Maximum Likelihood Theory

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and
Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Maximum Likelihood Theory

In our lessons, we will cover the following topics:

- Likelihood function
- Score vector
- Fisher information matrix
- Information matrix equality
- Cramer-Rao inequality
- Maximum Likelihood estimate/estimator
- Invariance
- Consistency
- Asymptotic normality
- Asymptotic efficiency
- Three "classical" tests based on the Likelihood

Lesson 1: The Likelihood Function

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Probabilistic model

Consider an observable random phenomenon that we want to study. Suppose that this phenomenon can be appropriately described by a random variable X with probability **density** function (*pdf*) (or probability **mass** function (*pmf*)) belonging to the family

$$\Phi = \{f(x;\theta);\ \theta \in \Theta\}$$

where Θ is a subset of the k-dimensional Euclidean space $\mathbb{R}^k$ called the **parametric space**. The family Φ is a **probabilistic model**.

A probabilistic model is a family of probability density functions (or probability mass functions in the case of discrete distributions)

Gaussian model

$$\Phi = \left\{ f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};\ \theta = (\mu, \sigma^2)' \in \Theta = \mathbb{R} \times \mathbb{R}^+ \right\}$$

Bernoulli model

$$\Phi = \left\{ f(x;\theta) = \theta^{x}(1-\theta)^{1-x};\ \theta \in \Theta = (0,1) \right\}$$

Poisson model

$$\Phi = \left\{ f(x;\theta) = \frac{e^{-\theta}\theta^{x}}{x!};\ \theta \in \Theta = (0,\infty) \right\}$$

Probabilistic model

Since the functional form of the density functions of the probabilistic model is known, we have that all the uncertainty concerning the random phenomenon is that concerning the parameter θ.

In order to get information on θ, we will consider a sample. What is a sample?

Sample

Definition 1. Let $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$ be a vector of n random variables identically distributed with *pdf* (or *pmf*) belonging to the family

$$\Phi = \{f(x;\theta);\ \theta \in \Theta\}.$$

We say that $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$ is a **sample** of size n from $f(x;\theta)$. The distribution of the sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$ is the joint distribution of the random variables $X_1, X_2, ..., X_n$ denoted by

$$f_{1,2,...,n}(\mathbf{x}_n;\theta) = f_{1,2,...,n}(x_1, x_2, ..., x_n;\theta)$$

If the n random variables $X_1, X_2, ..., X_n$ are independent, we say that $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$ is a **random sample** of size n from $f(x;\theta)$. In this case, we have that

$$f_{1,2,...,n}(x_1, x_2, ..., x_n;\theta) = f(x_1;\theta)f(x_2;\theta)...f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

An example

Let $\mathbf{X}_n = (X_1, X_2, ..., X_n)$ be a random sample from a $N(\mu, \sigma^2)$ distribution with μ and σ^2 unknown.
In this case

$$f(x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2} \quad \text{for} \quad i = 1, 2, ..., n$$

where $\theta = (\mu, \sigma^2)' \in \mathbb{R} \times \mathbb{R}^+$, and the joint distribution of the random sample is

$$\begin{aligned} f_{1,2,...,n}(\mathbf{x}_n; \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \end{aligned}$$

Likelihood Function

Definition 2. Let $\mathbf{X}_n = (X_1, X_2, ..., X_n)$ be a sample of size n from $f(x;\theta)$, $\theta \in \Theta$. Given a realization

$$\mathbf{x}_n = (x_1, x_2, ..., x_n)$$

of the sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)$, the function

$$L : \Theta \to [0, \infty)$$

defined by $L(\theta; \mathbf{x}_n) = f_{1,2,...,n}(\mathbf{x}_n; \theta)$ is called the **likelihood function**.

Thus, the likelihood function $L(\theta; \mathbf{x})$ is the function $f_{1,2,...,n}(\mathbf{x}_n; \theta)$, viewed as a function of the parameter θ with $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ fixed.

Likelihood Function

We note that if

$$\mathbf{x}_n = (x_1, x_2, ..., x_n)$$

is the realization of a random sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)$, then

$$f_{1,2,...,n}(\mathbf{x}_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = f(x_1; \theta) f(x_2; \theta) ... f(x_n; \theta)$$

and hence

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

An example

Let $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ be a realization of a random sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)$ from an $N(\mu, \sigma^2)$ distribution with μ and σ^2 unknown.

In this case $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$, and the likelihood function is

$$L(\mu, \sigma^2; \mathbf{x}_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

Another example

Let $\mathbf{x}_n = (x_1, x_2, ..., x_n)'$ be a realization of a random sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$ from a Bernoulli distribution with probability mass function

$$f(x;\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

The likelihood function is

$$L(\theta;\mathbf{x}_n) = \theta^{x_1}(1-\theta)^{(1-x_1)}\theta^{x_2}(1-\theta)^{(1-x_2)}...\theta^{x_n}(1-\theta)^{(1-x_n)}$$

$$= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{(n-\sum_{i=1}^{n} x_i)}$$

An example

Suppose that the realization of the random sample $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ is such that

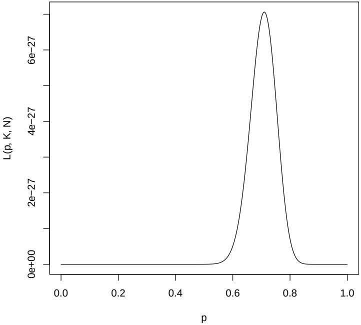
$$\sum_{i=1}^{n} x_i = 71$$

with $n = 100$.

The likelihood function is

$$L(\theta; \mathbf{x}_n) = \theta^{71}(1-\theta)^{29}$$

An example



Likelihood Function

Let $\mathbf{x}_n = (x_1, x_2, ..., x_n)'$ and $\mathbf{x}_n^* = (x_1^*, x_2^*, ..., x_n^*)'$ be two different realizations of a random sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)'$.

The likelihood function at the point

$$\mathbf{x}_n = (x_1, x_2, ..., x_n)'$$

is (generally) a different function from what it is at the point $\mathbf{x}_n^* = (x_1^*, x_2^*, ..., x_n^*)'$, that is

$$L(\theta; \mathbf{x}_n) \neq L(\theta; \mathbf{x}_n^*)$$

Likelihood Function

Consider a realization $\mathbf{x}_5 = (x_1, x_2, x_3, x_4, x_5)$ of a random sample $\mathbf{X}_5 = (X_1, X_2, X_3, X_4, X_5)'$ from a Bernoulli distribution with parameter θ.

Suppose $\mathbf{x}_5 = (1, 0, 1, 0, 1)'$. The likelihood function is:

$$L(\theta; (1, 0, 1, 0, 1)') = \theta^3(1-\theta)^2$$

Suppose $\mathbf{x}_5 = (1, 0, 1, 0, 0)'$. The likelihood function is:

$$L(\theta; (1, 0, 1, 0, 0)') = \theta^2(1-\theta)^3$$

Suppose $\mathbf{x}_5 = (1, 0, 0, 0, 0)'$. The likelihood function is:

$$L(\theta; (1, 0, 0, 0, 0)') = \theta(1-\theta)^4$$

Likelihood Function

The likelihood function at the point $\mathbf{x}_n = (x_1, x_2, ..., x_n)'$ is (generally) a different function from what it is at the point $\mathbf{x}_n^* = (x_1^*, x_2^*, ..., x_n^*)'$, that is

$$L(\theta; \mathbf{x}_n) \neq L(\theta; \mathbf{x}_n^*)$$

Generally, but not always!

Consider again two realizations of a random sample $\mathbf{X}_5 = (X_1, X_2, X_3, X_4, X_5)'$ from a Bernoulli distribution, $\mathbf{x}_5 = (1, 0, 0, 0, 0)'$ and $\mathbf{x}_5^* = (0, 0, 0, 0, 1)'$. We have that $\mathbf{x}_5 \neq \mathbf{x}_5^*$ but

$$L(\theta; \mathbf{x}_5) = L(\theta; \mathbf{x}_5^*) = \theta(1-\theta)^4$$

Likelihood Function

The likelihood function expresses the plausibilities of different parameters after we have observed $\mathbf{x}_n$. In particular, for $\theta = \theta^*$, the number $L(\theta^*; \mathbf{x}_n)$ is considered a measure of support that the observation $\mathbf{x}_n$ gives to the parameter θ^*.

