Long-run freq. properties of the score.

- Showed $E_\theta S(\theta) = 0$

  and $\mathrm{Var}_\theta S(\theta) = \mathcal{I}(\theta) = -E_\theta\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta)\right)$

  Expected Fisher information

  (How difficult estimation problem is in general.)

- $I(\theta) = -\frac{\partial^2}{\partial\theta^2}\log L(\theta|\underline{x})$ observed Fisher information

  says something about uncertainty about θ given the sample

- For exponential families, $I(\hat\theta) = \mathcal{I}(\hat\theta)$ and

  for canonical θ, $I(\theta) = \mathcal{I}(\theta)$ (whole function)

  $\Rightarrow$ Many possible estimates for $\mathrm{Var}_\theta S(\theta)$

  $\rightarrow$ inference

Cramér-Rao bound

Assume regularity conditions like

holds for "nice" distributions

(I) support of $p_\theta(x)$ does not depend on θ, and $\frac{\partial}{\partial\theta}\log p_\theta(x)$ exists and is finite.

(II) that for statistic $T(X)$ we have $E|T(X)| < \infty$ and integration / derivation can be interchanged

CR bound

Take $T(X)$ to be an unbiased estimator for $g(\theta)$ such that $E_\theta T(X) = g(\theta)$, and assume $Var_\theta[T(X)] < \infty \; \forall \theta$.

Then, provided I & II hold,

$$Var_\theta T(X) \geq \frac{(g'(\theta))^2}{I(\theta)}$$

Proof:

Take $S(\theta) = \frac{\partial}{\partial \theta} \log L(\theta | \underline{x}) = S(\theta|X)$

Now $E_\theta S(\theta) = 0$, $Var_\theta S(\theta) = I(\theta)$

Take covariance $Cov_\theta(S(\theta|X), T(X)) = E_\theta(S(\theta|X)T(X)) - \cancel{E_\theta S(\theta|X) E_\theta T(X)}$ $= 0$

Now $|Cov_\theta(S(\theta|X), T(X))|$ maximized for $T(X)$ and $S(\theta|X)$ linearly dependent, i.e. when $|Correlation_\theta(S(\theta|X), T(X))| = 1$

For other $T(X)$ we have

$$|Corr_\theta(S(\theta|X), T(X))| = \frac{|Cov_\theta(S(\theta|X), T(X))|}{\sqrt{Var_\theta(S(\theta|X)) Var_\theta[T(X)]}} \leq 1$$

$$\Rightarrow \frac{Cov_\theta(S(\theta|X), T(X))^2}{Var_\theta(S(\theta|X))} \leq Var_\theta T(X)$$

$$\Rightarrow Var_\theta T(X) \geq \frac{(Cov_\theta(S(\theta|X), T(X))^2}{I(\theta)}$$

Now, $\mathrm{Cov}_\theta(S(\theta|X), T(X)) = E_\theta(S(\theta|X)T(X))$ here

$$= \int S(\theta|x) T(x) p_\theta(x)\, dx$$

$$= \int \frac{\frac{\partial}{\partial\theta} p_\theta(x)}{p_\theta(x)} T(x)\, p_\theta(x)\, dx$$

$$\overset{\text{by (II)}}{=} \frac{\partial}{\partial\theta} \int p_\theta(x) T(x)\, dx = \frac{\partial}{\partial\theta} g(\theta) = g'(\theta)$$

And so $\mathrm{Var}_\theta T(x) \geq \dfrac{(g'(\theta))^2}{\mathcal{I}(\theta)}$

We achieve the boundary if $T(x)$ perfectly linearly related to the score.

$\Rightarrow$ $S(\theta) = \frac{\partial}{\partial\theta} \log p_\theta(x) = u(\theta) T(x) + v(\theta)$ for some functions u and v

$\Rightarrow$ $p_\theta(x) = \exp\{U(\theta)T(x) + V(\theta)\} \times$ normalizing constant

i.e. $p_\theta \sim$ exponential family.

