ECMA31000: Introduction to Empirical Analysis
Maximum Likelihood Estimation

Joe Hardwick

University of Chicago

Autumn 2021

Outline

- Last time:
  - Sample Analogue principle
  - Finite/large sample properties: Bias, MSE, Consistency, Asymptotic Distribution.
  - Method of Moments estimation.
- Today:
  - Maximum likelihood estimators.

MLE

- Let $\{X_i\}_{i=1}^{n}$ be an iid sample of random variables with marginal distribution F_{θ_0}, for some $\theta_0 \in \Theta$.
- We assume $\Theta \subset \mathbb{R}^d$, so our statistical model comprises a parametric class of distributions.
- Further assume that F_θ has density f_θ for each $\theta \in \Theta$.
- The <u>Likelihood Function</u>, $\ell_n(\theta)$ is the joint density of the sample $\{X_i\}_{i=1}^{n}$ evaluated at $(X_1, \ldots, X_n)$, and is regarded as a function of θ:

$$\ell_n(\theta) = \Pi_{1 \leq i \leq n} f_\theta(X_i) .$$

MLE

- A maximum likelihood estimator (MLE) $\hat{\theta}_n$ of θ_0 satisfies:

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \ell_n(\theta).$$

- This is equivalent to maximizing the log-likelihood function

$$L_n(\theta) = \frac{1}{n} \ln(\ell_n(\theta)) = \frac{1}{n} \sum_{i=1}^{n} \ln(f_\theta(X_i)),$$

  which is often easier to work with.
- There may be multiple MLEs or none.
- f_θ may also be a probability mass function.

MLE

- Under appropriate conditions:

$$L_n(\theta) \xrightarrow{p} L(\theta) := \mathrm{E}\left(\ln\left(f_\theta(X)\right)\right),$$

  where the expectation is taken with respect to f_{θ_0}.
- Reasonable to expect that

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \ell_n(\theta) \xrightarrow{p} \theta^* \in \arg\max_{\theta \in \Theta} \mathrm{E}\left(\ln\left(f_\theta(X)\right)\right).$$

MLE

- This is desirable because provided $\mathrm{P}\left(f_\theta\left(X\right) \neq f_{\theta_0}\left(X\right)\right) > 0$ when $\theta \neq \theta_0$:

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathrm{E}\left(\ln\left(f_\theta\left(X\right)\right)\right).$$

- In other words, the true θ_0 uniquely maximizes $\mathrm{E}\left(\ln\left(f_\theta\left(X\right)\right)\right)$.
- To see this, let

$$M\left(\theta\right) = L\left(\theta\right) - L\left(\theta_0\right) = \mathrm{E}\left(\ln\left(\frac{f_\theta\left(X\right)}{f_{\theta_0}\left(X\right)}\right)\right).$$

WTS: $M(\theta) \leq 0$ for any $\theta \in \Theta$.

$M(\theta) = 0$ iff $\theta = \theta_0$.

MLE

Expectation is taken w.r.t. the distribution generating the data ie f_{θ_0}.

$\int f_\theta(x)\,dx = 1$
$\forall \theta$.

- By Jensen's inequality:

$$M(\theta) \leq \ln \mathrm{E}\left(\frac{f_\theta(X)}{f_{\theta_0}(X)}\right) = \ln\left[\int \frac{f_\theta(x)}{\cancel{f_{\theta_0}(x)}} \cdot \cancel{f_{\theta_0}(x)}\,\mathrm{d}x\right] = 0,$$

with equality iff for some c

$$\mathrm{P}\left(\frac{f_\theta(X)}{f_{\theta_0}(X)} = c\right) = 1.$$

$f_\theta \neq f_{\theta_0}$

$c > 1$: $M(\theta) = E(\ln(c))$ $(c>1)$ > 0.

