Maximum Likelihood Estimation (Shortened Version)

Eric Zivot

Winter 2013

The Likelihood Function

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

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

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

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 p 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

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

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 *Linear Regression Model with Normal Errors*

Consider the linear regression

$$\begin{aligned} y_i &= \underset{(1\times k)}{x_i'} \underset{(k\times 1)}{\beta} + \varepsilon_i,\ i = 1, \ldots, n \\ & \varepsilon_i | x_i \sim \text{iid } N(0, \sigma^2) \end{aligned}$$

The pdf of $\varepsilon_i | x_i$ is

$$f(\varepsilon_i | x_i; \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\varepsilon_i^2\right)$$

The Jacobian of the transformation for ε_i to y_i is one so the pdf of $y_i | x_i$ is normal with mean $x_i'\beta$ and variance σ^2 :

$$\begin{aligned} f(y_i | x_i; \theta) &= (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right) \\ \theta &= (\beta', \sigma^2)' \text{ and } p = k + 1 \end{aligned}$$

Given an iid sample of n observations, y and X, the joint density of the sample is

$$\begin{aligned} f(\mathbf{y}|\mathbf{X};\theta) &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right) \\ \mathbf{y} &= (y_1,\ldots,y_n)',\ x_i = (x_{i1},\ldots,x_{in})' \\ \mathbf{X} &= [x_1,\ldots,x_k] \end{aligned}$$

The log-likelihood function is then

$$\begin{aligned} \ln L(\theta|\mathbf{y},\mathbf{X}) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) \\ &\quad -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

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,

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

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

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

In the notation of extremum estimators we have

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m(x_i;\theta), \; m(x_i;\theta) = \ln f(x_i;\theta)$$

Since the MLE is defined as a maximization problem, we would like know the conditions under which we may determine the MLE using the techniques of calculus.

A regular pdf $f(x;\theta)$ provides a sufficient set of such conditions. We say the $f(x;\theta)$ is regular if

1. The support of the random variables $X, S_X = \{x : f(x;\theta) > 0\}$, does not depend on θ

2. $f(x;\theta)$ is at least three times differentiable with respect to θ

3. The true value of θ lies in a compact set Θ

If $f(x;\theta)$ is regular then 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 $(p \times 1)$ the first order conditions define p, potentially nonlinear, equations in p unknown values:

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

The vector of derivatives of the log-likelihood function is called the *score* vector and is denoted

$$\underset{p\times 1}{S(\theta|\mathbf{x})} = \frac{\partial \ln L(\theta|\mathbf{x})}{\partial \theta}$$

By definition, the MLE satisfies

$$S(\hat{\theta}_{mle}|\mathbf{x}) = 0$$

Under random sampling

$$S(\theta|\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta} = \sum_{i=1}^{n} S(\theta|x_i)$$

where

$$S(\theta|x_i) = \frac{\partial \ln f(x_i;\theta)}{\partial \theta}$$

Example 2 *Linear regression example continued*

The log-likelihood is

$$\begin{aligned} \ln L(\theta|\mathbf{y},\mathbf{X}) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) \\ &\quad -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

The MLE of θ satisfies $S(\hat{\theta}_{mle}|\mathbf{y},\mathbf{X}) = 0$ where $S(\theta|\mathbf{y},\mathbf{X}) = \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{y},\mathbf{X})$ is the score vector. Now

$$\begin{aligned} \frac{\partial \ln L(\theta|\mathbf{y},\mathbf{X})}{\partial\beta} &= \frac{-1}{2\sigma^2}\frac{\partial}{\partial\beta}[\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta] \\ &= -(\sigma^2)^{-1}[-\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\beta] \\ \frac{\partial \ln L(\theta|\mathbf{y},\mathbf{X})}{\partial\sigma^2} &= -\frac{n}{2}(\sigma^2)^{-1} \\ &\quad +\frac{1}{2}(\sigma^2)^{-2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

Solving $\frac{\partial \ln L(\hat{\theta}_{mle}|\mathbf{y},\mathbf{X})}{\partial \beta} = 0$ for $\hat{\beta}_{mle}$ gives

$$\hat{\beta}_{mle} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{\beta}_{OLS}$$

