The Stata Journal (2003)
3, Number 1, pp. 1–31

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

Abstract. We discuss instrumental variables (IV) estimation in the broader context of the generalized method of moments (GMM), and describe an extended IV estimation routine that provides GMM estimates as well as additional diagnostic tests. Stand-alone test procedures for heteroskedasticity, overidentification, and endogeneity in the IV context are also described.

Keywords: st0030, instrumental variables, generalized method of moments, endogeneity, heteroskedasticity, overidentifying restrictions, clustering, intra-group correlation

1 Introduction

The application of the instrumental variables (IV) estimator in the context of the classical linear regression model, from a textbook context, is quite straightforward: if the error distribution cannot be considered independent of the regressors' distribution, IV is called for, using an appropriate set of instruments. But applied researchers often must confront several hard choices.

An omnipresent problem in empirical work is heteroskedasticity. Although the consistency of the IV coefficient estimates is not affected by the presence of heteroskedasticity, the standard IV estimates of the standard errors are inconsistent, preventing valid inference. The usual forms of the diagnostic tests for endogeneity and overidentifying restrictions will also be invalid if heteroskedasticity is present. These problems can be partially addressed through the use of heteroskedasticity-consistent or "robust" standard errors and statistics. The conventional IV estimator (though consistent) is, however, inefficient in the presence of heteroskedasticity. The usual approach today when facing heteroskedasticity of unknown form is to use the generalized method of moments (GMM), introduced by Hansen (1982). GMM makes use of the orthogonality conditions to allow for efficient estimation in the presence of heteroskedasticity of unknown form.

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 begin, therefore, with a short presentation of IV and GMM estimation in Section 2. We include here a

discussion 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. We discuss how clustering can be interpreted in the GMM context and how it can be dealt with in Stata to make efficient estimation, valid inference, and diagnostic testing possible.

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. In Section 3, 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"? This is the question we address in Section 4. "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, Hansen (1982), and show how the generalization of this test, the C or "difference-in-Sargan" test, can be used to test the validity of subsets of the instruments.

Although there may well be reason to suspect nonorthogonality 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. In Section 5, 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. The syntax diagrams for these commands are presented in the last section of the paper, and the electronic supplement presents annotated examples of their use.

2 IV and GMM estimation

The "Generalized Method of Moments" was introduced by L. Hansen in his celebrated 1982 paper. There are a number of good modern texts that cover GMM, and one recent prominent text, Hayashi (2000), presents virtually all the estimation techniques discussed in the GMM framework. A concise online text that covers GMM is Hansen (2000). The exposition below draws on Hansen (2000, Chapter 11); Hayashi (2000, Chapter 3); Wooldridge (2002, Chapter 8); Davidson and MacKinnon (1993); and Greene (2000).

We begin with the standard IV estimator, and then relate it to the GMM framework. We then consider the issue of clustered errors, and finally turn to OLS.

2.1 The method of instrumental variables

The equation to be estimated is, in matrix notation,

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

with typical row

$$y_i = X_i\beta + u_i \tag{2}$$

The matrix of regressors X is $n \times K$, where n is the number of observations. The error term u is distributed with mean zero and the covariance matrix Ω is $n \times n$. Three special cases for Ω that we will consider are

$$\text{Homoskedasticity:} \qquad \Omega = \sigma^2 I \tag{3}$$

$$\text{Heteroskedasticity:} \qquad \Omega = \begin{pmatrix} \sigma_1^2 & & & & 0 \\ & \ddots & & & \\ & & \sigma_i^2 & & \\ & & & \ddots & \\ 0 & & & & \sigma_n^2 \end{pmatrix} \tag{4}$$

$$\text{Clustering:} \qquad \Omega = \begin{pmatrix} \Sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \Sigma_m & & \\ & & & \ddots & \\ 0 & & & & \Sigma_M \end{pmatrix} \tag{5}$$

where Σ_m indicates an intra-cluster covariance matrix. For cluster m with t observations, Σ_m will be $t \times t$. Zero covariance between observations in the M different clusters gives the covariance matrix Ω, in this case, a block-diagonal form.

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:

$$\text{Regressors } X = [X_1 \;\; X_2] = [X_1 \;\; Z_2] = [\text{Endogenous} \quad \text{Exogenous}] \tag{6}$$

$$\text{Instruments } Z = [Z_1 \; Z_2] = [\text{Excluded} \quad \text{Included}] \tag{7}$$

The order condition for identification of the equation is $L \geq K$; there must be at least as many excluded instruments as there are endogenous regressors. If $L = K$, the equation is said to be "exactly identified"; if $L > K$, the equation is "overidentified".

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

$$\widehat{\beta}_{\text{IV}} = \{X'Z(Z'Z)^{-1}Z'X\}^{-1}X'Z(Z'Z)^{-1}Z'y = (X'P_Z X)^{-1}X'P_Z\, y \tag{8}$$

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 (1993, 220) 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 asymptotic distribution of the IV estimator under the assumption of conditional homoskedasticity (3) can be written as follows. Let

$$Q_{XZ} = E(X_i' Z_i) \tag{9}$$

$$Q_{ZZ} = E(Z_i' Z_i) \tag{10}$$

and let $\widehat{u}$ denote the IV residuals,

$$\widehat{u} \equiv y - X\widehat{\beta}_{\text{IV}} \tag{11}$$

Then, the IV estimator is asymptotically distributed as $\widehat{\beta}_{\text{IV}} \overset{A}{\sim} N\{\beta, V(\widehat{\beta}_{\text{IV}})\}$, where

$$V(\widehat{\beta}_{\text{IV}}) = \frac{1}{n}\sigma^2(Q_{XZ}'Q_{ZZ}^{-1}Q_{XZ})^{-1} \tag{12}$$

Replacing Q_{XZ}, Q_{ZZ} and σ^2 with their sample estimates

$$\overline{Q}_{XZ} = \frac{1}{n}X'Z \tag{13}$$

$$\overline{Q}_{ZZ} = \frac{1}{n} Z'Z \tag{14}$$

$$\widehat{\sigma}^2 = \frac{\widehat{u}'\widehat{u}}{n} \tag{15}$$

we obtain the estimated asymptotic variance–covariance matrix of the IV estimator:

$$V(\widehat{\beta}_{\text{IV}}) = \widehat{\sigma}^2 \{X'Z(Z'Z)^{-1}Z'X\}^{-1} = \widehat{\sigma}^2 (X'P_Z X)^{-1} \tag{16}$$

Note that some packages, including Stata's `ivreg`, include a degrees-of-freedom correction to the estimate of $\widehat{\sigma}^2$ by replacing n with $n-L$. This correction is not necessary, however, since the estimate of $\widehat{\sigma}^2$ would not be unbiased anyway (Greene 2000, 373). Our `ivreg2` routine defaults to the large-sample formulas for the estimated error variance and covariance matrix; the user can request the small-sample versions with the option `small`.

2.2 The generalized method of moments

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,

$$g_i(\widehat{\beta}) = Z_i'\widehat{u}_i = Z_i'(y_i - X_i\widehat{\beta}) \tag{17}$$

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:

$$E\{g_i(\beta)\} = 0 \tag{18}$$

Each of the L moment equations corresponds to a sample moment, and we write these L sample moments as

$$\overline{g}(\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Z_i'(y_i - X_i\widehat{\beta}) = \frac{1}{n} Z'\widehat{u} \tag{19}$$

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:

$$J(\widehat{\beta}) = n\overline{g}(\widehat{\beta})' W \overline{g}(\widehat{\beta}) \tag{20}$$

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

$$\frac{\partial J(\widehat{\beta})}{\partial \widehat{\beta}} = 0 \tag{21}$$

yields the GMM estimator:

$$\widehat{\beta}_{GMM} = (X'ZWZ'X)^{-1}X'ZWZ'y \tag{22}$$

Note that the results of the minimization, and hence the GMM estimator, will be the same for weighting matrices that differ by a constant of proportionality (we will make use of this fact below). Beyond this, however, there are as many GMM estimators as there are choices of weighting matrix W.

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

$$S = \frac{1}{n}E(Z'uu'Z) = \frac{1}{n}E(Z'\Omega Z) \tag{23}$$

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

$$V(\widehat{\beta}_{GMM}) = \frac{1}{n}(Q'_{XZ}WQ_{XZ})^{-1}(Q'_{XZ}WSWQ_{XZ})(Q'_{XZ}WQ_{XZ})^{-1} \tag{24}$$

The *efficient* GMM estimator is the GMM estimator with an optimal weighting matrix W, one that minimizes the asymptotic variance of the estimator. This is achieved by choosing $W = S^{-1}$. Substitute this into (22) and (24), and we obtain the efficient GMM estimator

$$\widehat{\beta}_{\text{EGMM}} = (X'ZS^{-1}Z'X)^{-1}X'ZS^{-1}Z'y \tag{25}$$

with asymptotic variance

$$V(\widehat{\beta}_{\text{EGMM}}) = \frac{1}{n}(Q'_{XZ}S^{-1}Q_{XZ})^{-1} \tag{26}$$

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. However, 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.

