GMM and SMM Summary

Based on

1. Hansen, L. 1982. "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica,* 50, p. 1029-54.
2. Lee, B.S. and B. Ingram. 1991 "Simulation estimation of time series models", *Journal of Econometrics*, 47, p. 197-205.

1 GMM

- Suppose there are n necessary conditions of the model:

$$E_t\left[u_{t+1}(x_{t+1}, b_0)\right] = 0 \tag{1}$$

  where

  - u_{t+1} is an $n \times 1$ vector of "errors"
  - x_{t+1} is a $k \times 1$ vector of data
  - b is an $\ell \times 1$ vector of parameters where b_0 stands for the true parameter vector

- Recall the following order conditions necessary for identification:

  - If $\ell < n$, the model is said to be overidentified.
  - If $\ell = n$, the model is said to be just identified.
  - If $\ell > n$, the model is said to be underidentified.

- If $\ell > n$, we cannot uniquely identify parameters. In that case we need more conditions. If z_t is a $q \times 1$ vector of variables in the econometricians information set, then from (1) and the law of iterated expectations we know

$$E_t\left[u_{t+1}(x_{t+1}, b_0) \otimes z_t\right] = 0 \otimes z_t = 0 \Longrightarrow E\left[u_{t+1}(x_{t+1}, b_0) \otimes z_t\right] = 0 \tag{2}$$

  is an $nq \times 1$ vector, which may satisfy the identification order condition $\ell \leq nq$.

- Letting $f(x_{t+1}, z_t, b) \equiv u_{t+1}(x_{t+1}, b) \otimes z_t$, define the moment $nq \times 1$ moment vector

$$g(b) \equiv E[f(x_{t+1}, z_t, b)] \tag{3}$$

  (i.e. the unconditional average error). By (2), $g(b_0) = 0$.

- The sample analogue of (3) is the $nq \times 1$ vector

$$g_T(b) \equiv \frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}, z_t, b) \tag{4}$$

The basic idea of GMM is that as $T \to \infty$, (1) implies $g_T(b_0) = 0$.

- Assuming that $g_T(b)$ is continuous in b, the GMM estimate of b solves

$$b_T = \arg\min_b J_T(b) \tag{5}$$

where $J_T(b) \equiv g_T'(b) W_T g_T(b)$ (which is $(1 \times nq)(nq \times nq)(nq \times 1)$) is a weighted sum of squared errors and W_T is an arbitrary weighting $(nq) \times (nq)$ matrix that can depend on the data.

- In the just identified case, the weighting matrix does not matter. To see this, consider the following $\ell = n = 2$ case:

$$\min_{b_1, b_2} w_1 g_1(b_1, b_2)^2 + w_2 g_2(b_1, b_2)^2$$

The foc are:

$$\begin{aligned} b_1 &: \quad 2w_1 g_1(b)\nabla_{b_1} g_1 + 2w_2 g_2(b)\nabla_{b_1} g_2 = 0 \\ b_2 &: \quad 2w_1 g_1(b)\nabla_{b_2} g_1 + 2w_2 g_2(b)\nabla_{b_2} g_2 = 0 \end{aligned}$$

Since there are 2 equations in 2 unknowns, one would think that $b(W)$. However, in this case b is not a function of W. Rewriting the 2 foc in matrix notation

$$[g_1(b)\ g_2(b)] \begin{bmatrix} w_1\nabla_{b_1} g_1 & w_1\nabla_{b_2} g_1 \\ w_2\nabla_{b_1} g_2 & w_2\nabla_{b_2} g_2 \end{bmatrix} = [0\ 0]$$

Then provided the 2×2 matrix is invertible, we have

$$[g_1(b) g_2(b)] = [0\ 0] \begin{bmatrix} w_1\nabla_{b_1} g_1 & w_1\nabla_{b_2} g_1 \\ w_2\nabla_{b_1} g_2 & w_2\nabla_{b_2} g_2 \end{bmatrix}^{-1} = [0\ 0].$$

1.1 Consistency

- Under certain conditions, Hansen 1982 (Theorem 2.1) proves that this estimator b_T exists and converges in probability to b_0.

- It is essential for consistency that the limit $J_\infty(b)$ have a unique minimum at the true parameter value b_0.This condition is related to identification; the distribution of the data at b_0 is different than that at any other possible parameter value.

- The conditions are:

- $W_T \to W$ in probability, where W is a positive semi-definite matrix
- $g(b) = 0$ (an $nq \times 1$ vector) only for $b = b_0$.
- $b_0 \in B$ (a compact set)
- $f(x, z, b)$ is continuous at each b
- $E[\sup_b \|f(x, z, b)\|] < \infty$.

