Testing for Underidentification

Frank Windmeijer

Dept. of Statistics and Nuffield College
University of Oxford

Stata Conference, Antwerp, 7 June 2021

Publication

- “Testing underidentification in linear models, with applications to dynamic panel and asset pricing models”,

  Journal of Econometrics, 2021

  `https://doi.org/10.1016/j.jeconom.2021.03.007`

Stata Commands

- Schaffer, M.E., Windmeijer, F., 2020.
  UNDERID: Postestimation tests of under- and over-identification after linear IV estimation.
  `https://ideas.repec.org/c/boc/bocode/s458805.html`
- Kleibergen, F., Schaffer, M.E., Windmeijer, F., 2020.
  RANKTEST: Stata module to test the rank of a matrix.
  `https://ideas.repec.org/c/boc/bocode/s456865.html`

Standard linear IV model specification

$$y = X\beta + u$$
$$X = Z\Pi + V$$

Standard 2SLS estimator for β is

$$\hat{\beta}_{2sls} = \left(X'P_Z X\right)^{-1} X'P_Z y$$

where $P_Z = Z\left(Z'Z\right)^{-1} Z'$.

With one endogenous regressor

$$x = Z\pi + v,$$

first-stage F is well-known test for underidentification, $H_0 : \pi = 0$, and also used as test for weak instruments under homoskedasticity.

Here, we focus on robust tests for underidentification when there are multiple endogenous variables, and show how they can be used in e.g. dynamic panel data models estimated by GMM.

Overidentification

The Sargan/Hansen test for overidentifying restrictions is a test for $H_0 : \mathbb{E}[z_i u_i] = 0$,

$$J = \hat{u}' Z \hat{W} Z' \hat{u}.$$

It is also a (robust) score test for $H_0 : \gamma = 0$ in

$$y = X\beta + Z_o\gamma + u,$$

where $Z_o$is any $k_z - k_x$ subset of the instruments.

Further, when using invariant estimators like LIML and CUE it is a test on the rank of $\Pi^* = [\pi_y \ \Pi]$, $H_0 : \mathrm{r}(\Pi^*) = k_x$, where π_y is the reduced form parameter vector

$$\begin{aligned} y &= X\beta + u = Z\Pi\beta + V\beta + u \\ &= Z\pi_y + w. \end{aligned}$$

Underidentification

Main result is that tests for underidentification, $H_0 : \mathrm{r}(\Pi) = k_x - 1$, are overidentification tests in an auxiliary model.

Let $X = [x_1 \;\; X_2]$, then the auxiliary model is given by

$$\begin{aligned} x_1 &= X_2\delta + \varepsilon \\ X_2 &= Z\Pi_2 + V_2 \end{aligned}$$

- Cragg-Donald (CD) tests are LIML-based Sargan or CUE-based Hansen tests for $H_0 : \mathbb{E}[z_i\varepsilon_i] = 0$.
- Kleibergen-Paap (KP) test is LIML-based robust score test for $H_0 : \gamma = 0$ in
$$x_1 = X_2\delta + Z_o\gamma + \varepsilon.$$

Underidentification

- Clearly, if z_i not correlated with ε_i then instruments not informative for x_1 after instrumenting X_2.
- So, if not rejecting the null of overidentifying restrictions in the auxiliary model, then underidentification is not rejected.

Alternative to KP test is Hansen J-test based on new two-step efficient invariant estimator

$$\hat{\delta}_{2L} = \left(\hat{\Pi}_{2L}' Z' Z W\left(\hat{\delta}_L\right) Z' X_2\right)^{-1} \hat{\Pi}_{2L}' Z' Z W\left(\hat{\delta}_L\right) Z' y,$$

where $\hat{\delta}_L$ and $\hat{\Pi}_{2L}$ are the LIML estimators for δ and Π_2 and $W\left(\hat{\delta}_L\right)$ is e.g.

$$W\left(\hat{\delta}_L\right) = \left(\sum_{i=1}^{n} \hat{\varepsilon}_{Li}^2 z_i z_i'\right)^{-1}$$

with $\hat{\varepsilon}_L = x_1 - X_2 \hat{\delta}_L$.

The statistic is

$$J_{2L} = J\left(\hat{\delta}_{2L}\right).$$

.

Underidentification

Alternatively, can compute non-invariant robust two-step Hansen tests for each variable:

$$x_j = X_{\{-j\}}\delta_j + \varepsilon_j$$
$$X_{\{-j\}} = Z\Pi_{\{-j\}} + V_{\{-j\}}$$

which is a robust extension to the Sanderson-Windmeijer (SW) conditional F-test procedure.

UNDERID

The stata command UNDERID (Schaffer and Windmeijer, 2020) incorporates all of the above:

- Non-robust 2SLS (SW) and LIML (CD)
- Robust LIML (KP and J2L) and CUE (CD)
- Robust two-step GMM (SW)

As post-estimation command after:

- ivregress, ivreg2
- xtivreg, xtivreg2, xthtaylor
- xtabond2, xtdpdgmm

- Problems of uninformative instruments in dynamic panel daa models well documented.
- Underidentification test not commonly reported/available. UNDERID changes that.

