ECON3389 Econometric Methods

Module 2
Maximum Likelihood Estimation

Alberto Cappello

Department of Economics, Boston College

Spring 2025

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*
- The MLE estimator of θ will find the parameter values that maximize $L(\theta)$

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*
- The MLE estimator of θ will find the parameter values that maximize $L(\theta)$
- In other words, it will find the parameter value that maximize the likelihood of the observed data being drawn from $f(x|\theta)$

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*
- The MLE estimator of θ will find the parameter values that maximize $L(\theta)$
- In other words, it will find the parameter value that maximize the likelihood of the observed data being drawn from $f(x|\theta)$
- Suppose $x_1, x_2, ..., x_n$ form a random sample from a normal distribution for which the mean μ is unknown and variance σ^2 is known

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*
- The MLE estimator of θ will find the parameter values that maximize $L(\theta)$
- In other words, it will find the parameter value that maximize the likelihood of the observed data being drawn from $f(x|\theta)$
- Suppose $x_1, x_2, ..., x_n$ form a random sample from a normal distribution for which the mean μ is unknown and variance σ^2 is known
- The likelihood function of μ is

$$L(\mu) = f_n(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{2}$$

Maximum Likelihood Estimation: Example 1

- **Example**: Suppose I toss a coin 100 times and get 56 heads. What is the MLE of the probability of heads in a single toss?

Maximum Likelihood Estimation: Example 1

- **Example**: Suppose I toss a coin 100 times and get 56 heads. What is the MLE of the probability of heads in a single toss?
- **Model**: $L(p) = L(p; n, x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{100}{56} p^{56} (1-p)^{44}$

Maximum Likelihood Estimation: Example 1

- **Example**: Suppose I toss a coin 100 times and get 56 heads. What is the MLE of the probability of heads in a single toss?
- **Model**: $L(p) = L(p; n, x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{100}{56} p^{56} (1-p)^{44}$
- $L(0.5) = 0.038$
- $L(0.52) = 0.058$
- $L(0.54) = 0.073$
- $L(0.56) = 0.081$
- $L(0.58) = 0.073$

Maximum Likelihood Estimation: Method

- **Method 1**
  In the previous example, it is easy for us to write a simple equation that describes the likelihood surface that can be differentiated to find the MLE estimate

Maximum Likelihood Estimation: Method

- **Method 1**
  In the previous example, it is easy for us to write a simple equation that describes the likelihood surface that can be differentiated to find the MLE estimate
- Step 1: Take the log

$$\ln L = \ln \binom{100}{56} + \ln(p^{56}(1-p)^{44}) \tag{3}$$

$$\ln L = 56 \ln(p) + 44 \ln(1-p) \tag{4}$$

$$\tag{5}$$

Maximum Likelihood Estimation: Method

- **Method 1**
  In the previous example, it is easy for us to write a simple equation that describes the likelihood surface that can be differentiated to find the MLE estimate
- Step 1: Take the log

$$\ln L = \ln \binom{100}{56} + \ln(p^{56}(1-p)^{44}) \tag{3}$$

$$\ln L = 56\ln(p) + 44\ln(1-p) \tag{4}$$

$$\tag{5}$$

- Step 2: Differentiate the log likelihood to find the optimal parameter

$$\frac{56}{p} - \frac{44}{(1-p)} = 0 \tag{6}$$

$$56(1-p) - 44p = 0 \tag{7}$$

$$p = \frac{56}{100} \tag{8}$$

Maximum Likelihood Estimation: Method

- **Method 2**
  In the previous example, we used the information that sum of bernoullis follows a binomial distribution to construct the overall likelihood surface

Maximum Likelihood Estimation: Method

- **Method 2**
  In the previous example, we used the information that sum of bernoullis follows a binomial distribution to construct the overall likelihood surface
- In many cases this might not be possible because we are not working with such distributions and the model is at the level of individual coin tosses

Maximum Likelihood Estimation: Method

- **Method 2**
  In the previous example, we used the information that sum of bernoullis follows a binomial distribution to construct the overall likelihood surface
- In many cases this might not be possible because we are not working with such distributions and the model is at the level of individual coin tosses
- In such cases we need to construct the overall likelihood surface using the individual likelihoods

Maximum Likelihood Estimation: Method

- **Method 2**
  In the previous example, we used the information that sum of bernoullis follows a binomial distribution to construct the overall likelihood surface
- In many cases this might not be possible because we are not working with such distributions and the model is at the level of individual coin tosses
- In such cases we need to construct the overall likelihood surface using the individual likelihoods
- Each individual coin toss follows a Bernoulli distribution. Suppose $X = 1$ when heads and 0 otherwise

$$L(p; x) = p^x (1 - p)^{1-x}$$

Maximum Likelihood Estimation: Method

- **Method 2**: Remember the data is given to us

Observation	Outcome (x)	Likelihood of outcome
1	1	p
2	0	1-p
3	1	p
...		
99	0	1-p
100	1	p
Total	56	?

