ECON6190 Section 12

Nov. 15, 2024

Yiwei Sun

Maximum Likelihood Estimation for Parametric Models

↳ X has pmf/pdf $f(x|\theta)$, with known f, θ is finite dimensional but unknown

$\boxed{\text{DEF}}$ (correct specification)

A model is correctly specified when there is a unique $\theta_0 \in \Theta$ s.t. $f(x|\theta_0)$ coincides with the true density of X.

$\boxed{\text{DEF}}$ (misspecification)

A model is misspecified if no value of $\theta \in \Theta$ ....

$\boxed{\text{DEF}}$ (likelihood function (of iid data))

$$\underbrace{L_n(\theta)}_{\text{joint density / pmf of the data}} = f(x_1, \cdots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$\boxed{\text{DEF}}$ (MLE)

A MLE $\hat{\theta}_{ML}$ maximizes

a function of parameter
= what value of θ maximizes the likelihood of observing sample $\{x_1, \cdots, x_n\}$

- the likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$
- the log-likehood function $\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$

$\boxed{\text{DEF}}$ (Expected likelihood function)

$$\ell(\theta) = E[\log \underbrace{f(x|\theta)}_{\text{pdf of the rv } X \text{, not indexed by } n}]$$

Theorem When a model is correctly specified, θ_0 maximizes $\ell(\theta)$.

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmax}} \, E[\log f(x|\theta)]$$

$$\hat{\theta}_{ML} = \underset{\theta \in \Theta}{\operatorname{argmax}} \, \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta)$$

→ sample analog

$\boxed{\text{DEF}}$ (Efficient score) $S = \frac{\partial}{\partial \theta} \log f(X|\theta_0)$

$\boxed{\text{DEF}}$ (Fisher information) $\mathcal{I}_\theta = E[SS']$

Theorem When a model is correctly speified, support X doesn't depend on θ, $\theta_0 \in \text{int}\,\Theta$, then $E[s]=0$, $var(s)=\mathcal{F}_\theta$.

Proof.
$$E[s] = E\left[\frac{\partial}{\partial\theta}\log f(x|\theta_0)\right]$$
Leibnize rule
$$= \frac{\partial}{\partial\theta}\underbrace{E[\log f(x|\theta_0)]}_{\ell(\theta_0)}$$
$$= \frac{\partial}{\partial\theta}\ell(\theta_0)$$
$$= 0 \qquad \text{b/c } \theta_0 \in \Theta \text{, interior sol given by FOC.}$$
$$var(s) = E[ss'] - (\underbrace{E[s]}_{0})^2 = \mathcal{F}_\theta.$$

Theorem (Information matrix equality)
$$\underbrace{E\left[\frac{\partial\log f(x|\theta_0)}{\partial\theta}\frac{\partial\log f(x|\theta_0)}{\partial\theta'}\right]}_{\text{Fisher information}} = \underbrace{-E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(x|\theta_0)\right]}_{\mathcal{H}_\theta \text{ expected Hessian}}$$

Theorem Assume model is correctly speified, support X doesn't depend on θ, $\theta_0 \in \text{int}\,\Theta$. If $\tilde\theta$ is an unbiased estimator of θ, then
$$var(\tilde\theta) \geq \underbrace{(n\mathcal{F}_\theta)^{-1}}_{\text{CRLB}}$$

An estimator is cramer-Rao efficient if it is unbiased and $var(\tilde\theta) = (n\mathcal{F}_\theta)^{-1}$.

Intuition: more curvature of log likelihood function wrt θ, more information data provides about θ b/c $var(\theta)$ is lower bounded by the inverse of $\mathcal{H}_\theta$. more curvature $\Rightarrow$ more preuise estimate of $\hat\theta_{ML}$

Asymptotic Properties of MLE

$var(\hat\theta_{ML}) = (n\mathcal{F}_\theta)^{-1}$

MLE is
- consistent
- asymptotically normal $\sqrt{n}(\hat\theta_{ML}-\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{F}_\theta^{-1})$
- asymptotically Cramer-Rao efficient.

4. Based on the notation in the slides on *Estimation*, let us prove the Information Matrix Equality

$$\mathbb{E}\left[\frac{\partial^2 \log f(X|\theta_0)}{\partial\theta\partial\theta'}\right] = -\mathbb{E}\left[\frac{\partial \log f(X|\theta_0)}{\partial\theta}\frac{\partial \log f(X|\theta_0)}{\partial\theta'}\right].$$

Let $f = f(x|\theta_0)$, ∇_j means derivative with respect to the j-th element $\theta^{(j)}$, and ∇_{jk} mean 2nd-order derivative with respect to $\theta^{(j)}$ and $\theta^{(k)}$. Suppose we can exchange the integral "$\int$" and derivatives "∇_j".

(a) By differentiating $\int f dx = 1$ with respect to $\theta^{(j)}$, show that $\mathbb{E}[\nabla_j \log f] = 0$.

(b) By differentiating $\mathbb{E}[\nabla_j \log f] = 0$ with respect to $\theta^{(k)}$, show that

$$\mathbb{E}[\nabla_{jk} \log f] + \mathbb{E}\left[(\nabla_j \log f)(\nabla_k \log f)\right] = 0,$$

which yields the Information Matrix Equality.

