Maximum Likelihood Estimation (MLE)

Class 2

Wonmun Shin

(wonmun.shin@sejong.ac.kr)

Department of Economics, Sejong University

* This lecture note is written based on Professor Chang-Jin Kim's lecture note .

Likelihood Function

Random Variable Y_i

$$Y_i = \beta_1 + \beta_2 X_i + e_i$$

- Classical assumptions (A1)-(A4) and *Normality assumption (A5)*

$$e_i \overset{iid}{\sim} N\left(0,\, \sigma^2\right)$$

$$\rightarrow Y_i \sim N\left(\beta_1 + \beta_2 X_i,\, \sigma^2\right)$$

  - $E\left(Y_i\right) = E\left(\beta_1 + \beta_2 X_i + e_i\right) = \beta_1 + \beta_2 X_i + E\left(e_i\right) = \beta_1 + \beta_2 X_i$
  - $Var\left(Y_i\right) = Var\left(\beta_1 + \beta_2 X_i + e_i\right) = Var\left(e_i\right) = \sigma^2$

Joint Probability Density Function

- Joint pdf (probability of density function) of $Y_1,\ Y_2,\ \cdots,\ Y_n$

$$f\left(Y_1,\ Y_2,\ \cdots,\ Y_n \mid \beta_1 + \beta_2 X_i,\ \sigma^2\right)$$

- When Y_i are independent

$$\begin{aligned} & f\left(Y_1,\ Y_2,\ \cdots,\ Y_n \mid \beta_1 + \beta_2 X_i,\ \sigma^2\right) \\ & = f\left(Y_1 \mid \beta_1 + \beta_2 X_i,\ \sigma^2\right) f\left(Y_2 \mid \beta_1 + \beta_2 X_i,\ \sigma^2\right) \cdots f\left(Y_n \mid \beta_1 + \beta_2 X_i,\ \sigma^2\right) \end{aligned}$$

Joint Probability Density Function

- pdf of a normal distribution

$$f\left(Y_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i-\beta_1-\beta_2X_i)^2}$$

- Therefore, a joint pdf of Y_i is

$$f\left(Y_1,\, Y_2,\, \cdots,\, Y_n \mid \beta_1+\beta_2X_i,\, \sigma^2\right) = \frac{1}{\sigma^n\left(\sqrt{2\pi}\right)^n} e^{-\frac{1}{2\sigma^2}\sum(Y_i-\beta_1-\beta_2X_i)^2}$$

  - The above is the **joint pdf** when we know β_1, β_2, and σ^2 and we don't know Y_i, that is, we can earn the probability of observing Y_i jointly from the above.
  - BTW, if $Y_1, Y_2, \cdots, Y_n$ are known or given but β_1, β_2, and σ^2 are unknown constants, then we call the above "**Likelihood Function**".

Main Idea of Maximum Likelihood Estimation (MLE)

Main Idea of MLE (Maximum Likelihood Estimation)

- **ML estimator**: a value that maximizes the likelihood function.
  - Recall what OLS is: a method minimizing the sum of squares of residuals
  - MLE finds an estimator whose probability of extracting the data is highest when we observe the data.
  - For example,

$$\text{Student A} \begin{cases} \textit{Average} & > 90 \\ \textit{Bad condition} & 50 \end{cases}$$

$$\text{Student B} \begin{cases} \textit{Average} & < 50 \\ \textit{Good condition} & 90 \end{cases}$$

    - If we observe a test score of 88, then who is a student of the score?
    - Good estimation might be "Student A" $\rightarrow$ Idea of **MLE**

Main Idea of MLE [cont'd]

- Suppose that a basket which has 8 balls (? red balls, ? blue balls).
  - Let a probability of picking up a red ball p. $\rightarrow$ p is a parameter.
  - Random variable X_1, X_2, and X_3

$$X_i = \begin{cases} 1 & \text{when red ball is out} \rightarrow \ P(X_i = 1) = p \\ 0 & \text{when blue ball is out} \rightarrow \ P(X_i = 0) = 1 - p \end{cases}$$

