11: Asymptotic Properties of Maximum Likelihood Estimators
ECON 837

Prof. Simon Woodcock, Spring 2014

We've discussed maximum likelihood estimation at several points already. It is a very popular estimation method. It assumes a parametric distribution for the population, which defines the joint density of the sample. This joint density, regarded as a function of unknown parameters for fixed data is called the **likelihood function**. A **maximum likelihood estimator** (MLE) is the set of parameter values at which the likelihood function achieves a maximum. The popularity of this approach is primarily because of the asymptotic properties of MLEs. These properties, discussed in detail below, are:

1. Consistency
2. Asymptotic normality
3. Invariance
4. Asymptotic efficiency.

We'll see the precise definition of each of these as we proceed.

Definition 1 *If* $\mathbf{x}_1, ..., \mathbf{x}_n$ *are a random sample of size* n *from a population with pdf* $p(\mathbf{x}|\boldsymbol{\theta})$, *the* ***likelihood function*** *for* $\boldsymbol{\theta}$ *given the data* $\mathbf{X}$ *is*

$$L(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}).$$

The ***log likelihood function*** *is almost always easier to work with. It is*

$$l(\boldsymbol{\theta}|\mathbf{X}) = \ln L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\boldsymbol{\theta}).$$

Definition 2 *A* ***maximum likelihood estimator*** *solves*

$$\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{X})$$

or equivalently,

$$\max_{\boldsymbol{\theta}} l\left(\boldsymbol{\theta}|\mathbf{X}\right).$$

Finding the MLE is just an optimization problem, something you should all be familiar with by now.

Consistency

Formally proving consistency of MLEs is not at all straightforward. We'll work through a heuristic proof below, that shows (intuitively) why MLEs are consistent. To begin, we'll just offer some justification for looking at the likelihood function at all. Proposition 3 shows that the expected value of the likelihood function is maximized at the true parameter values. This is not only an intuitive justification for the maximum likelihood approach, but also a useful intermediate result for the consistency proof that follows. Some notation: let $\boldsymbol{\theta}_0$ denote the true (unknown) value of the parameter vector we want to estimate, and let $\hat{\boldsymbol{\theta}}$ denote the MLE.

Proposition 3 *The expected value of the log likelihood is maximized at the true value of the parameters,* $\boldsymbol{\theta}_0$*. That is,*

$$E\left[l\left(\boldsymbol{\theta}_0|\mathbf{X}\right)\right] > E\left[l\left(\boldsymbol{\theta}|\mathbf{X}\right)\right] \quad \textit{for any } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \textit{ (including } \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}\textit{).}$$

Proof. We can restrict our attention to values of the likelihood function when parameters are in a neighborhood of $\boldsymbol{\theta}_0$. So let $\boldsymbol{\delta} \neq \mathbf{0}$ be an arbitrary vector, and consider the density in a neighborhood of $\boldsymbol{\theta}_0$, $p\left(\mathbf{x}_i|\boldsymbol{\theta}_0 + \boldsymbol{\delta}\right)$. We have

$$\begin{aligned} E\left[\frac{p\left(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta}\right)}{p\left(\mathbf{x}_i|\boldsymbol{\theta}_0\right)}\right] &= \int_{\mathbf{X}} \left(\frac{p\left(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta}\right)}{p\left(\mathbf{x}_i|\boldsymbol{\theta}_0\right)}\right) p\left(\mathbf{x}_i|\boldsymbol{\theta}_0\right) d\mathbf{x}_i \\ &= \int_{\mathbf{X}} p\left(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta}\right) d\mathbf{x}_i \\ &= 1 \end{aligned}$$

since $p(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta})$ is a density. Taking logs of both sides of this equation, we see that

$$\ln E\left[\frac{p(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta})}{p(\mathbf{x}_i|\boldsymbol{\theta}_0)}\right] = 0$$

and hence

$$E\left[\ln\frac{p(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta})}{p(\mathbf{x}_i|\boldsymbol{\theta}_0)}\right] < 0$$

by Jensen's inequality for concave functions. Rearranging,

$$E\left[\ln p(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta}) - \ln p(\mathbf{x}_i|\boldsymbol{\theta}_0)\right] < 0.$$