Likelihood Function

Consider a realization $\mathbf{x}_5 = (x_1, x_2, x_3, x_4, x_5)$ of a random sample $\mathbf{X}_5 = (X_1, X_2, X_3, X_4, X_5)'$ from a Bernoulli distribution with parameter θ.

Suppose $\mathbf{x}_5 = (1, 1, 1, 1, 1)'$ and consider two possible values of θ: $\theta_1 = 1/3$ and $\theta_2 = 2/3$. The plausibility of θ_1 is:

$$L(\theta_1; (1, 1, 1, 1, 1)') = \left(\frac{1}{3}\right)^5 = 0.004115226$$

The plausibility of θ_2 is:

$$L(\theta_2; (1, 1, 1, 1, 1)') = \left(\frac{2}{3}\right)^5 = 0.1316872$$

Clearly

$$L(\theta_2; (1, 1, 1, 1, 1)') > L(\theta_1; (1, 1, 1, 1, 1)')$$

Likelihood Function

Now, suppose $\mathbf{x}_5 = (0,0,0,0,0)'$. The plausibility of θ_1 is:

$$L(\theta_1; (0,0,0,0,0)') = \left(\frac{2}{3}\right)^5 = 0.1316872$$

The plausibility of θ_2 is:

$$L(\theta_2; (0,0,0,0,0)') = \left(\frac{1}{3}\right)^5 = 0.004115226$$

Clearly

$$L(\theta_1; (0,0,0,0,0)') > L(\theta_2; (0,0,0,0,0)').$$

Likelihood Function

In summary, the likelihood function is a fundamental concept in statistical inference that quantifies the plausibility of different parameter values given a set of observed data.

Key Concepts

- Probabilistic model
- Sample
- Sample realization
- Likelihood function

Lesson 2: Score vector and information matrix

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

The log-likelihood

Definition 3. Let $\mathbf{X}_n = (X_1, X_2, ..., X_n)$ be a sample of size n from $f(x;\theta)$, $\theta \in \Theta$. Given a realization

$$\mathbf{x}_n = (x_1, x_2, ..., x_n)$$

of the sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)$, the function

$$\ell : \Theta \to \mathbb{R}$$

defined by

$$\ell(\theta; \mathbf{x}_n) = \ln L(\mathbf{x}_n; \theta)$$

is called the **log-likelihood function**.

Log-likelihood function is a logarithmic transformation of the likelihood function.

The log-likelihood

We remember that if $\mathbf{x}_n$ is a realization of a **random** sample, then

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

and hence

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

An example

Let $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ be a realization of a random sample from a $N(\mu, \sigma^2)$ distribution with μ and σ unknown.
In this case $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$, and the likelihood function is

$$\begin{aligned} L(\mu, \sigma^2; \mathbf{x}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \\ &= \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \end{aligned}$$

and the log-likelihood function is given by

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

Score function

Definition 4. If the likelihood function, $L(\theta; \mathbf{x})$, is differentiable, then the gradient of the log-likelihood

$$s(\theta; \mathbf{x}) = \frac{\delta \ell(\theta; \mathbf{x})}{\delta \theta} = \frac{\delta \ln L(\theta; \mathbf{x})}{\delta \theta}$$

is called **the score function**.

The score function can be found through the chain rule:

$$\frac{\delta \ell(\theta; \mathbf{x})}{\delta \theta} = \frac{1}{L(\theta; \mathbf{x})} \frac{\delta L(\theta; \mathbf{x})}{\delta \theta}$$

An example

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a realization of a random sample from a $N(\mu, \sigma^2)$ distribution.In this case $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$.The log-likelihood function is given by

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

We have that

$$\frac{\delta\ell((\mu, \sigma^2), \mathbf{x})}{\delta\mu} = \frac{\Sigma(x_i - \mu)}{\sigma^2}$$

and

$$\frac{\delta\ell((\mu, \sigma^2), \mathbf{x})}{\delta\sigma^2} = \frac{\Sigma(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2}$$

Thus the score function is given by

$$s(\theta; \mathbf{x}) = \left(\frac{\Sigma(x_i - \mu)}{\sigma^2}, \frac{\Sigma(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2}\right)'$$

Score vector

Definition 5. Evaluating the score function at a specific value of θ and replacing the fixed values $\mathbf{x} = (x_1, x_2, ..., x_n)'$ by their corresponding random variables $\mathbf{X} = (X_1, X_2, ..., X_n)'$, the score function becomes a **random vector**

$$s(\theta; \mathbf{X}) = \frac{\delta \ell(\theta; \mathbf{X})}{\delta\theta} = \frac{\delta \ln f(\mathbf{X}; \theta)}{\delta\theta}.$$

We call this random vector **score vector**.

Score vector

Which is the expected value of the score vector?

The expected value of the score vector evaluated at the true parameter value equals zero.

Score vector

Theorem 1. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from a distribution with p.d.f. belonging to the family

$$\Phi = \{f(x;\theta);\ \theta \in \Theta\}$$

and let θ_0 be the true value of the parameter θ, then under certain regularity conditions

$$E\left[s(\theta_0;\mathbf{X})\right] = \int \frac{\delta \ln f(\mathbf{x};\theta_0)}{\delta\theta} f(\mathbf{x};\theta_0) d\mathbf{x} = \mathbf{0}$$

Here the single integral $\int ...d\mathbf{x}$, is used to indicate the multiple integration over all elements of $\mathbf{x}$.

Remark

We use often the phrase ‘under certain regularity conditions’. What are these regularity conditions?

These conditions mainly relate to differentiability of the density $f(x;\theta)$ and the ability to interchange differentiation and integration

$$\frac{\delta}{\delta\theta}\left[\int f(\mathbf{x};\theta)d\mathbf{x}\right] = \int \frac{\delta f(\mathbf{x};\theta)}{\delta\theta}d\mathbf{x}$$

Proof Theorem 1

Because $f(\mathbf{x};\theta)\ \forall\theta\in\Theta$ is a probability density function, we have that:

$$\int f(\mathbf{x};\theta)d\mathbf{x} = 1\ \ \forall\theta\in\Theta \quad (*)$$

Thus, differentiating (*) w.r.t. θ we get

$$\frac{\delta}{\delta\theta}\left[\int f(\mathbf{x};\theta)d\mathbf{x}\right] = \mathbf{0} \quad (**)$$

The regularity conditions guarantee that operations of differentiation and integration can be interchanged. Thus, we have

$$\frac{\delta}{\delta\theta}\left[\int f(\mathbf{x};\theta)d\mathbf{x}\right] = \int \frac{\delta f(\mathbf{x};\theta)}{\delta\theta}d\mathbf{x}$$

Proof Theorem 1

So, (**) can be rewritten as

$$\int \frac{\delta f(\mathbf{x};\theta)}{\delta\theta} d\mathbf{x} = \mathbf{0} \quad (***)$$

Proof Theorem 1

Because

$$\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} = \frac{1}{f(\mathbf{x};\theta)}\frac{\delta f(\mathbf{x};\theta)}{\delta\theta}.$$

we have that

$$\frac{\delta f(\mathbf{x};\theta)}{\delta\theta} = \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} f(\mathbf{x};\theta)$$

and hence

$$\int \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} f(\mathbf{x};\theta) d\mathbf{x} = \mathbf{0} \;\; \forall\theta \in \Theta. \quad (****)$$

Proof Theorem 1

On the other hand, we have that

$$\int \frac{\delta \ln f(\mathbf{x};\theta_0)}{\delta\theta} f(\mathbf{x};\theta_0) d\mathbf{x} = E\left[\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right] = E\left[s(\theta_0;\mathbf{X})\right].$$

By equation (****) it follows that

$$E\left[s(\theta_0;\mathbf{X})\right] = \mathbf{0}$$

The score vector evaluated at the true parameter value has mean zero.

