Econ 722 – Advanced Econometrics IV

Francis J. DiTraglia

University of Pennsylvania

Lecture #1 – Decision Theory

Lecture #2 – Model Selection I

Lecture #3 – Model Selection II

Lecture #4 – Asymptotic Properties

Lecture #5 – Andrews (1999) Moment Selection Criteria

Lecture #6 – Focused Moment Selection

Lecture #7 – High-Dimensional Regression I

Lecture #8 – High-Dimensional Regression II

Lecture #10 – Selective Inference

Optimal Inference After Model Selection (Fithian et al., 2017)

How Statistics is Done In Reality

Step 1: Selection – Decide what questions to ask.

"The analyst chooses a statistical model for the data at hand, and formulates testing, estimation, or other problems in terms of unknown aspects of that model."

Step 2: Inference – Answer the Questions.

"The analyst investigates the chosen problems using the data and the selected model."

Problem – "Data-snooping"

Standard techniques for (frequentist) statistical inference assume that we choose our questions before observing the data.

Simple Example: "File Drawer Problem"

$Y_i \sim$ iid $\mathrm{N}(\mu_i, 1)$ for $i = 1, \ldots, n$

- I want to know which $\mu_i \neq 0$, but I'm busy and n is big.
- My RA looks at each Y_i and finds the "interesting" ones, namely $\widehat{\mathcal{I}} = \{i : |Y_i| > 1\}$.
- I test $H_{0,i} : \mu_i = 0$ against the two-sided alternative at the 5% significance level for each $i \in \widehat{\mathcal{I}}$.

Two Questions

1. What is the probability of falsely rejecting $H_{0,i}$?
2. Among all $H_{0,i}$ that I test, what fraction are false rejections?

Simple Example: "File Drawer Problem"

$$\begin{aligned}
\mathbb{P}_{H_{0,i}}(\{\text{Reject } H_{0,i}\}) &= \mathbb{P}_{H_{0,i}}(\{\text{Test } H_{0,i}\} \cap \{\text{Reject } H_{0,i}\}) \\
&= \mathbb{P}_{H_{0,i}}(\{\text{Reject } H_{0,i}\}|\{\text{Test } H_{0,i}\})\mathbb{P}_{H_{0,i}}(\{\text{Test } H_{0,i}\}) \\
&= \mathbb{P}_{H_{0,i}}\left(\{|Y_i| > 1.96\}|\{|Y_i| > 1\}\right)\mathbb{P}_{H_{0,i}}(\{|Y_i| > 1\}) \\
&= \frac{2\Phi(-1.96)}{2\Phi(-1)} \times 2\Phi(-1) \\
&\approx 0.16 \times 0.32 \approx 0.05
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}_{H_{0,i}}(\{\text{Reject } H_{0,i}\}|\{\text{Test } H_{0,i}\}) &= \mathbb{P}_{H_{0,i}}\left(\{|Y_i| > 1.96\}|\{|Y_i| > 1\}\right) \\
&= \frac{\Phi(-1.96)}{\Phi(-1)} \approx 0.16
\end{aligned}$$

Simple Example: "File Drawer Problem"

Conditional vs. Unconditional Type I Error Rates

- The conditional probability of falsely rejecting $H_{0,i}$, given that I have tested it, is about 0.16.
- The unconditional probability of falsely rejecting $H_{0,i}$ is 0.05 since I only test a false null with probability 0.32.

Idea for Post-Selection Inference

Control the Type I Error Rate conditional on selection: "The answer must be valid, given that the question was asked."

Simple Example: "File Drawer Problem"

Conditional Type I Error Rate

Solve $\mathbb{P}_{H_{0,i}}(\{|Y_i| > c\}|\{|Y_i| > 1\}) = 0.05$ for c.

$$\begin{aligned}
\mathbb{P}_{H_{0,i}}(\{|Y_i| > c\}|\{|Y_i| > 1\}) = \frac{\Phi(-c)}{\Phi(-1)} &= 0.05 \\
c &= -\Phi^{-1}\big(\Phi(-1) \times 0.05\big) \\
c &\approx 2.41
\end{aligned}$$

Notice:

To account for the first-stage selection step, we need a larger critical value: 2.41 vs. 1.96. This means the test is less powerful.

Selective Inference vs. Sample-Splitting

Classical Inference

Control the Type I error under model M: $\mathbb{P}_{M,H_0}(\text{reject } H_0) \leq \alpha$.

Selective Inference

Control the Type I error under model M, given that M and H_0 were selected: $\mathbb{P}_{M,H_0}\big(\text{reject } H_0 | \{M, H_0 \text{ selected}\}\big) \leq \alpha$.

Sample-Splitting

Use different datasets to choose (M, H_0) and carry out inference: $\mathbb{P}_{M,H_0}\big(\text{reject } H_0 | \{M, H_0 \text{ selected}\}\big) = \mathbb{P}_{M,H_0}(\text{reject } H_0)$.

