Lecture 7 Point Estimation

Ma Xuejun

School of Mathematical Sciences

Soochow University

https://xuejunma.github.io

Outline

1. Introduction
2. Method of Moments
3. Maximum Likelihood Estimations
4. Bayes Estimators

Introduction

- Suppose $X_1, X_2, \cdots, X_n$ is a sample from $f(x|\theta)$, we want to find a statistic $W(X_1, X_2, \cdots, X_n)$ which is an estimator of θ.

Method of Moments

Let $X_1, X_2, \cdots, X_n$ be a sample from $f(x|\theta_1, \cdots, \theta_k)$. Define

$$
\begin{aligned}
m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, \qquad & \mu_1 &= EX^1 \\
m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \qquad & \mu_2 &= EX^2 \\
&\vdots \\
m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad & \mu_k &= EX^k
\end{aligned}
$$

Then we equate $m_1 = \mu_1(\theta_1, \cdots, \theta_k)$, $m_2 = \mu_2(\theta_1, \cdots, \theta_k)$, $\cdots$, $m_k = \mu_k(\theta_1, \cdots, \theta_k)$ to find $\hat{\theta}_1, \cdots, \hat{\theta}_k$.

Example 7.2.1 (Normal Method of Moments)

If $X_1, X_2, \cdots, X_n, \sim N(\theta, \sigma^2)$. Then we have $\mu_1 = \theta$, $\mu_2 = \sigma^2 + \theta^2$.
Using method of moments, we equate

$$\bar{X} = \theta, \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \theta^2$$

$$\Longrightarrow \hat{\theta} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Example 7.2.2 (Binomial Method of Moments)

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Bin(k, p), where both k and p are unknown. Equate the first two moments:

$$\begin{aligned} \bar{X} &= kp \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= kp(1-p) + k^2p^2. \end{aligned}$$

After a little algebra, we have

$$\hat{k} = \frac{\bar{X}^2}{\bar{X} - (1/n)\sum(X_i - \bar{X})^2} \quad \text{and} \quad \hat{p} = \bar{X}/\hat{k}.$$

Maximum Likelihood Estimations

Let $X_1, X_2, \cdots, X_n$ be i.i.d. from $f(x|\theta_1, \cdots, \theta_k)$. Let $L(\boldsymbol{\theta}|\boldsymbol{x}) = L(\theta_1, \cdots, \theta_k|x_1, \cdots, x_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \cdots, \theta_k)$ be the likelihood function.

Definition 7.2.4

For each sample point $\boldsymbol{x}$, let $\hat{\theta}(\boldsymbol{x})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, with $\boldsymbol{x}$ held fixed. Then $\hat{\theta}(\boldsymbol{x})$ is a maximum likelihood estimator of $\boldsymbol{\theta}$.

Remark Sometime maximizing $l(\boldsymbol{\theta}|\boldsymbol{x}) = \log(L(\boldsymbol{\theta}|\boldsymbol{x}))$ is much easier than maximizing $L(\boldsymbol{\theta}|\boldsymbol{x})$.

- **Theorem 7.2.10 (Invariance Property of MLE)** If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

  Remark $X_1, \cdots, X_n$ in most cases do not have to be identically distributed.

Bayes Estimators

Let $\boldsymbol{X} \sim f(\boldsymbol{x}|\theta)$, $\theta \sim \pi(\theta)$, where $\pi(\theta)$ is the prior distribution of θ. Then after observing $\boldsymbol{X} = \boldsymbol{x}$, the posterior distribution of θ is

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})}$$

where $m(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{X}$,

$$m(\boldsymbol{x}) = \int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta.$$

Example 7.2.14

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Bernoulli(p). Then $Y = \sum_{i=1}^{n} X_i$ is Bin(n, p).
We assume the prior distribution on p is Beta(α, β). Then

$$
\begin{aligned}
f(y,p) &= \left[\binom{n}{y} p^y (1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right] \\
&= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} \\
f(y) &= \int_0^1 f(y,p)dp = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}
\end{aligned}
$$

Therefore $f(p|y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$, which is Beta$(y+\alpha, n-y+\beta)$. Bayes estimator of p is the mean of $f(p|y)$, which is

$$
\hat{p}_B = \frac{y+\alpha}{\alpha+\beta+n}.
$$

Definition 7.2.6 (Conjugate Family)