Score vector

Remark. Consider a vector of parameters $\theta_1 \neq \theta_0$. We have that, in general,

$$E\left[s(\theta_1; \mathbf{X})\right]$$

can be different from the null vector $\mathbf{0}$

$$E\left[s(\theta_1; \mathbf{X})\right] = \int \frac{\delta \ln f(\mathbf{x}; \theta_1)}{\delta \theta} f(\mathbf{x}; \theta_0) d\mathbf{x} \neq \mathbf{0}$$

Score vector

The expected value of the score vector evaluated at the true parameter value equals zero.

$$E\left[s(\theta_0;\mathbf{X})\right]=\mathbf{0}$$

But the expected value of the score vector evaluated at a parameter different from the true parameter can be different from zero

$$E\left[s(\theta_1;\mathbf{X})\right]\neq\mathbf{0}$$

Information matrix

Definition 6. The variance-covariance matrix of the score vector, evaluated at the true parameter value,

$$Var\left[s(\theta_0;\mathbf{X})\right] = E\left[s(\theta_0;\mathbf{X})s(\theta_0;\mathbf{X})'\right] = E\left[\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta'}\right]$$

is called **Fisher information matrix** for θ_0 (or Fisher's information measure on θ_0 contained in the r.v. $\mathbf{X}$).

This matrix, denoted by $I_n(\theta_0)$, measures the amount of information about θ_0 contained (on average) in a realization $\mathbf{x}$ of the r.v. $\mathbf{X}$.

Information matrix

In summary, we have that

The Fisher information matrix is defined to be the variance of the score vector evaluated at the true parameter value θ_0

$$I_n(\theta_0) = Var\left[s(\theta_0; \mathbf{X})\right]$$

The Fisher information matrix is always positive semi-definite. It can be shown that the Fisher information matrix of regular probability distributions is positive definite, and therefore always invertible.

Information Number

If

$$\theta \in \Theta \subset \mathbb{R}$$

θ is scalar and the information matrix becomes a scalar that we call **information number**.

$$I_n(\theta_0) = E\left[\left(\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right)^2\right]$$

The Hessian of the log-likelihood

Consider the Hessian of the log-likelihood

$$H(\mathbf{x};\theta)=\frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta\delta\theta'}$$

The matrix of second partial derivatives.

$$H(\mathbf{x};\theta)=\frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta\delta\theta'}=\begin{bmatrix} \frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta_1^2} & \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta_1\delta\theta_2} & \cdots & \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta_1\delta\theta_k} \\ \frac{\ln f(\mathbf{x};\theta)}{\delta\theta_2\delta\theta_1} & \frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta_2^2} & \cdots & \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta_2\delta\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\ln f(\mathbf{x};\theta)}{\delta\theta_k\delta\theta_1} & \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta_k\delta\theta_2} & \cdots & \frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta_k^2} \end{bmatrix}$$

The Hessian of the log-likelihood

Evaluating the Hessian of the log-likelihood at a specific value of θ and replacing the fixed values $\mathbf{x} = (x_1, x_2, ..., x_n)'$ by their corresponding random variables $\mathbf{X} = (X_1, X_2, ..., X_n)'$, the Hessian becomes a **random matrix**

$$H(\mathbf{X};\theta) = \frac{\delta^2 \ln f(\mathbf{X};\theta)}{\delta\theta\delta\theta'}$$

In particular, we consider the Hessian of the log-likelihood evaluated at the true parameter θ_0, that is

$$H(\mathbf{X};\theta_0) = \frac{\delta^2 \ln f(\mathbf{X};\theta_0)}{\delta\theta\delta\theta'}$$

Information matrix equality

Theorem 2. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from a distribution with p.d.f. belonging to the family

$$\Phi = \{f(x;\theta);\ \theta \in \Theta\}$$

and let θ_0 be the true value of the parameter θ, then under some regularity conditions

$$I_n(\theta_0) = -E\left[\frac{\delta^2 \ln f(\mathbf{X};\theta_0)}{\delta\theta\delta\theta'}\right]$$

This theorem is called the **information matrix equality**. It provides an alternative expression for the information matrix. The information matrix for θ_0 equals the negative of the expected value of Hessian of the log-likelihood evaluated at the true parameter θ_0.

Information matrix

It is important to note that the results presented do not depend on the assumption of independence of the random variables $X_1, X_2, ..., X_n$. This assumption can be used in order to get the following result.

Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from a distribution with p.d.f. $f(x; \theta_0)$. We have that

$$I_n(\theta_0) = nI_1(\theta_0)$$

The information in a random sample of size n is n times that in a sample of size 1.

The asymptotic information matrix

The matrix

$$I_a(\theta_0) = \lim_{n\to\infty} I_n(\theta_0)/n$$

if it exists, is the **asymptotic information matrix** for θ_0.

The asymptotic information matrix

Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample from a distribution with p.d.f. $f(x; \theta_0)$. We have that

$$I_a(\theta_0) = \lim_{n\to\infty} \frac{I_n(\theta_0)}{n}$$

$$= \lim_{n\to\infty} \frac{nI_1(\theta_0)}{n}$$

$$= I_1(\theta_0)$$

The asymptotic information matrix is the Fisher information matrix for one observation.

Key Concepts

- Log-likelihood function
- Score function
- Score vector
- Fisher information matrix
- Asymptotic information matrix

Lesson 3: Cramér-Rao inequality

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Cramér-Rao inequality (scalar-parameter case)

Here we consider the case in which

$$\theta \in \Theta \subset \mathbb{R}$$

Thus θ is scalar and the information matrix becomes a scalar that we call information number.

$$I_n(\theta_0) = E\left[\left(\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right)^2\right]$$

Cramér-Rao inequality (scalar-parameter case)

Theorem 3. Let $\mathbf{X}_n = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x;\theta)$ depending on a real parameter $\theta \in \Theta \subset \mathbb{R}$. Let $T(\mathbf{X})$ be an unbiased estimator of θ. Then, subject to certain regularity conditions on $f(x;\theta)$, the variance of $T(\mathbf{X})$ satisfies the inequality

$$\mathrm{Var}[T(\mathbf{X})] \geq \frac{1}{E\left[\left(\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right)^2\right]}$$

where the derivative is evaluated at the true value of θ and the expectation is taken with respect to $f(x;\theta_0)$.

Proof

$T(\mathbf{X})$ is an unbiased estimator of θ, so

$$E[T(\mathbf{X})] = \int T(\mathbf{x}) f(\mathbf{x};\theta) d\mathbf{x} = \theta \ \ \forall \theta \in \Theta \tag{1}$$

Differentiating both sides of equation (1) with respect to θ, and interchanging the order of integration and differentiation, gives

$$\int T(\mathbf{x}) \frac{\delta f(\mathbf{x};\theta)}{\delta\theta} d\mathbf{x} = 1 \tag{2}$$

or

$$\int T(\mathbf{x}) \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} f(\mathbf{x};\theta) d\mathbf{x} = 1 \tag{3}$$

Proof

Because

$$\int T(\mathbf{x}) \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} f(\mathbf{x};\theta) d\mathbf{x} = E\left[T(\mathbf{X}) \frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right] \qquad (4)$$

by (3) it follows that

$$E\left[T(\mathbf{X}) \frac{\delta \ln f(\mathbf{X};\theta)}{\delta\theta}\right] = 1$$

Proof

On the other hand, since

$$E\left[\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right]=0$$

we have that

$$E\left[T(\mathbf{X})\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right]=\text{Cov}\left[T(\mathbf{X}),\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right]$$

Hence

$$\text{Cov}\left[T(\mathbf{X}),\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right]=1$$

Proof

Since the squared covariance cannot exceed the product of the two variances, we have

$$1 = \left(\mathrm{Cov}\left[T(\mathbf{X}), \frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right]\right)^2 \leq \mathrm{Var}\left[T(\mathbf{X})\right]\mathrm{Var}\left[\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right]$$

or

$$1 \leq \mathrm{Var}\left[T(\mathbf{X})\right] E\left[\left(\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right)^2\right]$$

It follows that

$$\mathrm{Var}[T(\mathbf{X})] \geq \frac{1}{E\left[\left(\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right)^2\right]}$$

Efficiency

Definition 7. An unbiased estimator $T(\mathbf{X})$ is red **more efficient** than another unbiased estimator, $T^*(\mathbf{X})$, if the variance of $T(\mathbf{X})$ is less than that $T^*(\mathbf{X})$. That is

$$\text{Var}(T(\mathbf{X}) < \text{Var}(T^*(\mathbf{X})$$

Efficiency

In many situations, there are numerous estimators for the unknown parameter θ. The usefulness of the Cramér-Rao inequality is that if one of these is known to attain the variance bound, there is no need to consider any other in order to seek a more efficient estimator.

