Instrumental variables and panel data methods in economics and finance

Christopher F Baum

Boston College and DIW Berlin

February 2009

Regression with Instrumental Variables

What are instrumental variables (IV) methods? Most widely known as a solution to *endogenous regressors*: explanatory variables correlated with the regression error term, IV methods provide a way to nonetheless obtain consistent parameter estimates.

Although IV estimators address issues of endogeneity, the violation of the zero conditional mean assumption caused by endogenous regressors can also arise for two other common causes: measurement error in regressors (errors-in-variables) and omitted-variable bias. The latter may arise in situations where a variable known to be relevant for the data generating process is not measurable, and no good proxies can be found.

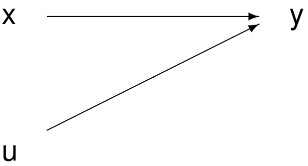
Regression with Instrumental Variables

What are instrumental variables (IV) methods? Most widely known as a solution to *endogenous regressors*: explanatory variables correlated with the regression error term, IV methods provide a way to nonetheless obtain consistent parameter estimates.

Although IV estimators address issues of endogeneity, the violation of the zero conditional mean assumption caused by endogenous regressors can also arise for two other common causes: measurement error in regressors (errors-in-variables) and omitted-variable bias. The latter may arise in situations where a variable known to be relevant for the data generating process is not measurable, and no good proxies can be found.

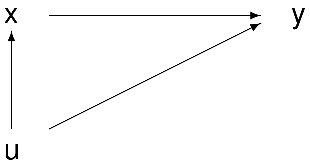
First let us consider a path diagram illustrating the problem addressed by IV methods. We can use ordinary least squares (OLS) regression to consistently estimate a model of the following sort.

Standard regression: $y = xb + u$
no association between x and u; OLS consistent



However, OLS regression breaks down in the following circumstance:

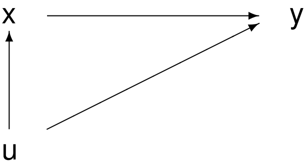
Endogeneity: $y = xb + u$
correlation between x and u; OLS inconsistent



The correlation between x and u (or the failure of the zero conditional mean assumption $E[u|x] = 0$) can be caused by any of several factors.

However, OLS regression breaks down in the following circumstance:

Endogeneity: $y = xb + u$
correlation between x and u; OLS inconsistent



The correlation between x and u (or the failure of the zero conditional mean assumption $E[u|x] = 0$) can be caused by any of several factors.

Endogeneity

We have stated the problem as that of *endogeneity*: the notion that two or more variables are jointly determined in the behavioral model. This arises naturally in the context of a *simultaneous equations model* such as a supply-demand system in economics, in which price and quantity are jointly determined in the market for that good or service.

A shock or disturbance to either supply or demand will affect both the equilibrium price and quantity in the market, so that by construction both variables are correlated with any shock to the system. OLS methods will yield inconsistent estimates of any regression including both price and quantity, however specified.

Endogeneity

We have stated the problem as that of *endogeneity*: the notion that two or more variables are jointly determined in the behavioral model. This arises naturally in the context of a *simultaneous equations model* such as a supply-demand system in economics, in which price and quantity are jointly determined in the market for that good or service.

A shock or disturbance to either supply or demand will affect both the equilibrium price and quantity in the market, so that by construction both variables are correlated with any shock to the system. OLS methods will yield inconsistent estimates of any regression including both price and quantity, however specified.

As a different example, consider a cross-sectional regression of public health outcomes (say, the proportion of the population in various cities suffering from a particular childhood disease) on public health expenditures *per capita* in each of those cities. We would hope to find that spending is effective in reducing incidence of the disease, but we also must consider the *reverse causality* in this relationship, where the level of expenditure is likely to be partially determined by the historical incidence of the disease in each jurisdiction.

In this context, OLS estimates of the relationship will be biased even if additional controls are added to the specification. Although we may have no interest in modeling public health expenditures, we must be able to specify such an equation in order to *identify* the relationship of interest, as we discuss henceforth.

As a different example, consider a cross-sectional regression of public health outcomes (say, the proportion of the population in various cities suffering from a particular childhood disease) on public health expenditures *per capita* in each of those cities. We would hope to find that spending is effective in reducing incidence of the disease, but we also must consider the *reverse causality* in this relationship, where the level of expenditure is likely to be partially determined by the historical incidence of the disease in each jurisdiction.

In this context, OLS estimates of the relationship will be biased even if additional controls are added to the specification. Although we may have no interest in modeling public health expenditures, we must be able to specify such an equation in order to *identify* the relationship of interest, as we discuss henceforth.

The solution provided by IV methods may be viewed as:

Instrumental variables regression: $y = xb + u$
z uncorrelated with u, correlated with x

$$\begin{tikzcd} z \arrow[r] & x \arrow[r] & y \\ & u \arrow[u] \arrow[ur] & \end{tikzcd}$$

The additional variable z is termed an *instrument* for x. In general, we may have many variables in x, and more than one x correlated with u. In that case, we shall need at least that many variables in z.

Choice of instruments

To deal with the problem of *endogeneity* in a supply-demand system, a candidate z will affect (e.g.) the quantity supplied of the good, but not directly impact the demand for the good. An example for an agricultural commodity might be temperature or rainfall: clearly exogenous to the market, but likely to be important in the production process.

For the public health example, we might use *per capita* income in each city as an instrument or z variable. It is likely to influence public health expenditure, as cities with a larger tax base might be expected to spend more on all services, and will not be directly affected by the unobserved factors in the primary relationship.

Choice of instruments

To deal with the problem of *endogeneity* in a supply-demand system, a candidate z will affect (e.g.) the quantity supplied of the good, but not directly impact the demand for the good. An example for an agricultural commodity might be temperature or rainfall: clearly exogenous to the market, but likely to be important in the production process.

