EC402 (20/21) - MT WK8 - ASYMPTOTICS TRILOGY: PART 3 "PUSHING LIMITS"

RAGVIR'S SPEAKING NOTES (UNOFFICIAL CONTENT)

(#1) BIG PICTURE

(a) This is the final part of my asymptotics trilogy. As usual, my focus is on $\bar{X}_N$, the ML estimator for μ.

We discuss asymptotic efficiency; that is, out of "CAN" estimators, which one is "Best"?

(b) Once we understand asymptotic efficiency, we will look at a very nice theorem on what I will call:

The "asymptotic gorgeousness of MLE"

It is exactly the result on page 59 (i.e. Section 22) of Vassilis' "Supplementary Technical Notes".

This will help tie all the basics together for you and help complete my trilogy.

(Reminder: Part (i) LLN, Part (ii) CLT, Part (iii) CRLB)

(#2) ASYMPTOTIC EFFICIENCY

- We saw in "gentle intro" that $\hat{\mu}_{MLE} := \bar{x}_N$ was an (asymptotically) unbiased estimator for μ

- Following on from this, let's restrict our attention now to (asymptotically) unbiased estimators. Even so, there are millions...

  e.g. $E(\bar{x}_N) = \mu$ for any N; $E(x_7) = \mu$ for any N let $N > 7$

- Obviously, $\bar{x}_N$ is a R.V. We've looked at its first moment properties (location), so perhaps now we can think about its second moment properties (dispersion)? Maybe $\bar{x}_N$ is preferable to some other (asymptotically) unbiased estimator because its variance is lower?

  Indeed, $\mathrm{Var}(\bar{x}_N) = \sigma^2 / N$

  and $\mathrm{Var}(x_7) = \sigma^2 > \sigma^2 / N$, $N > 7$.

  please ensure you understand this. it's item 1 from "status check" again.

IN OTHER WORDS, BASED ON THE VARIANCE CRITERION, IS $\bar{x}_N$ THE "BEST"?

WELL, ... [...think about this next part carefully...]

HOW CAN WE ANSWER THE ABOVE QUESTION WITHOUT COMPARING $\mathrm{Var}(\bar{x}_N)$ WITH ALL THE VARIANCES OF EVERY SINGLE (ASYMPTOTICALLY) UNBIASED ESTIMATOR IN THE ENTIRE UNIVERSE?!

[...do you see the problem?]

Alas, how can one ever claim that $\bar{x}_N$ is the "best" for μ?

Yes, my friends, someone figured it out! Impressive, huh?!

Enter rockstars Cramér and Rao...

(#3) CRLB

Theorem. Let $y_i \overset{\text{IID}}{\sim} F_{y_i}(\theta)$ for $\theta \in \mathbb{R}$ and $i = 1, \dots, N$. Let $U_N = h(Y)$ s.t. $E[U_N] = \theta$. Under mild regularity conditions,

$$\mathrm{Var}(U_N) \geq \frac{1}{\mathcal{I}(\theta; y)},$$

where $\mathcal{I}(\theta; y) := -E\left[\frac{\partial}{\partial \theta} s(\theta; y)\right]$.

[Note: In the IID case, we have $\mathcal{I}(\theta; y) = N\,\mathcal{I}(\theta; y_i)$.]

"Sigh, Raguir, you're really pushing me to my limits now."

Folks, I know... sorry... just let me finish the proof and I'll help make everything click into place. If you've come this far don't give up just yet! The proof is nothing more than a very quick application of the Cauchy-Schwarz inequality.

Proof. The Cauchy-Schwarz inequality implies that for RVs P, Q,

$$\left(\mathrm{Cov}(P,Q)\right)^2 \leq \mathrm{Var}(P)\,\mathrm{Var}(Q).$$

Let $P := U_N$ and $Q := s(\theta; y)$. Then,

$$\mathrm{Var}(U_N) \geqslant \frac{\left(\mathrm{Cov}(U_N, s(\theta;y))\right)^2}{\mathrm{Var}(s(\theta;y))}.$$

[Note: $\mathrm{Cov}\left(U_N, s(\theta;y)\right) = E\left[U_N \cdot s(\theta;y)\right]$ since

$E\left[s(\theta;y)\right] = \int_{\mathbb{R}^N} s(\theta;y) \cdot f_y(\theta)\, dy = \int_{\mathbb{R}^N} \frac{\partial f_y(\theta)/\partial\theta}{f_y(\theta)} \cdot f_y(\theta)\, dy = 0.$]

$\Rightarrow \mathrm{Cov}\left(U_N, s(\theta;y)\right) = \int_{\mathbb{R}^N} U_N\, s(\theta;y)\, f_y(\theta)\, dy = \int_{\mathbb{R}^N} U_N \frac{\partial f_y(\theta)}{\partial\theta}\, dy$

$= \partial E[U_N]/\partial\theta = 1$ since $E[U_N] = \theta$.

