EC771: Econometrics, Spring 2004

Greene, Econometric Analysis (5th ed, 2003)

Chapters 4–5: Properties of LS and IV estimators

We now consider the least squares estimator from the statistical viewpoint, rather than as merely an algebraic curve–fitting tool, returning to the assumptions that we spelled out in earlier presentations of the least squares approach.

The least squares method can be motivated by several of its advantages: including, of course, its ease of computation. Least squares, like many of the methods we use in econometrics, is a generalized method of moments (GMM) estimator, motivated by a set of orthogonality conditions on population moments. Let x

be the vector of independent variables in the population regression function, which may be stochastic or nonstochastic. We assume that these variables are exogenous: in statistical terms, that the population errors are orthogonal to the independent variables: $E[\epsilon|x] = 0$. If this conditional mean is zero, the covariance is also zero, and we have a vector of K moment conditions: one for each independent variable. Those moment conditions may be written as

$$E[x\epsilon] = E[x(y - x'\beta)] = 0,$$

or

$$E[xy] = E[xx']\beta.$$

This is a population relationship on the moments of the variables. A method of moments estimator replaces population moments with consistent estimates derived from the sample. The normal equations of least squares, $(X'X)b = X'y$ may be written as

$$\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right] = \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right] b$$

in which we estimate the population moments as the average of the sample moments over the n observations of the sample. Therefore, the least squares estimator may be considered as a method of moments estimator, in which the population relationship is reflected by that methodology applied to the sample.

We can also motivate the use of least squares as a solution to the problem of finding an optimal linear predictor: $x'\gamma$. The mean squared error of this predictor is

$$MSE = E[y - x'\gamma]^2$$

which may be decomposed as

$$MSE = E[y - E[y|x]]^2 + E[E[y|x] - x'\gamma]^2$$

in which the first term does not contain γ. Minimizing MSE thus only involves the second term. The first order condition for this

problem will lead us to the same population expression as above:

$$E[xy] = E[xx']\gamma$$

which leads to the same least squares estimator as that for β above. Thus, the least squares estimator solves the optimal linear predictor problem: even without specification of the form of the conditional mean function $E[y|x]$, since we did not use the assumption of linearity in the above derivation.

Unbiasedness of least squares

The LS estimator is unbiased:

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ b &= (X'X)^{-1}X'(X\beta + \epsilon) \\ b &= \beta + (X'X)^{-1}X'\epsilon \end{aligned}$$

Taking expectations of this expression, we find that $E[b|X] = E[b] = \beta$, since by the assumption of orthogonal regressors (exogeneity), the

expectation of $X'\epsilon$ is zero. For any particular set of observations X the least squares estimator has expectation β. If we average over the possible values of X (in the case where X is stochastic) the unconditional mean of b is β as well.

Best linear unbiased estimation

To discuss the precision of the LS estimators, we may consider X as a matrix of fixed constants in order to derive the sampling variance of b. Alternatively, we can perform the analysis conditional on X, and average over the possible values of X as above. Returning to the definition of the LS estimator,

$$b = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}X'\epsilon.$$

Since we can write $b = \beta + A\epsilon$, where $A = (X'X)^{-1}X'$, b is a linear function of the disturbances, and is thus a linear estimator; from

the above, we may also write $b = Ay$, so that the LS estimator is a linear function of the dependent variable. Since we have established unbiasedness, we may state that the LS b is a linear, unbiased estimator. What about its covariance matrix? Given exogeneity of the X variables and our assumption on spherical disturbances,

$$\begin{aligned} Var[b|X] &= E[(b-\beta)(b-\beta)'|X] \\ &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'E[\epsilon\epsilon'|X]X(X'X)^{-1} \\ &= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

Let us consider a more general approach to computing a minimum–variance linear unbiased estimator of β. Let $b_0 = Cy$ be another linear, unbiased estimator of β, which implies

$$E[Cy|X] = E[(CX\beta + C\epsilon)|X) = \beta,$$

which implies that $CX = I$. The covariance matrix of b_0 can be found, using the above derivation, as $\sigma^2 CC'$. Let