Next, solving $\frac{\partial \ln L(\hat{\theta}_{mle}|\mathbf{y},\mathbf{X})}{\partial \sigma^2} = 0$ for $\hat{\sigma}^2_{mle}$ gives

$$\begin{aligned} \hat{\sigma}^2_{mle} &= \frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\beta}_{mle})'(\mathbf{y} - \mathbf{X}\widehat{\beta}_{mle}) \\ &\neq \hat{\sigma}^2_{OLS} = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\widehat{\beta}_{OLS})'(\mathbf{y} - \mathbf{X}\widehat{\beta}_{OLS}) \end{aligned}$$

Properties of the Score Function

The $(p \times p)$ matrix of second derivatives of the log-likelihood is called the *Hessian*

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

The *information matrix* is defined as minus the expectation of the Hessian

$$I(\theta|\mathbf{x}) = -E[H(\theta|\mathbf{x})]$$

If we have random sampling then

$$H(\theta|\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial^2 \ln f(\theta|x_i)}{\partial\theta\partial\theta'} = \sum_{i=1}^{n} H(\theta|x_i)$$

and

$$I(\theta|\mathbf{x}) = -\sum_{i=1}^{n} E[H(\theta|x_i)] = -nE[H(\theta|x_i)] = nI(\theta|x_i)$$

Proposition 1 *Let $f(x_i;\theta)$ be a regular pdf. Then*

1. $E[S(\theta|x_i)] = \int S(\theta|x_i) f(x_i;\theta) dx_i = 0$

2. If θ is a scalar then

$$\begin{aligned} \text{var}(S(\theta|x_i)) &= E[S(\theta|x_i)^2] \\ &= \int S(\theta|x_i)^2 f(x_i;\theta) dx_i = I(\theta|x) \end{aligned}$$

   If θ is a vector then

$$\begin{aligned} \text{var}(S(\theta|x_i)) &= E[S(\theta|x_i) S(\theta|x)'] \\ &= \int S(\theta|x_i) S(\theta|x)' f(x_i;\theta) dx_i \\ &= I(\theta|x_i) \end{aligned}$$

Proof. For part 1, we have

$$\begin{aligned}
E[S(\theta|x_i)] &= \int S(\theta|x_i) f(x_i;\theta) dx_i \\
&= \int \frac{\partial \ln f(x_i;\theta)}{\partial \theta} f(x_i;\theta) dx_i \\
&= \int \frac{1}{f(x_i;\theta)} \frac{\partial}{\partial \theta} f(x_i;\theta) f(x_i;\theta) dx_i \\
&= \int \frac{\partial}{\partial \theta} f(x_i;\theta) dx_i \\
&= \frac{\partial}{\partial \theta} \int f(x_i;\theta) dx_i \\
&= \frac{\partial}{\partial \theta} \cdot 1 \\
&= 0.
\end{aligned}$$

The key part to the proof is the ability to interchange the order of differentiation and integration.

Part 2: Homework problem.

MLE as a Method of Moments Estimator

The MLE can be derived as a method of moments estimator based on the population moment

$$E[S(\theta|x_i)] = 0$$

The FOCs for maximizing the log-likelihood sets the sample score to zero

$$S(\hat{\theta}_{mle}|\mathbf{x}) = 0$$

For identification, we require

$$\begin{aligned} E[S(\theta|x_i)] &= 0 \text{ at } \theta = \theta_0 \\ E[S(\theta|x_i)] &\neq 0 \text{ at } \theta \neq \theta_0 \end{aligned}$$

Sufficient conditions for identification are given later on.

