Introductory Econometrics

Lecture 22: Maximum likelihood

Instructor: Ma, Jun

Renmin University of China

June 7, 2023

Definition of maximum likelihood: discrete sample

Let $(X_1, ..., X_n)$ be a random (i.i.d.) sample on a discrete population characterized by a vector of parameters $\theta = (\theta_1, ..., \theta_k)$ and let x_i be the observed value of X_i. Then we call

$$L(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} \Pr[X_i = x_i; \theta]$$

the likelihood function of θ given $(x_1, ..., x_n)$, and we call the value of θ that maximizes $L(\theta; X_1, ..., X_n)$ the maximum likelihood estimator.

- The purpose of estimation is to pick a probability distribution among many (usually infinite) probability distributions that could have generated given observations.
- Maximum likelihood estimation means choosing the probability distribution under which the observed values could have occurred with the highest probability.

An example

- Suppose $X \sim \text{Bin}(n, p_*)$ (binomial distribution) for some unknown p_* and n is known. Suppose our sample size is 1.
- The likelihood function at any parameter value $p \in (0,1)$ is given by
$$L(p;x) = C_n^x p^x (1-p)^{n-x}.$$
- We shall maximize $\log(L)$ rather than L because it is simpler. Since log is a monotonically increasing function, the value of the maximum likelihood estimator is unchanged by this transformation.
- We have
$$\log(L(p;x)) = \log\left(C_n^x\right) + x\log(p) + (n-x)\log(1-p).$$
- Solving the first order condition, the maximum likelihood estimator is $\hat{p} = X/n$.

Definition of maximum likelihood: continuous sample

- Let $(X_1, ..., X_n)$ be a random (i.i.d.) sample on a continuous with a density function $f(\cdot;\theta)$ where $\theta = (\theta_1, ..., \theta_k)$, and let x_i be the observed value of X_i. Then we call

$$L(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

  the likelihood function of θ given $(x_1, x_2, ..., x_n)$, and we call the value of θ that maximizes $L(\theta; X_1, ..., X_n)$ the maximum likelihood estimator.

- The function

$$\ell(\theta; x_1, ..., x_n) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

  is usually called the log-likelihood function.

An Example

- Let X_i, $i = 1, 2, \ldots, n$ be a random (i.i.d.) sample from the population $\mathrm{N}\left(\mu_*, \sigma_*^2\right)$ with some unknown (μ_*, σ_*).
- Then the likelihood function is given by

$$L\left(\mu, \sigma^2; x_1, x_2, \ldots, x_n\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(x_i - \mu\right)^2\right)$$

so that

$$\ell\left(\mu, \sigma^2; x_1, x_2, \ldots, x_n\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2.$$

- Equating the derivatives to zero, we have

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0$$

and

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu)^2 = 0.$$

- By solving the first order conditions, the maximum likelihood estimators are

$$\hat{\mu} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2,$$

which are simply the sample mean and the sample variance.

MLE in a general set-up

- Suppose now we have observations for a dependent variable Y and an explanatory variable (vector) X: $(Y_1, X_1), ..., (Y_n, X_n)$.
- Suppose we have a parametric model for the true joint density function of $(Y_1, X_1, ..., Y_n, X_n)$:

$$f(y_1, x_1, ..., y_n, x_n; \theta).$$

- In the case where the observations $(Y_1, X_1), ..., (Y_n, X_n)$ are i.i.d.,

$$f(y_1, x_1, ..., y_n, x_n; \theta) = \prod_{i=1}^{n} f_{Y|X}(y_i \mid x_i; \theta) f_X(x_i; \theta),$$

where $f_{Y|X}(\cdot \mid x; \theta)$ is the conditional density function of Y_i given $X_i = x$ and $f_X(\cdot; \theta)$ is the marginal density of X_i.

