Intermediate Econometrics

Maximum Likelihood Estimation

July 28, 2011

Maximum Likelihood (ML) Estimation

- To introduce the maximum likelihood approach to estimating unknown parameters, we consider a rather simple example.[1]

- Suppose that an urn contains a number of black and a number of white balls.

- It is known that the ratio of the numbers is $3/1$.

- However, it is not known whether the black or the white balls are more numerous.

- That is, the probability p of drawing a black ball is either $1/4$ or $3/4$.

[1] This is taken from Mood/Graybill/Boes: *Introduction to the Theory of Statistics*, McGraw-Hill, third edition, 1974.

- Define the *Bernoulli random variable* X as

$$X = \begin{cases} 1 & \text{if the ball is black} \\ 0 & \text{if the ball is white} \end{cases} \tag{1}$$

- Since p is the probability of drawing a black ball, the probability mass function of X is

$$\begin{aligned} f(x;p) &= \Pr(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases} && (2) \\ &= p^x(1-p)^{1-x}, \quad x \in \{0,1\}. && (3) \end{aligned}$$

- Now suppose we draw a sample of three balls *with replacement*[2] from the urn.

[2]This means that the probability does not change, as for each draw there is the same number of white and black balls in the urn.

- There are 8 possible outcomes of such an experiment:[3]

Table 1: Drawing balls from an urn

outcome	probability of outcome if...	
	$p = \frac{1}{4}$	$p = \frac{3}{4}$
(b, b, b)	0.0156	0.4219
(b, b, w)	0.0469	0.1406
(b, w, b)	0.0469	0.1406
(w, b, b)	0.0469	0.1406
(b, w, w)	0.1406	0.0469
(w, b, w)	0.1406	0.0469
(w, w, b)	0.1406	0.0469
(w, w, w)	0.4219	0.0156

- Now we use the outcome of the experiment to estimate the unknown parameter p (which is either $1/4$ or $3/4$ in our case).

[3] In the table, (w, b, w) describes the outcome that the first ball drawn is white, the second is black and the third is white. The probability of such an event is $p \times (1 - p)^2$.

- Suppose we observe (w, b, b).

- Then, applying maximum likelihood estimation, we would estimate p to be $3/4$, because such a sample is more likely to arise from a population with $p = \frac{3}{4}$ than from one with $p = \frac{1}{4}$.

- Following this logic, we can define the *maximum likelihood estimator* (MLE) for this example via

$$\widehat{p}_{ML} = \begin{cases} \frac{1}{4} & \text{if the number of black balls drawn is 0 or 1} \\ \frac{3}{4} & \text{if the number of black balls drawn is 2 or 3} \end{cases} \tag{4}$$

Maximum Likelihood (ML) Estimation

- Let us generalize the procedure considered so far:

- Suppose that a random sample of size n is drawn from the Bernoulli distribution with probability mass function

$$f(x;p) = p^x(1-p)^{1-x}, \quad x \in \{0,1\}, \quad p \in [0,1]. \tag{5}$$

- With sample values $x_1, x_2, \ldots, x_n$, the joint density of the random sample is

$$\prod_{i=1}^{n} f(x_i;p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \tag{6}$$

- Once we have observed the sample values $x_1, \ldots, x_n$, we may view (6) as a function of p:

$$L(p; x_1, \ldots, x_n) =: L(p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \tag{7}$$

- This is the **likelihood function**: The joint density function of n random variables *as a function of the parameters, with given values of the random variable outcomes* $x_1, \ldots, x_n$.

- Following the reasoning in our introductory example, we choose the maximum likelihood estimator (MLE) of p so as to maximize the probability of the sample at hand.

- That is, the maximum–likelihood estimator (MLE) is the value of the unknown parameter which maximizes the likelihood function.

- Typically, for ease of calculations, the *log–likelihood function* is considered; in our example:

$$\begin{aligned} \log L(p) &= \log\left\{\prod_{i=1}^{n} f(x_i;p)\right\} = \log\left\{p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}\right\} \\ &= y\log p + (n-y)\log(1-p), \end{aligned}$$

  where

$$y = \sum_{i=1}^{n} x_i. \tag{8}$$

- The MLE can be obtained by doing the maximization,

$$\frac{d\log L(p)}{dp} = \frac{y}{p} - \frac{n-y}{1-p} = 0 \Rightarrow \widehat{p}_{ML} = \frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}, \tag{9}$$

  which appears to be a rather intuitive result.

- As a *general method of estimation*, maximum likelihood of course needs a stronger justification than the intuitive argument presented so far.

- Now consider a continuous random variable X with *density* $f(x;\theta)$, where θ is a vector of parameters.

