ML method

In a ML method we search for the parameters, for which the sample is the most "likely".

It measures, how likely is the sample, with a likelihood function and searches for a parameter vector that maximizes the likelihood function.

ML method

In a ML method we search for the parameters, for which the sample is the most "likely".

It measures, how likely is the sample, with a likelihood function and searches for a parameter vector that maximizes the likelihood function.

Likelihood

Likelihood is a function $L(\theta)$ of a parameter vector θ. It gives a value of a joint density of a sample for a given parameter vector θ.

$$L(\theta|y) = f(y_1, ..., y_N; \theta)$$

If observations are independent, then it is a product of marginal densities

$$L(\theta|y) = \prod_{i=1}^{i=N} f(y_i; \theta)$$

Log likelihood

If observations are independent then it is more comfortable to look at the loglikelihood

$$l(\theta|y) = \sum_{i=1}^{i=N} \ln f(y_i;\theta)$$

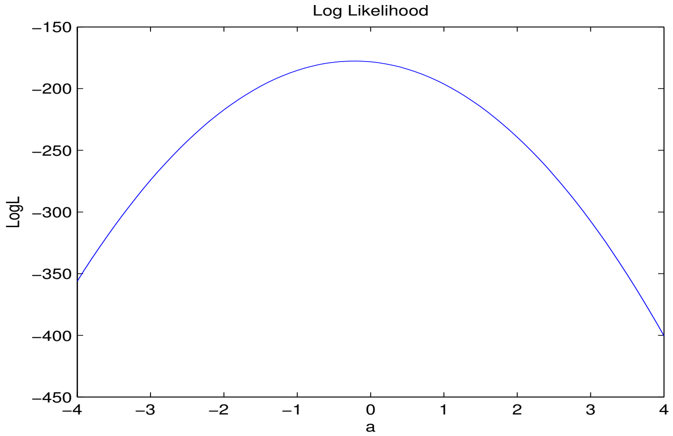
Example

Suppose, we want to model a process

$$y_n = \alpha + \varepsilon_n$$

where ε_n are i.i.d. with $\varepsilon_n \sim N(0, \sigma^2)$. What is the log-likelihood function $l(\theta|y)$?

Example ($\varepsilon_n \sim N(0,4)$)



Log likelihood

If observations are not i.i.d., for example

$$y_i = x_i\beta + u_i$$

then

$$l(\theta|y, X) = \sum_{i=1}^{i=N} \ln f(y_i|x_i; \theta)$$

Log likelihood

What is the log-likelihood for an AR(1) model with i.i.d residuals ($e_t \sim N(0, \sigma^2)$)?

$$y_t = \alpha y_{t-1} + e_t$$

Identification

The parameter vector θ is identifiable if for any other parameter vector θ^*

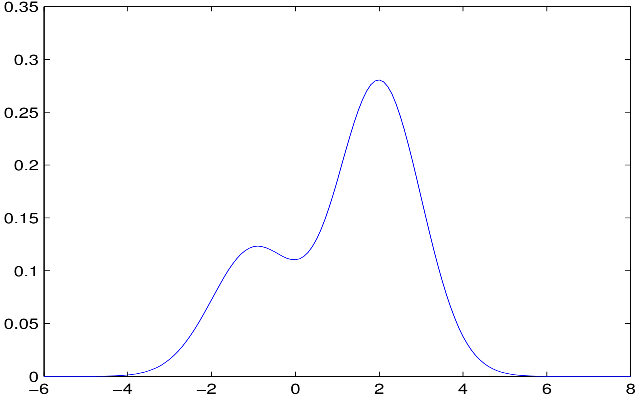
$$(\theta \neq \theta^*) \Rightarrow L(\theta|y) \neq L(\theta^*|y)$$

Identification - mixture of distributions

Let consider a model of a mixture of two normal distributions

$$y_t \sim \begin{cases} N(\alpha_1, 1) & \text{with probability } \gamma \\ N(\alpha_2, 1) & \text{with probability } 1-\gamma \end{cases}$$

Identification - mixture of distributions



Identification - mixture of distributions

The parameter vector is

$$\theta = (\alpha_1, \alpha_2, \gamma).$$

Let $\phi(y; \alpha, 1)$ denotes a density function of a normal distribution $N(\alpha, 1)$. Then the density of y is

$$f(y; \theta) = \gamma\phi(y; \alpha_1, 1) + (1 - \gamma)\phi(y; \alpha_2, 1)$$

Identification - mixture of distributions

Lets consider $\theta^* = (\alpha_2, \alpha_1, 1 - \gamma) \neq \theta$. Then

$$f(y; \theta^*) = (1 - \gamma)\phi(y; \alpha_2, 1) + \gamma\phi(y; \alpha_1, 1)$$

and

$$f(y; \theta) = f(y; \theta^*) \Rightarrow L(\theta|y) = L(\theta^*|y)$$

Model is not identifiable.

