Generalized Method of Moments and Optimal Instruments

Class Notes

Manuel Arellano

July 1992

Revised: February 26, 2009

A Generalized Method of Moments Estimation

Part A reviews the basic estimation theory of the generalized method of moments (GMM) and Part B deals with optimal instrumental variables.[1] For the most part, we restrict attention to *iid* observations.

A.1 Method of Moment Estimation Problems

Linear Regression Economists often use linear regression to quantify a relationship between economic variables. A linear regression between y and x is a relationship of the form

$$y = x'\beta + \varepsilon \tag{A.1}$$

where β and ε are chosen in such a way that ε is uncorrelated with x (which typically includes a constant term). Thus, the parameter vector β satisfies

$$E\left[x\left(y - x'\beta\right)\right] = 0. \tag{A.2}$$

If $E\left(xx'\right)$ has full rank then (A.2) has a unique solution

$$\beta = \left[E\left(xx'\right)\right]^{-1} E\left(xy\right). \tag{A.3}$$

An important property of linear regression is that it is an optimal predictor of y given x in the following sense:

$$\beta = \arg\min_b E\left[\left(y - x'b\right)^2\right]. \tag{A.4}$$

That is, it minimizes the expected squared linear prediction error. This is why $x'\beta$ is called a "best linear predictor" or "linear projection".

Moreover, if the conditional expectation of y given x is linear it turns out that it coincides with the linear predictor. If on the other hand $E\left(y \mid x\right)$ is nonlinear, the linear projection is an optimal approximation to it in the sense that

$$\beta = \arg\min_b E\left\{\left[E\left(y \mid x\right) - x'b\right]^2\right\}. \tag{A.5}$$

[1] Published in Arellano (2003) as Appendix A and Appendix B, respectively.

This is why sometimes the notation $E^*(y \mid x) = x'\beta$ is used, which emphasizes the proximity of the linear projection and conditional expectation concepts (e.g. Goldberger, 1991).[2]

Therefore, β is a useful quantity if we are interested in a linear prediction of y given x, or if we are interested in studying how the mean of y changes for different values of x, and we think that $E(y \mid x)$ is linear or approximately linear.

Linear regression may also be of interest as a structural or causal relationship between y and x if we have a priori reasons to believe that the unobservable determinants of y are uncorrelated with x.

Instrumental Variables If we are interested in a structural relationship between y, x, and an unobservable variable u

$$y = x'\delta + u, \tag{A.6}$$

such that u is correlated with at least some of the components of x, clearly $\delta \neq \beta$ in general.

In many situations of interest in econometrics, δ can be regarded as the solution to moment equations of the form

$$E\left[z\left(y - x'\delta\right)\right] = 0 \tag{A.7}$$

where z is a vector of instrumental variables that a priori can be assumed to be uncorrelated with u.[3]

If $E(zx')$ has full rank, the system of equations (A.7) has a unique solution. Moreover, if z and x are of the same dimension

$$\delta = \left[E\left(zx'\right)\right]^{-1} E(zy). \tag{A.8}$$

Examples of (A.7) include an equation from the classical supply and demand simultaneous system, and a regression model with measurement errors in the regressors.

Equations (A.2) and (A.7) can be described as "moment problems" because the parameters of interest solve moment equations.

The Analogy Principle According to the analogy principle, given a representative sample $\{y_i, x_i, z_i\}_{i=1}^N$, we choose as a candidate estimator for a population characteristic, the same characteristic defined in the sample (Manski, 1988). In this way, the sample linear regression coefficients solve

$$\frac{1}{N}\sum_{i=1}^{N} x_i\left(y_i - x_i'\widehat{\beta}\right) = 0 \tag{A.9}$$

[2] Nevertheless, whereas $E(y \mid x)$ is a characteristic of the conditional distribution of y given x, $E^*(y \mid x)$ is a characteristic of the joint distribution. That is, if we keep constant the distribution of $y \mid x$ but change the marginal distribution of x, $E(y \mid x)$ remains constant while $E^*(y \mid x)$ changes unless $E^*(y \mid x) = E(y \mid x)$.

[3] The vectors x and z may have elements in common. For example, a constant term.

giving rise to the standard OLS formula:

$$\widehat{\beta} = \left(\sum_{i=1}^{N} x_i x_i' \right)^{-1} \sum_{i=1}^{N} x_i y_i. \tag{A.10}$$

Similarly, the simple instrumental variable estimator, with as many instruments as explanatory variables, solves

$$\frac{1}{N} \sum_{i=1}^{N} z_i \left(y_i - x_i' \widehat{\delta} \right) = 0, \tag{A.11}$$

yielding

$$\widehat{\delta} = \left(\sum_{i=1}^{N} z_i x_i' \right)^{-1} \sum_{i=1}^{N} z_i y_i. \tag{A.12}$$

Generalized Moment Problems Suppose now that the number of instruments in z exceeds the number of explanatory variables in x. Let z and x be of orders r and k, respectively, and let $d = (y, x')'$. If we assume that $r > k$, the truth of (A.7) requires that the $r \times (k+1)$ matrix $E(zd')$ has reduced rank k. Otherwise it could not be the case that

$$E(zu) = E\left(zd'\right) \begin{pmatrix} 1 \\ \delta \end{pmatrix} = 0, \tag{A.13}$$

and at least some of the moment conditions would not hold.

However, even if $E(zd')$ has reduced rank in the population, its sample counterpart

$$\frac{1}{N} \sum_{i=1}^{N} z_i d_i'$$

will not have reduced rank in general because of sample error. Therefore, there will be no single value $\widehat{\delta}$ that satisfies the r equations (A.11), and different estimates of δ will be obtained from the solution to different subsets of k equations.

This situation modifies the nature of the estimation problem and makes assertion (A.7) empirically refutable. Following Sargan (1958), we consider estimators that solve k linear combinations of the r sample moment equations (A.11):

$$\frac{1}{N} \Gamma_N \sum_{i=1}^{N} z_i \left(y_i - x_i' \widehat{\delta} \right) = 0 \tag{A.14}$$

for an optimal choice of the $k \times r$ matrix of coefficients Γ_N. Moreover, we can test the overidentifying restrictions by testing whether the rank of the matrix $N^{-1} \sum_{i=1}^{N} z_i d_i'$ is significantly greater than k.

These issues are addressed in the following section in the context of a more general class of moment problems.

A.2 General Formulation

We consider parameters that are defined by a set of moment equations (or orthogonality conditions) of the form

$$E\psi(w,\theta)=0 \tag{A.15}$$

where:

w is a $p\times 1$ random vector,

ψ is a $r\times 1$ vector of functions,

θ is a $k\times 1$ vector of parameters such that $k\leq r$,

Θ is the parameter space (set of admissible values of θ).

We have a sample of N observations $(w_1,...,w_N)$ and estimation of θ is based on the sample counterpart of (A.15) given by

$$b_N(c)=\frac{1}{N}\sum_{i=1}^{N}\psi(w_i,c). \tag{A.16}$$

We choose as an estimator of θ the value of c that minimizes the quadratic distance of $b_N(c)$ from zero:

$$\begin{aligned}\widehat{\theta} &= \arg\min_{c\in\Theta}\left(\frac{1}{N}\sum_{i=1}^{N}\psi(w_i,c)\right)' A_N\left(\frac{1}{N}\sum_{i=1}^{N}\psi(w_i,c)\right) \\ &= \arg\min_{c\in\Theta} s(c)\end{aligned} \tag{A.17}$$

where A_N is an $r\times r$, possibly random, non-negative definite weight matrix, whose rank is greater than or equal to k. The statistic $\widehat{\theta}$ is a GMM estimator of θ.

If the problem is just identified we have that $r=k$, the weight matrix is irrelevant, and $\widehat{\theta}$ solves

$$b_N\left(\widehat{\theta}\right)=0. \tag{A.18}$$

A.3 Examples: 2SLS and 3SLS

Two Stage Least-Squares (2SLS) Let us consider the single equation model

$$y_i=x_i'\theta+u_i \tag{A.19}$$

together with the assumption

$$E(z_iu_i)=0 \tag{A.20}$$

where z_i is an $r\times 1$ vector of instruments and $r>k$. Thus, in this example $w_i=(y_i,x_i',z_i')'$, $\psi(w_i,\theta)=z_i(y_i-x_i'\theta)$, and the sample moment conditions are

$$b_N(c)=\frac{1}{N}\sum_{i=1}^{N}z_i\left(y_i-x_i'c\right)=\frac{1}{N}Z'(y-Xc) \tag{A.21}$$

where we are using the notation $y = (y_1, ..., y_N)'$, $X = (x_1, ..., x_N)'$, and $Z = (z_1, ..., z_N)'$. This is the example that was used in the introductory section. Since $r > k$ there is no solution to the system $b_N(c) = 0$.

The 2SLS estimator of θ minimizes the GMM objective function

$$b_N(c)' A_N b_N(c) = N^{-2} (y - Xc)' Z A_N Z' (y - Xc) \tag{A.22}$$

for $A_N = (Z'Z/N)^{-1}$. This choice of weight matrix will be motivated later in the GMM context. Let us note now that since the first-order conditions from (A.22) are

$$X'Z A_N Z' (y - Xc) = 0, \tag{A.23}$$

the form of the 2SLS estimator is

$$\widehat{\theta}_{2SLS} = \left[X'Z \left(Z'Z\right)^{-1} Z'X\right]^{-1} X'Z \left(Z'Z\right)^{-1} Z'y. \tag{A.24}$$

Moreover, this equals

$$\widehat{\theta}_{2SLS} = \left(\widehat{X}'\widehat{X}\right)^{-1} \widehat{X}'y \tag{A.25}$$

where $\widehat{X}$ is the matrix of fitted values in a multivariate regression of X on Z:

$$\widehat{X} = Z\widehat{\Pi}' \tag{A.26}$$

with $\widehat{\Pi} = X'Z (Z'Z)^{-1}$.

Thus, following the classic interpretation of 2SLS that justifies its name, in the first stage X is regressed on Z to obtain $\widehat{X}$, whereas in the second stage we obtain $\widehat{\theta}_{2SLS}$ as a regression of y on $\widehat{X}$.

Note also that we have

$$\widehat{\theta}_{2SLS} = \left(\widehat{X}'X\right)^{-1} \widehat{X}'y. \tag{A.27}$$

That is, the 2SLS estimator can also be interpreted as the simple IV estimator that uses $\widehat{X}$ as instrument. Specifically, it solves the k moment equations:

$$\sum_{i=1}^{N} \widehat{x}_i \left(y_i - x_i'\widehat{\theta}_{2SLS}\right) = 0 \tag{A.28}$$

where $\widehat{x}_i = \widehat{\Pi} z_i$. So, 2SLS uses as instruments the linear combinations of the z_i that best predict x_i.

