Empirical Economics

Instrumental Variables Regression

Andrew Proctor

andrew.proctor@phdstudent.hhs.se

October 12, 2018

1. Econometrics Concepts
2. Stata Commands

Econometrics Concepts

The endogeneity problem and the IV solution

Suppose we want to estimate:

$$y_i = \alpha + \beta x_i + u_i$$

- But we know that x_i is endogenous (that is, $Cov(x_i, u_i) \neq 0$) and we can't reasonably find control variables to remedy the problem. What can we do?
- One possibility is look for an "instrument" variable z_i that only affects our outcome y_i through its effect on x_i. So that:
  - z_i **is a *relevant* instrument:** $Cov(z_i, x_i) \neq 0$
  - z_i **is an *exogenous* instrument:** $Cov(z_i, u_i) = 0$

The instrumental variables equations

The resulting model is then:

$$x_i = \pi_0 + \pi_1 z_i + \epsilon_i \qquad \textbf{(first stage)}$$

$$y_i = \alpha + \beta x_i + u_i \qquad \textbf{(structural equation)}$$

And another key equation of interest is the relationship between y_i and z_i

$$y_i = \gamma_0 + \gamma_1 z_i + e_i \qquad \textbf{(reduced form for y)}$$

Deriving the IV Coefficients

There are several ways you can derive the IV estimator. Probably the easiest is to apply $Cov(Z, \cdot)$ to the endogenous regression for y_i.

$$Cov(z_i, y_i) = Cov(z_i, y_i)$$

Substitute in equation for $y_i = \alpha + \beta x_i + u_i$ on the RHS:

$$Cov(z_i, y_i) = Cov(z_i, \alpha + \beta x_i + u_i)$$

$$Cov(z_i, y_i) = Cov(z_i, \alpha) + Cov(z_i, \beta x_i) + Cov(z_i, u_i)$$

Deriving the IV Coefficients II

Using the rules for covariance, we get:

$$Cov(z_i, y_i) = \beta Cov(z_i, x_i) + Cov(z_i, u_i)$$

Since we assume $cov(z_i, u_i) = 0$, then:

$$Cov(z_i, y_i) = \beta Cov(z_i, x_i)$$

$$\beta^{IV} = \frac{Cov(z_i, x_i)}{Cov(z_i, y_i)}$$

IV as the Ratio of OLS Coefficients for the 1st Stage and Reduced Form

Both the numerator and demoninator of β^{IV} look similar to the formula for the usual OLS coefficient:

- If the model is $y_i = \alpha + \beta x_i + u_i$, then $B^{OLS} = \frac{Cov(y_i,x_i)}{Var(x_i)}$

So we have the following models and OLS coefficients:

- Model 1: $y_i = \gamma_0 + \gamma_1 z_i + \varepsilon_i$, OLS Coefficient: $\gamma_1 = \frac{Cov(y_i,z_i)}{Var(z_i)}$
- Model 2: $x_i = \pi_0 + \pi_1 z_i + e_i$, OLS Coefficient: $\pi_1 = \frac{Cov(x_i,z_i)}{Var(z_i)}$

IV as the Ratio of OLS Coefficients for the 1st Stage and Reduced Form II

Now clearly, model 1 is the "reduced form" equation for y_i and model 2 is the "first stage".

And dividing the reduced form coefficient by the first stage coefficient, we get:

$$\frac{\gamma_1}{\pi_1} = \frac{Cov(y_i, z_i)/Var(z_i)}{Cov(x_i, z_i)/Var(z_i)}$$

$$\frac{\gamma_1}{\pi_1} = \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)}$$

$$\frac{\gamma_1}{\pi_1} = \beta^{IV}$$

Linking IV back to regression on the explanatory variable

Hopefully, you can now get a sense that doing OLS regression in two stages (the "first stage" and "reduced form") can get you the IV estimates.

But it can perhaps seem weird that are explanatory variable, x_i never shows up on the right hand side of the regression equation.

