Research Methods in Political Science I

10. Maximum Likelihood Method

Yuki Yanai

School of Law and Graduate School of Law

December 9, 2015

KOBE UNIVERSITY

Today's Menu

1. What Is Likelihood?
   - Likelihood Functions
2. Examples of Likelihood Functions
   - Discrete Case (1): Binomial Distribution
   - Discrete Case (2): Bernoulli Distribution
   - Continuous Case: Normal Distribution
3. Using Likelihood Functions
   - Likelihood Ratio
   - Method of Maximum Likelihood

Likelihood Functions（尤度関数）

For a given parameter value θ, we express the probability of obtaining the data D (right-hand side of the equation) as a function of θ

$$L(\theta|D) = \Pr(D|\theta)$$

$\rightarrow$ **Likelihood of θ** corresponding to the data D

- D: Data
- θ: (vector of) parameter(s)

Sometimes treat equivalence class together

$$L(\theta|D) = k\Pr(D|\theta) \propto \Pr(D|\theta),$$

where k is a constant.

Likelihood（尤度）

“Likelihood of θ_i” is the value of $L(\theta|D)$ evaluated at $\theta = \theta_i$

- $L(\theta_1|D)$: When D is observed, how likely that the parameter value is θ_1
- $L(\theta_2|D)$: When D is observed, how likely that the parameter value is θ_2

Likelihood is *not* an absolute measure: Compared *within* a model, higher value implies higher likelihood

Likelihood is *not* probability: no repeated-sample interpretation is available

Bayes Rule and Likelihood

Bayes rule:

$$\begin{aligned}\Pr(\theta|D) &= \frac{\Pr(D|\theta)\Pr(\theta)}{\Pr(D)} \\ &\propto \Pr(D|\theta)\Pr(\theta) \\ &\propto L(\theta|D)\Pr(\theta)\end{aligned}$$

- $\Pr(\theta)$: prior probability of θ (probability distribution of θ before observing D)
- $\Pr(\theta|D)$: posterior probability of θ (probability distribution of θ updated by the observed info D)

Can't accept Bayesian logic (you should...) $\rightarrow$ use likelihood ($\neq$ probability)

Example: Coin Toss (Coin Flipping)

A coin: Pr(H) = θ and Pr(T) = $1 - \theta$

Flipping a coin 10 times, we observed 8 heads and 2 tails. What is the probability that we observe "head" by flipping the coin once.

- Data D:
  - the number of coin flips: $n = 10$
  - the number of heads: $x = 8$
- the parameter we estimate: θ
- the likelihood: $L(\theta|D) = \Pr(D|\theta)$

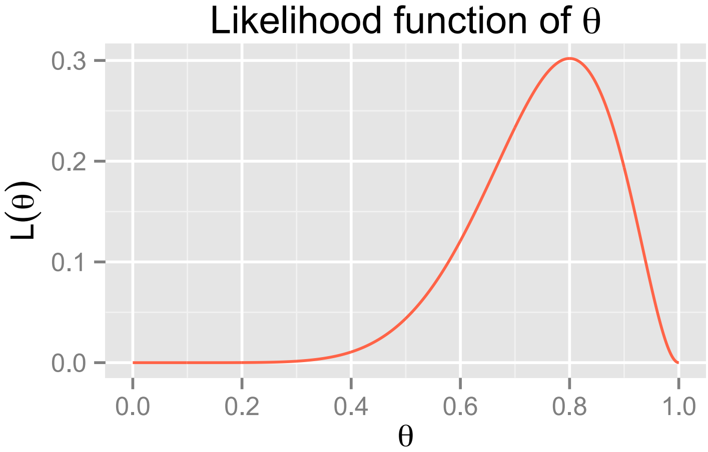
Specifying the Likelihood Function

$$\begin{aligned} L(\theta|D) &= \Pr(D|\theta) = \binom{10}{8}\theta^8(1-\theta)^{10-8} \\ &= 45\theta^8(1-\theta)^2 \end{aligned}$$

We'd like to find the value of θ that maximized $L(\theta|D)$: what is the most likely value of θ that generated the observed D

