# STA732

# Statistical Inference

## Lecture 23: Large-Sample Theory for Likelihood Ratio Tests

Yuansi Chen

Spring 2023

Duke University

https://www2.stat.duke.edu/courses/Spring23/sta732.01/

# Recap from Lecture 22

1. **Canonical linear model**

$$Z = \begin{pmatrix} Z_0 \\ Z_1 \\ Z_r \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} \mu_0 \\ \mu_1 \\ 0 \end{pmatrix}, \sigma \mathbb{I}_n \right)$$

   - $\sigma^2$ known, $d_1 = 1$, $\quad Z$-test: $\frac{Z_1}{\sigma}$
   - $\sigma^2$ unknown, $d_1 = 1$, $\quad t$-test: $\frac{Z_1}{\hat{\sigma}}$
   - $\sigma^2$ known, $d_1 \geq 1$, $\quad \chi^2$-test: $\frac{\|Z_1\|_2^2}{\sigma^2}$
   - $\sigma^2$ unknown, $d_1 \geq 1$, $\quad F$-test: $\frac{\|Z_1\|_2^2/d_1}{\hat{\sigma}^2}$

2. **General linear model**: find an orthonormal matrix $Q$ such that $Q^\top Y$ follows the canonical linear model

# Goal of Lecture 22

1. Wald test
2. Score test
3. Generalized likelihood ratio test

Chap. 17.1-3 of Keener or 12.4 in Lehmann and Romano

# Review the asymptotics of MLE

# Setup

$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)$, $p_\theta(\cdot)$ is "regular" enough (check the conditions in Thm 9.14 of Keener)

# Consistency of MLE on compact $\Omega$

Define

$$W_i(\theta) = \ell_1(\theta; X_i) - \ell_1(\theta_0; X_i)$$

$$\bar{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i$$

We know that

$$\mathbb{E}W_i(\theta) = -\mathcal{D}_{\mathsf{KL}}(\theta_0 \,\|\, \theta) \leq 0$$

and it becomes $= 0$ iff $P_\theta = P_{\theta_0}$.

**Consistency result**

If model is identifiable, $W_i$ continuous random function, then

- $\left\|\bar{W}_n - \mathbb{E}\bar{W}_n\right\|_\infty \xrightarrow{p} 0$ on compact $\Omega$.
- Then $\hat{\theta}_n \xrightarrow{p} \theta_0$ (convergence of argmax requires uniform convergence result in Thm 9.4 Keener)

# Asymptotic distribution of MLE

MLE satisfies

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

Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left(-\frac{1}{n}\nabla^2 \ell_n(\tilde{\theta}_n)\right)^{-1} \left(\frac{1}{\sqrt{n}}\nabla \ell_n(\theta_0)\right)$$

- $\left(-\frac{1}{n}\nabla^2 \ell_n(\tilde{\theta}_n)\right)^{-1} \xrightarrow{p} I_1(\theta_0)^{-1}$ (convergence of a random function evaluated on a random point requires uniform convergence result in Thm 9.4 Keener!)
- $\frac{1}{\sqrt{n}}\nabla \ell_n(\theta_0) \Rightarrow \mathcal{N}(0, I_1(\theta_0))$ (CLT)

By Slutsky's thm, $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, I_1(\theta_0)^{-1})$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, I_1(\theta_0)^{-1})$$

We can use the asymptotic distribution to compute confidence regions!

# Wald test

# Intuition for Wald-type confidence regions (1)

Assume we have an estimator $\hat{I}_n \succeq 0$ such that

$$\frac{1}{n}\hat{I}_n \xrightarrow{p} I_1(\theta_0)$$

Then we can use it as plug-in estimate for $I_1(\theta_0)$ in asymptotic distribution

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, I_1(\theta_0)^{-1})$,
then $(I_1(\theta_0))^{1/2}\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, \mathbb{I}_d)$,
by Slutsky's thm,

$$\hat{I}_n^{1/2}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, \mathbb{I}_d)$$

# Intuition for Wald-type confidence regions (2)

Under the null hypothesis $H_0 : \theta = \theta_0$, we have

$$\left\| \hat{I}_n^{1/2} (\hat{\theta}_n - \theta_0) \right\|_2^2 \Rightarrow \chi_d^2$$

We can construct a test that rejects for large value of $\left\| \hat{I}_n^{1/2} (\hat{\theta}_n - \theta_0) \right\|_2^2$:

$$\phi = \mathbf{1}_{\left\| \hat{I}_n^{1/2} (\hat{\theta}_n - \theta_0) \right\|_2^2 > \chi_d^2(\alpha)}$$

