Instrumental variables: Overview and advances

Christopher F Baum[1]

Boston College and DIW Berlin

UKSUG 13, London, September 2007

[1] Thanks to Austin Nichols for the use of his NASUG talks and Mark Schaffer for a number of useful suggestions.

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.

However, as Cameron and Trivedi point out in *Microeconometrics*, this method, *"widely used in econometrics and rarely used elsewhere, is conceptually difficult and easily misused."* (p.95)

My goal today is to present an overview of IV estimation—particularly for those of you from "elsewhere"—and lay out the benefits and pitfalls of the IV approach. I will discuss the latest enhancements to IV methods available in Stata 9.2 and 10, including the latest release of Baum, Schaffer, Stillman's widely used **ivreg2**, available for Stata 9.2 or better, and Stata 10's `ivregress`.

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.

However, as Cameron and Trivedi point out in *Microeconometrics*, this method, *"widely used in econometrics and rarely used elsewhere, is conceptually difficult and easily misused."* (p.95)

My goal today is to present an overview of IV estimation—particularly for those of you from "elsewhere"—and lay out the benefits and pitfalls of the IV approach. I will discuss the latest enhancements to IV methods available in Stata 9.2 and 10, including the latest release of Baum, Schaffer, Stillman's widely used **ivreg2**, available for Stata 9.2 or better, and Stata 10's `ivregress`.

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.

However, as Cameron and Trivedi point out in *Microeconometrics*, this method, *"widely used in econometrics and rarely used elsewhere, is conceptually difficult and easily misused."* (p.95)

My goal today is to present an overview of IV estimation—particularly for those of you from "elsewhere"—and lay out the benefits and pitfalls of the IV approach. I will discuss the latest enhancements to IV methods available in Stata 9.2 and 10, including the latest release of Baum, Schaffer, Stillman's widely used **ivreg2**, available for Stata 9.2 or better, and Stata 10's `ivregress`.

The discussion that follows is presented in much greater detail in three sources:

- Enhanced routines for instrumental variables/GMM estimation and testing. Baum, C.F., Schaffer, M.E., Stillman, S., Boston College Economics working paper no. 667, September 2007.
- *An Introduction to Modern Econometrics Using Stata*, Baum, C.F., Stata Press, 2006 (particularly Chapter 8).
- Instrumental variables and GMM: Estimation and testing. Baum, C.F., Schaffer, M.E., Stillman, S., *Stata Journal* 3:1–31, 2003. Boston College Economics working paper no. 545.

The discussion that follows is presented in much greater detail in three sources:

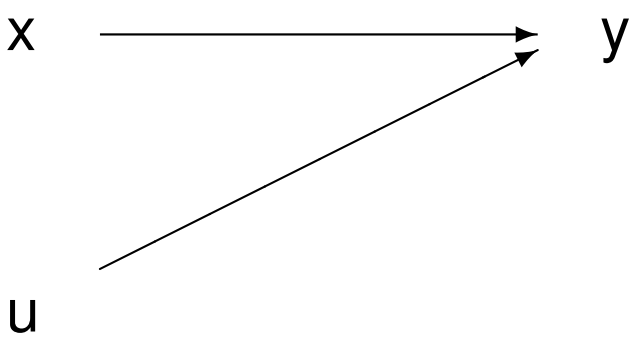
- Enhanced routines for instrumental variables/GMM estimation and testing. Baum, C.F., Schaffer, M.E., Stillman, S., Boston College Economics working paper no. 667, September 2007.
- *An Introduction to Modern Econometrics Using Stata*, Baum, C.F., Stata Press, 2006 (particularly Chapter 8).
- Instrumental variables and GMM: Estimation and testing. Baum, C.F., Schaffer, M.E., Stillman, S., *Stata Journal* 3:1–31, 2003. Boston College Economics working paper no. 545.

The discussion that follows is presented in much greater detail in three sources:

- Enhanced routines for instrumental variables/GMM estimation and testing. Baum, C.F., Schaffer, M.E., Stillman, S., Boston College Economics working paper no. 667, September 2007.
- *An Introduction to Modern Econometrics Using Stata*, Baum, C.F., Stata Press, 2006 (particularly Chapter 8).
- Instrumental variables and GMM: Estimation and testing. Baum, C.F., Schaffer, M.E., Stillman, S., *Stata Journal* 3:1–31, 2003. Boston College Economics working paper no. 545.

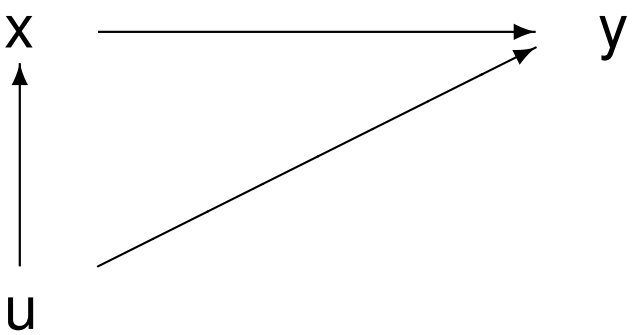
First let us consider a path diagram illustrating the problem addressed by IV methods. We can use ordinary least squares (OLS) regression to consistently estimate a model of the following sort.

