Maximum Likelihood Estimation

Eric Zivot

May 14, 2001
This version: November 15, 2009

1 Maximum Likelihood Estimation

1.1 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 $(k \times 1)$ vector of parameters that characterize $f(x_i;\theta)$. For example, if $X_i \tilde{} N(\mu, \sigma^2)$ then $f(x_i;\theta) = (2\pi\sigma^2)^{-1/2}\exp(-\frac{1}{2\sigma^2}(x_i-\mu)^2)$ and $\theta = (\mu, \sigma^2)'$. 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. The joint density[1] 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 functions 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

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

[1] If $X_1, \ldots, X_n$ are discrete random variables, then $f(x_1, \ldots, x_n;\theta) = \Pr(X_1 = x_1, \ldots, X_n = x_n)$ for a fixed value of θ.

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

Example 1 *Bernoulli Sampling*

Let $X_i \tilde{}$ Bernoulli(θ). That is, $X_i = 1$ with probability θ and $X_i = 0$ with probability $1 - \theta$ where $0 \leq \theta \leq 1$. 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

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

For a given value of θ and observed sample x, $f(x; \theta)$ gives the probability of observing the sample. For example, suppose $n = 5$ and $x = (0, \ldots, 0)$. Now some values of θ are more likely to have generated this sample than others. In particular, it is more likely that θ is close to zero than one. To see this, note that the likelihood function for this sample is

$$L(\theta|(0, \ldots, 0)) = (1-\theta)^5$$

This function is illustrated in figure xxx. The likelihood function has a clear maximum at $\theta = 0$. That is, $\theta = 0$ is the value of θ that makes the observed sample $x = (0, \ldots, 0)$ most likely (highest probability)

Similarly, suppose $x = (1, \ldots, 1)$. Then the likelihood function is

$$L(\theta|(1, \ldots, 1)) = \theta^5$$

which is illustrated in figure xxx. Now the likelihood function has a maximum at $\theta = 1$.

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

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

so that $\theta = (\mu, \sigma^2)'$. The likelihood function is given by

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

Figure xxx illustrates the normal likelihood for a representative sample of size $n = 25$. Notice that the likelihood has the same bell-shape of a bivariate normal density

Suppose $\sigma^2 = 1$. Then

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

Now

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \mu)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2 = \sum_{i=1}^{n}\left[(x_i - \bar{x})^2 + 2(x_i - \bar{x})(x - \bar{\mu}) + (\bar{x} - \mu)^2\right] \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\end{aligned}$$

so that

$$L(\mu|\mathbf{x}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]\right)$$

Since both $(x_i - \bar{x})^2$ and $(\bar{x} - \mu)^2$ are positive it is clear that $L(\mu|x)$ is maximized at $\mu = \bar{x}$. This is illustrated in figure xxx.

Example 3 *Linear Regression Model with Normal Errors*

Consider the linear regression

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

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 :

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

where $\theta = (\beta', \sigma^2)'$. Given an iid sample of n observations, $\mathbf{y}$ and $\mathbf{X}$, the joint density of the sample is

$$\begin{aligned}f(\mathbf{y}|\mathbf{X}; \theta) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)\end{aligned}$$

The log-likelihood function is then

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

Example 4 *AR(1) model with Normal Errors*

To be completed

1.2 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 previous example motives an estimator as the value of θ that makes the observed sample most likely. Formally, the maximum likelihood estimator, denoted $\hat{\theta}_{mle}$, is the value of θ that maximizes $L(\theta|\mathbf{x})$. That is, $\hat{\theta}_{mle}$ 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 the value of the θ that maximizes $\ln L(\theta|\mathbf{x})$ will also maximize $L(\theta|\mathbf{x})$. Therefore, we may also define $\hat{\theta}_{mle}$ as the value of θ that solves

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

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

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

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

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

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

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

By definition, the MLE satisfies

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

Under random sampling the score for the sample becomes the sum of the scores for each observation x_i :

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

where $S(\theta|x_i) = \frac{\partial \ln f(x_i;\theta)}{\partial \theta}$ is the score associated with x_i.

