6 Further Topics in (Cross-Sectional) Econometrics

6.1 Bootstrapping

For some (frequentist) point-estimators $\hat{\theta}$, we cannot obtain the finite-sample distribution. For example, in the linear regression model, without assuming Normality of erros, we only know the asymptotic distribution of $\hat{\beta} = (X'X)^{-1}X'Y$. The same holds for any estimator for which no closed form solution is available, as discussed in Chapter 5.

A bootstrap gives us a numerical approximation of the finite sample distribution. Recall that the finite sample distribution of $\hat{\theta}|\theta$ shows the variability in the point estimate $\hat{\theta}$ under different data samples $\{z_i\}_{i=1}^n$ of sample size n, i.e. it shows us the distribution of values for $\hat{\theta}$ we could have obtained had we drawn different samples of size n from the underlying population of data. Bootstrapping aims at approximating this randomness of the observed sample by using the particular sample we did obtain. It relies on the fact that all of our n observations are equally likely draws from the underlying population.

Algorithm 3 (Bootstrapping)**.**
For $m = 1 : M$,

1. *draw (with replacement) a sample of n_B observations from your data sample $\{z_i\}_{i=1}^n$.*
2. *using only this sample, $\{z_i^m\}_{i=1}^{n_B}$, compute the point estimator $\hat{\theta}^m$.*

The set $\{\hat{\theta}^m\}_{m=1}^M$ approximates the finite-sample distribution of $\hat{\theta}|\theta$ for a sample size n_B.

Taking $n_B = n$, we approxiamte the finite-sample distribution of $\hat{\theta}|\theta$ for our sample size of n observations.

6.2 Inference on Parameter-Transformations

Previous chapters discussed how to conduct inference on a parameter-vector θ using both frequentist as well as Bayesian methods. Based on that, what can we say about a function of θ, $g(\theta)$? For example, one might be interested in the predicted value $g(\beta) = x'\beta$ in the linear regression model or the partial effect under the Probit model, $g(\beta) = \phi(x'\beta)\beta$. The previous chapters – in particular Section 4.3 – discussed how to test hypotheses of the form $g(\theta) = 0$. This section discusses how to obtain a point estimator and coverage set for $g(\theta)$.

Bayesian inference yields a posterior distribution $p(\theta|Y)$, which allows us to construct point estimators and coverage sets for θ. Using this posterior, we can obtain the posterior for $h(\theta)$, $p(h(\theta)|Y)$, and compute point estimators and coverage sets for for $h(\theta)$. Sometimes, $p(h(\theta)|Y)$ can be obtained analytically. For example, in the linear regression model, we have $\beta|Y \sim N(\bar{\beta}, \sigma^2\bar{V})$, and therefore $x'\beta|Y \sim N(x'\bar{\beta}, \sigma^2 x'\bar{V}x)$.[1] In other cases, it must be obtained numerically. Based on draws $\{\theta^m\}_{m=1}^M$ from $p(\theta|Y)$, we can approximate $p(h(\theta)|Y)$ by the distribution of $\{h(\theta^m)\}_{m=1}^M$. This can be done arbitrarily well as $M \to \infty$. For example, in the Probit model, we could draw $\{\beta^m\}_{m=1}^M$ from $p(\beta|Y)$ and approximate $p(\phi(x'\beta)\beta|Y)$ by $\{\phi(x'\beta^m)\beta^m\}_{m=1}^M$.[2] In either case, given $p(g(\theta)|Y)$, we can take the posterior mean as a point estimator, and we can compute Bayesian Highest Posterior Density (HPD) credible sets by taking the set of values for $g(\theta)$ corresponding to the highest posterior mass (see Chapter 2).[3]

Frequentist inference yields – in absence of identification issues – a point estimator $\hat{\theta}$ with some finite-sample distribution, e.g. $\hat{\theta}|\theta \sim N(\theta, V)$. Sometimes, this distribution is exact, like in the linear regression model with Normal errors. In other cases, it is an approximation, obtained either based on the asymptotic distribution of $\hat{\theta}$ or by bootstrapping. An example is the linear regression model without assuming Normality of errors or any other model for which no closed form solution for $\hat{\theta} = \arg\min_{\theta\in\Theta} Q_n(\theta, Y^n)$ is available.

