Bertinoro, June 2014

Maximum Likelihood Statistical Inference

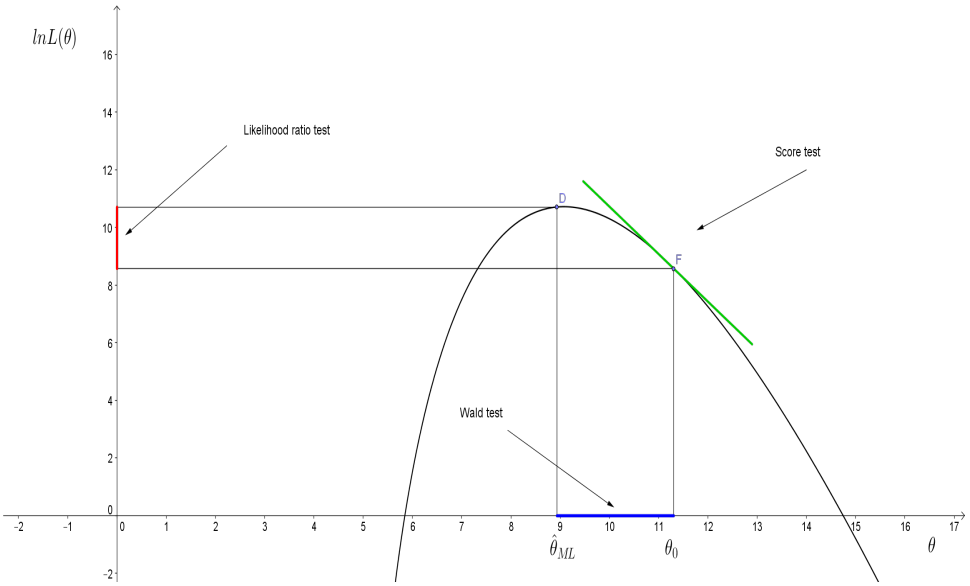

Figure 1: Holy Trinity

1. Likelihood Function
2. Score vector and information matrix
3. Cramer-Rao inequality
4. Maximum Likelihood estimator
5. The likelihood-based test procedures

1 Likelihood Function

Let $x_1, ..., x_n$ be n random variables (r.v.'s) with probability density functions (p.d.f.) $f_i(x_i; \theta)$ depending on a vector-valued parameter θ.

Often, though not always, $x_1, ..., x_n$ are assumed to be independent, identically distributed (i.i.d.) with a distribution whose probability density function is $f(x_i; \theta)$.

In this case we say that we have a **random sample** of n observations from the distribution with p.d.f. $f(x; \theta)$.

Then the joint p.d.f. of $x_1, ..., x_n$ is

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

When the sample $\mathbf{x} = (x_1, x_2, ..., x_n)$ is observed, the function of θ defined by $L(\mathbf{x}; \theta) = f(x; \theta)$ is called the **likelihood** of θ given the observations.

Thus the **likelihood function** $L(\mathbf{x}; \theta)$ is the joint p.d.f. of $x_1, ..., x_n$ viewed as a function of the unknown parameter θ.

It expresses the plausibilities of different parameters after we have observed $\mathbf{x}$, in the absence of any other information we may have about these different values. In particular, for $\theta = \theta_0$, the number $L(\theta_0$ is considered a measure of support that the observation $\mathbf{x}$ gives to the parameter θ_0.

Often we work with the natural logarithm of the likelihood function, the so-called **log-likelihood function**:

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

1.1 Examples

Example 1. Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be 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

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

and the log-likelihood function is given by

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

Example 2. Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a random sample from the exponential distribution with p.d.f.

$$f(x;\theta) = \begin{cases} \theta \mathrm{e}^{-\theta x} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

The likelihood function is

$$L(\mathbf{x};\theta) = \theta^n \mathrm{e}^{-\theta\sum_{i=1}^{n} x_i}$$

and the log-likelihood function is given by

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

2 Score vector and information matrix

2.1 Score vector

Now for some notation: given a differentiable single-valued function f, the function ∇f is defined as

$$\nabla f(\mathbf{x}) = \left(\frac{\delta f}{\delta x_1}(\mathbf{x}), ..., \frac{\delta f}{\delta x_n}(\mathbf{x})\right)'$$

and is known as the $n \times 1$ gradient vector of f.

On the other hand, by $\nabla^2 f$ we mean the $n \times n$ matrix of second partial derivatives of the function f defined as

$$\nabla^2 f = \begin{bmatrix} \frac{\delta^2 f(\mathbf{x})}{\delta x_1^2} & \frac{\delta f(\mathbf{x})}{\delta x_1 \delta x_2} & \cdots & \frac{\delta f(\mathbf{x})}{\delta x_1 \delta x_n} \\ \frac{\delta f(\mathbf{x})}{\delta x_2 \delta x_1} & \frac{\delta^2 f(\mathbf{x})}{\delta x_2^2} & \cdots & \frac{\delta f(\mathbf{x})}{\delta x_2 \delta x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta f(\mathbf{x})}{\delta x_n \delta x_1} & \frac{\delta f(\mathbf{x})}{\delta x_n \delta x_2} & \cdots & \frac{\delta^2 f(\mathbf{x})}{\delta x_n^2} \end{bmatrix}$$

This is also known as the Hessian matrix, which is often denoted as $H(\mathbf{x})$ or as $\frac{\delta^2 f(\mathbf{x})}{\delta \mathbf{x} \delta \mathbf{x}'}$.

Having introduced the log likelihood function, we can now define some concepts related to it.

Definition 1. (Score function) If the likelihood function is differentiable, then the gradient of the log-likelihood

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

is called **the score function**.

Example. Let $\mathbf{x} = (x_1, x_2, ..., x_n)'$ be a random sample from an $N(\mu, \sigma^2)$ distribution. Here $\theta = (\mu, \sigma^2)$. The score function is given by

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

Remark. The proofs of this section are based on a double interpretation of the function $f(\mathbf{x}; \theta)$ for a fixed θ.

- It must be considered a probability density function,

but at the same time, being a transformation of the random vector $\mathbf{x}$,

- it is a random variable itself, with expectation and variance.

The same double interpretation holds for the logarithm of $f(\mathbf{x}; \theta)$, as well as its derivatives. Thus if we consider the function $f(\mathbf{x}; \theta)$ like a random vector, the score function becomes a random vector. We call this random vector **score vector**.

Theorem 1. The score vector evaluated at the true parameter value has mean zero

Proof. As the function $f(\mathbf{x}; \theta)$ is a probability density functions we have that:

$$\int_{-\infty}^{+\infty} f(\mathbf{x}; \theta) d\mathbf{x} = 1 \ \ \forall \theta \tag{1}$$

This is a multiple integral with respect to $x_1, x_2, ..., x_n$

Thus, differentiating (1) w.r.t. θ we get

$$\frac{\delta}{\delta \theta} \left[\int_{-\infty}^{+\infty} f(\mathbf{x}; \theta) d\mathbf{x} \right] = 0 \tag{2}$$

We assume that f satisfies some regularity conditions that permit differentiation under integral (for instance, it is twice differentiable w.r.t. θ and the limits of integration are not functions of θ).

