14.32 Recitation – IV and Simultaneous Equations

Paul Schrimpf

May 12, 2009

1 IV

- Review omitted variable bias in OLS
  - Model: $y = x\beta + \epsilon$, (and for simpler formulas, let $\bar{x} = \bar{y} = 0$) then

$$\begin{aligned}\hat{\beta} =& \frac{\sum x_i y_i}{\sum_i x_i^2} \\ =& \frac{\sum x_i(x_i\beta + \epsilon)}{\sum x_i^2} \\ \operatorname{plim}\hat{\beta} =& \beta + \frac{\operatorname{plim}\frac{1}{n}\sum x_i\epsilon_i}{\operatorname{plim}\frac{1}{n}\sum x_i^2} \\ =& \beta + \frac{Cov(x,\epsilon)}{V(x)}\end{aligned}$$

  - Usual omitted variable bias formula; If $\epsilon_i = \beta_a a_i + u_i$ with $E[x_i u_i] = 0$, but not for a_i, then $\operatorname{plim}\hat{\beta} = \beta + \beta_a \frac{Cov(x,a)}{V(x)} = \beta + \beta_a\delta$ where δ is the coefficient from regressing a on x
- Need to use IV when we want to estimate a causal effect. A good way to think about causal effects is to imagine the ideal randomized experiment that you would perform to answer your question. Then think about the data you have. If you can tell stories about why OLS might not estimate the experimental effect, then you need an instrument.
- Consistency of IV:
  - Have z such that $E[z\epsilon] = 0$ (exogeneity) and $E[zx] \neq 0$ (relevance)
  - $\hat{\beta}^{IV} = \frac{\sum(z_i - \bar{z})y_i}{\sum(z_i - \bar{z})x_i}$[1]

$$\begin{aligned}\operatorname{plim}\hat{\beta}^{IV} =& \operatorname{plim}\frac{\frac{1}{n}\sum(z_i - \bar{z})(x_i\beta + \epsilon_i)}{\sum(z_i - \bar{z})x_i} \\ =& \beta + \frac{Cov(z,\epsilon)}{Cov(x,z)} = \beta\end{aligned}$$

- Asymptotic standard error:
  - With homoskedasticity, $E[\epsilon^2|z] = \sigma_\epsilon^2$ (notice now we need homoskedasticity as a function of z instead of x), then asymptotic $V(\hat{\beta}^I V) = \frac{\sigma_\epsilon^2}{n\sigma_x^2\rho_{xz}^2}$
    - How does this compare to the standard error of OLS?

[1] This is for scalar x and z. As an exercise, you may want to show that this formula is exactly the same as 2SLS

- Like OLS, standard error changes if there is heteroskedasticity or serial correlation.

- Test for endogeneity:
  - Suspect $E[x\epsilon]neq0$, but not sure. Have z such that $E[z\epsilon] = 0$
  - Asymptotically equivalent tests for $H_0 : E[x\epsilon] = 0$
    1. Regression version:
       (a) regress $x = \pi z + v$, save $\hat{v}$
       (b) regress $y = x\beta + hatv\delta + e$
       (c) test $\hat{\delta} = 0$
    2. Hausman test: compute $\hat{\beta}^{OLS}$ and $\hat{\beta}^{IV}$. Because under H_0, OLS is efficient and both estimators are consistent, $V(\hat{\beta}^{OLS}$ - $\hat{\beta}^{IV}) = V(\hat{\beta}^{IV}) - V(\hat{\beta}^{OLS})$[2]. Use this to do a t-test (or F-test if you have more than one x).

- Overidentification:
  - More instruments than endogenous variables. e.g. x_i scalar, and have $z_{1,i}$, $z_{2,i}$
  - Can test $H_0 : E[z_{1,i}\epsilon] = E[z_{2,i}\epsilon] = 0$ or perhaps more accurately can test whether instruments all give the same estimate of β
  - Equivalent forms of test
    1. Regression: estimate $\hat{\beta}^{2SLS}$ using all instruments, then regress residuals, $\hat{\epsilon}$ on z_1 and z_2. Test whether the coefficients are zero
    2. Hausman: compare $\hat{\beta}^{2SLS}$ using all z to $\hat{\beta}^{2SLS}_{z1}$ using just z_1

2 Simultaneous Equations

- Jointly determined variables, y_1 and y_2

$$\begin{aligned} y_1 =& \alpha_1 y_2 + z_1\beta_1 + u_1 \\ y_2 =& \alpha_1 y_1 + z_2\beta_2 + u_2 \end{aligned}$$

- Identification: when can we estimate α and β?
  - Can estimate α_1 and β_1 if we can find an instrument for y_2. To do this, we can look at the reduced form for y_2 – meaning write y_2 in terms of z

$$\begin{aligned} y_2 =& \alpha_1 (\alpha_1 y_2 + z_1\beta_1 + u_1) + z_2\beta_2 + u_2 \\ y_2 =& z_1 \frac{\alpha_1\beta_1}{1-\alpha_1\alpha_2} + z_2 \frac{\beta_2}{1-\alpha_1\alpha_2} + \frac{u_1\alpha_1 + u_2}{1-\alpha_1\alpha_2} \\ =& z_1\pi_{21} + z_2\pi_{22} + v_2 \end{aligned}$$

    To avoid dividing by 0, we need $\alpha_1\alpha_2 \neq 1$. For this to give us an instrument we need z_2 to not be colinear with z_1. In particular, we need z_2 and z_1 to not be the same variable, or if they are vectors, we need at least one component of z_2 to not be part of z_1 (the order condition). Finally, to satisfy the relevance condition we need the coefficient on the excluded component of z_2 to not be zero (the rank condition).

- Estimation:
  - 2SLS on each identified equation
  - If the system is overidentified 3SLS, which is a version of GLS that takes into account the correlation between u_1 and u_2 is asymptotically more efficient than 2SLS

[2]This equality is always true for the variance of the difference between an efficient estimator and another estimator.

3 Stata

- 2SLS to estimate $y = x\beta + w\delta + \epsilon$ using $z1$ and $z2$ as instruments for x

```
ivregress 2sls y (x = z1 z2) w
```

- Hausman test for endogeneity

```
ivregress 2sls y (x = z1 z2) w
estimates store biv
reg y x w
estimate store bols
hausman biv bols
```

Or,

```
ivregress 2sls y (x = z1 z2) w
estat endogenous
```

- Overid tests

```
// Sargan test (similar to regression based test described above)
ivregress 2sls y (x = z1 z2) w
estat overid
// Hausman test
ivregress 2sls y (x = z1 z2)w
estimates store ivall
ivregress 2sls y (x = z1) w
estimates store iv1
hausman iv1 ivall
```

- 3SLS – e.g.

$$\begin{aligned} y_1 &= \alpha_1 p + x\beta_1 + \epsilon_1 \\ y_2 &= \alpha_2 p + x\beta_2 + \epsilon_2 \\ p &= z_1\pi_1 + z_2\pi_2 + x\pi_x + u \end{aligned}$$

```
reg3 (y_1 p x) (y_2 p x) (p z_1 z_2 x)
// test a_1 = a_2
test [y_1]p=[y_2]p
```