Example 5 *Bernoulli example continued*

The log-likelihood function is

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

The score function for the Bernoulli log-likelihood is

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

The MLE satisfies $S(\hat{\theta}_{mle}|\mathbf{x}) = 0$, which after a little algebra, produces the MLE

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

Hence, the sample average is the MLE for θ in the Bernoulli model.

Example 6 *Normal example continued*

Since the normal pdf is regular, we may determine the MLE for $\theta = (\mu, \sigma^2)$ by maximizing the log-likelihood

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

The sample score is a (2×1) vector given by

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

where

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

Note that the score vector for an observation is

$$S(\theta|x_i) = \begin{pmatrix} \frac{\partial \ln f(\theta|x_i)}{\partial \mu} \\ \frac{\partial \ln f(\theta|x_i)}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} (\sigma^2)^{-1}(x_i - \mu) \\ -\frac{1}{2}(\sigma^2)^{-1} + \frac{1}{2}(\sigma^2)^{-2}(x_i - \mu)^2 \end{pmatrix}$$

so that $S(\theta|\mathbf{x}) = \sum_{i=1}^{n} S(\theta|x_i)$.

Solving $S(\hat{\theta}_{mle}|\mathbf{x}) = 0$ gives the *normal equations*

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

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

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

Hence, the sample average is the MLE for μ. Using $\hat{\mu}_{mle} = \bar{x}$ and solving the second equation for $\hat{\sigma}^2_{mle}$ gives

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

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

Example 7 *Linear regression example continued*

The log-likelihood is

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

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

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

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

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

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

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

1.3 Properties of the Score Function

The matrix of second derivatives of the log-likelihood is called the *Hessian*

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

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

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

The last result says that the sample information matrix is equal to n times the information matrix for an observation.

The following proposition relates some properties of the score function to the information matrix.

Proposition 8 *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

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

If θ is a vector then

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

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.

For part 2, consider the scalar case for simplicity. Now, proceeding as above we get

$$\begin{aligned}
E[S(\theta|x_i)^2] &= \int S(\theta|x_i)^2 f(x_i;\theta) dx_i = \int \left(\frac{\partial \ln f(x_i;\theta)}{\partial \theta} \right)^2 f(x_i;\theta) dx_i \\
&= \int \left(\frac{1}{f(x_i;\theta)} \frac{\partial}{\partial \theta} f(x_i;\theta) \right)^2 f(x_i;\theta) dx_i = \int \frac{1}{f(x_i;\theta)} \left(\frac{\partial}{\partial \theta} f(x_i;\theta) \right)^2 dx_i
\end{aligned}$$

Next, recall that $I(\theta|x_i) = -E[H(\theta|x_i)]$ and

$$-E[H(\theta|x_i)] = -\int \frac{\partial^2 \ln f(x_i;\theta)}{\partial\theta^2} f(x_i;\theta)dx_i$$

Now, by the chain rule

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

Then

$$\begin{aligned}-E[H(\theta|x_i)] &= -\int\left[-f(x_i;\theta)^{-2}\left(\frac{\partial}{\partial\theta}f(x_i;\theta)\right)^2 + f(x_i;\theta)^{-1}\frac{\partial^2}{\partial\theta^2}f(x_i;\theta)\right]f(x_i;\theta)dx_i \\ &= \int f(x_i;\theta)^{-1}\left(\frac{\partial}{\partial\theta}f(x_i;\theta)\right)^2 dx_i - \int\frac{\partial^2}{\partial\theta^2}f(x_i;\theta)dx_i \\ &= E[S(\theta|x_i)^2] - \frac{\partial^2}{\partial\theta^2}\int f(x_i;\theta)dx_i \\ &= E[S(\theta|x_i)^2].\end{aligned}$$

1.4 Concentrating the Likelihood Function

In many situations, our interest may be only on a few elements of θ. Let $\theta = (\theta_1, \theta_2)$ and suppose θ_1 is the parameter of interest and θ_2 is a *nuisance parameter* (parameter not of interest). In this situation, it is often convenient to *concentrate* out the nuisance parameter θ_2 from the log-likelihood function leaving a *concentrated log-likelihood* function that is only a function of the parameter of interest θ_1.