So, (2) can be written

$$\int_{-\infty}^{+\infty} \frac{\delta f(\mathbf{x}; \theta)}{\delta \theta} d\mathbf{x} = 0 \tag{3}$$

Integration will be confined to the region where f assumes nonzero (positive) values. Thus (3) can be written

$$\int \frac{\delta \ln f(\mathbf{x}; \theta)}{\delta \theta} f(\mathbf{x}; \theta) d\mathbf{x} = 0 \tag{4}$$

If derivative is computed at the *true* parameter value, so that $f(\mathbf{x}; \theta)$ is the probability density of the r.v. $\mathbf{x}$, we have that

$$\int \frac{\delta \ln f(\mathbf{x}; \theta)}{\delta \theta} f(\mathbf{x}; \theta) d\mathbf{x} = E \left[\frac{\delta \ln f(\mathbf{x}; \theta)}{\delta \theta} \right] \tag{5}$$

By equation (4) it follows that

$$E \left[\frac{\delta \ln f(\mathbf{x}; \theta)}{\delta \theta} \right] = 0 \tag{6}$$

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

2.2 Information matrix

Definition 2. The variance-covariance matrix of the score, evaluated at the true parameter value θ,

$$I(\theta;\mathbf{x}) = Var\left[\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right] = E\left[\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta'}\right] \tag{7}$$

is called **information matrix** (more precisely, Fisher's information measure on θ contained in the r.v. $\mathbf{x}$).

The following theorem establishes an important result.

Theorem 2.

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

The information matrix equals the negative of the expected value of the Hessian of the log likelihood evaluated at the true parameter value θ.

Proof. Further differentiation of (4) gives

$$\int \left[\frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta\delta\theta'} f(\mathbf{x};\theta) + \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\frac{\delta f(\mathbf{x};\theta)}{\delta\theta'}\right] d\mathbf{x} = 0 \tag{8}$$

that is

$$\int \frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta\delta\theta'} f(\mathbf{x};\theta) d\mathbf{x} + \int \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta'} f(\mathbf{x};\theta) d\mathbf{x} = 0 \tag{9}$$

Again, because derivatives are computed at the *true* parameter value, so that $f(\mathbf{x};\theta)$ is the probability density of the r. v. $\mathbf{x}$, the two terms in equation (9) are expectations, so

$$E\left[\frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta\delta\theta'}\right] + E\left[\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta'}\right] = 0 \tag{10}$$

The second term of the sum is the information matrix (7). Thus, from (10) we get an alternative expression for the information matrix

$$I(\theta) = E\left[-\frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta\delta\theta'}\right] \tag{11}$$

that is the expected Hessian of the log-density, with the opposite sign.

Suppose that we have a **random sample** $\mathbf{x} = (x_1, x_2, ..., x_n)'$ from a probability distribution with density function, $f(x;\theta)$, characterized by a parameter (vector) θ.

The joint density of the sample is

$$f(\mathbf{x},\theta) = f(x_1; x_2, ..., x_n; \theta) = f(x_1;\theta) f(x_2;\theta) ... f(x_n;\theta);$$

The log-density of the sample will be therefore the sum of the log-densities, that is

$$\ln f(\mathbf{x};\theta) = \ln f(x_1;\theta) + \ln f(x_2;\theta) + ... + \ln f(x_n;\theta)$$

while its first and second derivatives as well as their expectations will be sums of the corresponding derivatives or expectations.

$$\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} = \frac{\delta \ln f(x_1;\theta)}{\delta\theta} + \frac{\delta \ln f(x_2;\theta)}{\delta\theta} + ... + \frac{\delta \ln f(x_n;\theta)}{\delta\theta}$$

$$\frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta\delta\theta'} = \frac{\delta^2 \ln f(x_1;\theta)}{\delta\theta\delta\theta'} + \frac{\delta^2 \ln f(x_2;\theta)}{\delta\theta\delta\theta'} + ... + \frac{\delta^2 \ln f(x_n;\theta)}{\delta\theta\delta\theta'}$$

As a straightforward consequence, the expectation of the **score of the sample** will be zero

$$E\left[\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right] = E\left[\frac{\delta \ln f(x_1;\theta)}{\delta\theta}\right] + ... + E\left[\frac{\delta \ln f(x_n;\theta)}{\delta\theta}\right] = 0$$

while the **information in the whole sample** will be

$$E\left[-\frac{\delta^2 \ln f(\mathbf{x};\theta)}{\delta\theta\delta\theta'}\right] = E\left[-\frac{\delta^2 \ln f(x_1;\theta)}{\delta\theta\delta\theta'}\right] + ... + E\left[-\frac{\delta^2 \ln f(x_n;\theta)}{\delta\theta\delta\theta'}\right]$$

Let

$$i(\theta; x_i) = E\left[\frac{\delta \ln f(x;\theta)}{\delta\theta}\frac{\delta \ln f(x;\theta)}{\delta\theta'}\right]$$

This is by definition, the information provided by a single observation. We have that

$$i(\theta; x_i) = E\left[-\frac{\delta^2 \ln f(x;\theta)}{\delta\theta\delta\theta'}\right]$$

It follows that

$$I(\theta; \mathbf{x}) = ni(\theta; x_i)$$

The information in the whole sample is n time the information provided by a single observation.

Remark. If $f(\mathbf{x};\theta)$ is a strictly positive function whose integral is identically $= 1$ for any θ, but for no value of θ it is the probability density function of the r.v. $\mathbf{x}$, all the above identities involving integrals are still valid, but they cannot be interpreted as expected values.

3 Cramer-Rao inequality

The next theorem introduce the celebrated Cramer-Rao inequality.

Theorem 3. Let $\mathbf{x} = (x_1, ..., x_n)$ be a random sample of n observations from the distribution with p.d.f. $f(x;\theta)$ depending on a real parameter θ. 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

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

The variance of any unbiased estimator is greater than or equal to the inverse of information matrix.

Proof

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

$$E[T(\mathbf{x})] = \int T(\mathbf{x}) f(\mathbf{x};\theta) d\mathbf{x} = \theta \tag{12}$$

Differentiating both sides of equation (12) with respect to θ, and interchanging the order of integration and differentiation, gives

$$\int T(\mathbf{x}) \frac{\delta f(\mathbf{x};\theta)}{\delta\theta} d\mathbf{x} = 1 \tag{13}$$

or

$$\int T(\mathbf{x}) \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} f(\mathbf{x};\theta) d\mathbf{x} = 1 \tag{14}$$

Because

$$\int T(\mathbf{x}) \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} f(\mathbf{x};\theta) d\mathbf{x} = E\left[T(\mathbf{x}) \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right] \tag{15}$$

by (14) it follows that

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

On the other hand, since

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

we have that

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

Hence

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

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

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

or

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

It follows that

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

Definition 3 (Efficiency). An unbiased estimator is efficient if its variance is the lower bound of the inequality:

$$\text{Var}[T(\mathbf{x})] = \frac{1}{E\left[\left(\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right)^2\right]} = \frac{1}{ni(x;\theta)}$$

We now give a lemma which allows us to establish when the Cramer-Rao lower bound is attainable.

