Feynman Liang
STAT 210A
Lecture #24
11/15/2016

1 Review

$X_1, \dots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}(x)$, $\theta_0 \in \Theta \subset \mathbb{R}^d$.
Score $\nabla_\theta \ell(\theta_0, X)$ satisfies

- $\mathbb{E}_{\theta_0}[\nabla_\theta \ell(\theta_0, X)] = 0$
- $\text{Var}_{\theta_0}[\nabla_\theta \ell(\theta_0, X)] = J_1(\theta_0)$

CLT $\implies n^{-1/2} \nabla_\theta \ell(\theta_0; X) \overset{p_{\theta_0}}{\Rightarrow} N_d(0, J_1(\theta_0))$

1.1 Regularity assumptions

- p_θ "smooth" in θ
- $\hat{\theta} \xrightarrow{\text{p}} \theta_0$
- $\theta_0 \in \Theta^0$

1.2 Asymptotic distribution of MLE

$$\begin{aligned}
0 &= \nabla \ell(\hat{\theta}; X) && (1.1)\\
&\approx \nabla \ell(\theta_0; X) + \nabla^2 \ell(\theta_0; X)(\hat{\theta} - \theta_0) && (1.2)\\
\sqrt{n}(\hat{\theta} - \theta_0) &\approx \left[\underbrace{-\frac{1}{n}\nabla^2 \ell(\theta_0; X)}_{\xrightarrow{\text{p}} J_1(\theta_0)}\right]^{-1} \left[\underbrace{\frac{1}{\sqrt{n}}\nabla \ell(\theta_0; X)}_{\Rightarrow N(0, J_1(\theta_0))}\right] && (1.3)\\
&\Rightarrow N(0, J_1(\theta_0)^{-1}) && (1.4)
\end{aligned}$$

2 Wald-type conf regions/tests

Definition 2.1 (Matrix square root). If $A \succeq 0$, (symmetric) $A = U \underbrace{\Lambda}_{\text{diag}} U'$, then

$$\begin{aligned}
A^{1/2} &= U\Lambda^{1/2}U' && (2.1)\\
(A^{1/2})^2 &= U\Lambda^{1/2}U'U\Lambda^{1/2}U' = U\Lambda U' = A && (2.2)
\end{aligned}$$

If $n^{-1}\hat{J} \xrightarrow{p} J_1(\theta_0) \succ 0$, then $\hat{J}^{1/2}(\hat{\theta} - \theta_0) \Rightarrow N_d(0, I_d)$.

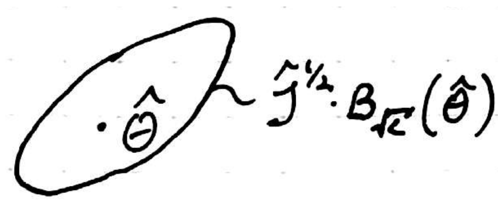
$$\|\hat{J}^{1/2}(\hat{\theta} - \theta_0)\|_2^2 \Rightarrow \chi_d^2 \tag{2.3}$$

Take $c = \chi_d^2(\alpha)$ (level-α cutoff)

$$\|\hat{J}^{1/2}(\hat{\theta} - \theta_0)\|_2^2 \leq c \tag{2.4}$$

$$\iff \hat{J}^{1/2}(\hat{\theta} - \theta_0) \in B(0) \tag{2.5}$$

$$\iff \theta_0 \in \hat{\theta} + \hat{J}^{-1/2} B_{\sqrt{c}}(0) \tag{2.6}$$

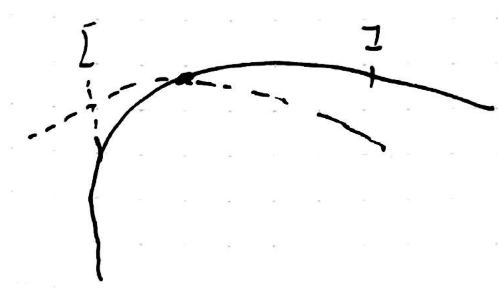


Popular choices for $\hat{J}$:

- $\hat{J} = nJ_1(\hat{\theta})$
- $\hat{J} = -\nabla^2 \ell(\hat{\theta}; X)$ ("observed Fisher information")

Observed Fisher information more often preferred because:

- Convenience — no expectations/probabilities, if using 2nd order solver for MLE then already computing Hessian
- For finite n — sometimes curvature of ℓ unusually high/low. When unusually high, more likely to have a more precise estimate. Correspondingly, the interval is smaller. Similar for when unusually low.



Example 2.2. $X \sim \text{Binom}(n, \theta)$. Observe $X = 1$.
$\hat{\theta} = n^{-1}$, $J(\theta) = n\theta(1-\theta) \implies J(\hat{\theta}) = 1 - n^{-1}$.
$\alpha = 0.05$.

$$\begin{aligned}\text{CI} &= \hat{\theta} \pm 1.96/\sqrt{J(\hat{\theta})} &(2.7)\\ &= n^{-1} \pm 1.96/\sqrt{1-n^{-1}} &(2.8)\end{aligned}$$

This CI extends outside $\Theta = [0, 1]$.

One drawback of the Wald interval is:

$$\ell(\theta; X) \to -\infty \text{ as } \theta \downarrow 0 \tag{2.9}$$

Example 2.3 (Logistic regression)**.**

$$P(y_i = 1 \mid X_i = x) = \frac{e^{\beta' x}}{1 + e^{\beta' x}} \quad x \in \mathbb{R}^d \tag{2.10}$$

1) Solve numerically for $\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^d} \ell(\beta)$

2) Find $\hat{J}^{-1} = (-\nabla^2 \ell(\hat{\beta}))^{-1}$

Can make confidence region for β or interval for coordinate β_i

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) &\Rightarrow N(0, J_1^{-1}) &(2.11)\\ \sqrt{n}(\hat{\beta}_i - \beta_i) &\Rightarrow N(0, (J_1^{-1})_{ii}) &(2.12)\end{aligned}$$

CI: $\hat{\beta}_i \pm \sqrt{(\hat{J}^{-1})_{ii}} z_{\alpha/2}$

3 Score test

$$n^{-1/2} \nabla \ell(\theta_0; X) \Rightarrow N_d(0, J_1(\theta_0)) \tag{3.1}$$

Reject $H_0 : \theta = \theta_0$ if

$$\| n^{-1/2} J_1(\theta_0)^{-1/2} \nabla \ell(\theta_0; X) \|^2 > \chi_d^2(\alpha) \tag{3.2}$$

where TODO: ?? $J_1(\theta_0)^{-1/2} = \nabla \ell(\theta_0; X)' [J(\theta_0)]^{-1} \nabla \ell(\theta_0; X)'$

Benefits:

- No need to estimate $J_1(\theta_0)$
- Don't need $\hat{\theta} \xrightarrow{p} \theta_0$
- Don't need to compute compute $\hat{\theta}$
- Often test statistic is convenient/easy to compute

Example 3.1 (Exponential Family)**.** $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\eta(x) = e^{\eta' T(X) - A(\eta)} h(x)$

$\nabla \ell(\eta; X) = \sum_i T(x_i) - \nabla A(\eta)$

$J(\eta) = n \nabla^2 A(\eta)$

Reject if $\left(\sum_i T(x_i) - \nabla A(\eta)\right)' (n \nabla^2 A(\eta))^{-1} (\cdots) > \chi_d^2(\alpha)$

Example 3.2 (Pearson's χ^2 test). $(N_1, \dots, N_d) \sim \text{Multinom}(n, (\pi_1, \dots, \pi_d))$
$H_0 : \pi = \pi_0 = (\pi_{0,1}, \dots, \pi_{0,d})$
(Note: constraint $\sum_i \pi_I = 1, \sum_i N_i = n$)
Test stat is $\sum_{i=1}^{d} \frac{(N_i - n\pi_{0,i})^2}{n\pi_{0,i}} \overset{H_0}{\Rightarrow} \chi^2_{d-1}$