Local identification

The parameter vector θ is locally identifiable if there exists a neighborhood Θ of θ such that for any other parameter vector $\theta^* \in \Theta$

$$(\theta \neq \theta^*) \Rightarrow L(\theta|y) \neq L(\theta^*|y)$$

Local identification

When model is only locally identifiable, then we can impose restrictions that will ensure that it becomes identifiable.
Example:
We can impose a restriction

$$\alpha_1 > \alpha_2$$

that will order the mixing distribution (no label switching).

ML estimator

The parameter vector $\hat{\theta}$ is a maximum likelihood estimate if its maximize the likelihood or the log likelihood function.

$$l(\hat{\theta}|y) = sup_{\theta} l(\theta|y)$$

ML regularity conditions

Condition 1:
The first three derivatives of $\ln f(y;\theta)$ with respect to θ are continuous and finite for almost all y and all θ. It ensures existence of a Taylor series approximation and the finite variance of the derivatives of $l(\theta)$

ML regularity conditions

Condition 2:
There exists $E(Dl(\theta))$ and $E(Dl^2(\theta))$, where $Dl(\theta)$ and $Dl^2(\theta)$ denotes the first and the second derivative of $l(\theta)$, respectively.

ML regularity conditions

Condition 3:
For all θ, the absolute value of a third derivative is less then a function that has a finite expectation.
This condition will allow to truncate the Taylor series (when constructing test statistics)

ML properties

ML properties:

1. Consistency: $\hat{\theta} \to^p \theta_0$
2. Asymptotic normality
3. Asymptotic efficiency
4. Invariance: $g(\hat{\theta}) = g(\hat{\theta})$ for continuous and continuously differentiable function.

ML properties

ML properties:

1. Consistency: $\hat{\theta} \rightarrow^{p} \theta_0$
2. Asymptotic normality
3. Asymptotic efficiency
4. Invariance: $g(\hat{\theta}) = g(\hat{\theta})$ for continuous and continuously differentiable function.

ML properties

ML properties:

1. Consistency: $\hat{\theta} \rightarrow^{p} \theta_0$
2. Asymptotic normality
3. Asymptotic efficiency
4. Invariance: $g(\hat{\theta}) = g(\hat{\theta})$ for continuous and continuously differentiable function.

ML properties

ML properties:

1. Consistency: $\hat{\theta} \to^{p} \theta_0$
2. Asymptotic normality
3. Asymptotic efficiency
4. Invariance: $g(\hat{\theta}) = g(\hat{\theta})$ for continuous and continuously differentiable function.

ML properties

ML properties:

1. Consistency: $\hat{\theta} \to^p \theta_0$
2. Asymptotic normality
3. Asymptotic efficiency
4. Invariance: $g(\hat{\theta}) = g(\hat{\theta})$ for continuous and continuously differentiable function.

Score vector

The score vector $s(\theta)$ is a vector of first derivatives of the log likelihood function with respect to the parameter vector θ.

$$s(\theta; y) = \frac{\partial l(\theta|y)}{\partial \theta}$$

Score vector

Example:
Lets consider the example with $y_n = \alpha + \varepsilon_n$ and $\varepsilon_n \sim N(0, 4)$.
What is the score vector $s(\alpha; y)$ and the ML estimator?

Score vector for θ_0

What is the expectation and the variance of the score vector for the true parameters θ_0?

$$E_0(s(\theta_0)) = 0$$

$$Var_0(s(\theta_0)) = -E(\frac{\partial^2 \ln f(y;\theta_0)}{\partial\theta_0\partial\theta_0'}) = -E_0(H(\theta_0))$$

Score vector for θ_0

Example:

$$E(s(\alpha_0)) = E(\frac{1}{4}\sum_{i=1}^{N}(y_n - \alpha_0)) = E(\frac{1}{4}\sum_{i=1}^{N}\varepsilon_n) = 0$$

and

$$Var(s(\alpha_0)) = \frac{N}{16}Var(\varepsilon) = \frac{N}{4} = E(-\frac{1}{4}\sum_{i=1}^{N}-1)$$

Asymptotic normality

We know that

$$s(\hat{\theta}) = 0$$

Expand it in a second-order Taylor series around true parameters θ_0

$$s(\hat{\theta}) = s(\theta_0) + H(\bar{\theta})(\hat{\theta} - \theta_0) = 0$$

$$\sqrt{N}(\hat{\theta} - \theta_0) = -H(\bar{\theta})^{-1}\sqrt{N}s(\theta_0)$$

Asymptotic normality

We know that

$$s(\hat{\theta}) = 0$$

Expand it in a second-order Taylor series around true parameters θ_0

$$s(\hat{\theta}) = s(\theta_0) + H(\bar{\theta})(\hat{\theta} - \theta_0) = 0$$

$$\sqrt{N}(\hat{\theta} - \theta_0) = -H(\bar{\theta})^{-1}\sqrt{N}s(\theta_0)$$

Asymptotic normality

Hessian is evaluated at $\bar{\theta}$ that lies between $\hat{\theta}$ and θ_0. Since $\hat{\theta} \to^p \theta_0$ then $H(\bar{\theta}) \to^p H(\theta_0)$

$$\sqrt{N}(\hat{\theta} - \theta_0) \to^p -H(\theta_0)^{-1}\sqrt{N}s(\theta_0)$$

Asymptotic normality

$$-H(\theta_0)^{-1}\sqrt{N}s(\theta_0) = [-\frac{1}{N}H(\theta_0)]^{-1}\sqrt{N}[\frac{1}{N}s(\theta_0)]$$