For the public health example, we might use *per capita* income in each city as an instrument or z variable. It is likely to influence public health expenditure, as cities with a larger tax base might be expected to spend more on all services, and will not be directly affected by the unobserved factors in the primary relationship.

But why should we not always use IV?

It may be difficult to find variables that can serve as valid instruments. Many variables that have an effect on included endogenous variables also have a direct effect on the dependent variable.

IV estimators are innately *biased*, and their finite-sample properties are often problematic. Thus, most of the justification for the use of IV is asymptotic. Performance in small samples may be poor.

The precision of IV estimates is lower than that of OLS estimates (least squares is just that). In the presence of *weak instruments* (excluded instruments only weakly correlated with included endogenous regressors) the loss of precision will be severe, and IV estimates may be no improvement over OLS. This suggests we need a method to determine whether a particular regressor must be treated as endogenous.

But why should we not always use IV?

It may be difficult to find variables that can serve as valid instruments. Many variables that have an effect on included endogenous variables also have a direct effect on the dependent variable.

IV estimators are innately *biased*, and their finite-sample properties are often problematic. Thus, most of the justification for the use of IV is asymptotic. Performance in small samples may be poor.

The precision of IV estimates is lower than that of OLS estimates (least squares is just that). In the presence of *weak instruments* (excluded instruments only weakly correlated with included endogenous regressors) the loss of precision will be severe, and IV estimates may be no improvement over OLS. This suggests we need a method to determine whether a particular regressor must be treated as endogenous.

But why should we not always use IV?

It may be difficult to find variables that can serve as valid instruments. Many variables that have an effect on included endogenous variables also have a direct effect on the dependent variable.

IV estimators are innately *biased*, and their finite-sample properties are often problematic. Thus, most of the justification for the use of IV is asymptotic. Performance in small samples may be poor.

The precision of IV estimates is lower than that of OLS estimates (least squares is just that). In the presence of *weak instruments* (excluded instruments only weakly correlated with included endogenous regressors) the loss of precision will be severe, and IV estimates may be no improvement over OLS. This suggests we need a method to determine whether a particular regressor must be treated as endogenous.

IV estimation as a GMM problem

Before discussing further the motivation for various weak instrument diagnostics, we define the setting for IV estimation as a Generalized Method of Moments (GMM) optimization problem. Economists consider GMM to be the invention of Lars Hansen in his 1982 *Econometrica* paper, but as Alistair Hall points out in his 2005 book, the method has its antecedents in Karl Pearson's *Method of Moments* [MM] (1895) and Neyman and Egon Pearson's *minimum Chi-squared estimator* [MCE] (1928). Their MCE approach overcomes the difficulty with MM estimators when there are more moment conditions than parameters to be estimated. This was recognized by Ferguson (*Ann. Math. Stat.* 1958) for the case of *i.i.d.* errors, but his work had no impact on the econometric literature.

The model:

$$y = X\beta + u, \quad u \sim (0, \Omega)$$

X $(N \times k)$. Define a matrix Z $(N \times \ell)$ where $\ell \geq k$. This is the Generalized Method of Moments IV (IV-GMM) estimator. The ℓ instruments give rise to a set of ℓ moments:

$$g_i(\beta) = Z_i' u_i = Z_i'(y_i - x_i\beta), \; i = 1, N$$

where each g_i is an ℓ-vector. The method of moments approach considers each of the ℓ moment equations as a sample moment, which we may estimate by averaging over N:

$$\bar{g}(\beta) = \frac{1}{N}\sum_{i=1}^{N} z_i(y_i - x_i\beta) = \frac{1}{N} Z' u$$

The GMM approach chooses $\hat{\beta}$ to solve $\bar{g}(\hat{\beta}_{GMM}) = 0$.

If $\ell = k$, the equation to be estimated is said to be *exactly identified* by the *order condition* for identification: that is, there are as many excluded instruments as included right-hand endogenous variables. The method of moments problem is then k equations in k unknowns, and a unique solution exists, equivalent to the standard IV estimator:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$$

In the case of *overidentification* ($\ell > k$) we may define a set of k instruments:

$$\hat{X} = Z'(Z'Z)^{-1}Z'X = P_Z X$$

which gives rise to the *two-stage least squares* (2SLS) estimator

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

which despite its name is computed by this single matrix equation.

If $\ell = k$, the equation to be estimated is said to be *exactly identified* by the *order condition* for identification: that is, there are as many excluded instruments as included right-hand endogenous variables. The method of moments problem is then k equations in k unknowns, and a unique solution exists, equivalent to the standard IV estimator:

$$\hat{\beta}_{IV} = (Z'X)^{-1} Z'y$$

In the case of *overidentification* ($\ell > k$) we may define a set of k instruments:

$$\hat{X} = Z'(Z'Z)^{-1} Z'X = P_Z X$$

which gives rise to the *two-stage least squares* (2SLS) estimator

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

which despite its name is computed by this single matrix equation.

The IV-GMM approach

In the 2SLS method with overidentification, the ℓ available instruments are “boiled down" to the k needed by defining the P_Z matrix. In the IV-GMM approach, that reduction is not necessary. All ℓ instruments are used in the estimator. Furthermore, a *weighting matrix* is employed so that we may choose $\hat{\beta}_{GMM}$ so that the elements of $\bar{g}(\hat{\beta}_{GMM})$ are as close to zero as possible. With $\ell > k$, not all ℓ moment conditions can be exactly satisfied, so a criterion function that weights them appropriately is used to improve the efficiency of the estimator.

The GMM estimator minimizes the criterion

$$J(\hat{\beta}_{GMM}) = N\, \bar{g}(\hat{\beta}_{GMM})'\, W\, \bar{g}(\hat{\beta}_{GMM})$$

where W is a $\ell \times \ell$ symmetric weighting matrix.

