Lecture 7. Maximum Likelihood Estimation.

Re-cap of last lecture

$x \sim f_x(x|\theta)$ — plug in random variable

$\ell(\theta|x) = \log f_x(x|\theta) \longrightarrow \max_\theta$

To find max likelihood, take max / derivative

$S(\theta) = \frac{\partial \ell}{\partial \theta}$

when multivariate have matrix

1st info eq. $ES(\theta_0) = 0$

I diff. at diff values of θ $\quad I(\theta) = \mathrm{Var}(S(\theta))$ $\hookrightarrow E(S(\theta_0)) = 0$

$= E(S^2(\theta)) - E[S(\theta)S'(\theta)]$

2nd info eq:

$$I(\theta) = -\mathcal{E}\left[\frac{\partial^2 l}{\partial\theta\,\partial\theta'}\right]$$

Rao-Cramer bound

$\hat{\theta} = W(X)$ you calculated $\hat{\theta}$ from data

unbiased $E\hat{\theta} = \theta$

Theorem 1 (Rao-Cramer bound).

Let X be a random data with distribution $f(x|\theta)$ and information $I(\theta)$. Let $W(X)$ be an estimator of θ such that

(1) $\frac{d}{d\theta} E_\theta[W(X)] = \int W(x) \frac{\partial f(x|\theta)}{\partial \theta} dx$

(2) $Var(W) < \infty$.

Then

$$Var(W) \geq \left(\frac{d}{d\theta} E_\theta[W(X)] \right)^2 \frac{1}{I(\theta)}.$$

$\Rightarrow Var(\hat{\theta}) \geq \frac{1}{I(\theta)}$

In particular, if $X = (X_1, ..., X_n)$ is an i.i.d. random sample and W is unbiased for θ, then

$Var(W) \geq \frac{1}{I(\theta)} = \frac{1}{nI_1(\theta)}$

data

$I_n = n I_1$

$Var(\hat{\theta}) \geq \frac{1}{n I_1(\theta)}$

ML Estimator

Parametric family:

- $X \sim f(x|\theta)$ with $\theta \in \Theta$, $f(\cdot|\theta)$ is pdf or pmf
- $\ell(\theta|X) = \log f(X|\theta)$ is log-likelihood function

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} \ell(\theta|x)$$

- F.O.C $s(\hat{\theta}_{ML}) = \frac{\partial \ell(\hat{\theta}_{ML}|X)}{\partial \theta} = 0$

$x = (x_1, \ldots, x_n)$ $\quad x_i \sim iii\ f_1(\cdot|\theta)$ || individual pdf.

pdf. $f(x|\theta) = \prod_i f_1(x_i|\theta)$

$$l(\theta) = \sum_{i=1}^{n} l_1(\theta|x_i) = \sum_{i=1}^{n} \log f_1(x_i|\theta)$$

Then take FOC $s(\hat{\theta}) = 0$

$$\frac{\partial l_n}{\partial \theta} = \sum_{i=1}^{n} (\theta|x_i) = \sum_{i=1}^{n} s_1(\theta|x_i)$$

ML Estimator

- If $X = (X_1, ..., X_n)$ i.i.d $f_1(\cdot|\theta_0)$ with $\theta_0 \in \Theta$
- Then the joint pdf is $f(x|\theta) = \prod_{i=1}^{n} f_1(x_i|\theta)$
- The log-likelihood is $\ell(\theta|x) = \sum_{i=1}^{n} \log f_1(x_i|\theta)$
- F.O.C. $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_1(\hat{\theta}_{ML}|x_i)}{\partial \theta} = 0$ FOC $\frac{1}{n}\sum_{i=1}^{n} (\hat{\theta}|x_i) = 0$

$E\, g_1(\theta_0)^* = 0$ ← population eg (1st info. equality)

* is a method of moment.

Theorem 2 (MLE consistency).

In the setting above, assume that

diff value of θ give diff. distn so diff pdfs.

