RECAP

- $L(\theta|X^n)$ captures the information in X^n about θ → and shape of $L(\theta|X^n)$ says something about the uncertainty about θ in X^n.

  As $n \to \infty$ the $L(\theta|X^n)$ becomes more and more centered around true θ.

- $L(\theta|X^n)$ is invariant to parameterizations ($\psi = g(\theta)$) <u>but</u> as we saw last lecture, for some parameterizations we may not attain the information bound (best possible variance)

- Our workhorse is the <u>score</u> $S(\theta) = \frac{\partial}{\partial\theta}\log L(\theta|X^n) = \frac{\partial}{\partial\theta}\ell(\theta|X^n)$ from which we attain the MLE $\hat\theta$ through the <u>score equation</u> $S(\theta) = 0$

- The precision in $\hat\theta$ is measured by the curvature of $\ell(\theta|X^n)$ at $\hat\theta$

$$I(\hat\theta) = -\frac{\partial^2}{\partial\theta^2}\log L(\theta|X^n)\Big|_{\hat\theta} \qquad \text{(observed information)}$$

- To work out the properties of $\hat\theta$, $\ell(\hat\theta|X^n)$ vs $\ell(\theta|X^n)$ etc we use <u>quadratic approximation</u> near $\hat\theta$.

$$\underbrace{\log\frac{\ell(\theta|X^n)}{\ell(\hat\theta|X^n)}}_{\text{normalized log likelihood}} \simeq -\frac{1}{2}I(\hat\theta)(\theta-\hat\theta)^2 \iff S(\theta) \simeq I(\hat\theta)(\hat\theta-\theta)$$

$$\Updownarrow$$

$$I(\hat\theta)^{-1/2}S(\theta) \simeq I(\hat\theta)^{\frac{1}{2}}(\hat\theta-\theta)$$

If approx. OK we call $\ell(\theta)$ <u>regular</u>

Exact if P_θ is Normal-mean model.

Using the likelihood

- Likelihood intervals

$$\left\{\theta : \frac{L(\theta|x^n)}{L(\hat\theta|x^n)} > c\right\}$$, i.e. $L(\theta|x^n)$ not too much smaller than maximum $L(\hat\theta|x^n)$

What is c?

Not easy to choose since $L(\theta|x^n)$ data dependent and we don't know sampling distribution of $\frac{L(\theta|x^n)}{L(\hat\theta|x^n)}$.

BUT if we trust the quad approx then

Wilks statistic $W = 2\log\frac{L(\hat\theta|x^n)}{L(\theta|x^n)} \sim \chi^2_1$ under null θ.

$\Rightarrow$ use cutoff $c = \exp\left\{-\frac{1}{2}\chi^2_1(1-\alpha)\right\}$

Can also use this for testing

$H_0: \theta = \theta_0$ Reject if $2\log\frac{L(\hat\theta|x^n)}{L(\theta_0|x^n)} > \chi^2_1(1-\alpha)$

- Alternative $\Rightarrow$ quad approximation to $\hat\theta$

$$\log\frac{L(\theta|x^n)}{L(\hat\theta|x^n)} \approx -\frac{1}{2}I(\hat\theta)(\theta-\hat\theta)^2$$

$LI: \left\{\theta : \frac{L(\theta|x^n)}{L(\hat\theta|x^n)} > c\right\} \Rightarrow$ CI: $\left\{\hat\theta \pm \underbrace{\sqrt{-2\log c}}_{\sqrt{\chi^2_1(1-\alpha)}} I(\hat\theta)^{-1/2}\right\}$ Wald

$\to$ quantile of normal

- When working with multiple parameter models

⇒ Fact $\log \frac{L(\theta|x^n)}{L(\hat{\theta}|x^n)}$ can get quite small just by chance

meaning cutoff for normalized log likelihood

should depend on # of parameters, o/w

we will reject $H_0: \theta = \theta_0$ too often

⇒ $2\log \underbrace{\frac{L(\hat{\theta}|x^n)}{L(\theta|x^n)}}_{W} \sim \chi^2_p$ under null $\theta = (\theta_1, \ldots, \theta_p)$

⇒ Fact Estimating more parameters always leads to more uncertainty

Demonstrated by comparing curvature of likelihood of one-parameter model and curvature of the profile likelihood

θ_1 – parameter of interest

θ_2 – other parameter

$L(\theta|x^n) = L(\theta_1, \theta_2|x^n)$ full likelihood

$\tilde{L}(\theta_1|x^n) = \max_{\theta_2} L(\theta_1, \theta_2|x^n)$ profile likelihood

