1 Outline

Agenda: Selective Inference

1. Want inference to hold when selecting promising leads/models after data snooping (accounting for "researcher’s degrees of freedom")
2. Not classical (subject still largely excluded from textbooks)

In the next two lecture, we will mainly look at two schemes of post selection inference:

- Post-selection inference (U. Penn group)
- Post-selection inference with the LASSO (Stanford group)

2 POSI (Post Selection Inference)


2.1 Problem Formulation

Let us consider a selection problem. We assume the data follows a linear model

\[ y = X\beta + z \]

In the classical setting, we specify a model and then fit the model with our data. However, in reality the data analyst selects model after viewing data, in which case a selection bias is introduced. Therefore, we hope to develop a scheme so that we can still provide inference about parameters in selected model.

Before proceeding, there are a few assumptions on the distributions:

- \( X \) is a \( n \times p \) design matrix
- \( \sigma \) is known (for convenience)

In reality \( \sigma \) is unknown and POSI requires an 'independent' estimate of \( \sigma \) - think \( p < n \) and \( \hat{\sigma}^2 = \text{MSE}_{\text{full model}} \). As an extension, we can also consider the case where \( \mu \notin \text{span}(X) \).
2.2 Classical Inference

In the classical setting, there is a fixed model $M \subset \{1, \ldots, p\}$. In this setting, the object of inference is the slopes after adjusting for variables in $M$ only:

$$\beta_M = X_M^\dagger \mu = \mathbb{E}X_M^\dagger y$$

where

- $X_M^\dagger = (X_M^T X_M)^{-1}X_M^T$
- $\hat{\beta}_M = X_M^\dagger y$ is the least square estimate.

With $M$ fixed, the sampling distribution is

$$\hat{\beta}_M \sim \mathcal{N}(\beta_M, \sigma^2(X_M^T X_M)^{-1})$$

And the corresponding z-score is

$$z_{j\cdot M} = \frac{\hat{\beta}_{j\cdot M} - \beta_{j\cdot M}}{\sigma \sqrt{(X_M^T X_M)^{-1}_{jj}}} = \frac{(y - \mu)^T X_{j\cdot M}}{\sigma \|X_{j\cdot M}\|} \sim \mathcal{N}(0, 1)$$

where $X_{j\cdot M}$ is equal to $\text{lm}(X[\cdot j] - X[\cdot \text{setdiff}(M, j)])$ resid. From this statistic we can have valid confidence intervals:

$$\hat{\beta}_{j\cdot M} \pm z_{1-\alpha/2}\|X_{j\cdot M}\|$$

If $\hat{\sigma}^2 = \text{MSE}_\text{Full}$, then $\hat{\beta}_{j\cdot M} \pm t_{n-p,1-\alpha/2}\hat{\sigma}\|X_{j\cdot M}\|$. The problem arises when there is model selection. Here we present an example from A. Buja:

$$y = \beta_0 x_0 + \sum_{j=1}^{10} \beta_j x_j + z_j \quad n = 250 \quad z_j \overset{iid}{\sim} \mathcal{N}(0, 1)$$

- We are interested in the confidence interval for $\beta_0$.
- Select model always including $x_0$ via BIC.

Figure 1 shows the marginal distribution of post-selection $t$-statistics (and the red line corresponds to $\mathcal{N}(0, 1)$). For $p = 30$, the coverage rate of confidence intervals constructed in the classical manner is 39% instead of 95%.

2.3 POSI scheme

Given a variable selection procedure $\hat{M}(y)$, we can come up with different sorts of selective inference, for example:

$$\mathbb{P}(j \in \hat{M} \text{ and } \beta_{j\cdot\hat{M}} \in C_{j\cdot\hat{M}}) \geq 1 - \alpha$$

$$\mathbb{P}(\beta_{j\cdot\hat{M}} \in C_{j\cdot\hat{M}}|j \in \hat{M}) \geq 1 - \alpha$$

$$\mathbb{P}(\forall j \in \hat{M} : \beta_{j\cdot\hat{M}} \in C_{j\cdot\hat{M}}) \geq 1 - \alpha$$
Here the object of inference random, and for different variable selection procedure, we have different $P(j \in \hat{M})$, thus different confidence intervals. It is not at all obvious how to construct such confidence intervals.

For POSI, the confidence interval corresponds to the universal validity for all selected procedures:

$$\forall M \quad P(\forall j \in \hat{M} : \beta_j \hat{M} \in C_{j \hat{M}}) \geq 1 - \alpha$$

There are a few pros and cons regarding this approach:

- **Pros**: simultaneous inference: strongest form of protection (no matter what the data scientist did)
- **Cons**: CI's can be very wide (later)
- **Merit**: got lots of people thinking...

We then may want to ask the question: is POSI doable?

Recall that

$$z_{j \hat{M}} = \frac{(y - \mu)^TX_{j \hat{M}}}{\sigma ||X_{j \hat{M}}||} \sim N(0, 1).$$

For any variable selection procedure,

$$\max_{j \in \hat{M}} |z_{j \hat{M}}| \leq \max_M \max_{j \in M} |z_{j M}|$$

From here we have the following theorem:
Theorem 1 (Universal guarantee).

\[ P(\max_{M} \max_{j \in M} z_{j \cdot M} \leq K_{1-\alpha/2}) \geq 1 - \alpha \]

where \( K_{1-\alpha/2} \) is POSI constant. Then with \( C_{j \cdot \hat{M}} = \hat{\beta}_{j \cdot \hat{M}} \pm K_{1-\alpha/2} \sigma \| X_{j \cdot \hat{M}} \| \)

\[ \forall \hat{M} \quad P(\forall j \in \hat{M} : \hat{\beta}_{j \cdot \hat{M}} \in C_{j \cdot \hat{M}}) \geq 1 - \alpha \]

The key in this universal guarantee theorem is to compute the POSI constant. The POSI constant is the quantile of the random variable

\[ \max_{M} \max_{j \in M} |z_{j \cdot M}| \]

The difficulty in calculating the quantile is that we have to look at \( 2^p \) models. As an alternative, we can try to develop (asymptotic) bounds.

It turns out that the range of POSI constant is as follows:

\[ \sqrt{2 \log(p)} \lesssim K_{1-\alpha}(X) \lesssim \sqrt{p} \]

- The lower bound is achieved for orthogonal designs
- The upper bound is achieved for SPAR1 designs
- POSI constant can get very large (but necessarily so)

Here are a few conclusions w.r.t POSI:

- Provides protection against all kinds of selection
- Can be conservative
- Is perhaps difficult to implement (how do you compute the POSI constant?)
- Alternative: split sample (for example independent samples; but not always possible)

A significant impact of POSI is that it asked important questions and stimulated lots of thinking/questioning/research.