$$D = C - (X'X)^{-1}X',$$

so $Dy = (b_0 - b)$. Then

$$Var[b_0|X] = \sigma^2 \left[(D + (X'X)^{-1}X')(D + (X'X)^{-1}X')'\right]$$

Since

$$CX = DX + (X'X)^{-1}(X'X) = I,$$

$DX = 0$, and

$$Var[b_0|X] = \sigma^2(X'X)^{-1} + \sigma^2 DD',$$

we find that the variance of any linear, unbiased estimator of β equals the variance of the LS estimator, $Var[b]$, plus a positive semidefinite matrix. The matrix D contains the differences between the elements of the arbitrary estimator b_0 and the LS estimator b. Iff those differences are uniformly zero, the variance of

b_0 will be minimized; but if that is so, $b_0 = b$. For all other choices of b_0, the estimator will be less precise than the corresponding LS estimator. Thus, b is the minimum variance linear unbiased estimator of β: the Gauss-Markov theorem, stating that b is "BLUE": the Best Linear Unbiased Estimator. The theorem also states that if we seek to estimate $w'\beta$, where w is an arbitrary vector of constants, the BLUE of $w'\beta$ will be $w'b$.

Estimating $\sigma^2(X'X)^{-1}$

In order to test hypotheses about β or to form confidence intervals, we need a sample estimate of the covariance matrix

$$Var[b|X] = \sigma^2(X'X)^{-1},$$

in particular, the population parameter σ^2. In the least squares approach, unlike the maximum likelihood approach to the problem, we

do not generate an estimate of σ^2 in the optimization. Since σ^2 is the expected value of ϵ_i^2 and e_i is an estimate of ϵ_i, by analogy we might calculate

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

which is, indeed, the maximum likelihood estimator of σ^2. But the least squares residuals, although consistent estimates of the corresponding ϵ_i, are imperfect estimates of their population counterparts in that we have replaced β with its unbiased estimate, b:

$$e_i = y_i - x_i'b = \epsilon_i - x_i'(b - \beta)$$

Although the expected value of e_i is ϵ_i, that does not necessarily hold for the expected value of the squares. In matrix form,

$$e = My = M[X\beta + \epsilon] = M\epsilon,$$

since $MX = 0$. An estimator based on the residual sum of squares is then

$$e'e = \epsilon' M \epsilon$$

with expected value

$$E[e'e|X] = E[\epsilon' M \epsilon|X]$$

The quantity $\epsilon' M \epsilon$ is a scalar, thus equal to its trace, and

$$E[tr(\epsilon' M \epsilon)|X] = E[tr(M \epsilon \epsilon')|X]$$

Since M is a function of X, this can be written as

$$tr(M E[\epsilon \epsilon'|X]) = tr(M \sigma^2 I) = \sigma^2 tr(M)$$

Since we know that

$$\begin{aligned} & tr(I - X(X'X)^{-1}X') \\ = tr(I_n) - & tr[(X'X)^{-1}(X'X)] \\ & = n - K \end{aligned}$$

We find that $E[e'e|X] = (n-K)\sigma^2$, so that an unbiased estimator of σ^2 will be

$$s^2 = \frac{e'e}{n-K}$$

and the standard error of regression, its positive square root, is s. We can now also calculate the estimated variance–covariance matrix of the estimated parameters:

$$Est.Var[b|X] = s^2(X'X)^{-1},$$

and the positive square roots of the diagonal elements of this matrix are the estimated standard errors s_i of the regression parameters b_i.

Normality and simple hypothesis testing

Thus far, we have not used the assumption of normality: even in deriving estimates of the precision of the LS coefficients above, we did not invoke normality. But to make use of those estimated standard errors, we must impose a specific distributional assumption: that the elements of b are distributed multivariate Normal.

How might we test an hypothesis about a specific coefficient? Assuming normality,

$$z = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}}$$

where S^{kk} is the k^{th} element on the main diagonal of $(X'X)^{-1}$. If we could compute this expression, z would have a standard Normal distribution under $H_0 : b_k = \beta_k^0$. Using s^2 rather than σ^2, we can derive a feasible test statistic:

$$\frac{(n-K)s^2}{\sigma^2} = \frac{e'e}{\sigma^2} = \left[\frac{\epsilon}{\sigma}\right]' M \left[\frac{\epsilon}{\sigma}\right]$$

is an idempotent quadratic form in a standard normal vector $\frac{\epsilon}{\sigma}$, which will have a χ^2 distribution with $tr(M) = (n-K)$ degrees of freedom. We may prove that this quadratic form is independent of the expression in s^2. That is, if ϵ is normally distributed, the least squares coefficient vector b is statistically independent of the residual vector e and all functions of e, including s^2.

