Lectures 16 & 17

Maximum Likelihood Estimation

Oscar Volpe

11/29/2021 & 12/01/2021

Terminology

Consider an *i.i.d.* sample $\{X_i\}_{i=1}^n$, where $X_i \sim F$. Suppose that F depends on some (unknown) parameter θ. Our goal is to find a good estimate for θ.

- *Question:* Under the assumed distribution F, which choice of θ makes the observed data $X_1, \ldots, X_n$ most likely to have occurred in nature?
- Answering this question gives us the maximum likelihood estimator $\hat{\theta}_n$.

Definition (Likelihood Function)

The *likelihood*, denoted $\ell_n(\theta)$, is the joint density of $X_1, \ldots, X_n$ under θ evaluated at the realized values $x_1, \ldots, x_n$. Thus, $\ell_n(\theta) = \prod_{i=1}^n f_\theta(x_i)$.

Definition (Log Likelihood Function)

The *log likelihood* $\mathcal{L}_n(\theta)$ is the natural log of $\ell_n(\theta)$, so $\mathcal{L}_n(\theta) = \log(\ell_n(\theta))$.

Solving for the MLE

We choose $\hat{\theta}_n$ to maximize the likelihood of having observed the data.

Definition (Maximum Likelihood Estimator)

The maximum likelihood estimator (MLE) equals $\hat{\theta}_n \in \text{argmax}_{\theta \in \Theta} \ell_n(\theta)$.

Since $\log(\cdot)$ is monotonic, we can equivalently write $\hat{\theta}_n \in \text{argmax}_{\theta \in \Theta} \mathcal{L}_n(\theta)$.

- Select $\hat{\theta}_n$ to maximize $\ell_n(\theta)$ or $\mathcal{L}_n(\theta)$. Choose whatever is easier.
- *Note:* $\mathcal{L}_n(\theta) = \sum_{i=1}^n \log f_\theta(x_i)$. Working with sums can be simpler!

Importantly, a maximizer of $\ell_n(\theta)$ need not exist and may not be unique.

- If $\hat{\theta}_n$ does not exist, then pick a "near" maximizer.
- If $\hat{\theta}_n$ is not unique, then choose any maximizer.

Example: A Biased Coin

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}$ Bernoulli(θ), where $\theta \in (0,1)$. Then, $\ell_n(\theta)$ equals:

$$\ell_n(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{n\bar{x}_n}(1-\theta)^{n(1-\bar{x}_n)}$$

The log likelihood function $\mathcal{L}_n(\theta)$ is given by:

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i) = n \times [\log(\theta)\bar{x}_n + \log(1-\theta)(1-\bar{x}_n)]$$

To solve for $\hat{\theta}_n$, we can take first-order and second-order conditions.

- *FOC:* $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} = \frac{\bar{x}_n}{\theta} - \frac{1-\bar{x}_n}{1-\theta} = 0$
- *SOC:* $\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2} = -\frac{\bar{x}_n}{\theta^2} - \frac{1-\bar{x}_n}{(1-\theta)^2} < 0$

These two conditions imply that $\hat{\theta}_n = \bar{x}_n$ is the unique ML estimator for θ.

Conditioning on Data

You have an *i.i.d.* sample $\{Y_i, X_i\}_{i=1}^n$, where $Y_i|X_i \sim F_{Y_i|X_i}$. Let $F_{Y_i|X_i}$ depend on some (unknown) parameter θ. Your goal is to estimate θ.

Definition (Conditional Likelihood Function)

The *conditional likelihood* $\ell_n(\theta|x)$ is the joint density of $\{Y_i\}_{i=1}^n$ given $\{X_i\}_{i=1}^n$ under θ evaluated at $\{y_i, x_i\}_{i=1}^n$. Thus, $\ell_n(\theta|x) = \prod_{i=1}^n f_\theta(y_i|x_i)$.

Definition (Log Likelihood Function)

The *conditional log likelihood* is given by $\mathcal{L}_n(\theta|x) = \log(\ell_n(\theta|x))$.

As before, the maximum likelihood estimator is the maximizer of $\ell_n(\theta|x)$.

