Lecture 7
Maximum Likelihood Estimation.

1 MLE

Let $f(\cdot|\theta)$ with $\theta \in \Theta$ be a parametric family. Let $X = (X_1, ..., X_n)$ be a random sample from distribution $f_1(\cdot|\theta_0)$ with $\theta_0 \in \Theta$. Then the joint pdf is $f(x|\theta) = \prod_{i=1}^n f_1(x_i|\theta)$ where $x = (x_1, ..., x_n)$. The log-likelihood is $\ell(\theta|x) = \sum_{i=1}^n \log f_1(x_i|\theta)$. The maximum likelihood estimator is, by definition,

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} \ell(\theta|x).$$

The FOC is

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial \ell_1(\hat{\theta}_{ML}|x_i)}{\partial \theta} = 0.$$

Note that the first information equality is $E[\partial \ell_1(\theta_0|X_i)] = 0$. Thus MLE is the method of moments estimator corresponding to the first information equality. So we can expect that the MLE is consistent. Indeed, the theorem below gives the consistency result for MLE:

Theorem 1 (MLE consistency)**.** *In the setting above, assume that (1) θ_0 is identifiable, i.e. for any $\theta \neq \theta_0$, there exists x such that $f(x|\theta) \neq f(x|\theta_0)$, (2) the support of $f(\cdot|\theta)$ does not depend on θ, and (3) θ_0 is an interior point of parameter space Θ. Then $\hat{\theta}_{ML} \to_p \theta_0$.*

The proof of MLE consistency will be given in 14.382 and 14.385. What the proof does, it shows that function $g(\theta) = E_{\theta_0}\ell_1(\theta|X_i)$ (here $X_i \sim f_1(x_i|\theta_0)$) is maximized at $\theta = \theta_0$ and random process $\frac{1}{n}\ell(\theta|X)$ converges to the function $g(\theta)$ in a uniform manner in probability. Then it argues that the maximizer of the process $\hat{\theta}_{ML}$ will converge to θ_0.

Once we know that the estimator is consistent, we can think about the asymptotic distribution of the estimator. The next theorem gives the asymptotic distribution of MLE:

Theorem 2 (MLE asymptotic normality)**.** *In the setting above, assume that conditions (1)-(3) in the MLE consistency theorem hold. In addition, assume that (4) $f_1(x_i|\theta)$ is thrice differentiable with respect to θ and we can interchange integration with respect to x and differentiation with respect to θ, and (5) $|\partial^3 \log f_1(x_i|\theta)/\partial\theta^3| \leq M(x)$ and $E[M(X_i)] < \infty$. Then*

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow N(0, I_1^{-1}(\theta_0))$$

Proof. This is a sketch of the proof as it misses and important step. By definition, $\frac{\partial \ell(\hat{\theta}_{ML},x)}{\partial \theta} = 0$. By the Taylor theorem with a remainder, there is some random variable $\tilde{\theta}$ with value between θ_0 and $\hat{\theta}_{ML}$ such that

$$\frac{\partial \ell(\hat{\theta}_{ML})|X}{\partial \theta} = \frac{\partial \ell(\theta_0|X)}{\partial \theta} + \frac{\partial^2 \ell(\tilde{\theta}|X)}{\partial \theta^2}(\hat{\theta}_{ML} - \theta_0).$$

So,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \frac{-\frac{1}{\sqrt{n}}\frac{\partial \ell(\theta_0|X)}{\partial \theta}}{\frac{1}{n}\frac{\partial^2 \ell(\tilde{\theta}|X)}{\partial \theta^2}}.$$

Since $\hat{\theta}_{ML} \to_p \theta_0$ and $\tilde{\theta}$ is between θ_0 and $\hat{\theta}_{ML}$, $\tilde{\theta} \to_p \theta_0$ as well. From $\tilde{\theta} \to_p \theta_0$, one can prove that

$$\frac{1}{n}\frac{\partial^2 \ell(\tilde{\theta}|X)}{\partial \theta^2} - \frac{1}{n}\frac{\partial^2 \ell(\theta_0|X)}{\partial \theta^2} = o_p(1).$$