**Remark**

- The test might not have the correct level. It only has asymptotic level $\alpha$
- The confidence region is an ellipsoid

$$\hat{\theta}_n + \hat{I}_n^{-1/2} \mathbb{B}(0, \chi_d^2(\alpha))$$

# Two options for $\hat{I}_n$

1. $I_n(\hat{\theta}_n)$ obtained by plugging in the MLE

$$\begin{aligned}\hat{I}_n &= I_n(\hat{\theta}_n) \\ &= \mathrm{Var}_\theta(\nabla \ell_n(\theta; X)) \mid_{\theta=\hat{\theta}_n}\end{aligned}$$

2. Observed Fisher information

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

**Remark:**

Both should have $\frac{1}{n}\hat{I}_n \xrightarrow{p} I_1(\theta_0)$ in "regular" model i.i.d. setting

# Wald interval for $\theta_j$

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, I_1(\theta_0)^{-1})$,
then by multiplying $(1, 0, \dots, 0)^\top$, we obtain

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j}) \Rightarrow \mathcal{N}(0, (I_1(\theta_0)^{-1})_{jj})$$

Using $\frac{1}{n}\hat{I}_n$ as plug-in estimate for $I_1(\theta_0)$, we obtain univariate interval

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

# Wald interval for $\theta_j$

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, I_1(\theta_0)^{-1})$,
then by multiplying $(1, 0, \dots, 0)^\top$, we obtain

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j}) \Rightarrow \mathcal{N}(0, (I_1(\theta_0)^{-1})_{jj})$$

Using $\frac{1}{n}\hat{I}_n$ as plug-in estimate for $I_1(\theta_0)$, we obtain univariate interval

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

**glm function in R uses the above intervals:**

with $\hat{I}_n = -\nabla^2 \ell_n(\hat{\theta}_n)$

# Confidence ellipsoid for $\theta_{0,S}$

Want to provide confidence ellipsoid for $\theta_{0,S} = (\theta_{0,j})_{j \in S}, |S| = k$

We have

$$\sqrt{n}\left(\hat{\theta}_{n,S} - \theta_{0,S}\right) \Rightarrow \mathcal{N}(0, \left(I_1(\theta_0)^{-1}\right)_{SS})$$

Then the confidence ellipsoid is

$$\hat{\theta}_{n,S} + \left((\hat{I}_n^{-1})_{SS}\right)^{1/2} \mathbb{B}(0, \chi_k(\alpha))$$

# Example: generalized linear model with fixed design

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

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

where $\eta_i = \beta^\top x_i$

**Link function**

Let $\mu_i(\beta) = \mathbb{E}_\beta Y_i$. If $f(\mu_i) = \beta^\top x_i$, then $f$ is called link function.

**Common examples**

- Logistic regression: $Y_i \overset{\text{ind.}}{\sim}$ Bernoulli $\left(\frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}\right)$
- Poisson log-linear model: $Y_i \overset{\text{ind.}}{\sim}$ Poisson$(e^{x_i^\top \beta})$

# Confidence interval in generalized linear model

$$
\begin{aligned}
\ell_n(\beta; Y) &= \sum_{i=1}^{n} (x_i^\top \beta) y_i - A(x_i^\top \beta) - \log h(y_i) \\
\nabla \ell_n(\beta; Y) &= \sum_{i=1}^{n} y_i x_i - A'(x_i^\top \beta) x_i \\
&= \sum_{i=1}^{n} \left(y_i - \mu_i(\beta)\right) x_i \\
-\nabla^2 \ell_n(\beta; Y) &= \sum_{i=1}^{n} A''(x_i^\top \beta) x_i x_i^\top \\
&= \sum_{i=1}^{n} \mathrm{Var}_\beta(y_i) x_i x_i^\top \\
&= \mathrm{Var}_\beta(\nabla \ell_n(\beta; Y))
\end{aligned}
$$

in GLM, $-\nabla^2 \ell_n(\beta; Y)$ is not random

Can estimate $\hat{I}_n$ by plug-in MLE

Apply our asymptotic directly (or do Taylor expansion from scracth)

$$\hat{I}_n^{1/2}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, \mathbb{I}_d)$$

# Pros and cons of Wald test

## Advantages

- Easy to invert, simple confidence regions
- Asympotically correct level

## Disadvantages

- Have to compute MLE
- Depends on parameterization
- Relies on second order Taylor expansion of $\ell_n$
- Need MLE to be consistent
- Confidence region might go outside of $\Omega$

# Score test

# Intuition for score test

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

We can bypass quadratic approximation by using the score as test statistics

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0) \Rightarrow \mathcal{N}(0, I_1(\theta_0))$$

# Score test

Reject $H_0 : \theta = \theta_0$ if

$$\left\| I_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0) \right\|_2^2 \geq \chi_d^2(\alpha)$$

if $d = 1$, we just use $Z$-test instead

# Score test

Reject $H_0 : \theta = \theta_0$ if

$$\left\| I_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0) \right\|_2^2 \geq \chi_d^2(\alpha)$$

if $d = 1$, we just use $Z$-test instead

## Advantages of score test

- No quadratic approximation
- No MLE

Disadvantage is that it might not be easy to invert the test

# Also, score test is invariant to reparameterization

Assume $d = 1, \theta = g(\xi)$ with $g'(\xi) > 0$,

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

show that the two test statistics are the same a.s.

