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.
- This problem set uses some large data. For the Census data, I have put the full dataset up on Blackboard, and I've also put a smaller version. For the CPS, only the small one would fit.

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?

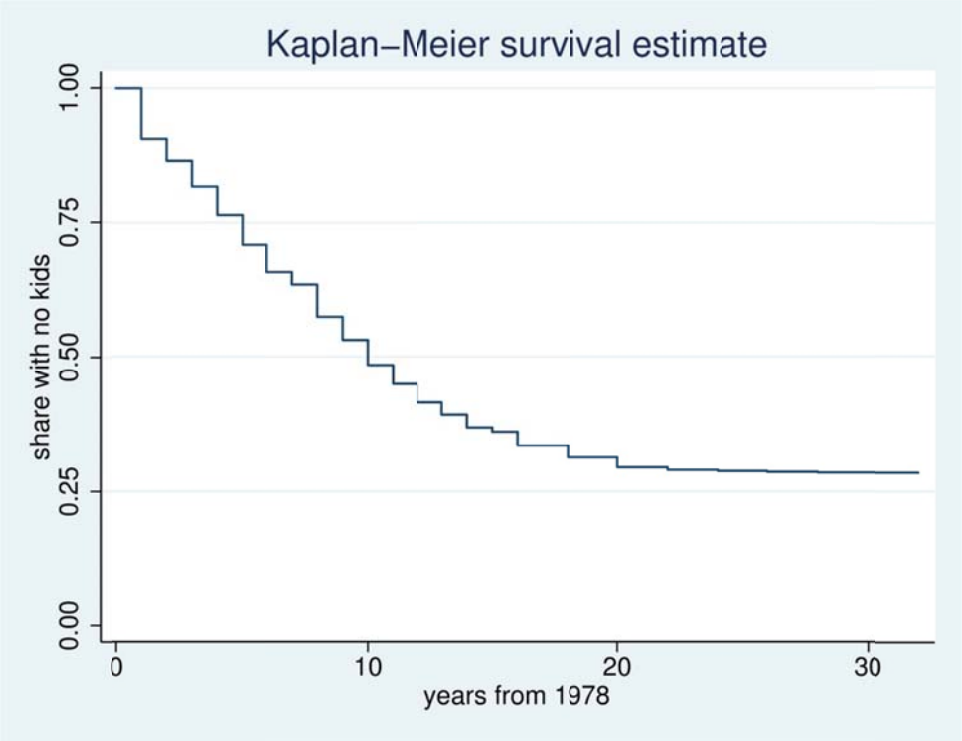
Of the 1979 population, 72 percent will eventually have children. As of 1979, 12 percent have children. Of the 1990 population, 45 percent have children. Of the population with no kids in 2000, 1.5 percent have kids in 2002. All of those who do have children are women.

(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

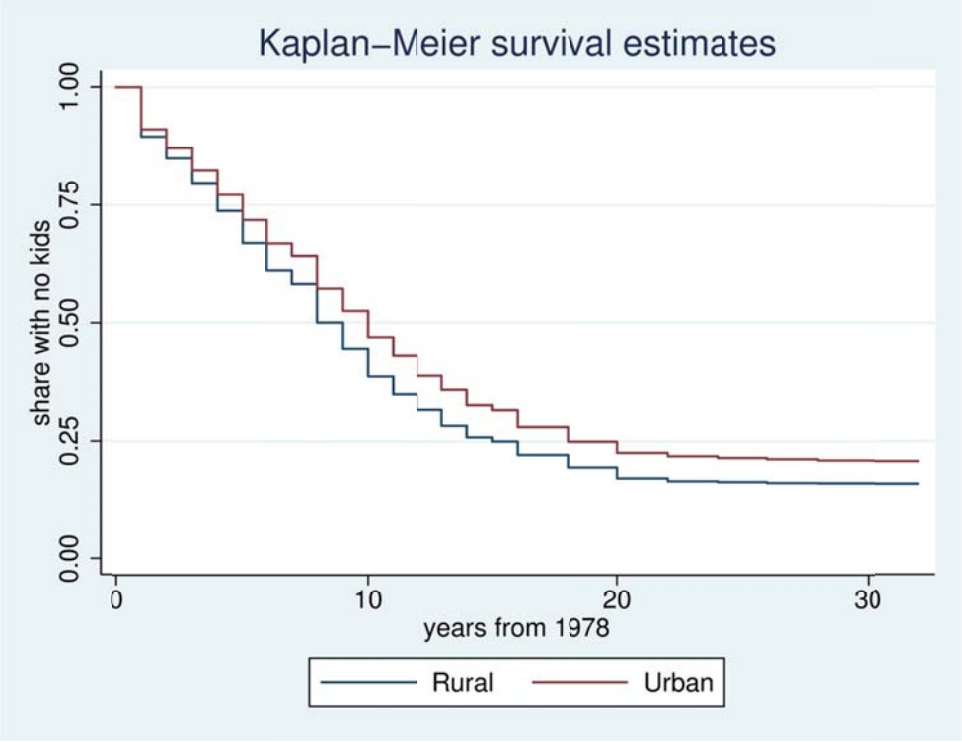
Some key commands are stset and sts graph.

