squared.

1. Technologies

I plan to use r with shiny. Inputs will include a chosen data set (I will include at least two), a dropdown for the step of the analysis (i.e. one of the three steps described above), a drop down to choose the dependent variable, and a slider to determine how many and which independent variables to show. If time permits, I would also like to add a way to split a continuous variable into a dummy variable based on a chosen split-value so that this new dummy variable could also be tested for interactions.

1. Datasets:

I will use at least these datasets. They come from my 621 class so I am familiar with them and know they have some good interactions.

* 1. Crime rate data: <https://raw.githubusercontent.com/ericonsi/CUNY_621/main/crime-training-data_modified.csv>

Crime levels (a binary variable denoting whether the crime rate does or does not exceed the median) for 466 towns in the Boston area, together with statistics on industrialization, pupil/teacher ratio, median home prices, etc.

* 1. Baseball team data: <https://raw.githubusercontent.com/ericonsi/CUNY_621/main/Baseball/moneyball-training-data.csv>

Outcomes for team baseball wins and losses based on hits, walks, strikeouts, etc., both made and given up.

Two mockup pages are included with the proposal so that it’s easier to understand/visualize.