651 Empirical Economics: Assignment 3

Per Hjertstrand[*] Andrew Proctor[†]

September 2017

Instructions

The assignment gives a maximum of **8 points** (but should take no longer than previous assignments). Answers should be provided in English. The key to a good score is to be clear and precise in your answers. For theory questions, be sure to show your work.

Answers to the written exercises should either be typed or, if handwritten, (legibly) scanned. Pictures of handwritten exercises taken from a camera (or smartphone) will not be accepted. Answers to the Stata-based exercises should be submitted in the form of a well-commented Stata **do-file and log file**. Your do-files should contain your code, output, and discussion of results (do not provide discussion in a seperate document. Please save your log files in the *text* format.

Finally, for this assignment, you will be expected to submit regression output tables (using the **estout** package) in a word document).

[*] Per.Hjertstrand@ifn.se

[†] Andrew.Proctor@phdstudent.hhs.se

Part A: Theory Exercises

Solve the following problems in Wooldridge (questions are from the 6^{th} edition. Please let Andrew know if you need access to the questions).

- **Chapter 15:** Problems 1, 2 3, 7 & 8

And the following (modified from Stock and Watson's *Introduction to Econometrics*):

SW Problem 1 Consider a Instrumental Regression model specified as:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 W_i + u_i \ ,$$

where X_i is correlated with u_i, W_i is uncorrelated with u_i, and Z_i is an instrument.

Which IV assumptions are violated when:

a Z_i is independent of (Y_i, X_i, W_i)?

b $Z_i = W_i$

c $W_i = 1 \ \ \forall i$

d $Z_i = X_i$

SW Problem 2 Consider the simple linear regression model

$$Y_i = \alpha + \beta X_i + u_i$$

Suppose MLR Assumptions 1-4 hold. Show that:

a X_i is a valid instrument.

b X_i is a relevant instrument.

c The IV estimator constructed using $Z_i = X_i$ is identical to the OLS estimator.

Part B: Stata Exercises

For this exercise, you will be replicating a classic research article using instrumental variables regression: Joshua Angrist and William Evans, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size" (*American Economic Review, 1998*).

In this paper, Evans and Angrist are interested in the effect of children on parents' labor supply. To overcome likely endogeneity of having children to a number of other possible determinants of labor supply, Angrist and Evans look at parents with at least two children and use whether or not the first two children were of the same sex as an instrument for the decision to have a third child.

Your task will be to replicate their results and interpret them. In particular, you will be replicating the descriptive statistics in columns 1-2 of Table 2, and the OLS and IV regression estimates in Table 7 (columns 1-2 & 4-5).

To download their initial dataset, go to: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/11288`. From there, download the file "m_d_806.tab". This is a *tab-delimited* file with data from the 1980 Census.

Using this dataset:

1 **Data Preparation:**

(a) Import the dataset.

(b) Format the data:

i. Rename the variables below as follows:

Original Name	**New Name**
yobm	yob_moth
agem	age_mother
weeksm	wks_wrked_moth
weeksd	wks_wrked_fath
ageqk	ageof1stchild
hoursd	hourswked_fath
hoursm	hourswked_moth
income1m	labinc_moth
income2m	selfempinc_moth
income1d	labinc_fath
income2d	selfempinc_fath

ii. Label the variables below as follows:

Variable	**Label**
wks_wrked_moth	“Weeks worked (moth)"
wks_wrked_fath	“Weeks worked (fath)"
labinc_moth	“Mother’s labor income"
selfempinc_moth	“Mother’s self-employment income"
labinc_fath	“Father’s labor income"
selfempinc_fath	“Father’s self-employment income"
age_mother	“mother’s age"
hourswked_moth	“Mother’s hours worked"

In addition, assign a descriptive label to any new variables you create in the remainder of your do-file.

iii. Use suitable commands to identify any string variables in the dataset and then identify any non-numeric values of these variables (which are indicative of missing values). Recode any of these texts ’missing values’ so that it conforms with Stata coding of missing values (this is a good opportunity for a *foreach* loop.

iv. For the *agemar* variable, recode any ’0’ values to missing.