To illustrate, consider the example of iid sampling from a normal distribution. Suppose the parameter of interest is μ and the nuisance parameter is σ^2. We wish to concentrate the log-likelihood with respect to σ^2 leaving a concentrated log-likelihood function for μ. We do this as follows. From the score function for σ^2 we have the first order condition

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

Solving for σ^2 as a function of μ gives

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

Notice that any value of $\sigma^2(\mu)$ defined this way satisfies the first order condition $\frac{\partial \ln L(\theta|\mathbf{x})}{\partial \sigma^2} = 0$. If we substitute $\sigma^2(\mu)$ for σ^2 in the log-likelihood function for θ we get the following concentrated log-likelihood function for μ :

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

Now we may determine the MLE for μ by maximizing the concentrated log-likelihood function $\ln L^2(\mu|\mathbf{x})$. The first order conditions are

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

which is satisfied by $\hat{\mu}_{mle} = \bar{x}$ provided not all of the x_i values are identical.

For some models it may not be possible to analytically concentrate the log-likelihood with respect to a subset of parameters. Nonetheless, it is still possible in principle to numerically concentrate the log-likelihood.

1.5 The Precision of the Maximum Likelihood Estimator

The likelihood, log-likelihood and score functions for a typical model are illustrated in figure xxx. The likelihood function is always positive (since it is the joint density of the sample) but the log-likelihood function is typically negative (being the log of a number less than 1). Here the log-likelihood is globally concave and has a unique maximum at $\hat{\theta}_{mle}$. Consequently, the score function is positive to the left of the maximum, crosses zero at the maximum and becomes negative to the right of the maximum.

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. On the other hand, 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.

The extreme case of a completely flat likelihood in θ is illustrated in figure xxx. Here, 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*. Formally, θ is identified if for all $\theta_1 \neq \theta_2$ there exists a sample $\mathbf{x}$ for which $L(\theta_1|\mathbf{x}) \neq L(\theta_2|\mathbf{x})$.

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.

1.5.1 The Cramer-Rao Lower Bound

If we restrict ourselves to the class of unbiased estimators (linear and nonlinear) then we define the best estimator as the one with the smallest variance. With linear estimators, the Gauss-Markov theorem tells us that the ordinary least squares (OLS) estimator is best (BLUE). When we expand the class of estimators to include linear and nonlinear estimators it turns out that we can establish an absolute lower bound on the variance of any unbiased estimator $\hat{\theta}$ of θ under certain conditions. Then if an unbiased estimator $\hat{\theta}$ has a variance that is equal to the lower bound then we have found the best unbiased estimator (BUE).

Theorem 9 *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

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

where $I(\theta|\mathbf{x}) = -E[H(\theta|\mathbf{x})]$ denotes the sample information matrix. Hence, the *Cramer-Rao Lower Bound* (CRLB) is the inverse of the information matrix. 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.

Example 10 *Bernoulli model continued*

To determine the CRLB the information matrix must be evaluated. The information matrix may be computed as

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

or

$$I(\theta|\mathbf{x}) = \text{var}(S(\theta|\mathbf{x}))$$

Further, due to random sampling $I(\theta|\mathbf{x}) = n \cdot I(\theta|x_i) = n \cdot \text{var}(S(\theta|x_i))$. Now, using the chain rule it can be shown that

$$\begin{aligned} H(\theta|x_i) &= \frac{d}{d\theta} S(\theta|x_i) = \frac{d}{d\theta}\left(\frac{x_i - \theta}{\theta(1-\theta)}\right) \\ &= -\left(\frac{1 + S(\theta|x_i) - 2\theta S(\theta|x_i)}{\theta(1-\theta)}\right) \end{aligned}$$

