14.385

Nonlinear Econometrics

Lecture 5.

Theory: Two-step estimators. Efficiency. One-step estimators.

Deriving Asymptotic Variances for Two-Step Estimators:

This is really easy if one considers any estimator in a GMM framework. Consider a two step estimator $\widehat{\beta}$ based on the moment equation:

$$E[g(z_i, \beta_0, \gamma_0)] = 0.$$

The preliminary estimate $\widehat{\gamma}$ of γ_0 is based on the moment equation:

$$E[h(z_i, \gamma_0)] = 0.$$

A formula for the asymptotic variance of such a two step estimator can be derived by noting that $\widehat{\theta} = (\widehat{\beta}', \widehat{\gamma}')'$ is a joint GMM estimator with

$$\tilde{g}(z, \theta) = \begin{pmatrix} g(z, \beta, \gamma) \\ h(z, \gamma) \end{pmatrix}$$

and **applying the general GMM formula** for its moment function with the block-diagonal weighting matrix.

Adaptive case. This is when the preliminary estimation of $\widehat{\gamma}$ has no first-order impact on asymptotic variance of $\widehat{\beta}$. That is, first order variance of $\widehat{\beta}$ is the same as if we knew the true value γ_0. It is often said in such situations that $\widehat{\beta}$ is **oracle-efficient**.

Adaptivity occurs when

$$G_\gamma = \nabla_\gamma E[g(z_i, \beta_0, \gamma_0)] = 0.$$

This means that the small changes in γ_0 have negligible first order impact on the root β_0 of the equation. Indeed, by the implicit function theorem, we have that

$$\nabla_\gamma \beta(\gamma_0) = G_\beta^{-1} G_\gamma$$

where

$$G_\beta = \nabla_\beta E[g(z_i, \beta_0, \gamma_0)].$$

A more extreme case, a very important one, is when even big changes do not affect the root β_0. In MLE, as the special case, the adaptive estimation of β is possible when the information matrix

$$J = -E[\nabla^2 \ln f(z, \theta_0)]$$

is block-diagonal with respect to β and γ.

Example 1 (Adaptive): Consider the population FOC for the log-likelihood in the classical regression model:

$$E[m(z, \beta, \sigma^2)] = E[\frac{(y_i - x_i\beta)x_i}{\sigma^2}] = 0.$$

Here,

$$G_{\sigma^2} = E[\frac{(y_i - x_i\beta_0)x_i}{\sigma^4}] = 0,$$

so consistent estimation of σ^2, or even inconsistent estimation of σ^2 has no impact on variance of $\widehat{\beta}$, the least squares. A similar situation occurs for the generalized least squares under correct specification, where preliminary estimation of weights has no first order impact on the asymptotic variance.

Example 2 (Non-adaptive): In PS2, you will work through the two-step estimation of the selection model, in which

$$E[y_i|x_i] = x_i'\beta + \alpha \cdot \lambda(x_i'\gamma), \quad \lambda(x_i'\gamma) = \frac{\phi(x_i'\gamma)}{\Phi(x_i'\gamma)},$$

which is the conditional expectation of a response variable selected on the basis of whether or not

$$y_{i0} = x_i'\gamma + \epsilon_i \geq 0.$$

Heckit: A convenient way to estimate this model is

1. to run a probit regression of the observed $d_i = 1(y_{i0} \geq 0)$ on $x_i'\gamma$, obtaining an estimated regressor
$$\lambda(x_i'\widehat{\gamma}),$$
2. to run ordinary least squares of y_i on x_i and $\lambda(x_i'\widehat{\gamma})$.

This example is clearly non-adaptive in general. Indeed, here we have

$$g(z_i, \beta, \alpha, \gamma) = d_i[x_i, \lambda(x_i'\gamma)]'[y_i - x_i'\beta - \alpha'\lambda(x_i'\gamma)]$$

and

$$m(z_i, \gamma) = \frac{[d_i - \Phi(x_i'\gamma)]}{[\Phi(x_i'\gamma)(1 - \Phi(x_i'\gamma)]};$$

and generally

$$G_\gamma = \nabla_\gamma E[g(z_i, \beta_0, \alpha_0, \gamma_0)] \neq 0.$$

The two-step estimator is computationally convenient, but is less efficient than the maximum likelihood estimator. However, starting with an initial, computationally convenient estimator, one can always gain efficiency by doing a **"one-step"** from an initial estimate, using a likelihood or other efficient criterion function.