(a) $$\nabla_j \int f\,dx \underset{\uparrow}{=} \int \nabla_j f\,dx = 0$$

exchange $\int, \nabla_j$

$$0 = \int (\nabla_j f)\,dx$$
$$= \int \boxed{(\nabla_j f)\,\frac{1}{f}} \cdot f\,dx \quad \left(\frac{1}{f}\cdot f = 1\right)$$

Notice: $\nabla_j \log f = \frac{1}{f}\nabla_j f$ by chain rule

$$= \int \underbrace{(\nabla_j \log f)}_{\text{a function of } x}\, f\,dx$$

density $f(x|\theta_0)$

def.

$$= E[\nabla_j \log f]$$

$f(x|\theta_0)$

→ "population" FOC of the j-th element

(b) $$0 = \nabla_k E[\nabla_j \log f]$$

plug in
$$= \nabla_k \int (\nabla_j \log f)\, f\,dx$$

exchange $\int, \nabla$
$$= \int \nabla_k \big((\nabla_j \log f) f\big)\,dx$$

Product rule
$$= \int \{(\nabla_{jk} \log f) f + (\nabla_j \log f)(\nabla_k f)\}\,dx$$

$\int$ is linear
$$= \underbrace{\int (\nabla_{jk} \log f) f\,dx}_{E[\nabla_{jk}\log f]} + \underbrace{\int (\nabla_j \log f)(\nabla_k f)\,dx}$$

$f(x|\theta_0)$

$\nabla_k \log f = \frac{1}{f}\nabla_k f$

$$= \int (\nabla_j \log f)\boxed{(\nabla_k f)\cdot\frac{1}{f}}\cdot f\,dx$$
$$= \int (\nabla_j \log f)(\nabla_k \log f)\cdot f\,dx$$
$$= E\left[(\nabla_j \log f)(\nabla_k \log f)\right]$$

3. Suppose X follows a normal distribution with unknown mean μ and variance $\sigma^2 > 0$. The density of X is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

Given a random sample $\{X_i, i = 1 \dots n\}$ drawn from X, find the MLE estimator for (μ, σ^2).

likelihood function: $L_n(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right)$ b/c random sampling

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}}$$

log-likelihood function $\ell_n(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$

MLE $(\hat{\mu}, \hat{\sigma}^2)$ should satisfy FOC:

$$\left.\frac{\partial \ell_n(\mu, \sigma^2)}{\partial \mu}\right|_{\substack{\mu=\hat{\mu},\\ \sigma^2=\hat{\sigma}^2}} = 2\left(\frac{1}{2\hat{\sigma}^2}\right)\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \quad \text{----} \ ①$$

$$\left.\frac{\partial \ell_n(\mu, \sigma^2)}{\partial \sigma^2}\right|_{\substack{\mu=\hat{\mu},\\ \sigma^2=\hat{\sigma}^2}} = -\frac{n}{2}\frac{1}{2\cancel{\pi}\hat{\sigma}^2}2\cancel{\pi} + \sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{2(\hat{\sigma}^2)^2} = 0 \quad \text{----} \ ②$$

take derivative with (σ^2) as a whole

①: we get $\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0$

$$\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

plug $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ into ②: $\frac{n}{\cancel{2}\cancel{\hat{\sigma}^2}} = \sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\cancel{2}(\hat{\sigma}^2)^{\cancel{2}}}$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n}x_i\right)^2.$$

To show $\hat{\mu}, \hat{\sigma}^2$ are the maximizer of $\ell_n(\mu, \sigma^2)$, need to check Hessian

Let $\theta = (\mu \ \ \sigma^2)$, $\hat{\theta} = (\hat{\mu} \ \ \hat{\sigma}^2)$

$$\frac{\partial^2 \ell_n(\theta)}{\partial\theta\partial\theta'} = \begin{pmatrix} \frac{\partial^2 \ell_n(\mu, \sigma^2)}{\partial\mu\partial\mu} & \frac{\partial^2 \ell_n(\mu, \sigma^2)}{\partial\mu\partial\sigma^2} \\ \frac{\partial^2 \ell_n(\mu, \sigma^2)}{\partial\sigma^2\partial\mu} & \frac{\partial \ell_n(\mu, \sigma^2)}{\partial\sigma^2\partial\sigma^2} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial}{\partial\mu}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{1}{\sigma^2}n\mu\right) & \frac{\partial}{\partial\sigma^2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\right) \\ \frac{\partial}{\partial\sigma^2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\right) & \frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2}\frac{1}{\sigma^2} + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^4}\right) \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 \end{pmatrix}$$

$$\left.\frac{\partial^2 l_n(\theta)}{\partial\theta\partial\theta'}\right|_{\theta=\hat\theta} = \begin{pmatrix} -\frac{n}{\hat\sigma^2} & 0 \\ 0 & -\frac{n}{2\hat\sigma^4} \end{pmatrix}$$ is ND.

b/c $\frac{1}{\hat\sigma^4}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n}x_i)$

$$= \frac{1}{\hat\sigma^4}\underbrace{\left(\sum_{i=1}^{n}x_i - \cancel{n}\left(\frac{1}{\cancel{n}}\sum_{i=1}^{n}x_i\right)\right)}_{=0} = 0$$

$$\frac{n}{2\hat\sigma^4} - \frac{1}{\hat\sigma^6}\sum_{i=1}^{n}(x_i-\hat\mu)^2$$

$$= \frac{n}{2\hat\sigma^4} - \frac{n}{\hat\sigma^{\cancel{6}\,4}}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat\mu)^2}_{=\cancel{\hat\sigma^2}} = \frac{n}{2\hat\sigma^4} - \frac{n}{\hat\sigma^4} = -\frac{n}{2\hat\sigma^4}$$