Econometric Issues in empirical research

Empirical research requires us to creatively blend substantive and technical issues.

- The first requirement is to pose an interesting question. The best questions have us learn something regardless of the answer the data gives us. Questions with no answers (or all you can say is "the data are uninformative") are best avoided.
- Finding data is a key and underappreciated skill. Good research requires good data. In economics/finance, we rarely get to generate original data. It usual comes from official sources, or as a by-product of some private activity. Interesting data are often the seminal impetus for good research questions.

- There’s no textbook on how to pose good questions, find good data, or "be creative". But I can provide a checklist of econometric issues that you should keep in mind.

1. Specification of the regression function

- Do the data have special features that you can exploit?
  - Restrictions on supp[y]: positive, dummy, count, truncated, censored. If so, is a regression model appropriate?
  - Time series, panel data, or multivariate y? If so, then there are potentially important possibilities to add variables to the regression function to eliminate LOV bias.

Suppose a regression model is appropriate

$$y = X\beta + u$$

- **Are you interested in the coefficient of the BLP or some other parameter**?
  - Are there important left out variables? Do you have measurement error? Is there simultaneity bias?
- Are there important nonlinearities?
- Is there heterogeneity in the coefficient vector?
- Are there dynamics?
- What are the error properties?
  - heteroskedasticity, serial correlation, cluster effects (system of equations, panel data, etc.)
  - specification tests
  - FGLS, HAC inference

2. Evaluation of the model

- Hypothesis tests. Should the model be simplified?
- Interpretation of the coefficients?
- How would you use the model?
  - Data summary
  - Prediction
  - Recommended action

So far, we've proceeded as if we are interested in β from the BLP. But it's not the only choice!

Ch 15. IV and 2SLS

- Consider the model

$$y_i = x_i\beta + u_i$$

  where $E(x_i' u_i) \neq 0$

- Why would we get $E(x_i' u_i) \neq 0$?
  1. LOV
  2. measurement error
  3. predetermined regressors and persistent errors
  4. simultaneity bias

Suppose we can find variables $z_i \in \mathbb{R}^p$ such that

i) $E(z_i' u_i) = 0$

ii) $E(z_i' x_i)$ has full rank K ($\Rightarrow p \geq K$)

- z_i may contain some of the variables in x_i as well as other variables not in x_i.
- We call the variables z_i satisfying (i) and (ii) *instruments.* We often say *valid instruments* to emphasize that they have these properties, rather than they are only *claimed* to do so
- In Finance, *instruments* is used to describe regressors, proxies, objects in the information set,....
- We can use Method of Moments and properties (i) and (ii) to generate the Instrumental Variables (IV) estimator. OLS is a special case.

Given the model

$$y = X\beta + u$$

the IV estimator solves the moment (normal) equations

$$(*) \quad Z'(y - X\widehat{\beta}_{IV}) = 0$$

- Why would this give a consistent estimator of the parameter of interest? Premultiply both sides of the model by $n^{-1}Z'$

$$\frac{1}{n}Z'y = \frac{1}{n}Z'X\beta + \frac{1}{n}Z'u$$

But $E(z_i'u_i) = 0$. So if we can invoke a LLN, the last term goes to zero in probability. So the coefficient vector β of interest will satisfy the IV moment equations asymptotically. And as long as $\frac{1}{n}Z'X$ stays nonsingular, there will be a unique solution to the moment equations.

Solution of IV moment equations

Case 1. If $p = K$, then w.p.1, $Z'X$ is nonsingular by (ii) and

$$\widehat{\beta}_{IV} = (Z'X)^{-1}Z'y$$

Case 2. If $p \geq K$, then w.p.1 we can find a matrix $Q \in R^{Kxp}$ such that $Q'Z'X$ is nonsingular and (premultiplying $*$ by Q')

$$\widehat{\beta}_{IV} = (Q'Z'X)^{-1}Q'Z'y$$

- Write $W = ZQ$. The formula above is just $\widehat{\beta}_{IV} = (W'X)^{-1}W'y$ Some call W the instruments and Z the *IV candidates*. If $p = K$, then Q must be nonsingular and the formula reduces to Case 1.