Since a Student t–distributed variable is the ratio of a standard normal variate to the square root of a χ^2 random variable which has been divided by its degrees of freedom, a bit of algebra will lead us to

$$t_{n-K} = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}}$$

in which we have essentially replaced σ^2 with its consistent estimate s^2. The resulting test statistic is distributed Student t with $(n-K)$ degrees of freedom under H_0. In particular, the ratio of the estimated coefficient to its estimated standard error is distributed t under the null hypothesis that the population coefficient equals zero. Likewise, we may construct a α–percent confidence interval for β_k as

$$Pr(b_k - t_{\alpha/2} s_{b_k} \leq \beta_k \leq b_k + t_{\alpha/2} s_{b_k}) = 1 - \alpha$$

where $(1-\alpha)$ is the desired level of confidence, and $t_{\alpha/2}$ is the appropriate critical value from

the t distribution with $(n-K)$ degrees of freedom.

How might we test the hypothesis that the regression as a whole is significant: in a model with a constant term, a joint test of the hypothesis that all regressors' slopes are zero? The ANOVA F statistic may be written as a transformation of the R^2 value:

$$F_{n-K}^{K-1} = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$$

which will have a Fisher F distribution under that null hypothesis. In later discussions, we will consider alternative joint hypotheses on combinations or subsets of the regression parameters.

Collinearity

Although the OLS estimator is minimum variance in the class of linear, unbiased estimators of the population parameters, that variance may still be unacceptably large under certain circumstances: such as a "high degree" of collinearity. If two regressors are perfectly correlated, of course, their sampling variances go to infinity, since the inverse matrix that enters the expression for their sampling variances cannot be calculated. What if that correlation is high, or in general, what if there are near–linear dependencies in the regressor matrix? When this is encountered, we may find that some data points have a great deal of "leverage": small changes in the data matrix may cause large changes in the parameter estimates. Although the overall fit of the regression (as measured by R^2 or $\bar{R}^2$) may be very

good, the coefficients may have very high standard errors, and perhaps even incorrect signs or implausibly large magnitudes.

These are all understandable consequences of near–linear dependencies in the regressor matrix. If we consider a k–variable regression model containing a constant and $(k-1)$ regressors, we may write the k^{th} diagonal element of $(X'X)^{-1}$ as:

$$\frac{1}{(1-R_k^2)S_{kk}}$$

where R_k^2 is the R^2 from the regression of variable k on all other variables in the model, and S_{kk} is the variation in the k^{th} variable about its mean. The estimated variance of the k^{th} coefficient estimate is s^2 times this quantity. Observations about this expression:

- The greater the correlation of x_k with the other regressors (including the constant term),

cet.par., the higher will be the estimated variance;

- The greater the variation in x_k about its mean, cet. par., the lower will be the estimated variance;

- The better the overall fit of the regression, the lower will be the estimated variance.

This expression is the rationale for the so–called VIF, or variance inflation factor, $(1 - R_k^2)^{-1}$. The VIF_k measures the degree to which the variance has been inflated due to the non–orthogonality of regressor k. Another measure, a summary for the regression equation, is the condition number of $X'X$, which can be expressed as the positive square root of the ratio of the largest to the smallest eigenvalue

of the matrix. Since a matrix which is near–computationally singular will have at least one very small eigenvalue, the condition number of such a matrix will be large (relative to the value of unity that would apply for I). Belsley, Kuh and Welsch (BKW) have written the seminal work on such regression diagnostics, and they suggest that the condition number should be calculated from a transformed data matrix in which each regressor has unit length. A rule of thumb would suggest that a condition number in excess of 20 might be cause for concern. But just as there is no objective measure of how small the determinant of $X'X$ might be to trigger instability in the estimates, it is difficult to come up with a particular value that would indicate a problem. One conclusion should be clear: the statement that "these estimates are flawed by collinearity among the regressors" is hardly sensible if the model fits well and its coefficient estimates are acceptably precise.

Some illustrations of the effects of contrived collinearity may be viewed in the accompanying handout. A number of these techniques, as well as several of BKW's proposed measures of "influential observations", may be found in the Stata documentation under "regression diagnostics."

