Endogenous Regressors and Instrumental Variables

James L. Powell
Department of Economics
University of California, Berkeley

Endogenous Regressors and Inconsistency of LS

The *endogenous regressor linear model*, a workhorse of econometric applications, assumes that the dependent variable and regressors are both random and satisfy the linear relation

$$y = X\beta + \varepsilon,$$

but the usual assumption of zero conditional mean of the error terms given the regressors is not satisfied, i.e,

$$E(X'\varepsilon) \neq 0 \Longrightarrow E(\varepsilon|X) \neq 0.$$

This is a more serious departure from the assumptions of the classical linear model than was the case for the Generalized Regression model, which maintained $E(\varepsilon|X) = 0$ but permitted nonconstant variances and/or nonzero correlations across error terms; unlike the Generalized Regression model, the classical least squares estimator will be inconsistent for β if the errors are correlated with the regressors. Writing

$$\hat{\beta}_{LS} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon,$$

application of an appropriate version of the Law of Large Numbers will imply that

$$\begin{aligned} \text{plim } \frac{1}{T}X'X &\equiv \text{plim } \frac{1}{T}\sum_t x_t x = \text{plim } \frac{1}{T}\sum_t E(x_t x_t') \\ &\equiv M_{xx}, \end{aligned}$$

which is routinely assumed to be invertible for the classical regression models, and

$$\begin{aligned} \text{plim } \frac{1}{T}X'\varepsilon &= \text{plim } \frac{1}{T}\sum_T x_t\varepsilon_t = \text{plim } \frac{1}{T}\sum_T E(x_t\varepsilon_t) \\ &\equiv M_{x\varepsilon}. \end{aligned}$$

Since $E(x_t\varepsilon_t) \neq 0$ generally implies $M_{x\varepsilon} \neq 0$ (at least for stationary data), it follows that

$$\begin{aligned} \text{plim } \hat{\beta}_{LS} &= \beta + \text{plim } (\frac{1}{T}X'X)^{-1}(\frac{1}{T}X'\varepsilon) \\ &= \beta + M_{xx}^{-1}M_{x\varepsilon} \\ &\equiv \beta^* \neq \beta. \end{aligned}$$

Of course, the inconsistency of $\hat{\beta}_{LS}$ for β is only a "problem" if you want to estimate the "primitive" parameter β rather than β^*, which is the vector of best linear predictor coefficients for y_t given x_t. Several examples below will outline cases in which β is a natural "structural" parameter for behavior even though it cannot be interpreted as the vector of best linear projection coefficients.

An alternative interpretation of the inconsistency of classical least squares lies in the fact that it solves a sample moment condition of the form

$$0 = \frac{1}{T} X'(y - X\hat{\beta}_{LS}) = \frac{1}{T} \sum_t x_t (y_t - x_t' \hat{\beta}_{LS})$$

(which are called the *normal equations*), but the analogous condition for the population data distribution is not satisfied, i.e.,

$$E(x_t(y_t - x_t'\beta)) = E(x_t \varepsilon_t) \neq 0.$$

Thus, least squares solves the "wrong" moment condition in the sample. The classical "solution" to this problem of endogenous regressors supposes that there is some L-dimensional vector of *instrumental variables*, denoted z_t below, which is observable and satisfies

$$E(z_t \varepsilon_t) \equiv M_{z\varepsilon} = 0$$

for all values of t. Thus, though the regressors x_t are correlated with the error terms ε_t, the instrumental variables (or "instruments") z_t are not. If only some of the components of x_t are correlated with the errors, the remaining components can be included in the instrument vector z_t; depending on the particular application, other transformations of observable random variables may be suitable instruments.

In addition to being uncorrelated with the error terms, it will be necessary that the instrumental variables z_t are "fully correlated" with the regressors x_t: that is, if

$$E[z_t x_t'] \equiv M_{zx}$$

is the $(L \times K)$ matrix of product-moments of z_t with x_t, we will need this matrix to be of full column rank, i.e.

$$rank(M_{zx}) = K.$$

This *rank condition*, which will be a sufficient condition for identification of the parameter vector β from the observable data, implies a weaker *order condition* which is a necessary condition for identification –

namely, that $L \geq K$, i.e., the number of instrumental variables must be at least as large as the number of regressors.

