14.384 Time Series Analysis, Fall 2008
Recitation by Paul Schrimpf
Supplementary to lectures given by Anna Mikusheva
November 21, 2008

Recitation 12

Review of the Asymptotics of Extremum Estimators

These notes review GMM, indirect inference, and MLE asymptotics. They discuss consistency, asymptotic normality, and testing. GMM, indirect inference, and MLE all involve minimizing some objective function, so they are collectively referred to as extremum estimators. A good reference for them is the Handbook of Econometrics chapter by Newey and McFadden (1994). I mostly copied this from notes from 385, so there might be references to stuff from lecture that we never covered in this course.

Types of Estimators

- Moment methods
  - GMM
  - IV
- Extremum methods
  - MLE
  - M-estimators
  - Quantile regression
  - **Minimum Distance**

Figure 1: Relationship among estimators

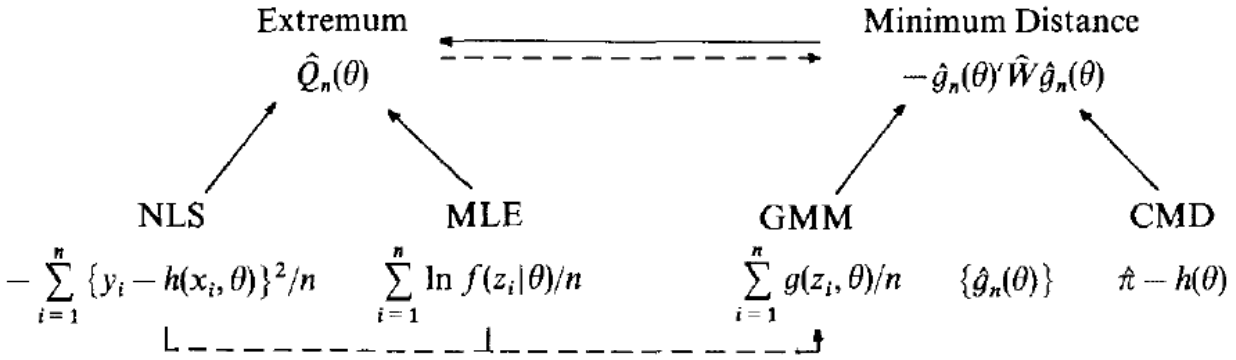

Figure 1.

Courtesy Elsevier, Inc., http://www.sciencedirect.com. Used with permission.

[a]

[a]Source: Newey and McFadden (1994)

Minimum Distance

We covered GMM and MLE in detail in lectures (in 385). Here we will go through minimum distance.

$$\hat{\theta} = \arg\min \hat{f}_n(\theta)'\hat{W}\hat{f}_n(\theta) \tag{1}$$

where plim $\hat{f}_n(\theta_0) = 0$.

Includes:

- GMM with $\hat{f}_n(\theta) = \frac{1}{n}\sum g(z_i,\theta)$ and $\hat{W} = \hat{A}$
- MLE with $\hat{f}_n(\theta) = \frac{1}{n}\sum \frac{\partial \ln f(z_i|\theta)}{\partial\theta}$ and $\hat{W} = I$
- Classical Minimum Distance (CMD): $\hat{f}_n(\theta) = \hat{\pi} - h(\theta)$ where $\hat{\pi} \xrightarrow{p} \pi_0 = h(\theta_0)$. Usually, π are reduced form parameters, θ are structural parameters, and $h(\theta)$ is a mapping from the structural parameters to the reduced form.
  - *Example:* Chamberlain (1982, HoE 1984) approach to panel data. Model:

    $$y_{it} = x_{it}\beta + c_i + e_{it} \ , \ E[e_{it}|x_i, c_i] = 0$$

    Reduced form: regress y_{it} on all $x_{i\cdot}$ to get π_t.
    $h(\theta)$: we know that $x_{i\cdot}\pi_t$ is the best linear predictor of y_{it} given $x_{i\cdot}$. We also know that

    $$\begin{aligned} BLP(y_{it}|x_{i\cdot}) =& BLP(x_{it}\beta + e_{it}|x_{i\cdot}) + BLP(c_i|x_{i\cdot}) \\ =& x_{it}\beta + x_{i\cdot}\lambda \end{aligned}$$

    So if we stack the π_t into a $t \times tk$ matrix π, we know that

    $$\pi = h(\beta,\lambda) = I_T \otimes \beta' + \iota_T\lambda'$$

    where β is $k \times 1$ and λ is $tk \times 1$.

