Instrumental Variables and GMM: Estimation and Testing

Christopher F Baum, Boston College

Mark E. Schaffer, Heriot–Watt University

Steven Stillman, New Zealand Department of Labour

March 2003

Stata Journal, 2003, 3-1, 1-31

http://fmwww.bc.edu/ec-p/WP545.pdf

Instrumental Variables and GMM: Estimation and Testing

In this paper, which has appeared in the current issue of *Stata Journal*, we describe several Stata routines that we have written to facilitate instrumental variables estimation, going beyond the capabilities of Stata's `ivreg` command. In the presentation today, I will mention the highlights of this paper, and encourage you to read it—either from the *Stata Journal*, or from the RePEc series set up for these meetings.

These routines—centered around our enhanced estimation routine `ivreg2`—implement a form of the Generalised Method of Moments (GMM) estimator, previously available in `ivgmm0`, that generates efficient estimates in the presence of heteroskedasticity of unknown form. In contrast, the conventional IV estimator with `robust` standard errors, although consistent, is relatively inefficient.

In the twenty years since it was first introduced, GMM has become a very popular tool among empirical researchers. It is also a very useful heuristic tool. Many standard estimators, including IV and OLS, can be seen as special cases of GMM estimators, and are often presented as such in first–year graduate econometrics texts. Most of the diagnostic tests we discuss in this paper can also be cast in a GMM framework.

We also discuss the problems encountered in the presence of intra–group correlation or “clustering”. If the error terms in the regression are correlated within groups, but not correlated across groups, then the consequences for IV estimation are similar to those of heteroskedasticity: the IV coefficient estimates are consistent, but their standard errors and the usual forms of the diagnostic tests are not.

Efficient GMM brings with it the advantage of consistency in the presence of arbitrary heteroskedasticity, but at a cost of possibly poor finite sample performance. If heteroskedasticity is in fact not present, then standard IV may be preferable. The usual Breusch–Pagan/Godfrey/Cook–Weisberg and White/Koenker tests for the presence of heteroskedasticity in a regression equation can be applied to an IV regression only under restrictive assumptions. We discuss the test of Pagan and Hall (1983) designed specifically for detecting the presence of heteroskedasticity in IV estimation, and its relationship to these other heteroskedasticity tests.

Even when IV or GMM is judged to be the appropriate estimation technique, we may still question its validity in a given application: are our instruments "good instruments"? "Good instruments" should be both relevant and valid: correlated with the endogenous regressors and at the same time orthogonal to the errors. Correlation with the endogenous regressors can be assessed by an examination of the significance of the excluded instruments in the first–stage IV regressions. We may cast some light on whether the instruments satisfy the orthogonality conditions in the context of an overidentified model: that is, one in which a surfeit of instruments are available. In that context we may test the overidentifying restrictions in order to provide some evidence of the instruments' validity. We present the variants of this test due to Sargan (1958), Basmann (1960) and, in the GMM context, L. Hansen (1982), and show how the generalization of this test, the C or "difference–in–Sargan" test, can be used test the validity of subsets of the instruments.

Although there may well be reason to suspect non–orthogonality between regressors and errors, the use of IV estimation to address this problem must be balanced against the inevitable loss of efficiency vis–à–vis OLS. It is therefore very useful to have a test of whether or not OLS is inconsistent and IV or GMM is required. This is the Durbin–Wu–Hausman (DWH) test of the endogeneity of regressors. We discuss how to implement variants of the DWH test, and how the test can be generalized to test the endogeneity of subsets of regressors. We then show how the Hausman form of the test can be applied in the GMM context, how it can be interpreted as a GMM test, when it will be identical to the Hansen/Sargan/C-test statistic, and when the two test statistics will differ.

We have written four Stata commands—`ivreg2`, `ivhettest`, `overid`, and `ivendog`—that, together with Stata's built-in commands, allow the user to implement all of the above estimators and diagnostic tests.

They are all meant for application in a cross–sectional data context. In our current research, we are considering the extension of these routines to a time–series context, and to a panel data setting.

The equation to be estimated is, in matrix notation,

$$y = X\beta + u, E(uu') = \Omega \tag{1}$$

The matrix of regressors X is $n \times K$, where n is the number of observations. Some of the regressors are endogenous, so that $E(X_i u_i) \neq 0$. We partition the set of regressors into $[X_1 \;\; X_2]$, with the K_1 regressors X_1 assumed under the null to be endogenous, and the $(K - K_1)$ remaining regressors X_2 assumed exogenous.

