8 Instrumental-variables estimators

8.1 Introduction

The zero-conditional-mean assumption presented in section 4.2 must hold for us to use linear regression. There are three common instances where this assumption may be violated in economic research: endogeneity (simultaneous determination of response variable and regressors), omitted-variable bias, and errors in variables (measurement error in the regressors). Although these problems arise for different reasons in microeconomic models, the solution to each is the same econometric tool: the *instrumental-variables* (IV) estimator, described in this chapter. The most common problem, endogeneity, is presented in the next section. The other two problems are discussed in chapter appendices. The following sections discuss the IV and two-stage least-squares (2SLS) estimators, identification and tests of overidentifying restrictions, and the generalization to generalized method-of-moments (GMM) estimators. The last three sections of the chapter consider testing for heteroskedasticity in the IV context, testing the relevance of instruments, and testing for endogeneity.

A variable is *endogenous* if it is correlated with the disturbance. In the model

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

x_j is endogenous if $\text{Cov}[x_j, u] \neq 0$. x_j is exogenous if $\text{Cov}[x_j, u] = 0$. The OLS estimator will be consistent only if $\text{Cov}[x_j, u] = 0$, $j = 1, 2, \ldots, k$. This zero-covariance assumption and our convention that x_1 is a constant imply that $E[u] = 0$. Following Wooldridge (2002, 2006), we use the zero-conditional-mean assumption

$$E[u|x_1, x_2, \ldots, x_k] = 0$$

which is sufficient for the zero-covariance condition.

Although the rest of this chapter uses economic intuition to determine when a variable is likely to be endogenous in an empirical study, it is the above definition of endogeneity that matters for empirical work.

8.2 Endogeneity in economic relationships

Economists often model behavior as simultaneous-equations systems in which economically endogenous variables are determined by each other and some additional economically exogenous variables. The simultaneity gives rise to empirical models with vari-

ables that do not satisfy the zero-conditional-mean assumption. Consider the textbook supply-and-demand paradigm. We commonly write

$$q^d = \beta_1 + \beta_2 p + \beta_3 inc \tag{8.1}$$

to indicate that the quantity demanded of a good (q^d) depends on its price (p) and the level of purchasers' income (inc). When $\beta_1 > 0$, $\beta_2 < 0$, and $\beta_3 > 0$, the demand curve in $[p, q]$ space slopes downward, and for any given price the quantity demanded will rise for a higher level of purchasers' income.

If this equation reflected an individual's demand function, we might argue that the individual is a price taker who pays the posted price if she chooses to purchase the good and has a fixed income at her disposal on shopping day. But we often lack *microdata*, or household-level data, for the estimation of this relationship for a given good. Rather, we have data generated by the market for the good. The observations on p and q are equilibrium prices and quantities in successive trading periods.

If we append an error term, u, to (8.1) and estimate OLS from these $[p, q]$ pairs, the estimates will be inconsistent. It does not matter whether the model is specified as above with q as the response variable or in inverse form with p as the response variable. In either case, the regressor is endogenous. Simple algebra shows that the regressor must be correlated with the error term, violating the zero-conditional-mean assumption. In (8.1), a shock to the demand curve must alter both the equilibrium price and quantity in the market. By definition, the shock u is correlated with p.

How can we use these market data to estimate a demand curve for the product? We must specify an *instrument* for p that is uncorrelated with u but highly correlated with p. In an economic model, this is termed the *identification problem*: what will allow us to *identify* or trace out the demand curve? Consider the other side of the market. Any factor in the supply function that does not appear in the demand function will be a valid instrument. If we are modeling the demand for an agricultural commodity, a factor like rainfall or temperature would suffice. Those factors are determined outside the economic model but may have an important effect on the yield of the commodity and thus the quantity that the grower will bring to market. In the economic model, these factors will appear in the *reduced-form* equations for both q and p: the algebraic solution to the simultaneous system.

To derive consistent estimates of (8.1), we must find an IV that satisfies two properties: the instrument z must be uncorrelated with u but must be highly correlated with p.[1] A variable that meets those two conditions is an IV or *instrument* for p that deals with the correlation of p and the error term. Because we cannot observe u, we cannot directly test the assumption of zero correlation between z and u, which is known as an orthogonality assumption. We will see that in the presence of multiple instruments, such a test can be constructed. But we can readily test the second assumption and should always do so by regressing the included regressor p on the instrument z:

1. The meaning of "highly correlated" is the subject of section 8.10.

$$p_i = \pi_1 + \pi_2 z_i + \zeta_i \tag{8.2}$$

If we fail to reject the null hypothesis H_0: $\pi_2 = 0$, we conclude that z is not a valid instrument. Unfortunately, rejecting the null of irrelevance is not sufficient to imply that the instrument is not "weak", as discussed in section 8.10.[2] There is no unique choice of an instrument here. We discuss below how we can construct an instrument if more than one is available.

If we decide that we have a valid instrument, how can we use it? Return to (8.1), and write it in matrix form in terms of $\mathbf{y}$ and $\mathbf{X}$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\boldsymbol{\beta}$ is the vector of coefficients $(\beta_1, \beta_2, \beta_3)'$ and $\mathbf{X}$ is $N \times k$. Define a matrix $\mathbf{Z}$ of the same dimension as $\mathbf{X}$ in which the endogenous regressor—p in our example above—is replaced by z. Then

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}'\mathbf{u}$$

The assumption that $\mathbf{Z}$ is unrelated to $\mathbf{u}$ implies that $1/N(\mathbf{Z}'\mathbf{u})$ goes to zero in probability as N becomes large. Thus we may define the estimator $\widehat{\boldsymbol{\beta}}_{\text{IV}}$ from

$$\begin{aligned} \mathbf{Z}'\mathbf{y} &= \mathbf{Z}'\mathbf{X}\,\widehat{\boldsymbol{\beta}}_{\text{IV}} \\ \widehat{\boldsymbol{\beta}}_{\text{IV}} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \end{aligned} \tag{8.3}$$

We may also use the zero-conditional-mean assumption to define a *method-of-moments* estimator of the IV model. In the linear regression model presented in section 4.2.1, the zero-conditional-mean assumption held for each of the k variables in $\mathbf{X}$, giving rise to a set of k moment conditions. In the IV model, we cannot assume that each $\mathbf{X}$ satisfies the zero-conditional-mean assumption: an endogenous x does not. But we can define a matrix $\mathbf{Z}$ as above in which each endogenous regressor will be replaced by its instrument, yielding a method-of-moments estimator for β:

$$\begin{aligned} \mathbf{Z}'\mathbf{u} &= 0 \\ \mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= 0 \end{aligned} \tag{8.4}$$

We may then substitute calculated moments from our sample of data into the expression and replace the unknown coefficients $\boldsymbol{\beta}$ with estimated values $\widehat{\boldsymbol{\beta}}$ in (8.4) to derive

$$\begin{aligned} \mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{IV}} &= 0 \\ \widehat{\boldsymbol{\beta}}_{\text{IV}} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \end{aligned}$$

The IV estimator has an interesting special case. If the zero-conditional-mean assumption holds, each explanatory variable can serve as its own instrument, $\mathbf{X} = \mathbf{Z}$, and

2. Bound, Jaeger, and Baker (1995) proposed the rule of thumb that this F statistic must be at least 10. In more recent work, table 1 of Stock, Wright, and Yogo (2002) provides critical values that depend on the number of instruments.

the IV estimator reduces to the OLS estimator. Thus OLS is a special case of IV that is appropriate when the zero-conditional-mean assumption is satisfied. When that assumption cannot be made, the IV estimator is consistent and has a large-sample normal distribution as long as the two key assumptions about the instrument's properties are satisfied. However, the IV estimator is not an unbiased estimator, and in small samples its bias may be substantial.

8.3 2SLS

Consider the case where we have one endogenous regressor and more than one potential instrument. In (8.1), we might have two candidate instruments: z_1 and z_2. We could apply the IV estimator of (8.3) with z_1 entering $\mathbf{z}$, and generate an estimate of $\widehat{\boldsymbol{\beta}}_{\text{IV}}$. If we repeated the process with z_2 entering $\mathbf{z}$, we would generate another $\widehat{\boldsymbol{\beta}}_{\text{IV}}$ estimate, and those two estimates would differ.

Obtaining the simple IV estimator of (8.3) for each candidate instrument raises the question of how we could combine them. An alternative approach, 2SLS, combines multiple instruments into one optimal instrument, which can then be used in the simple IV estimator. This optimal combination, conceptually, involves running a regression. Consider the auxiliary regression of (8.2), which we use to check that a candidate $\mathbf{z}$ is reasonably well correlated with the regressor that it is instrumenting. Merely extend that regression model,

$$p_i = \pi_1 + \pi_2 z_{i1} + \pi_3 z_{i2} + \omega_i$$

and generate the instrument as the predicted values of this equation: $\widehat{p}$. Given the mechanics of least squares, $\widehat{p}$ is an optimal linear combination of the information in z_1 and z_2. We may then estimate the parameters of (8.3), using the IV estimator with $\widehat{p}$ as a column of $\mathbf{Z}$.

2SLS is nothing more than the IV estimator with a decision rule that reduces the number of instruments to the exact number needed to estimate the equation and fill in the $\mathbf{Z}$ matrix. To clarify the mechanics, define matrix $\mathbf{Z}$ of dimension $N \times \ell$, $\ell \geq k$, of instruments. Then the first-stage regressions define the instruments as

$$\widehat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \tag{8.5}$$

Denote the projection matrix $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ as $\mathbf{P}_Z$. Then from (8.3),

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{2SLS}} &= (\widehat{\mathbf{X}}'\mathbf{X})^{-1}\widehat{\mathbf{X}}'\mathbf{y} \\ &= \{\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\}^{-1}\{\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\} \\ &= (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{y} \end{aligned} \tag{8.6}$$

where the "two-stage" estimator can be calculated in one computation using the data on $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{y}$. When $\ell = k$, 2SLS reduces to IV, so the 2SLS formulas presented below also cover the IV estimator.

