Lecture 4
Asymptotic Normality of Extremum Estimators[1]

1 Useful Results

Convergence in distribution

Convergence in distribution has two definitions, and the two definitions are equivalent. The first definition uses the distribution functions (df) of the random variables, while the second definition uses the first moment of bounded continuous functions of the random variables. The second definition is superior in many aspects: (1) It generalizes to random elements of metric spaces and allows us to prove results regarding convergence of stochastic processes; (2) it is often easier to work with than the df's of sums of random vectors.

Definition 1: A sequence of random vectors $Y_1, Y_2, ...$ is said to *converge in distribution* to a random vector Y if $F_n(y) \to F(y)$ for all continuity points y of the distribution function (df) $F(\cdot)$ of Y, where $F_n(\cdot)$ is the df of Y_n.

Notice that the property of convergence in distribution is solely a property of the *marginal distributions* of the random vectors $Y_1, Y_2, ...$ and not of the random vectors themselves or of their joint distribution. It is irrelevant whether Y_1, Y_2, Y_3 are independent or not. Only the marginal distributions of Y_1, Y_2, $Y_3, ...$ are important. For example, we could have $Y_n = Y$, $\forall n$ and then $Y_n \to_d Y$. Or, we could have Y_n are iid with the same distribution as Y and we still have $Y_n \to_d Y$.

Note the caveat in the definition that convergence is necessary only for *continuity* points y of the df $F(\cdot)$ of Y. This caveat is necessary if we want to be able to have discrete distributions as the limits of sequences of continuous random vectors. Since we sometimes want to do this, we set up the definition appropriately. We will give an example of this later.

Definition 2: A sequence of random vectors $Y_1, Y_2, ...$ in R^k is said to *converge weakly* (or to converge in distribution) to Y or to the distribution of Y, if $\forall f \in b\mathcal{C}$, $Ef(Y_n) \to Ef(Y)$ as $n \to \infty$, where $b\mathcal{C}$ is the class of all bounded, continuous, real-valued functions defined on $\mathbb{R}^k$, where $Y \in \mathbb{R}^k$.

Result: Definitions 1 and 2 of convergence in distribution are equivalent.

The **central limit theorem** (CLT) gives an example of weak convergence. It says: if $X_1, X_2, ...$

[1] The notes for this lecture is largely adapted from the notes of Donald Andrews on "Convergence in Distribution, Continuous Mapping Theorem and Delta Method" and "Asymptotic Normality of Extremum Estimators." I am grateful for Professor Andrews' generosity and elegant exposition. All errors are mine.

are iid with $EX_i'X_i < \infty$, then

$$\sqrt{n}(\overline{X}_n - EX_1) \xrightarrow{d} Z \sim N(\mathbf{0}, \text{Var}(X_1)) \text{ as } n \to \infty.$$

Note that the assumption of iid summands is far stronger than necessary. There are triangular array CLTs with non-identically distributed and non-independent summands. Heuristically, the requirements for the CLT to hold are that no finite number of summands can be dominant, the "amount of dependence" between X_i and X_j must die out sufficiently quickly as the difference in subscripts $|i-j|$ goes to infinity, and the tails of X_i cannot be too thick. For the moment, we do not discuss those cases in detail.

By using Taylor series expansions (or mean-value expansions), one can often use CLT results to prove weak convergence results for hosts of random vectors beyond the basic random vectors to which the CLT was applied originally.

We need **more motheds** to manipulate the convergence in distribution of one sequence into convergence in distribution of another sequence. There are three main results that allow us to do so:

(1) the continuous mapping theorem,

(2) the mean value theorem

(3) the generalized Slutsky's Theorem (a corollary of the continuous mapping theorem).

Continuous Mapping Theorem: *Suppose* $\{Y_n : n \geq 1\}$ *is a sequence of random* R^k*-vectors such that* $Y_n \to_d Y$ *as* $n \to \infty$*. If* $g : R^k \to R^\ell$ *is continuous on a set* C *with* $P(Y \in C) = 1$*, then* $g(Y_n) \to_d g(Y)$ as $n \to \infty$.

Comment: The requirement is that g must be continuous with Y-probability one. If $g(\cdot)$ is everywhere continuous, then there is no condition to check and we get instant results. For example,

(a) if $\sqrt{n}(\widehat{\theta}_n - \theta) \to_d Z \sim N(0,1)$, then $n(\widehat{\theta}_n - \theta)^2 \to_d Z^2 \sim \chi^2_1$,

(b) if $\sqrt{n}(\widehat{\theta}_n - \theta) \to_d Z \sim N(0, I_k)$, then $n(\widehat{\theta}_n - \theta)'(\widehat{\theta}_n - \theta) \to_d Z'Z \sim \chi^2_k$, and

(c) if $\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d Z \sim N(0, I_k)$, then $R(\sqrt{n}(\widehat{\theta}_n - \theta)) = \sqrt{n}(R\widehat{\theta}_n - R\theta) \to_d RZ \sim N(0, RR')$.

We will use the results from examples (b) and (c) when determining the asymptotic distribution of the Wald statistic.

One use of the continuous mapping theorem, in addition to its use in the examples above, is that it can be used to prove Slutsky's Theorem and numerous related results all in one go. To do this, we just need to establish two preliminary results:

Result 1: Let c be a nonrandom vector. If $Y_n \to_d Y$ and $W_n \to_p c$, then $(Y_n', W_n')' \to_d (Y', c')'$ as

$n \to \infty$.

Note this does **not** hold if c is a random vector. The reason is clear. If it did hold, we would be making the statement that the joint distribution of $(Y_n', W_n')'$ converges in distribution to the joint distribution of $(Y', c')'$, but we have made no assumptions regarding the joint distribution of Y_n and W_n, only about their marginal distributions.