- Indirect Inference: is mathematically the same as CMD, $\hat{f}_n(\theta) = \hat{\pi} - h(\theta)$ where $\hat{\pi} \xrightarrow{p} \pi_0 = h(\theta_0)$, but the justification is slightly different. We have an economic model, which we are not entirely certain is the true DGP (or perhaps is just difficult to compute the likelihood for), but we do believe can capture some important features of the data. These features of the data are summarized by the parameters of an easy to estimate auxillary model, π. $h(\theta)$ gives the estimates of the auxillary model that we would expect if our economic model were the true DGP and had parameters θ. $h(\theta)$ is often calculated through simulation.
  - *Example: DSGE* (taken from 14.384 notes, for a real application see e.g. Del Negro,Schorfheide, Smets and Wouters (2007)) Consider a simple RBC:

$$\begin{aligned} &\max E_0 \sum \omega^t \frac{c^{-\gamma} - 1}{\gamma} \\ &\text{s.t. } c_t + i_t = A\lambda_t k_t^{\alpha} \\ &\qquad k_{t+1} = (1-\delta)k_t + i_t \\ &\qquad \lambda_t = \rho\lambda_{t-1} + \epsilon_t \;\; \epsilon_t \sim N(0, \sigma^2) \end{aligned}$$

    This model has many parameters ($\theta = (\omega, \gamma, A, \alpha, \delta, \rho, \sigma^2)$) and it would be difficult to write down a likelihood or moment functions. Moreover, we don't really believe that this model is the true DGP and we don't want to use it to explain all aspects of the data. Instead we just want the model to explain some feature of the data, say the dynamics as captured by VAR coefficients. Also, although it is hard to write the likelihood function for this model, it is fairly easy to simulate the model. The we can use indirect inference as follows:

    1. Estimate (possibly misspecified) VAR from data. A VAR is simply OLS on:

$$Y_t = \pi_1 Y_{t-1} + ... + \pi_p Y_{t-p} + u_t$$

       where Y_t is the vector of observed variables at time t. In this example, Y_t might be c_t and i_t.
    2. Given β, simulate model, estimate VAR from simulations, repeat until minimize objective function

Consistency

A general theorem on consistency is:

Theorem 1. *If (i) $Q(\theta)$ is uniquely minimized at the true parameter value θ_0, (ii) Θ is compact, (iii) $Q(\cdot)$ is continuous, and (iv) $\sup_{\theta\in\Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0$, then $\hat{\theta} \xrightarrow{p} \theta_0$.*

We will now discuss applying this theorem to minimum distance. To do that, we need to verify each of the conditions:

1. (Identification): suppose $\hat{f}_n(\theta) \xrightarrow{p} f(\theta)$ and $\hat{W} \xrightarrow{p} W$, so that $Q(\theta) = f(\theta)'Wf(\theta)$. As with GMM, showing that this function has a unique minimum is difficult. A local identification condition is that $rank\frac{\partial f}{\partial \theta} = p$, where θ is $p \times 1$. Global identification is typically just assumed.
2. (Compactness): assume it.
3. (Continuity): depends on the particular application. For CMD and indirect inference, $f(\theta) = \pi - h(\theta)$ is continuous as long as $h(\theta)$ is continuous. Since $h(\theta)$ does not depend on the data at all, this condition is easily checked. In the panel data example, $h(\theta)$ is obviously continuous.
4. (Uniform Convergence): depends on the particular application. Recall lemma 3 from lecture 2. It was:

Lemma 2. *Suppose $\hat{Q}(\theta) \xrightarrow{p} Q(\theta)$ for each $\theta \in \Theta$. Then uniform convergence holds if for some $h > 0$, we have uniformly for $\theta, \theta' \in \Theta$*

$$|\hat{Q}(\theta) - \hat{Q}(\theta')| \leq B_T ||\theta - \theta'||^h \ , \ B_T = O_p(1)$$

For CMD and indirect inference,

$$\begin{aligned}
\hat{Q}(\theta) - \hat{Q}(\theta') =& |(\hat{\pi} - h(\theta))'\hat{W}(\hat{\pi} - h(\theta)) - (\hat{\pi} - h(\theta'))'\hat{W}(\hat{\pi} - h(\theta'))'| \\
=& |2(h(\theta) - h(\theta')\hat{W}\hat{\pi} + h(\theta)'\hat{W}h(\theta) - h(\theta')'\hat{W}h(\theta')| \\
\leq& |2(h(\theta) - h(\theta')\hat{W}\hat{\pi}| + |h(\theta)'\hat{W}h(\theta) - h(\theta')'\hat{W}h(\theta')| \\
\leq& |2(h(\theta) - h(\theta')\hat{W}\hat{\pi}| + |(h(\theta) - h(\theta'))'\hat{W}(h(\theta) - h(\theta'))|
\end{aligned}$$

so a sufficient condition is that $h(\theta)$ is Hölder continuous on Θ, i.e.

$$|h(\theta) - h(\theta')| \leq K||\theta - \theta'||^h$$

for some $h > 0$ and all $\theta, \theta' \in \Theta$. A sufficient condition for Hölder continuity is that $h()$ is differentiable with a bounded derivative because then

$$|h(\theta) - h(\theta')| \leq \sup_{\Theta} ||h'||||\theta - \theta'||$$

Clearly, this condition holds for the panel data example. It could also be checked in other applications.

- If $h(\theta)$ is computed through simulation, then some additional steps need to be taken to show consistency. Let $h_S(\theta)$ denote the value of $h(\theta)$ computed from S simulations. Typically, $h_S(\theta)$ will be some standard estimator and we will know that $h_S(\theta) \xrightarrow{p} h(\theta)$ as $S \to \infty$. For $\hat{Q}$ to converge uniformly, we need to promise that $S \to \infty$ as $T \to \infty$, and we will need the convergence of h_S to h to be uniform in addition to the conditions above.

Review of Asymptotic Normality

Recall the basic asymptotic normality theorem from lecture 3:

Theorem 3. Asymptotic Normality *If $\hat{\theta} \xrightarrow{p} \theta_0$ and*

(i) $\theta_0 \in int(\Theta)$

(ii) $\hat{Q}(\theta)$ *is twice continuously differentiable in a neighborhood, $\mathcal{N}$, of* θ_0

(iii) $\sqrt{n}\nabla\hat{Q}(\theta_0) \xrightarrow{d} N(0, \Omega)$

(iv) There is $J(\theta)$ *that is continuous at* θ_0 *and* $\sup_{\theta \in \mathcal{N}} \|\nabla^2\hat{Q}(\theta) - J(\theta)\| \xrightarrow{p} 0$

(v) $J = J(\theta_0)$ *is nonsingular*

then,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, J^{-1}\Omega J^{-1})$$

Make sure that you understand the reasoning behind this result – taking a mean-value expansion of the first order condition.

Asymptotic Linearity and Influence Functions

Another way of describing results on asymptotic normality is by considering asymptotically linear estimators and their influence functions. $\hat{\theta}$ is *asymptotically linear* with *influence funtion* $\psi(z)$ if:

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sum \psi(z_i)/\sqrt{n} + o_p(1) \tag{2}$$

with $E\psi(z) = 0$ and $E[\psi(z)\psi(z)'] < \infty$. Most common estimators are asymptotically linear. For example, MLE has influence function

$$\psi_{MLE}(z) = -H^{-1}\nabla \ln f(z|\theta_0)$$

We probably will not talk about asymptotic linearity or influence functions much in this course. Two places where influence functions come up are in calculating semi-parametric efficiency bounds, and in analyzing robustness (to outliers) of estimators.