Next, notice that

$$\tilde{I}(\theta;y) := -E\left[\frac{\partial}{\partial\theta} s(\theta;y)\right] = -E\left[\frac{\partial^2}{\partial\theta^2} l(\theta,y)\right] = E\left[s(\theta;y)^2\right]$$

$$= \mathrm{Var}(s\,\theta;y)$$

because

$$0 = \frac{\partial}{\partial\theta} E\left[s(\theta;y)\right] = \frac{\partial}{\partial\theta}\int_{\mathbb{R}^N} s(\theta;y) f_y(\theta)\, dy = \int_{\mathbb{R}^N} \frac{\partial}{\partial\theta} s(\theta;y) f_y(\theta)\, dy$$

$$= \int_{\mathbb{R}^N} \left[s(\theta;y)\frac{\partial}{\partial\theta} f_y(\theta) + f_y(\theta)\frac{\partial}{\partial\theta} s(\theta;y)\right] dy$$

$$= \int_{\mathbb{R}^N} \left[s(\theta;y)^2 f_y(\theta) + \frac{\partial}{\partial\theta} s(\theta;y) f_y(\theta)\right] dy$$

$$= E\left[s(\theta;y)^2\right] + E\left[\frac{\partial}{\partial\theta} s(\theta;y)\right].$$

This completes the proof.

(#4) "IN ENGLISH, PLEASE, RAGVIR?"

- The CRLB is the theoretical lower bound for the variance of any unbiased estimator.
- It is given by $1/\mathcal{I}(\theta;y)$ where $\mathcal{I}(\theta;y)$ is the Fisher Info.
  - Of course, I still need to explain $\mathcal{I}(\theta;y)$, and indeed there is quite a lot of intuition to be gained from it:

- So what is $\mathcal{I}(\theta;y)$?

Typically, if we plot $\mathcal{L}(\theta;y)$ against θ, we get a curve with a peak at the maximum. The sharper is the peak, the more is the information about θ that is contained in the sample. This is exactly the Fisher information.

Indeed, we showed above that it has an interpretation as $\text{Var}[s(\theta;y)]$ and/or $-E\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta;y)\right]$

(if these are high, what is the implication? Maybe sketch some diagrams?)

(#5) BACK TO OUR ORIGINAL STORY...

Alas, how can one ever claim that $\bar{x}_N$ is the "best" for μ?

Yes, my friends, someone figured it out! Impressive, huh?!
Enter rockstars Cramér and Rao...

Well, now we have a way to answer this. That is, if $\mathrm{Var}(\bar{x}_N)$ is indeed equal to the CRLB, job done.
There cannot possibly be another estimator with lower variance.
Make sense?

Let's see how this works...

Recall $x_i \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$ for $i=1,\dots,N$.

$$\mathcal{L}(\mu; x) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

$$\ell(\mu; x) = C - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 \text{, for some } C.$$

$$s(\mu; x) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu)$$

$$\text{MLE} \Rightarrow s(\hat{\mu}_{MLE}; x) = 0 \Rightarrow \sum_{i=1}^{N} x_i - N\hat{\mu}_{MLE} = 0$$

$$\therefore \hat{\mu}_{MLE} = \bar{x}_N.$$

with $\mathrm{Var}(\hat{\mu}_{MLE}) = \sigma^2/N$.

Further $\frac{\partial}{\partial\mu} s(\mu; x) = -N/\sigma^2 < 0$.

Then, $\mathcal{I}(\mu; x) = -E\left[-N/\sigma^2\right] = N/\sigma^2$, and

the CRLB is thus σ^2/N.

So, is $\bar{x}_N$ "best"?

Well, if you agree that $\underset{\text{Var}(\bar{x}_N)}{\sigma^2/N} = \underset{\text{CRLB}}{\sigma^2/N}$, it most certainly is!

In practice, we don't do it this way since it can be tricky in more complex settings. Instead, we use a result called the Cramér-Rao attainment theorem, but let's not worry about that now. I'm just mentioning it in case you encounter it in future studies.

(#6) Asymptotic Optimality of MLE

I'm pasting some old ST202 notes of mine for you below. If you want additional info, see Vassilis' notes - the reference is in the first page of this 'pack'.