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



However, OLS regression breaks down in the following circumstance:

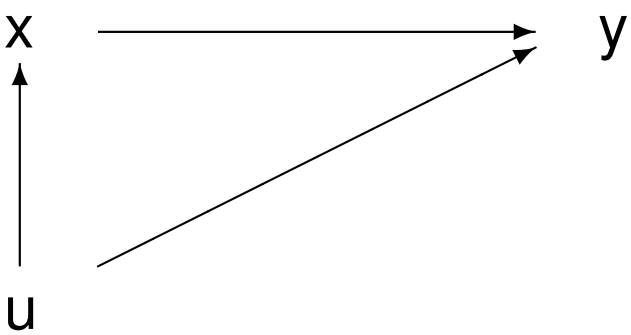
Endogeneity: $y = xb + u$
correlation between x and u; OLS inconsistent



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.

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.

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.

Although IV methods were first developed to cope with the problem of endogeneity in a simultaneous system, the correlation of regressor and error may arise for other reasons.

The presence of *measurement error in a regressor* will, in general terms, cause the same correlation of regressor and error in a model where behavior depends upon the true value of x and the statistician observes only a inaccurate measurement of x. Even if we assume that the magnitude of the measurement error is independent of the true value of x (often an inappropriate assumption) measurement error will cause OLS to produce biased and inconsistent parameter estimates of all parameters, not only that of the mismeasured regressor.

Although IV methods were first developed to cope with the problem of endogeneity in a simultaneous system, the correlation of regressor and error may arise for other reasons.

The presence of *measurement error in a regressor* will, in general terms, cause the same correlation of regressor and error in a model where behavior depends upon the true value of x and the statistician observes only a inaccurate measurement of x. Even if we assume that the magnitude of the measurement error is independent of the true value of x (often an inappropriate assumption) measurement error will cause OLS to produce biased and inconsistent parameter estimates of all parameters, not only that of the mismeasured regressor.

Another commonly encountered problem involves unobservable factors. Both *y* and *x* may be affected by *latent factors* such as ability. Consider a regression of (log) earnings (*y*) on years of schooling (*x*). The error term *u* embodies all other factors that affect earnings, such as the individual's innate ability or intelligence. But ability is surely likely to be correlated with educational attainment, causing a correlation between regressor and error. Mathematically, this is the same problem as that caused by endogeneity or measurement error.

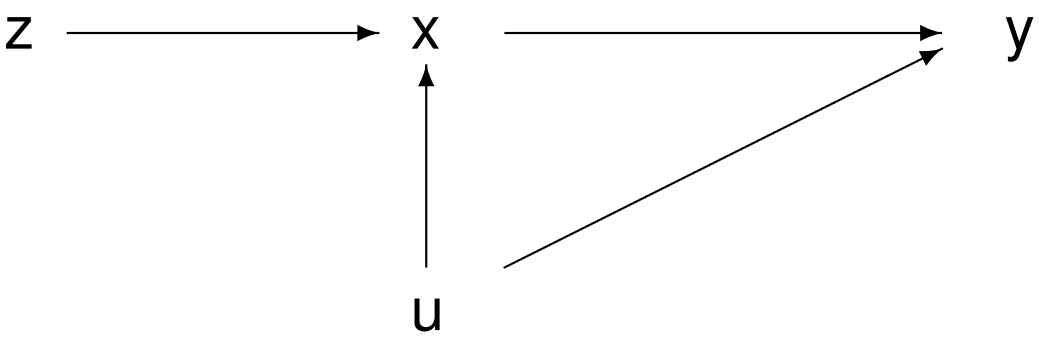
In a panel or longitudinal dataset, we could deal with this unobserved heterogeneity with the first difference or individual fixed effects transformations. But in a cross section dataset, we do not have that luxury, and must resort to other methods such as IV estimation.

Another commonly encountered problem involves unobservable factors. Both *y* and *x* may be affected by *latent factors* such as ability. Consider a regression of (log) earnings (*y*) on years of schooling (*x*). The error term *u* embodies all other factors that affect earnings, such as the individual's innate ability or intelligence. But ability is surely likely to be correlated with educational attainment, causing a correlation between regressor and error. Mathematically, this is the same problem as that caused by endogeneity or measurement error.

In a panel or longitudinal dataset, we could deal with this unobserved heterogeneity with the first difference or individual fixed effects transformations. But in a cross section dataset, we do not have that luxury, and must resort to other methods such as IV estimation.

The solution provided by IV methods may be viewed as:

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



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*.

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.

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.

For the problem of *measurement error in a regressor*, a common choice of instrument (z) is the rank of the mismeasured variable. Although the mismeasured variable contains an element of measurement error, if that error is relatively small, it will not alter the rank of the observation in the distribution.

In the case of *latent factors*, such as a regression of log earnings on years of schooling, we might be able to find an instrument (z) in the form of the mother's or father's years of schooling. More educated parents are more likely to produce more educated children; at the same time, the unobserved factors influencing the individual's educational attainment cannot affect prior events, such as their parent's schooling.

