Other Concepts of MLE

Mwebesa Edson

(M Biostat, MSc Stat, BSc Edu-Maths)

Score vector and information matrix

- **SCORE VECTOR**

In the theory of maximum likelihood estimation, the score vector (or simply, the score) is the gradient (i.e., the vector of first derivatives) of the log-likelihood function with respect to the parameters being estimated.

Definition

Definition Let θ be a $K \times 1$ parameter vector describing the distribution of a sample ξ. Let $L(\theta;\xi)$ be the likelihood function of the sample ξ, depending on the parameter θ. Let $l(\theta;\xi)$ be the log-likelihood function

$$l(\theta;\xi) = \ln[L(\theta;\xi)]$$

Then, the $K \times 1$ vector of first derivatives of $l(\theta;\xi)$ with respect to the entries of θ, denoted by

$$\nabla_\theta l(\theta;\xi)$$

is called the score vector.

The symbol ∇ is read nabla and is often used to denote the gradient of a function.

Example Of a Normal Distribution

In the next example, the likelihood depends on a 2×1 parameter. As a consequence, the score is a 2×1 vector.

Example Suppose the sample ξ is a vector of n draws x_1, ..., x_n from a normal distribution with mean μ and variance σ^2. As proved in the lecture on maximum likelihood estimation of the parameters of a normal distribution, the log-likelihood of the sample is

$$l(\mu,\sigma^2;x_1,\ldots,x_n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$

The two parameters (mean and variance) together form a 2×1 vector

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$$

The partial derivative of the log-likelihood with respect to μ is

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2; x_1, \ldots, x_n) = \frac{1}{\sigma^2} \sum_{j=1}^{n} (x_j - \mu)$$

and the partial derivative with respect to the variance σ^2 is

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2; x_1, \ldots, x_n) = -\frac{n}{2\sigma^2} + \left[\frac{1}{2} \sum_{j=1}^{n} (x_j - \mu)^2 \right] \frac{1}{(\sigma^2)^2}$$

The score vector is

$$\nabla_\theta l(\theta; \xi) = \nabla_\theta l(\theta; x_1, \ldots, x_n) = \begin{bmatrix} \frac{1}{\sigma^2} \sum_{j=1}^{n} (x_j - \mu) \\ -\frac{n}{2\sigma^2} + \left[\frac{1}{2} \sum_{j=1}^{n} (x_j - \mu)^2 \right] \frac{1}{(\sigma^2)^2} \end{bmatrix}$$

How the score is used to find the maximum likelihood estimator

The maximum likelihood estimator $\hat{\theta}$ of the parameter θ solves the maximization problem

$$\hat{\theta} = \arg\max_{\theta} l(\theta; \xi)$$

Under some regularity conditions, the solution of this problem can be found by solving the first order condition

$$\nabla_{\theta} l(\theta; \xi) = 0$$

that is, by equating the score vector to 0.

In Nutshell

- The first derivative of the log-likelihood function is called the **Score Function** also known as **Fisher's score function.**
- If we define the score function as

$$\mathbf{u}(\boldsymbol{\theta}) = \frac{\partial \log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}.$$

- Score is a vector of first partial derivatives, one for each element of θ.
- If the log-likelihood is concave, one can find the maximum likelihood estimator by setting the score to zero, i.e. by solving the system of equations:

$$\mathbf{u}(\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

Information Matrix

- The information matrix (also called Fisher information matrix) is the matrix of second cross-moments of the score vector. The later is the vector of first partial derivatives of the log-likelihood function with respect to its parameters.

Definition

Definition Let θ be a $K \times 1$ parameter vector characterizing the distribution of a sample ξ. Let $L(\theta;\xi)$ be the likelihood function of ξ, depending on the parameter θ. Let $l(\theta;\xi)$ be the log-likelihood function

$$l(\theta;\xi) = \ln[L(\theta;\xi)]$$

Denote by

$$\nabla_\theta l(\theta;\xi)$$

the score vector, that is, the $K \times 1$ vector of first derivatives of $l(\theta;\xi)$ with respect to the entries of θ. The information matrix $I(\theta)$ is the $K \times K$ matrix of second cross-moments of the score, defined by

$$I(\theta) = \mathrm{E}_\theta[\nabla_\theta l(\theta;\xi)\nabla_\theta l(\theta;\xi)^\top]$$

where the notation E_θ indicates that the expected value is taken with respect to the probability distribution associated to the parameter θ.