We will not discuss this result here since it requires knowledge of the concept of asymptotic equicontinuity which we do not cover in this class. You will learn it in 14.385. Note, however, that this result does not follow from the Continuous mapping theorem since we have a sequence of random functions $\ell(\theta|X)$ instead of just one non-random function. Suppose we believe in this result. Then, by the Law of large numbers,

$$\frac{1}{n}\frac{\partial^2 \ell(\theta_0|X)}{\partial \theta^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_1(X_i|\theta_0)}{\partial \theta^2} \to_p E\left[\frac{\partial^2 \log f_1(X_i|\theta_0)}{\partial \theta^2}\right] = -I_1(\theta_0).$$

Next, by the first information equality, $E\left[\frac{\partial \log f_1(X_i|\theta_0)}{\partial \theta}\right] = 0$ while $Var\left[\frac{\partial \log f_1(X_i|\theta_0)}{\partial \theta}\right] = I_1(\theta_0)$. Thus, by the Central limit theorem,

$$\frac{1}{\sqrt{n}}\frac{\partial \ell(\theta_0|X)}{\partial \theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f_1(X_i|\theta_0)}{\partial \theta} \Rightarrow N(0, I(\theta_0)).$$

Finally, by the Slutsky theorem,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow N(0, I^{-1}(\theta_0)).$$

□

One interpretation of MLE asymptotics is that the MLE is asymptotically efficient (hit Rao-Cramer bound in very large samples).

Example Let $X_1, ..., X_n$ be a random sample from a distribution with pdf $f(x|\lambda) = \lambda \exp(-\lambda x)$. This distribution is called exponential. Its log-likelihood for one draw is $\ell_1(\lambda|x_i) = \log \lambda - \lambda x_i$. So $\frac{\partial \ell_1(\lambda|x_i)}{\partial \lambda} = 1/\lambda - x_i$ and $\frac{\partial^2 \ell_1(\lambda|x_i)}{\partial \lambda^2} = -1/\lambda^2$. So Fisher information is $I_1(\lambda) = 1/\lambda^2$. Let us find the MLE for λ. The log-likelihood for the whole is $\ell(\theta|x) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$. The FOC is $\frac{n}{\hat{\lambda}_{ML}} - \sum_{i=1}^{n} x_i = 0$. So $\hat{\lambda}_{ML} = \frac{1}{\bar{X}_n}$. Its asymptotic distribution is given by $\sqrt{n}(\hat{\lambda}_{ML} - \lambda) \Rightarrow N(0, \lambda^2)$.

2 Inference using MLE

We will have a longer discussion about how to estimate the asymptotic variance of the MLE $(I_1(\theta_0))^{-1}$ later when we will discuss asymptotic tests. Right now I want to mention several suggestions.

First of all, if $I_1(\theta)$ is a continuous function in θ (which is needed for asymptotic results), then given that $\hat{\theta}_{ML}$ is consistent for θ_0, the quantity $\left(I_1(\hat{\theta}_{ML})\right)^{-1}$ is consistent for $\left(I_1^{-1}(\theta_0)\right)^{-1}$.

Second, by definition of Fisher information, it equals to the expectation of either negative second derivative of the likelihood or of the squared score. Instead of taking expectation one may approximate it by taking averages. For example

$$\hat{I} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ell_1(\hat{\theta}|X_i)}{\partial\theta^2}$$

will be a consistent estimator of the Fisher information.

The third idea to be used in this context is parametric bootstrap. Assume $\hat{\theta}_{ML}$ is the MLE we obtained from our sample of size n. For $b = 1, ..., B$ do the following:

- Simulate sample $X_b^* = (X_{1b}^*, ..., X_{nb}^*)$ as i.i.d. draws from $f_1(x_i|\hat{\theta}_{ML})$ (that is, assuming that $\hat{\theta}_{ML}$ is the true parameter value).
- Find MLE using sample X_b^*, denote it θ_b^*.

