

11/16/2023

Outline

- 1) Wald test
- 2) Score test
- 3) Generalized likelihood ratio test
- 4) Asymptotic Relative Efficiency

Likelihood-Based Inference

Setting $X_1, \dots, X_n \stackrel{iid}{\sim} p_\theta(x)$, $p_\theta(x)$ "smooth" in Θ

Assume $\mathbb{E}_\theta \nabla \ell_1(\theta; X_i) = 0$,

$$\text{Var}_\theta[\nabla \ell_1(\theta; X_i)] = -\mathbb{E}_\theta \nabla^2 \ell_1(\theta; X_i) = J_1(\theta) \succ 0,$$

$$\hat{\theta}_{MLE} \xrightarrow{P_\theta} \theta \quad (\text{consistent})$$

Then, if $\theta = \theta_0$:

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) \Rightarrow N_d(0, J_1(\theta_0))$$

$$\frac{1}{n} \nabla^2 \ell_n(\theta_0; X) \xrightarrow{P} J_1(\theta_0)$$

Used $0 = \nabla \ell_n(\hat{\theta}_n) \approx \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\theta_0) (\hat{\theta}_n - \theta_0)$

to get $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, J_1(\theta_0)^{-1})$

Can use this for inference on θ_0 !

Wald-Type Confidence Regions

Assume we have some estimator $\hat{J}_n \succ 0$ s. t.

$\frac{1}{n} \hat{J}_n \xrightarrow{P} J_1(\theta_0) \succ 0$. Then we can plug in:

If $\sqrt{n} (\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, J_1(\theta_0)^{-1})$

then $(J_1(\theta_0))^{1/2} \sqrt{n} (\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, I_d)$

so $\hat{J}_n^{1/2} (\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, I_d)$ (Slutsky)

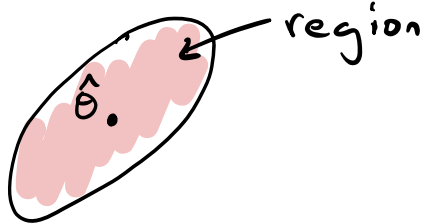
Leads to test of $H_0: \theta = \theta_0$:

$\|\hat{J}_n^{1/2} (\hat{\theta}_n - \theta_0)\|^2 \Rightarrow \chi_d^2$ (Reject if large)

So, $\mathbb{P}_{\theta_0}(\|\hat{J}_n^{1/2} (\hat{\theta}_n - \theta_0)\| \geq \underbrace{\chi_d^2(\alpha)}_{1-\alpha \text{ quantile}}) \rightarrow \alpha$

Note we reject θ_0 iff $\|\hat{J}_n^{1/2} (\hat{\theta}_n - \theta_0)\|^2 > \chi_d^2(\alpha)$

\Leftrightarrow reject θ_0 iff $\theta_0 \notin \underbrace{\hat{\theta}_n + \hat{J}_n^{-1/2} B(0)}_{\chi_d^2(\alpha)}$
confidence ellipsoid



More info \Leftrightarrow smaller ellipse (shrinks like $1/\sqrt{n}$)

Options for \hat{J}_n :

1) Most obvious is to "plug in" the MLE:

$$\hat{J}_n = J_n(\hat{\theta}_n) \quad (\text{MLE for } J_n(\theta))$$

$$= \text{Var}_{\theta}(\nabla \ell_n(\theta; X)) \Big|_{\theta = \hat{\theta}_n}$$

$$\text{(NB)} \neq \text{Var}_{\hat{\theta}_n}(\nabla \ell_n(\hat{\theta}_n(X); X)) = 0$$

$$\text{Or, } \hat{J}_n = -\mathbb{E}_{\theta} \nabla^2 \ell_n(\theta) \Big|_{\theta = \hat{\theta}_n}$$

2) Observed Fisher info:

$$\hat{J}_n = -\nabla^2 \ell_n(\hat{\theta}_n; X)$$

Remarks:

- Both have $\frac{1}{n} \hat{J}_n \xrightarrow{P} J_1(\theta_0)$ in "nice" iid sampling setting
- Both make sense outside of iid setting
- Heuristically, plug-in measures info about θ in "typical" data set but obs. info. measures info about θ in "this" data set

Wald interval for θ_j :

$$\text{If } \hat{\theta}_n \approx N_d(\theta_0, J_n(\theta_0)^{-1})$$

$$\text{then } \hat{\theta}_{n,j} \approx N_d(\theta_{0,j}, \underbrace{(J_n(\theta_0)^{-1})_{jj}}_{\text{s.e.}(\hat{\theta}_{n,j})^2})$$

Leads to univariate interval: $\text{s.e.}(\hat{\theta}_{n,j})^2$

$$C_j = \hat{\theta}_{n,j} \pm \text{s.e.}(\hat{\theta}_{n,j}) \cdot z_{\alpha/2}$$

$$= \hat{\theta}_{n,j} \pm \sqrt{(\hat{J}_n^{-1})_{jj}} \cdot z_{\alpha/2}$$

glm function in R uses these intervals /
 p -values, with $\hat{J}_n = -\nabla^2 \ell(\hat{\theta}_n)$

Conf. ellipsoid for $\theta_{0,s} = (\theta_{0,j})_{j \in s} : (|s|=k)$

$$\hat{\theta}_{n,s} \approx N_k(\theta_{0,s}, (J_n(\theta_0)^{-1})_{ss})$$

$$\leadsto C_s = \hat{\theta}_{n,s} + ((\hat{J}_n^{-1})_{ss})^{1/2} B_{\chi_k(\alpha)}(0)$$

More generally, if $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, \Sigma(\theta_0))$

and $\frac{1}{n} \hat{\Sigma}_n \xrightarrow{P_{\theta_0}} \Sigma(\theta_0)$ ($\hat{\theta}_n$ not nec. MLE)

then we can do the same things

Ex Generalized linear model with fixed x

$x_1, \dots, x_n \in \mathbb{R}^d$ fixed

$$Y_i \stackrel{\text{ind.}}{\sim} p_{\eta_i}(y) = e^{\eta_i y_i - A(\eta_i)} h(y_i)$$

$$\eta_i = \beta' x_i \quad (\text{canonical form})$$

$$\text{Let } \mu_i(\beta) = \mathbb{E}_{\beta} Y_i \quad (= \mu(\eta_i(\beta)))$$

(more general: $f(\mu_i) = \beta' x_i$ for link fun f)

Most common examples:

Logistic regression: $Y_i \stackrel{\text{ind.}}{\sim} \text{Bern}\left(\frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}}\right)$

Poisson log-linear model: $Y_i \stackrel{\text{ind.}}{\sim} \text{Pois}(e^{x_i' \beta})$

$$l_n(\beta; Y) = \sum_i (x_i' \beta) y_i - A(x_i' \beta) - \log h(y_i)$$

$$\nabla l_n(\beta; Y) = \sum_i y_i x_i - \dot{A}(x_i' \beta) \cdot x_i$$

$$= \sum_i (y_i - \mu_i(\beta)) x_i$$

$$-\nabla^2 l_n(\beta; Y) = \sum_i \ddot{A}(x_i' \beta) \cdot x_i x_i'$$

$$= \sum_i \text{Var}_{\beta}(y_i) \cdot x_i x_i'$$

$$= \text{Var}_{\beta}(\nabla l_n(\beta; Y))$$

(Not random)

$$(-\nabla^2 \ell_n(\beta))^{-1/2} \nabla \ell_n(\beta) \sim (0, \mathbb{I}_d) \quad \text{in finite samples}$$

$$\Rightarrow^* N_d(0, \mathbb{I}_d)$$

* Under regularity cond. on $X = \begin{pmatrix} -x_1' & - \\ & \vdots & \\ -x_n' & - \end{pmatrix}$

