Outline

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

Review

Consistency: \( W_i(\theta) = l_i(\theta; x_i) - l_i(\theta_0; x_i) \)

\[ \mathbb{E} W_i(\theta) = -D_{KL}(\theta_0 \| \theta) \leq 0 \], \( = 0 \) iff \( \theta_0 = \theta_{\theta_0} \)

\( \hat{\theta}_n \xrightarrow{p} \theta_0 \) if \( Wi \) cts, \( \| W_n - \mathbb{E} W_n \|_p \xrightarrow{P} 0 \) on compact \( \Theta \)

(If \( \Theta = \mathbb{R}^d \), need an extra argument)

Asym. normality: If \( \hat{\theta}_n \to \theta_0 \), \( l \) "smooth" near \( \theta_0 \)

\[ 0 = \nabla l_n(\hat{\theta}_n) = \nabla l_n(\theta_0) + \nabla^2 l_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \]

\[ \nabla^2 l_n(\hat{\theta}_n - \theta_0) = \left( -\frac{1}{n} \nabla^2 l_n(\hat{\theta}_n) \right)^{-1} \left( \frac{1}{n} \nabla l_n(\theta_0) \right) \]

\( \Rightarrow N_d(0, \nabla^2 l_n(\theta_0)^{-1}) \)
Likelihood-Based Inference

Setting \( X_1, \ldots, X_n \overset{iid}{\sim} \rho_\theta(x) \), \( \rho_\theta(x) \) "smooth" in \( \Theta \)

Assume \( \mathbb{E}_\theta \nabla l(\theta; x_i) = 0 \),

\[
\text{Var}_\theta \left[ \nabla l(\theta; x_i) \right] = -\mathbb{E}_\theta \nabla^2 l(\theta; x_i) = J_1(\theta) > 0.
\]

\( \hat{\theta} \xrightarrow{\text{MLE}} \theta \) (consistent)

Then, if \( \theta = \theta_0 \):

\[
\frac{1}{n} \nabla l_n(\theta_0; x) \Rightarrow \mathcal{N}_d(0, J_1(\theta_0))
\]

\[
\frac{1}{n} \nabla^2 J_n(\theta_0; x) \xrightarrow{p} J_1(\theta_0)
\]

Used \( O = \nabla l(\hat{\theta}_n) \approx \nabla l(\theta_0) + \nabla^2 l(\theta_0)(\hat{\theta}_n - \theta_0) \)
to get \( \frac{1}{n} (\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}_d(0, J_1(\theta_0)^{-1}) \)

Can use this for inference on \( \Theta_0 \)!
Wald-Type Confidence Regions

Assume we have some estimator \( \hat{\theta}_n \) such that
\[
\frac{1}{n} \hat{\theta}_n \xrightarrow{p} \theta_0.
\]
Then we can plug in:

If
\[
\sqrt{n} (\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, \Lambda_{\theta_0})
\]
then
\[
(\Lambda_{\theta_0})^{1/2} \sqrt{n} (\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, \Lambda_{\theta_0})
\]
so
\[
\sqrt{n} (\hat{\theta}_n - \theta_0) \Rightarrow N_d(0, \Lambda_{\theta_0})
\]
(Slutsky)

Leads to test of \( H_0: \theta = \theta_0 \):

\[
| \sqrt{n} (\hat{\theta}_n - \theta_0) |^2 \Rightarrow \chi_d^2 \quad \text{(Reject if large)}
\]

So,
\[
P_{\theta_0}( \sqrt{n} (\hat{\theta}_n - \theta_0) \leq \chi_d^2(\alpha) ) \Rightarrow 1 - \alpha \quad \text{quantile}
\]

Note we reject \( \theta_0 \) iff
\[
| \sqrt{n} (\hat{\theta}_n - \theta_0) |^2 \geq \chi_d^2(\alpha)
\]
\[
\Rightarrow \text{reject } \theta_0 \iff \theta_0 \notin \hat{\theta}_n + \sqrt{n}^{-1/2} B(0)
\]

More info \( \Rightarrow \) smaller ellipse (shrinks like \( \sqrt{n} \))
Options for $\hat{J}_n$:

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

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

\[
= \operatorname{Var}_\theta(\nabla l_n(\theta; x)) \bigg|_{\theta = \hat{\theta}_n}
\]

(NB) \neq \operatorname{Var}_{\hat{\theta}_n}(\nabla l_n(\hat{\theta}_n(x); x)) = 0

Or, \[\hat{J}_n = -\mathbb{E}_\theta \nabla^2 l_n(\theta) \bigg|_{\theta = \hat{\theta}_n}\]

2) Observed Fisher info:

\[
\hat{J}_n = -\nabla^2 l_n(\hat{\theta}_n; x)
\]

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 $\theta$ in "typical" data set but obs. info
takes measures info about $\theta$ in "this" data set
Wald interval for $\theta_j$:

If

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

then

$$\hat{\theta}_{n,j} \approx N_d(\theta_{0,j}, (J_n(\theta_0)^{-1})_{jj})$$

Leads to univariate interval:

$$C_j = \hat{\theta}_{n,j} \pm \text{s.e.}(\hat{\theta}_{n,j}) \cdot z_{1/2}$$

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

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

Conf. ellipsoid for $\Theta_s = (\theta_{0,s})_{ss}$ : ($\text{Is} = k$)

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

$$\Rightarrow C_s = \hat{\Theta}_{n,s} + ((\hat{\gamma}_n^{-1})_{ss})^{1/2} \beta_{k(\eta)}(0)$$

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

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

then we can do the same things
Ex Generalized linear model with fixed $X$

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

$y_i \overset{\text{ind.}}{\sim} ho_{y_i}(y) = e^{\eta_i} y_i - A(\eta_i) h(y_i)$

$\eta_i = \beta' x_i$ (canonical form)

Let $M_i(\beta) = \mathbb{E}_{\beta} y_i \quad (= m(\eta_i(\beta)))$

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

Most common examples:

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

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

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

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

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

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

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

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

(Not random)
\((- V_n^2 \ln(\beta))^{-1/2} V_n(\beta) \sim (0, I_d) \quad \text{in finite samples}\)

\[ \Rightarrow N_d(0, I_d) \]

* Under regularity cond. on \( X = (\begin{pmatrix} -x_1 \cdots -x_n \end{pmatrix}^T) \)

Taylor expansion of \( \ln \) leads to

\[ \frac{1}{\sqrt{n}} (\beta_n - \beta) \Rightarrow N_d(0, I_d) \]

**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: \( \nabla \ln \approx \text{Normal} \quad \text{and} \quad \ln_n \approx \text{quadratic} \)
4. Need MLE to be consistent
5. Confidence interval / ellipsoid might go outside \( \mathbb{R}^d \)!
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 statistic

$$\frac{1}{\sqrt{n}} \nabla l_n(\theta_0; x) \overset{P}{\rightarrow} N_d(0, J_1(\theta_0))$$
(or $J_n(\theta_0)^{-\frac{1}{2}} \nabla l_n(\theta_0; x) \overset{P}{\rightarrow} N_d(0, I_d)$)

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

$$\| J_n(\theta_0)^{-\frac{1}{2}} \nabla l_n(\theta_0; x) \|^2 \geq \chi^2_d(\alpha)$$

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

...can do 1-sided tests

Remarks

- No quadratic approx., no MLE
- No need to estimate Fisher info at $\theta_0$

Can be generalized to case with nuisance params

Typically estimate via MLE on $\Theta_0$. 
Score test is invariant to reparameterization.*

Assume \( d = 1 \), \( \Theta = g(S) \), \( j(S) > 0 \) \( \forall S \)

\[
q_S(x) = p_{g(S)}(x)
\]

\[
\ell^{(S)}(S; x) = \frac{d}{dS} \log p_{g(S)}(x)
\]

\[
= \ell^{(\Theta)}(g(S); X) \cdot j(S)
\]

\[
J^{(S)}(S) = J^{(\Theta)}(g(S)) \cdot j(S)^2
\]

\[
S_0 \quad \frac{\ell^{(S)}(S_0; x)}{\sqrt{J^{(S)}(S_0)}} \quad \text{a.s.} \quad \frac{\ell^{(\Theta)}(\Theta_0; x)}{\sqrt{J^{(\Theta)}(\Theta_0)}}
\]

if \( \Theta_0 = g(S_0) \)

**Example** \( S \)-parameter exp. fam:

\[
X_1, \ldots, X_n \overset{iid}{\sim} e^{\tau(x) - A(\gamma) h(x)}
\]

\[
\nabla \ell^o(\gamma; X) = \sum T(X_i) - \eta \mu(\gamma)
\]

\[
\frac{\| \sum (\gamma_0)^{-1/2} (\sum T(X_i) - \eta \mu(\gamma_0)) \|^2}{\sqrt{\text{Var}_{\gamma_0}(T(X_i))}} \overset{P}{\to} N(0, 1)
\]
Pearson's \( \chi^2 \) test (goodness of fit)

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

\[
= \frac{n! \pi_1^{N_1} \cdots \pi_d^{N_d}}{N_1! \cdots N_d!} \quad 1 \leq \sum N_i = n
\]

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

\[
\pi_j = \begin{cases} 
\frac{1}{\sum_{k=1}^{j} e^{\gamma_k}} & j = 1 \\
\frac{e^{\gamma_j}}{\sum_{k=1}^{j} e^{\gamma_k}} & j > 1 
\end{cases}
\]

\[
\nabla l(\gamma; N) = (N_2, \ldots, N_d) - (n \pi_2, \ldots, n \pi_d)
\]

\[
\text{Var}_q(\nabla l(\gamma)) = \begin{pmatrix}
n \pi_2 (1-\pi_2) & -n \pi_i \pi_j \\
-n \pi_i \pi_j & n \pi_d (1-\pi_d)
\end{pmatrix}
\]

\[
= n (\text{diag}(\pi_2:d) - \pi_2:d \pi_2:d^{-1})
\]

\[
\Rightarrow \nabla^2 l(\gamma) = \frac{1}{n} (\text{diag}(\pi_2:d)^{-1} - \pi_i^{-1} 11')
\]

