Topic 14: Maximum Likelihood Estimation

November, 2009

As before, we begin with a sample $X = (X_1, \dots, X_n)$ of random variables chosen according to one of a family of probabilities P_θ.

In addition, $\mathbf{f}(\mathbf{x}|\theta)$, $\mathbf{x} = (x_1, \dots, x_n)$ will be used to denote the density function for the data when θ is the true state of nature.

Definition 1. *The* **likelihood function** *is the density function regarded as a function of θ.*

$$\mathbf{L}(\theta|\mathbf{x}) = \mathbf{f}(\mathbf{x}|\theta),\ \theta \in \Theta. \tag{1}$$

The **maximum likelihood estimator (MLE)***,*

$$\hat{\theta}(\mathbf{x}) = \arg\max_\theta \mathbf{L}(\theta|\mathbf{x}). \tag{2}$$

Note that if $\hat{\theta}(\mathbf{x})$ is a maximum likelihood estimator for θ, then $g(\hat{\theta}(\mathbf{x}))$ is a maximum likelihood estimator for $g(\theta)$. For example, if θ is a parameter for the variance and $\hat{\theta}$ is the maximum likelihood estimator, then $\sqrt{\hat{\theta}}$ is the maximum likelihood estimator for the standard deviation. This flexibility in estimation criterion seen here is not available in the case of unbiased estimators.

Typically, maximizing the score function $\ln \mathbf{L}(\theta|\mathbf{x})$ will be easier.

1 Examples

Example 2 (Bernoulli trials)**.** *If the experiment consists of n Bernoulli trial with success probability θ, then*

$$\mathbf{L}(\theta|\mathbf{x}) = \theta^{x_1}(1-\theta)^{(1-x_1)} \cdots \theta^{x_n}(1-\theta)^{(1-x_n)} = \theta^{(x_1+\cdots+x_n)}(1-\theta)^{n-(x_1+\cdots+x_n)}.$$

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

$$\frac{\partial}{\partial\theta}\ln \mathbf{L}(\theta|\mathbf{x}) = n\left(\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta}\right).$$

This equals zero when $\theta = \bar{x}$. Check that this is a maximum. Thus,

$$\hat{\theta}(\mathbf{x}) = \bar{x}.$$

Example 3 (Normal data)**.** *Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n normal random variables,*

$$\mathbf{L}(\mu, \sigma^2|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_1-\mu)^2}{2\sigma^2}\right)\cdots\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_n-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2.$$