Let $\mathcal{F}$ denote the class of pdf or pmf $f(x|\theta)$(indexed by θ). A class Π of prior distribution is a conjugate family of $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all prior in Π, and all $x \in \mathcal{X}$

$$P(\theta) = \pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int_\theta f(\mathbf{y}|\theta)\pi(\theta)d\theta} \tag{12.3'}$$

- Estimating the normalising constant
- Markov chain Monte Carlo (MCMC)：Monte Carlo integration and Markov chain sampling

we estimated unknown parameters using the methods:

- Maximum likelihood ： Newton－Raphson(NR)
- MCMC

- NR比MCMC收敛的快。
- 对于多峰分布，NR只能找到一个，而MCMC可以找到多个。
- NW对于固定的似然初始值，它的路径是一样的；而MCMC即使初始值相同，它的路径也是随机的。

- **Target density** $P(\theta)$ is not always achievable because it may have a complex, or even unknown, form.
- Markov chains provide a method of drawing samples from target densities (regardless of their complexity).
- Using these conditional steps, we build up a chain of samples $(\theta^1, \cdots, \theta^M)$ after specifying a starting value θ^0

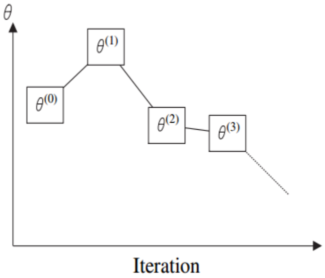

Figure 13.3 *A simple example of a Markov chain.*

Markov property:

$$P(\theta^i = a|\theta^{i-1}, \theta^{i-2}, \cdots, \theta^0) = P(\theta^i = a|\theta^{i-1})$$

An algorithm for creating a Markov chain for a target probability density $P(\theta)$ is:

1. Choose an initial value θ^0. The restriction on the initial value is that it needs to be within the distribution of $P(.)$, so that $P(\theta) > 0$
2. Create a new sample using $\theta^1 \sim \pi(\theta^1|\theta^0, \mathbf{y})$
3. Repeat step 2 M times, each time increasing both indices by 1.

Transitional density:$\pi(\theta^{i+1}|\theta^i)$,Normal distribution.

The Metropolis – Hastings sampler

- The Metropolis – Hastings sampler works by randomly **proposing a new value** θ^*
- If this proposed value is **accepted** (according to a criterion below),$\theta^{i+1} = \theta^*$
- If this proposed value is **rejected** (according to a criterion below),$\theta^{i+1} = \theta^i$
- Another proposal is made and the chain progresses by assessing this new proposal

- $\theta^* = \theta^i + Q$, Q ia called the **proposal density**,N(0,1) or U[-1.1]
- The acceptance criterion is:

$$\theta^{i+1} = \begin{cases} \theta^* \;\; if\, U < \alpha \\ \theta^i \;\; otherwise \end{cases}$$

where $U \sim U[-1,1]$ and

$$\alpha = \min\{\frac{\pi(\theta^*|\mathbf{y})}{\pi(\theta^i|\mathbf{y})} \cdot \frac{Q(\theta^i|\theta^*)}{Q(\theta^*|\theta^i)}, 1\}$$

where $P(\theta|\mathbf{y})$ is the probability of θ given the data $\mathbf{y}$ (the likelihood)
If the proposal density is symmetric($Q(a|b) = Q(b|a)$), then

$$\alpha = \min\{\frac{\pi(\theta^*|\mathbf{y})}{\pi(\theta^i|\mathbf{y})}, 1\}$$

例1：The Metropolis – Hastings sampler

设$Y_1, Y_2, \cdots, Y_n \sim^{iid} N(\mu, \sigma^2)$，$(\mu, \sigma^2)$的先验分布为$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$。计算$E(\mu|Y = y)$和$E(\sigma^2|Y = y)$。

解：(μ, σ^2)后验分布为

$$\pi(\mu, \sigma^2|Y = y) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left\{-\frac{\sum_{i=1}^{n}(y_1 - \mu)^2}{2\sigma^2}\right\}$$

