ECO 102: Topics in Economics
Ecole Polytechnique, Spring 2022
Professors: Geoffroy Barrows, Benoit Schmutz
Teaching Assistants: Arnault Chatelain, Maddalena Conte

TD 2 (part 2): labor force participation and IV

In this TD we will use IV to estimate the impact of family size on the labor force participation of women. As usual, download the folder TD_2_2.zip, write your answers directly in the Tex script and type your Stata commands in a do file.

Exercises

1. A refresher on IVs:

   (a) What is an Instrumental Variable? Why do we use them?
   Answer: When the $E[\epsilon|X] \neq 0$ which might lead to the biased result of OLS.Therefore, in order to solve this problem we introduce a IV, A valid instrument induces changes in the explanatory variable but has no independent effect on the dependent variable, allowing a researcher to uncover the causal effect of the explanatory variable on the dependent variable.

   Correction: It is a variable used to correct endogeneity issues with one of the explanatory variable. We use them to get correct estimates when we have an endogeneity issue.

   (b) What are the two conditions that should be met when using an IV?
   Answer:
   1. The instrument must be correlated with the endogenous explanatory variables
   2. The instrument cannot be correlated with the error term in the explanatory equation

2. Suppose that we have data on households, family members' labor force participation, and family composition. We would like to test if family size has an effect on mothers' labor force participation.

   (a) Write down a standard regression equation to identify this effect. Why is estimation by OLS likely to be biased?
   Answer: We applied the method on $LP = \beta_1 + \beta_2 * gender + b * family - size + u$ or we can also apply the model $LP = \alpha_1 + \alpha_2 * gender + \alpha_3 * family - size + \alpha_4 * gender * family - size + u$. LP means labor participation.
   We focus on the last model. The family-size is endogenous since we might have omitted variables and reverse causality.

   (b) What conditions would an IV need to satisfy in this case?
   Answer: The IV should have strong correlation with family-size and not related to error term.

(c) Would a variable indicating if the mother had twins after the first child be a good IV? Why?
Answer: This is good IV since it have strong correlation with family-size and not related to error term.

(d) Would a variable indicating if the mother had first two children of the same sex be a good IV? Why?
Answer: No it's a good IV since it's correlated to the family-size and since it's random which implies that this is not related to the error term.

3. Preparing the dataset. We will use data from the French Household Survey (Enquête Logement) for 2013.

(a) Use the household-dwelling file (*menlog*) and merge it with the individual file (*individu*). Hint: which variable identifies each observation in the master dataset? Note: the *Variables.pdf* document provides information on all variables in the datasets.

(b) Keep single-households dwellings (variable *nmen*). Among these, keep couples with their children (variables *mty1a* and *nlien*). Among these, keep families with 2 or more children (variable *mne*).
Explanation of the variables:
nmen = Number of households that actually live in the dwelling (from 1 to 5)

mty1a = type of household
1-one person living along
2-Multi-person household with no family
3-The main family is a single parent
4-The main family consists of a couple

nlien = Relationship of the individual with the reference person of the dwelling
1-Reference person
2-Reference person's spouse, married or common-law (the woman in the couple)
3-Child of the reference person or his or her spouse
4-Grandchild of the reference person or his or her spouse
5-Ascendant of the reference person or his or her spouse
6-Other relative of the reference person or his/her spouse
7-Friend of the reference person
8-Other non-family relationship (boarder, sub-tenant, lodger, servant, ...)

mne = Number of dependent children in the household
0 to 9-number of children
10-more than 10

(c) Identify twins among kids based on their age.

- Use both age variables (*nag* and *nag1*) to avoid rounding error and create an *age_kid* variable which identifies the age of children in the household (hint: use the variable *nlien* to identify children).
Explanation of the variables:
nag = Age of the individual at December 31, 2013 (from 0 to 140)
nag1 = Age of the individual at the date of servey (from 0 to 140)

- For each family and age of children, identify twins, i.e. duplicates by age of children. Hint: use the variable *idmen* to identify each family. Use the Stata function *cond* to identify duplicates.
  Explanation of the variables:
  idmen = Diffuse household identifier

(d) Identify families with same gender of first two kids.