Three Stage Least-Squares (3SLS) We turn to consider a system of g equations

$$\begin{aligned} y_{1i} &= x_{1i}'\theta_1 + u_{1i} \\ &\vdots \\ y_{gi} &= x_{gi}'\theta_g + u_{gi} \end{aligned} \tag{A.29}$$

whose errors are orthogonal to a common $r_0 \times 1$ vector of instruments z_i. Thus, in this example there are $r = gr_0$ moment conditions given by

$$\begin{aligned} E(z_i u_{1i}) &= 0 \\ &\vdots \\ E(z_i u_{gi}) &= 0. \end{aligned} \tag{A.30}$$

Convenient compact notations for these moments are:

$$E(u_i \otimes z_i) \equiv E\left(Z_i' u_i\right) \equiv E\left[Z_i'(y_i - X_i\theta)\right] = 0 \tag{A.31}$$

where $u_i = (u_{1i}, ..., u_{gi})'$, $Z_i = I_g \otimes z_i'$, $y_i = (y_{1i}, ..., y_{gi})'$, $\theta = \left(\theta_1', ..., \theta_g'\right)'$, and

$$X_i = \begin{pmatrix} x_{1i}' & & 0 \\ & \ddots & \\ 0 & & x_{gi}' \end{pmatrix}.$$

Accordingly, the sample orthogonality conditions are

$$\begin{aligned} b_N(c) &= \frac{1}{N}\sum_{i=1}^{N} Z_i'(y_i - X_i c) = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} z_i(y_{1i} - x_{1i}'c_1) \\ \vdots \\ z_i\left(y_{gi} - x_{gi}'c_g\right) \end{pmatrix} \\ &= \frac{1}{N}\begin{pmatrix} Z'(y_1 - X_1 c_1) \\ \vdots \\ Z'(y_g - X_g c_g) \end{pmatrix} = \frac{1}{N}\left(I_g \otimes Z'\right)(y - Xc) \end{aligned} \tag{A.32}$$

where $Z = (z_1, ..., z_N)'$ is an $N \times r_0$ matrix similar to that used in the 2SLS example, and we analogously define $y_1, ..., y_g$ and $X_1, ..., X_g$. Moreover, $y = \left(y_1', ..., y_g'\right)'$ and X is a block diagonal matrix with blocks $X_1, ..., X_g$.

The 3SLS estimator of θ minimizes the GMM criterion

$$b_N(c)' A_N b_N(c)$$

with weight matrix given by

$$A_N = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\widehat{\Omega} Z_i\right)^{-1} = \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{\Omega} \otimes z_i z_i'\right)^{-1} = N\left(\widehat{\Omega} \otimes Z'Z\right)^{-1} \tag{A.33}$$

where $\widehat{\Omega}$ is the 2SLS residual covariance matrix:

$$\widehat{\Omega} = \frac{1}{N}\sum_{i=1}^{N} \widehat{u}_i \widehat{u}_i' \tag{A.34}$$

and $\widehat{u}_i = y_i - X_i\widehat{\theta}_{2SLS}$.

Therefore:

$$\widehat{\theta}_{3SLS} = \left[\left(\sum_i X_i'Z_i\right) A_N \left(\sum_i Z_i'X_i\right)\right]^{-1} \left(\sum_i X_i'Z_i\right) A_N \left(\sum_i Z_i'y_i\right) \tag{A.35}$$

or

$$\widehat{\theta}_{3SLS} = \left[X'\left(\widehat{\Omega}^{-1} \otimes Z\left(Z'Z\right)^{-1}Z'\right)X\right]^{-1} X'\left(\widehat{\Omega}^{-1} \otimes Z\left(Z'Z\right)^{-1}Z'\right)y. \tag{A.36}$$

Moreover, in parallel with the earlier development for 2SLS, the 3SLS formula can be written as

$$\widehat{\theta}_{3SLS} = \left(\sum_i \widehat{X}_i'\widehat{\Omega}^{-1}\widehat{X}_i\right)^{-1} \sum_i \widehat{X}_i'\widehat{\Omega}^{-1}y_i \tag{A.37}$$

where $\widehat{X}_i$ is a block diagonal matrix with blocks $\widehat{x}_{1i}', ..., \widehat{x}_{gi}'$ and

$$\widehat{x}_{ji} = \widehat{\Pi}_j z_i \ \ (j = 1, ..., g) \tag{A.38}$$

with $\widehat{\Pi}_j = \sum_i x_{ji}z_i' \left(\sum_i z_i z_i'\right)^{-1}$.

Expression (A.37) corresponds to the interpretation of 3SLS on which its name is based. Namely, the first two stages coincide with those of 2SLS for each of the g equations, whereas in the third stage we obtain $\widehat{\theta}_{3SLS}$ as GLS of y_i on $\widehat{X}_i$ weighted by the inverse of $\widehat{\Omega}$. Note that replacing $\widehat{\Omega}$ by an identity matrix in (A.37) we obtain a compact expression for the 2SLS estimators of all the θ.

Finally, we also have

$$\widehat{\theta}_{3SLS} = \left(\sum_i \widehat{X}_i'\widehat{\Omega}^{-1}X_i\right)^{-1} \sum_i \widehat{X}_i'\widehat{\Omega}^{-1}y_i, \tag{A.39}$$

so that $\widehat{\theta}_{3SLS}$ can also be interpreted as a simple IV estimator of the full system that uses $\widehat{\Omega}^{-1}\widehat{X}_i$ as instrument and solves the moment conditions

$$\sum_{i=1}^{N} \widehat{X}_i'\widehat{\Omega}^{-1}\left(y_i - X_i\widehat{\theta}_{3SLS}\right) = 0. \tag{A.40}$$

A.4 Consistency of GMM Estimators

A general method for establishing consistency in the case of estimators that minimize a continuous function is provided by the following theorem of Amemiya (1985). Precursors of this type of theorem were employed by Sargan (1959) in his analysis of the asymptotic properties of nonlinear in parameters IV estimators, and by Jennrich (1969) and Malinvaud (1970) in their proofs of the consistency of nonlinear least-squares. A comprehensive discussion can be found in Newey and McFadden (1994).

Consistency Theorem Let us assume the following:

(a) The parameter space Θ is a compact subset of R^k such that $\theta \in \Theta$.

(b) The estimation criterion $s_N(w_1, ..., w_N, c) = s_N(c)$ is a continuous function in $c \in \Theta$ for all $(w_1, ..., w_N)$.

(c) $s_N(c)$ converges uniformly in probability to a nonstochastic function $s_\infty(c)$, which has a unique minimum at $c = \theta$, i.e.:

$$\sup_{\Theta} |s_N(c) - s_\infty(c)| \xrightarrow{p} 0 \text{ as } N \to \infty. \tag{A.41}$$

Let us define an estimator $\widehat{\theta}_N$ as the value that minimizes $s_N(c)$:

$$\widehat{\theta}_N = \arg\min_{\Theta} s_N(c). \tag{A.42}$$

Then $\text{plim}_{N\to\infty} \widehat{\theta}_N = \theta$. (Proof: See Amemiya, 1985, p. 107.)

Application to GMM Estimators In this case we have

$$s_N(c) = b_N(c)' A_N b_N(c) \tag{A.43}$$

with $b_N(c) = N^{-1} \sum_{i=1}^{N} \psi(w_i, c)$.

The previous theorem can be applied under the following assumptions:

1. Θ is compact, and $\psi(w, c)$ is continuous in $c \in \Theta$ for each w.
2. $A_N \xrightarrow{p} A_0$ and A_0 is positive semi-definite.
3. $E[\psi(w, c)]$ exists for all $c \in \Theta$ and $A_0 E[\psi(w, c)] = 0$ only if $c = \theta$ (identification condition).
4. $b_N(c)$ converges in probability uniformly in c to $E[\psi(w, c)]$.

Note that with these assumptions

$$s_\infty(c) = E[\psi(w, c)]' A_0 E[\psi(w, c)] \geq 0, \tag{A.44}$$

which has a unique minimum of zero at $c = \theta$.

Condition 4 states that $b_N(c)$ satisfies a uniform law of large numbers, and it ensures that the uniform convergence assumption of the theorem is satisfied. More primitive conditions for stationary or *iid* data are discussed by Hansen (1982), and Newey and McFadden (1994).

A.5 Asymptotic Normality

One way to establish the asymptotic normality of GMM estimators is to proceed as in the analysis of consistency. That is, to treat GMM as a special case within the class of estimators that minimize some objective function (*extremum* estimators). A general theorem for extremum estimators adapted from Amemiya (1985) is as follows.

Asymptotic Normality Theorem for Extremum Estimators Let us make the assumptions:

(a) We have $\widehat{\theta}_N = \arg\min_\Theta s_N(c)$ such that $\text{plim}_{N\to\infty} \widehat{\theta}_N = \theta$, where $s_N(c)$ has first and second derivatives in a neighbourhood of θ, and θ is an interior point of Θ.

(b) Asymptotic normality of the gradient:[4]

$$\sqrt{N}\frac{\partial s_N(\theta)}{\partial c} \xrightarrow{d} \mathcal{N}(0, \mathcal{W}) \tag{A.45}$$

(c) Convergence of the Hessian: for any $\widetilde{\theta}_N$ such that $\widetilde{\theta}_N \xrightarrow{p} \theta$

$$\frac{\partial^2 s_N\left(\widetilde{\theta}_N\right)}{\partial c \partial c'} \xrightarrow{p} H \tag{A.46}$$

where H is a non-singular non-stochastic matrix.

Then

$$\sqrt{N}\left(\widehat{\theta}_N - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}\mathcal{W}H^{-1}\right). \tag{A.47}$$

Proof. We can proceed as if $\widehat{\theta}_N$ were an interior point of Θ since consistency of $\widehat{\theta}_N$ for θ and the assumption that θ is interior to Θ implies that the probability that $\widehat{\theta}_N$ is not interior goes to zero as $N \to \infty$.

Using the mean value theorem (and writing $\widehat{\theta}$ for shortness):

$$0 = \frac{\partial s_N\left(\widehat{\theta}\right)}{\partial c_j} = \frac{\partial s_N(\theta)}{\partial c_j} + \sum_{\ell=1}^{k} \frac{\partial^2 s_N\left(\widetilde{\theta}_{[j]}\right)}{\partial c_j \partial c_\ell}\left(\widehat{\theta}_\ell - \theta_\ell\right) \quad (j = 1, ..., k) \tag{A.48}$$

where $\widehat{\theta}_\ell$ is the ℓ-th element of $\widehat{\theta}$, and $\widetilde{\theta}_{[j]}$ denotes a $k \times 1$ random vector such that $\left\|\widetilde{\theta}_{[j]} - \theta\right\| \leq \left\|\widehat{\theta} - \theta\right\|$. The expansion has to be made element by element since $\widetilde{\theta}_{[j]}$ may be different for each j.

[4] We use the notation $\partial s_N(\theta)/\partial c$ as an abbreviation of

$$\frac{\partial s_N(c)}{\partial c}_{|c=\theta}.$$

Note that $\widehat{\theta} \xrightarrow{p} \theta$ implies $\widetilde{\theta}_{[j]} \xrightarrow{p} \theta$. In view of assumption (c) this implies

$$\frac{\partial^2 s_N\left(\widetilde{\theta}_N\right)}{\partial c_j \partial c_\ell'} \xrightarrow{p} (j,\ell) \text{ element of } H. \tag{A.49}$$

Hence,

$$0 = \sqrt{N}\frac{\partial s_N(\theta)}{\partial c} + \left[H + o_p(1)\right]\sqrt{N}\left(\widehat{\theta} - \theta\right) \tag{A.50}$$

and

$$-H^{-1}\left[H + o_p(1)\right]\sqrt{N}\left(\widehat{\theta} - \theta\right) = H^{-1}\sqrt{N}\frac{\partial s_N(\theta)}{\partial c}. \tag{A.51}$$

Finally, using assumption (b) and Cramer's theorem the result follows. ■

Note that this theorem requires twice differentiability of the objective function. However, asymptotic normality for GMM can be easily proved when $b_N(c)$ only has first derivatives if we directly use the first-order conditions. An alternative specific result for GMM along these lines is as follows.