It follows immediately that

$$E\left[\sum_{i=1}^{n}\ln p(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta})\right] - E\left[\sum_{i=1}^{n}\ln p(\mathbf{x}_i|\boldsymbol{\theta}_0)\right] = E\left[l(\boldsymbol{\theta}_0+\boldsymbol{\delta}|\mathbf{X})\right] - E\left[l(\boldsymbol{\theta}_0|\mathbf{X})\right] < 0.$$

■

We'll use this intermediate result to sketch a proof of consistency of the MLE. A formal proof is quite detailed and not that instructive, but can easily be found in most advanced statistics textbooks. A formal proof also requires some regularity conditions that we'll state in a minute to prove asymptotic normality. First, a definition.

Definition 4 (Identification) *We say the parameter vector* $\boldsymbol{\theta}$ *is* ***identified*** *if for any* $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$ *we have* $L(\boldsymbol{\theta}^*|\mathbf{X}) \neq L(\boldsymbol{\theta}|\mathbf{X})$ *for some data* $\mathbf{X}$.

Proposition 5 (Consistency of MLE) *Under some regularity conditions, if the data* $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ *are iid and the parameter vector* $\boldsymbol{\theta}$ *is identified, then the MLE* $\hat{\boldsymbol{\theta}}$ *is consistent. That is, plim* $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

A Heuristic Proof. Since the data are iid, we know by Khinchine's WLLN that for any $\boldsymbol{\delta} \neq \mathbf{0}$ (and provided the expectation exists and is finite),

$$\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\ln p(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta}) - \ln p(\mathbf{x}_i|\boldsymbol{\theta}_0)\right)\right] = E\left[\ln p(\mathbf{x}_i|\boldsymbol{\theta}_0+\boldsymbol{\delta}) - \ln p(\mathbf{x}_i|\boldsymbol{\theta}_0)\right] < 0$$

where the inequality follows from Proposition 3. That is,

$$\text{plim}\left[\frac{1}{n}l\left(\boldsymbol{\theta}_0+\boldsymbol{\delta}|\mathbf{X}\right)\right] < \text{plim}\left[\frac{1}{n}l\left(\boldsymbol{\theta}_0|\mathbf{X}\right)\right] \tag{1}$$

so that asymptotically, the log-likelihood has a local maximum at $\boldsymbol{\theta}_0$ (in probability). Now consider the MLE $\hat{\boldsymbol{\theta}}_n$ when the sample size is n. Equation (1) tells us that

$$\text{plim}\left[\frac{1}{n}l\left(\hat{\boldsymbol{\theta}}_n|\mathbf{X}\right)\right] \leq \text{plim}\left[\frac{1}{n}l\left(\boldsymbol{\theta}_0|\mathbf{X}\right)\right] \tag{2}$$

where we have the weak inequality to account for the possibility that $\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0$. We also know that since $\hat{\boldsymbol{\theta}}_n$ maximizes $l\left(\boldsymbol{\theta}|\mathbf{X}\right)$,

$$\frac{1}{n}l\left(\hat{\boldsymbol{\theta}}_n|\mathbf{X}\right) \geq \frac{1}{n}l\left(\boldsymbol{\theta}_0|\mathbf{X}\right)$$

for every n, which implies

$$\text{plim}\frac{1}{n}l\left(\hat{\boldsymbol{\theta}}_n|\mathbf{X}\right) \geq \text{plim}\frac{1}{n}l\left(\boldsymbol{\theta}_0|\mathbf{X}\right). \tag{3}$$

(2) and (3) together imply

$$\text{plim}\frac{1}{n}l\left(\hat{\boldsymbol{\theta}}_n|\mathbf{X}\right) = \text{plim}\frac{1}{n}l\left(\boldsymbol{\theta}_0|\mathbf{X}\right).$$