- What is the intuition here?
  Answer: The percentage of first two children of boys or girls should be around 25% and the percentage of first two children have the same gender should be 50%
  Which we can also verify from the data (0 means all boys and 2 means all girls)

mgirl_12	Freq.	Percent	Cum.
0	5,765	28.19	28.19
1	9,971	48.76	76.96
2	4,712	23.04	100.00
Total	20,448	100.00	

Figure 1: Percentage of the first two children with same gender

- Hint: sort family members by age, identifying their order by creating a variable *nobs*. Identify the youngest 2 children, and whether their sex differs (use variable *nsexe*).
  Explanation of the variables:
  nsexe = the sex of individuals
  1-male
  2-female

(e) Identify families with twins at first birth.

- Hint: which conditions on *dup* and *nobs* do we need to identify twins at first birth?
- Tag all family members in families with twins at first birth. Hint: use the Stata *sum* function.
- Drop these families.
- Why do we do this?
  Answer: Cause our IV is the mother had twins after the first child so we have to drop the data points with the first child is twins.

(f) Identify families with twins after first birth.

- What is the intuition here?
  Answer: The size of data is larger compare with the twins at first birth.

- Hint: which conditions on *dup* and *nobs* do we need to identify twins after first birth?
- Tag all family members in families with twins after first birth. Hint: use the Stata *sum* function.

(g) Keep mothers in couples with men. Use *nlien* to identify parents in the household. Use *nsexe* to identify mothers.

(h) Keep couples with working ability: minimum and maximum age in the household must best 15 and 61 years respectively.
Keep the couples with working capacity.

(i) Generate the following variables

- *large_family*: for families with more than 2 children. Hint: use variable *mne1*.
- *mlarge*: interaction between *mother* and *large_family*.
- *mhh_twin*: interaction between *mother* and *hh_twin*.
- *msame_gender*: interaction between *mother* and *same_gender*.
- *non_work*: individuals that are not working, based on variable *ntravail*.
- *non_work_extended*: individuals working part-time, based on variables *ntpp* and *non_work*.
- *high_diploma*: individuals with high-school diploma, based on variable *ndiplo*.
- *ln_age*: natural logarithm of age, based on variable *nag*.

4. The variable *non_work_extended* will be our main outcome variable. Use t-tests to identify if there is a significant difference in *non_work_extended* between the various groups of interest (mothers versus fathers, large versus small families).

(a) What is a t-test? Why do we need a two-sample t-test in this case?
Answer: T-test, which also calls Student's t-test. We often use t-tests to infer differences in overall means or between overall means from small samples. T-test here is a way to test whether we can reject that the coefficient is equal to 0.

The two-sample t-test (also known as the independent samples t-test) is a method used to test whether the unknown population means of two groups are equal or not. we need two-sample t-test to compare two groups (e.g. mother versus father), and-study there difference in this question.

(b) Hint: use the Stata function *ttest* with *by* (type *help ttest* to explore the options available).

(c) What other options must we add if we want to compare different groups? Why?
Answer: We need to fix one condition, for example we can first set family scales to be large and run t-test on mother versus father, and then set family scales to be small and run the t-test again.

This is to exclude the effect of the other factor

(d) How do we interpret t-test outputs? What do these t-tests suggest?
From the ttest, we see that for fathers, family scale does not make a much difference on labor force participation. For mothers, larger family scale indicates a lower labor participation. Yet no matter what family scale it is, female's labor participation is always lower than male's.

5. We now run OLS regressions:

(a) Regress *non_work_extended* on *mother*, *large_family*, *mlarge*. How do we interpret these results? Why is it likely that the coefficient of interest is biased?