- For $\theta \neq \theta_0$: $c = 1$ is ruled out by assumption, $c > 1$ contradicts $M(\theta) \leq 0$, and $c < 1$ provides $M(\theta) < 0$.
- So, $M(\theta)$ is uniquely maximized at θ_0.

$c < 1$: $M(\theta) = E(\ln(c)) < 0$.

Example: Normal Distribution

- Suppose $X_i \sim \mathcal{N}(\mu, \sigma)$. $\theta = (\mu, \sigma^2)$ is unknown.
- The Likelihood function is given by

$$\ell_n(\theta) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right).$$

- The log-likelihood is given by

$$L_n(\theta) = \left(-\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right) / n$$

$$\ln \prod_{i=1}^{n} = \sum \ln\left((2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right)\right).$$

$$\sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(X_i - \mu)^2$$

Example: Normal Distribution

- For any value of $\sigma^2 > 0$, the value of μ which maximizes the log-likelihood is the value which minimizes

$$\sum_{i=1}^{n} (X_i - \mu)^2 .$$

  This is a strictly convex function in μ, so the solution is $\hat{\mu}_n = \bar{X}_n$.

- It remains to optimize over σ^2. Note that

$$\frac{\partial L_n(\theta)}{\partial \sigma^2} = \left(\frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right) . / n$$

Replaced μ with $\bar{X}_n$.

Solve $\frac{\partial L_n(\hat{\theta})}{\partial \sigma^2} = 0$.

Example: Normal Distribution

- Note that there is a unique solution to the FOC:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2.$$

- The second order condition reveals that this is a local maximum.
- In fact, it is the global max, because though $L_n(\theta)$ is not concave in σ^2, its derivative remains negative for all values of σ^2 larger than $\hat{\sigma}^2$.

Example: Bernoulli Distribution

- Let $\{X_i\}_{i=1}^{n}$ be an iid sample from the Bernoulli distribution with parameter $\theta \in (0, 1)$. Note that

$$f_\theta\left(x\right) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$
$$= \theta^x \left(1 - \theta\right)^{1-x}.$$

- The likelihood function is

$$\ell_n\left(\theta\right) = \Pi_{i \le n} \theta^{X_i} \left(1 - \theta\right)^{1-X_i}$$

Example: Bernoulli Distribution

- The log-likelihood function is

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} [X_i \ln(\theta) + (1 - X_i) \ln(1 - \theta)]$$
$$= \bar{X}_n \ln(\theta) + \left(1 - \bar{X}_n\right) \ln(1 - \theta).$$

- The first order condition is given by

$$\frac{\partial L_n\left(\hat{\theta}_n\right)}{\partial \theta} = \frac{\bar{X}_n}{\hat{\theta}_n} - \frac{1 - \bar{X}_n}{1 - \hat{\theta}_n} = 0.$$

- The solution is $\hat{\theta}_n = \bar{X}_n$.

Example: Bernoulli Distribution

- Note that
$$\frac{\partial^2 L_n(\theta)}{\partial \theta^2} = -\frac{\bar{X}_n}{\theta^2} - \frac{1-\bar{X}_n}{(1-\theta)^2} < 0$$
for all values of θ, so the log likelihood is concave and the FOC suffices for a maximum.

Example: Uniform Distribution

- If $\{X_i\}_{i\geq 1}$ are iid with $X_i \sim U[0,\theta]$, for some $\theta > 0$, then

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta, \\ 0 & \text{otherwise,} \end{cases}$$

  so

$$\begin{aligned} \ell_n(\theta) &= \Pi_{i=1}^n \left(\frac{1}{\theta} \mathbf{1}(0 \leq X_i \leq \theta) \right) \\ &= \frac{1}{\theta^n} \mathbf{1}\left(\max_{i\leq n} X_i \leq \theta \right). \end{aligned}$$

- To maximize the likelihood, we want the smallest value of θ such that $\mathbf{1}(\max_{i\leq n} X_i \leq \theta) = 1$.
- This yields $\hat{\theta}_{MLE} = \max_{i\leq n} X_i$.

