

**LECTURE 14**  
**MAXIMUM LIKELIHOOD ESTIMATION**

**Definition**

Suppose that the econometrician observes the data  $\{W_i : i = 1, \dots, n\}$ , where  $W_i$  is a random  $p$ -vector. Assume that the  $W_i$  are iid with PDF  $f(w_i; \theta)$ , where  $\theta \in \Theta \subset \mathbb{R}^k$  is an unknown vector of parameters. The set  $\Theta$  is usually assumed to be compact.

**Example** (Normal regression model). Let  $W_i = (Y_i, X_i^\top)^\top$ ,  $\theta = (\beta^\top, \sigma^2)^\top$ , where  $\beta \in \mathbb{R}^k$  and  $\sigma^2 > 0$ . Assume that  $Y_i = X_i^\top \beta + U_i$ , and  $U_i | X_i \sim N(0, \sigma^2)$ . Then,

$$f(y_i, x_i; \beta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - x_i^\top \beta)^2}{2\sigma^2}\right).$$

Since the observations are iid, the joint PDF of  $W_1, \dots, W_n$  is given by

$$\prod_{i=1}^n f(w_i; \theta).$$

The joint PDF gives the *likelihood* of the sample for a given value of  $\theta$ . The log of the joint PDF is

$$\sum_{i=1}^n \log f(w_i; \theta).$$

The log-likelihood function is defined as  $1/n$  times the log of the joint PDF evaluated at the random sample  $W_1, \dots, W_n$ :

$$\log L_n(\theta) = n^{-1} \sum_{i=1}^n \log f(W_i; \theta).$$

The maximum likelihood (ML) estimator is defined as

$$\hat{\theta}_n^{ML} = \arg \max_{\theta \in \Theta} \log L_n(\theta).$$

For a fixed set of observations, the ML estimate is the value of  $\theta$  for which we are most likely to observe the values of  $W_1, \dots, W_n$  obtained in the sample.

In the normal regression example,

$$\log L_n(\beta, \sigma^2) = -\frac{1}{2} \log \sigma^2 - \frac{1}{2} \log(2\pi) - \frac{1}{2\sigma^2 n} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2,$$

and

$$\hat{\beta}_n^{ML} = \left( \sum_{i=1}^n X_i X_i^\top \right)^{-1} \sum_{i=1}^n X_i Y_i,$$

$$\hat{\sigma}_{n,ML}^2 = n^{-1} \sum_{i=1}^n (Y_i - X_i^\top \hat{\beta}_n^{ML})^2.$$

In this case, the ML estimator of  $\beta$  is identical to the OLS estimator, since maximization of  $\log L_n$  with respect to  $\beta$  is equivalent to minimization of  $\sum_{i=1}^n (Y_i - X_i^\top \beta)^2$ .

# Asymptotic properties of the ML estimator

Let  $\theta_0$  be the true value of  $\theta$ .

## Consistency

By the WLLN, we should expect that for each value of  $\theta$ ,

$$\begin{aligned}\log L_n(\theta) &= n^{-1} \sum_{i=1}^n \log f(W_i; \theta) \\ &\rightarrow_p \mathbb{E}[\log f(W_i; \theta)] \\ &= \int (\log f(w; \theta)) f(w; \theta_0) dw.\end{aligned}$$

Consider the difference

$$\begin{aligned}&\mathbb{E}[\log f(W_i; \theta)] - \mathbb{E}[\log f(W_i; \theta_0)] \\ &= \mathbb{E} \left[ \log \frac{f(W_i; \theta)}{f(W_i; \theta_0)} \right] \\ &\leq \log \mathbb{E} \left[ \frac{f(W_i; \theta)}{f(W_i; \theta_0)} \right] \\ &= \log \int \frac{f(w; \theta)}{f(w; \theta_0)} f(w; \theta_0) dw \\ &= \log \int f(w; \theta) dw \\ &= \log 1 \\ &= 0.\end{aligned}$$