For example, if the sample ξ has a continuous distribution, then the likelihood function is

$$L(\theta;\xi) = f(\xi;\theta)$$

where $f(\xi;\theta)$ is the probability density function of ξ, parametrized by θ, and the information matrix is

$$I(\theta) = \int [\nabla_\theta \ln(f(\xi;\theta))][\nabla_\theta \ln(f(\xi;\theta))]^\top f(\xi;\theta) d\xi$$

The information matrix is the covariance matrix of the score

- Under mild regularity conditions, the expected value of the score is equal to zero:

$$\mathrm{E}_\theta[\nabla_\theta l(\theta;\xi)] = 0$$

- As a consequence,

$$\begin{aligned} I(\theta) &= \mathrm{E}_\theta[\nabla_\theta l(\theta;\xi)\nabla_\theta l(\theta;\xi)^\top] \\ &= \mathrm{E}_\theta[\{\nabla_\theta l(\theta;\xi) - \mathrm{E}_\theta[\nabla_\theta l(\theta;\xi)]\}\{\nabla_\theta l(\theta;\xi) - \mathrm{E}_\theta[\nabla_\theta l(\theta;\xi)]\}^\top] \\ &= \mathrm{Var}_\theta[\nabla_\theta l(\theta;\xi)] \end{aligned}$$

- that is, the information matrix is the covariance matrix of the score.

Information equality

- Under mild regularity conditions, it can be proved that

$$I(\theta) = -\mathrm{E}_{\theta}\left[\nabla^2_{\theta\theta} l(\theta; \xi)\right]$$

where

$$\nabla^2_{\theta\theta} l(\theta; \xi)$$

is the matrix of second-order cross-partial derivatives (so-called Hessian matrix) of the log-likelihood.

- This equality is called information equality.

Information matrix of the normal distribution

- As an example, consider a sample

$$\underline{\xi} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}$$

made up of the realizations of IID normal random variables with parameters μ and σ^2 (mean and variance).

- In this case, the information matrix is

$$I(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{bmatrix}$$

Proof

- The log-likelihood function is

$$l(\mu,\sigma^2;x_1,\ldots,x_n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j-1}^{n}(x_j-\mu)^2$$

as proved in the lecture on maximum likelihood estimation of the parameters of the normal distribution. The score s is a 2×1 vector whose entries are the partial derivatives of the log-likelihood with respect to μ and σ^2:

$$s_1 = \frac{\partial}{\partial\mu}l(\mu,\sigma^2;x_1,\ldots,x_n) = \frac{1}{\sigma^2}\sum_{j-1}^{n}(x_j-\mu)$$

$$s_2 = \frac{\partial}{\partial\sigma^2}l(\mu,\sigma^2;x_1,\ldots,x_n) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{j-1}^{n}(x_j-\mu)^2$$

- The information matrix is

$$I(\mu,\sigma^2) = \mathrm{E}_{\mu,\sigma^2}[ss^\top]$$
$$= \mathrm{E}_{\mu,\sigma^2}\left[\begin{bmatrix} s_1 \\ s_2 \end{bmatrix}\begin{bmatrix} s_1 & s_2 \end{bmatrix}\right]$$
$$= \mathrm{E}_{\mu,\sigma^2}\left[\begin{bmatrix} s_1^2 & s_1 s_2 \\ s_2 s_1 & s_2^2 \end{bmatrix}\right]$$

- We have

$$\mathrm{E}_{\mu,\sigma^2}[s_1^2]$$
$$= \mathrm{E}_{\mu,\sigma^2}\left[\frac{1}{(\sigma^2)^2}\left[\sum_{j=1}^{n}(x_j-\mu)\right]^2\right]$$
$$\boxed{\mathrm{A}} = \frac{1}{(\sigma^2)^2}\mathrm{E}_{\mu,\sigma^2}\left[\sum_{j=1}^{n}(x_j-\mu)^2\right]$$
$$\boxed{\mathrm{B}} = \frac{1}{(\sigma^2)^2}n\sigma^2 = \frac{n}{\sigma^2}$$

where: in step A we have used the fact that

$$\mathrm{E}_{\mu,\sigma^2}[(x_i-\mu)(x_j-\mu)] = \mathrm{E}_{\mu,\sigma^2}[x_i-\mu]\mathrm{E}_{\mu,\sigma^2}[x_j-\mu] = 0$$

for $i \neq j$ because the variables in the sample are independent and have mean equal to μ; in step B we have used the fact that

$$\mathrm{E}_{\mu,\sigma^2}\left[(x_j-\mu)^2\right] = \mathrm{Var}_{\mu,\sigma^2}[x_j] = \sigma^2$$