- With conditional MLE, $\hat{\theta}_n$ depends on $\{y_i\}_{i=1}^n$ as well as $\{x_i\}_{i=1}^n$.
- Again, a maximizer of $\log(\ell_n(\theta|x))$ may not always exist or be unique.
- An unconditional MLE is just a special case of a conditional MLE.

Example: Linear Regression

You collect *i.i.d.* data $\{Y_i, X_i\}_{i=1}^n$, where you assume $Y_i = X_i'\beta + U_i$. Suppose $U_i \sim N(0, \sigma^2)$. In this case, distribution of Y_i given X_i equals:

$$f_{\beta,\sigma^2}(Y_i|X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(Y_i - X_i'\beta)^2 \Big]$$

The conditional log-likelihood function for $\theta = (\beta, \sigma^2)'$ is given by:

$$\mathcal{L}_n(\theta|\{X_i\}_{i=1}^n) = -\frac{n}{2}\big[\log(2\pi) + \log(\sigma^2)\big] - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - X_i'\beta)^2$$

Taking first-order and second-order conditions, the MLE for θ equals:

$$\hat{\theta}_n = \begin{bmatrix} \hat{\beta}_n \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \big(\sum_{i=1}^n X_i X_i'\big)^{-1} \sum_{i=1}^n X_i Y_i \\ \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\hat{\beta}_n)^2 \end{bmatrix}$$

The Score

Definition (Score)

The *score* is the first derivative of the log likelihood: $s(\theta|x) = \frac{\partial \log f_\theta(y|x)}{\partial \theta}$.

- *Note:* if θ has multiple dimensions, then $s(\theta|x)$ is a column vector.
- We will use the score later on when testing restrictions about θ.

One property of the score is that $E[s(\theta|x)] = 0$. To see why, write:

$$E[s(\theta|x)] = \int \frac{\partial \log f_\theta(y|x)}{\partial \theta} f_\theta(y|x) dy = \int \frac{\frac{\partial f_\theta(y|x)}{\partial \theta}}{f_\theta(y|x)} f_\theta(y|x) dy = \int \frac{\partial f_\theta(y|x)}{\partial \theta} dy$$

Since $\int f_\theta(y|x) dy = 1$, under weak conditions: $\int \frac{\partial f_\theta(y|x)}{\partial \theta} dy = \frac{\partial}{\partial \theta} 1 = 0$.

The Fisher Information

The *Fisher information matrix* is equal to the variance of the score.

- It measures how much information (Y_i, X_i) carries about θ.
- If this matrix is "large", then the sample draws of (X_i, Y_i) will be more informative, and the ML estimator of θ becomes more precise.

Definition (Fisher Information Matrix)

The *Fisher information matrix* is defined as $\mathcal{I}(\theta) = E[s(\theta|x)s(\theta|x)']$

One useful property is that $\mathcal{I}(\theta) = E[s(\theta|x)s(\theta|x)'] = -E\left[\frac{\partial^2 \log f_\theta(y|x)}{\partial\theta\partial\theta'}\right]$.
To see why, recall that $0 = \int \frac{\partial \log f_\theta(y|x)}{\partial\theta} f_\theta(y|x)dy$. Differentiate w.r.t. θ'.

$$0 = \underbrace{\int \frac{\partial^2 \log f_\theta(y|x)}{\partial\theta\partial\theta'} f_\theta(y|x)dy}_{E\left[\frac{\partial^2 \log f_\theta(y|x)}{\partial\theta\partial\theta'}\right]} + \underbrace{\int \frac{\partial \log f_\theta(y|x)}{\partial\theta} \frac{\frac{\partial f_\theta(y|x)}{\partial\theta'}}{f_\theta(y|x)} f_\theta(y|x)dy}_{\mathcal{I}(\theta)}$$

Information Inequality

Given a sample $\{Y_i, X_i\}_{i=1}^n$, suppose $\hat{\theta}_n$ is an unbiased estimator for θ. Then, the variance of this estimator is bounded from below by $\mathcal{I}(\theta)^{-1}$.