Large–sample properties of OLS and IV estimators

To consider the asymptotic properties of least squares estimators, we leave the DGP for X unspecified: it may include any mixture of constants and random variables generated independently of the DGP producing ϵ. Two crucial assumptions:

- (x_i, ϵ_i), $i = 1, \ldots, n$ is a sequence of independent observations;

- plim $\frac{X'X}{n} = Q$, a positive definite matrix

The OLS estimator may then be written as

$$b = \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\epsilon}{n}\right).$$

Presuming the existence of Q^{-1},

$$\text{plim}\, b = \beta + Q^{-1}\,\text{plim}\,\left(\frac{X'\epsilon}{n}\right)$$

The plim of the last term may be written as the sample average of $x_i\epsilon_i = w_i$, and each term in that average has expectation zero. For non-stochastic x, this follows from the marginal distribution of ϵ; for stochastic x, the independence of the two DGPs provides the result. Thus, we may write

$$\text{plim}\, b = \beta + Q^{-1}\,\text{plim}\,\bar{w}.$$

Likewise, we may use the spherical distribution of ϵ to derive the conditional variance of this

expression:

$$Var[\bar{w}|X] = E[\bar{w}\bar{w}'|X] = \left(\frac{\sigma^2}{n}\right)\left(\frac{X'X}{n}\right),$$

or

$$Var[\bar{w}] = \left(\frac{\sigma^2}{n}\right) E\left(\frac{X'X}{n}\right)$$

The variance will collapse to zero if the expectation converges to a constant matrix, so that the leading scalar will dominate as n increases. Under this condition—on the "well–behavedness" of the regressor matrix—the limit of the variance of $\bar{w}$ is zero. Since the mean of $\bar{w}$ is identically zero and its variance converges to zero, we may state that $\bar{w}$ converges in mean square to zero:

$$\text{plim}\,\frac{x'\epsilon}{n} = 0,$$

or

$$\text{plim}\, b = \beta.$$

The assumptions we have used above are often too restrictive in the case of time–series data with trending variables and polynomials in the regressors. In that case, a weaker set of assumptions about X, the so–called Grenander conditions, state that the regressors are suitably "well-behaved" to lead to consistency of the OLS estimator. Those conditions are likely to be satisfied in empirical datasets.

Asymptotic normality of the OLS estimator

Our result on the plim of b allows us to write

$$\sqrt{n}(b-\beta) = \left(\frac{X'X}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\epsilon$$

Since this inverse matrix has a plim equal to Q^{-1}, the limiting distribution of the above quantity is that of $Q^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\epsilon$; we need to consider the limiting distribution of $\sqrt{n}(\bar{w}-E[\bar{w}])$, where the expectation is zero. What then is

the limiting distribution of $\sqrt{n}\bar{w}$? Assuming independence of the observations, $\bar{w}$ is the average of n independent random vectors $w_i = x_i\epsilon_i$ with means zero and variances (given spherical disturbances) of $\sigma^2 Q_i$. Thus the variance of $\sqrt{n}\bar{w}$ is $\sigma^2\bar{Q}_n = \frac{\sigma^2}{n}[Q_1 + Q_2 + \cdots + Q_n]$. As long as this sum is not dominated by any term and the regressors are "well behaved" as discussed above, the limit of this quantity is $\sigma^2 Q$. Thus if $x_i\epsilon_i$ are n independent vectors distributed with mean zero and finite variance $\sigma^2 Q$, we may write

$$\left(\frac{1}{\sqrt{n}}\right) X'\epsilon \xrightarrow{d} N[0, \sigma^2 Q].$$

We may premultiply this quantity by Q^{-1}, which then leads to the result that

$$\sqrt{n}(b - \beta) \xrightarrow{d} N[0, \sigma^2 Q^{-1}],$$

or in terms of b itself,

$$b \overset{a}{\sim} N\left[\beta, \frac{\sigma^2}{n} Q^{-1}\right].$$

The importance of this result is that if the regressors are "well behaved" and the observations are independent, then the asymptotic normality of the OLS estimator does not depend on normality of the disturbances, but rather on the central limit theorems used in the derivation. It will, of course, follow if the disturbances themselves are distributed normally.

To make this operational, we must estimate the two quantities in the covariance matrix: σ^2 by $\frac{e'e}{(n-K)}$ and $(1/n)Q^{-1}$ by $(X'X)^{-1}$. The former estimator can be demonstrated to be consistent, since it can be written

$$s^2 = \frac{1}{(n-K)}[\epsilon'\epsilon - \epsilon' X (X'X)^{-1} X'\epsilon]$$

$$= \frac{n}{n-K}\left[\frac{\epsilon'\epsilon}{n} - \left(\frac{\epsilon' X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\epsilon}{n}\right)\right].$$

Since the leading constant has a plim of one, and the second term in the brackets converges

to zero, we are concerned about the convergence of the first term. Under the weak conditions of finite moments of ϵ (two if they are identically distributed), we have that plim $s^2 = \sigma^2$, giving us the appropriate estimator for the asymptotic covariance matrix of b.