- What is the overall/joint likelihood of entries in the second column?

Maximum Likelihood Estimation: Method

- **Method 2**: Remember the data is given to us

Observation	Outcome (x)	Likelihood of outcome
1	1	p
2	0	1-p
3	1	p
...		
99	0	1-p
100	1	p
Total	56	?

- What is the overall/joint likelihood of entries in the second column?
- Each coin toss is independent

$$L(p) = p.p.p.p..(1-p)(1-p)...(1-p) \tag{9}$$

$$L(p) = p^{56}(1-p)^{44} \tag{10}$$

$$\ln L(p) = \ln(p^{56}(1-p)^{44}) \tag{11}$$

OLS as a special case of MLE

- Main assumption: The errors follow a normal distribution with mean 0 and variance σ^2

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

OLS as a special case of MLE

- Main assumption: The errors follow a normal distribution with mean 0 and variance σ^2

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

- The likelihood function of (β) is

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$
$$\ln(L(\beta)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

OLS as a special case of MLE

- Main assumption: The errors follow a normal distribution with mean 0 and variance σ^2

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

- The likelihood function of (β) is

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$
$$\ln(L(\beta)) = \sum_{i=1}^{n} \ln(\frac{1}{\sigma\sqrt{2\pi}}) - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

- The first term does not depend on (β) and the second term has a constant σ^2 that we can bring outside the summation

$$\ln(L(\beta)) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Maximum Likelihood Estimation

- Because we no longer have a direct connection between Y and our structural component $\boldsymbol{X}\boldsymbol{\beta}$, we need to specify our loss function in a different way. Using our link function, we can for every observation i wright down the probability of observing a certain value of Y_i given values of $\boldsymbol{X}_i$

Maximum Likelihood Estimation

- Because we no longer have a direct connection between Y and our structural component $\boldsymbol{X}\beta$, we need to specify our loss function in a different way. Using our link function, we can for every observation i wright down the probability of observing a certain value of Y_i given values of $\boldsymbol{X}_i$
- For example, for a logit model we have:

$$\Pr(Y = Y_i|\boldsymbol{X}_i) = \left(\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}\right)^{Y_i}\left(1-\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}\right)^{1-Y_i}$$

Maximum Likelihood Estimation

- Because we no longer have a direct connection between Y and our structural component $\boldsymbol{X}\beta$, we need to specify our loss function in a different way. Using our link function, we can for every observation i wright down the probability of observing a certain value of Y_i given values of $\boldsymbol{X}_i$
- For example, for a logit model we have:

$$\Pr(Y = Y_i | \boldsymbol{X}_i) = \left(\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}\right)^{Y_i}\left(1 - \frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}\right)^{1-Y_i}$$

- With the default assumption of i.i.d. observations we can wright down the joint probability or *likelihood function* of seeing our sample:

$$\ell(\beta) = \prod_{i=1}^{n} \Pr(Y = Y_i | \boldsymbol{X}_i)$$

Maximum Likelihood Estimation

- *Maximum likelihood* estimation (ML) is a method that chooses parameters $\boldsymbol{\beta}$ so as to minimize the loss function in form of the negative of the log likelihood function:

$$\widehat{\boldsymbol{\beta}}_{ML} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} - \ln \ell(\boldsymbol{\beta})$$

Maximum Likelihood Estimation

- *Maximum likelihood* estimation (ML) is a method that chooses parameters $\boldsymbol{\beta}$ so as to minimize the loss function in form of the negative of the log likelihood function:

$$\widehat{\boldsymbol{\beta}}_{ML} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} - \ln \ell(\boldsymbol{\beta})$$

- Under some general conditions $\widehat{\boldsymbol{\beta}}_{ML}$ is efficient, consistent and asymptotically normal, just like $\widehat{\boldsymbol{\beta}}_{OLS}$

- But unlike OLS, ML is a more general estimation procedure and allows one to recover structural parameters such as $\boldsymbol{\beta}$ in models that are far more flexible than standard MLR.