If $\hat{\theta}$ is consistent for θ and g is continuous at θ (the true value), by Slutsky's theorem (see Section 3.2), $g(\hat{\theta})$ is a consistent estimator of $g(\theta)$. Note that $g(\hat{\theta})$ is not necessarily unbiased. For example, if g is concave (convex), then there is a downward (upward) bias due to Jensen's inequality (see Section 1.1).

[1] These posteriors are conditional on σ^2.

[2] Note that this works both if we were able to derive $p(\beta|Y)$ analytically and if it is available only numerically.

[3] If $p(h(\theta)|Y)$ is obtained numerically, the former involves taking the mean of all draws, while the latter involves finding the $(1-\alpha)100\%$ draws that correspond to the highest values of the posterior. The latter can be done by computing a Kernel density estimate of the draws. If a connected interval is sought for, one can also sort the draws and look for the shortest connected interval with $(1-\alpha)100\%$ of the draws.

In order to construct confidence sets, we need to know the finite-sample distribution of $g(\hat{\theta})|g(\theta)$. For some, rather simple functions g, we can find this distribution analytically based on the finite-sample distribution of $\hat{\theta}|\theta$. For example, in the linear regression model, we have $\hat{\beta}|\beta \sim N(\beta, V)$ with $V = \sigma^2\mathbb{E}[(X'X)^{-1}]$, and therefore $x'\hat{\beta}|x'\beta \sim N(x'\beta, x'Vx)$. In other cases, we cannot find the finite-sample distribution of $g(\hat{\theta})|g(\theta)$, but only the asymptotic one, which enables us to construct coverage sets that are only asymptotically valid. For example, in the Probit model, suppose $\sqrt{n}(\hat{\beta} - \beta) \sim N(0, V)$. Because $g(\beta) = \phi(x'\beta)\beta$ is a continuous function with continuous first derivatives $G(\beta) = \frac{\partial g(\beta)}{\partial g(\beta)} = \phi(x'\beta)[I - x\beta']$ the Delta method (Proposition 11) tells us that $\sqrt{n}(g(\hat{\beta}) - g(\beta)) \xrightarrow{d} N(0, G(\beta)VG(\beta)')$.[4] Finally, for more complicated functions g, we can numerically approximate the finite-sample distribution of $g(\hat{\theta})|g(\theta)$ by drawing $\{\hat{\theta}^m\}_{m=1}^M$ from the finite-sample distribution of $\hat{\theta}|\theta$ and computing $\{g(\hat{\theta}^m)\}_{m=1}^M$. In turn, we can numerically construct confidence sets as laid out in Chapter 2.[5]

6.3 Generalized Method of Moments

The Generalized Method of Moments (GMM) finds θ s.t. a certain moment condition holds. Suppose we have r moment conditions for $k \leq r$ unknowns in θ. Formally, let $\mathbb{E}[h(y_i; \theta)] = 0$ iff $\theta = \theta_0$ for some function $h(y_i; \cdot) : \mathbb{R}^k \to \mathbb{R}^r$. For example, this could be the FOC from some theoretical model. For the linear regression model, we can write $\mathbb{E}[x_i(y_i - x_i'\beta)] = 0$ iff $\beta = \beta_0$.[6] Similarly, for the probit model, we have $\mathbb{E}[x_i(y_i - \Phi(x_i'\beta))] = 0$ iff $\beta = \beta_0$.

