

Economics 607 PROBLEM SET 2

Due Monday, October 25

In this problem set, we will replicate the key results from Angrist and Evans (1998), “Children and Their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size.” While the original work used data from 1980 and 1990, we will use data from the 5% 2000 Census Public Use Micro Sample. Please note that we will take a number of shortcuts along the way in order to make the project more manageable given our timeframe.

1. Download data and construct analysis sample

- a. Go to <http://usa.ipums.org/usa/> to begin the data extract process. All of the default options preceding variable selection are fine. Select “select data quality flags” and then download the following variables from the 5% 2000 Census in addition to those selected by default: SUBSAMP, HHTYPE, NFAMS, NSUBFAM, PERNUM, MOMLOC, STEPMOM, SPLOC, NCHILD, YNGCH, RELATE (detailed), AGE, SEX, MARST, RACE (general), HISPAN (general), EDUC (general), WKSWORK1, UHRSWORK, WORKEDYR, FTOTINC, INCWAGE, POVERTY *in addition to other outcome variables that might be affected by family size*. Note that you can reduce the size of the initial data set by looking at the sample restrictions below and making a few “case selections” for your download. (In general, however, it is best to start with the entirety of the data if possible.) Also note that one usually downloads data flags and throws out observations with imputed values.

Important: Each row of data represents an individual. Individuals in the same household will have the same value for serial. Within each household, individuals have a person number, pernum. As such, serial and pernum uniquely identify each person.

- b. Now create a .do file that will read in the data, restrict the sample, etc. I recommend that you rename the .do file provided by IPUMS ReadInData.do and then call on this program within the program you write with the command “do ReadInData.do.” After reading in the data, drop those in group quarters. You may find it useful to save the current data set as we will want to return to this exact data after part c (note that the **preserve** command is often a good alternative to saving).
- c. We will now focus on children or, more specifically, those with their mothers in their household. Limit the sample to those with a biological mother in the household using momloc and stepmom. To make the following steps easier, it will be useful to create a mother identifier:

gen momunique=string(serial)+"."+string(momloc);

We now want to manipulate the data so that we have a single observation per momunique.

- Use **egen** to create a variable for the number of siblings for each momunique
- Sort data by momunique and negative age before issuing the following command to generate birth order: **by momunique: gen birthorder=_n;**
[Note that this won’t be perfect because some siblings will have the same age—we’ll come back to this soon.]
- Limit the sample to 1st-3rd children
- We will now reshape the data so that there is one observation per mother. Begin by keeping only the household identifier, totkids, and child-specific information on sex, age, and birth order. Then use the **reshape wide** command:

reshape wide observation-specific-variables, i(groupID) j(withingroupID);

- d. After saving the child-info, we will want to merge it onto the main sample from part b. After creating a variable momunique in the main data, this is easy to do using the **merge** command matching on

momunique. After doing this merge, you may want to save your data again (or preserve it).

- e. Now, using the entirety of the data, limit the sample to individuals who have spouses in their household using sploc. In a similar fashion as in parts c and d, manipulate this data and merge it back onto the main data set so that the main data set will have variables for spouse's age, whether worked the previous year, weeks of work, usual hours of work, and earnings.
- f. You should now have the full census data set with children's info merged on mothers and spouses info merged onto spouses. Now limit the sample to women, aged 21-35, with non-zero sample weight, at least 2 kids, youngest child ≥ 1 , and those for whom we know which of the 1st two children was born first (i.e., they don't have the same age). Also make an indicator variable for those with 2nd children who might be twins (same age as 3rd).

2. Analysis – compare your results to the original study in each of the following parts

Note: Before using any variable in your analysis, you should look at it in detail to see if it needs cleaning. Also, don't forget to use sampling weights and robust standard errors where applicable.

- a. Replicate Table 2, using totkids from part 1d in place of “children ever born.” Also be sure to put money variables into 1995 dollars. Don't worry about the standard deviations.
- b. *After dropping women whose second children might be twins*, replicate the “fraction that had another child” columns of the lower panel of Table 3. Note that these estimates should be obtained from regressions.
- c. Replicate the portion of Table 4 not dealing with twins. Instead of years of education, consider indicators for less than 12 years of education, exactly 12 years, and more than 12 years.
- d. Replicate Table 8.
- e. Do something new. Perhaps explore different outcomes, different subsamples, etc.