Theorem 1

(a) let $Y_i \overset{\text{IID}}{\sim} F(y_i|\theta)$ for $i=1,\dots,n$; and

(b) let $g(\theta)$ be a continuous function of θ.

(c) let's also assume "some weak regularity conditions"

(on the differentiability of $f(y_i|\theta)$ and on the ability to interchange integration and differentiation)

THEN ...

23:56 Mon 9 Nov

moodle.lse.ac.uk

THEOREM 1

... as $n \to \infty$,

$$\sqrt{n}\left(g(\hat{\theta}) - g(\theta)\right) \xrightarrow{d} N\left(0, \frac{[g'(\theta)]^2}{\tilde{I}(\theta|y_i)}\right)$$

What does this mean?

Note: $\hat{\theta}$ is the MLE for θ.

INTERPRETATION

(i) When n is "large":

$$g(\hat{\theta}) \overset{\text{approx.}}{\sim} N\left(g(\theta), \frac{[g'(\theta)]^2}{n\,\mathcal{I}(\theta|y_i)}\right); \text{ or}$$

(ii) When n is "large":

$$\frac{g(\hat{\theta}) - g(\theta)}{\sqrt{\dfrac{[g'(\theta)]^2}{\mathcal{I}(\theta|y)}}} \overset{\text{approx.}}{\sim} N(0,1).$$

23:56 Mon 9 Nov

moodle.lse.ac.uk

INTERPRETATION

(1) Bias disappears as $n \to \infty$ ∴ ASYMPTOTIC UNBIASEDNESS!

Variance shrinks to 0 as $n \to \infty$

⇒ Convergence in M.S. ⇒ Convergence in prob.

∴ CONSISTENCY!

(2) Attainment of CRLB

∴ ASYMPTOTIC EFFICIENCY!

(3) Distribution is Gaussian

∴ ASYMPTOTIC NORMALITY!

(#7) THANKS FOR WATCHING!

That completes my asymptotics review, folks. I actually put a lot of thought into how to present all this stuff in a logical and helpful (?) way for you all.

Even so, my notes will never be a decent substitute for proper sources. They merely reflect how I organise this info. in my own mind. Nevertheless, I hope they serve as a decent starting point for you in case you're not confident with asymptotics!

All the best

R (09/11/20)

OPTIONAL PRACTICE QUESTION (ST202 HOMEWORK LEVEL)

Coverage $\begin{cases} \text{asymptotics;} \\ \text{pivots;} \\ \text{C.I.'s;} \\ \text{result on page 59 of Vassilis' notes;} \\ \text{MLE derivations;} \\ \text{review of bias, consistency, etc.} \end{cases}$

Say $Y_i \overset{iid}{\sim} \text{Bernoulli}(\theta)$ for $i=1,\dots,N$.

(1) Find an asymptotically unbiased & efficient estimator for θ.

(2) Construct a $100(1-\alpha)\%$ C.I. for $g(\theta) = \theta/(1-\theta)$.

DETAILED SOLUTION BELOW IF YOU NEED IT...

QUESTION 2(i)

$$\cdot\ f_{Y_i}(y_i) = \begin{cases} \theta^{y_i}(1-\theta)^{1-y_i}, & y_i \in \{0,1\} \\ 0, & \text{otherwise} \end{cases}$$

$$\Rightarrow f_Y(y) \overset{(\text{iid})}{=} \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

$$= \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}$$

$$\Rightarrow \ell(\theta|y) = \sum_{i=1}^{n} y_i \log(\theta) + \left(n - \sum_{i=1}^{n} y_i\right)\log(1-\theta)$$

QUESTION 2(i)

$$\cdot\ s(\theta|y) = \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{\left(n - \sum_{i=1}^{n} y_i\right)}{(1-\theta)}$$

$$\cdot\ \text{MLE} \Rightarrow \hat{\theta} = \overline{Y}$$

Check for a maximum:

$$\frac{\partial s(\theta|y)}{\partial \theta} = -\frac{\sum_{i=1}^{n} Y_i}{\theta^2} - \frac{\left(n - \sum_{i=1}^{n} Y_i\right)}{(1-\theta)^2}$$

Since $n \geq \sum_{i=1}^{n} Y_i$, $\frac{\partial s(\theta|y)}{\partial \theta} < 0$.

QUESTION 2(i)

Given the afore-mentioned theorem, $\hat{\theta} = \bar{Y}$ is is an asymptotically unbiased and efficient estimator for θ.

