Lecture 14
"GEE-GMM"

Throughout the course we have emphasized methods of estimation and inference based on the principle of maximum likelihood which required a complete specification of the probability model describing the mechanism generating the data. Even when we considered M-estimations and the idea of quasi-maximum likelihood in which the specified probability model was admittedly not *the true* data generating mechanism, it nevertheless presumed that we had specified a complete probabilistic model for the data, and KLIC was used to implicitly define the relevant population parameters of this model.

Occasionally, we have departed from this paradigm. In quantile regression, for example, we need not provide a fully specified probabilistic model; we require only a specification of the parametric model describing a single conditional quantile function. Of course, if we proceed to specify a parametric model for *all* conditional quantile functions we are back, essentially, to the framework underlying the MLE. Another example of a partially specified parametric model is, of course, the classical linear regression model which we may interpret as simply a specification of the parametric form of the conditional mean function. Strengthening the assumptions to specify a form for the conditional density at each design point, x_i, would complete the full probability model for the data, but this is certainly not necessary to justify ordinary regression M-estimators. These examples illustrate the method-of-moments estimation paradigm. A parametric model is specified by asserting the parametric form of certain observable functions of the data. Then under certain rank conditions on the Jacobian of these functions, they can be "inverted", or "solved" for the parameters of interest.

This approach to estimation has been advanced by Sims and elaborated by Hansen (1982) and proven highly successful, particularly in addressing specification problems in macroeconomics where fully parametric statistical models are difficult, but certain orthogonality conditions may be used to specify identifying moment conditions.

Actually, the prime examples of such specification of models and associated methods of estimation are classical regression and instrumental variables estimators in econometrics. Consider the linear model,

$$y_i = x_i'\beta + u_i \qquad\qquad u = 1, \ldots, n$$

and suppose we have instruments, $\{z_i\}$, so we claim that

$$Ez_iu_i = 0$$

i.e., that the IV's are orthogonal to the u_i's and that

$$Ez_ix_i'$$

is (asymptotically) invertible, requiring at the very least that there be as many coordinates to z_i as to β. If the latter condition is exactly satisfied so we have, β, exactly identified by the

orthogonality conditions we can define the IV estimator as

$$\begin{aligned}\hat{\beta} &= (\sum z_i x_i')^{-1} \sum z_i y_i \\ &= (Z'X)^{-1} Z'y.\end{aligned}$$

Under the classical condition that

$$Euu' = \sigma^2 I$$

it is easy to see that

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow \mathcal{N}(0, \sigma^2 (X'Z/n)^{-1} Z'Z/n (Z'X/n)^{-1})$$

provided these matrices are invertible and the Lindeberg condition

$$\max_i \| z_i \|^2 / \sum \| z_i \|^2 \to 0$$

holds. (The latter condition is needed for the CLT on $n^{1/2} \sum z_i u_i$ to apply.)

The condition that the matrix $Z'X/n$ be invertible, which is really an identifiability condition, is obviously quite restrictive. In general, we might hope to have *more* IV's than elements of β, i.e., have a parametric model which is overidentified. In this case we can not expect to satisfy the empirical counterpart of our population orthogonality conditions

$$EZ'u = 0$$

exactly, since we have a system of $q > p$ equations

$$Z'\hat{u} = Z'(y - X\beta) \equiv m(\beta)$$

in only $p = \dim(\beta)$ unknowns. It is reasonable to replace the infeasible requirement that $Z'\hat{u} = 0$, by the requirement that it be small in some reasonable norm, e.g.,

$$\hat{u}' Z A^{-1} Z' \hat{u} = \min!$$

The immediate question is then: how should we choose the matrix A and a natural response would be to set A equal to the inverse of the covariance matrix of the orthogonality conditions, i.e.,

$$A = \text{ Cov } (Z'u) = EZ'uu'Z = \sigma^2 Z'Z$$

This leads to the estimator

$$\begin{aligned}\hat{\beta} &= \text{ argmin } \hat{u}(\beta)' P_Z \hat{u}(\beta) \\ &= (X' P_Z X)^{-1} X' P_Z y\end{aligned}$$

where $P_Z = Z(Z'Z)^{-1}Z'$ and thus gives us the familiar two stage least squares estimator.

This approach can be extended in various ways. If $Euu' = \Omega$, then it is clear that we should take

$$A = Z'\Omega Z$$

and $\hat{\beta}$ is adapted accordingly. If the "response" function, i.e., the conditional mean function, is nonlinear in parameters, so for example

$$\hat{u}_i(\beta) = y_i - g(x_i, \beta)$$

we may still define $\hat{\beta}$ as the minimizer

$$\hat{\beta} = \text{ argmin } \hat{u}(\beta)' P_Z \hat{u}(\beta)$$

where we now have the nonlinear two stage least squares estimator. It is straightforward to derive the asymptotic behavior of this estimator using our usual strategy. Let

$$J = \nabla_\beta \hat{u}(\beta)$$

denote the Jacobian matrix of the model, so optimality would require that

$$J' P_Z \hat{u}(\beta) = 0$$

at $\beta = \hat{\beta}$. Expanding this condition around $\beta = \beta_0$, the true parameter and evaluating at $\beta = \hat{\beta}$ we obtain

$$0 = J' P_Z(\hat{u}(\beta_0) + J(\hat{\beta} - \beta_0)) + (\hat{\beta} - \beta_0)' H' P_Z \hat{u}(\beta_0) + R$$

where $H = \nabla^2 \hat{u}(\beta)$ and R is the remainder.

