Introduction to Maximum Likelihood Estimation

Eric Zivot

July 26, 2012

The Likelihood Function

Let $X_1, \ldots, X_n$ be an iid sample with pdf $f(x_i; \theta)$, where θ is a $(k \times 1)$ vector of parameters that characterize $f(x_i; \theta)$.

Example: Let $X_i \tilde{} N(\mu, \sigma^2)$ then

$$\begin{aligned} f(x_i; \theta) &= (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \\ \theta &= (\mu, \sigma^2)' \end{aligned}$$

The *joint density* of the sample is, by independence, equal to the product of the marginal densities

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

The joint density is an n dimensional function of the data $x_1, \ldots, x_n$ given the parameter vector θ and satisfies

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) &\geq 0 \\ \int \cdots \int f(x_1, \ldots, x_n; \theta) dx_1 \cdots dx_n &= 1. \end{aligned}$$

The *likelihood function* is defined as the joint density treated as a function of the parameters θ :

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

Notice that the likelihood function is a k dimensional function of θ given the data $x_1, \ldots, x_n$.

It is important to keep in mind that the likelihood function, being a function of θ and not the data, is not a proper pdf. It is always positive but

$$\int \cdots \int L(\theta|x_1, \ldots, x_n) d\theta_1 \cdots d\theta_k \neq 1.$$

To simplify notation, let the vector $\mathbf{x} = (x_1, \ldots, x_n)$ denote the observed sample. Then the joint pdf and likelihood function may be expressed as $f(\mathbf{x}; \theta)$ and $L(\theta|\mathbf{x})$, respectively.

Example 1 *Bernoulli Sampling*

Let $X_i \tilde{}$ Bernoulli(θ). That is,

$$\begin{aligned} X_i &= 1 \text{ with probability } \theta \\ X_i &= 0 \text{ with probability } 1-\theta \end{aligned}$$

The pdf for X_i is

$$f(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i},\ x_i = 0,1$$

Let $X_1, \ldots, X_n$ be an iid sample with $X_i \tilde{}$ Bernoulli(θ). The joint density / likelihood function is given by

$$f(\mathbf{x};\theta) = L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

Since X_i is a discrete random variable

$$f(\mathbf{x};\theta) = \Pr(X_1 = x_1, \ldots, X_n = x_n)$$

Example 2 *Normal Sampling*

Let $X_1, \ldots, X_n$ be an iid sample with $X_i\tilde{\ } N(\mu, \sigma^2)$. The pdf for X_i is

$$\begin{aligned} f(x_i;\theta) &= (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right), \\ \theta &= (\mu,\sigma^2)' \\ -\infty &< \mu<\infty,\ \sigma^2>0,\ -\infty<x_i<\infty \end{aligned}$$

The likelihood function is given by

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \end{aligned}$$

The Maximum Likelihood Estimator

Suppose we have a random sample from the pdf $f(x_i; \theta)$ and we are interested in estimating θ.

The maximum likelihood estimator, denoted $\hat{\theta}_{mle}$, is the value of θ that maximizes $L(\theta|\mathbf{x})$. That is,

$$\hat{\theta}_{mle} = \arg\max_{\theta} L(\theta|\mathbf{x})$$

Alternatively, we say that $\hat{\theta}_{mle}$ solves

$$\max_{\theta} L(\theta|\mathbf{x})$$

It is often quite difficult to directly maximize $L(\theta|\mathbf{x})$. It usually much easier to maximize the log-likelihood function $\ln L(\theta|\mathbf{x})$. Since $\ln(\cdot)$ is a monotonic function

$$\hat{\theta}_{mle} = \arg\max_{\theta} \ln L(\theta|\mathbf{x})$$

With random sampling, the log-likelihood has the particularly simple form

$$\ln L(\theta|\mathbf{x}) = \ln\left(\prod_{i=1}^{n} f(x_i;\theta)\right) = \sum_{i=1}^{n} \ln f(x_i;\theta)$$

Example 3 *Bernoulli example continued*

Given the likelihood function

$$L(\theta|\mathbf{x}) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i},$$

the log-likelihood is

$$\begin{aligned} \ln L(\theta|\mathbf{x}) &= \ln\left(\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}\right) \\ &= \left(\sum_{i=1}^{n} x_i\right)\ln(\theta) + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-\theta) \end{aligned}$$

Recall the results

$$\ln(x \cdot y) = \ln(x) + \ln(y), \;\; \ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y), \;\; \ln(x^y) = y\ln(x)$$

Example 4 *Normal example continued*

Given the likelihood function

$$\ln L(\theta|\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

the log-likelihood is

$$\ln L(\theta|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Recall the result

$$\ln(e^x) = x$$

Since the MLE is defined as the maximization problem, we can use the tools of calculus to determine its value. That is, we may find the MLE by differentiating $\ln L(\theta|\mathbf{x})$ and solving the first order conditions

$$\frac{\partial \ln L(\hat{\theta}_{mle}|\mathbf{x})}{\partial \theta} = \mathbf{0}$$

Since θ is $(k \times 1)$ the first order conditions define k, potentially nonlinear, equations in k unknown values:

$$\frac{\partial \ln L(\hat{\theta}_{mle}|\mathbf{x})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ln L(\hat{\theta}_{mle}|\mathbf{x})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln L(\hat{\theta}_{mle}|\mathbf{x})}{\partial \theta_k} \end{pmatrix} = \mathbf{0}$$

Review of Optimization Techniques: Unconstrained Optimization

Example: finding the minimum of a univariate function

$$y = f(x) = x^2$$
$$\min_x \; y = f(x)$$

First order conditions for a minimum

$$0 = \frac{df(x)}{dx} = \frac{d}{dx} 2x = 2 \cdot x$$
$$\Rightarrow x = 0$$

Second order conditions for a minimum

$$0 < \frac{d^2 f(x)}{dx^2} = \frac{d}{dx} 2 \cdot x = 2$$

- R function `optimize()`
  - Use to optimize (maximize or minimize) functions of one variable

- Excel solver
  - General optimizer for unconstrained and constrained optimization problems involving many variables
  - solver in Office 2010 is substantially improved and expanded

Example: Finding the minimum of a bivariate function

$$
\begin{aligned}
y &= f(x,z) = x^2 + z^2 \\
\min_{x,z}\ y &= f(x,z)
\end{aligned}
$$

First order conditions for a minimum

$$
\begin{aligned}
0 &= \frac{\partial f(x,z)}{\partial x} = \frac{\partial}{\partial x}\left(x^2 + z^2\right) = 2 \cdot x \\
0 &= \frac{\partial f(x,z)}{\partial z} = \frac{\partial}{\partial z}\left(x^2 + z^2\right) = 2 \cdot z \\
&\Rightarrow x = 0,\ z = 0
\end{aligned}
$$

Remark:

Second order conditions depend on the properties of the second derivative Hessian matrix

$$H(x,z) = \frac{\partial^2 f(x,z)}{\partial x \partial z} = \begin{bmatrix} \frac{\partial^2 f(x,z)}{\partial x^2} & \frac{\partial^2 f(x,z)}{\partial x \partial z} \\ \frac{\partial^2 f(x,z)}{\partial z \partial x} & \frac{\partial^2 f(x,z)}{\partial z^2} \end{bmatrix}$$

- R functions `nlminb()`, `optim()`
  - Use to optimize (maximize or minimize) functions of one or more variables variable
  - `nlminb()` uses Newton's method based on 1st and 2nd derivatives and can allow for box constraints on parameters
  - `optim()` can use 4 types of algorithms (secant method, Newton method, simplex method, simulated annealing)
- Excel solver

Example 5 *Bernoulli example continued*

To find the MLE for θ, we maximize the log-likelihood function

$$\ln L(\theta|\mathbf{x}) = \sum_{i=1}^{n} x_i \ln(\theta) + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-\theta)$$

The derivative of the log-likelihood is

$$\frac{\partial \ln L(\theta|\mathbf{x})}{\partial \theta} = \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\left(n - \sum_{i=1}^{n} x_i\right) = 0$$

The MLE satisfies $\frac{\partial \ln L(\theta|\mathbf{x})}{\partial \theta} = 0$ and solving for θ gives

$$\hat{\theta}_{mle} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Example 6 *Normal example continued*

To find the MLE for $\theta = (\mu, \sigma^2)'$, we maximize the log-likelihood function

$$\ln L(\theta|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

The derivative of the log-likelihood is a (2×1) vector given by

$$\frac{\partial \ln L(\theta|\mathbf{x})}{\partial \theta} = \begin{pmatrix} \frac{\partial \ln L(\theta|\mathbf{x})}{\partial \mu} \\ \frac{\partial \ln L(\theta|\mathbf{x})}{\partial \sigma^2} \end{pmatrix}$$

where

$$\begin{aligned}\frac{\partial \ln L(\theta|\mathbf{x})}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\\ \frac{\partial \ln L(\theta|\mathbf{x})}{\partial \sigma^2} &= -\frac{n}{2}(\sigma^2)^{-1}+\frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

Solving $\frac{\partial \ln L(\theta|\mathbf{x})}{\partial \theta}=0$ gives the *normal equations*

$$\frac{\partial \ln L(\hat{\theta}_{mle}|\mathbf{x})}{\partial \mu}=\frac{1}{\hat{\sigma}^2_{mle}}\sum_{i=1}^{n}(x_i-\hat{\mu}_{mle})=0$$

$$\frac{\partial \ln L(\hat{\theta}_{mle}|\mathbf{x})}{\partial \sigma^2}=-\frac{n}{2}(\hat{\sigma}^2_{mle})^{-1}$$

$$+\frac{1}{2}(\hat{\sigma}^2_{mle})^{-2}\sum_{i=1}^{n}(x_i-\hat{\mu}_{mle})^2=0$$

Solving the first equation for $\hat{\mu}_{mle}$ gives

$$\hat{\mu}_{mle} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

Solving the second equation for $\hat{\sigma}^2_{mle}$ gives

$$\hat{\sigma}^2_{mle} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{mle})^2.$$

Notice that $\hat{\sigma}^2_{mle}$ is not equal to the sample variance.

Invariance Property of Maximum Likelihood Estimators

One of the attractive features of the method of maximum likelihood is its invariance to one-to-one transformations of the parameters of the log-likelihood.

That is, if $\hat{\theta}_{mle}$ is the MLE of θ and $\alpha = h(\theta)$ is a one-to-one function of θ then $\hat{\alpha}_{mle} = h(\hat{\theta}_{mle})$ is the mle for α.

Example 7 *Normal Model Continued*

The log-likelihood is parameterized in terms of μ and σ^2 and

$$\begin{aligned} \hat{\mu}_{mle} &= \bar{x} \\ \hat{\sigma}^2_{mle} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{mle})^2 \end{aligned}$$

Suppose we are interested in the MLE for

$$\sigma = h(\sigma^2) = (\sigma^2)^{1/2}$$

which is a one-to-one function for $\sigma^2 > 0$.

The invariance property says that

$$\hat{\sigma}_{mle} = (\hat{\sigma}^2_{mle})^{1/2} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{mle})^2\right)^{1/2}$$

The Precision of the Maximum Likelihood Estimator

Intuitively, the precision of $\hat{\theta}_{mle}$ depends on the curvature of the log-likelihood function near $\hat{\theta}_{mle}$.

If the log-likelihood is very curved or "steep" around $\hat{\theta}_{mle}$, then θ will be precisely estimated. In this case, we say that we have a lot of *information* about θ.

If the log-likelihood is not curved or "flat" near $\hat{\theta}_{mle}$, then θ will not be precisely estimated. Accordingly, we say that we do not have much information about θ.

If the log-likelihood is completely flat in θ then the sample contains no information about the true value of θ because every value of θ produces the same value of the likelihood function. When this happens we say that θ is not *identified*.

The curvature of the log-likelihood is measured by its second derivative matrix (*Hessian*)

$$H(\theta|\mathbf{x}) = \frac{\partial^2 \ln L(\theta|\mathbf{x})}{\partial\theta\partial\theta'} = \begin{bmatrix} \frac{\partial^2 \ln L(\theta|\mathbf{x})}{\partial\theta_1\partial\theta_1} & \cdots & \frac{\partial^2 \ln L(\theta|\mathbf{x})}{\partial\theta_1\partial\theta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln L(\theta|\mathbf{x})}{\partial\theta_k\partial\theta_1} & \cdots & \frac{\partial^2 \ln L(\theta|\mathbf{x})}{\partial\theta_k\partial\theta_k} \end{bmatrix}$$

Since the Hessian is negative semi-definite, the *information* in the sample about θ may be measured by $-H(\theta|\mathbf{x})$. If θ is a scalar then $-H(\theta|\mathbf{x})$ is a positive number.

The expected amount of information in the sample about the parameter θ is the information matrix $I(\theta|\mathbf{x}) = -E[H(\theta|\mathbf{x})]$.

As we shall see, the Hessian and information matrix are directly related to the precision of the MLE.

Asymptotic Properties of Maximum Likelihood Estimators

Let $X_1, \ldots, X_n$ be an iid sample with probability density function (pdf) $f(x_i;\theta)$, where θ is a $(k \times 1)$ vector of parameters that characterize $f(x_i;\theta)$.

Under general regularity conditions, the ML estimator of θ is consistent and asymptotically normally distributed. That is,

$$\hat{\theta}_{mle} \xrightarrow{p} \theta \text{ as } n \to \infty$$

and for n large enough the Central Limit Theorem gives

$$\hat{\theta}_{mle} \sim N(\theta, I(\theta|\mathbf{x})^{-1})$$

Computing MLEs in R: the maxLik package

The R package maxLik has the function `maxLik()` for computing MLEs for any user-defined log-likelihood function

- uses the `optim()` function for maximizing the log-likelihood function
- Automatically computes standard errors by inverting the Hessian matrix

Remarks

- In practice we don't know $I(\theta|\mathbf{x}) = -E[H(\theta|\mathbf{x})]$ but we can estimate its value using $-H(\hat{\theta}_{mle}|\mathbf{x})$. Hence, the practically useful asymptotic normality result is

$$\hat{\theta}_{mle} \sim N(\theta, -H(\hat{\theta}_{mle}|\mathbf{x})^{-1})$$

- Estimated standard errors for the elements of $\hat{\theta}_{mle}$ are the square roots of the diagonal elements of $-H(\hat{\theta}_{mle}|\mathbf{x})^{-1}$:

$$\begin{aligned} \widehat{SE}(\hat{\theta}_{i,mle}) &= \sqrt{\left[-H(\hat{\theta}_{mle}|\mathbf{x})^{-1}\right]_{ii}} \\ \left[-H(\hat{\theta}_{mle}|\mathbf{x})^{-1}\right]_{ii} &= (i,i) \text{ element of } -H(\hat{\theta}_{mle}|\mathbf{x})^{-1} \end{aligned}$$

Optimality Properties of MLE (or why we care about MLE)

- Recall, a good estimator $\hat{\theta}$ has small bias and high precision (small $SE(\hat{\theta})$)

- The best estimator among all possible estimators has the smallest bias and smallest $SE(\hat{\theta})$

- In many cases, it can be shown that maximum likelihood estimator is the best estimator among all possible estimators (especially for large sample sizes)

MLE of the CER Model Parameters

Recall, the CER model matrix notation is

$$\begin{aligned} \mathbf{r}_t &= \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \sim GWN(\mathbf{0}, \boldsymbol{\Sigma}) \\ &\Rightarrow \mathbf{r}_t \sim iid\ N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

Given an iid sample $\mathbf{r} = \{\mathbf{r}_1, \ldots, \mathbf{r}_n\}$, the likelihood and log-likelihood functions for $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{r}) &= (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{t=1}^{n}(\mathbf{r}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{r}_t - \boldsymbol{\mu})\right\} \\ \ln L(\boldsymbol{\theta}|\mathbf{r}) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{n}(\mathbf{r}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{r}_t - \boldsymbol{\mu}) \end{aligned}$$

It can be shown that the MLEs for the elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\begin{aligned}
\hat{\mu}_{i,mle} &= \frac{1}{T}\sum_{t=1}^{T} r_{it},\ \ i = 1,\ldots,n \\
\hat{\sigma}^2_{i,mle} &= \frac{1}{T}\sum_{t=1}^{T}(r_{it} - \hat{\mu}_i)^2,\ \ i = 1,\ldots,n \\
\hat{\sigma}_{i,mle} &= \sqrt{\hat{\sigma}^2_{i,mle}},\ i = 1,\ldots,n \\
\hat{\sigma}_{ij,mle} &= \frac{1}{T}\sum_{t=1}^{T}(r_{it} - \hat{\mu}_i)(r_{jt} - \hat{\mu}_j),\ \ i,j = 1,\ldots,n \\
\hat{\rho}_{ij,mle} &= \frac{\hat{\sigma}_{ij,mle}}{\hat{\sigma}_{i,mle}\cdot\hat{\sigma}_{j,mle}},\ \ i,j = 1,\ldots,n
\end{aligned}$$

Remarks

- The MLEs for μ_i and ρ_{ij} are the same as the plug-in principle estimates

- The MLEs for σ_i^2, σ_i and σ_{ij} are almost equal to the plug-in principle estimates. They differ by a degrees of freedom adjustment ($\frac{1}{T}$ vs. $\frac{1}{T-1}$)

- The plug-in estimates for σ_i^2 and σ_{ij} are unbiased; the MLEs have a tiny bias that disappears in large samples.

- The formulas for the standard errors of the plug-in principle estimates come from the formulas for the standard errors of the MLEs