Let $\hat{\theta}_n = \arg\min_{\theta\in\Theta} Q_n(\theta, Y^n)$ for

$$Q_n(\theta, Y^n) = \frac{1}{2} g_n(\theta; Y^n)' W_n g_n(\theta; Y^n) \ , \quad \text{where } g_n(\theta; Y^n) = \frac{1}{n}\sum_i h(y_i; \theta) \ ,$$

and where $W_n \xrightarrow{p} W$ is an $r \times r$ symmetric, p.d. weighting matrix. Provided that there are no boundary issues, $\hat{\theta}_n$ solves the FOC

$$\frac{\partial\, Q_n(\theta; Y^n)}{\partial\, \theta} = \frac{\partial\, g_n(\theta; Y^n)'}{\partial\, \theta} \frac{\partial\, Q_n(\theta; Y^n)}{\partial\, g_n(\theta; Y^n)} = g_n^{(1)}(\theta; Y^n)' W_n g_n(\theta; Y^n) = 0 \ .$$

[4] Note that even if we knew the finite sample distribution of $\hat{\beta}$, we would not be able to derive the finite-sample distribution of $g(\hat{\beta})$ for this function g.

[5] First, note that if the finite sample distribution of $\hat{\theta}|\theta$ is only approximate, then this adds a second "layer" of approximation – and hence imprecision – in the computation of the finite-sample distribution of $g(\hat{\theta})|g(\theta)$. Second, note that this works even if the finite sample distribution of $\hat{\theta}|\theta$ is available only numerically, as is the case under bootstrapping.

[6] For any $\tilde{\beta} \neq \beta_0$, $\mathbb{E}[x_i(y_i - x_i'\tilde{\beta})] = \mathbb{E}[x_i(x_i'\beta + u_i - x_i'\tilde{\beta})] = \underbrace{\mathbb{E}[x_i u_i]}_{0} + \mathbb{E}[x_i x_i](\beta - \tilde{\beta}) \neq 0$.

We can verify consistency and asymptotic Normality of $\hat{\theta}_n$ by checking the conditions for extremum estimators laid out in Chapter 5. Thereby, $\mathbb{E}[h(y_i;\theta)] = 0$ iff $\theta = \theta_0$ and W being p.d. guarantee that $Q(\theta) = \frac{1}{2}\mathbb{E}[h(y_i;\theta)]'W\mathbb{E}[h(y_i;\theta)]$ is uniquely minimized at $\theta = \theta_0$. Also, $Q_n(\theta; Y^n) \xrightarrow{p} Q(\theta)$ uniformly if $\frac{1}{n}\sum_{i=1}^n h(y_i;\theta) = g_n(\theta; Y^n) \xrightarrow{p} \mathbb{E}[h(y_i;\theta)]$ uniformly, i.e. if $h(y_i;\theta)$ satisfies ULLN.

The asymptotic analysis of GMM estimators reveals insights about desirable properties of the weighting matrix and motivates a test for whether the GMM conditions are correctly specified. By CLT, $\sqrt{n}g_n(\theta_0; Y^n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h(\theta_0; y_i) \xrightarrow{d} N(0, S)$ and by WLLN, $g^{(1)}(\theta_0; Y^n) = \frac{1}{n}\sum_{i=1}^n h^{(1)}(\theta_0; y_i) \xrightarrow{p} D'$. Combining these two results, we have

$$\sqrt{n}Q_n^{(1)}(\theta_0; Y^n) = g^{(1)}(\theta_0; Y^n)'W_n\sqrt{n}g(\theta_0; Y^n) \xrightarrow{d} N(0, DWSWD') \ .$$

Also, $Q_n^{(2)}(\theta_0; Y^n) \xrightarrow{p} DWD'$ because

$$Q_n^{(2)}(\theta_0; Y^n) = \left[\frac{1}{n}\sum_{i=1}^n h^{(1)}(\theta_0; y_i)\right]' W_n \left[\frac{1}{n}\sum_{i=1}^n h^{(1)}(\theta_0; y_i)\right] + \left[\text{term involving } \frac{1}{n}\sum_{i=1}^n h(\theta_0; y_i) \xrightarrow{p} 0\right] \ .$$

