Economics 217

Exam #2

Due Thursday, February 23rd, noon

Instructions: Take home, notes and online resources are fine. You are required to submit a brief write-up showing your results (just like your homework), as well as a source code that is both runnable and readable. Credit will be given based on both materials. Partial credit will be given for code that shows comprehension of the material, even if the end result is incorrect.

Absolutely no discussing this exam with anybody, whether your classmates, instructors, other faculty, or anybody else that provides help. Any infractions will result in a failing grade for this exam, potentially the course, and possible dismissal from the program.

All questions should be directed to Alan or Jijian via email. All answers will be posted for the entire class to see.

Good luck, and have fun!!

Problem 1

For this question, please use the org data from the class website, we wish to estimate the following generalized additive model using R.

$$\log(hourslw) = s(age) + educ + wbho + month + u$$

Here, $hourslw$ is hours worked, $educ$ is a education factor variable, and $wbho$ is an factor variable describing the race of the respondent. Please treat "month" as a factor variable.

a. Please estimate the model described above, using the default smoothing parameters given by the R library that you use. Please display your results using a 2X2 plot, i.e. how each variable affects the outcome. Please label your plots. (10 points).

b. Using your plots from part 'a', please interpret the estimated relationships between all independent variables and $\log(hourslw)$. Where possible, please discuss the statistical significance of these relationships. (20 Points)

Problem 2

For this question, please use the data set with your name on it from the exam 2 webpage. This dataset was randomly created using the following equation

$$y = x^h \sin(Ax) + u$$

The variable x is evenly spaced between 0 and 10. The noise parameter u is random from a normal distribution with mean 0 and standard deviation 1. The term h is somewhere between 0 and 1, and A is somewhere between 1 and 2. Each student has a different value for both h and A.

Your job in this question is to estimate this function in a few different ways.

a. Using LOESS with degree=1 in R, use a leave-one-out (cross-validation) procedure to estimate the optimal span for non-parametric estimation. Please report this span, and plot your optimal non-parametric function using this span, including the original data on the same plot. (15 points)

b. Since I generated your dataset with noise, I would like to get a sense of the sampling variation in the non-parametric fit. Please write a wild-bootstrap procedure to provide a 90% confidence interval for your non-parametric fit. In doing so, use the optimal bandwidth you found in part 'a'. (15 points)

Suggestion: This is the hardest question on the exam, so I'd do it last. It is not required for part *c*. The hint is that you should run a new loess fit for each replication and then work with that....

Note: Please use at least 20 bootstrap replications for part b. It is possible that your confidence intervals will cross the original estimate (I'll tell you why later)

c. There are a number of ways to estimate h and A from the original function. Usually one would use non-linear least squares or something similar. However, for this question, I'd like you to use the logic from cross-validation - leave-one-out and test for out of sample prediction error - to find the optimal choices for h and A (15 points)

Precision: When finding the optimal h and A, please do so to the nearest tenth (e.g.. 1.1)