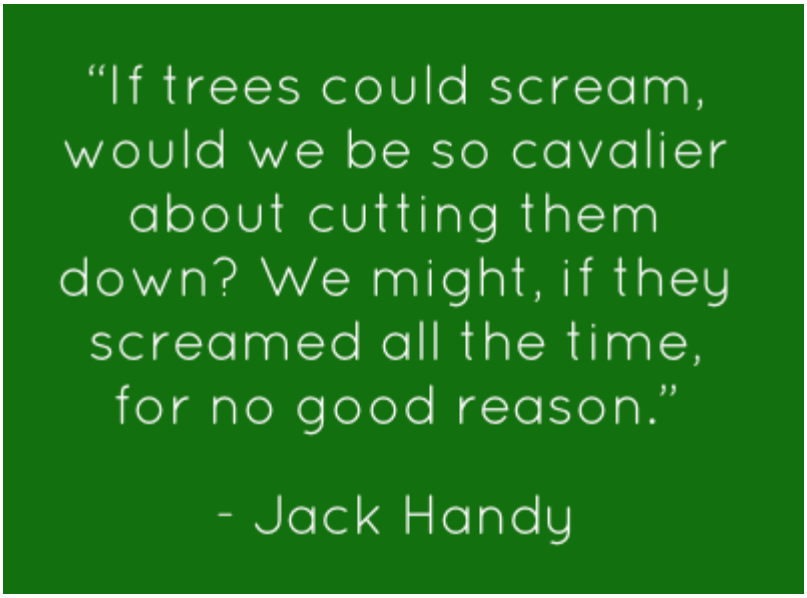
First Replication



1. Please use the data LATE_BETTER_THAN_NEVER.csv
   a. There is an outcome Y1 which is the treatment outcomes for everyone. There is an outcome Y0 which is the control outcomes for everyone. There is an treatment variable D which indicates whether individuals are in the treatment group (or control group). Make a histogram comparing the treatment and control outcomes for the treatment group, and then comparing the treatment and control outcomes for the control group.
   b. Now make a histogram comparing the treatment outcomes for the treatment group, and the control outcomes for the control group. How does the compare to your histograms for part a? Why?
   c. Finally, calculate the actual ATE by finding the average difference between Y1 and Y0 for the entire population.
   d. How does this compare to the coefficient from a linear regression where you only observe the Y1 outcome for treatment, Y0 for control, and a D variable for whether you are in treatment. What does this tell you about the importance of random assignment?

2. Suppose you are thinking about running an experiment. You hope to study whether assignment to Ben Hansen's metrics increases the odds of finding a job over taking Glen Waddell's class. The odd's of finding a job coming out of Glen's class is 70 percent.
   a. If you want a minimal detectable effect of increasing the odds of finding a job by 5 percent, how big would the entire sample need to be (assume the odds of ending up in either class if 50/50)?
   b. What is your minimal detectable effect if you have a sample size of 1000?

2 Regression Discontinuity and Drunk Driving Recidivism.

For this problem you will estimate a regression discontinuity design to test whether having a BAC over the legal limits. You can compare this to the paper by yours truly about DUI punishments we read earlier.

The file is DUI_deidentified

First things first. You must test for non-random sorting in the dataset.

a. Create a histogram of the running variable, BAC. Make sure you do it allowing for discrete bins. Is there evidence of clear sorting at the threshold?

b. Next run a regression discontinuity model. To do so, create a dummy variable for a BAC over .08. Include that dummy variable, and the rescaled BAC (BAC-.08) as a control, and also include an interaction between that dummy variable and the running variable in model. First use age, gender, accident at the scene and race as outcomes. Do those factors shift at .08?

c. Now run a regression of recidivism on the same regression discontinuity design. What is your estimated effect using a bandwidth of .05, and a rectangular kernel (no weighting). Create a visualization of this by graphing the mean recidivism rate against the running variable. Show this for the whole BAC distribution, and the range from .03 to .13. Please include a fitted line.

d. Do the same thing as part D but for the aggravated threshold of .151.

e. Now run this model for every possible bandwidth between .01 and .07. Store both the point estimates and lower and upper confidence intervals. Create a scatter plot of the confidence interval and the point estimates. Are the estimates robust? Create a visualization of this.