Models of Endogenous Regressors

Before turning to the details of estimation of the parameter β using observations on y_t, x_t, and the instrumental variables z_t, it may be informative to consider some examples of linear models with correlation between the regressors and error terms, and the ways that instrumental variables might be derived. While no general method for cooking up instruments is available, many models have "natural" instruments associated with them.

(1) *Autocorrelated Errors and Lagged Dependent Variables:* Here the linear model of interest is

$$y_t = x_t'\beta + \gamma y_{t-1} + \varepsilon_t,$$

where the error terms are assumed to be first-order autoregressive,

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

with u_t assumed to be i.i.d. with zero mean and variance σ_u^2, and are independent of $\{x_s\}$ for all s. Since

$$\begin{aligned} E(y_{t-1}\varepsilon_t) &= E(y_{t-1}(\rho\varepsilon_{t-1} + u_t)) \\ &= \rho E(y_{t-1}\varepsilon_{t-1}) \\ &= \rho E((\alpha + \beta x_{t-1} + \gamma_{t-2} + \varepsilon_{t-1})\varepsilon_{t-1}) \\ &= \rho\gamma E(y_{t-1}\varepsilon_{t-1}) + \rho E(\varepsilon_{t-1}), \end{aligned}$$

then assuming $\{x_t\}$ are stationary – implying the $\{y_t\}$ are stationary as well – it follows that

$$E(y_{t-1}\varepsilon_t) = E(y_{t-2}\varepsilon_{t-1}),$$

so that

$$\begin{aligned} E(y_{t-1}\varepsilon_t) &= \rho\gamma E(y_{t-1}\varepsilon_t) + \rho E(\varepsilon_t^2) \\ &= \rho E(\varepsilon_t^2)/(1 - \rho\gamma) \\ &= \rho\sigma_u^2/(1 - \rho^2)(1 - \rho\gamma), \end{aligned}$$

which is different from zero when $\rho \neq 0$. Thus, the regressor y_{t-1} is correlated with the error term ε_t. For instruments, we can use the current and lagged values of the regressors; for example, we can use $z_t = (x_t', x_{t-1}')'$ as instruments for the set of regressors $(x_t', y_{t-1})'$. Since the lagged values of the regressors, x_{t-1}, show up in the equation for y_{t-1}, it is easy to see that the IVs should be correlated with the right-hand side variables, and they are uncorrelated with the error terms ε_t by assumption. Of course, we could also use more lagged values, i.e., $z_t = (x_t, x_{t-1}, x_{t-2}, \ldots)$, but typically the extra lagged values of x_{t-1} aren't as highly correlated with y_{t-1}, and so might be less useful as IVs.

(2) *Omitted Variables:* In a "long regression" setting with for cross-section data,

$$\begin{aligned} y_i &= x_i'\beta + w_i'\gamma + u_i \\ &\equiv x_i'\beta + \varepsilon_i, \end{aligned}$$

with $\varepsilon_i \equiv w_i'\gamma + u_i$, we may only observe y_i and x_i, and only care about estimating β, but if x_i and the "missing regressors" w_i are correlated with x_i in the population, x_i and ε_i will be correlated in the "short regression" of y_i on x_i.

Often, it is hard to think of a convincing example where variables that are not included in the observable x_i (like geographic dummy variables) would be correlated with x_i but uncorrelated with w_i, so the instrumental variable strategy may problematic in this example. Occasionally, though, a "natural experiment" will provide a variable z_i which affects x_i directly, but clearly is independent of w_i. A well-known example in econometrics is J. Angrist's study of the effect of military service (a regressor in x_i that is possibly correlated with an unobserved "ability" variable w_i) on future earnings y_i. As an instrumental variable, Angrist used an indicator variable for whether individual i had a high or low draft lottery number during the Vietnam war years; this would clearly be correlated with military service but should be independent of individual unobserved ability. An earlier biostatistics study used the same instrumental variable to estimate the effect of military service on life expectancy. These and similar papers led to a paradigm shift toward "natural experiment" approaches to estimation of causal effects using microeconomic data over the past two decades.