In what follows, I'll assume Case 1, unless stated otherwise. The extension to Case 2 is straightforward.

Geometry of IV estimator

- Let $\widehat{y}_{IV} = X\widehat{\beta}_{IV}$ denote the fitted value from the IV estimation. By construction, $\widehat{y}_{IV} \in Sp(X)$, and

$$\widehat{y}_{IV} = X\widehat{\beta}_{IV} = X(Z'X)^{-1}Z'y \equiv P_{IV}y$$

- Notice that P_{IV} is idempotent

$$\begin{aligned} P_{IV}^2 &= (X(Z'X)^{-1}Z')X(Z'X)^{-1}Z' \\ &= X(Z'X)^{-1}Z' = P_{IV} \end{aligned}$$

  so $\widehat{y}_{IV}$ is a projection onto $Sp(X)$
- But P_{IV} is not symmetric, so the projection is not orthogonal
- R^2 makes no sense with IV estimation

Consistency of the IV estimator

$$\begin{aligned}\widehat{\beta}_{IV} &= (Z'X)^{-1}Z'y \\ &= \beta + (Z'X)^{-1}Z'u \\ &= \beta + (\frac{1}{n}\sum_{i=1}^{n} z_i' x_i)^{-1}(\frac{1}{n}\sum_{i=1}^{n} z_i' u_i)\end{aligned}$$

- By property (i) above and LLN, we have $plim \frac{1}{n}\sum_{i=1}^{n} z_i' u_i = 0$
- By property (ii) and LLN, we have $plim \frac{1}{n}\sum_{i=1}^{n} z_i' x_i$ is nonsingular

Therefore $plim\ \widehat{\beta}_{IV} = \beta$

Asymptotic Normality of IV Estimator

As with OLS, we must use a CLT to show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i' u_i \sim^a N(0, V_n)$$

where V_n is an appropriately chosen matrix (with my abuse of notation, it could even be random)

- If $E(u_i^2|z_1, z_2, \cdots) = \sigma^2$, then we obtain

  $$\widehat{\beta}_{IV} \sim^a N(\beta, \sigma^2(Z'X)^{-1}Z'Z(X'Z)^{-1}) = N(\beta, \sigma^2(X'P_zX)^{-1})$$

  If $K = 1, \sigma^2(X'P_zX)^{-1}) = \sigma^2(X'X)^{-1}/R^2_{0,(X,Z)}$
- If $E(u_i^2|z_1, z_2, \cdots) = \sigma_i^2$, and $\Omega \equiv diag(\sigma_i^2)$, then we obtain

  $$\widehat{\beta}_{IV} \sim^a N(\beta, \sigma^2(Z'X)^{-1}Z'\Omega Z(X'Z)^{-1})$$
- What happens if there is serial correlation?

Examples

- In general, finding valid instruments is HARD because we need to satisfy both (i) and (ii). Satisfying either is easy, but not both.
- Consider a wage regression

$$\ln wage = \beta_0 + \beta_1 educ + u \qquad u = \beta_2 ability + e$$

- We have 2 parameters and need 2 instruments. The choice $z_{1i} = 1$ is easy. But what could we pick for z_{2i}? We need a variable that is correlated with *educ* but uncorrelated with *ability* or anything that contributes to the unobserved disturbance *e*

- Candidates for z_2:
  - random numbers
    - satisfies (i) but not (ii)!
  - proxy variables for ability (eg. IQ)
    - satisfies (ii) but not (i)!
  - Mother's education
    - correlated with *educ*, is it correlated with *ability*?
  - number of siblings?
  - month of birth?
  - distance to nearest university?

