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 $\widehat{\mathcal{I}} = \{i \colon |Y_i| > 1\}$.
- I test $H_{0,i} \colon \mu_i = 0$ 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"

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

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

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

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

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 models M1 and M2 based on 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 $Y_\delta \sim N(\delta, 1)$ 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, $Y_\beta \sim N(\beta, 1)$
  - β is a *model-specific* parameter