LECTURE 9: ASYMPTOTICS II

MAXIMUM LIKELIHOOD ESTIMATION:

*Jensen's Inequality:* Suppose $X$ is a random variable with $\mathbb{E}(X) = \mu$, and $f$ is a convex function.

Then

$$\mathbb{E}(f(X)) \geq f(\mathbb{E}(X)).$$

See figure 9.1

This inequality will be used to get the consistency of the ML estimator.

Let $p(x|\theta)$ be the probability density function of $X$ given the parameter $\theta$.

Consider a random sample of $n$ observations and let

$$\ell(\theta|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \ln p(x_i|\theta)$$

be the log likelihood function.

Assume $\theta_0$ is the true value and that $\frac{d\ln p}{d\theta}$ exists in an interval including $\theta_0$.

We can get the consistency of the ML estimator immediately. We will use assumptions 1-3 to get the asymptotic normality of a consistent estimator in general and the ML estimator in particular.

**CONSISTENCY:**

Suppose $\overline{\theta}$ is an estimator for $\theta$. We would like to require that the probability of $\overline{\theta}$ being close to the true value of $\theta$ (i.e. $\theta_0$) should increase as the sample size increases. We state this idea formally by making use of the concept of convergence in probability associated with WLLN.

*Definition:* An estimator $\overline{\theta}$ is said to be consistent for $\theta$ if $\text{plim} \overline{\theta} = \theta_0$. 
**Proposition:** If \( \ln p \) is differentiable, then the ML equation

\[
\frac{d\ell}{d\theta} = 0 \text{ (first order condition)}
\]

has a root with probability 1 which is consistent for \( \theta \), i.e. ML estimator for \( \theta \) is consistent.

**Proof:** Using Jensen's inequality for concave functions

\[
E \ln \left[ \frac{p(\theta_0 - \delta)}{p(\theta_0)} \right] |_{\theta_0} < 0; \quad E \ln \left[ \frac{p(\theta_0 + \delta)}{p(\theta_0)} \right] |_{\theta_0} < 0 \quad \text{where } \delta \text{ is a small number.}
\]

(strict inequality unless \( p \) does not depend on \( \theta \))

To see this, note that

\[
E \ln \left[ \frac{p(\theta_0 + \delta)}{p(\theta_0)} \right] < \ln E \left[ \frac{p(\theta_0 + \delta)}{p(\theta_0)} \right] = \ln \int p(\theta_0 + \delta) dx = \ln 1 = 0.
\]

Then noting the definition of \( \ell(\theta) \) and using SLLN,

\[
\frac{1}{n} [\ell(\theta_0 \pm \delta) - \ell(\theta_0)] < 0
\]

\( \Rightarrow \) \( \ell(\theta) \) has a local maximum at \( \theta_0 \)

\( \Rightarrow \) first order condition is satisfied at \( \theta_0 \)

**Discussion?** Note that we have not shown that the MLE is a global max -- this requires more conditions.

**ASYMPTOTIC NORMALITY OF CONSISTENT ESTIMATORS:**

Note: In the following discussion, \( d \ln p/d\theta_0 \) means that the derivative \( d \ln p/d\theta \) is evaluated at the true value \( \theta_0 \).

**Assumption 1:**
exist in an interval including \( \theta_0 \).

**Assumption 2:**

\[
E \left( \frac{p'}{p} \bigg\vert \theta \right) = 0; \ E \left( \frac{p''}{p} \bigg\vert \theta \right) = 0; \ E \left( \frac{p'''}{p} \bigg\vert \theta \right) > 0
\]

where \( p' = dp/d\theta \) and \( p'' = d^2p/d\theta^2 \). These usually hold in the problems we will see. For example

\[
\int p(x|\theta)\,dx = 1 \Rightarrow \int p'\,dx = 0 = \int \left( \frac{p'}{p} \right)\,p\,dx = E \left( \frac{p'}{p} \bigg\vert \theta \right)
\]

which gives the first part of this assumption. Differentiating again gives the second part. As for

the third part, it is obviously the expectation of something positive.

**Assumption 3:**

\[
\left| \frac{d^3 \ln p}{d\theta^3} \right| < M(x) \text{ where } E[M(x)] < K.
\]

This is a purely technical assumption. It will basically control the expected error in Taylor expansions.

