36-705

12/06/2020

Youtube lecture 10/10/16

Lecture notes 9 - Asymptotic theory (cont.)

Test material for Test II covered here
- can bring cheatsheets (1 side)

3 question test
- convergence, delta
- point estimation
- sufficiency

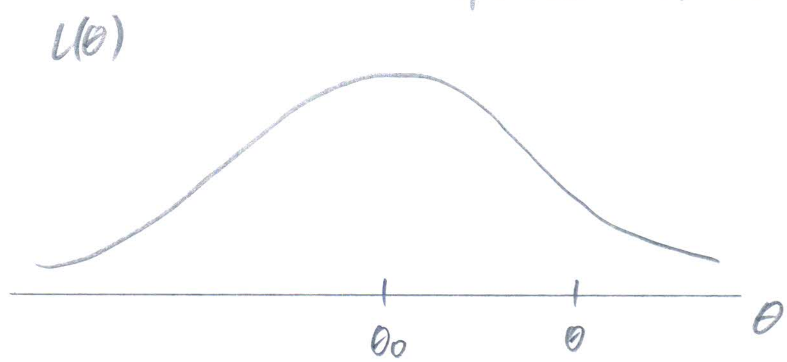
consistency of MLE: $\hat{\theta}_{n,MLE} \xrightarrow{p} \theta_0$

(*) Brief review (*)

$$P\left(\frac{L(\theta_0)}{L(\theta)} > 1\right) \rightarrow 1$$

as $n \rightarrow \infty$

θ_0 - true param value.



- Consistency: Estimator converges in probability to true parameter
- In general, prove via definition (convergence in prob); or via quadratic mean convergence, on a case-by-case basis.
- Above result was an attempt to establish whether MLE is, in general, consistent.
- Above result shows that $P(L(\theta_0) > L(\theta)) \rightarrow 1$ i.e. probability of likelihood evaluated at true parameter value θ_0 being greater than likelihood evaluated at an arbitrary param value approaches 1 in the limit as $n \rightarrow \infty$ (i.e. more samples collected.)

(*) That is close to, but not quite a proof of the consistency of MLE; again due to the distinction between pointwise and uniform convergence

(*) Consistency of MLE would involve showing that the maximum of the likelihood at points other than θ_0 is always smaller than that of $L(\theta_0)$ (?) - tighter expl. (A1) in terms of point-wise and uniform convergence.

(*) Proof of MLE consistency in general → beyond course scope.

(*) (th): In general MLE is consistent under regularity conditions:

1) Dimension of θ is fixed i.e. $\theta \in \mathbb{R}^d$ with d fixed
2) $p(x;\theta)$ is a smooth function of θ.
3) Identifiability of parameters. - intuitively, can't have two different values of parameter refer to the same distribution

i.e. $\theta_1 \neq \theta_2$

$p(x;\theta_1) \neq p(x;\theta_2)$ where "not equal" means "not equal" in distribution

(*) Practical takeaway: MLE is an effective estimation procedure when there is a large amount of data (large n) and small no. of parameters θ

(*) Nonparametric/high dimensional settings $\rightarrow$ this is not the case.

(*) MLE in general is also asymptotically normal

---

6. Asymptotic Normality of MLE

- We will prove that MLE satisfies:-

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$$

- or informally;

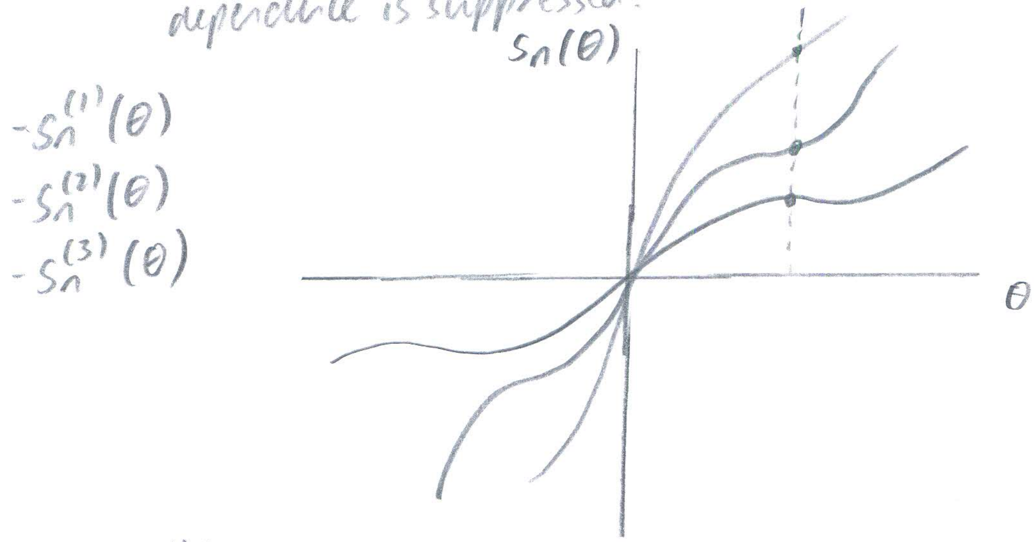
$$\hat{\theta}_n \approx N\left(\theta, \frac{1}{nI(\theta)}\right)$$