(uses \( (A + vv')^{-1} = A^{-1} - \frac{A^{-1} vv' A^{-1}}{1 + v' A^{-1} v} \))

Score test of \( H_0 : \pi = \pi_0 : \)

\[
\text{Score test of } H_0 : \pi = \pi_0 : \quad \text{Score test of } H_0 : \pi = \pi_0 : \quad \text{Score test of } H_0 : \pi = \pi_0:
\frac{1}{\sum_{j=1}^{d} (N_j - n \pi_{0,j})^2} \mathbb{P}_{\pi_0} \chi^2_{d-1}
\]

\( \text{Score test of } H_0 : \pi = \pi_0 : \quad \text{Score test of } H_0 : \pi = \pi_0 : \quad \text{Score test of } H_0 : \pi = \pi_0:
\]
Generalized LRT

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

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

$$l_n(\Theta_0) - l_n(\widehat{\Theta}_n) = \nabla l(\widehat{\Theta}_n) + \frac{1}{2}(\Theta - \widehat{\Theta}_n)' \nabla^2 l_n(\widehat{\Theta}_n)(\Theta - \widehat{\Theta}_n)$$

$$= -\frac{1}{2} \| \left( \frac{1}{n} \nabla^2 l_n(\widehat{\Theta}_n) \right)^{1/2} (\sqrt{n}(\Theta_0 - \widehat{\Theta}_n)) \|_2^2$$

$$\xrightarrow{p} J_1(\Theta_0) \Rightarrow N(0, J_1(\Theta_0))$$

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

Test stat: $2(l_n(\widehat{\Theta}_n; x) - l_n(\Theta_0; x)) \overset{p}{\rightarrow} \chi_d^2$
Composite vs. Composite:
\[ H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta \setminus \Theta_0, \]

Assume \( \Theta = \mathbb{R}^d \), \( \Theta_0 \) \( d_0 \)-dim manifold
1. \( \Theta_0 \in \text{relint}(\Theta_0) \)
2. \( \Theta_n \to \Theta_0 \)
3. Likelihood "smooth"

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

where \( \hat{\theta}_0 = \arg\min_{\Theta \in \Theta_0} \ell_n(\theta; x) \)

Why? Assume wlog \( \Theta_0 = \mathbb{O}, \mathbb{I}, (0) = \mathbb{I} \) (separam.)
Then \( \hat{\Theta}_n \approx N_d(\Theta_0, \frac{1}{n}\mathbb{I}) \)

And locally, \( \nabla^2 \ell_n(\theta) \approx n\mathbb{I} \) near \( \Theta_0 \)

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

\[ \hat{\Theta}_0 = \arg\min_{\Theta \in \Theta_0} \| \theta - \hat{\theta}_n \| = \text{Proj}_{\Theta_0}(\hat{\theta}_n) \]

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

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

\[ \Rightarrow \chi^2_{d-d_0} \]
Asymptotic Equivalence

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

\[
\ell_n(\theta) - \ell_n(\theta_0) \approx \ell_n(\theta_0)(\theta - \theta_0) + \frac{1}{2} J_n(\theta_0)(\theta - \theta_0)^2
\]

\[
\ell_n(\hat{\theta}_n) - \ell_n(\theta_0) \approx \frac{1}{2} J_n^{-1}\ell_n(\theta_0)^2 \quad \text{(GLRT)}
\]

\[
\hat{\theta}_n - \theta_0 \approx J_n^{-1}\ell_n(\theta_0) \quad \text{(Wald)}
\]

For large \(n\),

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

\[
\| J_n^{1/2}(\hat{\theta}_n - \theta_0) \|^2 \approx \| J_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0) \|^2 \quad \text{(Wald)}
\]

\[
J_n(\theta_0)^{1/2} \nabla \ell_n(\theta_0) \quad \text{(score)}
\]
Asymptotic Relative Efficiency (ARE)

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

\[
\sqrt{n} (\hat{\Theta}_n - \Theta) \Rightarrow N(0, \sigma_i^2)
\]

The ARE of \( \hat{\Theta}^{(2)} \) wrt \( \hat{\Theta}^{(1)} \) is \( \sigma_1^2/\sigma_2^2 \)

c.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 \( n \),

\[
\hat{\Theta}^{(1)}(x_1, \ldots, x_{2n}) \overset{D}{\sim} \hat{\Theta}^{(2)}(x_1, \ldots, x_n) \overset{D}{=} N(\Theta, \sigma_2^2)
\]

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