$L(\theta|x^n) \Rightarrow I(\hat{\theta}) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} \Rightarrow I(\hat{\theta})^{-1} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}$

(I_{11}: one-parameter problem, assuming $\theta_2 = \hat{\theta}_2$ known)

$\tilde{L}(\theta_1|x^n) \underset{\text{show}}{\overset{\text{can}}{\Rightarrow}} \tilde{I}(\hat{\theta}_1) = (I^{11})^{-1} < I_{11}$

- Inference using the likelihood

$S(\theta|X^n) = \frac{\partial}{\partial\theta} \log L(\theta|X^n)$ $\Rightarrow$ properties

- $E_\theta S(\theta|X^n) = 0$
- $Var_\theta S(\theta|X^n) = FI(\theta)$

Where $FI(\theta) = -E_\theta\left(\frac{\partial^2}{\partial\theta^2} \log L(\theta|X^n)\right)$

(provided $p_\theta(x)$ nice enough that we can change order of derivation and integration)

↳ holds for exponential families where

$$p_\theta(x) = \exp\{\eta(\theta)T(x) - A(\theta) + c(x)\}$$

↳ Where we also have that observed Fisher information @ $\hat\theta$ is equal to expected Fisher information @ $\hat\theta$

$$I(\hat\theta) = FI(\hat\theta)$$

and for canonical parameters where

$$p_\theta(x) = \exp\{\theta T(x) - A(\theta) + c(x)\}$$

the entire functions coincide

$$I(\theta) = FI(\theta) \quad \forall\theta.$$

- Using FI

Cramér-Rao Lower Bound

For estimator $T(X^n)$ where $E_\theta T(X^n) = g(\theta)$, we have that $Var_\theta T(X^n) \geq \frac{g'(\theta)^2}{nI(\theta)}$

$\Rightarrow$ bound can be attained for exponential family (for some parameterization $g(\theta)$ that is, not all possible parameterizations)

- Other variance bounds

UMVUEs – leaving MLEs and focusing on risk measure = MSE

Lehmann-Scheffé Thm

If $T(X^n)$ is a complete, sufficient statistic

complete: $E\,h(T) = 0$ possible only if $h(T) = 0$

sufficient: $X^n | T(X^n)$ indep of θ

then an unbiased estimator $S(X^n)$ for $g(\theta)$ can be turned into a UMVUE $T^*(X^n) = E(S(X^n) | T(X^n))$

Note, $Var\, T^*(X^n)$ may not attain CR bound but it will under exponential families for parameterizations where

$$P_\theta(x^n) = \exp\{\eta(\theta) T^*(x^n) - A(\theta) + c(x)\}$$

~~SO~~ for some parameterization in exponential family

the UMVUE attains the bound

Is MLE = UMVUE?

- No, but coincide sometimes. For exponential families, both estimators involve functions of $T(X)$ [$p_\theta(x) = \exp\{\eta(\theta)T(x) - A(\theta) + c(x)\}$]
- For large n, tend to coincide.
- For small n, MLE can be biased.
- MLEs tend to be easier to "find" (existence)

MLE – asymptotics

We may not understand small sample behaviour of MLEs

– or small sample behaviour may be poor

BUT for large enough n MLEs are "best"

Why?

Consistency $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$

Asymptotic normality Have $(n\bar{I}(\theta))^{-1/2} S(\theta) \xrightarrow{d} N(0,1)$

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

From CLT & Slutsky

Asymptotic efficiency

Note, $FI(\theta)$ depends on <u>true</u> value of θ

→ not so practical

⇒ we tend to use $FI(\hat{\theta})$ or $I(\hat{\theta})$ for testing, CI.

(Remember, for exponential family (canonical form $FI(\theta) = I(\theta)$ $\forall \theta$

and for exponential family $FI(\hat{\theta}) = I(\hat{\theta})$.

If not, $I(\hat{\theta})$ has better properties (coverage)

Practical use

- AN of $\hat{\theta}$ ⇒ inference on θ → CI
  → testing

- LRT ⇒ inference on nested models

$$W = 2\log \frac{L(\hat{\theta}|x^n)}{L(\hat{\theta}_{H_0}|x^n)} \sim \chi^2_{\Delta p} \text{ under } H_0$$

$\hat{\theta}_{H_0}$: max w. some parameter restrictions

where Δp is the number of restricted parameters under H_0

$H_0 : \theta_{H_0}$, e.g. $\begin{cases} \theta_1 = \theta_{10} \\ \theta_2 = ? \end{cases}$

(proof, from AN of $\hat{\theta}$ and conditional distribution of $\hat{\theta}_2 | \theta_1 = \theta_{10}$