Econometricks: Short guides to econometrics

Trick 06: The Maximum Likelihood Estimator

Davud Rostam-Afschar (Uni Mannheim)

Content

Suppose you have three credit cards. You forgot, which has money on it or not. Thus, the number credit cards with money, call it θ, 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 $y_i \sim Bernoulli(\theta/3)$. 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 θ_0 in many samples. Conversely, you can check each possible value of θ to find the probability of observing the sample $(y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1)$.

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

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

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

$$Prob(y_1 = y, y_2 = y, y_3 = y, y_4 = y|\theta) =$$
$$Prob(y_1)Prob(y_2)Prob(y_3)Prob(y_4).$$

This depends on θ, and is called **likelihood function**:

$$L(\theta|y_i) = Prob(y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1, \theta) =$$
$$\theta/3(1 - \theta/3)\theta/3\theta/3 = (\theta/3)^3(1 - \theta/3).$$

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

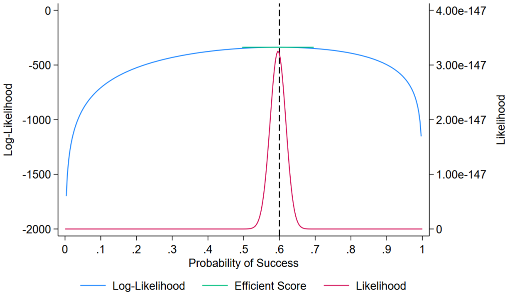
Trial	1	2	3	4
θ	0	1	2	3
$Prob(\cdot)$	0.0000	0.0247	0.0988	0.0000

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 θ that maximizes the likelihood of observing the actual sample to get an estimator for θ_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** $L_N(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})$ is the joint probability mass function or density $f(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\theta})$, viewed as a function of vector $\boldsymbol{\theta}$ given the data $(\boldsymbol{y}, \boldsymbol{X})$.
- Maximizing $L_N(\boldsymbol{\theta})$ is equivalent to maximizing the **log-likelihood function** $\mathcal{L}_N(\boldsymbol{\theta}) = \ln L_N(\boldsymbol{\theta})$. Because taking the logarithm is a monotonic transformation. A maximum for $L_N(\boldsymbol{\theta})$ corresponds with a maximum for $\mathcal{L}_N(\boldsymbol{\theta})$.

Specification of a Likelihood Function

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

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

- the joint density is

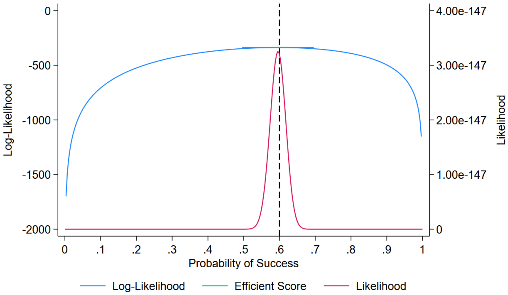
$$f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \Pi_{i=1}^{N} f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}),$$

- the log-likelihood function divided by N is

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

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

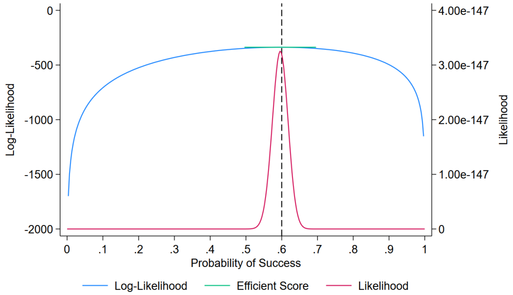
Maximum Likelihood Estimator



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

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

Maximum Likelihood Estimator



This estimator is an extremum estimator on based on the conditional density of y given $\boldsymbol{x}$. The gradient vector $\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is called the **score vector**, as it sums the first derivatives of the log density, and when evaluated at $\boldsymbol{\theta}_0$ it is called the **efficient score**.

How Were the Data Generated?

Definition

Simple Random Sampling.
$\{x_{i1}, \ldots, x_{iK}, y_i\}_{i=1}^{N}$ *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.