Overall,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, (DWD)^{-1}DWSWD'(DWD')^{-1}\right) \ .$$

We estimate this asymptotic variance using $\hat{D}_n' = \frac{1}{n}\sum_{i=1}^n h^{(n)}(\hat{\theta}_n; y_i) \xrightarrow{p} D'$ and $\hat{S}_n = \frac{1}{n}\sum_{i=1}^n h(\hat{\theta}_n i y_i)h(\hat{\theta}_n; y_i)'$.

Proposition 30 (GMM: Optimal Weighting Matrix)**.**
The limit weighting matrix $W = S^{-1}$ minimizes the asymptotic variance of $\hat{\theta}_n$.

The proof is in the Appendix. The intuition behind this result is that moment conditions with a lot of noise should be downweighted. For some applications, it is possible to directly compute $S(\theta)$, the probability limit of the covariance matrix of $\sqrt{n}g_n(\theta_0; y^h)$, which allows us to set $W_n = S(\theta)^{-1}$ and take into account the dependence of W_n on θ when computing $\hat{\theta}$ in a single step. Otherwise, we can use 2-step GMM:

- First, set $W = I$ and get a preliminary estimate $\tilde{\theta}_n$.
- Then, estimate S by $\tilde{S}_n$ based on $h(\tilde{\theta}_n; y_i)$, set $W = \tilde{S}_n^{-1}$ and compute $\hat{\theta}_n$.

To minimize the asymptotic variance, one should use as many moment conditions as possible (see Appendix for a proof). In practice, however, this is not a good strategy as less informative moment conditions introduce noise and can lead to estimates that are very "off"

in finite samples (even if one uses an optimal weighting matrix).

J-specification test For over-identified models – i.e. when $r > k$ – we can test whether the moment conditions are correctly specified. Formally, we test $\mathcal{H}_0 : \exists\ \theta_0$ s.t. $\mathbb{E}[h(\theta_0; Y^n)] = 0$ vs. $\mathcal{H}_1 : \nexists\ \theta_0$ s.t. $\mathbb{E}[h(\theta_0; Y^n)] = 0$. Under $\mathcal{H}_0$, $\frac{1}{\sqrt{n}} \sum_i h(\theta_0; y_i) = \sqrt{n} g_n(\theta_0; Y^n) \xrightarrow{d} N(0, S)$, which implies

$$n \cdot g_n(\theta_0; Y^n)' S^{-1} g_n(\theta_0; Y^n) \xrightarrow{d} \chi^2_r \ .$$

Since $\hat{\theta}_n$ is chosen such that k linear combinations of the $r \times 1$ vector $g_n(\hat{\theta}_n; Y^n)$ are set to zero (see FOC for $\hat{\theta}_n$), we get

$$J = n g_n(\hat{\theta}_n; Y^n) \hat{S}^{-1} g_n(\hat{\theta}_n; Y^n) \xrightarrow{d} X^2_{r-k} \ .$$

If $r = k$, there is nothing to test, but by construction $\hat{\theta}_n$ satisfies all $r = k$ moment conditions. This test is usually run as a justification to use GMM on a particular set of moment conditions derived from some theoretical model.

6.4 Instrumental Variable Estimation

Section 3.4 introduced the endogeneity problem in the context of the linear regression model: $y_i = x_i'\beta + u_i$, but $\mathbb{E}[x_i u_i] \neq 0$ is suspected. As discussed, this leads to an inconsistent OLS (and MLE) estimator. Instrumental variable (IV) estimation is a potential remedy.

An IV z_i should satisfy two conditions:

1. Relevance: z_i is (highly) correlated with the regressor x_i: $\mathbb{E}[z_i x_i'] \neq 0$.
2. Validity: z_i is uncorrelated with the error term u_i: $\mathbb{E}[z_i u_i] = 0$.