(3) *Measurement Error:* M. Friedman's classic model for the permanent income hypothesis posits an individual consumption function as

$$y_i = \alpha + \beta p_i + u_i,$$

where y_i is "measured consumption" and p_i is "permanent income" for individual i. As Friedman noted, we don't see permanent income, but only "measured" income, which is assumed to be of the form

$$x_i = p_i + v_i,$$

where v_i is "transitory income" (just like u_i is "transitory consumption"). There are two sorts of assumptions on the observable data that yield instrumental variables here. In one approach, the "repeated measurement" approach, it is assumed that some other variable related to permanent income, such as financial wealth "w_i", is observed. Supposing that wealth is linearly related to permanent income,

$$w_i = \gamma + \delta p_i + \eta_i,$$

where the "transitory components" (u_i, v_i, η_i) are independent of p_i, and further supposing that

$$\begin{aligned} E(u_i) &= E(v_i) = E(\eta_i) = 0, \\ E(u_i\eta_i) &= E(v_i\eta_i) = 0 \end{aligned}$$

(i.e., the shock to wealth is uncorrelated with transitory consumption or income), then

$$y_t = \alpha + \beta x_t + \varepsilon_t,$$

with

$$\varepsilon_i = u_i + \beta v_i,$$

so that $E(x_i\varepsilon_i) \neq 0$ but

$$E(w_i\varepsilon_i) = E((\gamma + \delta p_i + \eta_i)(u_i + \beta v_i)) = 0$$

and

$$E(w_i x_i) = E((\gamma + \delta p_i + \eta_i)(p_i + v_i)) = \delta E(p_i^2) \neq 0,$$

provided $\delta \neq 0$ (i.e., w_i is really related to p_i). So $z_i = (1, w_i)'$ can be used as a vector of instrumental variables for $(1, x_i)'$.

A different solution, proposed by A. Zellner, assumes a "causal model" relating permanent income to observable variables w_i (including, say, financial wealth, education, work experience, etc.). Writing this model as

$$p_i = w_i'\delta + \eta_i,$$

where $(u_i, v_i, \varepsilon_i)$ is assumed independent of w_i with $E(u_i) = E(v_i) = E(\varepsilon_i) = 0$, etc. Then w_i is clearly correlated with p_i if $\delta \neq 0$,

$$E(w_i x_i) = E(w_i(w_i'\delta + v_i + \eta_i)) = E(w_i w_i')\delta \neq 0,$$

but

$$E(w_i \varepsilon_i) = E(w_i(u_i + \beta v_i)) = 0,$$

so we can use $z_i \equiv (1, w_i')'$ as instruments for $(1, x_i)'$. Though the "repeated measurement" and "causal model" approaches are quite different – in the former, $E(p_i \eta_i) = 0, E(w_i \eta_i) \neq 0$, and vice versa in the latter – we get similar instruments in either case.

(3') *Rational Expectations Models (Measurement Error):* A variation on the previous measurement error model might be

$$y_t = \alpha + \beta E_t(x_{t+1}) + u_t,$$

with u_t i.i.d., independent of $\{x_s\}$, etc. For example, y_t might be "current investment," and x_t might be "current sales," so current investment would respond to current expectations of future sales. Assuming we observe only (y_t, x_t), where

$$x_{t+1} = E_t(x_{t+1}) + v_{t+1},$$

$$E(v_{t+1} | x_t, y_t, x_{t-1}, y_{t-1}, \ldots) = 0,$$

we can write

$$\begin{aligned} y_t &= \alpha + \beta x_{t+1} + \varepsilon_t, \\ \varepsilon_t &\equiv u_t + \beta v_{t+1}, \end{aligned}$$

and can use past values of y_t and current and past values of x_t as instrumental variables in this regression, i.e., $z_t = (1, x_t, y_{t-1}, x_{t-1}, y_{t-2}, \ldots)$.

