Replication Exercise
Angrist-Evans 1990

The big picture

Replication is an important part of academic research: the process of obtaining the raw data files used by another researcher, and retracing each of the steps they performed with that data to confirm that you can replicate their results. Replication is also a great way to learn the nuts and bolts of empirical work in hands-on way. You know the endpoint, but the challenge is to get there. (It's even more challenging when you don't know the end point. i.e., you are doing original research and feeling your way through the results.)

This particular replication exercise is also a good chance to develop your large data set handling skills. What do I mean by this? For smaller data sets, you can look at the data to fix, code, and merge it. But with hundreds of thousands of observations, you have to conceptualize each step in an abstract way. It's harder but more fun (well, that's a matter of taste) but also eventually much more reliable.

Angrist-Evans is one of the classic papers of modern labor economics. It looks at the effect of having an extra child on female labor force participation and earnings. By now, the first thought you hopefully have is that if that's the goal then you can regress a woman's labor force status on whether or not she has a child or the number of children that she has. But hopefully an instant later you have the second thought that you can't (shouldn't) run that regression because it is likely to be biased. Why? Read the paper and find out, but think omitted variables bias and simultaneity.

What you need is either a good experiment (unlikely in this case) or an instrumental variable. Angrist and Evan's IV strategy is to use the gender mix of the first two children as an IV for having a third child. Again, read the paper for details, but roughly the idea is that the gender mix of the first two children is random and parents have a preference for a gender mix in their children. So having either two boys or two girls as your first two kids increases the likelihood of having a third child (think of it as a randomly assigned encouragement).

Using micro census data for the US (IPUMS) they can figure out the gender mix of the first two kids, whether or not the mother had a third child, and whether or not she is working (and for what wage, etc.) Your task is to read the paper, think hard, and figure out how to do this using the same data.

Replication: the basics and some (free) hints

Replication involves three steps: obtain the raw data, clean / reshape the data to the format you need, and run the regressions.

Obtaining the raw data

The first step is to get the data. There are two ways to do this. The preferred way is to do this yourself. Begin by reading the paper. In a spreadsheet make note of each variable mentioned in the paper (don't forget to read the footnotes in the text and tables as well).

Then, you should create an account at: https://international.ipums.org/international/. Angrist and Evans use 1980 and 1990 Census data. We are going to focus on the former. So after creating your account, go to the select your data section, select USA 1980, and then select the variables you will need. From your spreadsheet search through the available variables till you find a match for all of your variables. Some of these won't be obvious at first pass. It may take a few attempts. As you find possible matches enter these on to your spreadsheet. You can save your data request and update it as you add

more variables. Once you think you have what you need, arrange to download it (it will probably be about 0.8 Gb).

Given our time constraints, I am giving you two options here. If you can download the raw data you need by the end of September, you can submit the spreadsheet and some summary stats to us. If these look good you'll get a check or check-plus that will count as bonus points on the assignment.

If you don't have the raw data by then, then I'll circulate my file. If you opt for this latter route, there are many things you can do while waiting for your data. Read the paper, and starting thinking through the third step of the assignment (running regressions).

Cleaning and shaping the data

In order to get going you will need to understand the structure of the data. You are getting "long format" data, which means each row in the data contains records for one individual, each column the different variables. This is household data, meaning that there is data on the respondent, the respondent's spouse, and the respondent's children. What you need to know next is how each household is identified, and how you can tell who the parents and children are. Some of the key variables are as follows:

MOMLOC is a constructed variable that indicates whether or not the person's mother lived in the same household and, if so, gives the person number of the mother (see PERNUM). The code "00" denotes that the mother is not in the sample.

SPLOC is a constructed variable that indicates whether or not the person's spouse lived in the same household and, if so, gives the person number (PERNUM) of the spouse. The code "00" denotes the spouse is not in the sample.

PERNUM numbers all persons within each household consecutively (starting with "1" for the first person record of each household).

SERIAL is an identifying number unique to each household in a given sample.

My final hint for now, and a big one, is to tell you about two commands that are going to be crucial for this exercise: reshape and merge.

The Stata command reshape, changes data from wide to long format or vice versa. For example, suppose your data is arranged with one row for each child, e.g., for the variable age. But instead of one row listing the age of each child, I want for each household a single row with age1=age of the first child, age2=age of the second child, etc. Reshape will do that for you. Of course, you'll have to figure each child's sequence number (i.e., first child, second child, third child, etc.)

You'll also need to figure out how to merge data sets. Why? you ask, since you already have a working data set. But think about the kind of regression you want to run, and what dataset format it implies. You want mother's work status on the left hand side and some combination of child and household characteristics on the right hand side. Currently you data have parents and children all stacked up in long format. You're going to need to reshape and merge the data.

Running regressions

Once you have data in the format you need (or while you're waiting for me to release the data), the next step is to generate the variables you need for Angrist and Evan's specification (e.g., a dummy for the first two kids being the same sex). Then you need to figure out what sample restrictions they use. These are described in the paper, and will be things like: women with at least one child living with their husband who are less than age 60, and children 17 or younger. You have to read their paper carefully, and really look at the footnotes and notes at the bottom of the tables. Since you have your spreadsheet going you can add a column for how Angrist and Evans transform or use a variable.

Something else you can do while waiting for your data is to figure out the specifications you will need. This often happens in real research projects when you have to wait for permission to access your data. How to proceed? Create fake data. By that I mean create the variables you think you need and start by populating them with normal random variables (if you want to be very fancy, you an match the descriptive stats given in the paper). Then run the specifications you need and create output tables – while the numbers will be wrong, you can get started on the work of getting the right specifications and formats.

Although the goal of a replication is normally to reproduce *exactly* their results, my attempts to do this have gotten me very close, but not to an exact replication. So don't pull out your hair if you can't exactly (by which I mean to each decimal place) replicate their finings. But you should be able to get to within one decimal place more or less of their findings. When you're close, you'll know it.

Focus on trying to replicate Table 2, Table 6, and Table 7 for the 1980 data.

Be warned: this is time consuming, at times frustrating, work, but when you succeed you will have mastered some of the key skills of data work, and have that "I just climbed a mountain" kind of feeling.

Grading scheme
C don't even manage to open the data
B- replicate basic summary statistics
B manage to get the data in the necessary format to run the specifications
B+ manage to run the same specifications, and get broadly similar results
A- replicate the results to the same degree of precision I managed to achieve
A replicate the results more precisely than I did *or* extending their results in an interesting way

The hint exchange
This is an individual assignment. You should work alone, not in groups. But if you get stuck, then I encourage you to ask each other for hints. Although I can't formally reward you for being helpful, at the end of the semester I will poll the class on who was the most helpful to his or her colleagues and find some appropriate way reward him or her.

Getting stuck and working your way out of a roadblock is part of the experience of advanced empirical work. I want you to find yourself in those situations and to have the satisfaction of getting out of them. And same time, there can be advantages to talking to other people and sometimes everyone needs a bit of help to get unstuck. I would like these two to balance out: try really hard to do things on your own. If you can't, then definitely ask for help!