The set of instrumental variables is Z and is $n \times L$; this is the full set of variables that are assumed to be exogenous, i.e., $E(Z_i u_i) = 0$. We partition the instruments into $[Z_1 \ Z_2]$, where the L_1 instruments Z_1 are excluded instruments, and the remaining $(L - L_1)$ instruments $Z_2 \equiv X_2$ are the included instruments/exogenous regressors:

$$\text{Regressors } X = [X_1 \ \ X_2] = [X_1 \ \ Z_2] = [\text{Endogenous} \quad \text{Exogenous}] \tag{2}$$

$$\text{Instruments } Z = [Z_1 \ Z_2] = [\text{Excluded} \quad \text{Included}] \tag{3}$$

Denote by P_Z the projection matrix $Z(Z'Z)^{-1}Z'$. The instrumental variables estimator of β is

$$\widehat{\beta}_{IV} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y = (X'P_ZX)^{-1}X'P_Z\,y \tag{4}$$

This estimator goes under a variety of names: the instrumental variables (IV) estimator, the generalized instrumental variables estimator (GIVE), or the two-stage least-squares (2SLS) estimator, the last reflecting the fact that the estimator can be calculated in a two–step procedure. We follow Davidson and MacKinnon and refer to it as the IV estimator rather than 2SLS because the basic idea of instrumenting is central, and because it can be (and in Stata, is more naturally) calculated in one step as well as in two.

The standard IV estimator is a special case of a Generalized Method of Moments (GMM) estimator. The assumption that the instruments Z are exogenous can be expressed as $E(Z_i u_i) = 0$. The L instruments give us a set of L moments,

$$g_i(\widehat{\beta}) = Z_i' \widehat{u}_i = Z_i'(y_i - X_i \widehat{\beta}) \tag{5}$$

where g_i is $L \times 1$. The exogeneity of the instruments means that there are L moment conditions, or orthogonality conditions, that will be satisfied at the true value of β. Each of the L moment equations corresponds to a sample moment, and we write these L sample moments as

$$\overline{g}(\widehat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} g_i(\widehat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Z_i'(y_i - X_i \widehat{\beta}) = \frac{1}{n} Z' \widehat{u} \tag{6}$$

The intuition behind GMM is to choose an estimator for β that solves $\overline{g}(\widehat{\beta}) = 0$.

If the equation to be estimated is exactly identified, so that $L = K$, then we have as many equations—the L moment conditions—as we do unknowns—the K coefficients in $\widehat{\beta}$. In this case it is possible to find a $\widehat{\beta}$ that solves $\overline{g}(\beta) = 0$, and this GMM estimator is in fact the IV estimator.

If the equation is overidentified, however, so that $L > K$, then we have more equations than we do unknowns, and in general it will not be possible to find a $\widehat{\beta}$ that will set all L sample moment conditions to exactly zero. In this case, we take an $L \times L$ weighting matrix W and use it to construct a quadratic form in the moment conditions. This gives us the GMM objective function:

$$J(\widehat{\beta}) = n\overline{g}(\widehat{\beta})'W\overline{g}(\widehat{\beta}) \tag{7}$$

A GMM estimator for β is the $\widehat{\beta}$ that minimizes $J(\widehat{\beta})$. Deriving and solving the K first order conditions

$$\frac{\partial J(\widehat{\beta})}{\partial \widehat{\beta}} = 0 \tag{8}$$

yields the GMM estimator:

$$\widehat{\beta}_{GMM} = (X'ZWZ'X)^{-1}X'ZWZ'y \tag{9}$$

What is the optimal choice of weighting matrix? Denote by S the covariance matrix of the moment conditions g:

$$S = \frac{1}{n}E(Z'uu'Z) = \frac{1}{n}E(Z'\Omega Z) \tag{10}$$

where S is an $L \times L$ matrix. The general formula for the distribution of a GMM estimator is

$$V(\widehat{\beta}_{GMM}) = \frac{1}{n}(Q'_{XZ}WQ_{XZ})^{-1}(Q'_{XZ}WSWQ_{XZ})(Q'_{XZ}WQ_{XZ})^{-1} \tag{11}$$

The *efficient* GMM estimator is the GMM estimator with an optimal weighting matrix W, one which minimizes the asymptotic variance of the estimator. This is achieved by choosing $W = S^{-1}$. Substitute this into Equation (9) and Equation (11) and we obtain the efficient GMM estimator

$$\widehat{\beta}_{EGMM} = (X'ZS^{-1}Z'X)^{-1}X'ZS^{-1}Z'y \tag{12}$$

with asymptotic variance

$$V(\widehat{\beta}_{EGMM}) = \frac{1}{n}(Q'_{XZ}S^{-1}Q_{XZ})^{-1} \tag{13}$$

Note the generality (the “G” of GMM) of the treatment thus far; we have not yet made any assumptions about Ω, the covariance matrix of the disturbance term. But the efficient GMM estimator is not yet a feasible estimator, because the matrix S is not known. To be able to implement the estimator, we need to estimate S, and to do this, we need to make some assumptions about Ω.

If we consider heteroskedasticity of unknown form (but no clustering), the squared IV residuals may be used as consistent estimates of the squared errors, and a consistent estimator of S is

$$\hat{S} = \frac{1}{n}(Z'\hat{\Omega}Z) \tag{14}$$

This works because, although we cannot hope to estimate the n diagonal elements of Ω with only n observations, they are sufficient to enable us to obtain a consistent estimate of the $L \times L$ matrix S.

To GMM or not to GMM?

The advantages of GMM over IV are clear: if heteroskedasticity is present, the GMM estimator is more efficient than the simple IV estimator, whereas if heteroskedasticity is not present, the GMM estimator is no worse asymptotically than the IV estimator.