- Typically, we assume that f_X is not parametrized, meaning that we leave the marginal density function unspecified.
- Then we take the logarithm of the density function to obtain

$$\sum_{i=1}^{n} \log f_{Y|X}\left(y_i \mid x_i; \theta\right) + \sum_{i=1}^{n} \log f_X\left(x_i\right).$$

- We define the log-likelihood function to be

$$\ell\left(\theta; Y_1, X_1, \ldots, Y_n, X_n\right) = \sum_{i=1}^{n} \log f_{Y|X}\left(Y_i \mid X_i; \theta\right) + \sum_{i=1}^{n} \log f_X\left(X_i\right).$$

- The MLE is the maximizer of $\ell\left(\theta; Y_1, X_1, \ldots, Y_n, X_n\right)$ with respect to θ:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \ell\left(\theta; Y_1, X_1, \ldots, Y_n, X_n\right) = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log f_{Y|X}\left(Y_i \mid X_i; \theta\right),$$

since $\sum_{i=1}^{n} \log f_X\left(X_i\right)$ is independent of θ.

An example: linear regression model with normal errors

- Now extend the sample mean example to the regression model:

$$Y_i = \beta_0 + \beta_1 X_i + U_i .$$

- Suppose the observations are i.i.d. and $U_i \mid X_i \sim \mathrm{N}(0,1)$. This implies $Y_i \mid X_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i, 1)$.
- The conditional density is

$$f_{Y|X}(y \mid x, \beta_0, \beta_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - \beta_0 - \beta_1 x)^2}{2} \right).$$

- The likelihood function is

$$\begin{aligned} & L(b_0, b_1; Y_1, X_1, \ldots, Y_n, X_n) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(Y_i - b_0 - b_1 X_i)^2}{2} \right) f_X(X_i), \end{aligned}$$

  where we left f_X unspecified.

- The log-likelihood function is

$$\ell\left(b_0, b_1; Y_1, X_1, \ldots, Y_n, X_n\right) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - b_0 - b_1 X_i\right)^2 + \sum_{i=1}^{n} \log f_X\left(X_i\right).$$

- Therefore, maximizing $\ell\left(b_0, b_1; Y_1, X_1, \ldots, Y_n, X_n\right)$ with respect to (b_0, b_1) is equivalent to minimizing

$$\sum_{i=1}^{n}\left(Y_i - b_0 - b_1 X_i\right)^2$$

with respect to (b_0, b_1). The minimizer is just the OLS.

Computation

- In both examples, it has been possible to solve the likelihood equation explicitly, equating the derivative of the log likelihood function to zero.
- The likelihood equation is often highly nonlinear in the parameters. It can be solved only by numerical method.
- The most common (numerical) method is the Newton-Raphson method, which can be used to maximize or minimize a general function, not just the likelihood function.
- Let $Q(\theta)$ be the function we want to maximize. Its quadratic Taylor expansion around an initial value θ_1 is given by

$$Q(\theta) \approx Q(\theta_1) + \left.\frac{\partial Q(\theta)}{\partial \theta}\right|_{\theta=\theta_1} (\theta - \theta_1) + \frac{1}{2} \left.\frac{\partial^2 Q(\theta)}{\partial \theta^2}\right|_{\theta=\theta_1} (\theta - \theta_1)^2$$

  where the derivatives are evaluated at θ_1.

- The second-round estimator of the iteration, denoted by θ_2, is the value of θ that maximizes the right hand of the quadratic approximation.
- Therefore,

$$\theta_2 = \theta_1 - \left(\frac{\partial^2 Q(\theta)}{\partial \theta^2}\bigg|_{\theta=\theta_1} \right)^{-1} \frac{\partial Q(\theta)}{\partial \theta}\bigg|_{\theta=\theta_1}.$$

- Next θ_2 can be used as the initial value to compute the third-round estimator, and the iteration should be repeated until it converges.
- Whether the iteration will converge to the global maximum, rather than some other stationary point, and if it does, how fast it converges depend upon the shape of Q and the initial value.

Cramér Rao Lower Bound (CRLB)

- Let $L(\theta; X_1, ..., X_n)$ be the likelihood function and let $\hat{\theta}$ be an unbiased estimator of θ_*, the true parameter. Then under general conditions, we have

$$\text{Var}\left[\hat{\theta}\right] \geq -\left(\text{E}\left[\left.\frac{\partial^2 \log L}{\partial \theta^2}\right|_{\theta=\theta_*}\right]\right)^{-1}$$

  where the right hand side is known to be the Cramer Rao lower bound.