Lemma 1. Under the same regularity conditions as for Cramer-Rao inequality, there exists an unbiased estimator $T(\mathbf{x})$ whose variance attains Cramer-Rao lower bound if and only if

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

Proof. We have

$$\left(\text{Cov}\left[T(\mathbf{x}), \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right]\right)^2 \leq \text{Var}\left[T(\mathbf{x})\right] \text{Var}\left[\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right]$$

The bound will be attained if and only if equality is achieved here. On the other hand equality occurs if and only if $T(\mathbf{x})$ and $\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}$ are linearly related, that is if and only if

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

where c and d are constants.

Taking expectations in this equations gives

$$0 = c + d\theta$$

that is

$$c = -d\theta.$$

It follows that

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

Now multiply the last equation by $\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}$ and take expectations

$$E\left[\left(\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right)^2\right] = dE\left[\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} T(\mathbf{x})\right] - d\theta E\left[\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right].$$

Since

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

and

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

we can conclude that

$$d = E\left[\left(\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right)^2\right] = I(\theta;\mathbf{x}).$$

Example Consider a random sample $x = (x_1, ..., x_2)'$ from the Poisson distribution. Does the sample mean $\bar{x}$ achieve the Cramer-Rao lower bound?

The likelihood is given by

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

and the log-likelihood is

$$l(\theta;\mathbf{x}) = -n\theta + \sum_{i=1}^{n} x_i \ln\theta - \ln\left(\prod_{i=1}^{n} x_i!\right).$$

Hence

$$\frac{\delta l}{\delta\theta} = -n\theta + \sum_{i=1}^{n} x_i/\theta = \frac{n}{\theta}(\bar{x} - \theta),$$

$$\frac{\delta^2 l}{\delta\theta^2} = -\sum_{i=1}^{n} x_i/\theta^2.$$

and

$$I(\theta;\mathbf{x}) = -E\left[\frac{\delta^2 l}{\delta\theta^2}\right] = \frac{n\theta}{\theta} = \frac{n}{\theta}.$$

Thus we can conclude that, in this case, the sample mean $\bar{x}$ achieves the Cramer-Rao lower bound.

3.1 Multidimensional Cramer-Rao inequality

The Cramer-Rao inequality (Theorem 3) can be generalized to a vector valued parameter θ.

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 $\mathrm{Var}(T(\mathbf{x}) - I^{-1}(\theta;\mathbf{x})$ is positive semi-definite.

When θ is a $(k \times 1)$ vector of parameters, analogously to before, we have

$$I_k = E\left[T(\mathbf{x})\frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta'}\right] = \mathrm{Cov}\left[T(\mathbf{x}), \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta'}\right]$$

Now, we consider the following $(2k \times 1)$ vector

$$\begin{bmatrix} T(\mathbf{x}) \\ \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} \end{bmatrix}$$

The variance-covariance matrix of this vector is given by

$$\text{Var}\begin{bmatrix} T(\mathbf{x}) \\ \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta} \end{bmatrix} = \begin{bmatrix} \text{Var}[T(\mathbf{x})] & \text{Cov}\left[T(\mathbf{x}), \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta'}\right] \\ \text{Cov}\left[T(\mathbf{x}), \frac{\delta \ln f(\mathbf{x};\theta)}{\delta\theta}\right] & I(\theta;\mathbf{x}) \end{bmatrix}$$

$$= \begin{bmatrix} \text{Var}[T(\mathbf{x})] & I_k \\ I_k & I(\theta;\mathbf{x}) \end{bmatrix}$$

which is positive semi definite, being a variance-covariance matrix $(2k \times 2k)$.

Thus, pre- and post-multiplication by a matrix and its transpose still provides a positive semi definite matrix. In particular, if the information matrix is not singular (i.e. the derivatives of the log-density are not linearly dependent), pre-multiplication by the $(k \times 2k)$ matrix

$$\left[I_k; -I^{-1}(\theta;\mathbf{x})\right]$$

and post-multiplication by its transpose produces

$$\text{Var}\left(T(\mathbf{x})\right) - I^{-1}(\theta;\mathbf{x})$$

which is a positive semi definite matrix.

3.2 Linear regression model with normal error terms

Consider the following linear regression model

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

where y is an $n \times 1$ vector of observations on the dependent variable, X is an $n \times k$ matrix of observations on the non stochastic exogenous variables, β is a $k \times 1$ vector of unknown regression coefficients, and ϵ is an $n \times 1$ vector consisting of errors. We assume that ϵ is distributed as multivariate normal with mean vector zero and variance covariance matrix $\sigma^2 I$, where I is an $n \times n$ identity matrix,

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

here

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

Being the Jacobian of the transformation $|\frac{\delta \epsilon_i}{\delta y_i}| = 1$ (so that the density $f(y_i;\theta) = f(\epsilon_i;\theta)$) the log-likelihood is

$$\ln L(y;\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

We note that the log-likelihood is a quadratic in the vector β. The score is

$$s(\theta) = \frac{\delta \ln L(y;\theta)}{\delta\theta} = \begin{bmatrix} \frac{\delta \ln L(y;\theta)}{\delta\beta} \\ \frac{\delta \ln L(y;\theta)}{\delta\sigma^2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sigma^2}(X'X\beta - X'y) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) \end{bmatrix}$$

and the Hessian matrix is

$$H(\theta) = \frac{\delta^2 \ln L(y;\theta)}{\delta\theta} = \begin{bmatrix} -\frac{1}{\sigma^2}X'X & \frac{1}{\sigma^4}(X'X\beta - X'y) \\ \frac{1}{\sigma^4}(X'X\beta - X'y) & \frac{n}{2\sigma^4} - \frac{1}{2\sigma^6}(y - X\beta)'(y - X\beta) \end{bmatrix}.$$

The expectation of the off-diagonal blocks of the Hessian matrix is zero, and the expectation of the last block is $n/(2\sigma^4 - (1/\sigma^6)n\sigma^2 = -n/(2\sigma^4)$. So, the information matrix is

$$ni(\theta) = E[-H(\theta)] = \begin{bmatrix} \frac{1}{\sigma^2}X'X & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{bmatrix}.$$

and its inverse (the Cramer-Rao bound for the covariance matrix of any unbiased estimator) is

$$[ni(\theta)]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma^4}{n} \end{bmatrix}.$$

The covariance matrix of coefficients estimated by OLS is $(X'X)^{-1}\sigma^2$, so OLS coefficients attain the Cramér-Rao bound.