Nevertheless, the use of GMM does come with a price. The problem, as Hayashi (2000) points out, is that the optimal weighting matrix $\widehat{S}$ at the core of efficient GMM is a function of fourth moments, and obtaining reasonable estimates of fourth moments may require very large sample sizes. The consequence is that the efficient GMM estimator can have poor small sample properties. In particular, Wald tests tend to over–reject the null.

If in fact the error is homoskedastic, IV would be preferable to efficient GMM. For this reason a test for the presence of heteroskedasticity when one or more regressors is endogenous may be useful in deciding whether IV or GMM is called for. Such a test was proposed by Pagan and Hall (1983), and we have implemented it in Stata as `ivhettest`.

The Breusch–Pagan/Godfrey/Cook–Weisberg and White/Koenker statistics are standard tests of the presence of heteroskedasticity in an OLS regression. The principle is to test for a relationship between the residuals of the regression and p indicator variables that are hypothesized to be related to the heteroskedasticity. Koenker (1981) noted that the power of this test is very sensitive to the normality assumption, and presented a version of the test that relaxed this assumption. Koenker's test statistic is based on the centered R^2 from an auxiliary regression of the squared residuals from the original regression on the indicator variables. When the indicator variables are the regressors of the original equation, their squares and their cross-products, Koenker's test is identical to White's (1980) nR_c^2 general test for heteroskedasticity. These tests are available in Stata, following estimation with `regress`, using our `ivhettest` as well as via `hettest` and `whitetst`.

As Pagan and Hall (1983) point out, the above tests will be valid tests for heteroskedasticity in an IV regression only if heteroskedasticity is present in that equation and *nowhere else in the system*. The other structural equations in the system (corresponding to the endogenous regressors X_1) must also be homoskedastic, even though they are not being explicitly estimated. Pagan and Hall derive a test which relaxes this requirement, and we have implemented their simpler test as `ivhettest`. The Pagan–Hall statistic has not been widely used in practice, perhaps because it is not a standard feature of most regression packages.

Testing the relevance of instruments

An instrumental variable must satisfy two requirements: it must be correlated with the included endogenous variable(s), and orthogonal to the error process. The former condition may be readily tested by examining the fit of the first stage regressions. To illustrate the pitfalls facing empirical researchers here, consider the following simple example.

The researcher has a model with two endogenous regressors and two excluded instruments. One of the two excluded instruments is highly correlated with each of the two endogenous regressors, but the other excluded instrument is just noise. The model is therefore basically unidentified: there is one good instrument but two endogenous regressors. But the Bound–Jaeger–Baker $F-$statistics and partial R^2 measures from the two first–stage regressions will not reveal this weakness. When multiple endogenous regressors are used, other statistics are required. One such statistic has been proposed by Shea (1997): a “partial R^2” measure that takes the intercorrelations among the instruments into account. We have implemented both the Bound et al. and Shea statistics as options on the `ivreg2` command.

We turn now to the second requirement for an instrumental variable. How can the instrument's independence from an unobservable error process be ascertained? If (and only if) we have a surfeit of instruments—i.e., if the equation is overidentified—then we can test the corresponding moment conditions: that is, whether the instruments are uncorrelated with the error process.

In the context of GMM, the overidentifying restrictions may be tested via the commonly employed J statistic of L. Hansen (1982). In the IV context, this statistic is known as the Sargan (1958) statistic, or the Basmann (1960) statistic. These statistics are produced by `ivreg2`.

The Hansen–Sargan tests for overidentification presented above evaluate the entire set of overidentifying restrictions. In a model containing a very large set of excluded instruments, such a test may have very little power. Another common problem arises when the researcher has prior suspicions about the validity of a subset of instruments, and wishes to test them.

In these contexts, a "difference–in–Sargan" statistic may usefully be employed. The C test allows us to test a subset of the original set of orthogonality conditions. The statistic is computed as the difference between two Sargan statistics (or, for efficient GMM, two J statistics): that for the (restricted, fully efficient) regression using the entire set of overidentifying restrictions, versus that for the (unrestricted, inefficient but consistent) regression using a smaller set of restrictions, in which a specified set of instruments are removed from the set. The C test is conducted in `ivreg2` by specifying the `orthog` option.

Lastly, we have also implemented a test for the endogeneity of the regressors, as an alternative to the use of the Hausman test. Application of the latter test (via `hausman`) requires estimating the model both via OLS and IV. The alternative approach involves estimating the less efficient but consistent model via `ivreg`, and using our `ivendog` routine to evaluate the exogeneity of some or all of the endogenous regressors. If `ivreg2` is employed, the same test may be calculated with the `orthog` option. The latter approach is of particular interest in the context of heteroskedasticity, for which `hausman` will often generate negative test statistics, and may miscalculate the degrees of freedom of the test (which may indeed be unknown; see the paper for the gory details).

In summary, the `ivreg2` suite of programs provide a number of state–of–the–art techniques for the estimation and testing of instrumental variables regression models. We appreciate your feedback on their features and utility.