Computation: Newton-Raphson Iteration

Goal: Using iterative scheme compute

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

Idea: Consider 2nd order TSE of $\ln L(\theta|\mathbf{x})$ about starting value $\hat{\theta}_1$

$$\begin{aligned}\ln L(\theta|\mathbf{x}) &= \ln L(\hat{\theta}_1|\mathbf{x}) + \frac{\partial \ln L(\hat{\theta}_1|\mathbf{x})}{\partial \theta'}(\theta - \hat{\theta}_1) \\ &\quad + \frac{1}{2}(\theta - \hat{\theta}_1)' \frac{\partial^2 \ln L(\hat{\theta}_1|\mathbf{x})}{\partial \theta \partial \theta'}(\theta - \hat{\theta}_1) + error\end{aligned}$$

Now maximize 2nd order TSE wrt θ. The FOCs are

$$\underset{p\times 1}{\mathbf{0}} = \frac{\partial \ln L(\hat{\theta}_1|\mathbf{x})}{\partial \theta} + \frac{\partial^2 \ln L(\hat{\theta}_1|\mathbf{x})}{\partial \theta \partial \theta'}(\hat{\theta}_2 - \hat{\theta}_1)$$

Solve for $\hat{\theta}_2$

$$\begin{aligned}\hat{\theta}_2 &= \hat{\theta}_1 - \left[\frac{\partial^2 \ln L(\hat{\theta}_1|\mathbf{x})}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial \ln L(\hat{\theta}_1|\mathbf{x})}{\partial\theta} \\ &= \hat{\theta}_1 - H(\hat{\theta}_1|\mathbf{x})^{-1} S(\hat{\theta}_1|\mathbf{x})\end{aligned}$$

This suggests the iterative scheme

$$\hat{\theta}_{n+1} = \hat{\theta}_n - H(\hat{\theta}_n|\mathbf{x})^{-1} S(\hat{\theta}_n|\mathbf{x})$$

Iteration stops when

$$S(\hat{\theta}_n|\mathbf{x}) \approx 0$$

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 (*Hessian*)

$$H(\theta|\mathbf{x}) = \frac{\partial^2 \ln L(\theta|\mathbf{x})}{\partial\theta\partial\theta'}$$

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 information matrix is directly related to the precision of the MLE.

Theorem 2 *Cramer-Rao Inequality*

Let $X_1, \ldots, X_n$ be an iid sample with pdf $f(x;\theta)$.

Let $\hat{\theta}$ be an unbiased estimator of θ; i.e., $E[\hat{\theta}] = \theta$.

If $f(x;\theta)$ is regular then

$$\text{var}(\hat{\theta}) \geq I(\theta|\mathbf{x})^{-1} = \text{CRLB}$$

where $I(\theta|\mathbf{x}) = -E[H(\theta|\mathbf{x})]$ denotes the sample information matrix.

Note: If θ is a vector then $\text{var}(\theta) \geq I(\theta|\mathbf{x})^{-1}$ means that $\text{var}(\hat{\theta}) - I(\theta|\mathbf{x})$ is positive semi definite

Result: If $E[\hat{\theta}] = \theta$ and $\text{var}(\hat{\theta}) = I(\theta|\mathbf{x})^{-1}$ then $\hat{\theta}$ is the Best Unbiased Estimator (BUE)

Example 3 *Linear regression model continued*

The score vector is given by

$$\begin{aligned} S(\theta|\mathbf{y},\mathbf{X}) &= \begin{pmatrix} -(\sigma^2)^{-1}[-\mathbf{X}'\mathbf{y}+\mathbf{X}'\mathbf{X}\beta] \\ -\frac{n}{2}(\sigma^2)^{-1}+\frac{1}{2}(\sigma^2)^{-2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \end{pmatrix} \\ &= \begin{pmatrix} -(\sigma^2)^{-1}(-\mathbf{X}'\boldsymbol{\varepsilon}) \\ -\frac{n}{2}(\sigma^2)^{-1}+\frac{1}{2}(\sigma^2)^{-2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \end{pmatrix} \\ \boldsymbol{\varepsilon} &= \mathbf{y}-\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Now

$$\begin{aligned} E[\boldsymbol{\varepsilon}] &= 0 \\ E[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}] &= n\sigma^2 \text{ since } \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi^2(n) \end{aligned}$$

and so that

$$E[S(\theta|\mathbf{y},\mathbf{X})] = \begin{pmatrix} -(\sigma^2)^{-1}\left(-\mathbf{X}'E[\boldsymbol{\varepsilon}]\right) \\ -\frac{n}{2}(\sigma^2)^{-1} + \frac{1}{2}(\sigma^2)^{-2}E[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