- The second condition (known as **Global Identification**) is hard to verify. A simpler necessary but not sufficient condition is known as **Local Identification**. If $g(b)$ is continuously differentiable in a neighborhood of b_0, then the matrix $\nabla_b g(b)$ (which is $(nq \times \ell)$) must have full column rank (i.e. there are ℓ linearly independent columns).

- Hansen 1982 (Theorem 3.1) establishes asymptotic normality of the estimator.

1.2 Efficiency

- While the above result shows that the GMM estimator is consistent for arbitrary weighting matrices (e.g. $W = I$), it is not necessarily efficient. Hansen (1982, Theorem 3.2) shows that the statistically optimal weighting matrix $W^* = S^{-1}$ where the asymptotic variance covariance matrix is:

$$S = \sum_{j=-\infty}^{\infty} E\left[f(x_t, z_{t-1}, b_0) f(x_{t-j}, z_{t-j-1}, b_0)'\right]. \tag{6}$$

- Why does this weighting matrix make sense? Some moments will have more variance than others. This downweights errors from high variance moments.

- Hansen (1982, Theorem 3.2) shows that the asymptotic distribution of the estimator when $W^* = S^{-1}$ is given by

$$\sqrt{T}(b_T - b_0) \to N(0, \left[\nabla_b g(b_0)' S^{-1} \nabla_b g(b_0)\right]^{-1}) \tag{7}$$

  where $\nabla_b g(b_0)' S^{-1} \nabla_b g(b_0)$ is an $(\ell \times nq)(nq \times nq)(nq \times \ell)$ matrix.

- The problem is that we do not know S^{-1} nor $g(b)$. If the errors are serially uncorrelated, then a consistent estimate of the asymptotic var-covar matrix S is given by

$$\widehat{S}_T = \frac{1}{T} \sum_{t=1}^{T} f(x_{t+1}, z_t, \widehat{b}_T) f(x_{t+1}, z_t, \widehat{b}_T)'$$

  where b_T is a consistent estimate of b_0.[1] In this case, the distribution of the estimator is given by

$$\sqrt{T}(b_T - b_0) \to N(0, \left[\nabla_b g_T(b_T)' S_T^{-1} \nabla_b g_T(b_T)\right]^{-1}).$$

[1] For the case in which $f(x_{t+1}, z_t, b_0)$ is serially correlated you can use the Newey-West (1987) correction.

- Notice that the precision of the estimates is related to $\nabla_b g_T(b_T)$. If the objective is very sensitive to changes in the parameters (i.e. $\nabla_b g_T(b_T)$ is high), then there will be a low variance of the estimate (since $\nabla_b g_T(b_T)' S_T^{-1} \nabla_b g_T(b_T)$ is inverted). If the objective is not very sensitive to changes in the parameters, it will produce a high variance for the estimates. Simply put, this suggests that if you find big standard errors, it is because the objective is not very sensitive to changes in the parameters so it is hard to find the true unique minimum. This is how local identification is linked to standard errors.

- To implement this, use a two step procedure: (i) the first stage estimate of b minimizes a quadratic form of the sample mean of errors for $W = I$, which is consistent; (ii) estimate a var-covar matrix of the residuals S_T from the first stage to form $W_T = S_T^{-1}$ in a second stage minimization of $g_T'(b) S_T^{-1} g_T(b)$. This is like the two step procedure in Generalized Least Squares.

- Testing Overidentifying Restrictions: Hansen (1982, Lemma 4.2) shows

$$T g_T'(b_T) S_T^{-1} g_T(b_T) \to \chi^2(nq - \ell). \tag{8}$$

  That is, the minimized value of the objective function is distributed as a chi-squared r.v. with degrees of freedom equal to the #moments-#parameters. This is known as the J-test.

- You often want to compare one model to another. If one model can be expressed as a special or "restricted" case of the other "unrestricted" model, we can conduct something like a likelihood ratio test. If we use the same S matrix (usually that of the unrestricted model), then J_T(restricted) must rise. If the restricted model is really true, it should not rise too much. Thus

$$T J_T(\text{restricted}) - T J_T(\text{unrestricted}) \sim \chi^2(\#\text{ of restrictions})$$

  This is Newey-West's (1987, IER) D-test.

2 SMM

- Let $\{x_t\}_{t=1}^T$ be a realization of an $k \times 1$ vector valued stationary and ergodic stochastic process generating the observed data (e.g. detrended GDP).

- Let $\{y_t(b)\}_{t=1}^T$ be a realization of an $k \times 1$ vector valued stationary stochastic and ergodic process generating the simulated data (e.g. GDP generated by the model) where b is an $\ell \times 1$ vector of parameters. In general we may take H simulations of length T.