- When we discuss the definition and computation, the likelihood function was always evaluated at the observed values of the sample, since we are concerned with the definition and computation only. When we are concerned with the properties of the maximum likelihood estimator, we need to evaluate the likelihood function at the random variables X_1, X_2,...,X_n, which makes the likelihood function itself random.

Proof of the Cramér Rao Theorem

First we notice

$$\mathrm{E}\left[\left.\frac{\partial^2 \log L}{\partial \theta^2}\right|_{\theta=\theta_*}\right] = n \cdot \mathrm{E}\left[\left.\frac{\partial^2 \log f\left(X;\theta\right)}{\partial \theta^2}\right|_{\theta=\theta_*}\right].$$

Define $\mathrm{S}\left(\theta\right)$ as the score function:

$$\mathrm{S}\left(\theta\right) = \frac{\partial \log L\left(\theta; X_1, ..., X_n\right)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log f\left(X_i;\theta\right)}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{f\left(X_i;\theta\right)} \frac{\partial f\left(X_i;\theta\right)}{\partial \theta}.$$

We have the following equalities (noticing that $\int f\left(x;\theta\right) \mathrm{d}x = 1$ for all θ)

$$\int \frac{\partial \log f\left(x;\theta\right)}{\partial \theta} f\left(x;\theta\right) \mathrm{d}x = \int \frac{\partial f\left(x;\theta\right)}{\partial \theta} \mathrm{d}x = \frac{\partial}{\partial \theta} \int f\left(x;\theta\right) \mathrm{d}x = 0,$$

where the integral and partial derivative have been interchanged. Therefore we have $\mathrm{E}\left[\mathrm{S}\left(\theta_*\right)\right] = 0$.

By standard rules of differentiation, we have

$$\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} = \frac{\partial}{\partial \theta}\frac{\partial \log f(x;\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial \theta}\right)$$

and also

$$\begin{aligned}\frac{\partial}{\partial \theta}\left(\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial \theta}\right) &= -\frac{1}{f(x;\theta)^2}\left(\frac{\partial f(x;\theta)}{\partial \theta}\right)^2 + \frac{1}{f(x;\theta)}\frac{\partial^2 f(x;\theta)}{\partial \theta^2} \\ &= -\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2 + \frac{1}{f(x;\theta)}\frac{\partial^2 f(x;\theta)}{\partial \theta^2}.\end{aligned}$$

Therefore we have

$$\begin{aligned}\mathrm{E}\left[\left.\frac{\partial^2 \log L(\theta;X_1,\ldots,X_n)}{\partial \theta^2}\right|_{\theta=\theta_*}\right] &= -n\cdot \mathrm{E}\left[\left(\left.\frac{\partial \log f(X;\theta)}{\partial \theta}\right|_{\theta=\theta_*}\right)^2\right] \\ &\quad + n\cdot \mathrm{E}\left[\left.\frac{1}{f(X;\theta_*)}\frac{\partial^2 f(X;\theta)}{\partial \theta^2}\right|_{\theta=\theta_*}\right]\end{aligned}$$

We have for any θ

$$\int \frac{\partial^2 f(x;\theta)}{\partial \theta^2} \mathrm{d}x = \frac{\partial^2}{\partial \theta^2} \int f(x;\theta)\, \mathrm{d}x = 0$$

since $\int f(\theta;x)\, \mathrm{d}x = 1$ for all θ. Therefore we have

$$\mathrm{E}\left[\frac{1}{f(X;\theta_*)} \left. \frac{\partial^2 f(X;\theta)}{\partial \theta^2} \right|_{\theta=\theta_*} \right] = 0$$

and

$$\mathrm{E}\left[\left. \frac{\partial^2 \log f(X;\theta)}{\partial \theta^2} \right|_{\theta=\theta_*} \right] = -\mathrm{E}\left[\left(\left. \frac{\partial \log f(X;\theta)}{\partial \theta} \right|_{\theta=\theta_*} \right)^2 \right].$$