Assuming i.i.d. disturbances, a consistent large-sample estimator of the VCE of the 2SLS estimator is

$$\text{Var}[\widehat{\boldsymbol{\beta}}_{\text{2SLS}}] = \widehat{\sigma}^2\{\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\}^{-1} = \widehat{\sigma}^2(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1} \tag{8.7}$$

where $\widehat{\sigma}^2$ is computed as

$$\widehat{\sigma}^2 = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{N}$$

calculated from the 2SLS residuals

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$$

defined by the *original* regressors and the estimated 2SLS coefficients.[3]

The point of using the 2SLS estimator is the consistent estimation of $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$ in a model containing response variable $\mathbf{y}$ and regressors $\mathbf{X}$, some of which are correlated with the disturbance process $\mathbf{u}$. The predictions of that model involve the original regressors $\mathbf{X}$, not the instruments $\widehat{\mathbf{X}}$. Although from a pedagogical standpoint we speak of 2SLS as a sequence of first-stage and second-stage regressions, we should never perform those two steps by hand. If we did so, we would generate predicted values $\{\widehat{\mathbf{X}}\}$ from first-stage regressions of endogenous regressors on instruments and then run the second-stage OLS regression using those predicted values. Why should we avoid this? Because the second stage will yield the incorrect residuals,

$$\widehat{\mathbf{u}}_i = \mathbf{y}_i - \widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}}_{\text{2SLS}} \tag{8.8}$$

rather than the correct residuals,

$$\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$$

which would be calculated by `predict` after a 2SLS estimation. Statistics computed from the incorrect residuals, such as an estimate of σ^2 and the estimated standard error for each $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$ in (8.7), will be inconsistent since the $\widehat{X}$ variables are not the true explanatory variables (see Davidson and MacKinnon 2004, 324). Using Stata's 2SLS command `ivreg` avoids these problems, as I now discuss.

8.4 The ivreg command

The `ivreg` command has the following partial syntax:

`ivreg` *depvar* [*varlist1*] (*varlist2* = *instlist*) [*if*] [*in*] [, *options*]

where *depvar* is the response variable, *varlist2* contains the endogenous regressors, *instlist* contains the excluded instruments, and the optional *varlist1* contains any exogenous regressors included in the equation. In our example from the demand for an agricultural commodity in (8.1), we could specify

3. Some packages, including Stata's `ivreg`, include a degrees-of-freedom correction to the estimate of $\widehat{\sigma}^2$ by replacing N with $N - k$. This correction is unnecessary since the estimate of $\widehat{\sigma}^2$ would not be unbiased anyway (Greene 2000, 373).

```
. ivreg q inc (p = rainfall temperature)
```

to indicate that q is to be regressed on inc and p with rainfall and temperature as excluded instruments. Stata reports that the instruments used in estimation include inc rainfall temperature, considering that inc is serving as its own instrument. Just as with regress, a constant term is included in the equation by default. If a constant appears in the equation, it also implicitly appears in the instrument list used to specify **Z**, the matrix of instruments in the first-stage regression. The first-stage regression (one for each endogenous regressor) may be displayed with the first option.

In a situation with multiple endogenous regressors such as

```
. ivreg y x2 (x3 x4 = za zb zc zd)
```

novice users of instrumental variables often ask, "How do I tell Stata that I want to use za, zb as instruments for x3, and zc, zd as instruments for x4?" You cannot, but not because of any limitation of Stata's ivreg command. The theory of 2SLS estimation does not allow such designations. *All* instruments—included and excluded—must be used as regressors in *all* first-stage regressions. Here both x3 and x4 are regressed on **z**: x2 za zb zc zd and a constant term to form the $\widehat{\mathbf{X}}$ matrix.

We noted above that summary statistics such as Root MSE should be calculated from the appropriate residuals using the original regressors in **X**. If we compare the Root MSE from ivreg and the Root MSE from regress on the same model, the former will inevitably be larger. It appears that taking account of the endogeneity of one or more regressors has cost us something in goodness of fit: least squares is least squares. The minimum sum of squared errors from a model including $\{y\ \mathbf{x}\}$ is by definition that computed by regress. The 2SLS estimator calculated by ivreg is a least-squares estimator, but the criterion minimized involves the improper residuals of (8.8). The 2SLS method is fitting **y** to $\widehat{\mathbf{X}}$ by least squares to generate consistent estimates $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$, thereby minimizing sum of squared errors with respect to $\widehat{\mathbf{X}}$. As long as $\widehat{\mathbf{X}} \neq \mathbf{X}$, those $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$ estimates cannot also minimize the sum of squared errors calculated by [R] **regress**.

Before I present an example of ivreg, we must define *identification* of a structural equation.

8.5 Identification and tests of overidentifying restrictions

The parameters in an equation are said to be identified when we have sufficient valid instruments so that the 2SLS estimator produces unique estimates. In econometrics, we say that an equation is identified, if the parameters in that equation are identified.[4] Equation (8.6) shows that $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$ is unique only if $(\mathbf{Z}'\mathbf{Z})$ is a nonsingular $\ell \times \ell$ matrix and $(\mathbf{Z}'\mathbf{X})$ has full rank k. As long as the instruments are linearly independent, $(\mathbf{Z}'\mathbf{Z})$ will be a nonsingular $\ell \times \ell$ matrix, so this requirement is usually taken for granted. That

4. This terminology comes from literature on estimating the structural parameters in systems of simultaneous equations.

$(\mathbf{Z}'\mathbf{X})$ be of rank k is known as the *rank condition*. That $\ell \geq k$ is known as the *order condition*. Because the exogenous regressors in $\mathbf{X}$ serve as their own instruments, the order condition is often stated as requiring that there be at least as many instruments as endogenous variables. The order condition is necessary, but not sufficient, for the rank condition to hold.

If the rank condition fails, the equation is said to be *underidentified*, and no econometric procedure can produce consistent estimates. If the rank of $(\mathbf{Z}'\mathbf{X})$ is k, the equation is said to be *exactly identified*. If the rank of $(\mathbf{Z}'\mathbf{X}) > k$, the equation is said to be *overidentified*.

The rank condition requires only that there be enough correlation between the instruments and the endogenous variables to guarantee that we can compute unique parameter estimates. For the large-sample approximations to be useful, we need much higher correlations between the instruments and the regressors than the minimal level required by the rank condition. Instruments that satisfy the rank condition but are not sufficiently correlated with the endogenous variables for the large-sample approximations to be useful are known as *weak instruments*. We discuss weak instruments in section 8.10.

The parameters of exactly identified equations can be estimated by IV. The parameters of overidentified equations can be estimated by IV, after combining the instruments as in 2SLS. Although overidentification might sound like a nuisance to be avoided, it is actually preferable to working with an exactly identified equation. Overidentifying restrictions produce more efficient estimates in large samples. Furthermore, recall that the first essential property of an instrument is statistical independence from the disturbance process. Although we cannot test the validity of that assumption directly, we can assess the adequacy of instruments in an overidentified context with a *test of overidentifying restrictions*.

In such a test, the residuals from a 2SLS regression are regressed on all exogenous variables: both included exogenous regressors and excluded instruments. Under the null hypothesis that all instruments are uncorrelated with u, an LM statistic of the $N \times R^2$ form has a large-sample $\chi^2(r)$ distribution, where r is the number of *overidentifying restrictions*: the number of excess instruments. If we reject this hypothesis, we cast doubt on the suitability of the instrument set. One or more of the instruments do not appear to be uncorrelated with the disturbance process. This Sargan (1958) or Basmann (1960) test is available in Stata as the `overid` command (Baum, Schaffer, and Stillman 2003). This command can be installed from `ssc` for use after estimation with `ivreg`. I present an example of its use below.

8.6 Computing IV estimates

I illustrate how to use `ivreg` with a regression from Griliches (1976), a classic study of the wages of a sample of 758 young men.[5] Griliches models their wages as a function of several continuous factors: `s`, `expr`, and `tenure` (years of schooling, experience, and job tenure, respectively); `rns`, an indicator for residency in the South; `smsa`, an indicator for urban versus rural; and a set of year dummies since the data are a set of pooled cross sections. The endogenous regressor is `iq`, the worker's IQ score, which is considered as a potentially mismeasured version of ability. Here we do not consider that `wage` and `iq` are simultaneously determined, but rather that `iq` cannot be assumed independent of the error term: the same correlation that arises in the context of an endogenous regressor in a structural equation.[6] The IQ score is instrumented with four factors excluded from the equation: `med`, the mother's level of education; `kww`, the score on another standardized test; `age`, the worker's age; and `mrt`, an indicator of marital status. I present the descriptive statistics with `summarize` and then fit the IV model.

```
. use http://www.stata-press.com/data/imeus/griliches, clear
(Wages of Very Young Men, Zvi Griliches, J.Pol.Ec. 1976)
. summarize lw s expr tenure rns smsa iq med kww age mrt, sep(0)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lw	758	5.686739	.4289494	4.605	7.051
s	758	13.40501	2.231828	9	18
expr	758	1.735429	2.105542	0	11.444
tenure	758	1.831135	1.67363	0	10
rns	758	.2691293	.4438001	0	1
smsa	758	.7044855	.456575	0	1
iq	758	103.8562	13.61867	54	145
med	758	10.91029	2.74112	0	18
kww	758	36.57388	7.302247	12	56
age	758	21.83509	2.981756	16	30
mrt	758	.5145119	.5001194	0	1

We use the `first` option for `ivreg` to evaluate the degree of correlation between these four factors and the endogenous regressor `iq`:

5. These data were later used by Blackburn and Neumark (1992). I am grateful to Professor Fumio Hayashi for his permission to use the version of the Blackburn–Neumark data circulated as `grilic` with his econometrics textbook (Hayashi 2000).

6. Measurement error problems are discussed in appendix B to this chapter.

```
. ivreg lw s expr tenure rns smsa _I* (iq=med kww age mrt), first
First-stage regressions
```

Source	SS	df	MS		
Model	47176.4676	15	3145.09784	Number of obs =	758
Residual	93222.8583	742	125.637275	F(15, 742) =	25.03
				Prob > F =	0.0000
Total	140399.326	757	185.468066	R-squared =	0.3360
				Adj R-squared =	0.3226
				Root MSE =	11.209

iq	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
s	2.497742	.2858159	8.74	0.000	1.936638	3.058846
expr	-.033548	.2534458	-0.13	0.895	-.5311042	.4640082
tenure	.6158215	.2731146	2.25	0.024	.0796522	1.151991
rns	-2.610221	.9499731	-2.75	0.006	-4.475177	-.7452663
smsa	.0260481	.9222585	0.03	0.977	-1.784499	1.836595
_Iyear_67	.9254935	1.655969	0.56	0.576	-2.325449	4.176436
_Iyear_68	.4706951	1.574561	0.30	0.765	-2.620429	3.56182
_Iyear_69	2.164635	1.521387	1.42	0.155	-.8221007	5.15137
_Iyear_70	5.734786	1.696033	3.38	0.001	2.405191	9.064381
_Iyear_71	5.180639	1.562156	3.32	0.001	2.113866	8.247411
_Iyear_73	4.526686	1.48294	3.05	0.002	1.615429	7.437943
med	.2877745	.1622338	1.77	0.077	-.0307176	.6062665
kww	.4581116	.0699323	6.55	0.000	.3208229	.5954003
age	-.8809144	.2232535	-3.95	0.000	-1.319198	-.4426307
mrt	-.584791	.946056	-0.62	0.537	-2.442056	1.272474
_cons	67.20449	4.107281	16.36	0.000	59.14121	75.26776

Instrumental variables (2SLS) regression

Source	SS	df	MS		
Model	59.2679161	12	4.93899301	Number of obs =	758
Residual	80.0182337	745	.107407025	F(12, 745) =	45.91
				Prob > F =	0.0000
Total	139.28615	757	.183997556	R-squared =	0.4255
				Adj R-squared =	0.4163
				Root MSE =	.32773

lw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
iq	.0001747	.0039374	0.04	0.965	-.0075551	.0079044
s	.0691759	.013049	5.30	0.000	.0435587	.0947931
expr	.029866	.006697	4.46	0.000	.0167189	.0430132
tenure	.0432738	.0076934	5.62	0.000	.0281705	.058377
rns	-.1035897	.0297371	-3.48	0.001	-.1619682	-.0452111
smsa	.1351148	.0268889	5.02	0.000	.0823277	.1879019
_Iyear_67	-.052598	.0481067	-1.09	0.275	-.1470388	.0418428
_Iyear_68	.0794686	.0451078	1.76	0.079	-.009085	.1680222
_Iyear_69	.2108962	.0443153	4.76	0.000	.1238984	.2978939
_Iyear_70	.2386338	.0514161	4.64	0.000	.1376962	.3395714
_Iyear_71	.2284609	.0441236	5.18	0.000	.1418396	.3150823
_Iyear_73	.3258944	.0410718	7.93	0.000	.2452642	.4065247
_cons	4.39955	.2708771	16.24	0.000	3.867777	4.931323