That is, the log likelihood evaluated at the MLE shares the same probability limit as the log-likelihood evaluated at $\boldsymbol{\theta}_0$. When certain regularity conditions are satisfied (defined below) and $\boldsymbol{\theta}$ is identified, this also implies plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. ■

A formal proof of consistency actually proceeds a bit differently. It proceeds by showing that the FOCs for maximizing the log likelihood, evaluated at $\boldsymbol{\theta}_0$, converge in probability to zero. With sufficient regularity conditions, we can invert the FOCs everywhere, so that the MLEs exist and are unique. The next step is to show that the difference between the FOCs evaluated at the MLEs and the FOCs evaluated at $\boldsymbol{\theta}_0$ converges in probability to zero. When the FOCs are sufficiently smooth, this implies that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Asymptotic Normality

Let's lay out those regularity conditions. After a couple of intermediate results, we'll prove the asymptotic normality of MLEs.

R1 The first three derivatives of $\ln p(\mathbf{x}_i|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist in an interval that includes $\boldsymbol{\theta}_0$, are continuous, and finite for almost all $\mathbf{x}_i$ and all $\boldsymbol{\theta}$.

R2 Some expectations that tell us we can differentiate under certain integrals:

$$E\left[\frac{p'}{p}|\boldsymbol{\theta}\right] = \mathbf{0}, \quad E\left[\frac{p''}{p}|\boldsymbol{\theta}\right] = \mathbf{0}, \quad E\left[\frac{(p')^2}{p}|\boldsymbol{\theta}\right] > \mathbf{0}$$

where $p = p(\mathbf{x}_i|\boldsymbol{\theta})$, $p' = dp(\mathbf{x}_i|\boldsymbol{\theta})/d\boldsymbol{\theta}$, and $p'' = d^2p(\mathbf{x}_i|\boldsymbol{\theta})/d\boldsymbol{\theta}d\boldsymbol{\theta}'$. Note that in almost all situations these conditions will be satisfied, since for example we can get the first expectation

$$1 = \int p(\mathbf{x}_i|\boldsymbol{\theta})\, d\mathbf{x} \Rightarrow \mathbf{0} = \int p' d\mathbf{x} = \int \frac{p'}{p} p d\mathbf{x} = E\left[\frac{p'}{p}|\boldsymbol{\theta}\right] \tag{4}$$

whenever p is sufficiently well-behaved that we can differentiate under the integral. We get the second expectation by differentiating again. The third one is clearly the expectation of something positive.

R3 A purely technical assumption that will control the expected error in some Taylor series expansions. It just says that the third derivative of the log likelihood is bounded by an integrable function.

$$\frac{\partial^3 \ln p(\mathbf{x}_i|\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k \partial\theta_l} < M(\mathbf{x}), \text{ where } E[M(\mathbf{x})] < K < \infty.$$

A **very** useful equality when doing maximum likelihood estimation is the **information matrix** equality. As we'll see, this provides an estimate of the asymptotic variance of the MLEs. The equality we'll establish gives us a couple of different ways to do this. Return to

equation (4) above, which was just a derivation of one of the expectations in R2. Given that

$$p' = \frac{p'}{p}p = \frac{d\ln p}{d\boldsymbol{\theta}}p$$

we have

$$\mathbf{0} = \int p' d\mathbf{x}_i = \int \frac{p'}{p} p d\mathbf{x}_i = \int \frac{d\ln p}{d\boldsymbol{\theta}} p d\mathbf{x}_i. \tag{5}$$