Taylor expansion of ℓ_n leads to

$$\hat{J}_n^{-1/2} (\hat{\beta}_n - \beta) \Rightarrow N_d(0, \mathbb{I}_d)$$

Advantages of Wald test:

- 1) Easy to invert, simple conf. regions
- 2) Asymptotically correct

Disadvantages:

- 1) Have to compute MLE
- 2) Depends on parameterization
- 3) Relies on two approximations:
 $\nabla \ell_n \approx \text{Normal}$ and $\ell_n \approx \text{quadratic}$
- 4) Need MLE to be consistent
- 5) Confidence interval/ellipsoid might go outside (H) !

Score Test

Test $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$

We can bypass quadratic approximation entirely by using score as test stat

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) \stackrel{P_{\theta_0}}{\Rightarrow} N_d(0, J_1(\theta_0))$$

$$(\text{or } J_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0; X) \stackrel{P_{\theta_0}}{\Rightarrow} N_d(0, I_d))$$

So, we can reject $H_0: \theta = \theta_0$ if

$$\| J_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0; X) \|_2^2 \geq \chi_d^2(\alpha)$$

$$d=1: \frac{\dot{\ell}_n(\theta_0)}{\sqrt{J_n(\theta_0)}} \Rightarrow N(0, 1),$$

can do 1-sided tests

Remarks

- No quadratic approx., no MLE
- No need to estimate Fisher info at θ_0

Can be generalized to case with nuisance params
Typically estimate via MLE on Θ_0

Score test is invariant to reparameterization*

Assume $d=1$, $\theta = g(\xi)$, $\dot{g}(\xi) > 0 \forall \xi$

$$q_{\xi}(x) = p_{g(\xi)}(x)$$

$$\begin{aligned}\dot{\ell}^{(\xi)}(\xi; X) &= \frac{d}{d\xi} \log p_{g(\xi)}(x) \\ &= \dot{\ell}^{(\theta)}(g(\xi); X) \cdot \dot{g}(\xi)\end{aligned}$$

$$J^{(\xi)}(\xi) = J^{(\theta)}(g(\xi)) \cdot \dot{g}(\xi)^2$$

$$s_0 \quad \frac{\dot{\ell}^{(\xi)}(s_0; X)}{\sqrt{J^{(\xi)}(s_0)}} \stackrel{\text{a.s.}}{=} \frac{\dot{\ell}^{(\theta)}(\theta_0; X)}{\sqrt{J^{(\theta)}(\theta_0)}}$$

$$\text{if } \theta_0 = g(s_0)$$

Ex s -parameter exp. fam:

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} e^{\eta^T T(x) - A(\eta)} h(x)$$

$$\bigvee \ell_n(\eta; X) = \sum T(x_i) - n\mu(\eta)$$

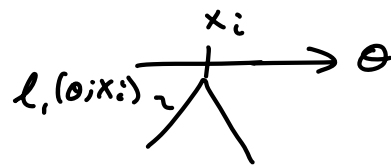
$$\| \Sigma_n(\eta_0)^{-1/2} (\sum T(x_i) - n\mu(\eta_0)) \|^2 \Rightarrow \chi_d^2$$

$$\frac{\sum T(x_i) - n\mu(\eta_0)}{\sqrt{n \text{Var}_{\eta_0}(T(x_i))}} \stackrel{P_{\eta_0}}{\Rightarrow} N(0, 1)$$

Ex $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Laplace}(\theta) = \frac{1}{2} e^{-|x-\theta|}$