  - Realization: $X_1 = 1$, $X_2 = 1$, and $X_3 = 0$
  - Probability is $p \times p \times (1 - p) \rightarrow \underbrace{L(p) = p^2 (1 - p)}_{\textit{Likelihood function}}$

Main Idea of MLE (Maximum Likelihood Estimation)

- Consider the possible two candidates:
  - *Basket A*: 2 red balls, 6 blue balls $\rightarrow p_A = \frac{1}{4}$
  - *Basket B*: 6 red balls, 2 blue balls $\rightarrow p_B = \frac{3}{4}$
- When we observe $X_1 = 1$, $X_2 = 1$, and $X_3 = 0$, which candidate is more likely?
  - $L\left(p_A = \frac{1}{4}\right) = \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{3}{64}$
  - $L\left(p_B = \frac{3}{4}\right) = \frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{9}{64}$
  - Therefore, *Basket B* is more likely, and $\hat{p} = \frac{3}{4}$ is a better estimate.
- Let's compute the ML estimate.

$$\max_p L(p) = p^2 (1 - p) = p^2 - p^3$$

  - (FOC) $2p - 3p^2 = p(2 - 3p) = 0 \qquad \therefore \hat{p}_{MLE} = \frac{2}{3}$
  - (SOC) $2 - 6\hat{p}_{MLE} = 2 - 6\left(\frac{2}{3}\right) = 2 - 4 = -2 < 0$

ML Estimator in Simple Regression

Log Likelihood Function

- Note The form of a joint pdf and a likelihood function is identical.
- Joint pdf

$$f\left(Y_1,\ Y_2,\ \cdots,\ Y_n \mid \beta_1+\beta_2 X_i,\ \sigma^2\right)=\frac{1}{\sigma^n\left(\sqrt{2\pi}\right)^n}e^{-\frac{1}{2\sigma^2}\sum(Y_i-\beta_1-\beta_2 X_i)^2}$$

- Likelihood function

$$L\left(\beta_1,\ \beta_2,\ \sigma^2 \mid Y_1,\ Y_2,\ \cdots,\ Y_n\right)=\frac{1}{\sigma^n\left(\sqrt{2\pi}\right)^n}e^{-\frac{1}{2\sigma^2}\sum(Y_i-\beta_1-\beta_2 X_i)^2}$$

- We can take a (natural) log. ($\because$ logarithm is monotonically increasing)

$$\begin{aligned}\ln L &= -n\ln\sigma-\frac{n}{2}\ln(2\pi)-\frac{1}{2\sigma^2}\sum\left(Y_i-\beta_1-\beta_2X_i\right)^2\\ &= -\frac{n}{2}\ln\sigma^2-\frac{n}{2}\ln 2\pi-\frac{1}{2\sigma^2}\sum\left(Y_i-\beta_1-\beta_2X_i\right)^2\end{aligned}$$

ML Estimators

- First order conditions
  - (wrt β_1) $-\frac{1}{\sigma^2}\sum(Y_i-\beta_1-\beta_2X_i)(-1)=0 \qquad (1)$
  - (wrt β_2) $-\frac{1}{\sigma^2}\sum(Y_i-\beta_1-\beta_2X_i)(-X_i)=0 \qquad (2)$
  - (wrt σ^2) $-\frac{n}{2\sigma^2}-\frac{1}{2}\sum(Y_i-\beta_1-\beta_2X_i)^2\left(-\frac{1}{\sigma^4}\right)=0 \qquad (3)$
- Let ML estimators $\hat{\beta}_1^{MLE}$, $\hat{\beta}_{2,}^{MLE}$, and $\hat{\sigma^2}^{MLE}$
  - From (1)

$$\Sigma Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

  - From (2)

$$\Sigma Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

ML Estimators [cont'd]

- The above equations are identical to normal equations of OLS.

