Econometrics

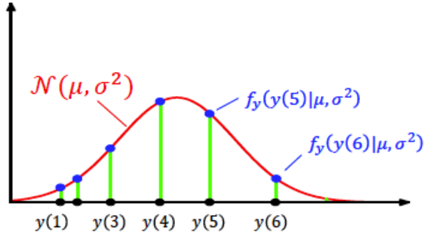
Maximum Likelihood Estimation: Prerequisites

November 2023 – January 2024

Maximum Likelihood Estimation

- Estimation procedure: given a probabilistic model we **estimate** its parameters in a consistent with the **observed** data way.

Assume to have 6 i. i. d. observations $\mathcal{D} = \{y(1), y(2), \ldots, y(6)\}$, where $y(i) \sim \mathcal{N}(\mu, \sigma^2)$



The **pdf** of a single random variable is

$$f(y(i)|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y(i) - \mu}{\sigma}\right)^2\right]$$

Maximum Likelihood Estimation: The likelihood function

- Consider a random sample $Y = (y_1, \ldots, y_n)$, i.e., i.i.d observations drawn (or following) the Gaussian distribution $N(\mu, \sigma^2)$. Notation: $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$.
- The likelihood of the observed data for **fixed** values of the parameters μ, σ^2 is the joint pdf of the data vector Y

Known data
unknown parameters

$$L(\mu, \sigma^2; Y) = f_Y(y_1, \ldots, y_n; \mu, \sigma^2) = \prod_{i=1}^{n} N(y_i; \mu, \sigma^2)$$

unknown data
known param.

- The product above is the product of the blue dots in the previous Figure.
- Notice the notation $L(\mu, \sigma^2; Y)$ $f_Y(Y; \mu, \sigma^2)$ is seen as **function of the parameters** of the model when is considered as likelihood function; for given data and a grid for the values of the parameters we can draw the Figure of the likelihood.

Maximum Likelihood Estimation: Example

Let's assume $\sigma^2 = 1$ known and aim to find the MLE of μ.

- The MLE of μ is that value of μ that maximizes the likelihood $L(\mu; Y) = \prod_{i=1}^{n} N(y_i; \mu, 1)$

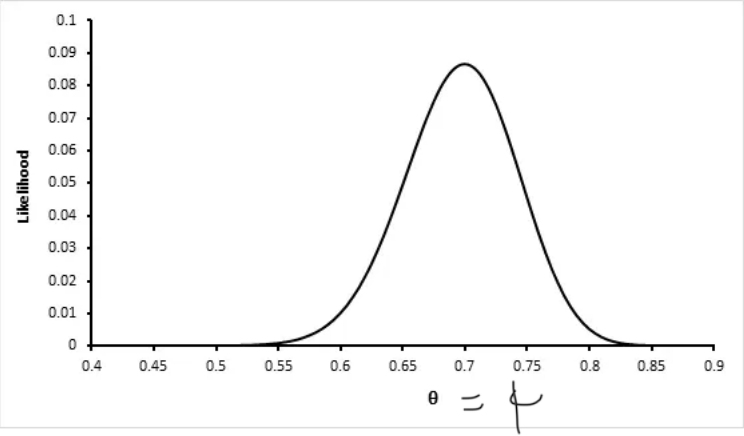


Figure 1: Distribution of likelihood corresponding to different θ values

Maximum Likelihood Estimation

We generally work with the log-likelihood and aim to solve the maximisation problem

$$\max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}; Y) = \ln \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i; \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} \in \mathbb{R}^k$ is the vector with parameters of interest.

- In the case of the Gaussian example $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $f(y_i; \boldsymbol{\theta}) = N(y_i; \mu, \sigma^2)$.

Maximum Likelihood Estimation

We generally work with the log-likelihood and aim to solve the maximisation problem

$$\max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}; Y) = \ln \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i; \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} \in \mathbb{R}^k$ is the vector with parameters of interest.

- In the case of the Gaussian example $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $f(y_i; \boldsymbol{\theta}) = N(y_i; \mu, \sigma^2)$.
- Homework: $\hat{\boldsymbol{\theta}}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} y_i$

↳ How? See next slides

Maximum Likelihood Estimation: Procedure

- Use F.O.C. to find location of possible maximum

$$\frac{\partial \ln L(\boldsymbol{\theta}; Y)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_2} \\ \cdots \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_k} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix} = \mathbf{0}_{k \times 1}$$

Maximum Likelihood Estimation: Procedure

- Use F.O.C. to find location of possible maximum

$$\frac{\partial \ln L(\boldsymbol{\theta}; Y)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_2} \\ \dots \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_k} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix} = \mathbf{0}_{k \times 1}$$

- Use S.O.C. to ensure global maximum.

$$\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{pmatrix} \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_2^2} & \dots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_k} \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_2} & \dots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_k^2} \end{pmatrix} = \boldsymbol{H}$$

$\boldsymbol{H}$ is the Hessian matrix which should be negative definite at $\hat{\boldsymbol{\theta}}$

Maximum Likelihood Estimation: Properties

Prerequisites:

- Score function: $E\left[\frac{\partial \ln L(\boldsymbol{\theta}; Y)}{\partial \boldsymbol{\theta}}\right]$
- Information matrix: $I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = -E[\boldsymbol{H}]$

Maximum Likelihood Estimation: Properties

Prerequisites:

- Score function: $E\left[\frac{\partial \ln L(\boldsymbol{\theta}; Y)}{\partial \boldsymbol{\theta}}\right]$
- Information matrix: $I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = -E[\boldsymbol{H}]$
- Cramer-Rao lower bound:

Theorem

The variance of an unbiased estimator of a parameter θ will always be at least as large as

$$(I(\theta))^{-1} = \left(-E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]\right)^{-1} = \left(E\left[\frac{\partial \ln L(\theta)}{\partial \theta}\right]^2\right)^{-1}$$

i.e.

$$Var(\theta) \geq (I(\theta))^{-1}$$

Maximum Likelihood Estimation: Properties

Proposition: Under regularity conditions (on $f\left(y|X;\theta\right)$), the MLE estimator $\hat{\theta}$ has the following asymptotic properties:

M1: Consistency: $p\lim\left(\hat{\theta}\right)=\theta_0$

M2: Asymptotic normality: $\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,\left(\frac{I(\theta_0)}{n}\right)^{-1}\right)$,

or in practice $\hat{\theta}\overset{a}{\sim}N\left(\theta_0,I\left(\theta_0\right)^{-1}\right)$, where $I\left(\theta_0\right)=E\left[-\partial^2\ln L/\partial\theta\partial\theta'\right]$.

M3: Asymptotic efficiency: $\hat{\theta}$ achieves the CRLB and it's asymptotically efficient.

M4: Invariance: MLE of $\gamma_0=c\left(\theta_0\right)$ is $\hat{\gamma}=c\left(\hat{\theta}\right)$.

The underlying assumption above is that $f\left(\cdot\right)$ is the true density.