Econ 551: Lecture Note 9
Asymptotic Properties of Nonlinear Estimators

Professor John Rust

Background: So far in Econ 551 we have focused on the asymptotic properties of nonlinear least squares and maximum likelihood estimators under the *IID* sampling assumption (i.e. that the data $\{(y_1,x_1),\ldots,(y_N,x_N)\}$ are independent and identically distributed draws from some unknown joint population distribution $F(y,x)$). However this basic asymptotic framework can be generalized to a much wider class of *M-estimators* (where "M" is intended as a mnemonic for "Maximization") where the estimator $\hat{\theta}$ of some unknown parameter vector θ^* is the solution to an optimization problem, just as in least squares or maximum likelihood. We can also dispense with the *IID* sampling assumption and allow the data $\{(y_1,x_1),\ldots,(y_N,x_N)\}$ to be a realization of a strictly stationary and ergodic stochastic process. These notes will also discuss the closely related class of *Z estimators* and *GMM estimators.*

M-Estimators These are defined in terms of a population optimization condition for the "true parameter" θ^*, i.e. we assume there is some function $\psi(y,x,\theta)$ whose expectation is uniquely maximized at the "true" value of the parameter, θ^*:

$$\theta^* = \underset{\theta\in\Theta}{argmax}\, E\{\psi(\tilde{y},\tilde{x},\theta)\} \tag{1}$$

where the expectation is taken with respect to the invariant distribution of (y_t,x_t) (which doesn't depend on t due to the assumption of strict stationarity), and the function ψ is twice continuously differentiable in θ for each (y,x) and measurable in (y,x) for each θ. We assume the parameter space Θ is a compact subset of R^K and that θ^* is uniquely identified as an interior point of Θ. The M-estimator is then given by a sample analog optimization condition for $\hat{\theta}$. That is, for any strictly stationary and ergodic stochastic process, averages of functions of the values of the process converge to the "long run expectation", i.e. the expectation with respect to the marginal or invariant distribution of the process, we can apply the *Analogy principle* and compute $\hat{\theta}$ as

$$\hat{\theta} = \underset{\theta\in\Theta}{argmax}\, \frac{1}{N}\sum_{i=1}^{N}\psi(y_i,x_i,\theta) \tag{2}$$

Note that whether we are taking min or max is inessential, since $argmax\, f(x) = argmin -f(x)$. Note that the class of M-estimators encompass both maximum likelihood ($\psi(y,x,\theta) = \log[f(y|x,\theta)]$) and linear and nonlinear least squares ($\psi(y,x,\theta) = -[y-f(x,\theta)]^2$) as special cases.

Z-Estimators There is a closely related class of estimators calls *Z-Estimators* (with the "Z" denoting "Zero") where the parameters are solutions or zeros to system of nonlinear equations. Generally the first order condition to an M-estimator defines an associated Z-estimator. Given a function $h(y,x,\theta)$, we assume the true parameter θ^* is the unique solution to the following population unconditional moment restrictions or *orthogonality condition*

$$\theta^* \quad \text{solves} \quad 0 = H(\theta) \equiv E\{h(\tilde{y},\tilde{x},\theta)\} \tag{3}$$

The Z-estimator $\hat{\theta}$ is defined as a solution to the sample analog of the population moment condition in equation (3):

$$\hat{\theta} \quad \text{solves} \quad 0 = H_N(\theta) \equiv \frac{1}{N}\sum_{i=1}^{N}h(y_i,x_i,\theta) \tag{4}$$

Here h is a $J \times 1$ vector functions of (y, x, θ). Note that an M-estimator with function $\psi(y, x, \theta)$ implies an associated Z-estimator with function $h(y, x, \theta) = \partial\psi(y, x, \theta)/\partial\theta$.

GMM Estimators and Minimum Distance Estimators Given a Z-estimator one can define an associated estimator, a *GMM estimator* (for Generalized Methods of Moments) that is basically similar to an M-estimator, or more precisely, a type of *Minimum Distance Estimator.* If there are more orthogonality conditions than parameters, i.e. if $J > K$, then it will generally not be possible to find an exact zero to the sample orthogonality condition (4) and so it is convenient to transform the Z-estimator into an M-estimator using a $J \times J$ positive definite *weighting matrix* W. In the limiting population case, it is easy to see that θ^* is a solution to (3) if and only if θ^* is the unique minimizer of

$$\theta^* = \underset{\theta \in \Theta}{argmin}\, H(\theta)' W H(\theta) \tag{5}$$

Once again we appeal to the analogy principle to define the GMM estimator by replacing $H(\theta)$ with its sample analog $H_N(\theta)$ and replacing W by any positive definite (possibly stochastic) weighting matrix W_N that converges in probability to W:

$$\hat{\theta} = \underset{\theta \in \Theta}{argmin}\, H_N(\theta)' W_N H_N(\theta) \tag{6}$$

This estimator is also known as a *minimum distance estimator* since the quadratic form $x' W_N x$ defines (the square of) a *norm* or distance function on R^J, (i.e. the distance between two vectors x and y in R^J under this norm is $sqrt(x-y)' W_N (x-y)$). Thus, the GMM estimator is defined as the parameter estimate $\hat{\theta}$ that make the sample orthogonality conditions $H_N(\theta)$ as close as possible to zero in this norm.

Example 1 Consider the linear model $y = x\theta + \epsilon$. Note that OLS esimator is a type of GMM estimator with the orthogonality condition $E\{h(y, x, \theta)\} = E\{x'(y - x\theta)\} = E\{x'\epsilon\}$ when $\theta = \theta^*$. In this case the parameter θ^* is said to be *just-identified* since there are as many orthogonality conditions J as parameters K. Assuming that the $K \times K$ matrix $E\{\tilde{x}'\tilde{x}\}$ is invertible, the population moment condition can be solved to show that θ^* must equal the standard formula for the coefficients of the best linear predictor of $\tilde{y}$ given $\tilde{x}$:

$$0 = H(\theta) \equiv E\{\tilde{x}'(\tilde{y} - \tilde{x}\theta)\} \Longrightarrow \theta^* = E\{\tilde{x}'\tilde{x}\}^{-1} E\{\tilde{x}'\tilde{y}\} \tag{7}$$

It is straightforward to show that if the matrix $\sum_{i=1}^N x_i' x_i$ is invertible, that the GMM estimator $\hat{\theta}$ for this moment condition reduces to the OLS estimator, $\hat{\theta} = [\sum_{i=1}^N x_i' x_i]^{-1} [\sum_{i=1}^N x_i' y_i]$ regardless of the choice of a positive weighting matrix W_N since the OLS estimates set $H_N(\hat{\theta}) = 0$ identically.

Exercise 1 Consider a linear structural model $y = x\theta + \epsilon$ but where some of the x variables are suspected of being endogenous, i.e. $E\{x'\epsilon\} \neq 0$. Suppose there are $J \geq K$ *instrumental variables* z, i.e. the (y, x, z) satisfy the following orthogonality condition at θ^*:

$$0 = H(\theta^*) = E\{z'(y - x\theta^*)\} \tag{8}$$

Show that the GMM estimator for this orthogonality condition coincides with the two stage least squares estimator.