```
Instrumented:  iq
Instruments:   s expr tenure rns smsa _Iyear_67 _Iyear_68
               _Iyear_69 _Iyear_70 _Iyear_71 _Iyear_73 med
               kww age mrt
```

The first-stage regression results suggest that three of the four excluded instruments are highly correlated with `iq`. The exception is `mrt`, the indicator of marital status. However, the endogenous regressor `iq` has an IV coefficient that cannot be distinguished from zero. Conditioning on the other factors included in the equation, `iq` does not seem to play an important role in determining the wage. The other coefficient estimates agree with the predictions of theory and empirical findings.

Are the instruments for `iq` appropriately uncorrelated with the disturbance process? To answer that, we compute the test for overidentifying restrictions:

```
. overid
Tests of overidentifying restrictions:
Sargan N*R-sq test        87.655  Chi-sq(3)     P-value = 0.0000
Basmann test              97.025  Chi-sq(3)     P-value = 0.0000
```

The above test signals a strong rejection of the null hypothesis that the instruments are uncorrelated with the error term and suggests that we should not be satisfied with this specification of the equation. We return to this example in the next section.

In the following sections, I present several topics related to the IV estimator and a generalization of that estimator. These capabilities are not provided by Stata's `ivreg` but are available in the extension of that routine known as `ivreg2` (Baum, Schaffer, and Stillman 2003, 2005).[7]

8.7 ivreg2 and GMM estimation

In defining the simple IV estimator and the 2SLS estimator, we assumed the presence of i.i.d. errors. As for linear regression, when the errors do not satisfy the i.i.d. assumption, the simple IV and 2SLS estimators produce consistent but inefficient estimates whose large-sample VCE must be estimated by a robust method. In another parallel to the linear regression case, there is a more general estimator based on the GMM that will produce consistent and efficient estimates in the presence of non-i.i.d. errors. Here I describe and illustrate this more general estimation technique.

The equation of interest is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \qquad E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \boldsymbol{\Omega}$$

The matrix of regressors $\mathbf{X}$ is $N \times k$, where N is the number of observations. The error term $\mathbf{u}$ is distributed with mean zero, and its covariance matrix $\boldsymbol{\Omega}$ is $N \times N$. We consider four cases for $\boldsymbol{\Omega}$: homoskedasticity, conditional heteroskedasticity, clustering, and the combination of heteroskedasticity and autocorrelation. The last three cases correspond to those described in section 6.1.

7. I am deeply indebted to collaborators Mark E. Schaffer and Steven Stillman for their efforts in crafting the software in the `ivreg2` suite of programs and its description in our cited article. Much of this section is adapted from that article and subsequent joint work.

Some of the regressors are endogenous, so $E[\mathbf{x}u] \neq \mathbf{0}$. We partition the set of regressors into $\{\mathbf{x}_1\ \mathbf{x}_2\}$ with the k_1 regressors $\mathbf{x}_1$ considered endogenous and the $(k-k_1)$ remaining regressors $\mathbf{x}_2$ assumed to be exogenous.

The matrix of instrumental variables $\mathbf{Z}$ is $N \times \ell$. These variables are assumed to be exogenous: $E[\mathbf{z}u] = \mathbf{0}$. We partition the instruments into $\{\mathbf{z}_1\ \mathbf{z}_2\}$ where the ℓ_1 instruments $\mathbf{z}_1$ are excluded instruments and the remaining $(\ell-\ell_1)$ instruments $\mathbf{z}_2 \equiv \mathbf{x}_2$ are the included instruments or exogenous regressors.

8.7.1 The GMM estimator

The standard IV and 2SLS estimators are special cases of the GMM estimator. As with the simple IV case discussed in section 8.2, the assumption that the instruments $\mathbf{z}$ are exogenous can be expressed as a set of *moment conditions* $E[\mathbf{z}u] = \mathbf{0}$. The ℓ instruments give us a set of ℓ moments:

$$g_i(\boldsymbol{\beta}) = \mathbf{Z}_i'u_i = \mathbf{Z}_i'(y_i - \mathbf{x}_i\boldsymbol{\beta})$$

where g_i is $\ell \times 1$.[8] Just as in method-of-moments estimators of linear regression and simple IV, each of the ℓ moment equations corresponds to a sample moment. We write these ℓ sample moments as

$$\overline{g}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i'(y_i - \mathbf{x}_i\boldsymbol{\beta}) = \frac{1}{N}\mathbf{Z}'\mathbf{u}$$

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

If the equation to be estimated is *exactly identified* ($\ell = k$), we have just as many moment conditions as we do unknowns. We can exactly solve the ℓ moment conditions for the k coefficients in $\widehat{\boldsymbol{\beta}}_{\text{GMM}}$. Here there is a unique $\widehat{\boldsymbol{\beta}}_{\text{GMM}}$ that solves $\overline{g}(\widehat{\boldsymbol{\beta}}_{\text{GMM}}) = 0$. This GMM estimator is identical to the standard IV estimator of (8.3).

If the equation is *overidentified*, $\ell > k$, we have more equations than we do unknowns. We will not be able to find a k-vector $\widehat{\boldsymbol{\beta}}_{\text{GMM}}$ that will set all ℓ sample moment conditions to zero. We want to choose $\widehat{\boldsymbol{\beta}}_{\text{GMM}}$ so that the elements of $\overline{g}(\widehat{\boldsymbol{\beta}}_{\text{GMM}})$ are as close to zero as possible. We could do so by minimizing $\overline{g}(\widehat{\boldsymbol{\beta}}_{\text{GMM}})'\overline{g}(\widehat{\boldsymbol{\beta}}_{\text{GMM}})$, but this method offers no way to produce more efficient estimates when the errors are not i.i.d. For this reason, the GMM estimator chooses the $\widehat{\boldsymbol{\beta}}_{\text{GMM}}$ that minimizes

$$J(\widehat{\boldsymbol{\beta}}_{\text{GMM}}) = N\ \overline{g}(\widehat{\boldsymbol{\beta}}_{\text{GMM}})'\mathbf{W}\overline{g}(\widehat{\boldsymbol{\beta}}_{\text{GMM}}) \tag{8.9}$$

where $\mathbf{W}$ is an $\ell \times \ell$ *weighting matrix* that accounts for the correlations among the $\overline{g}(\widehat{\boldsymbol{\beta}}_{\text{GMM}})$ when the errors are not i.i.d.

8. Because these conditions imply that $(\mathbf{Z}, \mathbf{u})$ will be uncorrelated, they are often termed *orthogonality conditions* in the literature.

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

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

yields the GMM estimator of an overidentified equation:

$$\widehat{\beta}_{\text{GMM}} = (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{y} \tag{8.10}$$

The results of the minimization—and hence the GMM estimator—will be identical for all weighting matrices $\mathbf{W}$ that differ by a constant of proportionality. We use knowledge of this fact below. However, there are as many GMM estimators as there are choices of weighting matrix $\mathbf{W}$. For an exactly identified equation, $\mathbf{W} = \mathbf{I}_N$. The weighting matrix only plays a role in the presence of overidentifying restrictions.

The *optimal* weighting matrix is that which produces the most efficient estimate.[9] Hansen (1982) showed that this process involves choosing $\mathbf{W} = \mathbf{S}^{-1}$, where $\mathbf{S}$ is the covariance matrix of the moment conditions g:

$$\mathbf{S} = E[\mathbf{Z}'\mathbf{u}\mathbf{u}'\mathbf{Z}] = E[\mathbf{Z}'\mathbf{\Omega}\mathbf{Z}] \tag{8.11}$$

where $\mathbf{S}$ is an $\ell \times \ell$ matrix. Substitute this matrix into (8.10) to obtain the efficient GMM estimator:

$$\widehat{\beta}_{\text{EGMM}} = (\mathbf{X}'\mathbf{Z}\mathbf{S}^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{S}^{-1}\mathbf{Z}'\mathbf{y}$$

Note the generality (the G of GMM) of this approach. We have made no assumptions about $\mathbf{\Omega}$, the covariance matrix of the disturbance process.[10] But the efficient GMM estimator is not a feasible estimator since the matrix $\mathbf{S}$ is not known. To implement the estimator, we need to estimate $\mathbf{S}$, so we must make some assumptions about $\mathbf{\Omega}$, as we discuss next.

Assume that we have developed a consistent estimator of $\mathbf{S}$, denoted $\widehat{\mathbf{S}}$. Generally such an estimator will involve the 2SLS residuals. Then we may use that estimator to define the feasible efficient two-step GMM estimator (FEGMM) implemented in `ivreg2` when the `gmm` option is used.[11] In the first step, we use standard 2SLS estimation to generate parameter estimates and residuals. In the second step, we use an assumption about the structure of $\mathbf{\Omega}$ to produce $\widehat{\mathbf{S}}$ from those residuals and define the FEGMM:

$$\widehat{\beta}_{\text{FEGMM}} = (\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{y}$$

8.7.2 GMM in a homoskedastic context

If we assume that $\mathbf{\Omega} = \sigma^2 I_N$, the optimal weighting matrix implied by (8.11) will be proportional to the identity matrix. Since no weighting is involved in calculating

9. Efficiency is described in section 4.2.3.

10. Aside from some conditions that guarantee $\frac{1}{\sqrt{N}}\mathbf{Z}'u$ is a vector of well-behaved random variables.

11. This estimator goes under various names: two-stage instrumental variables (2SIV), White (1982); two-step two-stage least squares, Cumby, Huizinga, and Obstfeld (1983); heteroskedastic two-stage least squares (H2SLS), Davidson and MacKinnon (1993, 599).

(8.10), the GMM estimator is merely the standard IV estimator in point and interval form. The IV estimator of (8.6) and (8.7) is the FEGMM estimator under conditional homoskedasticity of $\mathbf{\Omega}$.

8.7.3 GMM and heteroskedasticity-consistent standard errors

One of the most commonly encountered problems in economic data is heteroskedasticity of unknown form, as described in section 6.2. We need a heteroskedasticity-consistent estimator of $\mathbf{S}$. Such an $\widehat{\mathbf{S}}$ is available by using the standard sandwich approach to robust covariance estimation described in section 6.1.2. Define the 2SLS residuals as $\widehat{u}_i$ and the ith row of the instrument matrix as $\mathbf{Z}_i$. Then a consistent estimator of $\mathbf{S}$ is given by

$$\widehat{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{Z}_i'\mathbf{Z}_i$$

The residuals can come from any consistent estimator of $\boldsymbol{\beta}$ because efficiency of the parameter estimates used to compute the $\widehat{u}_i$ is not required. In practice, 2SLS residuals are almost always used. For more details, see Davidson and MacKinnon (1993, 607–610).

If the regression equation is exactly identified with $\ell = k$, the results from `ivreg2, gmm` will be identical to those of `ivreg2, robust` or from `ivreg` with the `robust` option. For overidentified models, the GMM approach makes more efficient use of the information in the ℓ moment conditions than the standard 2SLS approach that reduces them to k instruments in $\widehat{\mathbf{X}}$. The 2SLS estimator can be considered a GMM estimator with a suboptimal weighting matrix when the errors are not i.i.d.

To compare GMM with 2SLS, we reestimate the wage equation displayed earlier by using the `gmm` option. This step automatically generates heteroskedasticity-robust standard errors. By default, `ivreg2` reports large-sample `z` statistics for the coefficients.

(*Continued on next page*)