- $\theta = 0 \to L(\theta) = 0$: nope
- $\theta = 0.2 \to L(\theta) = 0.000073$: likely?
- $\theta = 0.6 \to L(\theta) = 0.12$: likely?
- $\theta = 0.8 \to L(\theta) = 0.30$: likely?
- $\theta = 0.9 \to L(\theta) = 0.19$: likely?
- $\theta = 1 \to L(\theta) = 0$: nope

Likelihood Function $L(\theta|D)$



Maximum of a Likelihood Function

Easy to find the maximum of a likelihood function in this example

$$L(\theta|D) = 45\theta^8(1-\theta)^2 = 45(\theta^{10} - 2\theta^9 + \theta^8)$$

First-order condition:

$$\begin{aligned}
\frac{d}{d\theta}L(\theta|D) = 90(5\theta^9 - 9\theta^8 + 4\theta^7) &= 0 \\
\Leftrightarrow 5\theta^9 - 9\theta^8 + 4\theta^7 &= 0 \\
\Leftrightarrow \theta^7(\theta-1)(5\theta-4) &= 0 \\
\Leftrightarrow \theta &= \frac{4}{5} \qquad (\because \theta \neq 0, 1)
\end{aligned}$$

Example

A coin: Pr(H) = θ and Pr(T) = $1-\theta$

Flipping a coin 10 times, we observed the result {H, H, T, H, H, H, H, H, H, T}. What is θ?

- Data D:

$$\begin{aligned} D &= \{H,H,T,H,H,H,H,H,H,T\} \\ &= \{1,1,0,1,1,1,1,1,1,0\} \end{aligned}$$

- the parameter we estimate: θ
- the likelihood: $L(\theta|D) = \Pr(D|\theta)$

Specifying the Likelihood Function (1)

Assuming each coin flip is independent,

$$L(\theta|D) = \Pr(D|\theta) = \prod_{i=1}^{10} \Pr(D_i|\theta) = \prod_{i=1}^{10} L_i(\theta|D_i),$$

where $D = \{D_1, D_2, \ldots, D_{10}\}$.

For each Bernoulli trial i,

$$L_i(\theta|D_i) = \Pr(D_i|\theta) = \theta^{D_i}(1-\theta)^{1-D_i}.$$

Thus,

$$L(\theta|D) = \prod_{i=1}^{10}[\theta^{D_i}(1-\theta)^{1-D_i}].$$

Specifying the Likelihood Function (2)

D_i is either 0 or 1:

$$L_i(\theta|D_i=1) = \theta^1(1-\theta)^0 = \theta,$$
$$L_i(\theta|D_i=0) = \theta^0(1-\theta)^1 = 1-\theta.$$

Therefore,

$$L(\theta|D) = \prod_{i=1}^{10} L_i(\theta|D_i) = \theta^8(1-\theta)^2$$

First-order condition for the maximum:

$$\frac{d}{d\theta}L(\theta|D) = 2\theta^7(\theta-1)(5\theta-4) = 0$$
$$\therefore \theta = \frac{4}{5} \qquad (\because \theta \neq 0, 1)$$

Log Likelihood

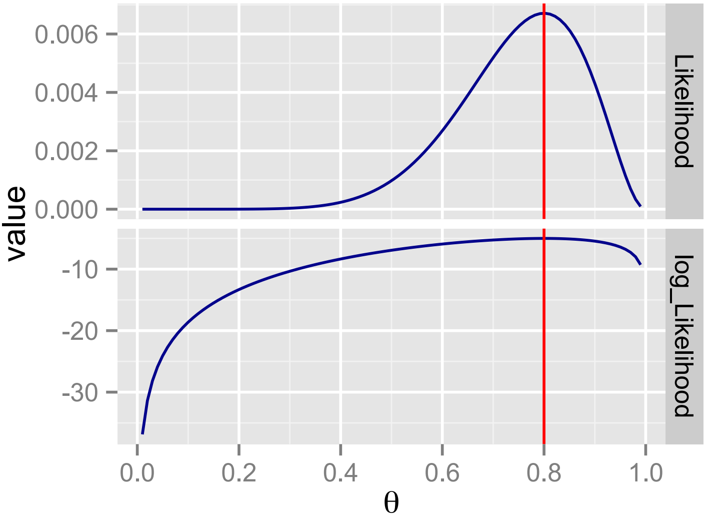
Natural logarithm is an increasing function:
$x_1 < x_2 \Rightarrow \log(x_1) < \log(x_2)$
$\rightarrow$ We can find the maximum of the likelihood by finding the maximum of the log-likelihood

$$\begin{aligned}\log[L(\theta|D)] &= \log\left(\prod_{i=1}^{10}[\theta^{D_i}(1-\theta)^{1-D_i}]\right) \\ &= \sum_{i=1}^{10}\log[\theta^{D_i}(1-\theta)^{1-D_i}] = 8\log\theta + 2\log(1-\theta)\end{aligned}$$

