Introduction to Maximum Likelihood Estimation

Olanrewaju Akande

Department of Statistical Science, Duke University

March 28, 2016

Lesson Plan

- Questions
- Introduction to ML estimation
- Illustrations
- Recap

Objectives

By the end of class, you should:

- Understand the concept of ML estimation
- Know how to set up a ML estimation problem
- Be able to find ML estimators for common distributions

Motivation

There are several ways of constructing estimators for parameters such as method of moments and maximum likelihood estimation.

ML estimation is very widely used and usually efficient.

The MLE approach chooses the value for the parameter that "maximizes the likelihood" of seeing the observed data.

This is the same as asking the following question: "given an assumed probability density (or mass) function for the data, what values of the parameters (out of the range of possible values) make the observed data least surprising?"

An Intuitive Illustration

Suppose we have three biased coins A, B and C: for A, $p(Head) = 0.8$, for B, $p(Head) = 0.4$ and for C, $p(Head) = 0.1$.

Suppose we randomly choose one coin, toss it ten times and see seven heads, which coin do you guess we rolled and why?

Most people would guess A and the idea is that we are less likely to see that many heads using B or C.

This is essentially the intuition behind the MLE approach – our guess of the parameter should be that for which the observed data are least surprising.

An Intuitive Illustration

Suppose we have three biased coins A, B and C: for A, $p(Head) = 0.8$, for B, $p(Head) = 0.4$ and for C, $p(Head) = 0.1$.

Suppose we randomly choose one coin, toss it ten times and see seven heads, which coin do you guess we rolled and why?

Most people would guess A and the idea is that we are less likely to see that many heads using B or C.

This is essentially the intuition behind the MLE approach – our guess of the parameter should be that for which the observed data are least surprising.

Introduction

Recall the function that links the probability of random variables to parameters:

$$f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m).$$

When the $x_1, \ldots, x_n$ are treated as variables and the the parameters $\theta_1, \ldots, \theta_m$ are treated as constants, this is the **joint density (mass) function**.

But when the $x_1, \ldots, x_n$ are treated as constants (the values observed in the sample) and the $\theta_1, \ldots, \theta_m$ are treated as variables, this is the **likelihood function**,

$$L(\theta_1, \ldots, \theta_m) = f(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n).$$

Comments

In general, one can show that maximum likelihood estimators

- have bias that goes to zero for large sample sizes
- have approximately minimum variance for large sample sizes
- often have approximately normal distributions for large sample sizes.

Additionally, if we are interested in estimating some function $h(\theta)$ rather than θ itself, the ML estimate of $h(\theta)$ is $h(\hat{\theta}_{MLE})$. This is not generally true for unbiased estimators or minimum variance unbiased estimators.

Note: When maximizing the likelihood function, it is often easier to maximize the log of the likelihood function. Since the log is a monotonic function, its maximum must occur for the same value of θ as does the likelihood function.

Illustrations

Example 1: Let $x_1, \ldots, x_n$ be an observed random sample from an exponential distribution with parameter λ. We want to find the ML estimate of λ.

First, we find the likelihood function:

$$\begin{aligned} f(x_1, \ldots, x_n; \lambda) &= \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) \\ &= \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i). \end{aligned}$$

Next, we solve this to find the value of λ that maximizes the likelihood function.

Illustrations

Example 1: Let $x_1, \ldots, x_n$ be an observed random sample from an exponential distribution with parameter λ. We want to find the ML estimate of λ.

First, we find the likelihood function:

$$\begin{aligned} f(x_1, \ldots, x_n; \lambda) &= \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) \\ &= \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i). \end{aligned}$$

Next, we solve this to find the value of λ that maximizes the likelihood function.

Illustrations

Taking logs, let $\ell(\lambda) = \ln f(x_1, ..., x_n; \lambda)$. Then

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

Next we take the derivative with respect to λ, set it to 0, and solve:

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Thus, the mle of λ is

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = 1/\bar{x}.$$

Is this MLE biased? **Yes**

$$\mathbf{E}[\hat{\lambda}] = \mathbf{E}[1/\bar{X}] \neq 1/\left(\mathbf{E}[\bar{X}]\right)$$

So it is a biased estimator. In fact, one can show $\mathbb{E}[\hat{\lambda}] = \frac{n}{n-1}\lambda$.

Illustrations

Taking logs, let $\ell(\lambda) = \ln f(x_1, ..., x_n; \lambda)$. Then

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

Next we take the derivative with respect to λ, set it to 0, and solve:

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Thus, the mle of λ is

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = 1/\bar{x}.$$

Is this MLE biased? **Yes**

$$\mathbb{E}[\hat{\lambda}] = \mathbb{E}[1/\bar{X}] \neq 1/\left(\mathbb{E}[\bar{X}]\right)$$

So it is a biased estimator. In fact, one can show $\mathbb{E}[\hat{\lambda}] = \frac{n}{n-1}\lambda$.