- Is there a way we can think about IV/2SLS where we are regressing y_i on x_i? **Yes!**

Linking IV back to regression on the explanatory variable

To begin, let's go back to the reduced form equation for y_i:

$$y_i = \gamma_0 + \gamma_1 z_i + e_i$$

Since IV estimation tells us $\frac{\gamma_1}{\pi_1} = \beta^{IV}$, then really what we have hear is:

$$y_i = \gamma_0 + \beta \pi_1 z_i + e_i$$

And what is $\pi_1 z_i$? $\hat{x}_i$

$$y_i = \gamma_0 + \beta \overbrace{\pi_1 z_i}^{\hat{x}_i} + e_i$$

- Where $\hat{x}_i$ is the predicted value of x_i from the first stage regression.

Reconceptualizing IV with $\hat{x}_i$

Using this idea of predicted values of x_i, let's reevaluate what exactly IV/2SLS is doing:

1. We begin by recognizing that for most given variables (eg attributes of an individual), the value of that variable for a given individual is probably not random.
   - A person's education is related to their parental income, IQ, etc.
   - A country's quality of government is related to their income, geography, civil climate, etc.
2. So we find a variable that influences the value of the *endogenous* variable, but which is otherwise as good as random with regards to the dependent variable,
   - This variable is called the *instrument*.

Reconceptualizing IV with $\hat{x}_i$

3. We then use that instrument to isolate "good" (ie random) variation in x to run the regression we care about.
4. We run regression of y_i on the "good" variation in x_i ($\hat{x}_i$) to produce the causal estimates of the relationship between y and x.

Let's formally write up this process in the next slide.

2SLS as regression using predicted values

First Stage:

Begin by running the first stage regression of the model:

$$x_i = \pi_0 + \pi_1 z_i + e_i$$

From this regression, we get:

$$\hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$$

$$x_i = \underbrace{\hat{x}_i}_{\text{"Good" (Exogenous) variation in x}} + \underbrace{\hat{\epsilon}_i}_{\text{"Bad" Variation in X}}$$

2SLS as regression using predicted values

Second Stage: In the second stage, run a regression of the form:

$$y_i = \alpha + \beta \hat{x}_i + u_i^*$$

Since we have now estimated the effect of x using only the good varaibles (omitting the "bad" variation that is correlated / endogenous to unobserved factors), our estimated β should be correct if the IV assumptions are satisfied.

Why are the standard errors wrong?

As noted in the lecture, one thing to be aware of is that running 2SLS by hand in Stata will produce the right coefficients, but the wrong standard errors?

Why?

- It should be clear that there is uncertainty in $\hat{x}_i$, unlike our usual assumption that x is a fixed value.
  - To see this, note that $\hat{x}_i$ depends on the coefficient $\hat{\pi}_1$, which has it's own variance.
- Without any special adjustments though, if you generate $\hat{x}_i$ in Stata and then use it in a regression, Stata won't know that it's any different from a normal regressor.

Why are the standard errors wrong? II

Let's see if we can be a little more formal about how the results are different:

What we want is the variance of β from the true structural equation:

$$y_i = \alpha + \beta x_i + u_i$$

The usual form of the variance if we could directly run this equation (ie if x not endogenous) is:

$$V(\beta) = \frac{\sigma_u^2}{Var(X_i)}$$

Why are the standard errors wrong? III

Since our x is endogenous, however, we isolate the good variation using $\hat{x}_i$, and run the 2^{nd} Stage regression:

$$y_i = \alpha + \beta \hat{x}_i + u_i^*$$

To see what's wrong here, relate this back to the structural equation:

$$y_i = \alpha + \beta x_i + u_i$$

$$y_i = \alpha + \beta(\hat{x}_i + \hat{e}_i) + u_i$$

$$y_i = \alpha + \beta \hat{x}_i + \beta \hat{e}_i + u_i$$

That is:

$$y_i = \alpha + \beta \hat{x}_i + u_i^*, \quad u_i^* = \beta \hat{e}_i + u_i$$

So the variance of β in the 2nd stage is: $Var(\beta^{2nd}) = \frac{\sigma^2_{u_i^*}}{Var(X_i)}$

Fixing the standard errors

As a result, we are using the wrong error variance in the 2nd stage.