For the problem of *measurement error in a regressor*, a common choice of instrument (z) is the rank of the mismeasured variable. Although the mismeasured variable contains an element of measurement error, if that error is relatively small, it will not alter the rank of the observation in the distribution.

In the case of *latent factors*, such as a regression of log earnings on years of schooling, we might be able to find an instrument (z) in the form of the mother's or father's years of schooling. More educated parents are more likely to produce more educated children; at the same time, the unobserved factors influencing the individual's educational attainment cannot affect prior events, such as their parent's schooling.

What if we do not have data on parents' educational attainment? In a seminal (and highly criticized) 1991 paper in the *Quarterly Journal of Economics*, Angrist and Krueger (AK) used quarter of birth as an instrument for educational attainment, defining an indicator variable for those born in the first calendar quarter. Although arguably independent of innate ability, how could this factor be correlated with educational attainment?

AK argue that compulsory school attendance laws in the U.S. (and varying laws across states) cause some individuals to attend school longer than others depending on when they enter primary school, which is in turn dependent on their birth date. We can test whether this relationship holds by regressing years of schooling on the indicator variable.

Example: OLS vs IV

What if we do not have data on parents' educational attainment? In a seminal (and highly criticized) 1991 paper in the *Quarterly Journal of Economics*, Angrist and Krueger (AK) used quarter of birth as an instrument for educational attainment, defining an indicator variable for those born in the first calendar quarter. Although arguably independent of innate ability, how could this factor be correlated with educational attainment?

AK argue that compulsory school attendance laws in the U.S. (and varying laws across states) cause some individuals to attend school longer than others depending on when they enter primary school, which is in turn dependent on their birth date. We can test whether this relationship holds by regressing years of schooling on the indicator variable.

