In-Class Lab 11

ECON 4223

October 17, 2023

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 `ICL11_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)
```

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 `modelsummary` package. We can export this to a word document format if we'd like:

```
df %>% datasummary_skim(histogram=F,output="myfile.docx")
```

```
## [1] "myfile.docx"
```

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 `modelsummary` to examine the output. It puts the standard errors of each variable in parentheses under the estimated coefficient.

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

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))
```

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

	(1)	(2)	(3)
(Intercept)	−4.138	6.363	−3.388
	(0.241)	(0.087)	(0.548)
educ	−0.091		−0.171
	(0.006)		(0.053)
age	0.332		0.324
	(0.017)		(0.018)
	−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	15 681.2	24 249.6	15 864.3
BIC	15 713.1	24 268.7	15 896.2
Log.Lik.	−7835.592	−12 121.779	
F	1915.196	62.620	
RMSE	1.46	3.90	1.49

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

```
## save_kable will have the best result with magick installed.
```

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

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.