Asymptotic Normality Theorem for GMM We make the following assumptions in addition to those used for consistency:

5. θ is in the interior of Θ, and $\psi(w,c)$ is (once) continuously differentiable in Θ.

6. The quantity $D_N(c) = \partial b_N(c)/\partial c'$ converges in probability uniformly in c to a non-stochastic matrix $D(c)$, and $D(c)$ is continuous at $c = \theta$.

7. $\sqrt{N}b_N(\theta)$ satisfies a central limit theorem:

$$\sqrt{N}b_N(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(w_i,\theta) \xrightarrow{d} \mathcal{N}(0,V_0). \tag{A.52}$$

8. For $D_0 = D(\theta)$, $D_0'A_0D_0$ is non-singular.

Then

$$\sqrt{N}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}(0,W_0) \tag{A.53}$$

where W_0 is given by the sandwich formula

$$W_0 = \left(D_0'A_0D_0\right)^{-1}D_0'A_0V_0A_0D_0\left(D_0'A_0D_0\right)^{-1}. \tag{A.54}$$

Proof. The GMM estimator satisfies the first-order conditions

$$D_N'\left(\widehat{\theta}\right)A_Nb_N\left(\widehat{\theta}\right) = 0. \tag{A.55}$$

Moreover, in view of condition 6 and the consistency of $\widehat{\theta}$:

$$D_0' A_0 \sqrt{N} b_N \left(\widehat{\theta}\right) = o_p(1). \tag{A.56}$$

Next, using a first-order expansion, we have

$$D_0' A_0 \left[\sqrt{N} b_N(\theta) + D_N(\theta) \sqrt{N} \left(\widehat{\theta} - \theta\right)\right] = o_p(1). \tag{A.57}$$

Hence

$$\left(D_0' A_0 D_0\right) \sqrt{N} \left(\widehat{\theta} - \theta\right) = -D_0' A_0 \sqrt{N} b_N(\theta) + o_p(1) \tag{A.58}$$

and

$$\sqrt{N} \left(\widehat{\theta} - \theta\right) = -\left(D_0' A_0 D_0\right)^{-1} D_0' A_0 \sqrt{N} b_N(\theta) + o_p(1). \tag{A.59}$$

Finally, from the central limit theorem for $\sqrt{N} b_N(\theta)$ the result follows. ■

Note that the conditions of this result imply the first two conditions of the asymptotic normality theorem for extremum estimators but not the third one, which requires twice differentiability of $b_N(c)$. From a different angle, the theorem for extremum estimators can be regarded as a special case of a result based on GMM-like first-order conditions with estimating equation given by $\partial s_N\left(\widehat{\theta}\right)/\partial c = 0$ (cf. Hansen, 1982; Newey and McFadden, 1994).

As long as the relevant moments exist, V_0 is given by

$$\begin{aligned} V_0 &= \lim_{N\to\infty} Var\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(w_i, \theta)\right) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} E\left[\psi(w_i, \theta) \psi(w_j, \theta)'\right]. \end{aligned} \tag{A.60}$$

With independent observations, V_0 reduces to

$$V_0 = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E\left[\psi(w_i, \theta) \psi(w_i, \theta)'\right], \tag{A.61}$$

and with *iid* observations

$$V_0 = E\left[\psi(w_i, \theta) \psi(w_i, \theta)'\right]. \tag{A.62}$$

Given our focus on *iid* observations, in the sequel we assume that V_0 is given by (A.62).[5] Similarly, we take the $r \times k$ matrix D_0 to be given by

$$D_0 = E\left(\frac{\partial \psi(w_i, \theta)}{\partial c'}\right). \tag{A.63}$$

[5] Depending on the context we may suppress the i subscript for convenience and simply write $E\left[\psi(w, \theta) \psi(w, \theta)'\right]$.

A.6 Estimating the Asymptotic Variance

To obtain a consistent estimate of W_0 we just replace A_0, D_0, and V_0 in (A.54) by their sample counterparts A_N, $\widehat{D}$, and $\widehat{V}$, where the last two are given by

$$\widehat{D} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi\left(w_i,\widehat{\theta}\right)}{\partial c'} \tag{A.64}$$

$$\widehat{V} = \frac{1}{N}\sum_{i=1}^{N}\psi\left(w_i,\widehat{\theta}\right)\psi\left(w_i,\widehat{\theta}\right)'. \tag{A.65}$$

In this way we obtain

$$\widehat{W}_N = \left(\widehat{D}'A_N\widehat{D}\right)^{-1}\widehat{D}'A_N\widehat{V}A_N\widehat{D}\left(\widehat{D}'A_N\widehat{D}\right)^{-1}. \tag{A.66}$$

Thus, the concluding result from the discussion so far is

$$\left(\widehat{W}_N/N\right)^{-1/2}\left(\widehat{\theta}_N-\theta\right)\xrightarrow{d}\mathcal{N}\left(0,I\right), \tag{A.67}$$

which justifies the approximation of the joint distribution of the random vector $\left(\widehat{W}_N/N\right)^{-1/2}\left(\widehat{\theta}_N-\theta\right)$ by a $\mathcal{N}\left(0,I\right)$ when N is large.

The squared root of the diagonal elements of $\widehat{W}_N/N$ are the asymptotic standard errors of the components of $\widehat{\theta}_N$, and $\widehat{W}_N/N$ itself is sometimes referred to as the *estimated asymptotic variance matrix* of $\widehat{\theta}_N$.

Example: 2SLS with *iid* observations In this case we have

$$\psi\left(w_i,\theta\right)=z_i\left(y_i-x_i'\theta\right)=z_iu_i. \tag{A.68}$$

The expressions for A_0, D_0, and V_0 are given by

$$A_0 = \left[E\left(z_iz_i'\right)\right]^{-1} \tag{A.69}$$

$$D_0 = E\left(z_ix_i'\right) \tag{A.70}$$

$$V_0 = E\left(u_i^2z_iz_i'\right), \tag{A.71}$$

and their sample counterparts are

$$A_N = \left(\frac{1}{N}\sum_{i=1}^{N}z_iz_i'\right)^{-1}=N\left(Z'Z\right)^{-1} \tag{A.72}$$

$$\widehat{D} = \frac{1}{N}\sum_{i=1}^{N}z_ix_i'=\frac{1}{N}Z'X \tag{A.73}$$

$$\widehat{V} = \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2z_iz_i'=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-x_i'\widehat{\theta}_{2SLS}\right)^2z_iz_i'. \tag{A.74}$$

Hence, the estimated asymptotic variance matrix of 2SLS is

$$\begin{aligned}\widehat{W}_N/N &= \left(\widehat{X}'\widehat{X}\right)^{-1}(X'Z)\,(Z'Z)^{-1}\left(\sum_{i=1}^{N}\widehat{u}_i^2 z_i z_i'\right)(Z'Z)^{-1}\,(Z'X)\left(\widehat{X}'\widehat{X}\right)^{-1} \\ &= \left(\widehat{X}'\widehat{X}\right)^{-1}\left(\sum_{i=1}^{N}\widehat{u}_i^2\widehat{x}_i\widehat{x}_i'\right)\left(\widehat{X}'\widehat{X}\right)^{-1} \qquad \text{(A.75)}\end{aligned}$$

where as before $\widehat{X} = Z\,(Z'Z)^{-1}\,(Z'X) = Z\widehat{\Pi}'$ with rows $\widehat{x}_i = \widehat{\Pi}z_i$.

Homoskedastic Case Under conditional homoskedasticity u_i^2 is mean independent of z_i:

$$E\left(u_i^2 \mid z_i\right) = \sigma^2, \qquad \text{(A.76)}$$

in which case the variance matrix of the 2SLS moment conditions particularizes to:

$$V_0 \equiv E\left[E\left(u_i^2 \mid z_i\right) z_i z_i'\right] = \sigma^2 E\left(z_i z_i'\right). \qquad \text{(A.77)}$$

Hence, in this case

$$W_0 = \sigma^2\left\{E\left(x_i z_i'\right)\left[E\left(z_i z_i'\right)\right]^{-1}E\left(z_i x_i'\right)\right\}^{-1} = \sigma^2\left[E\left(\Pi z_i z_i'\Pi'\right)\right]^{-1} \qquad \text{(A.78)}$$

where $\Pi = E\left(x_i z_i'\right)\left[E\left(z_i z_i'\right)\right]^{-1}$.

Therefore, letting the 2SLS residual variance be

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2, \qquad \text{(A.79)}$$

an alternative consistent estimate of W_0 under homoskedasticity is

$$\widetilde{W}_N/N = \widehat{\sigma}^2\left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1} = \widehat{\sigma}^2\left(\widehat{X}'\widehat{X}\right)^{-1}, \qquad \text{(A.80)}$$

which is the standard formula for the 2SLS estimated asymptotic variance matrix.

Standard errors and tests of hypothesis calculated from (A.75) are robust to heteroskedasticity, whereas those calculated from (A.80) are not. However, $\widehat{W}_N$ depends on fourth-order moments of the data whereas $\widetilde{W}_N$ only depends on second-order moments. This fact may imply that under homoskedasticity, for given N, the quality of the asymptotic approximation is poorer for the robust statistics. So in practice there may be a finite-sample trade-off in the choice between (A.75) and (A.80).

A.7 Optimal Weight Matrix

So far we have not considered the problem of choosing the matrix A_N, but clearly given the form of the asymptotic variance of $\widehat{\theta}$, different choices of A_0 will give rise to GMM estimators with different precision (at least in large samples).

It is a matter of deciding which linear combinations of a given set of orthogonality conditions are optimal for the estimation of θ. This is a different question from the problem of constructing optimal orthogonality conditions from conditional moments, which will be considered in Appendix B. The following optimality result takes the specification of orthogonality conditions as given.

The result is that an optimal choice of A_N is such that A_0 is equal to V_0^{-1} up to an arbitrary positive multiplicative constant k:

$$A_0 = kV_0^{-1}. \tag{A.81}$$

For a GMM estimator with such A_0, the asymptotic covariance matrix is given by

$$\left(D_0'V_0^{-1}D_0\right)^{-1}. \tag{A.82}$$

We can prove that this is an optimal choice showing that for any other A_0:[6]

$$\left(D_0'A_0D_0\right)^{-1}D_0'A_0V_0A_0D_0\left(D_0'A_0D_0\right)^{-1} - \left(D_0'V_0^{-1}D_0\right)^{-1} \geq 0. \tag{A.83}$$

To see this, note that this difference is equivalent to

$$\overline{D}\left[I - H\left(H'H\right)^{-1}H'\right]\overline{D}' \tag{A.84}$$

where

$$\overline{D} = \left(D_0'A_0D_0\right)^{-1}D_0'A_0V_0^{1/2} \tag{A.85}$$

$$H = V_0^{-1/2}D_0. \tag{A.86}$$

Moreover, (A.84) is a positive semi-definite matrix since $\left[I - H\left(H'H\right)^{-1}H'\right]$ is idempotent.