2.3 GMM and heteroskedastic errors

Let us start with one of the most commonly encountered cases in cross-section analysis: heteroskedasticity of unknown form, but no clustering (4). We need a heteroskedasticity-consistent estimator of S. Such an $\widehat{S}$ is available by using the standard "sandwich" approach to robust covariance estimation. Denote by $\widehat{\Omega}$ the diagonal matrix of squared residuals,

$$\widehat{\Omega} = \begin{pmatrix} \widehat{u}_1^2 & & & & 0 \\ & \ddots & & & \\ & & \widehat{u}_i^2 & & \\ & & & \ddots & \\ 0 & & & & \widehat{u}_n^2 \end{pmatrix} \tag{27}$$

where $\widehat{u}_i$ is a consistent estimate of u_i. Then a consistent estimator of S is

$$\widehat{S} = \frac{1}{n}(Z'\widehat{\Omega}Z) \tag{28}$$

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.

The $\widehat{u}$ used for the matrix in (27) can come from any consistent estimator of β; efficiency is not required. In practice, the most common choice for estimating $\widehat{u}$ is the IV residuals. This gives us the algorithm for the *feasible efficient two-step* GMM *estimator*, as implemented in `ivreg2,gmm` and `ivgmm0`:[1]

1. Estimate the equation using IV.
2. Form the residuals $\widehat{u}$. Use these to form the optimal weighting matrix $\widehat{W} = \widehat{S}^{-1} = \left\{1/n(Z'\widehat{\Omega}Z)\right\}^{-1}$.
3. Calculate the efficient GMM estimator $\widehat{\beta}_{\text{EGMM}}$ and its variance-covariance matrix using the estimated optimal weighting matrix and (25), (26), and (13). This yields

$$\widehat{\beta}_{\text{EGMM}} = \{X'Z(Z'\widehat{\Omega}Z)^{-1}Z'X\}^{-1}X'Z(Z'\widehat{\Omega}Z)^{-1}Z'y \tag{29}$$

   with asymptotic variance

$$V(\widehat{\beta}_{\text{EGMM}}) = \{X'Z(Z'\widehat{\Omega}Z)^{-1}Z'X\}^{-1} \tag{30}$$

A variety of other feasible GMM procedures are also possible. For example, the procedure above can be iterated by obtaining the residuals from the two-step GMM estimator, using these to calculate a new $\widehat{S}$, using this in turn to calculate the three-step feasible efficient GMM estimator, and so forth, for as long as the user wishes or until the estimator converges; this is the "iterated GMM estimator".[2]

[1] This estimator goes under various names: "2-stage instrumental variables"(2SIV), White (1982); "2-step 2-stage least squares", Cumby et al. (1983); and "heteroskedastic 2-stage least squares" (H2SLS), Davidson and MacKinnon (1993, 599).

[2] Another approach is to choose a different consistent but inefficient Step 1 estimator for the calculation of residuals used in Step 2. One common alternative to IV as the initial estimator is to use the residuals from the GMM estimator that uses the identity matrix as the weighting matrix. Alternatively, one may work directly with the GMM objective function. Note that the estimate of the optimal weighting matrix is derived from some $\widehat{\beta}$. Instead of first obtaining an optimal weighting matrix and then taking it as given when maximizing (20), we can write the optimal weighting matrix as a function of $\widehat{\beta}$, and choose $\widehat{\beta}$ to maximize $J(\widehat{\beta}) = n\overline{g}_n(\widehat{\beta})'W(\widehat{\beta})\overline{g}_n(\widehat{\beta})$. This is the "continuously updated GMM" of Hansen et al. (1996); it requires numerical optimization methods.

2.4 GMM, IV and homoskedastic versus heteroskedastic errors

Let us now see what happens if we impose the more restrictive assumption of conditional homoskedasticity on Ω (3). This means that the S matrix simplifies to

$$S = \frac{1}{n}E(Z'\Omega Z) = \sigma^2\frac{1}{n}E(Z'Z) \tag{31}$$

The expectation term in (31) can be estimated by $(1/n)Z'Z$, but what about σ^2? As we noted above, the GMM estimator will be the same for weighting matrices that differ by a constant of proportionality. We can therefore obtain the efficient GMM estimator under conditional homoskedasticity if we simply ignore σ^2 and use as our weighting matrix

$$\widehat{W} = \left(\frac{1}{n}Z'Z\right)^{-1} \tag{32}$$

Substituting (32) into (22), we find that it reduces to the formula for the IV estimator in (8). To obtain the variance of the estimator, however, we *do* need an estimate of σ^2. If we use the residuals of the IV estimator to calculate $\widehat{\sigma}^2 = (1/n)\widehat{u}'\widehat{u}$, we obtain

$$\widehat{S} = \widehat{\sigma}^2\frac{1}{n}Z'Z \tag{33}$$

Finally, if we now set

$$\widehat{W} = \widehat{S}^{-1} = \left(\widehat{\sigma}^2\frac{1}{n}Z'Z\right)^{-1} \tag{34}$$

and substitute (34) into the formula for the asymptotic variance of the efficient GMM estimator (26), we find that it reduces to the formula for the asymptotic variance of the IV estimator (12). In effect, under the assumption of conditional homoskedasticity, the (efficient) iterated GMM estimator is the IV estimator, and the iterations converge after one step.[3]

What are the implications of heteroskedasticity for the IV estimator? Recall that in the presence of heteroskedasticity, the IV estimator is inefficient but consistent, whereas the standard estimated IV covariance matrix is inconsistent. Asymptotically correct inference is still possible, however. In these circumstances the IV estimator is a GMM estimator with a suboptimal weighting matrix, and hence the general formula for the asymptotic variance of a general GMM estimator, (24), still holds. The IV weighting matrix $\widehat{W}$ remains as in (32); what we need is a consistent estimate of $\widehat{S}$. This is easily done, using exactly the same method employed in two-step efficient GMM. First, form

[3] It is worth noting that the IV estimator is not the only such efficient GMM estimator under conditional homoskedasticity. Instead of treating $\widehat{\sigma}^2$ as a parameter to be estimated in a second stage, what if we return to the GMM criterion function and minimize by simultaneously choosing $\widehat{\beta}$ *and* $\widehat{\sigma}^2$? The estimator that solves this minimization problem is in fact the Limited Information Maximum Likelihood estimator (LIML). In effect, under conditional homoskedasticity, the continuously updated GMM estimator is the LIML estimator. Calculating the LIML estimator does not require numerical optimization methods, since it can be calculated as the solution to an eigenvalue problem; for example, see Davidson and MacKinnon (1993, 644–51).

the "hat" matrix $\widehat{\Omega}$ as in (27), using the IV residuals, and use this matrix to form the $\widehat{S}$ matrix as in (28). Substitute this $\widehat{S}$, the (suboptimal) IV weighting matrix $\widehat{W}$ (32), and the sample estimates of Q_{XZ} (13) and Q_{ZZ} (14) into the general formula for the asymptotic variance of a GMM estimator (24), and we obtain an estimated variance–covariance matrix for the IV estimator that is robust to the presence of heteroskedasticity:

$$\text{Robust } V(\widehat{\beta}_{\text{IV}}) = (X'P_Z X)^{-1}\{X'Z(Z'Z)^{-1}(Z'\widehat{\Omega}Z)(Z'Z)^{-1}Z'X\}(X'P_Z X)^{-1} \quad (35)$$

This is in fact the usual Eicker–Huber–White "sandwich" robust variance–covariance matrix for the IV estimator, available from `ivreg` or `ivreg2` with the `robust` option.

2.5 Clustering, robust covariance estimation, and GMM

We turn now to the third special form of the disturbance covariance matrix Ω, clustering. Clustering arises very frequently in cross-section and panel data applications. For example, it may be reasonable to assume that observations on individuals drawn from the same family (cluster) are correlated with each other, but observations on individuals from different families are not. In the panel context, it may be reasonable to assume that observations on the same individual (cluster) in two different time periods are correlated, but observations on two different individuals are not.

As specified in (5), the form of clustering is very general. The intra-cluster correlation Σ_m can be of any form, be it serial correlation, random effects, or anything else. The Σ_m's may, moreover, vary from cluster to cluster (the cluster analog to heteroskedasticity). Even in these very general circumstances, however, efficient estimation and consistent inference is still possible.

As usual, what we need is a consistent estimate of S. Denote by u_m the vector of disturbances for cluster m; if there are t observations in the cluster, then u_m is $t \times 1$. Let $\widehat{u}_m$ be some consistent estimate of u_m. Finally, define $\widehat{\Sigma}_m \equiv \widehat{u}_m \widehat{u}_m'$. If we now define $\widehat{\Omega}_C$ as the block-diagonal form

$$\widehat{\Omega}_C = \begin{pmatrix} \widehat{\Sigma}_1 & & & 0 \\ & \ddots & & \\ & & \widehat{\Sigma}_m & \\ & & & \ddots \\ 0 & & & \widehat{\Sigma}_M \end{pmatrix} \quad (36)$$

then an estimator of S that is consistent in the presence of arbitrary intra-cluster correlation is

$$\widehat{S} = \frac{1}{n}(Z'\widehat{\Omega}_C Z) \quad (37)$$

The earliest reference to this approach to robust estimation in the presence of clustering of which we are aware is White (1984, 135–136). It is commonly employed in the context of panel data estimation; see Wooldridge (2002, 193), Arellano (1987), and

Kézdi (2002). It is the standard Stata approach to clustering, implemented in, for example, `_robust`, `regress` and `ivreg2`.[4]

The cluster-robust covariance matrix for IV estimation is obtained exactly as in the preceding subsection except using $\widehat{S}$ as defined in (37). This generates the robust standard errors produced by `ivreg` and `ivreg2` with the `cluster` option. Similarly, GMM estimates that are efficient in the presence of arbitrary intra-cluster correlation are obtained exactly as in Subsection 2.3, except using the cluster-robust estimate of $\widehat{S}$. This efficient GMM estimator is a useful alternative to the fixed or random effects IV estimators available from Stata's `xtivreg` because it relaxes the constraint imposed by the latter estimators that the correlation of individual observations within a group is constant.

