Stata How-to:
Instrumental Variables using 2SLS

2021-12-27

Contents

Imagine we would like to estimate the following model:

$$y_i = \alpha + \beta x_i + e_i \tag{1}$$

However, we suspect an endogeneity problem; in other words, that residuals residuals are systematically different for different values of the variable x, a case of selection bias. Technically, the conditional expectation of the residual, $\mathbb{E}[e_i | x_i]$, still depends on x_i even after conditioning on x_i, and would therefore be different for different values of x_i. We therefore know that OLS yields a biased estimate of the true causal effect of x on y.

To address this, assume we find a variable z that fulfills the 3 requirements to be used as an instrument for x.

1. Basic syntax

The command for performing a estimation based on instrumental variables using two-stage least squares is **`ivregress 2sls`**.

```
ivregress 2sls y (x = z), robust
```

where y is our dependent variable, x is our initial explanatory variable, and z is the variable we use to instrument for x.

If instead of one, we have *two* instruments z_1 and z_2 for variable x, we would instead run:

```
ivregress 2sls y (x = z1 z2), robust
```

It is important to specify **`2sls`** in the command, as they are other ways to perform an instrumental variables approach (beyond the scope of ECO372).

2. Additional controls

Naturally, we can also include control variables in the IV estimation. If residuals still depend on x after controlling for w_1 and w_2,[1] we would run:

```
ivregress 2sls y (x = z1 z2) w1 w2, robust
```

As with OLS regressions, we can easily turn a categorical variable into a series of dummies using the `i.` operator:

```
ivregress 2sls y (x = z1 z2) w1 w2 i.group, robust
```

It is also possible to add heterogeneous effects in an IV model, although interpretation becomes less straightforward (and beyond the scope of ECO372).

3. Example

Let's use the `auto`. We use it to see how car prices (`price`) depends on their range, measured in miles-per-gallon (`mpg`). (Cars with higher `mpg` can drive longer distances before having to refuel.)

```
sysuse auto, clear // Clears memory and loads a default dataset included in Stata
```

We could estimate directly the correlation between fuel consumption and price. However, we suspect that people who live far from their place of work have lower income (and therefore a lower willingness to pay), and at the same time like cars that have low fuel consumption (high `mpg`). That is, even when taking into account their miles-per-gallon, cars might be less expensive for high `mpg` because they attract a subpopulation of less wealthy customers. A clear case of selection bias.

Let's say we use the weight of the car as an instrument for their miles-per-gallon (lighter cars should use less fuel, and therefore travel more miles-per-gallon). This is unlikely to satisfy requirements 2 and 3 for an IV estimation, but let's ignore that for the sake of this example (it is however important to check when working with real data).

We would then run:

```
ivregress 2sls price (mpg=weight), robust
```

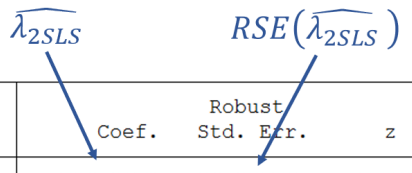
And if we still want to control for domestic versus foreign cars:

```
ivregress 2sls price (mpg=weight) i.foreign, robust
```

Stata then reports the two-stage least squares λ which, provided requirements for IV are satisfied, identifies the causal LATE of `mpg` on `price`.

[1] That is, if $\mathbb{E}[e_i \mid x_i, w_{1i}, w_{2i}]$ still depends on x_i.

```
Instrumental variables (2SLS) regression          Number of obs   =         74
                                                  Wald chi2(2)    =      31.66
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.1406
                                                  Root MSE        =     2715.8
```



price	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
mpg	-504.0671	94.02956	-5.36	0.000	-688.3616	-319.7725
foreign						
Foreign	2805.276	725.1719	3.87	0.000	1383.965	4226.587
_cons	16066.52	2021.968	7.95	0.000	12103.54	20029.51

```
Instrumented:  mpg
Instruments:   1.foreign weight
```

To check the strength of the instrument, we need to run the first stage seperately. This can be done as a separate regression (including the same controls):

```
regress mpg weight i.foreign, robust
```

or directly in the **ivregress** command, with the option **first**:

```
ivregress 2sls price (mpg=weight) i.foreign, robust first
```