$$\therefore \hat{\beta}_1^{MLE} = \hat{\beta}_1^{OLS}, \quad \hat{\beta}_2^{MLE} = \hat{\beta}_2^{OLS}$$

$$i.e.\ \hat{\beta}_1^{MLE} = \bar{Y} - \hat{\beta}_2^{MLE} \bar{X}$$

$$\hat{\beta}_2^{MLE} = \frac{\sum x_i y_i}{\sum x_i^2}$$

- From (3)

$$-\frac{n}{2\hat{\sigma^2}} + \frac{1}{2\hat{\sigma^4}} \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2 = 0$$

$$\rightarrow \frac{n}{2\hat{\sigma^2}} = \frac{\sum \hat{e}_i^2}{2\hat{\sigma^4}}$$

$$\rightarrow \hat{\sigma^2} = \frac{\sum \hat{e}_i^2}{n}$$

- Therefore, $\hat{\sigma^2}^{MLE} = \frac{\sum \hat{e}_i^2}{n}$ (Note that $\hat{\sigma^2}^{MLE} \neq \hat{\sigma^2}^{OLS}$)

Properties of ML Estimators

1. Unbiasedness and Consistency

- We know that $\hat{\beta}_1^{OLS}$, $\hat{\beta}_2^{OLS}$, and $\hat{\sigma^2}^{OLS}$ are unbiased and consistent estimators.
- Therefore, $\hat{\beta}_1^{MLE}$ and $\hat{\beta}_2^{MLE}$ are unbiased and consistent estimators.
- However, $\hat{\sigma^2}^{MLE}$ is not unbiased estimator.

$$E\left(\hat{\sigma^2}^{MLE}\right) = \frac{1}{n}E\left(\sum \hat{e}_i^2\right) = \frac{n-2}{n}\sigma^2 = \sigma^2 - \frac{2}{n}\sigma^2 \neq \sigma^2$$

  - $Bias\left(\hat{\sigma^2}^{MLE}\right) = E\left(\hat{\sigma^2}^{MLE}\right) - \sigma^2 = -\frac{2}{n}\sigma^2$

- Note **ML estimator is *generally* NOT unbiased estimator.**
- But ML estimator is consistent estimator!
  - $\lim_{n\to\infty} Bias\left(\hat{\sigma^2}^{MLE}\right) = 0$

Properties of ML Estimators [cont'd]

2. Efficiency

- Recall, OLS estimator is BLUE (Best Linear Unbiased Estimator).
- Also, OLS estimator is MVUE if we assume Normality.
- Since we got MLE under normality assumption, MLE is most efficient estimator.
  - But MLE is not MVUE. Why? MLE is not always satisfying unbiasedness.
- Note that MLE is a consistent estimator (when sample is sufficiently large), and its variance is smallest.
- So ***in asymptotic world, MLE is theoretically best***.
- *(optional)* When n is large, the variance of MLE converges to $\frac{1}{I(\theta)}$.
  - $I(\theta)$: Fisher information $= -E\left[\frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta'}\right]$
  - $I(\theta)^{-1} \leq \Sigma$ where Σ is the variance of any consistent parameter
  - Therefore, MLE is most efficient, and we call $I(\theta)^{-1}$ as Cramer-Rao lower bound.

ML Estimators: Pros and Cons

- MLE is very **GENERAL** estimation method.
  - When we maximize the likelihood function, it is easy to impose any constraint.
- Invariance: MLE is invariant with respect to transformation.
  - Suppose that $\hat{\theta}$ is the MLE of θ. Then, for any function g, $\hat{\eta} = g\left(\hat{\theta}\right)$ is the MLE of $\eta = g(\theta)$.
  - For example, a random variable $X \sim N(\theta, 1)$
    - MLE of θ is $\hat{\theta} = X$
    - MLE of $\eta = \theta^2$ is $\hat{\eta} = \hat{\theta}^2 = X^2$
- Drawbacks
  - Arbitrariness problem in choosing pdf of r.v.
  - May not exist or may not be unique. *(algorithm for computation is not that easy.)*