**Proposition:** Let

\[
- E \left[ \frac{d^2 \ln p}{d\theta_0^2} \right] = E \left[ \left( \frac{d \ln p}{d\theta_0} \right)^2 \right] = i(\theta_0).
\]

Let \( \overline{\theta} \) be a consistent estimator for \( \theta \). Then

\[
\sqrt{n} \left[ (\overline{\theta} - \theta_0) i(\theta_0) - \frac{1}{n} \frac{d\ell}{d\theta_0} \right] \rightarrow 0
\]

in probability.
Proof: From the first order condition, we get the following expansion:

\[
0 = \frac{d\ell}{d\theta_0} + (\bar{\theta} - \theta_0)\frac{d^2\ell}{d\theta_0^2} + \frac{(\bar{\theta} - \theta_0)^2}{2} \frac{d^3\ell}{d\theta_0^3}
\]

\[
\Rightarrow \sqrt{n}(\bar{\theta} - \theta_0) = -\frac{1}{\sqrt{n}} \frac{d\ell}{d\theta_0} + \frac{1}{n} \left[ \frac{d^2\ell}{d\theta_0^2} + \frac{(\bar{\theta} - \theta_0) d^3\ell}{d\theta_0^3} \right]
\]

Taking the probability limit we note that the first expression in the denominator converges to -\(i(\theta_0)\) and the second expression in the denominator converges to 0 (because of consistency of \(\bar{\theta}\) and assumption 3), we get the required result. ■

Proposition: (Asymptotic Normality)

\[
\sqrt{n}(\bar{\theta} - \theta_0) \sim q(0,i(\theta_0)^{-1})
\]

Proof: Note that

\[
\sqrt{n}(\bar{\theta} - \theta_0)i(\theta_0)\text{has the same asymptotic distribution as } \frac{1}{\sqrt{n}} \frac{d\ell}{d\theta_0} = \sqrt{n} \left( \frac{1}{n} \sum \frac{d\ln p}{d\theta_0} \right)
\]

(To see that this is true, check property 1 under "Some properties of convergence in probability and convergence in distribution" in Lecture 8 and the previous proposition.)

We know that

\[
E\left[ \frac{d\ln p}{d\theta_0} \right] = 0
\]

since

\[
\int p(x|\theta_0)dx = 1 \Rightarrow \int p'dx = 0 = E\left[ \frac{d\ln p}{d\theta_0} \right]
\]
Note that differentiating $\int p \, d \ln p \, dx = 0$ again implies that

$$\int p \left( \frac{p'}{p} \right) d \ln p \, dx + \int p d^2 \ln p \, dx = 0.$$ 

The first term is just the variance of $d \ln p/d\theta_0$ and the second expression is $-i(\theta_0)$. Thus,

$$V \left[ \frac{d \ln p}{d \theta_0} \right] = i(\theta_0).$$

Now we use the Central Limit Theorem for

$$\sqrt{n} \left( \frac{1}{n} \sum \frac{d \ln p}{d \theta_0} \right) \text{ with } \mathbb{E} \left[ \frac{d \ln p}{d \theta_0} \right] = 0, \quad V \left[ \frac{d \ln p}{d \theta_0} \right] = i(\theta_0)$$

to obtain

$$\sqrt{n} (\bar{\theta} - \theta_0) i(\theta_0) \sim q(0,i(\theta_0)).$$

Thus (using $z \sim N(0, \Sigma) \Rightarrow Az \sim N(0,A\Sigma A')$)

$$\sqrt{n} (\bar{\theta} - \theta_0) \sim q(0,i(\theta_0))^{-1} \quad \blacksquare$$

**Basic result:** Approximate the distribution of $(\bar{\theta} - \theta_0)$ by $q \left( \frac{i(\theta_0)^{-1}}{n} \right)$. 

($i^{-1}$ refers to the reciprocal, not the function inverse.) Of course, $i(\theta_0)^{-1}$ is consistently estimated by $i(\bar{\theta})^{-1}$ under our assumptions. (why?)

**Applications:**

1. This will give the exact distribution in estimating a normal mean. Check this.
2. Consider a regression model with $Ey = X\beta$, $Vy = \sigma^2 I$ and $y \sim$ normal. Check that the asymptotic distribution of $\hat{\beta}$ is equal to its exact distribution.
MISCELLANEOUS USEFUL RESULTS:

1. **Consistency of continuous functions of ML estimators:**

Suppose \( \hat{\theta} \) is the ML estimator.

Recall that \( \text{plim} \hat{\theta} = \theta_0 \Rightarrow \text{plim} g(\hat{\theta}) = g(\theta_0). \)

(Choice of parametrization is irrelevant in this regard.)

2. **Asymptotic variances of differentiable (+ a little more) functions of asymptotically normal estimators: The \( \delta \) - method**

Suppose \( \hat{\theta} \sim q(\theta_0, \sigma^2) \). Taylor expansion of \( g(\hat{\theta}) \) around \( \theta_0 \) gives

\[
g(\hat{\theta}) = g(\theta_0) + g'(\theta_0)(\hat{\theta} - \theta_0) + \text{more}.
\]

Remember that consistency implies \( \text{plim} g(\hat{\theta}) = g(\theta_0) \). Thus,

\[
V(g(\hat{\theta})) = \text{E}(g(\hat{\theta}) - g(\theta_0))^2 = (g'(\theta_0))^2V(\hat{\theta}) \text{ using the Taylor expansion above and ignoring "more". Hence } g(\hat{\theta}) \sim q(g(\theta_0), V(g(\hat{\theta}))) \text{ asymptotically.}
\]

This method of estimating standard errors is **very useful**. It is called the \( \delta \) - method. In the K-dimensional case with \( \hat{\theta} \sim q(\theta_0, \Sigma) \) we have \( g(\hat{\theta}) \sim q(g(\theta_0), (\partial g / \partial \theta_0)\Sigma(\partial g / \partial \theta_0)'), \) Note that \( \partial g / \partial \theta_0 \) means that the derivative is evaluated at the true value \( \theta_0. \)

Note: Do not use the ambiguous term "asymptotically unbiased" estimators.

