TA Session for Econometrics II 2025

2: Maximum likelihood Estimation

Jukina HATAKEYAMA

The University of Osaka, Department of Economics

September 16, 2025

1 Introduction

2 Maximum Likelihood Estimation
- Likelihood function and Log-likelihood function
- Fisher Information Matrix
- Cauchy–Schwarz Inequality
- Cramér–Rao Lower Bound
- Cramér–Rao Lower Bound (Vector Case)
- Asymptotic Normality of MLE

3 Simple exercise: Linear Regression with Gaussian Errors

Maximum Likelihood Estimation (MLE)

- A method to estimate parameters by choosing the distribution that makes the observed data most "likely".
- Treats the likelihood of the data as a function of the parameters and finds the values that maximise it.
- One of the most widely used estimation techniques in statistics and econometrics.

Definition of the MLE

Definition

The maximum likelihood estimator $\hat{\theta}$ is

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x_1, \ldots, x_n).$$

- Choose the parameter value that maximises the likelihood.
- In practice, solve

$$\frac{d}{d\theta}\ell(\theta) = 0$$

to obtain the estimator.

Intuition of MLE

- Normally: parameters are fixed, data are random variables.
- In Maximum Likelihood Estimation:
  - The observed data are treated as fixed.
  - The parameter(s) are regarded as variables.
- The likelihood measures how plausible the data are under different parameter values.
- The MLE is the parameter value that maximises this likelihood.

The Likelihood Function

- The probability (density or mass) function considered as a function of the parameter:

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

- Data $(x_1, \ldots, x_n)$ are fixed; the parameter θ varies.
- For convenience, we often use the log-likelihood:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

Example: Bernoulli Distribution

- Independent trials with success probability p.
- Likelihood function:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

- Log-likelihood:

$$\ell(p) = \sum_{i=1}^{n} \Big(x_i \log p + (1-x_i)\log(1-p) \Big)$$

- Maximisation yields

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(the sample mean).

Normal Distribution Example

- Suppose $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$.
- The likelihood is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

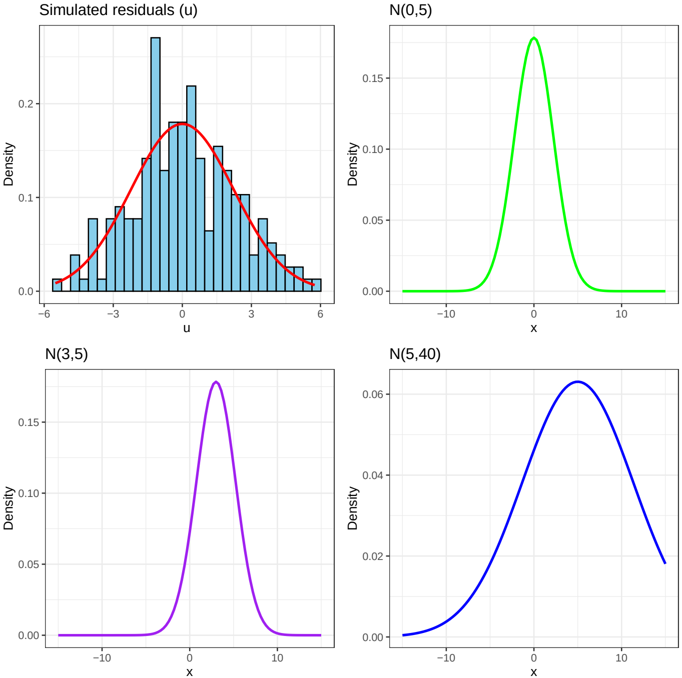
- Here the x_i are fixed observed values; we vary (μ, σ^2) to see where L is largest.

Result of the Maximisation

- Solving the maximisation problem gives

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

- Interpretation:
  - The MLE chooses the parameters that make the observed data most plausible.
  - For the Normal distribution, these turn out to be the sample mean and the (biased) sample variance.



Simulated residuals (u)
Density
u
N(0,5)
Density
x
N(3,5)
Density
x
N(5,40)
Density
x

Likelihood Function

Assume random variables, $X_1, \ldots, X_N$, are mutually independent and identically distributed (i.e. Gaussian distribution).

We denote the probability density function of $\{X_i\}_{i=1}^N$ as $f(x;\theta)$, where $x = (x_1, \ldots, x_N)$ and θ is a parameter vector.