The information for an observation is then

$$\begin{aligned} I(\theta|x_i) &= -E[H(\theta|x_i)] = \frac{1 + E[S(\theta|x_i)] - 2\theta E[S(\theta|x_i)]}{\theta(1-\theta)} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

since

$$E[S(\theta|x_i)] = \frac{E[x_i] - \theta}{\theta(1-\theta)} = \frac{\theta - \theta}{\theta(1-\theta)} = 0$$

The information for an observation may also be computed as

$$\begin{aligned} I(\theta|x_i) &= \text{var}(S(\theta|x_i) = \text{var}\left(\frac{x_i - \theta}{\theta(1-\theta)}\right) \\ &= \frac{\text{var}(x_i)}{\theta^2(1-\theta)^2} = \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

The information for the sample is then

$$I(\theta|\mathbf{x}) = n \cdot I(\theta|x_i) = \frac{n}{\theta(1-\theta)}$$

and the CRLB is

$$CRLB = I(\theta|\mathbf{x})^{-1} = \frac{\theta(1-\theta)}{n}$$

This the lower bound on the variance of any unbiased estimator of θ.

Consider the MLE for θ, $\hat{\theta}_{mle} = \bar{x}$. Now,

$$\begin{aligned} E[\hat{\theta}_{mle}] &= E[\bar{x}] = \theta \\ \text{var}(\hat{\theta}_{mle}) &= \text{var}(\bar{x}) = \frac{\theta(1-\theta)}{n} \end{aligned}$$

Notice that the MLE is unbiased and its variance is equal to the CRLB. Therefore, $\hat{\theta}_{mle}$ is *efficient*.

Remarks

- If $\theta = 0$ or $\theta = 1$ then $I(\theta|\mathbf{x}) = \infty$ and $\text{var}(\hat{\theta}_{mle}) = 0$ (why?)
- $I(\theta|\mathbf{x})$ is smallest when $\theta = \frac{1}{2}$.
- As $n \to \infty$, $I(\theta|\mathbf{x}) \to \infty$ so that $\text{var}(\hat{\theta}_{mle}) \to 0$ which suggests that $\hat{\theta}_{mle}$ is consistent for θ.

Example 11 *Normal model continued*

The Hessian for an observation is

$$H(\theta|x_i) = \frac{\partial^2 \ln f(x_i;\theta)}{\partial\theta\partial\theta'} = \frac{\partial S(\theta|x_i)}{\partial\theta'} = \begin{pmatrix} \frac{\partial^2 \ln f(x_i;\theta)}{\partial\mu^2} & \frac{\partial^2 \ln f(x_i;\theta)}{\partial\mu\partial\sigma^2} \\ \frac{\partial^2 \ln f(x_i;\theta)}{\partial\sigma^2\partial\mu} & \frac{\partial^2 \ln f(x_i;\theta)}{\partial(\sigma^2)^2} \end{pmatrix}$$

Now

$$\begin{aligned} \frac{\partial^2 \ln f(x_i;\theta)}{\partial\mu^2} &= -(\sigma^2)^{-1} \\ \frac{\partial^2 \ln f(x_i;\theta)}{\partial\mu\partial\sigma^2} &= -(\sigma^2)(x_i - \mu) \\ \frac{\partial^2 \ln f(x_i;\theta)}{\partial\sigma^2\partial\mu} &= -(\sigma^2)(x_i - \mu) \\ \frac{\partial^2 \ln f(x_i;\theta)}{\partial(\sigma^2)^2} &= \frac{1}{2}(\sigma^2)^{-2} - (\sigma^2)^{-3}(x_i - \mu)^2 \end{aligned}$$

so that

$$\begin{aligned} I(\theta|x_i) &= -E[H(\theta|x_i)] \\ &= \begin{pmatrix} (\sigma^2)^{-1} & E[(x_i - \mu)](\sigma^2)^{-2} \\ E[(x_i - \mu)](\sigma^2)^{-2} & \frac{1}{2}(\sigma^2)^{-2} - (\sigma^2)^{-3}E[(x_i - \mu)^2] \end{pmatrix} \end{aligned}$$