Then writing

$$\sqrt{n}(\hat{\beta} - \beta_0) = (J' P_Z J/n + H' P_Z \hat{u}(\beta_0)/n)^{-1} n^{-1/2} J' P_Z \hat{u}(\beta_0) + o_p(1)$$

and noting that the Hessian term tends to zero we have the linear representation,

$$\sqrt{n}(\hat{\beta} - \beta_0) = (J' P_Z J/n)^{-1} n^{-1/2} J' P_Z \hat{u}(\beta_0) + o_p(1)$$

and assuming that $E\hat{u}(\beta_0)\hat{u}(\beta_0)' = \sigma^2 I$ we have

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightsquigarrow \mathcal{N}(0, \sigma^2 J' P_Z J/n)$$

How would this change if $\sigma^2 I$ were Ω? What should be used in place of P_Z for this case? If we continue to use P_Z, we obtain

$$V = (J P_Z J)^{-1} J' P_Z \Omega P_Z J (J P_Z J)^{-1}$$

for the limiting form of the covariance matrix. However, as above, if we replace P_Z by

$$P_Z^* = Z(Z'\Omega Z)^{-1} Z'$$

then we see that V collapses to

$$V^* = (J' P_Z^* J)^{-1}$$

It is easy to see that $V^* << V$, in the sense of positive definite matrices so P_Z^* is clearly preferred to P_Z. The difficulty, of course, is that we need to estimate Ω or at least the lower dimensional matrix $Z'\Omega Z$. This is very much like the familiar Eicker-White problem and can be handled analogously. In the case of dependence the approach of Newey-West can be adopted.

Inference in GMM estimation

Not surprisingly, GMM inference can be conducted using any of the three approaches already considered in the discussion of likelihood based inference. Clearly Wald tests based on the large-sample theory outlined above are possible. The GMM criterion function can itself be used as a

quasi-likelihood for constructing LR-type tests, and finally we may construct LM tests based on the gradient of the GMM criterion evaluated at the restricted estimate. Of these approaches, the LR approach would appear to offer the most attractive strategy, particularly, in situations in which the objective function may be highly non-quadratic.

Information theoretic approaches to GMM inference

Imbens, Johnson and Spady(1998) considered some alternative approaches to GMM inference based on information theoretic ideas. Following their notation, suppose we have iid observations $\{z_i\}_{i=1}^n$ from F and moment conditions

$$E\psi(Z, \theta_0) = 0$$

where $\theta_0 \in \Re^p$ constitutes a unique solution to the $q >> p$ equations

$$E\psi(Z, \theta) = 0$$

This may be viewed as a GMM problem with criterion function

$$Q(\theta) = n^{-2}\psi' W^{-1}\psi$$

where $\psi = (\psi(z_i, \theta)$ and (optimally) we would try to choose $W = E\psi\psi'$. As we have seen, the minimizer of $Q(\cdot)$, say $\hat{\theta}$, satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, H^{-1}JH^{-1})$$

where $H = E\nabla\psi$, and $J = E\psi\psi'$, and furthermore

$$nQ(\hat{\theta}) \rightsquigarrow \chi^2_{q-p}$$

which provides a test of over-identifying restrictions. Usually, as we have seen this is accomplished in two steps. In the first step we replace W by some preliminary estimate, such as I_q, and in the second step we use this first stage estimator, say $\tilde{\theta}$ to obtain

$$\hat{W} = n^{-1}\sum \psi(z_i, \tilde{\theta})\psi(z_i, \tilde{\theta})'$$

and minimize again to obtain $\hat{\theta}$.

An alternative to this two-step procedure is the empirical likelihood (Owen (1988) and Qin and Lawless (1994)) estimator which solves

$$\begin{aligned} &\max_{\pi,\theta} \sum n^{-1}(\log \pi_i - \log(n^{-1})) \\ &\text{subject to} \\ &\qquad \sum \psi(z_i, \theta)\pi_i = 0 \\ &\qquad \sum \pi_i = 1. \end{aligned}$$

The solution to this problem may be seen to be the solution to the system of equations

$$0 = \begin{pmatrix} \sum t'\nabla\psi(z_i, \theta)/(1 + t'\psi(z_i, \theta) \\ \sum \psi(z_i, \theta)/(1 + t'\psi(z_i, \theta) \end{pmatrix}$$

where $t \in \Re^q$ denotes a vector of q Lagrange multiplier parameters which is called the tilting parameter. This can be seen as follows by concentrating with respect to the π_i's. Differentiating wrt π_i we obtain

$$\frac{n^{-1}}{\pi_i} - \lambda - t'\psi_i = 0$$

multiply by π_i and summing,

$$\sum n^{-1} - \sum \lambda \pi_i - \sum t'\psi_i \pi_i = 0$$

Note that the last term is zero and consequently $\lambda = 1$ so we have from the first equation

$$\pi_i = \frac{n^{-1}}{1 + t'\psi_i}$$

substituting back into the constraint and differentiating wrt to θ and t yields the estimating equations as given. An alternative suggested by Imbens, Johnson and Spady is the exponential tilting estimator which replaces the empirical likelihood term by

$$\begin{aligned} \max_{\pi,\theta} \sum \pi_i (\log(n^{-1}) - \log \pi_i) \\ \text{s.t.} \qquad \sum \psi_i \pi_i = 0 \\ \sum \pi_i = 1 \end{aligned}$$