Selective Inference in Exponential Family Models

Questions

1. Recipe for selective inference in realistic examples?
2. How to construct the "best" selective test in a given example?
3. How does selective inference compare to sample-splitting?

Fithian, Sun & Taylor (2017)

- Use classical theory for exponential family models (Lehmann & Scheffé).
- Computational procedure for UMPU selective test/CI after arbitrary model/hypothesis selection.
- Sample-splitting is typically inadmissible (wastes information).
- Example: post-selection inference for high-dimensional regression

A Prototype Example of Selective Inference

This is my own example, but uses the same idea that underlies Fithian et al.

- Choose between two models on a parameter δ.
  - If $\delta \neq 0$, choose M1; if $\delta = 0$, choose M2
  - E.g. δ is the endogeneity of X, M1 is IV and M2 is OLS
- Observe $Y_\delta \sim N(\delta, \sigma^2_\delta)$ and use this to choose a model.
  - Selection Event: $A \equiv \{|Y_\delta| > c\}$, for some critical value c
  - If A, then choose M1. Otherwise, choose M2.
- After choosing a model, carry out inference for β.
  - Under a particular model M, $Y_\beta \sim N(\beta, \sigma^2_\beta)$
  - β is a *model-specific* parameter: could be meaningless or not even exist under a different model.
- If Y_β and Y_δ are correlated (under model M), we need to account for conditioning on A when carrying out inference for β.

All Calculations are Under a Given Model M

Key Idea

Under whichever model M ends up being selected, there is a joint normal distribution for Y_β and Y_δ *without* conditioning on A.

WLOG unit variances, ρ known

$$\begin{bmatrix} Y_\beta \\ Y_\delta \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \beta \\ \delta \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

As long as we can consistently estimate the variances of Y_β and Y_δ along with their covariance, this is not a problem.

Selective Inference in a Bivariate Normal Example

$$\boxed{\begin{bmatrix} Y_\beta \\ Y_\delta \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \beta \\ \delta \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad A \equiv \{|Y_\delta| > c\}}$$

Two Cases

1. Condition on A occurring
2. Condition on A *not* occurring

Problem

If δ were known, we could directly calculate how conditioning on A affects the distribution of Y_β, but δ is unknown!

Solution

Condition on a sufficient statistic for δ.

Conditioning on a Sufficient Statistic

Theorem

If U is a sufficient statistic for δ, then the joint distribution of (Y_β, Y_δ) given U does not depend on δ.

In Our Example

Residual $U = Y_\delta - \rho Y_\beta$ from a projection of Y_δ onto Y_β is sufficient for δ.

Straightforward Calculation

$$\left[\begin{array}{c} Y_\beta \\ Y_\delta \end{array}\right] \Bigg| (U = u) = \left[\begin{array}{c} \beta + Z \\ u + \rho(\beta + Z) \end{array}\right], \quad Z \sim \text{N}(0,1)$$

Notice that this is a singular normal distribution

The Distribution of $Y_\beta|(A, U = u)$

$$\boxed{\begin{bmatrix} Y_\beta \\ Y_\delta \end{bmatrix} \Bigg| (U = u) = \begin{bmatrix} \beta + Z \\ u + \rho(\beta + Z) \end{bmatrix}, \quad Z \sim \mathrm{N}(0, 1)}$$

Start with case in which A occurs so we select $M1$. Under $H_0\colon \beta = \beta_0$,

$$\begin{aligned} \mathbb{P}_{\beta_0}\left(Y_\beta \leq y | A, U = u\right) &= \frac{\mathbb{P}_{\beta_0}(\{Y_\beta \leq y\} \cap A | U = u)}{\mathbb{P}_{\beta_0}(A|U = u)} \\ &= \frac{\mathbb{P}\left(\{Z \leq y - \beta_0\} \cap \{|u + \rho(\beta_0 + Z)| > c\}\right)}{\mathbb{P}\left(|u + \rho(\beta_0 + Z)| > c\right)} \end{aligned}$$

$\mathbb{P}(A|U = u)$ under H_0: $\beta = \beta_0$

$$
\begin{aligned}
P_D(A) &\equiv P_{\beta_0}(A|U = u) \\
&= \mathbb{P}\left(|u + \rho(\beta_0 + Z)| > c\right) \\
&= \mathbb{P}\left[u + \rho(\beta_0 + Z) > c\right] + \mathbb{P}\left[u + \rho(\beta_0 + Z) < -c\right] \\
&= \mathbb{P}\left[\rho(\beta_0 + Z) > c - u\right] + \mathbb{P}\left[u + \rho(\beta_0 + Z) < -c - u\right] \\
&= 1 - \Phi\left(\frac{c - u}{\rho} - \beta_0\right) + \Phi\left(\frac{-c - u}{\rho} - \beta_0\right)
\end{aligned}
$$