The likelihood function is defined as:

$$L(\theta; x) := f(x; \theta),$$

where

$$f(x;\theta) = \prod_{i=1}^{N} f(x_i;\theta)$$

when the random variables are i.i.d.

The likelihood function is a joint distribution function. Therefore, it is generally non-linear.

To simplify the calculation, we take its logarithm.

Let $\ell(\theta; x) := \log(L(\theta; x))$. Then we have the following equivalence:

$$\max_{\theta} L(\theta; x) \iff \max_{\theta} \ell(\theta; x).$$

ML estimator must satisfy the following two conditions:

1. $\partial_{\theta} \ell(\hat{\theta}; X) = 0$,
2. The Hessian, $\partial^2_{\theta\theta^\top} \ell(\hat{\theta}; X)$, is a negative definite matrix.

Fisher Information Matrix

For a parameter vector θ, the **Fisher information matrix** is defined as:

$$I(\theta) := \mathbb{E}\left[-\partial^2_{\theta\theta^\top}\ell(\theta; X)\right],$$

where $\ell(\theta; X)$ is the log-likelihood function.

Intuitively, $I(\theta)$ measures the amount of information that the observed data X contain about the parameter θ.

Derivation (1)

Derivation:
Assume the domain of x does not depend on θ and that the first derivative of the likelihood function exists.
Since the likelihood is a probability distribution in x for fixed θ, we have:

$$\int L(\theta; x)\, dx = 1.$$

Differentiating with respect to θ, we obtain:

$$\int \partial_\theta L(\theta; x)\, dx = 0.$$

Derivation (2)

Using

$$\frac{\partial \log X}{\partial X} = \frac{1}{X},$$

we obtain:

$$\int \left(\partial_\theta \ell(\theta; x)\right) \, L(\theta; x) \, dx = 0.$$

This is equivalent to:

$$\mathbb{E}[\partial_\theta \ell(\theta; x)] = 0.$$

This follows from the fact that the likelihood function is a probability distribution in x and from the definition of expectation.

Equivalence of Definitions

Differentiating the above equation with respect to θ again, we obtain:

$$
\begin{aligned}
0 &= \partial_{\theta^\top} \int \left(\partial_\theta \ell(\theta; x)\right) \, L(\theta; x) \, dx \\
&= \int \left(\partial^2_{\theta\theta^\top} \ell(\theta; x)\right) \, L(\theta; x) \, dx + \int \left(\partial_\theta \ell(\theta; x)\right) \, \left(\partial_{\theta^\top} L(\theta; x)\right) \, dx \\
&= \int \left(\partial^2_{\theta\theta^\top} \ell(\theta; x)\right) \, L(\theta; x) \, dx + \int \left(\partial_\theta \ell(\theta; x)\right) \, \left(\partial_{\theta^\top} \ell(\theta; x)\right) \, L(\theta; x) \, dx \\
&= \mathbb{E}\left[\partial^2_{\theta\theta^\top} \ell(\theta; x)\right] + \mathbb{E}\left[\partial_\theta \ell(\theta; x) \, \partial_{\theta^\top} \ell(\theta; x)\right].
\end{aligned}
$$

Recalling the definition of the variance, $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$, we obtain:

$$
-\mathbb{E}\left[\partial^2_{\theta\theta^\top} \ell(\theta; x)\right] = \mathbb{E}\left[\partial_\theta \ell(\theta; x) \, \partial_{\theta^\top} \ell(\theta; x)\right] = \mathrm{Var}(\partial_\theta \ell(\theta; x)).
$$

Remark on the Parameter θ

- The parameter θ is **not** a random variable. It represents a fixed but unknown constant (the true value of the parameter).
- The randomness lies in the data X, which are treated as random variables.
- Consequently, the expectation in the Fisher information matrix

$$I(\theta) = \mathbb{E}\big[-\partial^2_{\theta\theta^\top}\ell(\theta; X)\big]$$

is taken with respect to the distribution of X, **not** over θ.

Cauchy–Schwarz Inequality (Scalar Case)

For any two random variables X and Y, the **Cauchy–Schwarz inequality** states that:

$$|\mathrm{Cov}(X, Y)|^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y).$$

Equivalently, using expectations:

$$|\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]|^2 \leq \mathbb{E}[(X - \mathbb{E}[X])^2]\,\mathbb{E}[(Y - \mathbb{E}[Y])^2].$$

Interpretation: The absolute value of the covariance between two random variables cannot exceed the product of their standard deviations.

