14.310x Data Analysis for Social Scientists
Endogeneity, Instrumental Variables, and Experimental Design

Welcome to your final homework assignment! You will have one week to work through the assignment. We have provided this PDF copy of the assignment so that you can print and work through the assignment offline.

Good luck!

2LS Estimates: Questions 1 – 9

In this part of the problem set, we are going to replicate part of the results of Joshua Angrist and William Evans' article "*Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size.*" Here is the abstract of the study:

> Research on the labor-supply consequences of childbearing is complicated by the endogeneity of fertility. This study uses parental preferences for a mixed sibling-sex composition to construct instrumental variables (IV) estimates of the effect of childbearing on labor supply. IV estimates for women are significant but smaller than ordinary least-squares estimates. The IV are also smaller for more educated women and show no impact of family size on husbands' labor supply. A comparison of estimates using sibling-sex composition and twins instruments implies that the impact of a third child disappears when the child reaches age 13. (JEL J13, J22)

The purpose of this exercise is to study how fertility affects female labor supply. In order to do this, we are going to compare female labor supply in households with two children versus households with three children. Since fertility decisions are endogenous, we are going to use two sets of instruments: whether there is a multiple pregnancy in the second pregnancy and sex composition of the first two children. This latter instrument was the one proposed by Angrist & Evans (1998). Intuitively, parents are more likely to have a third child when the first two have the same sex. Assuming that whether the first two children have the same sex is random, we can use this variable as an instrument for the number of children in the household.

We have provided you with the data set census80.csv that corresponds to an extract of the 1980 US Census. It has been restricted to the set of families with two or three children and with mother's age between 21 and 35 years. The data set contains the following variables:

- **workedm**: whether the mother works.
- **weeksm**: number of weeks the mother works.
- **whitem**: mother is White.
- **blackm**: mother is Black.
- **hispm**: mother is Hispanic.
- **othracem**: mother is of other race.

- **sex1st**: sex of the first child (0 corresponds to male and 1 to female).
- **sex2nd**: sex of the second child (0 corresponds to male and 1 to female).
- **ageq2nd**: age in quarters of the second child.
- **ageq3rd**: age in quarters of the third child.
- **numberkids**: number of children in the household.

Load the data into R, follow our instructions, and answer the following questions.

Question 1
Use the command `summary` to summarize the variables in the data. Using your output, fill in the following information:

Please round all answers to the second decimal place, i.e. if the answer is 6.6728, round to 6.67 and if it is 6.6788, round to 6.68.

a. Fraction of mothers that work

b. 3rd quartile of weeks worked

c. Proportion of Hispanic mothers

d. Median age of the second child in quarters

Question 2
Use the variable **ageq2nd** and the variable **ageq3rd** to construct an indicator variable on whether there was a multiple pregnancy during the mother's second pregnancy. What is the proportion of households with a multiple pregnancy in the second pregnancy?

Please round your answer to the fourth decimal place, i.e. if your answer is 0.12435, please round to 0.1244, and if it is 0.12433, please round to 0.1243.

Question 3
Use the variable **sex1st** and **sex2nd** to construct an indicator variable on whether the first and second born children have the same sex. What is the proportion of households in which the first two children have the same sex?

Please provide your answer to the fourth decimal place, i.e. exactly how it appears in the output.

Now let's set up the model we want to estimate. In particular, we are interested in estimating the following equation:

$$labor\ supply_h = \alpha_0 + \alpha_1 \mathbf{1}_{3\ children_h} + \alpha_2 black\ mother_h + \alpha_3 hispanic\ mother_h + \alpha_4 other\ race_h + \varepsilon_h$$ **(equation 1)**

where $labor\ supply_h$ corresponds to a labor supply variable of the mother in household h, $\mathbf{1}_{3\ children_h}$ is an indicator on whether there are three children born in the households, and the other variables correspond to the race categories. Finally, ε_h corresponds to an error term.

Question 4
Run this model through OLS using whether the mom works and the number of weeks she works as the dependent variables. According to your estimates, which of the following statements are correct? Select all that apply.

- ☐ According to the OLS estimates, having a third child reduces the likelihood that the mother works by 8.39 percentage points.
- ☐ According to the OLS estimates, having a third child reduces the likelihood that the mother works by 3.94 percentage points.
- ☐ According to the OLS estimates, having a third child reduces the number of weeks a mother decides to work by 8.39 weeks.
- ☐ According to the OLS estimates, having a third child reduces the number of weeks a mother decides to work by 3.94 weeks.

Since fertility is an endogenous variable, we want to use the multiple pregnancy and the same sex variable as instruments for having three children in the household. We are going to estimate the first-stage using each variable separately. Run a regression for each of these instruments using the indicator of having three children as the dependent variable and controlling for the race of the mother.

Question 5
According to your estimates, by having a multiple pregnancy during the second pregnancy, by how many percentage points does the likelihood of having a third child increase?
Please round your answer to the second decimal place, i.e. if your answer is 51.2322, please round to 51.23, and if it is 51.2382, please round to 51.24.

