Lecture 11: Maximum Likelihood Estimators

- Maximum likelihood estimation
- Properties of maximum likelihood estimation
- Derivation

Objectives

By the end of class, you should:

- Understand the concept of maximum likelihood (ML) estimation
- Know how to set up ML estimation problems
- 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 (ML) 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:

1. for A, $p(Head) = 0.75$
2. for B, $p(Head) = 0.45$
3. for C, $p(Head) = 0.3$

Suppose we randomly choose one coin, toss it 100 times and see 65 heads, which coin do you guess we rolled? Why?

An Intuitive Illustration

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 is least surprising.

In our simple illustration, this is in fact a binomial experiment, thus, a $Bin(n, p)$ distribution where n is known to be 100 but we are unsure what p is and the only possible options for p are $0.75, 0.45$ and 0.3. The MLE approach would choose $p = 0.75$.

Essentially, we are looking at the binomial pmf in terms of a fixed x but variable p, whereas, previously we would look at a fixed p and variable x.

Introduction

Recall the function that calculates the probability of a random sample of size n given parameter values is the joint density (or mass function):

$$f(x_1, \ldots, x_n)$$

To be more explicit, we will sometimes show the dependence on parameters by writing

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

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**,

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

Maximum Likelihood Estimation

Suppose we have a random sample of i.i.d. random variables X_1, X_2, ..., X_n with a PMF or PDF $f_\theta(x)$ which depends on a parameter θ. The joint PMF/PDF is

$$f_\theta(x_1, x_2, \ldots, x_n) = f_\theta(x_1) f_\theta(x_2) \ldots f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

Upon observing the data, we can substitute $x_1, x_2, \ldots, x_n$ by the actual values in the sample, so $f_\theta(x_1, x_2, \ldots, x_n)$ becomes a function of θ alone. Seeing $f_\theta(x_1, x_2, \ldots, x_n)$ as a function of θ, we write

$$\mathcal{L}(\theta) = f_\theta(x_1, x_2, \ldots, x_n),$$

and call $\mathcal{L}(\theta)$ the **likelihood** of the data. The **Maximum Likelihood Estimator of θ (MLE)** is the value $\widehat{\theta}$ that maximizes the likelihood. Products are typically hard to maximize, so we usually take logarithms and maximize the **log-likelihood** $\ell(\theta) = \log \mathcal{L}(\theta)$ instead.

Properties of MLE

Maximum-likelihood estimators have no optimum properties for finite samples, in the sense that (when evaluated on finite samples) other estimators may have greater concentration around the true parameter-value. However, maximum likelihood estimation possesses a number of attractive **limiting properties**: As the sample size increases to infinity, sequences of maximum likelihood estimators have these properties:

- MLEs are **consistent**
- have bias that goes to zero as $n \to \infty$ (**asymptotic unbiasedness**)
- often have **approximately Normal** distributions for large sample
- **Efficiency**, i.e. it achieves the Cramér-Rao lower bound when the sample size tends to infinity. This means that no consistent estimator has lower asymptotic mean squared error than the MLE.

Additionally, if $\widehat{\theta}$ is the MLE for θ and if g is a one-to-one function, then $g(\widehat{\theta})$ is the MLE for $g(\theta)$. (**Invariance**) This is not generally true for unbiased estimators or minimum variance unbiased estimators.

Illustrations

Example 1: Let $X_1, X_2, \ldots, X_n$ be a random sample from a discrete distribution with support $\{0, 1, 2\}$. Suppose that θ, the parameter of interest, can only take on the values $\theta = 0$ and $\theta = 1$. The PMFs for $\theta = 0$ and $\theta = 1$ are:

	$\theta = 0$	$\theta = 1$
$X = 0$	0.1	0.2
$X = 1$	0.3	0.5
$X = 2$	0.6	0.3

Suppose we observe a random sample of size 4 and the values are 0, 0, 1, 2. What is the MLE of θ? The likelihood at $\theta = 0$ is

$$\begin{aligned}\mathcal{L}(0) &= P_{\theta=0}(X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 2) \\ &= P_{\theta=0}(X = 0)^2 P_{\theta=0}(X = 1) P_{\theta=0}(X = 2) = 0.1^2 \times 0.3 \times 0.6 = 0.0018\end{aligned}$$

and analogously, $\mathcal{L}(1) = 0.2^2 \times 0.5 \times 0.3 = 0.006$, so the MLE in this case is $\widehat{\theta} = 1$.

Illustrations

Example 2: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$ with p unknown, and suppose that $x_1, x_2, \ldots, x_n$ have been observed, then:

$$\begin{aligned}\mathcal{L}(p) &= \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\ &= p^{S_n}(1-p)^{n-S_n}\end{aligned}$$

where $S_n = \sum_{i=1}^n x_i$. The likelihood function can be evaluated at fixed values of p, say, $p = 0.35$, $\mathcal{L}(0.35)$ is the likelihood of observing the data if the true value of p were 0.35. Since the MLE $\widehat{p}$ is the value of p that maximizes the likelihood, we can say that the maximum likelihood estimator is the value of p that is "most likely" to have generated the data.

Illustrations

The log-likelihood is

$$\ell(p) = S_n \log p + (n - S_n) \log (1 - p)$$

The first order derivative of the log-likelihood is

$$\ell'(p) = \frac{S_n}{p} - \frac{(n - S_n)}{(1 - p)}$$

and setting it to 0, and we find that the MLE is the **sample proportion**:

$$\widehat{p} = \frac{S_n}{n} = \overline{X_n}$$

which is a maximum since $\ell''(\widehat{p}) < 0$.

Illustrations

Example 3: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$, then:

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = C\lambda^{S_n} e^{-n\lambda}$$

where $S_n = \sum_{i=1}^{n} x_i$ and $C = 1/\prod_{i=1}^{n} x_i!$. The log-likelihood is

$$\ell(\lambda) = \log(C) + S_n \log\lambda - n\lambda,$$

so the derivative of the log-likelihood is

$$\ell'(\lambda) = S_n/\lambda - n$$

Therefore, we can find $\widehat{\lambda} = S_n/n = \overline{X}_n$. It is easy to check that $\ell''(\widehat{\lambda}) < 0$, so $\widehat{\lambda}$ is indeed a maximum.

Illustrations

Example 4: Let $x_1, \ldots, x_n$ be observed values of a random sample from an exponential distribution with parameter λ. **By the way, the exponential pdf is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ where $\lambda > 0$.** We want to find the ML estimate of λ.

First, we find the likelihood function:

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

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

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = 1/\bar{x}, \quad \frac{d^2\ell(\lambda)}{d\lambda^2}|_{\lambda=\hat{\lambda}} = -\frac{n}{\hat{\lambda}^2} < 0$$

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 find the distribution of $1/\bar{X}$

Illustrations

Example 5: Sometimes differentiating the likelihood (or log-likelihood) isn't the way to go. Consider estimating θ in a Unif$(0, \theta)$ distribution. **We remember the form of the pdf from the previous class.**

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 6:

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?

For a normal distribution, the pdf is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right].$$

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

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

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

Notes:

- We need to check that we are maximizing the log-likelihood
- The joint minimization wrt (with respect 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

Suggested reading:

- D.S. Sec. 7.5, 7.6