Economics 217

Exam #2

Due Friday, February 19th, 8pm

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 your classmates, instructors, or other faculty. 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 data set "Wages1983.csv" from the Exam 2 website (the link I sent over email), we wish to estimate the following generalized additive model using R.

$$log(wage) = s(educ) + s(exper) + s(feduc) + s(meduc) + u$$

Here, *wage* is the wage of respondent in 1983, *educ* is their current years of education, *exper* is their current experience in the workforce, and *meduc* and $feduc$ are years of education for the respondent's mother and father, respectively.

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

b. Using your plots from part 'a', please interpret the estimated relationships between *educ*, *exper*, *meduc*, and $feduc$ and $log(wage)$. Where possible, please discuss the statistical significance of these relationships. For example, "respondents with education between 4 and 5 years earn a significantly lower wage than the sample average, holding other variables fixed". It may be helpful to use the 'abline' function to pinpoint different points on your Figure. (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

$$y = \beta_0 + x + \beta_1 \mathbf{1}(x > h)(x - h) + u$$

This is a regression kink. Recall that $\mathbf{1}(x > h) = 1$ when $x > h$, and $\mathbf{1}(x > h) = 0$ otherwise. So, when $x \leq h$, the term $\mathbf{1}(x > h)(x - h)$ is zero. Once $x > h$, $\mathbf{1}(x > h)(x - h) = (x - h)$ and β_1 measures the change in slope when $x > h$.

Your job in this question is to find the kink in your dataset. The question is two parts.

a. Use a leave-one-out (cross validation) procedure to estimate the location of the kink. Please report this estimate, $\widehat{h}$, and the estimates for β_0 and β_1 at this optimally placed kink. (15 points)

b. Since I generated your dataset with noise, I would like to get a sense of the noise in the estimate of the kink, $\widehat{h}$. Please write a bootstrap procedure (resample data, not residuals) to provide a 90% confidence interval for $\widehat{h}$. (15 points)

Notes/Hints for this question:

Precision: When finding the optimal kink, please do so to the nearest tenth (e.g.. 4.1)

Bootstrap Replications: Please use at least 20 bootstrap replications for part b. Ideally one would use more, but I want this to be manageable for all computer speeds. On my laptop, using conservative (ie. slow) programming, part b takes 10 minutes to run with 20 replications. This could certainly be sped up, and while not required for the question, I do expect that many of you will figure out a way to speed up the procedure. While testing your code, use fewer bootstrap replications to save time waiting.