

Lecture Note 8

Maximum Likelihood Estimation—Part II

Uniqueness (continued):

If we can interchange differentiation and integration (i.e., the integration w.r.t. E_0), then θ_0 satisfies

$$E_0 \left[\frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta} \right] = 0,$$

and to be a unique maximizer, the solution has to be unique.

In words, this equation states that the expected score, when evaluated at the population parameter θ_0 , is zero.

While we will not discuss it here in greater details, the condition that sufficient for the SLLN discuss before is also sufficient for the interchange of the expectation and differentiation.

Example: Logit Model (continued):

$$\begin{aligned} E_0 \left[\frac{\partial \ln f(y_i, x_i; \theta)}{\partial \theta} \right] &= E_0 \left[\left(y_i \frac{1}{1 + e^{x_i' \theta}} - (1 - y) \frac{e^{x_i' \theta}}{1 + e^{x_i' \theta}} \right) x_i \right], \\ &= E_0 \left[\left(y_i - \frac{e^{x_i' \theta}}{1 + e^{x_i' \theta}} \right) x_i \right], \end{aligned}$$

and clearly

$$E_0 \left[\left(y_i - \frac{e^{x_i' \theta_0}}{1 + e^{x_i' \theta_0}} \right) x_i \right] = 0.$$

Weak consistency of MLE:

Theorem:

If for some parameter space Θ with $\theta_0 \in \Theta$ the normalized log likelihood converges uniformly to the expected log likelihood and the population parameter maximizes the expected log-likelihood, then

$$\hat{\theta}_n \xrightarrow{p} \theta_0.$$

Proof:

Denote the normalized log likelihood by $M_n(\theta)$, and the expected log-likelihood (with respect to the population distribution) as $M(\theta)$. Then

$$M_n(\hat{\theta}_n) \geq M_n(\theta_0).$$

Since $M_n(\theta_0) \xrightarrow{p} M(\theta_0)$, there exists a sequence on non-negative random variables $\{Z_n\}$, with $Z_n \xrightarrow{p} 0$, and

$$|M_n(\theta_0) - M(\theta_0)| \leq Z_n.$$

Hence, we have

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &\leq M_n(\theta_0) - M(\hat{\theta}_n) + Z_n \\ &\leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + Z_n, \end{aligned} \tag{8.1}$$

and the right-hand side of (8.1) converges to 0 in probability.

Recall that we assume that (Assumption Uniqueness) for all $\varepsilon > 0$,

$$\sup_{|\theta - \theta_0| \geq \varepsilon} E_0 [\ln f(y, x; \theta)] \leq E_0 [\ln f(y, x; \theta_0)]. \tag{8.2}$$

So, we have that for every $\varepsilon > 0$, there exist $\eta > 0$, such that for all $|\theta - \theta_0| \geq \varepsilon$,

$$M(\theta) < M(\theta_0) - \eta.$$

Hence,

$$\Pr(|\theta_n - \theta_0| \geq \varepsilon) \leq \Pr\left(M(\hat{\theta}_n) < M(\theta_0) - \eta\right),$$

and the probability on the right-hand side of the equation converges to 0.

Q.E.D.

Remark: This theorem gives sufficient conditions for weak convergence. There are alternative sufficient conditions that are discussed in the literature (e.g. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998).

Asymptotic Normality of MLE:

Because $\hat{\theta}_n \xrightarrow{p} \theta_0$, the distribution of the MLE $\hat{\theta}_n$ becomes concentrated around the true population parameter θ_0 , at least when the sample is large, and $\hat{\theta}_n - \theta_0$ has a degenerate distribution at the point 0 when $n \rightarrow \infty$. Hence, we consider the quantity

$$\sqrt{n} \left(\hat{\theta}_n - \theta_0 \right),$$

and we hope that the distribution does not become degenerate, even when $n \rightarrow \infty$.

Assumption Continuity:

We assume that the functions $f(y, x; \theta)$, $\partial \ln f(y, x; \theta) / \partial \theta$, and $\partial^2 \ln f(y, x; \theta) / \partial \theta \partial \theta'$, are continuous function of θ . (This make this assumption to facilitate the proof, but one can actually relax this assumption.)

Consider the following Taylor-series expansion around the population parameter θ_0 :

$$0 = \frac{1}{n} \frac{\partial \ln L_n(\theta_0)}{\partial \theta} + \frac{1}{n} \frac{\partial^2 \ln L_n(\hat{\theta}_n^*)}{\partial \theta \partial \theta'} (\hat{\theta}_n - \theta_0), \quad (8.3)$$

where $\hat{\theta}_n^*$ is on the line segment connecting $\hat{\theta}_n$ and θ_0 .