But the OLS estimator of σ^2 does not attain the bound. In fact, remembering that $\hat{\sigma}^2/\sigma^2$ is a random variable χ^2_{n-k} divided by $n-k$, and that the variance of the χ^2_{n-k} is $2(n-k)$, we get:

$$V(\hat{\sigma}^2) = \left[\frac{\sigma^2}{n-k}\right]^2 V\left(\chi^2_{n-k}\right) = \frac{2\sigma^4}{n-k}$$

which is larger than the Cramér-Rao bound (however, it is not possible to find an unbiased estimator of σ^2 with a smaller variance; see Rao, 1973, 5a.2).

Remark. If the Hessian (**??**) is *estimated*, that is it is computed at the OLS estimated parameters $\hat{\beta}$ and $\hat{\sigma}^2$, the off diagonal blocks are zero ($X'X\hat{\beta} - X'y = -X'\hat{u} = 0$).

Remark. Obviously, the *good* properties of the OLS estimator just described are no more valid if some elements of x_i are correlated with u_i. In principle, the likelihood should be re-specified, to take explicitly into account the correlation, and maximum likelihood would be different from the simple OLS estimator.

4 Maximum Likelihood estimator

Let $\mathbf{x} = (x_1, ..., x_n)'$ be a random sample of n observations from the distribution with p.d.f. $f(x;\theta)$ depending on a parameter θ. The set of values that θ could take is called the parametric space and denoted by Θ.

Definition 4. A Maximum Likelihood estimator of $\theta \in \Theta$ is a measurable function

$$\hat{\theta} : \mathbb{R}^n \longrightarrow \Theta$$

such that

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

Remark. Of course it is possible, if, for example, Θ is an open set, that no maximum likelihood estimator exists.

Proposition 1 (Sufficient condition for existence). If the parameter space Θ is compact and if the likelihood function $L(\mathbf{x};\theta)$ 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 $L(\mathbf{x};\theta)$ is strictly concave in Θ, then the MLE is unique when it exists.

Since the natural logarithm is a monotonic transformation, it follows that

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

$$\Updownarrow$$

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

The values that maximize $L(\mathbf{x};\theta)$ are the same as those that maximize $\ln L(\mathbf{x};\theta)$.

It is usually possible to assume that the MLE emerges as a solution of the equation

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

This is called the **likelihood equation**.

The likelihood equation often have to be solved numerically. A standard method of solving the likelihood equation is Newton's method or an adaptation of it.

Remark. The method of maximum likelihood is applicable mainly in situations where the true distribution on the sample space is known apart from the values of a finite number of unknown real parameters.

Example 1. Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a random sample from an $N(\mu, \sigma^2)$ distribution. The log-likelihood function is

$$\ell(\theta \mid \boldsymbol{x}) = -\frac{\Sigma(x_i-\mu)^2}{2\sigma^2} - \frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi).$$

Taking the first derivative (gradient), we get

$$\frac{\partial \ell}{\partial \theta} = (\frac{\Sigma(x_i-\mu)}{\sigma^2}, \frac{\Sigma(x_i-\mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2}).$$

Setting

$$\frac{\partial \ell}{\partial \theta} = 0$$

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.

Example 2. Consider the following linear regression model

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

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

Here

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

We have seen that, in this case, the log-likelihood is

$$\ln L(y;\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

The score is

$$s(\theta) = \frac{\delta \ln L(y;\theta)}{\delta\theta} = \begin{bmatrix} \frac{\delta \ln L(y;\theta)}{\delta\beta} \\ \frac{\delta \ln L(y;\theta)}{\delta\sigma^2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sigma 2}(X'X\beta - X'y) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) \end{bmatrix}$$

The maximum likelihood estimators for β and σ^2 are obtained solving the following system

$$\frac{\delta \ln L(y;\theta)}{\delta\beta} = -\frac{1}{\sigma 2}(X'X\beta - X'y) = 0$$

$$\frac{\delta \ln L(y;\theta)}{\delta\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta)$$

The maximum likelihood (ML) estimator for β is the ordinary least squares (OLS) estimator given by

$$b = (X'X)^{-1}X'y$$

The ML estimator of σ^2 is given by

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

where $e = y - Xb$.

Substituting b and $\hat{\sigma}^2$ in the log-likelihood function and exponentiating gives the maximum of the likelihood function as

$$L(b, \hat{\sigma}^2) = (2\pi)^{-n/2}\exp\left(-\frac{n}{2}\right)(\hat{\sigma}^2)^{-n/2} = \left(\frac{2\pi}{n}\right)^{-n/2}\exp\left(-\frac{n}{2}\right)(e'e)^{-n/2}$$

4.1 Maximum Likelihood estimator: consistency

Here, we consider θ a single parameter, that is $\Theta \subset \mathbb{R}$.

Theorem 4. Let $x_1, ..., x_n$ be i.i.d. with probability density satisfying suitable regularity conditions. Then there exists a sequence $\hat{\theta}_n = \hat{\theta}_n(x_1, ..., x_n)$ of local maxima of the likelihood function $L(x_1, ..., x_n; \theta)$ which is consistent.

Proof. We denote with θ_0 the true parameter value. Applying first order Taylor expansion to the score, with initial point θ_0 we get

$$\frac{\delta \ln f(x_i;\theta)}{\delta\theta} = \left[\frac{\delta \ln f(x_i;\theta)}{\delta\theta}\right]_{\theta_0} + \left[\frac{\delta^2 \ln f(x_i;\theta)}{\delta\theta^2}\right]_{\theta_0} (\theta-\theta_0) + \mathrm{R}(x_i;\theta;\theta_0)$$

Summing for $i = 1, 2, ..., n$ and dividing by n (averaging) we get

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

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta \ln f(x_i;\theta)}{\delta\theta}\right]_{\theta_0} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta^2 \ln f(x_i;\theta)}{\delta\theta^2}\right]_{\theta_0} (\theta-\theta_0)$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathrm{R}(x_i;\theta;\theta_0) \tag{16}$$

Some suitable form of the weak law of large numbers (WLLN) ensures that

$$\mathrm{plim}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta \ln f(x_i;\theta)}{\delta\theta}\right]_{\theta_0} = 0$$

and

$$\mathrm{plim}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta^2 \ln f(x_i;\theta)}{\delta\theta^2}\right]_{\theta_0} = -i(\theta_0;x)$$

Thus, for a conveniently large n, the first term on the right hand side of (16) will be negligible, while the second term will be negative if $(\theta-\theta_0)$ is positive, and will be positive if $(\theta-\theta_0)$ is negative.

Concerning the residual term, for large n and small $(\theta-\theta_0)$, regularity conditions and Taylor expansion properties ensure that its contribution is negligible with respect to the other terms.

The consequence is that, when n is large enough, analyzing an arbitrarily small interval around θ_0, the left hand side of (16) is positive on the left of θ_0, negative on the right: thus, arbitrarily close to θ_0 there is a point where the log-likelihood has a local maximum (and the score is zero).

Remark 1.. Theorem 3 asserts the consistence not of the MLE but of a suitable sequence of local maxima of the likelihood

Remark 2. Theorem 3 does not guarantee the existence of a local maximum for all $(x_1, ..., x_n)$.