Asymptotic Normality of Minimum Distance

In the last recitation we talked about minimum distance estimators, which have the form:

$$\hat{\theta} = \arg\min \hat{f}_n(\theta)'\hat{W}\hat{f}_n(\theta)$$

GMM and MLE fit into this framework, as well as classical minimum distance (CMD) and indirect inference. CMD and indirect inference use $\hat{f}_n(\theta) = \hat{\pi} - h(\theta)$ where $\hat{\pi} \xrightarrow{p} \pi_0 = h(\theta_0)$. Let's specialize the generic asymptotic normality theorem to minimum distance. Conditions (i)-(v) above become:

Theorem 4. Asymptotic Normality for Minimum Distance *If* $\hat{\theta} \xrightarrow{p} \theta_0$ *and*

(i) $\theta_0 \in int(\Theta)$

(ii) $\hat{f}_n(\theta)$ *is continuously differentiable in a neighborhood,* $\mathcal{N}$*, of* θ_0

(iii) $\sqrt{n}\hat{f}_n(\theta_0) \xrightarrow{d} N(0, \Omega)$

- *For CMD and indirect inference* $\sqrt{n}\hat{f}(\theta_0) = (\hat{\pi} - \pi_0) + o_p(1)$*, so it is enough that* $\sqrt{n}(\hat{\pi} - \pi_0) \xrightarrow{d} N(0, \Omega)$

(iv) There is $G(\theta)$ *that is continuous at* θ_0 *and* $\sup_{\theta \in \mathcal{N}} \|\nabla \hat{f}_n(\theta) - G(\theta)\| \xrightarrow{p} 0$

- *For CMD and indirect inference,* $\nabla \hat{f}_n(\theta) = \nabla h(\theta)$*, so it is enough that* $h(\theta)$ *is continuously differentiable.*

(v) $\hat{W} \xrightarrow{p} W$ *is positive semi-definite and* $G'WG$ *is nonsingular*

then,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1})$$

Proof. Verify that these conditions are the same as in theorem (3). □

The primary difference compared to the basic asymptotic normality theorem is that twice differentiability of the objective function is restated as once differentiability of the distance function.

Example 5. *Chamberlain Panel Data*: See recitation 1 notes for setup. π are unrestricted OLS coefficients. $h(\theta) = I_T \otimes \beta' + \imath_T \lambda'$. We know that OLS is asymptotically normal, so condition (iii) is satisfied. $h()$ is linear, so conditions (ii) and (iv) hold. Suppose we choose $\hat{W} = \hat{\Omega}^{-1}$, where Ω is the usual OLS estimate of the variance of π. We know that $\hat{Omega} \xrightarrow{p} \Omega$ is positive semi-definite. $G'WG$ will be nonsingular as long as the model is identified. If there are K regressors, and T time periods, then there are T^2K elements in π. There are K unknowns in β and TK unknowns in λ. Hence, an order condition is that $T \geq 2$.

Variance Matrix Estimation

To use results on asymptotic normality for inference, we need to be able to consistently estimate the asymptotic variance matrix. The Hessian term, H, for MLE and Jacobian, G, for GMM can simply be estimated by evaluating the derivative of the sample objective function at $\hat{\theta}$. Estimation of the middle term, the variance of the gradient, depends on whether there is dependence in the data. For iid data, $\Omega = E[\nabla \hat{q}_i(\theta_0) \nabla \hat{q}_i(\theta_0)']$, which when $Q(\theta) = \sum q_i(\theta)$, can be estimated by

$$\hat{\Omega} = \frac{1}{n} \sum \nabla \hat{q}_i(\hat{\theta}) \nabla \hat{q}_i(\hat{\theta})'$$

Lemma 4.3 from Newey and McFadden gives precise conditions for when $\hat{\Omega} \xrightarrow{p} \Omega$