Conditional Maximum Likelihood

- Let $\{Y_i, X_i\}_{i=1}^{n}$ be an iid sample of $(K+1) \times 1$ random vectors. Suppose the conditional distribution of $Y|X$ is given by F_{θ_0} for some $\theta_0 \in \Theta$.
- Suppose this conditional distribution has density $f_\theta\left(y|x\right)$.
- The <u>conditional likelihood</u> of $(Y_1, \ldots, Y_n)$ given $(X_1, \ldots, X_n)$ is the conditional density evaluated at the sample points, regarded as a function of θ:

$$\ell_n\left(\theta\right) = \Pi_{i \leq n} f_\theta\left(Y_i|X_i\right).$$

- A <u>conditional maximum likelihood estimator</u> $\hat{\theta}_n$ of θ_0 is given by

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \ell_n\left(\theta\right).$$

Conditional Maximum Likelihood

- This is equivalent to maximizing the conditional log-likelihood function

$$L_n(\theta) = \frac{1}{n} \ln(\ell_n(\theta)) = \frac{1}{n} \sum_{i=1}^{n} \ln(f_\theta(Y_i | X_i)),$$

which is often easier to work with.

- Taking X to be a constant random variable shows that this concept generalizes (unconditional) ML discussed previously.

Example: Normal Regression

- Suppose $\{Y_i, X_i\}_{i=1}^n$ is an iid sample of 2×1 random vectors, and
$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}\right).$$
- We wish to estimate the best predictor of Y given X. A property of the bivariate normal distribution is that
$$Y|X \sim \mathcal{N}\left(\mu_y + \frac{\sigma_y}{\sigma_x}\rho\left(X - \mu_x\right), \left(1 - \rho^2\right)\sigma_y^2\right),$$
so $\mathrm{E}\left(Y|X\right) = \mu_y + \frac{\sigma_y}{\sigma_x}\rho\left(X - \mu_x\right) \equiv \beta_0 + \beta_1 X$: The best predictor is a linear function of X.

$\beta_1 = \frac{\sigma_y}{\sigma_x}\rho$.

Note: $E\left(Y - E(Y|X)\right)^2 = E\left(Y - \beta_0 - \beta_1 X\right)^2$, so best linear predictor is the best predictor. Minimizing gives $\beta_1 = \frac{\sigma_y}{\sigma_x}\rho$.

The least squares estimates minimize the

Example: Normal Regression

sample criterion $\frac{1}{n}\sum(y_i-\beta_0-\beta_1 x_i)^2$
yielding $\hat{\beta}_1^{OLS} = \frac{\hat{\sigma}_y}{\hat{\sigma}_x}\hat{\rho}$.

- Note: Restricting attention to the feature $\mathrm{E}(Y|X)$ means we don't require an estimate of all unknown parameters.
- Suppose $\{Y_i, X_i\}_{i=1}^n$ is an iid sample of 2×1 random vectors, and $Y|X \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma^2)$.
- It follows that for some U:

$$Y = \beta_0 + \beta_1 X + U; \qquad \mathrm{E}(U|X) = 0,$$

because $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$, so we can simply define $U = Y - \mathrm{E}(Y|X)$.

- The unknown parameter vector is $\theta = (\beta_0, \beta_1, \sigma^2)$.

Example: Normal Regression

- Conditional density given by

$$f_\theta(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x)^2\right),$$

which yields the log-likelihood

$$L_n(\theta) = \left(-\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right). / n$$

- Similar arguments establish that $\hat{\beta}_0, \hat{\beta}_1$ must be chosen to minimize

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2.$$

$\therefore$ MLE = OLS.