Calculate the sample variance of $(\theta_1^*, ..., \theta_B^*)$, it gives the bootstrap approximation to $(nI_1(\theta_0))^{-1}$. You may also do bootstrap- bias correction using similar procedure.

3 When MLE asymptotic theory fails us...

Example A word of caution. For asymptotic normality of MLE, we should have common support. Let us see what might happen otherwise. Let $X_1, ..., X_n$ be a random sample from $U[0, \theta]$. Then $\hat{\theta}_{ML} = X_{(n)}$. So $\sqrt{n}(\hat{\theta}_{ML} - \theta)$ is always nonpositive. So it does not converge to mean zero normal distribution. In fact, $E[X_{(n)}] = (n/(n+1))\theta$ and $V(X_{(n)}) = \theta^2 n/((n+1)^2(n+2)) \approx \theta^2/n^2$. On the other hand, if the theorem worked, we would have $V(X_{(n)}) \approx 1/(nI(\theta))$. The MLE happens to be "super-consistent" here, means it converges to the true value at a faster speed than the regular parametric speed of $1/\sqrt{n}$.

Example Now, let us consider what might happen if the true parameter value θ_0 were on the boundary of Θ. Let $X_1, ..., X_n$ be a random sample from distribution $N(\mu, 1)$ with $\mu \geq 0$. As an exercise, check that $\hat{\mu}_{ML} = \overline{X}_n$ if $\overline{X}_n \geq 0$ and 0 otherwise. Suppose that $\mu_0 = 0$. Then $\sqrt{n}(\hat{\mu}_{ML} - \mu_0)$ is always nonnegative. So it does not converge to mean zero normal distribution.

Example Finally, note that it is implicitly assumed both in the consistency and asymptotic normality theorems that parameter space Θ is fixed, i.e. independent of n. In particular, the number of parameters should not depend on n. Indeed, let

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$$

for $i = 1, ..., n$, and $X_1, ..., X_n$ be mutually independent. One can show that if the sample size (n) increases to infinity, the MLE for σ^2 is inconsistent in this case, though a consistent estimator for σ^2 exists.

What is interesting, though we won't show it here is that a bootstrap does not help in this cases, that is, the bootstrap approximation to the distribution of $\hat{\theta}_{ML}$ is not close to the true finite-sample distribution of $\hat{\theta}_{ML}$.

4 Pseudo-MLE

Let us have a sample $X = (X_1, ..., X_n)$ i.i.d from some distribution. We do not know what distribution it is, let's assume it has pdf $g(x_i)$. But we wrongly assumed a specific parametric family, that is, we assumed $X_i \sim f_1(x_i|\theta)$. What would happen if we do MLE. Apparently, MLE will be estimating a "pseudo-true" parameter value θ_0 with minimizes in some sense the distance between $g(\cdot)$ and family $f(\cdot|\theta)$. In particular:

$$\theta_0 = \arg\max_\theta \int \log[f_1(x_i|\theta)] g(x_i) dx_i = \arg\max_\theta E \log f_1(X_i|\theta).$$

Parameter θ_0 may be of interest or may be not. Under some regularity condition $\hat{\theta}_{ML} \to^p \theta_0$, and in most parts the logic of the proof of theorem about normality will hold. However, the the information equality would fail. Define

$$\Sigma_1 = E\left[\left(\frac{\partial \log f_1(X_i|\theta_0)}{\partial \theta_0}\right)^2\right],$$

$$\Sigma_2 = -E\left[\frac{\partial^2 \log f_1(X_i|\theta_0)}{\partial \theta_0^2}\right],$$

where expectations in both cases are taken assuming that $X_i \sim g(\cdot)$. If g is not in the parametric family, then in general $\Sigma_1 \neq \Sigma_2$. But using the logic of the proof, we can prove that

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow N(0, \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1})$$

This asymptotic variance $\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}$ is often called White's due to White's (1980) paper and thus White's standard errors.

MIT OpenCourseWare
https://ocw.mit.edu

14.381 Statistical Method in Economics

Fall 2018

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms