Problem Set 2

PPPA 6022
Due in class, on paper, March 5

Some overall instructions:

- Please use a do-file (or its SAS or SPSS equivalent) for this work – do not program interactively!
- I have provided Stata datasets, but you should feel free to do the analysis in whatever software you prefer. If you need to transfer to another format, use StatTransfer.
- Make formal tables to present your results – don't use statistical software output. Make sure you discuss the answers.

1. Hazard Models

For this problem, we are interested in how covariates impact the rate at which people are likely to have children. We are using data from the National Longitudinal Survey of Youth 1979, which you can read more about at www.nlsinfo.org. For our purposes, you should know that it's a panel of individuals who were 14 to 22 years old in 1979. They have been followed at regular intervals since the survey's inception. I've downloaded the data and reformatted them so they are easily useable for this problem set (don't think it would be this simple on your own!). I didn't download many interesting and useful variables, so don't think of this as the extent of the data. You may find the page on the weight variable helpful:
https://www.nlsinfo.org/investigator/pages/search.jsp#R2141300

(a) Summary statistics warm-up (to help you understand the data set-up): Of the 1979 population, what share will ever have kids? What share of the 1979 population has kids in 1979? What share of the 1990 population has kids? Of the population with no kids in 2000, what share has kids in 2002? What proportion of this population (those who have kids in 2002, with no kids in 2000) are male?

(b) Draw an overall survival curve for the likelihood of having kids. Recall that for the Worcester Heart Survey data we looked at, the survival curve was for death. Here, the "death" equivalent is having kids. Condition on not having kids, is the likelihood of having kids greater between 1979 and 1989, or between 1989 and 1999?

For hazard analysis in Stata, you may find this page helpful:
http://www.ats.ucla.edu/stat/examples/asa/test_proportionality.htm
Some key commands are stset and sts graph.

(c) Draw the same survival curve, separating into two curves: one for urban, and one for rural. What does this tell us about the likelihood of entering parenthood by urban status?

(d) Estimate a Cox proportional hazard model, where the depending variable is having kids. Use urban/rural, weight and gender as control variables. Present the results in a table, and explain the

effect of each variable. Then find the change in the hazard ratio for a 10 lb change in weight on the likelihood of having children.

2. Instrumental Variables

For this problem, we are revisiting a classic: Angrist and Kreuger. We use a random sample (chosen by me) from the 1980 public use micro data file (five percent of long-form respondents; this is the 1980 version of data we used last class). Documentation for the version we're using is at www.ipums.org.

Note that A&K keep only white and black men born between 1930 and 1959. Unfortunately, I didn't include race in my download, so ignore the race restriction.

Some of additional variables are not an exact match. We don't have a continuous education variable like A&K (not sure why not), so make educ into a continuous variable as best you can. We don't have weeks worked, so ignore restrictions relating to that. Use incwage as the dependent variable, rather than weekly earnings.

(a) Replicate the first two rows of A&K's Table 1, but don't worry about de-trending the data as A&K do.

(b) Do the A&K first stage, using two sets of instruments: (a) quarter of birth, (b) quarter of birth * birth year. Do the first stage to do the analysis in Table 5, column 8. Make a table to report the F for the instruments and the additional R2 from the instruments in each regression; you don't need to report all the coefficients. Interpret whether these instrument seem "good" in a weak instrument sense.

(c) Use your previous specification to make two predicted value variables for education. Do two A&K second stages, one with each predicted value. Then do a parallel 2SLS analysis using Stata's ivregress (or the equivalent). Compare the coefficients and errors on the variable of interest. What are your findings about education? Why are the coefficients and errors the same or not?