Note that this just reverses the prior KL divergence expression. A similar argument to the one just employed yields the modified estimating equation,

$$0 = \begin{pmatrix} \sum t'\nabla\psi(z_i,\theta)\exp(t'\psi(z_i,\theta)) \\ \sum \psi(z_i,\theta)\exp(t'\psi(z_i,\theta)) \end{pmatrix}$$

In this approach

$$\pi_i = e^{t'\psi_i} / \sum_j e^{t'\psi_j}$$

and our problem becomes

$$\begin{aligned} \max\{-\sum t'\psi_i \frac{\exp(t'\psi_i)}{\sum \exp(t'\psi_i)} + \log(\sum_j \exp(t'\psi_j))\} \\ \text{s.t.} \qquad \sum \psi_i \frac{\exp(t'\psi_i)}{\sum \exp(t'\psi_i)} = 0 \end{aligned}$$

When the constraints are satisfied the first term is identically zero so we can focus on the second term which we may recognize as the cumulant generating function of ψ. Thus the problem may be written more compactly as,

$$\max_{t,\theta} K(t,\theta) \qquad \text{s.t.} \qquad K_t(t,\theta) = 0$$

where $K_t = \frac{\partial}{\partial t} K(t,\theta)$, higher order derivatives will be denoted similarly $K_{t\theta}, K_{\theta\theta}$, etc. Note that at a solution both $K_t(t,\theta) = 0$ and $K_\theta(t,\theta) = 0$. In practice the constrained problem may be replaced by the unconstrained problem

$$\max_{t,\theta} K(t,\theta) - \frac{1}{2} A K_t' W^{-1} K_t$$

where A is a "large" scalar and W is a positive definite matrix of conformable dimension. A sensible choice of W is

$$W(t,\theta) = K_{tt} + K_t K_t'$$

evaluated at some preliminary estimates of t, θ. The form of this matrix is not crucial since at a solution $K_t = 0$ and the contribution of the penalty term vanishes. A connection with our original formulation of GMM may be made here by noting that

$$\hat{\theta}_{gmm} = \text{ argmax } \{-\frac{1}{2} A K_t(0,\theta)' W(0,\hat{\theta})^{-1} K_t(0,\theta)\}$$

for some preliminary consistent choice of $\hat{\theta}$, since

$$\begin{aligned} W(0,\theta) &= K_{tt}(0,\theta) + K_t(0,\theta) K_t(0,\theta)' \\ &= n^{-1} \sum \psi(z_i,\theta)\psi(z_i,\theta)' \end{aligned}$$

An important recent development in this direction involves the use of EL methods combined with Bayesian MCMC ideas to impose prior information in addition to that provided by the EL framework. An interesting example of this approach is the paper of Yang and He (2012) which treats quantile regression models in this way.

Testing Overidentifying Moment Conditions

We have already discussed what IJS call average moment tests based on

$$T_n = nQ(\hat{\theta}_{gmm}) \rightsquigarrow \chi^2_{q-p}.$$

A second version of this test is based on iterating the estimation of the weight matrix W in GMM until it is consistent with the one based on the final estimates of θ.

Since the Lagrange multiplier parameter t measures the "marginal cost" of imposing the moment conditions in terms of the sacrifice in the objective function, it seems natural to consider tests based on $\hat{t}$ as well. The large sample theory of $(\hat{\theta}, \hat{t})$ is given by

$$\sqrt{n} \begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{t} \end{pmatrix} \rightsquigarrow \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (H'J^{-1}H)^{-1} & 0 \\ 0 & J^{-1}(I - H(H'J^{-1}H)^{-1}H')J^{-1} \end{pmatrix} \right)$$

so we may consider the test statistic

$$T_n = \hat{t}' V_n^- \hat{t}$$

where $V_n = J^{-1}(I - H(H'J^{-1}H)^{-1}H')J^{-1}$ and V_n^- denotes any generalized inverse of V_n. Such tests perform extremely well in monte carlo relative the classical average moment tests which are commonly used.

Reference

Imbens, G.W., Johnson, P. and Spady, R.H., (1998). "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333-59.

Qin, J. and Lawless, J. (1994). Generalized estimating equations, *Annals of Statistics*.

Owen, A. (2001) *Empirical Likelihood*, Chapman-Hall/CRC Press.

Yang, Y. and X. He (2012) Bayesian empirical likelihood for quantile regression *Annals of Statistics*, 40, 1102-1131.