Economics 217

Homework #3

Due Friday, February 17th

Problem 1

This question will use the Gehan survival data, available on the course website.

a. Using GLM, please estimate the survival model that is featured at the end of lecture module 4 (Yes, I'm am asking you to code and execute exactly what is in the notes). Please interpret both the constant and the coefficient on the treatment dummy variable.(10 points)

b. Just like the leave-one-out estimator from the non-parametric part of the course, now run the model from part 'a', but run the model N times, dropping one observation on each iteration (don't forget to put the observation back). Please collect all coefficients of interest and summarize their distribution, both in text and visually with a figure. Where does the original estimate lie within this distribution? (10 points)

c. We justified the model in 'a' by appealing to the issue of censoring, and needing to adjust for censoring in the likelihood function. As an exploratory exercise, use the gam function in R to study the relationship between time to relapse, censoring and being in the treatment group. (10 points)

Problem 2

For this question we will use the loess and gam functions in R to study monthly fluctuations in real wages. From the website, please download WageTimeSeries.csv, which reports the average real wage in each month of 1983, 1988,1993, 1998, 2003, 2008, and 2013.

a. Using loess, please estimate the relationship between the real wage, rw, and month of the year, $month$. Please plot your answer, using a span of 1 for the estimation. (10 points)

b. Using gam, please estimate the relationship between the real wage, rw, month of the year, $month$, and $year$. Be careful how you treat $month$ and $year$ in the estimation. Please provide confidence intervals for your estimates on the Figure. (10 points)

c. Back to using loess, repeat the same exercise as in (2a), but please write a cross-validation procedure to find the optimal degree of smoothing (span, in the function). You may not use a "canned" package from the R library to run the cross-validation. Please plot your optimal figure, as well as provide results as to why you chose the degree of smoothing that you did. (10 points)

Problem 3

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.