Part 3: Properties of MLEs

Summary

In this Part we explore some of the important properties possessed by maximum likelihood estimators.

Briefly stated, we will see that, asymptotically, these estimators cannot be beat, in the sense that they have minimal MSE, and also are approximatley normally distributed with variance that is easy to calculate. This makes quantifying the error in the estimators straightforward.

Invariance of the MLE

The invariance result can be stated as follows: If $\widehat{\theta}_{\text{MLE}}$ is the MLE for θ, then $g(\widehat{\theta}_{\text{MLE}})$ is the MLE for $g(\theta)$.

This is a useful, important result because it is often the case that we are truly interested in estimating some function of the parameter θ, not θ itself.

Using this, we can say, for example, that if $X_1, X_2, \ldots, X_n$ are i.i.d. normal(μ, σ^2), then the MLE for σ is

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

and the MLE for the signal to noise ratio μ/σ is

$$\overline{X} \Big/ \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

One consequence of this property is that estimation is not affected by how one chooses to parameterize a model.

For example, some choose to represent the Exponential distribution using the "mean parameterization" which leads to $E(X) = \mu$, whereas we have used the "rate parameterization" which leads to $E(X) = 1/\lambda$. Because of invariance, these two approaches will lead to the same estimates, i.e., $\widehat{\mu} = 1/\widehat{\lambda}$.

Exercise: Does unbiasedness have this invariance property?

No. If $\hat{\theta}$ is an unbiased estimator for θ, no guarantee that $g(\hat{\theta})$ is unbiased for $g(\theta)$. For example, S is a biased estimator for σ, even when s^2 is unbiased for σ^2.

Exercise: Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. Exponential(λ). What is the MLE of the tail probability $P(X_i > a)$? assuming $a > 0$.

In exponential case, $P(X_i > a) = 1 - F_{X_i}(a) = e^{-\lambda a}$ (if $a > 0$)
The MLE for λ is $\hat{\lambda} = 1/\bar{X}$. So MLE for $e^{-\lambda a}$ is $e^{-a/\bar{X}}$.

Consistency of the MLE

An estimator $\widehat{\theta}$ is consistent for θ if as the sample size n increases, the estimator converges to the true value of the parameter.

Formally, we would write: For any $\epsilon > 0$,

true value of parameter Θ

$$\lim_{n \to \infty} P(|\widehat{\theta} - \theta_0| > \epsilon) = 0$$

"Convergence in prob."

This is a fairly low standard placed on an estimator, and, fortunately, under certain "regularity conditions," the MLE is consistent.

Asymptotic Normality of the MLE

Another great property of the MLE: The MLE is asymptotically normal, with variance that is not difficult to calculate, under certain "regularity conditions." (See page 516 from Casella and Berger, Second Edition.)

This will allow an easy route to:

1. Calculate standard errors for MLEs
2. Constructing confidence intervals for parameters
3. Constructing confidence intervals for general functions of parameters (via the Delta method)
4. Constructing hypothesis tests for parameters

From Casella and Berger, Second Edition

Although estimators such as d_n, called *superefficient*, can be constructed in some generality, they are more of a theoretical oddity than a practical concern. This is because the values of θ for which the variance goes below the bound are a set of Lebesgue measure 0. However, the existence of superefficient estimators serves to remind us to always be careful in our examination of assumptions for establishing properties of estimators (and to be careful in general!).

10.6.2 Suitable Regularity Conditions

The phrase "under suitable regularity conditions" is a somewhat abused phrase, as with enough assumptions we can probably prove whatever we want. However, "regularity conditions" are typically very technical, rather boring, and usually satisfied in most reasonable problems. But they are a necessary evil, so we should deal with them. To be complete, we present a set of regularity conditions that suffice to rigorously establish Theorems 10.1.6 and 10.1.12. These are not the most general conditions but are sufficiently general for many applications (with a notable exception being if the MLE is on the boundary of the parameter space). Be forewarned, the following is not for the fainthearted and can be skipped without sacrificing much in the way of understanding.

These conditions mainly relate to differentiability of the density and the ability to interchange differentiation and integration (as in the conditions for Theorem 7.3.9). For more details and generality, see Stuart, Ord, and Arnold (1999, Chapter 18), Ferguson (1996, Part 4), or Lehmann and Casella (1998, Section 6.3).

The following four assumptions are sufficient to prove Theorem 10.1.6, consistency of MLEs:

($A1$) We observe $X_1, \ldots, X_n$, where $X_i \sim f(x|\theta)$ are iid.

($A2$) The parameter is *identifiable*; that is, if $\theta \neq \theta'$, then $f(x|\theta) \neq f(x|\theta')$.

($A3$) The densities $f(x|\theta)$ have common support, and $f(x|\theta)$ is differentiable in θ.

($A4$) The parameter space Ω contains an open set ω of which the true parameter value θ_0 is an interior point. True value not on boundary of parameter space

The next two assumptions, together with ($A1$)–($A4$) are sufficient to prove Theorem 10.1.12, asymptotic normality and efficiency of MLEs.

($A5$) For every $x \in \mathcal{X}$, the density $f(x|\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ, and $\int f(x|\theta)\,dx$ can be differentiated three times under the integral sign.

($A6$) For any $\theta_0 \in \Omega$, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x|\theta)\right| \leq M(x) \quad \text{for all } x \in \mathcal{X}, \quad \theta_0 - c < \theta < \theta_0 + c,$$

with $\mathrm{E}_{\theta_0}[M(X)] < \infty$.

To start, we focus on the case where the parameter θ is one-dimensional.

The formal result can be stated as follows:

If $X_1, X_2, \ldots, X_n$ are i.i.d. $f(x; \theta_0)$, then, under suitable regularity conditions, as n increases,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \longrightarrow N\left(0, I^{-1}(\theta_0)\right)$$

where

$$I(\theta) = E_\theta\left(-\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right)$$

is the Fisher Information.

In the definition of $I(\theta)$, X is random, and the subscript on E_θ is meant to emphasize that X has distribution specified by $f(x; \theta)$.

Comment: It would be technically more correct to define the Fisher information as

$$I(\theta) = E_\theta\left(-\frac{\partial^2 \log f(X;t)}{\partial t^2}\bigg|_{t=\theta}\right)$$

since this makes a distinction between the θ which is serving as the input into the function $I(\theta)$ and the θ used in the two derivatives of $\log f(X;\theta)$.

You do not typically see it written this way (even in advanced texts) because it just makes things more messy without much conceptual gain.

Exercise: Explain the practical value of the above result.

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \longrightarrow N\left(0, I^{-1}(\theta_0)\right) \quad \text{as } n \to \infty$$

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \approx N\left(0, I^{-1}(\theta_0)\right) \quad \text{for } n \text{ "large"}$$

$$\hat{\theta}_{MLE} \approx N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

So, $E(\hat{\theta}_{MLE}) \approx \theta_0$, i.e. approx. unbiased for θ

and $V(\hat{\theta}_{MLE}) \approx \frac{1}{nI(\theta_0)} \approx \frac{1}{nI(\hat{\theta})}$

$$\text{SE of } \hat{\theta}_{MLE} \approx \frac{1}{\sqrt{nI(\hat{\theta})}}$$

Also, $\hat{\theta}_{MLE} \approx$ Normal.

Exercise: Show that

$$nI(\theta) = E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log L(\theta)\right]$$

$\ell(\theta)$

$$nI(\theta) = n\,E_\theta\left[-\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right]$$

$$= \sum_{i=1}^{n} E_\theta\left[-\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta}\right] \quad \text{since } X_1, X_2, \ldots, X_n \text{ are iid}$$