Efficiency

Definition 8. An unbiased estimator $T(\mathbf{X})$ is red **efficient** if its variance is the lower bound of the inequality, that is

$$\mathrm{Var}[T(\mathbf{X})] = \frac{1}{I_n(\theta_0)}$$

Remark

We have seen that the quantity

$$I_n(\theta_0) = E\left[\left(\frac{\delta \ln f(\mathbf{X};\theta_0)}{\delta\theta}\right)^2\right]$$

is called Fisher information number.
Now, we are able to explain the reason of this terminology.

It is well known that there is an inverse relationship between the variance of an efficient estimator and the information contained in the sample, concerning the unknown parameter. The bigger is this information, the lower it will be the variance. On the other hand, there is also an inverse relationship between the variance of an efficient estimator and the quantity $I_n(\theta_0)$. Thus, we can conclude that $I_n(\theta_0)$ is a measure of the information about the unknown parameter contained in a sample.

Multidimensional Cramer-Rao inequality

The Cramer-Rao inequality (Theorem 3) can be generalized to a vector valued parameter $\theta \in \Theta \subset \mathbb{R}^k$.

The generalization of the Cramer-Rao inequality states that, again subject to regularity conditions, the variance-covariance matrix of the unbiased estimator $T(\mathbf{X})$, the $k \times k$ matrix $\mathrm{Var}(T(\mathbf{X}))$, is such that $\mathrm{Var}(T(\mathbf{X})) - I_n^{-1}(\theta_0)$ is positive semi-definite.

Thus $I_n^{-1}(\theta_0)$ is in a sense a 'lower bound' for the variance matrix of an unbiased estimator of θ.

Lesson 4: Maximum Likelihood estimator

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Maximum Likelihood Estimate

First, we consider the question of whether estimation of the unknown parameter θ is possible at all: **the question of identification**.

Definition 9 (Identification). The parameter θ is **identified (estimable)** if for any other parameter $\theta^* \neq \theta$, for some sample $\mathbf{x}$, we have

$$L(\theta;^* \mathbf{x}) \neq L(\theta; \mathbf{x}).$$

In the follow, we assume that θ is identified.

How can we estimate the parameter θ? Given that the likelihood function represents the plausibility of the various $\theta \in \Theta$ given the realization $\mathbf{x}$, it is natural to chose as estimate of θ the most plausible element of Θ.

Maximum Likelihood Estimate

Definition 10. Let $\mathbf{x} = (x_1, ..., x_n)$ be a realization of a sample from a distribution with p.d.f. $f(x;\theta)$ depending on an unknown parameter $\theta \in \Theta$. A **Maximum Likelihood Estimate** $\hat{\theta}_n(\mathbf{x}) = \hat{\theta}_n(x_1, ..., x_n)$ is an element of Θ that maximizes the value of $L(\theta; \mathbf{x})$, i.e.,

$$L(\hat{\theta}_n(\mathbf{x}); \mathbf{x}) = \max_{\theta\in\Theta} L(\theta; \mathbf{x})$$

or

$$\hat{\theta}_n(\mathbf{x}) = \text{argmax}_{\theta\in\Theta} L(\theta; \mathbf{x})$$

Maximum Likelihood estimate

There are examples where the MLE is not unique or even does not exist

Proposition 1 (Sufficient condition for existence). If the parameter space Θ is **compact** and if the likelihood function $L(\theta; \mathbf{x})$ is **continuous** on Θ, then there exists an MLE.

Proposition 2 (Sufficient condition for uniqueness of MLE). If the parameter space Θ is **convex** and if the likelihood function $L(\theta; \mathbf{x})$ is **strictly concave** in Θ, then the MLE is unique when it exists.

Maximum Likelihood estimate

Maximizing the likelihood function is mathematically equivalent to maximizing the log-likelihood function since the logarithm function is a strictly increasing function. The values that maximize $L(\theta; \mathbf{x})$ are the same as those that maximize $\ln L(\theta; \mathbf{x})$

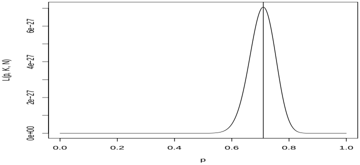
$$L(\hat{\theta}; \mathbf{x}) = \max_{\theta \in \Theta} L(\theta; \mathbf{x})$$

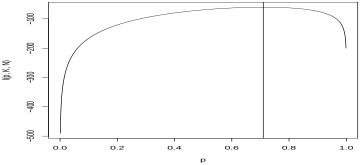
if and only if

$$\ln L(\hat{\theta}; \mathbf{x}) = \max_{\theta \in \Theta} \ln L(\theta; \mathbf{x})$$

The log-likelihood function is usually simpler to optimize.

Maximum Likelihood estimate





Maximum Likelihood estimate

We remember that if $\mathbf{x}_n$ is a realization of a **random** sample, then

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

and hence

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

The convenience of the log likelihood arises from the fact that it is typically much easier to differentiate a sum than a product.

Maximum Likelihood estimate

In the case where $L(\theta; \mathbf{x})$ is differentiable the MLE can be derived as a solution of the equation

$$\frac{\delta \ln L(\theta; \mathbf{x})}{\delta \theta} = \mathbf{0}$$

called **the likelihood equation**.

Maximum Likelihood estimate

- The likelihood equation represents the **first-order necessary condition** for the maximization of the log-likelihood function.
- The **second-order necessary condition** for a point to be the local maximum of the log-likelihood function is that the Hessian be negative semi-definite at the point.

Five steps for finding MLE

1. Find Likelihood function $L(\theta; \mathbf{x})$.
2. Get natural log of Likelihood function $l(\theta; \mathbf{x}) = ln(L(\theta; \mathbf{x})$.
3. Differentiate log-Likelihood function with respect to θ.
4. Set derivative to zero.
5. Solve for θ.

Maximum Likelihood Estimate: an example

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a realization of a random sample from an $N(\mu, \sigma^2)$ distribution with μ and σ unknown.
In this case $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$, and the likelihood function is

$$L(\mu, \sigma^2; \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$

The log-likelihood function is given by

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

Maximum Likelihood Estimate: an example

Taking the first derivative (gradient), we get

$$\frac{\partial l(\theta; \mathbf{x})}{\partial \theta} = \left(\frac{\Sigma(x_i - \mu)}{\sigma^2}, \frac{\Sigma(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} \right)'.$$

Setting

$$\frac{\partial l(\theta; \mathbf{x})}{\partial \theta} = 0$$

and solve for $\theta = (\mu, \sigma^2)$ we have

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = (\overline{x}, \frac{n-1}{n} s^2),$$

where $\overline{x} = \Sigma x_i / n$ is the sample mean and $s^2 = \Sigma(x_i - \overline{x})^2/(n-1)$ is the sample variance.
It is not difficult to verify that these values of μ and σ^2 yield an absolute (not only a local) maximum of the log-likelihood function, so that they are maximum likelihood estimates.