Differentiating again gives

$$\begin{aligned} \mathbf{0} &= \int \frac{d^2 \ln p}{d\boldsymbol{\theta} d\boldsymbol{\theta}'} p d\mathbf{x}_i + \int \frac{d\ln p}{d\boldsymbol{\theta}} \frac{dp}{d\boldsymbol{\theta}'} d\mathbf{x}_i \\ &= \int \frac{d^2 \ln p}{d\boldsymbol{\theta} d\boldsymbol{\theta}'} p d\mathbf{x}_i + \int \frac{d\ln p}{d\boldsymbol{\theta}} \frac{d\ln p}{d\boldsymbol{\theta}'} p d\mathbf{x}_i \\ &= E\left[\frac{d^2 \ln p}{d\boldsymbol{\theta} d\boldsymbol{\theta}'}\right] + E\left[\frac{d\ln p}{d\boldsymbol{\theta}} \frac{d\ln p}{d\boldsymbol{\theta}'}\right] \end{aligned} \tag{6}$$

Definition 6 (Information Matrix Equality) *The* ***information matrix*** *is* $\mathbf{i}(\boldsymbol{\theta}_0)$, *where*

$$\frac{1}{n}\mathbf{i}(\boldsymbol{\theta}_0) = -E\left[\left.\frac{d^2 \ln p(\mathbf{x}_i|\boldsymbol{\theta})}{d\boldsymbol{\theta} d\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = E\left[\left(\left.\frac{d\ln p(\mathbf{x}_i|\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)\left(\left.\frac{d\ln p(\mathbf{x}_i|\boldsymbol{\theta})}{d\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)\right]$$

or equivalently,

$$\mathbf{i}(\boldsymbol{\theta}_0) = -E\left[\left.\frac{d^2 l(\boldsymbol{\theta}|\mathbf{X})}{d\boldsymbol{\theta} d\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = E\left[\left(\left.\frac{dl(\boldsymbol{\theta}|\mathbf{X})}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)\left(\left.\frac{dl(\boldsymbol{\theta}|\mathbf{X})}{d\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)\right].$$

The following proposition is of fundamental importance. It gives us the asymptotic distribution of **any** maximum likelihood estimator. In many cases we are unable to derive an exact finite sample distribution for the MLEs, and the asymptotic distribution is all we have for inference.

Proposition 7 (Asymptotic Normality) $\hat{\boldsymbol{\theta}} \overset{a}{\sim} N\left(\boldsymbol{\theta}_0, \left[\mathbf{i}(\boldsymbol{\theta}_0)\right]^{-1}\right)$. *That is,*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \left[\frac{1}{n}\mathbf{i}(\boldsymbol{\theta}_0)\right]^{-1}\right).$$

Proof. Let

$$\begin{aligned}
\mathbf{g}\left(\hat{\boldsymbol{\theta}}\right) &= \left.\frac{dl\left(\boldsymbol{\theta}|\mathbf{X}\right)}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \sum_{i=1}^{n}\left.\frac{d\ln p\left(\mathbf{x}_i|\boldsymbol{\theta}\right)}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \equiv \sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right) \\
\mathbf{H}\left(\hat{\boldsymbol{\theta}}\right) &= \left.\frac{d^2 l\left(\boldsymbol{\theta}|\mathbf{X}\right)}{d\boldsymbol{\theta}d\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \sum_{i=1}^{n}\left.\frac{d^2\ln p\left(\mathbf{x}_i|\boldsymbol{\theta}\right)}{d\boldsymbol{\theta}d\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \equiv \sum_{i=1}^{n}\mathbf{H}_i\left(\hat{\boldsymbol{\theta}}\right)
\end{aligned}$$

denote the gradient and Hessian of the log likelihood. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ solves the first order condition $\mathbf{g}\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{0}$. A first order Taylor expansion of $\mathbf{g}\left(\hat{\boldsymbol{\theta}}\right)$ around $\boldsymbol{\theta}_0$ gives

$$\mathbf{g}\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{g}\left(\boldsymbol{\theta}_0\right) + \mathbf{H}\left(\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + \text{remainder} = 0.$$