- 2 definitions:- score function and Fisher information.

score function : $S_n(\theta) \equiv S_n(\theta, X_1, X_2, \ldots, X_n) = l'(\theta) = \dfrac{\partial \log p(X_1, \ldots, X_n; \theta)}{\partial \theta}$

- IID data:- $S_n(\theta) = \displaystyle\sum_{n=1}^{N} \frac{\partial \log p(X_i;\theta)}{\partial \theta}$

- $S_n(\theta) = S_n(\theta, X_1, \ldots, X_n)$ is a function of data and parameters; although formal dependence is suppressed.



- Score function $S_n(\theta)$ is a function of θ.
- But it is a random function that depends on dataset $X_1, \ldots, X_n$.

- Graphic shows score function for 3 datasets of size n.
- $S_n^{(i)}(\theta)$ $i=1,\ldots 3$ (e.g. through simulation)

- We can then also conceive of mean and variance of the score function.

Fisher Information

(*) The Fisher Information is defined to be:-

$$I_n(\theta) = Var_\theta(S_n(\theta))$$

([redacted] is defined as (variance of the score function))

(*) - In context of graphic, it can help to visualise this generatively. In that we can think of computing the variance of the score function over the 3 datasets we have at a particular value of θ. Doing this for all values of θ yields the Fisher information ([redacted] 2D slice of graph)

- Also note that:-

$$Var(\hat{\theta}_{MLE}) \approx \frac{1}{I_n(\theta)}$$

- under MLE; variance of estimator is approx equal to Fisher info.

LW: called 'information' $\longrightarrow$ if $I_n(\theta)$ is large; there is a lot of information, and $Var(\hat{\theta}_{MLE})$ is small, ie. our estimator is precise.

LW: Provides theoretical understanding of MLE; but also a means of computing standard errors (standard dev of estimators).

Theorem 7

- visualise by looking at graph

- under regularity conditions; expected value of score fn is 0

$$\mathbb{E}_\theta[S_n(\theta)] = 0$$

$$\mathbb{E}_\theta[S_n(\theta)] = \int \dots \int \left(\frac{\partial \log p(x_1, \dots, x_n; \theta)}{\partial \theta} \right) p(x_1, \dots, x_n; \theta)\, dx_1 \dots dx_n = 0$$

(*) Integral/exp. is wrt to the joint distribution of the data $p(x^n; \theta)$

- noting that $p(x^n; \theta) = \prod_{n=1}^{N} p(x_n; \theta)$

(*) We assume that the model is correct and that the data is generated from a joint distribution under a particular value of θ.

Proof:

$$\mathbb{E}_\theta[S_n(\theta)] = \int \cdots \int \frac{\partial \log p(x_1, \ldots, x_n; \theta)}{\partial \theta} p(x_1, \ldots, x_n; \theta)\, dx_1 \ldots dx_n$$

$$= \int \cdots \int \frac{\partial}{\partial p} \log p \cdot \frac{\partial p}{\partial \theta} \cdot p \, dx_1, \ldots, dx_n$$

$$= \int \cdots \int \frac{1}{p(x_1, \ldots, x_n; \theta)} \cdot \frac{\partial p(x_1, \ldots, x_n; \theta)}{\partial \theta} \cdot p(x_1, \ldots, x_n; \theta)\, dx_1 \ldots dx_n$$

$$= \int \cdots \int \frac{\partial}{\partial \theta} p(x_1, \ldots, x_n; \theta)\, dx_1 \cdots dx_n$$

(*) Regularity conditions allow us to exchange order of operators.

(*) We require that the set over which density is defined to not depend on θ.

· E.g. Normal is defined over $\mathbb{R}$; independent of params μ, σ^2 (univ.)

$X \sim \text{unif}(0, \theta) \longrightarrow$ range of random variable X depends on the parameter θ; violates condition.

(*) Notes:- If support of p depends on θ; then $\int \cdots \int$ and $\frac{\partial}{\partial \theta}$ cannot be switched.