Therefore, in order to obtain an optimal estimator we need a consistent estimate of V_0 up to scale. In general, this will require us to obtain a preliminary suboptimal GMM estimator, which is then used in the calculation of an estimate like (A.65).

Example: 2SLS Under conditional homoskedasticity, $V_0 = \sigma^2 E\left(z_iz_i'\right)$, in which case the 2SLS limiting weight matrix (A.69) is a multiple of V_0^{-1}, and therefore 2SLS is optimal. Moreover, a consistent estimate of the 2SLS asymptotic variance is given by (A.80).

However, if the conditional variance of u_i given z_i depends on z_i, then 2SLS is suboptimal in the GMM class, and $\widetilde{W}_N$ is not a consistent estimate of the asymptotic variance of 2SLS.

Under heteroskedasticity, we can still do valid asymptotic inference with 2SLS since $\widehat{W}_N$ in (A.75) remains a consistent estimate of W_0.

A two-step optimal GMM estimator is given by

$$\widetilde{\theta} = \left[X'Z\left(\sum_{i=1}^N \widehat{u}_i^2 z_iz_i'\right)^{-1}Z'X\right]^{-1}X'Z\left(\sum_{i=1}^N \widehat{u}_i^2 z_iz_i'\right)^{-1}Z'y. \tag{A.87}$$

[6] The weak inequality notation $B \geq 0$ applied to a matrix here denotes that B is positive semi-definite.

This estimator is of the same form as 2SLS but it replaces the 2SLS weight matrix $(Z'Z)^{-1}$ by the robust choice $\left(\sum_{i=1}^{N}\widehat{u}_i^2 z_i z_i'\right)^{-1}$ based on 2SLS residuals (cf. White, 1982).

Semi-parametric Asymptotic Efficiency We obtained (A.82) as the best asymptotic variance that can be achieved by an estimator within the GMM class. An interesting theoretical question is whether a different type of estimator based on the same information could be more efficient asymptotically than optimal GMM. The answer is that no additional efficiency gains are possible since, as shown by Chamberlain (1987), (A.82) is a semi-parametric information bound. That is, (A.82) is the best one can do if all that is known about the distribution of w is that it satisfies the moment restrictions in (A.15).

Chamberlain's argument proceeds as follows. Suppose that the w_i are *iid* observations with a multinomial distribution with known finite support given by $\{\xi_1, ..., \xi_q\}$ and corresponding probabilities $\pi_1, ..., \pi_q$. Suppose all that is known about these probabilities is that they add up to one

$$\sum_{j=1}^{q}\pi_j = 1 \tag{A.88}$$

and that they satisfy the moment restrictions (A.15):

$$\sum_{j=1}^{q}\psi\left(\xi_j, \theta\right)\pi_j = 0. \tag{A.89}$$

Since this is a parametric likelihood problem, we can obtain the asymptotic Cramer-Rao information bound for θ. Chamberlain (1987) showed that this bound corresponds to (A.82). Thus the optimal GMM variance is the lower bound on asymptotic variance that can be achieved in the multinomial case, regardless of knowledge of the support of the distribution of w.

Next, Chamberlain argued that any distribution can be approximated arbitrarily well by a multinomial. He used a formal approximation argument to show that the restriction to finite support is not essential, thus characterizing (A.82) as a semi-parametric information bound.

A.8 Testing the Overidentifying Restrictions

When $r > k$ there are testable restrictions implied by the econometric model. Estimation of θ sets to zero k linear combinations of the r sample orthogonality conditions $b_N(c)$. So, when the model is right, there are $r-k$ linearly independent combinations of $b_N\left(\widehat{\theta}\right)$ that should be close to zero but are not exactly equal to zero.

The main result here is that a minimized optimal GMM criterion scaled by N has an asymptotic chi-square distribution with $r-k$ degrees of freedom:

$$Ns\left(\widehat{\theta}\right) = Nb_N\left(\widehat{\theta}\right)'\widehat{V}^{-1}b_N\left(\widehat{\theta}\right) \xrightarrow{d} \chi^2_{r-k} \tag{A.90}$$

where $\widehat{\theta}$ is an optimal estimator and $\widehat{V}$ is a consistent estimate of V_0.

A statistic of this form is called a Sargan test statistic in the instrumental-variable context, and more generally a J or a Hansen test statistic (cf. Sargan, 1958, 1959; and Hansen, 1982).

As a sketch of the argument, note that factoring $\widehat{V}^{-1} = \widehat{C}\widehat{C}'$, in view of (A.52)

$$\sqrt{N}\widehat{C}'b_N(\theta) \xrightarrow{d} \mathcal{N}(0, I_r). \tag{A.91}$$

Moreover, letting $\widehat{G} = \widehat{C}'\widehat{D}$ we have

$$\begin{aligned}\sqrt{N}\left(\widehat{\theta}-\theta\right) &= -\left(\widehat{D}'\widehat{V}^{-1}\widehat{D}\right)^{-1}\widehat{D}'\widehat{V}^{-1}\sqrt{N}b_N(\theta)+o_p(1)\\ &= -\left(\widehat{G}'\widehat{G}\right)^{-1}\widehat{G}'\sqrt{N}\widehat{C}'b_N(\theta)+o_p(1).\end{aligned} \tag{A.92}$$

Now, using a first-order expansion for $b_N\left(\widehat{\theta}\right)$ and combining the result with (A.92):

$$\begin{aligned}h &\equiv \sqrt{N}\widehat{C}'b_N\left(\widehat{\theta}\right) = \sqrt{N}\widehat{C}'b_N(\theta)+\widehat{C}'\widehat{D}\sqrt{N}\left(\widehat{\theta}-\theta\right)+o_p(1)\\ &= \left[I_r-\widehat{G}\left(\widehat{G}'\widehat{G}\right)^{-1}\widehat{G}'\right]\sqrt{N}\widehat{C}'b_N(\theta)+o_p(1).\end{aligned} \tag{A.93}$$

Since the limit of $\left[I_r-\widehat{G}\left(\widehat{G}'\widehat{G}\right)^{-1}\widehat{G}'\right]$ is idempotent and has rank $r-k$, $h'h \xrightarrow{d} \chi^2_{r-k}$, from which (A.90) follows.

Incremental Sargan Tests Let us consider a partition

$$\psi(w,\theta) = \begin{pmatrix}\psi_1(w,\theta)\\ \psi_2(w,\theta)\end{pmatrix} \tag{A.94}$$

where $\psi_1(w,\theta)$ and $\psi_2(w,\theta)$ are of orders r_1 and r_2, respectively.

Suppose that $r_1 > k$ and that we wish to test the restrictions

$$E\psi_2(w,\theta) = 0 \tag{A.95}$$

taking $E\psi_1(w,\theta) = 0$ as a maintained hypothesis.

From the earlier result we know that

$$Ns_1\left(\widehat{\theta}_{[1]}\right) = Nb_{1N}\left(\widehat{\theta}_{[1]}\right)'\widehat{V}_1^{-1}b_{1N}\left(\widehat{\theta}_{[1]}\right) \xrightarrow{d} \chi^2_{r_1-k} \tag{A.96}$$

where $b_{1N}(c) = N^{-1}\sum_{i=1}^N \psi_1(w_i,c)$, $\widehat{V}_1$ is a consistent estimate of the covariance matrix $E\left[\psi_1(w,\theta)\psi_1(w,\theta)'\right]$, and $\widehat{\theta}_{[1]}$ is the minimizer of $s_1(c)$.

Then it can be shown that

$$S_d = Ns\left(\widehat{\theta}\right) - Ns_1\left(\widehat{\theta}_{[1]}\right) \xrightarrow{d} \chi^2_{r_2} \tag{A.97}$$

and that S_d is asymptotically independent of $Ns_1\left(\widehat{\theta}_{[1]}\right)$.

Therefore, the incremental statistic S_d can be used to test (A.95), having previously tested the validity of $E\psi_1(w,\theta)=0$ or maintaining their validity a priori.

To prove (A.97), note that in view of (A.93) we have:

$$\begin{aligned} h &\equiv \sqrt{N}\widehat{C}'b_N\left(\widehat{\theta}\right) = \left[I_r - G\left(G'G\right)^{-1}G'\right]\sqrt{N}C'b_N(\theta) + o_p(1) \\ h_1 &\equiv \sqrt{N}\widehat{C}_1'b_{1N}\left(\widehat{\theta}_{[1]}\right) = \left[I_{r_1} - G_1\left(G_1'G_1\right)^{-1}G_1'\right]\sqrt{N}C_1'b_{1N}(\theta) + o_p(1) \end{aligned} \tag{A.98}$$

where G and C denote the probability limits of $\widehat{G}$ and $\widehat{C}$, respectively, and we are using similar definitions of G_1, C_1, $\widehat{G}_1$, and $\widehat{C}_1$ applied to the first r_1 moments. Thus, we have $G=C'D_0$ and $G_1=C_1'D_{10}$, with $D_0=(D_{10}',D_{20}')'$.

Next, consider an orthogonal transformation of the two blocks of moments:

$$\psi_i^* = \begin{pmatrix} \psi_{1i} \\ \psi_{2i}^* \end{pmatrix} = \begin{pmatrix} I & 0 \\ -H_{21} & I \end{pmatrix}\begin{pmatrix} \psi_{1i} \\ \psi_{2i} \end{pmatrix} = H\psi_i \tag{A.99}$$

where for shortness we are writing $\psi_i=\psi(w_i,\theta)$, etc., and

$$H_{21} = E\left(\psi_{2i}\psi_{1i}'\right)\left[E\left(\psi_{1i}\psi_{1i}'\right)\right]^{-1}. \tag{A.100}$$

Also, let us denote $b_N^*(\theta)=Hb_N(\theta)$, $D_0^*=HD_0$, and $C^{*\prime}=C'H^{-1}$. With these notations, we can rewrite (A.98) as:

$$h = \left[I_r - G\left(G'G\right)^{-1}G'\right]\sqrt{N}C^{*\prime}b_N^*(\theta) + o_p(1). \tag{A.101}$$

Clearly, G is unaffected by the transformation since $G=C'D_0=C^{*\prime}D_0^*$. Moreover, because of block-orthogonality, $C^{*\prime}$ is block diagonal with elements C_1' and $C_2^{*\prime}$, say. Hence, G_1 contains the top r_1 rows of G:

$$G = C^{*\prime}D_0^* = \begin{pmatrix} C_1' & 0 \\ 0 & C_2^{*\prime} \end{pmatrix}\begin{pmatrix} D_{10} \\ D_{20}^* \end{pmatrix} = \begin{pmatrix} G_1 \\ G_2^* \end{pmatrix}. \tag{A.102}$$

Therefore, letting $M=I_r-G(G'G)^{-1}G'$ and

$$M_1 = \begin{pmatrix} \left[I_{r_1} - G_1\left(G_1'G_1\right)^{-1}G_1'\right] & 0 \\ 0 & 0 \end{pmatrix},$$

we can write

$$\begin{pmatrix} h_1 \\ 0 \end{pmatrix} = M_1\sqrt{N}C^{*\prime}b_N^*(\theta) + o_p(1) \tag{A.103}$$

and

$$h'h - h_1'h_1 = Nb_N^*(\theta)'C^*(M-M_1)C^{*\prime}b_N^*(\theta) + o_p(1). \tag{A.104}$$

Finally, notice that $(M - M_1)$ is symmetric and idempotent with rank $r - r_1$, and also

$$(M - M_1) M_1 = 0, \tag{A.105}$$

from which (A.97) and the asymptotic independence between S_d and $Ns_1\left(\widehat{\theta}_{[1]}\right)$ follow.