The IV-GMM approach

In the 2SLS method with overidentification, the ℓ available instruments are "boiled down" to the k needed by defining the P_Z matrix. In the IV-GMM approach, that reduction is not necessary. All ℓ instruments are used in the estimator. Furthermore, a *weighting matrix* is employed so that we may choose $\hat{\beta}_{GMM}$ so that the elements of $\bar{g}(\hat{\beta}_{GMM})$ are as close to zero as possible. With $\ell > k$, not all ℓ moment conditions can be exactly satisfied, so a criterion function that weights them appropriately is used to improve the efficiency of the estimator.

The GMM estimator minimizes the criterion

$$J(\hat{\beta}_{GMM}) = N\, \bar{g}(\hat{\beta}_{GMM})' W\, \bar{g}(\hat{\beta}_{GMM})$$

where W is a $\ell \times \ell$ symmetric weighting matrix.

Solving the set of FOCs, we derive the IV-GMM estimator of an overidentified equation:

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

which will be identical for all W matrices which differ by a factor of proportionality. The *optimal* weighting matrix, as shown by Hansen (1982), chooses $W = S^{-1}$ where S is the covariance matrix of the moment conditions to produce the most *efficient* estimator:

$$S = E[Z'uu'Z] = lim_{N\to\infty}\, N^{-1}[Z'\Omega Z]$$

With a consistent estimator of S derived from 2SLS residuals, we define the feasible IV-GMM estimator as

$$\hat{\beta}_{FEGMM} = (X'Z\,\hat{S}^{-1}Z'X)^{-1}X'Z\,\hat{S}^{-1}Z'y$$

where *FEGMM* refers to the *feasible efficient* GMM estimator.

Solving the set of FOCs, we derive the IV-GMM estimator of an overidentified equation:

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

which will be identical for all W matrices which differ by a factor of proportionality. The *optimal* weighting matrix, as shown by Hansen (1982), chooses $W = S^{-1}$ where S is the covariance matrix of the moment conditions to produce the most *efficient* estimator:

$$S = E[Z'uu'Z] = lim_{N\to\infty}\ N^{-1}[Z'\Omega Z]$$

With a consistent estimator of S derived from 2SLS residuals, we define the feasible IV-GMM estimator as

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

where *FEGMM* refers to the *feasible efficient* GMM estimator.

The derivation makes no mention of the form of Ω, the variance-covariance matrix (*vce*) of the error process u. If the errors satisfy all classical assumptions are *i.i.d.*, $S = \sigma_u^2 I_N$ and the optimal weighting matrix is proportional to the identity matrix. The IV-GMM estimator is merely the standard IV (or 2SLS) estimator.

If there is heteroskedasticity of unknown form, we usually compute *robust* standard errors in any Stata estimation command to derive a consistent estimate of the *vce*. In this context,

$$\hat{S} = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 Z_i' Z_i$$

where $\hat{u}$ is the vector of residuals from any consistent estimator of β (e.g., the 2SLS residuals). For an overidentified equation, the IV-GMM estimates computed from this estimate of S will be more efficient than 2SLS estimates.

The derivation makes no mention of the form of Ω, the variance-covariance matrix (*vce*) of the error process u. If the errors satisfy all classical assumptions are *i.i.d.*, $S = \sigma_u^2 I_N$ and the optimal weighting matrix is proportional to the identity matrix. The IV-GMM estimator is merely the standard IV (or 2SLS) estimator.

If there is heteroskedasticity of unknown form, we usually compute *robust* standard errors in any Stata estimation command to derive a consistent estimate of the *vce*. In this context,

$$\hat{S} = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 Z_i' Z_i$$

where $\hat{u}$ is the vector of residuals from any consistent estimator of β (e.g., the 2SLS residuals). For an overidentified equation, the IV-GMM estimates computed from this estimate of S will be more efficient than 2SLS estimates.

We must distinguish the concept of IV/2SLS estimation with robust standard errors from the concept of estimating the same equation with IV-GMM, allowing for arbitrary heteroskedasticity. Compare an overidentified regression model estimated (a) with IV and classical standard errors and (b) with robust standard errors. Model (b) will produce the same point estimates, but different standard errors in the presence of heteroskedastic errors.

However, if we reestimate that overidentified model using the GMM two-step estimator, we will get different point estimates because we are solving a different optimization problem: one in the ℓ-space of the instruments (and moment conditions) rather than the k-space of the regressors, and $\ell > k$. We will also get different standard errors, and in general smaller standard errors as the IV-GMM estimator is more efficient. This does not imply, however, that summary measures of fit will improve.

We must distinguish the concept of IV/2SLS estimation with robust standard errors from the concept of estimating the same equation with IV-GMM, allowing for arbitrary heteroskedasticity. Compare an overidentified regression model estimated (a) with IV and classical standard errors and (b) with robust standard errors. Model (b) will produce the same point estimates, but different standard errors in the presence of heteroskedastic errors.

However, if we reestimate that overidentified model using the GMM two-step estimator, we will get different point estimates because we are solving a different optimization problem: one in the ℓ-space of the instruments (and moment conditions) rather than the k-space of the regressors, and $\ell > k$. We will also get different standard errors, and in general smaller standard errors as the IV-GMM estimator is more efficient. This does not imply, however, that summary measures of fit will improve.

If errors are considered to exhibit arbitrary intra-cluster correlation in a dataset with M clusters, we may derive a *cluster-robust* IV-GMM estimator using

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

where

$$\hat{u}_j = (y_j - x_j \hat{\beta}) X' Z (Z'Z)^{-1} z_j$$

The IV-GMM estimates employing this estimate of S will be both robust to arbitrary heteroskedasticity and intra-cluster correlation, equivalent to estimates generated by Stata's `cluster(`*varname*`)` option. For an overidentified equation, IV-GMM cluster-robust estimates will be more efficient than 2SLS estimates.