Since $\hat{\theta}_n \xrightarrow{p} \theta_0$ it follows also that

$$\hat{\theta}_n^* \xrightarrow{p} \theta_0.$$

Multiplying both side of (8.3) by \sqrt{n} gives then

$$\frac{1}{n} \frac{\partial^2 \ln L_n(\hat{\theta}_n^*)}{\partial \theta \partial \theta'} \sqrt{n} (\hat{\theta}_n - \theta_0) = -\frac{1}{\sqrt{n}} \frac{\partial \ln L_n(\theta_0)}{\partial \theta}. \quad (8.4)$$

Consider first the right-hand side of (8.4),

$$\frac{1}{\sqrt{n}} \frac{\partial \ln L_n(\theta_0)}{\partial \theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta}.$$

Note that

$$\frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta}, \quad i = 1, 2, 3, \dots$$

is a sequence of random variable for which

$$\begin{aligned} E_0 \left[\frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta} \right] &= 0, \quad \text{and} \\ V_0 \left(\frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta} \right) &= E_0 \left[\frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta} \frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta'} \right] \\ &= I(\theta_0), \end{aligned}$$

where the matrix $I(\theta_0)$ is called the *Fisher information matrix*.

Note now that

$$\frac{\partial^2 \ln f(y_i, x_i; \theta)}{\partial \theta \partial \theta'} = \frac{1}{f(y_i, x_i; \theta)} \frac{\partial^2 f(y_i, x_i; \theta)}{\partial \theta \partial \theta'} - \frac{\partial \ln f(y_i, x_i; \theta)}{\partial \theta} \frac{\partial \ln f(y_i, x_i; \theta)}{\partial \theta'}. \quad (8.5)$$

If we assume that

$$E_0 \left[\sup_{\theta \in \Theta} \left| \frac{\partial^2 f(y_i, x_i; \theta)}{\partial \theta \partial \theta'} \right| \right] < \infty,$$

so that the expectation of (8.5) exists, and it is given by

$$E_0 \left[\frac{\partial^2 f(y_i, x_i; \theta)}{\partial \theta \partial \theta'} \right] = - E_0 \left[\frac{1}{f(y_i, x_i; \theta)} \frac{\partial^2 f(y_i, x_i; \theta)}{\partial \theta \partial \theta'} \right] \\ + E_0 \left[\frac{\partial \ln f(y_i, x_i; \theta)}{\partial \theta} \frac{\partial \ln f(y_i, x_i; \theta)}{\partial \theta'} \right].$$

At $\theta = \theta_0$ we have

$$E_0 \left[\frac{1}{f(y_i, x_i; \theta_0)} \frac{\partial^2 f(y_i, x_i; \theta_0)}{\partial \theta \partial \theta'} \right] = \int_{y_i} \int_{x_i} \frac{1}{f(y_i, x_i; \theta_0)} \frac{\partial^2 f(y_i, x_i; \theta_0)}{\partial \theta \partial \theta'} f(y_i, x_i; \theta_0) dx_i dy_i, \\ = \int_{y_i} \int_{x_i} \frac{\partial^2 f(y_i, x_i; \theta_0)}{\partial \theta \partial \theta'} dx_i dy_i, \\ = \frac{\partial^2}{\partial \theta \partial \theta'} \left[\int_{y_i} \int_{x_i} f(y_i, x_i; \theta_0) dx_i dy_i \right],$$

where the last equality holds if

$$E_0 \left[\sup_{\theta \in \Theta} \left| \frac{\partial f(y_i, x_i; \theta)}{\partial \theta} \right| \right] < \infty,$$

so that we can interchange the integration with the differentiation.

Note that since

$$\int_{y_i} \int_{x_i} f(y_i, x_i; \theta_0) dx_i dy_i = 1,$$

it follows that

$$E_0 \left[\frac{1}{f(y_i, x_i; \theta_0)} \frac{\partial^2 f(y_i, x_i; \theta_0)}{\partial \theta \partial \theta'} \right] = 0.$$

Hence we obtained that

$$E_0 \left[- \frac{\partial^2 \ln f(y_i, x_i; \theta_0)}{\partial \theta \partial \theta'} \right] = E_0 \left[\frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta} \frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta'} \right], \quad (8.6) \\ = I(\theta_0).$$

The equality in (8.6) is call the *information matrix equality*. That is, the information matrix $I(\theta_0)$ has two representations:

1. The *Hessian* representation:

$$I(\theta_0) = E_0 \left[- \frac{\partial^2 \ln f(y_i, x_i; \theta_0)}{\partial \theta \partial \theta'} \right].$$

2. The *Outer Product Gradient (OPG)* representation:

$$I(\theta_0) = E_0 \left[\frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta} \frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta'} \right].$$

The information matrix equality has two direct implications:

(i) The information matrix is finite, that is

$$I(\theta_0) < \infty,$$

(ii) The information matrix is positive definite because the integrand in the OPG is positive definite.

Hence, $I^{-1}(\theta_0)$ exists.