(1) θ_0 *is identifiable, i.e. for any* $\theta \neq \theta_0$, *there exists* x *such that* $f(x|\theta) \neq f(x|\theta_0)$,

(2) *the support of* $f(\cdot|\theta)$ *does not depend on* θ,

(3) θ_0 *is an interior point of parameter space* Θ.

Then $\hat{\theta}_{ML} \to_p \theta_0$.

→ not on the boundary

* Outline of proof:

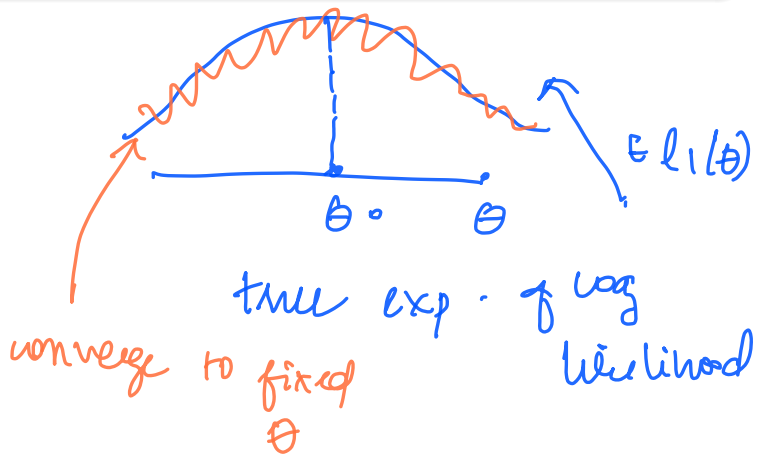
1) $\frac{1}{n}\sum_i \ell_1(\theta|X_i) \xrightarrow{\text{converge}} E\ell_1(\theta)$

averages do converge to

2) $\theta_0 = \text{(true)} \ \arg\max_\theta E\ell_1(\theta)$

Taking this two,

3) $\hat{\theta}_{ML} = \arg\max \frac{1}{n} E\ell_1(\theta|x_i) \xrightarrow{P} \theta_0$



eg: consistency of GMM proved using this method.

$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow N(0, I_1^{-1})$

Theorem 3 (MLE asymptotic normality).

In the setting above, assume that conditions (1)-(3) in the MLE consistency theorem hold. In addition, assume that

(4) $f_1(x_i|\theta)$ *is thrice differentiable with respect to* θ *and we can interchange integration with respect to* x *and differentiation with respect to* θ,

(5) $|\partial^3 \log f_1(x_i|\theta)/\partial\theta^3| \leq M(x)$ *and* $E[M(X_i)] < \infty$.

Then

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow N(0, I_1^{-1}(\theta_0))$$

Sketch of proof

$$0 = \frac{d\ell_n}{d\theta}(\hat{\theta}_{MLE}) \quad \text{— FOC we have}$$

$$= \frac{d\ell}{d\theta}(\theta_0) + \frac{d^2\ell}{d\theta d\theta}(\theta^*) \cdot (\hat{\theta}_{MLE} - \theta_0) \quad \text{— Taylor Exp}$$

θ^* some in between point between θ_0 & $\hat{\theta}_{MLE}$

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\frac{\partial\ell}{\partial\theta}(\theta_0)\frac{1}{\sqrt{n}}}{\frac{\partial^2\ell}{\partial\theta\partial\theta}(\theta^*)\frac{1}{n}}$$

② — always on numerator to normalize

① — always.

Now want normalizing:

solving once iid X_i's & converge to $\hat{\theta}$

① $\frac{1}{n}\frac{\partial^2 \ln(\theta^*)}{\partial\theta\partial\theta}$ = canonical accum = $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ell_1(\hat{\theta}|X_i)}{\partial\theta\partial\theta}_{\theta_0}$ random but close to θ_0

so denominator ① converges to constant.