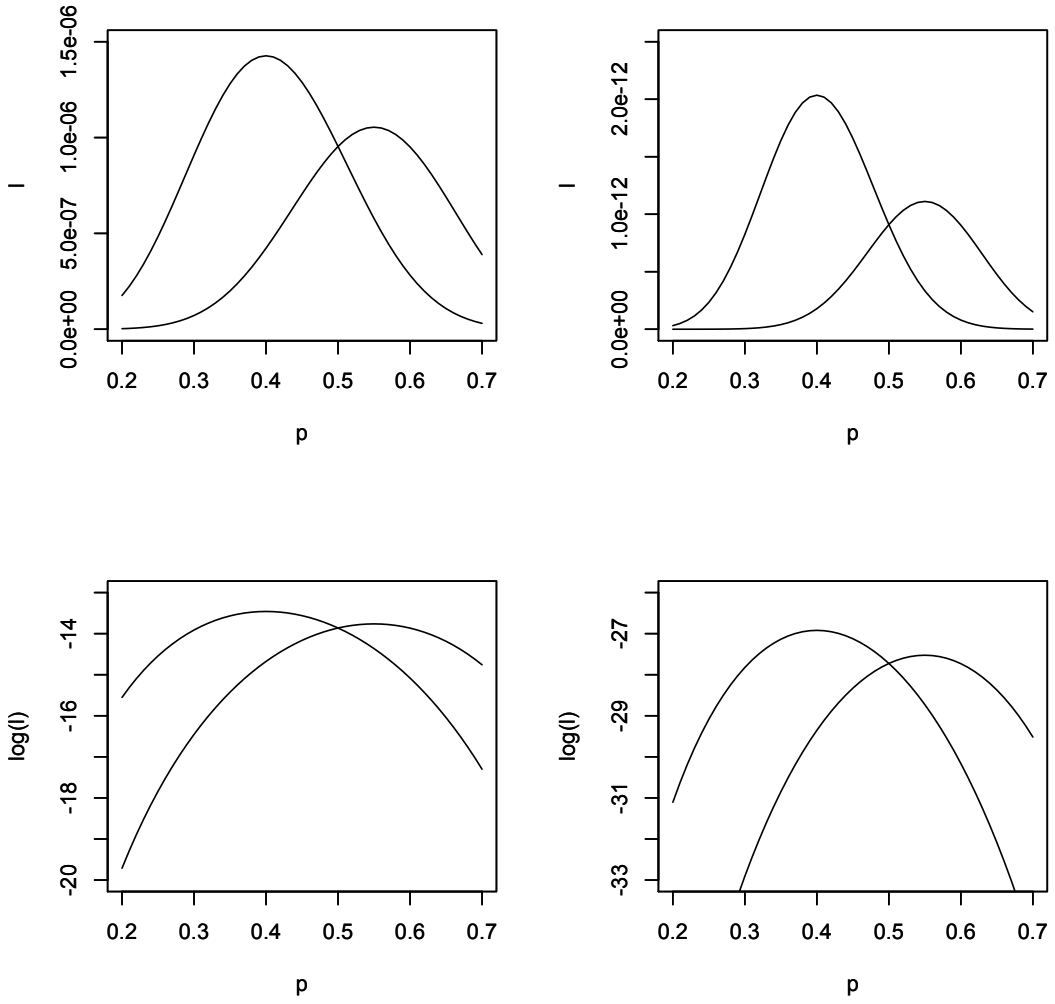

Figure 1: Likelihood function (top row) and its logarithm, the score function, (bottom row) for Bernouli trials. The left column is based on 20 trials having 8 and 11 successes. The right column is based on 40 trials having 16 and 22 successes. Notice that the maximum likelihood is approximately 10^{-6} for 20 trials and 10^{-12} for 40. Note that the peaks are more narrow for 40 trials rather than 20.

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

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

Because the second partial derivative with respect to μ is negative,

$$\hat{\mu}(\mathbf{x}) = \bar{x}$$

is the maximum likelihood estimator.

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

Recalling that $\hat{\mu}(\mathbf{x}) = \bar{x}$, we obtain

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

Note that the maximum likelihood estimator is a biased estimator.

Example 4 (Linear regression)**.** *Our data is n observations with one explanatory variable and one response variable. The model is that*

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where the ϵ_i are independent mean 0 normal random variable. The (unknown) variance is σ^2. The likelihood function

$$\mathbf{L}(\alpha, \beta, \sigma^2|\mathbf{y}, \mathbf{x}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

$$\ln \mathbf{L}(\alpha, \beta, \sigma^2|\mathbf{y}, \mathbf{x}) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

This, the maximum likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$ also the least square estimator. The predicted value for the response variable

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i.$$

The maximum likelihood estimator for σ^2 is

$$\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{k=1}^{n} (y_i - \hat{y}_i)^2.$$