Corollary 1. Under the assumptions of Theorem 4, if the likelihood equation has a unique root $\hat{\theta}_n = \hat{\theta}_n(x_1, ..., x_n)$ for each n and all $(x_1, ..., x_n)$, then

i. $\hat{\theta}_n$ is a consistent estimator

and

ii. with probability tending to 1 as $n \to \infty$, $\hat{\theta}_n$ is the MLE.

4.2 Maximum Likelihood estimator: asymptotic normality

In this subsection we will utilize the following results.

Proposition 3. If $x_n \xrightarrow{D} x$ and plim$y_n = c$. Then, $x_n y_n \xrightarrow{D} cx$. That is the limiting distribution of $x_n y_n$ is the distribution of cx.

Proposition 4. If $\mathbf{y} \sim N(\mu, \Sigma)$ and C is a $(p \times n)$ constant matrix of rank p, then $C\mathbf{y} \sim N(C\mu, C\Sigma C')$.

Considering again θ a vector of parameters, that is $\Theta \subset \mathbb{R}^p$. If

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

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta \ln f(x_i;\theta)}{\delta\theta}\right]_{\theta_0} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta^2 \ln f(x_i;\theta)}{\delta\theta^2}\right]_{\theta_0}(\theta-\theta_0)$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathrm{R}(x_i;\theta;\theta_0)$$

is computed at $\hat{\theta}$, the left hand side is zero.

Thus we have

$$-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta^2 \ln f(x_i;\theta)}{\delta\theta^2}\right]_{\theta_0}\left(\hat{\theta}-\theta_0\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta \ln f(x_i;\theta)}{\delta\theta}\right]_{\theta_0}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathrm{R}(x_i;\hat{\theta};\theta_0)$$

Multiplying by $\sqrt{n}$ we get

$$-\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta^2 \ln f(x_i;\theta)}{\delta\theta^2}\right]_{\theta_0}\left(\hat{\theta}-\theta_0\right) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta \ln f(x_i;\theta)}{\delta\theta}\right]_{\theta_0}$$

$$+\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathrm{R}(x_i;\hat{\theta};\theta_0)$$

It follows that

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) = \left[-\sum_{i=1}^{n}\frac{\delta^2 \ln f(x_i;\theta)}{\delta\theta^2}\right]_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\delta \ln f(x_i;\theta)}{\delta\theta}\right]_{\theta_0}$$

$$+\left[-\sum_{i=1}^{n}\frac{\delta^2 \ln f(x_i;\theta)}{\delta\theta^2}\right]_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{R}(x_i;\hat{\theta};\theta_0)$$

When $n \to \infty$ (and therefore $\hat{\theta} \to \theta_0$) still the contribution of the residual term becomes negligible.

Thus, for a conveniently large n, we have

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)=\left[-\sum_{i=1}^{n}\frac{\delta^2 \ln f(x_i;\theta)}{\delta\theta^2}\right]_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\delta \ln f(x_i;\theta)}{\delta\theta}\right]_{\theta_0}$$

The term with second order derivatives, it converges in probability to the information matrix $I(\theta_0)$ and some suitable form of the Central Limit Theorem (CLT) ensures that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\delta \ln f(x_i;\theta)}{\delta\theta}\right]_{\theta_0}\xrightarrow{D}N(0,I(\theta_0))$$

Thus, by Propositions 3 and 4, we can conclude that

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{D}N(0,I(\theta_0)^{-1})\quad(+)$$

The practical consequence of (+) is that, when n is large enough, $\sqrt{n}(\hat{\theta}-\theta_0)$ has approximately a normal distribution with zero mean and $I(\theta_0)^{-1}$ variance-covariance matrix. Thus $(\hat{\theta}-\theta_0)$ has approximately a normal distribution with zero mean and $I(\theta_0)^{-1}/n$ variance-covariance matrix, that is

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

Practical estimation of the information matrix can be done in two different ways, using the sample analogues of the *expectations* on the right hand sides of (7) or (11): each expectation is replaced by the sample average, and derivatives are computed at $\hat{\theta}$

1. Hessian estimator of $I(\theta_0)=\frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\partial^2 lnf(x_i,\theta)}{\partial\theta\partial\theta'}\right]_{\hat{\theta}}=\frac{1}{n}\left[-\frac{\partial^2 lnL(x,\theta)}{\partial\theta\partial\theta'}\right]_{\hat{\theta}}$
2. Outer Product estimator of $I(\theta_0)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\frac{\partial lnf(x_i,\theta)}{\partial\theta'}\right]_{\hat{\theta}}$

As a consequence, also the practical estimation of the variance-covariance matrix of $\hat{\theta}$ can be done in two different ways: using the Hessian or using the Outer Product matrix

1. $\widehat{Var}(\hat{\theta})=(H)^{-1}=\left[-\frac{\partial^2 lnL(x,\theta)}{\partial\theta\partial\theta'}\right]_{\hat{\theta}}^{-1}$
2. $\widehat{Var}(\hat{\theta})=(OP)^{-1}=\left\{\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\frac{\partial lnf(x_i,\theta)}{\partial\theta'}\right]_{\hat{\theta}}\right\}^{-1}$

Summarizing, we have that, under suitable regularity conditions, the maximum likelihood estimator is

- consistent
- asymptotically normal
  - with mean equal to the true parameter value
  - and variance-covariance matrix equal to the inverse of the information matrix.

Therefore, if an estimator is an ML estimator and the regularity conditions are satisfied, it is not necessary to show that it is consistent or derive its asymptotic distribution.

These properties provide the main justification of the method of maximum likelihood.

4.3 Nonlinear regression model: maximum likelihood and nonlinear least squares

Let the model and the vector of parameters be

$$y_i = q(x_i, \beta) + u_i \qquad u_i \ i.i.d. \ N(0, \sigma^2) \qquad i = 1, 2, ..., n \qquad \theta = \begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix} \quad (17)$$

where q is a nonlinear function of the explanatory variables x_i and of the coefficients β, satisfying some *regularity* conditions (continuity and differentiability). Almost all properties of the linear regression with normal errors apply to the nonlinear regression as well. The only difference is that, unlike the linear case, estimation of coefficients usually requires the application of a numerical technique (e.g. Newton-Raphson or similar), as it cannot be done in closed form.

