

# Econometricks: Short guides to econometrics

Trick 06: The Maximum Likelihood Estimator

Davud Rostam-Afschar (Uni Mannheim)

# Content

1. From Probability to Likelihood
2. The Econometric Model
3. Properties of the Maximum Likelihood Estimator

## The Likelihood Principle

Suppose you have three credit cards. You forgot, which has money on it or not. Thus, the number credit cards with money, call it  $\theta$ , might be 0, 1, 2, or 3. You can try your cards 4 times at random to check if you can make a payment.

The checks are random variables  $y_1, y_2, y_3$ , and  $y_4$ . They are

$$y_i = \begin{cases} 1, & \text{if the } i\text{th card has money on it,} \\ 0, & \text{otherwise.} \end{cases}$$

Since you chose  $y_i$ 's uniformly, they are i.i.d. and  $y_i \sim \text{Bernoulli}(\theta/3)$ . After checking, we find  $y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1$ . We observe 3 cards with money and 1 without.

The number credit cards with money could still be 0, 1, 2, or 3.

**Which is most likely?**

## From Probability to Likelihood

You could test for the true  $\theta_0$  in many samples. Conversely, you can check each possible value of  $\theta$  to find the probability of observing the sample ( $y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1$ ).

Since  $y_i \sim \text{Bernoulli}(\theta/3)$ , we have

$$\text{Prob}(y_i = y) = \begin{cases} \theta/3, & \text{for } y = 1, \\ 1 - \theta/3, & \text{for } y = 0. \end{cases}$$

Since  $y_i$ 's are independent, the joint PMF of  $y_1, y_2, y_3$ , and  $y_4$  can be written as

$$\begin{aligned} \text{Prob}(y_1 = y, y_2 = y, y_3 = y, y_4 = y | \theta) = \\ \text{Prob}(y_1) \text{Prob}(y_2) \text{Prob}(y_3) \text{Prob}(y_4). \end{aligned}$$

This depends on  $\theta$ , and is called **likelihood function**:

$$\begin{aligned} L(\theta | y_i) = \text{Prob}(y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1, \theta) = \\ \theta/3(1 - \theta/3)\theta/3\theta/3 = (\theta/3)^3(1 - \theta/3). \end{aligned}$$

## The Likelihood Principle

Values of the Likelihood  $L(\theta|y_i)$  for different  $\theta$

Trial	1	2	3	4
$\theta$	0	1	2	3
$Prob(\cdot)$	0.0000	0.0247	0.0988	0.0000

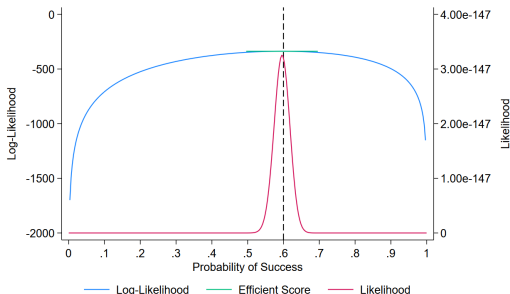
The probability of the observed sample for  $\theta = 0$  and  $\theta = 3$  is zero. This makes sense because our sample included both cards with and without money. The observed data is most likely to occur for  $\theta = 2$ .

**Likelihood principle:** choose  $\theta$  that maximizes the likelihood of observing the actual sample to get an estimator for  $\theta_0$ .

The likelihood is the probability from

- ▶ probability mass function if discrete
- ▶ probability distribution function if continuous

## From Likelihood to Log-Likelihood



- ▶ The **likelihood function**  $L_N(\theta|\mathbf{y}, \mathbf{X})$  is the joint probability mass function or density  $f(\mathbf{y}, \mathbf{X}|\theta)$ , viewed as a function of vector  $\theta$  given the data  $(\mathbf{y}, \mathbf{X})$ .
- ▶ Maximizing  $L_N(\theta)$  is equivalent to maximizing the **log-likelihood function**  $\mathcal{L}_N(\theta) = \ln L_N(\theta)$ . Because taking the logarithm is a monotonic transformation. A maximum for  $L_N(\theta)$  corresponds with a maximum for  $\mathcal{L}_N(\theta)$ .