Example: 2SLS We may consider a test of the validity of a subset of instruments for the model

$$y = X\theta + u, \tag{A.106}$$

where the $N \times r$ data matrix of instruments is partitioned as $Z = \left(Z_1 \vdots Z_2\right)$, Z_1 is $N \times r_1$ and Z_2 is $N \times r_2$.

Thus,

$$\sum_{i=1}^{N} \psi(w_i, \theta) = \begin{pmatrix} Z_1' u \\ Z_2' u \end{pmatrix}. \tag{A.107}$$

In this example S_d performs a test of the validity of the additional instruments Z_2 given the validity of Z_1.

If $k = 1$ and $r = 2$, x_i, z_{1i}, and z_{2i} are scalar variables, and the single parameter θ satisfies the two moment conditions

$$\begin{aligned} E[z_{1i}(y_i - \theta x_i)] &= 0 \\ E[z_{2i}(y_i - \theta x_i)] &= 0. \end{aligned} \tag{A.108}$$

So the Sargan test is testing just one overidentifying restriction, which can be written as the equality of two simple IV estimating coefficients:

$$\frac{E(z_{1i} y_i)}{E(z_{1i} x_i)} = \frac{E(z_{2i} y_i)}{E(z_{2i} x_i)}. \tag{A.109}$$

Irrelevance of Unrestricted Moments Let us suppose that the sample moment vector consists of two components

$$b_N(\theta) = \begin{pmatrix} b_{1N}(\theta_1) \\ b_{2N}(\theta) \end{pmatrix} \begin{matrix} r_1 \times 1 \\ r_2 \times 1 \end{matrix} \tag{A.110}$$

corresponding to a partition $\theta = \left(\theta_1', \theta_2'\right)' \in \Theta_1 \times \Theta_2$ of dimensions k_1 and k_2, respectively. The first component of $b_N(\theta)$ depends only on θ_1 whereas the second depends on both θ_1 and θ_2. We assume that $r_1 \geq k_1$, but $r_2 = k_2$. Moreover, $\partial b_{2N}(c) / \partial c_2'$ is non-singular for all c, so that $b_{2N}(\theta)$ are effectively unrestricted moments.

Suppose that we are primarily interested in the estimation of θ_1. We wish to compare two different GMM estimators of θ_1. The first one is a joint estimator of θ_1 and θ_2 using all the moments:

$$\widehat{\theta} = \begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} = \arg\min_c b_N(c)' V^{-1} b_N(c) = \arg\min_c s_N(c). \tag{A.111}$$

The other is a separate estimator of θ_1 based on the first r_1 moments:

$$\widetilde{\theta}_1 = \arg\min_{c_1} b_{1N}(c_1)' V_{11}^{-1} b_{1N}(c_1) = \arg\min_c s_N^*(c_1) \tag{A.112}$$

where V_{11} consists of the first r_1 rows and columns of V.

The result is that as long as $b_{2N}(\theta)$ are unrestricted moments:[7]

$$\widehat{\theta}_1 = \widetilde{\theta}_1. \tag{A.113}$$

Moreover, since $s_N\left(\widehat{\theta}\right) = s_N^*\left(\widetilde{\theta}_1\right)$, provided V is an optimal weight matrix, the Sargan test statistics of $b_N(\theta)$ and $b_{1N}(\theta_1)$ coincide.

To see this, we need to show that $s_N^*(c_1)$ coincides with $s_N(c)$ concentrated with respect to c_2. Let us write

$$s_N(c) = b_{1N}(c_1)' V^{11} b_{1N}(c_1) + 2 b_{1N}(c_1)' V^{12} b_{2N}(c) + b_{2N}(c)' V^{22} b_{2N}(c) \tag{A.114}$$

where

$$V^{-1} = \begin{pmatrix} V^{11} & V^{12} \\ V^{21} & V^{22} \end{pmatrix},$$

and let $\widehat{\theta}_2(c_1)$ be the minimizer of $s_N(c)$ with respect to c_2 for given c_1.

In general, $\widehat{\theta}_2(c_1)$ satisfies the first-order conditions

$$\frac{\partial s_N(c)}{\partial c_2} = 2\left(\frac{\partial b_{2N}(c)}{\partial c_2'}\right)' \left[V^{22} b_{2N}(c) + V^{21} b_{1N}(c_1)\right] = 0, \tag{A.115}$$

but if $b_{2N}(c)$ are unrestricted, $\widehat{\theta}_2(c_1)$ satisfies

$$b_{2N}\left(c_1, \widehat{\theta}_2(c_1)\right) = -\left(V^{22}\right)^{-1} V^{21} b_{1N}(c_1). \tag{A.116}$$

Therefore, the concentrated criterion is given by

$$\begin{aligned} s_N\left(c_1, \widehat{\theta}_2(c_1)\right) &= b_{1N}(c_1)' \left[V^{11} - V^{12}\left(V^{22}\right)^{-1} V^{21}\right] b_{1N}(c_1) \\ &= b_{1N}(c_1)' V_{11}^{-1} b_{1N}(c_1), \end{aligned} \tag{A.117}$$

which coincides with $s_N^*(c_1)$ in view of the formulae for partitioned inverses.

[7] A similar result for minimum distance is in Chamberlain (1982, Proposition 9b).

B Optimal Instruments in Conditional Models

B.1 Introduction

So far the starting point of our discussion has been an $r \times 1$ vector of orthogonality conditions of the form

$$E\psi(w, \theta) = 0. \tag{B.1}$$

Given these moment restrictions we obtained asymptotically efficient GMM estimators of θ.

However, we are often interested in models that imply an infinite number of orthogonality conditions. In particular, this is the case with models defined by conditional moment restrictions. For example, the linear regression model

$$E(y \mid x) = x'\theta \tag{B.2}$$

implies that

$$E\left[h(x)\left(y - x'\theta\right)\right] = 0 \tag{B.3}$$

for any function h such that the expectation exists, and therefore in general an infinite set of unconditional moment restrictions.

Note, however, that the number of restrictions is finite if x is discrete and only takes a finite number of different values. For example, suppose that x is a single $0-1$ binary variable. Let $h_0(x)$ and $h_1(x)$ be the indicator functions of the events $x = 0$ and $x = 1$, respectively, so that $h_0(x) = 1 - x$ and $h_1(x) = x$. Clearly, any other function of x will be a linear combination of these two. Therefore, in this case (B.2) only implies two moment restrictions:

$$\begin{aligned} E[h_0(x)(y - x\theta)] &= 0 \\ E[h_1(x)(y - x\theta)] &= 0. \end{aligned} \tag{B.4}$$

Similarly, if x is discrete with q points of support $\left(\xi_1, ..., \xi_q\right)$, the conditional moment restriction (B.2) implies q unconditional moments:

$$E[h_j(x)(y - x\theta)] = 0 \,(j = 1, ..., q) \tag{B.5}$$

where $h_j(x) = 1\left(x = \xi_j\right)$ (cf. Chamberlain, 1987).[8]

The question that we address here is whether it is possible to find a finite set of optimal orthogonality conditions that give rise to asymptotically efficient estimators, in the sense that their asymptotic variance cannot be reduced by using additional orthogonality conditions.

[8] We use $1(A)$ to denote the indicator function of event A, such that $1(A) = 1$ if A is true and $1(A) = 0$ otherwise.

We begin by solving the problem for the linear regression model which is the most familiar context, and next we use the same procedure for increasingly more complex models. The most general case that we consider, a set of nonlinear simultaneous implicit equations, nests all the others as special cases.

In all cases we assume the identification of the parameters that we wish to estimate. Moreover, except in a cursory way, we do not consider explicitly specific feasible estimators. Instead, the focus of our discussion is in finding the optimal instruments for each type of model.

We only consider optimal instruments for *iid* observations. The analysis of optimal instruments for dependent observations is more complicated. Moreover, the *iid* assumption is sufficiently general to cover panel data models in a fixed T, large N setting.

Amemiya (1977) obtained the optimal instruments for a nonlinear simultaneous equation model with homoskedastic and serially uncorrelated errors. The form of the optimal instruments for a conditional mean model with dependent observations was derived by Hansen (1985). Chamberlain (1987) found that the optimal IV estimator attains the semi-parametric efficiency bound for conditional moment restrictions. Newey (1993) provides a survey of the literature and discussion of nonparametric estimation of the optimal instruments in the *iid* case.

B.2 Linear Regression

The model is

$$y = x'\theta + u \tag{B.6}$$

$$E(u \mid x) = 0, \tag{B.7}$$

where y is a scalar variable, and x and θ are $k \times 1$.

Let $z = z(x)$ denote a $p \times 1$ vector of functions of x such that $p \geq k$. Then z is a vector of valid instruments since

$$E\left[z\left(y - x'\theta\right)\right] = 0. \tag{B.8}$$

The optimal GMM estimator based on a given set of orthogonality conditions $E\psi(w, \theta) = 0$ and *iid* observations has asymptotic variance

$$\left(D_0' V_0^{-1} D_0\right)^{-1} \tag{B.9}$$

where $D_0 = E\left[\partial\psi(w, \theta)/\partial c'\right]$ and $V_0 = E\left[\psi(w, \theta)\psi(w, \theta)'\right]$ (see Section A.7). In our case $\psi(w, \theta) = z(y - x'\theta)$, and therefore

$$D_0 = -E\left(zx'\right) \tag{B.10}$$

$$V_0 = E\left(u^2 zz'\right) = E\left[\sigma^2(x) zz'\right] \tag{B.11}$$

where

$$E\left(u^2 \mid x\right) = \sigma^2\left(x\right). \tag{B.12}$$

Hence, the expression for (B.9) is

$$\left\{E\left(xz'\right) E\left[\sigma^2\left(x\right) zz'\right]^{-1} E\left(zx'\right)\right\}^{-1} \tag{B.13}$$

The optimal instruments in this case are

$$z^*\left(x\right) = \frac{x}{\sigma^2\left(x\right)}. \tag{B.14}$$

Setting $z = z^*\left(x\right)$ the asymptotic variance (B.13) for the optimal instruments takes the form[9]

$$\left[E\left(\frac{xx'}{\sigma^2\left(x\right)}\right)\right]^{-1}. \tag{B.15}$$

To show that $z^*\left(x\right)$ are the optimal instruments we prove that for any other z:

$$E\left(\frac{xx'}{\sigma^2\left(x\right)}\right) - E\left(xz'\right) E\left[\sigma^2\left(x\right) zz'\right]^{-1} E\left(zx'\right) \geq 0. \tag{B.16}$$

Letting $x^\dagger = x/\sigma\left(x\right)$, $z^\dagger = \sigma\left(x\right) z$, and $w = \left(x^{\dagger\prime}, z^{\dagger\prime}\right)'$, the *lhs* of (B.16) can be rewritten as

$$E\left(x^\dagger x^{\dagger\prime}\right) - E\left(x^\dagger z^{\dagger\prime}\right)\left[E\left(z^\dagger z^{\dagger\prime}\right)\right]^{-1} E\left(z^\dagger x^{\dagger\prime}\right) = H'E\left(ww'\right) H \tag{B.17}$$

where

$$H' = \left(I \vdots - E\left(x^\dagger z^{\dagger\prime}\right)\left[E\left(z^\dagger z^{\dagger\prime}\right)\right]^{-1}\right). \tag{B.18}$$

Clearly, $E\left(ww'\right) \geq 0$ since for any a of the same dimension as w, we have $a'E\left(ww'\right) a = E\left(\zeta^2\right) \geq 0$ with $\zeta = a'w$. Therefore, $H'E\left(ww'\right) H \geq 0$ also, which shows that (B.15) is a lower bound for variances of the form (B.13).

Thus, for example, if we consider an optimal GMM estimator that uses an augmented instrument set

$$z = \begin{pmatrix} z^*\left(x\right) \\ h\left(x\right) \end{pmatrix}$$

for some $h\left(x\right)$, there is no improvement in the asymptotic variance which remains equal to (B.15).

The direct implication of this result is that the estimator $\widetilde{\theta}$ that solves

$$\sum_{i=1}^{N} z^*\left(x_i\right)\left(y_i - x_i'\widetilde{\theta}\right) = 0 \tag{B.19}$$

[9] The optimal choice of instruments is up to a multiplicative constant, since any $b\left(x\right)c$ for constant $c \neq 0$ does not change the asymptotic variance.

is optimal. Note that the optimal instrument has the same dimension as θ, so that no further weighting of the moments is required.

Of course, we have just reviewed the Gauss–Markov result and $\widetilde{\theta}$ is nothing more than the unfeasible GLS estimator of θ:

$$\widetilde{\theta} = \left(\sum_{i=1}^{N} \frac{x_i x_i'}{\sigma^2 (x_i)} \right)^{-1} \sum_{i=1}^{N} \frac{x_i y_i}{\sigma^2 (x_i)}. \tag{B.20}$$

This estimator is unfeasible because the form of $\sigma^2 (.)$ is unknown.

Homoskedasticity Under homoskedasticity $\sigma^2 (x) = \sigma^2$ for all x and the optimal variance (B.15) becomes

$$\sigma^2 \left[E \left(xx' \right) \right]^{-1}. \tag{B.21}$$

The optimal instruments are the x themselves since σ becomes an irrelevant constant, so that OLS is optimal.

Note that all we are saying is that OLS attains the asymptotic variance bound when $\sigma^2 (x)$ happens to be constant, but this constancy is not taken into account in the calculation of the bound. If we incorporate the homoskedasticity assumption in estimation the bound for θ may be lowered as we show later in Section B.6.

Feasible GLS The efficiency result suggests to consider feasible GLS estimators that use estimated optimal instruments $\widehat{z}^* (x_i) = x_i / \widehat{\sigma}^2 (x_i)$ where $\widehat{\sigma}^2 (x_i)$ is an estimate of $\sigma^2 (x_i)$. If there is a known (or presumed) functional form of the heteroskedasticity $\sigma^2 (x_i, \gamma)$, we can set $\widehat{\sigma}^2 (x_i) = \sigma^2 (x_i, \widehat{\gamma})$ using a consistent estimator $\widehat{\gamma}$ of γ. For example, we can use squared OLS residuals $\widehat{u}_i$ to obtain a regression estimate of γ of the form

$$\widehat{\gamma} = \arg \min_{b} \sum_{i=1}^{N} \left[\widehat{u}_i^2 - \sigma^2 (x_i, b) \right]^2. \tag{B.22}$$

Under correct specification and appropriate regularity conditions, it is well known that feasible and unfeasible GLS have the same asymptotic distribution. Alternatively, $\widehat{\sigma}^2 (x_i)$ could be a nonparametric estimator of the conditional variance. A nearest neighbour estimate that lead to an asymptotically efficient feasible GLS was discussed by Robinson (1987).

B.3 Nonlinear Regression

The model is

$$y = f (x, \theta) + u \tag{B.23}$$

$$E(u \mid x) = 0 \tag{B.24}$$

where y is a scalar variable, θ is $k \times 1$, and $f(x, \theta)$ is some differentiable nonlinear function of x and θ.

As before, we consider an arbitrary vector of instruments $z = z(x)$ and the moments

$$E[z(y - f(x, \theta))] = 0. \tag{B.25}$$

The only difference with the linear case is that now

$$D_0 = -E\left(z f_1'\right) \tag{B.26}$$

where

$$f_1 \equiv f_1(x, \theta) = \frac{\partial f(x, \theta)}{\partial c}. \tag{B.27}$$

Therefore, the expression for the asymptotic variance (B.9) is

$$\left\{E\left(f_1 z'\right) E\left[\sigma^2(x) zz'\right]^{-1} E\left(z f_1'\right)\right\}^{-1}. \tag{B.28}$$

Following the steps of the linear case, the optimal instruments are

$$z^*(x) = \frac{f_1(x, \theta)}{\sigma^2(x)} \tag{B.29}$$

and the corresponding variance

$$\left[E\left(\frac{f_1(x, \theta) f_1(x, \theta)'}{\sigma^2(x)}\right)\right]^{-1}. \tag{B.30}$$

This variance is achieved by the unfeasible IV estimator $\widetilde{\theta}$ that solves the nonlinear sample moment equations:

$$q_N(c) = \sum_{i=1}^{N} \frac{f_1(x_i, \theta)}{\sigma^2(x_i)} [y_i - f(x_i, c)] = 0. \tag{B.31}$$

This estimator is unfeasible on two accounts. In common with the linear case, $\widetilde{\theta}$ depends on the conditional variance $\sigma^2(x_i)$, which is an unknown function of x_i. But it also depends on the vector of partial derivatives $f_1(x_i, \theta)$ evaluated at θ, which are known functions of x_i and the unknown true values of the parameters. It can be shown that substituting θ in (B.31) by a consistent estimator we still get an asymptotically efficient estimator. Alternatively, instead of keeping the optimal instrument fixed we can update it in the estimation. This is precisely what nonlinear least-squares does, which we discuss next.

Nonlinear Least-Squares The generalized nonlinear least-squares estimator minimizes

$$\min_{c} \sum \frac{[y_i - f(x_i, c)]^2}{\sigma^2(x_i)} \tag{B.32}$$

and its first-order conditions are

$$\sum_{i=1}^{N} \frac{f_1(x_i, c)}{\sigma^2(x_i)} [y_i - f(x_i, c)] = 0, \tag{B.33}$$

which are similar to (B.31) except for the replacement of θ by c in the optimal instruments. It can be easily shown that the estimator that solves (B.33) is asymptotically equivalent to $\widetilde{\theta}$. Of course, to obtain a feasible estimator one still has to replace $\sigma^2(x_i)$ by an estimate as in the linear case.

Finally, note that in the homoskedastic case, the variance term becomes irrelevant and we obtain the ordinary nonlinear least-squares formulae.

B.4 Nonlinear Structural Equation

The model is

$$\rho(y, x, \theta) = u \tag{B.34}$$

$$E(u \mid x) = 0 \tag{B.35}$$

where now y is a vector of endogenous variables and $\rho(.,.,.)$ is a scalar function that usually will only depend on a subset of the conditioning variables x.

Again, considering an arbitrary instrument vector $z = z(x)$ with moments $E[z\rho(y, x, \theta)] = 0$ we obtain an expression for V_0 equal to (B.11) and

$$D_0 = E\left[z\rho_1(y, x, \theta)'\right] \tag{B.36}$$

where

$$\rho_1(y, x, \theta) = \frac{\partial \rho(y, x, \theta)}{\partial c}. \tag{B.37}$$

Note that $\rho_1(y, x, \theta)$ may depend on y and hence it is not a valid instrument in general. So we consider instead its conditional expectation given x

$$b(x) = E[\rho_1(y, x, \theta) \mid x]. \tag{B.38}$$

Moreover, by the law of iterated expectations

$$D_0 = E\left[zb(x)'\right]. \tag{B.39}$$

Now we can argue as in the case of the regression models, so that the optimal instruments are

$$z^*(x) = \frac{b(x)}{\sigma^2(x)} = E\left(\frac{\partial \rho(y, x, \theta)}{\partial c} \mid x\right) \sigma^{-2}(x), \tag{B.40}$$

and the optimal variance

$$\left[E\left(\frac{b\left(x\right)b\left(x\right)'}{\sigma^{2}\left(x\right)}\right)\right]^{-1}. \tag{B.41}$$

This variance is achieved by the unfeasible IV estimator $\widetilde{\theta}$ that satisfies the sample moment equations:

$$\sum_{i=1}^{N}\frac{b\left(x_{i}\right)}{\sigma^{2}\left(x_{i}\right)}\rho\left(y_{i},x_{i},\widetilde{\theta}\right)=0. \tag{B.42}$$

The difference with nonlinear regression is that now both $b\left(x_{i}\right)$ and $\sigma^{2}\left(x_{i}\right)$ are unknown functions of x_{i}. A parametric approach to feasible estimation is to specify functional forms for $b\left(x_{i}\right)$ and $\sigma^{2}\left(x_{i}\right)$, and substitute suitable estimates in (B.42). 2SLS can be regarded as an example of this approach and this is discussed below. On the other hand, there are two nonparametric approaches to feasible estimation. One is the "plug-in" method that replaces $b\left(x_{i}\right)$ and $\sigma^{2}\left(x_{i}\right)$ in (B.42) by nonparametric regression estimates. Another is to consider a GMM estimator based on an expanding set of instruments as N tends to infinity for a pre-specified class of functions (cf. Newey, 1990, 1993, for discussion and references).

Linear Structural Equation Letting $y=\left(y_{1},y_{2}'\right)'$, $x=\left(x_{1}',x_{2}'\right)'$, and $w=\left(y_{2}',x_{1}'\right)'$, we have

$$\rho\left(y,x,\theta\right)=y_{1}-w'\theta \tag{B.43}$$

where y_{1} is the (first) element of y, and w contains the remaining components of y and the conditioning variables that are included in the equation.

In this case

$$b\left(x\right)=-E\left(w\mid x\right)=-\left(\begin{array}{c}E\left(y_{2}\mid x\right)\\ x_{1}\end{array}\right), \tag{B.44}$$

so that the unfeasible optimal IV estimator is

$$\widetilde{\theta}=\left(\sum_{i=1}^{N}\frac{b\left(x_{i}\right)}{\sigma^{2}\left(x_{i}\right)}w_{i}'\right)^{-1}\sum_{i=1}^{N}\frac{b\left(x_{i}\right)}{\sigma^{2}\left(x_{i}\right)}y_{1i}. \tag{B.45}$$

If $E\left(w\mid x\right)$ is linear and $\sigma^{2}\left(x\right)$ is constant:

$$\begin{aligned}E\left(w\mid x\right) &= \Pi x\\ \sigma^{2}\left(x\right) &= \sigma^{2}\end{aligned}$$

where $\Pi=E\left(wx'\right)\left[E\left(xx'\right)\right]^{-1}$, the asymptotic variance (B.41) becomes

$$\sigma^{2}\left[\Pi E\left(xx'\right)\Pi'\right]^{-1}=\sigma^{2}\left\{E\left(wx'\right)\left[E\left(xx'\right)\right]^{-1}E\left(xw'\right)\right\}^{-1}, \tag{B.46}$$

which is the 2SLS asymptotic variance under homoskedasticity.[10]

In effect, 2SLS is the IV estimator that uses the sample linear projection $\widehat{\Pi}x$ as an estimate of the optimal instrument. It achieves the variance bound when $E(w \mid x)$ is linear and $\sigma^2(x)$ is constant, but this information is not used in the specification of the estimation problem.