Outline of the adaptivity argument: The two-step estimator satisfies

$$\hat{g}(\hat{\beta}, \hat{\gamma}) := E_n[g(z_i, \hat{\beta}, \hat{\gamma})] = 0.$$

Expand for $\tilde{G}_\beta = \nabla_\beta \hat{g}(\bar{\beta}, \bar{\gamma})$ and $\tilde{G}_\gamma = \nabla_\gamma \hat{g}(\bar{\beta}, \bar{\gamma})$, and

$$\hat{g}(\hat{\beta}, \hat{\gamma}) = \hat{g}(\beta_0, \gamma_0) + \tilde{G}_\beta(\hat{\beta} - \beta_0) + \tilde{G}_\gamma(\hat{\gamma} - \gamma_0)$$

and solve

$$\sqrt{n}(\hat{\beta} - \beta_0) = \tilde{G}_\beta^{-1}(\hat{g}(\beta_0, \gamma_0) + \underbrace{\tilde{G}_\gamma}_{?} \underbrace{\sqrt{n}(\hat{\gamma} - \gamma_0)}_{O_p(1)})$$

If $G_\gamma = 0$, then $\tilde{G}_\gamma \to_p 0$ by ULLN under regularity conditions, and preliminary estimation has no effect on variance of $\hat{\beta}$.

If $G_\gamma \neq 0$, then $\tilde{G}_\gamma \to_p G_\gamma$ under regularity conditions, and preliminary estimation has a non-trivial effect on variance of $\hat{\beta}$. In such a case, simply treat the two-step estimator in GMM framework and use appropriate formulas for this case.

Reference: e.g. Newey and McFadden pp 2150–2152.

Asymptotic Efficiency:

Here we emphasize two key results:

1) Asymptotic variance minimizing choice of any $\widehat{A}$ is $\widehat{A} \xrightarrow{p} \Omega^{-1}$.

2) MLE is the efficient estimator in the class of GMM estimators, i.e. that $\nabla \ln f(z|\theta)$ is the optimal choice of moment functions $g(z, \theta)$.

EFFICIENT DISTANCE MATRIX. *Note that if $A = \Omega^{-1}$ the asymptotic variance reduces to $(G'\Omega^{-1}G)^{-1}$ .We have that if $G'AG$ and Ω are nonsingular then*

$$(G'AG)^{-1}G'A\Omega AG(G'AG)^{-1} - (G'\Omega^{-1}G)^{-1} \geq 0$$

Thus, the asymptotic variance of the GMM estimator is minimized when $A = \Omega^{-1}$

Proof: Efficiency result is actually implied by Gauss-Markov theorem for the linear model. Consider the classical normal linear regression model with m observations:

$$Y_{(m\times 1)} = G\beta + \epsilon, \epsilon|G \sim N(0, \Omega),$$

so that

$$E[Y|G] = G\beta, \quad Var(Y|G) = \Omega.$$

Consider the GLS estimator

$$(\widehat{\beta} - \beta) = (G'\Omega^{-1}G)^{-1}G\Omega^{-1}\epsilon$$

and the WLS

$$(\tilde{\beta} - \beta) = (G'AG)^{-1}G'A\epsilon.$$

GLS has the variance matrix

$$(G'\Omega^{-1}G)^{-1}$$

and WLS has the variance matrix

$$(G'AG)^{-1}G'A\Omega AG(G'AG)^{-1}.$$

By Gauss-Markov Theorem, the former matrix is smaller than the latter. □

Thus, in large samples, the GMM estimator is **equivalent in distribution** to WLS for the normal regression model. The optimally weighted GMM is equal in distribution to the GLS that uses optimal weights.

Remarks. 1) The above proof is in the spirit of Le Cam's **limits of experiments**. In large samples, the problem of choice of weight matrix is like the choice of the weight matrix in the "experiment" with m-observation normal regression model with design matrix G and disturbances having general variance matrix Ω.

2) A more direct proof without "tricks" can be done as follows.

$$\begin{array}{rcl} \text{Defs} & : & L'L = \Omega,\ H = (L')^{-1}G,\ F = (G'AG)^{-1}G'AL'. \\ \text{Note} & : & H'H = G'\Omega^{-1}G, FH = I. \\ \text{Proof} & : & (G'AG)^{-1}G'A\Omega A'G(G'AG)^{-1} - (G'\Omega^{-1}G)^{-1} \\ & = & FF' - (H'H)^{-1} = FF' - FH(H'H)^{-1}H'F' \\ & = & F(I - H(H'H)^{-1}H')F' \geq 0. \end{array}$$

Notes: All that is needed for efficiency is that the limit of $\hat{A}$ is Ω^{-1}. Thus, any consistent estimator of Ω^{-1} leads to an asymptotically efficient GMM estimator.