Being the Jacobian of the transformation $\partial u_i / \partial y_i = 1$ (so that the density $f(y_i, \theta) = f(u_i, \theta)$) the log-likelihood is

$$lnL(y, \theta) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - q(x_i, \beta)]^2 \quad (18)$$

the score is

$$\frac{\partial lnL(y, \theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial lnL}{\partial \beta} \\ \frac{\partial lnL}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{n} [y_i - q(x_i, \beta)] \frac{\partial q(x_i, \beta)}{\beta} \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} [y_i - q(x_i, \beta)]^2 \end{bmatrix} \quad (19)$$

thus the system of first order conditions is

$$\begin{bmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{n} [y_i - q(x_i, \beta)] \frac{\partial q(x_i, \beta)}{\beta} = 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} [y_i - q(x_i, \beta)]^2 = 0 \end{bmatrix} \quad (20)$$

Solution of the last equation gives

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - q(x_i, \beta)]^2 \quad (21)$$

that can be substituted into (18) producing the *concentrated* log-likelihood

$$lnL(y, \beta) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} - \frac{n}{2} ln \left\{ \frac{1}{n} \sum_{i=1}^{n} [y_i - q(x_i, \beta)]^2 \right\} \quad (22)$$

There is no more the parameter σ^2, so the concentrated log-likelihood has to be maximized only with respect to the coefficients β. From equation (22) it is clear that the maximum of the concentrated log-likelihood is the minimum of the sum of squared errors $[y_i - q(x_i, \beta)]^2$; thus maximum likelihood is *nonlinear least squares*.

After β has been estimated minimizing (with some numerical technique) the sum of squared errors, the estimate of σ^2 is obtained from (21); it is the average of the squared residuals, analogously to the linear regression case.

Remark. Rather than minimizing the sum of squared residuals, one could minimize "$n/2$ ln of the average of the squared residuals" (as in equation 22), using the Newton-Raphson procedure (at least in the last iterations). The coefficient estimates would obviously be the same, but there would be no need of any further calculation to estimate the variance-covariance matrix of the coefficients: it would simply be the inverse of the last iteration's Hessian matrix.

However, from a computational viewpoint, convergence of the Newton-Raphson procedure is usually faster when the method is applied to the sum of squared residuals. Thus, it might be more convenient to split the procedure in two parts: first compute coefficients minimizing the sum of squared residuals; then, when convergence has been achieved, compute (and invert) the Hessian of "$n/2$ ln of the average of the squared residuals" as an estimate of the coefficients variance-covariance matrix.

5 The likelihood-based test procedures

Consider $(x_1, ..., x_n)$, a random sample from a distribution with p.d.f. $f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}^p$, and suppose that we wish to test

$$H_0 : \theta \in \Theta_0$$

against

$$H_1 : \theta \in \Theta - \Theta_0$$

In the present framework Θ_0 is defined by an equality restriction

$$g(\theta) = 0$$

where g is a $r \times 1$ vector of functions and $1 \leq r \leq p$, that is

$$\Theta_0 = \{\theta \in \Theta | g(\theta) = 0\}$$

g is assumed to be differentiable at all interior points of Θ, and the $(r \times p)$ Jacobian matrix

$$G(\theta) = \frac{\delta g}{\delta \theta'}$$

is assumed to have full rank r, at least in an open neighborhood of true parameter θ_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

The Lagrange multiplier test was introduced in Rao (1948) as an alternative to the likelihood ratio test of Neyman and Pearson (1928) and Wald (1943) test. These three tests are known as the *Holy Trinity*

5.1 The Likelihood Ratio Test

Let the likelihood function be $L(\theta; \mathbf{x})$, then, in order to test the hypothesis, we can consider the **likelihood ratio** defined by

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

The numerator measures the highest 'support' $\mathbf{x}$ renders to $\theta \in \Theta_0$ and the denominator measures the maximum value of the likelihood function. By definition $\lambda(\mathbf{x})$ can never exceed unity and smaller it is the less H_0 is 'supported' by the data. This suggests that the critical region based on $\lambda(\mathbf{x})$ must be of the form

$$\lambda(\mathbf{x}) < k_\alpha \qquad 0 \leq k_\alpha \leq 1$$

where k_α is such that

$$\sup_{\theta \in \Theta_0} P_\theta \left(\lambda(\mathbf{x}) < k_\alpha\right) = \alpha$$

where the function $P_\theta \left(\lambda(\mathbf{x}) < k_\alpha\right)$, defined on Θ_0, describes the probabilities of the Type I first error.

A Problem. The exact null distribution of $\lambda(\mathbf{x})$ is often difficult to obtain. However, under certain regularity conditions, the distribution of minus twice the log likelihood ratio

$$LR = -2\ln\lambda(\mathbf{x})$$

converges to a chi-square distribution where the degrees of freedom are determined as the number r of restrictions on θ required to define Θ_0.

If the restriction $\theta \in \Theta_0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function

In other terms, we have that

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

and hence

$$LR = -2\ln\lambda(\mathbf{x}) \approx 0$$

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

5.2 The Wald test

A practical shortcoming of the likelihood ratio test is that it usually requires estimation of both the restricted and unrestricted parameter vectors. 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. In particular, the Wald test involves only the unrestricted estimate of θ and consequently is convenient when the restricted estimate of θ is difficult to compute.

In order to test the null hypothesis

$$H_0 : \ g(\theta) = 0$$

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 informal argument underlying the Wald test, is as follows. We calculate unrestricted maximum-likelihood estimates of the unknown parameters (which for large samples are likely to be near the corresponding true parameters). If the restrictions are true, then the unrestricted estimates should come close to satisfying the restrictions and hence $g(\hat{\theta})$ should be close to zero. Therefore large values of the Wald statistic provide evidence against the null hypothesis. We reject the hypothesis if W is significantly different from zero.

Under some regularity conditions and under the null hypothesis $H_0 : g(\theta) = 0$, 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.

To summarize, the Wald test is based on measuring the extent to which the unrestricted estimates fail to satisfy the hypothesized restrictions.

5.3 Lagrange multiplier test

Wald tests offer the advantage of only requiring estimates of the unconstrained model, whereas likelihood ratio tests require estimates of both the unconstrained and the constrained models. A third class of tests, Lagrange multiplier tests, only require estimates of the constrained model.

Suppose that we maximize the log-likelihood subject to the set of constraints $g(\theta) = 0$. Let λ be a vector of Lagrange multipliers and define the Lagrangian function

$$L^*(\theta) = \ln L(\theta) + \lambda' g(\theta)$$

The set of first order conditions can be expressed as:

$$\frac{\delta L^*(\theta)}{\delta \theta} = \frac{\delta \ln L(\theta)}{\delta \theta} + G'\lambda = 0$$

$$\frac{\delta L^*(\theta)}{\delta \lambda} = g(\theta) = 0$$

where

$$G = G(\theta) = \frac{\delta g(\theta)}{\delta \theta'}.$$

The solution of this system provides the maximum likelihood estimate $\tilde{\theta}$ of θ subject to the $r \times 1$ vector of constraints $g(\theta) = 0$, and the estimate $\tilde{\lambda}$ of λ. If the restriction is valid, then the restricted estimate, $\tilde{\theta}$, should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. Hence, the test can be based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction. The score test statistic is defined as

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

If the constraints are true, $s(\tilde{\theta})$ should be near to the null vector, so that the region of rejection of the null hypothesis $H_0 : g(\theta) = 0$ is associated with large values of LM.

Under the null hypothesis $H_0 : g(\theta) = 0$, ML follows asymptotically a chi-square distribution with r degrees of freedom. The null hypothesis is rejected if ML exceeds the appropriate critical value from the chi-squared tables.