We saw in Section A.7 that if there is heteroskedasticity the two-step GMM estimator (A.87) is asymptotically more efficient than 2SLS. In the current notation two-step GMM solves

$$\sum_{i=1}^{N} \widehat{\Gamma} x_i \left(y_{1i} - w_i' \widehat{\theta}_{GMM2} \right) = 0 \tag{B.47}$$

where

$$\widehat{\Gamma} = \sum_{i=1}^{N} w_i x_i' \left(\sum_{i=1}^{N} \widehat{u}_i^2 x_i x_i' \right)^{-1}. \tag{B.48}$$

Thus, both GMM2 and 2SLS are using linear combinations of x as instruments.[11] Under heteroskedasticity, GMM2 is combining optimally a non-optimal set of orthogonality conditions. So, it is more efficient than 2SLS but inefficient relative to the IV estimator that uses $E(w \mid x)/\sigma^2(x)$ as instruments.

B.5 Multivariate Nonlinear Regression

We now consider a multivariate nonlinear regression

$$\begin{aligned} y_1 &= f_{[1]}(x, \theta) + u_1 \\ &\vdots \\ y_g &= f_{[g]}(x, \theta) + u_g \end{aligned} \tag{B.49}$$

with $E(u_j \mid x) = 0$ $(j = 1, ..., g)$, or in compact notation

$$y = f(x, \theta) + u \tag{B.50}$$

$$E(u \mid x) = 0 \tag{B.51}$$

where θ is $k \times 1$, and y, $f(x, \theta)$, and u are $g \times 1$ vectors.

This is a nonlinear system of "seemingly unrelated regression equations" (SURE) that places no restrictions on the second moments of the errors. There may be correlation among the errors of

[10] Note that replacing Π by $\widehat{\Pi}$ has no effect on the asymptotic distribution.

[11] $\widehat{\Gamma}$ is a consistent estimate of the coefficients of a linear projection of $E(w \mid x)/\sigma(x)$ on $\sigma(x)x$:

$$\Gamma = E\left[E(w \mid x)x'\right]\left\{E\left[\sigma^2(x)xx'\right]\right\}^{-1} = E\left(wx'\right)\left[E\left(u^2xx'\right)\right]^{-1}.$$

different equations, and their conditional variances and covariances may be heteroskedastic. Let the conditional variance matrix of u given x be

$$E\left(uu' \mid x\right) = \Omega\left(x\right), \tag{B.52}$$

which we assume to be nonsingular with probability one.

Let $Z = Z\left(x\right)$ denote a $g \times p$ matrix of functions of x such that $p \geq k$. Then we can form the moment conditions

$$E\left\{Z'\left[y - f\left(x, \theta\right)\right]\right\} = 0. \tag{B.53}$$

Proceeding as in the previous cases we have

$$D_0 = -E\left(Z'F_1\right) \tag{B.54}$$

where F_1 is the $g \times k$ matrix of partial derivatives

$$F_1 \equiv F_1\left(x, \theta\right) = \frac{\partial f\left(x, \theta\right)}{\partial c'}. \tag{B.55}$$

Moreover,

$$V_0 = E\left(Z'uu'Z\right) = E\left(Z'\Omega\left(x\right)Z\right). \tag{B.56}$$

Therefore, the optimal GMM variance based on (B.53) is

$$\left\{E\left(F_1'Z\right)\left[E\left(Z'\Omega\left(x\right)Z\right)\right]^{-1}E\left(Z'F_1\right)\right\}^{-1}. \tag{B.57}$$

The optimal instruments are

$$Z^*\left(x\right) = \Omega^{-1}\left(x\right)F_1 \tag{B.58}$$

in which case the asymptotic variance is

$$\left[E\left(F_1'\Omega^{-1}\left(x\right)F_1\right)\right]^{-1}. \tag{B.59}$$

To show that $Z^*\left(x\right)$ are the optimal instruments we just use a multivariate version of the argument employed in Section B.2. We need to prove that for any other Z:

$$E\left(F_1'\Omega^{-1}\left(x\right)F_1\right) - E\left(F_1'Z\right)\left[E\left(Z'\Omega\left(x\right)Z\right)\right]^{-1}E\left(Z'F_1\right) \geq 0. \tag{B.60}$$

Letting $F_1^\dagger = \Omega^{-1/2}\left(x\right)F_1$, $Z^\dagger = \Omega^{1/2}\left(x\right)Z$, and $W = \left(F_1^\dagger : Z^\dagger\right)$, the *lhs* of (B.60) can be rewritten as

$$E\left(F_1^{\dagger\prime}F_1^\dagger\right) - E\left(F_1^{\dagger\prime}Z^\dagger\right)\left[E\left(Z^{\dagger\prime}Z^\dagger\right)\right]^{-1}E\left(Z^{\dagger\prime}F_1^\dagger\right) = H'E\left(W'W\right)H \tag{B.61}$$

where

$$H' = \left(I \vdots - E\left(F_1^{\dagger\prime} Z^\dagger \right) \left[E\left(Z^{\dagger\prime} Z^\dagger \right) \right]^{-1} \right). \tag{B.62}$$

Since $H'E\left(W'W\right)H \geq 0$, (B.59) is a variance bound.[12]

The unfeasible optimal IV estimator solves

$$\sum_{i=1}^{N} F_1'\left(x_i,\theta\right)\Omega^{-1}\left(x_i\right)\left[y_i - f\left(x_i,c\right)\right] = 0. \tag{B.63}$$

Under homoskedasticity $\Omega\left(x\right) = \Omega$, but in the multivariate case the error variance still plays a role in the construction of the optimal instruments.

Multivariate Nonlinear Least-Squares This estimator minimizes

$$\sum_{i=1}^{N} \left[y_i - f\left(x_i,c\right)\right]' \Omega^{-1}\left(x_i\right)\left[y_i - f\left(x_i,c\right)\right] \tag{B.64}$$

with first-order conditions given by

$$\sum_{i=1}^{N} F_1'\left(x_i,c\right)\Omega^{-1}\left(x_i\right)\left[y_i - f\left(x_i,c\right)\right] = 0. \tag{B.65}$$

The same remarks we made for the single-equation case apply here. The estimators that solve (B.63) and (B.65) can be shown to be asymptotically equivalent. Moreover, a feasible estimator replaces $\Omega\left(x_i\right)$ by an estimated variance matrix. Under homoskedasticity this is simply given by

$$\widehat{\Omega} = \frac{1}{N}\sum_{i=1}^{N} \widehat{u}_i \widehat{u}_i'$$

where the $\widehat{u}_i$ are preliminary consistent residuals.

B.6 Nonlinear Simultaneous Equation System

Finally, we consider a system of implicit nonlinear simultaneous equations

$$\begin{aligned} \rho_{[1]}\left(y,x,\theta\right) &= u_1 \\ &\vdots \\ \rho_{[g]}\left(y,x,\theta\right) &= u_g \end{aligned} \tag{B.66}$$

[12] Note that $\Omega^{-1}\left(x\right)F_1\left(x,\theta\right)C$, where C is any $k \times k$ non-singular matrix of constants, are also optimal instruments, and that the variance bound does not depend on C.

with $E(u_j \mid x) = 0$ $(j = 1, ..., g)$. The structural model (B.34) can be regarded as a single equation from this system, so that the notation in both cases is similar, i.e. y and x are vectors of endogenous and conditioning variables, respectively. In a compact notation we have

$$\rho(y, x, \theta) = u \tag{B.67}$$

$$E(u \mid x) = 0 \tag{B.68}$$

where $\rho(y, x, \theta)$ and u are $g \times 1$ vectors.

As in the multivariate regression case, we take an arbitrary $g \times p$ instrument matrix $Z = Z(x)$ and form the moments

$$E\left[Z'\rho(y, x, \theta)\right] = 0. \tag{B.69}$$

In this case the expression for V_0 is the same as (B.56) and D_0 is given by

$$D_0 = E\left[Z'P_1(y, x, \theta)\right] = E\left\{Z'E\left[P_1(y, x, \theta) \mid x\right]\right\} = E\left[Z'B(x)\right] \tag{B.70}$$

where $P_1(y, x, \theta)$ is the $g \times k$ matrix of partial derivatives

$$P_1(y, x, \theta) = \frac{\partial \rho(y, x, \theta)}{\partial c'}, \tag{B.71}$$

and $B(x)$ denotes their conditional expectations given x:

$$B(x) = E\left[P_1(y, x, \theta) \mid x\right]. \tag{B.72}$$

The components of $P_1(y, x, \theta)$ are not valid instruments in general because of their dependence on y. So we consider instead $B(x)$, which are optimal predictors of $P_1(y, x, \theta)$ given x.

The optimal instruments are

$$Z^*(x) = \Omega^{-1}(x) B(x) \tag{B.73}$$

and the corresponding variance bound is

$$\left\{E\left[B(x)' \Omega^{-1}(x) B(x)\right]\right\}^{-1}. \tag{B.74}$$

This variance is achieved by the unfeasible IV estimator $\widetilde{\theta}$ that satisfies the sample moment equations:

$$\sum_{i=1}^{N} B(x_i)' \Omega^{-1}(x_i) \rho\left(y_i, x_i, \widetilde{\theta}\right) = 0. \tag{B.75}$$

The optimal IV estimator can also be expressed as the minimizer of a quadratic objective function. Since the number of moments equals the number of parameters, the choice of weight matrix is statistically irrelevant, but a computationally useful objective function is

$$\left(\sum_{i=1}^{N} \rho(y_i, x_i, c)' Z_i^*\right) \left(\sum_{i=1}^{N} Z_i^{*\prime} Z_i^*\right)^{-1} \left(\sum_{i=1}^{N} Z_i^{*\prime} \rho(y_i, x_i, c)\right) \tag{B.76}$$

where $Z_i^* = \Omega^{-1}(x_i) B(x_i)$.

The same comments we made for the single structural equation case regarding strategies to feasible estimation apply also here, so we shall not elaborate further.

Linear Simultaneous Equation System The 3SLS estimator considered in Section A.3 is an example of a feasible IV estimator that adopts a parametric specification of the optimal instruments. In the 3SLS context, $\rho(y_i, x_i, c)$ is a linear system, $B(x_i)$ contains sample linear projections, and $\Omega(x_i)$ is replaced by the unconditional covariance matrix of 2SLS residuals.