It is important to note here that, just as we require a reasonable number of diagonal elements (observations) for the usual "hat" matrix $\widehat{\Omega}$, we also require a reasonable number of diagonal elements (clusters) for $\widehat{\Omega}_C$. An extreme case is where the number of clusters M is $\leq K$. When this is the case, $\text{rank}(\widehat{S}) = M \leq K = \text{rank}(Z'Z)$. At this point, `ivreg2` will either refuse to report standard errors (in the case of IV estimation) or exit with an error message (in the case of GMM estimation). But users should take care that, if the `cluster` option is used, then it ought to be the case that $M >> K$.[5]

2.6 GMM, OLS, and Heteroskedastic OLS (HOLS)

Our final special case of interest is OLS. It is not hard to see that under conditional homoskedasticity and the assumption that all the regressors are exogenous, OLS is an efficient GMM estimator. If the disturbance is heteroskedastic, OLS is no longer efficient, but correct inference is still possible through the use of the Eicker–Huber–White "sandwich" robust covariance estimator, and this estimator can also be derived using the general formula for the asymptotic variance of a GMM estimator with a suboptimal weighting matrix, (24).

A natural question is whether a more efficient GMM estimator exists, and the answer is "yes" (Chamberlain (1982) and Cragg (1983)). If the disturbance is heteroskedastic, there are no endogenous regressors, and the researcher has available additional moment conditions, i.e., additional variables that do not appear in the regression but that are known to be exogenous, then the efficient GMM estimator is that of Cragg (1983), dubbed "heteroskedastic OLS" (HOLS) by Davidson and MacKinnon (1993, 600). It can be obtained in precisely the same way as feasible efficient two-step GMM, except now the first-step inefficient but consistent estimator used to generate the residuals is OLS rather than IV. This estimator can be obtained using `ivreg2` by specifying the `gmm` option, an

[4] There are other approaches to dealing with clustering that put more structure on the Ω matrix and hence are more efficient but less robust. For example, the Moulton (1986) approach to obtaining consistent standard errors is in effect to specify an "error components" (a.k.a. "random effects") structure in (36): Σ_m is a matrix with diagonal elements $\sigma_u^2 + \sigma_v^2$ and off-diagonal elements σ_v^2. This is then used with (24) to obtain a consistent estimate of the covariance matrix.

[5] Stata's official `ivreg` is perhaps excessively forgiving in this regard, and will indicate error only if $M \leq L$; i.e., the number of regressors exceeds the number of clusters.

empty list of endogenous regressors, and the additional exogenous variables in the list of excluded instruments. If the `gmm` option is omitted, OLS estimates are reported.

2.7 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 (215), 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 (good news for the unscrupulous investigator in search of large t statistics, perhaps, but not for the rest of us). 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`. We describe this test in the next section.

3 Testing for heteroskedasticity

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. Breusch and Pagan (1979), Godfrey (1978), and Cook and Weisberg (1983) separately derived the same test statistic. This statistic is distributed as χ^2 with p degrees of freedom under the null of no heteroskedasticity, and under the maintained hypothesis that the error of the regression is normally distributed. 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, also distributed as χ^2_p under the null, is easily obtained as nR^2_c, where R^2_c is 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 nR^2_c general test for heteroskedasticity (White 1980). 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.[6] Pagan and Hall derive a test that relaxes this requirement. Under the null of homoskedasticity in the IV regression, the Pagan–Hall statistic is distributed as χ^2_p, irrespective of the presence of heteroskedasticity elsewhere in the system. A more general form of this test was separately proposed by White (1982). Our implementation is of the simpler Pagan–Hall statistic, available with the command `ivhettest` after estimation by `ivreg`, `ivreg2`, or `ivgmm0`. We present the Pagan–Hall test here in the format and notation of the original White (1980 and 1982) tests, however, to facilitate comparisons with the other tests noted above.[7]

Let Ψ be the $n \times p$ matrix of indicator variables hypothesized to be related to the heteroskedasticity in the equation, with typical row Ψ_i. These indicator variables must be exogenous, typically either instruments or functions of the instruments. Common choices would be

1. The levels, squares, and cross-products of the instruments Z (excluding the constant), as in the White (1980) test. This is the default in `ivhettest`.
2. The levels only of the instruments Z (excluding the constant). This is available in `ivhettest` by specifying the `ivlev` option.
3. The "fitted value" of the dependent variable. This is *not* the usual fitted value of the dependent variable, $X\widehat{\beta}$. It is, rather, $\widehat{X}\widehat{\beta}$; i.e., the prediction based on the IV estimator $\widehat{\beta}$, the exogenous regressors Z_2, and the fitted values of the endogenous regressors $\widehat{X}_1$. This is available in `ivhettest` by specifying the `fitlev` option.
4. The "fitted value" of the dependent variable and its square (`fitsq` option).

The trade-off in the choice of indicator variables is that a smaller set of indicator variables will conserve degrees of freedom, at the cost of being unable to detect heteroskedasticity in certain directions.

(*Continued on next page*)

[6] For a more detailed discussion, see Pagan and Hall (1983) or Godfrey (1988, 189–90).

[7] We note here that the original Pagan–Hall paper has a serious typo in the presentation of their nonnormality-robust statistic. Their equation (58b), 195, is missing the term (in their terminology) $-2\mu_3\psi(\widehat{X}'\widehat{X})^{-1}\widehat{X}'D(D'D)^{-1}$. The typo reappears in the discussion of the test by Godfrey (1988). The correction published in Pesaran and Taylor (1999) is incomplete, as it applies only to the version of the Pagan–Hall test with a single indicator variable.

Let

$$\begin{aligned}
&\overline{\Psi} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \Psi_i && \text{dimension} = n \times p \\
&\widehat{D} \equiv \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \Psi_i'(\widehat{u}_i^2 - \widehat{\sigma}^2) && \text{dimension} = n \times 1 \\
&\widehat{\Gamma} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} (\Psi_i - \widehat{\Psi})' X_i \widehat{u}_i && \text{dimension} = p \times K \\
&\widehat{\mu}_3 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \widehat{u}_i^3 \\
&\widehat{\mu}_4 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \widehat{u}_i^4 \\
&\widehat{X} = P_z X
\end{aligned} \tag{38}$$

If u_i is homoskedastic and independent of Z_i, then Pagan and Hall (1983) (Theorem 8) show that under the null of no heteroskedasticity,

$$n\widehat{D}'\widehat{B}^{-1}\widehat{D} \overset{A}{\sim} \chi_p^2 \tag{39}$$

where

$$\begin{aligned}
&\widehat{B} = B_1 + B_2 + B_3 + B_4 \\
&B_1 = (\widehat{\mu}_4 - \widehat{\sigma}^4)\tfrac{1}{n}(\Psi_i - \overline{\Psi})'(\Psi_i - \overline{\Psi}) \\
&B_2 = -2\widehat{\mu}^3 \tfrac{1}{n}\Psi'\widehat{X}(\tfrac{1}{n}\widehat{X}'\widehat{X})^{-1}\widehat{\Gamma}' \\
&B_3 = B_2' \\
&B_4 = 4\widehat{\sigma}^2 \tfrac{1}{n}\widehat{\Gamma}'(\tfrac{1}{n}\widehat{X}'\widehat{X})^{-1}\widehat{\Gamma}
\end{aligned} \tag{40}$$

This is the default statistic produced by `ivhettest`. Several special cases are worth noting:

- If the error term is assumed to be normally distributed, then $B_2 = B_3 = 0$ and $B_1 = 2\widehat{\sigma}^4(1/n)(\Psi_i - \overline{\Psi})'(\Psi_i - \overline{\Psi})$. This is available from `ivhettest` with the `phnorm` option.
- If the rest of the system is assumed to be homoskedastic, then $B_2 = B_3 = B_4 = 0$ and the statistic in (39) becomes the White/Koenker nR_c^2 statistic. This is available from `ivhettest` with the `nr2` option.
- If the rest of the system is assumed to be homoskedastic and the error term is assumed to be normally distributed, then $B_2 = B_3 = B_4 = 0$, $B_1 = 2\widehat{\sigma}^4(1/n)(\Psi_i - \overline{\Psi})'(\Psi_i - \overline{\Psi})$, and the statistic in (39) becomes the Breusch–Pagan/Godfrey/Cook–Weisberg statistic. This is available from `ivhettest` with the `bpg` option.

All of the above statistics will be reported with the `all` option. `ivhettest` can also be employed after estimation via OLS or HOLS using `regress` or `ivreg2`. In this case, the default test statistic is the White/Koenker nR_c^2 test.