(c) Angrist and Evans are interested in estimating the effect of children on their parents’ labor supply for two samples, (1) all women aged 21-35, and (2) only those women aged 21-35 who were married when giving birth. To determine whether or not parents were married when giving birth, we will need to create some variables related to timing of birth and marriage.[1]

i. First, subtract 1 from the qtrmar variable if the value for this variable is both not missing and greater than zero.

ii. Next, create a new “year married” variable.
   - This should be equal to *yob_moth* plus *agemar* if *qtrbthm* is less than or equal to *qtrmar.*
   - If *qtrbthm* is greater than *qtrmar*, this variable should be equal to *yob_moth* plus *agemar* plus 1.

iii. Next, create a more detailed timing of marriage variable, equal to the year and quarter of marriage. To do this, the new marriage timing variable should be equal to your year marriaged variable plus ($qtrmar/4$).

iv. Now, create a similarly detailed timing variable for when parents had their first child. To do this, the timing of first birth variable should be equal to $yobk + (qtrbkid/4)$.

v. Finally, create an indicator variable for whether or not parents were married when giving birth to their first child.

[1] Make sure you are handling missing values of variables appropriately throughout

- This variable should initially be equal to zero.
- Assign it the value of 1 if the detailed timing variable for marriage is greater than the detailed timing variable for birth of the first child.

(d) Create indicators for sex of children to be used in the analysis.

i. In the dataset, *sexk* refers to the sex of the first child and *sex2nd* refers to the sex of the second child, with values of 0 indicating male and 1 indicating female. Use this to create an indicators for:

- The first child is a male.
- The second child is a male.
- Whether the first two children were male.
- Whether the first two childrens were female.
- Whether the first two children were of the same sex.

ii. Finally, generate an indicator variable for whether the parents had more than 2 children using information about the total number of children, found in the *kidcount* variable.

(e) Create indicators for race/ethnicity of the mother using the *racem* variable.

- Create an indicator for whether the mother is black/African-American (which is denoted by a value of 2 in the *racem* variable).
- Create an indicator for whether the mother is Hispanic/Latino (which is denoted by a value of 12 in the *racem* variable).
- Create an indicator for whether the mother is white (which is denoted by a value of 1 in the *racem* variable).
- Create an indicator for whether the mother is another race/ethnicity. This should be equal to 1 minus the values for the black indicator, Hispanic indicator, and white indicator.

(f) Create the labor supply variables to be used as dependent variables in the analysis:

i. Because we want to estimate the effect of having children on income, we will need to find an appropriate inflation factor to make the earnings reported be interpretable in 2017 US Dollars.

- To do this, go to the Bureau of Labor Statistics website, and select retrieve data for "CPI for All Urban Consumers (CPI-U) 1982-84=100."
- In the table that follows, then modify the "From" value on the year range to include 1979 (the year for which earnings were reported in this dataset). Also select the option, "include annual average."

    - Calculate the inflation factor as the 2017 “Half1” (ie first half of the year) price index, divided by the 1979 “Annual” price index.

  ii. For each the father and the mother (i.e. 2 different variables), create an indicator variable equal to 1 if the respective parent worked at least 1 week (based on the values from *wks_wrked_moth* and *wks_wrked_fath* variables).

  iii. Create a new variable equal to the total income for a “mother” by adding *selfempinc_moth* to *labinc_moth*, if *selfempinc_moth* is positive. If *selfempinc_moth* is not positive, mothers’ total income should be equal to only the labor income.

  iv. Create a total income variable for fathers using the same procedure.

  v. For each mother’s total income and father’s total income, convert these values to 2017 USD by multiplying the variables by the inflation factor you derived above.

  vi. Create a new variable equal to family income in 2017 USD by setting it equal to *faminc* times the inflation factor.

  vii. Now make a new variable equal to the log of family income. More specifically, this should be equal to the log of the max between (1) family income in 2017 USD and (2) 1. In this manner, the minimum that logfamilyincome can achieve is 0.

  viii. Finally, create a variable equal to family income minus mother’s labor income.