Using the results[2]

$$\begin{aligned} E[(x_i - \mu)] &= 0 \\ E\left[\frac{(x_i - \mu)^2}{\sigma^2}\right] &= 1 \end{aligned}$$

we then have

$$I(\theta|x_i) = \begin{pmatrix} (\sigma^2)^{-1} & 0 \\ 0 & \frac{1}{2}(\sigma^2)^{-2} \end{pmatrix}$$

The information matrix for the sample is then

$$I(\theta|\mathbf{x}) = n \cdot I(\theta|x_i) = \begin{pmatrix} n(\sigma^2)^{-1} & 0 \\ 0 & \frac{n}{2}(\sigma^2)^{-2} \end{pmatrix}$$

[2] $(x_i - \mu)^2/\sigma^2$ is a chi-square random variable with one degree of freedom. The expected value of a chi-square random variable is equal to its degrees of freedom.

and the CRLB is

$$CRLB = I(\theta|\mathbf{x})^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$$

Notice that the information matrix and the CRLB are diagonal matrices. The CRLB for an unbiased estimator of μ is $\frac{\sigma^2}{n}$ and the CRLB for an unbiased estimator of σ^2 is $\frac{2\sigma^4}{n}$.

The MLEs for μ and σ^2 are

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

Now

$$\begin{aligned} E[\hat{\mu}_{mle}] &= \mu \\ E[\hat{\sigma}^2_{mle}] &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

so that $\hat{\mu}_{mle}$ is unbiased whereas $\hat{\sigma}^2_{mle}$ is biased. This illustrates the fact that mles are not necessarily unbiased. Furthermore,

$$\text{var}(\hat{\mu}_{mle}) = \frac{\sigma^2}{n} = CRLB$$

and so $\hat{\mu}_{mle}$ is efficient.

The MLE for σ^2 is biased and so the CRLB result does not apply. Consider the unbiased estimator of σ^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Is the variance of s^2 equal to the CRLB? No. To see this, recall that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

Further, if $X \sim \chi^2(n-1)$ then $E[X] = n-1$ and $\text{var}(X) = 2(n-1)$. Therefore,

$$\begin{aligned} s^2 &= \frac{\sigma^2}{(n-1)}X \\ &\Rightarrow \text{var}(s^2) = \frac{\sigma^4}{(n-1)^2}\text{var}(X) = \frac{\sigma^4}{(n-1)} \end{aligned}$$

Hence, $\text{var}(s^2) = \frac{\sigma^4}{(n-1)} > CRLB = \frac{\sigma^4}{n}$.

Remarks

- The diagonal elements of $I(\theta|\mathbf{x}) \to \infty$ as $n \to \infty$
- $I(\theta|\mathbf{x})$ only depends on σ^2

Example 12 *Linear regression model continued*

The score vector is given by

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

where $\varepsilon = y - X\beta$. Now $E[\varepsilon] = 0$ and $E[\varepsilon'\varepsilon] = n\sigma^2$ (since $\varepsilon'\varepsilon/\sigma^2 \sim \chi^2(n)$) so that

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

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

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

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, $\hat{\beta}_{mle}$ is unbiased and $\text{var}(\beta_{mle}|X) = \sigma^2(X'X)^{-1} = CRLB$ for an unbiased estimator for β. Hence, $\hat{\beta}_{mle}$ is the most efficient unbiased estimator (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. What about the unbiased estimator $s^2 = (n-k)^{-1}(y - X\hat{\beta}_{OLS})'(y - X\hat{\beta}_{OLS})$? It can be shown that $\text{var}(s^2|X) = \frac{2\sigma^4}{n-k} > \frac{2}{n}\sigma^4 = CRLB$ for an unbiased estimator of σ^2. Hence s^2 is not the most efficient unbiased estimator of σ^2.

1.6 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 13 *Normal Model Continued*

The log-likelihood is parameterized in terms of μ and σ^2 and we have the MLEs

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

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

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

1.7 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 (see Hayashi Chapter 7), 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

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

That is,

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

Alternatively,

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

where $I(\theta|\mathbf{x}) = nI(\theta|x_i) =$ information matrix for the sample.