(4) *Keynesian Cross:* In this shopworn example, discussed by virtually all introductory econometrics texts, an aggregate consumption function

$$c_t = \alpha + \beta y_t + \varepsilon_t$$

is paired with an income identity

$$y_t = c_t + i_t,$$

where c_t, y_t, and i_t are aggregate consumption, income, and "autonomous investment." Assuming

$$E[\varepsilon_t] = 0 = E(\varepsilon_t i_s)$$

(so that investment is determined by "animal spirits," and does not respond to consumption or income), we can take $z_t \equiv (1, i_t)'$ as IV's for $(1, y_t)'$ in estimating the consumption function. This is possibly the simplest example of a "simultaneous equations model," in which the variables c_t and y_t appearing in one equation (the consumption function) are jointly determined as the solution of a system of equations, rather than the left-hand variable (here, c_t) being determined by the "independent" causal effects of the right-hand-side regressors and errors.

(4') *Supply and Demand Model:* In another simple canonical model of simultaneity, quantity q_t and price p_t are determined to equate supply and demand in a particular market. Writing the demand function as

$$q_t = \alpha + \beta p_t + \gamma y_t + u_t,$$

it is assumed that some other variable y_t besides price (say, aggregate income) shifts the demand function, but does not affect supply. Similarly, the inverse supply function is assumed to be

$$p_t = \delta + \phi q_t + \psi w_t + v_t,$$

where some "supply shifter" variable w_t (e.g., "weather") is included in the inverse supply function but is excluded from the demand equation. Assuming (y_t, w_t) are *exogenous*, i.e.,

$$E \begin{bmatrix} y_t u_t & y_t v_t \\ w_t u_t & w_t v_t \end{bmatrix} = 0,$$

then $z_t = (1, y_t, w_t)'$ will be uncorrelated with (u_t, v_t), assuming $E(u_t) = E(v_t) = 0$; if $\psi \neq 0$, $E(w_t p_t) \neq 0$, and if $\gamma \neq 0$, $E(q_t y_t) \neq 0$, so we can use $z_t = (1, y_t, w_t)'$ as instrumental variables for $x_t^d \equiv (1, p_t, y_t)'$ in the demand equation and for $x_t^s \equiv (1, q_t, w_t)'$ in the inverse supply equation.

Just-Identification and Instrumental Variables Estimation

The foregoing examples illustrate how L-dimensional instrumental variables z_t satisfying the conditions $E[z_t \varepsilon_t] = M_{z\varepsilon} = 0$ and $rank(E[z_t x_t']) = rank(M_{zx}) = K$ can be obtained in certain applications. Now,

assuming such variables exist, we turn to the question of what to do with them, i.e., how to use them to get consistent estimators of β. In the special case where $L = K$ (known as the "just identified" case, since the order condition for identification is barely satisfied), we can define the *instrumental variables* (IV) *estimator* of β for the linear model as

$$\hat{\beta}_{IV} \equiv (Z'X)^{-1}Z'y = \left[\frac{1}{T}\sum_t z_t x_t'\right]^{-1} \cdot \left[\frac{1}{T}\sum_t z_t y_t\right],$$

which is generally well-defined because $Z'X$ is a square $(K \times K)$ matrix. This is an obvious generalization of the classical least squares estimator $\hat{\beta}_{LS}$, with "Z'" replacing "X'" in that formula throughout. It is easy to show consistency of this estimator if the conditions $M_{z\varepsilon} = 0$ and $\text{rank}(M_{zx}) = K$ are satisfied; again writing the IV estimator $\hat{\beta}$ in terms of the true parameter and error terms,

$$\hat{\beta}_{IV} = \beta + (\frac{1}{T}Z'X)^{-1}(\frac{1}{T}Z'\varepsilon),$$

a law of large numbers and Slutsky's theorem will imply that}

$$\text{plim } \left(\frac{1}{T}Z'X\right)^{-1} \equiv \text{plim } \left(\frac{1}{T}\sum_t z_t x_t'\right)^{-1} = M_{zx}^{-1}$$

and

$$\text{plim } \frac{1}{T}Z'\varepsilon = \text{plim } \frac{1}{T}\sum_t z_t\varepsilon_t = \lim \frac{1}{T}\sum_t E(z_t\varepsilon_t) \equiv M_{z\varepsilon} = 0,$$

so

$$\hat{\beta}_{IV} \to^p \beta + (M_{zx})^{-1}M_{z\varepsilon} = \beta,$$

as required for consistency.