Below is the survival curve we discussed in class.



The slope of the survival function is steeper for the first decade (1979-1999) than the second (1999-2009), meaning that people were transitioning more quickly into having children in the first decade relative to the second.

(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?



This picture shows that, for any year, people rural areas are more likely to enter parenthood than urban residents. The difference is small in the early years, and increases in the late 1980s. The slopes of the two curves seem roughly similar.

(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.

	Cox Model
Weight, lbs	1.002*
	(0.0004)
1{Urban}	0.872***
	(0.0314)
1{Male}	0.647***
	(0.0212)
Observations	62,382

In a precise sense, there is no association between weight and the likelihood of entering parenthood; the hazard ratio is almost exactly one. Note that we can reject coefficients substantially different than one. Sometimes we refer to this type of finding as a "precise zero" (except that here it is a precise one). We find nothing, and it's not that we can't say anything about the result – we can say rather precisely that this variable is not associated with the rate of entry in having children.

Statistically, men enter parenthood more slowly than women. The coefficient tells us that men are more than 30 percent less likely to enter parenthood at any given time.

We also see a statistically significant difference between the behavior of urban residents relative to rural ones. Urban residents are roughly 15 percent less likely to enter parenthood.

To calculate the effect of a 10lb change in weight on the likelihood of entering parenthood, recall that
$HR_{1\text{ lb change}} = \exp(\beta) = 1.002$
This implies that $\beta \approx 0$ (= 0.002)
$HR_{10\text{ lb change}} = \exp(10 * \beta) = 1$ (or, more precisely $\exp(10*0.002) = 1.02$)
Virtually no change!

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.

See columns 1 and 2 of the table at the end. Even without de-trending the data, the results are very similar to A&K's original results. Men born in the first quarter of the year, and to a lesser extent men born in the second quarter, have less education.

*(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.*

See columns 3 and 4 of the table at the end. The F-tests for these instruments are in both cases quite low. The F-test value for using three instruments (column 3) is 3.4. This is below levels that would now be considered acceptable for instrument strength. The F-test value using quarter of birth*birth year is even lower, at 1.4. In both cases, the R2 for the regression increases by

0.001 when I add the instruments. In other words, while the instruments may be individually significant (at least in the first case), they do not explain a substantial amount of the variation in the endogenous variable.

(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?

This regression finds that an additional year of education increases wages by a whopping 17 percent; much larger than the estimates in A&K. This coefficient is significant at the five percent level.

The coefficients using `ivregress` and doing the regression manually are exactly the same – as they should be. Mechanically, the IV coefficient is generated by using the instrumented variable.

However, the standard error for the IV estimation is not correctly calculated using the OLS formula. In addition, the IV standard error should be always larger than the OLS standard error. In my example, the OLS standard error is actually a tad larger (0.081 vs 0.080) than the IV standard error. I expect that this anomaly is driven by rounding errors, since the difference between the values is quite small.

Table for Question 2

	Question 2(a)		Question 2(b): 1st stgs		Question 2 (c)			
	Table 1, row 1	Table 1, row 2	3 instrumts	bq * birth year	Using predicted value	ivregress	Using predicted value	ivregress
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1{birth quarter=1}	-0.138***	-0.070*	-0.102*					
	(0.039)	(0.031)	(0.041)					
1{birth quarter=2}	-0.098*	-0.047	-0.103*					
	(0.039)	(0.031)	(0.041)					
1{birth quarter=3}	0.018	-0.005	-0.012					
	(0.039)	(0.030)	(0.040)					
Predicted value,					0.171*	0.171*	0.171*	0.171*
years of education					(0.081)	(0.080)	(0.081)	(0.080)
F test: instruments	7.629	2.453	3.407	1.356				
p-value of F test	0.000	0.061	0.033	0.103				
R-squared	0.000	0.000	0.034	0.034	0.056	0.070	0.056	0.070
Observations	51,162	71,816	43,163	43,163	43,163	43,163	43,163	43,163