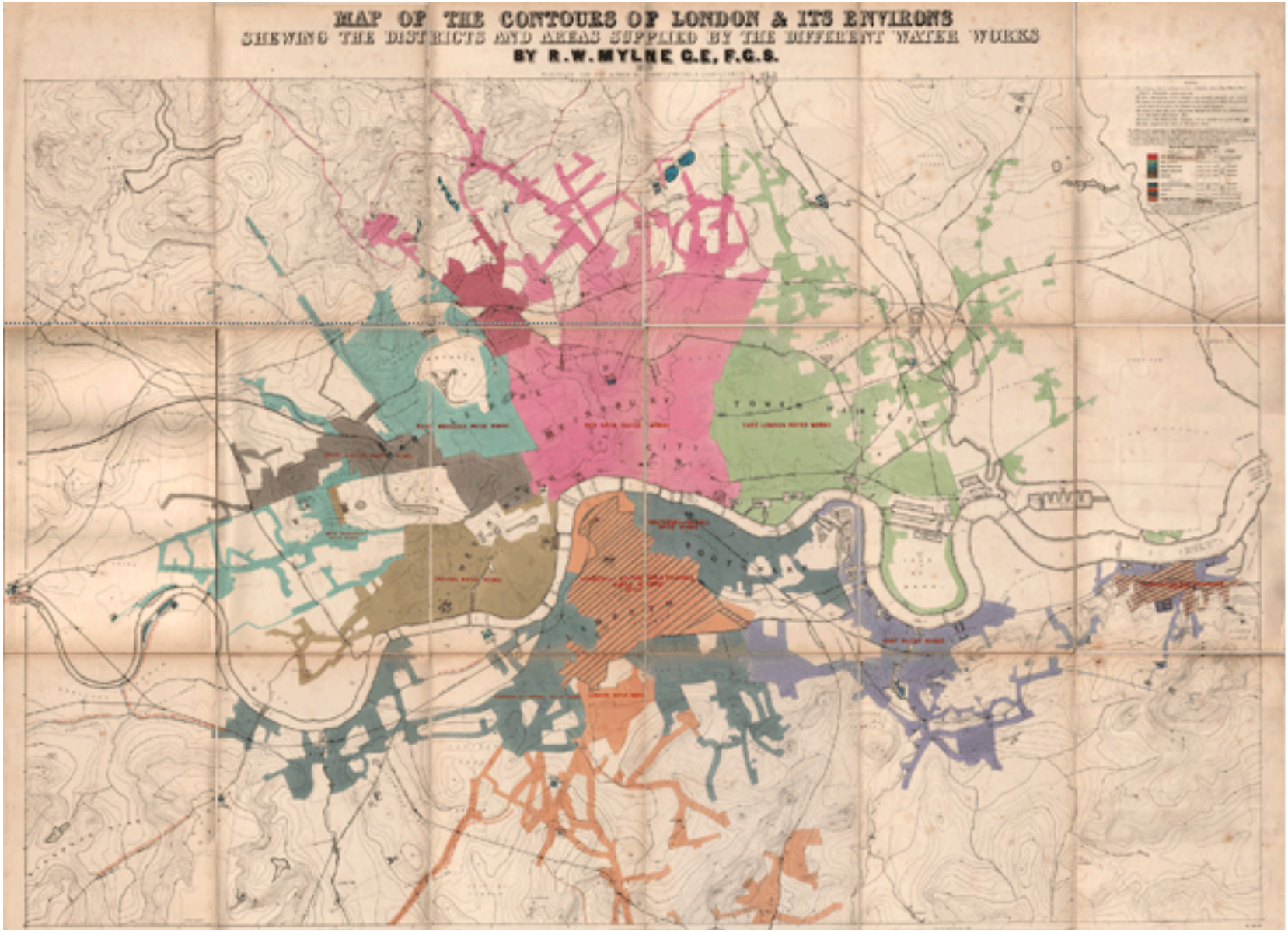
Example: OLS vs IV

An interesting example—particularly as I walked this morning over the site of the Fleet Ditch—is provided by Paul Grootendorst in his research paper “A review of instrumental variables estimation in the applied health sciences." He suggests that IV methods were developed in 1855 by John Snow in *On the Mode of Communication of Cholera*. [`http://www.ph.ucla.edu/EPI/snow/snowbook.html`]. I excerpt from his paper below.

Snow hypothesized that cholera was waterborne. But he could not merely examine water purity and its correlation with the incidence of cholera, for those who drank impure water were more likely to be poor, to live in crowded tenements and to live in an environment contaminated in many ways. What could serve as an instrument?

An interesting example—particularly as I walked this morning over the site of the Fleet Ditch—is provided by Paul Grootendorst in his research paper “A review of instrumental variables estimation in the applied health sciences." He suggests that IV methods were developed in 1855 by John Snow in *On the Mode of Communication of Cholera*. [`http://www.ph.ucla.edu/EPI/snow/snowbook.html`]. I excerpt from his paper below.

Snow hypothesized that cholera was waterborne. But he could not merely examine water purity and its correlation with the incidence of cholera, for those who drank impure water were more likely to be poor, to live in crowded tenements and to live in an environment contaminated in many ways. What could serve as an instrument?


MAP OF THE CONTOURS OF LONDON & ITS ENVIRONS
SHEWING THE DISTRICTS AND AREAS SUPPLIED BY THE DIFFERENT WATER WORKS
BY R.W. MYLNE C.E, F.G.S.

The instrument Snow proposed: *the identity of the water company supplying households with drinking water*. Londoners received water directly from the Thames. The Lambeth water company drew water from the river upstream of the main sewage discharge; the Southwark and Vauxhall company drew water just below the main discharge.

Snow mentions that "The pipes of each Company go down all the streets, and into nearly all the courts and alleys. ... No fewer than 300,000 people ... of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice and, in most cases, without their knowledge; one group supplied with water containing the sewage of London...the other group having water quite free from such impurity."

The instrument Snow proposed: *the identity of the water company supplying households with drinking water*. Londoners received water directly from the Thames. The Lambeth water company drew water from the river upstream of the main sewage discharge; the Southwark and Vauxhall company drew water just below the main discharge.

Snow mentions that “The pipes of each Company go down all the streets, and into nearly all the courts and alleys. ... No fewer than 300,000 people ... of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice and, in most cases, without their knowledge; one group supplied with water containing the sewage of London...the other group having water quite free from such impurity."

Demonstrably, the identity of the water suppliers (and the lack of public perception of their relative quality) is correlated with water purity and through that mechanism influences the incidence of waterborne disease. It is likely to be uncorrelated with other factors influencing cholera (such as the health status of those living in certain neighborhoods) given that the suppliers competed throughout the city.

Although econometricians may believe that IV methods were the product of Sewall Wright's analysis of agricultural supply and demand in the 1920s, or the work of the Cowles Commission in the 1950s, they may have far predated that era!

Demonstrably, the identity of the water suppliers (and the lack of public perception of their relative quality) is correlated with water purity and through that mechanism influences the incidence of waterborne disease. It is likely to be uncorrelated with other factors influencing cholera (such as the health status of those living in certain neighborhoods) given that the suppliers competed throughout the city.

Although econometricians may believe that IV methods were the product of Sewall Wright's analysis of agricultural supply and demand in the 1920s, or the work of the Cowles Commission in the 1950s, they may have far predated that era!

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.

Instruments may be *weak*: satisfactorily exogenous, but only weakly correlated with the endogenous regressors. As Bound, Jaeger, Baker (NBER TWP 1993, *JASA* 1995) argue "the cure can be worse than the disease."

Staiger and Stock (*Econometrica*, 1997) formalized the definition of weak instruments. Many researchers conclude from their work that if the first-stage F statistic exceeds 10, their instruments are sufficiently strong. This criterion does not necessarily establish the absence of a weak instruments problem.

Stock and Yogo (Camb. U. Press festschrift, 2005) further explore the issue and provide useful rules of thumb for evaluating the weakness of instruments. `ivreg2` and Stata 10's `ivregress` now present Stock–Yogo tabulations based on the Cragg–Donald statistic.

Instruments may be *weak*: satisfactorily exogenous, but only weakly correlated with the endogenous regressors. As Bound, Jaeger, Baker (NBER TWP 1993, *JASA* 1995) argue "the cure can be worse than the disease."