The concept of the cluster-robust covariance matrix has been extended by Cameron, Gelbach and Miller (NBER TWP327, 2006) and Thompson (SSRN, 2009) to define *two-way clustering*. This allows for arbitrary within-cluster correlation in two cluster dimensions. For instance, in a panel dataset, we may want to allow observations belonging to each individual unit to be arbitrarily correlated, and we may want to allow observations coming from a particular time period to be arbitrarily correlated.

Heretofore, a common tactic involved (one-way) clustering by unit (e.g., firm), and the introduction of a set of time (e.g., year) dummies. Although the latter account for any macro-level heterogeneity, they do not relax the assumption of independence across units at each point in time, which may be highly unlikely. Thus, it may be beneficial to utilize two-way clustering in estimating the covariance matrix.

The concept of the cluster-robust covariance matrix has been extended by Cameron, Gelbach and Miller (NBER TWP327, 2006) and Thompson (SSRN, 2009) to define *two-way clustering*. This allows for arbitrary within-cluster correlation in two cluster dimensions. For instance, in a panel dataset, we may want to allow observations belonging to each individual unit to be arbitrarily correlated, and we may want to allow observations coming from a particular time period to be arbitrarily correlated.

Heretofore, a common tactic involved (one-way) clustering by unit (e.g., firm), and the introduction of a set of time (e.g., year) dummies. Although the latter account for any macro-level heterogeneity, they do not relax the assumption of independence across units at each point in time, which may be highly unlikely. Thus, it may be beneficial to utilize two-way clustering in estimating the covariance matrix.

The IV-GMM approach may also be used to generate *HAC standard errors*: those robust to arbitrary heteroskedasticity and autocorrelation. Although the best-known *HAC* approach in econometrics is that of Newey and West, using the Bartlett kernel (per Stata's `newey`), that is only one choice of a *HAC* estimator that may be applied to an IV-GMM problem. Baum–Schaffer–Stillman's `ivreg2` (from the SSC Archive) and Stata 10's `ivregress` provide several choices for kernels. For some kernels, the kernel *bandwidth* (roughly, number of lags employed) may be chosen automatically in either command.

The *HAC* options may be combined with one- or two-way clustering.

The IV-GMM approach may also be used to generate *HAC standard errors*: those robust to arbitrary heteroskedasticity and autocorrelation. Although the best-known *HAC* approach in econometrics is that of Newey and West, using the Bartlett kernel (per Stata's `newey`), that is only one choice of a *HAC* estimator that may be applied to an IV-GMM problem. Baum–Schaffer–Stillman's `ivreg2` (from the SSC Archive) and Stata 10's `ivregress` provide several choices for kernels. For some kernels, the kernel *bandwidth* (roughly, number of lags employed) may be chosen automatically in either command.

The *HAC* options may be combined with one- or two-way clustering.

The ivreg2 command

The estimators we have discussed are available from Baum, Schaffer and Stillman's *ivreg2* package, revised February 2010 (`ssc describe ivreg2`). The `ivreg2` command has the same basic syntax as Stata's older `ivreg` command:

```
ivreg2 depvar [varlist1] (varlist2=instlist) ///
        [if] [in]  [, options]
```

The ℓ variables in `varlist1` and `instlist` comprise Z, the matrix of instruments. The k variables in `varlist1` and `varlist2` comprise X. Both matrices by default include a units vector.

The ivreg2 command

The estimators we have discussed are available from Baum, Schaffer and Stillman's *ivreg2* package, revised February 2010 (`ssc describe ivreg2`). The `ivreg2` command has the same basic syntax as Stata's older `ivreg` command:

```
ivreg2 depvar [varlist1] (varlist2=instlist) ///
          [if] [in]  [, options]
```

The ℓ variables in `varlist1` and `instlist` comprise Z, the matrix of instruments. The k variables in `varlist1` and `varlist2` comprise X. Both matrices by default include a units vector.

By default `ivreg2` estimates the IV estimator, or 2SLS estimator if $\ell > k$. If the `gmm2s` option is specified in conjunction with `robust`, `cluster()` or `bw()`, it estimates the IV-GMM estimator.

With the `robust` option, the *vce* is heteroskedasticity-robust.

With the `cluster(`*varname*`)` or `cluster(`*varname1 varname2*`)` option, the *vce* is cluster-robust.

With the `robust` and `bw( )` options, the *vce* is *HAC* with the default Bartlett ("Newey–West") kernel. Other `kernel( )` choices lead to alternative *HAC* estimators. In `ivreg2`, both `robust` and `bw( )` options must be specified to produce *HAC*. Estimates produced with `bw( )` alone are robust to arbitrary autocorrelation but assume homoskedasticity.

By default `ivreg2` estimates the IV estimator, or 2SLS estimator if $\ell > k$. If the `gmm2s` option is specified in conjunction with `robust`, `cluster()` or `bw()`, it estimates the IV-GMM estimator.

With the `robust` option, the *vce* is heteroskedasticity-robust.

With the `cluster(`*varname*`)` or `cluster(`*varname1 varname2*`)` option, the *vce* is cluster-robust.

With the `robust` and `bw( )` options, the *vce* is *HAC* with the default Bartlett ("Newey–West") kernel. Other `kernel( )` choices lead to alternative *HAC* estimators. In `ivreg2`, both `robust` and `bw( )` options must be specified to produce *HAC*. Estimates produced with `bw( )` alone are robust to arbitrary autocorrelation but assume homoskedasticity.

By default `ivreg2` estimates the IV estimator, or 2SLS estimator if $\ell > k$. If the `gmm2s` option is specified in conjunction with `robust`, `cluster()` or `bw()`, it estimates the IV-GMM estimator.