Test $H_0: \theta \leq 0$ vs $H_1: \theta > 0$ (right-tailed)

$$l_n(\theta; x) = - \sum_{i=1}^n |x_i - \theta| - n \log(2)$$



$$\dot{l}_n(\theta; x) = \sum_{i=1}^n \text{sgn}(x_i - \theta)$$

$$\text{sgn}(z) = \begin{cases} +1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$$

$$\dot{l}_n(0; x) = \sum_{i=1}^n \text{sgn}(x_i)$$

$$= \#\{i: x_i > 0\} - \#\{i: x_i < 0\}$$

$$= 2 \underbrace{\#\{i: x_i > 0\}}_{Y \sim \text{Binom}(n, 1/2)} - n$$

sign test

Note: this test is \approx the exact NP LRT for

$H_0: \theta = 0$ vs $H_1: \theta = \varepsilon$, for $\varepsilon \downarrow 0$

Intuition: Maximize power for nearby alternatives,

since we'll have power ≈ 1 for $\theta \gg 1/\sqrt{n}$

More generally, one-sided score test is "almost"

UMP for nearby alternatives.

$$\log \frac{p_{\theta_0 + \varepsilon}(x)}{p_{\theta_0}(x)} \approx \varepsilon \dot{l}_n(\theta_0; x) \quad \text{for small } \varepsilon > 0$$

Ex Pearson's χ^2 test (goodness of fit)

$$N = (N_1, \dots, N_d) \sim \text{Multinom}(n, (\pi_1, \dots, \pi_d))$$

$$= \frac{n! \pi_1^{N_1} \dots \pi_d^{N_d}}{N_1! \dots N_d!} \mathbb{1}_{\{\sum N_i = n\}}$$

Note $\sum \pi_j = 1$ so this is a full-rank $(d-1)$ -parameter exp. family, e.g.

$$\pi_j = \begin{cases} \frac{1}{1 + \sum_{k=2}^d e^{\gamma_k}} & j=1 \\ \frac{e^{\gamma_j}}{1 + \sum_{k=2}^d e^{\gamma_k}} & j>1 \end{cases}$$

$$\nabla \ell_n(\gamma; N) = (N_2, \dots, N_d) - (n\pi_2, \dots, n\pi_d)$$

$$\text{Var}_\gamma(\nabla \ell(\gamma)) = \begin{pmatrix} n\pi_2(1-\pi_2) & & -n\pi_2\pi_3 \\ & \ddots & \\ -n\pi_2\pi_3 & & n\pi_d(1-\pi_d) \end{pmatrix}$$

$$= n(\text{diag}(\pi_{2:d}) - \pi_{2:d}\pi_{2:d}')^{\cdot}$$

$$\Rightarrow J_n(\gamma)^{-1} = \frac{1}{n} \cdot (\text{diag}(\pi_{2:d}))^{-1} - \pi_{2:d}^{-1} \mathbb{1}\mathbb{1}'$$

(uses $(A + uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u}$)

Score test of $H_0: \pi = \pi_0$:

don't really need asy. approx

$$\nabla \ell_n(\gamma_0) J_n^{-1}(\gamma_0) \nabla \ell_n(\gamma_0) \stackrel{\text{(algebra)}}{=} \sum_{j=1}^d \frac{(N_j - n\pi_{0j})^2}{n\pi_{0j}} \Rightarrow \chi_{d-1}^2$$

Generalized LRT

Test $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$

Taylor expand around $\hat{\theta}_n$:

$$\begin{aligned} \ell_n(\theta_0) - \ell_n(\hat{\theta}_n) &= \cancel{\nabla \ell(\hat{\theta}_n)} + \frac{1}{2}(\theta_0 - \hat{\theta}_n)' \nabla^2 \ell_n(\tilde{\theta}_n) (\theta_0 - \hat{\theta}_n) \\ &= -\frac{1}{2} \cdot \left\| \underbrace{\left(-\frac{1}{n} \nabla^2 \ell_n(\tilde{\theta}_n) \right)^{1/2}}_{\xrightarrow{P} J_1(\theta_0)} \underbrace{(\sqrt{n}(\theta_0 - \hat{\theta}_n))}_{\Rightarrow N(0, J_1(\theta_0)^{-1})} \right\|_2^2 \\ &\Rightarrow -\frac{1}{2} \chi_d^2 \end{aligned}$$