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

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

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

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

The Score has Expected Value Zero

Likelihood Equation:

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

Example

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

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

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

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

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

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

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

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
  $\rightarrow \mathcal{L}_N(\boldsymbol{\theta})$ 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 equals the expectation of the Hessian $\boldsymbol{H}$:

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

Example

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

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

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

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

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

The Information Matrix in Practice

The variance of the sum of random score vector is:
Information matrix equality:

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

After taking the expected value, $\widehat{\boldsymbol{\theta}}$ 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:

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

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

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

Properties of the MLE

- *Small sample properties of* $\hat{\boldsymbol{\theta}}$
  - may be biased
  - may have unknown distribution
  - variance may be biased, even towards zero

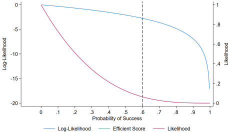
- *Large sample properties of* $\hat{\boldsymbol{\theta}}$
  - consistent
  - approx. normal
  - asymptotically efficient
  - invariant

Law of Large Numbers

Consistency

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

Consistency

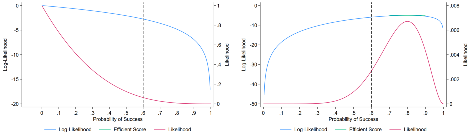


(a) N=3

Law of Large Numbers

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

Consistency

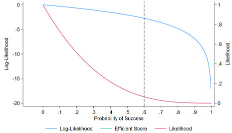


(a) N=3 (b) N=10

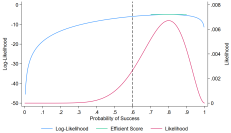
Law of Large Numbers
As N increases, the distribution of $\hat{\boldsymbol{\theta}}$ becomes more tightly centered around $\boldsymbol{\theta}$.

Consistency

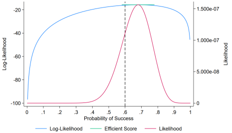
Law of Large Numbers
As N increases, the distribution of $\hat{\boldsymbol{\theta}}$ becomes more tightly centered around $\boldsymbol{\theta}$.



(a) N=3



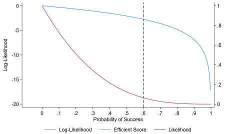
(b) N=10



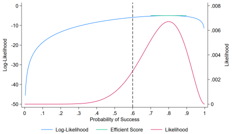
(c) N=25

Consistency

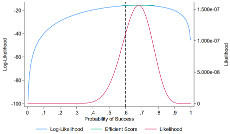
Law of Large Numbers
As N increases, the distribution of $\hat{\theta}$ becomes more tightly centered around θ.



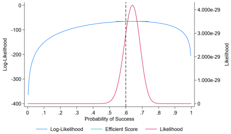
(a) N=3



(b) N=10



(c) N=25



(d) N=100

Likelihood Inequality

Likelihood Inequality

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

Consistency

Likelihood Inequality

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

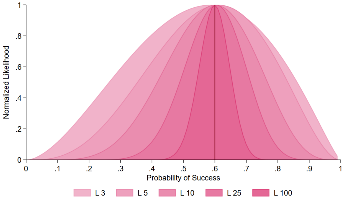
The expected value of the log-likelihood is maximized at the true value of the parameters.

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.

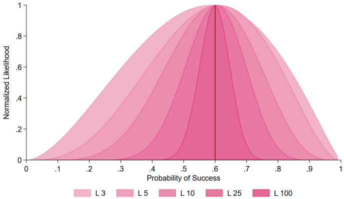


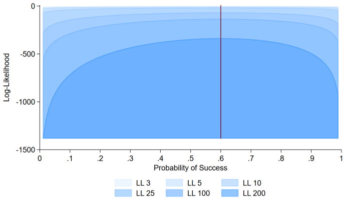
Consistency

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.





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.