If we know that $Y_n \to_d Y$, $W_n \to_d W$, and (Y_n, W_n) are independent random variables, then it is true that $(Y_n', W_n')' \to_d (Y', W')'$, where Y and W are independent random variables. This follows easily using the df definition of convergence in distribution, because the joint df of independent random variable is just the product of their marginal df's.

The reason Result 1 holds even though we know nothing about the joint distribution of Y_n and W_n, is that since W_n converges in probability to a constant, W_n is asymptotically a constant and, hence, is asymptotically independent of Y_n.

Proof of Result 1: Suppose W_n is a scalar. For $w < c$,

$$F_{(Y_n,W_n)}(y,w) = P(Y_n \leq y,\ W_n \leq w) \leq P(W_n \leq w) \to 0 \text{ as } n \to \infty.$$

For $w > c$,

$$\begin{aligned} F_{(Y_n,W_n)}(y,w) &= P(Y_n \leq y) + P(W_n \leq w) - P(Y_n \leq y \text{ or } W_n \leq w) \\ &\to F_Y(y) + 1 - 1 \\ &= F_Y(y) \end{aligned}$$

if y is a continuity point of Y. Hence, the limit of the df of $F_{(Y_n,W_n)}(y,w)$ equals the df of $(Y', c')'$ at all continuity points of $(Y', c')'$, as desired.

The Generalized Slutsky Theorem implied by Result 2 and the continuous mapping theorem, we get:

Theorem (Generalized Slutsky Theorem): If $Y_n \to_p c$, $W_n \to_d W$, and $g(\cdot,\cdot)$ is continuous with (c, W) probability one (i.e., $g(\cdot,\cdot)$ is continuous at (c,w), $\forall w$ in S such that $P(W \in S) = 1$), then $g(Y_n, W_n) \to_d g(c, W)$ as $n \to \infty$.

The theorem implies that if $Y_n \to_p c$ and $W_n \to_d W$, then
(i) $Y_n + W_n \to_d c + W$,
(ii) $Y_n \cdot W_n \to_d c \cdot W$,
(iii) $W_n / Y_n \to_d W/c$ provided $c \neq 0$, and
(iv) if Y_n is a square weighting matrix such that $Y_n \to_p C$, then $W_n' Y_n W_n \to_d W'CW$.

2 Asymptotic Normality of Extremum Estimators

This section provides conditions under which extremum estimators are asymptotically normally distributed. The results given are not the most general that can be obtained. In particular, we consider cases where the criterion function $\hat{Q}_n(\theta)$ is smooth (i.e., twice differentiable in θ). Many examples satisfy this condition, but some do not. It is possible to obtain asymptotic normality of an extremum estimator with this assumption replaced by weaker assumptions. Also, we only consider the cases in which the estimators have normal asymptotic distribution (or smooth functions of normal distribution by the delta method). Extremum estimators do not always converge weakly to normal distributions. Notable cases include the case in which the true value of a parameter is on the boundary of the parameter space, and the case in which the true value of a parameter is not identified. Partially identified models will be discussed later on in this course.

Heuristics. The idea of deriving asymptotic normality of an estimator in a nonlinear context is to mimic what happens in the linear context. Consider an OLS estimator:

$$\hat{\beta}_n = \arg\min_{\beta \in B} \hat{Q}_n(\beta) = \arg\min_{\beta \in B} ||Y - X\beta||^2/2. \tag{1}$$

The first order condition is a linear equation in $\hat{\beta}_n$:

$$0 = X'(Y - X\hat{\beta}_n). \tag{2}$$

Because the equation is linear, we can easily solve for $\hat{\beta}_n$: $\hat{\beta}_n = (X'X)^{-1} X'Y$. After solving for $\hat{\beta}_n$, studying the asymptotic properties of $\hat{\beta}_n$ is straightforward.

Now consider a nonlinear setting:

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \hat{Q}_n(\theta). \tag{3}$$

The first order condition is nonlinear in $\hat{\theta}_n$:

$$0 = \frac{\partial \hat{Q}\left(\hat{\theta}_n\right)}{\partial \theta}. \tag{4}$$

We cannot solve for $\hat{\theta}_n$ explicitly. However, we can linearize the first order condition using mean-value expansion. The first question is that around which point we do the expansion. From last lecture, we know $\hat{\theta}_n \to_p \theta_0$ for some $\theta_0 \in \Theta$. The point θ_0 thus is the appropriate center point. The mean-value expansion gives us:

$$0 = \frac{\partial \hat{Q}(\theta_0)}{\partial \theta} + \frac{\partial^2 \hat{Q}\left(\check{\theta}_n\right)}{\partial \theta \partial \theta'} \left(\hat{\theta}_n - \theta_0\right). \tag{5}$$

Now, we can "solve for" $\hat{\theta}_n$ explicitly:

$$\hat{\theta}_n - \theta_0 = -\left[\frac{\partial^2 \hat{Q}\left(\tilde{\theta}_n\right)}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial \hat{Q}\left(\theta_0\right)}{\partial\theta}. \tag{6}$$