Example: Normal Regression

- This yields the well known OLS estimators:

$$\hat{\beta}_1 = \frac{\sum_i X_i Y_i - n\bar{X}_n \bar{Y}_n}{\sum_i \left(X_i - \bar{X}_n\right)^2}; \qquad \hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n.$$

- Finally:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 .$$

- In summary, if (Y, X) are bivariate normal, the maximum likelihood estimators of the parameters $\left(\beta_0, \beta_1, \sigma^2\right)$ which characterize the conditional distribution of $Y|X$ are the OLS estimates of (β_0, β_1), and a (biased) estimate of σ^2 in a linear regression of Y on X and a constant.

Properties of MLE

- It is not always possible to explicitly characterize a MLE, but there are general results which guarantee that a maximizer (or near maximizer) of the likelihood function will be consistent and asymptotically normal.
- Under certain regularity conditions (not satisfied by $U[0,\theta]$):

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I(\theta_0)^{-1}\right)$$

  where $I(\theta_0)$ is the Fisher Information, given by

$$I(\theta_0) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln\left(f_{\theta_0}(Y|X)\right)\right].$$

- This asymptotic variance is optimal in the sense that no smaller asymptotic variance can be attained by any 'regular' estimator. For this reason the MLE is called 'asymptotically efficient'.

What if $\hat{\theta}$ isn't (a function of) a sample average?

- Sometimes an estimator won't (quite) contain a sample average, but we still might be able to create one by transforming it. See PSET 4.
- Sometimes, analyse the distribution directly:
- E.g. If $X_i \sim U[0, \theta]$, $\hat{\theta}_{MLE} = \max_{i \leq n} X_i$, which is not a sample average, but a maximum of a collection of random variables.
- In PSET 4 we derive consistency and the limiting exponential distribution directly as $n \to \infty$.
- Under regularity conditions, we may also appeal to asymptotic normality results about maximizers of objective functions, known as "extremum estimators".

Example

- Let $\{X_i\}_{i=1}^n$ be iid draws from $\exp(\lambda)$, with pdf

$$f_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & x > 0, \\ 0 & x \leq 0. \end{cases}$$

- The likelihood function is:

$$\begin{aligned} \ell_n(\lambda) &= \prod_{i=1}^n \lambda \exp(-\lambda X_i) \\ &= \lambda^n \exp\left(-\lambda \sum_{i=1}^n X_i\right) \end{aligned}$$

- The log-likelihood is

$$\ell_n(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^n X_i,$$

which is a concave function of λ.

Example

- The first order condition is sufficient for a maximum, so:

$$\frac{n}{\hat{\lambda}_{ML}} - \sum_{i=1}^{n} X_i = 0,$$

which yields

$$\hat{\lambda}_{ML} = \frac{1}{\bar{X}_n}.$$

Example

- This is the same as the Method of Moments estimator based on $\mathrm{E}(X)$. To see this, note that

$$\mathrm{E}(\ln(f_\lambda(X))) = \ln\lambda - \lambda\mathrm{E}(X),$$

  which is concave and maximized by setting $\frac{1}{\lambda} = \mathrm{E}(X)$.
- The FOC is

$$\begin{aligned}\frac{\partial}{\partial\lambda}\mathrm{E}(\ln(f_{\lambda_0}(X))) &= \mathrm{E}\left(\frac{\partial}{\partial\lambda}\ln(f_{\lambda_0}(X))\right)\\ &= \mathrm{E}\left(\frac{1}{\lambda_0} - X\right) = 0.\end{aligned}$$

- The method of moments estimator based on this equality satisfies:

$$\frac{1}{\hat{\lambda}_{MM}} = \bar{X}_n.$$

Example: Bias

- The <u>bias</u> of $\hat{\lambda}_{ML}$ is given by $\text{Bias}\left(\hat{\lambda}_{ML}\right) = \mathrm{E}\left(\hat{\lambda}_{ML}\right) - \lambda$.
- By Jensen's inequality, since $g(x) = \frac{1}{x}$ is a convex function:

$$\mathrm{E}\left(\hat{\lambda}_{ML}\right) = \mathrm{E}\left(\frac{1}{\bar{X}_n}\right) > \frac{1}{\mathrm{E}\left(\bar{X}_n\right)} = \frac{1}{(1/\lambda)} = \lambda,$$

so $\hat{\lambda}_{ML}$ is <u>biased upward</u>.