Why is the name of this test Lagrange multiplier test? We note that the vector $\tilde{\theta}$ maximizes the Lagrangian function $L^*(\theta)$, and so it satisfies the equations

$$\frac{\delta L^*(\theta)}{\delta \theta} = 0.$$

It follows that

$$\frac{\delta \ln L(\tilde{\theta})}{\delta \theta} = \tilde{G}'\tilde{\lambda}.$$

Thus an alternative expression of the ML statistic is given by

$$ML = \tilde{\lambda}'\tilde{G}I(\tilde{\theta})^{-1}\tilde{G}'\tilde{\lambda}$$

This form motives the name of the test.

Much of the justification of the LM test depends on the fact that it bases only on parameter estimates from the restricted model. That makes it attractive in situations where the unconstrained model is difficult or impossible to estimate.

5.4 Likelihood ratio test, Wald test and Lagrange multiplier test in linear regression model

In this subsection we will illustrate these tests in the framework of linear regression model

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

$$H_0 : R\beta - q = 0$$

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.

5.4.1 The likelihood ratio statistic

The likelihood ratio test requires estimates of both the unrestricted and the restricted models. We have seen that if ϵ is distributed as multivariate normal with mean vector zero and variance covariance matrix $\sigma^2 I$, then the maximum likelihood (ML) estimator for β is the ordinary least squares (OLS) estimator given by

$$b = (X'X)^{-1}X'y$$

and the ML estimator of σ^2 is given by

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

where $e = y - Xb$.

The restricted ML estimator is obtained as the solution to

$$\min_\beta S(\beta) = (y - X\beta)'(y - X\beta) \quad \text{subject to} \quad R\beta = q.$$

A Lagrangean function for this problem can be written

$$L^*(\beta,\lambda) = (y - X\beta)'(y - X\beta) + 2\lambda'(R\beta - q)$$

where λ is a $J \times 1$ vector of Lagrange multipliers. The solutions b_* and λ_* will satisfy the necessary condition

$$\frac{\delta L^*}{\delta b_*} = -2X'(y - Xb_*) + 2R'\lambda_* = 0$$

$$\frac{\delta L^*}{\delta \lambda_*} = 2(Rb_* - q) = 0$$

The first equation yields

$$b_* = b - (X'X)^{-1}R'\lambda_*$$

Premultiplying by R gives

$$Rb_* = Rb - R(X'X)^{-1}R'\lambda_*$$

hence

$$\lambda_* = \left[R(X'X)^{-1}R'\right]^{-1}(Rb - q)$$

and so

$$b_* = b - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(Rb - q)$$

The restricted ML estimator of σ^2 is given by

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

where $e_* = y - Xb_*$. The LR statistic is given by:

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

$$= -2[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_r^2) - \frac{1}{2\hat{\sigma}_r^2}(y - Xb_*)'(y - Xb_*) + \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\hat{\sigma}^2) + \frac{1}{2\hat{\sigma}^2}(y - Xb)'(y - Xb)].$$

$$= -2[-\frac{n}{2}\ln(\hat{\sigma}_r^2) - \frac{n}{2} + \frac{n}{2}\ln(\hat{\sigma}^2) + \frac{n}{2}].$$

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

5.4.2 The Wald statistic

We remember that, in general, the Wald statistic is given by

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

Being

$$g(\theta) = [R, \mathbf{0}]\begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix} - q,$$

We have that

$$W = (Rb - q)'\left[[R, \mathbf{0}]\begin{bmatrix} \hat{\sigma}^2(X'X)^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}[R, \mathbf{0}]'\right]^{-1}(Rb - q)$$

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

5.4.3 The Lagrange multiplier statistic

Finally, we consider the LM statistic given by:

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

To evaluate the score vector at the restricted estimator $\tilde{\theta} = (b_*, \hat{\sigma}_r^2)'$, we replace ϵ with $e_* = y - Xb_*$ and σ^2 by $\hat{\sigma}_r^2$. Thus

$$s(\tilde{\theta}) = \begin{bmatrix} \frac{1}{\hat{\sigma}_r^2} X'e_* \\ 0 \end{bmatrix}$$

We have

$$\begin{aligned} ML &= \begin{bmatrix} \frac{1}{\hat{\sigma}_r^2} X'e_* \; 0 \end{bmatrix} \begin{bmatrix} \hat{\sigma}_r^2 (X'X)^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\hat{\sigma}_r^4}{n} \end{bmatrix} \begin{bmatrix} \frac{1}{\hat{\sigma}_r^2} X'e_* \\ 0 \end{bmatrix} \\ &= \frac{e_*' X (X'X)^{-1} X'e_*}{\hat{\sigma}_r^2}. \end{aligned}$$

Being

$$e_* = y - Xb_* = y - Xb - X(b_* - b)$$

and

$$b_* - b = -(X'X)^{-1} R' \left[R(X'X)^{-1} R' \right]^{-1} (Rb - q),$$

we have that

$$e_* = e + (X'X)^{-1} R' \left[R(X'X)^{-1} R' \right]^{-1} (Rb - q).$$

Hence

$$e_*' X (X'X)^{-1} X'e_* = (Rb - q)' \left[R(X'X)^{-1} R' \right]^{-1} (Rb - q)$$

It follows that

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

5.4.4 The asymptotically equivalence of W and LM statistics

Consider the following expressions of W and LM statistics, respectively

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

and

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

We note that the W and LM statistics differ only by different estimates of σ^2. This implies that we can express the LM test statistic as

$$LM = \frac{\hat{\sigma}^2}{\hat{\sigma}_r^2} W$$

If we make use of the fact that the likelihood ratio is given by

$$\lambda(\beta, \sigma^2) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_r^2} \right)^{n/2},$$

we can express the LM statistic as

$$LM = [\lambda(\beta, \sigma^2)]^{2/n} W$$

This makes it obvious that the LM and W statistics are asymptotically equivalent.

5.4.5 The *LR*, *LM*, and *W* tests as functions of the *F* test

An alternative approach to testing the null hypothesis $H_0 : R\beta - q = 0$ is to estimate the restricted and unrestricted models and compute the following F statistic

$$F = \frac{(RRSS - URSS)/J}{URSS/(n-k)}$$

where $RRSS$ stands for the sum of the squared residuals of the restricted model and $URSS$ is the same for the unrestricted model (For example, see Maddala (19??, p. 458). The idea behind this test is intuitive. The F-statistic compares the residual sum of squares computed with and without the restrictions imposed. If the restrictions are valid, there should be little difference in the two residual sum of squares and the F-value should be small. If RRSS is different from URSS, then we reject these restrictions.

Evans and Savin [4, p. 740] have shown that all the LR, LM, and W tests are monotonic functions of the F test statistic. Here we offer a proof of this fact.