Maximum-Likelihood Estimation of the Classical Linear Regression Model

Consider the classical linear regression model

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + \epsilon_t \quad t = 1, 2, ..., T$$

Let us assume that the disturbances ϵ_t are distributed normally, independently and identically with $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) = \sigma_2$ for all t. The equation above can be written in summary form as

$$y_t = x_t'\beta + \epsilon_t \quad t = 1, 2, ..., T$$

where $x_t' = [x_{t1}, x_{t2}, ..., x_{tk}]$, and $\beta = [\beta_1, \beta_2, ..., \beta_k]'$.

Maximum-Likelihood Estimation of the Classical Linear Regression Model

Then, if the vectors x_t are taken as data, the observations y_t $t = 1, 2, ..., T$ have density functions $N(x_t'\beta, \sigma^2)$ and the likelihood function of β and σ^2, based on the sample $y = (y_1, y_2, ..., y_T)$ is

$$L(\beta, \sigma^2; y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right]$$

where

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_T' \end{bmatrix}$$

Maximum-Likelihood Estimation of the Classical Linear Regression Model

The logarithm of this function is

$$l(\beta, \sigma^2; y) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2)) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

Maximum Likelihood Estimate of the linear regression model

Therefore the first order conditions for a maximum are:

$$\frac{\delta l}{\delta \beta} = \frac{1}{\sigma^2}(X'(y - X\beta) = 0$$

$$\frac{\delta l}{\delta \sigma^2} = -\frac{T}{\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) = 0$$

From the first of the two conditions it is evident that the maximum likelihood estimator of β coincides with that of the ordinary least squares.

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

Maximum Likelihood Estimate

Sometimes it is not possible to find an explicit solution of the likelihood equation and so we have to use iterative algorithms to maximize $l(\theta; \mathbf{x})$, as the Newton-Raphson or the Fisher-scoring, which at any iteration update the parameter θ in appropriate way until convergence.

A summary example

Let $\mathbf{X}_n = (X_1, X_2, ..., X_n)$ be a random sample from a $N(\mu, \sigma^2)$ distribution with μ and σ^2 unknown.In this case

$$f(x_i;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \quad \text{for} \quad i = 1, 2, ..., n$$

where $\theta = (\mu, \sigma^2)' \in \mathbb{R} \times \mathbb{R}^+$, and the joint probability density function of the random sample is

$$\begin{aligned} f_{1,2,...,n}(\mathbf{x}_n; \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \end{aligned}$$

The likelihood function is

$$L(\mu, \sigma^2; \mathbf{x}_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

A summary example

The log-likelihood function is given by

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

We have that

$$\frac{\delta\ell(\mu, \sigma^2; \mathbf{x})}{\delta\mu} = \frac{\Sigma(x_i - \mu)}{\sigma^2}$$

and

$$\frac{\delta\ell(\mu, \sigma^2; \mathbf{x})}{\delta\sigma^2} = \frac{\Sigma(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2}$$

Thus the score function is given by

$$s(\theta; \mathbf{x}) = \left(\frac{\Sigma(x_i - \mu)}{\sigma^2}, \frac{\Sigma(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2}\right)'$$

A summary example

Setting

$$\left(\frac{\Sigma(x_i-\mu)}{\sigma^2}, \frac{\Sigma(x_i-\mu)^2}{2\sigma^4}-\frac{n}{2\sigma^2}\right)' = (0,0)'$$

we obtain the following system of two equations in two unknowns μ and σ^2:

$$\frac{\Sigma(x_i-\mu)}{\sigma^2} = 0$$

$$\frac{\Sigma(x_i-\mu)^2}{2\sigma^4}-\frac{n}{2\sigma^2} = 0$$

A summary example

Solving for μ and σ^2 we obtain the maximum likelihood estimator

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)' = (\overline{x}, \frac{n-1}{n} s^2)',$$

where

$$\overline{x} = \Sigma x_i / n$$

is the sample mean and

$$s^2 = \Sigma (x_i - \overline{x})^2 / (n-1)$$

is the sample variance.

A summary example

Let $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ be a realization of a random sample $\mathbf{X}_n = (X_1, X_2, ..., X_n)$ from a $N(\mu, \sigma^2)$. The score function is given by

$$s(\theta; \mathbf{x}) = \left(\frac{\Sigma(x_i - \mu)}{\sigma^2}, \frac{\Sigma(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} \right)'$$

Evaluating the score function at a specific value of $\theta = (\mu, \sigma^2)'$ and replacing the fixed values $\mathbf{x} = (x_1, x_2, ..., x_n)'$ by their corresponding random variables $\mathbf{X} = (X_1, X_2, ..., X_n)'$, the score vector

$$s(\theta; \mathbf{X}) = \left(\frac{\Sigma(X_i - \mu)}{\sigma^2}, \frac{\Sigma(X_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} \right)'$$

A summary example

In particular, we consider the score vector evaluated at the true parameter $\theta_0 = (\mu_0, \sigma_0^2)'$

$$s(\theta_0; \mathbf{X}) = \left(\frac{\Sigma(X_i - \mu_0)}{\sigma_0^2}, \frac{\Sigma(X_i - \mu_0)^2}{2\sigma_0^4} - \frac{n}{2\sigma_0^2} \right)'$$

We note that

$$E\left[\frac{\Sigma(X_i - \mu_0)}{\sigma_0^2}\right] = \frac{\Sigma(E(X_i) - \mu_0)}{\sigma^2} = \frac{\Sigma(\mu_0 - \mu_0)}{\sigma_0^2} = 0$$

and

$$E\left[\frac{\Sigma(X_i - \mu_0)^2}{2\sigma_0^4} - \frac{n}{2\sigma_0^2}\right] = \frac{\Sigma E(X_i - \mu_0)^2}{2\sigma_0^4} - \frac{n}{2\sigma_0^2} = \frac{n\sigma_0^2}{2\sigma_0^4} - \frac{n}{2\sigma_0^2} = 0.$$

Thus

$$E(s(\theta_0; \mathbf{X})) = \mathbf{0}.$$

The expected value of the score vector evaluated at the true parameter value equals zero

A summary example

The variance-covariance matrix of the score vector evaluated at the true parameter (the Fisher information matrix) is

$$\mathrm{Var}\left(s(\theta_0;\mathbf{X})\right)=\begin{pmatrix}\mathrm{Var}\left(\frac{\partial\ell(\theta_0;\mathbf{X})}{\partial\mu}\right) & \mathrm{Cov}\left(\frac{\partial\ell(\theta_0;\mathbf{X})}{\partial\mu},\frac{\partial\ell(\theta_0;\mathbf{X})}{\partial\sigma^2}\right)\\ \mathrm{Cov}\left(\frac{\partial\ell(\theta_0;\mathbf{X})}{\partial\mu},\frac{\partial\ell(\theta_0;\mathbf{X})}{\partial\sigma^2}\right) & \mathrm{Var}\left(\frac{\partial\ell(\theta_0;\mathbf{X})}{\partial\sigma^2}\right)\end{pmatrix}$$

$$=\begin{pmatrix}\frac{n}{\sigma_0^2} & 0\\ 0 & \frac{n}{2\sigma_0^4}\end{pmatrix}.$$

A summary example

The Hessian matrix evaluated at the true parameter is

$$H(\mu_0, \sigma_0^2) = \begin{pmatrix} -\frac{n}{\sigma_0^2} & -\frac{1}{\sigma_0^4}\sum_{i=1}^n (X_i - \mu_0) \\ -\frac{1}{\sigma_0^4}\sum_{i=1}^n (X_i - \mu_0) & \frac{n}{2\sigma_0^4} - \frac{1}{\sigma_0^6}\sum_{i=1}^n (X_i - \mu_0)^2 \end{pmatrix}$$

The expected value of $H(\mu_0, \sigma_0^2)$ is

$$E[H(\mu_0, \sigma_0^2)] = \begin{pmatrix} -\frac{n}{\sigma_0^2} & 0 \\ 0 & -\frac{n}{2\sigma_0^4} \end{pmatrix}$$

Thus

$$I(\mu_0, \sigma_0^2) = -E[H(\mu_0, \sigma_0^2)]$$

This is the information matrix equality. The information matrix equals the negative of the expected value of Hessian of the log-likelihood evaluated at the true parameter $\theta_0 = (\mu_0, \sigma_0^2)$.