Example: Consistency

- Recall: An estimator $\hat{\theta}_n : (X_1, \dots, X_n) \to \mathbb{R}$ of θ is consistent if
$$\hat{\theta}_n \xrightarrow{p} \theta.$$
- Step 1: See that $\hat{\lambda}_{ML}$ contains a sample average (we can apply SLLN!):
$$\hat{\lambda}_{ML} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i}.$$
Conclude that
$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathrm{E}(X) = \frac{1}{\lambda}.$$

Example: Consistency

- Step 2: Use the continuous mapping theorem. The condition we have to check is that $g(x) = \frac{1}{x}$ is continuous at the limit $\frac{1}{\lambda}$ with probability 1. Since the limit random variable is a constant, and $\lambda > 0$, this is satisfied, so

$$g\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \stackrel{a.s.}{\to} g\left(\frac{1}{\lambda}\right) = \lambda.$$

- Since $\stackrel{a.s.}{\to}$ implies $\stackrel{p}{\to}$, we conclude that $\hat{\lambda}_{ML}$ is a consistent estimator of λ.

Example: Asymptotic distribution

- We are searching for constants $r \geq 0$ and c such that

$$n^r \left(\hat{\lambda}_{ML} - c \right) \xrightarrow{d} Y$$

  for some non-degenerate Y.
- Step 1: Check $r = 0$: We previously established $\hat{\lambda}_{ML} \xrightarrow{p} \lambda$. If $r = 0$,

$$\hat{\lambda}_{ML} - \lambda \xrightarrow{p} 0,$$

  which is degenerate.

Example: Asymptotic distribution

- Step 2: Find c given that $r > 0$. In PSET 3 we showed that if

$$n^r \left(\hat{\lambda}_{ML} - c\right) \xrightarrow{d} Y$$

then $\hat{\lambda}_{ML} \xrightarrow{p} \lambda$. So, if there exist constants $r > 0$ and c such that this holds, $c = \lambda$.

- Step 3: Find $r > 0$ and Y. Looking for $r > 0$ such that

$$n^r \left(\hat{\lambda}_{ML} - \lambda\right) \xrightarrow{d} Y.$$

Example: Asymptotic distribution

- Step 3a: How did we establish consistency? Was there a sample average? (Yes! Use the CLT):
- Note that

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\lambda}\right) \xrightarrow{d} \mathcal{N}\left(0, Var\left(X_i\right)\right)$$

by the CLT. Can show that

$$Var\left(X_i\right) = \mathrm{E}\left(X_i^2\right) - \mathrm{E}\left(X_i\right)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

using integration by parts (twice), yielding

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\lambda}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right).$$

Example: Asymptotic distribution

- Step 3b: Compare CLT result with what we actually want:
- Note that

$$\sqrt{n}\left(\hat{\lambda}_{ML}-\lambda\right)=\sqrt{n}\left(\frac{1}{\bar{X}_n}-\frac{1}{(1/\lambda)}\right),$$

so we try the Delta Method with $g\left(x\right)=\frac{1}{x}$. This yields

$$\sqrt{n}\left(g\left(\bar{X}_n\right)-g\left(\frac{1}{\lambda}\right)\right)\xrightarrow{d}\mathcal{N}\left(0,g'\left(\frac{1}{\lambda}\right)^2\cdot\frac{1}{\lambda^2}\right).$$

- Simplify the variance to obtain

$$\sqrt{n}\left(\hat{\lambda}_{ML}-\lambda\right)\xrightarrow{d}\mathcal{N}\left(0,\lambda^2\right).$$