STA732
Statistical Inference

Lecture 22: Testing in general linear model

Yuansi Chen

Spring 2023

Duke University

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

Recap from Lecture 21

- Showed conditional tests are UMPU for multi-param exponential family with nuisance param

$$p_{\theta,\eta}(x) = h(x) \exp\left(\theta U(x) + \eta^\top T(x) - A(\theta, \eta)\right)$$

1. $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$
2. $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

The UMPU tests condition on $T(x)$, rejects for conditionally large/extreme $U(x)$

Goal of Lecture 22

1. χ^2, t, F distributions
2. Canonical linear model
3. General linear model

Chap. 13.5-8 of Keener or Chap. 7 of Lehmann and Romano

Distributions related to Gaussian

Chi-square distribution

Suppose $Z_1, \dots, Z_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then

$$V = \sum_{i=1}^{d} Z_i^2 \sim \chi_d^2$$

Note that

$$\chi_d^2 = \text{Gamma}\left(\frac{d}{2}, 2\right)$$

with shape and scale parametrization Gamma(k, θ) has density $f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$.

$$\mathbb{E}[V] = d, \quad \mathrm{Var}[V] = 2d$$

Asymptotically, $d \to \infty$

$$\frac{V}{d} \xrightarrow{p} 1$$

t-distribution

Suppose $Z \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2_d$ independent, then

$$\frac{Z}{\sqrt{V/d}} \sim t_d$$

t-distribution with degree of freedom ν has density

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Asymptotically, $d \to \infty$

$$\frac{Z}{\sqrt{V/d}} \Rightarrow \mathcal{N}(0, 1)$$

F-distribution

Suppose $V_1 \sim \chi^2_{d_1}$ and $V_2 \sim \chi^2_{d_2}$, V_1 and V_2 independent, then

$$\frac{V_1/d_1}{V_2/d_2} \sim F_{d_1,d_2}$$

Asymptotically, $d_2 \to \infty$

$$\frac{V_1/d_1}{V_2/d_2} \Rightarrow \frac{1}{d_1}\chi^2_{d_1}$$

Note that if $T \sim t_d$ then $T^2 \sim F_{1,d}$

Geometric interpretation of 1-sample t-test

$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, μ, σ^2 unknown. We show the UMPU test for $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ rejects for extreme value

$$\frac{\sqrt{n}\bar{X}}{\sqrt{S^2}}$$

geometric interpretation?

Canonical linear model

Canonical linear model

Suppose

$$Z = \begin{pmatrix} Z_0 \\ Z_1 \\ Z_r \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \mu_1 \\ 0 \end{pmatrix}, \sigma^2 \mathbb{I}_n \right)$$

where dimensions of Z_0, Z_1, Z_r are $d_0, d_1 = d - d_0, d_r = n - d$.

$\mu_0 \in \mathbb{R}^{d_0}, \mu_1 \in \mathbb{R}^{d_1}, \sigma^2 > 0$

Testing $H_0 : \mu_1 = 0$ vs $H_1 : \mu_1 \neq 0$

The density of Z is exponential family

$$p(z) \propto \exp\left(\frac{\mu_1}{\sigma^2}^\top z_1 + \frac{\mu_0}{\sigma^2}^\top z_0 - \frac{1}{2\sigma^2}\left\|z\right\|_2^2\right)$$

We distinguish four cases

1. σ^2 known, $d_1 = 1$
2. σ^2 unknown, $d_1 = 1$
3. σ^2 known, $d_1 \geq 1$
4. σ^2 unknown, $d_1 \geq 1$

1. σ^2 known, $d_1 = 1$

Idea:

- μ_0 is the only nuisance parameter
- Z_r is irrelevant
- So condition on Z_0, build a conditional test that rejects for extreme value of Z_1
- Observe that Z_1 is independent of Z_0.
- Finally, test that rejects for extreme value of Z_1

Test statistic is $Z_1 \sim \mathcal{N}(\mu_1, \sigma^2)$. Under the null

$$\frac{Z_1}{\sigma} \sim \mathcal{N}(0, 1)$$

2. σ^2 unknown, $d_1 = 1$

Idea:

- μ_0, σ are both nuisance parameters
- So condition on Z_0 and $\|Z\|_2^2$, build conditional tests that reject for extreme value of Z_1

The test statistics could be $\frac{Z_1}{\sqrt{\|Z_r\|_2^2/d_r}}$ (as it is increasing on Z_1 conditioned on $\|Z\|_2^2$ and Z_0, independent of $\|Z\|_2^2$ and Z_0). Under the null,

$$\frac{Z_1}{\sqrt{\|Z_r\|_2^2/d_r}} \sim t_{d_r}$$

t-test!

3. σ^2 known, $d_1 \geq 1$

test that rejects for extreme value of $\|Z_1\|_2$

Under the null,

$$\frac{\|Z_1\|_2^2}{\sigma^2} \sim \chi^2_{d_1}$$

χ^2-test!

4. σ^2 unknown, $d_1 \geq 1$

conditional test (conditioned on Z_0 and $\|Z\|_2^2$) that rejects for extreme value of $\|Z_1\|_2^2$

The test statistic could be $\frac{\|Z_1\|_2^2/d_1}{\|Z_r\|_2^2/d_r}$ (independence?) Under the null,

$$\frac{\|Z_1\|_2^2/d_1}{\|Z_r\|_2^2/d_r} \sim F_{d_1,d_r}$$

F-test!

Summary of four cases

Note that

$$\frac{\|Z_r\|_2^2}{d_r} \sim \frac{\sigma^2}{d_r}\chi^2_{d_r}$$

serves an unbiased estimator of σ^2.

$$\mathbb{E}\hat{\sigma}^2 = \sigma^2, \quad \mathrm{Var}(\hat{\sigma}^2) = 2\sigma^2/d_r$$

1. σ^2 known, $d_1 = 1$, Z-test: $\frac{Z_1}{\sigma}$
2. σ^2 unknown, $d_1 = 1$, t-test: $\frac{Z_1}{\hat{\sigma}}$
3. σ^2 known, $d_1 \geq 1$, χ^2-test: $\frac{\|Z_1\|_2^2}{\sigma^2}$
4. σ^2 unknown, $d_1 \geq 1$, F-test: $\frac{\|Z_1\|_2^2/d_1}{\hat{\sigma}^2}$

How to test $H_0 : \mu_1 = \mu_1^0 \in \mathbb{R}^d$

If μ_1^0 is not 0, μ_1 is not a natural parameter
But we can still translate the problem

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

We can do the same tests with $Z_1 - \mu_1^0$ replacing Z_1.

Intervals for the canonical linear model

Invert the tests

1. σ^2 known, $d_1 = 1$: $\frac{Z_1 - \mu_1^0}{\sigma} \sim \mathcal{N}(0,1)$ CI: $Z_1 \pm \sigma z_{\alpha/2}$
2. σ^2 unknown, $d_1 = 1$: $\frac{Z_1 - \mu_1^0}{\hat{\sigma}} \sim t_{d_r}$ CI: $Z_1 \pm \hat{\sigma} t_{d_r}(\alpha/2)$
3. σ^2 known, $d_1 \geq 1$: $\frac{\|Z_1 - \mu_1^0\|_2^2}{\sigma^2} \sim \chi^2_{d_1}$ CI: $Z_1 + \sigma\sqrt{C_{\chi^2}(\alpha)}\mathbb{B}(0,1)$
4. σ^2 unknown, $d_1 \geq 1$: $\frac{\|Z_1 - \mu_1^0\|_2^2/d_1}{\hat{\sigma}^2} \sim F_{d_1,d_r}$ CI: $Z_1 + \hat{\sigma}\sqrt{C_F(\alpha)}\mathbb{B}(0,1)$

General linear model

Strategy for testing in general linear model

Generic setup

Observe $Y \sim \mathcal{N}(\theta, \sigma^2 \mathbb{I}_n), \sigma^2 > 0$

Test $\theta \in \Theta_0$ vs $\theta \in \Theta \backslash \Theta_0$, where $\Theta_0 \subseteq \Theta$, and Θ are subspaces of $\mathbb{R}^n$. $\dim(\Theta_0) = d_0$, $\dim(\Theta) = d = d_0 + d_1$.