The Pagan–Hall statistic has not been widely used in practice, perhaps because it is not a standard feature of most regression packages. For a discussion of the relative merits of the Pagan–Hall test, including some Monte Carlo results, see Pesaran and Taylor (1999). Their findings suggest caution in the use of the Pagan–Hall statistic particularly in small samples; in these circumstances, the nR_c^2 statistic may be preferred.

4 Testing the relevance and validity of instruments

4.1 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. The first stage regressions are reduced form regressions of the endogenous variables X_1 on the full set of instruments Z; the relevant test statistics here relate to the explanatory power of the excluded instruments Z_1 in these regressions. A statistic commonly used, as recommended by, for example, Bound et al. (1995), is the R^2 of the first-stage regression with the included instruments "partialled-out".[8] Alternatively, this may be expressed as the F test of the joint significance of the Z_1 instruments in the first-stage regression. However, for models with multiple endogenous variables, these indicators may not be sufficiently informative.

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 et al. (1995) F statistics and partial R^2 measures from the two first-stage regressions will not reveal this weakness. Indeed, the F statistics will be statistically significant, and without further investigation, the researcher will not realize that the model cannot be estimated in this form. To deal with this problem of "instrument irrelevance", either additional relevant instruments are needed, or one of the endogenous regressors must be dropped from the model. The statistics proposed by Bound et al. (1995) are able to diagnose instrument relevance only in the presence of a single endogenous regressor. When multiple endogenous regressors are used, other statistics are required.

[8] More precisely, this is the "squared partial correlation" between the excluded instruments Z_1 and the endogenous regressor in question. It is defined as $(RSS_{Z_2} - RSS_Z)/TSS$, where RSS_{Z_2} is the residual sum of squares in the regression of the endogenous regressor on Z_2, and RSS_Z is the RSS when the full set of instruments is used.

One such statistic has been proposed by Shea (1997): a "partial R^2" measure that takes the intercorrelations among the instruments into account.[9] For a model containing a single endogenous regressor, the two R^2 measures are equivalent. The distribution of Shea's partial R^2 statistic has not been derived, but it may be interpreted like any R^2. As a rule of thumb, if an estimated equation yields a large value of the standard (Bound et al. 1995) partial R^2 and a small value of the Shea measure, one may conclude that the instruments lack sufficient relevance to explain all the endogenous regressors, and the model may be essentially unidentified.

The Bound et al. (1995) measures and the Shea partial R^2 statistic can be obtained via the `first` or `ffirst` options on the `ivreg2` command.

The consequence of excluded instruments with little explanatory power is increased bias in the estimated IV coefficients (Hahn and Hausman 2002b). If their explanatory power in the first stage regression is nil, the model is in effect unidentified with respect to that endogenous variable; in this case, the bias of the IV estimator is the same as that of the OLS estimator, IV becomes inconsistent, and nothing is gained from instrumenting (*ibid.*). If the explanatory power is simply "weak",[10] conventional asymptotics fail. What is surprising is that, as Staiger and Stock (1997) and others have shown, the "weak instrument" problem can arise even when the first stage tests are significant at conventional levels (5% or 1%) and the researcher is using a large sample. One rule of thumb is that for a single endogenous regressor, an F statistic below 10 is cause for concern (Staiger and Stock 1997, 557). Since the size of the IV bias is increasing in the number of instruments (Hahn and Hausman 2002b), one recommendation when faced with this problem is to be parsimonious in the choice of instruments. For further discussion, see, for example, Staiger and Stock (1997), Hahn and Hausman (2002a, 2002b), and the references cited therein.

4.2 Overidentifying restrictions in GMM

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 described in (17); that is, whether the instruments are uncorrelated with the error process. This condition will arise when the order condition for identification is satisfied in inequality: the number of instruments excluded from the equation exceeds the number of included endogenous variables. This test can and should be performed as a standard diagnostic in any overidentified instru-

[9] The Shea partial R^2 statistic may be easily computed according to the simplification presented in Godfrey (1999), who demonstrates that Shea's statistic for endogenous regressor i may be expressed as $R^2_p = \nu^{\text{OLS}}_{i,i}/\nu^{\text{IV}}_{i,i} \left\{(1 - R^2_{\text{IV}})/(1 - R^2_{\text{OLS}})\right\}$, where $\nu_{i,i}$ is the estimated asymptotic variance of the coefficient.

[10] One approach in the literature, following Staiger and Stock (1997), is to define "weak" as meaning that the first stage reduced form coefficients are in a $N^{1/2}$ neighborhood of zero, or equivalently, holding the expectation of the first stage F statistic constant as the sample size increases. See also Hahn and Hausman (2002b).

mental variables estimation.[11] These are tests of the joint hypotheses of correct model specification and the orthogonality conditions, and a rejection may properly call either or both of those hypotheses into question.

In the context of GMM, the overidentifying restrictions may be tested via the commonly employed J statistic of Hansen (1982). This statistic is none other than the value of the GMM objective function (20), evaluated at the efficient GMM estimator $\widehat{\beta}_{\text{EGMM}}$. Under the null,

$$J(\widehat{\beta}_{\text{EGMM}}) = n\overline{g}(\widehat{\beta})'\widehat{S}^{-1}\overline{g}(\widehat{\beta}) \quad \overset{A}{\sim} \quad \chi^2_{L-K} \tag{41}$$

In the case of heteroskedastic errors, the matrix $\widehat{S}$ is estimated using the $\widehat{\Omega}$ matrix (27), and the J statistic becomes

$$J(\widehat{\beta}_{\text{EGMM}}) = \widehat{u}Z'(Z'\widehat{\Omega}Z)^{-1}Z\widehat{u}' \quad \overset{A}{\sim} \quad \chi^2_{L-K} \tag{42}$$

With clustered errors, the $\widehat{\Omega}_C$ matrix (37) can be used instead, and this J will be consistent in the presence of arbitrary intra-cluster correlation.

The J statistic is distributed as χ^2 with degrees of freedom equal to the number of overidentifying restrictions $L-K$ rather than the total number of moment conditions L because, in effect, K degrees of freedom are used up in estimating the coefficients of β. J is the most common diagnostic utilized in GMM estimation to evaluate the suitability of the model. A rejection of the null hypothesis implies that the instruments are not satisfying the orthogonality conditions required for their employment. This may be either because they are not truly exogenous, or because they are being incorrectly excluded from the regression. The J statistic is calculated and displayed by `ivreg2` when the `gmm`, `robust`, or `cluster` options are specified. In the last case, the J statistic will be consistent in the presence of arbitrary intra-cluster correlation. This can be quite important in practice: Hoxby and Paserman (1998) have shown that the presence of intra-cluster correlation can readily cause a standard overidentification statistic to over-reject the null.

4.3 Overidentifying restrictions in IV

In the special case of linear instrumental variables under conditional heteroskedasticity, the concept of the J statistic considerably predates the development of GMM estimation techniques. The `ivreg2` procedure routinely presents this test, labeled as Sargan's statistic (Sargan 1958) in the estimation output.

Just as IV is a special case of GMM, Sargan's statistic is a special case of Hansen's J under the assumption of conditional homoskedasticity. Thus, if we use the IV optimal weighting matrix (34) together with the expression for J (41), we obtain

$$\text{Sargan's statistic} = \frac{1}{\widehat{\sigma}^2}\widehat{u}'Z(Z'Z)^{-1}Z'\widehat{u} = \frac{\widehat{u}'Z(Z'Z)^{-1}Z'\widehat{u}}{\widehat{u}'\widehat{u}/n} = \frac{\widehat{u}'P_Z\widehat{u}}{\widehat{u}'\widehat{u}/n} \tag{43}$$

[11] Thus Davidson and MacKinnon (1993, 236): "Tests of overidentifying restrictions should be calculated routinely whenever one computes IV estimates." Sargan's own view, cited in Godfrey (1988), 145, was that regression analysis without testing the orthogonality assumptions is a "pious fraud".

It is easy to see from (43) that Sargan's statistic has an nR_u^2 form (where R_u^2 is the uncentered R^2), and it can be easily calculated this way by regressing the IV equation's residuals upon all instruments Z (both the included exogenous variables and those instruments that do not appear in the equation). The nR_u^2 of this auxiliary regression will have a χ^2_{L-K} distribution under the null hypothesis that all instruments are orthogonal to the error. This auxiliary regression test is that performed by `overid` after `ivreg`, and the statistic is also automatically reported by `ivreg2`.[12] A good discussion of this test is presented in Wooldridge (2002, 123).

The literature contains several variations on this test. The main idea behind these variations is that there is more than one way to consistently estimate the variance in the denominator of (43). The most important of these is that of Basmann (1960). Independently of Sargan, Basmann proposed an $F(L-K, n-L)$ test of overidentifying restrictions,

$$\text{Basmann's } F \text{ statistic} = \frac{\widehat{u}' P_Z \widehat{u}/(L-K)}{\widehat{u}' M_Z \widehat{u}/(n-L)} \tag{44}$$

where $M_Z \equiv I - P_Z$ is the "annihilator" matrix and L is the total number of instruments. Note that since $\widehat{u}' M_Z \widehat{u} = \widehat{u}'\widehat{u} - \widehat{u}' P_Z \widehat{u}$, the same artificial regression can be used to generate both the Basmann and the Sargan statistics.