```
. ivreg2 lw s expr tenure rns smsa _I* (iq=med kww age mrt), gmm
GMM estimation
------------------------------------------------------------------------------

                                                      Number of obs =       758
                                                      F( 12,   745) =     49.67
                                                      Prob > F      =    0.0000
Total (centered) SS     =  139.2861498                Centered R2   =    0.4166
Total (uncentered) SS   =  24652.24662                Uncentered R2 =    0.9967
Residual SS             =  81.26217887                Root MSE      =     .3274

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |  -.0014014   .0041131    -0.34   0.733    -.009463    .0066602
           s |   .0768355   .0131859     5.83   0.000     .0509915    .1026794
        expr |   .0312339   .0066931     4.67   0.000     .0181157    .0443522
      tenure |   .0489998   .0073437     6.67   0.000     .0346064    .0633931
         rns |  -.1006811   .0295887    -3.40   0.001    -.1586738   -.0426884
        smsa |   .1335973   .0263245     5.08   0.000     .0820021    .1851925
    _Iyear_67 |  -.0210135   .0455433    -0.46   0.645    -.1102768    .0682498
    _Iyear_68 |   .0890993    .042702     2.09   0.037     .0054049    .1727937
    _Iyear_69 |   .2072484   .0407995     5.08   0.000     .1272828     .287214
    _Iyear_70 |   .2338308   .0528512     4.42   0.000     .1302445    .3374172
    _Iyear_71 |   .2345525   .0425661     5.51   0.000     .1511244    .3179805
    _Iyear_73 |   .3360267   .0404103     8.32   0.000     .2568239    .4152295
        _cons |   4.436784   .2899504    15.30   0.000     3.868492    5.005077
------------------------------------------------------------------------------
Anderson canon. corr. LR statistic (identification/IV relevance test):  54.338
                                                      Chi-sq(4) P-val =  0.0000
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):        74.165
                                                      Chi-sq(3) P-val =  0.0000
------------------------------------------------------------------------------
Instrumented:         iq
Included instruments: s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69
                      _Iyear_70 _Iyear_71 _Iyear_73
Excluded instruments: med kww age mrt
------------------------------------------------------------------------------
```

We see that the endogenous regressor `iq` still does not play a role in the equation. The Hansen J statistic displayed by `ivreg2` is the GMM equivalent of the Sargan test produced by `overid` above. The independence of the instruments and the disturbance process is called into question by this strong rejection of the J test null hypothesis.

8.7.4 GMM and clustering

When the disturbances have within-cluster correlation, `ivreg2` can compute the cluster–robust estimator of the VCE, and it can optionally use the cluster–robust estimator of $\widehat{\mathbf{S}}$ to produce more efficient parameter estimates when the model is overidentified. A consistent estimate of $\mathbf{S}$ in the presence of within-cluster correlated disturbances is

$$\widehat{\mathbf{S}} = \sum_{j=1}^{M} \widehat{\mathbf{u}}_j' \widehat{\mathbf{u}}_j$$

where

$$\widehat{\mathbf{u}}_j = (y_j - \mathbf{x}_j\widehat{\boldsymbol{\beta}})\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_j$$

and y_j is the jth observation on y, $\mathbf{x}_j$ is the jth row of $\mathbf{X}$, and $\mathbf{z}_j$ is the jth row of $\mathbf{Z}$. The fact that we are summing over the M clusters instead of the N observations leads to the intuition that we essentially do not have N observations: we have M, where M is the number of clusters. That number, M, must exceed ℓ if we are to estimate the equation, since $M - \ell$ is the effective degrees of freedom of the cluster estimator. If we are using a sizable number of instruments, this constraint may bind.[12]

Specifying the `cluster()` option will cause `ivreg2` to compute the cluster–robust estimator of the VCE. If the equation is overidentified, adding the `gmm` option will cause `ivreg2` to use the cluster–robust estimate of $\mathbf{S}$ to compute more efficient parameter estimates.

8.7.5 GMM and HAC standard errors

When the disturbances are conditionally heteroskedastic and autocorrelated, we can compute HAC estimates of the VCE and if the equation is overidentified, we can optionally use an HAC estimate of $\mathbf{S}$ to compute more efficient parameter estimates. The `ivreg2` routine will compute Newey–West estimates of the VCE using the Bartlett-kernel weighting when the `robust` and `bw()` options are specified. When there are no endogenous regressors, the results will be the same as those computed by `newey`. If some of the regressors are endogenous, then specifying the `robust` and `bw()` options will cause `ivreg2` to compute an HAC estimator of the VCE. If the equation is overidentified and the `robust` and `gmm` options are specified, the resulting GMM estimates will be more efficient than those produced by 2SLS.

The number specified in the `bw()` (bandwidth) option should be greater than that specified in the `lag()` option in `newey`. The `ivreg2` routine lets us choose several alternative kernel estimators (see the `kernel()` option) as described in the online help for that command.[13]

To illustrate, we estimate a Phillips curve relationship with annual time-series data for the United States, 1948–1996. The descriptive statistics for consumer price inflation (`cinf`) and the unemployment rate (`unem`) are as follows:

```
. use http://www.stata-press.com/data/imeus/phillips, clear
. summarize cinf unem if cinf < .
```

Variable	Obs	Mean	Std. Dev.	Min	Max
cinf	48	-.10625	2.566926	-9.3	6.6
unem	48	5.78125	1.553261	2.9	9.7

12. Official `ivreg` is more forgiving and will complain only if $M < k$. On the other hand, `ivreg2` insists that $M > \ell$.

13. If we do not doubt the homoskedasticity assumption but want to deal with autocorrelation of unknown form, we should use the AC correction without the H correction for arbitrary heteroskedasticity. `ivreg2` allows us to select H, AC, or HAC VCEs by combining the `robust` (or `gmm`), `bw()`, and `kernel()` options.

A Phillips curve is the relation between price or wage inflation and the unemployment rate. In Phillips' model, these variables should have a negative relationship with lower unemployment leading to inflationary pressures. Since both variables are determined within the macroeconomic environment, we cannot consider either as exogenous.

Using these data, we regress the rate of consumer price inflation on the unemployment rate. To deal with simultaneity, we instrument the unemployment rate with its second and third lags. Specifying `bw(3)`, `gmm`, and `robust` causes `ivreg2` to compute the efficient GMM estimates.

```
. ivreg2 cinf (unem = l(2/3).unem), bw(3) gmm robust
GMM estimation
---------------

Heteroskedasticity and autocorrelation-consistent statistics
  kernel=Bartlett; bandwidth=3
  time variable (t):  year
                                                    Number of obs =        46
                                                    F(  1,    44) =      0.39
                                                    Prob > F      =    0.5371
Total (centered) SS     =  217.4271745              Centered R2   =   -0.1266
Total (uncentered) SS   =  217.4900005              Uncentered R2 =   -0.1262
Residual SS             =  244.9459113              Root MSE      =     2.308