Lemma 6. Newey and McFadden Lemma 4.3 *If z_i is iid, $a(z, \theta)$ is continuous at θ_0 and there is a neighborhood, $\mathcal{N}$, of θ_0, such that $E\left[\sup_{\theta \in \theta_0} \|a(z, \theta)\|\right] < \infty$, then for any $\tilde{\theta} \xrightarrow{p} \theta_0$, we have $\frac{1}{n} \sum a(z_i, \tilde{\theta}) \xrightarrow{p} E[a(z, \theta_0)]$.*

When the data is not iid, $\Omega \neq E[\nabla \hat{q}_i(\theta_0) \nabla \hat{q}_i(\theta_0)']$, and some other estimator must be used. The same ideas that apply to OLS apply here. For example, if there is clustering, then

$$\hat{\Omega} = \frac{1}{C} \sum_c \frac{1}{n_c} \sum_i \sum_j \nabla \hat{q}_i(\hat{\theta}) \nabla \hat{q}_j(\hat{\theta})'$$

is a consistent estimator for Ω. If there is serial correlation, then Newey-West or some similar estimator can be used. You can learn more about this time series if you want.

GMM

The above remarks apply to GMM with $g(z_i, \theta)$ in place of $\nabla \hat{q}_i(\theta)$.

MLE

For MLE, we know that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1} \Omega H^{-1})$$

where $H = E[\nabla^2 \ln f(z|\theta)]$ and $\Omega = E[\nabla \ln f(z|\theta) \nabla \ln f(z|\theta)']$. In lecture 4, we saw that when the likelihood is correctly specified the information equality holds, $\Omega = H^{-1}$. This suggests the following estimators for the asymptotic variance:

- *Hessian*: $\hat{H}^{-1} = \left(\frac{1}{n} \sum \frac{\partial^2 \ln f_i}{\partial \theta \partial \theta'} |_{\theta = \hat{\theta}} \right)^{-1}$
  - In principle, when doing conditional MLE you can also use the expected conditional hessian:

    $$\hat{H}_E = \frac{1}{n} \sum E \left[\frac{\partial^2 \ln f(y_i|x_i, \hat{\theta})}{\partial \theta \partial \theta'} | x_i \right]$$

    but it is often difficult to compute this expectation
- *Outer product of gradients*: $\hat{Omega} = \frac{1}{n} \sum \nabla \ln f(z|\hat{\theta}) \nabla \ln f(z|\hat{\theta})'$
- *Sandwich*: $\hat{H}^{-1} \hat{\Omega} \hat{H}^{-1}$
  - Could use $\hat{H}_E$ in place of $\hat{H}$
  - Since this estimator does not use the information equality, it is consistent even if the likelihood is misspecified (as long as $\hat{\theta}$ remains consistent)

Hypothesis Testing

Suppose we want to test a hypothesis of the form:

$$H_0 : r(\theta) = 0$$

where $r : \mathbb{R}^k \to \mathbb{R}^q$ is differentiable. First we will discuss the familiar likelihood setup, then we will talk about testing in GMM. Before discussing these test statistics, it will be useful to review the delta method and to derive the asymptotic distribution of a constrained extremum estimator.

Delta Method Suppose $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$. Let $f(\theta)$ be continuously differentiable. Then $\sqrt{n}(f(\hat{\theta}) - f(\theta_0)) \xrightarrow{d} N(0, f'(\theta_0)Vf'(\theta_0)')$.

Asymptotic Normality of Constrained Estimators Suppose $\hat{\theta}$ solves:

$$\hat{\theta} = \arg\min Q(\theta) \text{ s.t. } r(\theta) = 0$$

The first order condition for this problem is:

$$0 = \begin{pmatrix} \nabla Q(\hat{\theta}_R) + \lambda r'(\hat{\theta}_R) \\ r(\hat{\theta}_R) \end{pmatrix}$$