Staiger and Stock (*Econometrica*, 1997) formalized the definition of weak instruments. Many researchers conclude from their work that if the first-stage *F* statistic exceeds 10, their instruments are sufficiently strong. This criterion does not necessarily establish the absence of a weak instruments problem.

Stock and Yogo (Camb. U. Press festschrift, 2005) further explore the issue and provide useful rules of thumb for evaluating the weakness of instruments. `ivreg2` and Stata 10's `ivregress` now present Stock–Yogo tabulations based on the Cragg–Donald statistic.

Instruments may be *weak*: satisfactorily exogenous, but only weakly correlated with the endogenous regressors. As Bound, Jaeger, Baker (NBER TWP 1993, *JASA* 1995) argue "the cure can be worse than the disease."

Staiger and Stock (*Econometrica*, 1997) formalized the definition of weak instruments. Many researchers conclude from their work that if the first-stage *F* statistic exceeds 10, their instruments are sufficiently strong. This criterion does not necessarily establish the absence of a weak instruments problem.

Stock and Yogo (Camb. U. Press festschrift, 2005) further explore the issue and provide useful rules of thumb for evaluating the weakness of instruments. `ivreg2` and Stata 10's `ivregress` now present Stock–Yogo tabulations based on the Cragg–Donald statistic.

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.

We consider the model

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

with X $(N \times k)$ and 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 an estimate that solves $\bar{g}(\hat{\beta}_{GMM}) = 0$.

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.

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.

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.

Example: IV and IV(robust) vs IV-GMM

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.

Example: IV and IV(robust) vs IV-GMM

If errors are considered to exhibit arbitrary intra-cluster correlation in a dataset with M clusters ($M << N$), 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 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. `ivreg2` 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 both commands.

In `ivreg2` (but not in `ivregress`) you may also specify a *vce* that is robust to autocorrelation while maintaining the assumption of conditional homoskedasticity: that is, *AC* without the *H*.

The estimators we have discussed are available from Baum, Schaffer and Stillman's *ivreg2* package (`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 estimators we have discussed are available from Baum, Schaffer and Stillman's *ivreg2* package (`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*`)` option, the *vce* is cluster-robust.

With the `robust` and `bw( )` options, the *vce* is *HAC* with the default Bartlett kernel, or "Newey–West". Other `kernel( )` choices lead to alternative *HAC* estimators. In `ivreg2`, both `robust` and `bw( )` options must be specified for *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*`)` option, the *vce* is cluster-robust.

With the `robust` and `bw( )` options, the *vce* is *HAC* with the default Bartlett kernel, or "Newey–West". Other `kernel( )` choices lead to alternative *HAC* estimators. In `ivreg2`, both `robust` and `bw( )` options must be specified for *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*`)` option, the *vce* is cluster-robust.

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

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 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 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 routinely calculated by `ivreg2` when the `gmm` 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 routinely calculated by `ivreg2` when the `gmm` 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.

Example: Tests of overidentifying restrictions

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.

Example: Tests of overidentifying restrictions

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.

Example: C (GMM distance) test of a subset of overidentifying restrictions

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.

Example: C (GMM distance) test of a subset of overidentifying restrictions

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.

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?

Example: Test of exclusion of an instrument

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?

Example: Test of exclusion of an instrument

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 there is no analytical validity in such a computation in a limited-information single equation context. You may do so by using `reg3`, but must then specify the entire simultaneous system.

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 there is no analytical validity in such a computation in a limited-information single equation context. You may do so by using `reg3`, but must then specify the entire simultaneous system.

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` routinely reports this LR statistic.

Example: Analysis of first stage regressions

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` routinely reports this LR statistic.

Example: Analysis of first stage regressions

The C–D 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 C–D 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 canonical correlations may also be used to test a set of instruments for *redundancy* by considering their statistical significance in the first stage regressions. This can be calculated, in robust form, as a K–P LM test. The `redundant( )` option of `ivreg2` allows a set of excluded instruments to be tested for relevance, with the null hypothesis that they do not contribute to the asymptotic efficiency of the equation.

In this example, we add `mrt` (marital status) to the equation, and test it for redundancy. It barely rejects the null hypothesis.

Example: Test of redundancy of instruments

The canonical correlations may also be used to test a set of instruments for *redundancy* by considering their statistical significance in the first stage regressions. This can be calculated, in robust form, as a K–P LM test. The `redundant( )` option of `ivreg2` allows a set of excluded instruments to be tested for relevance, with the null hypothesis that they do not contribute to the asymptotic efficiency of the equation.

In this example, we add `mrt` (marital status) to the equation, and test it for redundancy. It barely rejects the null hypothesis.

Example: Test of redundancy of instruments

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.

Example: Stock–Yogo critical values for C–D or K–P test

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.

Example: Stock–Yogo critical values for C–D or K–P test

The Anderson–Rubin (*Ann. Math. Stat.*, 1949) test for the significance of endogenous regressors in the structural equation is robust to the presence of weak instruments, and may be "robustified" for non-$i.i.d$. errors if an alternative *VCE* is estimated. The test essentially substitutes the reduced-form equations into the structural equation and tests for the joint significance of the excluded instruments in Z_1.