- Moreover,

$$
\begin{aligned}
& \mathrm{E}_{\mu,\sigma^2}\left[s_2^2\right] \\
&= \mathrm{E}_{\mu,\sigma^2}\left[\left(-\frac{n}{2\sigma^2}+\frac{1}{2\left(\sigma^2\right)^2}\sum_{j-1}^{n}(x_j-\mu)^2\right)^2\right] \\
\boxed{\mathrm{A}} &= \mathrm{E}_{\mu,\sigma^2}\left[\left(\frac{n}{2\sigma^2}\right)^2-\frac{n}{2\left(\sigma^2\right)^3}\sum_{j-1}^{n}(x_j-\mu)^2+\frac{1}{4\left(\sigma^2\right)^4}\sum_{j-1}^{n}\sum_{i-1}^{n}(x_j-\mu)^2(x_i-\mu)^2\right] \\
&= \left(\frac{n}{2\sigma^2}\right)^2-\frac{n}{2\left(\sigma^2\right)^3}\mathrm{E}_{\mu,\sigma^2}\left[\sum_{j-1}^{n}(x_j-\mu)^2\right]+\frac{1}{4\left(\sigma^2\right)^4}\mathrm{E}_{\mu,\sigma^2}\left[\sum_{j-1}^{n}\sum_{i-1}^{n}(x_j-\mu)^2(x_i-\mu)^2\right]
\end{aligned}
$$

$$
\begin{aligned}
\boxed{B} &= \left(\frac{n}{2\sigma^2}\right)^2 - \frac{n}{2\left(\sigma^2\right)^3} n\sigma^2 + \frac{1}{4\left(\sigma^2\right)^4}\left[n3\left(\sigma^2\right)^2 + n(n-1)\left(\sigma^2\right)^2\right] \\
&= \frac{n^2}{4\left(\sigma^2\right)^2} - \frac{n^2}{2\left(\sigma^2\right)^2} + \frac{n^2+2n}{4\left(\sigma^2\right)^2} \\
&= \frac{n^2 - 2n^2 + n^2 + 2n}{4\left(\sigma^2\right)^2} \\
&= \frac{2n}{4\left(\sigma^2\right)^2} \\
&= \frac{n}{2\left(\sigma^2\right)^2}
\end{aligned}
$$

where: in steps A and B we have used the independence of the observations in the sample and in step B we have used the fact that the fourth central moment of the normal distribution is equal to $3\left(\sigma^2\right)^2$. Finally,

$$
\begin{aligned}
&\mathrm{E}_{\mu,\sigma^2}[s_1 s_2] \\
&= \mathrm{E}_{\mu,\sigma^2}\left[\left(\frac{1}{\sigma^2} \sum_{j=1}^{n} (x_j - \mu) \right) \left(-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{j=1}^{n} (x_j - \mu)^2 \right) \right] \\
&= -\frac{n}{2\sigma^2} \frac{1}{\sigma^2} \sum_{j=1}^{n} \mathrm{E}_{\mu,\sigma^2}[x_j - \mu] + \frac{1}{2(\sigma^2)^2} \frac{1}{\sigma^2} \mathrm{E}_{\mu,\sigma^2}\left[\left(\sum_{j=1}^{n} (x_j - \mu) \right) \left(\sum_{j=1}^{n} (x_j - \mu)^2 \right) \right] \\
\boxed{\text{A}} &= \frac{1}{2(\sigma^2)^2} \frac{1}{\sigma^2} \sum_{j=1}^{n} \mathrm{E}_{\mu,\sigma^2}\left[(x_j - \mu)^3 \right] \\
\boxed{\text{B}} &= 0
\end{aligned}
$$

where: in step A we have used the facts that $\mathrm{E}_{\mu,\sigma^2}[x_j] = \mu$ and that

$$\mathrm{E}_{\mu,\sigma^2}\left[(x_i - \mu)(x_j - \mu)^2 \right] = \mathrm{E}_{\mu,\sigma^2}[x_i - \mu] \mathrm{E}_{\mu,\sigma^2}\left[(x_j - \mu)^2 \right] = 0$$

for $i \neq j$ because the variables in the sample are independent; in step B we have used the fact that the third central moment of the normal distribution is equal to zero.