Theorem (Cramér-Rao Lower Bound)

Let $\hat{\theta}_n$ be an estimator of θ satisfying $E(\hat{\theta}_n) = \theta$. Then $Var(\hat{\theta}_n) \geq \mathcal{I}(\theta)^{-1}$.

This inequality shows how $\mathcal{I}(\theta)$ relates to an estimator's precision.

- A "smaller" $\mathcal{I}(\theta)$ is associated with greater variability of $\hat{\theta}_n$.
- An unbiased estimator with variance $\mathcal{I}(\theta)^{-1}$ will be *efficient*.
- The proof of this result relies on the Cauchy-Schwartz Inequality.

A Useful Result

Theorem (Delta Method)

Let $\{X_n\}_{n=1}^n$ and X be random vectors, and c a constant, in $\mathbb{R}^k$. Let τ_n be a sequence of constants such that $\tau_n \to \infty$ and $\tau_n(X_n - c) \xrightarrow{d} X$. Then, for any continuous function $g : \mathbb{R}^k \to \mathbb{R}^m$, $\tau_n(g(X_n) - g(c)) \xrightarrow{d} Dg(c)X$.

- Here, $Dg(c)$ is an $m \times k$ matrix of partials of $g(\cdot)$ evaluated at c.
  - *Note:* if $g : \mathbb{R} \to \mathbb{R}$, then $Dg(c) = g'(c)$.
- To prove it, take a first-order Taylor expansion of $g(X_n)$ about c.

Important Special Case

The application of this theorem most relevant to us states that:

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, \Sigma) \implies \sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \to N(0, Dg(\theta)\Sigma Dg(\theta)')$$

We will use this result to derive the limiting distribution of the MLE.

Asymptotic Distribution of the MLE *(Part 1)*

You have an *i.i.d.* sample $\{Y_i, X_i\}_{i=1}^n$, where $Y_i|X_i \sim F_{Y_i|X_i}$. Let $F_{Y_i|X_i}$ depend on some (unknown) parameter θ that you want to estimate.

Theorem (Asymptotic Normality)

Suppose that $\hat{\theta}_n$ is the MLE of θ. Then $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \mathcal{I}(\theta)^{-1})$.

Proof of the Theorem

Let $s_n(\theta|x) = \frac{1}{n}\sum_{i=1}^n \frac{\partial \log f_\theta(y_i|x_i)}{\partial\theta}$ and $H_n(\theta|x) = \frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \log f_\theta(y_i|x_i)}{\partial\theta\partial\theta'}$. Applying the Central Limit Theorem, we know that:

$$\sqrt{n}s_n(\theta|x) \xrightarrow{d} N(0, E[s(\theta|x)s(\theta|x)']) \stackrel{d}{=} N(0, \mathcal{I}(\theta))$$

Note that $\hat{\theta}_n$ solves $s_n(\theta|x) = 0$. A Taylor approximation around θ gives us:

$$0 = s_n(\hat{\theta}_n) \approx s_n(\theta) + H_n(\theta|x)(\hat{\theta}_n - \theta)$$

Asymptotic Distribution of the MLE *(Part 2)*

As $s_n(\theta) \approx -H_n(\theta|x)[\hat{\theta}_n - \theta]$, we know $\sqrt{n}(\hat{\theta}_n - \theta) \approx -\sqrt{n} H_n(\theta|x)^{-1} s_n(\theta)$.

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta) &\approx -\sqrt{n} H_n(\theta|x)^{-1} s_n(\theta) \\ &\xrightarrow{d} N(0, \mathcal{I}(\theta)^{-1}\mathcal{I}(\theta)\mathcal{I}(\theta)^{-1}) \stackrel{d}{=} N(0, \mathcal{I}(\theta)^{-1}) \end{aligned}$$

The line above follows from the Delta Method. So, the theorem is true.

□

As $n \to \infty$, $\hat{\theta}_n$ is distributed normally with mean 0 and variance $\mathcal{I}(\theta)^{-1}$.

- By the Cramér-Rao Lower Bound, the MLE is *asymptotically efficient*. It has the smallest asymptotic variance among all unbiased estimators.
- This observation helps to justify our use of maximum likelihood.