If a single endogenous regressor appears in the equation, alternative test statistics robust to weak instruments and non-$i.i.d$. errors are provided by Moreira and Poi (*Stata J.*, 2003) and Mikusheva and Poi (*Stata J.*, 2006) as the `condivreg` and `condtest` commands.

The Anderson–Rubin (*Ann. Math. Stat.*, 1949) test for the significance of endogenous regressors in the structural equation is robust to the presence of weak instruments, and may be "robustified" for non-*i.i.d.* errors if an alternative *VCE* is estimated. The test essentially substitutes the reduced-form equations into the structural equation and tests for the joint significance of the excluded instruments in Z_1.

If a single endogenous regressor appears in the equation, alternative test statistics robust to weak instruments and non-*i.i.d.* errors are provided by Moreira and Poi (*Stata J.*, 2003) and Mikusheva and Poi (*Stata J.*, 2006) as the `condivreg` and `condtest` commands.

LIML and GMM-CUE

OLS and IV estimators are special cases of *k-class estimators*: OLS with $k = 0$ and IV with $k = 1$. Limited-information maximum likelihood (LIML) is another member of this class, with k chosen optimally in the estimation process. Like any ML estimator, LIML is invariant to normalization. In an equation with two endogenous variables, it does not matter whether you specify y_1 or y_2 as the left-hand variable. One of the other virtues of the LIML estimator is that it has been found to be more resistant to weak instruments problems than the IV estimator. On the down side, it makes the distributional assumption of normally distributed (and *i.i.d.*) errors. `ivreg2` produces LIML estimates with the `liml` option, and `liml` is a subcommand for Stata 10's `ivregress`.

If the *i.i.d.* assumption of LIML is not reasonable, you may use the GMM equivalent: the *continuously updated* GMM estimator, or CUE estimator. In `ivreg2`, the `cue` option combined with `robust`, `cluster` and/or `bw( )` options specifies that non-*i.i.d.* errors are to be modeled. GMM-CUE requires numerical optimization via Stata's `ml` command, and may require many iterations to converge.

`ivregress` provides an iterated GMM estimator, which is not the same estimator as GMM-CUE.

Example: LIML and GMM-CUE

When you may (and may not!) use IV

You now know that you may only use IV methods when you can plausibly specify the necessary instruments. Beyond that important concern, two cases come to mind that are FAQs on Statalist.

A common inquiry: what if I have an endogenous regressor that is a dummy variable? Should I, for instance, fit a probit model to generate the "hat values", estimate the model with OLS including those "hat values" instead of the 0/1 values, and puzzle over what to do about the standard errors?

(An aside: you really do not want to do two-stage least squares "by hand", for one of the things that you must then deal with is getting the correct *VCE* estimate. The *VCE* and *RMSE* computed by the second-stage regression are not correct, as they are generated from the "hat values", not the original regressors. But back to our question).

When you may (and may not!) use IV

You now know that you may only use IV methods when you can plausibly specify the necessary instruments. Beyond that important concern, two cases come to mind that are FAQs on Statalist.

A common inquiry: what if I have an endogenous regressor that is a dummy variable? Should I, for instance, fit a probit model to generate the "hat values", estimate the model with OLS including those "hat values" instead of the 0/1 values, and puzzle over what to do about the standard errors?

(An aside: you really do not want to do two-stage least squares "by hand", for one of the things that you must then deal with is getting the correct *VCE* estimate. The *VCE* and *RMSE* computed by the second-stage regression are not correct, as they are generated from the "hat values", not the original regressors. But back to our question).

When you may (and may not!) use IV

You now know that you may only use IV methods when you can plausibly specify the necessary instruments. Beyond that important concern, two cases come to mind that are FAQs on Statalist.

A common inquiry: what if I have an endogenous regressor that is a dummy variable? Should I, for instance, fit a probit model to generate the "hat values", estimate the model with OLS including those "hat values" instead of the 0/1 values, and puzzle over what to do about the standard errors?

(An aside: you really do not want to do two-stage least squares "by hand", for one of the things that you must then deal with is getting the correct *VCE* estimate. The *VCE* and *RMSE* computed by the second-stage regression are not correct, as they are generated from the "hat values", not the original regressors. But back to our question).

Should I fit a probit model to generate the "hat values", estimate the model with OLS including those "hat values" instead of the 0/1 values, and puzzle over what to do about the standard errors?

No, you should just estimate the model with `ivreg2` or `ivregress`, treating the dummy endogenous regressor like any other endogenous regressor. This yields consistent point and interval estimates of its coefficient. There are other estimators (notably in the field of selection models or treatment regression) that explicitly deal with this problem, but they impose additional conditions on the problem. If you can use those methods, fine. Otherwise, just run IV. This solution is also appropriate for count data.

Another solution to the problem of an endogenous dummy (or count variable), as discussed by Cameron and Trivedi, is due to Basmann (*Econometrica*, 1957). Obtain fitted values for the endogenous regressor with appropriate nonlinear regression (logit or probit for a dummy, Poisson regression for a count variable) using all the instruments (included and excluded). Then do regular linear IV using the fitted value as an instrument, but the original dummy (or count variable) as the regressor. This is also a consistent estimator, although it has a different asymptotic distribution than does that of straight IV.