- 《统计模拟及其R实现》P203
- 为了比较方法，假设$Y_1, Y_2, \cdots, Y_n \sim^{iid} N(2, 4^2)$

The Gibbs sampler

- The Gibbs sampler is another way of generating a Markov chain.
- It splits the parameters into a number of components and then updates each one in turn.
- For the beetle mortality example, a Gibbs sampler to update the two unknown parameters would be:
    1. Assign an initial value to the two unknowns: β_1^0 and β_2^0
    2. (a) Generate $\beta_2^1 \sim \pi(\beta_2|\mathbf{y}, \beta_1^0)$
       (b) Generate $\beta_1^1 \sim \pi(\beta_1|\mathbf{y}, \beta_2^0)$
    3. Repeat the step 2 M times, each time increasing the sample indices by 1.

上面举例比较简单，复杂的见《统计模拟及其R实现》P199 例8.4

例1续：The Gibbs sampler

设$Y_1, Y_2, \cdots, Y_n \sim^{iid} N(\mu, \sigma^2)$，$(\mu, \sigma^2)$的先验分布为$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$。计算$E(\mu|Y = y)$和$E(\sigma^2|Y = y)$ 。

解：(μ, σ^2)后验分布为

$$\pi(\mu, \sigma^2|Y = y) \propto (\frac{1}{\sigma^2})^{\frac{n}{2}+1} exp\{-\frac{\sum_{i=1}^{n}(y_1 - \mu)^2}{2\sigma^2}\}$$

则：

$$\pi(\mu|\sigma^2, y) \propto exp\{-\frac{\sum_{i=1}^{n}(y_1 - \mu}{2\sigma^2}\} \propto N(\overline{y}, \frac{\sigma^2}{n}) \quad (1)$$

$$\pi(\sigma^2|\mu, y) \propto (\frac{1}{\sigma^2}) exp\{-\frac{\sum_{i=1}^{n}(y_1 - \mu)^2}{2\sigma^2}\} \propto IG(\frac{n}{2}, \frac{\sum_{i=1}^{n}(y_1 - \mu)^2}{2}) \quad (2)$$

考虑Logit模型

例： 考虑54个老人智力得分。

Table 7.8 *Symptoms of senility (s=1 if symptoms are present and s=0 otherwise) and WAIS scores (x) for N=54 people.*

x	s	x	s	x	s	x	s	x	s
9	1	7	1	7	0	17	0	13	0
13	1	5	1	16	0	14	0	13	0
6	1	14	1	9	0	19	0	9	0
8	1	13	0	9	0	9	0	15	0
10	1	16	0	11	0	11	0	10	0
4	1	10	0	13	0	14	0	11	0
14	1	12	0	15	0	10	0	12	0
8	1	11	0	13	0	16	0	4	0
11	1	14	0	10	0	10	0	14	0
7	1	15	0	11	0	16	0	20	0
9	1	18	0	6	0	14	0		

注：中科大张伟平《计算统计讲义》

考虑Logit模型：

$$Y_i \sim Bin(1,\pi_i),\ \ log\frac{\pi_i}{1-\pi_i} = \beta_0+\beta_1 x_i,\ \ i=1,2,\cdots,54$$

则似然函数为：

$$\begin{aligned} f(\mathbf{y}|\beta_0,\beta_1) &= \prod_{i=1}^{n}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)^{1-y_i} \\ &= exp\{\sum_{i=1}^{n}[(\beta_0+\beta_1 x_i)y_i - log(1+e^{\beta_0+\beta_1 x_i})]\} \end{aligned}$$

考虑β_0,β_1的先验分布为独立的正态分布：

$$\beta_j \sim N(\mu_j,\sigma_j^2)$$

从而后验分布为:

$$
\begin{aligned}
f(\beta_0,\beta_1|\mathbf{y}) \propto & f(\mathbf{y}|\beta_0,\beta_1)\pi(\beta_0,\beta_1) \\
\propto & exp\{\sum_{i=1}^{n}[(\beta_0+\beta_1 x_i)y_i - log(1+e^{\beta_0+\beta_1 x_i})] \\
& -\frac{(\beta_0-\mu_0)^2}{\sigma_0^2}-\frac{(\beta_1-\mu_1)^2}{\sigma_1^2}\}
\end{aligned}
$$

Diagnostics of chain convergence

- Assumption:the sample densities for the unknown parameters were good estimates of the target densities.
- If this assumption is incorrect, then inferences could be invalid.
- We can only make valid inference when a chain has **converged** to the target density.
  - Chain history
  - Chain autocorrelation
  - Multiple chains

Chain history

A chain that has converged should show a **reasonable degree of randomness** between iterations, signifying that the Markov chain has found an area of high likelihood and is integrating over the target density (known as mixing).