First-order condition for a maximum:

$$\begin{aligned}\frac{d}{d\theta}\log[L(\theta|D)] = \frac{8}{\theta} - \frac{2}{1-\theta} &= 0 \\ \Leftrightarrow \theta &= \frac{4}{5}\end{aligned}$$

Likelihood and Log-Likelihood



A Problem for Continuous Distributions (1)

$\Pr(X = x|\theta) = 0$: always gives us zero likelihood

- observed value has error ε (precision limit)
- observed value x: $x \in (x - \varepsilon/2, x + \varepsilon/2)$
- Suppose $p(x|\theta)$ is the PDF of a continuous random variable x, if ε is small enough

$$\begin{aligned} L(\theta|X) &= \Pr[X \in (x - \varepsilon/2, x + \varepsilon/2)] \\ &= \int_{x-\varepsilon/2}^{x+\varepsilon/2} p(X|\theta)dx \approx \varepsilon p(X|\theta) \end{aligned}$$

A Problem for Continuous Distributions (2)

- When we compare θ values within a model, we can multiply them by a constant (we treat equivalence class together)

  $\rightarrow$ we can ignore ε on the right-hand side of the equation above

We use PDF to construct likelihood functions of continuous variables

$$L(\theta|X) \propto p(X|\theta),$$

where $p(X|\theta)$ is the PDF of X given θ

Example: Normal Distribution

Example

Suppose that a random variable x is normally distributed, $x_i \sim N(\theta, \sigma^2), i = 1, 2, \ldots, n$, and σ^2 is known. What is the likelihood function of θ corresponding to the observed x

- PDF of N(θ, σ^2)

$$p(x|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]$$

Specifying Likelihood Function

- Likelihood of θ for each x_i is

$$L_i(\theta|x_i,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\theta)^2}{2\sigma^2}\right]$$

- The log-likelihood for the whole data is

$$\begin{aligned}\log L(\theta) &= \log\left[\prod_{i=1}^{n} L_i(\theta|x_i,\sigma^2)\right] \\ &= \sum_{i=1}^{n} \log L_i(\theta|x_i,\sigma^2) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\end{aligned}$$

Likelihood Ratio（尤度比）

How to compare two likelihoods $L(\theta_1|D)$ and $L(\theta_2|D)$?

- If a random variable x has one-to-one relationship with another variable y,

$$\frac{L(\theta_2|y)}{L(\theta_1|y)} = \frac{L(\theta_2|x)}{L(\theta_1|x)}$$

- Important: ratio of $L(\theta_1|D)$ to $L(\theta_2|D)$ (not the difference: consider why)
- Meaningless to evaluate a single likelihood alone: what if we multiply the likelihood function by a positive constant k?
- Generally, we can use a function $f(L)$ instead of L if $f'(.) > 0$: we prefer log-likelihood to likelihood
- Can ignore the terms without a parameter

Maximum Likelihood Estimate（MLE: 最尤推定値）

MLE: the maximum of the likelihood function $\rightarrow$ point estimate of maximum likelihood method

- MLE is the simplest summary of ML method
- MLE represents **only a part** of ML inference
- MLE is not sufficient to reveal characteristics of a likelihood function $\rightarrow$ **inference should be based on the likelihood function itself**
- MLE can be analytically obtained by solving the score equation
- MLE is usually obtained by numerical methods

Score Function and Fisher Information

- Score function: first derivative of the log-likelihood function

$$S(\theta) \equiv \frac{\partial}{\partial \theta} \log L(\theta)$$

- MLE $\hat{\theta}$ is obtained by the score equation: $S(\theta) = 0$
- the curvature at $\hat{\theta}$ is denoted by $I(\hat{\theta})$:

$$I(\hat{\theta}) \equiv -\frac{\partial^2}{\partial \theta^2} \log L(\hat{\theta})$$

