ECON 282, Professor Hogendorn

Assignment 5

The data for this assignment is on the website. It is data I generated. There are 98 x variables in the dataset which are all normal random variables with mean 0 and standard deviation 1. There is also a key variable – don't forget to remove it when you run regressions. Finally there is a y variable, which was generated by the following function

$$y = -X1 + X2 + X3 + X4 + X5 + 2 \times X2^2 + X5^2 + 2 \times X1 \times X3 + 2 \times X4 \times X5$$

Thus there are 5 variables that actually matter for y, and in two cases the square matters as well. There are also two interaction terms. Five of the other variables have high correlations with $X1$ through $X5$, but they don't actually have any causal effect.

To make things challenging, I have renamed all the variables, so you have no way to know if, for example, the $X4$ in the dataset is the original $X4$ in the equation above.

1. Divide the dataset into a 70% training sample and a 30% testing sample. Save the training and testing data.
2. Use ctree to determine what makes y likely to be greater or less than the mean of y. Look at the tree and write down any interaction effects that seem to appear.
3. Add those interaction terms to your training data, and then use LASSO to see which coefficients seem to determine the y variable.
4. Run an OLS regression using the results of LASSO. If some of the LASSO-selected variables have high p-values, that could be because of those 5 correlated variables I added. You could try re-

moving the variables one at a time to see if you can improve your regression model.

5. Now visualize the residuals from your regression as a function of each of the remaining x variables. See any nonlinearities? If so, you may have found one of the variables that has a nonlinear effect. Add squared terms to see.

6. Run your final regression on your training data, and then give it a try on the testing data to see how well it really works.