------------------------------------------------------------------------------
             |               Robust
        cinf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |   .1949334   .3064662     0.64   0.525    -.4057292     .795596
       _cons |  -1.144072   1.686995    -0.68   0.498    -4.450522    2.162378
------------------------------------------------------------------------------
Anderson canon. corr. LR statistic (identification/IV relevance test):  13.545
                                                    Chi-sq(2) P-val =    0.0011
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.589
                                                    Chi-sq(1) P-val =    0.4426
------------------------------------------------------------------------------
Instrumented:         unem
Excluded instruments: L2.unem L3.unem
------------------------------------------------------------------------------
```

The hypothesized relationship is not borne out by these estimates, as many researchers have found. The original relationship over the period ending in the late 1960s broke down badly in the presence of 1970s supply shocks and high inflation. To focus on the IV technique, we see that the Hansen J test statistic indicates that the instruments are appropriately uncorrelated with the disturbance process. If the first and second lags of `unem` are used, the J test rejects its null with a p-value of 0.02. The first lag of `unem` appears to be inappropriate as an instrument in this specification.

8.8 Testing overidentifying restrictions in GMM

Just as for 2SLS (see section 8.5), the validity of the overidentifying restrictions imposed on a GMM estimator can be tested. The test, which can and should be performed as a

standard diagnostic in any overidentified model,[14] has a null hypothesis of correct model specification and valid overidentifying restrictions. A rejection calls either or both of those hypotheses into question.

With GMM, the overidentifying restrictions may be tested by the commonly used J statistic of Hansen (1982).[15] This statistic is merely the value of the GMM objective function (8.9), evaluated at the efficient GMM estimator $\widehat{\beta}_{\text{EGMM}}$. Under the null,

$$J(\widehat{\beta}_{\text{EGMM}}) = N\ \overline{g}(\widehat{\beta}_{\text{EGMM}})'\widehat{\mathbf{S}}^{-1}\overline{g}(\widehat{\beta}_{\text{EGMM}}) \ \overset{A}{\sim}\ \chi^2_{\ell-k}$$

where the matrix $\widehat{\mathbf{S}}$ is estimated using the two-step methods described above.

The J statistic is asymptotically distributed as χ^2 with degrees of freedom equal to the number of overidentifying restrictions $\ell - k$ rather than the total number of moment conditions, ℓ. In effect, k degrees of freedom are spent in estimating the coefficients β. Hansen's J is the most common diagnostic used in GMM estimation to evaluate the suitability of the model. A rejection of the null hypothesis implies that the instruments do not satisfy the required orthogonality conditions—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` or `robust` options is specified.[16]

8.8.1 Testing a subset of the overidentifying restrictions in GMM

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 little power. Another common problem arises when you have suspicions about the validity of a subset of instruments and want to test them.

In these contexts, you can use a *difference-in-Sargan* test.[17] 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 J statistics. The first is computed from the fully efficient regression using the entire set of overidentifying restrictions. The second is that of the inefficient but consistent regression using a smaller set of restrictions in which a specified set of instruments are removed from the instrument list. For excluded instruments, this

14. 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".

15. For conditional homoskedasticity (see section 8.7.2), this statistic is numerically identical to the Sargan test statistic discussed above.

16. Despite the importance of testing the overidentifying restrictions, the J test is known to overreject the null hypothesis in certain circumstances. Using the "continuous updating" GMM estimator discussed in the help file for `ivreg2` may produce rejection rates that are closer to the level of the test. See Hayashi (2000, 218) for more information.

17. See Hayashi (2000, 218–221 and 232–234) or Ruud (2000, chap. 22), for comprehensive presentations. The test is known under other names as well; e.g., Ruud (2000) calls it the distance difference statistic, and Hayashi (2000) follows Eichenbaum, Hansen, and Singleton (1988) and dubs it the C statistic. I use the latter term.

step is equivalent to dropping them from the instrument list. For included instruments, the C test places them in the list of included endogenous variables, treating them as endogenous regressors. The order condition must still be satisfied for this form of the equation. Under the null hypothesis that the specified variables are proper instruments, the difference-in-Sargan C test statistic is distributed χ^2 with degrees of freedom equal to the loss of overidentifying restrictions or the number of suspect instruments being tested.[18]

Specifying `orthog(`*instlist*`)` with the suspect instruments causes `ivreg2` to compute the C test with *instlist* as the excluded instruments. The equation must still be identified with these instruments removed (or placed in the endogenous regressor list) to compute the C test. If the equation excluding suspect instruments is exactly identified, the J statistic for that equation will be zero and the C statistic will coincide with the statistic for the original equation. This property illustrates how the J test of overidentifying restrictions is an *omnibus* test for the failure of any of the instruments to satisfy the orthogonality conditions. At the same time, the test requires that the investigator believe the nonsuspect instruments to be valid (see Ruud 2000, 577).

Below we use the C statistic to test whether `s`, years of schooling, is a valid instrument in the wage equation estimated above by 2SLS and GMM. In those examples, the Sargan and J tests of overidentifying restrictions signaled a problem with the instruments used.

18. 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. For 2SLS, 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 2SLS regression is used to calculate the Sargan statistic for the regression with nonsuspect instruments as well. The equivalent procedure in GMM is to use the $\widehat{\mathbf{S}}$ matrix from the original estimation to calculate both J statistics. More precisely, $\widehat{\mathbf{S}}$ from the original equation is used to form the first J statistic, and the submatrix of $\widehat{\mathbf{S}}$ with rows/columns corresponding to the reduced set of instruments is used to form the J statistic for the second equation (see Hayashi 2000, 220).

```
. ivreg2 lw s expr tenure rns smsa _I* (iq=med kww age mrt), gmm orthog(s)
GMM estimation
------------------

                                                Number of obs =       758
                                                F( 12,   745) =     49.67
                                                Prob > F      =    0.0000
Total (centered) SS     =  139.2861498          Centered R2   =    0.4166
Total (uncentered) SS   =  24652.24662          Uncentered R2 =    0.9967
Residual SS             =  81.26217887          Root MSE      =     .3274

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |  -.0014014   .0041131    -0.34   0.733    -.009463    .0066602
           s |   .0768355   .0131859     5.83   0.000     .0509915    .1026794
        expr |   .0312339   .0066931     4.67   0.000     .0181157    .0443522
      tenure |   .0489998   .0073437     6.67   0.000     .0346064    .0633931
         rns |  -.1006811   .0295887    -3.40   0.001    -.1586738   -.0426884
        smsa |   .1335973   .0263245     5.08   0.000     .0820021    .1851925
   _Iyear_67 |  -.0210135   .0455433    -0.46   0.645    -.1102768    .0682498
   _Iyear_68 |   .0890993    .042702     2.09   0.037     .0054049    .1727937
   _Iyear_69 |   .2072484   .0407995     5.08   0.000     .1272828     .287214
   _Iyear_70 |   .2338308   .0528512     4.42   0.000     .1302445    .3374172
   _Iyear_71 |   .2345525   .0425661     5.51   0.000     .1511244    .3179805
   _Iyear_73 |   .3360267   .0404103     8.32   0.000     .2568239    .4152295
       _cons |   4.436784   .2899504    15.30   0.000     3.868492    5.005077
------------------------------------------------------------------------------
Anderson canon. corr. LR statistic (identification/IV relevance test):  54.338
                                                Chi-sq(4) P-val =    0.0000
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):        74.165
                                                Chi-sq(3) P-val =    0.0000
-orthog- option:
Hansen J statistic (eqn. excluding suspect orthog. conditions):         15.997
                                                Chi-sq(2) P-val =    0.0003
C statistic (exogeneity/orthogonality of suspect instruments):          58.168
                                                Chi-sq(1) P-val =    0.0000
Instruments tested:  s
------------------------------------------------------------------------------
Instrumented:         iq
Included instruments: s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69
                      _Iyear_70 _Iyear_71 _Iyear_73
Excluded instruments: med kww age mrt
------------------------------------------------------------------------------
```

The C test rejects its null, indicating that the suspect instrument, `s`, fails the test for overidentifying restrictions. The significant J statistic of 15.997 for the equation excluding suspect instruments implies that treating s as endogenous still results in an unsatisfactory equation. The remaining instruments do not appear to be independent of the error distribution.

Now we use the `orthog()` option to test whether a subset of the excluded instruments are appropriately exogenous. We include age and the marital status indicator (`age` and `mrt`) in the option's *varlist*. The equation estimated without suspect instruments merely drops those instruments from the list of excluded instruments:

```
. ivreg2 lw s expr tenure rns smsa _I* (iq=med kww age mrt), gmm orthog(age mrt)
GMM estimation
------------------

                                                      Number of obs =       758
                                                      F( 12,   745) =     49.67
                                                      Prob > F      =    0.0000
Total (centered) SS     =  139.2861498                Centered R2   =    0.4166
Total (uncentered) SS   =  24652.24662                Uncentered R2 =    0.9967
Residual SS             =  81.26217887                Root MSE      =     .3274

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |  -.0014014   .0041131    -0.34   0.733    -.009463    .0066602
           s |   .0768355   .0131859     5.83   0.000    .0509915    .1026794
        expr |   .0312339   .0066931     4.67   0.000    .0181157    .0443522
      tenure |   .0489998   .0073437     6.67   0.000    .0346064    .0633931
         rns |  -.1006811   .0295887    -3.40   0.001   -.1586738   -.0426884
        smsa |   .1335973   .0263245     5.08   0.000    .0820021    .1851925
   _Iyear_67 |  -.0210135   .0455433    -0.46   0.645   -.1102768    .0682498
   _Iyear_68 |   .0890993    .042702     2.09   0.037    .0054049    .1727937
   _Iyear_69 |   .2072484   .0407995     5.08   0.000    .1272828     .287214
   _Iyear_70 |   .2338308   .0528512     4.42   0.000    .1302445    .3374172
   _Iyear_71 |   .2345525   .0425661     5.51   0.000    .1511244    .3179805
   _Iyear_73 |   .3360267   .0404103     8.32   0.000    .2568239    .4152295
       _cons |   4.436784   .2899504    15.30   0.000    3.868492    5.005077
------------------------------------------------------------------------------

Anderson canon. corr. LR statistic (identification/IV relevance test):   54.338
                                                      Chi-sq(4) P-val =    0.0000
------------------------------------------------------------------------------

Hansen J statistic (overidentification test of all instruments):         74.165
                                                      Chi-sq(3) P-val =    0.0000

-orthog- option:
Hansen J statistic (eqn. excluding suspect orthog. conditions):           1.176
                                                      Chi-sq(1) P-val =    0.2782
C statistic (exogeneity/orthogonality of suspect instruments):           72.989
                                                      Chi-sq(2) P-val =    0.0000
Instruments tested:  age mrt
------------------------------------------------------------------------------

Instrumented:         iq
Included instruments: s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69
                      _Iyear_70 _Iyear_71 _Iyear_73
Excluded instruments: med kww age mrt
------------------------------------------------------------------------------
```

The equation estimated without suspect instruments, free of the two additional orthogonality conditions on `age` and `mrt`, has an insignificant J statistic, whereas the C statistic for those two instruments is highly significant. These two instruments do not appear valid in this context. To evaluate whether we have found a more appropriate specification, we reestimate the equation with the reduced instrument list:

```
. ivreg2 lw s expr tenure rns smsa _I* (iq=med kww), gmm
GMM estimation
------------------------------------------------------------------------------

                                                      Number of obs =       758
                                                      F( 12,   745) =     30.77
                                                      Prob > F      =    0.0000
Total (centered) SS     =  139.2861498                Centered R2   =    0.1030
Total (uncentered) SS   =  24652.24662                Uncentered R2 =    0.9949
Residual SS             =  124.9413508                Root MSE      =      .406