The **unbiased estimator** *is*

$$\hat{\sigma}^2_U = \frac{1}{n-2} \sum_{k=1}^{n} (y_i - \hat{y}_i)^2.$$

For the measurements on the lengths in centimeters of the femur and humerus for the five specimens of Archeopteryx*, we have the following R output for linear regression.*

```
> femur<-c(38,56,59,64,74)
> humerus<-c(41,63,70,72,84)
> summary(lm(humerus~femur))

Call:
```

```
lm(formula = humerus ~ femur)

Residuals:
      1       2       3       4       5
-0.8226 -0.3668  3.0425 -0.9420 -0.9110

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.65959    4.45896  -0.821 0.471944
femur        1.19690    0.07509  15.941 0.000537 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.982 on 3 degrees of freedom
Multiple R-squared: 0.9883,Adjusted R-squared: 0.9844
F-statistic: 254.1 on 1 and 3 DF,  p-value: 0.0005368
```

The residual standard error of 1.982 centimeters is obtained by squaring the 5 residuals, dividing by $3 = 5 - 2$ *and taking a square root.*

Example 5 (Uniform random variables)**.** *If our data* $X = (X_1, \ldots, X_n)$ *are a simple random sample drawn from uniformly distributed random variable whose maximum value* θ *is unknown, then each random variable has density*

$$f(x|\theta) = \begin{cases} 1/\theta & \text{if } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the likelihood

$$\mathbf{L}(\theta|\mathbf{x}) = \begin{cases} 1/\theta^n & \text{if, for all } i,\ 0 \le x_i \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, to maximize $\mathbf{L}(\theta|\mathbf{x})$*, we should minimize the value of* θ^n *in the first alternative for the likelihood. This is achieved by taking*

$$\hat{\theta}(\mathbf{x}) = \max_{1 \le i \le n} x_i.$$

However,

$$\hat{\theta}(X) = \max_{1 \le i \le n} X_i < \theta$$

and the maximum likelihood estimator is biased.

For $0 \le x \le \theta$*, the distribution of* $X_{(n)} = \max_{1 \le i \le n} X_i$ *is*

$$F_{(n)}(x) = P\{\max_{1 \le i \le n} X_i \le x\} = P\{X_1 \le x\}^n = (x/\theta)^n.$$

Thus, the density

$$f_{(n)}(x) = \frac{nx^{n-1}}{\theta^n}.$$

The mean

$$E_\theta X_{(n)} = \frac{n}{n+1}\theta.$$

and thus

$$d(X) = \frac{n+1}{n} X_{(n)}$$

is an unbiased estimator of θ*.*

2 Asymptotic Properties

Much of the attraction of maximum likelihood estimators is based on their properties for large sample sizes.

1. **Consistency** If θ_0 is the state of nature, then

   $$\mathbf{L}(\theta_0|X) > \mathbf{L}(\theta|X)$$

   if and only if

   $$\frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(X_i|\theta_0)}{f(X_i|\theta)} > 0.$$

   By the strong law of large numbers, this sum converges to

   $$E_{\theta_0}\left[\ln \frac{f(X_1|\theta_0)}{f(X_1|\theta)}\right].$$

   which is greater than 0. From this, we obtain

   $$\hat{\theta}(X) \to \theta_0 \quad \text{as } n \to \infty.$$

   We call this property of the estimator **consistency**.

2. **Asymptotic normality and efficiency** Under some assumptions that is meant to insure some regularity, a central limit theorem holds. Here we have

   $$\sqrt{n}(\hat{\theta}(X) - \theta_0)$$

   converges in distribution as $n \to \infty$ to a normal random variable with mean 0 and variance $1/I(\theta_0)$, the Fisher information for one observation. Thus

   $$\text{Var}_{\theta_0}(\hat{\theta}(X)) \approx \frac{1}{nI(\theta_0)},$$

   the lowest possible under the Crámer-Rao lower bound. This property is called **asymptotic efficiency**.

3. **Properties of the log likelihood surface**. For large sample sizes, the variance of an MLE of a single unknown parameter is approximately the negative of the reciprocal of the the Fisher information

   $$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} \ln L(\theta|X)\right].$$

   Thus, the estimate of the variance given data $\mathbf{x}$

   $$\hat{\sigma}^2 = -1 \Big/ \frac{\partial^2}{\partial\theta^2} \ln L(\hat{\theta}|\mathbf{x}).$$

   the negative reciprocal of the second derivative, also known as the curvature, of the log-likelihood function evaluated at the MLE.