The delta method

The delta method may be used to generate estimated variances and covariances for functions of b. Let $f(b)$ be a set of J continuous and continuously differentiable functions of b, and define

$$C(b) = \frac{\partial f(b)}{\partial b'}$$

be the $J \times K$ matrix whose j^{th} row is the vector of derivatives of the j^{th} function with respect to b'. We can then write

$$\text{plim } f(b) = f(\beta)$$

and

$$\text{plim}\, C(b) = \frac{\partial f(\beta)}{\partial \beta'} = \Gamma.$$

Using a linear Taylor series,

$$f(b) = f(\beta) + \Gamma(b - \beta) + \ldots$$

If plim $b = \beta$, the higher–order terms are negligible as $n \to \infty$. So we can write

$$f(b) \overset{a}{\sim} N\left[f(\beta), \Gamma \frac{\sigma^2}{n} Q^{-1} \Gamma'\right].$$

so that the operational estimated asymptotic covariance matrix will be

$$C[s^2(X'X)^{-1}]C'.$$

If any of the J functions are nonlinear, the unbiasedness of b may not carry over to $f(b)$. Nevertheless, $f(b)$ will be a consistent estimator of $f(\beta)$, with a consistent estimate of the asymptotic covariance matrix. This is the rationale for the widely–employed delta method, implemented in Stata as the `testnl` command.

Asymptotic efficiency

What about asymptotic efficiency, the large–sample counterpart of the Gauss–Markov result? In finite samples, we can prove that OLS is BLUE under a set of assumptions on X and ϵ (which do not require normality of the ϵ). As we noted in examining maximum likelihood estimators, OLS is also a MLE if ϵ is distributed multivariate normal. Since MLEs are asy. efficient among consistent and asy. normally distributed estimators, we can state that OLS estimators are asy. efficient in the presence of normally distributed ϵ. Conversely, if the error distribution is not normal, this result of asy. efficiency does not follow.

Heterogeneity in x_i and dependent observations

The assumptions made to establish these asymptotic results include exogeneity of the regressors (violations of which we will discuss in the

next section), spherically distributed errors, and independence of the observations. The latter assumption is often called into question in the context of a panel, or longitudinal data set, where we have multiple observations on each of a number of units. It is surely likely that the xs will be correlated across observations within individuals: indeed, some may be constant for each individual. The regressors are likely to include both stochastic terms (such as family income, or firms' revenues) and non–stochastic regressors such as an individual "fixed effect". The asymptotics for such a setup are often described as "large n, small T": that is, very commonly we have a large number of observations (individuals, families, or firms) with a time–series of modest length on each individual. In this case, we hold T fixed and let n increase, treating each individual's observations as a unit. The conditions necessary to establish convergence are those related to $n \to \infty$.

This is the setup, for instance, underlying the "dynamic panel data" estimator of Arellano and Bond, which is commonly applied to panels of firm–level balance sheet and income statement data. There are, of course, instances of "small n, large T" panels, for which the asymptotics consider holding n fixed and letting T increase; and in the econometric theory underlying a number of panel unit–root tests, there are asymptotics in which both n and T are allowed to increase.

A second cause for violation of the assumption of independence is that of regressors which are prior values of the dependent variable: that is, lagged dependent variables. We continue to assume that the disturbances are $i.i.d.$, but even with this assumption, it is obvious that the regressor vectors are correlated across observations. Since every observation y_t is dependent on the history of the disturbances,

we cannot assume strict exogeneity: by construction, current ϵ is correlated with future x. The finite–sample results presented earlier no longer hold in this case of stochastic regressors which cannot be considered independent of the error process, past, present and future; only asymptotic results remain.