This implies that

$$\mathrm{Var}\left[\mathrm{S}(\theta_*)\right] = -\mathrm{E}\left[\left. \frac{\partial^2 \log L(\theta; X_1, \ldots, X_n)}{\partial \theta^2} \right|_{\theta=\theta_*} \right].$$

For all θ, we have

$$\begin{aligned}
&\int \cdots \int \hat{\theta} \frac{\partial \log L\left(\theta; x_1, \ldots, x_n\right)}{\partial \theta} L\left(\theta; x_1, \ldots, x_n\right) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&\qquad = \int \cdots \int \hat{\theta} \frac{\partial L\left(\theta; x_1, \ldots, x_n\right)}{\partial \theta} \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&\qquad\qquad = \frac{\partial}{\partial \theta} \int \cdots \int \hat{\theta} L\left(\theta; x_1, \ldots, x_n\right) \mathrm{d}x_1 \cdots \mathrm{d}x_n = 1.
\end{aligned}$$

The covariance of $\hat{\theta}$ and $\mathrm{S}\left(\theta_*\right)$ is (since $\mathrm{E}\left[S\left(\theta_*\right)\right] = 0$)

$$\mathrm{Cov}\left[\hat{\theta}, S\left(\theta_*\right)\right] = \mathrm{E}\left[\hat{\theta} \left. \frac{\partial \log L\left(\theta; X_1, \ldots, X_n\right)}{\partial \theta} \right|_{\theta=\theta_*}\right] = 1.$$

By the Cauchy-Schwarz inequality, we have

$$\left|\mathrm{Cov}\left[\hat{\theta}, S\left(\theta_*\right)\right]\right| \leq \sqrt{\mathrm{Var}\left[\hat{\theta}\right]} \cdot \sqrt{\mathrm{Var}\left[S\left(\theta_*\right)\right]}.$$

Therefore, we have

$$\mathrm{Var}\left[\hat{\theta}\right] \geq \mathrm{Var}\left[S\left(\theta_*\right)\right]^{-1}.$$

CRLB Example

Let $X \sim \mathrm{Bin}(n, p_*)$ (n is known). We have

$$\frac{\partial^2 \log L(p; X)}{\partial p^2} = -\frac{X}{p^2} - \frac{n - X}{(1-p)^2}.$$

Therefore we obtain

$$\mathrm{CRLB} = \frac{p_*(1-p_*)}{n}$$

since $\mathrm{E}[X] = np_*$. In this case, the maximum likelihood estimator

$$\hat{p} = \frac{X}{n}$$

has variance

$$\mathrm{Var}[\hat{p}] = \mathrm{CRLB}.$$

Therefore it is the best unbiased estimator.

Consistency of Maximum Likelihood

- The maximum likelihood estimator can be shown to be consistent under general conditions. Let us define

$$Q_n(\theta) = \frac{1}{n}\log L(\theta; X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i; \theta).$$

- By LLN, we know that for each θ,

$$Q_n(\theta) \longrightarrow_p Q(\theta)$$

  where $Q(\theta) = \mathrm{E}[\log f(X; \theta)]$.
- The maximum likelihood estimator is defined to be the maximizer of $Q_n(\theta)$. We expect the maximizer should converge to the maximizer of its limit $Q(\theta)$ in probability.
- We can show that $Q(\theta)$ is maximized at θ_* (the true density of X is $f(\cdot; \theta_*)$).