To determine the Hessian and information matrix we need the second derivatives of $\ln L(\theta|\mathbf{y},\mathbf{X})$:

$$\begin{aligned}
\frac{\partial^2 \ln L(\theta|\mathbf{y},\mathbf{X})}{\partial\beta\partial\beta'} &= \frac{\partial}{\partial\beta'}\left(-(\sigma^2)^{-1}[-\mathbf{X}'\mathbf{y}+\mathbf{X}'\mathbf{X}\beta]\right) \\
&= -(\sigma^2)^{-1}\mathbf{X}'\mathbf{X} \\
\frac{\partial^2 \ln L(\theta|\mathbf{y},\mathbf{X})}{\partial\beta\partial\sigma^2} &= \frac{\partial}{\partial\sigma^2}\left(-(\sigma^2)^{-1}[-\mathbf{X}'\mathbf{y}+\mathbf{X}'\mathbf{X}\beta]\right) \\
&= -(\sigma^2)^{-2}\mathbf{X}'\boldsymbol{\varepsilon} \\
\frac{\partial^2 \ln L(\theta|\mathbf{y},\mathbf{X})}{\partial\sigma^2\partial\beta'} &= -(\sigma^2)^{-2}\boldsymbol{\varepsilon}'\mathbf{X} \\
\frac{\partial^2 \ln L(\theta|\mathbf{y},\mathbf{X})}{\partial\left(\sigma^2\right)^2} &= \frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2}(\sigma^2)^{-1}+\frac{1}{2}(\sigma^2)^{-2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right) \\
&= \frac{n}{2}(\sigma^2)^{-2}-(\sigma^2)^{-3}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}
\end{aligned}$$

Therefore,

$$H(\theta|y,X) = \begin{pmatrix} -(\sigma^2)^{-1}X'X & -(\sigma^2)^{-2}X'\varepsilon \\ -(\sigma^2)^{-2}\varepsilon'X & \frac{n}{2}(\sigma^2)^{-2} - (\sigma^2)^{-3}\varepsilon'\varepsilon \end{pmatrix}$$

and

$$\begin{aligned} I(\theta|y,X) &= -E[H(\theta|y,X)] \\ &= \begin{pmatrix} -(\sigma^2)^{-1}X'X & -(\sigma^2)^{-2}X'E[\varepsilon] \\ -(\sigma^2)^{-2}E[\varepsilon]'X & \frac{n}{2}(\sigma^2)^{-2} - (\sigma^2)^{-3}E[\varepsilon'\varepsilon] \end{pmatrix} \\ &= \begin{pmatrix} (\sigma^2)^{-1}X'X & 0 \\ 0 & \frac{n}{2}(\sigma^2)^{-2} \end{pmatrix} \end{aligned}$$

Notice that the information matrix is block diagonal in β and σ^2. The CRLB for unbiased estimators of θ is then

$$I(\theta|y,X)^{-1} = \begin{pmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2}{n}\sigma^4 \end{pmatrix}$$

Do the MLEs of β and σ^2 achieve the CRLB?

First, $\hat{\beta}_{mle} = \hat{\beta}_{OLS}$ is unbiased and

$$\text{var}(\hat{\beta}_{mle}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = CRLB$$

Therefore, $\hat{\beta}_{mle}$ is BUE.

This is an improvement over the Gauss-Markov theorem which says that $\hat{\beta}_{mle} = \hat{\beta}_{OLS}$ is the most efficient *linear* and unbiased estimator (BLUE).

Next, note that $\hat{\sigma}^2_{mle}$ is not unbiased (why?) so the CRLB result does not apply.

Consider the unbiased estimator

$$s^2 = (n-k)^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}_{OLS})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{OLS})$$

It can be shown that

$$\text{var}(s^2|\mathbf{X}) = \frac{2\sigma^4}{n-k} > \frac{2}{n}\sigma^4 = CRLB$$

Hence s^2 is not the most efficient unbiased estimator of σ^2.