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0240417   .0060961     3.94   0.000     .0120936    .0359899
           s |   .0009181   .0194208     0.05   0.962    -.0371459     .038982
        expr |   .0393333   .0088012     4.47   0.000     .0220833    .0565834
      tenure |   .0324916   .0091223     3.56   0.000     .0146122     .050371
         rns |  -.0326157   .0376679    -0.87   0.387    -.1064433     .041212
        smsa |    .114463   .0330718     3.46   0.001     .0496434    .1792825
    _Iyear_67 |  -.0694178   .0568781    -1.22   0.222    -.1808968    .0420613
    _Iyear_68 |   .0891834   .0585629     1.52   0.128    -.0255977    .2039645
    _Iyear_69 |   .1780712   .0532308     3.35   0.001     .0737407    .2824016
    _Iyear_70 |    .139594   .0677261     2.06   0.039     .0068533    .2723346
    _Iyear_71 |   .1730151   .0521623     3.32   0.001      .070779    .2752512
    _Iyear_73 |    .300759   .0490919     6.13   0.000     .2045407    .3969772
        _cons |   2.859113   .4083706     7.00   0.000     2.058721    3.659504
------------------------------------------------------------------------------
Anderson canon. corr. LR statistic (identification/IV relevance test):  35.828
                                                      Chi-sq(2) P-val =  0.0000
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.781
                                                      Chi-sq(1) P-val =  0.3768
------------------------------------------------------------------------------
Instrumented:         iq
Included instruments: s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69
                      _Iyear_70 _Iyear_71 _Iyear_73
Excluded instruments: med kww
------------------------------------------------------------------------------
```

In these results, we find that in line with theory, `iq` appears as a significant regressor for the first time and the equation's J statistic is satisfactory. The regressor `s`, which appeared in an earlier test to be inappropriately considered exogenous, plays no role in this form of the estimated equation.[19]

8.9 Testing for heteroskedasticity in the IV context

This section discusses the Pagan and Hall (1983) test for heteroskedasticity in 2SLS models and the `ivhettest` command (Baum, Schaffer, and Stillman 2003), which im-

19. You might wonder why the J statistic for this equation is not equal to that of the equation lacking the suspect instruments in the C test above. As explained in an earlier footnote, a positive C statistic is guaranteed by computing both J statistics using the estimated error variance of the full equation. Those two error variances differ—the equation above has a larger `Root MSE`—so that the J statistics differ as well.

plements this test in Stata. The idea behind the test—similar to that of the Breusch–Pagan (Breusch and Pagan 1979) and White tests for heteroskedasticity discussed in section 6.2.1—is that if any of the exogenous variables can predict the squared residuals, the errors are conditionally heteroskedastic.[20] Under the null of conditional homoskedasticity in the 2SLS regression, the Pagan–Hall statistic is distributed as χ^2_p, irrespective of the presence of heteroskedasticity elsewhere in the system.[21]

The `ivhettest` command follows the abbreviated syntax:

`ivhettest` [*varlist*] [, *options*]

where the optional *varlist* specifies the exogenous variables to be used to model the squared errors. Common choices for those variables include the following:

1. The levels only of the instruments Z (excluding the constant). This choice is available in `ivhettest` by specifying the `ivlev` option, which is the default option.
2. The levels and squares of the instruments Z, available as the `ivsq` option.
3. The levels, squares, and cross products of the instruments Z (excluding the constant), as in the White (1980) test: available as the `ivcp` option.
4. The fitted value of the response variable.[22] This choice is available in `ivhettest` by specifying the `fitlev` option.
5. The fitted value of the response variable and its square, available as the `fitsq` option.
6. A user-defined set of variables may also be provided.

The tradeoff in the choice of variables to be used is that a smaller set of variables will conserve degrees of freedom, at the cost of being unable to detect heteroskedasticity in certain directions.

The Pagan–Hall statistic has not been widely used, perhaps because it is not a standard feature of most regression packages.[23] However, from an analytical standpoint, it is clearly superior to the techniques more commonly used since it is robust to the presence of heteroskedasticity elsewhere in a system of simultaneous equations and to non–normally distributed disturbances.[24]

20. The Breusch–Pagan and White tests for heteroskedasticity (Breusch and Pagan 1979) discussed in section 6.2.1 can be applied in 2SLS models, but Pagan and Hall (1983) point out that they will be valid 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 must also be homoskedastic even though they are not being explicitly estimated.

21. A more general form of this test was separately proposed by White (1982).

22. This fitted value is not the usual fitted value of the response variable, $X\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}$. It is, rather, $\widehat{X}\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}$, i.e., the prediction based on the IV estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}$, the exogenous regressors Z_2, and the fitted values of the endogenous regressors $\widehat{X}_1$.

23. Although we discuss its use in 2SLS, `ivhettest` may also be used after `regress`.

24. White's general test (White 1980), or its generalization by Koenker (1981), also relaxes the assumption of normality underlying the Breusch–Pagan test.

We compute several of the tests for heteroskedasticity appropriate in the IV context with `ivhettest` from the last regression reported above. The default setting uses the levels of the instruments as associated variables. Results from the `fitsq` option are also displayed.

```
. ivhettest, all
IV heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    Pagan-Hall general test statistic   :   8.645  Chi-sq(13) P-value = 0.7992
    Pagan-Hall test w/assumed normality :   9.539  Chi-sq(13) P-value = 0.7311
    White/Koenker nR2 test statistic    :  13.923  Chi-sq(13) P-value = 0.3793
    Breusch-Pagan/Godfrey/Cook-Weisberg :  15.929  Chi-sq(13) P-value = 0.2530
. ivhettest, fitsq all
IV heteroskedasticity test(s) using fitted value (X-hat*beta-hat) & its square
Ho: Disturbance is homoskedastic
    Pagan-Hall general test statistic   :   0.677  Chi-sq(2) P-value = 0.7127
    Pagan-Hall test w/assumed normality :   0.771  Chi-sq(2) P-value = 0.6799
    White/Koenker nR2 test statistic    :   0.697  Chi-sq(2) P-value = 0.7056
    Breusch-Pagan/Godfrey/Cook-Weisberg :   0.798  Chi-sq(2) P-value = 0.6710
```

None of the tests signal any problem of heteroskedasticity in the estimated equation's disturbance process.

8.10 Testing the relevance of instruments

As discussed above, an instrumental variable must not be correlated with the equation's disturbance process and it must be highly correlated with the included endogenous regressors. We may test the latter condition by examining the fit of the first-stage regressions. The first-stage regressions are reduced-form regressions of the endogenous regressors, $\mathbf{x}_1$, on the full set of instruments, $\mathbf{z}$. The relevant test statistics here relate to the explanatory power of the excluded instruments, $\mathbf{z}_1$, in these regressions. A statistic commonly used, as recommended by Bound, Jaeger, and Baker (1995), is the R^2 of the first-stage regression with the included instruments partialled out.[25] This test may be expressed as the F test of the joint significance of the $\mathbf{z}_1$ instruments in the first-stage regression. But the distribution of this F statistic is nonstandard.[26] Also, for models with multiple endogenous variables, these indicators may not be sufficiently informative.

To grasp the pitfalls facing empirical researchers here, consider the following simple example. You have 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. Your model is basically *underidentified*. You have one valid instrument but two endogenous regressors. The Bound, Jaeger, and Baker F statistics and partial R^2 measures from the two

25. More precisely, this is the squared partial correlation between the excluded instruments $\mathbf{z}_1$ and the endogenous regressor in question. It is defined as $(\text{RSS}_{\mathbf{z}_2} - \text{RSS}_{\mathbf{z}})/\text{TSS}$, where $\text{RSS}_{\mathbf{z}_2}$ is the residual sum of squares in the regression of the endogenous regressor on $\mathbf{z}_2$ and $\text{RSS}_{\mathbf{z}}$ is the RSS when the full set of instruments is used.

26. See Bound, Jaeger, and Baker (1995) and Stock, Wright, and Yogo (2002).

first-stage regressions will not reveal this weakness. Indeed, the F statistics will be statistically significant, and without investigation you may not realize that the model cannot be estimated in this form. To deal with this problem of instrument irrelevance, either more relevant instruments are needed or one of the endogenous regressors must be dropped from the model. The statistics proposed by Bound, Jaeger, and Baker can diagnose instrument relevance only in the presence of one endogenous regressor. 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.[27] For a model containing one 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, Jaeger, and Baker 1995) partial R^2 and a small value of the Shea measure, you should conclude that the instruments lack sufficient relevance to explain all the endogenous regressors. Your model may be essentially underidentified. The Bound, Jaeger, and Baker measures and the Shea partial R^2 statistic are provided by the `first` or `ffirst` options of the `ivreg2` command.

A more general approach to the problem of instrument relevance was proposed by Anderson (1984) and discussed in Hall, Rudebusch, and Wilcox (1996).[28] Anderson's approach considers the *canonical correlations* of the $\mathbf{X}$ and $\mathbf{Z}$ matrices. These measures, r_i, $i = 1, \ldots, k$ represent the correlations between linear combinations of the k columns of $\mathbf{X}$ and linear combinations of the ℓ columns of $\mathbf{Z}$.[29] If an equation to be estimated by instrumental variables is identified from a numerical standpoint, all k of the canonical correlations must be significantly different from zero. Anderson's likelihood-ratio test has the null hypothesis that the smallest canonical correlation is zero and assumes that the regressors are distributed multivariate normal. Under the null, the test statistic is distributed χ^2 with $(\ell - k + 1)$ degrees of freedom, so that it may be calculated even for an exactly identified equation. A failure to reject the null hypothesis calls the identification status of the estimated equation into question. The Anderson statistic is displayed in `ivreg2`'s standard output.

The canonical correlations between $\mathbf{X}$ and $\mathbf{Z}$ may also be used to test a set of instruments for *redundancy* following Hall and Peixe (2000). In an overidentified context with $\ell \geq k$, if some of the instruments are redundant then the large-sample efficiency of the estimation is not improved by including them. The test statistic is a likelihood-ratio statistic based on the canonical correlations with and without the instruments being

27. 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 $R^2_p = (\nu_{i,i,\mathrm{OLS}})/(\nu_{i,i,\mathrm{IV}}) \left\{(1 - R^2_{\mathrm{IV}})/(1 - R^2_{\mathrm{OLS}})\right\}$, where $\nu_{i,i}$ is the estimated asymptotic variance of the coefficient.

28. Hall, Rudebusch, and Wilcox state that the test is closely related to the minimum-eigenvalue test statistic proposed by Cragg and Donald (1993). This test is displayed with the `first` or `ffirst` option of `ivreg2`: see the following example.

29. The squared canonical correlations may be calculated as the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})$; see Hall, Rudebusch, and Wilcox (1996, 287).

tested. Under the null hypothesis that the specified instruments are redundant, the statistic is distributed as χ^2 with degrees of freedom equal to the number of endogenous regressors times the number of instruments being tested. Like the Anderson test, the redundancy test assumes that the regressors are distributed multivariate normal. This test is available in `ivreg2` with the `redundant()` option.

I illustrate the weak-instruments problem with a variation on the log wage equation using only `age` and `mrt` as instruments.