- Jensen's inequality: Let X be a random variable and g be a strictly concave function. That is,

$$g\left(\lambda a + (1-\lambda) b\right) > \lambda g\left(a\right) + (1-\lambda) g\left(b\right)$$

for any $a < b$ and $0 < \lambda < 1$. Then

$$\mathrm{E}\left[g\left(X\right)\right] < g\left(\mathrm{E}\left[X\right]\right).$$

- Take g to be log and X to be $f\left(X;\theta\right) / f\left(X;\theta_*\right)$ for arbitrary θ. If $\theta \neq \theta_*$.

$$\mathrm{E}\left[\log\left(\frac{f\left(X;\theta\right)}{f\left(X;\theta_*\right)}\right)\right] < \log\left(\mathrm{E}\left[\frac{f\left(X;\theta\right)}{f\left(X;\theta_*\right)}\right]\right).$$

- But we have

$$\mathrm{E}\left[\frac{f\left(X;\theta\right)}{f\left(X;\theta_*\right)}\right] = \int \frac{f\left(x;\theta\right)}{f\left(x;\theta_*\right)} f\left(x;\theta_*\right) \mathrm{d}x = 1, \text{ for all } \theta.$$

Asymptotic Normality of Maximum Likelihood

- Under general conditions, we have

$$\sqrt{n}\left(\hat{\theta}_{ML}-\theta_*\right) \to_d \mathrm{N}\left(0, V_{ML}\right),$$

where

$$V_{ML} = -n\left(\mathrm{E}\left[\left.\frac{\partial^2 \log L}{\partial \theta^2}\right|_{\theta=\theta_*}\right]\right)^{-1} = \left(\mathrm{E}\left[\left(\left.\frac{\partial \log f\left(X;\theta\right)}{\partial \theta}\right|_{\theta=\theta_*}\right)^2\right]\right)^{-1}$$

- This means

$$\hat{\theta}_{ML} \overset{a}{\sim} \mathrm{N}\left(\theta_*, -\left(\mathrm{E}\left[\left.\frac{\partial^2 \log L}{\partial \theta^2}\right|_{\theta=\theta_*}\right]\right)^{-1}\right)$$

when n is very large, i.e. the maximum likelihood estimator attains the Cramer-Rao lower bound asymptotically.

- Loosely speaking, the maximum likelihood estimator has the smallest asymptotic variance among all the consistent estimators $\hat{\theta}$ such that $\sqrt{n}\left(\hat{\theta}-\theta_*\right)$ is asymptotically normal.

An Example

- Let X have density belonging to the family

$$f(x;\mu) = \begin{cases} \left(1 + \frac{1-2\mu}{\mu-1}\right) x^{\frac{1-2\mu}{\mu-1}} & x \in (0,1) \\ 0 & x \notin (0,1), \end{cases}$$

  for $0 < \mu < 1$, with true density $f(x;\mu_*)$.
- It can be shown that $\mu = \int x f(x;\mu)\,\mathrm{d}x$, i.e., in this parametrization, μ is also the population mean.
- We observe a random sample $X_1, \ldots, X_n$. The log-maximum likelihood function is

$$\log L(\mu; X_1, \ldots, X_n) = n \log\left(\frac{\mu}{1-\mu}\right) + \frac{1-2\mu}{\mu-1} \sum_{i=1}^{n} \log(X_i).$$

- Differentiating with respect to μ:

$$\frac{\partial \log L}{\partial \mu} = \frac{n}{\mu\,(1-\mu)} + \frac{1}{(1-\mu)^2} \sum_{i=1}^{n} \log\left(X_i\right).$$

- Solving the first order condition, the maximum likelihood estimator is

$$\hat{\mu} = \frac{n}{n - \sum_{i=1}^{n} \log\left(X_i\right)}.$$

- It can be shown that

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{(1-2\mu)\, n}{\mu^2 (1-\mu)^2} + \frac{2}{(1-\mu)^3} \sum_{i=1}^{n} \log (X_i)$$

  and

$$\int \log (x)\, f\,(x;\mu)\, \mathrm{d}x = \frac{\mu - 1}{\mu}.$$

- Therefore the asymptotic variance of the maximum likelihood estimator is $\mu_*^2 (1-\mu_*)^2$.
- For this example, the asymptotic variance of the sample mean is $\mu_* (1-\mu_*)^2 / (2-\mu_*)$.
- It can be shown that

$$\frac{\mu_* (1-\mu_*)^2}{2-\mu_*} - \mu_*^2 (1-\mu_*)^2 > 0$$

  for $0 < \mu_* < 1$.