

Lecture 5: Maximum Likelihood Estimation

Chris Conlon

Sunday 5th October, 2025

NYU Stern

Consider a linear regression with $\varepsilon_i \mid X_i \sim \mathcal{N}(0, \sigma^2)$

$$Y_{it} = X_i \beta + \varepsilon_i$$

We've discussed the **least squares estimator**:

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

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

Review: What is a Likelihood?

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

$$\Pr(y_1, \dots, y_n, x_1, \dots, x_n \mid \theta)$$

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

$$\Pr(y_1, \dots, y_n, x_1, \dots, x_n \mid \theta) = \prod_{i=1}^N \Pr(y_i, x_i \mid \theta) \propto \prod_{i=1}^N \Pr(y_i \mid x_i, \theta) = \mathbb{L}(\mathbf{y} \mid \mathbf{x}, \theta)$$

We call this $\mathbb{L}(\mathbf{y} \mid \mathbf{x}, \theta)$ the **likelihood** of the observed data.

MLE: Example

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

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

Where

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

Take the FOC's of: $\ell(\beta, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (Y_i - X_i\beta)^2$:

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

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

$$\sigma_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - X_i\beta)^2$$

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

MLE: General Case

1. Start with the **joint density of the data** Z_1, \dots, Z_N with density $f_Z(z, \theta)$
2. Construct the likelihood function of the sample z_1, \dots, z_n

$$\mathbb{L}(\mathbf{z} \mid \theta) = \prod_{i=1}^N f_Z(z_i, \theta)$$

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

$$\ell(\mathbf{z} \mid \theta) = \sum_{i=1}^N \ln f_Z(z_i, \theta)$$

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

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0 \text{ or } \nabla_{\theta} \ell(\theta) = 0 \text{ or } \sum_{i=1}^N \frac{\partial \ln f_Z(z_i, \theta)}{\partial \theta} = 0.$$

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

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

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

- ▶ In practice the computer always **minimizes negative log-likelihood** so that $\hat{\theta}_{MLE} = \arg \min_{\theta} -\ell(\theta)$.

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

$$\lim_{N \rightarrow \infty} \Pr \left(\left\| \hat{\theta}_{MLE} - \theta_0 \right\| > \epsilon \right) = 0$$

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

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

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

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

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

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

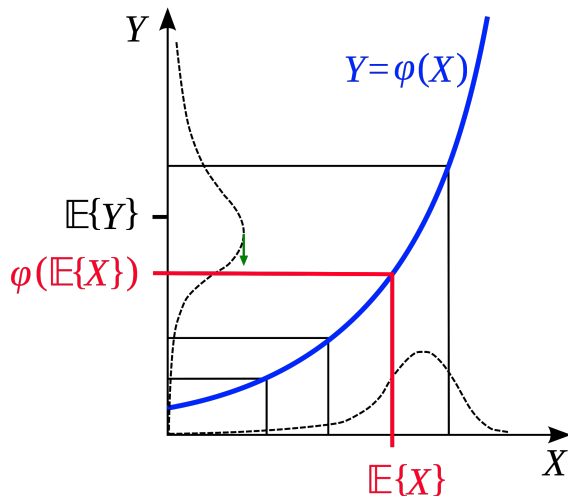
- ▶ We already have shown consistency and AN for linear regression with normally distributed errors...

- ▶ $Z_i = (Y_i, X_i)$ where 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 $\phi(z)$ be a convex function.

Then $\mathbb{E}[\phi(Z)] \geq \phi(\mathbb{E}[Z])$, 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

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

- ▶ The RHS is

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

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

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

Therefore

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

$$\mathbb{E} [\ln f_Z(Z; \theta_0)] \geq \mathbb{E} [\ln f_Z(Z; \theta)]$$

- ▶ 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.

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

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

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.

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

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

It turns out this provides a bound on the variance

$$\text{Var}(\hat{\theta}(Z)) \geq \mathcal{J}(\theta_0)^{-1}$$

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

Tradeoffs

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

Exponential Regression Example (Careful: two parameterizations here!)

For $y_i > 0$, $f_{y|x}(y_i | x_i, \beta) = (x_i \beta) e^{-(x_i \beta) y_i}$ so that $\mathbb{E}[y_i] = \frac{1}{x_i \beta}$

With log likelihood

$$\ell(\beta) = \sum_{i=1}^N \ln f_{y|x}(y_i | x_i, \beta) = \sum_{i=1}^N \log(x'_i \beta) - (x'_i \beta) y_i$$

And Score, Hessian, and Information Matrix:

$$\begin{aligned}\frac{\partial \ell_i}{\partial \theta} &= \mathcal{S}_i(y_i, x_i, \beta) = x'_i \left(\frac{1}{x_i \beta} - y_i \right) \\ \frac{\partial^2 \ell_i}{\partial \theta^2} &= \mathcal{H}_i(y_i, x_i, \beta) = -\frac{x'_i x_i}{(x'_i \beta)^2} \\ \mathcal{J}(\beta_0) &= \sum_i \frac{x'_i x_i}{(x_i \beta)^2}\end{aligned}$$

Thanks!
