

## Lab 8: Instrumental variables

It's our final lab of the semester!

### MATERIALS

- `voucher.dta`
- Do-file template `labtemplate.do`

### OBJECTIVES

By the end of this lab, you should be able to complete the following tasks in Stata:

- Estimate instrumental variable specifications and interpret them.
- Output regression results using `outreg2`

### KEY COMMANDS

#### Conducting instrumental variables regressions with `ivregress`

We can estimate an instrumental variables regression with `ivregress`

General form:

```
ivregress estimator depvar [varlist1] (varlist2 = varlist_iv) [if] [in]
      [weight] [, options]
```

- `estimator` is where we will type `2sls`
- `depvar` is your dependent variable
- You can include other explanatory variables before or after the parentheses, '[varlist1]
- In the parentheses, write you endogenous ( $x$ ) then your instrument ( $z$ ) - these can be lists!
- The rest of it is just as you're used to

Example:

To estimate the following two-stage least squares equation:

$$rent = \beta_0 + \beta_1 \widehat{hsngval} + \beta_2 pcturban + u$$

where  $\widehat{hsngval}$  is predicted from the following first-stage equation

$$hsngval = \alpha_0 + \alpha_1 faminc + \alpha_2 pcturban + v$$

```
webuse hsng2
```

```
ivregress 2sls rent (hsngval = faminc ) pcturban, robust
```

You can add `, first` to report the first-stage results:

```
'ivregress 2sls rent (hsngval = faminc ) pcturban, robust first'
```

### Outputting your results with `outreg2`

We are very good at reading raw Stata output. But, raw stata output has no place in our papers. How do we make it pretty? There are lots of ways, including `putexcel`, which lets you create customizable excel tables with your outputs (good for descriptive statistics), and `estout`, which does the same thing but is more regression oriented.

Personally, I like `outreg2`, because it's easy to set up and use. So that's what we'll use!

`outreg2` is a user-created package, which means you have to install it:

```
ssc install outreg2
```

You only have to do this once.

You'll run `outreg2` after estimating a regression. It takes your results and saves them to a table. You can run it multiple time and generate columns of results within the same excel sheet, which is pretty handy! The general format of `outreg2` is this:

```
// You can copy and paste this into stata, and it should work!
//But note that it will save to your working directory
```

```
sysuse auto,clear
```

```
// Specification 1
regress mpg foreign weight headroom trunk length turn displacement
outreg2 using myfile.xls, replace
```

```
// Specification 2 (add on)
regress mpg foreign weight headroom trunk length turn displacement,robust
outreg2 using myfile.xls, append
```

You can customize, with lots of options! (see `help outreg2`, or check out [these resources](#))

What sort of things?

- Export directly to Word
  - `outreg2 using myfile, word replace`

- Add summary statistics and p-values
  - See [here](#) for more details
- Add notes
  - `outreg2 using myfile, addnote(Dummy variables not shown)`
- Report only some variables
  - `outreg2 using myfile, keep(mpg foreign)`
- Modify number of decimal places
  - `outreg2 using myfile, dec(5)`
- You can use a loop to make a whole set of columns!

An example:

```
sysuse auto,clear
local r "replace"
forval num=1/5 {
    regress mpg weight headroom if rep78=='num'
    sum mpg if rep78 == 'num'
    local mean = 'r(mean)'
    outreg2 using myfile.xls, 'r' keep(headroom) title("Sample Graph")
    nocons addtext("Rep78", 'num') addstat("Mean", 'mean') auto(2) bracket

local r "append"
}
```

## EXERCISES

Today we're going to work with [voucher.dta](#), a data set of student performance from Rouse (1998). She measures the impact of private school vouchers on student achievement. The final measure of student performance we're interested in is `mnce`, their math test scores in 1994 (after up to four years in the private school). We also have some measures of baseline performance, their math test score in 1990 (`mnce90`). The variable `choicelyrs` is the number of years enrolled in a private school, and `selectyrs` is the number of years a student was *selected* to receive a voucher to fund enrolling in a private school.

1. In your do-file, start a log and open `voucher.dta`.
2. Summarize your data. Of the 990 students in the sample, how many were never awarded a voucher? How many had a voucher for all four years? How many actually attended a choice school for four years?

3. Predict the relationship between choice school attendance and math scores by regressing math scores `mnce` (dependent variable) on number of years enrolled in a choice school `choicelyrs` (independent variable). What do you find? Is this what you expect? What happens if you add in the variables `black`, `hispanic`, and `female`? Write your results in equation form.
4. Why might `choicelyrs` be endogenous? Explain:
5. Now, estimate a regression of `choicelyrs` (dependent variable) on `selectyrs` (independent variable), including race/ethnicity and gender controls. Why is this a reasonable choice of an instrument? What is the F-statistic on `selectyrs`? (Hint: You can use the `testparm` command for a hypothesis test with just one coefficient)
6. Based on the previous regression, use the `predict` command to generate a predicted `choicelyrs`. Estimate the regression of `mnce` on `choicelyrs`, including race/ethnicity and gender controls. Write the estimated equation. How does your result compare to your OLS estimate?)
7. Re-estimate a regression of `mnce` (dependent variable) on `choicelyrs` (independent variable) using `selectyrs` as an instrument for `choicelyrs`. However, this time, estimate the equation in one command line using `ivregress 2sls`. How do your results change, if at all?
8. Repeat your IV analysis, but this time include a control for baseline achievement by adding `mnce90`. Write the results in equation form below. Do you find these results convincing? Explain.
9. We can also use multiple instruments for multiple endogenous variables. The variables `choicelyrs1`, `choicelyrs2`, etc. are dummy variables indicating the different number of years a student could have been in a choice school. Similarly, `selectyrs1`, `selectyrs2`, etc. have a similar definition, but for being selected from the lottery.

Estimate the following equation using IV.

$$mnce = \beta_0 + \beta_1 choicelyrs_1 + \beta_2 choicelyrs_2 + \beta_3 choicelyrs_3 + \beta_4 choicelyrs_4 + \beta_5 black + \beta_6 hispanic + \beta_7 female + \beta_8 mnce90 + u$$

10. Finally, go back through your regressions in your do-file. After each regression (there should be six: OLS without controls, OLS with controls, IV by hand, IV using `ivregress`, IV with `mnce90`, and IV with multiple instruments), add a line of code to output the results to a word or excel file using `outreg2`.

**Include a table with your results with your submission** - there should be six columns in one table. Note that you can use the `append` option to add each regression as a new column, rather than a new file.

References: Rouse, Cecilia Elena (1998), "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *The Quarterly Journal of Economics* 113(2), 553-602.