## Specification of a Likelihood Function

The conditional likelihood  $L_N(\theta) = f(\mathbf{y}, \mathbf{X}|\theta)/f(\mathbf{X}|\theta) = f(\mathbf{y}|\mathbf{X}, \theta)$  does not require the specification of the marginal distribution of  $\mathbf{X}$ .

For observations  $(y_i, x_i)$  independent over  $i$  and distributed with  $f(\mathbf{y}|\mathbf{X}, \theta)$ ,

- ▶ the joint density is

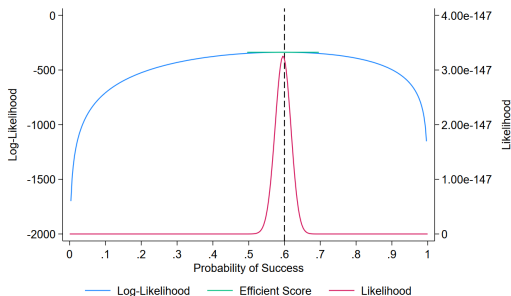
$$f(\mathbf{y}|\mathbf{X}, \theta) = \prod_{i=1}^N f(y_i|x_i, \theta),$$

- ▶ the log-likelihood function divided by  $N$  is

$$\frac{1}{N} \mathcal{L}_N(\theta) = \frac{1}{N} \sum_{i=1}^N \ln f(y_i|x_i, \theta).$$

Model	Range of $y$	Density $f(y)$	Common Parametrization
Bernoulli	0 or 1	$p^y(1-p)^{1-y}$	$p = \frac{e^{-x'\beta}}{1+e^{-x'\beta}}$
Poisson	0, 1, 2, ...	$e^{-\lambda} \lambda^y / y!$	$\lambda = e^{x'\beta}$
Exponential	$(0, \infty)$	$\lambda e^{-\lambda y}$	$\lambda = e^{x'\beta}$ or $1/\lambda = e^{x'\beta}$
Normal	$(-\infty, \infty)$	$(2\pi\sigma^2)^{-1/2} e^{-(y-\mu)^2/2\sigma^2}$	$\mu = x'\beta, \sigma^2 = \sigma^2$

# Maximum Likelihood Estimator

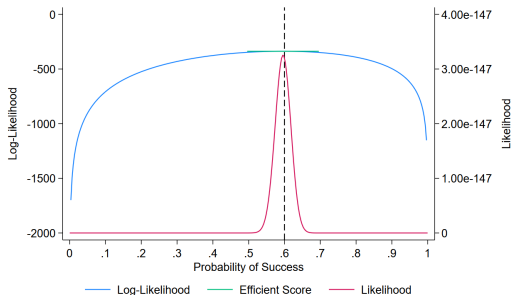


The **maximum likelihood estimator** (MLE) is the estimator that maximizes the (conditional) log-likelihood function  $\mathcal{L}_N(\theta)$ . The MLE is the local maximum that solves the first-order conditions

$$\frac{1}{N} \frac{\partial \mathcal{L}_N(\theta)}{\partial \theta} = \frac{1}{N} \sum_{i=1}^N \frac{\partial \ln f(y_i | x_i, \theta)}{\partial \theta} = \mathbf{0}.$$



# Maximum Likelihood Estimator



This estimator is an extremum estimator based on the conditional density of  $y$  given  $x$ . The gradient vector  $\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta}$  is called the **score vector**, as it sums the first derivatives of the log density, and when evaluated at  $\theta_0$  it is called the **efficient score**.

## How Were the Data Generated?

### Definition

#### **Simple Random Sampling.**

$\{x_{i1}, \dots, x_{iK}, y_i\}_{i=1}^N$  *i.i.d. (independent and identically distributed)*

This assumption means that

- ▶ observation  $i$  has no information content for observation  $j \neq i$
- ▶ all observations  $i$  come from the same distribution

This assumption is guaranteed by simple random sampling provided there is no systematic non-response or truncation.