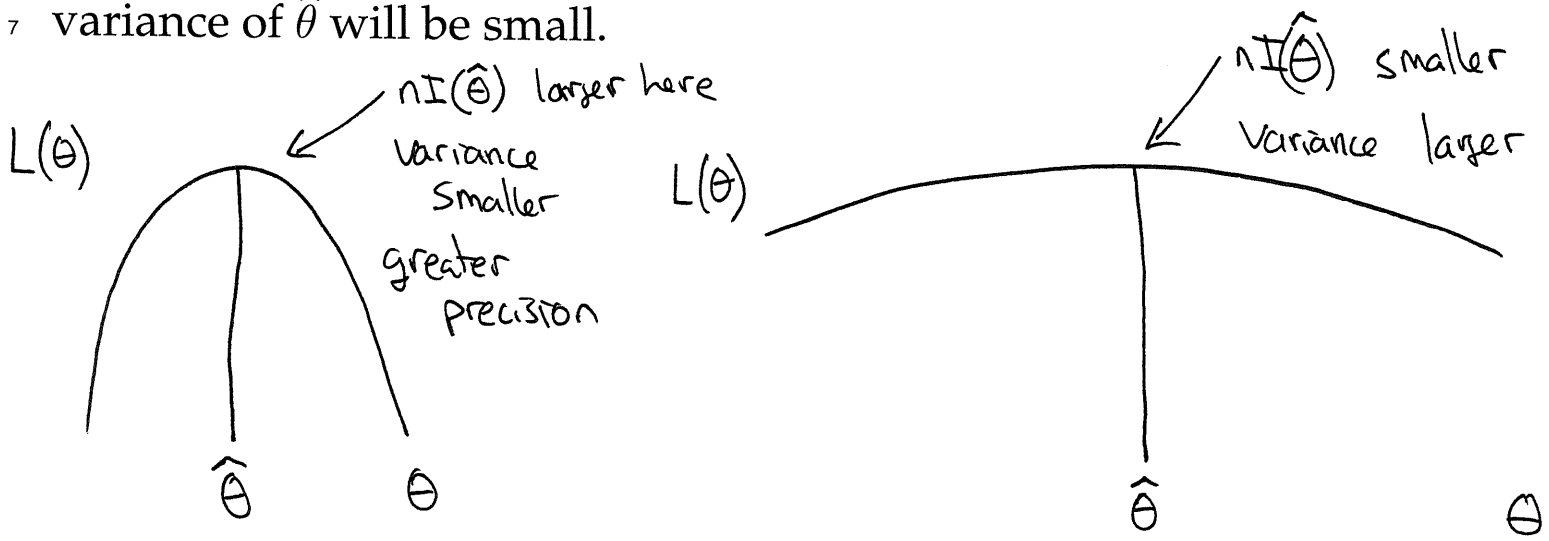
$$= E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log f(X_i;\theta)\right]$$

$$= E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right]$$

The previous results show that the variance of the MLE is approximately equal to the reciprocal of the expected value of the negative second derivative of the log likelihood at its peak.

And, "the negative of the second derivative" is the amount of curvature at the peak of the likelihood.

If the log likelihood has a "sharp" peak, then $nI(\hat{\theta})$ will be large, and the variance of $\hat{\theta}$ will be small.



Exercise: Assume $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, where σ^2 is known. We know that $\overline{X}$ is the MLE for μ. Derive the Fisher Information for μ, and use it to approximate the standard error and distribution of $\overline{X}$.

$$\ell(\mu) = -\frac{\sum(X_i-\mu)^2}{2\sigma^2} + \text{constant}$$ (see Slide 30 of previous part)

$$\frac{\partial \ell(\mu)}{\partial \mu} = \frac{\sum(X_i-\mu)}{\sigma^2} = \frac{\sum X_i - n\mu}{\sigma^2}$$

$$\frac{\partial^2 \ell(\mu)}{\partial \mu} = -\frac{n}{\sigma^2}$$

$$nI(\mu) = E_\mu\left[-\frac{\partial^2}{\partial \mu^2}\ell(\mu)\right] = E_\mu\left[\frac{n}{\sigma^2}\right] = n/\sigma^2$$

(note that $I(\mu) = 1/\sigma^2$)

The asymptotic normality of MLE tells me that $\bar{X}$ is approx. normal with mean μ_0 and variance

$$\frac{1}{nI(\mu_0)} = 1/n/\sigma^2 = \frac{\sigma^2}{n}$$

In fact, we already knew that $\bar{X}$ is exactly normal with mean μ_0 and variance σ^2/n

Exercise: Assume that $X_1, X_2, \ldots, X_n$ are i.i.d. Exponential(λ). Derive the Fisher Information for λ and the approximate distribution of its MLE.