The difference between Sargan's and Basmann's statistics is that the former uses an estimate of the error variance from the IV regression estimated with the full set of overidentifying restrictions, whereas the latter uses an estimate from a regression without the overidentifying restrictions being imposed.[13] Either method will generate a consistent estimator of the error variance under the null of instrument validity, and hence the two statistics are asymptotically equivalent.

By default the Sargan nR_u^2 statistic and a χ^2 version of Basmann's statistic (without the numerator degrees of freedom) are reported in the `overid` output. An alternative form of the Sargan statistic that uses a small-sample correction, replacing the estimate of the error variance $\widehat{u}'\widehat{u}/n$ with $\widehat{u}'\widehat{u}/(n-K)$, may be requested via the `dfr` option; this is also the version of the Sargan statistic reported by `ivreg2` for IV estimation when the `small` option is used. "Pseudo-F" forms of the Sargan and Basmann tests, obtained by dividing the numerator $\widehat{u}' P_Z \widehat{u}$ by $L-K$, may be requested via the `f` option. The `all` option displays all five statistics.

Neither the Sargan nor the Basmann statistics computed for an IV regression is valid in the presence of conditional heteroskedasticity. In this case, a heteroskedasticity-robust overidentification statistic can be calculated for an IV regression by applying a general result in the literature for a test of overidentification for a GMM estimator with

[12] Note that Stata's `regress` reports an uncentered R^2 only if the model does not contain a constant, and a centered R^2 otherwise. Consequently, `overid` calculates the uncentered R^2 itself; the uncentered total sum of squares of the auxiliary regression needed for the denominator of R_u^2 is simply the residual sum of squares of the original IV regression.

[13] See Davidson and MacKinnon (1993, 235–236). The Basmann statistic uses the error variance from the estimate of their equation (7.54), and the pseudo-F form of the Basmann statistic is given by equation (7.55); the Sargan statistic is given by their (7.57).

a suboptimal weighting matrix, which is what IV amounts to in these circumstances.[14] It does not seem to have been noted in the literature that in the IV case, this "robustified Sargan statistic" is numerically identical to the J statistic computed from feasible efficient two-step GMM for that equation. Thus, if robust inference is sought in an instrumental variables model, one may calculate the test for overidentification via a standard J statistic. When the `robust` and/or `cluster` options are used with `ivreg2` to estimate an IV regression with robust standard errors, the Hansen J statistic for feasible efficient two-step GMM is automatically reported.

4.4 Testing a subset of the overidentifying restrictions

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.[15] The test is known under other names as well; e.g., Ruud (2000) calls it the "distance difference" statistic, and Hayashi (2000) follows Eichenbaum et al. (1988) and dubs it the C statistic; we will use the latter term. 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. For excluded instruments, this is equivalent to dropping them from the instrument list. For included instruments, the C test hypothecates placing them in the list of included endogenous variables; in essence, treating them as endogenous regressors. The C test, distributed χ^2 with degrees of freedom equal to the loss of overidentifying restrictions (i.e., the number of suspect instruments being tested), has the null hypothesis that the specified variables are proper instruments.

Although the C statistic can be calculated as the simple difference between the Hansen–Sargan statistics for two regressions, this procedure can generate a negative test statistic in finite samples. In the IV context, this problem can be avoided, and the C statistic guaranteed to be nonnegative if the estimate of the error variance $\widehat{\sigma}^2$ from the original (restricted, more efficient) IV regression is used to calculate the Sargan statistic for the unrestricted IV regression as well. The equivalent procedure in the GMM context is to use the $\widehat{S}$ matrix from the original estimation to calculate both J statistics. More precisely, $\widehat{S}$ from the restricted estimation is used to form the restricted J statistic, and the submatrix of $\widehat{S}$ with rows/columns corresponding to the unrestricted estimation is used to form the J statistic for the unrestricted estimation; see Hayashi (2000, 220).

[14] See Ahn (1995), Proposition 1, or, for an alternative formulation, Wooldridge (1995), Procedure 3.2.

[15] See Hayashi (2000, 218–221 and 232–234) or Ruud (2000, Chapter 22), for comprehensive presentations.

The C test is conducted in `ivreg2` by specifying the `orthog` option, and listing the instruments (either included or excluded) to be challenged. The equation must still be identified with these instruments either removed or reconsidered as endogenous if the C statistic is to be calculated. Note that if the unrestricted equation is exactly identified, the Hansen–Sargan statistic for the unrestricted equation will be zero and the C statistic will coincide with the Hansen–Sargan statistic for the original (restricted) equation, and this will be true *irrespective* of the instruments used to identify the unrestricted estimation. This illustrates how the Hansen–Sargan overidentification test is an "omnibus" test for the failure of *any* of the instruments to satisfy the orthogonality conditions, but at the same time requires that the investigator believe that at least *some* of the instruments are valid; see Ruud (2000, 577).

4.5 Tests of overidentifying restrictions as Lagrange multiplier (score) tests

The Sargan test can be viewed as analogous to a Lagrange multiplier (LM) or score test.[16] In the case of OLS, the resemblance becomes exact. Consider the model $Y = X_2\beta+u$, for which the researcher wishes to test whether the additional variables Z_1 can be omitted; both X_2 and Z_1 are assumed to be exogenous. The LM test statistic of this hypothesis is obtained as nR_u^2 from a regression of the OLS residuals $\widehat{u}$ on X_2 and Z_1. It is easy to see that this is in fact the same procedure used to obtain the Sargan statistic for the special case of no endogenous regressors: $X = X_2$ and $Z = [Z_1\ X_2]$. This result carries over into GMM estimation using Cragg's HOLS: the J statistic for the HOLS estimator is a heteroskedasticity-robust LM-type test of the hypothesis that Z_1 can be omitted from the estimation.

When `ivreg2` is used to generate OLS estimates, the Sargan statistic reported is an LM test of the variables in the IV *varlist*. If the `gmm` option is chosen, HOLS estimates are reported along with a robust LM statistic. As usual, the `cluster` option generates a statistic that is robust to arbitrary intra-cluster correlation.

If the estimation method is OLS but the error is not homoskedastic, then the standard LM test is no longer valid. A heteroskedasticity-robust version is, however, available.[17] The robust LM statistic for OLS is numerically equivalent to the J statistic from feasible efficient two-step GMM, i.e., HOLS, a result that again does not seem to have been noted in the literature.

5 Testing for endogeneity of the regressors

There may well be reason to suspect nonorthogonality between regressors and errors—which can arise from several sources, after all, including classical errors-in-variables. Turning to IV or efficient GMM estimation for the sake of consistency must be balanced

[16] For a detailed discussion of the relationship between the different types of tests in a GMM framework, see Ruud (2000, Chapter 22).

[17] See Wooldridge (2002, 58–61) and Wooldridge (1995) for more detailed discussion.

against the inevitable loss of efficiency. As Wooldridge states, "... an important cost of performing IV estimation when x and u are uncorrelated: the asymptotic variance of the IV estimator is always larger, and sometimes much larger, than the asymptotic variance of the OLS estimator." (Wooldridge 2003, 490) Naturally, this loss of efficiency is a price worth paying if the OLS estimator is biased and inconsistent; thus, a test of the appropriateness of OLS, and the necessity to resort to instrumental variables or GMM methods, would be very useful. The intuition for such a test may also be couched in terms of the number of orthogonality conditions available. May all or some of the included endogenous regressors be appropriately treated as exogenous? If so, these restrictions can be added to the set of moment conditions, and more efficient estimation will be possible.

5.1 Durbin–Wu–Hausman tests for endogeneity in IV estimation

Many econometrics texts discuss the issue of "OLS versus IV" in the context of the Durbin–Wu–Hausman (DWH) tests, which involve estimating the model via both OLS and IV approaches and comparing the resulting coefficient vectors. In the Hausman form of the test, a quadratic form in the differences between the two coefficient vectors, scaled by the precision matrix, gives rise to a test statistic for the null hypothesis that the OLS estimator is consistent and fully efficient.

Denote by $\widehat{\beta}^c$ the estimator that is consistent under both the null and the alternative hypotheses, and by $\widehat{\beta}^e$ the estimator that is fully efficient under the null but inconsistent if the null is not true. The Hausman (1978) specification test takes the quadratic form

$$\begin{aligned} H &= n(\widehat{\beta}^c - \widehat{\beta}^e)' D^- (\widehat{\beta}^c - \widehat{\beta}^e) \\ \text{where} \qquad & \\ D &= \left\{ V(\widehat{\beta}^c) - V(\widehat{\beta}^e) \right\} \end{aligned} \tag{45}$$

where $V(\widehat{\beta})$ denotes a consistent estimate of the asymptotic variance of β, and the operator $^-$ denotes a generalized inverse.

A Hausman statistic for a test of endogeneity in an IV regression is formed by choosing OLS as the efficient estimator $\widehat{\beta}^e$ and IV as the inefficient but consistent estimator $\widehat{\beta}^c$. The test statistic is distributed as χ^2 with K_1 degrees of freedom, this being the number of regressors being tested for endogeneity. The test is perhaps best interpreted not as a test for the endogeneity or exogeneity of regressors per se, but rather as a test of the consequence of employing different estimation methods on the same equation. Under the null hypothesis that OLS is an appropriate estimation technique, only efficiency should be lost by turning to IV; the point estimates should be qualitatively unaffected.

The Hausman statistic comes in several flavors, depending on which estimates of the asymptotic variances are used. An obvious possibility would be to use $V(\widehat{\beta}_{\text{IV}})$ and $V(\widehat{\beta}_{\text{OLS}})$ as generated by standard IV and OLS estimation; this would be the result if Stata's `hausman` command were used without any options. This is actually rarely done because, although asymptotically valid, it has the drawback of possibly generating a

negative Hausman statistic in finite samples.[18] Avoiding this problem is straightforward, however. Recall that the standard asymptotic covariances for IV and OLS are

$$V(\widehat{\beta}_{\text{IV}}) = \widehat{\sigma}^2_{\text{IV}}(X'P_ZX)^{-1} \qquad V(\widehat{\beta}_{\text{OLS}}) = \widehat{\sigma}^2_{\text{OLS}}(X'X)^{-1} \tag{46}$$

Under the null, both the IV and the OLS estimates of the error variance are consistent estimators of σ, and either can be used to form the Hausman statistic. If a common estimate of σ is used, then the generalized inverse of D is guaranteed to exist and a positive test statistic is guaranteed.[19]

If the Hausman statistic is formed using the OLS estimate of the error variance, then the D matrix in (45) becomes

$$D = \widehat{\sigma}^2_{\text{OLS}}\left\{(X'P_ZX)^{-1} - (X'X)^{-1}\right\} \tag{47}$$

This version of the endogeneity test was first proposed by Durbin (1954) and separately by Wu (1973) (his T_4 statistic) and Hausman (1978). It can be obtained within Stata by using `hausman` with the `sigmamore` option in conjunction with estimation by `regress`, `ivreg` and/or, `ivreg2`.

If the Hausman statistic is formed using the IV estimate of the error variance, then the D matrix becomes

$$D = \widehat{\sigma}^2_{\text{IV}}\left\{(X'P_ZX)^{-1} - (X'X)^{-1}\right\} \tag{48}$$

This version of the statistic was proposed separately by Wu (1973) (his T_3 statistic) and Hausman (1978). It can be obtained within Stata by using `hausman` with the (undocumented) `sigmaless` option.

Use of `hausman` with the `sigmamore` or `sigmaless` options avoids the additional annoyance that because Stata's `hausman` tries to deduce the correct degrees of freedom for the test from the rank of the matrix D, it may sometimes come up with the wrong answer. It will correctly report K_1 degrees of freedom for the test if a common estimate of the error variance is used, i.e., in either the Durbin (47) or Wu T_3 (48) forms of the statistic,[20] but not if both $V(\widehat{\beta}_{\text{IV}})$ and $V(\widehat{\beta}_{\text{OLS}})$ are used to form D. What will happen in this case is that `hausman` will report the correct χ^2 statistic, but with degrees of freedom equal to K rather than K_1, and the user will have to calculate the correct $p-$value by hand.

[18] Readers should also bear in mind here and below that the estimates of the error variances may or may not have small-sample corrections, according to the estimation package used and the options chosen. If one of the variance–covariance matrices in D uses a small-sample correction, then so should the other.

[19] The matrix difference in (47) and (48) has rank K_1; see Greene (2000, 384–385). Intuitively, the variables being tested are those not shared by X and Z, namely the K_1 endogenous regressors X_1. The Hausman statistic for the endogeneity test can also be expressed in terms of a test of the coefficients of the endogenous regressors alone and the rest of the βs removed. In this alternate form, the matrix difference in the expression equivalent to (47) is positive definite and a generalized inverse is not required. See Bowden and Turkington (1984, 50–51).

[20] This works in the former two cases because the matrix difference in (47) and (48) has rank K_1; see note 19 above.

Although these different flavors of the DWH endogeneity test are asymptotically equivalent, they will differ numerically, and may perform differently in finite samples. Given the choice between forming the Hausman statistic using either $\widehat{\sigma}^2_{\text{OLS}}$ or $\widehat{\sigma}^2_{\text{IV}}$, the standard choice is the former (the Durbin statistic) because under the null, both are consistent but the former is more efficient. The Durbin flavor of the test has the additional advantage of superior performance when instruments are weak (Staiger and Stock 1997).

5.2 Extensions: Testing a subset of the regressors for endogeneity, and heteroskedastic-robust testing for IV and GMM estimation

In some contexts, the researcher may be certain that one or more regressors in X_1 is endogenous but may question the endogeneity of the others. In such a context, the DWH tests above are easily modified to apply to a subset of the endogenous regressors.

Consider dividing the set of endogenous regressors into two subsets, X_{1A} and X_{1B}, where only the second set of variables is to be tested for endogeneity. In the tests using the Hausman statistic formulation, (45), the less efficient but consistent estimator $\widehat{\beta}^c$ remains the IV estimator $\widehat{\beta}_{\text{IV}}$, but the fully efficient estimator is now the IV estimator $\widehat{\beta}_{\text{IVB}}$ from the regression in which X_{1A} is still treated as endogenous but X_{1B} is treated as exogenous. A positive test statistic can again be guaranteed if the estimate of the error variance $\widehat{\sigma}$ used in the matrix D is from either of the two IV estimations, since both are consistent under the null. Again, use of the $\widehat{\sigma}^2$ from the more efficient estimation is traditional.

The Hausman statistic framework of (45) for tests of the endogeneity of regressors is available both for IV estimation with robust standard errors and for efficient GMM estimation. The procedure is essentially the same as in the standard IV versus OLS case discussed above: estimate the equation twice, once with the regressors being tested as exogenous (the more efficient estimator) and once with the same regressors treated as endogenous (the less efficient but consistent estimator), and form the Hausman statistic using the estimated coefficients and (robust) covariance matrices.

If Stata's `hausman` command is used to form the statistic this way, the mildly annoying problem of a possibly negative Hausman statistic can arise, and furthermore, `hausman` will report the correct statistic but with the wrong degrees of freedom (K instead of the correct K_1). The way to guarantee a nonnegative test statistic is the same method used with the C test: the equivalent of the `sigmamore` option of `hausman` would be to use the $\widehat{S}$ matrix from the more efficient estimation to form the covariance matrix for the less efficient but consistent estimation as well; see Section 4.4. Unfortunately, this feature is not available with `hausman`,[21] nor can it easily be computed by hand, but it is available via the `orthog` option of `ivreg2`, as we shall see at the very end of this section.

[21] Users beware: the `sigmamore` option following a `robust` estimation will not only fail to accomplish this, but it will also generate an invalid test statistic.

5.3 Durbin–Wu–Hausman tests as GMM tests of orthogonality conditions

Readers at this point may be wondering about the relationship between the GMM tests of orthogonality conditions implemented by the Hansen–Sargan, Basmann, and C tests as discussed in Sections 4.2–4.4, and the Durbin–Wu–Hausman tests. The potential resemblance is even closer once we note that the application of the Hausman test is not limited to testing the endogeneity of regressors.

A Hausman test, like the C test, can be used to test a variety of combinations of the orthogonality conditions, not only those involving regressors but those involving excluded instruments as well. Denote by L^e and L^c the number of total instruments in, respectively, the restricted (efficient) and the unrestricted (consistent but inefficient) regressions. $L^e - L^c$ is therefore the number of orthogonality conditions being tested. Also denote by K_1^c the number of endogenous regressors in the unrestricted regression. Then it can be shown that under conditional homoskedasticity, the Hausman statistic based on IV or GMM estimates, $\widehat{\beta}^e$ and $\widehat{\beta}^c$, respectively, will be distributed as χ^2 with degrees of freedom $= Min[L^e - L^c,\ K_1^c]$. In the conditional heteroskedasticity case, the degrees of freedom will be $L^e - L^c$ if $L^e - L^c \leq K_1^c$ but unknown otherwise (making the test impractical).[22]

What, then, is the difference between the GMM C test and the Hausman specification test? In fact, because the two estimators being tested are both GMM estimators, the Hausman specification test is a test of linear combinations of orthogonality conditions (Ruud 2000, 578-584). When the particular linear combination of orthogonality conditions being tested is the same for the C test and for the Hausman test, the two test statistics will be numerically equivalent. We can state this more precisely as follows: If $L^e - L^c \leq K_1^c$, the C statistic and the Hausman statistic are numerically equivalent.[23] If $L^e - L^c > K_1^c$, the two statistics will be numerically different, the C statistic will have $L^e - L^c$ degrees of freedom, and the Hausman statistic will have K_1^c degrees of freedom in the conditional homoskedasticity case (and an unknown number of degrees of freedom in the conditional heteroskedasticity case).

One commonly encountered case in which the two statistics are exactly equivalent is in fact the one with which we began our discussion of DWH tests, namely when we want to test the endogeneity of regressors. An example of when the two test statistics differ would arise when the investigator has suspicions about a large number of excluded instruments. In this case, the number of instruments being tested, $L^e - L^c$, may be larger than the K_1^c endogenous regressors in the less efficient estimation.

The intuition behind the circumstances in which the two statistics will differ follows from what is being tested. The Hausman test is a vector of contrasts test that detects changes in the coefficients of the regressors treated as endogenous in the consistent but inefficient specifications. When the number of moment conditions being tested is larger

[22] See Hausman and Taylor (1981) and Newey (1985), summarized by Hayashi (2000, 233–34).