Cauchy–Schwarz Inequality (Vector Case)

For two vectors $a, b \in \mathbb{R}^n$, the inequality states:

$$|a^\top b|^2 \leq (a^\top a)(b^\top b).$$

Interpretation: The absolute value of the inner product of two vectors is bounded by the product of their lengths (Euclidean norms).

In terms of random vectors $X, Y \in \mathbb{R}^p$:

$$\mathrm{Cov}(X, Y)\,\mathrm{Var}(Y)^{-1}\,\mathrm{Cov}(X, Y)^\top \preceq \mathrm{Var}(X).$$

This is a matrix generalisation used in multivariate Cramér–Rao inequalities.

Geometric Interpretation

- Consider vectors a and b in $\mathbb{R}^n$.
- Let θ be the angle between a and b.
- Then $a^\top b = \|a\| \, \|b\| \cos\theta$.
- Hence, $|a^\top b| \leq \|a\| \, \|b\|$, which is exactly the Cauchy–Schwarz inequality.

Remark: Equality holds if and only if a and b are linearly dependent.

Cramér–Rao Lower Bound (Scalar Case)

Let $\hat{\theta}$ be an unbiased estimator of the (unknown but fixed) parameter θ.

Then its variance satisfies:

$$\mathrm{Var}(\hat{\theta}) \ \geq \ \frac{1}{I(\theta)}.$$

Interpretation: The bound is evaluated at the true value of θ and shows that no unbiased estimator can achieve a variance smaller than the reciprocal of the Fisher information.

Derivation (Scalar Case)

Suppose an unbiased estimator of θ is given by $s(X) = \hat{\theta}$.

By definition of unbiasedness:

$$\mathbb{E}[s(X)] = \theta.$$

Writing the expectation as an integral over the likelihood:

$$\mathbb{E}[s(X)] = \int s(x)\, L(\theta; x)\, dx.$$

Differentiating with respect to θ:

$$\partial_\theta \mathbb{E}[s(X)] = \int s(x)\, \partial_\theta \ell(\theta; x)\, L(\theta; x)\, dx.$$

Since $\mathbb{E}[\partial_\theta \ell(\theta; X)] = 0$, this can be written as

$$\partial_\theta \mathbb{E}[s(X)] = \mathrm{Cov}\big(s(X),\, \partial_\theta \ell(\theta; X)\big).$$

Cauchy–Schwarz Inequality

Assume $s(X)$ is scalar. Using the Cauchy–Schwarz inequality:

$$
\begin{aligned}
(\partial_\theta \mathbb{E}[s(X)])^2 &= (\mathrm{Cov}(s(X), \partial_\theta \ell(\theta; X)))^2 \\
&\leq \mathrm{Var}(s(X))\,\mathrm{Var}(\partial_\theta \ell(\theta; X)).
\end{aligned}
$$

Rearranging:

$$
\mathrm{Var}(s(X)) \geq \frac{(\partial_\theta \mathbb{E}[s(X)])^2}{\mathrm{Var}(\partial_\theta \ell(\theta; X))} = \frac{1}{\mathrm{Var}(\partial_\theta \ell(\theta; X))}.
$$

Finally, using $\mathrm{Var}(\partial_\theta \ell(\theta; X)) = I(\theta)$, we recover:

$$
\mathrm{Var}(s(X)) \geq \frac{1}{I(\theta)}.
$$

Vector Case

Suppose $s(X) = \hat{\theta}$ is an unbiased estimator of the p-dimensional parameter vector $\theta \in \mathbb{R}^p$.

By definition of unbiasedness:

$$\mathbb{E}[s(X)] = \theta.$$

Differentiating with respect to θ:

$$\partial_{\theta^\top} \mathbb{E}[s(X)] = \partial_{\theta^\top} \theta = I_p,$$

where I_p is the $p \times p$ identity matrix.

