36705 07/07/2020

(L17)

Youtube Lecture - 17/10/16

Lecture Notes 9: Asympt. Norm (cont.)

Review

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

Theorem 12 where $\hat{\theta}_n$ is MLE

$$\hat{\theta}_n = \theta + O_p\left(\frac{1}{\sqrt{n}}\right)$$

- Regularity conditions: (for asymp. Norm.)
  - Density three times diff. wrt θ
  - Key exception is uniform, where range depends on θ
- Good proof:- (from 10/10/16)
  - Not: $l(\hat{\theta}) = l(\hat{\theta}_n) = l(\hat{\theta}_{MLE})$

$$l'(\hat{\theta}) = l'(\theta) + (\hat{\theta} - \theta) l''(\theta) + \dots = 0$$

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

- Consider A, B:-

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

$$= \sqrt{n}(\bar{S} - 0) \quad (*)$$

- $S(\theta, X_i)$ - score fn based on single X_i
- $S(\theta, X_i) = \frac{\partial}{\partial \theta} \log p(X_i, \theta)$
- (*) Also, derivative of log-like over all data $(X_1, \dots, X_n)$ is sum of derivatives of log-likelihood over a single instance wrt param. (due to IID data)

$$l'(\theta) = \sum_{i=1}^{N} \frac{\partial \log p(X_i; \theta)}{\partial \theta} = \sum_{i=1}^{N} S(X_i; \theta)$$

- $\bar{S}_n = \frac{1}{n} \sum_{i=1}^{n} S(\theta, X_i)$

- so we know form

(*) : sample mean / average of IID r.v.s centred and rescaled

- think CLT

<u>via CLT:-</u>

$$A = \sqrt{n}(\bar{S} - 0) \xrightarrow{d} N(0, \underbrace{I(\theta)}_{(\cdot)})$$

-(·) $\mathbb{E}[S(\theta, X_i)] = 0$

$\text{var}[S(\theta, X_i)] = I(\theta)$

(*) key notational trick you've seen; but is explained explicitly:-

$Y \sim N(0, \sigma^2)$ - in general for mean 0, σ^2 variance Normal r.v.

$\Rightarrow Y = \sigma Z$ where $Z \sim N(0,1)$

(*) can always write a general Normal r.v. as a standard deviation times standard Normal.

Hence:

$$A = \sqrt{n}(\bar{S} - 0) \xrightarrow{d} N(0, I(\theta))$$

$$= \sqrt{I(\theta)}\, Z$$

(A1) - Tidy up notation for consistency e.g. indexing
- Esphere.

<u>Consider B:</u>

$$B = \frac{1}{n} l''(\theta) \xrightarrow{p} -\mathbb{E}\left[\frac{\partial^2 \log p(X, \theta)}{\partial \theta^2}\right] = I(\theta)$$

$$\cdot\, l'' = \frac{\partial^2}{\partial \theta^2} \underbrace{l(\theta)}_{\text{sum over } n \text{ observations}}$$

(*) classic trick - exchange $\frac{\partial^2}{\partial \theta^2}$, Σ ; so we have $l''(\theta)$ is sum over n <u>2nd derivatives</u>

(*) Hence $\frac{1}{n} l''(\theta)$ is an average ('sample mean')

(*) so $B \xrightarrow{p} I(\theta)$

Putting this together :-

$A \xrightarrow{d} \sqrt{I(\theta)}\, Z$ $\quad B \xrightarrow{p} I(\theta) = c$ (constant) (and note $\xrightarrow{p} \Rightarrow \xrightarrow{d}$)

A2 - some steps missing
- clarify
- Do we need cont. mapping here to extend Slutsky to division?

Hence by Slutsky:

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

(*) Definitions of score, Fisher info arise naturally

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

informally: $\hat{\theta} \approx N\left(0, \frac{1}{nI(\theta)}\right) = N\left(0, \frac{1}{I_n(\theta)}\right)$

(*) v. important:
- If you want variance of MLE, approximate as above, it is approx $\frac{1}{I_n(\theta)}$
- standard deviation of error i.e. approx standard error is:-

$$se(\hat{\theta}) \approx \sqrt{\frac{1}{I_n(\theta)}} = \sqrt{\frac{1}{nI(\theta)}}$$ use Slutsky

(*) we do not know θ in practice, so we use estimates, to get estimated standard error

in practice

$$\hat{se} = \widehat{se(\hat{\theta})} = \sqrt{\frac{1}{I_n(\hat{\theta})}}$$

(*) via Slutsky, we insert an estimate of 'this' - which? - as ratio is going to 1 in prob.
- use Fisher information evaluated at MLE $\hat{\theta}$ - $I_n(\hat{\theta})$
As long as Fisher info is smooth function (contin function), then by continuous mapping theorem $\hat{\theta} \xrightarrow{p} \theta$ will imply

that $\hat{se}$ is a consistent estimator of se.

(*) we can also write :-

$$\hat{se} = \sqrt{\frac{1}{I_n(\hat{\theta})}} = \sqrt{\frac{1}{nI(\hat{\theta})}}$$

(due to property of fisher info relating individual-samp obs)

(51): This is the most common method in science for computing standard errors

---

Theorem 14

- let τ be a smooth function of θ (via delta method)
- Then:- $\sqrt{n}(\tau(\hat{\theta}_n) - \tau(\theta)) \xrightarrow{d} N\left(0, \frac{(\tau'(\theta))^2}{I(\theta)}\right)$

- (*) Say interested in function of θ, $\tau(\theta)$
- The standard error of $\hat{\tau} = \tau(\hat{\theta})$ is:-

$$se = \sqrt{\frac{|\tau'(\theta)|^2}{nI(\theta)}} = \sqrt{\frac{|\tau'(\theta)|^2}{I_n(\theta)}}$$

- The estimated standard error :-

$$\hat{se}(\hat{\tau}) = \sqrt{\frac{|\tau'(\hat{\theta})|^2}{I_n(\hat{\theta})}}$$

- i.e.: can get MLE of parameter, function of that param
standard error for MLE of param; function of param

Example 15

- $X_1, \dots, X_n \sim Exp(\theta)$
- $p(x;\theta) = \theta e^{-\theta x}$ $\quad x > 0$

(*) Exponential distn family (not exp. family) comes up a lot in lifetimes of components etc.

$$L(\theta) = \prod_{i=1}^{n} p(x_i;\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{N} x_i} = \theta^n e^{-n\theta \bar{x}_n}$$

Hence $\ell(\theta) = -n\theta\bar{x}_n + n\log\theta$

$$s(\theta) = \ell'(\theta) = \frac{n}{\theta} - n\bar{x}_n$$

- MLE $\hat{\theta} = \frac{1}{\bar{x}_n}$

· $\ell''(\theta) = \frac{-n}{\theta^2}$ so $I_n(\theta) = \mathbb{E}[-\ell''(\theta)] = \frac{n}{\theta^2}$ (?)

✓ (13) Review/clarify

Hence:- $\hat{\theta} \approx N\left(\theta, \frac{\theta^2}{n}\right)$ $\hat{se} = \frac{\hat{\theta}}{\sqrt{n}}$

Example 16 - Bernoulli

- $X_1, \dots, X_n \sim Ber(p)$

MLE: $\hat{p} = \bar{X}$

Fisher info: $(n=1)$ $I(p) = \frac{1}{p(1-p)}$ so $\sqrt{n}(\hat{p}-p) \xrightarrow{d} N(0, p(1-p))$

NW: This result can be attained via application of CLT; where the above machinery (asympt. Normality) gets useful is when the MLE is a complicated non-linear function

(14) - this deduction (simple manip. of normal)
- asymptotic variance is $\frac{p(1-p)}{n}$

(*) informally; $\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$

- Asymptotic variance $\frac{p(1-p)}{n}$, estimated via $\frac{\hat{p}(1-\hat{p})}{n}$

- Estimated standard error of MLE: $\hat{se} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- Suppose we want to estimate

$$\tau = \frac{p}{(1-p)}$$

- we know MLE $\hat{\tau} = \frac{\hat{p}}{1-\hat{p}}$ so $\frac{\partial}{\partial p} \frac{p}{1-p} = \frac{1}{(1-p)^2}$

- The estimated standard error :-

$$\hat{se}(\hat{\tau}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \times \frac{1}{(1-\hat{p})^2}} = \sqrt{\frac{\hat{p}}{n(1-\hat{p})^3}}$$

- LN 9.1 typo p9 :- $\hat{se} = \sqrt{\frac{|\tau'(\hat{\theta})|^2}{I_n(\hat{\theta})}}$

- LW: optimality of MLE (claimed by Fisher)
- There exist better estimators for more complex distns; but ok in one param settings. For finite dim models with reg. conditions

LW: Any well behaved estimator $\tilde{\theta}$ satisfies

$$\tilde{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \psi(X_i) + o_p(n^{-1/2}) \quad \text{or} \quad \sqrt{n}(\tilde{\theta}-\theta) \xrightarrow{d} N(0, V(\theta))$$

But also: $V(\theta) \geq \frac{1}{I(\theta)}$

- this is the sense in which MLE is 'optimal'

LW: Proven by LeCam in 70s/80s
- see van der Vaart, asympt. stats
- We really mean MLE is optimal sense of efficient i.e. smallest possible variance
- Asymptotic variance is small compared to other estimators.
- A lot of technical apparatus to describe some of insights on regularity, efficiency.