To generate asymptotic results for this case, we must modify the assumption of strict exogeneity with

$$E[\epsilon_t | x_{t-s}] = 0 \ \forall s \geq 0.$$

This states that the disturbance at period t is an innovation, uncorrelated with the past history of the x process. It cannot be uncorrelated with the future of the process, since it will become part of those future values. We further must assume that the series in x are stationary (at least in terms of covariance stationarity), which assumes that they have finite,

non–time–varying second moments which depend only on the temporal displacement between their values; and that the autocorrelation of the series is damped (so that the dependence between observations declines with the temporal displacement, and sample estimates of the autocovariance function will be suitable estimates of their population counterparts). The combination of these conditions is equivalent to stating that the regressors are stationary and ergodic. Under these conditions, consistency of the OLS estimator can be proven.

Next, we consider the appropriate strategy if the assumption of exogeneity of the regressors is untenable: for instance, in the context of a simultaneous equations model.

The method of instrumental variables

The equation to be estimated is, in matrix notation,

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

with typical row

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

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

Homoskedasticity:$\Omega = \sigma^2 I$

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

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

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

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

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

$$\text{Regr } X = [X_1 \;\; X_2] = [X_1 \;\; Z_2] = [\text{Endog} \quad \text{Exog}]$$

$$\text{Inst } Z = [Z_1 \; Z_2] = [\text{Excluded} \quad \text{Included}]$$

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

Denote by P_Z the projection matrix $Z(Z'Z)^{-1}Z'$. The instrumental variables or two–stage least squares (2SLS) estimator of β is

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

The asymptotic distribution of the IV estimator under the assumption of conditional homoskedasticity can be written as follows. Let

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

$$Q_{ZZ} = E(Z_i' Z_i)$$

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

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

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

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

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

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

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

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

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

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

Note that some packages, including Stata's `ivreg`, include a degrees–of–freedom correction to the estimate of $\widehat{\sigma}^2$ by replacing n with $n-L$. This correction is not necessary, however, since the estimate of $\widehat{\sigma}^2$ would not be unbiased anyway (cf. Greene, 2000, p. 373.)

The Generalized Method of Moments

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

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

where g_i is $L \times 1$. The exogeneity of the instruments means that there are L moment conditions, or orthogonality conditions, that will be satisfied at the true value of β:

$$E(g_i(\beta)) = 0$$

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

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

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

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

If the equation is overidentified, however, so that $L > K$, then we have more equations than we do unknowns, and in general it will not be possible to find a $\widehat{\beta}$ that will set all L sample moment conditions to exactly zero. In this

case, we take an $L \times L$ weighting matrix W and use it to construct a quadratic form in the moment conditions. This gives us the GMM objective function:

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

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

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

yields the GMM estimator:

$$\widehat{\beta}_{GMM} = (X'ZWZ'X)^{-1}X'ZWZ'y \qquad (3)$$

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

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

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

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

$$V(\widehat{\beta}_{GMM}) = \frac{1}{n}(Q'_{XZ}WQ_{XZ})^{-1}(Q'_{XZ}WSWQ_{XZ})(Q'_{XZ}WQ_{XZ})^{-1} \quad (4)$$

The *efficient* GMM estimator is the GMM estimator with an optimal weighting matrix W, one which minimizes the asymptotic variance of the estimator. This is achieved by choosing $W = S^{-1}$. Substitute this into Equation (3) and Equation (4) and we obtain the efficient GMM estimator

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

with asymptotic variance

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

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

GMM and heteroskedastic errors

Let us start with one of the most commonly encountered cases in cross–section analysis: heteroskedasticity of unknown form, but no clustering. We need a heteroskedasticity–consistent estimator of S. Such an $\widehat{S}$ is available by using

the standard "sandwich" approach to robust covariance estimation. Denote by $\hat{\Omega}$ the diagonal matrix of squared residuals:

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

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

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

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

The $\hat{u}$ used for the matrix can come from any consistent estimator of β; efficiency is not required. In practice, the most common choice

for estimating $\widehat{u}$ is the IV residuals. This gives us the algorithm for the *feasible efficient two-step GMM estimator*, as implemented in `ivreg2,gmm`.

1. Estimate the equation using IV.

2. Form the residuals $\widehat{u}$. Use these to form the optimal weighting matrix $\hat{W} = \hat{S}^{-1} = \left(\frac{1}{n}(Z'\hat{\Omega}Z)\right)^{-1}$.

