

Lecture Note 9

Maximum Likelihood Estimation—Part III

MLE in misspecified model:

The observed joint distribution of (y, x) is determined by

1. **Model:** The conditional pdf of y , conditional on x , $f(y|x; \theta)$; and
2. **Survey design:** Marginal density of x , $g(x)$.

Example:

y = Car ownership (1 if own a car, and 0 otherwise).

x = Household income.

There are cases in which the survey design is such that it may overrepresent (or under-represent) some income classes.

The model and survey design imply that the joint pdf of (y, x) is

$$f(y, x; \theta) = f(y|x; \theta)g(x). \quad (8.1)$$

The population joint pdf of (y, x) is

$$f(y, x; \theta) = f(y|x; \theta)h(x). \quad (8.2)$$

In the derivation of the last two lecture notes it was assumed that $h(\cdot) = g(\cdot)$. Hence, for $\theta = \theta_0$ the “model” joint pdf and the population joint pdf for (y, x) coincide. The MLE is then a consistent estimator for the population parameter vector θ_0 .

Now we will consider the case in which

$$f(y|x; \theta) = h(y|x; \theta),$$

but

$$g(x) \neq h(x),$$

because, for example, of bias sampling of x .

Recall that we defined θ_0 as:

$$\theta_0 = \arg \max_{\theta \in \Theta} E_0 [\ln f(y, x; \theta)],$$

with E_0 denote the expectation w.r.t. the population joint pdf of (y, x) and the “model” joint pdf of (y, x) , i.e., $f(y, x; \theta)$.

If $g(x)$ does not depend on θ , then we can replace $g(\cdot)$ by $h(\cdot)$ without changing the maximand.

Conclusion: *The MLE is consistent estimator for the population parameter vector θ_0 if*

$$f(y|x; \theta_0) = h(y|x; \theta_0),$$

that is, for $\theta = \theta_0$ the model (the conditional pdf of y , conditional on x) must be equal to the population conditional pdf.

Now consider a different case in which $f(y|x; \theta_0) \neq h(y|x; \theta_0)$. That is,

The model: $f(y|x; \theta)$.

The population: $h(y|x)$.

As before, the MLE is given by

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ln f(y_i, x_i; \theta).$$

Under the same conditions as in the previous two lecture notes

$$\frac{1}{n} \sum_{i=1}^n \ln f(y_i, x_i; \theta) \xrightarrow{p} E_0 [\ln f(y_i, x_i; \theta)],$$

uniformly for $\theta \in \Theta$, where the expectation is taken with respect to the *population joint density*, that is

$$h(y, x) = h(y|x)h(x).$$

Define the *quasi-true value*, say θ_q , by

$$\theta_q = \arg \max_{\theta \in \Theta} E_0 [\ln f(y|x; \theta)],$$

and assume that it is unique. Then, using the same proof as before we can show that

$$\hat{\theta}_n \longrightarrow \theta_q.$$

Consider now

$$E_0 [\ln h(y, x; \theta) - \ln f(y, x; \theta)] = \int_y \int_x [\ln h(y, x; \theta) - \ln f(y, x; \theta)] dx dy. \quad (8.3)$$

Note that the expectation in (8.3) is a weighted difference of log-densities. It can be interpreted therefore as a measure of the distance between the densities. This equation is, in fact, the definition of the *Kullback-Leibler* distance (information criterion) between the densities $h(y, x; \theta)$ and $f(y, x; \theta)$. We denote this measure by $K(h, f)$.

Note that: (i) $K(h, f) \geq 0$; and (ii) if $h = f$, then $K(h, f) = 0$.

Hence, the quasi-true value defined by θ_q minimizes the Kullback-Leibler distance to the population joint pdf.

Conclusion: The MLE converges to the value that makes the model the best approximation to the true model, in the Kullback-Leibler sense.

Asymptotic distribution of θ_q :

Consider the following Taylor expansion, similar to that we have in the previous lecture note:

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

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

If θ_q is the quasi-true value, then we still have

$$E_0 \left[\frac{\partial \ln L_n(\theta_q)}{\partial \theta} \right] = 0,$$

and hence

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial \ln L_n(\theta_q)}{\partial \theta} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \ln f(y_i, x_i; \theta_q)}{\partial \theta} \\ &\xrightarrow{D} N(0, I(\theta_q)), \end{aligned} \quad (8.5)$$

where

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

Furthermore, note that

$$\hat{\theta}_n^* \xrightarrow{p} \theta_q, \quad (8.6)$$

and

$$-\frac{1}{n} \frac{\partial^2 \ln L_n(\theta)}{\partial \theta \partial \theta'} \xrightarrow{p} E_0 \left[-\frac{\partial^2 \ln f(y_i, x_i; \theta)}{\partial \theta \partial \theta'} \right] = A(\theta), \quad (8.7)$$

uniformly for $\theta \in \Theta$. However,

$$I(\theta_q) \neq A(\theta_q).$$

Using the results from (8.5), (8.6), and (8.7) we have that

$$\begin{aligned} \frac{1}{n} \frac{\partial^2 \ln L_n(\hat{\theta}_n^*)}{\partial \theta \partial \theta'} &\xrightarrow{p} A(\theta_q), \quad \text{and} \\ \frac{1}{\sqrt{n}} \frac{\partial \ln L_n(\theta_q)}{\partial \theta} &\xrightarrow{D} N(0, I(\theta_q)), \end{aligned}$$

and hence

$$\sqrt{n} (\hat{\theta}_n - \theta_q) \xrightarrow{D} N\left(0, \xrightarrow{p} A^{-1}(\theta_q) I(\theta_q) A^{-1}(\theta_q)\right). \quad (8.8)$$

Note that the asymptotic covariance matrix has the *sandwich (or Eicker-White)* formula.

If the information equality holds (which is not the case when the model is misspecified) then the covariance matrix simplifies to the usual MLE form.