3. $\hat{\theta}_{mle}$ is efficient in the class of consistent and asymptotically normal estimators. That is,

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

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

Remarks:

1. The consistency of the MLE requires the following

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

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

2. Asymptotic normality of θ_{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:

$$\begin{aligned} 0 &= S(\hat{\theta}_{mle}|\mathbf{x}) = S(\theta_0) + H(\bar{\theta}|\mathbf{x})(\hat{\theta}_{mle} - \theta_0),\ \bar{\theta} = \lambda\hat{\theta}_{mle} + (1-\lambda)\theta_0 \\ &\Rightarrow H(\bar{\theta}|\mathbf{x})(\hat{\theta}_{mle} - \theta_0) = -S(\theta_0) \\ &\Rightarrow \sqrt{n}(\hat{\theta}_{mle} - \theta_0) = -\left(\frac{1}{n}H(\bar{\theta}|\mathbf{x})\right)^{-1}\sqrt{n}\left(\frac{1}{n}S(\theta_0)\right) \end{aligned}$$

Now

$$\begin{aligned} H(\bar{\theta}|\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n} H(\bar{\theta}|x_i) \xrightarrow{p} E[H(\theta_0|x_i)] = -I(\theta_0|x_i) \\ \frac{1}{\sqrt{n}}S(\theta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S(\theta_0|x_i) \xrightarrow{d} N(0, I(\theta_0|x_i)) \end{aligned}$$

Therefore

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

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

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

The first estimate requires second derivatives of the log-likelihood, whereas the second estimate only requires first derivatives. Also, the second estimate is guaranteed to be positive semi-definite in finite samples. The estimate of $\text{avar}(\sqrt{n}(\hat{\theta}_{mle} - \theta))$ then takes the form

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

To prove consistency of the MLE, one must show that $Q_0(\theta) = E[\ln f(x_i|\theta)]$ is uniquely maximized at $\theta = \theta_0$. To do this, let $f(x, \theta_0)$ denote the true density and let $f(x, \theta_1)$ denote the density evaluated at any $\theta_1 \neq \theta_0$. Define the Kullback-Leibler Information Criteria (KLIC) as

$$K(f(x,\theta_0), f(x,\theta_1)) = E_{\theta_0}\left[\ln \frac{f(x,\theta_0)}{f(x,\theta_1)}\right] = \int \ln \frac{f(x,\theta_0)}{f(x,\theta_1)} f(x,\theta_0) dx$$

where

$$\begin{aligned} \ln \frac{f(x,\theta_0)}{f(x,\theta_1)} &= \infty \text{ if } f(x,\theta_1) = 0 \text{ and } f(x,\theta_0) > 0 \\ K(f(x,\theta_0), f(x,\theta_1)) &= 0 \text{ if } f(x,\theta_0) = 0 \end{aligned}$$

The KLIC is a measure of the ability of the likelihood ratio to distinguish between $f(x, \theta_0)$ and $f(x, \theta_1)$ when $f(x, \theta_0)$ is true. The Shannon-Komogorov Information Inequality gives the following result:

$$K(f(x,\theta_0), f(x,\theta_1)) \geq 0$$

with equality if and only if $f(x, \theta_0) = f(x, \theta_1)$ for all values of x.

Example 14 *Asymptotic results for MLE of Bernoulli distribution parameters*

Let $X_1, \ldots, X_n$ be an iid sample with $X \sim$Bernoulli(θ). Recall,

$$\begin{aligned} \hat{\theta}_{mle} &= \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \\ I(\theta|x_i) &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

The asymptotic properties of the MLE tell us that

$$\begin{aligned} &\hat{\theta}_{mle} \xrightarrow{p} \theta \\ &\sqrt{n}(\hat{\theta}_{mle} - \theta) \xrightarrow{d} N\left(0, \theta(1-\theta)\right) \end{aligned}$$