Expanding around θ_0 and $\lambda_0 = 0$ gives:

$$\begin{aligned}
0 &= \begin{pmatrix} \nabla Q(\theta_0) + \lambda_0 r'(\theta_0) \\ r(\theta_0) \end{pmatrix} + \begin{pmatrix} \hat{\theta}_R - \theta_0 \\ \hat{\lambda} - \lambda_0 \end{pmatrix} \begin{pmatrix} \nabla^2 Q(\bar{\theta}) & r'(\bar{\theta})' \\ r'(\bar{\theta}) & r(\theta_0) \end{pmatrix} \\
\sqrt{n} \begin{pmatrix} \hat{\theta}_R - \theta_0 \\ \hat{\lambda} \end{pmatrix} &= \begin{pmatrix} \nabla^2 Q(\bar{\theta}) & r'(\bar{\theta})' \\ r'(\bar{\theta}) & 0 \end{pmatrix}^{-1} \begin{pmatrix} \nabla Q(\theta_0) \\ 0 \end{pmatrix} \\
&= \sqrt{n} \begin{pmatrix} (\nabla^2 Q)^{-1} - (\nabla^2 Q)^{-1}\bar{R}'(\bar{R}(\nabla^2 Q)^{-1}\bar{R}')^{-1}\bar{R}(\nabla^2 Q)^{-1} & (\nabla^2 Q)^{-1}\bar{R}'(\bar{R}(\nabla^2 Q)^{-1}\bar{R}')^{-1} \\ (\bar{R}(\nabla^2 Q)^{-1}\bar{R}')^{-1}\bar{R}(\nabla^2 Q)^{-1} & -(\bar{R}(\nabla^2 Q)^{-1}\bar{R}')^{-1} \end{pmatrix} \begin{pmatrix} \nabla Q(\theta_0) \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} \left((\nabla^2 Q)^{-1} - (\nabla^2 Q)^{-1}\bar{R}'(\bar{R}(\nabla^2 Q)^{-1}\bar{R}')^{-1}\bar{R}(\nabla^2 Q)^{-1}\right)(\sqrt{n}\nabla Q(\theta_0)) \\ (\bar{R}(\nabla^2 Q)^{-1}\bar{R}')^{-1}\bar{R}(\nabla^2 Q)^{-1}(\sqrt{n}\nabla Q(\theta_0)) \end{pmatrix}
\end{aligned}$$

where $\bar{R} = r'(\bar{\theta})$. This gives us the following conclusions:

Theorem 7. *Under the conditions of theorem 3 andj $r(\theta)$ is continuously differentiable in a neighborhood of θ_0, we have:*

$$\sqrt{n}(\hat{\theta_R} - \theta_0) = \left(J^{-1} - J^{-1}R'(RJ^{-1}R')^{-1}RJ^{-1}\right)(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) \tag{3}$$

$$\sqrt{n}\hat{\lambda} = (RJ^{-1}R')^{-1}RJ^{-1}(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) \tag{4}$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = J^{-1}(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) \tag{5}$$

$$\sqrt{n}(\hat{\theta} - \hat{\theta}_R) = J^{-1}R'(RJ^{-1}R')^{-1}RJ^{-1}(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) \tag{6}$$

$$\sqrt{n}\nabla Q(\hat{\theta}_R) = -R'(RJ^{-1}R')^{-1}RJ^{-1}(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) \tag{7}$$

Proof. (3) and (4) are direct consequences of the previous reasoning. (5) comes from theorem 3. (6) is simply the difference of (3) and (5). (7) comes from pluggin (4) into the first order condition. □

This theorem tells us the asymptotic variance of various quantities that will be used in our test statistics. For example from (6), we know that

$$\sqrt{n}(\hat{\theta} - \hat{\theta}_R) \xrightarrow{d} N\left(0, J^{-1}R'(RJ^{-1}R')^{-1}RJ^{-1}\Omega J^{-1}R'(RJ^{-1}R')^{-1}RJ^{-1}\right)$$

ML Testing

When doing MLE, we have the usual trinity of tests: Wald, Lagrange multiplier, and likelihood-ratio. Throughout we will write $Avar(\hat{\theta})$ to denote the asymptotic variance of $\hat{\theta}$. It should be straightforward to calculate using theorem 7. Each statistic can be thought of as a measure of the difference between the restricted and unrestricted objective functions. The Likelihood ratio uses the actual difference. The Wald statistic uses a quadratic expansion at the unrestricted estimate, $\hat{\theta}$, to approximate the restricted objective value. The Lagrange multiplier uses a quadratic expansion at the restricted estimate, $\hat{\theta}_R$, to approximate the unrestricted objective value.