Therefore, the asymptotic distribution of $\hat{\theta}_n - \theta_0$ can be obtained by studying the asymptotic distributions of $\frac{\partial^2 \hat{Q}(\tilde{\theta}_n)}{\partial\theta\partial\theta'}$ and $\frac{\partial \hat{Q}(\theta_0)}{\partial\theta}$. Hopefully,

$$d_n \frac{\partial \hat{Q}\left(\theta_0\right)}{\partial\theta} \quad \to \quad {}_d Z \tag{7}$$

$$\frac{\partial^2 \hat{Q}\left(\tilde{\theta}_n\right)}{\partial\theta\partial\theta'} \quad \to \quad {}_p B, \tag{8}$$

for some increasing sequence of numbers $\{d_n\}$, some mean-zero random variable Z and some invertible deterministic matrix B. Then by the generalized Slutsky Theorem we have

$$d_n\left(\hat{\theta}_n - \theta_0\right) \to_d BZ. \tag{9}$$

The two paragraphs above gives the heuristic derivation of the extremum estimator. The obviously leaves a lot of questions open. For example: (1) when does (4) hold? (2) is $\hat{Q}_n()$ smooth enough to have second derivatives? (3) does (7) hold? (4) what is B and is it invertible? (5) when does (8) hold? Before considering specific examples, we cannot really answer these questions in detail. Instead, we make high level assumptions that we will verify later in the examples.

Asymptotic normality of the extremum estimator (EE) $\widehat{\theta}_n$ holds under the following two high level assumptions — one concerning the criterion function (CF) and the other concerning the estimator itself.

Assumption CF: (i) θ_0 *is in the interior of* Θ.

(ii) $\hat{Q}_n(\theta)$ *is twice continuously differentiable on some neighborhood* $\Theta_0 \subset \Theta$ *of* θ_0 (*with probability one*).

(iii) $\sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) \to_d N(0, \Omega_0)$.

(iv) for any sequence $\tilde{\theta}_n \to_p \theta_0$, $\frac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\tilde{\theta}_n) - B_0 \to_p 0$ *for some non-stochastic* $d \times d$ *matrix* B_0 *that is nonsingular.*

Assumption EE2: (i) $\widehat{\theta}_n \to_p \theta_0$.

(ii) $\frac{\partial}{\partial\theta}\hat{Q}_n(\widehat{\theta}_n) = o_p(n^{-1/2})$.

Assumption CF will be verified in the examples later.

Assumption EE2(i) assumes that we have already established consistency of $\widehat{\theta}_n$, perhaps by

using the results of Lecture 3. Assumption EE2(ii) requires that the first-order conditions for minimizing the criterion function $\hat{Q}_n(\theta)$ hold approximately. This assumption allows one some leeway in computing the estimator, since it may be difficult and/or costly to find a value $\widehat{\theta}_n$ that exactly satisfies the first-order condition.

Theorem 4.1: *Under Assumptions* CF *and* EE2,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, B_0^{-1}\Omega_0 B_0^{-1}).$$

Note that B_0 must be symmetric given Assumption CF.

Proof of Theorem 4.1: Using CF(ii) and EE2(ii), element-by-element mean value expansions of $\frac{\partial}{\partial\theta}\hat{Q}_n(\widehat{\theta}_n)$ about θ_0 yield

$$o_p(n^{-1/2}) = \frac{\partial}{\partial\theta}\hat{Q}_n(\widehat{\theta}_n) = \frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) + \frac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\tilde{\theta}_n)(\widehat{\theta}_n - \theta_0), \tag{10}$$

where $\tilde{\theta}_n$ lies between $\widehat{\theta}_n$ and θ_0 (and, hence, satisfies $\tilde{\theta}_n \to_p \theta_0$) and $\tilde{\theta}_n$ may differ across the rows of $\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n^*)$. By Assumption CF(iv), and EE2(i),

$$\frac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\tilde{\theta}_n) = B_0 + o_p(1). \tag{11}$$

Multiplying (10) by $\sqrt{n}$ and substituting (11) into (10) gives

$$o_p(1) = \sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) + (B_0 + o_p(1))\sqrt{n}(\widehat{\theta}_n - \theta_0). \tag{12}$$

Rearranging (12) yields

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &= -(B_0 + o_p(1))^{-1}\sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) + o_p(1) \\ &= -B_0^{-1}\sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) + o_p(1) \\ &\to_d N(0, B_0^{-1}\Omega_0 B_0^{-1})\end{aligned} \tag{13}$$

using CF(iii) and CF(iv) (i.e., nonsingularity of B_0). □

The following sufficient condition makes it easier to verify Assumption CF.

Assumption CF(iv)* $\sup_{\theta\in\Theta_0}\left|\left|\frac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\theta) - B(\theta)\right|\right| \to_p 0$ *for some non-stochastic $d \times d$ matrix-valued function $B(\theta)$ that is continuous at θ_0 and for which $B_0 = B(\theta_0)$ is nonsingular.*

The following Lemma shows that Assumption CF(iv)* is a sufficient condition for Assumption CF(iv).

Lemma 4.1: *Assumption* CF(iv)* *implies Assumption* CF(iv).

The proof of Lemma 4.1 uses the following lemma. Let "wp $\to 1$" abbreviate "with probability that goes to one as $n \to \infty$."

Lemma 4.2: *Suppose* (i) $\widehat{\beta}_n \to_p \beta_0 \in R^s$, (ii) $\sup_{\beta \in B(\beta_0,\varepsilon)} |L_n(\beta) - L(\beta)| \to_p 0$ *for some* $\varepsilon > 0$, *and* (iii) *the non-stochastic function* $L(\beta)$ *is continuous at* β_0. *Then,*

$$L_n(\widehat{\beta}_n) \xrightarrow{p} L(\beta_0).$$

Proof of Lemma 4.2: We have

$$\begin{aligned} & |L_n(\widehat{\beta}_n) - L(\beta_0)| \\ = \; & |L_n(\widehat{\beta}_n) - L(\widehat{\beta}_n) + L(\widehat{\beta}_n) - L(\beta_0)| \\ \leq \; & |L_n(\widehat{\beta}_n) - L(\widehat{\beta}_n)| + |L(\widehat{\beta}_n) - L(\beta_0)| \\ \leq \; & \sup_{\beta \in B(\beta_0,\varepsilon)} |L_n(\beta) - L(\beta)| + |L(\widehat{\beta}_n) - L(\beta_0)| \\ \xrightarrow{p} \; & 0, \end{aligned}$$

where the first inequality holds by the triangle inequality, the second inequality holds wp $\to 1$ because $\widehat{\beta}_n \in B(\beta_0, \varepsilon)$ wp $\to 1$ by (i), and the convergence to zero holds using (i), (ii), and (iii). □