Example: Regression on an endogenous dummy

A second FAQ: what if my equation includes a nonlinear function of an endogenous regressor? For instance, from Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, p. 231, we might write the supply and demand equations for a good as

$$\begin{aligned} \log q^s &= \gamma_{12} \log(p) + \gamma_{13}[\log(p)]^2 + \delta_{11} z_1 + u_1 \\ \log q^d &= \gamma_{22} \log(p) + \delta_{22} z_2 + u_2 \end{aligned}$$

where we have suppressed intercepts for convenience. The exogenous factor z_1 shifts supply but not demand. The exogenous factor z_2 shifts demand but not supply. There are thus two exogenous variables available for identification.

This system is still *linear in parameters*, and we can ignore the log transformations on p, q. But it is, in Wooldridge's terms, *nonlinear in endogenous variables*, and identification must be treated differently.

A second FAQ: what if my equation includes a nonlinear function of an endogenous regressor? For instance, from Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, p. 231, we might write the supply and demand equations for a good as

$$\begin{aligned} \log q^s &= \gamma_{12} \log(p) + \gamma_{13}[\log(p)]^2 + \delta_{11} z_1 + u_1 \\ \log q^d &= \gamma_{22} \log(p) + \delta_{22} z_2 + u_2 \end{aligned}$$

where we have suppressed intercepts for convenience. The exogenous factor z_1 shifts supply but not demand. The exogenous factor z_2 shifts demand but not supply. There are thus two exogenous variables available for identification.

This system is still *linear in parameters*, and we can ignore the log transformations on p, q. But it is, in Wooldridge's terms, *nonlinear in endogenous variables*, and identification must be treated differently.

If we used these equations to obtain $\log(p) = y_2$ as a function of exogenous variables and errors (the reduced form equation), the result would not be linear. $E[y_2|z]$ would not be linear unless $\gamma_{13} = 0$, assuming away the problem, and $E[y_2^2|z]$ will not be linear in any case. We might imagine that y_2^2 could just be treated as an additional endogenous variable, but then we need at least one more instrument. Where do we find it?

Given the nonlinearity, other functions of z_1 and z_2 will appear in a linear projection with y_2^2 as the dependent variable. Under linearity, the reduced form for y_2 involves z_1, z_2 and combinations of the errors. Square that reduced form, and $E[y_2^2|z]$ is a function of z_1^2, z_2^2 and $z_1 z_2$ (and the expectation of the squared composite error). Given that this relation has been derived under assumptions of linearity and homoskedasticity, we should also include the levels of z_1, z_2 in the projection (first stage regression).

If we used these equations to obtain $\log(p) = y_2$ as a function of exogenous variables and errors (the reduced form equation), the result would not be linear. $E[y_2|z]$ would not be linear unless $\gamma_{13} = 0$, assuming away the problem, and $E[y_2^2|z]$ will not be linear in any case. We might imagine that y_2^2 could just be treated as an additional endogenous variable, but then we need at least one more instrument. Where do we find it?

Given the nonlinearity, other functions of z_1 and z_2 will appear in a linear projection with y_2^2 as the dependent variable. Under linearity, the reduced form for y_2 involves z_1, z_2 and combinations of the errors. Square that reduced form, and $E[y_2^2|z]$ is a function of z_1^2, z_2^2 and $z_1 z_2$ (and the expectation of the squared composite error). Given that this relation has been derived under assumptions of linearity and homoskedasticity, we should also include the levels of z_1, z_2 in the projection (first stage regression).

The supply equation may then be estimated with instrumental variables using z_1, z_2, z_1^2, z_2^2 and $z_1 z_2$ as instruments. You could also use higher powers of the exogenous variables.

The mistake that may be made in this context involves what Wooldridge calls the **forbidden regression**: trying to mimic 2SLS by substituting fitted values for some of the endogenous variables inside the nonlinear functions. Nether the conditional expectation of the linear projection nor the linear projection operator passes through nonlinear functions, and such attempts "...rarely produce consistent estimators in nonlinear systems." (p. 235)

The supply equation may then be estimated with instrumental variables using z_1, z_2, z_1^2, z_2^2 and $z_1 z_2$ as instruments. You could also use higher powers of the exogenous variables.

The mistake that may be made in this context involves what Wooldridge calls the **forbidden regression**: trying to mimic 2SLS by substituting fitted values for some of the endogenous variables inside the nonlinear functions. Nether the conditional expectation of the linear projection nor the linear projection operator passes through nonlinear functions, and such attempts "...rarely produce consistent estimators in nonlinear systems." (p. 235)

In our example above, imagine regressing y_2 on exogenous variables, saving the predicted values, and squaring them. The "second stage" regression would then regress $\log(q)$ on $\hat{y}, \hat{y}^2, z_1$.

This two-step procedure does not yield the same results as estimating the equation by 2SLS, and it generally cannot produce consistent estimates of the structural parameters. The linear projection of the square is not the square of the linear projection, and the "by hand" approach assumes they are identical.