With the `robust` option, the *vce* is heteroskedasticity-robust.

With the `cluster(`*varname*`)` or `cluster(`*varname1 varname2*`)` option, the *vce* is cluster-robust.

With the `robust` and `bw( )` options, the *vce* is *HAC* with the default Bartlett ("Newey–West") kernel. Other `kernel( )` choices lead to alternative *HAC* estimators. In `ivreg2`, both `robust` and `bw( )` options must be specified to produce *HAC*. Estimates produced with `bw( )` alone are robust to arbitrary autocorrelation but assume homoskedasticity.

Example of IV and IV-GMM estimation

We illustrate with a wage equation estimated from the Griliches dataset (`griliches76`) of very young men's wages. Their log(wage) is explained by completed years of schooling, experience, job tenure and IQ score.

The IQ variable is considered endogenous, and instrumented with three factors: their mother's level of education (`med`), their score on a standardized test (`kww`) and their `age`. The estimation in `ivreg2` is performed with

```
ivreg2 lw s expr tenure (iq = med kww age)
```

where the parenthesized expression defines the *included endogenous* and *excluded exogenous* variables. You could also use official Stata's `ivregress 2sls`.

```
. esttab, label stat(rmse) mtitles(IV IVrob IVGMMrob) nonum
```

	IV	IVrob	IVGMMrob
iq score	-0.00509	-0.00509	-0.00676
	(-1.06)	(-1.01)	(-1.34)
completed years of~g	0.122***	0.122***	0.128***
	(7.68)	(7.51)	(7.88)
experience, years	0.0357***	0.0357***	0.0368***
	(5.15)	(5.10)	(5.26)
tenure, years	0.0405***	0.0405***	0.0443***
	(4.78)	(4.51)	(4.96)
Constant	4.441***	4.441***	4.523***
	(14.22)	(13.21)	(13.46)
rmse	0.366	0.366	0.372

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

These three columns compare standard IV (2SLS) estimates, IV with robust standard errors, and IV-GMM with robust standard errors, respectively. Notice that the coefficients' point estimates change when IV-GMM is employed, and that their *t*-statistics are larger than those of robust IV. Note also that the IQ score is not significant in any of these models.

Tests of overidentifying restrictions

If and only if an equation is *overidentified*, we may test whether the excluded instruments are appropriately independent of the error process. That test should always be performed when it is possible to do so, as it allows us to evaluate the validity of the instruments.

A test of *overidentifying restrictions* regresses the residuals from an IV or 2SLS regression on all instruments in Z. Under the null hypothesis that all instruments are uncorrelated with u, the test has a large-sample $\chi^2(r)$ distribution where r is the number of overidentifying restrictions.

Tests of overidentifying restrictions

If and only if an equation is *overidentified*, we may test whether the excluded instruments are appropriately independent of the error process. That test should always be performed when it is possible to do so, as it allows us to evaluate the validity of the instruments.

A test of *overidentifying restrictions* regresses the residuals from an IV or 2SLS regression on all instruments in Z. Under the null hypothesis that all instruments are uncorrelated with u, the test has a large-sample $\chi^2(r)$ distribution where r is the number of overidentifying restrictions.

Under the assumption of *i.i.d.* errors, this is known as a *Sargan test*, and is routinely produced by `ivreg2` for IV and 2SLS estimates. It can also be calculated after `ivreg` estimation with the `overid` command, which is part of the `ivreg2` suite. After `ivregress`, the command `estat overid` provides the test.

If we have used IV-GMM estimation in `ivreg2`, the test of overidentifying restrictions becomes *J*: the GMM criterion function. Although *J* will be identically zero for any exactly-identified equation, it will be positive for an overidentified equation. If it is “too large”, doubt is cast on the satisfaction of the moment conditions underlying GMM.

The test in this context is known as the *Hansen test* or *J test*, and is calculated by `ivreg2` when the `gmm2s` option is employed.

The Sargan–Hansen test of overidentifying restrictions should be performed routinely in any overidentified model estimated with instrumental variables techniques. Instrumental variables techniques are powerful, but if a strong rejection of the null hypothesis of the Sargan–Hansen test is encountered, you should strongly doubt the validity of the estimates.

If we have used IV-GMM estimation in `ivreg2`, the test of overidentifying restrictions becomes J: the GMM criterion function. Although J will be identically zero for any exactly-identified equation, it will be positive for an overidentified equation. If it is “too large”, doubt is cast on the satisfaction of the moment conditions underlying GMM.

The test in this context is known as the *Hansen test* or *J test*, and is calculated by `ivreg2` when the `gmm2s` option is employed.

The Sargan–Hansen test of overidentifying restrictions should be performed routinely in any overidentified model estimated with instrumental variables techniques. Instrumental variables techniques are powerful, but if a strong rejection of the null hypothesis of the Sargan–Hansen test is encountered, you should strongly doubt the validity of the estimates.

For instance, let's rerun the last IV-GMM model we estimated and focus on the test of overidentifying restrictions provided by the Hansen J statistic. The model is overidentified by two degrees of freedom, as there is one endogenous regressor and three excluded instruments. We see that the J statistic strongly rejects its null, casting doubts on the quality of these estimates.

Let's reestimate the model excluding `age` from the instrument list and see what happens. We will see that the sign and significance of the key endogenous regressor changes as we respecify the instrument list, and the p-value of the J statistic becomes large when `age` is excluded.

For instance, let's rerun the last IV-GMM model we estimated and focus on the test of overidentifying restrictions provided by the Hansen J statistic. The model is overidentified by two degrees of freedom, as there is one endogenous regressor and three excluded instruments. We see that the J statistic strongly rejects its null, casting doubts on the quality of these estimates.