This instrumental variables estimator can be interpreted as a *method of moments* estimator, because $\hat{\beta}_{IV}$ solves a variant of the normal equations, the *IV estimating equations*

$$0 = \frac{1}{T}Z'(y - X\hat{\beta}_{IV}) = \frac{1}{T}\sum_t z_t(y_t - x_t'\hat{\beta}_{IV}),$$

which correspond to the correct population moment conditions

$$E[z_t(y_t - x_t'\beta)] = E(z_t\varepsilon_t) \equiv M_{z\varepsilon} = 0.$$

In effect, the IV estimator sets the sample covariance of the instrumental variables and residuals equal to zero (assuming a constant term is included in the set of instruments).

Assuming some central limit theorem can be invoked to show that

$$\frac{1}{\sqrt{T}} Z'\varepsilon \equiv \frac{1}{\sqrt{T}} \sum_t z_t \varepsilon_t \rightarrow^d \mathcal{N}(0, V_0)$$

for some matrix V_0 (with $V_0 = E[\varepsilon_t^2 z_t z_t']$ if the original observations on y_t and x_t are i.i.d.), then it is easy to show that the IV estimator is approximately normally distributed,

$$\sqrt{T}(\hat{\beta} - \beta_0) = (\frac{1}{T} Z'X)^{-1} \frac{1}{\sqrt{T}} Z'\varepsilon \rightarrow^d \mathcal{N}(0, M_{zx}^{-1} V_0 M_{xz}^{-1}),$$

where

$$M_{xz} \equiv M_{zx}' = \text{plim } \frac{1}{T} \sum_t x_t z_t' \equiv \text{plim } \hat{M}_{xz}.$$

A consistent estimator of M_{zx} is, of course, $\hat{M}_{zx}$; with a consistent estimator $\hat{V}$ of V_0 using either the Eicker-White (for serially uncorrelated data) or Newey-West (for serially correlated data) methods applied to $(y_t - x_t'\hat{\beta}_{IV})z_t$, large-sample confidence regions and hypothesis tests can be constructed using normal sampling theory.

Overidentification and Two-Stage Least Squares (2SLS)

If there are more instrumental variables than regressors, i.e., $L > K$ (called the "overidentified" case), this method-of-moments approach must be generalized, since the IV estimating equations would require solution of an overdetermined system of linear equations (i.e., more equations than unknown $\hat{\beta}$ components) which will not exist except in rare (probability zero) cases. Of course, we could always "throw out" some of the "extra" instruments to make $L = K$. A more general strategy is to premultiply the L-vector z_t by some $(K \times L)$ matrix $\hat{\Pi}'$ – which could, in general, be estimated (random) – then use the K-vector $\hat{\Pi}' z_t$ as instruments. A special case would be $\hat{\Pi}' = (I_K, 0)$, which would delete the last $L - K$ components of z_t; in general, we will require that the square matrix $\hat{\Pi}' Z'X$ will have full rank K, which implies that the rank of $\hat{\Pi}$ must be K (with probability one). Since the dimension of $Z\hat{\Pi}$ is the same as for X, namely, $N \times K$, we can just substitute $Z\hat{\Pi}$ into the IV estimator formula to define a *generalized instrumental variable* or *generalized method of moments* estimator of β:

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

whose asymptotic distribution will depend on the particular sequence of $\hat{\Pi}$ matrices used, except in the just-identified case $L = K$, where this formula reduces to the previous IV formula (because the $K \times K$

matrix $\hat{\Pi}$ must then be invertible by assumption). Assuming

$$\hat{\Pi} \rightarrow^p \Pi$$

for some fixed, full-rank $(L \times K)$ matrix Π, and assuming that $\frac{1}{T}Z'X \rightarrow^p M_{zx}$ and $\frac{1}{\sqrt{T}}Z'\varepsilon \rightarrow^d \mathcal{N}(0, V_0)$ as before, then it is easy to verify that the asymptotic distribution of the GIV estimator is

$$\sqrt{T}(\hat{\beta} - \beta) \rightarrow^d \mathcal{N}(0, [\Pi' M_{zx}]^{-1}(\Pi' V_0 \Pi)[M_{xz}\Pi]^{-1}),$$

which depends on Π. As before, we can consistently estimate this asymptotic covariance matrix using either the Eicker-White or Newey-West approaches, depending on whether the data are serially- correlated.