[23] We also need to assume, of course, that the two tests use the same estimate of the error variance, $\widehat{\sigma}^2$, or the same $\widehat{S}$ matrix.

than the number of endogenous regressors that will be affected by them, the Hausman test will have fewer degrees of freedom than the C test. This means an investigator faces a trade-off when deciding which of the two tests to use: when the two tests differ, the Hausman test is a test of linear combinations of moment conditions, and is more powerful than the C test at detecting violations on restrictions of these linear combinations, but the latter test will be able to detect other violations of moment conditions that the former test cannot. As Ruud (2000), 585, points out, one of the appealing features of the Hausman test is that its particular linear combination of moment conditions also determines the consistency of the more efficient GMM estimator.

There is an interesting semantic issue here: is there a difference between an "endogeneity test" and a test of "exogeneity" or "orthogonality"? The answer is, in the IV context, "not really". The DWH endogeneity tests are usually presented in textbooks as tests of "endogeneity", and the Hansen–Sargan–Basmann C tests are usually presented as tests of the "validity" or "exogeneity" of instruments—and we have adopted these conventions here—but they are all really just tests of orthogonality conditions. The reason for the different terminology relates, instead, to the circumstances in which the researcher is operating and in particular his/her starting point. Say we start with an IV estimation in which two regressors x_{1A} and x_{1B} are treated as endogenous and there are five excluded instruments. We suspect that we do not need to be instrumenting x_{1B}, and so we employ the Hausman form of the DWH endogeneity test to see whether or not we can increase our set of orthogonality conditions from 5 to 6. Now consider a second researcher whose priors are somewhat less conservative; s/he starts with a specification in which x_{1A} is still treated as endogenous but x_{1B} is exogenous. S/he does, however, have the same suspicions about x_{1B}, and so s/he employs a C test of its orthogonality to see whether or not s/he needs to reduce the set of orthogonality conditions from 6 to 5. The two tests are numerically the same, and are testing the same hypothesis—the exogeneity of x_{1B}—and the only difference is the starting point of the researchers.

5.4 DWH endogeneity tests in practice

There are a variety of ways of conducting a DWH endogeneity test in Stata for the standard IV case with conditional homoskedasticity. Three equivalent ways of obtaining the Durbin flavor of the Durbin–Wu–Hausman statistic (47) are

1. Fit the less efficient but consistent model using IV, followed by the command `hausman, save`. Then, fit the fully efficient model by OLS (or by IV if only a subset of regressors is being tested for endogeneity), followed by `hausman, sigmamore`.
2. Fit the fully efficient model using `ivreg2`, specifying the regressors to be tested in the `orthog` option.
3. Fit the less efficient but consistent model using `ivreg`, then use `ivendog` to conduct an endogeneity test. This program will take as its argument a *varlist* consisting of the subset of regressors to be tested for endogeneity; if the *varlist* is empty, the full set of endogenous regressors is tested.

The latter two methods are of course more convenient than the first, as the test can be done in one step.

Yet another asymptotically equivalent flavor of the DWH test is available for standard IV estimation under conditional homoskedasticity, and is included in the output of `ivendog`. This is the test statistic introduced by Wu (1973) (his T_2), and separately shown by Hausman (1978) to be calculated straightforwardly through the use of auxiliary regressions. We will refer to it as the Wu–Hausman statistic.[24]

Consider a simplified version of our basic model (1) with a single endogenous regressor x_1,

$$y = \beta_1 x_1 + X_2\beta_2 + u, \tag{49}$$

with $X_2 \equiv Z_2$ assumed exogenous (including the constant, if one is specified) and with excluded instruments Z_1 as usual. The auxiliary regression approach involves estimating the reduced form (first-stage) regression for x_1 :

$$x_1 = Z_1\Gamma_1 + X_2\Gamma_2 + v = Z\Gamma + v \tag{50}$$

We are concerned with testing that $x_1 \perp u$. Since by assumption each z in Z is uncorrelated with u, the first stage regression implies that this condition is equivalent to a test of $v \perp u$. Exogeneity of the z's implies that $\widehat{v}$—the residuals from OLS estimation of the first-stage regression (50)—will be a consistent estimator of v. Thus, we augment (49) with $\widehat{u}$ and re-estimate it with OLS. A $t-$test of the significance of $\widehat{v}$ in this auxiliary regression is then a direct test of the null hypothesis—in this context, that $\theta = 0$:

$$y = \beta_1 x_1 + X_2\beta_2 + \theta\widehat{v} + \epsilon \tag{51}$$

The Wu–Hausman test may be readily generalized to multiple endogenous variables, since it merely requires the estimation of the first-stage regression for each of the endogenous variables, and augmentation of the original model with their residual series. The test statistic then becomes an F test, with numerator degrees of freedom equal to the number of included endogenous variables. One advantage of the Wu–Hausman F statistic over the other DWH tests for IV versus OLS is that with certain normality assumptions, it is a finite sample test exactly distributed as F (see Wu (1973) and Nakamura and Nakamura (1981)). Wu (1974)'s Monte Carlo studies also suggest that this statistic is to be preferred to the statistic using just $\widehat{\sigma}^2_{\text{IV}}$.

A version of the Wu–Hausman statistic for testing a subset of regressors is also available, as Davidson and MacKinnon (1993, 241–242) point out. The modified test involves estimating the first-stage regression for each of the K_{1B} variables in X_{1B} in order to generate a residual series. These residual series $\widehat{V}_B$ are then used to augment the original model,

$$y = X_{1A}\delta + X_{1B}\lambda + X_2\beta + \widehat{V}_B\Theta + \epsilon \tag{52}$$

which is then estimated via instrumental variables, with only X_{1A} specified as included endogenous variables. The test for endogeneity of the variables in X_{1B} is then a test for

[24] A more detailed presentation of the test can be found in Davidson and MacKinnon (1993, 237–242).

the joint significance of the Θ estimates; rejection of that null hypothesis implies that instruments must be provided for the X_{1B} variables.

An inconvenient complication here is that an ordinary F test for the significance of Θ in this auxiliary regression will *not* be valid, because the unrestricted sum of squares needed for the denominator is wrong, and obtaining the correct SSR requires further steps; see Davidson and MacKinnon (1993, Chapter 7). Only in the special case where the efficient estimator is OLS will an ordinary F test yield the correct test statistic. The auxiliary regression approach to obtaining the Wu–Hausman statistic described above has the further disadvantage of being computationally expensive and practically cumbersome when there are more than a few endogenous variables to be tested, because a residual series must be constructed separately for every endogenous variable being tested.

We have taken a different and simpler approach to programming the Wu–Hausman statistic in `ivendog`. The Durbin flavor of the Durbin–Wu–Hausman statistic (47) can be written as

$$\text{Durbin DWH: } \chi^2(K_{1B}) = \frac{Q^*}{USSR/n} \tag{53}$$

and the Wu–Hausman F statistic can be written as

$$\text{Wu-Hausman: } F(K_{1B}, n-K-K_{1B}) = \frac{Q^*/K_{1B}}{(USSR-Q^*)/(n-K-K_{1B})} \tag{54}$$

where Q^* is the difference between the restricted and unrestricted sums of squares given by the auxiliary regression (51) or (52), and $USSR$ is the sum of squared residuals from the efficient estimate of the model.[25] From the discussion in the preceding section, however, we know that for tests of the endogeneity of regressors, the C statistic and the Hausman form of the DWH test are numerically equal, and when the error variance from the more efficient estimation is used, the Hausman form of the DWH test is the Durbin flavor. We also know from the discussion in Sections (4.3) and (4.4) that the C statistic is simply the difference of two Sargan statistics, one for the unrestricted/consistent estimation and one for the restricted/efficient estimation, and we can use the estimate of the error variance from the more efficient estimation for both. Finally, we can see from (53) and (54) that the Wu–Hausman F statistic can be easily calculated from the same quantities needed for the DWH statistic.

This means that the Wu–Hausman F statistic in (54) does not need to be calculated using the traditional auxiliary regression method, with all the first-stage regressions and generation of residual series as described above. Instead, it can be calculated using only three additional regressions: one to estimate the restricted/efficient model, and two artificial regressions to obtain the two Sargan statistics. More precisely, we can write

[25] See Wu (1973) or Nakamura and Nakamura (1981). Q^* can also be interpreted as the difference between the sums of squares of the second-stage estimation of the efficient model with and without the residual series, and with the fitted values for the variables X_{1A} maintained as endogenous. If the efficient model is OLS, then the second-stage estimation is simply OLS augmented by the residual series.

$$\text{Durbin DWH: } \chi^2(K_{1B}) = \frac{\widehat{u}_e' P_{Z,X_{1B}} \widehat{u}_e - \widehat{u}_c' P_Z \widehat{u}_c}{\widehat{u}_e' \widehat{u}_e / n} \tag{55}$$

$$\text{W–H: } F(K_{1B}, n-K-K_{1B}) = \frac{(\widehat{u}_e' P_{Z,X_{1B}} \widehat{u}_e - \widehat{u}_c' P_Z \widehat{u}_c)/K_{1B}}{\{\widehat{u}_e' \widehat{u}_e - (\widehat{u}_e' P_{Z,X_{1B}} \widehat{u}_e - \widehat{u}_c' P_Z \widehat{u}_c)\}/(n-K-K_{1B})} \tag{56}$$

where $\widehat{u}_e$ and $\widehat{u}_c$ refer to the residuals from the restricted/efficient and the unrestricted/consistent estimations, respectively, and $P_{Z,X_{1B}}$ is the projection matrix of the instruments Z augmented by the regressors X_{1B} whose endogeneity is being tested.