$\mathbb{P}(\{Y_\beta \leq y\} \cap A | U = u)$ under $H_0\colon \beta = \beta_0$

$$
\begin{aligned}
P_N(A) &\equiv \mathbb{P}(\{Y_\beta \leq y\} \cap A | U = u) \\
&= \mathbb{P}\left(\{Z \leq y - \beta_0\} \cap \{|u + \rho(\beta_0 + Z)| > c\}\right) \\
&= \begin{cases}
\Phi\left(y - \beta_0\right), & y < (-c - u)/\rho \\
\Phi\left(\dfrac{-c - u}{\rho} - \beta_0\right), & (-c - u)/\rho \leq y \leq (c - u)/\rho \\
\Phi(y - \beta_0) - \Phi\left(\dfrac{c - u}{\rho} - \beta_0\right) + \Phi\left(\dfrac{-c - u}{\rho} - \beta_0\right), & y > (c - u)/\rho
\end{cases}
\end{aligned}
$$

$F_{\beta_0}(y|A, U = u)$

Define $\ell(u) = (-c - u)/\rho$, $r(u) = (c - u)/\rho$. We have:

$$F_{\beta_0}(y|A, U = u) = P_N(A)/P_D(A)$$

where

$$P_D(A) \equiv 1 - \Phi\left(r(u) - \beta_0\right) + \Phi\left(\ell(u) - \beta_0\right)$$

$$P_N(A) \equiv \begin{cases} \Phi\left(y - \beta_0\right), & y < \ell(u) \\ \Phi\left(\ell(u) - \beta_0\right), & \ell(u) \leq y \leq r(u) \\ \Phi(y - \beta_0) - \Phi\left(r(u) - \beta_0\right) + \Phi\left(\ell(u) - \beta_0\right), & y > r(u) \end{cases}$$

Note that $F_{\beta_0}(y|A, U = u)$ has a *flat region* where $\ell(u) \leq y \leq r(u)$

$Q_{\beta_0}(p|A, U = u)$

Inverting the CDF from the preceding slide:

$$Q_{\beta_0}(p|A, U = u) = \begin{cases} \beta_0 + \Phi^{-1}\left(p \times P_D(A)\right), & p < p^* \\ \beta_0 + \Phi^{-1}\left[p \times P_D(A) + \Phi\left(r(u) - \beta_0\right) - \Phi\left(\ell(u) - \beta_0\right)\right], & p \geq p^* \end{cases}$$

where

$$\begin{aligned} p^* &\equiv \Phi\left(\ell(u) - \beta_0\right)/P_D(A) \\ P_D(A) &\equiv 1 - \Phi\left(r(u) - \beta_0\right) + \Phi\left(\ell(u) - \beta_0\right) \\ \ell(u) &\equiv (-c - u)/\rho \\ r(u) &\equiv (c - u)/\rho \end{aligned}$$

The Distribution of $Y_\beta | (A^c, U = u)$

$$\boxed{\left[\begin{array}{c} Y_\beta \\ Y_\delta \end{array}\right] \Bigg| (U = u) = \left[\begin{array}{c} \beta + Z \\ u + \rho(\beta + Z) \end{array}\right], \quad Z \sim \mathrm{N}(0,1)}$$

If A does not occur, when we select $M2$. Under $H_0\colon \beta = \beta_0$,

$$\begin{aligned}
\mathbb{P}_{\beta_0}\left(Y_\beta \leq y | A^c, U = u\right) &= \frac{\mathbb{P}_{\beta_0}(\{Y_\beta \leq y\} \cap A^c | U = u)}{\mathbb{P}_{\beta_0}(A^c | U = u)} \\
&= \frac{\mathbb{P}\left(\{Z \leq y - \beta_0\} \cap \{|u + \rho(\beta_0 + Z)| < c\}\right)}{\mathbb{P}\left(|u + \rho(\beta_0 + Z)| < c\right)}
\end{aligned}$$

$F_{\beta_0}(y|A^c, U = u)$

As above, define $\ell(u) = (-c - u)/\rho$, $r(u) = (c - u)/\rho$. We have:

$$F_{\beta_0}(y|A^c, U = u) = P_N(A^c)/P_D(A^c)$$

where

$$P_D(A^c) \equiv \Phi\left(r(u) - \beta_0\right) - \Phi\left(\ell(u) - \beta_0\right)$$

$$P_N(A^c) \equiv \begin{cases} 0, & y < \ell(u) \\ \Phi(y - \beta_0) - \Phi\left(\ell(u) - \beta_0\right), & \ell(u) \leq y \leq r(u) \\ \Phi\left(r(u) - \beta_0\right) - \Phi\left(\ell(u) - \beta_0\right), & y > r(u) \end{cases}$$

Notice that this is a CDF with a bounded support set: $y \in [\ell(u), r(u)]$

$Q_{\beta_0}(p|A^c, U = u)$