LECTURE 14
MAXIMUM LIKELIHOOD ESTIMATION

Definition

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

Example (Normal regression model)**.** Let $W_i = (Y_i, X_i')'$, $\theta = (\beta', \sigma^2)'$, where $\beta \in R^k$ and $\sigma^2 \in R$. Assume that $Y_i = X_i'\beta + U_i$, and $U_i|X_i \sim N(0, \sigma^2)$. Then, $f(y_i, x_i, \beta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-(y_i - x_i'\beta)^2 / (2\sigma^2)\right)$.

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

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

The joint PDF gives us the *likelihood* of our sample given the value of θ. Log of joint PDF is

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

We define the log-likelihood function as $1/n$ log of joint PDF evaluated at random sample $W_1, \ldots, W_n$:

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

The maximum likelihood (ML) estimator is defined as

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

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

In the normal regression example,

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

and

$$\begin{aligned} \widehat{\beta}_{n,ML} &= \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i Y_i, \\ \widehat{\sigma}^2_{n,ML} &= n^{-1} \sum \left(Y_i - X_i' \widehat{\beta}_n^{ML} \right)^2. \end{aligned}$$

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

Asymptotic properties of the ML estimator

Let θ_0 be the true value of θ.

Consistency

By the WLLN, we should expect that for each value of θ,

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

Next, consider

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

The inequality above follows from the fact that log is a concave function and the Jensen's inequality: if f is concave, then $Ef(X)\leq f(EX)$. The inequality is in fact strict provided that $P(f(W_i,\theta_0)\neq f(W_i,\theta))>0$ for all $\theta\neq\theta_0$. As a result, θ_0 uniquely maximizes $E\log f(W_i,\theta)$, and, under additional technical assumptions, we have that

$$\begin{aligned}\widehat{\theta}_n^{ML} &= \arg\max_{\theta\in\Theta}\log L_n(\theta)\\ &\to_p \arg\max_{\theta\in\Theta}E\log f(W_i,\theta)\\ &= \theta_0.\end{aligned}$$

Asymptotic normality

The ML estimator solves the first order conditions

$$0=n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f\left(W_i,\widehat{\theta}_n^{ML}\right).$$

Using the mean value theorem element-by-element, we obtain

$$0=n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)+n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\log f(W_i,\theta_n^*)\left(\widehat{\theta}_n^{ML}-\theta_0\right), \tag{1}$$

where θ_n^* lies between $\widehat{\theta}_n^{ML}$ and θ_0. Note that, since $\widehat{\theta}_n^{ML}\to_p\theta_0$, we have that $\theta_n^*\to_p\theta$. Re-arranging (1) gives

$$n^{1/2}\left(\widehat{\theta}_n^{ML}-\theta_0\right)=-\left(n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\log f(W_i,\theta_n^*)\right)^{-1}n^{-1/2}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(W_i,\theta_0). \tag{2}$$

Under additional technical assumptions,

$$n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\log f\left(W_i,\theta_n^*\right)\rightarrow_p E\frac{\partial^2}{\partial\theta\partial\theta'}\log f\left(W_i,\theta_0\right). \tag{3}$$

Next, consider $\partial\log f\left(W_i,\theta_0\right)/\partial\theta$. Assuming that we can change the order of integration and differentiation,

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

Thus, by the CLT, we should expect that

$$n^{-1/2}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f\left(W_i,\theta_0\right)\rightarrow_d N\left(0,E\frac{\partial}{\partial\theta}\log f\left(W_i,\theta_0\right)\frac{\partial}{\partial\theta'}\log f\left(W_i,\theta_0\right)\right). \tag{4}$$

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

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

where

$$V=\left(E\frac{\partial^2}{\partial\theta\partial\theta'}\log f\left(W_i,\theta_0\right)\right)^{-1}E\frac{\partial}{\partial\theta}\log f\left(W_i,\theta_0\right)\frac{\partial}{\partial\theta'}\log f\left(W_i,\theta_0\right)\left(E\frac{\partial^2}{\partial\theta\partial\theta'}\log f\left(W_i,\theta_0\right)\right)^{-1}.$$

Next,

$$\begin{aligned}
E\frac{\partial^2}{\partial\theta\partial\theta'}\log f\left(W_i,\theta_0\right) &= E\frac{\partial}{\partial\theta'}\frac{\partial f\left(W_i,\theta_0\right)/\partial\theta}{f\left(W_i,\theta_0\right)}\\
&= E\frac{\partial^2 f\left(W_i,\theta_0\right)/\partial\theta\partial\theta'}{f\left(W_i,\theta_0\right)}-E\frac{\partial f\left(W_i,\theta_0\right)/\partial\theta}{f\left(W_i,\theta_0\right)}\frac{\partial f\left(W_i,\theta_0\right)/\partial\theta'}{f\left(W_i,\theta_0\right)}\\
&= -E\frac{\partial}{\partial\theta}\log f\left(W_i,\theta_0\right)\frac{\partial}{\partial\theta'}\log f\left(W_i,\theta_0\right), \qquad (5)
\end{aligned}$$

since,

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

The result in (5) is called the *information equality*. It implies that

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

where

$$
I\left(\theta_0\right) = -E\frac{\partial^2}{\partial\theta\partial\theta'}\log f\left(W_i,\theta_0\right)
$$

is called the *information matrix.* Thus, we have that

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

Remarks

- In general, ML estimation requires numerical maximization of the log-likelihood function.
- Hypothesis testing can be performed using the Wald type statistic. Suppose that $H_0 : h\left(\theta_0\right) = 0$, where $h : R^k \to R^q$. The Wald statistic is given by

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

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

$$
LR_n = 2\left(\log L_n\left(\widehat{\theta}_n^{ML}\right) - \log L_n\left(\widetilde{\theta}_n^{ML}\right)\right),
$$

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

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

  Under the null, $LR_n \to_d \chi^2_q$, and asymptotically equivalent to the Wald statistic.
- The ML estimator is efficient. Let $\widehat{\theta}_n$ be an estimator such that $n^{1/2}\left(\widehat{\theta}_n - \theta_0\right) \to_d N\left(0, \Sigma\left(\theta_0\right)\right)$, then $\Sigma\left(\theta_0\right) - I^{-1}\left(\theta_0\right)$ is positive semi-definite. Thus, the asymptotic variance of the ML estimator is as small as, or smaller than, the variance of any consistent and asymptotically normal estimator.
- The ML estimation relies on very strong assumptions - the true PDF is known up to the value of parameters. If the PDF is misspecified, the estimator is called the quasi-ML estimator. In some cases, the quasi-ML estimator is still consistent. The OLS estimator is an example. It is still consistent even if the data is not normally distributed.