Using the score function $U(\theta) = \partial_\theta \ell(\theta; X)$:

$$\partial_{\theta^\top} \mathbb{E}[s(X)] = \operatorname{Cov}(s(X), U(\theta)) \operatorname{Var}(U(\theta))^{-1} \operatorname{Var}(U(\theta)).$$

Fisher Information Inequality (Vector Case)

By the matrix version of the Cauchy–Schwarz inequality:

$$\mathrm{Var}(s(X)) \succeq \mathrm{Cov}(s(X), U(\theta))\,\mathrm{Var}(U(\theta))^{-1}\,\mathrm{Cov}(s(X), U(\theta))^{\top}.$$

Substituting $\mathrm{Cov}(s(X), U(\theta)) = I_p$ and $\mathrm{Var}(U(\theta)) = I(\theta)$, we obtain:

$$\mathrm{Var}(s(X)) \succeq I(\theta)^{-1}.$$

Interpretation: No unbiased estimator of θ can have a covariance matrix smaller (in the positive semidefinite sense) than the inverse of the Fisher information matrix.

Asymptotic Normality of MLE

Let $\hat{\theta}_{\mathrm{MLE}}$ be the maximum likelihood estimator of a parameter vector θ.

Under regularity conditions, as the sample size $N \to \infty$:

$$\sqrt{N}\,(\hat{\theta}_{\mathrm{MLE}} - \theta) \quad \xrightarrow{d} \quad N\Big(0, I(\theta)^{-1}\Big),$$

where $I(\theta)$ is the Fisher information matrix.

Interpretation: The MLE is asymptotically unbiased and its distribution approaches a multivariate normal distribution centred at the true parameter, with covariance given by the inverse Fisher information.

Remarks

- This result follows from the general theory of **M-estimators**, of which the MLE is a special case.
- The proof relies on Taylor expansion of the score function and the Law of Large Numbers / Central Limit Theorem.
- Practically, it justifies the use of normal-based confidence intervals and Wald tests for large samples:

$$\hat{\theta}_{\mathrm{MLE}} \pm z_{\alpha/2} \sqrt{\mathrm{diag}(I(\hat{\theta}_{\mathrm{MLE}})^{-1})}.$$

- Detailed derivations are omitted, as they follow directly from standard M-estimator theory.

Linear Regression with Gaussian Errors

Consider the linear regression model:

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

where

- $y \in \mathbb{R}^n$ is the vector of observations,
- $X \in \mathbb{R}^{n \times p}$ is the design matrix,
- $\beta \in \mathbb{R}^p$ is the vector of parameters,
- $u \sim N(0, \sigma^2 I_n)$ is the error term.

Likelihood via Change of Variable

Suppose $u \sim N(0, \sigma^2 I_n)$. Its density is

$$f_u(u) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} u^\top u \right].$$

Consider the linear transformation

$$y = X\beta + u \quad \Longleftrightarrow \quad u = y - X\beta.$$

By the change of variable formula:

$$f_y(y) = f_u(u) \left| \det\left(\frac{\partial u}{\partial y} \right) \right|$$

where the Jacobian is $|\det(I_n)| = 1$.

Resulting Likelihood Function

Substituting $u = y - X\beta$ gives the likelihood function:

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

This matches the likelihood derived previously, confirming that

$$\ell(\beta, \sigma^2; y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta).$$

Interpretation: The change of variable formula justifies the likelihood for the observed data y from the distribution of the errors u.

Likelihood Function

Under the Gaussian assumption, the likelihood function is:

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

The log-likelihood function is

$$\ell(\beta, \sigma^2; y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta).$$

MLE of β

Differentiate the log-likelihood with respect to β and set to zero:

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2} X^\top (y - X\beta) = 0.$$

Solving for β, we obtain the MLE:

$$\hat{\beta}_{\mathrm{MLE}} = (X^\top X)^{-1} X^\top y.$$

This is exactly the ordinary least squares (OLS) estimator.

MLE of σ^2

Differentiate the log-likelihood with respect to σ^2 and set to zero:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y - X\beta)^\top (y - X\beta) = 0.$$

Solve for σ^2:

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}(y - X\hat{\beta}_{\mathrm{MLE}})^\top (y - X\hat{\beta}_{\mathrm{MLE}}).$$

Note: This differs from the unbiased OLS estimator by a factor of $n/(n-p)$.

Summary

- Under Gaussian errors, MLE of β coincides with OLS: $\hat{\beta}_{\mathrm{MLE}} = (X^\top X)^{-1} X^\top y$
- MLE of σ^2 is $\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}(y - X\hat{\beta})^\top (y - X\hat{\beta})$
- Log-likelihood can be used to construct confidence intervals and likelihood-ratio tests
- Provides a concrete example of MLE in a simple linear model