By we get $\xrightarrow{P}$

$$E\frac{\partial^2 \ell_1}{\partial\theta\partial\theta}(\theta_0|X_i) = -I_1$$

in future info

② $\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial\theta}(\theta_0) = \frac{1}{\sqrt{n}}\sum\frac{\partial \ell_1}{\partial\theta}(\theta_0|X_i)$

iid random variable

$= \frac{1}{\sqrt{n}}\sum \underbrace{S_1(\theta_0|X_i)}_{Y_i}$

$EY_i = ES_1 = 0$

$\mathrm{Var}\,Y_i = \mathrm{Var}(S_1) = I_1$

$\frac{1}{\sqrt{n}}\sum_i y_i \xrightarrow{CLT} \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} y_i - 0\right) \Rightarrow N(0, I_1)$

② numerator converges to Gaussian

So, $\sqrt{n}(\hat{\theta}-\theta_0) \Rightarrow -\frac{1}{I_1}\cdot N(0, I_1) = N\left(0, \frac{I_1}{I_1^2}\right) = N\left(0, \frac{1}{I_1}\right)$

Examples

"kinda" : $\hat{\theta}_{ML} - \theta_0 \overset{\sim}{} N\left(0, \frac{1}{I_1}\right) = N\left(0, \frac{1}{I_n}\right)$

$Var\left(\hat{\theta}_{ML} - \theta_0\right) \times \frac{1}{I_n}$

MLE is asymptotically efficient.

Examples 1) $x_1, \dots, x_n \sim iid \ \exp(\lambda)$ — expo

$x \geq 0$
$\lambda \geq 0$

$x_i \sim \exp(\lambda)$ $\quad f_1(x_i|\lambda) = \lambda \exp(-\lambda x)$

$l_1(\lambda) = \log\lambda - \lambda x_i$ — one obs log likelihood

$l_n(\lambda) = n\log\lambda - \lambda \sum x_i$ — for all. $\longrightarrow \max_\lambda$

Taking FOC: $\frac{n}{\lambda} - \sum x_i$ & so $\hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$

Q) Exam what is the distbn of $\hat{\lambda}$

$$E X_i = \frac{1}{\lambda} = \int_0^\infty x \lambda \exp(-\lambda x)\, dx$$

Var $(X_i) = \frac{1}{\lambda_2}$

Can calc by int. by parts twice.

*One way of getting asymptotics:

$$\bar{X} \xrightarrow{P} E X_i = \frac{1}{\lambda}, \text{ by CLT} \qquad \hat{\lambda}_{ML} = \frac{1}{\bar{X}} \xrightarrow{P} \frac{1}{E X_i} = \frac{1}{1/\lambda} = \lambda$$

$$\hat{\lambda}_{ML} \longrightarrow \lambda_0 \text{ (consistent)}$$

By Delta method, $g(x) = \frac{1}{x}$

$$\sqrt{n}\left(g(\bar{X}) - g\left(\frac{1}{\lambda}\right)\right) = \sqrt{n}(\hat{\lambda} - \lambda_0) \Longrightarrow N(0, ?)$$

calc. variance.

$I_1 = 1/\lambda_2$

$$\frac{\partial l_1}{\partial \theta} = \frac{1}{\lambda} - x_i \qquad \frac{\partial^2 l_1}{\partial \theta \partial \theta} = -\frac{1}{\lambda^2}$$