- For continuous random variables, the density $f(x;\theta)$ cannot be interpreted as a probability.

- We have

$$\Pr(a \leq X \leq b) = \int_a^b f(x;\theta)dx = F(b;\theta) - F(a;\theta) \tag{10}$$

  instead, where F is the cumulative distribution function (cdf) of X,

$$F(b;\theta) = \int_{-\infty}^b f(x;\theta)dx = \Pr(X \leq b). \tag{11}$$

- Still, however, regions where the density is high have a higher probability than low–density regions.

- Thus, proceeding with maximum likelihood estimation as before may still give rise to reasonable results (which, of course, remains to be seen).

- To illustrate maximum likelihood estimation for continuous distributions, suppose we have a random sample of size n from a normal distribution with mean μ and variance σ^2, with density ($\theta = (\mu, \sigma^2)$)

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{12}$$

- When the sample values are given by $x_1, \ldots, x_n$, the likelihood function is

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \right) \\ &= \left(\frac{1}{2\pi}\right)^{n/2} (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}. \end{aligned}$$

- The log–likelihood function is

$$\log L(\mu, \sigma^2) = \sum_{i=1}^{n} \log f(x_i; \mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{13}$$

- Maximizing (13) with respect to the unknown parameters shows that the MLE is

$$\widehat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}, \quad \widehat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{14}$$

- That is, the MLE of the normal mean is the sample mean, which we know is unbiased and consistent by the law of large numbers.

- The MLE of σ^2 is *not* unbiased, since (with one regressor, i.e., a constant), the unbiased estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{15}$$

- However, $\widehat{\sigma}^2_{ML}$ is consistent for σ^2, since

$$\widehat{\sigma}^2_{ML} = \underbrace{\frac{n-1}{n}}_{\substack{n \to \infty \\ \longrightarrow \\ 1}} \tilde{\sigma}^2 \xrightarrow{p} \sigma^2. \tag{16}$$

- This is a general property of MLEs: Although they are not unbiased in many situations, they are consistent under rather general conditions.

- This is an attractive property, since in many situations unbiased estimators do not exist, and consistency may be the best that can be achieved.

- Moreover, the MLE typically has an asymptotic normal distribution; namely, for a scalar parameter θ, subject to some technical conditions,

$$\sqrt{n}(\widehat{\theta}_{ML} - \theta) \xrightarrow{d} \mathsf{N}(0, 1/I(\theta)), \tag{17}$$

where

$$I(\theta) = \mathsf{E}\left[\left(\frac{d \log f(x;\theta)}{d\theta}\right)^2\right] = -\mathsf{E}\left(\frac{d^2 \log f(x;\theta)}{d\theta^2}\right). \qquad (18)$$

- Thus, in large samples, we treat $\widehat{\theta}_{ML}$ as

$$\widehat{\theta}_{ML} \overset{a}{\sim} \text{Normal}\left(\theta, \frac{1}{nI(\widehat{\theta}_{ML})}\right), \qquad (19)$$

  where $\overset{a}{\sim}$ means approximately in large samples.

- Approximation (19) can be used to construct confidence intervals and conduct t tests of significance.

- If the expectation in (18) cannot be calculated explicitly, we can estimate it via its sample analogue,

$$\widehat{I}(\widehat{\theta}_{ML}) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(x;\widehat{\theta}_{ML})}{\partial\theta^2} = -\frac{1}{n}\frac{\partial^2 \log L(\widehat{\theta}_{ML})}{\partial\theta^2}. \qquad (20)$$

- In case of a p–dimensional parameter vector $\theta = [\theta_1, \theta_2, \ldots, \theta_p]$, a similar result holds. In this case, under general conditions, we may treat the MLE as

$$\widehat{\theta}_{ML} \overset{a}{\sim} \mathsf{N}(\theta, n^{-1} I(\theta)^{-1}), \tag{21}$$

  where the covariance matrix $I(\theta)^{-1}$ is a (symmetric) $p \times p$ matrix with elements

$$I(\theta)_{ij} = -\mathsf{E}\left[\frac{\partial^2 f(x;\theta)}{\partial\theta_i \partial\theta_j}\right], \quad i, j = 1, \ldots, p. \tag{22}$$

- Again, if the integration involved in calculating the expectation in (22) cannot be done analytically, it can be estimated by means of sample analogues analogously to (20).

- The ML estimator is also efficient. That is, the asymptotic variance of the MLE is not larger than the asymptotic variance of any other consistent and asymptotically normally distributed estimator.

- Due to these properties, maximum likelihood estimation is one of the most frequently applied approaches to estimation in econometrics.