4 Generalized likelihood ratio test (LRT)

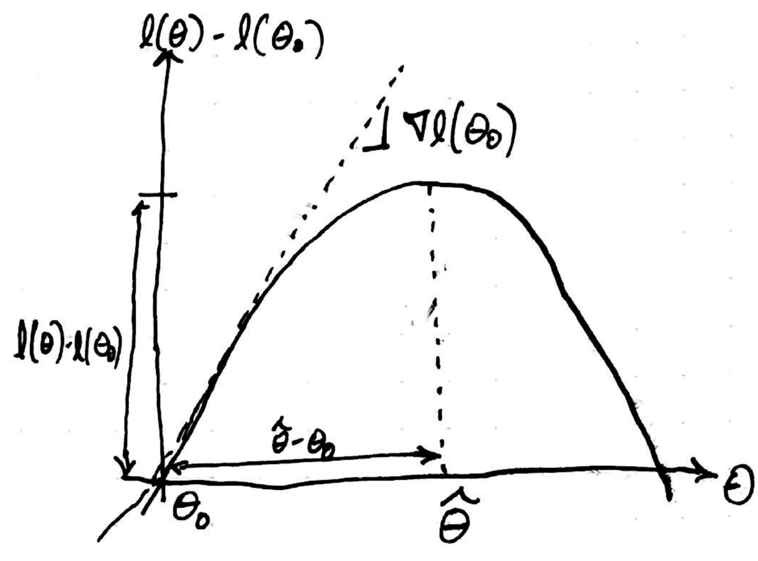
Expand around $\hat{\theta}$:

$$\ell(\theta_0 - \ell(\hat{\theta}) \approx \nabla \ell(\hat{\theta})'(\theta_0 - \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})'\nabla^2 \ell(\hat{\theta})(\theta_0 - \hat{\theta}) \tag{4.1}$$

$$2(\ell(\hat{\theta}) - \ell(\theta_0)) = (\hat{\theta} - \theta_0)[-\nabla^2 \ell(\hat{\theta})](\hat{\theta} - \theta_0) \tag{4.2}$$

$$= \underbrace{[\sqrt{n}(\hat{\theta} - \theta_0)]'}_{\Rightarrow N(0, J_1(\theta_0)^{-1})} \underbrace{\left[-\frac{1}{n}\nabla^2 \ell(\hat{\theta})\right]}_{\overset{\mathrm{P}}{\to} J_1(\theta_0)} \underbrace{[\sqrt{n}(\hat{\theta} - \theta_0)]}_{\Rightarrow N(0, J_1(\theta_0)^{-1})} \tag{4.3}$$

$$\Rightarrow \chi^2_d \tag{4.4}$$



Three different test statistics all tell the same thing asymptotically, not equivalent for finite n.

If Θ_0 is d_0-dim manifold inside Θ, $\theta_0 \in \text{relint}(\Theta_0) \cap \Theta^\circ$, then $2(\ell(\hat{\theta}) - \ell(\hat{\theta}_0)) \Rightarrow \chi^2_{d-d_0}$.

Basic idea: near θ_0 we have

$$\ell(\theta) \approx \ell(\hat{\theta}) - \frac{1}{2}\|J(\theta_0)^{1/2}(\theta - \hat{\theta})\|^2 \tag{4.5}$$

Assume param s.t. $J(\theta_0) = I_d$, then

$$\hat{\theta}_0 = \underset{\theta \in \Theta}{\arg\min} \|\theta - \hat{\theta}\|^2 = \prod_{Theta_0} (\hat{\theta}) \tag{4.6}$$

TODO: Fig 24.4

GRLT $\approx \|\hat{\theta} - \mathrm{Proj}_{\Theta_0}(\hat{\theta})\|^2 \approx \chi^2_{d-d_0}$