To fix, simply correct for the right error variance:

$$Var(\beta^{2nd}) = \frac{\sigma^2_{u_i^*}}{Var(X_i)}$$

$$Var(\beta^{2nd}) * \frac{\sigma^2_{u_i}}{\sigma^2_{u_i^*}} = \frac{\sigma^2_u}{Var(X_i)} = Var(\beta^{Struct})$$

Error variance of the structural equation

Of course, to apply the correction term $\frac{\sigma^2_{u_i}}{\sigma^2_{u_i^*}}$, we need to know how to estimate $\sigma^2_{u_i}$.

From the structural equation:

$$y_i = \alpha + \beta x_i + u_i$$

And so:

$$u_i = y_i - \alpha + \beta x_i$$

Hence, from the usual formula for variance:

$$\sigma^2_{u_i} = \frac{\sum(y_i - \alpha + \beta x_i)^2}{N}$$

Error variance of the structural equation II

The only thing left to note is that this true variance relies on the actual structural equation variables x (not $\hat{x}_i$) and y, but also the actual structural estimates of the coefficients.

- Of course, the true coefficients should come from the the 2^{nd} stage, hence the error variance is:

$$\sigma^2_{u_i} = \frac{\sum(y_i - \hat{\alpha}^{2nd} + \hat{\beta}^{2nd} x_i)^2}{N}$$

I will demonstrate correcting the manual two stage estimates in Stata.

Stata Commands

Performing IV regression using the -ivreg2- package

There are multiple regression commands/packages to perform IV/2SLS in Stata, but the most comprehensive is **ivreg2**.

- Like **estout**, you will need to install the package from the Stata repository by typing in the console:

ssc install ivreg2

Performing IV regression using the -ivreg2- package II

The basic syntax of **ivreg2** is:

```
ivreg2 DEPVAR CONTROL_VARS (ENDOGVAR = INSTRUMENTS)
```

- Common options for **ivreg2** are:
  - You can specify heteroskedastic-robust or clustered standard errors, in the same manner as **regress** or **xtregress**
  - Include first-stage results by indicating **first**.
  - Include the reduced-form results by indicating **rf**.
  - Omit results for extra identification tests by indicating **noid**.

Tests of instrument relevance with **ivreg2**

The **ivreg2** regression command comes with a bunch of useful tools to help evaluate the IV strategy.

- When using the **first** option, **ivreg2** reports F-test of excluded instruments (ie your z variable(s)).
- If your tests are relevant, then we should reject the F-test null that the effect of the variables is jointly zero.
- **ivreg2** also reports an *underidentification test*, which is a formal hypothesis test of relevance.
    - The null hypothesis of this test is that the instruments are not relevant.

Testing instrument weakness with ivreg2

- As mentioned in lecture, a rule of thumber for whether instruments are *weak* is that the F-test for these instruments should be at least equal to 10 (relative bias is usually about 10% of OLS in this case).
- ivreg2 also reports a more formal *weak identification test*.
  - This test reports a test statistic and then critical values for different maximal sizes (probability of falsely rejecting) and relative biases.
  - If the test statistic is below the critical value, then the size or bias property is worse than the listed level.

Falsification Tests of Instrument Exogeneity with ivreg2 II

- **ivreg2** also present a falsification test of instrumental exogeneity that is sometimes useful, called the **test of overidentifying restrictions**.
- It is possible to use more than one instrument for an endogenous variable. In this case, it is possible to test whether there is any evidence that the exogeneity restriction is violated for at least one instrument, under the null that both instruments are exogenous.
- Rejecting the null hypothesis indicates evidence against the exogeneity of at least one instrument.
- Failing to reject the null though does not mean the instruments are valid - it is just a *falsification* test.