Wald

Wald test statistics look at the distance between θ or $r(\theta)$ in the restricted and unrestricted models. One version of the Wald statistic is motivated by asking whether $r(\hat{\theta}) = 0$? It uses the test statistic:

$$\begin{aligned} W_1 =& nr(\hat{\theta})' AVar(r(\hat{\theta}))^{-1} r(\hat{\theta}) \\ & \text{(delta method)} \qquad (8) \\ =& nr(\hat{\theta})'(r'(\hat{\theta}) AVar(\hat{\theta}) r'(\hat{\theta}))^{-1} r(\hat{\theta}) \xrightarrow{d} \chi^2_q \qquad (9) \end{aligned}$$

Another variant of the Wald test looks at the distance between restricted and unrestricted estimates of θ:

$$W_2 = n(\hat{\theta} - \hat{\theta}_R)'(Avar(\hat{\theta} - \hat{\theta}_R))^{-1}(\hat{\theta} - \hat{\theta}_R) \xrightarrow{d} X^2_q \qquad (10)$$

Lagrange Multiplier

The Lagrange multiplier test is based on the fact that under H_0, the Lagrange multiplier of the restricted optimization problem should be near 0. The first order condition from the restricted ML is:

$$\frac{1}{n} \sum \nabla \ln f(z|\hat{\theta}_R) = \hat{\lambda} r'(\hat{\theta}_R)$$

which suggests the test statistic:

$$LM_1 = \frac{1}{n} \left(\sum \nabla \ln f(z|\hat{\theta}_R) \right)' Avar(\nabla \ln f(z|\hat{\theta}_R))^{-1} \left(\sum \nabla \ln f(z|\hat{\theta}_R) \right) \xrightarrow{d} \chi^2_q$$

Equivalently, we could look at the estimated Lagrange Multiplier,

$$LM_2 = n\hat{\lambda}' Avar(\hat{\lambda})^{-1} \hat{\lambda}$$

Likelihood Ratio

The Likelihood ratio statistic compares the restricted and unrestricted likelihoods.

$$LR = 2(L_N(\hat{\theta}) - L_N(\hat{\theta}_R)) \xrightarrow{d} \chi^2_q$$

To prove this, expand $L_N(\hat{\theta}_R)$ around $L_N(\hat{\theta})$:

$$\begin{aligned} LR =& 2(L_N(\hat{\theta}) - L_N(\hat{\theta}_R)) \\ =& 2\left(L_N(\hat{\theta}) - L_N(\hat{\theta}) - \nabla L_N(\hat{\theta})(\hat{\theta} - \hat{\theta}_R) - (\hat{\theta} - \hat{\theta}_R)' \nabla^2 L_N(\bar{\theta})(\hat{\theta} - \hat{\theta}_R) \right) \\ =& (\hat{\theta} - \hat{\theta}_R)' \nabla^2 L_N(\bar{\theta})(\hat{\theta} - \hat{\theta}_R) \end{aligned}$$

GMM

The same three test types of test stastistics work for GMM. The Wald and Lagrange Multiplier statistics are particularly identical to the ML case. The likelihood ratio statistic is replaced by the *distance metric*:

$$DM = 2n(Q_n(\hat{\theta}) - Q_n(\hat{\theta}_R))$$

Asymptotics of Simulated Estimators

This is a quick and dirty discussion of the asymptotics of simulated extremum estimators. See Kenneth Train's book and the references therein for more details and rigor. Train's book can be found at http://elsa.berkeley.edu/books/choice2.html.

