Outline

1) Wald test

2) Score test

3) Generalized likelihood ratio test

4) Asymptotic Relative Efficiency

Likelihood-Based Inference

Setting: $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta(x)$, $p_\theta(x)$ "smooth" in θ

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

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

$$\hat{\theta}_{MLE} \overset{P_\theta}{\longrightarrow} \theta \qquad \text{(consistent)}$$

Then, if $\theta = \theta_0$:

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

$$\frac{1}{n} \nabla^2 J_n(\theta_0; X) \overset{P}{\longrightarrow} J_1(\theta_0)$$

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

Can use this for inference on θ_0!

Wald-Type Confidence Regions

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

$$\frac{1}{n}\hat{J}_n \overset{p}{\longrightarrow} J_1(\theta_0) \geq 0.$$ Then we can plug in:

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

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

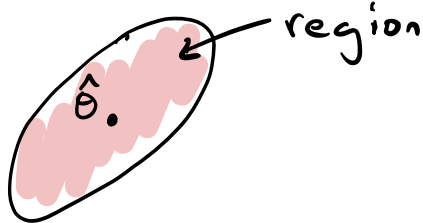so $\hat{J}_n^{1/2}(\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, I_d)$ (Slutsky)

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

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

So, $\mathbb{P}_{\theta_0}\left(\hat{J}_n^{1/2}(\hat{\theta}_n - \theta_0) \geq \underbrace{\chi_d^2(\alpha)}_{\color{blue}1-\alpha \text{ quantile}}\right) \longrightarrow \alpha$

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

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

region
$\hat{\theta}$

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

Options for $\hat{\mathcal{J}}_n$:

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

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

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

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

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

2) Observed Fisher info:

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

Remarks:

- Both have $\frac{1}{n}\hat{\mathcal{J}}_n \xrightarrow{p} J_1(\theta_0)$ 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:

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

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

Leads to univariate interval:

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

$$= \hat{\Theta}_{nj} \pm \sqrt{(\hat{J}_n^{-1})_{jj}} \cdot z_{\alpha/2}$$

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

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

$$\hat{\Theta}_{n,S} \approx N_k(\theta_{0,S}, (J_n(\theta_0)^{-1})_{SS})$$

$$\rightsquigarrow C_S = \hat{\Theta}_{n,S} + ((\hat{J}_n^{-1})_{SS})^{1/2} B_{\chi_k(\alpha)}(0)$$

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

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

then we can do the same things

<u>Ex</u>   Generalized linear model with fixed x

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

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

$\eta_i = \beta' x_i$ (<u>canonical form</u>)

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

(more general: $f(\mu_i) = \beta' x_i$ for <u>link fcn</u> f)

Most common examples:
  Logistic regression: $Y_i \overset{ind.}{\sim} \text{Bern}\left(\dfrac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)$

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

$\ell_n(\beta; Y) = \sum_i (x_i'\beta) Y_i - A(x_i'\beta) - \log h(Y_i)$

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

$\qquad = \sum_i (Y_i - \mu_i(\beta)) x_i$

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

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

$\qquad = \text{Var}_\beta(\nabla \ell_n(\beta; Y))$ (Not random)

$$\left(-\nabla^2 \ell_n(\hat{\beta})\right)^{-1/2} \nabla \ell_n(\hat{\beta}) \sim (0, I_d) \quad \text{in finite samples}$$

$$\overset{*}{\Longrightarrow} N_d(0, I_d)$$

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

Taylor expansion of ℓ_n leads to

$$\hat{J}_n^{1/2} (\hat{\beta}_n - \beta) \Longrightarrow N_d(0, I_d)$$

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:
   $$\nabla \ell_n \approx \text{Normal} \quad \underline{\text{and}} \quad \ell_n \approx \text{quadratic}$$

4) Need MLE to be consistent

5) Confidence interval / ellipsoid might go outside $\ominus$!

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

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

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

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

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

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

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

$$q_s(x) = p_{g(s)}(x)$$

$$\dot{\ell}^{(s)}(s;X) = \frac{d}{ds} \log p_{g(s)}(x)$$

$$= \dot{\ell}^{(\theta)}(g(s);X) \cdot \dot{g}(s)$$

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

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

if $\theta_0 = g(s_0)$

__Ex__   s-parameter exp. fam:

$$X_1,\ldots,X_n \overset{iid}{\sim} e^{\eta' T(x) - A(\eta)} h(x)$$

$$\nabla \ell_n(\eta;X) = \sum T(X_i) - n\mu(\eta)$$

$$\| J_n(\eta_0)^{-1/2} (\sum T(X_i) - n\mu(\eta_0)) \|^2 \Rightarrow \chi^2_d$$

$$\frac{\sum T(X_i) - n\mu(\eta_0)}{\sqrt{n \, Var_{\eta_0}(T(X_i))}} \overset{P_{\eta_0}}{\Rightarrow} N(0,1)$$

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

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

$$= \frac{n! \, \pi_1^{N_1} \cdots \pi_d^{N_d}}{N_1! \cdots N_d!} \, \mathbb{1}\{\textstyle\sum N_i = n\}$$

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

$$\pi_j = \begin{cases} \dfrac{1}{1 + \sum_{k>1} e^{\eta_k}} & j = 1 \\[4mm] \dfrac{e^{\eta_j}}{1 + \sum_{k>1} e^{\eta_k}} & j > 1 \end{cases}$$

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

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

$$= n\left(\text{diag}(\pi_{2:d}) - \pi_{2:d}\,\pi_{2:d}'\right)$$

$$\Rightarrow J_n(\eta)^{-1} = \frac{1}{n} \cdot \left(\text{diag}(\pi_{2:d})^{-1} - \pi_1^{-1}\, \mathbb{1}\mathbb{1}'\right)$$

$\left(\text{uses } (A + uv')^{-1} = A^{-1} - \dfrac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u}\right)$

Score test of $H_0 : \pi = \pi_0$: don't really need asy. approx

(algebra)

$$\nabla \ell_n(\gamma_0) J_n^{-1}(\gamma_0) \nabla \ell_n(\gamma_0) = \sum_{j=1}^{d} \frac{(N_j - n\pi_{0j})^2}{n\pi_{0j}} \overset{P_{\pi_0}}{\Rightarrow} \chi^2_{d-1}$$

Generalized LRT

Test $\quad H_0: \Theta = \Theta_0 \quad$ vs. $\quad H_1: \Theta \neq \Theta_0$

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

$$\ell_n(\Theta_0) - \ell_n(\hat{\Theta}_n) = \nabla \ell(\hat{\Theta}_n)^{\,0} + \frac{1}{2}(\Theta_0 - \hat{\Theta}_n)' \nabla^2 \ell_n(\tilde{\Theta}_n)(\Theta_0 - \hat{\Theta}_n)$$

$$= -\frac{1}{2} \cdot \left\| \underbrace{\left(-\frac{1}{n} \nabla^2 \ell_n(\tilde{\Theta}_n)\right)^{1/2}}_{\xrightarrow{P} J_1(\Theta_0)} \underbrace{\left(\sqrt{n}(\Theta_0 - \hat{\Theta}_n)\right)}_{\Rightarrow N(0, J_1(\Theta_0)^{-1})} \right\|_2^2$$

$$\Rightarrow -\frac{1}{2} \chi_d^2$$

Test stat: $\quad 2\left(\ell_n(\hat{\Theta}_n; X) - \ell_n(\Theta_0; X)\right) \xRightarrow{P_{\Theta_0}} \chi_d^2$

Composite vs. Composite:

$H_0: \theta \in \textcircled{H}_0$ vs $H_1: \theta \in \textcircled{H} \setminus \textcircled{H}_0$,

Assume · $\textcircled{H} = \mathbb{R}^d$, $\textcircled{H}_0$ d_0-dim manifold

· $\theta_0 \in \text{relint}(\textcircled{H}_0)$

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

· Likelihood "smooth"

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

where $\hat{\theta}_0 = \underset{\theta \in \textcircled{H}_0}{\arg\min} \, \ell_n(\theta; X)$

Why? Assume wlog $\theta_0 = 0$, $J_1(0) = I_d$ (reparam.)

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

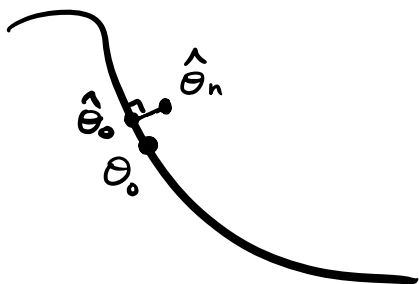And locally, $\nabla^2 \ell_n(\theta) \approx n I_d$ near θ_0

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

$\hat{\theta}_0 \approx \underset{\theta \in \textcircled{H}_0}{\arg\min} \|\theta - \hat{\theta}_n\| = \text{Proj}_{\textcircled{H}_0}(\hat{\theta}_n)$