LW: Asymptotic theory gives use useful info about approx. distn of an estimator, and optimality. But also for comparing estimators

8. Relative efficiency

(*) A natural way to compare well-behaved estimators (i.e. have asymptotic normal distributions).

(*) Done through asymptotic relative efficiency (ARE).

- For estimators W_n, V_n, if

$$\sqrt{n}(W_n - \tau(\theta)) \xrightarrow{d} N(0, \sigma_W^2)$$

$$\sqrt{n}(V_n - \tau(\theta)) \xrightarrow{d} N(0, \sigma_V^2)$$

then asymptotic rel. efficiency is:-

$$ARE(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}$$ (ratio of asymptotic variances)

Example 17

$X_1, \dots, X_n \sim \text{Poisson}(\lambda)$

- MLE of λ is $\bar{X}$

- Suppose you want to estimate $\tau = e^{-\lambda}$ (e.g. situations where you require $P(X_i = 0)$), then $\hat{\tau} = e^{-\hat{\lambda}}$ (MLE) (e.g. no. of crashes/week)

- A few methods for estimating $P(X_i = 0)$.

- Define:- $Y_i = \mathbb{I}(X_i = 0)$ $W_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ $\mathbb{E}[W_n] = \tau$

(*) How to find limiting distn of $\hat{\tau}$? (AS) review.

- From λ - variance is Fisher info.
- Take a function of λ - use Delta method (test q.)

(*) $\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, \lambda e^{-2\lambda})$ (via Delta method)

(*) Can also use CLT.

$$\sqrt{n}(W_n - \tau) \xrightarrow{d} N(0, e^{-\lambda}(1 - e^{-\lambda}))$$

(*)

$$ARE = \frac{\lambda}{e^\lambda - 1} \leq 1$$

- as MLE has smallest variance

- can trial different values of λ to search for other efficient est.

(*) If model is wrong, mle is also not very good
(i.e. not Poisson)

LW: There is a tradeoff between robustness and efficiency; how much you trust the model ; tempers claim of mle optimality.
Must include caveat that it is optimal IF the model is 'correct'

---

9. Multivariate case

- $\underline{\theta} = (\theta_1, \ldots, \theta_K)$ $|\underline{\theta}|$ is fixed

$I^{-1}(\theta)$ - Fisher inverse matrix info.

- In this case:-

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

- Approximate standard error $\hat{se}(\hat{\theta}_j) = \sqrt{\frac{I^{-1}_{j,j}(\hat{\theta})}{n}}$ (?)

- If $\tau = g(\underline{\theta})$ with $g: \mathbb{R}^K \to \mathbb{R}$; then by delta method:-

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, (g')^T I^{-1} g')$$ where g' is gradient of g eval at θ

LW: E.g. what's MLE of $\frac{\mu}{\sigma}$?

- MLE $= \frac{\hat{\mu}}{\hat{\sigma}}$

- standard error : $g(\mu, \sigma) = \frac{\mu}{\sigma}$

- compute gradient i.e. $\frac{\partial g}{\partial \mu}, \frac{\partial g}{\partial \sigma}$; multiply by inverse Fisher info; get limiting distn of funct. of params.

LW: many of parametric models covered seem to be 'nice', as they are in exponential family. Not to be confused with exponential distr (although this is an exp. family)

10. Exp. families

A density of the form:

(A6) - crossover with 10-708

$$p(x;\theta) = c(\theta)h(x)e^{\theta^T t(x)}$$

- MLE is obtained by solving:-

$$\mathbb{E}_\theta[t(X)] = \frac{1}{n}\sum_{i=1}^{n} t(X_i)$$

- In general $t(x)$ is a vector, with same dim as θ
- Note: $\frac{1}{n}\sum_{i=1}^{n} t(X_i)$ is a minimal sufficient statistic

- Fisher info: $I(\theta) = a''(\theta)$; $a(\theta) = -\log c(\theta)$

LW: favors non-parametric stats
- everything said so far depends on whether parametric model is correct

LW: A parametric model is 'never correct'
- deal with this by finding estimators that are robust (i.e. allow deviations from correctness of models)
- (other than nonparametric methods)

11. Robustness

LW: will not go through
- LW calculates ARE of (median, MLE) = 0.64

- But if data is not Normal, there are outliers; median is better → more robust

(JT): The whole subfield of robust statistics/estimators concerns tradeoff between efficiency and robustness

LW: we deal with this issue of model correctness through non-parametric methods (soon)

point estim ⟶ hypoth test ⟶ confidence int.

- we've seen much on finding good point estimators; next hypothesis testing