- logit(π_i)=β_1+$\beta_2 x_1$
- logit(π_i)=β_1+$\beta_2(x_1 - \overline{x})$

Note: By centring the dose covariate we have greatly improved the convergence because centring reduces the correlation between the parameter estimates β_1 and β_2

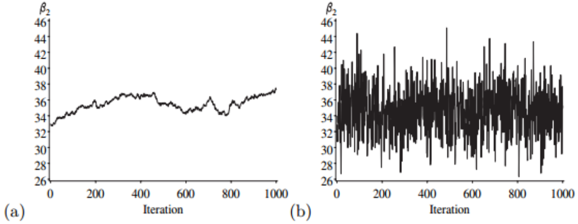

Figure 13.11 *Example of a chain showing (a) poor convergence and (b) reasonable convergence (first 1,000 iterations using Gibbs sampling). Estimate for* β_2 *using the logit link using two different parameterizations.*

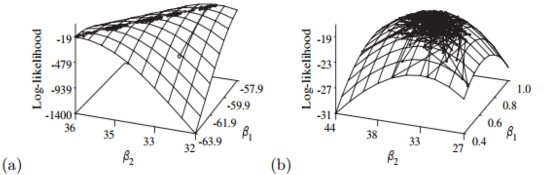

Figure 13.12 *Three-dimensional plots of the log-likelihood and 200 Gibbs samples for the beetle mortality data using the logit link function and (a) uncentered dose or (b) centered dose. The initial value is shown as an open circle and subsequent estimates as closed circles.*

13.5.2 Chain autocorrelation

- Autocorrelation is a useful diagnostic because it summarizes the dependence between neighbouring samples.
- Ideally we would like neighbouring samples to be completely independent, as this would be the most efficient chain possible.
- In practice we usually accept some autocorrelation, but large values (greater than 0.4) can be problematic.
- **Autocorrelation function (ACF)**

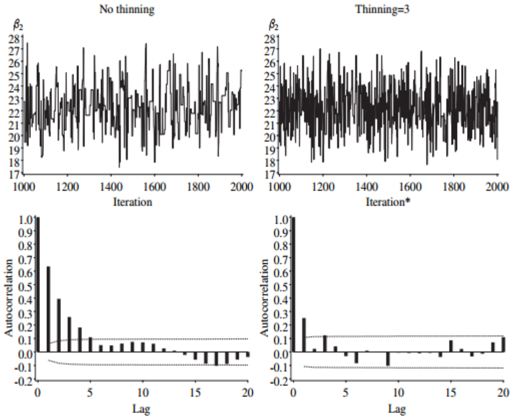

Figure 13.13 *Reduction in autocorrelation of Metropolis–Hastings samples after thinning. Chain history (top row) and ACF (bottom row) for the estimate of* β_2 *from the beetle mortality example using the extreme value model and a centered dose.*

Multiple chains

Advantage:

- Using multiple chains is a good way to assess convergence.
- If we start multiple chains at widely varying starting values and each chain converges to the same solution, this would increase our confidence in this solution.
- This method is particularly good for assessing the influence of initial values.

Drawback:It may be difficult to generate suitably varied starting values, particularly for complex problems with **many unknown parameters** and **multi-dimensional likelihoods**.

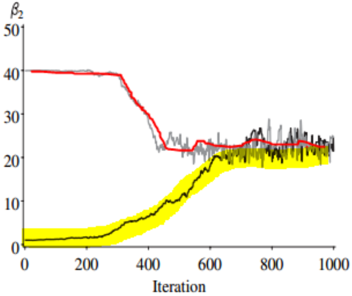

Figure 13.14 *Two chains with different starting values. Estimates of β_2 using Metropolis–Hastings sampling for the extreme value model using the beetle mortality data.*