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 4 *Normal Linear Regression Model Continued*

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

$$\begin{aligned}\hat{\beta}_{mle} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \hat{\sigma}^2_{mle} &= \frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\beta}_{mle})'(\mathbf{y} - \mathbf{X}\widehat{\beta}_{mle})\end{aligned}$$

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

$$\hat{\sigma}_{mle} = (\hat{\sigma}^2_{mle})^{1/2} = \left(\frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\beta}_{mle})'(\mathbf{y} - \mathbf{X}\widehat{\beta}_{mle})\right)^{1/2}$$

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 $(p \times 1)$ vector of parameters that characterize $f(x_i; \theta)$.

Under general regularity conditions (to be discussed below), the ML estimator of θ has the following asymptotic properties

1. $\hat{\theta}_{mle} \xrightarrow{p} \theta$

2. $\sqrt{n}(\hat{\theta}_{mle} - \theta) \xrightarrow{d} N(0, I(\theta|x_i)^{-1})$, where

$$I(\theta|x_i) = -E\left[H(\theta|x_i)\right] = -E\left[\frac{\partial \ln f(x_i; \theta)}{\partial \theta \partial \theta'}\right]$$

That is,

$$\text{avar}(\sqrt{n}(\hat{\theta}_{mle} - \theta)) = I(\theta|x_i)^{-1}$$

Alternatively,

$$\hat{\theta}_{mle} \overset{A}{\sim} N\left(\theta, \frac{1}{n} I(\theta|x_i)^{-1}\right) = N(\theta, I(\theta|\mathbf{x})^{-1})$$

where

$$I(\theta|\mathbf{x}) = nI(\theta|x_i)$$

Recall, with iid data

$$\frac{1}{n} I(\theta|x_i)^{-1} = [nI(\theta|x_i)]^{-1} = I(\theta|\mathbf{x})^{-1}$$

Remarks:

1. For a wide class of consistent and asymptotically normal estimators which include GMM estimators, $\hat{\theta}_{mle}$ is efficient in the class. That is,

$$\text{avar}(\sqrt{n}(\tilde{\theta} - \theta)) - \text{avar}(\sqrt{n}(\hat{\theta}_{mle} - \theta)) \geq 0$$

   for any consistent and asymptotically normal estimator $\tilde{\theta}$ in the class.

2. For some weird cases, it is possible to find a consistent and asymptotically normal estimator that has smaller variance than the MLE (see Amemiya, 1985 Example 4.2.4)

Recall, the MLE is an extremum estimator of the form

$$\begin{aligned} Q_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n} m(x_i, \theta) \\ m(x_i, \theta) &= \ln f(x_i; \theta) \end{aligned}$$

The consistency of the MLE requires

1. Continuity of $Q_n(\theta)$ and $Q_0(\theta)$

2. $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i|\theta) \xrightarrow{p} E[\ln f(x_i|\theta)] = Q_0(\theta)$ uniformly in θ

3. Compact parameter space Θ

4. $Q_0(\theta)$ is uniquely maximized at $\theta = \theta_0$.

Now,

1. $Q_n(\theta)$ and $Q_0(\theta)$ will be continuous provided $f(x_i;\theta)$ is continuous

2. Uniform convergence of $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i|\theta)$ to $E[\ln f(x_i|\theta)] = Q_0(\theta)$ is satisfied if

$$E\left[\sup_{\theta} |\ln f(x_i|\theta)|\right] < \infty$$

3. To establish that $Q_0(\theta)$ is uniquely maximized at $\theta = \theta_0$, we need that $E[\ln f(x_i|\theta)]$ is uniquely maximized at $\theta = \theta_0$

Result: $E[\ln f(x_i|\theta)]$ is uniquely maximized at $\theta = \theta_0$ provided

$$\Pr(f(x_i|\theta) \neq f(x_i|\theta_0)) > 0 \text{ for all } \theta \neq \theta_0$$

Sketch of proof.

Let $f(x_i|\theta)$ be a parametric family of hypothetical pdfs with true density function $f(x_i|\theta_0)$.

Suppose $E[\ln f(x_i|\theta)]$ exists and is finite for all θ.

Assume $f(x_i|\theta) > 0$ for all x_i and suppose $\Pr(f(x_i|\theta) \neq f(x_i|\theta_0)) > 0$ for all $\theta \neq \theta_0$.