Question 6
According to your estimates, when the first two children are of the same sex, by how many percentage points does the likelihood of having a third child increase?
Please round your answer to the third decimal place, i.e. if your answer is 51.2322, please round to 51.232, and if it is 51.2382, please round to 51.238.

Question 7
Now, run the IV regression using whether the mother works as the dependent variable and multiple pregnancy as the instrument. According to this model, by how many percentage points does the likelihood that the mother works change when a third child is born?
Note: if it decreases by 7.25%, input -7.25. If it increases by 7.25%, input 7.25. Please round your answer to the third decimal place, i.e. 51.2322, please round to 51.232, and if it is 51.2382, please round to 51.238

Question 8
Now, run the IV regression using whether the mother works as the dependent variable and same-sex variable as the instrument. According to this model, by how many percentage points does the likelihood that the mother works change when a third child is born?
Note: if it decreases by 7.25%, input -7.25. If it increases by 7.25%, input 7.25. Please round your answer to the third decimal place, i.e. 51.2322, please round to 51.232, and if it is 51.2382, please round to 51.238

Question 9
As you should see, the following relationship holds between the point estimates of the three strategies that we have used: $\hat{\alpha}_1^{IV-multiple} \leq \hat{\alpha}_1^{OLS} \leq \hat{\alpha}_1^{IV-same\ sex}$. Assuming a model of heterogeneous effects, what might explain these differences?

- Women whose first two children are of the same sex are very different from women whose first two children are of different sex.
- Fertility doesn't seem to be a relevant variable when women take labor supply decisions.
- IV estimates are local treatment effects. Thus, we are identifying the effect of fertility over women who have a third child when the relevant instrument changes.
- Women with a multiple pregnancy in the second pregnancy are very different than women with no-multiple pregnancy.
- The instruments seem to be not valid since they show an opposite sign of the bias.

Experimental Design: Questions 10 – 17
During the lecture, Prof. Duflo discussed that thinking clearly about experimental design allows us to identify parameters beyond treatment effects, for example, General Equilibrium Effects as in the French Unemployment experiment. Another potential advantage of designing carefully experiments is the identification of potential mechanisms that drive a causal relationship. In this set of questions, we are going to discuss the identification of mechanisms. We are going to study Bursztyn et al.'s (2014) article *"Understanding Mechanisms Underlying Peer Effects: Evidence from a Field Experiment on Financial Decisions"*

For now, assume you are interested in establishing whether there is social influence on financial decisions, and that you have the following experimental design:

- You start your research project partnering with a financial company.
- You identify investor pairs using referrals to a financial company. One of the investors referred the other one to the company.
- You randomize among the pair who is investor number 1 and who is investor number 2.
- You offer to one of the investors (number 1) the possibility of purchasing a new financial asset.

- When you offer the financial asset to the second investor (number 2), you randomize whether or not you share the decision of the first investor.

Using this experimental design, you decide to estimate the following model:

$$decision_p = \beta_0 + \beta_1 information_p + \varepsilon_p \textbf{ (equation 4)}$$

where $decision_p$ is a dummy variable that indicates whether investor 2 in the pair p takes the same decision as her peer; $information_p$ indicates whether pair p elongs to the treatment group and investor 2 received information on the decision of investor 1; finally, ε_{ij} is an error term.

Question 10
Does this experimental design allow you to identify the causal effect of what peers do on financial decisions?

- Yes
- No

A researcher points out that equation 4 is not exploiting all the information in the data. She suggests that you can estimate the following model, which will allow you to identify not only the causal effect of knowing the peer's decision, but also the causal effect of having a peer who doesn't purchase the asset:

$$purchase_{p2} = \beta_0 + \beta_{1 purchase_{p1}} + \beta_2 information_p + \beta_3 purchase_{p1}\ x\ information_p + \varepsilon_p \textbf{ (equation 5)}$$

where $purchase_{p2}$ is a dummy variable that indicates whether investor 2 in pair p purchased the asset; $purchase_{p1}$ indicates whether investor 1 purchased the asset; $information_p$ indicates whether the pari p belongs to the treatment group of sharing information; $purchase_{p1}\ x\ information_p$ is the interaction; finally, ε_p is an error term.

Question 11
Which parameter allows you to identify the causal effect of having a peer who doesn't purchase the asset?

- β_0
- β_1
- β_2
- β_3
- It is not possible to tell in this setting.

Question 12
Which parameter allows you to identify heterogeneous effects of social influence by investor's 1 decision (whether she decided to purchase the asset or not)?

- β_0
- β_1

- β_2
- β_3
- It is not possible to tell in this setting.

Economic theory has identified two potential mechanisms of social influence on financial decisions. When someone learns that her peers have purchased an asset, she can be influenced via:

1. **Social learning:** she learned some information of the asset via the decision of her peers.
2. **Social utility:** she is influenced by the fact that her peers hold the asset, even under a setting where information remains constant.