Illustrations

Taking logs, let $\ell(\lambda) = \ln f(x_1, ..., x_n; \lambda)$. Then

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

Next we take the derivative with respect to λ, set it to 0, and solve:

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Thus, the mle of λ is

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = 1/\bar{x}.$$

Is this MLE biased? **Yes**

$$\mathbb{E}[\hat{\lambda}] = \mathbb{E}[1/\bar{X}] \neq 1/\left(\mathbb{E}[\bar{X}]\right)$$

So it is a biased estimator. In fact, one can show $\mathbb{E}[\hat{\lambda}] = \frac{n}{n-1}\lambda$.

Illustrations

Example 2: Sometimes differentiating the likelihood (or log-likelihood) isn't the way to go. Consider estimating θ in a Unif$(0, \theta)$ distribution.

The joint density function is

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) &= \prod_{i=1}^{n} f(x_i; \theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} \\ &= \frac{1}{\theta^n} \end{aligned}$$

Therefore, the likelihood has the form $L(\theta) = \frac{1}{\theta^n}$ for $0 \leq x_i \leq \theta$ $(i = 1, \ldots, n)$ and 0 otherwise.

Illustrations

It can be seen that the MLE of θ must be a value θ for which $\theta \geq x_i$ for $i = 1, \ldots, n$ and which maximizes $1/\theta^n$ among all such values.

Since $1/\theta^n$ is a decreasing function of θ, the estimate will be smallest possible value of θ such that $\theta \geq x_i$ for $i = 1, \ldots, n$. In fact, this value is x_{max}, and it follows that the MLE of θ is $\hat{\theta} = max(X_1, X_2, \ldots, X_n)$

Thus the maximum likelihood estimate of θ is the sample maximum. This is slightly biased but the bias goes to zero as the sample size increases.

Illustrations

Example 3:

Let $x_1, .., x_n$ be an observed random sample from a normal distribution with unknown mean μ and unknown standard deviation σ. What are the MLEs?

First, the likelihood function:

$$\begin{aligned} f(x_1, \ldots, x_n; \mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]. \end{aligned}$$

$$\ell(x_1, \ldots, x_n; \mu, \sigma) = n \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Illustrations

Take partial derivatives to compute $\hat{\mu}$ and $\hat{\sigma}$ and solve

$$\begin{aligned} 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} \\ 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \sigma}. \end{aligned}$$

For μ,

$$\begin{aligned} 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ &= \sum_{i=1}^{n} (x_i - \mu) = \left(\sum_{i=1}^{n} x_i \right) - n\mu. \end{aligned}$$

and solving this for μ shows that the maximum likelihood estimate is $\hat{\mu} = \bar{x}$.

Illustrations

Take partial derivatives to compute $\hat{\mu}$ and $\hat{\sigma}$ and solve

$$\begin{aligned} 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} \\ 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \sigma}. \end{aligned}$$

For μ,

$$\begin{aligned} 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ &= \sum_{i=1}^{n} (x_i - \mu) = \left(\sum_{i=1}^{n} x_i \right) - n\mu. \end{aligned}$$

and solving this for μ shows that the maximum likelihood estimate is $\hat{\mu} = \bar{x}$.

Illustrations

Now, to find the mle for σ, we take the derivative of the log-likelihood with respect to σ, set it to 0, and solve:

$$\begin{aligned} 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \sigma} \\ &= \frac{\partial}{\partial \sigma}\left[n \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] \\ &= \frac{\partial}{\partial \sigma}\left[-n \ln \sqrt{2\pi} - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] \\ &= \frac{-n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2. \end{aligned}$$

In the penultimate step we used properties of the logarithm to simplify the first term.

$$n \ln \frac{1}{\sqrt{2\pi}\sigma} = -n \ln \sqrt{2\pi}\sigma = -n \ln \sqrt{2\pi} - n \ln \sigma.$$

Illustrations

Now, to find the mle for σ, we take the derivative of the log-likelihood with respect to σ, set it to 0, and solve:

$$\begin{aligned} 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \sigma} \\ &= \frac{\partial}{\partial \sigma}\left[n \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] \\ &= \frac{\partial}{\partial \sigma}\left[-n \ln \sqrt{2\pi} - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] \\ &= \frac{-n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2. \end{aligned}$$

In the penultimate step we used properties of the logarithm to simplify the first term.

$$n \ln \frac{1}{\sqrt{2\pi}\sigma} = -n \ln \sqrt{2\pi}\sigma = -n \ln \sqrt{2\pi} - n \ln \sigma.$$

Illustrations

Thus

$$\frac{n}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \quad \textbf{and} \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Substitute in the mle of μ, so

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Note:

- We need to check that we are maximizing the log-likelihood instead of minimizing it or finding an inflection point;
- The joint minimization wrt to both parameters requires solving a set of simultaneous equations, which is why we can substitute $\bar{x}$ for μ in finding the mle for σ.
- We haven't addressed the problem of multiple local maxima.

Recap

You should have a basic understanding of and intuition behind maximum likelihood estimation.

You should be able to:

- Set up ML estimation problems
- Derive MLEs for common distributions

Conclusion

Questions?