ECON 282, Professor Hogendorn

Final Project Part 2

1. Describe your dataset in words, and don't forget to take credit for any data wrangling you did to get it into shape.

2. Explain what dependent variable you chose and why. Also explain if there are any specific independent variables that you are viewing as treatment effects. Alternatively, you may just be trying to find what all influences the dependent variable.

3. Use theory to work on your data. Are there any variables that should be omitted for theoretical reasons (can't be causal, bad control, etc.). Are there transformations that are needed for variables to be comparable across observations (percentages, ratios, differences from the mean, etc.). Are there variables you need to construct from the data (averages, counts, medians, etc.). Would it be better to make some variables into categories (e.g. categories 12-17, 18-25, 25-34 versus just age in years)?

4. Split into training and testing samples.

5. What, if any, concerns do you have about omitted variable bias? Look for evidence of it by comparing conditional means. Draw a causal diagram to illustrate at least one omitted variable situation and why you think it will or will not cause bias.

6. Use your theoretical knowledge of the data generating process plus ctree to construct interaction effects. (E.g. "young men" rather than just separate variables for age and sex.)

7. Try some initial regressions and visualize the residuals to see if nonlinear terms should be added. You can use theory for this or just use the Ivan method and add them for everything. But either way visualize first.

8. Use LASSO to narrow down the variables that you will actually use in your model. Remember that you may have theoretical reasons to keep in variables that LASSO does not choose or to remove variables that it does choose.

9. Try your regression on the training data and make any final changes based on t-statistics, R^2, etc. You can test up or test down a little here if you want.

10. One you are completely finished, run your regression on the testing data and interpret the results as best you can. If you had a particular treatment variable of interest, you can run one testing regression with it and one omitting it. This may give you better insight on how it affects your model.

Please submit your final project as a printed paper (or pdf if necessary). All your writing should be in a serifed, variable-space font such as Times Roman. All your R Code and output should be in a monospaced font such as Courier or Monaco. Be judicious in what you include from R so that the whole thing does not get longer than about 5-10 pages, but definitely include all the results from your final regression.