[Some work to show that CR bound can be achieved for larger class of models, but if we assume regularity conditions on p_θ (smoothness, differentiability etc $\Rightarrow$ only exponential family remains]

Multiparameter case

$\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_p)$, $S(\underset{\sim}{\theta}) = \nabla_\theta \log L(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \log L(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \log L(\theta) \end{pmatrix}$

Observed FI has elements $I_{ij}(\theta) = \frac{-\partial^2}{\partial \theta_i \partial \theta_j} \log L(\theta)$

Expected FI ..... $E(I_{ij}(\theta))$

(Can still show $\begin{cases} E_\theta S(\underset{\sim}{\theta}) = \underset{\sim}{0} \\ Var_\theta S(\underset{\sim}{\theta}) = FI(\underset{\sim}{\theta}) \end{cases}$

$p \times p$ matrix

and <u>CR-bound</u> ; now have that for

scalar $T(X)$ where $E_\theta T(X) = g(\underset{\sim}{\theta})$ (some (scalar) function of $\underset{\sim}{\theta}$)

$\Rightarrow Var_\theta(T(X)) \geq (\nabla_\theta g(\theta))' FI(\theta)^{-1} (\nabla_\theta g(\theta))$

Example, special case $g(\underset{\sim}{\theta}) = (\theta_1, 0 \quad 0)$

$\Rightarrow Var_\theta(\hat{\theta}_1) \geq (FI(\theta))^{-1}_{11}$

Note $(FI(\theta))^{-1}_{11} \neq (FI(\theta_1))^{-1}$

FI for p parameter problem

FI for one-parameter problem

In fact $\frac{1}{\mathcal{I}(\theta_1)} \leq (\mathcal{I}(\underline{\theta})^{-1})_{11}$

meaning it is easier to estimate one parameter and estimation precision of θ_1 affected by estimation of the parameters to estimate.

Into asymptotics

We have $E_\theta S(\theta) = 0$ $\quad Var_\theta S(\theta) = \mathcal{I}(\theta)$

and quadratic approximation

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

$$\Rightarrow S(\theta) \approx I(\hat\theta)(\hat\theta - \theta)$$

$$\Rightarrow I(\hat\theta)^{1/2} S(\theta) \approx I(\hat\theta)^{1/2}(\hat\theta - \theta)$$

Sampling distribution of this?

Sampling distribution of this

Above, $S(\theta) = \sum_i \underbrace{\frac{\partial}{\partial \theta} \log f_\theta(x_i)}_{y_i} = \sum y_i$ for iid r.v y_i

Know $E_\theta(y_i) = 0$, $E_\theta(\sum y_i) = 0$

Know $Var_\theta(y_i) = FI(\theta)$, $Var_\theta(\sum y_i) = n\,Var(y_i) = n\,FI(\theta)$

where $FI(\theta) = E_\theta\left[-\frac{\partial^2}{\partial \theta^2} \log p_\theta(X)\right]$

Central Limit Theorem

For iid y_i with mean μ and variance σ^2

define $S_n = \sum_{i=1}^{n} y_i$.

Then $\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{y} - \mu)}{\sigma} \to^d N(0,1)$

Use here $\Rightarrow$ $\frac{S(\theta)}{\sqrt{n}} \to^d N(0, FI(\theta))$

or $(n\,FI(\theta))^{-1/2} S(\theta) \to^d N(0,1)$

but we have

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

Can we just replace $I(\theta)$ by $FI(\theta)$?

Well, we have

by WLLN, $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log p_\theta(x_i) \to E_\theta\left(\frac{\partial^2}{\partial\theta^2} \log p_\theta(\xi)\right)$

and so $\frac{I(\theta)}{n} \to^p \bar{I}(\theta) \quad \forall\theta$

Moreover, Slutsky's theorem states that if

$X_n \to^d X$ and $A_n \to^p a$ (a constant)

then $A_n X_n \to^d aX$

Here $I(\theta)^{-1/2} S(\theta) = \underbrace{(n\bar{I}(\theta))^{-1/2} S(\theta)}_{\to^d N(0,1)} \cdot \underbrace{\left(\frac{I(\theta)}{n\bar{I}(\theta)}\right)^{1/2}}_{\to^p 1}$

$\Rightarrow$ results on $\hat{\theta}$ that follow

- $\sqrt{n\bar{I}(\theta_0)}\,(\hat{\theta}-\theta_0) \to^d N(0,1)$
- $\sqrt{n\bar{I}(\hat{\theta})}\,(\hat{\theta}-\theta_0) \to^d N(0,1)$
- $\sqrt{I(\theta_0)}\,(\hat{\theta}-\theta_0) \to^d N(0,1)$
- $\sqrt{I(\hat{\theta})}\,(\hat{\theta}-\theta_0) \to^d N(0,1)$

all can be used for inference

$\Downarrow$

but computable / practical use

$\hat{\theta} \sim N\left(\theta, \frac{1}{\sqrt{n\bar{I}(\hat{\theta})}}\right)$

or

$\hat{\theta} \sim N\left(\theta, \frac{1}{\sqrt{I(\hat{\theta})}}\right)$