Proof of Lemma 4.1: Lemma 4.1 is a direct application of Lemma 4.2. □

3 Examples

We now provide sufficient conditions for Assumption CF and discuss the form of the asymptotic covariance matrix, $B_0^{-1}\Omega_0 B_0^{-1}$, for each of the five examples introduced above.

(1) ML Estimator (with iid observations): We have

$$\begin{aligned} & \frac{\partial}{\partial \theta}\hat{Q}_n(\theta) = -\frac{1}{n}\textstyle\sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(W_i, \theta), \\ & \frac{\partial^2}{\partial \theta \partial \theta'}\hat{Q}_n(\theta) = -\frac{1}{n}\textstyle\sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta'} \log f(W_i, \theta), \\ & \Omega_0 = E\frac{\partial}{\partial \theta} \log f(W_i, \theta_0) \frac{\partial}{\partial \theta'} \log f(W_i, \theta_0), \text{ and} \\ & B(\theta) = -E\frac{\partial^2}{\partial \theta \partial \theta'} \log f(W_i, \theta). \end{aligned} \tag{14}$$

From (13), the linear expansion of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &= -B_0^{-1}\sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) + o_p(1) \\ &= -\left(-E\frac{\partial^2}{\partial\theta\partial\theta'}\log f(W_i,\theta_0)\right)^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^n\frac{\partial}{\partial\theta}\log f(W_i,\theta_0) + o_p(1). \qquad (15)\end{aligned}$$

Assumption CF(ii) holds if $f(w,\theta)$ is twice continuously differentiable in θ on some neighborhood $\Theta_0 \subset \Theta$ of θ_0 for all w in the support $\mathcal{W}$ of W_i.

Assumption CF(iii) holds by the central limit theorem (CLT) for iid random vectors with finite second moment provided

$$E\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) = 0 \text{ and } E\left|\left|\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)\right|\right|^2 < \infty. \qquad (16)$$

The former condition holds by the first order conditions for minimization of $Q(\theta)$ over Θ, provided θ_0 is an interior point of Θ. That is,

$$0 = \frac{\partial}{\partial\theta}Q(\theta_0) = -\frac{\partial}{\partial\theta}E\log f(W_i,\theta_0) = -E\frac{\partial}{\partial\theta}\log f(W_i,\theta_0) = E\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0), \qquad (17)$$

provided the inter-change of E and $\partial/\partial\theta$ in the third equality is justified. (Sufficient conditions for interchange are that $\log f(w,\theta)$ is continuously differentiable in θ on a neighborhood Θ_0 of θ_0 for all $w \in \mathcal{W}$ and $E\sup_{\theta\in\Theta_0}\left|\left|\frac{\partial}{\partial\theta}\log f(W_i,\theta)\right|\right| < \infty$. Sufficiency holds by the dominated convergence theorem with the mean value theorem applied to obtain the dominating function.) Note that (16) holds by definition of θ_0 (as the value that minimizes $Q(\theta)$), whether or not the model is correctly specified.

The third condition in (14) is equivalent to requiring the **information matrix** at θ_0 to be well defined. The information matrix is

$$\mathcal{I}_0 = E\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)\frac{\partial}{\partial\theta'}\log f(W_i,\theta_0). \qquad (18)$$

Assumption CF(iv) holds if Assumption CF(iv)* holds by Lemma 3.1. The latter assumption holds if $\left\{\frac{\partial^2}{\partial\theta\partial\theta'}\log f(W_i,\theta) : i \geq 1\right\}$ satisfies a uniform WLLN over $\theta \in \Theta_0$, B_0 is nonsingular, and $B(\theta)$ is continuous at θ_0. By the ULLN in Lecture 2 and Lemma 3.2, the convergence holds and $B(\theta)$ is continuous at θ_0, if $\frac{\partial^2}{\partial\theta\partial\theta'}\log f(w,\theta)$ is continuous in θ on Θ_0 $\forall w \in \mathcal{W}$ (as assumed above),

$$E\sup_{\theta\in\Theta_0}\left|\left|\frac{\partial^2}{\partial\theta\partial\theta'}\log f(W_i,\theta)\right|\right| < \infty, \qquad (19)$$

and Θ_0 is compact.

The *information matrix equality* holds if the parametric model is *correctly specified* and one can switch the order of differentiation and integration in the definition of $B_0 = B(\theta_0)$. (The latter holds under weak assumptions.) The information matrix equality is

$$B_0 = \Omega_0. \tag{20}$$

Hence, in this case, the asymptotic covariance matrix of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is the inverse of the information matrix

$$B_0^{-1}. \tag{21}$$

The information matrix equality is derived as follows: We differentiate the equality $1 = \int f(w,\theta)d\mu(w)$ with respect to θ to get[2]

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int f(w,\theta) d\mu(w) \\ &= \int \frac{\partial}{\partial \theta} f(w,\theta) d\mu(w) \\ &= \int \frac{\partial}{\partial \theta} \log f(w,\theta) \cdot f(w,\theta) d\mu(w). \end{aligned} \tag{22}$$