Question 13

Instead of estimating the model in equation 5, you could use the following one:
$purchase_{p2} = \beta_0 + \beta_1 no\ purchase_{p1} + \beta_2 information_p + \varepsilon_{ij}$ (equation 6)
where $no\ purchase_{p1}$ indicates whether investor 1 of pair p declined to purchase the asset.

Would any of the models given by equations 4, 5 or 6 allow you to separately identify the channels of social learning and social utility?

- Yes
- No
- I can't tell from the given information.

Bursztyn et al. (2014) conduct an experiment in which they precisely try to separately identify these channels. Figure 1 presents the experimental design of their paper. Here is a brief summary of their experimental design:

(a) Partner with a financial company.

(b) Identify peer-pairs of investors using referrals to a financial company.

(c) Randomize who is investor 1 and investor 2 in each pair.

(d) Offer investor 1 the possibility of participating in a lottery to purchase a new financial asset.

(e) On those pairs in which investor 1 decided to participate in the lottery, randomize whether she can or can't purchase the asset.

(f) On the pairs in which investor 1 couldn't purchase the asset, randomize whether investor 2 learns the decision of her peer:

- No information **(group A).**
- Information that individual 1 decided to participate in the lottery and was unsuccessful in purchasing the asset **(group B).**

(g) On the pairs in which investor 1 could purchase the asset, randomize whether investor 2 learns the decision of her peer:

- No information **(group A)**
- Information that individual 2 decided to participate in the lottery and was successful in purchasing the asset **(group C)**.

(h) Have an additional group of individuals with no information: investors 2 in pairs in which investor 1 declined to purchase the asset **(group Aneg).**

(i) Their main outcome is whether investor 2 decides to purchase the asset or not.

If you want more information on this paper, view it on WorldCat.

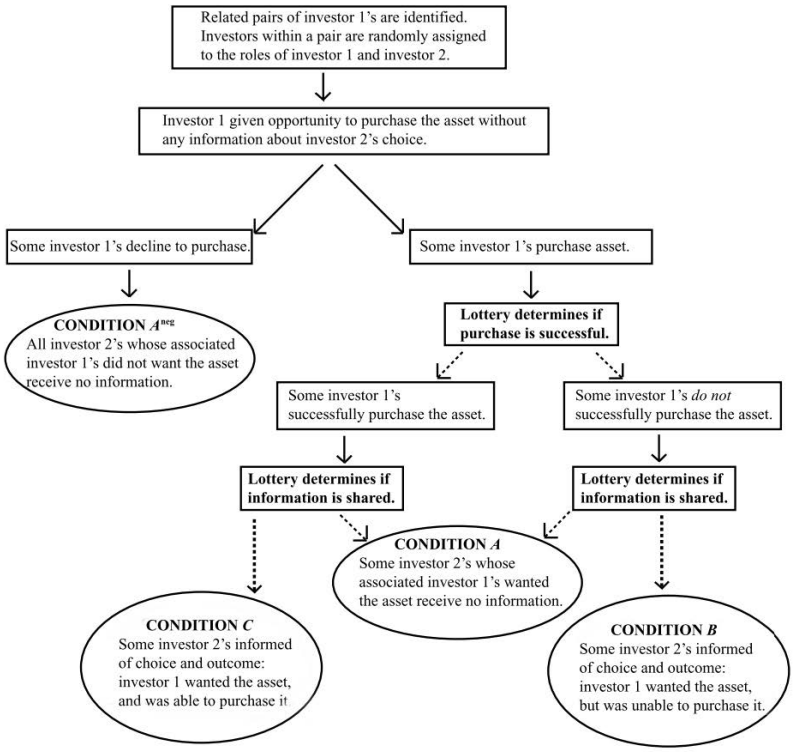

FIGURE 1.—Experimental design "roadmap."

Question 14

Which comparison between these groups will correspond to the treatment effect of social influence (social learning + social utility) in equation 6?

- Group A vs. Group A^{neg}
- Group C. vs Group B

- Group B vs. Group A
- Group C vs. Group A
- It is not possible to tell with this experimental design.

Question 15
Which comparison between these groups will correspond to the treatment effect of social learning without social utility?

- Group A vs. Group A^{neg}
- Group C. vs Group B
- Group B vs. Group A
- Group C vs. Group A
- It is not possible to tell with this experimental design.

Question 16
Which comparison between these groups will correspond to the treatment effect of social utility conditional on social learning?

- Group A vs. Group A^{neg}
- Group C. vs Group B
- Group B vs. Group A
- Group C vs. Group A
- It is not possible to tell with this experimental design.

Question 17
Which comparison between these groups will correspond to the treatment effect of social utility without social learning?

- Group A vs. Group A^{neg}
- Group C. vs Group B
- Group B vs. Group A
- Group C vs. Group A
- It is not possible to tell with this experimental design.

14.310x Data Analysis for Social Scientists
Spring 2023