Now, by the CLT we have that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta} \xrightarrow{D} N(0, I(\theta_0)). \quad (8.7)$$

Also, because $\hat{\theta}_n^* \xrightarrow{p} \theta_0$, we have that

$$\frac{1}{n} \frac{\partial^2 \ln L_n(\hat{\theta}_n^*)}{\partial \theta \partial \theta'} \xrightarrow{p} E_0 \left[\frac{\partial^2 \ln f(y_i, x_i; \theta_0)}{\partial \theta \partial \theta'} \right]. \quad (8.8)$$

(Recall that we already have that

$$\frac{1}{n} \frac{\partial^2 \ln L_n(\theta)}{\partial \theta \partial \theta'} \xrightarrow{p} E_0 \left[\frac{\partial^2 \ln f(y_i, x_i; \theta)}{\partial \theta \partial \theta'} \right],$$

uniformly on Θ .)

Hence we have from (8.4), (8.7), (8.8), and Slutsky's theorem that

$$\begin{aligned} \sqrt{n} (\hat{\theta}_n - \theta_0) &\xrightarrow{D} I^{-1}(\theta_0) \frac{1}{\sqrt{n}} \frac{\partial \ln L_n(\theta_0)}{\partial \theta} \\ &\xrightarrow{D} N(0, I^{-1}(\theta_0)), \end{aligned} \quad (8.9)$$

where we have used the information matrix equality to simplify the final result.

Conclusion: The asymptotic distribution of the MLE $\hat{\theta}_n$ is normal!

Recall now that in order to be able to do inference we need a consistent estimator for $I(\theta_0)$.

Here we have two such estimates based on the two presentation for $I(\theta_0)$:

1. The Hessian estimator:

$$\begin{aligned} -\frac{1}{n} \frac{\partial^2 \ln L_n(\theta_0)}{\partial \theta \partial \theta'} &= -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \ln f(y_i, x_i; \theta_0)}{\partial \theta \partial \theta'} \\ &\xrightarrow{p} I(\theta_0). \end{aligned}$$

2. The OPG representation:

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta} \frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta'} \xrightarrow{p} I(\theta_0).$$

Note that the statement

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0)),$$

implies that the sample variance for $\hat{\theta}_n$ is given by

$$V(\hat{\theta}_n) = \frac{1}{n} I^{-1}(\theta_0),$$

and hence an estimator for the sample variance is

$$\hat{V}(\hat{\theta}_n) = \left(-\frac{\partial^2 \ln L_n(\theta_0)}{\partial \theta \partial \theta'} \right)^{-1},$$

or alternatively

$$\hat{V}(\hat{\theta}_n) = \left(\sum_{i=1}^n \frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta} \frac{\partial \ln f(y_i, x_i; \theta_0)}{\partial \theta'} \right)^{-1}.$$

Both estimates are inverse of non-normalized sums that diverge to ∞ . Hence, the estimates of the sample variance converge to 0.

Example: Logit Model (continued):

The information matrix for this particular application is obtained as follows. Note first that

$$\begin{aligned} \ln f(y_i|x_i; \theta) &= y_i \ln \frac{e^{x_i' \theta}}{1 + e^{x_i' \theta}} - (1 - y) \ln \frac{1}{1 + e^{x_i' \theta}}, \\ \frac{\partial \ln f(y_i|x_i; \theta)}{\partial \theta} &= \left(y_i - \frac{e^{x_i' \theta}}{1 + e^{x_i' \theta}} \right) x_i, \quad \text{and} \\ \frac{\partial^2 \ln f(y_i|x_i; \theta)}{\partial \theta \partial \theta'} &= -\frac{e^{x_i' \theta}}{(1 + e^{x_i' \theta})^2} x_i x_i'. \end{aligned}$$

Hence,

$$\begin{aligned} I(\theta) &= E_0 \left[\frac{\partial \ln f(y_i, x_i; \theta)}{\partial \theta} \frac{\partial \ln f(y_i, x_i; \theta)}{\partial \theta'} \right] \\ &= E_0 \left[\left(y_i - \frac{e^{x_i' \theta}}{1 + e^{x_i' \theta}} \right)^2 x_i x_i' \right] \\ &= E_0 \left[\frac{e^{x_i' \theta_0}}{(1 + e^{x_i' \theta_0})^2} x_i x_i' \right] + E_0 \left[\left(\frac{e^{x_i' \theta_0}}{1 + e^{x_i' \theta_0}} - \frac{e^{x_i' \theta}}{1 + e^{x_i' \theta}} \right)^2 x_i x_i' \right]. \end{aligned}$$

It is easy to check that for $\theta = \theta_0$ we have

$$I(\theta_0) = E_0 \left[-\frac{\partial^2 \ln f(y_i, x_i; \theta_0)}{\partial \theta \partial \theta'} \right].$$

Consistent estimators for $I(\theta_0)$ are provided by:

1. The OPG estimator:

$$\hat{I}_n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \frac{e^{x_i' \hat{\theta}_n}}{1 + e^{x_i' \hat{\theta}_n}} \right)^2 x_i x_i'.$$

2. The Hessian estimator:

$$\hat{I}_n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n \frac{e^{x_i' \hat{\theta}_n}}{(1 + e^{x_i' \hat{\theta}_n})^2} x_i x_i'.$$

The inverse of the two estimates above give the estimate of the normal asymptotic covariance matrix for $\hat{\theta}_n$.