Lecture 2: Maximum Likelihood and Friends

Chris Conlon

February 6, 2021

NYU Stern

Review: What is a Likelihood?

Suppose we write down the joint distribution of our data (y_i, x_i) for $i = 1, \ldots, n$.

$$Pr(y_1, \ldots, y_n, x_1, \ldots, x_n; \theta)$$

If (y_i, x_i) are I.I.D then we can write this as:

$$Pr(y_1, \ldots, y_n, x_1, \ldots, x_n; \theta) = \prod_{i=1}^{N} Pr(y_i, x_i; \theta) \propto \prod_{i=1}^{N} Pr(y_i | x_i; \theta) = L(\mathrm{y}|\mathrm{x}; \theta)$$

We call this $L(\mathrm{y}|\mathrm{x}; \theta)$ the likelihood of the observed data.

MLE: Example

Consider a linear regression with $\varepsilon_i | X_i \sim N(0, \sigma^2)$

$$Y_{it} = X_i' \beta_i + \varepsilon_i$$

We've discussed the least squares estimator:

$$\widehat{\beta}_{ols} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i' \beta)^2$$

$$\widehat{\beta}_{ols} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$$

MLE: Example

If we know the distribution of ε_i we can construct a maximum likelihood estimator

$$(\widehat{\beta}_{MLE}, \widehat{\sigma}^2_{MLE}) = \arg\min_{\beta,\sigma^2} L(\beta, \sigma^2)$$

Where

$$\begin{aligned}
L(\beta, \sigma^2) &= \prod_{i=1}^{N} p(y_i | X_i, \beta, \sigma^2) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - X_i'\beta)^2\right] \\
\ell(\beta, \sigma^2) &= \sum_{i=1}^{N} -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - X_i'\beta)^2
\end{aligned}$$

MLE: FOC's

Take the FOC's

$$\ell(\beta, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - X_i'\beta)^2$$

Where

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - X_i'\beta) = 0 \rightarrow \widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$$

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \sigma^2} = -N\frac{1}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i - X_i'\beta)^2 = 0$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i'\beta)^2$$

Note: the unbiased estimator uses $\frac{1}{N-K-1}$.

MLE: General Case

1. Start with the joint density of the data $Z_1, \ldots, Z_N$ with density $f_Z(z;\theta)$
2. Construct the likelihood function of the sample $\boldsymbol{z} = (z_1, \ldots, z_n)$

$$L(\boldsymbol{z};\theta) = \prod_{i=1}^{N} f_Z(z_i;\theta)$$

3. Construct the log likelihood (this has the same arg max)

$$\ell(\boldsymbol{z};\theta) = \sum_{i=1}^{N} \ln f_Z(z_i;\theta)$$

4. Take the FOC's to find $\widehat{\theta}_{MLE}$

$$\theta : \frac{\partial \ell(\theta)}{\partial \theta} = 0$$

MLE in Detail

Basic Setup: we know $F(z;\theta_0)$ but not θ_0. We know $\theta_0 \in \Theta \subset \mathbb{R}^K$.

- Begin with a sample of z_i from $i = 1, \ldots, N$ which are I.I.D. with CDF $F(z;\theta_0)$.
- The MLE chooses

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \ln f_Z(z_i;\theta)$$

MLE: Technical Details

1. Consistency. When is it true that for $\epsilon > 0$?

$$\lim_{N\to\infty} \Pr\left(\left\|\hat{\theta}_{mle} - \theta_0\right\| > \varepsilon\right) = 0$$

2. Asymptotic Normality. What else do we need to show?

$$\sqrt{N}\left(\hat{\theta}_{mle} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, -\left[E\frac{\partial^2}{\partial\theta\partial\theta'}\left(Z_i, \theta_0\right)\right]^{-1}\right)$$

3. Optimization. How to we obtain $\widehat{\theta}_{MLE}$ anyway?

MLE: Example # 1

- $Z_i \sim N(\theta_0, 1)$ and $\Theta = (-\infty, \infty)$. In this case:

$$\ell(\theta) = -N \cdot \ln(2\pi) - \sum_{i=1}^{N} (z_i - \theta)^2 / 2$$

- MLE is $\widehat{\theta}_{MLE} = \overline{z}$ which is consistent for $\theta_0 = E[Z_i]$
- Asymptotic distribution is $\sqrt{N}(\overline{z} - \theta_0) \sim N(0, 1)$.
- Calculating mean is easy!

MLE: Example # 2

- $Z_i = (Y_i, X_i)$— X_i has finite mean and variance (but arbitrary distribution)
- $(Y_i | X_i = x) \sim N(x'\beta_0, \sigma_0^2)$

$$\widehat{\beta}_{MLE} = (X'X)^{-1}X'Y$$
$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum(y_i - x_i\widehat{\beta}_{MLE})^2$$

- We already have shown consistency and AN for linear regression with normally distributed errors...

MLE: Example # 3

- $Z_i = (Y_i, X_i)$— X_i has finite mean and variance (but arbitrary distribution)
- $Pr(Y_i = 1 | X_i = x) = \frac{e^{x'\theta_0}}{1+e^{x'\theta_0}}$
- Solution is the logit model.
- No simple MLE solution, establishing properties is not obvious...