Test stat: $2(\ell_n(\hat{\theta}_n; X) - \ell_n(\theta_0; X)) \xrightarrow{P_{\theta_0}} \chi_d^2$

Composite vs. ComPOSITE:

$$H_0: \theta \in \Theta_0 \quad \text{vs} \quad H_1: \theta \in \Theta \setminus \Theta_0,$$

Assume $\cdot \Theta = \mathbb{R}^d$, Θ_0 d_0 -dim manifold

$$\cdot \theta_0 \in \text{relint}(\Theta_0)$$

$$\cdot \hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$$

\cdot Likelihood "smooth"

$$\text{Then } 2(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_0)) \Rightarrow \chi_{d-d_0}^2$$

$$\text{Where } \hat{\theta}_0 = \underset{\theta \in \Theta_0}{\text{argmin}} \ell_n(\theta; X)$$

Why? Assume wlog $\theta_0 = 0$, $J_\ell(0) = I_d$ (reparam.)

$$\text{Then } \hat{\theta}_n \approx N_d(\theta_0, \frac{1}{n} I_d)$$

And locally, $\nabla^2 \ell_n(\theta) \approx n I_d$ near θ_0

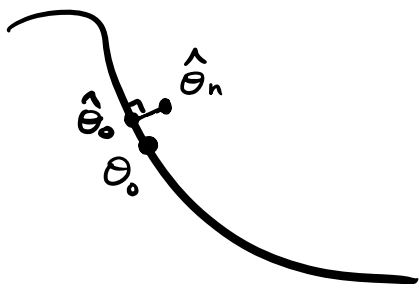
$$\ell_n(\theta) - \ell_n(\hat{\theta}_n) \approx \frac{n}{2} \|\theta - \hat{\theta}_n\|^2$$

$$\hat{\theta}_0 \approx \underset{\theta \in \Theta_0}{\text{argmin}} \|\theta - \hat{\theta}_n\| = \text{Proj}_{\Theta_0}(\hat{\theta}_n)$$

$$2(\ell_n(\hat{\theta}_0) - \ell_n(\hat{\theta}_n)) \approx n \|\hat{\theta}_n - \text{Proj}_{\Theta_0}(\hat{\theta}_n)\|^2$$