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

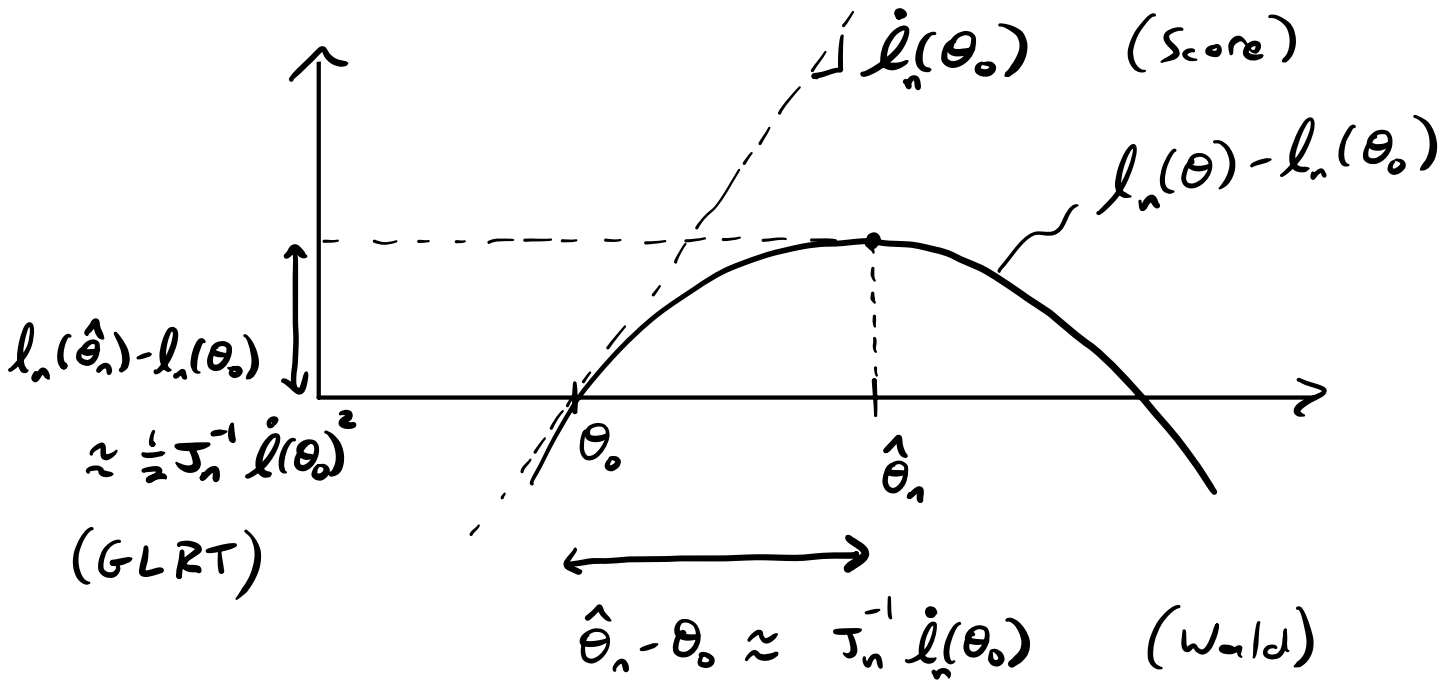$= n \|\text{Proj}_{\textcircled{H}_0^{\perp}}(\hat{\theta}_n)\|^2$

$\Rightarrow \chi^2_{d-d_0}$

Asymptotic Equivalence

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

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

$\dot{\ell}_n(\theta_0)$ (Score)

$\ell_n(\theta) - \ell_n(\theta_0)$

$\ell_n(\hat{\theta}_n) - \ell_n(\theta_0)$

$\approx \frac{1}{2} J_n^{-1} \dot{\ell}(\theta_0)^2$

(GLRT)

θ_0

$\hat{\theta}_n$

$$\hat{\theta}_n - \theta_0 \approx J_n^{-1} \dot{\ell}_n(\theta_0) \qquad \text{(Wald)}$$

For large n,

$$\underset{\text{(GLRT)}}{\ell_n(\hat{\theta}_n) - \ell_n(\theta_0)} \approx \| J_n(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0) \|^2$$

$$\approx$$

$$\underset{\text{(wald)}}{\| \hat{J}_n^{1/2}(\hat{\theta}_n - \theta_0) \|^2} \qquad \underset{\text{(Score)}}{\| J_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0) \|^2}$$

Asymptotic Relative Efficiency (ARE)

Suppose $\hat{\Theta}_n^{(i)}$ $i = 1, 2$ are two asy. Normal

estimators of $\Theta_0 \in \mathbb{R}$, with

$$\sqrt{n}\left(\hat{\Theta}_n^{(i)} - \Theta_0\right) \Rightarrow N(0, \sigma_i^2)$$

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,

$$\hat{\Theta}_{\lfloor \gamma n \rfloor}^{(1)}(X_1, \ldots, X_{\lfloor \gamma n \rfloor}) \overset{D}{\approx} \hat{\Theta}_n^{(2)}(X_1, \ldots, X_n) \approx N\left(\Theta, \frac{\sigma_2^2}{n}\right)$$

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