It suggests that having a large family and being a mother increases the probability of not working by 13%. This is likely to be biased since there likely to have omitted variables such as age differences.

Source	SS	df	MS	Number of obs	=	8,834
				F(3, 8830)	=	109.87
Model	65.7573209	3	21.919107	Prob > F	=	0.0000
Residual	1761.57616	8,830	.199498999	R-squared	=	0.0360
				Adj R-squared	=	0.0357
Total	1827.33348	8,833	.206875748	Root MSE	=	.44665

non_work_e~d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mother	.1317347	.0111472	11.82	0.000	.1098836	.1535857
large_family	-.0032333	.0150848	-0.21	0.830	-.0328031	.0263364
mlarge	.1045837	.0213332	4.90	0.000	.0627658	.1464017
_cons	.2130178	.0078823	27.02	0.000	.1975667	.2284688

Figure 2: Regression of *non_work_extended* on *mother*, *large_family*

(b) Add some additional control variables (*ln_age*, *high_diploma*). What do you see?
By adding some variables, we can still get that having a large family and being a mother increases the probability of not working by 13%, since this time we add *ln_age*, *high_diploma* variables.

Source	SS	df	MS	Number of obs	=	8,834
				F(5, 8828)	=	110.16
Model	107.316489	5	21.4632978	Prob > F	=	0.0000
Residual	1720.017	8,828	.194836542	R-squared	=	0.0587
				Adj R-squared	=	0.0582
Total	1827.33348	8,833	.206875748	Root MSE	=	.4414

non_work_e~d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mother	.1338593	.0111773	11.98	0.000	.1119491	.1557695
large_family	-.0056398	.0149134	-0.38	0.705	-.0348735	.0235939
mlarge	.0968249	.0210891	4.59	0.000	.0554853	.1381645
ln_age	-.1929583	.0258833	-7.45	0.000	-.2436955	-.1422211
high_diploma	-.1229138	.0095271	-12.90	0.000	-.1415892	-.1042385
_cons	.9986777	.0976566	10.23	0.000	.807248	1.190107

Figure 3: Regressions with additional control variables (*ln_age*, *high_diploma*)

(c) Run this same linear regression, but absorb household fixed effects. Hint: use the Stata command *reghdfe*, with the option *absorb*. Why might we want to add fixed effects in this case? What do you see?
Adding fixed effect can help to control for omitted variable bias due to unobserved heterogeneity across households,for example sorting of couples.
We see that being a mother increases the probability of not working by 11 percent.

6. We now run IV regressions. Use the Stata command *ivreghdfe*, with the options *absorb* and *first* (the latter option reports the individual first-stage regressions separately).

HDFE Linear regression — Number of obs = 8,834
Absorbing 1 HDFE group — F(4, 4413) = 137.01
Prob > F = 0.0000
R-squared = 0.7089
Adj R-squared = 0.4173
Within R-sq. = 0.1105
Root MSE = 0.3472

non_work_e~d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mother	.1154929	.0096823	11.93	0.000	.0965107	.1344751
mlarge	.0995697	.0166063	6.00	0.000	.0670129	.1321264
ln_age	-.3446347	.0650957	-5.29	0.000	-.4722549	-.2170146
high_diploma	-.0442253	.0141325	-3.13	0.002	-.0719322	-.0165185
_cons	1.524657	.2438769	6.25	0.000	1.046536	2.002778

Absorbed degrees of freedom:

Absorbed FE	Categories	- Redundant	= Num. Coefs
idmen	4417	0	4417

Figure 4: IV regression

(a) Instrument *mlarge* with *mhh_twin*. Look at results for the first stage: what do they tell you? Look at main results (IV 2SLS estimation): what do they tell you?

(b) Instrument *mlarge* with *mhh_twin* and *msame_gender*. Look at results for the first stage: what do they tell you? Look at main results (IV 2SLS estimation): what do they tell you?

7. (Optional) Why might these IVs not work?