Hence, we have from above:-

$$\mathbb{E}_\theta[S_n(\theta)] = \frac{\partial}{\partial \theta} \underbrace{\int \cdots \int p(x_1, \ldots, x_n; \theta)\, dx_1 \ldots dx_n}_{=1}$$

$$= 0. \quad \blacksquare$$

(*) Some nuances from notes:-

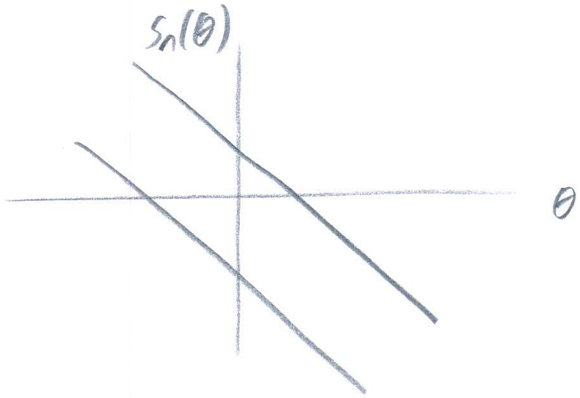
- If the expectation value is taken at the same θ as we evaluate $S_n(\theta)$; then $\mathbb{E}_\theta[S_n(\theta)] = 0$

- when the θs mismatch, i.e. $\theta_1 \neq \theta_2$, then

$\mathbb{E}_{\theta_1}[S_n(\theta_2)] \neq 0.$

(*) example 8

- let $X_1, \dots, X_n \sim N(\theta, 1)$
- $L(\theta) \propto e^{-\frac{1}{2}\sum_{n=1}^{N}(X_i-\theta)^2}$
- $\ell(\theta) = -\frac{1}{2}\sum_{n=1}^{N}(X_i-\theta)^2$
- $S_n(\theta) = \sum_{n=1}^{N}(X_i-\theta) = n(\bar{X}-\theta)$ (multiplying and dividing by n)



- note $\mathbb{E}_\theta[S_n(\theta)] = 0$

- $S_n(\theta)$ is a random function.
- $S_n(\theta) = f(\theta, x_1, \dots, x_n)$
- different datasets (i.e. different values of $\bar{X}$) will shift $S_n(\theta)$

(*) further properties of Fisher information

- As Fisher info $:= I_n = Var(S_n(\theta))$ and $\mathbb{E}_\theta[S_n(\theta)] = 0$; we have

$$I_n(\theta) = \mathbb{E}_\theta[S_n^2(\theta)]$$

lemma 9

$I(\theta) = I_1(\theta)$ notationally

- IID case
- $I_n(\theta) = nI(\theta)$ where $I(\theta)$ is F.I for $n=1$ (via definition of $\ell(\theta)$ as sum for IID.)

- ~~w~~ Helps simplify calculations (additive charact. of information)
- LW: follows since log-likelihood, and hence score, is the sum of n-independent terms.

(*) lemma 10

- under reg. conditions:-

$$I_n(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ell_n(\theta)\right]$$

(*) - Take log likelihood, 2nd deriv, negative; provides another means of computing $I_n(\theta)$, in particular when combined with lemma 9.

Proof Take $n=1$

-Note:- $\int p = 1 \Rightarrow \int p' = 0 \Rightarrow \int p'' = 0 \Rightarrow \int \frac{p''}{p} p = 0$

$$\Rightarrow \mathbb{E}\left(\frac{p''}{p}\right) = 0$$

· let $\ell = \log p$ and $S = \ell' = \frac{p'}{p}$; then $\ell'' = (p''/p) - \left(\frac{p'}{p}\right)^2$

(*) Above identities ~~from~~ rely on reg. conditions.

$$Var(S) = \mathbb{E}[S^2] - (\underbrace{\mathbb{E}[S]}_{=0})^2 = \mathbb{E}[S^2] = \mathbb{E}\left[\left(\frac{p'}{p}\right)^2\right] = \mathbb{E}\left[\left(\frac{p'}{p}\right)^2\right] - \underbrace{\mathbb{E}\left[\left(\frac{p''}{p}\right)\right]}_{=0}$$

$$= -\mathbb{E}\left[\left(\frac{p''}{p}\right) - \left(\frac{p'}{p}\right)^2\right] = -\mathbb{E}(\ell'')$$

$\blacksquare$

IW: Score fn is derivative of likelihood
Fisher information is variance of score function; or above form.