  This is positive because the second-order differential coefficient at the maximum is negative
- $I(\hat{\theta})$: observed Fisher information: the larger the value, the less uncertain the location of the maximum θ

Score Function and Fisher Information (Eg 1-1)

Normal Distribution

A random variable x is normally distributed, $x_i \sim N(\theta, \sigma^2), i = 1, 2, \ldots, n$, where σ^2 is known. Obtain the MLE and the observed Fisher information of θ for the observed x.

- Ignoring the terms without θ,

$$\log L(\theta|x, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2.$$

- Score function is

$$S(\theta) = \frac{\partial}{\partial \theta} \log L(\theta|x, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \theta).$$

Score Function and Fisher Information (Eg 1-2)

- Differentiating the log-likelihood twice and changing the sign , we get the observed Fisher information:

$$I(\hat{\theta}) = \frac{n}{\sigma^2}$$

  - $\text{Var}(\hat{\theta}) = \sigma^2/n = I^{-1}(\hat{\theta})$: the higher the information value, the smaller the variance of the estimate
  - $\text{se}(\hat{\theta}) = \sigma/\sqrt{n} = I^{-1/2}(\hat{\theta})$

Score Function and Fisher Information (Eg 2-1)

Binomial Distribution

Running the Bernoulli trial with the success probability θ n times, x successes and $n-x$ failures have been observed. Obtain the MLE and the observed Fisher information of θ for x.

- Ignoring the constant term, the log-likelihood is

$$\log L(\theta) = x \log \theta + (n-x)\log(1-\theta).$$

- Score function is

$$S(\theta) = \frac{\partial}{\partial \theta} \log L(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

- Solving $S(\theta) = 0$, we get

$$\hat{\theta} = \frac{x}{n}.$$

Score Function and Fisher Information (Eg 2-2)

- Differentiating the log-likelihood twice and changing the sign, we get

$$I(\theta) \equiv -\frac{\partial^2}{\partial \theta^2} \log L(\theta) = \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}.$$

- Therefore, the observed Fisher information is

$$I(\hat{\theta}) = \frac{n}{\hat{\theta}(1-\hat{\theta})} = \frac{n^3}{x(n-x)}.$$

Quadratic Approximation

- When we can approximate the log-likelihood function by a quadratic function (called “regular” likelihood), we need at least two statistics to show the characteristics of the function
  1. location of the maximum (MLE): point estimate
  2. curvature at the maximum: uncertainty
- When the likelihood is approximately normal,

$$\log \frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

- This is exact for the normal likelihood

$$\log \frac{L(\theta)}{L(\hat{\theta})} = -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

Likelihood Intervals

MLE doesn't tell the uncertainty of estimation $\rightarrow$ interval estimation is desirable

- Likelihood interval: a set of θ that satisfies the following.

$$\left\{\theta : \frac{L(\theta)}{L(\hat{\theta})} > c\right\}$$

  - $c \in (0, 1)$: an arbitrary threshold
  - $L(\theta)/L(\hat{\theta})$: Normalized likelihood function

Example of Likelihood Interval

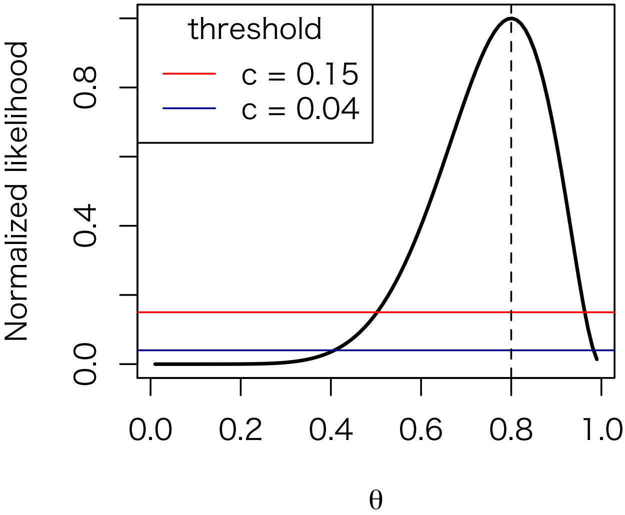
Got $x = 8$ heads by flipping a coin with the head probability θ $n = 10$ times

- $c = 0.15$: likelihood interval is $(0.50, 0.96)$
- $c = 0.04$: likelihood interval is $(0.41, 0.98)$

Problems in likelihood intervals

- How should we choose the value for c?
- How should we interpret a given interval?