Let's reestimate the model excluding `age` from the instrument list and see what happens. We will see that the sign and significance of the key endogenous regressor changes as we respecify the instrument list, and the p-value of the J statistic becomes large when `age` is excluded.

Example: Test of overidentifying restrictions

```
. esttab, label stat(j jdf jp) mtitles(age no_age) nonum
```

	age	no_age
iq score	-0.00676	0.0181**
	(-1.34)	(2.97)
completed years of~g	0.128***	0.0514**
	(7.88)	(2.63)
experience, years	0.0368***	0.0440***
	(5.26)	(5.58)
tenure, years	0.0443***	0.0303***
	(4.96)	(3.48)
Constant	4.523***	2.989***
	(13.46)	(7.58)
j	49.84	0.282
jdf	2	1
jp	1.50e-11	0.595

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

```
. sjlog close
```

We may be quite confident of some instruments' independence from *u* but concerned about others. In that case a *GMM distance* or *C* test may be used. The `orthog( )` option of `ivreg2` tests whether a *subset* of the model's overidentifying restrictions appear to be satisfied.

This is carried out by calculating two Sargan–Hansen statistics: one for the full model and a second for the model in which the listed variables are (a) considered endogenous, if included regressors, or (b) dropped, if excluded regressors. In case (a), the model must still satisfy the order condition for identification. The difference of the two Sargan–Hansen statistics, often termed the *GMM distance* or *C statistic*, will be distributed χ^2 under the null hypothesis that the specified orthogonality conditions are satisfied, with d.f. equal to the number of those conditions.

We may be quite confident of some instruments' independence from u but concerned about others. In that case a *GMM distance* or *C* test may be used. The `orthog( )` option of `ivreg2` tests whether a *subset* of the model's overidentifying restrictions appear to be satisfied.

This is carried out by calculating two Sargan–Hansen statistics: one for the full model and a second for the model in which the listed variables are (a) considered endogenous, if included regressors, or (b) dropped, if excluded regressors. In case (a), the model must still satisfy the order condition for identification. The difference of the two Sargan–Hansen statistics, often termed the *GMM distance* or *C statistic*, will be distributed χ^2 under the null hypothesis that the specified orthogonality conditions are satisfied, with d.f. equal to the number of those conditions.

A variant on this strategy is implemented by the `endog( )` option of `ivreg2`, in which one or more variables considered endogenous can be tested for exogeneity. The *C* test in this case will consider whether the null hypothesis of their exogeneity is supported by the data.

If all endogenous regressors are included in the `endog( )` option, the test is essentially a test of whether IV methods are required to estimate the equation. If OLS estimates of the equation are consistent, they should be preferred. In this context, the test is equivalent to a *Hausman test* comparing IV and OLS estimates, as implemented by Stata's `hausman` command with the `sigmaless` option. Using `ivreg2`, you need not estimate and store both models to generate the test's verdict.

A variant on this strategy is implemented by the `endog( )` option of `ivreg2`, in which one or more variables considered endogenous can be tested for exogeneity. The *C* test in this case will consider whether the null hypothesis of their exogeneity is supported by the data.

If all endogenous regressors are included in the `endog( )` option, the test is essentially a test of whether IV methods are required to estimate the equation. If OLS estimates of the equation are consistent, they should be preferred. In this context, the test is equivalent to a *Hausman test* comparing IV and OLS estimates, as implemented by Stata's `hausman` command with the `sigmaless` option. Using `ivreg2`, you need not estimate and store both models to generate the test's verdict.

For instance, with the model above, we might question whether IV techniques are needed. We can conduct the C test via:

```
ivreg2 lw s expr tenure (iq=med kww), gmm2s robust endog(iq)
```

where the `endog(iq)` option tests the null hypothesis that `iq` is properly exogenous in this model. The test statistic has a p-value of 0.0108, suggesting that the data overwhelmingly reject the use of OLS in favor of IV. At the same time, the J statistic (with a p-value of 0.60) indicates that the overidentifying restrictions are not rejected.

The weak instruments problem

Instrumental variables methods rely on two assumptions: the excluded instruments are distributed independently of the error process, and they are sufficiently correlated with the included endogenous regressors. Tests of overidentifying restrictions address the *first* assumption, although we should note that a rejection of their null may be indicative that the exclusion restrictions for these instruments may be inappropriate. That is, some of the instruments have been improperly excluded from the regression model's specification.

The specification of an instrumental variables model asserts that the excluded instruments affect the dependent variable only *indirectly*, through their correlations with the included endogenous variables. If an excluded instrument exerts both direct and indirect influences on the dependent variable, the exclusion restriction should be rejected. This can be readily tested by including the variable as a regressor.

In our earlier example we saw that including `age` in the excluded instruments list caused a rejection of the *J* test. We had assumed that `age` could be treated as excluded from the model. Is that assumption warranted?

If `age` is entered as a regressor, it has a t-statistic over 8. Thus, its rejection as an excluded instrument may well reflect the misspecification of the equation, omitting age.

The specification of an instrumental variables model asserts that the excluded instruments affect the dependent variable only *indirectly*, through their correlations with the included endogenous variables. If an excluded instrument exerts both direct and indirect influences on the dependent variable, the exclusion restriction should be rejected. This can be readily tested by including the variable as a regressor.

In our earlier example we saw that including `age` in the excluded instruments list caused a rejection of the J test. We had assumed that `age` could be treated as excluded from the model. Is that assumption warranted?

If `age` is entered as a regressor, it has a t-statistic over 8. Thus, its rejection as an excluded instrument may well reflect the misspecification of the equation, omitting age.

The specification of an instrumental variables model asserts that the excluded instruments affect the dependent variable only *indirectly*, through their correlations with the included endogenous variables. If an excluded instrument exerts both direct and indirect influences on the dependent variable, the exclusion restriction should be rejected. This can be readily tested by including the variable as a regressor.