Differentiating again gives

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta'} \int \frac{\partial}{\partial \theta} \log f(w,\theta) \cdot f(w,\theta) d\mu(w) \\ &= \int \frac{\partial}{\partial \theta'} \left(\frac{\partial}{\partial \theta} \log f(w,\theta) \cdot f(w,\theta) \right) d\mu(w) \\ &= \int \frac{\partial^2}{\partial \theta \partial \theta'} \log f(w,\theta) \cdot f(w,\theta) d\mu(w) + \int \frac{\partial}{\partial \theta} \log f(w,\theta) \cdot \frac{\partial}{\partial \theta'} f(w,\theta) d\mu(w) \\ &= \int \frac{\partial^2}{\partial \theta \partial \theta'} \log f(w,\theta) \cdot f(w,\theta) d\mu(w) \\ &\quad + \int \frac{\partial}{\partial \theta} \log f(w,\theta) \cdot \frac{\partial}{\partial \theta'} \log f(w,\theta) \cdot f(w,\theta) d\mu(w). \end{aligned} \tag{23}$$

Now, if the model is correctly specified, then the density of W_i is $f(w,\theta_0)$ and the equation above yields

$$\begin{aligned} 0 &= E \frac{\partial^2}{\partial \theta \partial \theta'} \log f(W_i,\theta_0) + E \frac{\partial}{\partial \theta} \log f(W_i,\theta_0) \cdot \frac{\partial}{\partial \theta'} \log f(W_i,\theta_0) \\ &= -B_0 + \Omega_0. \end{aligned} \tag{24}$$

[2] μ is a dominating measure on the space of w. For example, μ can be the Lebesgue measure if W_i is a continuous random vector on the Euclidean space. When μ is the Lebesgue measure, a perhaps more familiar form of the integral $\int \cdot d\mu(w)$ is $\int \cdot dw$.

(2) LS Estimator: We have

$$\frac{\partial}{\partial\theta}\hat{Q}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i,\theta))\frac{\partial}{\partial\theta}g(X_i,\theta), \tag{25}$$

$$\frac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \left(\frac{\partial}{\partial\theta}g(X_i,\theta)\frac{\partial}{\partial\theta'}g(X_i,\theta) - (Y_i - g(X_i,\theta))\frac{\partial^2}{\partial\theta\partial\theta'}g(X_i,\theta)\right),$$

$$\Omega_0 = EU_i^2 \frac{\partial}{\partial\theta}g(X_i,\theta_0)\frac{\partial}{\partial\theta'}g(X_i,\theta_0), \text{ where } U_i = Y_i - g(X_i,\theta_0),$$

$$B(\theta) = E\frac{\partial}{\partial\theta}g(X_i,\theta)\frac{\partial}{\partial\theta'}g(X_i,\theta) - E(Y_i - g(X_i,\theta))\frac{\partial^2}{\partial\theta\partial\theta'}g(X_i,\theta), \text{ and}$$

$$B_0 = E\frac{\partial}{\partial\theta}g(X_i,\theta_0)\frac{\partial}{\partial\theta'}g(X_i,\theta_0) - E(Y_i - g(X_i,\theta_0))\frac{\partial^2}{\partial\theta\partial\theta'}g(X_i,\theta_0). \tag{26}$$

By iterated expectations,

$$\Omega_0 = E\sigma^2(X_i)\frac{\partial}{\partial\theta}g(X_i,\theta_0)\frac{\partial}{\partial\theta'}g(X_i,\theta_0), \text{ where } \sigma^2(X_i) = E(U_i^2|X_i). \tag{27}$$

From (13), the linear expansion of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &= -B_0^{-1}\sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) + o_p(1) \\ &= -B_0^{-1}\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n U_i\frac{\partial}{\partial\theta}g(X_i,\theta_0) + o_p(1).\end{aligned} \tag{28}$$

Assumption CF(ii) holds if $g(x,\theta)$ is twice continuously differentiable in θ on some neighborhood $\Theta_0 \subset \Theta$ of θ_0 for all x in the support $\mathcal{X}$ of X_i.

Assumption CF(iii) holds by the CLT provided

$$E\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) = 0 \text{ and } EU_i^2\left|\left|\frac{\partial}{\partial\theta}g(X_i,\theta_0)\right|\right|^2 < \infty. \tag{29}$$

The first condition holds because, by definition, θ_0 minimizes $Q(\theta)$ over Θ. Since θ_0 is assumed to be an interior point of Θ, the first order conditions for the minimization of $Q(\theta)$ give

$$\begin{aligned}0 &= \frac{\partial}{\partial\theta}Q(\theta_0) = \frac{\partial}{\partial\theta}E(Y_i - g(X_i,\theta_0))^2/2 \\ &= E\frac{\partial}{\partial\theta}(Y_i - g(X_i,\theta_0))^2/2 = E\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0),\end{aligned} \tag{30}$$

provided the interchange of E and $\partial/\partial\theta$ in the third equality is justified. (Sufficient conditions are that $g(x,\theta)$ is continuously differentiable in a neighborhood Θ_0 of θ_0 for all x in $\mathcal{X}$ and $E\sup_{\theta\in\Theta_0}\left|\left|(Y_i - g(X_i,\theta))\frac{\partial}{\partial\theta}g(X_i,\theta)\right|\right| < \infty$.) As in the ML example, $E\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) = 0$ by definition of θ_0 whether or not the model is correctly specified.