```
. ivreg2 lw s expr tenure rns smsa _I* (iq = age mrt), ffirst redundant(mrt)

Summary results for first-stage regressions
-------------------------------------------

                Shea
Variable     | Partial R2     |    Partial R2    F(  2,   744)    P-value
iq           |   0.0073       |      0.0073           2.72          0.0665

Underidentification tests:
                                                    Chi-sq(2)      P-value
Anderson canon. corr. likelihood ratio stat.          5.52          0.0632
Cragg-Donald N*minEval stat.                          5.54          0.0626
Ho: matrix of reduced form coefficients has rank=K-1 (underidentified)
Ha: matrix has rank>=K (identified)

Weak identification statistics:
Cragg-Donald (N-L)*minEval/L2 F-stat       2.72

Anderson-Rubin test of joint significance of
endogenous regressors B1 in main equation, Ho:B1=0
  F(2,744)=      43.83     P-val=0.0000
  Chi-sq(2)=     89.31     P-val=0.0000

Number of observations N           =        758
Number of regressors   K           =         13
Number of instruments  L           =         14
Number of excluded instruments L2  =          2
```

(*Continued on next page*)

```
Instrumental variables (2SLS) regression
```

		Number of obs =	758
		F(12, 745) =	3.95
		Prob > F =	0.0000
Total (centered) SS =	139.2861498	Centered R2 =	-6.4195
Total (uncentered) SS =	24652.24662	Uncentered R2 =	0.9581
Residual SS =	1033.432656	Root MSE =	1.168

lw	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
iq	-.0948902	.0433073	-2.19	0.028	-.1797708	-.0100095
s	.3397121	.125526	2.71	0.007	.0936856	.5857386
expr	-.006604	.028572	-0.23	0.817	-.062604	.0493961
tenure	.0848854	.0327558	2.59	0.010	.0206852	.1490856
rns	-.3769393	.1584438	-2.38	0.017	-.6874834	-.0663952
smsa	.2181191	.1022612	2.13	0.033	.0176908	.4185474
_Iyear_67	.0077748	.1733579	0.04	0.964	-.3320005	.3475501
_Iyear_68	.0377993	.1617101	0.23	0.815	-.2791466	.3547452
_Iyear_69	.3347027	.1666592	2.01	0.045	.0080568	.6613487
_Iyear_70	.6286425	.2486186	2.53	0.011	.141359	1.115926
_Iyear_71	.4446099	.182733	2.43	0.015	.0864599	.8027599
_Iyear_73	.439027	.1542401	2.85	0.004	.136722	.7413321
_cons	10.55096	2.821406	3.74	0.000	5.02111	16.08082

```
Anderson canon. corr. LR statistic (identification/IV relevance test):    5.522
                                                  Chi-sq(2) P-val =    0.0632
-redundant- option:
LR IV redundancy test (redundancy of specified instruments):              0.002
                                                  Chi-sq(1) P-val =    0.9685
Instruments tested:   mrt
```

```
Sargan statistic (overidentification test of all instruments):            1.393
                                                  Chi-sq(1) P-val =    0.2379
```

```
Instrumented:         iq
Included instruments: s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69
                      _Iyear_70 _Iyear_71 _Iyear_73
Excluded instruments: age mrt
```

In the first-stage regression results, Shea's partial R^2 statistic is very small for this equation, and the Cragg–Donald statistic marginally rejects its null hypothesis of underidentification. The Anderson canonical correlation statistic fails to reject its null hypothesis at the 5% level, suggesting that although we have more instruments than coefficients the instruments may be inadequate to identify the equation. The `redundant(mrt)` option indicates that `mrt` provides no useful information to identify the equation. This equation may be only exactly identified.

The consequence of excluded instruments with little explanatory power is increased bias in the estimated IV coefficients (Hahn and Hausman 2002b) and worsening of the large-sample approximations to the finite-sample distributions. If these instruments' explanatory power in the first-stage regression is nil, the model is in effect unidentified with respect to that endogenous variable. Here the large-sample bias of the IV

estimator is the same as that of the OLS estimator, IV becomes inconsistent, and nothing is gained from instrumenting (Hahn and Hausman 2002b). 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 one endogenous regressor, an F statistic less than 10 is cause for concern (Staiger and Stock 1997, 557). The magnitude of large-sample bias of the IV estimator increases with the number of instruments (Hahn and Hausman 2002b). Given that, one recommendation when faced with a weak-instrument problem is to be parsimonious in the choice of instruments. For further discussion, see Staiger and Stock (1997); Hahn and Hausman (2002a,b); Stock, Wright, and Yogo (2002); Chao and Swanson (2005); and references therein.

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

There may well be reason to suspect a failure of the zero-conditional-mean assumption presented in section 4.2 in many regression models. Turning to IV or efficient GMM estimation for the sake of consistency must be balanced against the inevitable loss of efficiency. As Wooldridge states, "[there is an] important cost of performing IV estimation when $\mathbf{x}$ 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 2006, 516; emphasis added). This loss of efficiency is a price worth paying if the OLS estimator is biased and inconsistent. A test of the appropriateness of OLS and the necessity to resort to IV or GMM methods would be useful.[30] The intuition for such a test may also be couched in the number of orthogonality conditions available. Can 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.

Many econometrics texts discuss the issue of OLS versus IV in the context of the Durbin–Wu–Hausman (DWH) tests. These tests involve fitting the model by 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 $\widehat{\boldsymbol{\beta}}^c$ the estimator that is consistent under both the null and the alternative hypotheses, and by $\widehat{\boldsymbol{\beta}}^e$ 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

$$H = (\widehat{\boldsymbol{\beta}}_c - \widehat{\boldsymbol{\beta}}_e)' \mathbf{D}^- (\widehat{\boldsymbol{\beta}}_c - \widehat{\boldsymbol{\beta}}_e)$$

30. As discussed in Baum, Schaffer, and Stillman (2003, 11), GMM may have poor small-sample properties. If the zero-conditional-mean assumption cannot be refuted, we should use linear regression rather than IV or GMM, especially in small samples.

where

$$\mathbf{D} = \mathrm{Var}[\widehat{\boldsymbol{\beta}}_c] - \mathrm{Var}[\widehat{\boldsymbol{\beta}}_e]$$

$\mathrm{Var}[\widehat{\boldsymbol{\beta}}]$ denotes a consistent estimate of the asymptotic variance of $\boldsymbol{\beta}$, 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 $\widehat{\boldsymbol{\beta}}_e$ and IV as the inefficient but consistent estimator $\widehat{\boldsymbol{\beta}}_c$. The test statistic is distributed as χ^2 with k_1 degrees of freedom: 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 using 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.

There are many ways to conduct a DWH endogeneity test in Stata for the standard IV case with conditional homoskedasticity. Three equivalent ways of obtaining the Durbin component of the DWH statistic in Stata are

1. Fit the less efficient but consistent model using IV, followed by the command `estimates store iv` (where `iv` is a name of your choice that is attached to this set of estimates; see the discussion of stored estimates in section 4.4). Then fit the fully efficient model with `regress` (or with `ivreg` if only a subset of regressors is being tested for endogeneity), followed by
   `hausman iv ., constant sigmamore`.[31]
2. Fit the fully efficient model using `ivreg2` and specify the regressors to be tested in the `orthog()` option.
3. Fit the less efficient but consistent model using `ivreg` and use `ivendog` to conduct an endogeneity test. The `ivendog` command takes 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 last two methods are more convenient than the first because the test can be done in one step. Furthermore, the `hausman` command will often generate a negative χ^2 statistic, rendering the test infeasible. Stata's documentation describes this result as a small-sample problem in which the variance of the difference of the coefficient vectors is not necessarily positive definite in finite samples.[32] The different commands implement distinct versions of the tests, which although asymptotically equivalent can lead to different inference from finite samples.

31. You should disregard the note produced by `hausman` regarding the rank of the differenced matrix. As the documentation of the `sigmamore` option indicates, this is the proper setting for a test of exogeneity comparing linear regression and IV estimates.
32. The description of `hausman` suggests that a generalized Hausman test can be performed by `suest`. However, this command does not support the `ivreg` estimator.

I first illustrate using the `hausman` command for the wage equation:

```
. quietly ivreg2 lw s expr tenure rns smsa _I* (iq=med kww), small
. estimates store iv
. quietly regress lw s expr tenure rns smsa _I* iq
. hausman iv ., constant sigmamore
Note: the rank of the differenced variance matrix (1) does not equal the number
      of coefficients being tested (13); be sure this is what you expect, or
      there may be problems computing the test.  Examine the output of your
      estimators for anything unexpected and possibly consider scaling your
      variables so that the coefficients are on a similar scale.
```

	Coefficients (b) iv	(B) .	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
iq	.0243202	.0027121	.021608	.0046882
s	.0004625	.0619548	-.0614923	.0133417
expr	.039129	.0308395	.0082896	.0017985
tenure	.0327048	.0421631	-.0094582	.0020521
rns	-.0341617	-.0962935	.0621318	.0134804
smsa	.1140326	.1328993	-.0188667	.0040934
_Iyear_67	-.0679321	-.0542095	-.0137226	.0029773
_Iyear_68	.0900522	.0805808	.0094714	.002055
_Iyear_69	.1794505	.2075915	-.028141	.0061056
_Iyear_70	.1395755	.2282237	-.0886482	.0192335
_Iyear_71	.1735613	.2226915	-.0491302	.0106595
_Iyear_73	.2971599	.3228747	-.0257148	.0055792
_cons	2.837153	4.235357	-1.398204	.3033612

```
                b = consistent under Ho and Ha; obtained from ivreg2
     B = inconsistent under Ha, efficient under Ho; obtained from regress

  Test:  Ho:  difference in coefficients not systematic

              chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =       21.24
            Prob>chi2 =      0.0000
            (V_b-V_B is not positive definite)
```

The comparison here is restricted to the point estimate and estimated standard error of the endogenous regressor, `iq`; the `hausman` test statistic rejects exogeneity of this variable. The command also warns of difficulties computing a positive-definite covariance matrix. The large χ^2 value indicates that estimation of the equation with `regress` yields inconsistent results.

I now illustrate the second method, using `ivreg2` and the `orthog()` option. You should notice the peculiar syntax of the parenthesized list in which no variable is identified as endogenous. This argument (and the equals sign) is still required to signal to Stata that `med kww` are to be considered as instruments in the unrestricted equation in which `iq` is considered endogenous. This treatment causes `ivreg2` to perform the reported estimation using linear regression and consider the alternative model to be IV.[33]

33. I use the `small` option to ensure that the χ^2 statistic takes on the same value in the second and third methods.