Alternatively,

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

An estimate of the asymptotic variance of $\hat{\theta}_{mle}$ is

$$\text{avar}(\hat{\theta}_{mle}) = \frac{\hat{\theta}_{mle}(1-\hat{\theta}_{mle})}{n} = \frac{\bar{x}(1-\bar{x})}{n}$$

Example 15 *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

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

and the information matrix for the sample is

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

The asymptotic results for MLE tell us that

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

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

1.8 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, $f(x_i|\theta)$, of an observation in order to form the log-likelihood function

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

where $\theta \in \mathbb{R}^p$. The MLE satisfies the first order condtions

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

For general models, the first order condtions are p nonlinear equations in p unknowns. Under regularity conditions, the MLE is consistent, asymptotically normally distributed, and efficient in the class of asymptotically normal estimators:

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

where $I(\theta|x_i) = -E[H(\theta|x_i)] = E[S(\theta|x_i)S(\theta|x_i)']$.

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

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

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

$$J(\theta, \hat{S}^{-1}) = n g_n(\theta)' \hat{S}^{-1} g_n(\theta)$$

where $S = E[g(x_i, \theta)g(x_i, \theta)']$. 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, the efficient GMM estimator is consistent, asymptotically normally distributed, and efficient in the class of asymptotically normal GMM estimators for a given set of moment conditions:

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

where $G = E\left[\frac{\partial g_n(\theta)}{\partial \theta'}\right]$.

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

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

which implies that $\hat{\theta}_{gmm} = \hat{\theta}_{mle}$. Since

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

the asymptotic variance of the GMM estimator becomes

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

which is the asymptotic variance of the MLE.

1.9 Hypothesis Testing in a Likelihood Framework

Let $X_1, \ldots, X_n$ be iid with pdf $f(x, \theta)$ and assume that θ is a scalar. The hypotheses to be tested are

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

A statistical test is a decision rule based on the observed data to either reject H_0 or not reject H_0.

1.9.1 Likelihood Ratio Statistic

Consider the likelihood ratio

$$\lambda = \frac{L(\theta_0|\mathbf{x})}{L(\hat{\theta}_{mle}|\mathbf{x})} = \frac{L(\theta_0|\mathbf{x})}{\max_\theta L(\theta|\mathbf{x})}$$

which is the ratio of the likelihood evaluated under the null to the likelihood evaluated at the MLE. By construction $0 < \lambda \leq 1$. If $H_0 : \theta = \theta_0$ is true, then we should see

$\lambda \approx 1$; if $H_0 : \theta = \theta_0$ is not true then we should see $\lambda < 1$. The likelihood ratio (LR) statistic is a simple transformation of λ such that the value of LR is large if $H_0 : \theta = \theta_0$ is true, and the value of LR is small when $H_0 : \theta = \theta_0$ is not true. Formally, the LR statistic is

$$\begin{aligned} LR &= -2\ln\lambda = -2\ln\frac{L(\theta_0|\mathbf{x})}{L(\hat{\theta}_{mle}|\mathbf{x})} \\ &= -2[\ln L(\theta_0|\mathbf{x}) - \ln L(\hat{\theta}_{mle}|\mathbf{x})] \end{aligned}$$

From Figure xxx, notice that the distance between $\ln L(\hat{\theta}_{mle}|\mathbf{x})$ and $\ln L(\theta_0|\mathbf{x})$ depends on the curvature of $\ln L(\theta|\mathbf{x})$ near $\theta = \hat{\theta}_{mle}$. If the curvature is sharpe (i.e., information is high) then LR will be large for θ_0 values away from $\hat{\theta}_{mle}$. If, however, the curvature of $\ln L(\theta|\mathbf{x})$ is flat (i.e., information is low) the LR will be small for θ_0 values away from $\hat{\theta}_{mle}$.