## How Were the Data Generated?

I.i.d. data simplify the maximization as the joint density of the two variables is simply the product of the two marginal densities.

For example with a normal joint pdf with two observations

$$f(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2) = \frac{1}{2\pi\sigma^2} e^{-\frac{[(y_1-\mu)^2+(y_2-\mu)^2]}{2\sigma^2}}.$$

With dependent observations we would have to maximize the following likelihood function, where  $\rho$  is the correlation:

$$\frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} e^{-\frac{[(y_1-\mu)^2+(y_2-\mu)^2-(y_1-\mu)(y_2-\mu)]}{2\sigma^2(1-\rho^2)}}.$$

The Score has Expected Value Zero

Likelihood Equation:

$$E_f \left[ \mathbf{g}(\boldsymbol{\theta}) \right] = E_f \left[ \frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] = \int \frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\mathbf{x}, \boldsymbol{\theta}) dy = \mathbf{0}.$$

Example

$$\int f(y|\boldsymbol{\theta}) dy = 1. \quad \frac{\partial}{\partial \boldsymbol{\theta}} \int f(y|\boldsymbol{\theta}) dy = 0.$$

$$\int \frac{\partial f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dy = 0.$$

$$\partial \ln f(y|\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = [\partial f(y|\boldsymbol{\theta}) / \partial \boldsymbol{\theta}] / [f(y|\boldsymbol{\theta})]$$

$$\frac{\partial f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{\theta}).$$

$$\int \frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{\theta}) dy = 0.$$

## Fisher Information

The information matrix is the expectation of the outer product of the score vector,

$$\mathcal{I} = E_f \left[ \frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right].$$

The Fisher information  $\mathcal{I}$  is equals the variance of the score, since  $\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$  has mean zero.

- ▶ Large values of  $\mathcal{I}$  mean that small changes in  $\boldsymbol{\theta}$  lead to large changes in the log-likelihood  
→  $\mathcal{L}_N(\boldsymbol{\theta})$  contains considerable information about  $\boldsymbol{\theta}$ ,
- ▶ Small values of  $\mathcal{I}$  mean that the maximum is shallow and there are many nearby values of  $\boldsymbol{\theta}$  with a similar log-likelihood.

## Information Matrix Equality

The Fisher information  $\mathcal{I}$  is equal to the expectation of the Hessian  $\mathbf{H}$ :

$$-E_f \left[ \mathbf{H}(\boldsymbol{\theta}) \right] = -E_f \left[ \frac{\partial^2 \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = E_f \left[ \frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right].$$

### Example

For vector moment function, e.g.,  $\mathbf{m}(y, \boldsymbol{\theta}) = \frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$  with  $E[\mathbf{m}(y, \boldsymbol{\theta})] = 0$ ,

$$\int \mathbf{m}(y, \boldsymbol{\theta}) f(y|\boldsymbol{\theta}) dy = 0.$$

$$\int \left( \frac{\partial \mathbf{m}(y, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(y|\boldsymbol{\theta}) + \mathbf{m}(y, \boldsymbol{\theta}) \frac{\partial f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right) dy = 0.$$

$$\int \left( \frac{\partial \mathbf{m}(y, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(y|\boldsymbol{\theta}) + \mathbf{m}(y, \boldsymbol{\theta}) \frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(y|\boldsymbol{\theta}) \right) dy = 0.$$

$$E \left[ \frac{\partial \mathbf{m}(y, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right] = -E \left[ \mathbf{m}(y, \boldsymbol{\theta}) \frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right] = 0.$$

## The Information Matrix in Practice

The variance of the sum of random score vector is:

**Information matrix equality:**

$$\text{Var} \left[ \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\theta}) \right] = \text{Var}[\mathbf{g}(\boldsymbol{\theta})] = -E_f[\mathbf{H}(\boldsymbol{\theta})] = -E \left[ \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right].$$

After taking the expected value,  $\hat{\boldsymbol{\theta}}$  is substituted for  $\boldsymbol{\theta}$ . Problem: Taking the expected value of the second derivative matrix is frequently infeasible.