    - To do this, set the new variable equal to family income in 2017 USD (not the log-version) minus (*labinc_moth* * the inflation factor).
    - Once again transform this variable into logged income using the same procedure as in (vii).

(g) Now define the samples in the analysis:

  i. To do this, we first need to generate a couple more age variables for the parents. First create a year of birth variable for the father.

    - If *qtrbthd* indicates the father was born in ‘Q0’, set the year of birth for the father equal to 80 - *aged*.
    - Otherwise, set the year of birth to 79 - *aged*.

  ii. Then create a variable for age in quarters for each the mother and the father:

    - The mother age in quarters variable should be equal to: 4 * (80 - *yob_moth*) - *qtrbthm* - 1.

- For father age in quarters variable should be equal to: 4 * (80 - [year of birth father] - *qtrbthd*,
  where you should replace [year of birth father] with the name of your variable for this data.

iii. Now create variables for the age at the birth of the first child, for each the mother and the father (that is, again two separate variables).

- Using the **floor** command, set the age at first birth equal to: the floor (rounded down number of ((\[age] - $ageof1stchild)/4)$, where the age of the mother is indicated by *ageqm* and age of the father is indicated by *ageqd*.

iv. Create a "main sample" indicator variable.

- Set this variable equal to 1 if all of the following condition's are met: mother's age is in the range 21 to 35, *kidcount* is greater than or equal to 2, $ageq2nd$ is greater than 4, age at first birth for the mother is greater than or equal to 15, and all of the following variables are equal to zero: $asex$,$aage$,$aqtrbrth$,$asex2nd$,$aage2nd$, $aqtrbrth$.

v. Create a "married sample" indicator variable.

- Set this variable equal to 1 if all of the following condition's are met: the "main sample" indicator is equal to 1, the value of the *aged* variable is not missing, *timesmar* is equal to 1, *marital* is equal to zero, the "unmarried birth" indicator is equal to zero, and finally the "age at first birth" variables are greater than or equal to 15 for both the mother and the father.

2 **Exploratory Analysis:**

(a) Using the the dataset you have prepared (and adequately labeled), generate the summary statistics for columns 1 and 2 of Table 2 in Angrist and Evans (omitting the statistics for 'twins').

3 **Regression Analysis:**

(a) First, comment on why researchers like Angrist and Evans are concerned that a variable indicating "# of children" might be endogenous in an earnings or labor supply regression.

(b) Discuss why having two children of the same sex might work as an instrument for the decision for have additional children (you can consult Angrist and Evans' paper if you like, but you shouldn't need to do so).

(c) Run the OLS and IV regressions (including the "first stage") in Column 1-2 and 4-5 of Table 7 (using **eststo** to store you results).[2]

(d) Comment on what the F-test for the excluded instruments indicates about the instrument.

(e) Fully interpret the estimates for the main explanatory variable in each of the regressions of column (2).

(f) Comment on how your results change from the OLS to IV specifications (i.e. in terms of magnitude and significance).

(g) Should one expect the over-identification test reported in the regression output to tell us anything about the over-identifying restrictions? Why or why not?

(h) Either in the same document as your theory answers, or in a separate document, present 4 regression tables using **esttab**.

- The first table should correspond to the regressions of Table 7 column 1 (but including all the explanatory variables).
- The second table should correspond to the regressions in column (2).
- The third table should correspond to the regressions in column (4).
- And the fourth table should correspond to the regressions in column (5).

4 For three person groups only:

(a) Also perform the regressions corresponding to columns (3) and (6) in Table 7 (i.e. repeat the IV regressions using as instruments both the "first two children males" and "first two children females" variables.

(b) What are the potential advantages of having multiple instruments in IV regression?

(c) What is a potential concern with having multiple instruments in an IV regression?

(d) What do the results of the F-test for the excluded instruments indicate about the IV strategy here?

(e) What do the results for the over-identification test indicate about the IV strategy here?

Good luck!

[2] Be sure to use realistic assumptions about the error structure in your regression code.