Under CLT

$$\sqrt{N}[\frac{1}{N}s(\theta_0)] \rightarrow^{d} N(0, -E_0(\frac{1}{N}H(\theta_0)))$$

Asymptotic normality

Therefore,

$$\sqrt{N}(\hat{\theta} - \theta_0) \rightarrow^d N(0, [-E_0(\frac{1}{N}H(\theta_0))]^{-1})$$

$$\hat{\theta} \rightarrow^d N(\theta_0, [-E_0(H(\theta_0))]^{-1}) = N(\theta_0, I(\theta_0)^{-1})$$

Where $I(\theta_0) = -E_0(H(\theta_0))$ is called Information matrix

Asymptotic normality

Therefore,

$$\sqrt{N}(\hat{\theta} - \theta_0) \rightarrow^d N(0, [-E_0(\frac{1}{N}H(\theta_0))]^{-1})$$

$$\hat{\theta} \rightarrow^d N(\theta_0, [-E_0(H(\theta_0))]^{-1}) = N(\theta_0, I(\theta_0)^{-1})$$

Where $I(\theta_0) = -E_0(H(\theta_0))$ is called Information matrix

Information matrix equity

Information matrix equity

$$Var_0(s(\theta_0)) = E_0(s(\theta_0)s(\theta_0)') = -E_0(H(\theta_0))$$

Secondly,

$$E_0(s(\theta_0)s(\theta_0)') = \sum_{n=1}^{N} E_0(s_n(\theta_0)s_n(\theta_0)')$$

Information matrix equity

Information matrix equity

$$Var_0(s(\theta_0)) = E_0(s(\theta_0)s(\theta_0)') = -E_0(H(\theta_0))$$

Secondly,

$$E_0(s(\theta_0)s(\theta_0)') = \sum_{n=1}^{N} E_0(s_n(\theta_0)s_n(\theta_0)')$$

Information matrix equity

Finally,

$$-E_0(H(\theta_0)) = \sum_{n=1}^{N} E_0(s_n(\theta_0)s_n(\theta_0)')$$

Useful for estimating the variance of a score vector (Hessian is often too complicated to calculate).

Estimators of Information matrix

1. If the form of an expected Hessian $H(\theta)$ is known then
$$\hat{I}_1 = -E(H(\hat{\theta}))$$
2. If we know the analytical form of a Hessian, then
$$\hat{I}_2 = -H(\hat{\theta})$$
3. If we do not know the form of a Hessian, we can use an information matrix equity and estimate (BHHH estimator)
$$\hat{I}_3 = \sum_{n=1}^{N} s_n(\hat{\theta}) s_n(\hat{\theta})'$$
It is called as the *BHHH* or the *outer product estimator*.

Estimators of Information matrix

1. If the form of an expected Hessian $H(\theta)$ is known then

$$\hat{I}_1 = -E(H(\hat{\theta}))$$

2. If we know the analytical form of a Hessian, then

$$\hat{I}_2 = -H(\hat{\theta})$$

3. If we do not know the form of a Hessian, we can use an information matrix equity and estimate (BHHH estimator)

$$\hat{I}_3 = \sum_{n=1}^{N} s_n(\hat{\theta}) s_n(\hat{\theta})'$$

It is called as the *BHHH* or the *outer product estimator*.

Estimators of Information matrix

1. If the form of an expected Hessian $H(\theta)$ is known then
$$\hat{I}_1 = -E(H(\hat{\theta}))$$
2. If we know the analytical form of a Hessian, then
$$\hat{I}_2 = -H(\hat{\theta})$$
3. If we do not know the form of a Hessian, we can use an information matrix equity and estimate (BHHH estimator)
$$\hat{I}_3 = \sum_{n=1}^{N} s_n(\hat{\theta}) s_n(\hat{\theta})'$$
It is called as the *BHHH* or the *outer product estimator*.

Estimators of Information matrix

1. If the form of an expected Hessian $H(\theta)$ is known then
$$\hat{I}_1 = -E(H(\hat{\theta}))$$
2. If we know the analytical form of a Hessian, then
$$\hat{I}_2 = -H(\hat{\theta})$$
3. If we do not know the form of a Hessian, we can use an information matrix equity and estimate (BHHH estimator)
$$\hat{I}_3 = \sum_{n=1}^{N} s_n(\hat{\theta}) s_n(\hat{\theta})'$$
It is called as the *BHHH* or the *outer product estimator*.

Estimators of Information matrix

Example:

$$\hat{I}_2 = \frac{N}{4}$$

and

$$\hat{I}_3 = \frac{1}{16}\sum_{n=1}^{N}(y_n - \hat{\alpha})^2$$

Estimators of Information matrix

Example:

$$\hat{I}_2 = \frac{N}{4}$$

and

$$\hat{I}_3 = \frac{1}{16}\sum_{n=1}^{N}(y_n - \hat{\alpha})^2$$