Specifically, we have

$$\rho(y, x, \theta) = y_1 - W\theta \tag{B.77}$$

where y_1 is a $g \times 1$ (sub) vector of the endogenous variables y whose coefficients are normalized to one, $\theta = \left(\theta_1' ... \theta_g'\right)'$ and W is a $g \times k$ block diagonal matrix containing both endogenous and exogenous explanatory variables:[13]

$$W = \begin{pmatrix} w_1' & & 0 \\ & \ddots & \\ 0 & & w_g' \end{pmatrix} \tag{B.78}$$

In this case

$$B(x) = -E(W \mid x) = -\begin{pmatrix} E(w_1' \mid x) & & 0 \\ & \ddots & \\ 0 & & E\left(w_g' \mid x\right) \end{pmatrix} \tag{B.79}$$

so that the unfeasible optimal IV estimator solves

$$\theta = \left(\sum_{i=1}^{N} B(x_i)' \Omega^{-1}(x_i) W_i\right)^{-1} \sum_{i=1}^{N} B(x_i)' \Omega^{-1}(x_i) y_{1i}. \tag{B.80}$$

If $E(W \mid x)$ is linear and $\Omega(x)$ is constant:

$$\begin{aligned} E(W \mid x) &= X\Pi^{\dagger} \\ \Omega(x) &= \Omega \end{aligned}$$

where $X = (I_g \otimes x')$ and $\Pi^{\dagger} = [E(X'X)]^{-1} E(X'W)$, the variance bound (B.74) coincides with the asymptotic variance of 3SLS:

$$\left[\Pi^{\dagger\prime} E\left(X'\Omega^{-1}X\right)\Pi^{\dagger}\right]^{-1}. \tag{B.81}$$

[13] If $y_1 - W\theta$ represents the errors of a complete system then $y_1 = y$, but distinguishing between y and y_1 our notation also accommodates incomplete linear systems.

Indeed, the 3SLS estimator solves

$$\sum_{i=1}^{N} \widehat{\Pi}^{\dagger\prime} X_i' \widehat{\Omega}^{-1} \left(y_{1i} - W_i \widehat{\theta}_{3SLS} \right) = 0 \tag{B.82}$$

where $\widehat{\Pi}^{\dagger} = \left(\sum_i X_i' X_i\right)^{-1} \sum_i X_i' W_i$ and $\widehat{\Omega}$ is the sample covariance matrix of 2SLS residuals.

Homoskedastic Linear Regression Nonlinear simultaneous equations are a useful motivation for (B.66), hence the title of this section. However, the conditional moment restrictions framework has broader applicability. Here we consider a linear regression model subject to homoskedasticity as an example of (B.66).

The model is

$$y = x'\beta + u \tag{B.83}$$

$$E\left(u \mid x\right) = 0 \tag{B.84}$$

$$E\left(u^2 \mid x\right) = \sigma^2. \tag{B.85}$$

Thus, we have $\theta = \left(\beta', \sigma^2\right)'$ and

$$\rho\left(y, x, \theta\right) = \begin{pmatrix} y - x'\beta \\ \left(y - x'\beta\right)^2 - \sigma^2 \end{pmatrix}. \tag{B.86}$$

Moreover,

$$P_1\left(y, x, \theta\right) = \frac{\partial \rho\left(y, x, \theta\right)}{\partial c'} = -\begin{pmatrix} x' & 0 \\ 2ux' & 1 \end{pmatrix} \tag{B.87}$$

$$B\left(x\right) = E\left(\frac{\partial \rho\left(y, x, \theta\right)}{\partial c'} \mid x\right) = -\begin{pmatrix} x' & 0 \\ 0 & 1 \end{pmatrix}. \tag{B.88}$$

Also,

$$\Omega\left(x\right) = \begin{pmatrix} \sigma^2 & E\left(u^3 \mid x\right) \\ E\left(u^3 \mid x\right) & E\left(u^4 \mid x\right) - \sigma^4 \end{pmatrix}. \tag{B.89}$$

If $E\left(u^3 \mid x\right) = 0$, the variance bound becomes

$$\left\{E\left[B\left(x\right)' \Omega^{-1}\left(x\right) B\left(x\right)\right]\right\}^{-1} = \begin{pmatrix} \sigma^{-2} E\left(xx'\right) & 0 \\ 0 & E\left[1/Var\left(u^2 \mid x\right)\right] \end{pmatrix}^{-1}. \tag{B.90}$$

Thus, there is no efficiency gain from incorporating the homoskedasticity assumption in estimation since we obtain the same bound that we got using (B.84) only. However, if $E\left(u^3 \mid x\right) \neq 0$ there is a lower variance bound for β than the one given in (B.21) (cf. MaCurdy, 1982).

Simultaneous System with Heteroskedasticity of Known Form Extending the argument in the previous example, let us consider now a nonlinear simultaneous system with a heteroskedastic conditional covariance matrix of known parametric form.[14] The model is

$$\rho^{\dagger}(y, x, \beta) = u \tag{B.91}$$

$$E(u \mid x) = 0 \tag{B.92}$$

$$E\left(uu' \mid x\right) = \Omega^{\dagger}(x, \gamma). \tag{B.93}$$

Thus, we have $\theta = \left(\beta', \gamma'\right)'$, $c = (b', g')'$, and[15]

$$\rho(y, x, \beta) = \begin{pmatrix} \rho^{\dagger}(y, x, \beta) \\ vech\left[\rho^{\dagger}(y, x, \beta)\,\rho^{\dagger}(y, x, \beta)' - \Omega^{\dagger}(x, \gamma)\right] \end{pmatrix}. \tag{B.94}$$

In this case we have[16]

$$P_1(y, x, \theta) = \begin{pmatrix} P_1^{\dagger}(y, x, \theta) & 0 \\ LK\left[u \otimes P_1^{\dagger}(y, x, \theta)\right] & -G_1(x, \gamma) \end{pmatrix}, \tag{B.95}$$

where

$$\begin{aligned} P_1^{\dagger}(y, x, \theta) &= \frac{\partial \rho^{\dagger}(y, x, \beta)}{\partial b'} \\ G_1(x, \gamma) &= \frac{\partial vech\Omega^{\dagger}(x, \gamma)}{\partial g'}, \end{aligned}$$

and L and K are matrices of constants such that $vech(uu') = Lvec(uu')$, and $\left(u \otimes P_1^{\dagger}\right) + \left(P_1^{\dagger} \otimes u\right) = K\left(u \otimes P_1^{\dagger}\right)$, respectively. Also,

$$B(x) = E\left[P_1(y, x, \theta) \mid x\right] = \begin{pmatrix} E\left[P_1^{\dagger}(y, x, \theta) \mid x\right] & 0 \\ LKE\left[u \otimes P_1^{\dagger}(y, x, \theta) \mid x\right] & -G_1(x, \gamma) \end{pmatrix}. \tag{B.96}$$

Note that now in general $E\left[u \otimes P_1^{\dagger}(y, x, \theta) \mid x\right] \neq 0$, since $P_1^{\dagger}(y, x, \theta)$ depends on y, and therefore its elements may be correlated with those of u. This is in contrast with the regression case, where $\rho^{\dagger}(y, x, \beta) = y - f(x, \beta)$, so that $P_1^{\dagger}$ does not depend on y.

[14] Note that in our context, homoskedasticity (with or without covariance restrictions) is just a special case of heteroskedasticity of known parametric form.

[15] The *vech* operator stacks by rows the lower triangle of a square matrix. It is used to avoid redundant elements, given the symmetry of the covariance matrix.

[16] We are using

$$\frac{\partial vec(uu')}{\partial b'} = \left(\frac{\partial u}{\partial b'} \otimes u\right) + \left(u \otimes \frac{\partial u}{\partial b'}\right).$$

Moreover, using the fact that $vech\left(uu'\right) = L\left(u \otimes u\right)$,

$$\Omega\left(x\right) = \begin{pmatrix} \Omega^{\dagger}\left(x,\gamma\right) & E\left(uu' \otimes u' \mid x\right) L' \\ LE\left(uu' \otimes u \mid x\right) & LE\left(uu' \otimes uu' \mid x\right) L' \end{pmatrix}. \tag{B.97}$$

This matrix is block diagonal if the conditional third-order moments of the us are zero. However, even if $E\left(uu' \otimes u \mid x\right) = 0$, the variance bound is not block diagonal between β and γ because $B\left(x\right)$ is not block diagonal. Therefore, there is an efficiency gain from incorporating the conditional covariance restrictions in the estimation of β. There is, of course, a trade-off between robustness and efficiency, since estimates of β that exploit the covariance restrictions may be inconsistent if these restrictions turn out to be false.

Finally, note that if the covariance matrix depends on both β and γ, so that

$$E\left(uu' \mid x\right) = \Omega^{\dagger}\left(x,\beta,\gamma\right), \tag{B.98}$$

the off-diagonal term of $B\left(x\right)$ has an additional non-zero term which is given by $-\partial vech\Omega^{\dagger}\left(x,\beta,\gamma\right)/\partial b'$. In such case there is an obvious efficiency gain from incorporating the covariance structure in the estimation of β, even if $P_1^{\dagger}$ does not depend on endogenous variables.

References

[1] Amemiya, T. (1977): "The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model", *Econometrica*, 45, 955–968.

[2] Amemiya, T. (1985): *Advanced Econometrics*, Blackwell, Oxford.

[3] Arellano, M. (2002): "Sargan's Instrumental Variables Estimation and the Generalized Method of Moments", *Journal of Business & Economic Statistics*, 20, 450–459.

[4] Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press, Oxford.

[5] Chamberlain, G. (1982): "Multivariate Regression Models for Panel Data", *Journal of Econometrics*, 18, 5–46.

[6] Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics*, 34, 305–334.

[7] Goldberger, A. S. (1991): *A Course in Econometrics*, Harvard University Press.

[8] Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50, 1029–1054.

[9] Hansen, L. P. (1985): "A Method of Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators", *Journal of Econometrics*, 30, 203–238.

[10] Jennrich, R. I. (1969): "Asymptotic Properties of Non-Linear Least Squares Estimators", *Annals of Mathematical Statistics*, 40, 633–643.

[11] MaCurdy, T. E. (1982): "Using Information on the Moments of Disturbances to Increase the Efficiency of Estimation", NBER Technical Paper 22, Cambridge, MA.

[12] Malinvaud, E. (1970): "The Consistency of Nonlinear Regressions", *Annals of Mathematical Statistics*, 41, 956–969.

[13] Manski, C. (1988): *Analog Estimation Methods in Econometrics*, Chapman and Hall, London.

[14] Newey, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models", *Econometrica*, 58, 809–837.

[15] Newey, W. K. (1993): "Efficient Estimation of Models with Conditional Moment Restrictions", in Maddala, G. S., C. R. Rao, and H. D. Vinod (eds.), *Handbook of Statistics*, Vol. 11, Elsevier Science.

[16] Newey, W. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing", in Engle, R. F. and D. L. McFadden (eds.) *Handbook of Econometrics,* IV, Ch. 36, North-Holland.

[17] Robinson, P. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form", *Econometrica*, 55, 875–891.

[18] Sargan, J. D. (1958): "The Estimation of Economic Relationships Using Instrumental Variables", *Econometrica*, 26, 393–415.

[19] Sargan, J. D. (1959): "The Estimation of Relationships with Autocorrelated Residuals by the Use of Instrumental Variables", *Journal of the Royal Statistical Society. Series B*, 21, 91–105.

[20] White, H. (1982): "Instrumental Variables Regression with Independent Observations", *Econometrica*, 50, 483–499.