# 14.385 Recitation 2

Paul Schrimpf

September 19, 2008

## 1 Review of Asymptotic Normality

Recall the basic asymptotic normality theorem from lecture 3:

**Theorem 1.** 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* $\theta_0$

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

*(iv) There is* $J(\theta)$ *that is continuous at* $\theta_0$ *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.

### 1.1 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:

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

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.

### 1.2 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 2.** 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* $\theta_0$

*(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* $\theta_0$ *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 (1). □

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* 3. *Chamberlain Panel Data*: See recitation 1 notes for setup. $\pi$ 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 $\Omega$ is the usual OLS estimate of the variance of $\pi$. 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 $\pi$. There are $K$ unknowns in $\beta$ and $TK$ unknowns in $\lambda$. Hence, an order condition is that $T \geq 2$.

## 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 4.** Newey and McFadden Lemma 4.3 *If* $z_i$ *is iid,* $a(z, \theta)$ *is continuous at* $\theta_0$ *and there is a neighborhood,* $\mathcal{N}$*, of* $\theta_0$*, 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 $\Omega$. 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.

### 2.1 GMM

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

### 2.2 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)

## 3 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 $\theta_0$ 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 5.** *Under the conditions of theorem 1 andj $r(\theta)$ is continuously differentiable in a neighborhood of $\theta_0$, we have:*

$$
\begin{aligned}
\sqrt{n}(\hat{\theta_R} - \theta_0) = &\left(J^{-1} - J^{-1}R'(RJ^{-1}R')^{-1}RJ^{-1}\right)(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) && (2) \\
\sqrt{n}\hat{\lambda} = &(RJ^{-1}R')^{-1}RJ^{-1}(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) && (3) \\
\sqrt{n}(\hat{\theta} - \theta_0) = &J^{-1}(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) && (4) \\
\sqrt{n}(\hat{\theta} - \hat{\theta}_R) = &J^{-1}R'(RJ^{-1}R')^{-1}RJ^{-1}(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) && (5) \\
\sqrt{n}\nabla Q(\hat{\theta}_R) = &-R'(RJ^{-1}R')^{-1}RJ^{-1}(\sqrt{n}\nabla Q(\theta_0)) + o_p(1) && (6)
\end{aligned}
$$

*Proof.* (2) and (3) are direct consequences of the previous reasoning. (4) comes from theorem 1. (5) is simply the difference of (2) and (4). (6) comes from pluggin (3) 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 (5), we know that

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

## 3.1 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 5. 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.

### 3.1.1 Wald

Wald test statistics look at the distance between $\theta$ 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:

$$
\begin{aligned}
W_1 = &nr(\hat{\theta})' AVar(r(\hat{\theta}))^{-1} r(\hat{\theta}) \\
&\text{(delta method)} && (7) \\
= &nr(\hat{\theta})'(r'(\hat{\theta})AVar(\hat{\theta})r'(\hat{\theta}))^{-1} r(\hat{\theta}) \xrightarrow{d} \chi^2_q && (8)
\end{aligned}
$$

Another variant of the Wald test looks at the distance between restricted and unrestricted estimates of $\theta$:

$$W_2 = n(\hat{\theta} - \hat{\theta}_R)'(Avar(\hat{\theta} - \hat{\theta}_R))^{-1}(\hat{\theta} - \hat{\theta}_R) \xrightarrow{d} X^2_q \qquad (9)$$

### 3.1.2 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,

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

### 3.1.3 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 $L_N(\hat{\theta})$:

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

## 3.2 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*:

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