Problem set 5

Jason Parker

Due: April 19th 2019 at 11:59pm

In this problem set, the data set is generated in the posted `ps5.R` file. Your solution should have 2 parts, the code you used (a `.R` file) and the text of your solution (any file format). Please format your code and text neatly in some way similar to the example that I have posted in box.

Question 1

For this question, use the data generated in my `ps5.R` file that was posted in box. In this simulation, we are looking at how family income affects student SAT scores. We are primarily interested in the following two linear models:

$$SAT_i = \beta_0 + \beta_1 income_i + e_i \tag{1}$$

$$SAT_i = \beta_1 income_i + \beta_2 1(group_i = 1) + \beta_3 1(group_i = 2) + \beta_4 1(groupi = 3) + e_i \tag{2}$$

The first model is pooled and the second is a within-groups model.

1. Study and describe the data generating process in your own words.
2. Assume for now that the data generating process is unknown, but the groups are still known. Run the pooled OLS model, the fixed-effects model, and individual models for each group separately. Why are the signs different between these different models?
3. Run three recursive partitioning models and plot the results. Model SAT using income, group, and both variables. What insights can you learn from these models?
4. Run a glmtree model for SAT. Define the glmtree that best fits this data (that you remember from the data generating process).
5. For all the rest of the questions, pretend that the groups are unknown to us as well. Your job is to find the relationships there. Using both the variables, find the optimal number of groups using k-means estimation (ignore scaling). Fit the k-means model and showing the correct means.
6. How often are you able to correctly identify the cluster of the data? Does k-means do a good job here? Try fitting with hierarchical clustering to see if you get a better result.
7. From this point, run the pooled model and the fixed-effects model using your endogeneously selected groups. Are you able to find the relationships you know exist from the data generating process?
8. Re-run the k-means estimation using only the income variable. How accurate is the estimation now? Are you able to find the relationships from the data generating process now?
9. Re-run the k-means estimation using both variables, but now scaling the variables beforehand. How is the estimation now?