Lesson 5: Properties of the Maximum Likelihood Estimator

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Maximum Likelihood estimator

Definition 11. Let $\mathbf{X} = (X_1, ..., X_n)$ be a sample from a distribution with p.d.f. $f(x; \theta)$ depending on an unknown parameter $\theta \in \Theta$. An estimator $\hat{\theta}_n(\mathbf{X}) = \hat{\theta}_n(X_1, ..., X_n)$ of θ is a **Maximum Likelihood Estimator** if for any particular realization $\mathbf{x} = (x_1, ..., x_n)$, the resulting estimate $\hat{\theta}_n(\mathbf{x}) = \hat{\theta}_n(x_1, ..., x_n) \in \Theta$ is a Maximum Likelihood estimate i.e.,

$$L(\hat{\theta}_n(\mathbf{x}); \mathbf{x}) = \max_{\theta \in \Theta} L(\theta; \mathbf{x})$$

Maximum Likelihood estimator: Properties

We will present some properties of MLE's in the context in which θ a single parameter, that is $\Theta \subset \mathbb{R}$.
Under certain regularity conditions, the maximum likelihood estimator possesses many appealing properties:

1. The maximum likelihood estimator is equivariant
2. The maximum likelihood estimator is consistent
3. The maximum likelihood estimator is asymptotically normal
4. The maximum likelihood estimator is asymptotically efficient

Maximum Likelihood estimator: Invariance

One of the most attractive properties of MLE's is invariance.

Let $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X})$ be a MLE of θ. If $g : \Theta \to \mathbb{R}$ is a continuous function, then a MLE of $g(\theta)$ exists and is given by $g(\hat{\theta}_n(\mathbf{X}))$.

For example, if $g(\theta) = \theta^2$ its MLE is $g(\hat{\theta}_n) = \hat{\theta}_n^2$.

Maximum Likelihood estimator: consistency

Definition 12. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x; \theta)$ depending on a real parameter $\theta \in \Theta$. An estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ is said to be **consistent** for θ if

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1 \quad \forall \theta \in \Theta$$

and we write $\hat{\theta}_n \xrightarrow{P} \theta$.

Maximum Likelihood estimator: consistency

Theorem 4. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x;\theta)$ depending on a real parameter $\theta \in \Theta$. Under suitable regularity conditions, the ML estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ is a consistent estimator for θ.

Maximum Likelihood estimator: asymptotic normality

Here, we consider θ a vector of parameters

Definition 13. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x;\theta)$ depending on a vector of parameters $\theta \in \Theta \subset \mathbb{R}^k$. An estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ for θ, with covariance matrix $\mathbf{V}_n(\theta)$, is said to be **asymptotically normal** if

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \overset{D}{\rightarrow} N(\mathbf{0}, \mathbf{V}(\theta))$$

where $\mathbf{V}(\theta) = \lim_{n\to\infty}\mathbf{V}_n(\theta)$

In other terms, an estimator is said to have an asymptotic normal distribution if

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right)$$

converges in distribution to a random vector with multivariate distribution $N(\mathbf{0}, \mathbf{V}(\theta))$

Maximum Likelihood estimator: asymptotic normality

We note that if $\hat{\theta}_n$ is asymptotically normal, then $\hat{\theta}_n$ is approximately distributed as a normal random vector with mean θ and covariance matrix

$$\frac{1}{n}\mathbf{V}(\theta).$$

The matrix $\frac{1}{n}\mathbf{V}(\theta)$ is called asymptotic variance.

Maximum Likelihood estimator: asymptotic normality

Theorem 5. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x;\theta)$ depending on a vector of parameters $\theta \in \Theta \subset \mathbb{R}^k$. Under suitable regularity conditions, the ML estimator $\hat{\theta}_n = \hat{\theta}_n(x_1, ..., x_n)$ is asymptotically normal. That is

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \overset{D}{\to} N(\mathbf{0}, I_a(\theta_0)^{-1})$$

where

$$I_a(\theta_0) = \lim_{n\to\infty} I_n(\theta_0)/n \quad (\textit{asymptotic information matrix})$$

and θ_0 is the true parameter value.

Maximum Likelihood estimator: asymptotic normality

Because

$$I_a(\theta_0) = \lim_{n\to\infty} \frac{I_n(\theta_0)}{n} = I_1(\theta_0),$$

we have that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \overset{D}{\to} N(\mathbf{0}, I_1(\theta_0)^{-1})$$

The asymptotic variance matrix is the inverse of the Fisher information matrix for one observation.

Maximum Likelihood estimator: asymptotic normality

The practical consequence of this result is that in large samples, when n is large enough, the ML estimator $\hat{\theta}$ has approximately a normal distribution with mean vector θ_0 and variance-covariance matrix $I_1(\theta_0)^{-1}/n$, in symbols

$$\hat{\theta} \ \ approx. \ \sim N\left[\theta_0, I_1(\theta_0)^{-1}/n\right].$$

Maximum Likelihood estimator: asymptotic efficiency

Definition 14. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x;\theta)$ depending on a vector of parameters $\theta \in \Theta \subset \mathbb{R}^k$. A consistent and asymptotically normal estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ for θ, with asymptotic variance $(1/n)\mathbf{V}(\theta)$, is said to be **asymptotically efficient** if the asymptotic variance of any other consistent, asymptotically normally distributed estimator exceeds $(1/n)\mathbf{V}(\theta)$ by a nonnegative definite matrix.

Maximum Likelihood estimator: asymptotic efficiency

Theorem 6. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from the distribution with p.d.f. $f(x;\theta)$ depending on a vector of parameters $\theta \in \Theta \subset \mathbb{R}^k$. Under suitable regularity conditions, the ML estimator $\hat{\theta}_n = \hat{\theta}_n(x_1, ..., x_n)$ is asymptotically efficient.

The MLE has the "smallest" variance among all consistent asymptotically normal estimators.

Maximum Likelihood estimator: Unbiasedness and efficiency

It is possible to show that, under some regularity conditions, if $\hat{\theta}_n(\mathbf{X})$ is an unbiased estimator of θ whose variance achieves the Cramer-Rao bound, then the likelihood equation has a unique solution equal to $\hat{\theta}_n(\mathbf{x})$.

In other terms, when there exists an unbiased estimator whose variance attains the lower bound, this estimator is identical to the ML estimator.

Lesson 6: The likelihood-based test procedures

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

Statistical Hypothesis Testing: General Aspects

As usual, we consider a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from $f(x; \theta)$. In such a context, a statistical hypothesis is a statement that asserts the unknown parameter θ belongs to a given subset Θ_0 of the parameter space Θ.

$$H : \theta \in \Theta_0, \;\; \Theta_0 \subset \Theta$$

An hypothesis of this type is called parametric hypothesis. A parametric hypothesis can be of two types:

1. Simple
2. Composite

A parametric hypothesis

$$H : \theta \in \Theta_0, \;\; \Theta_0 \subset \Theta$$

is called simple if and only if $\Theta_0 = \{\theta_0\}$ (It is a singleton set). Parametric hypotheses that are not simple are called composite.

Statistical Hypothesis Testing: General Aspects

Often, we will consider the following system of hypotheses:

$$H_0 : \theta \in \Theta_0, \;\; \Theta_0 \subset \Theta$$

versus

$$H_1 : \theta \in \Theta_0^c.$$

H_0 is called null hypothesis and H_1 is called alternative hypothesis.