Strategy for testing in general linear model

Generic setup

Observe $Y \sim \mathcal{N}(\theta, \sigma^2 \mathbb{I}_n), \sigma^2 > 0$
Test $\theta \in \Theta_0$ vs $\theta \in \Theta \backslash \Theta_0$, where $\Theta_0 \subseteq \Theta$, and Θ are subspaces of $\mathbb{R}^n$. $\dim(\Theta_0) = d_0$, $\dim(\Theta) = d = d_0 + d_1$.

Basic strategy: rotate into canonical form

$$Q = \left[\underbrace{Q_0}_{\text{orthonormal basis for } \Theta_0} \quad \underbrace{Q_1}_{\text{orthonormal basis for } \Theta \cap \Theta_0^\perp} \quad \underbrace{Q_r}_{\text{o.b. completed}} \right] \in \mathbb{R}^{n \times n}$$

$$Z = Q^\top Y \sim \mathcal{N}\left(\begin{pmatrix} Q_0^\top \theta \\ Q_1^\top \theta \\ 0 \end{pmatrix}, \sigma^2 \mathbb{I}_n \right), \quad H_0 : Q_1^\top \theta = 0$$

Example 1: linear regression

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

$$Y_i = x_i^\top \beta + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

Then

$$Y \sim \mathcal{N}(X\beta, \sigma^2 \mathbb{I}_n),$$

where $X \in \mathbb{R}^{n \times d}$. Assume X has full column rank.
$\theta = X\beta \in \Theta = \text{span}\,(X_1, \dots, X_d)$

$$H_0 : \beta_1 = \dots = \beta_{d_1} = 0, \quad (1 \le d_1 \le d)$$

is equivalent to $\theta \in \text{span}\left(X_{d_1+1}, \dots, X_d\right)$.

Relevant statistics

$$\begin{aligned}\hat{\beta}_{\mathsf{OLS}} &= \arg\min \|Y - X\beta\|_2^2 \\ &= (X^\top X)^{-1} X^\top Y\end{aligned}$$

$$\begin{aligned}\|Z_r\|_2^2 &= \left\|Y - \mathsf{Proj}_\Theta(Y)\right\|_2^2 \\ &= \left\|Y - X\hat{\beta}_{\mathsf{OLS}}\right\|_2^2 \\ &= \sum_{i=1}^n (Y_i - x_i^\top \hat{\beta}_{\mathsf{OLS}})^2 = \mathsf{RSS}\end{aligned}$$

$$\|Z_1\|_2^2 + \|Z_r\|_2^2 = \left\|Y - \mathsf{Proj}_{\Theta_0}(Y)\right\|_2^2 = \mathsf{RSS}_0 (\text{null RSS})$$

$$\text{F-statistic} = \frac{\|Z_1\|_2^2 / (d - d_0)}{\|Z_r\|_2^2 / (n - d)} = \frac{(\mathsf{RSS}_0 - \mathsf{RSS})/(d - d_0)}{\mathsf{RSS}/(n - d)}$$

$d_1 = 1$, reparametrization explained

Let $X_0 = (X_2, \ldots, X_d) \in \mathbb{R}^{d_0 \times n}$

Let $X_{1\perp} = X_1 - \text{Proj}_{\Theta_0}(X_1) = X_1 - X_0 \underbrace{(X_0^\top X_0)^{-1} X_0^\top X_1}_{\gamma}$

Reparameterization

$$\theta = X\beta \Leftrightarrow \theta = X_{1\perp}\beta_1 + X_0 \underbrace{(\beta_{-1} + \gamma)}_{\delta}$$