Assumption CF(iv) holds if Assumption CF(iv)* holds by Lemma 4.1. The latter assumption holds by the ULLN in Lecture 2 and Lemma 3.2 provided that $\frac{\partial}{\partial\theta}g(x,\theta)$ and $\frac{\partial^2}{\partial\theta\partial\theta'}g(x,\theta)$ are continuous in θ on Θ_0 $\forall x \in \mathcal{X}$ (as is assumed above),

$$E \sup_{\theta\in\Theta_0} \left[\left|\left| \frac{\partial}{\partial\theta} g(X_i,\theta) \right|\right|^2 + \left|\left| (Y_i - g(X_i,\theta)) \frac{\partial^2}{\partial\theta\partial\theta'} g(X_i,\theta_0) \right|\right| \right] < \infty,$$

Θ_0 is compact, and B_0 is nonsingular.

If the model is correctly specified (i.e., $E(U_i|X_i) = 0$ a.s.) or $g(X_i,\theta)$ is linear in θ (i.e., $g(X_i,\theta) = X_i'\theta$), then the second summand of B_0 is zero, which gives

$$B_0 = E \frac{\partial}{\partial\theta} g(X_i,\theta_0) \frac{\partial}{\partial\theta'} g(X_i,\theta_0). \tag{31}$$

If, in addition, the errors $\{U_i : i \geq 1\}$ are homoskedastic, i.e., $\sigma^2(X_i) = \sigma^2$ a.s. for some $\sigma^2 > 0$, then

$$\Omega_0 = \sigma^2 B_0$$

and the asymptotic covariance matrix of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is

$$\sigma^2 B_0^{-1}. \tag{32}$$

(3) GMM Estimator (with iid observations): We have

$$\begin{aligned}
&\tfrac{\partial}{\partial\theta}\hat{Q}_n(\theta) = \left[\tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{\partial}{\partial\theta'} g(W_i,\theta)\right]' A_n' A_n \tfrac{1}{n}\textstyle\sum_{i=1}^n g(W_i,\theta) \\
&\left[\tfrac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\theta)\right]_{mj} = \tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{\partial}{\partial\theta_m} g(W_i,\theta)' A_n' A_n \tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{\partial}{\partial\theta_j} g(W_i,\theta) \\
&\qquad\qquad + \tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{\partial^2}{\partial\theta_m\partial\theta_j} g(W_i,\theta)' A_n' A_n \tfrac{1}{n}\textstyle\sum_{i=1}^n g(W_i,\theta), \text{ for } m,j = 1,...,d, \\
&\Omega_0 = \Gamma_0' A' A V_0 A' A \Gamma_0, \text{ where} \\
&V_0 = E g(W_i,\theta_0) g(W_i,\theta_0)', \ \Gamma_0 = E \tfrac{\partial}{\partial\theta'} g(W_i,\theta_0), \text{ and } A_n \xrightarrow{p} A, \\
&[B(\theta)]_{mj} = E \tfrac{\partial}{\partial\theta_m} g(W_i,\theta)' A' A E \tfrac{\partial}{\partial\theta_j} g(W_i,\theta) \\
&\qquad\qquad + E \tfrac{\partial^2}{\partial\theta_m\partial\theta_j} g(W_i,\theta)' A' A E g(W_i,\theta), \text{ for } m,j = 1,...,d, \text{ and} \\
&B_0 = \Gamma_0' A' A \Gamma_0.
\end{aligned} \tag{33}$$

From (13), the linear expansion of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &= -B_0^{-1}\sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) + o_p(1) \\ &= -(\Gamma_0' A' A \Gamma_0)^{-1}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'} g(W_i,\theta_0)\right]' A_n' A_n \frac{1}{n^{1/2}}\sum_{i=1}^n g(W_i,\theta_0) + o_p(1) \\ &= -(\Gamma_0' A' A \Gamma)^{-1}\Gamma_0' A' A \frac{1}{n^{1/2}}\sum_{i=1}^n g(W_i,\theta_0) + o_p(1). \end{aligned} \tag{34}$$

Assumption CF(ii) holds if $g(w,\theta)$ is twice continuously differentiable in θ on some neighborhood $\Theta_0 \subset \Theta$ of θ_0 for all w in the support $\mathcal{W}$ of W_i.

Assumption CF(iii) holds by the CLT applied to $\frac{1}{\sqrt{n}}\sum_{i=1}^n g(W_i,\theta_0)$ since $Eg(W_i,\theta_0) = 0$, by the WLLN applied to $\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'} g(W_i,\theta_0)$, and by the assumption that $A_n \to_p A$. The CLT and WLLN require

$$E||g(W_i,\theta_0)||^2 < \infty \text{ and } E\left|\left|\frac{\partial}{\partial\theta'} g(W_i,\theta_0)\right|\right| < \infty.$$

Assumption CF(iv) holds if Assumption CF(iv)* holds by Lemma 4.1. The latter assumption holds if $\{g(W_i,\theta) : i \geq 1\}$, $\left\{\frac{\partial}{\partial\theta'} g(W_i,\theta) : i \geq 1\right\}$, and $\left\{\frac{\partial^2}{\partial\theta_m\partial\theta_j} g(W_i,\theta) : i \geq 1\right\}$ for $m,j = 1,...,d$ satisfy uniform WLLNs over Θ_0, Γ_0 is full rank, A is nonsingular, and $Eg(W_i,\theta)$, $E\frac{\partial}{\partial\theta'} g(W_i,\theta)$, and $E\frac{\partial^2}{\partial\theta_m\partial\theta_j} g(W_i,\theta)$ are continuous at θ_0 $\forall m,j = 1,...,d$. By the ULLN in Lecture 2 is satisfied and $B(\theta)$ is continuous at θ_0 by Lemma 3.2, if $g(w,\theta)$, $\frac{\partial}{\partial\theta} g(w,\theta)$, and $\frac{\partial^2}{\partial\theta_m\partial\theta_j} g(w,\theta)$ are continuous in θ on Θ_0 $\forall w \in \mathcal{W}$ $\forall m,j = 1,...,d$ (as is assumed above),

$$E\sup_{\theta\in\Theta_0}\left|\left|\frac{\partial}{\partial\theta} g(W_i,\theta)\right|\right| < \infty,\ E\sup_{\theta\in\Theta_0}\left|\left|\frac{\partial^2}{\partial\theta_m\partial\theta_j} g(W_i,\theta)\right|\right| < \infty \tag{35}$$

$\forall m,j = 1,...,d$, and Θ_0 is compact.