Example Revisited: A Biased Coin

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, where $\theta \in (0, 1)$. We showed $\hat{\theta}_n = \bar{x}_n$ is the maximum likelihood estimator of θ. We compute $s(\theta)$ and $\mathcal{I}(\theta)$ as:

$$s(\theta) = \frac{\partial \log f_\theta(x)}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2 \log f_\theta(x)}{\partial \theta^2}\right] = E\left[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}\right] = \frac{1}{\theta(1-\theta)}$$

By our previous result, the limiting distribution of $\hat{\theta}_n = \bar{X}_n$ is equal to:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \theta(1-\theta))$$

Recall that $\text{Var}(X_i) = \theta(1-\theta)$. Since $\hat{\theta}_n$ is the sample mean of X_i, this result could have also been derived by applying the Central Limit Theorem.

Example Revisited: Linear Regression

Let $\{Y_i, X_i\}_{i=1}^n$ be an *i.i.d.*, where $Y_i = X_i'\beta + U_i$ and $U_i \sim N(0, \sigma^2)$. We showed that the maximum likelihood estimator of θ equals:

$$\hat{\theta}_n = \begin{bmatrix} \hat{\beta}_n \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \sum_{i=1}^n X_i Y_i \\ \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\hat{\beta}_n)^2 \end{bmatrix}$$

The score and Fisher information matrix of θ are given by:

$$s(\theta|x) = \begin{bmatrix} \frac{1}{\sigma^2}(xy - xx'\beta) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y - x'\beta)^2 \end{bmatrix}$$

$$\mathcal{I}(\theta) = E\begin{bmatrix} \frac{1}{\sigma^2}xx' & \frac{1}{\sigma^4}(xy - xx'\beta) \\ \frac{1}{\sigma^4}(xy - xx'\beta) & -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}(y - x'\beta)^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}E(xx') & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

So, as $n \to \infty$, we find that: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \begin{bmatrix} \sigma^2 E(xx')^{-1} & 0 \\ 0 & 2\sigma^4 \end{bmatrix}\right)$.

- *Note:* $\sigma^2 E(xx')^{-1}$ is the asymptotic variance of the OLS estimator $\hat{\beta}_n$ with homoskedastic errors, and asymptotic variance of $\hat{\sigma}^2$ is $2\sigma^4$.

1. The Likelihood Function
   - Unconditional Likelihood Functions
   - Conditional Likelihood Functions

2. Properties of MLE
   - Score and Information Matrix
   - Cramér-Rao Lower Bound
   - Asymptotic Distribution

3. **Inference**

Hypothesis Testing Setup

Suppose $\theta \in \mathbb{R}^k$, and let $g : \mathbb{R}^k \to \mathbb{R}^p$ be continuously differentiable. We wish to test a restriction of the form $H_0 : g(\theta) = 0$ vs. $H_1 : g(\theta) \neq 0$.

- Let $\tilde{\theta}_n$ be the (constrained) maximizer of $\ell_n(\theta)$ among θ satisfying H_0.
- Under H_0, we expect that $\tilde{\theta}_n$ should be "close" to $\hat{\theta}_n$.

To assess H_0 vs. H_1, we introduce three types of tests.

(1) *Wald Test:* compare $g(\hat{\theta}_n)$ with zero
   - If H_0 holds, then $g(\theta) = 0$. So, $g(\hat{\theta}_n)$ should be "close" to zero.

(2) *Lagrange Multiplier Test:* compare $\frac{\partial \mathcal{L}_n(\tilde{\theta}_n)}{\partial \theta}$ with zero
   - If H_0 holds, then $\tilde{\theta}_n$ is the maximizer of $\mathcal{L}_n(\theta)$. So, $\frac{\partial \mathcal{L}_n(\tilde{\theta}_n)}{\partial \theta} \approx 0$.

(3) *Likelihood Ratio Test:* compare $\mathcal{L}_n(\hat{\theta}_n)$ with $\mathcal{L}_n(\tilde{\theta}_n)$
   - If H_0 holds, then the likelihood functions for the constrained and unconstrained maximizers should be similar, i.e. $\mathcal{L}_n(\hat{\theta}_n) \approx \mathcal{L}_n(\tilde{\theta}_n)$.