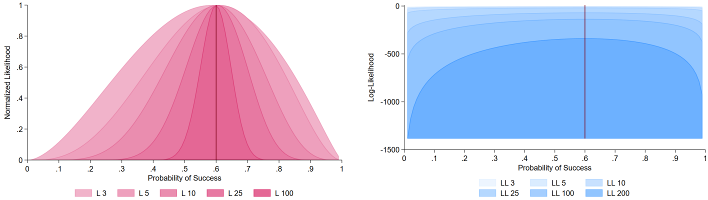


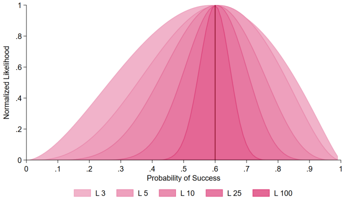
Figure 2: $\hat{\theta}$, Likelihood and Log-Likelihood as $n \to \infty$. True $\theta = 0.6$.

Consistency

Likelihood Inequality

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

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



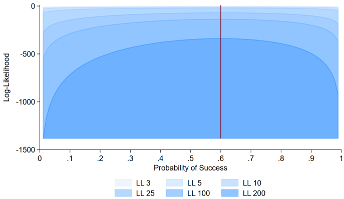


Figure 2: $\hat{\theta}$, Likelihood and Log-Likelihood as $n \to \infty$. True $\theta = 0.6$.

$$\lim_{n \to \infty} P(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \epsilon) = \mathbf{0}.$$

Consistency

Likelihood Inequality

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

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

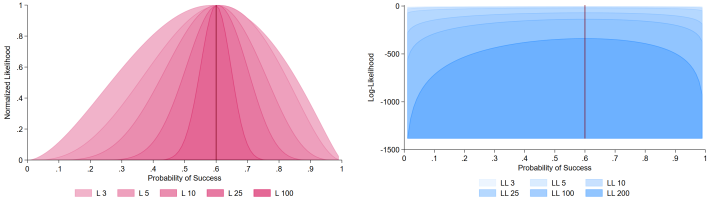


Figure 2: $\hat{\theta}$, Likelihood and Log-Likelihood as $n \rightarrow \infty$. True $\theta = 0.6$.

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

Approximate Normality

Central Limit Theorem

As N becomes large,

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

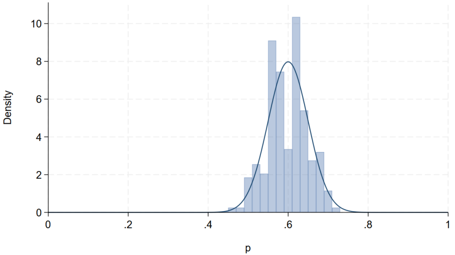


Figure 3: Sampling distribution of $\hat{\boldsymbol{\theta}}$ drawn from Bernoulli distribution and normal distribution at $N = 100$. True $\boldsymbol{\theta} = 0.6$.

Efficiency

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

$$\operatorname{Var}\left(\hat{\boldsymbol{\theta}}\right) \geq \frac{1}{\mathcal{I}\left(\boldsymbol{\theta}\right)}.$$

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 $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ uniform in compact intervals of $\boldsymbol{\theta}_0$.
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,

$$\mathrm{E}\left[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \;\middle|\; \boldsymbol{\theta}\right] = \int \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \, f(y; \boldsymbol{\theta}) \, dy = 0 \text{ regardless of the value of } \boldsymbol{\theta}.$$

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

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

Example

For each $\boldsymbol{\theta}$, 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 $f(y; \boldsymbol{\theta})$, one can verify that

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

Using these two facts, we get

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

Factoring the integrand gives $\int \left(\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \sqrt{f}\right) \left(\sqrt{f} \, \frac{\partial \ln f}{\partial \boldsymbol{\theta}}\right) \, dy = 1.$
Squaring the expression in the integral, the Cauchy-Schwarz inequality yields

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

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

Invariance

The MLE of $\boldsymbol{\gamma} = \boldsymbol{c}\left(\boldsymbol{\theta}\right)$ is $\widehat{\boldsymbol{\theta}} = \boldsymbol{c}(\widehat{\boldsymbol{\theta}})$ if $\boldsymbol{c}\left(\boldsymbol{\theta}\right)$ is a continuous and continuous differentiable function.

- This simplifies the log-likelihood,
- This allows a function of $\widehat{\boldsymbol{\theta}}$ 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 μ is $\bar{x}$. But the likelihood equation for θ^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.