If the number k of moment conditions equals the dimension d of θ_0, then Γ_0 and A are nonsingular square matrices, $B_0^{-1} = \Gamma_0^{-1}A^{-1}(A')^{-1}(\Gamma_0')^{-1}$, and the asymptotic covariance matrix of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ simplifies:

$$B_0^{-1}\Omega_0 B_0^{-1} = \Gamma_0^{-1}V_0(\Gamma_0^{-1})'. \tag{36}$$

(4) MD Estimator: We have

$$\frac{\partial}{\partial\theta}\hat{Q}_n(\theta) = -\left(\frac{\partial}{\partial\theta'} g(\theta)\right)' A_n' A_n(\widehat{\pi}_n - g(\theta)) \tag{37}$$

$$\left[\frac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\theta)\right]_{mj} = \frac{\partial}{\partial\theta_m} g(\theta)' A_n' A_n \frac{\partial}{\partial\theta_j} g(\theta) - \frac{\partial^2}{\partial\theta_m\partial\theta_j} g(\theta)' A_n' A_n(\widehat{\pi}_n - g(\theta))$$

$$\text{for } m,j = 1,...,d.$$

We assume

$$\sqrt{n}(\widehat{\pi}_n - \pi_0) \xrightarrow{d} N(0, V_0) \text{ and } A_n \xrightarrow{p} A. \tag{38}$$

This assumption can be established using Theorem 4.1 if $\widehat{\pi}_n$ is an extremum estimator.

Given (38), we have

$$\begin{array}{l} \Omega_0 = \Gamma_0' A' A V_0 A' A \Gamma_0, \text{ where } \Gamma_0 = \frac{\partial}{\partial \theta'} g(\theta_0), \text{ and} \\ [B(\theta)]_{mj} = \frac{\partial}{\partial \theta_m} g(\theta)' A' A \frac{\partial}{\partial \theta_j} g(\theta) - \frac{\partial^2}{\partial \theta_m \partial \theta_j} g(\theta)' A' A (\pi_0 - g(\theta)) \text{ for } m, j = 1, ..., d. \end{array} \tag{39}$$

Assumption CF(ii) holds if $g(\theta)$ is twice differentiable on some neighborhood $\Theta_0 \subset \Theta$ of θ_0.

CF(iii) holds by (38) and (39) provided the restrictions $\pi_0 = g(\theta_0)$ hold (where θ_0 is the probability limit of $\widehat{\theta}_n$). Note that $\pi_0 = g(\theta_0)$ implies that B_0 simplifies, since the second summand of $[B(\theta_0)]_{mj}$ equals zero:

$$B_0 = \Gamma_0' A' A \Gamma_0. \tag{40}$$

CF(iv) holds under the assumptions given above, provided Γ_0 and A are full rank.

Using (13), the linear expansion for $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is obtained by substituting the linear expansion for $\sqrt{n}(\widehat{\pi}_n - \pi_0)$ into

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_n - \theta_0) &= -B_0^{-1} \sqrt{n} \frac{\partial}{\partial \theta} \hat{Q}_n(\theta_0) + o_p(1) \\ &= (\Gamma_0' A' A \Gamma_0)^{-1} \Gamma_0' A' A \sqrt{n}(\widehat{\pi}_n - \pi_0) + o_p(1). \end{aligned} \tag{41}$$

(5) TS Estimator: We have

$$\begin{array}{l} \frac{\partial}{\partial \theta} \hat{Q}_n(\theta) = \left(\frac{\partial}{\partial \theta'} G_n(\theta, \widehat{\tau}_n)\right)' A_n' A_n G_n(\theta, \widehat{\tau}_n), \\ \left[\frac{\partial^2}{\partial \theta \partial \theta'} \hat{Q}_n(\theta)\right]_{mj} = \frac{\partial}{\partial \theta_m} G_n(\theta, \widehat{\tau}_n)' A_n' A_n \frac{\partial}{\partial \theta_j} G_n(\theta, \widehat{\tau}_n) + \frac{\partial^2}{\partial \theta_m \partial \theta_j} G_n(\theta, \widehat{\tau}_n)' A_n' A_n G_n(\theta, \widehat{\tau}_n) \\ \qquad\qquad \text{for } m, j = 1, ..., d. \end{array} \tag{42}$$

For brevity, we do not give sufficient conditions for Assumption CF for this example. We show what B_0 and Ω_0 equal in this example by making intermediate, rather than primitive assumptions.