A special case worth noting is when the unrestricted/consistent estimation is exactly identified. In that case, the Sargan statistic for that equation is zero, and hence $\widehat{u}_c' P_Z \widehat{u}_c = 0$. It is easy to see from the above that the Durbin flavor of the Durbin–Wu–Hausman χ^2 test statistic becomes identical to the Sargan statistic (43) for the restricted/efficient estimation, and the Wu–Hausman F statistic becomes identical to Basmann's F statistic (44).[26]

Whereas we have available a large smorgasbord of alternative but asymptotically equivalent tests of endogeneity in the case of IV estimation under conditional homoskedasticity, there is much less choice when estimating either IV with a robust covariance matrix or efficient GMM. As noted above, the use of `hausman` to test regressors for endogeneity in the context of heteroskedasticity-robust or efficient GMM estimation will sometimes generate negative test statistics, and the degrees of freedom printed out for the statistic will be wrong. If $L^e - L^c > K_1^c$, there is the additional problem that the degrees of freedom of the Hausman statistic are unknown. All these problems can be avoided and a valid endogeneity test statistic obtained simply through the use of the C statistic: estimate the restricted/fully efficient model with `ivreg2`, specifying the regressors to test for endogeneity in the `orthog` option.

6 Syntax diagrams

`ivreg2` *depvar* [*varlist1*] (*varlist2=varlist_iv*) [*weight*] [`if` *exp*] [`in` *range*]
[, `gmm orthog(`#`) small level(`#`) hascons noconstant robust`
`cluster(`*varname*`) first ffirst noheader nofooter eform(`*string*`)`
`depname(`*varname*`) mse1 plus`]

`ivhettest` [*varlist*] [, `ivlev ivsq fitlev fitsq ph phnorm nr2 bpg all`]

`overid` [, `chi2 dfr f all`]

`ivendog` [*varlist*]

[26] This is another way of illustrating that the estimate of the error variance used in Basmann's statistic comes from an estimation without any overidentifying restrictions being imposed; see the discussion of (44) above.

For a description of the options accepted by these commands, type `help ivreg2`, `help ivhettest`, `help overid`, or `help ivendog`.

7 Acknowledgments

We are grateful to Jeffrey Wooldridge, Fumio Hayashi, Barbara Sianesi, Arthur Lewbel, and seminar participants at Heriot–Watt University for their constructive suggestions. We also thank an associate editor of the Stata Journal, and members of the Stata user community who helped to test the software. Any remaining errors are our own.

8 References

Ahn, S. C. 1995. Robust GMM Tests for Model Specification. *Arizona State University* (Working Paper).

Arellano, M. 1987. Computing robust standard errors for within–groups estimators. *Oxford Bulletin of Economics and Statistics* 49: 431–434.

Basmann, R. 1960. On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association* 55(292): 650–659.

Bound, J., D. Jaeger, and R. Baker. 1995. Problems with instrumental variable estimation when the correlation between the instruments and the endogenous explanatory variables is weak. *Journal of the American Statistical Association* 90: 443–450.

Bowden, R. J. and D. A. Turkington. 1984. *Instrumental Variables*. Cambridge: Cambridge University Press.

Breusch, T. S. and A. R. Pagan. 1979. A simple test for heteroskedasticity and random coefficient variation. *Econometrica* 47: 1287–1294.

Chamberlain, G. 1982. Multivariate regression models for panel data. *Journal of Econometrics* 18: 5–46.

Cook, R. D. and S. Weisberg. 1983. Diagnostics for heteroscedasticity in regression. *Biometrika* 70: 1–10.

Cragg, J. 1983. More efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* 51: 751–763.

Cumby, R. E., J. Huizinga, and M. Obstfeld. 1983. Two-step two-stage least squares estimation in models with rational expectations. *Journal of Econometrics* 21: 333–355.

Davidson, R. and J. G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. 2d ed. New York: Oxford University Press.

Durbin, J. 1954. Errors in variables. *Review of the International Statistical Institute* 22: 23–32.

Eichenbaum, M. S., L. P. Hansen, and K. J. Singleton. 1988. A time series analysis of representative agent models of consumption and leisure. *Quarterly Journal of Economics* 103(1): 51–78.

Godfrey, L. G. 1978. Testing for multiplicative heteroskedasticity. *Journal of Econometrics* 8: 227–236.

—. 1988. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge: Cambridge University Press.

—. 1999. Instrument relevance in multivariate linear models. *Review of Economics & Statistics* 81(3): 550–552.

Greene, W. 2000. *Econometric Analysis*. 4th ed. Upper Saddle River, NJ: Prentice–Hall.

Hahn, J. and J. Hausman. 2002a. A new specification test for the validity of instrumental variables. *Econometrica* 70(1): 163–89.

—. 2002b. Notes on bias in estimators for simultaneous equation models. *Economics Letters* 75(2): 237–41.

Hansen, B. E. 2000. Econometrics. Manuscript, Madison, WI. *http://www.ssc.wisc.edu/~bhansen/notes/notes.htm*.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50(3): 1029–1054.

Hansen, L., J. Heaton, and A. Yaron. 1996. Finite sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* 14(3): 262–280.

Hausman, J. 1978. Specification tests in econometrics. *Econometrica* 46(6): 1251–1271.

Hausman, J. A. and W. E. Taylor. 1981. A generalized specification test. *Economics Letters* 8: 239–245.

Hayashi, F. 2000. *Econometrics*. Princeton, NJ: Princeton University Press.

Hoxby, C. and M. D. Paserman. 1998. Overidentification tests with grouped data. *National Bureau of Economic Research* (Technical Paper 223).

Kézdi, G. 2002. The Economic Returns to Education: Finite–Sample Properties of an IV Estimator. *University of Michigan* (Working Paper).

Koenker, R. 1981. A note on Studentizing a test for heteroskedasticity. *Journal of Econometrics* 17: 107–112.

Moulton, B. R. 1986. Random group effects and the precision of regression estimates. *Journal of Econometrics* 32: 385–397.

Nakamura, A. and M. Nakamura. 1981. On the relationships among several specification error tests presented by Durbin, Wu, and Hausman. *Econometrica* 49(6): 1583–1588.

Newey, W. 1985. Generalized method of moments specification testing. *Journal of Econometrics* 29: 229–256.

Pagan, A. R. and D. Hall. 1983. Diagnostic tests as residual analysis. *Econometric Reviews* 2(2): 159–218.

Pesaran, M. H. and L. W. Taylor. 1999. Diagnostics for IV regressions. *Oxford Bulletin of Economics & Statistics* 61(2): 255–281.

Ruud, P. A. 2000. *An Introduction to Classical Econometric Theory*. Oxford: Oxford University Press.

Sargan, J. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26(3): 393–415.

Shea, J. 1997. Instrument relevance in multivariate linear models: A simple measure. *Review of Economics & Statistics* 79(2): 348–352.

Staiger, D. and J. H. Stock. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65(3): 557–586.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–838.

—. 1982. Instrumental variables regression with independent observations. *Econometrica* 50(2): 483–499.

—. 1984. *Asymptotic Theory for Econometricians*. Orlando, FL: Academic Press.

Wooldridge, J. M. 1995. Score diagnostics for linear models estimated by two stage least squares. In *Advances in Econometrics and Quantitative Economics: Essays in honor of Professor C. R. Rao*, eds. G. S. Maddala, P. C. B. Phillips, and T. N. Srinivasan, 66–87. Cambridge, MA: Blackwell Publishers.

—. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

—. 2003. *Introductory Econometrics: A Modern Approach*. 2d ed. New York: South-Western College Publishing.

Wu, D.-M. 1973. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 41(4): 733–750.

—. 1974. Alternative tests of independence between stochastic regressors and disturbances: Finite sample results. *Econometrica* 42(3): 529–546.

About the Authors

Christopher F. Baum is an associate professor of economics at Boston College. He is an associate editor of *Computational Economics* and *The Stata Journal*, and serves on the Advisory Council of the Society for Computational Economics. Baum founded and manages the Boston College Statistical Software Components (SSC) archive at RePEc (*http://repec.org*). His recent research has focused on the effects of uncertainty on international trade flows, bank lending behavior, and firms' cash holdings.

Mark E. Schaffer is professor of economics and Director of the Centre for Economic Reform and Transformation (CERT) at Heriot–Watt University, Edinburgh, Scotland. He is also a Research Fellow at the Centre for Economic Policy Research (CEPR), the Institute for the Study of Labor (IZA), and the William Davidson Institute. His research interests include various aspects of firm and household behavior in the transition countries of Eastern Europe, the former USSR, and East Asia.

Steven Stillman is a Senior Economic Researcher in the Labour Market Policy Group of the New Zealand Department of Labour. He is also an affiliated Research Fellow at the Institute for the Study of Labor (IZA) and the William Davidson Institute. His current research examines the effect of public policy and institutions on various dynamic aspects of household well-being in New Zealand, Russia, and the United States.