The idea is to use z_i to extract the part of the information in x_i that is uncorrelated with u_i. This is best illustrated in the two-stage least squares (2SLS) estimation procedure. Suppose z_i and x_i are both scalars.

- First, estimate $x_i = z_i'\gamma + e_i$ to get $\hat{\gamma} = (Z'Z)^{-1}Z'X$ and $\hat{X} = P_Z X = Z\hat{\gamma}$. As Z is uncorrelated with U, so is $\hat{X}$.
- Then, estimate $y_i = \hat{x}_i'\beta + u_i^*$ to get $\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'P_Z X)^{-1}X'P_Z Y$.

Ideally, z_i should be as highly correlated as possible with x_i in order to preserve as much variation of x_i in $\hat{x}_i$ as possible. Note that $u_i^* \neq u_i$. Correspondingly, we estimate σ_u^2 based on the regressors x_i and not $\hat{x}_i$: $\hat{\sigma}_u^2 = \frac{1}{n}\sum_i (y_i - x_i'\hat{\beta}_{2SLS})^2$.

GMM Analysis of IV Estimation Suppose we have k regressors in x_i and r IVs in z_i. The vector z_i includes all regressors in x_i which satisfy the exogeneity condition – i.e. all regressors j for which $\mathbb{E}[x_{i,j}u_i] = 0$ – and enough actual IVs so that $r \geq k$. Based on the validity condition $\mathbb{E}[z_i u_i] = 0$, we can define the GMM-moment condition $\mathbb{E}[h(\beta, w_i)] = 0$ for $h(\beta, w_i) = z_i(y_i - x_i'\beta)$, where $w_i = (y_i, x_i', z_i')'$. If $\mathbb{E}[z_i x_i']_{r\times k}$ has full rank, it is easy to verify that $\mathbb{E}[h(\beta, w_i)] = 0$ iff $\beta = \beta_0$.

We get $g_n(\beta, W^n) = \frac{1}{n}\sum_i z_i(y_i - x_i'\beta) = \frac{1}{n}Z'(Y - X\beta)$ and the GMM objective function

$$Q_n(\beta, W^n) = \frac{1}{2}\frac{1}{n^2}(Y - X\beta)'ZW_nZ'(Y - X\beta) \ .$$

This leads to the FOC $\frac{1}{n^2}X'ZW_nZ'(Y - X\beta) = 0$ and therefore to

$$\hat{\beta}_{IV} = (X'ZW_nZ'X)^{-1}X'ZW_nZ'Y \ .$$

Note that we get the 2SLS estimator by setting $W_n = (\frac{1}{n}Z'Z)^{-1}$. Also, in just-identified models – i.e. under $r = k$ – we get $\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y$.

It is easy to show that under the validity assumption, $\hat{\beta}_{IV}$ is consistent for β and

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N\left(0, (DWD)^{-1}DWSWD'(DWD')^{-1}\right) \ ,$$

with $D = \mathbb{E}[z_i x_i']$ and $S = \mathbb{E}[z_i u_i (z_i u_i)']$. Under 2SLS, we have $W = \mathbb{E}[z_i z_i']^{-1}$. If in addition, u_i are homoskedastic – i.e. $\mathbb{V}[u_i|x_i] = \sigma^2$ – then $S = \sigma^2\mathbb{E}[z_i z_i']$ and the asymptotic variance simplifies to $\sigma^2(DWD)^{-1}$.

If $r > k$, i.e. if we have more IVs than covariates in x_i, we can apply the J-specification test. For IV estimation, this amounts to testing whether indeed all IVs are exogenous.