- To further appreciate the usefulness of such general results, note that often closed–form expressions for the estimator are not available (as for logit and probit models).

- In the two examples considered above, the MLEs have been the sample mean in the first case (Bernoulli example) and the sample mean and the sample variance in the second case (normal distribution example), i.e., ML estimation gave rise to closed–from estimators.

- The OLS estimator in the linear regression model likewise has a closed–form solution.

- When a closed–form solution is available, the properties of these solutions can often be analyzed directly; e.g., we know that the sample mean is unbiased and consistent for the population mean (by the law of large numbers) and approximately normal in large samples (by the central limit theorem); similar results have been obtained for the OLS estimator.

- When closed–from solutions are not available, however, such direct analysis is not possible, and results about the general properties of an estimation strategy are rather valuable.

- As a simple example for this case, consider ML estimation of the mean μ of a logistic distribution, which has density function

$$f(x;\mu) = \frac{\exp\{x-\mu\}}{(1+\exp\{x-\mu\})^2}, \quad -\infty < x < \infty. \tag{23}$$

- For a random sample of size n with sample values $x_1, \ldots, x_n$, the log–likelihood function is

$$\begin{aligned} \log L(\mu) &= \prod_{i=1}^{n} f(x_i;\mu) = \log\left\{\prod_{i=1}^{n} \frac{\exp\{x_i-\mu\}}{(1+\exp\{x_i-\mu\})^2}\right\} && (24) \\ &= \sum_i (x_i-\mu) - 2\sum_{i=1}^{n} \log\left(1+\exp\{x-\mu\}\right). && (25) \end{aligned}$$

- Therefore the MLE is the solution of the likelihood equation

$$0 = \frac{d \log L(\mu)}{d\mu} = -n + 2\sum_{i=1}^{n} \frac{\exp\{x_i - \mu\}}{1 + \exp\{x_i - \mu\}} \tag{26}$$

$$= \sum_{i=1}^{n} \left[-1 + 2\frac{\exp\{x_i - \mu\}}{1 + \exp\{x_i - \mu\}} \right] \tag{27}$$

$$= \sum_{i=1}^{n} \frac{\exp\{x_i - \mu\} - 1}{1 + \exp\{x_i - \mu\}}. \tag{28}$$

- This equation cannot be solved in closed–from, so numerical methods have to be used to find the MLE $\widehat{\mu}$.

- Statistical inference for the mean can be based on (19), where the asymptotic variance follows from and (19)

$$\frac{d^2 \log f(x;\mu)}{d\mu^2} = \frac{d}{d\mu}\left(\frac{\exp\{x-\mu\}-1}{\exp\{x-\mu\}+1}\right) \quad (29)$$

$$= -2\frac{\exp\{x-\mu\}}{(\exp\{x-\mu\}+1)^2}, \quad (30)$$

so that

$$\begin{aligned} I(\theta) &= -\mathsf{E}\left[\frac{d^2 \log f(x;\mu)}{d\mu^2}\right] = -\int_{-\infty}^{\infty} \frac{d^2 \log f(x;\mu)}{d\mu^2} f(x;\mu)dx \\ &= 2\int_{-\infty}^{\infty} \frac{\exp\{x-\mu\}^2}{(1+\exp\{x-\mu\})^4}dx \overset{y=x-\mu}{=} 2\int_{-\infty}^{\infty} \frac{e^{2y}}{(1+e^y)^4}dy \\ &\overset{z=e^y}{=} = 2\int_0^{\infty} \frac{z^2}{(1+z)^4}\frac{dz}{z} \overset{s=1+z}{=} 2\int_1^{\infty} \frac{s-1}{s^4}ds = 2\int_1^{\infty}\left(\frac{1}{s^3}-\frac{1}{s^4}\right)ds \\ &= 2\left(\frac{1}{2}-\frac{1}{3}\right) = \frac{2}{6} = \frac{1}{3}. \end{aligned}$$

- Although we don't have a closed–form solution for the ML estimator $\widehat{\mu}_{ML}$, i.e., the solution of (26), we know from (17) that

$$\sqrt{n}(\widehat{\mu}_{ML} - \mu) \xrightarrow{d} \mathsf{N}\left(0, 1/I(\theta)\right) = \mathsf{N}(0, 3), \tag{31}$$

so for statistical inference we treat $\widehat{\mu}_{ML}$ as (cf. Equation (19))

$$\widehat{\mu}_{ML} \overset{a}{\sim} \mathsf{N}\left(\mu, \frac{3}{n}\right), \tag{32}$$

which can form the basis for the calculation of confidence intervals and t tests about μ.