7 Further Topics in (Cross-Sectional) Econometrics

7.1 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 6.

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.

7.2 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, $f(\theta)$? For example, one might be interested in the predicted value $f(\beta) = x'\beta$ in the linear regression model or the partial effect under the Probit model, $f(\beta) = \phi(x'\beta)\beta$. We might be interested in constructing a point estmator for $f(\theta)$, conducting hypothesis tests of the form $\mathcal{H}_0 : f(\theta) \in \mathcal{F}_0$ vs. $\mathcal{H}_0 : f(\theta) \in \mathcal{F}_1$ and finding a coverage set for $f(\theta)$.

Bayesian Inference Bayesian methods yield a posterior distribution of θ, $p(\theta|Y)$, which allows us yields point estimators and allows us to conduct hypothesis tests and construct coverage sets for θ. Using $p(\theta|Y)$, we can obtain the posterior distribution of $f(\theta)$, $p(f(\theta)|Y)$, and do the same for $f(\theta)$.

Sometimes, $p(f(\theta)|Y)$ 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 $p(f(\theta)|Y)$ by the distribution of $\{f(\theta^m)\}_{m=1}^M$. 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 $p(f(\theta)|Y)$, we can take the posterior mean as a point estimator,[3] we can conduct hypothesis tests by comparing some $\mathbb{P}[f(\theta) \in \mathcal{F}_0|Y]$ to some $\mathbb{P}[f(\theta) \in \mathcal{F}_1|Y]$, and we can compute Bayesian Highest Posterior Density (HPD) credible sets by taking the set of values for $f(\theta)$ corresponding to the highest posterior mass (see Chapter 2).[4]

[1] T hese 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] This presumes a quadratic loss function. See Section 2.2.1.

[4] If $p(f(\theta)|Y)$ 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.

Frequentist Inference Frequentist methods yield – 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.

A consistent estimator of $f(\theta_0)$ is given by $f(\hat{\theta})$, provided that $\hat{\theta}$ is consistent for θ_0 and f is continuous at θ_0. This follows from Slutsky's theorem (see Section 3.2). Note that $f(\hat{\theta})$ is not necessarily unbiased. For example, if f is concave (convex), then there is a downward (upward) bias due to Jensen's inequality (see Section 1.1).

To test a hypothesis such as $\mathcal{H}_0 : f(\theta) \in \mathcal{F}_0$, we can have several approaches available. For one, we might be able to transform the hypothesis into a statement of the form $\mathcal{H}_0 : \theta \in \Theta_0$, enabling us to use the tests discussed in the previous chapters. At least for a point hypothesis $\mathcal{H}_0 : f(\theta) = f_0$, this should be possible, as we can rewrite it as $\mathcal{H}_0 : g(\theta) = 0$ for $g(\theta) = f(\theta) - f_0$. In turn, we can turn these tests into confidence sets for θ and therefore $f(\theta)$.

Alternatively, we can test hypotheses and construct confidence sets related to $f(\theta)$, by using the finite-sample distribution of $f(\hat{\theta})|f(\theta)$. For some, rather simple functions f, we can find this distribution analytically based on the finite-sample distribution of $\hat{\theta}|\theta$. For more general, but nevertheless continuous functions f with continuous first derivates, we can find the asymptotic distribution of $f(\hat{\theta})|f(\theta)$, which enables us to conduct tests and construct confidence sets that are only asymptotically valid. Examples follow below. If all else fails, we can numerically approximate the finite-sample distribution of $f(\hat{\theta})|f(\theta)$ by drawing $\{\hat{\theta}^m\}_{m=1}^M$ from the finite-sample distribution of $\hat{\theta}|\theta$ and computing $\{f(\hat{\theta}^m)\}_{m=1}^M$. In turn, we can numerically conduct hypothesis tests and construct confidence sets as laid out in Chapter 2.[5]

An example of the first case is $f(\beta) = x'\beta$ in the linear regression model, provided that we condition on X. We have $\hat{\beta}|X, \beta \sim N(\beta, V)$ with $V = \sigma^2(X'X)^{-1}$, and therefore we know that $x'\hat{\beta}|X, \beta \sim N(x'\beta, x'Vx)$. An example of the second case is $f(\beta) = \phi(x'\beta)\beta$ in the probit model. We have $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V)$, and because $f(\beta) = \phi(x'\beta)\beta$ is a continuous function with continuous first derivatives $F(\beta) = \frac{\partial f(\beta)}{\partial f(\beta)} = \phi(x'\beta)[I - x\beta']$, the Delta method

[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 $f(\hat{\theta})|f(\theta)$. 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.