   If the curvature is small, then the likelihood surface is flat around its maximum value (the MLE). If the curvature is large and thus the variance is small, the likelihood is strongly curved at the maximum.

   For a multidimensional parameter space $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, Fisher information $I(\theta)$ is a matrix, the ij-th entry is

   $$I(\theta_i, \theta_j) = E_\theta\left[\frac{\partial}{\partial\theta_i} \ln L(\theta|X) \frac{\partial}{\partial\theta_j} \ln L(\theta|X)\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln L(\theta|X)\right]$$

Example 6. *To obtain the maximum likelihood estimate for the gamma family of random variables, write*

$$\mathbf{L}(\alpha,\beta|\mathbf{x}) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}x_1^{\alpha-1}e^{-\beta x_1}\right)\cdots\left(\frac{\beta^\alpha}{\Gamma(\alpha)}x_n^{\alpha-1}e^{-\beta x_n}\right).$$

$$\ln \mathbf{L}(\alpha,\beta|\mathbf{x}) = n(\alpha\ln\beta - \ln\Gamma(\alpha)) + (\alpha-1)\sum_{i=1}^n \ln x_i - \beta\sum_{i=1}^n x_i.$$

To determine the parameters that maximize the likelihood, solve the equations

$$\frac{\partial}{\partial\alpha}\ln \mathbf{L}(\hat\alpha,\hat\beta|\mathbf{x}) = n(\ln\hat\beta - \frac{d}{d\alpha}\ln\Gamma(\hat\alpha)) + \sum_{i=1}^n \ln x_i = 0, \quad \overline{\ln x} = \frac{d}{d\alpha}\ln\Gamma(\hat\alpha) - \ln\hat\beta$$

and

$$\frac{\partial}{\partial\beta}\ln \mathbf{L}(\hat\alpha,\hat\beta|\mathbf{x}) = n\frac{\hat\alpha}{\hat\beta} - \sum_{i=1}^n x_i = 0, \quad \bar x = \frac{\hat\alpha}{\hat\beta}.$$

To compute the Fisher information matrix note that

$$I(\alpha,\beta)_{11} = -\frac{\partial^2}{\partial\alpha^2}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}) = n\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha), \quad I(\alpha,\beta)_{22} = -\frac{\partial^2}{\partial\beta^2}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}) = n\frac{\alpha}{\beta^2},$$

$$I(\alpha,\beta)_{12} = -\frac{\partial^2}{\partial\alpha\partial\beta}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}) = -n\frac{1}{\beta}.$$

This give a Fisher information matrix

$$I(\alpha,\beta) = n\begin{pmatrix} \frac{d^2}{d\alpha^2}\ln\Gamma(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2}\end{pmatrix}.$$

The inverse

$$I(\alpha,\beta)^{-1} = \frac{1}{n\alpha(\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha)-1)}\begin{pmatrix}\alpha & \beta \\ \beta & \beta^2\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha)\end{pmatrix}.$$

For the example for the distribution of fitness effects $\alpha = 0.23$ *and* $\beta = 5.35$ *and* $n = 100$*, and*

$$I(0.23,5.35)^{-1} = \frac{1}{100(0.23)(19.12804)}\begin{pmatrix}0.23 & 5.35 \\ 5.35 & 5.35^2(20.12804)\end{pmatrix} = \begin{pmatrix}0.0001202 & 0.01216 \\ 0.01216 & 1.3095\end{pmatrix}.$$

$$\mathit{Var}_{(0.23,5.35)}(\hat\alpha) \approx 0.0001202, \quad \mathit{Var}_{(0.23,5.35)}(\hat\beta) \approx 1.3095.$$

Compare this to the empirical values of 0.0662 *and* 2.046 *for the method of moments*