Economics 217

Homework #3

Due Tuesday, February 26th

Problem 1

In this question we will utilize bootstrap procedures to evaluate the differential recovery after the great recession for California and Nevada.

a. To begin our study of the differential recovery, please use the Org dataset from the website, and restrict the sample to include California and Nevada for the years 2008 and 2013. Then, estimate the following difference-in-difference regression:

$$\log(rw) = \beta_0 + \beta_1 D_{ca} + \beta_2 D_{2013} + \beta_3 D_{2013} D_{ca} + controls + u \tag{1}$$

where D_{ca} is a dummy variable identifying California, D_{2013} is a dummy variable identifying observations from 2013, and controls are a set of other controls that may affect the real wage. For these controls, please use age and education of the respondent.

Please run a simple regression and interpret the coefficient on β_3. Please construct a 95% confidence interval.(10 points)

b. For this question, please run 1000 bootstrap replications of the difference-in-difference regression, with each replication being of the same size as the original dataset. Please use the data resampling technique (as opposed to residual resampling). Is the 95% confidence interval larger or smaller than part 'a'? (10 points)

c. As you may recall from 216, for a difference-in-difference regression to be appropriate, there should be no differential pre-tends in the data. Please propose a regression to test for the presence of pre-trends, and estimate this regression. Please test for the presence of pre-trends using a 95% confidence interval constructed via a residual resampling procedure with 1000 replications. (10 points)

Problem 2

In this question we will compare the out-of-sample predictive power of k-nearest neighbors and classification trees. Specifically we will predict labor market outcomes using the Org dataset in 2013. To fit the models, use data from California. Begin by creating a character variable measuring three employment outcomes: "out of the labor force", "unemployed", and "employed".

a. Using both techniques, fit a model predicting the labor market outcomes as a function of education, gender, race, marital status, number of children, and whether the respondent is a

citizen. Using data from outside of California in 2013, which model has the best accuracy in out of sample predictions? (20 points)

b. Please generate a visual representation of the optimal classification tree, and interpret the figure. (10 points)