In Nutshell-The Information Matrix

- The score is a random vector with some interesting statistical properties. In particular, the score evaluated at the true parameter value $\boldsymbol{\theta}$ has mean zero

$$\mathrm{E}[\mathbf{u}(\boldsymbol{\theta})] = \mathbf{0}$$

and variance-covariance matrix given by the *information matrix*:

$$\mathrm{var}[\mathbf{u}(\boldsymbol{\theta})] = \mathrm{E}[\mathbf{u}(\boldsymbol{\theta})\mathbf{u}'(\boldsymbol{\theta})] = \mathbf{I}(\boldsymbol{\theta}).$$

- Under mild regularity conditions, the information matrix can also be obtained as minus the expected value of the second derivatives of the log-likelihood:

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}[\frac{\partial^2 \log \mathrm{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}].$$

- The matrix of negative observed second derivatives is sometimes called the *observed information matrix*.

Likelihood Ratio Tests

Notation. We'll assume that the probability density (or mass) function of X is $f(x;\theta)$ where θ represents one or more unknown parameters. Then:

(1) Let Ω (greek letter "omega") denote the total possible parameter space of θ, that is, the set of all possible values of θ as specified in totality in the null and alternative hypotheses.

(2) Let $H_0 : \theta \in \omega$ denote the null hypothesis where ω (greek letter "omega") is a subset of the parameter space Ω.

(3) Let $H_A : \theta \in \omega'$ denote the alternative hypothesis where ω' is the complement of ω with respect to the parameter space Ω.

Example

If the total parameter space of the mean μ is $\Omega = \{\mu: -\infty < \mu < \infty\}$ and the null hypothesis is specified as $H_0: \mu = 3$, how should we specify the alternative hypothesis so that the alternative parameter space is the complement of the null parameter space?

Solution. If the null parameter space is $\omega = \{\mu: \mu = 3\}$, then the alternative parameter space is everything that is in $\Omega = \{\mu: -\infty < \mu < \infty\}$ that is not in ω. That is, the alternative parameter space is $\omega' = \{\mu: \mu \neq 3\}$. And, so the alternative hypothesis is:

$$H_A: \mu \neq 3$$

In this case, we'd be interested in deriving a two-tailed test.

Example

If the alternative hypothesis is H_A: $\mu > 3$, how should we (technically) specify the null hypothesis so that the null parameter space is the complement of the alternative parameter space?

Solution. If the alternative parameter space is $\omega' = \{\mu: \mu > 3\}$, then the null parameter space is $\omega = \{\mu: \mu \leq 3\}$. And, so the null hypothesis is:

$$H_0 : \mu \leq 3$$

Now, the reality is that some authors do specify the null hypothesis as such, even when they mean H_0: $\mu = 3$. Ours don't, and so we won't. (That's why I put that "technically" in parentheses up above.) At any rate, in this case, we'd be interested in deriving a one-tailed test.

Definition. Let:

(1) $L(\hat{\omega})$ denote the maximum of the likelihood function with respect to θ when θ is in the null parameter space ω.

(2) $L(\hat{\Omega})$ denote the maximum of the likelihood function with respect to θ when θ is in the entire parameter space Ω.

Then, the **likelihood ratio** is the quotient:

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$$

And, to test the null hypothesis $H_0 : \theta \in \omega$ against the alternative hypothesis $H_A : \theta \in \omega'$, the **critical region for the likelihood ratio test** is the set of sample points for which:

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} \leq k$$

where $0 < k < 1$, and k is selected so that the test has a desired significance level α.

Example

A food processing company packages honey in small glass jars. Each jar is supposed to contain 10 fluid ounces of the sweet and gooey good stuff.

Previous experience suggests that the volume X, the volume in fluid ounces of a randomly selected jar of the company's honey is normally distributed with a known variance of 2. Derive the likelihood ratio test for testing, at a significance level of $\alpha = 0.05$, the null hypothesis H_0: $\mu = 10$ against the alternative hypothesis H_A: $\mu \neq 10$.

Solution. Because we are interested in testing the null hypothesis H_0: $\mu = 10$ against the alternative hypothesis H_A: $\mu \neq 10$ for a normal mean, our total parameter space is:

$$\Omega = \{\mu : -\infty < \mu < \infty\}$$

and our null parameter space is:

$$\omega = \{10\}$$

Now, to find the likelihood ratio, as defined above, we first need to find $L(\hat{\omega})$. Well, when the null hypothesis H_0: $\mu = 10$ is true, the mean μ can take on only one value, namely, $\mu = 10$. Therefore:

$$L(\hat{\omega}) = L(10)$$

We also need to find $L(\hat{\Omega})$ in order to define the likelihood ratio. To find it, we must find the value of μ that maximizes $L(\mu)$. Well, we did that back when we studied maximum likelihood as a method of estimation. We showed that $\hat{\mu} = \bar{x}$ is the maximum likelihood estimate of μ. Therefore:

$$L(\hat{\Omega}) = L(\bar{x})$$

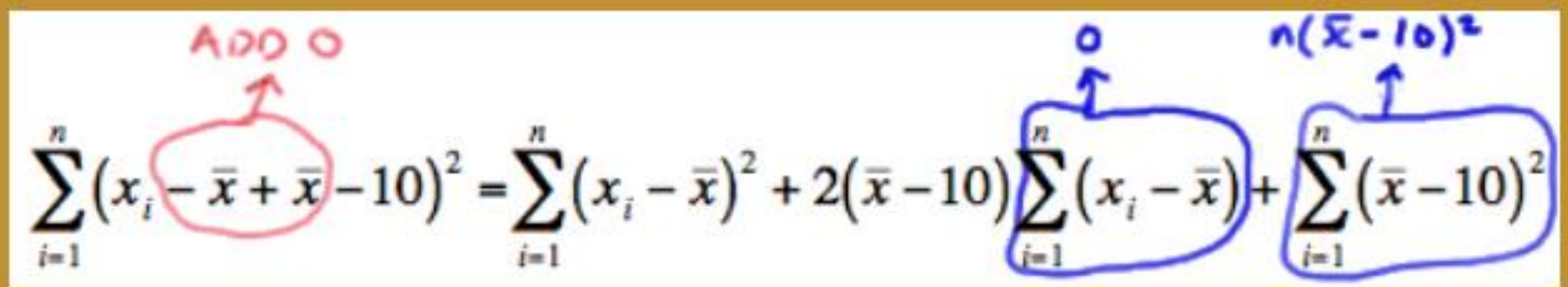
Now, putting it all together to form the likelihood ratio, we get:

$$\lambda = \frac{L(10)}{L(\bar{x})} = \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{2}} \exp\left[-\frac{(x_i - 10)^2}{2(2)}\right]}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{2}} \exp\left[-\frac{(x_i - \bar{x})^2}{2(2)}\right]}$$

which simplifies to:

$$\lambda = \frac{\exp\left[-\frac{1}{4}\sum_{i=1}^{n}(x_i-10)^2\right]}{\exp\left[-\frac{1}{4}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]}$$

Now, let's step aside for a minute and focus just on the summation in the numerator. If we "add 0" in a special way to the quantity in parentheses:



we can show that the summation can be written as:

$$\sum_{i=1}^{n}(x_i - 10)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - 10)^2$$

Therefore, the likelihood ratio becomes:

$$\lambda = \frac{exp\left[-\frac{1}{4}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{n}{4}(\bar{x} - 10)^2\right]}{exp\left[-\frac{1}{4}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]} = \frac{exp\left[-\frac{1}{4}\sum(x_i - \bar{x})^2\right] exp\left[-\frac{n}{4}(\bar{x} - 10)^2\right]}{exp\left[-\frac{1}{4}\sum(x_i - \bar{x})^2\right]}$$

which greatly simplifies to:

$$\lambda = exp\left[-\frac{n}{4}(\bar{x} - 10)^2\right]$$

Now, the likelihood ratio test tells us to reject the null hypothesis when the likelihood ratio λ is small, that is, when:

$$\lambda = exp\left[-\frac{n}{4}(\bar{x} - 10)^2\right] \leq k$$

where k is chosen to ensure that, in this case, $\alpha = 0.05$. Well, by taking the natural log of both sides of the inequality, we can show that $\lambda \leq k$ is equivalent to:

$$-\frac{n}{4}(\bar{x} - 10)^2 \leq \ln k$$

which, by multiplying through by $-4/n$, is equivalent to:

$$(x - 10)^2 \geq -\frac{4}{n}\ln k$$

which is equivalent to:

$$\frac{|\bar{X} - 10|}{\sigma/\sqrt{n}} \geq \frac{\sqrt{-(4/n)\ln k}}{\sigma/\sqrt{n}} = k*$$

Aha! We should recognize that quantity on the left-side of the inequality! We know that:

$$Z = \frac{\bar{X} - 10}{\sigma/\sqrt{n}}$$

follows a standard normal distribution when H_0: $\mu = 10$. Therefore we can determine the appropriate k^* by using the standard normal table. We have shown that the likelihood ratio test tells us to reject the null hypothesis H_0: $\mu = 10$ in favor of the alternative hypothesis H_A: $\mu \neq 10$ for all sample means for which the following holds:

$$\frac{|\bar{X} - 10|}{\sqrt{2}/\sqrt{n}} \geq z_{0.025} = 1.96$$

Doing so will ensure that our probability of committing a Type I error is set to $\alpha = 0.05$, as desired.

- END