The inequality follows from Jensen's inequality applied to the concave function  $\log$ : if  $g$  is concave, then  $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$ . The inequality is strict provided that  $\Pr(f(W_i; \theta) \neq f(W_i; \theta_0)) > 0$  for all  $\theta \neq \theta_0$ . As a result,  $\theta_0$  uniquely maximizes  $\mathbb{E}[\log f(W_i; \theta)]$ , and, under additional regularity conditions,

$$\begin{aligned}\hat{\theta}_n^{ML} &= \arg \max_{\theta \in \Theta} \log L_n(\theta) \\ &\rightarrow_p \arg \max_{\theta \in \Theta} \mathbb{E}[\log f(W_i; \theta)] \\ &= \theta_0.\end{aligned}$$

## Asymptotic normality

The ML estimator solves the first-order conditions

$$0 = n^{-1} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(W_i; \hat{\theta}_n^{ML}).$$

Applying the mean value theorem element by element, we obtain

$$0 = n^{-1} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(W_i; \theta_0) + n^{-1} \sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_n^*) (\hat{\theta}_n^{ML} - \theta_0), \quad (1)$$

where  $\theta_n^*$  lies between  $\hat{\theta}_n^{ML}$  and  $\theta_0$ . Since  $\hat{\theta}_n^{ML} \rightarrow_p \theta_0$ , we have  $\theta_n^* \rightarrow_p \theta_0$ . Rearranging (1) gives

$$n^{1/2} (\hat{\theta}_n^{ML} - \theta_0) = - \left( n^{-1} \sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_n^*) \right)^{-1} n^{-1/2} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(W_i; \theta_0). \quad (2)$$

Under regularity conditions,

$$n^{-1} \sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_n^*) \rightarrow_p \mathbb{E} \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_0) \right]. \quad (3)$$

Consider  $\partial \log f(W_i; \theta_0) / \partial \theta$ . Assuming the interchange of integration and differentiation is valid,

$$\begin{aligned} \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f(W_i; \theta_0) \right] &= \mathbb{E} \left[ \frac{\partial f(W_i; \theta_0) / \partial \theta}{f(W_i; \theta_0)} \right] \\ &= \int \frac{\partial f(w; \theta_0) / \partial \theta}{f(w; \theta_0)} f(w; \theta_0) dw \\ &= \int \frac{\partial f(w; \theta_0)}{\partial \theta} dw \\ &= \frac{\partial}{\partial \theta} \int f(w; \theta_0) dw \\ &= \frac{\partial}{\partial \theta} 1 \\ &= 0. \end{aligned}$$

Thus, by the CLT, we should expect that

$$n^{-1/2} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(W_i; \theta_0) \rightarrow_d N \left( 0, \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f(W_i; \theta_0) \frac{\partial}{\partial \theta^\top} \log f(W_i; \theta_0) \right] \right). \quad (4)$$

Combining (2), (3), and (4), we obtain

$$\begin{aligned} n^{1/2} (\hat{\theta}_n^{ML} - \theta_0) &\rightarrow_d - \left( \mathbb{E} \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_0) \right] \right)^{-1} N \left( 0, \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f(W_i; \theta_0) \frac{\partial}{\partial \theta^\top} \log f(W_i; \theta_0) \right] \right) \\ &= N(0, V), \end{aligned}$$

where

$$V = \left( \mathbb{E} \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_0) \right] \right)^{-1} \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f(W_i; \theta_0) \frac{\partial}{\partial \theta^\top} \log f(W_i; \theta_0) \right] \left( \mathbb{E} \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_0) \right] \right)^{-1}.$$