Under general regularity conditions, if $H_0 : \theta = \theta_0$ is true then

$$LR \xrightarrow{d} \chi^2(1)$$

In general, the degrees of freedom of the chi-square limiting distribution depends on the number of restrictions imposed under the null hypothesis. The decision rule for the LR statistic is to reject $H_0 : \theta = \theta_0$ at the $\alpha \times 100\%$ level if $LR > \chi^2_{1-\alpha}(1)$, where $\chi^2_{1-\alpha}(1)$ is the $(1-\alpha) \times 100\%$ quantile of the chi-square distribution with 1 degree of freedom.

1.9.2 Wald Statistic

The Wald statistic is based directly on the asymptotic normal distribution of $\hat{\theta}_{mle}$:

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

where $\hat{I}(\hat{\theta}_{mle}|\mathbf{x})$ is a consistent estimate of the sample information matrix. An implication of the asymptotic normality result is that the usualy t-ratio for testing $H_0 : \theta = \theta_0$

$$\begin{aligned} t &= \frac{\hat{\theta}_{mle} - \theta_0}{\widehat{SE}(\hat{\theta}_{mle})} = \frac{\hat{\theta}_{mle} - \theta_0}{\sqrt{\hat{I}(\hat{\theta}_{mle}|\mathbf{x})^{-1}}} \\ &= \left(\hat{\theta}_{mle} - \theta_0\right)\sqrt{\hat{I}(\hat{\theta}_{mle}|\mathbf{x})} \end{aligned}$$

is asymptotically distributed as a standard normal random variable. Using the continuous mapping theorem, it follows that the square of the t-statistic is asymptotically distributed as a chi-square random variable with 1 degree of freedom. The Wald

statistic is defined to be simply the square of this t-ratio

$$\begin{aligned} Wald &= \frac{\left(\hat{\theta}_{mle} - \theta_0\right)^2}{\hat{I}(\hat{\theta}_{mle}|\mathbf{x})^{-1}} \\ &= \left(\hat{\theta}_{mle} - \theta_0\right)^2 \hat{I}(\hat{\theta}_{mle}|\mathbf{x}) \end{aligned}$$

Under general regularity conditions, if $H_0 : \theta = \theta_0$ is true, then

$$Wald \xrightarrow{d} \chi^2(1)$$

The intuition behind the Wald statistic is illustrated in Figure xxx. If the curvature of $\ln L(\theta|\mathbf{x})$ near $\theta = \hat{\theta}_{mle}$ is big (high information) then the squared distance $\left(\hat{\theta}_{mle} - \theta_0\right)^2$ gets blown up when constructing the Wald statistic. If the curvature of $\ln L(\theta|\mathbf{x})$ near $\theta = \hat{\theta}_{mle}$ is low, then $\hat{I}(\hat{\theta}_{mle}|x)$ is small and the squared distance $\left(\hat{\theta}_{mle} - \theta_0\right)^2$ gets attenuated when constucting the Wald statistic.

1.9.3 Lagrange Multiplier/Score Statistic

With ML estimation, $\hat{\theta}_{mle}$ solves the first order conditions

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

If $H_0 : \theta = \theta_0$ is true, then we should expect that

$$0 \approx \frac{d \ln L(\theta_0|\mathbf{x})}{d\theta} = S(\theta_0|\mathbf{x})$$

If $H_0 : \theta = \theta_0$ is not true, then we should expect that

$$0 \neq \frac{d \ln L(\theta_0|\mathbf{x})}{d\theta} = S(\theta_0|\mathbf{x})$$

The Lagrange multiplier (score) statistic is based on how far $S(\theta_0|\mathbf{x})$ is from zero.

Recall the following properties of the score $S(\theta|x_i)$. If $H_0 : \theta = \theta_0$ is true then

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

Further, it can be shown that

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

so that

$$S(\theta_0|\mathbf{x}) \sim N(0, I(\theta_0|\mathbf{x}))$$

This result motivates the statistic

$$LM = \frac{S(\theta_0|\mathbf{x})^2}{I(\theta_0|\mathbf{x})} = S(\theta_0|\mathbf{x})^2 I(\theta_0|\mathbf{x})^{-1}$$

Under general regularity conditions, if $H_0 : \theta = \theta_0$ is true, then

$$LM \xrightarrow{d} \chi^2(1)$$