*Why do we use ML estimators?*

Under our assumptions which provide lots of smoothness, ML estimators are asymptotically efficient - attaining (asymptotically) the Cramer-Rao lower bound on variance. *(What is the relation to the Gauss-Markov property?)*
Proposition: (Cramer-Rao bound for unbiased estimators) Let \( p \) be the likelihood function.

(Note the slight change of notation from the beginning of this lecture. There \( p \) was the density of one observation. Here we have in mind that \( p \) is the joint density of the data, regarded as a function of the parameter.) Suppose \( \theta^* \) is an unbiased estimator of \( \theta \). Then

\[
V(\theta^*) \geq \left[ V\left( \frac{d \ln p}{d \theta_0} \right) \right]^{-1} = i(\theta_0)^{-1}
\]

Proof: Note that

\[
E(\theta^*) = \theta_0 = \int \theta^* p \, dx
\]

from unbiasedness. Here the integral is with respect to all of the \( x \)'s. Note that \( \theta^* \) is a function of \( x \) but not \( \theta_0 \).

Differentiating above equality with respect to \( \theta_0 \), we get

\[
1 = \int \theta^* p' \, dx = \int \theta^* \left( \frac{p'}{p} \right) p \, dx
\]

\[
= E\left[ \theta^* \left( \frac{p'}{p} \right) \right] = E\left[ \theta^* \frac{d \ln p}{d \theta_0} \right] = \text{cov} \left[ \theta^*, \frac{d \ln p}{d \theta_0} \right] \quad (why?)
\]

The Cauchy-Schwartz inequality implies that \([\text{cov}(X,Y)]^2 \leq V(X) V(Y)\). Thus,

\[
\left[ \text{cov} \left[ \theta^*, \frac{d \ln p}{d \theta_0} \right] \right]^2 = 1 \leq V(\theta^*) \left( \frac{d \ln p}{d \theta_0} \right) \Rightarrow V(\theta^*) \geq \left[ V\left( \frac{d \ln p}{d \theta_0} \right) \right]^{-1}
\]

Note:

\[
E\left( \frac{d^2 \ln p}{d \theta^2} \right) = -E\left[ \left( \frac{d \ln p}{d \theta} \right)^2 \right] = -V\left( \frac{d \ln p}{d \theta} \right)
\]
Proof:

\[
\frac{d^2 \ln p}{d\theta^2} = \frac{d}{d\theta} \left( \frac{p'}{p} \right) = \frac{pp'' - (p')^2}{p^2} = \frac{p''}{p} \left( \frac{p'}{p} \right)
\]

\[\Rightarrow E \left( \frac{d^2 \ln p}{d\theta^2} \right) = -E \left[ \left( \frac{d \ln p}{d\theta} \right)^2 \right] \quad \text{(why?)}\]

Thus we have an expression for the variance of the first derivative of \( \ln p \) in terms of the second derivative - a property we have seen before.

**LINEAR MODEL:**

The assumption of "fixed in repeated samples" is rarely useful in economics. The basic assumption is that the distribution of \( X \) satisfies

\[
\text{plim} \left[ \frac{XX}{n} \right] = Q \quad \text{where } Q \text{ is positive definite,}
\]

and does not depend on parameters.

Our density of observables is \( p(y,x|\theta) \); usually, we assume that this is \( p(y|x,\theta) \cdot p(x) \). (Why is this restrictive?) Then the ML estimator depends on the conditional distribution. It is useful to go through the asymptotics applied to the linear model.

Recall that \( \hat{\beta} = \beta + \left( X'X \right)^{-1}X'\varepsilon = \beta + \left[ X'X/n \right]^{-1}X'\varepsilon/n \). If \( \text{plim} \left[ X'X/n \right] = Q \) and \( \text{plim} \left[ X'\varepsilon/n \right] = 0 \), then \( \text{plim} \hat{\beta} = \beta \) (i.e. \( \hat{\beta} \) is a consistent estimator of \( \beta \)). Recall that if also

\[
n^{1/2} \left[ X'\varepsilon / n \right] \overset{D}{\rightarrow} q(0,\sigma^2Q), \text{ then } n^{1/2} \left[ \hat{\beta} - \beta \right] \overset{D}{\rightarrow} q(0,\sigma^2Q^{-1}).
\]