Statistical Hypothesis Testing: General Aspects

Once the system of hypotheses is formulated, we need to make a decision regarding whether to reject the null hypothesis or not. This can be done by partitioning the sample space C (i.e., the set of all possible sample realizations) into two subsets C_1 and C_0, and deciding to reject H_0 if the sample realization $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in C_1$. The subset C_1 is called critical region of the test while C_0? is called acceptance region. The critical region is the set of all points of the sample space C for which the null hypothesis is rejected

$$C_1 = \{\mathbf{x} \in C : H_0 \text{ is rejected}\}$$

Statistical Hypothesis Testing: General Aspects

When we decide whether to reject or not the null hypothesis H_0, we can incur in two kinds of errors:

1. Type I error is made if H_0 is rejected when it is true
2. Type II error is made if H_0 is accepted when it is false

Statistical Hypothesis Testing: General Aspects

We observe that the probability to commit a Type I error (the probability of rejecting the null hypothesis when the null hypothesis is true) varies as θ varies within Θ_0. This probability is given by

$$P_\theta(C_1) = \int_{C_1} f(\mathbf{x};\theta)d\mathbf{x} \;\; \forall\theta \in \Theta_0$$

where the notation P_θ is used to indicate the fact that the probability is calculated using the joint probability density function $f(\mathbf{x};\theta)$ associated to the parameter θ.

Statistical Hypothesis Testing: General Aspects

The maximum probability of committing a Type I error is

$$\sup_{\theta \in \Theta_0} P_\theta(C_1)$$

This probability is called size of the critical region C_1 or level of the test.

We note that the critical region can be defined in terms of a test statistic $s = h(x_1, x_2, ..., x_n)$, requiring that $s \in B \subset \mathbb{R}$. In this case in order to denote the size of the critical region we use the notation

$$\sup_{\theta \in \Theta_0} P_\theta(s \in B)$$

The likelihood-based test procedures

Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from a distribution with p.d.f. $f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$.
Consider the vector-valued function $g : \Theta \subset \mathbb{R}^k \longrightarrow \mathbb{R}^r$

$$g(\theta) = [g_1(\theta), g_2(\theta), ..., g_r(\theta)]',$$

with $1 \leq r \leq k$.
It is assumed to be differentiable at all interior points of Θ, and the $(r \times k)$ Jacobian matrix

$$G(\theta) = \frac{\delta g}{\delta \theta'} = \left[\begin{array}{ccc} \frac{\delta g_1(\theta)}{\delta \theta_1} & ... & \frac{\delta g_1(\theta)}{\delta \theta_k} \\ . & . & . \\ \frac{\delta g_r(\theta)}{\delta \theta_1} & ... & \frac{\delta g_r(\theta)}{\delta \theta_k} \end{array} \right].$$

is assumed to have full rank r.

The likelihood-based test procedures

We want to test

$$H_0 : g(\theta) = \mathbf{0}$$

against

$$H_1 : g(\theta) \neq \mathbf{0}$$

A number of different test procedures based on ML estimators can be used.

1. Likelihood ratio test
2. Wald test
3. Lagrange multiplier test

To emphasize their key role in Statistical Inference, Rao (2005) named them **the Holy Trinity**. All three tests are asymptotically equivalent, in the sense that all the test statistics tend to the same random variable (under the null hypothesis) as the sample size tends to infinity.

The Likelihood Ratio Test

Let the likelihood function be $L(\theta; \mathbf{x})$. Consider the so-called **likelihood ratio**, defined by

$$\lambda(\mathbf{x}) = \frac{\max_{\theta \in \Theta_0} L(\theta; \mathbf{x})}{\max_{\theta \in \Theta} L(\theta; \mathbf{x})} = \frac{L(\tilde{\theta}; \mathbf{x})}{L(\hat{\theta}; \mathbf{x})}$$

where $\tilde{\theta}$ is the value of $\theta \in \Theta_0 = \{\theta \in \Theta | g(\theta) = \mathbf{0}\}$ that maximizes $L(\theta; \mathbf{x})$ and $\hat{\theta}$ is the MLE of θ. In other terms, $\tilde{\theta}$ is the restricted maximum likelihood estimator and $\hat{\theta}$ is the MLE of θ obtained without regard to the restrictions:

$$\tilde{\theta} = \arg\max_{\theta \in \Theta_0} L(\theta; \mathbf{x})$$

and

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{x})$$

The Likelihood Ratio Test

Now, we consider the test statistic, $LR = -2\ln\lambda(\mathbf{x})$. We note that if the restriction $g(\theta) = \mathbf{0}$ is valid, then the restricted estimate, $\tilde{\theta}$, should be near the point that maximizes the likelihood without any restrictions , that is $\tilde{\theta} \approx \hat{\theta}$. In fact, if $H_0 : g(\theta) = \mathbf{0}$ is true, then the restriction $g(\theta) = \mathbf{0}$ is valid at the true parameter value θ_0, that is $g(\theta_0) = \mathbf{0}$ and hence $\theta_0 \in \Theta_0$. This implies that for large sample sizes $\tilde{\theta} \approx \theta_0$. On the other hand, we have that when n is big $\hat{\theta} \approx \theta_0$ ($\hat{\theta}$ is a consistent estimator). Thus we can conclude that $\tilde{\theta} \approx \hat{\theta}$
The condition $\tilde{\theta} \approx \hat{\theta}$ implies that $\lambda(\mathbf{x}) \approx 1$. It follows that

$$LR = -2\ln\lambda(\mathbf{x}) = -2(\ln L(\tilde{\theta}; \mathbf{x}) - \ln L(\hat{\theta}; \mathbf{x})) \approx 0.$$

Therefore large values of the LR statistic provide evidence against the null hypothesis. We reject the null hypothesis when LR is "large".

The Likelihood Ratio Test

More precisely, we reject the null hypothesis, $H_0 : g(\theta) = \mathbf{0}$, if $LR \geq k_\alpha$ where k_α is such that

$$\sup_{\theta \in \Theta_0} P_\theta(LR \in [k_\alpha, +\infty)) = \alpha$$

and α is a fixed value belonging to $[0, 1]$ interval.
Now, since it is possible to show that under H_0, the distribution of the test statistic $LR = -2\ln\lambda(\mathbf{x})$ converges to a chi-square distribution where the degrees of freedom are determined as the number r of restrictions on θ, the critical value, k_α, is found from the chi-squared tables.

The Likelihood Ratio Test

Summarizing, we reject the null hypothesis

$$H_0 : g(\theta) = \mathbf{0}$$

if $LR = -2\ln\lambda(\mathbf{x}) \geq k_\alpha$, where k_α is the 100(1-α) percentile point of a Chi-Square distribution with r degree of freedom.

The Wald test

A shortcoming of the likelihood ratio test is that it requires computation of an MLE, $\hat{\theta}$, and a restricted MLE, $\tilde{\theta}$, In complex models, one or the other of these estimates may be very difficult to compute. Fortunately, there are two alternative testing procedures, the Wald test and the Lagrange multiplier test, that circumvent this problem.

The Wald test

In order to test the null hypothesis

$$H_0: \; g(\theta) = \mathbf{0}$$

Wald (1943) proposed the following quadratic form in the vector $g(\hat{\theta})$

$$W = g(\hat{\theta})' \left[G(\hat{\theta}) I(\hat{\theta})^{-1} G(\hat{\theta})' \right]^{-1} g(\hat{\theta})$$

The so-called Wald test statistic.

The Wald test

The informal argument underlying the Wald test, is as follows. We start remember that the MLE estimator is consistent estimator. This means that when n is big $\hat{\theta} \approx \theta_0$. If $H_0 : g(\theta) = \mathbf{0}$ is true, then the restriction $g(\theta) = \mathbf{0}$ is valid at the true parameter value θ_0, that is $g(\theta_0) = \mathbf{0}$. Since $\hat{\theta} \approx \theta_0$ (for large sample), we have that $g(\hat{\theta}) \approx \mathbf{0}$ and hence

$$W = g(\hat{\theta})' \left[G(\hat{\theta}) I(\hat{\theta})^{-1} G(\hat{\theta})' \right]^{-1} g(\hat{\theta}) \approx 0$$

Therefore large values of the Wald statistic provide evidence against the null hypothesis. We reject the hypothesis if W is significantly different from zero.