Dynamic panel data model example from Acemoglu et al. (2008)

$$d_{it} = \alpha d_{i,t-1} + \beta inc_{i,t-1} + \tau_t + \eta_i + u_{it},$$

where d_{it} is a measure of democracy and inc_{it} is log income per capita.

The data used is a panel of 127 countries observed over the period 1960-2000 at 5-year and 10-year frequencies.

The parameters are estimated using the Arellano-Bond GMM estimator for the model in first-differences,

$$d_{it} - d_{i,t-1} = \alpha\,(d_{i,t-1} - d_{i,t-2}) + \beta\,(inc_{i,t-1} - inc_{i,t-2}) + \tilde{\tau}_t + u_{it} - u_{i,t-1},$$

with the instruments for $t = 3, \ldots, T$, specified as

$$Z_i = \begin{bmatrix} d_{i1} & 0 & 0 & 0 & inc_{i1} \\ 0 & d_{i1}\; d_{i2} & 0 & 0 & inc_{i2} \\ 0 & 0 & \ddots & 0 & \vdots \\ 0 & 0 & 0 & d_{i1} \ldots d_{i,T-2} & inc_{i,T-2} \end{bmatrix}.$$

Table: Estimation results, Arellano-Bond estimator

	one-step GMM		iterated GMM	
Frequency	d_{-1}	inc_{-1}	d_{-1}	inc_{-1}
5 year	0.489	-0.129	0.744	-0.009
$k_z - k_x = 44$	(0.095)	(0.088)	(0.128)	(0.039)
$n = 127$, $\#obs = 838$				
10 year	0.227	-0.318	0.288	-0.280
$k_z - k_x = 15$	(0.125)	(0.183)	(0.146)	(0.202)
$n = 118$, $\#obs = 338$				

Notes: From Acemoglu et al. (2008, Table 2) and Hansen and Lee (2021, Table 4). Misspecification-robust standard errors of Hansen and Lee (2021) in brackets.

Stata command using xtdpdgmm (Kripfganz, 2019):

```
xtdpdgmm d l.d l.y i.year, gmm(d,lag(2 .))  iv(l2.y
i.year) m(d) nocons vce(rob)
underid, sw
```

Or using xtabond2 (Roodman, 2009)

```
xtabond2 d l.d l.y i.year, gmm(d,lag(2 .))  iv(l2.y,p)
iv(i.year) rob nol
underid, sw nopartial
```

This gives as output:

```
Underidentification test:
Cragg-Donald robust CUE-based (LM version)
Test statistic robust to heteroskedasticity and
clustering on country
j = 51.62 Chi-sq( 45) p-value=0.2309

2-step GMM J underidentification stats by regressor:
j = 67.00 Chi-sq( 45) p-value=0.0183 L.d
j = 56.27 Chi-sq( 45) p-value=0.1209 L.y
```

Table: P-values of robust underidentification test statistics

	CD	SW	
Frequency		d_{-1}	inc_{-1}
5 year	0.231	0.018	0.121
10 year	0.026	0.000	0.022

General Rank Tests

The same methodology can be used for testing $H_0 : \mathrm{r}(B) = k_y - 1$ in standard linear model

$$Y = XB + V,$$

with $k_x \geq k_y$, i.e. specify auxiliary model

$$\begin{aligned} y_1 &= Y_2\delta + \varepsilon \\ Y_2 &= XB_2 + V_2 \end{aligned}$$

and test $H_0 : \mathbb{E}[x_i\varepsilon_i] = 0$.

When $k_x \leq k_y$, then test for $H_0 : \mathrm{r}(B') = k_x - 1$, is same as test for $H_0 : \mathrm{r}(C) = k_x - 1$ in

$$X = YC + V$$

with auxiliary model

$$\begin{aligned} x_1 &= X_2\delta + \varepsilon \\ X_2 &= YC_2 + V_2 \end{aligned}$$

General Rank Tests

- The updated RANKTEST command (Kleibergen, Schaffer, Windmeijer 2020) incorporates these developments, in particular now has the CUE version of the CD rank test, and the J2L alternative to the KP test.
- Examples of applications to asset pricing models in paper and help file.
- Direct access to CUE estimator and standard asymptotic variance estimator.
- Gives access to I-test of Arellano, Hansen and Sentana (2012)
- Can fully replicate the overspecification/multiple solutions CUE analysis of Manresa, Peñaranda and Sentana (2017)

Conclusion

- Unifying framework for testing underidentification/rank of a matrix
- These are tests for overidentification in auxiliary model
- This simple structure makes it possible to construct tests for underidentification where they have not been applied before, like dynamic panel data models estimated by GMM
- RANKTEST and UNDERID have incorporated these insights for the Stata user community to use