Next,

$$\begin{aligned} \mathbb{E} \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_0) \right] &= \mathbb{E} \left[ \frac{\partial}{\partial \theta^\top} \frac{\partial f(W_i; \theta_0) / \partial \theta}{f(W_i; \theta_0)} \right] \\ &= \mathbb{E} \left[ \frac{\partial^2 f(W_i; \theta_0) / \partial \theta \partial \theta^\top}{f(W_i; \theta_0)} \right] - \mathbb{E} \left[ \frac{\partial f(W_i; \theta_0) / \partial \theta}{f(W_i; \theta_0)} \frac{\partial f(W_i; \theta_0) / \partial \theta^\top}{f(W_i; \theta_0)} \right] \\ &= - \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f(W_i; \theta_0) \frac{\partial}{\partial \theta^\top} \log f(W_i; \theta_0) \right], \end{aligned} \quad (5)$$

since

$$\begin{aligned} \mathbb{E} \left[ \frac{\partial^2 f(W_i; \theta_0) / \partial \theta \partial \theta^\top}{f(W_i; \theta_0)} \right] &= \int \frac{\partial^2 f(w; \theta_0) / \partial \theta \partial \theta^\top}{f(w; \theta_0)} f(w; \theta_0) dw \\ &= \int \frac{\partial^2 f(w; \theta_0)}{\partial \theta \partial \theta^\top} dw \\ &= \frac{\partial^2}{\partial \theta \partial \theta^\top} \int f(w; \theta_0) dw \\ &= \frac{\partial^2}{\partial \theta \partial \theta^\top} 1 \\ &= 0. \end{aligned}$$

Equation (5) is the *information equality*. This equality implies that

$$\begin{aligned} V &= - \left( \mathbb{E} \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_0) \right] \right)^{-1} \\ &= I^{-1}(\theta_0), \end{aligned}$$

where

$$I(\theta_0) = - \mathbb{E} \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(W_i; \theta_0) \right]$$

is the *information matrix*. Thus,

$$n^{1/2} \left( \hat{\theta}_n^{ML} - \theta_0 \right) \rightarrow_d N(0, I^{-1}(\theta_0)).$$

## Remarks

- ML estimation generally requires numerical maximization of the log-likelihood function.
- Hypothesis testing can be performed using a Wald-type statistic. Suppose that  $H_0 : h(\theta_0) = 0$ , where  $h : \mathbb{R}^k \rightarrow \mathbb{R}^q$ . The Wald statistic is given by

$$W_n = nh \left( \hat{\theta}_n^{ML} \right)^\top \left( \frac{\partial h \left( \hat{\theta}_n^{ML} \right)}{\partial \theta^\top} I^{-1} \left( \hat{\theta}_n^{ML} \right) \frac{\partial h^\top \left( \hat{\theta}_n^{ML} \right)}{\partial \theta} \right)^{-1} h \left( \hat{\theta}_n^{ML} \right).$$

One should reject the null if  $W_n > \chi_{q,1-\alpha}^2$ . Alternatively, the null hypothesis  $h(\theta_0) = 0$  can be tested using the *likelihood ratio* statistic

$$LR_n = 2n \left( \log L_n \left( \hat{\theta}_n^{ML} \right) - \log L_n \left( \tilde{\theta}_n^{ML} \right) \right),$$

where  $\tilde{\theta}_n^{ML}$  is the null-restricted ML estimator:

$$\tilde{\theta}_n^{ML} = \arg \max_{\theta \in \Theta, h(\theta)=0} \log L_n(\theta).$$

Under the null,  $LR_n \rightarrow_d \chi_q^2$ , and  $LR_n$  is asymptotically equivalent to the Wald statistic.

- The ML estimator is efficient in the following sense: let  $\hat{\theta}_n$  be any estimator such that  $n^{1/2} \left( \hat{\theta}_n - \theta_0 \right) \rightarrow_d N(0, \Sigma(\theta_0))$ . Then the matrix  $\Sigma(\theta_0) - I^{-1}(\theta_0)$  is positive semidefinite. The ML estimator therefore has the smallest asymptotic variance among all consistent and asymptotically normal estimators. This is the Cramér–Rao lower bound.
- ML estimation relies on a strong assumption: the true PDF is known up to the values of the parameters. If the PDF is misspecified, the estimator is called the quasi-ML estimator. In some cases, the quasi-ML estimator remains consistent for certain parameters. For instance, the normal quasi-ML estimator in the linear regression model coincides with OLS, which remains consistent for  $\beta$  even if the errors are not normally distributed.