The traditional choice of the "combination" matrix $\hat{\Pi}$ is

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

so that $\hat{\Pi}$ is the $(L \times K)$ matrix of least-squares coefficients for the regression of the columns of X on the matrix Z, which should certainly yield a matrix $Z\hat{\Pi}$ of fitted values of this regression which will be "correlated" with X. The resulting estimator of β is called the *two-stage least squares* (2SLS) *estimator*, and has the algebraic form

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

$$= (\hat{X}'X)^{-1}\hat{X}'y,$$

where $\hat{X} = Z\hat{\Pi}$ is the matrix of predicted values of X from the regression of X on Z. In this procedure, we first fit a "reduced form" regression for x_t in the first stage,

$$x_t = \Pi' z_t + v_t,$$

and use the fitted values as IVs in the second-stage regression. By a law of large numbers, we expect that

$$\hat{\Pi} \rightarrow^p \Pi_0 \equiv \{E(z_t z_t')\}^{-1}E(z_t x_t) \equiv M_{zz}^{-1}M_{zx};$$

substituting this into the previous formula for the asymptotic distribution of the GIV estimator gives a mighty unwieldy general formula for the asymptotic distribution of the 2SLS estimator,

$$\sqrt{T}(\hat{\beta} - \beta) \rightarrow^d \mathcal{N}(0, [M_{xz}M_{zz}^{-1}M_{zx}]^{-1}(M_{xz}M_{zz}^{-1}V_0 M_{zz}^{-1}M_{zx})[M_{xz}M_{zz}^{-1}M_{zx}]^{-1}).$$

If the error terms ε_t are assumed to be independent of z_t, as was traditionally assumed, this expression simplifies quite a bit; then

$$V_o = E[\varepsilon_t^2 z_t z_t'] = E[\varepsilon_t^2] \cdot E[z_t z_t'] = \sigma^2 M_{zz},$$

so that

$$\sqrt{T}(\hat{\beta} - \beta) \rightarrow^d \mathcal{N}(0, \sigma^2 [M_{xz} M_{zz}^{-1} M_{zx}]^{-1})$$

when the errors are independent of the instruments. The constant variance σ^2 could be estimated by the sample average of the squared values of the residuals $e_t = y_t - x_t' \hat{\beta}_{2SLS}$, as for the classical regression model, and

$$\text{plim } \frac{1}{T} \hat{X}' \hat{X} = M_{xz} M_{zz}^{-1} M_{zx}.$$

Because $Z(Z'Z)^{-1}Z'$ is an idempotent (projection) matrix, there are several other algebraically equivalent ways of writing the 2SLS estimator. In Theil's interpretation of the 2SLS estimator,

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = [(X'Z(Z'Z)^{-1}Z')(Z(Z'Z)^{-1}Z'X)]^{-1}\hat{X}'y,$$

so β is estimated by getting the fitted values of X on Z, then regressing y on $\hat{X}$ using the classical least squares formula. Another interpretation, suggested by Basmann, uses the fact that

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

$$= (X'P_{zz}X)^{-1}X'P_{zz}y = (\hat{X}'X)^{-1}\hat{X}'y,$$

for $P_{zz} \equiv Z(Z'Z)^{-1}Z'$ and $\hat{y} = P_{zz}y$ the vector of fitted values of the least-squares regression of y on Z. Other interpretations of 2SLS, some to be found on homework problems, exist.