The Wald test

Under some regularity conditions and under the null hypothesis $H_0 : g(\theta) = \mathbf{0}$, W follows asymptotically a chi-square distribution with r degrees of freedom.

The null hypothesis is rejected if W exceeds the appropriate critical value from the chi-squared tables.

The Wald test involves only the unrestricted estimate of θ and consequently is convenient when the restricted estimate of θ is difficult to compute.

The Lagrange multiplier test or score test

The Lagrange multiplier test or Score test is an alternative to the LR and Wald tests. The motivation for it is that on occasion it can be easier to maximize the log likelihood subject to $g(\theta) = \mathbf{0}$ than to simply maximize it without restrictions .
Let $\tilde{\theta}$ be a restricted MLE (i.e. a maximizer of $\ell(\theta)$ subject to $g(\theta) = \mathbf{0}$).
The score test statistic is defined as

$$LM = s(\tilde{\theta})' I(\tilde{\theta})^{-1} s(\tilde{\theta})$$

The Lagrange multiplier test or Score test

We have seen that if the restriction $g(\theta) = \mathbf{0}$ is valid, then $\tilde{\theta} \approx \hat{\theta}$. Thus $s(\tilde{\theta}) \approx s(\hat{\theta})$. Being $s(\hat{\theta}) = \mathbf{0}$, we have that $s(\tilde{\theta}) \approx \mathbf{0}$. It follows that

$$LM = s(\tilde{\theta})' I(\tilde{\theta})^{-1} s(\tilde{\theta}) \approx \mathbf{0}$$

So that the region of rejection of the null hypothesis $H_0 : g(\theta) = \mathbf{0}$ is associated with large values of LM.
Under the null hypothesis $H_0 : g(\theta) = \mathbf{0}$, LM follows asymptotically a chi-square distribution with r degrees of freedom. The null hypothesis is rejected if LM exceeds the appropriate critical value from the chi-squared tables.

The Holy Trinity

$$H_0 : g(\theta) = \mathbf{0}$$

$$LR = -2\ln\left(\frac{L(\tilde{\theta};\mathbf{x})}{L(\hat{\theta};\mathbf{x})}\right)$$

$$W = g(\hat{\theta})'\left[G(\hat{\theta})I(\hat{\theta})^{-1}G(\hat{\theta})'\right]^{-1}g(\hat{\theta})$$

$$LM = s(\tilde{\theta})'I(\tilde{\theta})^{-1}s(\tilde{\theta})$$

The case of a scalar parameter

We refer to the usual hypotheses

$$H_0 : \theta = \theta_0$$

vs

$$H_1 : \theta \neq \theta_0$$

The case of a scalar parameter

With this specification, the ‘Holy Trinity’ becomes

$$LR = 2[\ell(\hat{\theta}; \mathbf{x}) - \ell(\theta_0; \mathbf{x})]$$

$$W = (\hat{\theta} - \theta_0)^2 I(\hat{\theta})$$

$$LM = s(\theta_0; \mathbf{x})^2 I(\theta_0)^{-1}$$

The likelihood-based test procedures: conclusions

The LR test compares the log likelihoods of a model with values of the parameter θ constrained to some value to a model where θ is freely estimated.
In contrast, the Wald test compares the parameter estimate $\hat{\theta}$ to θ_0; θ_0 is the value of θ under the null hypothesis.
Finally, the score test looks at the slope of the log likelihood when θ is constrained. That is, it looks at how quickly the likelihood is changing at the (null) hypothesized value of θ.

The likelihood-based test procedures: conclusions

The following figure illustrates what each of the three tests does.

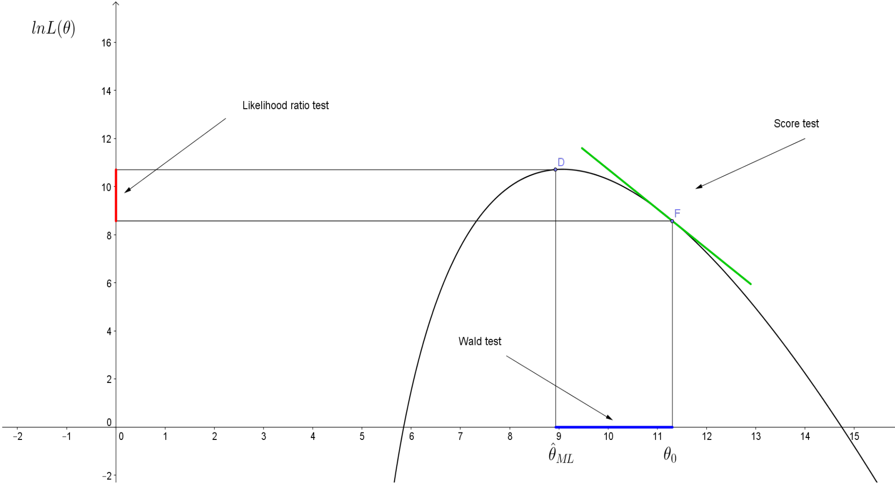


Figure: Holy Trinity. Figure based on a figure in Fox (1997, p. 570)

The likelihood-based test procedures: conclusions

These three tests are asymptotically equivalent under the null hypothesis. The choice among them is typically made on the basis of ease of computation.

1. The Likelihood Ratio test requires computation of an unrestricted MLE, $\hat{\theta}$, and a restricted MLE, $\tilde{\theta}$.
2. The Wald test requires only computation of an unrestricted MLE, $\hat{\theta}$.
3. The Lagrange Multiplier test requires only computation of a restricted MLE, $\tilde{\theta}$.

If an unrestricted MLE, $\hat{\theta}$, and a restricted MLE, $\tilde{\theta}$, both are simple to compute, then is convenient to use the Likelihood Ratio test. In some problems, one of these estimators may be much easier to compute than the other. If it is easier to calculate the unrestricted estimator, then we use the Wald test. If it is easier to calculate the restricted estimator, then we use the Lagrange Multiplier test.

Lesson 7: Likelihood ratio test, Wald test and Lagrange multiplier test in linear regression model

Umberto Triacca

Università dell'Aquila
Department of Computer Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
umberto.triacca@univaq.it

The linear regression model

Consider the following linear regression model

$$y = X\beta + \epsilon,$$

$$\epsilon \sim N(0, \sigma^2 I).$$

Here

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

The null hypothesis

Consider a set of J linear restrictions on the coefficient vector β of the form

$$H_0 : R\beta - q = \mathbf{0}$$

where R is a known $J \times k$ constant matrix of rank $J(< k)$, and q is a $J \times 1$ vector of known constants.

The likelihood ratio statistic

The LR statistic is given by:

$$LR = -2\ln \frac{\max_{R\beta=q,\sigma^2} L(\beta,\sigma^2)}{\max_{\beta,\sigma^2} L(\beta,\sigma^2)}$$

$$= n\left(\ln\hat{\sigma}_r^2 - \ln\hat{\sigma}^2\right).$$

where

$$\hat{\sigma}^2 = \frac{1}{n}e'e$$

with $e = y - Xb$

and

$$\hat{\sigma}_r^2 = \frac{1}{n}e_*'e_*$$

with $e_* = y - Xb_*$.

The Wald statistic

We have that

$$W = (Rb - q)' \left[\hat{\sigma}^2 R(X'X)^{-1}R'\right]^{-1}(Rb - q).$$

The Lagrange multiplier statistic

We have

$$LM = \frac{e_*' X(X'X)^{-1}X'e_*}{\hat{\sigma}_r^2}.$$

The W, LR and LM Inequality

An interesting relationship among the three tests statistics, when the model is linear, is the following:

$$W \geq LR \geq LM$$

That is, the Wald test statistic will always be greater than the LR test statistic, which will, in turn, always be greater than the test statistic from the score test.