3. Calculate the efficient GMM estimator $\hat{\beta}_{EGMM}$ and its variance-covariance matrix using the estimated optimal weighting matrix.This yields

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

with asymptotic variance

$$V(\hat{\beta}_{EGMM}) = (X'Z(Z'\hat{\Omega}Z)^{-1}Z'X)^{-1}$$

GMM, IV and homoskedastic vs. heteroskedastic errors

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

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

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

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

Substituting into Equation (3), we find that it reduces to the formula for the IV estimator in

Equation (3). To obtain the variance of the estimator, however, we *do* need an estimate of σ^2. If we use the residuals of the IV estimator to calculate $\widehat{\sigma}^2 = \frac{1}{n}\widehat{u}'\widehat{u}$, we obtain

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

Finally, if we now set

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

and substitute into the formula for the asymptotic variance of the efficient GMM estimator we find that it reduces to the formula for the asymptotic variance of the IV estimator. In effect, under the assumption of conditional homoskedasticity, the (efficient) iterated GMM estimator is the IV estimator, and the iterations converge after one step. It is worth noting that the IV estimator is not the only such efficient GMM estimator under conditional homoskedasticity. Instead of treating $\widehat{\sigma}^2$ as a parameter to be estimated in a second stage,

what if we return to the GMM criterion function and minimize by simultaneously choosing $\widehat{\beta}$ *and* $\widehat{\sigma}^2$? The estimator that solves this minimization problem is in fact the Limited Information Maximum Likelihood estimator (LIML). In effect, under conditional homoskedasticity, the continuously updated GMM estimator is the LIML estimator. Calculating the LIML estimator does not require numerical optimization methods; it can be calculated as the solution to an eigenvalue problem. The latest version of `ivreg2` (Baum, Schaffer and Stillman) supports LIML and k–class estimation methods.

What are the implications of heteroskedasticity for the IV estimator? Recall that in the presence of heteroskedasticity, the IV estimator is inefficient but consistent, whereas the standard estimated IV covariance matrix is inconsistent. Asymptotically correct inference is

still possible, however. In these circumstances the IV estimator is a GMM estimator with a sub–optimal weighting matrix, and hence the general formula for the asymptotic variance of a general GMM estimator The IV weighting matrix $\hat{W}$ remains, what we need is a consistent estimate of $\hat{S}$. This is easily done, using exactly the same method employed in two–step efficient GMM. First, form the "hat" matrix $\hat{\Omega}$ using the IV residuals, and use this matrix to form the $\hat{S}$ matrix. Substitute this $\hat{S}$, the (sub–optimal) IV weighting matrix $\hat{W}$ and the sample estimates of Q_{XZ} and Q_{ZZ} into the general formula for the asymptotic variance of a GMM estimator (4), and we obtain an estimated variance–covariance matrix for the IV estimator that is robust to the presence of heteroskedasticity:

$$\text{Robust } V(\hat{\beta}_{IV}) = (X'P_ZX)^{-1} \times (X'Z(Z'Z)^{-1}(Z'\hat{\Omega}Z)(Z'Z)^{-1}Z'X)(X'P_ZX)^{-1}$$

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

We may also deal with non–independence in the error distribution by using a kernel estimator to produce autocorrelation–consistent standard errors. The "Newey–West" standard errors are HAC, that is, heteroskedasticity– and autocorrelation–consistent. If one does not doubt the homoskedasticity assumption, but wants to deal with autocorrelation, one should use the "AC" correction without the "White" piece. Thus, the latest `ivreg2` allows selection of H, AC, or HAC standard errors by combining the `robust`, `bandwidth` and `kernel` options. One may use the Bartlett kernel, as do Newey and West, but a number of alternative kernel estimators are available that will likewise produce a positive definite estimated covariance matrix. See `help ivreg2` for details.

Testing the relevance of instruments

An instrumental variable must satisfy two requirements: it must be correlated with the included endogenous variable(s), and orthogonal to the error process. The former condition may be readily tested by examining the fit of the first stage regressions. The first stage regressions are reduced form regressions of the endogenous variables X_1 on the full set of instruments Z; the relevant test statistics here relate to the explanatory power of the excluded instruments Z_1 in these regressions.

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

Equation (): that is, whether the instruments are uncorrelated with the error process. This condition will arise when the order condition for identification is satisfied in inequality: the number of instruments excluded from the equation exceeds the number of included endogenous variables. This test can and should be performed as a standard diagnostic in any overidentified instrumental variables estimation. These are tests of the joint hypotheses of correct model specification and the orthogonality conditions, and a rejection may properly call either or both of those hypotheses into question.