(Proposition 11) tells us that $\sqrt{n}(f(\hat{\beta}) - f(\beta)) \xrightarrow{d} N(0, F(\beta)VF(\beta)')$.[6]

7.3 Generalized Method of Moments

The Generalized Method of Moments (GMM) finds θ s.t. a certain moment condition $\mathbb{E}[h(y_i;\theta)] = 0$ holds. For example, this could be the first-order condition (FOC) from some theoretical model.[7]

Suppose we have r moment conditions for $k \leq r$ unknowns in θ, i.e. $\theta \in \mathbb{R}^k$ while $h(y_i;\cdot) : \mathbb{R}^k \to \mathbb{R}^r$. Let θ_0 be the unique value of θ that solves the moment condition:

$$\mathbb{E}[h(y_i;\theta_0)] = 0 \quad \text{and} \quad \mathbb{E}[h(y_i;\theta_0)] \neq 0 \; \forall \; \theta \neq \theta_0 \; .$$

For example, for the linear regression model, we have $\mathbb{E}[x_i(y_i - x_i'\beta)] = 0$ iff $\beta = \beta_0$.[8] Similarly, for the probit model, we have $\mathbb{E}[x_i(y_i - \Phi(x_i'\beta))] = 0$ iff $\beta = \beta_0$.

Let

$$\hat{\theta} = \arg\min_{\theta \in \Theta} Q_n(\theta, Y^n) \; , \quad Q_n(\theta, Y^n) = \frac{1}{2} g_n(\theta; Y^n)' W_n g_n(\theta; Y^n) \; ,$$

where $g_n(\theta; Y^n) = \frac{1}{n}\sum_{i=1}^n h(y_i;\theta)$ and where W_n is an $r \times r$ symmetric, p.d. weighting matrix with probability limit W. Assuming that the optimum is interior, $\hat{\theta}$ solves the FOC

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

Sometimes, we can solve this FOC analytically for $\hat{\theta}$. In other cases we use numerical optimization techniques from Section 8.1.

We can verify consistency and asymptotic Normality of $\hat{\theta}$ by checking the conditions for extremum estimators laid out in Chapter 6. That discussion makes clear that we need to establish that $Q_n(\theta; Y^n) \xrightarrow{p} Q(\theta)$ uniformly, that $Q(\theta)$ is continuous, and that it is uniquely minimized at $\theta = \theta_0$. In this, we are greatly helped by the particular form that Q_n takes under GMM. First, $g_n(\theta; Y^n) \xrightarrow{p} \mathbb{E}[h(y_i;\theta)]$ uniformly guarantees that $Q_n(\theta; Y^n) \xrightarrow{p} Q(\theta) =$

[6] 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 $f(\hat{\beta})$ for this function f, only its asymptotic one.

[7] With time series data, the canonical example is the Euler equation from a dynamic macroeconomic model:

$$c_t^{-\sigma} = \beta \mathbb{E}_t[c_{t+1}^{-\sigma}(1 + r_{t+1})] \quad \Rightarrow \quad \mathbb{E}\left[c_{t+1}^{-\sigma}\beta(1 + r_{t+1}) - c_t^{-\sigma}\right] = 0 \; ,$$

where c_t is the household's consumption, r_{t+1} is the return on risky assets, β is the discount factor and σ is the risk aversion coefficient.

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

$\frac{1}{2}\mathbb{E}[h(y_i;\theta)]'W\mathbb{E}[h(y_i;\theta)]$ uniformly. This occurs if $h(y_i;\theta)$ satisfies the ULLN. Second, $Q(\theta)$ is continuous if $h(y_i;\theta)$ is continuous. Third, given that W is p.d., $\mathbb{E}[h(y_i;\theta)]$ being uniquely minimized by θ_0 guarantees that $Q(\theta)$ is.

The asymptotic analysis of GMM estimators reveals insights about desirable properties of the weighting matrix W_n and the optimal number of moment conditions, and it motivates a test for whether the GMM conditions are correctly specified. The derivations in the Appendix show that

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

where $D' = \mathbb{E}[h^{(1)}(\theta_0; y_i)]$ and $S = \mathbb{E}[h(\theta_0; y_i)h(\theta_0; y_i)']$. We can estimate this asymptotic variance by replacing D' and S with $\hat{D}' = \frac{1}{n}\sum_{i=1}^{n} h^{(n)}(\hat{\theta}; y_i)$ and $\hat{S} = \frac{1}{n}\sum_{i=1}^{n} h(\hat{\theta}_n i y_i)h(\hat{\theta}_n; y_i)'$, respectively.