Let

$$q_1 = X_{1\perp} / \|X_{1\perp}\|_2 \, , Q_1 = \left(q_1\right) \in \mathbb{R}^{n \times 1}, Q_0 = U \in \mathbb{R}^{n \times d_0}$$

where U is obtained from SVD of $X_0 = U \Lambda V^\top$ not hard to see that $X_0 \left(X_0^\top X_0\right)^{-1} X_0^\top = UU^\top$

Check that we do obtain canonical linear model

Let $\hat{\beta}_1 = X_{1\perp}^\top Y / \|X_{1\perp}\|_2^2$
t-statistic is

$$\frac{q_1^\top Y}{\sqrt{\mathsf{RSS}/(n-d)}} = \frac{\hat{\beta}_1}{\hat{\mathsf{s.e.}}(\hat{\beta}_1)}$$

because s.e.$(\hat{\beta}_1) = \sigma / \|X_{1\perp}\|_2$ and RSS$/(n-d)$ is an unbiased estimate of σ^2.

Example 2: two-sample t-test with equal variance

Suppose $Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), Y_{m+1}, \dots, Y_{m+n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\nu, \sigma^2)$, parameter

$$\theta = \mathbb{E}[Y] = \begin{pmatrix} \mu \mathbf{1}_m \\ \nu \mathbf{1}_n \end{pmatrix}$$

Hypothesis

$$H_0 : \mu = \nu \quad \Leftrightarrow \theta \in \text{span}(\mathbf{1}_{n+m})$$

$d_0 = 1, d_1 = 1, d_r = n + m - 2$

$$q_1 = \frac{1}{\sqrt{\frac{1}{m} + \frac{1}{n}}} \begin{pmatrix} 1/m \\ \vdots \\ 1/m \\ -1/n \\ \vdots \\ -1/n \end{pmatrix}$$

Hence the t-statistic is

$$\frac{\frac{1}{m}\sum_{i=1}^{m} Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_{m+i}}{\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\mathsf{RSS}/(n+m-2)}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\hat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

Example 3: One-way ANOVA with fixed effects

Suppose

$$Y_{k,i} \overset{\text{ind.}}{\sim} \mu_k + \epsilon_{k,i}, \quad \epsilon_{k,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

$k = 1, \ldots, m$, $i = 1, \ldots, n$

Testing $H_0 : \mu_1 = \ldots = \mu_m = \mu$

$d_0 = 1$, $d_1 = m - 1$, $d_r = mn - m$

$$\bar{Y}_k = \frac{1}{n}\sum_i Y_{k,i}, \qquad S_k^2 = \frac{1}{n-1}\sum_i \left(Y_{k,i} - \bar{Y}_k\right)^2$$

$$\bar{Y} = \frac{1}{mn}\sum_k \sum_i Y_{k,i}, \qquad S_0^2 = \frac{1}{mn-1}\sum_k \sum_i \left(Y_{k,i} - \bar{Y}\right)^2$$

$$\mathsf{RSS} = \sum_{k,i} \left(Y_{k,i} - \bar{Y}_k\right)^2$$

$$\mathsf{RSS}_0 = \sum_{k,i} \left(Y_{k,i} - \bar{Y}\right)^2$$

$$\mathsf{RSS}_0 - \mathsf{RSS} = n\sum_k \left(\bar{Y}_k - \bar{Y}\right)^2$$

$$\text{F-statistic} = \frac{(\mathsf{RSS}_0 - \mathsf{RSS})/(d_1)}{\mathsf{RSS}/(d_r)} = \frac{\frac{n}{m-1}\sum_k \left(\bar{Y}_k - \bar{Y}\right)^2}{\frac{1}{mn-m}\sum_{k,i} \left(Y_{k,i} - \bar{Y}_k\right)^2}$$

between variance / within variance

Summary

Testing in general linear model

- The relevant distributions are $\mathcal{N}, \chi^2, t, F$
- The basic strategy is to find a change of basis so the problem is transformed to the canonical linear model

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

Testing $H_0 : \mu_1 = 0$ vs $H_1 : \mu_1 \neq 0$

What is next?

- Likelihood ratio test in large sample

Thank you