In our earlier example we saw that including `age` in the excluded instruments list caused a rejection of the J test. We had assumed that `age` could be treated as excluded from the model. Is that assumption warranted?

If `age` is entered as a regressor, it has a t-statistic over 8. Thus, its rejection as an excluded instrument may well reflect the misspecification of the equation, omitting age.

To test the *second* assumption—that the excluded instruments are sufficiently correlated with the included endogenous regressors—we should consider the goodness-of-fit of the “first stage” regressions relating each endogenous regressor to the entire set of instruments.

It is important to understand that the theory of single-equation (“limited information”) IV estimation requires that all columns of X are conceptually regressed on all columns of Z in the calculation of the estimates. We cannot meaningfully speak of “this variable is an instrument for that regressor” or somehow restrict which instruments enter which first-stage regressions. Stata’s `ivregress` or `ivreg2` will not let you do that because such restrictions only make sense in the context of estimating an entire system of equations by full-information methods (for instance, with `reg3`).

To test the *second* assumption—that the excluded instruments are sufficiently correlated with the included endogenous regressors—we should consider the goodness-of-fit of the "first stage" regressions relating each endogenous regressor to the entire set of instruments.

It is important to understand that the theory of single-equation ("limited information") IV estimation requires that all columns of X are conceptually regressed on all columns of Z in the calculation of the estimates. We cannot meaningfully speak of "this variable is an instrument for that regressor" or somehow restrict which instruments enter which first-stage regressions. Stata's `ivregress` or `ivreg2` will not let you do that because such restrictions only make sense in the context of estimating an entire system of equations by full-information methods (for instance, with `reg3`).

The `first` and `ffirst` options of `ivreg2` present several useful diagnostics that assess the first-stage regressions. If there is a single endogenous regressor, these issues are simplified, as the instruments either explain a reasonable fraction of that regressor's variability or not. With multiple endogenous regressors, diagnostics are more complicated, as each instrument is being called upon to play a role in each first-stage regression.

With sufficiently weak instruments, the asymptotic identification status of the equation is called into question. An equation identified by the order and rank conditions in a finite sample may still be *effectively unidentified*.

The `first` and `ffirst` options of `ivreg2` present several useful diagnostics that assess the first-stage regressions. If there is a single endogenous regressor, these issues are simplified, as the instruments either explain a reasonable fraction of that regressor's variability or not. With multiple endogenous regressors, diagnostics are more complicated, as each instrument is being called upon to play a role in each first-stage regression.

With sufficiently weak instruments, the asymptotic identification status of the equation is called into question. An equation identified by the order and rank conditions in a finite sample may still be *effectively unidentified*.

As Staiger and Stock (*Econometrica*, 1997) show, the weak instruments problem can arise even when the first-stage t- and F-tests are significant at conventional levels in a large sample. In the worst case, the bias of the IV estimator is the same as that of OLS, IV becomes inconsistent, and instrumenting only aggravates the problem.

Beyond the informal "rule-of-thumb" diagnostics such as $F > 10$, `ivreg2` computes several statistics that can be used to critically evaluate the strength of instruments. We can write the first-stage regressions as

$$X = Z\Pi + v$$

With X_1 as the endogenous regressors, Z_1 the excluded instruments and Z_2 as the included instruments, this can be partitioned as

$$X_1 = [Z_1 Z_2]\ [\Pi_{11}' \Pi_{12}']' + v_1$$

The rank condition for identification states that the $L \times K_1$ matrix Π_{11} must be of full column rank.

We do not observe the true Π_{11}, so we must replace it with an estimate. Anderson's (John Wiley, 1984) approach to testing the rank of this matrix (or that of the full Π matrix) considers the *canonical correlations* of the X and Z matrices. If the equation is to be identified, all K of the canonical correlations will be significantly different from zero.

The squared canonical correlations can be expressed as eigenvalues of a matrix. Anderson's *CC* test considers the null hypothesis that the minimum canonical correlation is zero. Under the null, the test statistic is distributed χ^2 with $(L - K + 1)$ d.f., so it may be calculated even for an exactly-identified equation. Failure to reject the null suggests the equation is unidentified. `ivreg2` reports this Lagrange Multiplier (LM) statistic.

We do not observe the true Π_{11}, so we must replace it with an estimate. Anderson's (John Wiley, 1984) approach to testing the rank of this matrix (or that of the full Π matrix) considers the *canonical correlations* of the X and Z matrices. If the equation is to be identified, all K of the canonical correlations will be significantly different from zero.

The squared canonical correlations can be expressed as eigenvalues of a matrix. Anderson's *CC* test considers the null hypothesis that the minimum canonical correlation is zero. Under the null, the test statistic is distributed χ^2 with $(L - K + 1)$ d.f., so it may be calculated even for an exactly-identified equation. Failure to reject the null suggests the equation is unidentified. `ivreg2` reports this Lagrange Multiplier (LM) statistic.

The Cragg–Donald statistic is a closely related test of the rank of a matrix. While the Anderson *CC* test is a LR test, the C–D test is a Wald statistic, with the same asymptotic distribution. The C–D statistic plays an important role in Stock and Yogo's work (see below). Both the Anderson and C–D tests are reported by `ivreg2` with the `first` option.

Recent research by Kleibergen and Paap (KP) (*J. Econometrics*, 2006) has developed a robust version of a test for the rank of a matrix: e.g. testing for *underidentification*. The statistic has been implemented by Kleibergen and Schaffer as command `ranktest`. If non-*i.i.d.* errors are assumed, the `ivreg2` output contains the K–P `rk` statistic in place of the Anderson canonical correlation statistic as a test of underidentification.