```
. ivreg2 lw s expr tenure rns smsa _I* iq (=med kww), orthog(iq) small
Ordinary Least Squares (OLS) regression
                                                  Number of obs =      758
                                                  F( 12,   745) =    46.86
                                                  Prob > F      =   0.0000
Total (centered) SS     =  139.2861498            Centered R2   =   0.4301
Total (uncentered) SS   =  24652.24662            Uncentered R2 =   0.9968
Residual SS             =  79.37338879            Root MSE      =    .3264
------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |    .0619548   .0072786     8.51   0.000     .0476658    .0762438
        expr |    .0308395   .0065101     4.74   0.000     .0180592    .0436198
      tenure |    .0421631   .0074812     5.64   0.000     .0274763    .0568498
         rns |   -.0962935   .0275467    -3.50   0.001    -.1503719   -.0422151
        smsa |    .1328993   .0265758     5.00   0.000     .0807268    .1850717
    _Iyear_67 |   -.0542095   .0478522    -1.13   0.258    -.1481506    .0397317
    _Iyear_68 |    .0805808   .0448951     1.79   0.073    -.0075551    .1687168
    _Iyear_69 |    .2075915   .0438605     4.73   0.000     .1214867    .2936963
    _Iyear_70 |    .2282237   .0487994     4.68   0.000      .132423    .3240245
    _Iyear_71 |    .2226915   .0430952     5.17   0.000     .1380889     .307294
    _Iyear_73 |    .3228747   .0406574     7.94   0.000     .2430579    .4026915
           iq |    .0027121   .0010314     2.63   0.009     .0006873    .0047369
        _cons |    4.235357   .1133489    37.37   0.000     4.012836    4.457878
------------------------------------------------------------------------------
Sargan statistic (Lagrange multiplier test of excluded instruments):    22.659
                                                  Chi-sq(2) P-val =     0.0000
-orthog- option:
Sargan statistic (eqn. excluding suspect orthogonality conditions):      1.045
                                                  Chi-sq(1) P-val =     0.3067
C statistic (exogeneity/orthogonality of suspect instruments):          21.614
                                                  Chi-sq(1) P-val =     0.0000
Instruments tested:  iq
------------------------------------------------------------------------------
Included instruments: s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69
                      _Iyear_70 _Iyear_71 _Iyear_73 iq
Excluded instruments: med kww
------------------------------------------------------------------------------
```

The second method's C test statistic from `ivendog` agrees qualitatively with that from `hausman`. I now illustrate the third method's use of `ivendog`:

```
. quietly ivreg lw s expr tenure rns smsa _I* (iq=med kww)
. ivendog
Tests of endogeneity of: iq
H0: Regressor is exogenous
    Wu-Hausman F test:                 21.83742  F(1,744)     P-value = 0.00000
    Durbin-Wu-Hausman chi-sq test:     21.61394  Chi-sq(1)    P-value = 0.00000
```

The test statistic is identical to that provided by the C statistic above. All forms of the test agree that estimation of this equation with linear regression yields inconsistent results. The regressor `iq` must be considered endogenous in the fitted model.

Exercises

1. Following the discussion in section 8.3, use the Griliches data in section 8.6 to estimate two-stage least squares "by hand". Compare the residuals and s^2 with those computed by `ivreg` on the same equation.
2. When we presented robust linear regression estimates, the estimated coefficients and summary statistics were unchanged; only the VCE was affected. Compare the estimates displayed in section 8.6 with those of section 8.7.3. Why do the coefficient estimates and summary statistics such as `R-squared` and `Root MSE` differ?
3. Using the Griliches data, estimate the equation

```
. ivreg2 lw s expr rns smsa (iq=med kww age mrt) if year==67, gmm
```

   What comments can you make about these estimates? Reestimate the equation, adding the `cluster(age)` option. What is the rationale for clustering by age? Evaluate this form of the equation versus that estimated without clustering. What are its problems?
4. Following the discussion in section 8.7.5, refit the Phillips curve model (a) without the `gmm` option and (b) without the `gmm` and `robust` options. How do these estimates—corresponding to 2SLS–HAC and 2SLS–AC—compare with the GMM–HAC estimates displayed in the text?
5. Refit the Phillips curve model using lags 1, 2, and 3 of `unem` as instruments for the unemployment rate. What do you find?
6. Does the Phillips curve require an IV estimator, or can it be consistently estimated with linear regression? Refit the model of section 8.7.5, using the `orthog()` option of `ivreg2` to decide whether linear regression is satisfactory using the DWH framework.
7. Does the Phillips curve exhibit heteroskedasticity in the time dimension? Refit the model of section 8.7.5 without the `robust` option, and use the options of `ivhettest` to test this hypothesis.

8.A Appendix: Omitted-variables bias

The OLS estimator cannot produce consistent estimates if the zero-conditional-mean assumption (4.2) is violated. I illustrate an alternative solution by considering the omitted-variables problem discussed above in section 5.2: an unobserved but relevant omitted explanatory factor. Consider the relationship among high schools' average Scholastic Aptitude Test (SAT) scores (`sat`),[34] expenditure per pupil (`spend`), and the poverty rate in each district (`poverty`):

$$\texttt{sat} = \beta_1 + \beta_2 \texttt{expend} + \beta_3 \texttt{poverty} + u_i \tag{8.12}$$

We cannot estimate this equation because we do not have access to poverty rates at the school-district level. However, that factor is thought to play an important role in educational attainment, proxying for the quality of the student's home environment. If we had a *proxy variable* available, we could substitute it for `poverty`, for example, the median income in the school district. Whether this strategy would succeed depends on how highly the proxy variable is correlated with the unobserved `poverty`. If no proxy is available, we might estimate the equation, ignoring `poverty`:

$$\log(\texttt{sat}_i) = \beta_1 + \beta_2 \texttt{expend}_i + v_i$$

The disturbance process v_i in this equation is composed of $(\beta_3 \texttt{poverty}_i + u_i)$. If `expend` and `poverty` are correlated—as they are likely to be—regression will yield biased and inconsistent estimates of β_1 and β_2 because the zero-conditional-mean assumption is violated.

To derive consistent estimates of this equation, we must find an IV, as discussed in section 8.2. Many potential variables could be uncorrelated with the unobservable factors influencing SAT performance (including `poverty`) and highly correlated with `expend`.[35] What might be an appropriate instrument for `expend`? Perhaps we could measure each school district's student–teacher ratio (`stratio`). This measure is likely to be (negatively) correlated with district expenditure. If states' education policy mandates that student–teacher ratios fall within certain bounds, `stratio` should not be correlated with district poverty rates.

8.B Appendix: Measurement error

I introduced the concept of measurement error in section 5.3 and now discuss its consequences. Measurement error could appear in the response variable. Say that the true relationship explains y^*, but we observe $y = y^* + \epsilon$, where ϵ is a mean-zero-error process. Then ϵ becomes a component of the regression error term, worsening the fit of the

34. The SAT is the most common standardized test taken by U.S. high school students for college entrance.

35. We are not searching for a proxy variable for `poverty`. If we had a good proxy for $\texttt{poverty}_i$, it would not make a satisfactory instrumental variable. Correlation with $\texttt{poverty}_i$ implies correlation with the composite error process v_i.

estimated equation. We assume that ϵ is not systematic in that it is not correlated with the independent variables x. Then measurement error does no real harm—it merely weakens the model without introducing bias in either point or interval estimates.[36]

On the other hand, measurement error in a regressor is a far more serious problem. Say that the true model is

$$y = \beta_1 + \beta_2 x_2^* + u$$

but that x_2^* is not observed: we observe $x_2 = x_2^* + \epsilon_2$. We assume that $E[\epsilon_2] = 0$. What should we assume about the relationship between ϵ_2 and x_2^*? First, let us assume that ϵ_2 is not correlated with the observed measure x_2: larger values of x_2 do not give rise to systematically larger or smaller errors of measurement, which we can write as $\text{Cov}[\epsilon_2, x_2] = 0$. But if so, $\text{Cov}[\epsilon_2, x_2^*] \neq 0$: that is, the error of measurement must be correlated with the true explanatory variable x_2^*. We can then write the estimated equation in which x_2^* is replaced with the observable x_2 as

$$y = \beta_1 + \beta_2 x_2 + (u - \beta_2 \epsilon_2) \tag{8.13}$$

Since both u and ϵ_2 have zero mean and, by assumption, are uncorrelated with x_2, the presence of measurement error merely inflates the error term. $\text{Var}[u - \beta_2 \epsilon_2] = \sigma_u^2 + \beta_2^2 \sigma_{\epsilon_2}^2$ given a zero correlation of u, ϵ. Measurement error in x_2^* does not damage the regression of y on x_2—it merely inflates the error variance, as does measurement error in the response variable.

However, this is not the case that is usually considered in applied econometrics as *errors in variables*. It is more reasonable to assume that the measurement error is uncorrelated with the true explanatory variable: $\text{Cov}[\epsilon_2, x_2^*] = 0$. For instance, we might assume that the discrepancy between reported income and actual income is not a function of actual income. If so, $\text{Cov}[\epsilon_2, x_2] = \text{Cov}[\epsilon_2, (x_2^* + \epsilon_2)] \neq 0$ by construction, and the regression of (8.13) will have a nonzero correlation between its explanatory variable x_2 and the composite error term. This result violates the zero-conditional-mean assumption of (4.2). The covariance of $(x_2, u - \beta_2 \epsilon_2) = -\beta_2 \text{Cov}[\epsilon_2, x_2] = -\beta_2 \, \sigma_{\epsilon_2}^2 \neq 0$, causing the OLS regression of y on x_2 to be biased and inconsistent. In this simple case of one explanatory variable measured with error, we can determine the nature of the bias because $\widehat{\beta}_2$ consistently estimates

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_2 &= \beta_2 + \frac{\text{Cov}[x_2, u - \beta_2 \epsilon_2]}{\text{Var}[x_2]} \\ &= \beta_2 \left(\frac{\sigma_{x_2}^2}{\sigma_{x_2}^2 + \sigma_{\epsilon_2}^2} \right) \end{aligned}$$

This expression demonstrates that the OLS point estimate will be *attenuated*—biased toward zero even in large samples—because the bracketed expression of squared quantities must be a fraction. In the absence of measurement error, $\sigma_{\epsilon_2}^2 \to 0$, and the OLS coefficient becomes consistent and unbiased. As $\sigma_{\epsilon_2}^2$ increases relative to the variance

36. If the magnitude of the measurement error in y is correlated with one or more of the regressors in x, the point estimates will be biased.

in the (correctly measured) explanatory variable, the OLS estimate becomes more and more unreliable, shrinking toward zero.

We conclude that in a multiple regression equation in which one of the regressors is subject to measurement error, if the measurement error is uncorrelated with the true (correctly measured) explanatory variable, then the OLS estimates will be biased and inconsistent for all the regressors, not merely for the coefficient of the regressor measured with error. We cannot predict the direction of bias with multiple regressors. Realistically, more than one regressor in an economic model may be subject to measurement error. In a household survey, both reported income and reported wealth may be measured incorrectly. Since measurement error violates the zero-conditional-mean assumption in the same sense as simultaneity bias or omitted-variables bias, we can treat it similarly.

8.B.1 Solving errors-in-variables problems

We can use the IV estimator to deal with the errors-in-variables model discussed in section 8.B. To deal with measurement error in one or more regressors, we must be able to specify an instrument for the mismeasured x variable that satisfies the usual assumptions. The instrument must not be correlated with the disturbance process u but must be highly correlated with the mismeasured x. If we could find a second measurement of x—even one that is prone to measurement error—we could use it as an instrument, since it would presumably be well correlated with x itself. If it is generated by an independent measurement process, it will be uncorrelated with the original measurement error. For instance, we might have data from a household survey that inquired about each family's disposable income, consumption, and saving. The respondents' answers about their saving last year might well be mismeasured since it is much harder to track saving than, say, earned income. We could say the same for their estimates of how much they spent on various categories of consumption. But using income and consumption data, we could derive a second (mismeasured) estimate of saving, which we could use as an instrument to mitigate the problems of measurement error in the direct estimate.