* Second way:

$$\sqrt{n}(\hat{\lambda}_{ML} - \lambda) \to N\left(0, \frac{1}{I_1(\lambda)}\right)$$

so $\sqrt{n}(\hat{\lambda}_{ML} - \lambda) \Longrightarrow N(0, \lambda^2)$

How to calculate Fisher Information

- If $I_1(\theta)$ is continuous in θ (which is needed for asymptotic results), then $\left(I_1(\hat{\theta}_{ML})\right)^{-1}$ is consistent for $(I_1(\theta_0))^{-1}$;
- $\hat{I} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ell_1(\hat{\theta}|X_i)}{\partial \theta^2}$ is consistent for $I_1(\theta_0)$
- $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial \ell_1(\hat{\theta}|X_i)}{\partial \theta}\right)^2$ is consistent for $I_1(\theta_0)$

usually $\hat{I} \to I_1(\theta_0)$

$I_1(\theta_0) = E\left(\left[\frac{\partial \ell_1(\theta_0|X_i)}{\partial \theta}\right]^2\right) = -E\left[\frac{\partial^2 \ell_1}{\partial\theta\,\partial\theta}(\theta_0|X_i)\right]$

Challenge 1 — Exp is the integral. Take avg: $E \approx \frac{1}{n}\sum(\,|X_i)$

Challenge 2 is using at θ_0: $\theta_0 \approx \tilde{\theta}_{ML}$ - bec ML is consistent

White Sconer

$I_{(2)}^{-1}\ I_{(1)}\ I_{(2)}^{-1}$

We had from previous proof: $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow N\left(0, \frac{I_{(1)}}{I_{(2)}^2}\right)$

How to calculate Fisher Information

Another idea is parametric bootstrap:

- For $b = 1, ..., B$ simulate sample $X_b^* = (X_{1b}^*, ..., X_{nb}^*)$ as i.i.d. draws from $f_1(x_i|\hat{\theta}_{ML})$ draw sample from $\hat{\theta}_{ML}$
- Find MLE using sample X_b^*, denote it $\hat{\theta}_b^*$
- Calculate the sample variance of $(\hat{\theta}_1^*, ..., \hat{\theta}_B^*)$

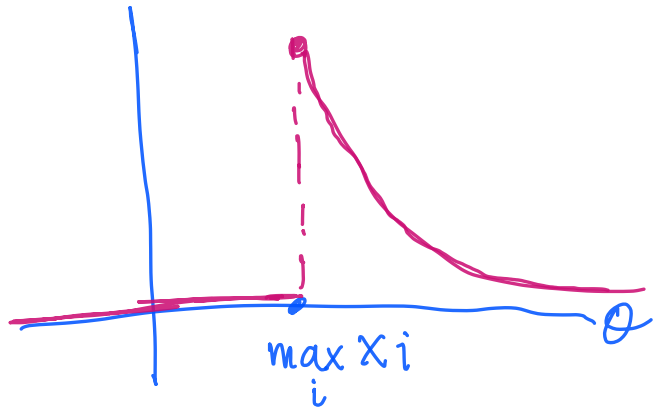
$$\widehat{Var}(\hat{\theta}_{ML}) \simeq \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \bar{\theta}^*\right)^2$$

$\bar{\theta}^* \approx \hat{\theta}_{MM}$

When MLE asymptotic theory fails us

Example 1 If support depends on the parameter.
Consider $X_1, ..., X_n$ be i.i.d. from $U[0, \theta]$.

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } x \in [0, \theta] \\ 0 & \text{otherwise} \end{cases}$$



joint pdf $f(x) = \frac{1}{\theta^n} \mathbb{1}\{\max_i x_i \le \theta\}$ ← decaying

$\hat{\theta}_{ML} = \max_i X_i \xrightarrow{P} \theta_0$ as $n \to \infty$ (consistent)

- $\hat{\theta}_{ML} \le \theta_0$ (always on one side) so not be gaussian

- super consistent: est. $n(\hat{\theta}_{ML} - \theta_0) \Rightarrow$ special-distbn (not normal) ← like exp

Violation of MLE theory: assumptions violated.

When MLE asymptotic theory fails us

when parameter on boundary

Example 2 If the true parameter value θ_0 is on the boundary of Θ.
Consider $X_1, ..., X_n$ i.i.d. $N(\mu, 1)$ with $\mu \geq 0$.