MLE is optimal GMM.

This follows from MLE attaining the Cramer-Rao lower variance bound and from asymptotic unbiasedness of MLE. (In fact, MLE is efficient among all estimators.)

The asymptotic efficiency of MLE among GMM is based on a generalized information matrix equality. The moment conditions being satisfied for all possible θ means that

$$\int g(z,\theta) f(z|\theta) dz = 0$$

is an identity in θ. Assuming that differentiation under the integral is allowed, we can differentiate this identity to obtain,

$$\begin{aligned} 0 &= \int [\nabla g(z,\theta_0)] f(z|\theta_0) dz \\ &+ \int g(z,\theta_0) [\nabla \ln f(z|\theta_0)] f(z|\theta_0) dz \\ &= G + E[gs'],\ G = E[\nabla g(z,\theta_0)], \end{aligned}$$

where

$$g = g(z,\theta_0)$$

and

$$s = \nabla \ln f(z|\theta_0) = \nabla f(z|\theta_0)/f(z|\theta_0).$$

That is, we have the generalized information equality

$$G = -E[gs'].$$

If we set

$$g(z, \theta_0) = \nabla \ln f(z, \theta_0),$$

we get the usual **information matrix equality**:

$$J = -E[\nabla^2 \ln f(z, \theta_0)] = E[\underbrace{\nabla \ln f(z, \theta_0)}_{s} \underbrace{\nabla \ln f(z, \theta_0)'}_{s'}],$$

that is, the information matrix is equal to the variance of the score.

Theorem: (MLE is optimal GMM): *If* $G + E[gs'] = 0$, $E[ss'] = J$ *is nonsingular, and* $G'\Omega^{-1}G$ *is nonsingular then*

$$(G'\Omega^{-1}G)^{-1} - J^{-1} \geq 0.$$

Proof: Consider the moment function $h = -G'\Omega^{-1}g$. Then by $G = -E[gs']$ we have

$$E[hh'] = G'\Omega^{-1}G = -G'\Omega^{-1}E[gs'] = E[hs'],$$

that is the variance of moment function h is equal to its covariance with the score s. This means that the score s equals h plus some noise, i.e., the score s has bigger variance than h. Indeed, the variance matrix of $h - s$ is

$$E[hh'] - 2E[hs'] + E[ss'] = E[ss'] - E[hh'] \geq 0,$$

which says that $J - (G'\Omega^{-1}G) \geq 0$. This in turn implies the result. $\square$

Comment: Maximum likelihood is also asymptotically efficient more generally, not just in GMM class; MLE is efficient in class of estimators satisfying certain regularity conditions.

Comment: We can differentiate $\int k(z,\theta)dz$ in θ under the integral sign, if e.g., first, $\nabla_\theta k(z,\theta)$ is continuous in θ for each z, and, second, $\int \sup_\theta \|\nabla_\theta k(z,\theta)\| dz < \infty$. This allows us to apply the dominated convergence theorem, and conclude that

$$\nabla_\theta \int k(z,\theta)dz = \int \nabla_\theta k(z,\theta)dz.$$

Theoretical exercise: verify this.

Iteration and "One Step" Estimation:

This material is important for

- numerical work,
- obtaining efficient estimators,
- and bootstrap.

In numerical computations, starting from

an **initial** estimate $\tilde{\theta}$,

there are two common ways of iteration to obtain the

next step estimate $\bar{\theta}$.

1. Newton-Raphson Step. Minimize the quadratic approximation for $\hat{Q}(\theta)$:

$$\hat{Q}(\theta) \approx \hat{Q}(\tilde{\theta}) + \nabla\hat{Q}(\tilde{\theta})'(\theta - \tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})'\nabla^2\hat{Q}(\tilde{\theta})(\theta - \tilde{\theta}).$$

minimize RHS and get FOC

$\Longrightarrow \nabla\hat{Q}(\tilde{\theta}) + \nabla^2\hat{Q}(\tilde{\theta})(\bar{\theta} - \tilde{\theta}) = 0$

solve the FOC

$\Longrightarrow \bar{\theta} = \tilde{\theta} - [\nabla^2\hat{Q}(\tilde{\theta}]^{-1}\nabla\hat{Q}(\tilde{\theta})$

(Draw picture to see how it works.)