The Cragg–Donald statistic is a closely related test of the rank of a matrix. While the Anderson *CC* test is a LR test, the C–D test is a Wald statistic, with the same asymptotic distribution. The C–D statistic plays an important role in Stock and Yogo's work (see below). Both the Anderson and C–D tests are reported by `ivreg2` with the `first` option.

Recent research by Kleibergen and Paap (KP) (*J. Econometrics*, 2006) has developed a robust version of a test for the rank of a matrix: e.g. testing for *underidentification*. The statistic has been implemented by Kleibergen and Schaffer as command `ranktest`. If non-*i.i.d*. errors are assumed, the `ivreg2` output contains the K–P `rk` statistic in place of the Anderson canonical correlation statistic as a test of underidentification.

Stock and Yogo (Camb. U. Press festschrift, 2005) propose testing for weak instruments by using the F-statistic form of the C–D statistic. Their null hypothesis is that the estimator is weakly identified in the sense that it is subject to bias that the investigator finds unacceptably large.

Their test comes in two flavors: maximal relative bias (relative to the bias of OLS) and maximal size. The former test has the null that instruments are weak, where weak instruments are those that can lead to an asymptotic relative bias greater than some level b. This test uses the finite sample distribution of the IV estimator, and can only be calculated where the appropriate moments exist (when the equation is suitably overidentified: the m^{th} moment exists iff $m < (L - K + 1)$). The test is routinely reported in `ivreg2` and `ivregress` output when it can be calculated, with the relevant critical values calculated by Stock and Yogo.

Stock and Yogo (Camb. U. Press festschrift, 2005) propose testing for weak instruments by using the F-statistic form of the C–D statistic. Their null hypothesis is that the estimator is weakly identified in the sense that it is subject to bias that the investigator finds unacceptably large.

Their test comes in two flavors: maximal relative bias (relative to the bias of OLS) and maximal size. The former test has the null that instruments are weak, where weak instruments are those that can lead to an asymptotic relative bias greater than some level b. This test uses the finite sample distribution of the IV estimator, and can only be calculated where the appropriate moments exist (when the equation is suitably overidentified: the m^{th} moment exists iff $m < (L - K + 1)$). The test is routinely reported in `ivreg2` and `ivregress` output when it can be calculated, with the relevant critical values calculated by Stock and Yogo.

The second test proposed by Stock and Yogo is based on the performance of the Wald test statistic for the endogenous regressors. Under weak identification, the test rejects too often. The test statistic is based on the rejection rate r tolerable to the researcher if the true rejection rate is 5%. Their tabulated values consider various values for r. To be able to reject the null that the size of the test is unacceptably large (versus 5%), the Cragg–Donald F statistic must exceed the tabulated critical value.

The Stock–Yogo test statistics, like others discussed above, assume *i.i.d.* errors. The Cragg–Donald F can be robustified in the absence of *i.i.d.* errors by using the Kleibergen–Paap `rk` statistic, which `ivreg2` reports in that circumstance.

The second test proposed by Stock and Yogo is based on the performance of the Wald test statistic for the endogenous regressors. Under weak identification, the test rejects too often. The test statistic is based on the rejection rate r tolerable to the researcher if the true rejection rate is 5%. Their tabulated values consider various values for r. To be able to reject the null that the size of the test is unacceptably large (versus 5%), the Cragg–Donald F statistic must exceed the tabulated critical value.

The Stock–Yogo test statistics, like others discussed above, assume *i.i.d.* errors. The Cragg–Donald F can be robustified in the absence of *i.i.d.* errors by using the Kleibergen–Paap `rk` statistic, which `ivreg2` reports in that circumstance.

Further reading

There are many important considerations relating to the use of IV techniques, including LIML (limited-information maximum likelihood estimation) and GMM-CUE (continuously updated GMM estimates). For more details, please see

- Enhanced routines for instrumental variables/GMM estimation and testing. Baum CF, Schaffer ME, Stillman S, *Stata Journal* 7:4, 2007. Boston College Economics working paper no. 667, available from `http://ideas.repec.org`.
- *An Introduction to Modern Econometrics Using Stata*, Baum CF, Stata Press, 2006 (particularly Chapter 8).
- Instrumental variables and GMM: Estimation and testing. Baum CF, Schaffer ME, Stillman S, *Stata Journal* 3:1–31, 2003. Freely available from `http://stata-journal.com`.

Further reading

There are many important considerations relating to the use of IV techniques, including LIML (limited-information maximum likelihood estimation) and GMM-CUE (continuously updated GMM estimates). For more details, please see

- Enhanced routines for instrumental variables/GMM estimation and testing. Baum CF, Schaffer ME, Stillman S, *Stata Journal* 7:4, 2007. Boston College Economics working paper no. 667, available from `http://ideas.repec.org`.
- *An Introduction to Modern Econometrics Using Stata*, Baum CF, Stata Press, 2006 (particularly Chapter 8).
- Instrumental variables and GMM: Estimation and testing. Baum CF, Schaffer ME, Stillman S, *Stata Journal* 3:1–31, 2003. Freely available from `http://stata-journal.com`.

Further reading

There are many important considerations relating to the use of IV techniques, including LIML (limited-information maximum likelihood estimation) and GMM-CUE (continuously updated GMM estimates). For more details, please see

- Enhanced routines for instrumental variables/GMM estimation and testing. Baum CF, Schaffer ME, Stillman S, *Stata Journal* 7:4, 2007. Boston College Economics working paper no. 667, available from `http://ideas.repec.org`.
- *An Introduction to Modern Econometrics Using Stata*, Baum CF, Stata Press, 2006 (particularly Chapter 8).
- Instrumental variables and GMM: Estimation and testing. Baum CF, Schaffer ME, Stillman S, *Stata Journal* 3:1–31, 2003. Freely available from `http://stata-journal.com`.