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

The asymptotic analysis in the Appendix also reveals that to minimize the asymptotic variance, one should use as many moment conditions as possible. 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 the optimal weighting matrix is used).

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

$$\mathcal{H}_0 : \exists\ \theta_0 \text{ s.t. } \mathbb{E}[h(\theta_0; Y^n)] = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \nexists\ \theta_0 \text{ s.t. } \mathbb{E}[h(\theta_0; Y^n)] = 0 .$$

The test-statistic is easily derived, as under $\mathcal{H}_0$, $\frac{1}{\sqrt{n}}\sum_i h(\theta_0; y_i) = \sqrt{n}g_n(\theta_0; Y^n) \xrightarrow{d} N(0, S)$. Therefore,

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

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 $\hat{\theta}_n$ above), we get

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

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

7.4 Instrumental Variable Estimation

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

Suppose we have a variable $z_i \in \mathbb{R}^r$. The r variables in z_i are valid IVs if z_i satisfies two conditions:

1. Relevance: $\mathbb{E}[z_i x_i'] \neq 0$.
2. Exogeneity: $\mathbb{E}[z_i u_i] = 0$, i.e. z_i is uncorrelated with the error term u_i.

The first condition is often stated as requireing that z_i is correlated with the regressor x_i.

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. For illustration purposes, suppose x_i is a scalar and derive $\hat{\beta}_{2SLS}$ in two steps:

1. Estimate $x_i = z_i'\gamma + e_i$ to get $\hat{\gamma} = (Z'Z)^{-1}Z'X$ and $\hat{X} = Z\hat{\gamma} = P_Z X$.
2. Estimate $y_i = \hat{x}_i'\beta + u_i^*$ to get $\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'P_Z X)^{-1}X'P_Z Y$.

Provided that our model is correctly specified and the above two assumptions are satisfied, $\hat{\beta}_{2SLS}$ is consistent: by WLLN,

$$\begin{aligned}
\hat{\beta}_{2SLS} &= \left[X'P_Z X\right]^{-1} X'P_Z Y \\
&= \beta + \left[X'P_Z X\right]^{-1} X'P_Z U \\
&= \beta + \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'U \\
&\xrightarrow{p} \beta + Q_{xz}^{-1}\mathbb{E}[x_i z_i']\mathbb{E}[z_i z_i']\mathbb{E}[z_i u_i] = \beta \, ,
\end{aligned}$$

[9]

where $Q_{xz} = \mathbb{E}[x_i z_i'] \mathbb{E}[z_i z_i'] \mathbb{E}[x_i z_i']'$. To compute $\hat{\beta}_{2SLS}$, we need $Z'Z$ to be of full rank. Based on the analogous discussion of $X'X$ in Section 3.1, $rank(Z'Z) = r$ requires us to have $n > r$, i.e. more observations than IVs, and it prevents perfect multicollinearity of IVs.[10]

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. To see this, note that the conditional variance of $\hat{\beta}_{2SLS}$ is

$$\mathbb{V}[\hat{\beta}_{2SLS}|X, Z] = (X'P_Z X)^{-1} X'P_Z \mathbb{E}[UU'|X, Z] P_Z X (X'P_Z X)^{-1} .$$

Under homoskedasticity, $\mathbb{E}[UU'|X, Z] = \sigma^2 I$ and we obtain $\mathbb{V}[\hat{\beta}_{2SLS}|X, Z] = \sigma^2 (X'P_Z X)^{-1}$. This variance is larger than that of $\hat{\beta}_{OLS}$, $\mathbb{V}[\hat{\beta}_{OLS}|X] = \sigma^2 (X'X)^{-1}$. To see this, note that

$$\Delta = X'X - X'P_Z X = X'M_Z X = (M_Z X)'(M_Z X)$$

is p.d. It is the sum of squared residuals from the first-stage regression of X on Z. Therefore, the more variation in X is explained by Z, i.e. the higher the R^2 in the first-stage regression, the smaller is the efficiency loss of 2SLS compared to OLS.