Define

$$a(x_i) = \frac{f(x_i|\theta)}{f(x_i|\theta_0)}$$

Then by assumption

$$\begin{aligned} \Pr(a(x_i) &\neq 1) = \Pr\left(\frac{f(x_i|\theta)}{f(x_i|\theta_0)} \neq 1\right) \\ &= \Pr(f(x_i|\theta) \neq f(x_i|\theta_0)) > 0 \end{aligned}$$

Recall Jensen's inequality: If $c(x)$ is a strictly concave function and x is a non-constant random variable then

$$E[c(x)] < c(E[x])$$

Since $\ln(x)$ is strictly concave and $a(x_i)$ is non-constant, Jensen's inequality gives

$$E\left[\ln a(x_i)\right] = E\left[\ln \frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right] < \ln E\left[\frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right]$$

Now,

$$\begin{aligned} E\left[\frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right] &= \int \frac{f(x_i|\theta)}{f(x_i|\theta_0)} f(x_i|\theta_0) dx_i \\ &= \int f(x_i|\theta) dx_i = 1 \end{aligned}$$

So that

$$E\left[\ln \frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right] < \ln(1) = 0$$

It follows that

$$E\left[\ln f(x_i|\theta) - \ln f(x_i|\theta_0)\right] < 0 \text{ for all } \theta \neq \theta_0$$

which implies that

$$E[\ln f(x_i|\theta)] < E[\ln f(x_i|\theta_0)] \text{ for all } \theta \neq \theta_0$$

and so $E[\ln f(x_i|\theta)]$ is uniquely maximized at $\theta = \theta_0$

Remarks

- The *Kullback-Leibler Information Criterion* is defined as

$$\text{KLIC}(\theta, \theta_0) = -E\left[\ln \frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right]$$

which measures the expected distance between $f(x_i|\theta)$ and $f(x_i|\theta_0)$. Here

$$\begin{aligned} \text{KLIC}(\theta_0, \theta_0) &= 0 \\ \text{KLIC}(\theta, \theta_0) &> 0 \text{ for } \theta \neq \theta_0 \end{aligned}$$

- The MLE can be seen as the estimator that minimizes the sample version of KLIC. It gives the value of θ that minimizes the expected distance between $f(x_i|\theta)$ and $f(x_i|\theta_0)$

Asymptotic Normality

$$\hat{\theta}_{mle} = \arg\max_{\theta} \ln L(\theta|\mathbf{x}) \Rightarrow S(\hat{\theta}_{mle}|\mathbf{x}) = 0$$

Asymptotic normality of $\hat{\theta}_{mle}$ follows from an exact first order Taylor's series expansion of the first order conditions for a maximum of the log-likelihood about θ_0:

$$\begin{aligned} 0 &= S(\hat{\theta}_{mle}|\mathbf{x}) = S(\theta_0|\mathbf{x}) + H(\bar{\theta}|\mathbf{x})(\hat{\theta}_{mle} - \theta_0), \\ \bar{\theta}_i &= \lambda_i \hat{\theta}_{mle,i} + (1 - \lambda_i)\theta_{0,i} \end{aligned}$$

Re-arranging gives

$$H(\bar{\theta}|\mathbf{x})(\hat{\theta}_{mle} - \theta_0) = -S(\theta_0|\mathbf{x})$$

$$\Rightarrow \sqrt{n}(\hat{\theta}_{mle} - \theta_0) = -\left(\frac{1}{n}H(\bar{\theta}|\mathbf{x})\right)^{-1}\left(\frac{1}{\sqrt{n}}S(\theta_0|\mathbf{x})\right)$$

Now

$$\frac{1}{n}H(\bar{\theta}|\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} H(\bar{\theta}|x_i) \xrightarrow{p} E[H(\theta_0|x_i)] = -I(\theta_0|x_i)$$

assuming $\frac{1}{n}H(\theta|\mathbf{x})$ converges in probability to $E[H(\theta|x_i)]$ uniformly in θ.