We could use this Taylor expansion to derive asymptotic normality of $\hat{\boldsymbol{\theta}}$, but we would need to worry about the rate at which the remainder converges to zero (in probability). Alternately, we can use a little trick that relies on the Mean Value Theorem (MVT). [Recall the MVT for real-valued functions: if f is defined and continuous on $[a,b]$ and differentiable on (a,b), then there exists at least one $c \in (a,b)$ such that $f'(c) = (f(b) - f(a))/(b-a)$]. By the MVT, there exists $\bar{\boldsymbol{\theta}} = \lambda\hat{\boldsymbol{\theta}} + (1-\lambda)\boldsymbol{\theta}_0$ for $\lambda \in (0,1)$ such that

$$\mathbf{H}\left(\bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \mathbf{g}\left(\hat{\boldsymbol{\theta}}\right) - \mathbf{g}\left(\boldsymbol{\theta}_0\right)$$

and so we'll work with

$$\mathbf{g}\left(\boldsymbol{\theta}_0\right) + \mathbf{H}\left(\bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \mathbf{0}$$

instead. Rearranging, we get

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \left[-\mathbf{H}\left(\bar{\boldsymbol{\theta}}\right)\right]^{-1}\sqrt{n}\mathbf{g}\left(\boldsymbol{\theta}_0\right)$$

or

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \left[-\frac{1}{n}\mathbf{H}\left(\bar{\boldsymbol{\theta}}\right)\right]^{-1}\left(\sqrt{n}\frac{1}{n}\mathbf{g}\left(\boldsymbol{\theta}_0\right)\right).$$

We know plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, and hence plim $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$ also. Since the Hessian is continuous (R1), it follows that plim $\mathbf{H}\left(\bar{\boldsymbol{\theta}}\right) =$ plim $\mathbf{H}\left(\boldsymbol{\theta}_0\right)$, and therefore if the limiting distribution exists,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \left[-\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{g}\left(\boldsymbol{\theta}_0\right)\right).$$

Since

$$\frac{1}{\sqrt{n}}\mathbf{g}\left(\boldsymbol{\theta}_0\right) = \sqrt{n}\frac{1}{n}\mathbf{g}\left(\boldsymbol{\theta}_0\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\boldsymbol{\theta}_0\right)\right)$$

is just $\sqrt{n}$ times a sample mean, we can apply a Central Limit Theorem to find its asymptotic distribution. We know $E\left[\mathbf{g}_i\left(\boldsymbol{\theta}_0\right)\right] = \mathbf{0}$ (have a look at eq (5) above) and

$$Var\left[\mathbf{g}_i\left(\boldsymbol{\theta}_0\right)\right] = E\left[\mathbf{g}_i\left(\boldsymbol{\theta}_0\right)\mathbf{g}_i\left(\boldsymbol{\theta}_0\right)'\right] = -E\left[\mathbf{H}_i\left(\boldsymbol{\theta}_0\right)\right] = \frac{1}{n}\mathbf{i}\left(\boldsymbol{\theta}_0\right)$$

by the information matrix equality. Therefore,

$$\frac{1}{\sqrt{n}}\mathbf{g}\left(\boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \frac{1}{n}\mathbf{i}\left(\boldsymbol{\theta}_0\right)\right).$$

Since $-\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)$ is yet another sample mean, we can apply a WLLN to show

$$\text{plim}\left[-\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right] = -E\left[\mathbf{H}_i\left(\boldsymbol{\theta}_0\right)\right] = \frac{1}{n}\mathbf{i}\left(\boldsymbol{\theta}_0\right).$$

Putting all this together, we have

$$\left[-\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{g}\left(\boldsymbol{\theta}_0\right)\right) \xrightarrow{d} N\left(\mathbf{0}, \left[\frac{1}{n}\mathbf{i}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\frac{1}{n}\mathbf{i}\left(\boldsymbol{\theta}_0\right)\left[\frac{1}{n}\mathbf{i}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\right)$$

since if $\mathbf{z} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$ we know $\mathbf{Az} \sim N\left(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\right)$. Therefore

$$\left[-\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\left(\sqrt{n}\frac{1}{n}\mathbf{g}\left(\boldsymbol{\theta}_0\right)\right) \xrightarrow{d} N\left(\mathbf{0}, \left[\frac{1}{n}\mathbf{i}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\right)$$

and hence

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \left[\frac{1}{n}\mathbf{i}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\right)$$

also. More simply,

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N\left(\boldsymbol{\theta}_0, \left[\mathbf{i}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\right).$$