By the usual arguments, the asymptotic analysis reveals that $\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, V_{2SLS})$ with the huge but simple-to-derive expression for V_{2SLS}:

$$V_{2SLS} = Q_{xz}^{-1} \mathbb{E}[x_i z_i'] \mathbb{E}[z_i z_i'] \mathbb{E}[z_i z_i' u_i^2] \mathbb{E}[z_i z_i']' \mathbb{E}[x_i z_i']' Q_{xz}^{-1} .$$

As usual, we can estimate it by replacing u_i with $\hat{u}_i$ and expectation operators with population means. Thereby, it is important to note that $u_i \neq u_i^*$, i.e. o obtain $\hat{u}_i$, we use regressors x_i and not $\hat{x}_i$: $\hat{u}_i = y_i - x_i' \hat{\beta}_{2SLS}$. Under homoskedasticity, V_{2SLS} simplifies to $V_{2SLS} = \sigma^2 Q_{xz}^{-1}$, which we estimate using $\hat{\sigma}^2 = \frac{1}{n} \sum_i u_i^2$.

Note that the 2SLS estimator does not actually have to be carried out in two steps. It simply involves running a regression of Y on $P_Z X$ instead of X. Relatedly, 2SLS can trivially be applied for more than one regressor in x_i. The only thing that changes is that X is not necessarily an $n \times 1$ vector but more generally an $n \times k$ matrix, where $k \geq 1$. Typically, not all variables in x_i are suspected to be endogenous. While one could in principle replace

[9] To see this, note that we can write out

$$\left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'U = \left[\left(\sum_{i=1}^n x_i z_i'\right)\left(\sum_{i=1}^n z_i z_i'\right)\left(\sum_{i=1}^n x_i z_i'\right)'\right]^{-1} \left(\sum_{i=1}^n x_i z_i'\right)\left(\sum_{i=1}^n z_i z_i'\right)\left(\sum_{i=1}^n z_i u_i\right) ,$$

and we can divide each of the sums by n.

[10] We also need and $X'P_Z X = (P_Z X)'(P_Z X)$ to be of full rank. This condition, $rank(X'P_Z X) = k$, is somewhat more subtle. It is related to the relevance condition above and, intuitively, requires Z to explain some variation in X. With $k = 1$ regressor in x_i, it is equivalent to saying that $\hat{X} = P_Z X$ is not zero.

only the endogenous variables $x_{i,m}$ with their predicted values $\hat{x}_{i,m}$ obtained in regressions on IVs, with possibly different IVs for different variables, a more straightforward approach is to let z_i include not only the actual IVs (all of them), but also all exogenous variables from x_i, and construct the above 2SLS estimator.[11] This guarantees that $r \geq k$, i.e. we have at least as many IVs as endogenous variables.

The use of IVs has been popularized by causal inference methods in the context of the estimation of so-called local average treatment effects (LATEs). It is discussed in Section 13.3.1, which contains examples of IV-analyses.

The rest of this section discusses the challenges posed by weak IVs and GMM-based IV estimation. Likelihood-based IV estimation is touched upon in the Appendix.

Weak Identification in IV Models If the the correlation between x_i and z_i is rather low, we speak of weak IVs. Under weak IVs, the finite sample distribution of $\hat{\beta}$ may not resemble the asymptotic one at all. This needs to be taken into account when conducting inference, i.e. testing hypotheses and constructing confidence sets.

The asymptotic analysis can be adjusted to take into account the weak correlation of x_i and z_i. However, this is more interesting from a methodological than an empirical point of view. The approach is sketched in the Appendix.

In absence of an asymptotic distribution that is useful for approximating the finite sample distribution, we can conduct inference using its numerical approximation via bootstrapping (Section 7.1). Alternatively, we can construct a confidence set for β under weak IVs using the inference procedure of Anderson and Rubin (1949). It is based on the insight that, for $\beta = \beta_0$, the auxiliary regression $y_i - x_i'\beta = \delta z_i + v_i$ should yield $\delta = 0$, because $y_i - x_i'\beta_0 = u_i$ and u_i and z_i are uncorrelated. Suppose for simplicity that z_i is a scalar. 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 $\mathscr{H}_0 : \delta = 0$. As δ is a scalar, we can use the t-test $t_\delta(\beta_0) = \hat{\delta}(\beta_0)/\sqrt{\hat{\sigma}_v^2/Z'Z} \xrightarrow{d} N(0,1)$. A confidence set for β is obtained by taking all β_0 for which $\mathscr{H}_0 : \delta = 0$ cannot be rejected. Note that this can give very large, unbounded and even disconnected or empty confidence sets.

[11] It is straightforward to see that for the exogenous variables in x_i, predicted values will equal the actual values, as regressing a variable on itself (and other covariates) generates $R^2 = 1$. Regarding changing IVs for different endogenous variables: if indeed a set of IVs is irrelevant for one of the endgenous variables, their coefficient will be close to zero, not impacting the predicted value much.