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

$$J(\widehat{\beta}_{EGMM}) = n\overline{g}(\widehat{\beta})'\widehat{S}^{-1}\overline{g}(\widehat{\beta}) \overset{A}{\sim} \chi^2_{L-K} \quad (5)$$

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

$$J(\hat{\beta}_{EGMM}) = \hat{u}Z'(Z'\hat{\Omega}Z)^{-1}Z\hat{u}' \overset{A}{\sim} \chi^2_{L-K} \quad (6)$$

The J statistic is distributed as χ^2 with degrees of freedom equal to the number of overidentifying restrictions $L - K$ rather than the total number of moment conditions L because, in effect, K degrees of freedom are used up in estimating the coefficients of β. J is the most common diagnostic utilized in GMM estimation to evaluate the suitability of the model. A rejection of the null hypothesis implies that the instruments are not satisfying the orthogonality conditions required for their employment. This may be either because they are not truly exogenous, or because they are being incorrectly excluded from the regression. The J statistic is calculated and displayed by `ivreg2`

when the `gmm`, `robust`, or `cluster` options are specified.

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

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

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

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

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

For excluded instruments, this is equivalent to dropping them from the instrument list. For included instruments, the C test hypothecates placing them in the list of included endogenous variables: in essence, treating them as endogenous regressors. The C test, distributed χ^2 with degrees of freedom equal to the loss of overidentifying restrictions (i.e., the number of suspect instruments being tested), has the null hypothesis that the specified variables are proper instruments.

Although the C statistic can be calculated as the simple difference between the Hansen–Sargan statistics for two regressions, this procedure can generate a negative test statistic in finite samples. In the IV context this problem can be avoided and the C statistic guaranteed to be non–negative if the estimate of the error variance $\hat{\sigma}^2$ from the original (restricted, more efficient) IV regression is used to calculate the

Sargan statistic for the unrestricted IV regression as well. The equivalent procedure in the GMM context is to use the $\widehat{S}$ matrix from the original estimation to calculate both J statistics. More precisely, $\widehat{S}$ from the restricted estimation is used to form the restricted J statistic, and the submatrix of $\widehat{S}$ with rows/columns corresponding to the unrestricted estimation is used to form the J statistic for the unrestricted estimation.

The C test is conducted in `ivreg2` by specifying the `orthog` option, and listing the instruments (either included or excluded) to be challenged. The equation must still be identified with these instruments either removed or reconsidered as endogenous if the C statistic is to be calculated. Note that if the unrestricted equation is exactly identified, the Hansen–Sargan statistic for the unrestricted equation will be zero and

the C statistic will coincide with the Hansen–Sargan statistic for the original (restricted) equation, and this will be true *irrespective* of the instruments used to identify the unrestricted estimation. This illustrates how the Hansen–Sargan overidentification test is an "omnibus" test for the failure of *any* of the instruments to satisfy the orthogonality conditions, but at the same time requires that the investigator believe that at least *some* of the instruments are valid.

Durbin–Wu–Hausman tests for endogeneity in IV estimation

Many econometrics texts discuss the issue of "OLS vs. IV" in the context of the Durbin–Wu–Hausman (DWH) tests, which involve estimating the model via both OLS and IV approaches and comparing the resulting coefficient vectors. In the Hausman form of the test,

a quadratic form in the differences between the two coefficient vectors, scaled by the precision matrix, gives rise to a test statistic for the null hypothesis that the OLS estimator is consistent and fully efficient.

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

$$H = n(\widehat{\beta}^c - \widehat{\beta}^e)' D^- (\widehat{\beta}^c - \widehat{\beta}^e)$$

where

$$D = \left(V(\widehat{\beta}^c) - V(\widehat{\beta}^e) \right) \tag{7}$$

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

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

There are a variety of ways of conducting a DWH endogeneity test in Stata for the standard IV case with conditional homoskedasticity. Three equivalent ways of obtaining the

Durbin flavor of the Durbin–Wu–Hausman statistic are:

1. Estimate the less efficient but consistent model using IV, followed by the command `hausman, save`. Then estimate the fully efficient model by OLS (or by IV if only a subset of regressors is being tested for endogeneity), followed by `hausman, sigmamore`.

2. Estimate the fully efficient model using `ivreg2`, specifying the regressors to be tested in the `orthog` option.

3. Estimate the less efficient but consistent model using `ivreg`, then use `ivendog` to conduct an endogeneity test. This program will take as its argument a $varlist$ consisting of the subset of regressors to be tested

for endogeneity; if the $varlist$ is empty, the full set of endogenous regressors is tested.

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