Economics 217

Homework #3

Due Tuesday, February 20th

Problem 1

This question will use the "Veteran" survival data, which is now available on the course website. It is a standard survival dataset originating from the Veterans Administration that measures the initial state of patient health and the effects of a chemotherapy treatment on survival from lung cancer. It contains the following variables:

- ID: ID of the patient
- TIME: time of start of the observation period (if Y=0, first occurrence), death (if Y=1) or censoring (if Y=0, second occurrence)
- Y: 0 to indicate the initiation observation or censoring, and 1 to indicate death
- trt: treatment type
- celltype: histological type of the tumor
- karno: Karnofsky performance score that describes the overall patients status at the beginning of the study
- diagtime: Time between diagnosis and start of the study
- age: age of the patient
- priortherapy: indicates if the patient has received another therapy before the current one

To be clear, each patient is observed twice. On the first occurrence, (TIME=0,Y=0) for all patients (the initial observation). The second observation is either censoring (TIME>0,Y=0) or death (TIME>0,Y=1).

a. First, let's clean the data. We are only interested in the basic survival questions as a function of treatment/control and a few other covariates. So, using the dataset, drop observations for which TIME=0. Please summarize and briefly discuss the variables TIME, Y, trt, and celltype. Do any of these variables suggest we should clean the dataset further? If so, please clean the dataset. (10 points)

b. Please plot Kaplan Meier estimates of survival as a function of treatment and control groups. Please label your figures and briefly discuss your answer. (10 points)

c. Please estimate a proportional hazards model, hazard being a function of treatment/control, cell type, age, and prior therapy. Please summarize and briefly interpret your results. (10 points)

Problem 2

For this question we will use the loess and gam functions in R to study the relationship between real average hourly wages and hours worked (using the Org dataset).

a. Using loess, please estimate the relationship between the log real wage, rw and hours worked. On a second plot, please use log of hours worked. Please interpret your figures. (10 points)

b. Using gam, please estimate the relationship between the log real wage and log hours worked controlling for education and age. Please provide confidence intervals for your estimates on the Figure, and interpret your results. (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 relationship between real wages and educational attainment using the ORG dataset.

a. To begin our study of real wages, restrict the sample to CA in 2013, and then estimate the following linear regression:

$$\log(rw) = \beta_0 + \beta_1 D_{HS} + \beta_2 D_{SC} + \beta_3 D_C + \beta_4 D_{AD} + \beta_5 age + u \tag{1}$$

where $\{HS, SC, C, AD\}$ represent a maximum education level of high school, some college, college, and an advanced degree, respectively. Age is the age 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 this 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 in previous work, we'd like to test for the effect of removing all education variables. For this, you should use an F-test. However, rather than running a simple F-test, I want you to use residual resampling to bootstrap the F statistic using 1000 replications. Illustrate the variation in your F-statistic using a Figure, and discuss the implications of your results.