In-Class Lab 12

ECON 4223 (Prof. Tyler Ransom, U of Oklahoma)

October 20, 2020

The purpose of this in-class lab is to use R to practice with instrumental variables estimation. The lab should be completed in your group. To get credit, upload your .R script to the appropriate place on Canvas.

For starters

You may need to install the packages `AER`, `flextable` and `modelsummary`. (`AER` may have already been installed when you previously installed `car` and `zoo`.)

Open up a new R script (named `ICL12_XYZ.R`, where `XYZ` are your initials) and add the usual "preamble" to the top:

```
# Add names of group members HERE
library(tidyverse)
library(wooldridge)
library(broom)
library(AER)
library(magrittr)
library(modelsummary)
library(vtable)
```

Load the data

We're going to use data on fertility of Botswanian women.

```
df <- as_tibble(fertil2)
```

Summary statistics

Let's look at summary statistics of our data by using the `vtable` package. We can export this to a word document format if we'd like:

```
df %>% sumtable(out="return")
```

```
##    Variable    N   Mean Std. Dev. Min Pctl. 25 Pctl. 75  Max
## 1  mnthborn 4361  6.331     3.323   1        3        9   12
## 2  yearborn 4361 60.434     8.683  38       55       68   73
## 3       age 4361 27.405     8.685  15       20       33   49
## 4  electric 4358   0.14     0.347   0        0        0    1
## 5     radio 4359  0.702     0.458   0        0        1    1
## 6        tv 4359  0.093      0.29   0        0        0    1
## 7   bicycle 4358  0.276     0.447   0        0        1    1
## 8      educ 4361  5.856     3.927   0        3        8   20
```

```
## 9        ceb 4361  2.442     2.407   0         1         4   13
## 10 agefbrth 3273 19.011     3.092  10        17        20   38
## 11 children 4361  2.268     2.222   0         0         4   13
## 12 knowmeth 4354  0.963     0.188   0         1         1    1
## 13  usemeth 4290  0.578     0.494   0         0         1    1
## 14  monthfm 2079   6.27      3.62   1         3         9   12
## 15   yearfm 2079 76.912      7.76  50        72        83   88
## 16    agefm 2079 20.686     5.002  10        17        23   46
## 17 idlnchld 4241  4.616     2.219   0         3         6   20
## 18    heduc 1956  5.145     4.803   0         0         8   20
## 19    agesq 4361 826.46   526.923 225       400      1089 2401
## 20    urban 4361  0.517       0.5   0         0         1    1
## 21 urb_educ 4361  3.469     4.294   0         0         7   20
## 22   spirit 4361  0.422     0.494   0         0         1    1
## 23  protest 4361  0.228     0.419   0         0         0    1
## 24 catholic 4361  0.102     0.303   0         0         0    1
## 25 frsthalf 4361   0.54     0.498   0         0         1    1
## 26    educ0 4361  0.208     0.406   0         0         0    1
## 27 evermarr 4361  0.477       0.5   0         0         1    1
```

1. What do you think is going on when you see varying numbers of observations across the different variables?

Determinants of fertility

Suppose we want to see if education causes lower fertility (as can be seen when comparing more- and less-educated countries):

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

where *children* is the number of children born to the woman, *educ* is years of education, and *age* is age (in years).

2. Interpret the estimates of the regression:

```
est.ols <- lm(children ~ educ + age + I(age^2), data=df)
```

(Note: include `I(age^2)` puts the quadratic term in automatically without us having to use `mutate()` to create a new variable called `age.sq`.)

We can also use stargazer to examine the output. It puts the standard errors of each variable in parentheses under the estimated coefficient.

```
modelsummary(est.ols)
```

	Model 1
(Intercept)	-4.138
	(0.241)
age	0.332
	(0.017)
educ	-0.091
	(0.006)
I(age^2)	-0.003
	(0.000)
Num.Obs.	4361
R2	0.569
R2 Adj.	0.568
AIC	15681.2
BIC	15713.1
Log.Lik.	-7835.592

Instrumenting for endogenous education

We know that education is endogenous (i.e. people choose the level of education that maximizes their utility). A possible instrument for education is $firsthalf$, which is a dummy equal to 1 if the woman was born in the first half of the calendar year, and 0 otherwise.

Let's create this variable:

```
df %<>% mutate(firsthalf = mnthborn<7)
```

We will assume that $firsthalf$ is uncorrelated with u.

3. Check that $firsthalf$ is correlated with $educ$ by running a regression. (I will suppress the code, since it should be old hat) Call the output `est.iv1`.

IV estimation

Now let's do the IV regression:

```
est.iv <- ivreg(children ~ educ + age + I(age^2) | firsthalf + age + I(age^2), data=df)
```

The variables on the right hand side of the `|` are the instruments (including the x's that we assume to be exogenous, like age). The endogenous x is the first one after the `~`.

Now we can compare the output for each of the models:

```
modelsummary(list(est.ols,est.iv1,est.iv))
```

	Model 1	Model 2	Model 3
(Intercept)	-4.138	6.363	-3.388
	(0.241)	(0.087)	(0.548)
age	0.332		0.324
	(0.017)		(0.018)
educ	-0.091		-0.171
	(0.006)		(0.053)
I(age^2)	-0.003		-0.003
	(0.000)		(0.000)
firsthalfTRUE		-0.938	
		(0.118)	
Num.Obs.	4361	4361	4361
R2	0.569	0.014	0.550
R2 Adj.	0.568	0.014	0.550
AIC	15681.2	24249.6	
BIC	15713.1	24268.7	
Log.Lik.	-7835.592	-12121.779	

We can also save the output of `modelsummary()` to an image, a text file or something else:

```r
modelsummary(list(est.ols,est.iv1,est.iv), output="results.jpg")
```

4. Comment on the IV estimates. Do they make sense? Discuss why the IV standard error is so much larger than the OLS standard error.