- Assumption (ii), $E(z_i' x_i)$ has full column rank, is testable. And it *should* be tested!
- Assumption (i), $E(z_i' u_i) = 0$ is an identifying assumption. It can't be tested. You have to use extraneous knowledge (theory or natural experiment) to justify it.
- In Case II, $p > K$, we have more instruments than we "need". We can test if all p instruments are valid while maintaining that $W = ZQ$ are valid instruments (Sargan-Hansen test). The idea is to see if IV residuals are orthogonal to Z (taking into account that they will be orthogonal to W by construction)
- If we can find enough valid instruments, we can also test if OLS is valid (Hausman-Wu test)

Properties of IV with poor instruments

For the simple regression model

$$y = \beta_0 + \beta_1 x + u$$

- The IV estimator satisfies

$$plim\ \hat{\beta}_{1,IV} = \beta_1 + \frac{cov(z'u)}{cov(z'x)} = \beta_1 + \frac{corr(z'u)}{corr(z'x)}\frac{\sigma_u}{\sigma_x}$$

- The OLS estimator goes to

$$plim\ \hat{\beta}_1 = corr(x'u)\frac{\sigma_u}{\sigma_x}$$

- So if $corr(z'x)$ is small, the inconsistency using OLS may be smaller even though $0 \neq corr(z'u) << corr(x'u)$
- Poor instruments also generate very large standard errors. In limit, as $corr(z'x) \downarrow 0$, standard asymptotic theory provides a very poor approximation even with HUGE sample size

2SLS (Two-stage least squares)

Suppose we have $p > K$. For each Q

$$\widehat{\beta}_{IV} \sim^{a} N(\beta, (Q'Z'X)^{-1}Q'Z'\Omega ZQ(X'ZQ)^{-1})$$

- Assume $\Omega = \sigma^2 I$; to minimize the variance of the asymptotic distribution (avar), choose

$$Q = (Z'Z)^{-1}Z'X$$

the resulting estimator is called the *2SLS estimator*

Proof: If $Q = (Z'Z)^{-1}Z'X$, the formula for the avar simplifies to

$$(Q'Z'X)^{-1}Q'Z'ZQ(X'ZQ)^{-1} = (X'P_Z X)^{-1}$$

So we need to show (using ≥ 0 to denote a nonnegative definite matrix)

$$(Q'Z'X)^{-1}Q'Z'ZQ(X'ZQ)^{-1} - (X'P_ZX)^{-1} \geq 0$$

But we know that $A^{-1} - B^{-1} \geq 0$ iff $B - A \geq 0$. So the equation above is equivalent to

$$(X'P_ZX) - (X'ZQ)(Q'Z'ZQ)^{-1}(Q'Z'X) \geq 0$$

which can be rewritten as

$$X'(P_Z - P_{ZQ})X \geq 0$$

So the result follows immediately from $Sp(ZQ) \subset Sp(Z)$

- Define $\widehat{X} = Z(Z'Z)^{-1}Z'X = P_Z X$.
- The 2SLS estimator can be written in as

$$\begin{aligned}\widehat{\beta}_{2SLS} &= (\widehat{X}'X)^{-1}\widehat{X}'y \\ &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y \\ &= (X'P_Z X)^{-1}X'P_Z y \\ &\sim^a N(\beta, \sigma^2(X'P_Z X)^{-1})\end{aligned}$$

(1) More instruments reduce $(X'P_Z X)^{-1}$

(2) Can't use R^2 or SSR to test $H_0 : R\beta = r$, but Wald test still works.

A related approach to combining instruments optimally when $p > K$ uses a nonsingular matrix $V \in \mathbb{R}^{pxp}$ to weight the p moments in $(*)$

$$\min_{\beta}(y - X\beta)'ZVZ'(y - X\beta)$$

The first-order conditions

$$X'ZVZ'(y - X\widehat{\beta}_{IV,V}) = 0$$

The problem then is to choose the weighting matrix V to minimize that avar of the resulting estimator.

Compared to our approach of choosing Q, this appears to restrict attention to matrices of the from $Q' = X'ZV$. But the restriction is not costly as this alternative approach yields the same optimal estimator with $V = (Z'Z)^{-1}$

Testing for Endogeneity (Hausman-Wu)

- Suppose we have the model

$$y_1 = y_2\beta_1 + Z_1\beta_2 + u \equiv X\beta + u$$

Where we believe $E(Z_1'u) = 0$ but we suspect $E(y_2'u) \neq 0$

- Posit (using obvious notation if y_2 has many columns)

$$y_2 = Z_1\pi_1 + Z_2\pi_2 + v \equiv Z\pi + v$$

- Let $\hat{v} = M_Z y_2$. Run the regression

$$(*) \quad y_1 = y_2\beta_1 + Z_1\beta_2 + \hat{v}\delta + u$$

and test $H_0 : \delta = 0$ to decide if $E(y_2'u) = 0$. Why?

$$(*) \quad y_1 = (P_Z y_2)\beta_1 + Z_1\beta_2 + \hat{v}(\beta_1 + \delta) + u$$

A test of $\delta = 0 \Leftrightarrow$ test that OLS=2SLS.

- By construction, $(*)$ gives 2SLS estimates.

Testing Overidentification (Sargan-Hansen)

- If $p > K$ we could use any K independent linear combinations of the instrumental variable candidates to estimate β.
- Suppose we have homoskedastic errors. Then 2SLS gives the K lin. comb. with the smallest covariance matrix. We can test if the remaining linear combinations are orthogonal to the disturbance.
- Construct the 2SLS residuals $\hat{u}_{2SLS}$
- Regress $\hat{u}_{2SLS}$ on the IV candidates Z. Under $H_0 : E(z_i' u_i) = 0$, we have $nR^2 \sim^a \chi^2(p - K)$
- According to the usual first-order theory, adding valid instruments can't hurt and usually helps. But in small samples, significant bias obtains if we have too many instruments. In the limit, if $p = n$ then $\hat{\beta}_{IV} = \hat{\beta}$!

2SLS with heteroskedasticity

- We can use the Breusch-Pagan test with $\hat{u}_{2SLS}$ in place of $\hat{u}$.
- If we know the form of the heteroskedasticity, we can weight the observations to gain efficiency.
- We can get heteroskedasticity robust standard errors using a direct analog of the White covariance matrix estimator:

$$\sqrt{n}\,(Z'X)(\hat{\beta}_{IV} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i' u_i \sim^a N(0, V_n)$$

estimate V_n consistently with $\hat{V}_n = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_{i,2SLS}^2 z_i' z_i$

and with usual abuse of notation, we get

$$\hat{\beta}_{IV} \sim^a N(\beta, n(Z'X)^{-1} V_n (X'Z)^{-1})$$

2SLS with time series

- Unit roots and near unit roots raise the same issues for IV estimation as we discussed for OLS. Consider adding trends (seasonals) to the regression; first differencing data, etc.
- The Breusch-Pagan test extends in the obvious way:
  1. Estimate $y_t = x_t\beta + u_t$ by 2SLS (instruments z_t)
  2. Estimate $y_t = x_t\beta + \rho_1\hat{u}_{t-1,2SLS} + \cdots\rho_p\hat{u}_{t-p,2SLS} + e_t$ (using the same instruments z_t as in step 1., but adding the lagged IV residuals to the instrument set)
  3. Use an F-test (or HC analog) to test the null hypothesis $\rho_1 = \cdots = \rho_p = 0$

- If we find serial correlation
  1. We can compute the analog to the N-W HAC
  2. We can add lags to get a DC model
  3. We can attempt FGLS by quasi-differencing and then estimating the model

     $$\tilde{y}_t = \tilde{x}_t\beta + \tilde{u}_t \qquad \tilde{y}_t = \rho(L)y_t, \text{ etc}$$

     But this raises the question: Which instruments to use? It's natural to try $\tilde{z}_t = \rho(L)z_t$ but this raises the issue that unless the regressors are strictly exogenous, $E(z_t'u_t) = 0 \nRightarrow E(\tilde{z}_t'\tilde{u}_t) = 0$