(*) extension to vectors (cosmetic)

$\underline{\theta} = (\theta_1, \dots, \theta_k)$

- $L_n(\underline{\theta})$ and $\ell_n(\underline{\theta})$ do not change
- But score $S_n(\underline{\theta})$ is affected → it is derivative of $\ell(\theta)$ for univariate case.
- score function is now a vector of functions:-

$$\underline{S}_n(\underline{\theta}) = \begin{bmatrix} \frac{\partial \ell_n(\underline{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell_n(\underline{\theta})}{\partial \theta_k} \end{bmatrix} \quad \underline{S}_n(\underline{\theta}) \in \mathbb{R}^k \quad (\underline{S}_n(\underline{\theta}))_k = \frac{\partial \ell_n(\underline{\theta})}{\partial \theta_k}$$

- Fisher information is a variance-covariance matrix of the score vector.
- $\underline{I}_n(\underline{\theta}) \in \mathbb{R}^{k \times k}$

$$\underline{I}_n(\underline{\theta}) = \operatorname{Cov}(S_n(\underline{\theta})) \qquad I_n(r,s) = -\mathbb{E}_\theta\left[\frac{\partial^2 \ell(\theta)}{\partial\theta_r \partial\theta_s}\right]$$

(*) Example

$X_1, \dots, X_n \sim \text{Bern}(p)$

likelihood:- $L(p) = p^S(1-p)^{n-S} \quad S = \sum_{n=1}^{N} X_n$

log-likelihood: $S \log p + (n-S)\log(1-p) = \ell(p)$

score fn: $S_n(p) = \frac{\partial \ell(p)}{\partial p} = \frac{S}{p} - \frac{n-S}{1-p}$

2nd deriv $\ell(p)$: $\ell''(p) = -\frac{S}{p^2} - \frac{n-S}{(1-p)^2}$

Fisher info: $I_n(p) = -\mathbb{E}[\ell''(p)]$

$$= \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p} + \frac{n}{(1-p)} = \frac{n}{p(1-p)}$$

NB: Score function is
a random function (of data)
Fisher info is just a function of the parameter

(*) Example II

- suppose $X_1, \dots, X_n \sim N(\mu, \gamma) \quad \gamma = \sigma^2$ (?)

(A1) - verify this yourself as part of review:-

$$S_n(\mu, \gamma) = \begin{bmatrix} \frac{1}{\gamma}\sum(x_i - \mu) \\ -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum(x_i - \mu)^2 \end{bmatrix} \qquad I_n(\mu, \gamma) = \begin{bmatrix} \frac{n}{\gamma} & 0 \\ 0 & \frac{n}{2\gamma^2} \end{bmatrix}$$

(*) score and F.I. cannot be computed without regularity conditions.

6. Asymptotic Normality of MLE

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$$

· we will prove this **Theorem 12**

$$\hat{\theta} \approx N\left(\theta, \frac{1}{nI(\theta)}\right) = N\left(\theta, \frac{1}{I_n(\theta)}\right)$$

- we will also prove the following for MLE estimators:-

$$\hat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \psi^*(X_i) + o_p(n^{-1/2}) \quad \text{where } \psi^*(x) = \frac{S(\theta, X_i)}{I(\theta)} \qquad (1)$$

· where ψ^* is an influence function

· any well-behaved estimator can be written as (1) for some ψ and that $var(\psi) \geq var(\psi^*)$ later

(*) influence function with smallest variance gives MLE. (later)

- w: computation of confidence intervals relies on asymptotic normality

Proof of Theorem 12:

- $\hat{\theta}_{MLE}$ (MLE est.)

- we know:-

$$l'(\hat{\theta}_{MLE}) = 0$$

· expand $l'(\hat{\theta}_{MLE})$ about true value of param:-

$$0 = l'(\hat{\theta}_{MLE}) = l'(\theta) + (\hat{\theta} - \theta) l''(\theta) + \ldots$$

$$\Rightarrow \sqrt{n}(\hat{\theta}_n - \theta) \approx \frac{\frac{1}{\sqrt{n}} l'(\theta)}{-\frac{1}{\sqrt{n}} l''(\theta)} \equiv \frac{A}{B}$$

?

(A2) - don't understand reasoning for this

$$A = \frac{1}{\sqrt{n}} l'(\theta) = \sqrt{n} \times \frac{1}{n}\sum_{i=1}^{n} S(\theta, X_i) = \sqrt{n}(\bar{S} - 0)$$

$$A \xrightarrow{d} N(0, I(\theta))$$

- And also

$$B \xrightarrow{p} -E(\ell'') = I(\theta)$$

- via Slutsky:-

$$\frac{A}{B} \xrightarrow{d} \frac{\sqrt{I(\theta)}\, Z}{I(\theta)} = \frac{Z}{\sqrt{I(\theta)}} = N\left(0, \frac{1}{I(\theta)}\right)$$

So:-

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right)$$

?

A2 - don't understand this argument; review. $B \xrightarrow{p}$ (need to move on).

■