Jensen's Inequality

Let $g(z)$ be a convex function. Then $\mathbb{E}[g(Z)] \geq g(\mathbb{E}[Z])$, with equality only in the case of a linear function.

More Technical Details

Define Y as the ratio of the density at θ to the density at the true value θ_0 both evaluated at Z

$$Y = \frac{f_Z(Z;\theta)}{f_Z(Z;\theta_0)}$$

- Let $g(a) = -\ln(a)$ so that $g'(a) = \frac{-1}{a}$ and $g''(a) = \frac{1}{a^2}$.
- Then by Jensen's Inequality $\mathbb{E}[-\ln Y] \geq -\ln \mathbb{E}[Y]$.
- This gives us

$$\mathbb{E}_z\left[-\ln\left(\frac{f_Z(Z;\theta)}{f_Z\left(Z;\theta_0\right)}\right)\right] \geq -\ln\left(\mathbb{E}_z\left[\frac{f_Z(Z;\theta)}{f_Z\left(Z;\theta_0\right)}\right]\right)$$

- The RHS is

$$\mathbb{E}_z\left[\frac{f_Z(Z;\theta)}{f_Z\left(Z;\theta_0\right)}\right] = \int \frac{f_Z(z;\theta)}{f_Z\left(z;\theta_0\right)} \cdot f_Z\left(z;\theta_0\right) dz = \int f_Z(z;\theta) dz = 1$$

More Technical Details

Because $\log(1) = 0$ this implies:

$$\mathbb{E}_z\left[-\ln\left(\frac{f_Z(Z;\theta)}{f_Z\left(Z;\theta_0\right)}\right)\right] \geq 0$$

Therefore

$$-\mathbb{E}\left[\ln f_Z(Z;\theta)\right] + \mathbb{E}\left[\ln f_Z\left(Z;\theta_0\right)\right] \geq 0$$
$$\mathbb{E}\left[\ln f_Z\left(Z;\theta_0\right)\right] \geq \mathbb{E}\left[\ln f_Z(Z;\theta)\right]$$

- We maximize the expected value of the log likelihood at the true value of θ!
- Helpful to work with $\mathbb{E}[\log f(z;\theta)]$ sometimes.

Information Matrix Equality

We can relate the Fisher Information to the Hessian of the log-likelihood

$$\mathcal{I}(\theta_0) = -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial\theta\partial\theta}(z;\theta_0)\right] = \mathbb{E}\left[\frac{\partial \ln f}{\partial\theta}(z;\theta_0) \times \frac{\partial \ln f}{\partial\theta}(z;\theta_0)'\right]$$

- This is sometimes known as the outer product of scores.
- This matrix is negative definite
- Recall that $\mathbb{E}\left[\frac{\partial \ln f}{\partial\theta}(z;\theta_0)\right] \approx 0$ at the maximum

Proof

$$1 = \int_z f_Z(z;\theta)dz \Rightarrow 0 = \frac{\partial}{\partial\theta}\int_z f_Z(z;\theta)dz$$

With some regularity conditions

$$0 = \int_z \frac{\partial f_Z}{\partial\theta}(z;\theta)dz = \underbrace{\int_z \frac{\partial \ln f_Z}{\partial\theta}(z;\theta)\cdot f_Z(z;\theta)dz}_{\mathbb{E}\left[\frac{\partial \ln f_Z}{\partial\theta}(z;\theta_0)\right]}$$

- This gives us the FOC we needed.
- Can get information identity with another set of derivatives.

The Cramer-Rao Bound

We can relate the Fisher Information to the Hessian of the log-likelihood

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial\theta\partial\theta'}(Z;\theta)\right]$$

It turns out this provides a bound on the variance

$$\mathrm{Var}(\hat{\theta}(Z)) \geq \mathcal{I}\left(\theta_0\right)^{-1}$$

Because we can't do better than Fisher Information we know that MLE is most efficient estimator!

MLE: Discussion

Tradeoffs

- How does this compare to GM Theorem?
- If MLE is most efficient estimate, why ever use something else?

Exponential Example

$$f_{Y|X}(y|x, \beta_0) = e^{x'\beta_0} \exp\left(-ye^{x'\beta_0}\right)$$

With log likelihood

$$\ell(\beta) = \sum_{i=1}^{N} \ln f_{Y|X}\left(y_i|x_i, \beta\right) = \sum_{i=1}^{N} X_i'\beta - y_i \cdot \exp\left(x_i'\beta\right)$$

And Score, Hessian, and Information Matrix:

$$\begin{aligned} \mathcal{S}_i(y_i, x_i, \beta) &= x_i'\left(1 - y_i \exp\left(x_i'\beta\right)\right) \\ \mathcal{H}_i(y_i, x_i, \beta) &= -y_i x_i x_i' \exp\left(x_i'\beta\right) \\ \mathcal{I}\left(\beta_0\right) &= \mathbb{E}\left[YXX' \exp\left(X'\beta_0\right)\right] = \mathbb{E}\left[XX'\right] \end{aligned}$$

Thanks!