2SLS: Testing

Econometrics I

Ricardo Mora

Department of Economics
Universidad Carlos III de Madrid
Master in Industrial Economics and Markets

Outline

Motivation

Two Stage Least Squares

$$y_1 = \beta_0 + \beta_1 z_1 + \beta_2 y_2 + u$$

$$cov(z_1, u) = 0, cov(z_2, u) = 0$$

First stage: $\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2$
where $\hat{\pi}_j$ are OLS estimates.

$$\sum \left(y_1 - \hat{\beta}_0 - \hat{\beta}_2 y_2 - \hat{\beta}_1 z_1 \right) = 0$$
$$\sum z_1 \left(y_1 - \hat{\beta}_0 - \hat{\beta}_2 y_2 - \hat{\beta}_1 z_1 \right) = 0$$
$$\sum \hat{y}_2 \left(y_1 - \hat{\beta}_0 - \hat{\beta}_2 y_2 - \hat{\beta}_1 z_1 \right) = 0$$

Identification

- the 2SLS estimator exploits in the sample the orthogonality conditions from all exogenous regressors and the instruments
- when we have more orthogonality conditions than parameters, they cannot simultaneously be satisfied in small samples (almost surely)
- it can be shown that 2SLS satisfies a linear combination of all orthogonality conditions & that the weight of each condition depends on how good the instrument is
- suppose we have k_{y_2} endogenous variables and k_{z_2} instruments
  - if $k_{z_2} = k_{y_2}$, the model is said to be "just-identified"
  - if $k_{z_2} \geq k_{y_2}$, the model is said to be "over-identified"

Model Specification

- after estimation, we can conduct the usual tests on the estimated parameters
- in addition, we can perform other tests related two the model specification
  - we can test whether y_2 is actually endogenous
  - if the model is over-identified, we can test for the validity of the over-identifying conditions

Testing for Regressor Endogeneity

Testing for regressor endogeneity

$$H_0 : cov\left(y_2, u\right) = 0$$
$$H_1 : cov\left(y_2, u\right) \neq 0$$

- Under H_0,
  - $\hat{\beta}^{OLS}$ and $\hat{\beta}^{2SLS}$ are consistent
  - $\hat{\beta}^{OLS}$ is more efficient
- Under H_1, only $\hat{\beta}^{2SLS}$ is consistent

The Hausman Test

In the simple regression model ($k=1$)

$$H=\frac{\left(\hat{\beta}^{2SLS}-\hat{\beta}^{OLS}\right)^2}{A\hat{v}ar\left(\hat{\beta}^{2SLS}-\hat{\beta}^{OLS}\right)}$$

In the general case

$$H=\left(\hat{\beta}^{2SLS}-\hat{\beta}^{OLS}\right)^t\left[A\hat{v}ar\left(\hat{\beta}^{2SLS}-\hat{\beta}^{OLS}\right)\right]^{-1}\left(\hat{\beta}^{2SLS}-\hat{\beta}^{OLS}\right)$$

- the Hausman test compares $\hat{\beta}^{2SLS}$ with $\hat{\beta}^{OLS}$: a large difference suggests $\hat{\beta}^{OLS}$ is inconsistent
- under the null, $H\xrightarrow{a}\chi^2_k$ where k is the number of regressors

The Hausman test with *iid* errors

- If errors are *iid*, then $\hat{\beta}^{OLS}$ is the fully efficient estimator under the null
- Hausman proved that, in that case, $Avar\left(\hat{\beta}^{2SLS} - \hat{\beta}^{OLS}\right) = Avar\left(\hat{\beta}^{2SLS}\right) - Avar\left(\hat{\beta}^{OLS}\right)$

$$H = \left(\hat{\beta}^{2SLS} - \hat{\beta}^{OLS}\right)^t \left[A\hat{v}ar\left(\hat{\beta}^{2SLS}\right) - A\hat{v}ar\left(\hat{\beta}^{OLS}\right)\right]^{-1} \left(\hat{\beta}^{2SLS} - \hat{\beta}^{OLS}\right)$$