GMM-Based IV Estimation Suppose we have k regressors and r IVs: $x_i \in \mathbb{R}^k$ and $z_i \in \mathbb{R}^r$, where $r \geq k$.[12] Based on the validity condition $\mathbb{E}[z_i u_i] = 0$, we can define the GMM-moment condition

$$\mathbb{E}[h(\beta, w_i)] = 0 \ , \quad h(\beta, w_i) = z_i(y_i - x_i'\beta) \ ,$$

where $w_i = (y_i, x_i', z_i')'$. If $\mathbb{E}[z_i x_i']$ has full rank, it is easy to verify that the moment condition is satisfied only by β_0, i.e. $\mathbb{E}[h(\beta, w_i)] = 0$ iff $\beta = \beta_0$.

The GMM-based analysis of IV methods translates the IV approach into the known framework of GMM estimation, with several advantages. First, it suggests a way to reduce the asymptotic variance of 2SLS by choosing the optimal weighting matrix. Second, it enables the J-specification test, which in the context of IV estimation amounts to testing whether indeed all IVs are exogenous. Third, it renders IV estimation more flexible, enabling, for example, explicitly the use of different IVs for different variables or for different observations.[13]

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

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

The 2SLS estimator is obtained with the weighting matrix $W_n = \left[\frac{1}{n}\sum_{i=1}^n z_i z_i'\right]^{-1}$. In just-identified models – i.e. under $r = k$ – we get $\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y$, regardless of W_n.

Using the same arguments as above, $\hat{\beta}_{IV}$ is consistent for β provided that the model is correctly specified and we have valid instruments, i.e. $y_i = x_i'\beta + u_i$ and $\mathbb{E}[z_i u_i] = 0$. By the analysis in Section 7.3, we have

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

with $D = \mathbb{E}[z_i x_i']$ and $S = \mathbb{E}[z_i u_i (z_i u_i)']$. Under 2SLS, we have $W = \mathbb{E}[z_i z_i']^{-1}$, which gives the above asymptotic variance V_{2SLS}. However, a lower asymptotic variance can

[12] Recall that z_i includes all exogenous variables from x_i along with the actual IVs. Therefore, this condition requires us to have at least as many IVs as endogenous variables.

[13] This is required, for example, in the Fixed Effects (FE)-IV estimation method of Arellano and Bond (1991) in the context of panel data (see Section 12.3). There, the endogenous variables are instrumentalized by different sets of IVs, whereby the IVs from one observation are not just irrelevant for another observation, but even violate the exogeneity assumption and therefore must not be used.

be obtained by taking an optimal weighting matrix s.t. $W_n \xrightarrow{p} S^{-1}$, for example $W_n = \left[\frac{1}{n}\sum_{i=1}^n z_i z_i' \hat{u}_i^2\right]^{-1}$. It is easy to verify that the two coincide under homoskedasticity and amount to $\sigma^2(D\mathbb{E}[z_i z_i']^{-1}D)^{-1}$.

Appendix

GMM Estimation: Asymptotic Analysis and Results

By CLT,

$$\sqrt{n}g_n(\theta_0; Y^n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h(\theta_0; y_i) \xrightarrow{d} N(0, S) \ ,$$

where $S = \mathbb{E}[h(\theta_0; y_i)h(\theta_0; y_i)']$, and we used the fact that $\mathbb{E}[h(\theta_0; y_i)] = 0$. By WLLN,

$$g^{(1)}(\theta_0; Y^n) = \frac{1}{n}\sum_{i=1}^n h^{(1)}(\theta_0; y_i) \xrightarrow{p} D' \ ,$$

where $D' = \mathbb{E}[h^{(1)}(\theta_0; y_i)]$. 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

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

Putting all pieces together, we get

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

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 $m \leq r$. Also, let

$$\tilde{\theta}_n = \arg\min_\theta \frac{1}{2} g_n(\theta; Y^n)' B' W B g_n(\theta; Y^n) \ ,$$

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. ■

Asymptotic Analysis under Weak IVs

Consider again our two 2SLS regressions

$$x_i = z_i'\gamma + e_i \quad \text{and} \quad y_i = \hat{x}_i'\beta + u_i^* \ .$$

A better approximation of the finite sample distribution of IV-based estimators of β is obtained by making the parameter γ "local to zero", i.e. assuming that $\gamma = \tilde{\gamma}/\sqrt{n}$ 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 that $r = k = 1$ and suppose that u_i is homoskedastic. We then 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 $\hat{\beta}$.

Likelihood-Based 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.