$\geq$ is like a constraint

$X_1, \dots, X_n \sim iid\ N(\mu, 1)$

$$\ln(\mu) = \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum(x_i - \mu)^2 \to \max_{\mu \geq 0}$$

over parameter space

FOC: $\hat{\mu} = \bar{x}$

true restricted optimum $\hat{\mu}_2 = \begin{cases} \bar{x} & ; \text{if } \bar{x} > 0 \\ 0 & \text{otherwise} \end{cases}$

if $\mu_0 = 0$, $\sqrt{n}(\hat{\mu}_{m2} - \mu_0) = \begin{cases} \end{cases}$

with $p = 1/2$ $\bar{x} < 0$ and $\hat{\mu}_{ML} = 0$

"parameter on boundary problem"
"moment inequality problem"

|| still active area of research

When MLE asymptotic theory fails us

μ_i hard to get with σ^2 can

problem of FE.

Example 3 Incidental parameter problem (number of parameters grow)
Consider

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right)$$

for $i = 1, \ldots, n$, Two independent scores for two indep. tests

$$l_n = n \log(2\pi) - n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i \left[(X_{1i} - \mu_i)^2 + (X_{2i} - \mu_i)^2 \right]$$

$\hookrightarrow$ $\max_{\{\mu_i, \sigma^2\}}$

$$\hat{\mu}_i = \frac{X_{1i} + X_{2i}}{2}$$

$$X_{ji} - \hat{\mu}_i = X_{1i} - \frac{X_{1i} + X_{2i}}{2} = \frac{X_{1i} + X_{2i}}{2}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4} \sum \left\{ \underbrace{(X_{1i} - \hat{\mu}_i)^2}_{\frac{(X_{1i} - X_{2i})^2}{4}} + \underbrace{(X_{2i} - \hat{\mu}_i)^2}_{\frac{(X_{1i} - X_{2i})^2}{4}} \right\}$$

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_i \left\{ (x_{1i} - \hat{\mu}_i)^2 + (x_{2i} - \hat{\mu}_i)^2 \right\} = \frac{1}{2n} \sum_i \frac{2}{4} (x_{1i} - x_{2i})^2$$

$$\hat{\sigma}^2 = \frac{1}{4n} \sum (x_{1i} - x_{2i})^2$$ — now see whether $\hat{\sigma}^2$ is consistent or not.

$$\rightarrow \frac{1}{4} E(x_{1i} - x_{2i})^2 = \frac{1}{4} \operatorname{var}(x_{1i} - x_{2i})$$

$$= \frac{1}{4} (\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}$$

So $\hat{\sigma}^2 \longrightarrow \frac{\sigma^2}{2}$ not to σ^2 so the est. is not consistent.

* when some more parameters to est, not be consistent the degrees of freedom also increase.

Quasi-MLE

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$

- $X = (X_1, ..., X_n)$ i.i.d from $g(x_i)$.
- We assumed $X_i \sim f_1(x_i|\theta)$.
- What would happen if we do MLE?
- MLE estimates a "pseudo-true" parameter value θ_0:

$$\theta_0 = \arg\max_\theta \int \log[f_1(x_i|\theta)] g(x_i) dx_i = \arg\max_\theta \mathbb{E}_g \log f_1(X_i|\theta).$$

$$=$$

$$\ell = \sum_{i=1}^{n} x_i \log p + (1-x_i)\log(1-p) - \log(\sigma x_i) - \left[\frac{1}{2}\log(2\pi) - \frac{(y_i - u(x_i))^2}{2\sigma(x_i)^2} \right]$$

$x_i \in \{0,1\}$

$$\ell(\sigma(0), \sigma(1), u(0), u(1)) =$$

$$\sum_{i=1}^{n} x_i \log p + (1-x_i)\log(1-p) - \log(\sigma x_i) - \frac{1}{2}\log(2\pi) - \frac{(y_i - u(0))}{2\sigma(0)}$$