2. Gauss-Newton. Use an approximation obtained by linear approximation for the first-order condition, e.g. GMM:

$$\widehat{Q}(\theta) \approx (\widehat{g}(\tilde{\theta}) + \tilde{G}(\theta - \tilde{\theta}))' A (\widehat{g}(\tilde{\theta}) + \tilde{G}(\theta - \tilde{\theta}))'$$

the approximate first order condition

$$\Longrightarrow \tilde{G}' A [\widehat{g}(\tilde{\theta}) + \tilde{G}(\bar{\theta} - \tilde{\theta})] = 0.$$

solve first order condition

$$\Longrightarrow \bar{\theta} = \tilde{\theta} - (\tilde{G}' A \tilde{G})^{-1} \tilde{G} A \widehat{g}(\tilde{\theta})$$

"Theorem": Under regularity conditions, if the initial guess is a $\sqrt{n}$ consistent estimate, i.e.

$$(\tilde{\theta} - \theta_0) = O_p(\frac{1}{\sqrt{n}}),$$

then

$$\sqrt{n}(\bar{\theta} - \theta_0) = \sqrt{n}(\widehat{\theta} - \theta_0) + o_p(1),$$

for

$$\widehat{\theta} = \arg\min_{\theta} \widehat{Q}(\theta).$$

Thus, under regularity conditions, "one-step" estimators are equivalent to the extremum estimator up to the first order. Further iterations do not increase (first-order) asymptotic efficiency.

Remark: you can supply your own regularity conditions and details for the proof to formalize this.

"Proof": The proof can be outlined as follows:

a) For Newton-Raphson:

$$\begin{aligned}
&\sqrt{n}(\bar{\theta} - \theta_0) \\
&= \sqrt{n}(\tilde{\theta} - \theta_0) - \nabla^2 \hat{Q}(\tilde{\theta})^{-1} \sqrt{n} \nabla \hat{Q}(\tilde{\theta}) \\
&= \sqrt{n}(\tilde{\theta} - \theta_0) - \nabla^2 \hat{Q}(\tilde{\theta})^{-1} [\sqrt{n} \nabla \hat{Q}(\theta_0) + \nabla^2 \hat{Q}(\theta^*) \sqrt{n}(\tilde{\theta} - \theta_0)] \\
&= \underbrace{(I - \nabla^2 \hat{Q}(\tilde{\theta})^{-1} \nabla^2 \hat{Q}(\theta^*))}_{o_p(1)} \underbrace{\sqrt{n}(\tilde{\theta} - \theta_0)}_{O_p(1)} - \underbrace{\nabla^2 \hat{Q}(\tilde{\theta})^{-1} \sqrt{n} \nabla \hat{Q}(\theta_0)}_{\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)} \\
&= o_p(1) + \sqrt{n}(\hat{\theta} - \theta_0)
\end{aligned}$$

b) For Gauss-Newton:

$$\begin{aligned}
&\sqrt{n}(\bar{\theta} - \theta_0) = \sqrt{n}(\tilde{\theta} - \theta_0) - (\tilde{G}' A \tilde{G})^{-1} \tilde{G} A \sqrt{n} \hat{g}(\tilde{\theta}) \\
&= \sqrt{n}(\tilde{\theta} - \theta_0) - (\tilde{G}' A \tilde{G})^{-1} \tilde{G}' A \sqrt{n} [\hat{g}(\theta_0) + G^*(\tilde{\theta} - \theta_0)] \\
&= \underbrace{(I - (\tilde{G}' A \tilde{G})^{-1} \tilde{G}' A G^*)}_{o_p(1)} \underbrace{\sqrt{n}(\tilde{\theta} - \theta_0)}_{O_p(1)} - \underbrace{(\tilde{G}' A \tilde{G})^{-1} \tilde{G}' A \sqrt{n} \hat{g}(\theta_0)}_{\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)} \\
&= o_p(1) + \sqrt{n}(\hat{\theta} - \theta_0). \quad \square
\end{aligned}$$

Now you can see why "one-step" estimators and their properties are very important for

- numerical work (the one-step theorem is a positive result on the possibility of obtaining statistically well-behaved estimates),

- obtaining efficient estimators (we can simply do "one-steps" on efficient criterion functions starting with computationally convenient initial estimates),

- and bootstrap (in bootstrap samples, instead of recomputing the extremum estimate, we can do "one-steps" from the initial sample extremum estimate).