There exist two alternatives which are asymptotically equivalent:

- Ignore the expected value operator:

$$\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'}.$$

- Berndt-Hall-Hausman (BHHH) algorithm  
Never take a second derivative and sum over the outer product of the scores:  
(first derivatives per observation):

$$\check{\mathbf{I}}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^n \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' = \sum_{i=1}^n \left( \frac{\partial \ln f(y_i, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right) \left( \frac{\partial \ln f(y_i, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right)'.$$

## Properties of the MLE

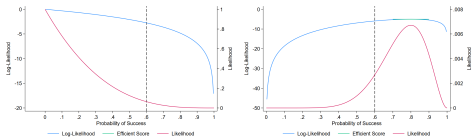
- ▶ *Small sample properties of  $\hat{\theta}$* 
  - ▶ may be biased
  - ▶ may have unknown distribution
  - ▶ variance may be biased, even towards zero
  
- ▶ *Large sample properties of  $\hat{\theta}$* 
  - ▶ consistent
  - ▶ approx. normal
  - ▶ asymptotically efficient
  - ▶ invariant



# Consistency

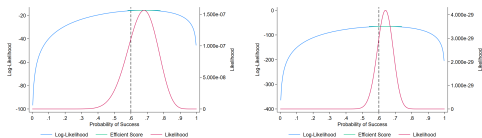
## Law of Large Numbers

As  $N$  increases, the distribution of  $\hat{\theta}$  becomes more tightly centered around  $\theta$ .



(a)  $N=3$

(b)  $N=10$



(c)  $N=25$

(d)  $N=100$

## Consistency

### Likelihood Inequality

$$E[(1/N) \ln L(\hat{\theta})] \geq E[(1/N) \ln L(\theta)].$$

The expected value of the log-likelihood is maximized at the true value of the parameters.

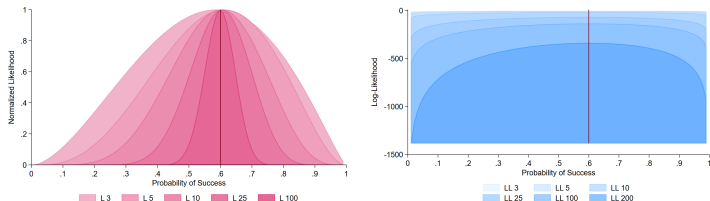


Figure 2:  $\hat{\theta}$ , Likelihood and Log-Likelihood as  $n \rightarrow \infty$ . True  $\theta = 0.6$ .

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \epsilon) = \mathbf{0}. \quad \lim_{n \rightarrow \infty} E[\hat{\theta}] = \theta.$$

## Approximate Normality

### Central Limit Theorem

As  $N$  becomes large,

$$\hat{\theta} \approx N\left[\theta, -\left(E\left[\frac{\partial^2 \mathcal{L}_N(\theta)}{\partial \theta \partial \theta'}\right]\right)^{-1}\right].$$

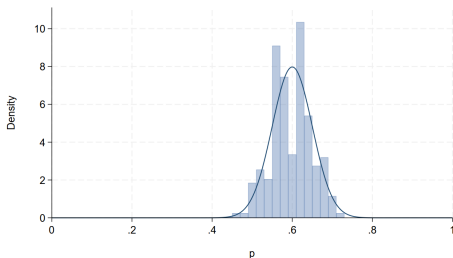


Figure 3: Sampling distribution of  $\hat{\theta}$  drawn from Bernoulli distribution and normal distribution at  $N = 100$ . True  $\theta = 0.6$ .

## Efficiency

The precision of the estimate  $\hat{\theta}$  is limited by the Fisher information  $\mathcal{I}$  of the likelihood.

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

For large samples, this is the so-called Cramér-Rao lower bound for the variance matrix of consistent asymptotically normal estimators with convergence to normality of  $\sqrt{N}(\hat{\theta} - \theta_0)$  uniform in compact intervals of  $\theta_0$ .

Under the strong assumption of correct specification of the conditional density, the MLE has the **smallest asymptotic variance** among root- $N$  consistent estimators.