■

The basic result here is that we can approximate the sampling distribution of $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ by a normal distribution with mean $\mathbf{0}$ and variance $\left[\frac{1}{n}\mathbf{i}\left(\boldsymbol{\theta}_0\right)\right]^{-1}$. Of course we need to estimate this variance to do any sort of inference, but a consistent estimate is provided by $\left[\frac{1}{n}\mathbf{i}\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1}$ (why?). How accurate is this approximation of the sampling distribution? Because it is based on a first-order Taylor expansion of the gradient vector around $\boldsymbol{\theta}_0$, i.e., a (local) linear approximation, a normal approximation to the sampling distribution of $\hat{\boldsymbol{\theta}}$ is about as accurate as a (local) linear approximation to a function.

Notice that we normalized the quantity $\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ by $\sqrt{n}$. This stabilizing transformation "blows up" $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ so that it has a well-defined limiting distribution, which $\hat{\boldsymbol{\theta}}$ itself does not. [Because plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, it follows that $\hat{\boldsymbol{\theta}} \xrightarrow{d} \boldsymbol{\theta}_0$. That is, the limiting distribution of $\hat{\boldsymbol{\theta}}$ assigns unit mass to $\boldsymbol{\theta}_0$.]

Some things for you to verify

1. The asymptotic distribution defined above is the **exact** distribution of the MLE of the mean of a normally distributed random variable.
2. The asymptotic distribution of the MLE of $\boldsymbol{\beta}$ in the linear regression model under normality equals its exact distribution.

Invariance

The invariance property tells us that continuous functions of the MLEs are consistent estimators of continuous functions of the unknown parameters. This is true for any consistent estimator (Slutsky's theorem).

Proposition 8 (Invariance) *Let $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimator of $\boldsymbol{\theta}$, whose true value is $\boldsymbol{\theta}_0$. Then for any continuous function h we have*

$$\textit{plim}\ h\left(\hat{\boldsymbol{\theta}}\right) = h\left(\boldsymbol{\theta}_0\right).$$

Proof. By Slutsky's Theorem, plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$ implies plim $h\left(\hat{\boldsymbol{\theta}}\right) = h\left(\boldsymbol{\theta}_0\right)$. ■

The delta method

We can actually derive the complete asymptotic distribution of any continuous and differentiable function of MLEs (or any other consistent and asymptotically normal estimator). This result is known as the **delta method**. It is **very useful**, since we can use it to derive the asymptotic variance (or standard error) of even very complicated nonlinear functions of estimated parameters. For simplicity of exposition, suppose we have a single parameter $\hat{\theta} \overset{a}{\sim} N\left(\theta_0, \sigma^2/n\right)$, so that $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \sigma^2\right)$. Let h be a continuous and differentiable function of θ. We're interested in the asymptotic distribution of $h\left(\hat{\theta}\right)$. Consider the first order Taylor expansion of $h\left(\hat{\theta}\right)$ around θ_0 :

$$h\left(\hat{\theta}\right) = h\left(\theta_0\right) + h'\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right) + \text{remainder}.$$

Ignoring the remainder term (we can appeal to the MVT like we did above), we know that:

$$\sqrt{n}\left[h\left(\hat{\theta}\right) - h\left(\theta_0\right)\right] \xrightarrow{d} \sqrt{n}h'\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right)$$

Since $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \sigma^2\right)$, we further know

$$\sqrt{n}\left[h\left(\hat{\theta}\right) - h\left(\theta_0\right)\right] \xrightarrow{d} \sqrt{n}h'\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, h'\left(\theta_0\right)^2\sigma^2\right)$$

because $h'\left(\theta_0\right)$ is a constant. Therefore,

$$h\left(\hat{\theta}\right) \overset{a}{\sim} N\left(h\left(\theta_0\right), h'\left(\theta_0\right)^2\sigma^2/n\right). \tag{7}$$

Replacing unknown parameters with consistent estimates (e.g., MLEs), expression (7) gives us a way to approximate the sampling distribution of $h\left(\hat{\theta}\right)$ in large samples.