We assume the following:

$$
\begin{aligned}
&\sqrt{n}\begin{pmatrix} G_n(\theta_0,\tau_0) \\ \widehat{\tau}_n - \tau_0 \end{pmatrix} \to_d \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} V_{10} & V_{20} \\ V_{20}' & V_{30} \end{pmatrix}\right), \\
&\frac{\partial}{\partial\theta'} G_n(\theta_0,\widehat{\tau}_n) \to_p \frac{\partial}{\partial\theta'} G(\theta_0,\tau_0) = \Gamma_0, \\
&A_n \to_p A, \\
&\frac{\partial}{\partial\tau'} G_n(\theta_0,\tau_n^*) \to_p \frac{\partial}{\partial\tau'} G(\theta_0,\tau_0) = \Lambda_0, \\
&\sup_{\theta\in\Theta_0} \left|\left| \frac{\partial^2}{\partial\theta\partial\theta'} \hat{Q}_n(\theta) - B(\theta) \right|\right| \to_p 0, \text{ where} \\
&[B(\theta)]_{mj} = \frac{\partial}{\partial\theta_m} G(\theta,\tau_0)' A' A \frac{\partial}{\partial\theta_j} G(\theta,\tau_0) + \frac{\partial^2}{\partial\theta_m\partial\theta_j} G(\theta,\tau_0)' A' A G(\theta,\tau_0), \text{ and} \\
&B_0 = \Gamma_0' A' A \Gamma_0.
\end{aligned}
\tag{43}
$$

where τ_n^* is any random vector that satisfies $\tau_n^* \to_p \tau_0$. The first convergence result in (43) is verified by applying the multivariate CLT to the linear expansion for $n^{1/2}(\widehat{\tau}_n - \tau_0)$ coupled with (i) the normalized sample average $n^{-1/2}\sum_{i=1}^n g(W_i,\theta_0)$ in the GMM case and (ii) the linear expansion for $n^{1/2}(\widehat{\pi}_n - g(\theta_0,\tau_0)) = n^{1/2}(\widehat{\pi}_n - \pi_0)$ in the minimum distance case. The second and fourth convergence results in (43) are verified by using Lemma 4.1.

To find the asymptotic distribution $N(0,\Omega_0)$ of $\sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0)$, as required for CF(ii), we carry out element-by-element mean value expansions of $\sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0)$ about τ_0 and use the assumptions of (43):

$$
\begin{aligned}
\sqrt{n}\tfrac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) &= \left(\frac{\partial}{\partial\theta'} G_n(\theta_0,\widehat{\tau}_n)\right)' A_n' A_n \sqrt{n} G_n(\theta_0,\widehat{\tau}_n) \\
&= (\Gamma_0 + o_p(1))' A_n' A_n \left(\sqrt{n} G_n(\theta_0,\tau_0) + \frac{\partial}{\partial\tau'} G_n(\theta_0,\tau_n^*)\sqrt{n}(\widehat{\tau}_n - \tau_0)\right) \\
&= (\Gamma_0 + o_p(1))' A_n' A_n [I_k \vdots \Lambda_0 + o_p(1)] \sqrt{n}\begin{pmatrix} G_n(\theta_0,\tau_0) \\ \widehat{\tau}_n - \tau_0 \end{pmatrix} \\
&\xrightarrow{d} \Gamma_0' A' A (Z_1 + \Lambda_0 Z_2) \\
&\sim N(0,\Omega_0), \text{ where} \\
\Omega_0 &= \Gamma_0' A' A (V_{10} + \Lambda_0 V_{20}' + V_{20}\Lambda_0' + \Lambda_0 V_{30}\Lambda_0') A' A \Gamma_0
\end{aligned}
\tag{44}
$$

and τ_n^* lies between $\widehat{\tau}_n$ and τ_0 and may differ across rows of $\frac{\partial}{\partial\tau'}G_n(\theta_0,\tau_n^*)$.

Note that if $\Lambda_0 = 0$, as occurs with the feasible GLS estimator of linear and nonlinear regression models, then Ω_0 simplifies to an expression that is the same as one would get if τ_0 replaced $\widehat{\tau}_n$ in $\hat{Q}_n(\theta)$. In this case, the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is the same whether τ_0 is known or

is estimated. In general, however, $\Lambda_0 \neq 0$ and the estimator of τ_0 by $\widehat{\tau}_n$ affects the limit distribution of $\widehat{\theta}_n$.

Using (13), the linear expansion for $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is obtained by substituting the linear expansion for $\sqrt{n}(G_n(\theta_0, \tau_0)', (\widehat{\tau}_n - \tau_0)')'$ into the following expression:

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &= -B_0^{-1}\sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0) + o_p(1) \\ &= (\Gamma_0' A' A \Gamma_0)^{-1}\Gamma_0' A' A[I_k \vdots \Lambda_0]\sqrt{n}\begin{pmatrix} G_n(\theta_0, \tau_0) \\ \widehat{\tau}_n - \tau_0 \end{pmatrix} + o_p(1). \qquad (45)\end{aligned}$$

To show Assumption CF(iv) in this example and in others, the following lemma is useful.

Lemma 4.3: *Suppose* (i) $\widehat{\beta}_n \to_p \beta_0 \in R^s$, (ii) $\sup_{\gamma\in\Gamma}\sup_{\beta\in B(\beta_0,\varepsilon)}|L_n(\gamma,\beta) - L(\gamma,\beta)| \to_p 0$ *for some* $\varepsilon > 0$, *and* (iii) $L(\gamma,\beta)$ *is continuous in* β *at* β_0 *uniformly over* $\gamma \in \Gamma$ *(i.e.,* $\lim_{\beta\to\beta_0}\sup_{\gamma\in\Gamma}|L(\gamma,\beta) - L(\gamma,\beta_0)| = 0$*.) Then,*

$$\sup_{\gamma\in\Gamma}|L_n(\gamma,\widehat{\beta}_n) - L(\gamma,\beta_0)| \xrightarrow{p} 0.$$

Proof of Lemma 4.3: (Problem Set Question #4. Hint: similar to Lemma 4.2)

Note that condition (iii) of the Lemma holds if Γ is compact and $L(\gamma,\beta)$ is continuous in (γ,β) on $\Gamma \times B(\beta_0, \varepsilon)$.