Likelihood-Based Analysis of IV Estimation IV models can also be estimated using ML or Bayesian estimation. For this, we construct the likelihood of the model

$$y_i = x_i'\beta + u_i \ , \quad x_i = z_i\gamma + \varepsilon \ ,$$

where $\mathbb{E}[x_i u_i] \neq 0$, $\mathbb{E}[z_i u_i] = 0$ and therefore $\mathbb{E}[\varepsilon_i u_i] \neq 0$. To proceed, we need to specify a distribution for $[u_i, \varepsilon_i]'$. For example,

$$\begin{bmatrix} u_i \\ \varepsilon_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{uu} & \Sigma_{u\varepsilon} \\ \Sigma_{u\varepsilon} & \Sigma_{\varepsilon\varepsilon} \end{bmatrix}\right) \ , \quad \text{implying} \quad \varepsilon_i \sim N(0, \Sigma_{\varepsilon\varepsilon}) \ , \quad \text{and} \ \ u_i|\varepsilon_i \sim N(\mu_{u|\varepsilon}, \Sigma_{u|\varepsilon}) \ ,$$

with $\mu_{u|\varepsilon} = \mu_u + \Sigma_{u\varepsilon}\Sigma_{\varepsilon\varepsilon}^{-1}(\varepsilon_i - \mu_\varepsilon)$ and $\Sigma_{u|\varepsilon} = \Sigma_{uu} - \Sigma_{u\varepsilon}\Sigma_{\varepsilon\varepsilon}^{-1}\Sigma_{u\varepsilon}$. Let $\theta = (\beta, \gamma, \Sigma)$. This leads to the likelihood

$$\begin{aligned}
\mathcal{L}(\theta|Y,X,Z) &= p(X,Y,Z|\theta) \\
&= p(Y|X,Z,\theta)p(X|Z,\theta)p(Z|\theta) \\
&\propto p(Y|X,Z,\theta)p(X|Z,\theta) \\
&= \prod_i p(y_i|x_i,z_i,\theta)p(x_i|z_i,\beta,\gamma,\Sigma) \\
&= \prod_i p(u_i|x_i,z_i,\theta)\big|_{u_i=y_i-x_i'\beta} p(\varepsilon_i|z_i,\theta)\big|_{\varepsilon_i=x_i'-z_i\gamma} \\
&= \prod_i (2\pi)^{-\frac{1}{2}}|\Sigma_{u|\varepsilon}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}[u_i - \Sigma_{u\varepsilon}\Sigma_{\varepsilon\varepsilon}^{-1}\varepsilon_i]\Sigma_{u|\varepsilon}^{-1}[u_i - \Sigma_{u\varepsilon}\Sigma_{\varepsilon\varepsilon}^{-1}\varepsilon_i]'\right\} \\
&\qquad \cdot (2\pi)^{-\frac{k}{2}}|\Sigma_{\varepsilon\varepsilon}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\varepsilon_i\Sigma_{\varepsilon\varepsilon}^{-1}\varepsilon_i\right\}.
\end{aligned}$$

where $u_i = y_i - x_i'\beta$ and $\varepsilon_i = x_i - z_i'\gamma$. Compared to the linear regression model, here the conditional mean of y_i is not simply $x_i'\beta$, but it includes a correction for endogeneity, $\mu_{u|\varepsilon}$. The intuition is that (x_i, z_i) provide information on ε_i, which provides information on u_i.

Weak Identification in IV Models If the the correlation between x_i and z_i is not very high, we speak of weak IVs. In this case, the finite sample distribution of $\hat{\beta}$ does not resemble the asymptotic one at all. A better approximation of the finite sample distribution is obtained by making the parameter γ "local to zero", i.e. by writing $\gamma = \tilde{\gamma}/\sqrt{n}$ for some fixed $\tilde{\gamma}$. In other words, we let the correlation of z_i and x_i go to zero as $n \to \infty$. For simplicity, assume we have $r = k = 1$ such that $\hat{\beta}$ satisfies $\frac{1}{n}\sum_i z_i(y_i - x_i\hat{\beta}) = 0$, and suppose u_i is homoskedastic. Then we get