Already found $\hat{\lambda}_{MLE} = 1/\bar{x}$, $f(x;\lambda) = \lambda e^{-\lambda x}$, $x > 0$

$\log f(x;\lambda) = \log \lambda - \lambda x$, $x > 0$

$$\frac{\partial \log f(x;\lambda)}{\partial \lambda} = \frac{1}{\lambda} - x$$

$$\frac{\partial^2 \log f(x;\lambda)}{\partial \lambda^2} = -\frac{1}{\lambda^2}$$

So, $I(\lambda) = E_\lambda\left[-\frac{\partial^2 \log f(X;\lambda)}{\partial \lambda^2}\right] = E_\lambda\left[+\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}$

So, $\hat{\lambda}_{MLE}$ is approx. normal with mean λ_0 and variance $\frac{1}{nI(\lambda_0)} = \lambda_0^2/n$. In practice, use $\hat{\lambda}^2/n$ as approx. variance.

Exercise: Assume that $X_1, X_2, \ldots, X_n$ are i.i.d. Poisson(λ). Derive the Fisher Information for λ and the approximate distribution of its MLE.

Approx. variance is $\hat{\lambda}/n$

Efficiency of the MLE

It is possible to show that the lower bound on the variance of any unbiased estimator is $1/nI(\theta_0)$. This is called the Cramer-Rao lower bound.

Since the MLE is asymptotically unbiased, we know that we are achieving the Cramer-Rao lower bound, at least asymptotically.

Hence, we are minimizing the MSE (asymptotically). This is objectively the best rationale for utilizing maximum likelihood estimators.

The Delta Method

Often, our objective is not to estimate θ, but instead some function of θ, call it $g(\theta)$. For example, often want to estimate the probability of some event A. This probability is not a parameter, but it is a function of θ.

From the invariance property we know that the MLE of $g(\theta)$ is $g(\widehat{\theta})$. But, we also need to calculate a SE or construct a confidence interval for this estimate.

One approach would be to "reparameterize" the model in terms of $g(\theta)$, and then find $I(g(\theta))$, and then use $1/nI(g(\theta))$ as the variance, etc. Fortunately, there is an easier way...

Recall the Delta Method:

> If Y_n is approximately $N(\mu, \sigma_n^2)$ and $\sigma_n^2 \to 0$ as n increases, then $g(Y_n)$ is approximately $N(g(\mu), (g'(\mu))^2\sigma_n^2)$, assuming that $g'(\mu)$ exists and is not zero.

Here we will use this result with the MLE for θ in place of Y_n.

Then, $\mu = \theta_0$ and $\sigma_n^2 = 1/nI(\theta_0)$.

We can conclude that $g(\widehat{\theta})$ is approximately

$$N\left(g(\theta_0),\ g'(\theta_0)^2\left(\frac{1}{nI(\theta_0)}\right)\right)$$

In practice, we use the approximate variance:

$$g'(\widehat{\theta})^2\left(\frac{1}{nI(\widehat{\theta})}\right)$$

Exercise: Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. Exponential(λ). We seek to estimate $\tau = P(X_i > a)$ for some constant $a > 0$. Construct the MLE and find its standard error. Also construct a confidence interval for τ.

By invariance prop, $\hat{\tau}_{MLE} = e^{-a\hat{\lambda}}$, i.e. $g(\lambda) = e^{-a\lambda}$

We know $I(\lambda) = 1/\lambda^2$ from Slide 14

Seek to use Δ-method

$g'(\lambda) = -ae^{-a\lambda}$

So, the Δ-method says that $\hat{\tau}_{MLE} = g(\hat{\lambda}_{MLE})$ is approx. normal with mean $g(\lambda_0) = e^{-a\lambda_0}$ and variance

$$(g'(\lambda_0))^2 \left(\frac{1}{nI(\lambda_0)}\right) = a^2 e^{-2a\lambda_0} \left(\frac{\lambda_0^2}{n}\right)$$

In practical terms, $\hat{\tau}_{MLE}$ is approx. normal with variance $\left(a^2 e^{-2a/\bar{x}}\right)\left(\frac{1}{n\bar{x}^2}\right)$

and the approx. SE is

$$\widehat{SE} = \left(a e^{-a/\bar{x}}\right)\left(\frac{1}{\sqrt{n}\,\bar{x}}\right)$$

A $100(1-\alpha)\%$ CI for τ is

$$\hat{\tau} \pm z_{\alpha/2}\,\widehat{SE}$$

Since $P\left(-z_{\alpha/2} \leq \frac{\hat{\tau} - \tau_0}{\widehat{SE}} \leq z_{\alpha/2}\right) = 1 - \alpha$