Let $Q_n(\theta)$ denote the exact objective function. Let $\tilde{Q}_n(\theta)$ denote the simulated objective function. Assume that $\hat{\theta} = \arg\min Q_n(\theta)$ is consistent and asymptotically normal. We want to understand the behavior of $\tilde{\theta} = \arg\min \tilde{Q}_n(\theta)$. For specificity, assume that in simulating, we make R draws for each observation, and these draws are independent across observations.

Consistency

For consistency, the key condition to check is that $\tilde{Q}(\theta) = \operatorname{plim} \tilde{Q}_n(\theta)$ is uniquely minimized at θ_0. Consider the first order condition:

$$\nabla\tilde{Q}_n(\theta) = \nabla Q_n(\theta) + \left(E_r(\nabla\tilde{Q}_n(\theta)) - \nabla Q_n(\theta)\right) + \left(\nabla\tilde{Q}_n(\theta) - E_r(\nabla\tilde{Q}_n(\theta))\right)$$

where E_r denotes an expectation taken over our simulated draws. If we can show that the second and third terms on the right vanish as $n \to \infty$, then we will have consistency. The third term is easy. Since we are making R independent draws for each observation, as long as R is fixed or increasing with N, $\nabla\tilde{Q}_n(\theta)$ satisfies an LLN and converges to its expectation. The second term depends on how we are simulating. If R increases with N, then it also vanishes because of an LLN. Furthermore, even if R is fixed with N, it will be zero, if our simulations result in an unbiased estimate of the gradient. In the mixed logit example above, the simulation of choice probabilities is unbiased. Therefore, NLLS, for which the first order condition is linear in $\tilde{P}$, is consistent with fixed R. However, MLE, for which the first order condition involves $\frac{1}{\tilde{P}}$, is consistent only if R increases with N. For this reason, people sometimes suggest using the method of simulated scores (MSS) instead of MSL. MSS call for simulated the score in an unbiased way and doing GMM on the simulated score.

Asymptotic Normality

As always, we start by taking an expansion of the first order condition:

$$\sqrt{n}(\tilde{\theta} - \theta_0) = (\nabla^2\tilde{Q}_n(\bar{\theta}))^{-1}(\sqrt{n}\nabla\tilde{Q}_n(\theta_0))$$

If $\tilde{\theta}$ is consistent, then $(\nabla^2\tilde{Q}_n(\bar{\theta}))^{-1} \xrightarrow{p} (\nabla^2 E\tilde{Q}(\theta_0))^{-1}$. The main thing to worry about is the behavior of the gradient. As above, it helps to break it into three pieces:

$$\sqrt{n}\nabla\tilde{Q}_n(\theta_0) = \sqrt{n}\nabla Q_n(\theta) + \sqrt{n}\left(E_r(\nabla\tilde{Q}_n(\theta)) - \nabla Q_n(\theta)\right) + \sqrt{n}\left(\nabla\tilde{Q}_n(\theta) - E_r(\nabla\tilde{Q}_n(\theta))\right)$$

Let's start with the third term. Suppose we have iid observations so that $\nabla\tilde{Q}_n = \sum_{i=1}^n \nabla\tilde{q}_{i,R}$. Let S be the variance of $\nabla\tilde{q}_{i,1}$. Then the variance of $\nabla\tilde{q}_{i,R}$ is S/R, and

$$\sqrt{n}\left(\nabla\tilde{Q}_n(\theta) - E_r(\nabla\tilde{Q}_n(\theta))\right) \xrightarrow{d} N(0, S/R)$$

Now, on to the second term. As above, it is zero if our simulations are unbiased. If our simulations are biased, then it is $O(\frac{1}{R})$. If R is fixed, then our estimator is inconsistent. If $\frac{\sqrt{n}}{R} \to 0$, then this term vanishes, and our estimator has the same asymptotic distribution as when using the exact objective function. If R grows with n, but slower that $\sqrt{n}$, then $\tilde{\theta}$ is consistent, but not asymptotically normal.

14.384 Time Series Analysis

Fall 2013