QUESTION 2(ii)

- Relying again on the afore-mentioned theorem, we can say that:

$$\frac{g(\hat{\theta}) - g(\theta)}{\sqrt{\frac{[g'(\theta)]^2}{\mathcal{I}(\theta|y)}}} \overset{\text{approx.}}{\sim} N(0,1) \quad \text{when } n \text{ is "large".}$$

- let's work out all the individual elements of the above quantity given $g(\theta) = \theta/(1-\theta)$

QUESTION 2(ii)

- $g(\hat{\theta}) = \hat{\theta} / (1-\hat{\theta})$

[Recall that this is the ML estimator of $g(\theta)$ due to the invariance property of ML estimation]

- $$g'(\theta) = \theta(1-\theta)^{-2} + (1-\theta)^{-1}$$

$$= \frac{\theta}{(1-\theta)^2} + \frac{(1-\theta)}{(1-\theta)^2}$$

$$= 1/(1-\theta)^2$$

QUESTION 2(ii)

- $$\mathcal{I}(\theta|y) = -E\left[\frac{\partial s(\theta|y)}{\partial\theta}\right]$$

$$= E\left[\frac{\sum_{i=1}^{n} y_i}{\theta^2} + \frac{\left(n - \sum_{i=1}^{n} y_i\right)}{(1-\theta)^2}\right]$$

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

$$= n\left[\frac{1}{\theta} + \frac{1}{1-\theta}\right] = \frac{n}{\theta(1-\theta)}$$

QUESTION 2(ii)

- In other words, when n is "large",

$$\frac{\hat{\theta}/(1-\hat{\theta}) - \theta/(1-\theta)}{\sqrt{\dfrac{\left[1/(1-\theta)^2\right]^2}{\left[n/\theta(1-\theta)\right]}}} \overset{\text{approx.}}{\sim} N(0,1)$$

- Is this a (useable) pivotal quantity for $\theta/(1-\theta)$?

QUESTION 2(ii)

- Not quite. We do not know the denominator, which is the square root of

$$\left[\frac{1}{(1-\theta)^4} \cdot \frac{\theta(1-\theta)}{n}\right] = \frac{\theta}{n(1-\theta)^3}$$

since it contains the unknown parameter θ.

- However, we are working within an asymptotic $(n \to \infty)$ framework and we know that ML is consistent so let's use an ML estimator as an approximation.

QUESTION 2(ii)

So, let's consider

$$T = \frac{\hat{\theta}/(1-\hat{\theta}) - \theta/(1-\theta)}{\sqrt{\hat{\theta}/n(1-\hat{\theta})^3}} \overset{\text{approx.}}{\sim} N(0,1)$$

for "large" n

as a pivotal quantity which we can use to construct an asymptotically valid $100(1-\alpha)\%$ C.I. for $\theta/(1-\theta)$, the unknown odds ratio.

QUESTION 2(ii)

- We can find q_1 and q_2 s.t.

$P(q_1 < T < q_2) = 1-\alpha$ for some α.

- In particular, $\begin{cases} q_1 = z_{\alpha/2} = -z_{1-\alpha/2} \\ q_2 = z_{1-\alpha/2} \end{cases}$

(using the $N(0,1)$ distribution, which is symmetric)

QUESTION 2(ii)

$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{\hat{\theta}/(1-\hat{\theta}) - \theta/(1-\theta)}{\sqrt{\hat{\theta}/n(1-\hat{\theta})^3}} < Z_{1-\alpha/2}\right)$$

$$= P\left(\frac{-\hat{\theta}}{1-\hat{\theta}} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}}{n(1-\hat{\theta})^3}} < \frac{-\theta}{1-\theta} < \frac{-\hat{\theta}}{1-\hat{\theta}} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}}{n(1-\hat{\theta})^3}}\right)$$

$$= P\left(\frac{\hat{\theta}}{1-\hat{\theta}} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}}{n(1-\hat{\theta})^3}} > \frac{\theta}{1-\theta} > \frac{\hat{\theta}}{1-\hat{\theta}} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}}{n(1-\hat{\theta})^3}}\right)$$

QUESTION 2(ii)

Thus, a $100(1-\alpha)\%$ asymptotic C.I. for $\theta/(1-\theta)$ is:

$$\left(\frac{\bar{y}}{1-\bar{y}} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{y}}{n(1-\bar{y})^3}}\ ,\ \frac{\bar{y}}{1-\bar{y}} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{y}}{n(1-\bar{y})^3}}\right)$$