- this test can be directly computed from the $2SLS$ and the OLS regressions (the `hausman` command in `Stata`)
- *iid* is a very strong assumption & sometimes the small sample approximation is not positive definite
- we can implement an alternative test

The Durwin-Wu-Hausman test

1. First step:
   `regress` y_2 z_1
   `predict` $\hat{v}$`, res`
2. Second step
   regress y_1 z_1 y_2 $\hat{v}$

- If $cov(y_2, u) \neq 0$, $plim\, cov_N(\hat{v}, y_1) \neq 0$ and the coefficient for $\hat{v}$ in second step would be significant (in this case, the second step is like adding to the original regression the missing variable which captures the correlation between y_2 and u)
- If $cov(y_2, u) = 0$, $plim\, cov_N(\hat{v}, u) = plim\, cov_N(\hat{v}, y_1) = 0$ and the slope for $\hat{v}$ in the second step should not be statistically significantly different from 0

The Durwin-Wu-Hausman test

1. First step:
   `regress` y_2 z_1
   `predict` $\hat{v}$`, res`
2. Second step
   regress y_1 z_1 y_2 $\hat{v}$

- If $cov\left(y_2, u\right) \neq 0$, $plim\, cov_N\left(\hat{v}, y_1\right) \neq 0$ and the coefficient for $\hat{v}$ in second step would be significant (in this case, the second step is like adding to the original regression the missing variable which captures the correlation between y_2 and u)
- If $cov\left(y_2, u\right) = 0$, $plim\, cov_N\left(\hat{v}, u\right) = plim\, cov_N\left(\hat{v}, y_1\right) = 0$ and the slope for $\hat{v}$ in the second step should not be statistically significantly different from 0

Testing for over-identifying restrictions

Can we test how good instruments are?

- if there is just one instrument z_j for each endogenous variable y_j
  - we say the model is just-identified
  - we can't test whether z_j is uncorrelated with the error u
- if we have multiple instruments for at least one endogenous variable
  - we say the model is over-identified
  - we can test if "the over-identifying instruments" are good instruments
- this is called testing for over-identifying restrictions

A Simple Test for Over-identifying Restrictions

1. Estimate the model using 2SLS and obtain the residuals $\hat{u}$
2. regress $\hat{u}$ z_1 z_2^1 $z_2^2 \rightarrow R^2$
3. $S = nR^2$ where n is the sample size

under the null that all instruments are uncorrelated with the error

$$S \rightarrow \chi^2_q$$

- where q is the number of extra instruments

2SLS and Stata

Stata and Two Stage Least Squares

- Stata does 2*SLS* the estimation for you to get the correct (robust) standard errors
  - help `ivregress` (`ivreg`, `ivreg2` for `Stata 9`)
- also use `test` command to test for linear restrictions
  - help `ivregress postestimation`
- you need at least as many instruments as the number of endogenous variables

ivregress 2sls *depvar varlist1* (*varlist2=instruments*),vce(robust)

- Demand function:

ivreg 2sls *quantity demand_shifters* (*price=supply_shifters*), vce(robust)

- Supply function:

ivreg 2sls *quantity supply_shifters* (*price=demand_shifters*), vce(robust)

- Wages and education

ivreg 2sls *wages exp exp2* (*educ* = *fed med*), vce(robust)

- Savings and Income

ivreg 2sls *sav* (*income=house_size car_price*),vce(robust)

Postestimation commands after `ivregress 2sls`

- `estat endogenous`: are regressors in the model exogenous?
  1. with an unadjusted VCE: the Durbin (1954) and Wu-Hausman statistics
  2. with a robust VCE, a robust score test (Wooldrigde 1995) and a robust regression-based test
  3. if the test statistic is significant, the variables must be treated as endogenous
- `estat overid`: tests of over-identifying restrictions.
  1. Sargan's (1958) and Basmann's (1960) chi-squared tests are reported, as is Wooldridge's (1995) robust score test
  2. a statistically significant test statistic indicates that the instruments may not be valid.

Summary

- we can test for endogeneity and also for the validity of the extra instruments