In the k-dimensional case with $\hat{\boldsymbol{\theta}} \overset{a}{\sim} N\left(\boldsymbol{\theta}_0, \frac{1}{n}\boldsymbol{\Sigma}\right)$ we have

$$\mathbf{h}\left(\hat{\boldsymbol{\theta}}\right) \overset{a}{\sim} N\left(\mathbf{h}\left(\boldsymbol{\theta}_0\right), \frac{1}{n}\frac{\partial \mathbf{h}\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}_0'}\boldsymbol{\Sigma}\frac{\partial \mathbf{h}\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}_0}\right)$$

where $\partial \mathbf{h}/\partial \boldsymbol{\theta}_0$ means the derivative is evaluated at the true value $\boldsymbol{\theta}_0$.

Asymptotic Efficiency

Beyond the properties of MLEs that we have discussed so far, perhaps the most important justification for maximum likelihood estimation is that MLEs are asymptotically efficient. That is, they attain (asymptotically) the Cramer-Rao lower bound on variance. (*how does this relate to the GMT?*)

Proposition 9 (Cramer-Rao Lower Bound for Unbiased Estimators) *Let L be the likelihood function. Suppose that $\boldsymbol{\theta}^*$ is an unbiased estimator of $\boldsymbol{\theta}$. Then*

$$Var\left[\boldsymbol{\theta}^*\right] \geq \left[Var\left[\frac{\partial \ln L}{\partial \boldsymbol{\theta}_0}\right]\right]^{-1} = \left[\mathbf{i}\left(\boldsymbol{\theta}_0\right)\right]^{-1}.$$

Proof. We'll do the proof for the one-dimensional case. Note first that L is just the joint density of the data (p) regarded as a function of parameters. Therefore, unbiasedness implies

$$E\left[\theta^*\right] = \int \theta^* p d\mathbf{x} = \theta_0.$$

Note that θ^* is a function of $\mathbf{x}$, but not θ_0. Differentiating this equality with respect to θ_0, and letting $p' = dp/d\theta_0$ we get

$$\begin{aligned}
1 &= \int \theta^* p' d\mathbf{x} = \int \theta^* \left(\frac{p'}{p}\right) p d\mathbf{x} \\
&= E\left[\theta^* \left(\frac{p'}{p}\right)\right] = E\left[\theta^* \frac{d\ln p}{d\theta_0}\right] = E\left[\theta^* \frac{d\ln L}{d\theta_0}\right] && (p = L) \\
&= Cov\left[\theta^*, \frac{d\ln L}{d\theta_0}\right]. && \left(E\left[\tfrac{d\ln L}{d\theta_0}\right] = 0\right)
\end{aligned}$$

The Cauchy-Schwartz inequality implies $\left[Cov\left[X,Y\right]\right]^2 \leq Var\left[X\right] Var\left[Y\right]$, and hence

$$\left[Cov\left[\theta^*, \frac{d\ln L}{d\theta_0}\right]\right]^2 = 1 \leq Var\left[\theta^*\right] Var\left[\frac{d\ln L}{d\theta_0}\right]$$

which implies

$$Var\left[\theta^*\right] \geq \left[Var\left[\frac{d\ln L}{d\theta_0}\right]\right]^{-1} = \left[\mathbf{i}\left(\boldsymbol{\theta}_0\right)\right]^{-1}.$$

■

So the Cramer-Rao lower bound on the variance of an unbiased estimator is just the inverse of the information matrix. We've seen already that this is the variance of MLE; hence maximum likelihood estimators achieve the Cramer-Rao lower bound. It is also possible to show that if an unbiased estimator θ^* achieves the Cramer-Rao lower bound, then θ^* is the MLE.