Wald Test

We know that $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \mathcal{I}(\theta)^{-1})$. By the Delta Method:

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} N(0, Dg(\theta)\mathcal{I}(\theta)^{-1}Dg(\theta)')$$

Under H_0, we should expect $ng(\hat{\theta}_n)[Dg(\theta)\mathcal{I}(\theta)^{-1}Dg(\theta)']^{-1}g(\hat{\theta}_n) \xrightarrow{d} \chi^2_p$. Therefore, we choose $T_n = ng(\hat{\theta}_n)\hat{\Sigma}^{-1}g(\hat{\theta}_n)$ as a test statistic, where:

$$\hat{\Sigma} = Dg(\hat{\theta}_n)\Big(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f_{\hat{\theta}_n}(y_i|x_i)}{\partial\theta\partial\theta'}\Big)^{-1} Dg(\hat{\theta}_n)'$$

- WLLN and CMT imply that $\hat{\Sigma}$ is consistent for $Dg(\theta)\mathcal{I}(\theta)^{-1}Dg(\theta)'$.
- Our test is $\mathbb{I}\{T_n > c_n\}$, where c_n is the $(1-\alpha)$th quantile of a χ^2_p.

Lagrange Multiplier Test

Step 1. Choose $\tilde{\theta}_n$ to maximize $\mathcal{L}_n(\theta)$ such that $f(\theta) = 0$.

- Write down the FOC for the Lagrangian: $s_n(\tilde{\theta}_n|x) - \frac{\partial g(\tilde{\theta}_n)}{\partial \theta}\lambda_n = 0$.
- Pre-multiply by $\frac{\partial g(\tilde{\theta}_n)}{\partial \theta}\mathcal{I}(\theta)^{-1}$ and solve for λ_n.

Step 2. By the Central Limit Theorem and the Delta Method:

$$\begin{aligned}\sqrt{n}\lambda_n &= \sqrt{n}\Big[\frac{\partial g(\tilde{\theta}_n)}{\partial \theta}\mathcal{I}(\theta)^{-1}\frac{\partial g(\tilde{\theta}_n)}{\partial \theta}\Big]^{-1}\frac{\partial g(\tilde{\theta}_n)}{\partial \theta}\mathcal{I}(\theta)^{-1}s_n(\tilde{\theta}_n|x) \\ &\xrightarrow{d} N(0, \Big[\frac{\partial g(\tilde{\theta}_n)}{\partial \theta}\mathcal{I}(\theta)^{-1}\frac{\partial g(\tilde{\theta}_n)}{\partial \theta}\Big]^{-1})\end{aligned}$$

Step 3. Derive the Lagrange Multiplier Test statistic to be:

$$T_n = ns_n(\tilde{\theta}_n|x)'\Big(-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_{\tilde{\theta}_n}(y_i|x_i)}{\partial\theta\partial\theta'}\Big)^{-1}s_n(\tilde{\theta}_n|x)$$

Our test is $\mathbb{I}\{T_n > c_n\}$, where c_n is the $(1-\alpha)$th quantile of a χ^2_p.

Likelihood Ratio Test

If H_0 is true, then $\ell_n(\tilde{\theta}_n) = \ell_n(\hat{\theta}_n)$. So, under H_0, we should expect:

$$\frac{\ell_n(\hat{\theta}_n)}{\ell_n(\tilde{\theta}_n)} \approx 1 \quad \Longleftrightarrow \quad \mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\tilde{\theta}_n) \approx 0$$

It can be shown that $2[\mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\tilde{\theta}_n)] \xrightarrow{d} \chi^2_p$. So, choose T_n as:

$$T_n = 2[\mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\tilde{\theta}_n)]$$

Our test is $\mathbb{I}\{T_n > c_n\}$, where c_n is the $(1-\alpha)$th quantile of a χ^2_p.

- By the Neyman-Pearson Lemma, the likelihood ratio test is *uniformly most powerful* for simple hypothesis tests $H_0 : \theta = c$ vs. $H_1 : \theta \neq c$.
- For this reason, likelihood ratio tests are often quite convenient.