Furthermore, assume that

$$\frac{1}{\sqrt{n}}S(\theta_0|\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S(\theta_0|x_i) \xrightarrow{d} N(0, I(\theta_0|x_i))$$

which will occur if $\{S(\theta_0|x_i)\}$ is an ergodic-stationary process with

$E[S(\theta_0|x_i)S(\theta_0|x_i)'] = I(\theta_0|x_i).$

Then

$$\begin{aligned} & \sqrt{n}(\hat{\theta}_{mle} - \theta_0) \xrightarrow{d} I(\theta_0|x_i)^{-1} N(0, I(\theta_0|x_i)) \\ = \quad & N(0, I(\theta_0|x_i)^{-1}) \end{aligned}$$

Alternatively

$$\hat{\theta}_{mle} \overset{A}{\sim} N\left(\theta, I(\theta_0|\mathbf{x})^{-1}\right)$$

where

$$I(\theta_0|\mathbf{x}) = nI(\theta_0|x_i)$$

Remark

Since $I(\theta|x_i) = -E[H(\theta|x_i)] = \text{var}(S(\theta|x_i))$ is generally not known, it must be estimated. The most common estimates for $I(\theta|x_i)$ are

$$\begin{aligned}
\hat{I}_1(\hat{\theta}_{mle}|x_i) &= -\frac{1}{n}\sum_{i=1}^{n} H(\hat{\theta}_{mle}|x_i) \xrightarrow{p} -E\left[H(\theta_0|x_i)\right] = I(\theta_0|x_i) \\
\hat{I}_2(\hat{\theta}_{mle}|x_i) &= \frac{1}{n}\sum_{i=1}^{n} S(\hat{\theta}_{mle}|x_i)S(\hat{\theta}_{mle}|x_i)' \xrightarrow{p} E[S(\theta_0|x_i)S(\theta_0|x_i)'] \\
&= I(\theta_0|x_i)
\end{aligned}$$

Then

$$\begin{aligned}
\hat{I}_1(\hat{\theta}_{mle}|\mathbf{x}) &= n\hat{I}_1(\hat{\theta}_{mle}|x_i) = -H(\hat{\theta}_{mle}|\mathbf{x}) \\
\hat{I}_2(\hat{\theta}_{mle}|\mathbf{x}) &= n\hat{I}_2(\hat{\theta}_{mle}|x_i) = \sum_{i=1}^{n} S(\hat{\theta}_{mle}|x_i)S(\hat{\theta}_{mle}|x_i)'
\end{aligned}$$

Example 5 *Asymptotic results for MLE of linear regression model parameters*

In the linear regression with normal errors

$$\begin{aligned} y_i &= x_i'\beta + \varepsilon_i,\ i = 1, \ldots, n \\ & \quad \varepsilon_i | x_i \sim iid\ N(0, \sigma^2) \end{aligned}$$

the MLE for $\theta = (\beta', \sigma^2)'$ is

$$\begin{pmatrix} \hat{\beta}_{mle} \\ \hat{\sigma}^2_{mle} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ n^{-1}(\mathbf{y} - \mathbf{X}\widehat{\beta}_{mle})'(\mathbf{y} - \mathbf{X}\widehat{\beta}_{mle}) \end{pmatrix}$$

and the information matrix for the sample is

$$I(\theta|\mathbf{x}) = \begin{pmatrix} \sigma^{-2}\mathbf{X}'\mathbf{X} & 0 \\ 0 & \frac{n}{2}\sigma^{-4} \end{pmatrix}$$

The asymptotic results for MLE tell us that

$$\theta_{mle} \overset{A}{\sim} N(\theta, I(\theta|\mathbf{x})^{-1}$$

or

$$\begin{pmatrix} \hat{\beta}_{mle} \\ \hat{\sigma}^2_{mle} \end{pmatrix} \overset{A}{\sim} N\left(\begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & 0 \\ 0 & \frac{2}{n}\sigma^4 \end{pmatrix} \right)$$

Further, the block diagonality of the information matrix implies that $\hat{\beta}_{mle}$ is asymptotically independent of $\hat{\sigma}^2_{mle}$.