In our example above, imagine regressing y_2 on exogenous variables, saving the predicted values, and squaring them. The "second stage" regression would then regress $\log(q)$ on $\hat{y}, \hat{y}^2, z_1$.

This two-step procedure does not yield the same results as estimating the equation by 2SLS, and it generally cannot produce consistent estimates of the structural parameters. The linear projection of the square is not the square of the linear projection, and the "by hand" approach assumes they are identical.

We illustrate the forbidden regression with a variation on the log wage model estimated in earlier examples. Although the second-stage OLS regression will yield the wrong standard errors (as any 2SLS “by hand" estimates will) we find that the forbidden regression appears to produce significant coefficients for the nonlinear relationship. Unfortunately, those estimates are inconsistent, and as you can see quite far from the NL-IV estimates generated by the proper instrumenting procedure.

Example: The forbidden regression

Testing for *i.i.d.* errors in IV

In the context of an equation estimated with instrumental variables, the standard diagnostic tests for heteroskedasticity and autocorrelation are generally not valid.

In the case of heteroskedasticity, Pagan and Hall (*Econometric Reviews*, 1983) showed that the Breusch–Pagan or Cook–Weisberg tests (`estat hettest`) are generally not usable in an IV setting. They propose a test that will be appropriate in IV estimation where heteroskedasticity may be present in more than one structural equation. Mark Schaffer's `ivhettest`, part of the `ivreg2` suite, performs the Pagan–Hall test under a variety of assumptions on the indicator variables. It will also reproduce the Breusch–Pagan test if applied in an OLS context.

Testing for *i.i.d.* errors in IV

In the context of an equation estimated with instrumental variables, the standard diagnostic tests for heteroskedasticity and autocorrelation are generally not valid.

In the case of heteroskedasticity, Pagan and Hall (*Econometric Reviews*, 1983) showed that the Breusch–Pagan or Cook–Weisberg tests (`estat hettest`) are generally not usable in an IV setting. They propose a test that will be appropriate in IV estimation where heteroskedasticity may be present in more than one structural equation. Mark Schaffer's `ivhettest`, part of the `ivreg2` suite, performs the Pagan–Hall test under a variety of assumptions on the indicator variables. It will also reproduce the Breusch–Pagan test if applied in an OLS context.

In the same token, the Breusch–Godfrey statistic used in the OLS context (`estat bgodfrey`) will generally not be appropriate in the presence of endogenous regressors, overlapping data or conditional heteroskedasticity of the error process. Cumby and Huizinga (*Econometrica*, 1992) proposed a generalization of the BG statistic which handles each of these cases.

Their test is actually more general in another way. Its null hypothesis of the test is that the regression error is a moving average of known order $q \geq 0$ against the general alternative that autocorrelations of the regression error are nonzero at lags greater than q. In that context, it can be used to test that autocorrelations beyond any q are zero. Like the BG test, it can test multiple lag orders. The C–H test is available as Baum and Schaffer's `ivactest` routine, part of the `ivreg2` suite.

In the same token, the Breusch–Godfrey statistic used in the OLS context (`estat bgodfrey`) will generally not be appropriate in the presence of endogenous regressors, overlapping data or conditional heteroskedasticity of the error process. Cumby and Huizinga (*Econometrica*, 1992) proposed a generalization of the BG statistic which handles each of these cases.

Their test is actually more general in another way. Its null hypothesis of the test is that the regression error is a moving average of known order $q \geq 0$ against the general alternative that autocorrelations of the regression error are nonzero at lags greater than q. In that context, it can be used to test that autocorrelations beyond any q are zero. Like the BG test, it can test multiple lag orders. The C–H test is available as Baum and Schaffer's `ivactest` routine, part of the `ivreg2` suite.

Panel data IV estimation

The features of `ivreg2` are also available in the routine `xtivreg2`, which is a "wrapper" for `ivreg2`. This routine of Mark Schaffer's extends Stata's `xtivreg`'s support for the fixed effect (`fe`) and first difference (`fd`) estimators. The `xtivreg2` routine is available from `ssc`.

Just as `ivreg2` may be used to conduct a Hausman test of IV vs. OLS, Schaffer and Stillman's `xtoverid` routine may be used to conduct a Hausman test of random effects vs. fixed effects after `xtreg, re` and `xtivreg, re`. This routine can also calculate tests of overidentifying restrictions after those two commands as well as `xthtaylor`. The `xtoverid` routine is also available from `ssc`.

Panel data IV estimation

The features of `ivreg2` are also available in the routine `xtivreg2`, which is a "wrapper" for `ivreg2`. This routine of Mark Schaffer's extends Stata's `xtivreg`'s support for the fixed effect (`fe`) and first difference (`fd`) estimators. The `xtivreg2` routine is available from `ssc`.

Just as `ivreg2` may be used to conduct a Hausman test of IV vs. OLS, Schaffer and Stillman's `xtoverid` routine may be used to conduct a Hausman test of random effects vs. fixed effects after `xtreg, re` and `xtivreg, re`. This routine can also calculate tests of overidentifying restrictions after those two commands as well as `xthtaylor`. The `xtoverid` routine is also available from `ssc`.