# Generalized likelihood ratio test

# GLRT in simple vs composite two-sided testing

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

Taylor expansion around $\hat{\theta}_n$ gives

$$\begin{aligned}
\ell_n(\theta_0) - \ell_n(\hat{\theta}_n) &= \nabla \ell(\hat{\theta}_n) + \frac{1}{2}(\theta_0 - \hat{\theta}_n)^\top \nabla^2 \ell_n(\tilde{\theta}_n)(\theta_0 - \hat{\theta}_n) \\
&= 0 - \frac{1}{2} \left\| \left( -\frac{1}{n} \nabla^2 \ell_n(\tilde{\theta}_n) \right)^{1/2} (\sqrt{n}(\theta_0 - \hat{\theta}_n)) \right\|_2^2 \\
&\Rightarrow -\frac{1}{2} \chi_d^2
\end{aligned}$$

why?

## Test statistic in GLRT

$$2 \log(\lambda) = 2 \left( \ell_n(\hat{\theta}_n) - \ell_n(\theta_0) \right) \Rightarrow \chi_d^2$$

# GLRT in composite vs composite

Testing $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \in \Omega \backslash \Omega_0$

The generalized likelihood ratio is

$$\lambda = \frac{\sup_{\Omega_1} L(\theta)}{\sup_{\Omega_0} L(\theta)}$$

The test statistic is

$$2 \log(\lambda) = 2 \left( \ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_0) \right)$$

where $\hat{\theta}_0 = \arg\max_{\theta \in \Omega_0} \ell_n(\theta)$

# Asympotitic distribution of $2\log(\lambda)$

**Asymptotic distribution of $2\log(\lambda)$, see 17.2 Keener**

Assume $\Omega = \mathbb{R}^d$, $\Omega_0$ $d_0$-dim subspace. $\theta_0$ in interior of $\Omega_0$, $\hat{\theta}_n$ is consistent, $p_\theta(\cdot)$ is "regular" (as in the asymptotic of MLE), then

$$2\log(\lambda) = 2\left(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_0)\right) \Rightarrow \chi^2_{d-d_0}$$

where $\hat{\theta}_0 = \arg\max_{\theta\in\Omega_0} \ell_n(\theta)$

# Intuition for the asymptotic distribution

(See rigorous derivation in 17.2 Keener)

Assume $\theta_0 = 0$, $I_0(0) = \mathbb{I}_d$ (after reparameterization), then

- $\hat{\theta}_n \approx \mathcal{N}(\theta_0, \frac{1}{n}\mathbb{I}_d)$
- locally, $\nabla^2 \ell_n(\theta) \approx n\mathbb{I}_d$ near $\theta_0$
- $\ell_n(\theta) - \ell_n(\hat{\theta}_n) \approx \frac{n}{2}\left\|\theta - \hat{\theta}_n\right\|_2^2$
- $\hat{\theta}_0 \approx \arg\min_{\theta \in \Omega_0} \left\|\theta - \hat{\theta}\right\|_2^2 = \text{Proj}_{\Omega_0}(\hat{\theta}_n)$
-

$$2\left(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_0)\right) \approx n\left\|\hat{\theta}_n - \text{Proj}_{\Omega_0}(\hat{\theta}_n)\right\|_2^2$$

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

# Asymptotic equivalence of the three tests

How close are the three tests asymptotically?

- **Wald test:** $\left\|\hat{I}_n^{1/2}(\hat{\theta}_n - \theta_0)\right\|_2^2$
- **Score test:** $\left\|I_n(\theta_0)^{-1/2}\nabla\ell_n(\theta_0)\right\|_2^2$
- **GLRT:** $\ell_n(\hat{\theta}_n) - \ell_n(\theta_0)$

all are related to (for large $n$)

$$\left\|I_n(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0)\right\|_2^2$$

# Summary

- **Wald test**: test statistic based on quadratic approx
- **Score test**: test statistic using score
- **Generalized likelihood ratio test**: $2\log(\lambda)$
  We intuitively derived its asympotitic distribution

Read Page 362 of Keener for strengths and weaknesses

# What is next?

- Final exam on Monday, May 1
- Office hours in the week of April 24
    - Yuansi: usual lecture hours in Old Chem 223B
    - Christine: Tuesday 1:00-2:00, Friday 2:20-3:20 in Old Chem 203B

Thank you