Example of Likelihood Interval



Interval Estimation Based on Probability: Normal (1)

- Using the log-likelihood function of a normal mean derived above,

$$\log \frac{L(\theta)}{L(\hat{\theta})} = -\frac{n}{2\sigma^2}(\bar{x} - \theta)^2$$

- Because $\bar{x} \sim N(\theta, \sigma^2/n)$,

$$\frac{n}{\sigma^2}(\bar{x} - \theta)^2 \sim \chi_1^2$$

- That is,

$$W \equiv 2\log \frac{L(\hat{\theta})}{L(\theta)} \sim \chi_1^2$$

- W: Wilks's likelihood ratio statistic

(If n is large enough, other distributions can be approximated by χ^2)

Interval Estimation Based on Probability: Normal (2)

- Consider the probability of θ taking a value in a specific interval

$$\begin{aligned}\Pr\left(\frac{L(\theta)}{L(\hat{\theta})} > c\right) &= \Pr\left(2\log\frac{L(\hat{\theta})}{L(\theta)} < -2\log c\right) \\ &= \Pr(\chi_1^2 < -2\log c)\end{aligned}$$

- Here, we choose c by setting $0 < \alpha < 1$

$$c = \exp\left(-\frac{1}{2}\chi^2_{1,(1-\alpha)}\right),$$

where $\chi^2_{1,(1-\alpha)}$ is $100(1-\alpha)$ percentile of χ_1^2

Interval Estimation Based on Probability: Normal (3)

- Then,

$$\Pr\left(\frac{L(\theta)}{L(\hat{\theta})} > c\right) = \Pr(\chi^2_1 < \chi^2_{1,(1-\alpha)}) = 1 - \alpha.$$

- This gives us an interval comparable to $100(1-\alpha)$ percent CI
- Especially, $\alpha = 0.05$ when $c = 0.15$ and $\alpha = 0.01$ when $c = 0.04$

We can use the likelihood interval with $c = 0.15$ ($c = 0.04$) as a substitute of 95% (99%) CI

Likelihood Ratio Test（尤度比検定）

- Consider a null hypothesis H_0: $\theta = \theta_0$
- We reject the null if the following likelihood ratio is “too small”

$$\frac{L(\theta_0)}{L(\hat{\theta})}$$

- How small is “too small”? $\rightarrow$ requires probabilistic thinking
- Using Wilks’s likelihood ratio, if the likelihood ratio of the null is c, the p value is

$$p = \Pr(\chi^2_1 > -2\log c).$$

- This isn’t always true, unfortunately.

Standard Error

- When the log-likelihood can be approximated by a quadratic function,

$$\log \frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

- Thus, the interval that satisfies $\{\theta : L(\theta)/L(\hat{\theta}) > c\}$ is approximately

$$\theta \pm \sqrt{-2\log c} \cdot I(\hat{\theta})^{-1/2}.$$

- Generally, the standard error of the MLE $\hat{\theta}$ is

$$\text{se}(\hat{\theta}) = I(\hat{\theta})^{-\frac{1}{2}}.$$

Wald Statistic

- Using the se of the MLE, Wald statistic z is

$$z = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}.$$

- As $|z|$ grows, the likelihood of the null $\theta = \theta_0$ and the p value get smaller.
- 95% Wald interval is

$$\hat{\theta} \pm 1.96\text{se}(\hat{\theta})$$

- Strength of Wald intervals: symmetric about $\hat{\theta}$
- Weakness of Wald intervals: approximation doesn't work unless the log-likelihood is well approximated by a quadratic function

probability-based likelihood intervals are preferred in most cases

Next Week

Maximum Likelihood Method (cont.)

- Logistic (logit) regression by maximum likelihood method
- Probit regression