Generalized Method of Moments (GMM)

In the general case where the errors are heteroskedastic and/or serially correlated, so that $V_0 \neq \sigma^2 M_{zz}$, the 2SLS estimator will not have the smallest asymptotic covariance matrix. To obtain an efficient estimator, we want to choose the combination matrix $\Pi = \text{plim } \hat{\Pi}$ to minimize the asymptotic covariance matrix

$$AV(\hat{\beta}_{GIV}(\Pi)) = [\Pi' M_{zx}]^{-1} \Pi' V_0 \Pi [M_{xz}\Pi]^{-1}$$

of the generalized IV estimator (in the matrix sense). The solution can be found by transforming the model to an asymptotic version of the Generalized Regression model, and then finding the appropriate GLS

estimator for the transformed model. First, premultiply the linear model $y = X\beta + \varepsilon$ by the normalized matrix $(1/\sqrt{T})Z'$, so that

$$\begin{aligned}\tilde{y} &\equiv \frac{1}{\sqrt{T}}Z'y \\ &= \left(\tfrac{1}{\sqrt{T}}Z'X \right)\beta + \left(\tfrac{1}{\sqrt{T}}Z'\varepsilon \right) \\ &\equiv \tilde{X}\beta + \tilde{\varepsilon},\end{aligned}$$

which defines a new linear model relating the L-dimensional vector $\tilde{y}$ to the $(L \times K)$ matrix $\tilde{X}$ of transformed regressors. By the Central Limit Theorem and the assumption $M_{z\varepsilon} = 0 = E[\tilde{\varepsilon}]$, we know that

$$\tilde{\varepsilon} \to \mathcal{N}(0, V_0),$$

and we can also show that the asymptotic covariance of $\tilde{X}$ and $\tilde{\varepsilon}$ is zero, so for large T the transformed variables $\tilde{y}$ and $\tilde{X}$ obey the Generalized Regression model (with approximately normal errors, no less!). The appropriate (infeasible) GLS estimator of β for this model is

$$\begin{aligned}\tilde{\beta}_{GLS} &\equiv (\tilde{X}'V_0^{-1}\tilde{X})^{-1}\tilde{X}'V_0^{-1}\tilde{y} \\ &= \left(\tfrac{1}{\sqrt{T}}X'ZV_0^{-1}\tfrac{1}{\sqrt{T}}Z'X \right)^{-1} \left(\tfrac{1}{\sqrt{T}}X'ZV_0^{-1}\tfrac{1}{\sqrt{T}}Z'y \right) \\ &= [X'ZV_0^{-1}Z'X]^{-1}X'ZV_0^{-1}Z'y \\ &\equiv \hat{\beta}_{GMM},\end{aligned}$$

which is called the *optimal generalized method of moments* estimator. This is in the form of a generalized IV estimator, with combination matrix of the form.

$$\hat{\Pi}^* = V_0^{-1}(\frac{1}{T}Z'X).$$

To construct a feasible version of this estimator, we would need a consistent estimator of

$$\begin{aligned}V_0 &= AsyVar(\tilde{\varepsilon}) \\ &= AsyVar\left(\frac{1}{\sqrt{T}}Z'\varepsilon\right),\end{aligned}$$

which we could obtain by applying the Eicker-White or Newey-West procedures to the residuals from a 2SLS fit of y on X using Z as instruments. For either the infeasible or feasible versions of this GMM estimator, the asymptotic distribution will be normal, and of the form

$$\sqrt{T}(\hat{\beta}_{GMM} - \beta) \rightarrow^d \mathcal{N}(0, [M_{xz} V_0^{-1} M_{zx}]^{-1}),$$

which reduces to the asymptotic distribution of 2SLS in the special case that ε is independent of Z. In general, though, we can show that the asymptotic covariance matrix of 2SLS will be at least as large (in the matrix sense) as that for GMM using the same arguments that show that the asymptotic variance of LS exceeds that of GLS for the Generalized Regression model.

Many variations on 2SLS and GMM exist. For example, we may have a system of equations

$$y_j = X_j \beta_j + \varepsilon_j, \qquad j = 1, \ldots, J,$$

with contemporaneous correlation across the components of ε_j over j. A combination of 2SLS and Zellner's Seemingly Unrelated Equations (SUR) estimation method yield something known as *three-stage least squares* (3SLS), which is efficient if the error terms are normal and independent of Z; otherwise, the GMM and SUR approaches can be combined to get more efficient joint estimators of all the unknown coefficient vectors $\{\beta_j\}$. The IV and GMM approaches are also naturally applicable to nonlinear systems of equations, but the estimators do not have nice closed-form expressions for such models, and the derivation of their asymptotic properties is much more complicated.