$$= n \|\text{Proj}_{\Theta_0^\perp}(\hat{\theta}_n)\|^2$$

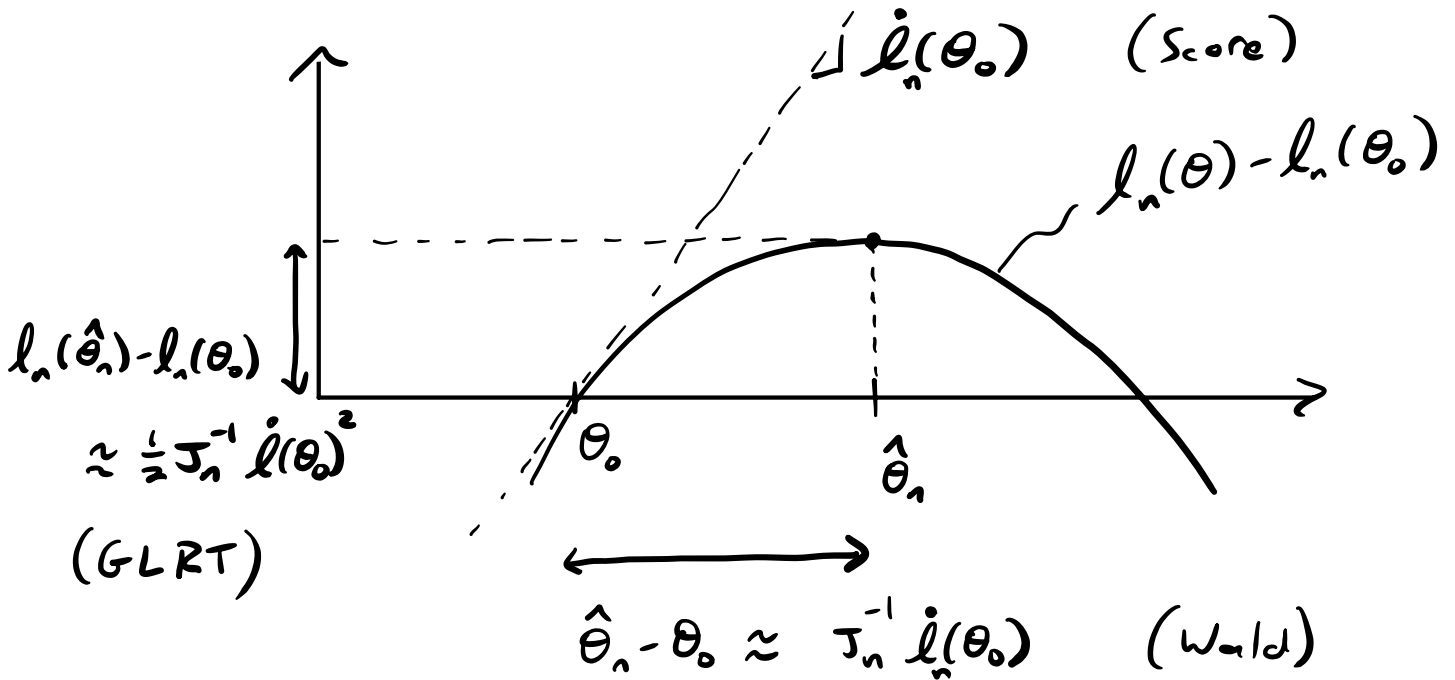
$$\Rightarrow \chi_{d-d_0}^2$$



Asymptotic Equivalence

Recall quadratic approx. picture ($d=1$):

$$l_n(\theta) - l_n(\theta_0) \approx \dot{l}_n(\theta_0)(\theta - \theta_0) + \frac{1}{2} J_n(\theta_0)(\theta - \theta_0)^2$$



For large n ,

$$\begin{aligned}
 l_n(\hat{\theta}_n) - l_n(\theta_0) &\approx \|J_n(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0)\|^2 \\
 \text{(GLRT)} &\quad \approx \quad \approx \\
 &\| \hat{J}_n^{1/2}(\hat{\theta}_n - \theta_0) \|^2 && \|J_n(\theta_0)^{-1/2} \nabla l_n(\theta_0)\|^2 \\
 &\text{(Wald)} && \text{(score)}
 \end{aligned}$$

Asymptotic Relative Efficiency (ARE)

Suppose $\hat{\theta}_n^{(i)}$ $i=1,2$ are two asy. Normal estimators of $\theta_0 \in \mathbb{R}$, with

$$\sqrt{n} (\hat{\theta}_n^{(i)} - \theta_0) \Rightarrow N(0, \sigma_i^2)$$

The ARE of $\hat{\theta}^{(2)}$ wrt $\hat{\theta}^{(1)}$ is σ_1^2 / σ_2^2

e.g. if $\sigma_2^2 = 2\sigma_1^2$ then $\hat{\theta}^{(2)}$ is 50% as efficient

Interpretation: Suppose $\sigma_1^2 / \sigma_2^2 = \gamma \in (0,1)$

Then for large n ,

$$\hat{\theta}_{\lfloor \gamma n \rfloor}^{(1)}(X_1, \dots, X_{\lfloor \gamma n \rfloor}) \stackrel{D}{\approx} \hat{\theta}_n^{(2)}(X_1, \dots, X_n) \approx N(\theta, \frac{\sigma_2^2}{n})$$

Using $\hat{\theta}^{(2)}$ is like throwing away $100(1-\gamma)\%$ of the data and then using $\hat{\theta}^{(1)}$