### Example

Since the MLE is unbiased,

$$\mathbb{E}[\hat{\theta} - \theta \mid \theta] = \int (\hat{\theta} - \theta) f(y; \theta) dy = 0 \text{ regardless of the value of } \theta.$$

This expression is zero independent of  $\theta$ , so its partial derivative with respect to  $\theta$  must also be zero. By the product rule, this partial derivative is also equal to

$$0 = \frac{\partial}{\partial \theta} \int (\hat{\theta} - \theta) f(y; \theta) dy = \int (\hat{\theta} - \theta) \frac{\partial f}{\partial \theta} dy - \int f dy.$$

## Efficiency

### Example

For each  $\theta$ , the likelihood function is a probability density function, and therefore  $\int f dy = 1$ . By using the chain rule on the partial derivative of  $\ln f$  and then dividing and multiplying by  $f(y; \theta)$ , one can verify that

$$\frac{\partial f}{\partial \theta} = f \frac{\partial \ln f}{\partial \theta}.$$

Using these two facts, we get

$$\int (\hat{\theta} - \theta) f \frac{\partial \ln f}{\partial \theta} dy = 1.$$

Factoring the integrand gives  $\int ((\hat{\theta} - \theta) \sqrt{f}) (\sqrt{f} \frac{\partial \ln f}{\partial \theta}) dy = 1$ .

Squaring the expression in the integral, the Cauchy-Schwarz inequality yields

$$1 = \left( \int [(\hat{\theta} - \theta) \sqrt{f}] \cdot \left[ \sqrt{f} \frac{\partial \ln f}{\partial \theta} \right] dy \right)^2 \leq \left[ \int (\hat{\theta} - \theta)^2 f dy \right] \cdot \left[ \int \left( \frac{\partial \ln f}{\partial \theta} \right)^2 f dy \right].$$

The first factor is the expected mean-squared error (the variance) of the estimator  $\hat{\theta}$ , the second factor is the Fisher Information.

## Invariance

The MLE of  $\gamma = \mathbf{c}(\theta)$  is  $\hat{\theta} = \mathbf{c}(\hat{\theta})$  if  $\mathbf{c}(\theta)$  is a continuous and continuous differentiable function.

- ▶ This simplifies the log-likelihood,
- ▶ This allows a function of  $\hat{\theta}$  to serve as MLE if it is desired to analyze the function of an MLE.

## Example

Suppose that the normal log-likelihood is parameterized in terms of the precision parameter,  $\theta^2 = 1/\sigma^2$ . The log-likelihood becomes

$$\ln L(\mu, \sigma^2) = -(N/2) \ln(2\pi) + (N/2) \ln \theta^2 - \frac{\theta^2}{2} \sum_{i=1}^N (y_i - \mu)^2.$$

The MLE for  $\mu$  is  $\bar{x}$ . But the likelihood equation for  $\theta^2$  is now

$$\frac{\partial \ln L(\mu, \theta^2)}{\partial \theta^2} = 1/2 \left[ N/\theta^2 - \sum_{i=1}^N (y_i - \mu)^2 \right] = 0,$$

which has solution  $\hat{\theta}^2 = N / \sum_{i=1}^N (y_i - \mu)^2 = 1/\hat{\sigma}^2$ .

## Invariance

The MLE is also equivariant with respect to certain **transformations of the data**.

If  $y = c(x)$  where  $c$  is one to one and does not depend on the parameters to be estimated, then the density functions satisfy

$$f_Y(y) = \frac{f_X(x)}{|c'(x)|},$$

and hence the likelihood functions for  $x$  and  $y$  differ only by a factor that does not depend on the model parameters.

### Example

The MLE parameters of the log-normal distribution are the same as those of the normal distribution fitted to the logarithm of the data.

## References I

- CAMERON, A. C., AND P. K. TRIVEDI (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press, 3 edn., Section 4.1–4.5, 4.8–4.9.
- GREENE, W. H. (2011): *Econometric Analysis*. Prentice Hall, 5 edn.
- PISHRO-NIK, H. (2014): *Introduction to Probability, Statistics, and Random Processes*. Kappa Research LLC.