We start considering the sum of the squared residuals of the restricted model. It is given by

$$e_*'e_* = e'e + (b_* - b)'X'X(b_* - b)$$

and hence

$$e_*'e_* - e'e = (Rb - q)' \left[R(X'X)^{-1}R'\right]^{-1} (Rb - q)$$

This appears in the numerator of the F statistic. Inserting the remaining parts, we obtain

$$F = \frac{\left((Rb - q)' \left[R(X'X)^{-1}R'\right]^{-1} (Rb - q)\right)/J}{e'e/(n-k)}$$

On the other hand, we have seen that the W test statistics is given by:

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

thus we can conclude that

$$W = \frac{nJ}{n-k}F.$$

The Wald statistic is a strictly increasing function of the F statistic.

Now we consider the LM statistic given by:

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

where $\hat{\sigma}_r^2 = e_*'e_*/n$. We have that

$$W = \frac{n\,(e_*'e_* - e'e)}{e'e}$$

and

$$LM = \frac{n\,(e_*'e_* - e'e)}{e_*'e_*} = \frac{n\,(e_*'e_* - e'e)}{e'e + e_*'e_* - e'e} = \frac{\frac{n\left(e_*'e_* - e'e\right)}{e'e}}{1 + \frac{e_*'e_* - e'e}{e'e}} = \frac{W}{1 + W/n}$$

Thus

$$LM = \frac{\frac{nJ}{n-k}F}{1 + (\frac{nJ}{n-k}F)/n} = \frac{nJF}{m - k + JF}$$

Also the LM statistic is a strictly increasing function of the F statistic.

Finally, we consider the LR statistic given by:

$$LR = -2\ln\frac{\max_{R\beta=q,\sigma^2} L(\beta,\sigma^2)}{\max_{\beta,\sigma^2} L(\beta,\sigma^2)} = -2\ln\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_r^2}\right)^{n/2} = n\left(\ln\hat{\sigma}_r^2 - \ln\hat{\sigma}^2\right)$$

This test statistic can also be written as follows

$$LR = n\left(\ln\hat{\sigma}_r^2 - \ln\hat{\sigma}^2\right) = n\left(\ln e_*'e_* - \ln n + \ln n - \ln e'e\right) = n\left(\ln e_*'e_* - \ln e'e\right)$$

$$= n\ln\frac{e_*'e_*}{e'e} = n\ln\left(1 + \frac{e_*'e_* - e'e}{e'e}\right) = n\ln\left(1 + \frac{W}{n}\right)$$

It follows that

$$LR = n\ln\left(1 + \frac{J}{n-k}F\right)$$

Hence the W, LM and LR statistics are functions of the F statistic.

The fact that each test statistic is a function of the F statistic implies that the three exact tests are equivalent: in the n dimensional sample space the three exact tests have the same critical region. In other words, when the exact W test accepts H_0 at significance level α the exact LM and LR tests also accept H_0 at level α and similarly if the exact W test rejects H_0. As a consequence, the exact tests have the same power function-the power function of the F test. Hence there can be no conflict between the exact tests.

5.4.6 The W, LR and LM Inequality

An interesting relationship between 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. This inequality was obtained by Berndt and Savin (1977). In order to prove this result, we remember that

$$LM = \frac{W}{1 + W/n}$$

and that

$$LR = n\ln\left(1 + \frac{W}{n}\right)$$

It follows that

$$\frac{LM}{n} = \frac{W/n}{1 + W/n}$$

and that

$$\frac{LR}{n} = \ln\left(1 + \frac{W}{n}\right).$$

Then we use the fact that $y \geq \ln(1+y) = y \geq (1+y)$ for $y = W/n$.

We have that although the asymptotic Wald, likelihood ratio, and Lagrange Multiplier tests have identical limiting chi-square distributions, a numerical inequality holds, yielding conflicting inference, especially for small samples.

Finally, we note that $LR = W = LM$ for the null $H_0 : \theta = \theta_0$ if the log-likelihood is quadratic. In fact, we consider a quadratic likelihood function given by

$$l(\theta) = \kappa - \frac{1}{2}(\theta - \hat{\theta})' A (\theta - \hat{\theta})$$

where $\hat{\theta}$ is a statistic, κ a constant and A is a known positive definite matrix. The score vector is

$$\frac{\delta l(\theta)}{\delta \theta} = s(\theta) = -A(\theta - \hat{\theta}),$$

and the Hessian matrix is

$$\frac{\delta^2 l(\theta)}{\delta \theta \delta \theta'} = -A.$$

Since the Hessian matrix is a constant, we have $I(\theta) = A$. Further, it is clear that $\hat{\theta}$ is the MLE for θ. Thus we have

$$LR = \ln L(\theta_0) - \ln L(\hat{\theta}) = -2\left[-\frac{1}{2}(\theta - \hat{\theta})' A (\theta - \hat{\theta}) + \frac{1}{2}(\theta - \theta_0)' A (\theta - \theta_0)\right]$$

$$= (\hat{\theta} - \theta_0)' A (\hat{\theta} - \theta_0).$$

$$W = g(\hat{\theta})' \left[G(\hat{\theta}) I(\hat{\theta})^{-1} G(\hat{\theta})'\right]^{-1} g(\hat{\theta}) = (\hat{\theta} - \theta_0)' A (\hat{\theta} - \theta_0).$$

and

$$ML = s(\theta_0)' I(\tilde{\theta})^{-1} s(\theta_0) = (\hat{\theta} - \theta_0)' A (\hat{\theta} - \theta_0).$$

We can therefore conclude that $LR = W = LM$.

5.5 The likelihood-based test procedures: conclusions

In summary:

Likelihood ratio test. Estimate θ with MLE, estimate again by imposing the H_0 restrictions, test if $\ln L(\tilde{\theta}; \mathbf{x}) - \ln L(\hat{\theta}; \mathbf{x}) = 0$.

Wald test. Estimate θ with MLE, check if $g(\hat{\theta}) = 0$.

Lagrange multiplier test. Estimate θ under the H_0 restrictions, check if $s(\tilde{\theta}) = 0$

1. Under regularity conditions and under the null hypothesis $H_o : g(\hat{\theta}) = 0$; each of the above statistics LR, W and ML, follows asymptotically a chi-square distribution with r degrees of freedom. Thus, in large samples, if the null hypothesis is true, the likelihood ratio, Wald, and Lagrange multiplier tests all tend to the same answer.

2. In small samples, the statistics may lead to conflicting conclusions, even when the null hypothesis is true.

3. However, even in large samples, the three tests can differ in their power against various alternatives.

4. The choice among the three statistics is often based on computational convenience.

6 References

Calzolari, G., (2012) *Econometric notes*, www.ds.unifi.it/didattica/materiale_didat/calzolari

Capuccio, N. and R. Orsi, (2005) *Econometria*, Il Mulino, Bologna.

Garthwaite, P.H., Jolliffe, I.T. and Jones, B, (1995) *Statistical Inference*, Prentice-Hall, London.

Greene, W. H., (2000), *Econometric analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Silvey, S. D., (1975), *Statistical Inference*, Chapman and Hall, New York.

Spanos, A, (1986), *Statistical Foundations of Econometric Modelling*, Cambridge University Press