A natural estimate of $I(\theta|\mathbf{x})$ is

$$\hat{I}(\hat{\theta}_{mle}|\mathbf{x}) = \begin{pmatrix} \hat{\sigma}^{-2}_{mle}\mathbf{X}'\mathbf{X} & 0 \\ 0 & \frac{n}{2}\hat{\sigma}^{-4}_{mle} \end{pmatrix}.$$

Relationship Between ML and GMM

Let $X_1, \ldots, X_n$ be an iid sample from some underlying economic model.

To do ML estimation, you need to know the pdf, $f(x_i|\theta_0)$, of an observation in order to form the log-likelihood function

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

The MLE satisfies the p first order condtions

$$\begin{aligned} 0 &= \frac{\partial \ln L(\hat{\theta}_{mle}|\mathbf{x})}{\partial \theta} = S(\hat{\theta}_{mle}|\mathbf{x}) \\ &= \sum_{i=1}^{n} S(\hat{\theta}_{mle}|x_i) = \sum_{i=1}^{n} \frac{\partial \ln f(x_i; \hat{\theta}_{mle})}{\partial \theta} \end{aligned}$$

Recall, the population moment conditions

$$\begin{aligned} E[S(\theta_0|x_i)] &= 0 \\ E[S(\theta_0|x_i)S(\theta_0|x_i)'] &= I(\theta_0|x_i) \end{aligned}$$

Under regularity conditions we know that

$$\hat{\theta}_{mle} \overset{A}{\sim} N\left(\theta_0, \frac{1}{n} I(\theta_0|x_i)^{-1}\right)$$

$$I(\theta_0|x_i) = -E[H(\theta_0|x_i)] = E[S(\theta_0|x_i)S(\theta_0|x_i)']$$

To do GMM estimation, you need to know $k \geq p$ population moment condtions

$$E[g(x_i, \theta_0)] = 0$$

The GMM estimator matches sample moments with the population moments. The sample moments are

$$g_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(x_i, \theta)$$

If $k > p$, the efficient GMM estimator minimizes the objective function

$$\begin{aligned} J(\theta, \hat{S}^{-1}) &= n g_n(\theta)' \hat{S}^{-1} g_n(\theta) \\ S &= E[g(x_i, \theta) g(x_i, \theta)'] \end{aligned}$$

The first order conditions are

$$\frac{\partial J(\hat{\theta}_{gmm}, S^{-1})}{\partial \theta} = G_n'(\hat{\theta}_{gmm})\hat{S}^{-1}g_n(\hat{\theta}_{gmm}) = 0$$

Under regularity conditions,

$$\hat{\theta}_{gmm} \overset{A}{\sim} N\left(\theta_0, \frac{1}{n}(G'S^{-1}G)^{-1}\right)$$

$$G = E\left[\frac{\partial g(x_i, \theta_0)}{\partial \theta'}\right]$$

The asymptotic efficiency of the MLE in the class of consistent and asymptotically normal estimators implies that

$$\text{avar}(\hat{\theta}_{mle}) - \text{avar}(\hat{\theta}_{gmm}) \leq 0$$

That is, the efficient GMM estimator is generally less efficient than the ML estimator.

The GMM estimator will be equivalent to the ML estimator if the moment conditions happen to correspond with the score associated with the pdf of an observation. That is, if

$$g(x_i, \theta) = S(\theta | x_i)$$

In this case, there are p moment conditions and the model is just identified.

The GMM estimator then satisfies the sample moment equations

$$\begin{aligned} g_n(\hat{\theta}_{gmm}) &= \frac{1}{n} S(\hat{\theta}_{gmm}|\mathbf{x}) = 0 \\ &\Rightarrow \hat{\theta}_{gmm} = \hat{\theta}_{mle} \end{aligned}$$

Furthermore,

$$\begin{aligned} G &= E\left[\frac{\partial S(\theta_0|x_i)}{\partial \theta'}\right] = E[H(\theta_0|x_i)] = -I(\theta_0|x_i) \\ S &= E[S(\theta_0|x_i)S(\theta_0|x_i')] = I(\theta_0|x_i) \end{aligned}$$

Therefore, the asymptotic variance of the GMM estimator is the same as the asymptotic variance of the MLE

$$(G'S^{-1}G)^{-1} = I(\theta_0|x_i)^{-1}$$