- Let $M_T(x)$ be an $n \times 1$ vector of data moments (e.g. standard deviation of detrended GDP) and $M_N(y(b))$ be a $n \times 1$ vector of model moments of the simulated data where $N = H \cdot T$.

- Assume that $M_T(x) \overset{a.s.}{\to} \mu(x)$ as $T \to \infty$ and that $M_N(y(b)) \overset{a.s.}{\to} \mu(y(b))$ as $N \to \infty$ where $\mu(x)$ and $\mu(y(b))$ are the population moments.

- Furthermore, under the null that the model is correct at the true parameter vector b_0, then $\mu(x) = \mu(y(b_0))$. If you understand this equality you understand everything you need to know about economics. It says there is a link between data and theory.

- In summary, x_t is observed data (which we may not even have), y_n is simulated data, $M_T(x)$ is observed moments (which we will assume we have), $M_N(y(b))$ is simulated moments, and the reason we can use the model to say something about the data we don't have is that if the model is true, then the asymptotic models have to be equal at the true parameter values (i.e. $\mu(x) = \mu(y(b_0))$).

- Given a symmetric $n \times n$ weighting matrix W_T (which may depend on the data - hence the subscript T), Lee and Ingram show that under certain conditions the simulation estimator $\widehat{b}_{TN}$ which minimizes the weighted sum of squared errors of the model moments from the data moments - ie. the solution to

$$\widehat{b}_{TN} = \arg\min_b [M_T(x) - M_N(y(b))]' W_T [M_T(x) - M_N(y(b))] \qquad (9)$$

  - is a consistent and asymptotically normal estimator of b_0.

- Basically, SMM is just GMM where the errors are just the difference between the data moment and the model moment $g_{TN} = M_T - M_N(y(b))$, i.e. the difference between the data moment and the model moment.

- Since the solution to this problem is essentially a special case of Hansen's (1982) GMM estimator, the conditions are from his paper: (i) x and $y(b)$ are independent; (ii) the model must be identified; and (iii) $M_N(y(b))$ must be continuous in the mean.

- You can think of the estimation as being conducted in a sequence of two steps (or calls of functions):

1. For any given value of b,say b^i,

   (a) simulate artificial data from the model (in the context of the growth model, this would be H draws of $\{\varepsilon_t\}_{t=1}^T$ which implies a realization of technology shocks and then via decision rules (which depend on parameters b^i) a realization of a variable of interest like real output $y(b^i)$), and

   (b) compute a moment (i.e. $M_N(y(b^i))$), and evaluate the objective function $J_{TN}(b^i) = [M_T(x) - M_N(y(b^i))]' W [M_T(x) - M_N(y(b^i))]$; and

2. choose a new value for the parameters, say b^{i+1}, for which $J_{TN}(b^{i+1}) \leq J_{TN}(b^i)$.

- A minimization routine constructs this sequence of smaller and smaller $J_{TN}(b^i)$ for you. **You must use the same random draw throughout each simulation** of the artificial data or else you wouldn't know whether the change in the objective function were coming from a change in the parameter or a change in the draw.

- As before, to obtain the optimal weighting matrix, you use a two stage procedure. See Section 4.2.3 on Asymptotic Properties of Indirect Inference Estimators in Gourieroux and Monfort (1996, Simulation Based Econometric Methods, Oxford University Press) for justification of this process.

  - In the first stage, minimize $J^1_{TN}(b)$ constructed using $W = I$.
  - Now the second stage is different since above, you had a vector of residuals to construct the variance-covariance matrix. Here, you don't have enough "data" to get a var-covar matrix.
  - Since the resulting estimate $\widehat{b}^1_{TN}$ of b_0 is consistent, generate H repetitions of model moments (from T length simulated samples) analogous to the data moments in order to construct an estimate of the var-covar matrix $\widehat{S}_T$ of the "data moments".
  - Use $W_T = \widehat{S}_T^{-1}$ to construct the second stage J^2_{TN} and obtain the corrected estimate $\widehat{b}^2_{TN}$.

- Once you have the optimal weighting matrix, you can generate standard errors as in (7) of Hansen (1982).

- Alternatively, you can run a Monte Carlo experiment to compute standard errors of the estimates. Recall that each estimate of $\widehat{b}_{TN}$ is derived for a given HT draw of the shocks ε_t to the underlying data generation process. Different draws will generate different estimates of $\widehat{b}_{TN}$. You can generate a histogram and summary statistics (mean and sd of $\widehat{b}_{TN}$) which are interpretable as the point estimate and standard error of b.