$$\hat{\beta} = \frac{\frac{1}{n}\sum_i z_i y_i}{\frac{1}{n}\sum_i z_i x_i} = \beta + \frac{\frac{1}{\sqrt{n}}\sum_i z_i u_i}{\frac{1}{n}\sum_i z_i^2\tilde{\gamma} + \frac{1}{\sqrt{n}}\sum_i z_i\varepsilon_i} \xrightarrow{d} \beta + \frac{N(0,\sigma_u^2\mathbb{E}[z_i^2])}{\tilde{\gamma}\mathbb{E}[z_i^2] + N(0,\sigma_\varepsilon^2\mathbb{E}[z_i^2])},$$

which can be used to approximate the finite sample distribution of $\hat{\beta}$.

We can construct a confidence set for β under weak IVs using the inference procedure by Anderson and Rubin (1949). It is based on the insight that, for the right β, the auxiliary regression $y_i - x_i'\beta = \delta z_i + v_i$ should yield $\delta = 0$, because $y_i - x_i'\beta = u_i$ and u_i and z_i are uncorrelated. For a given β_0, we get

$$\sqrt{n}\hat{\delta}(\beta_0) = \sqrt{n}(Z'Z)^{-1}Z'(Y - X\beta_0) = (Z'Z)^{-1}\sqrt{n}Z'U \xrightarrow{d} N\left(0, \frac{\sigma_u^2}{\mathbb{E}(z_i^2)}\right),$$

which allows us to test $\mathcal{H}_0 : \delta = 0$. For example, if δ is a scalar, we can use the t-test $t_\delta(\beta_0) = \hat{\delta}(\beta_0)/\sqrt{\tilde{\sigma}_v^2/Z'Z} \xrightarrow{d} N(0,1)$. A confidence set for β is obtained by taking all β_0 for which $\mathcal{H}_0 : \delta = 0$ cannot be rejected. Note that this can give very large, unbounded and even disconnected or empty confidence sets.

Appendix

Claim. *The limit weighting matrix $W = S^{-1}$ minimizes the asymptotic variance of the GMM estimator $\hat{\theta}_n$.*

Proof: Under this W, we have $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, (DS^{-1}D')^{-1})$. We want to show that

$$\Delta = DS^{-1}D' - DWD'(DWSWD')^{-1}DWD'$$

is p.d..

We can always find Λ s.t. $S = \Lambda\Lambda'$. Also, define $K = \Lambda' WD'$ with $M_K = K(K'K)^{-1}K'$. Then we get

$$\begin{aligned}\Delta &= D\Lambda^{-1\prime}(I - \Lambda' UD'(DW\Lambda\Lambda' WD')^{-1}DW\Lambda)\Lambda^{-1}D' \\ &= D'\Lambda^{-1\prime}M_K\Lambda^{-1}D' \\ &= D\Lambda^{-1\prime}M_K(D\Lambda^{-1\prime}M_K)' \,,\end{aligned}$$

which is p.d. ■

Claim. *To minimize the asymptotic variance, one should use as many moment conditions as possible.*

Proof: Let B be a $m \times r$ matrix, with $m < r$. Also, let

$$\tilde{\theta}_n = \underset{p}{\text{argmin}} \frac{1}{2} g_n(\theta; Y^n)' B' W B g_n(\theta; Y^n) \,,$$

meaning that we only use m of the r moment conditions.

Suppose we use the efficient $W = (BSB')^{-1}$. Then we get that using $B = I$ – i.e. using all r moment conditions – gives the lowest asymptotic variance, because

$$\begin{aligned}\Delta &= DS^{-1}D' - DB'(BSB')^{-1}BD' \\ &= D\Lambda^{-1}[I - \Lambda' B'(B\Lambda\Lambda' B')^{-1}B\Lambda)\Lambda^{-1}D'\end{aligned}$$

is p.d. ■