Post-Selection Inference

PS690 Computational Methods in Social Science

Jiawei Fu
Department of Political Science
Duke University

September 2, 2025

Overview

Classical Inference

- Consider the low-dimension linear regression model ($p < n$),

$$Y = X'\beta + e$$

- To make inference of the OLS estimator $\hat{\beta} = (\frac{1}{n}\sum_{i=1}^{n} X_i X_i')^{-1}(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i)$, we derive the distribution of the estimator.
- Multiply be $\sqrt{n}$,

$$\sqrt{n}(\hat{\beta} - \beta) = (\frac{1}{n}\sum_{i=1}^{n} X_i X_i')^{-1}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i)$$

- By CLT, the second part $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i \rightarrow_d N(0, \Omega)$, where $\Omega = \mathbb{E}[XX'e^2]$
- By WLLN, the first part $\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \rightarrow_p \mathbb{E}[XX']$
- Therefore, $\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, V_\beta)$, where $V_\beta = \mathbb{E}[XX']^{-1}\Omega\mathbb{E}[XX']$

Two types of Inference revolving Post-Selection

- Assume a high-dimensional structural model: $Y = X\beta + \epsilon$.
- One natural target of inference is the *structural (population) coefficient* β_j (j^{th} component).
- Let $\mathcal{M}$ denote the universe of all possible models. For $M \in \mathcal{M}$, we can also define the *submodel coefficient*, $\beta_{j \cdot M}$, which depends on the submodel M.
- In practice, researcher often use data to select a model, $\widehat{M}$, and obtain the corresponding estimator $\hat{\beta}_{\widehat{M}}$. Therefore $\widehat{M}$ is random.
- We can use estimator $\hat{\beta}_{\widehat{M}}$ to construct CI for target $\beta_{\widehat{M}}$ (note it is random), which focus on the selected (random) model $\widehat{M}$, rather than the population parameter β_j.
- We can also use post-selected estimator $\hat{\beta}_{\widehat{M}}$ to conduct statistical inference for population parameter β_j. In this view, model selection is simply a kind of regularization which provide a lower dimensional estimator for the high dimensional parameter.

Example

- For example, let $M = \{1, 2, ..., p\}$ be the index set of all the predictors, and assume $Y = X\beta + \epsilon$.
- The structural parameter is β_M:

$$\beta_M := argmin_{\beta \in \mathbb{R}^{|M|}} \mathbb{E}[(Y_i - X'_{i,M}\beta)^2]$$

- Researchers use data to select some predictors (through different methods like forward stepwise regression and lasso), $\widehat{M} \subseteq M$; and we can get the submodel OLS estimator $\hat{\beta}_{\widehat{M}} = (X'_{\widehat{M}} X_{\widehat{M}})^{-1} X_{\widehat{M}} Y$
- The target of this submodel OLS estimator is $\beta_{\widehat{M}}$: Hope to find confidence interval $\widehat{CI}_{\widehat{M}}$ satisfying

$$\liminf_{n \to \infty} \mathbb{P}[\beta_{\widehat{M}} \in \widehat{CI}_{\widehat{M}}] \geq 1 - \alpha$$

Problem of Ignorance of Model selection

- Without selection, we know $\hat{\beta}_M$ behaves nicely: asymptotically normal at a $\sqrt{n}-$rate. (Recall the OLS estimator on page 3).
- However, for $\hat{\beta}_{\widehat{M}}$ with a data-driven choice of $\widehat{M}$, there is also some randomness through $\widehat{M}$.
- Due to data exploration, $\hat{\beta}_{\widehat{M}}$ generally does not have a normal distribution and can be quite biased, even asymptotically.

Problem of Ignorance of Model selection

- Consider we select the model by forward stepwise regression.
- In the simulation, we create three predictors $X = (X_1, X_2, X_3)$, and the response Y is drawn from $N(1, 9)$, independent of X. Therefore, the true $\beta_M = 0$.

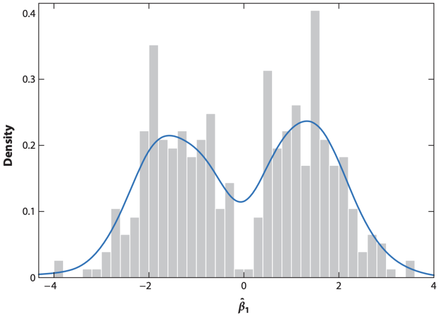

Figure: [Kuchibhotla et al., 2022]

The bimodal distributions are expected because X_1 is selected by the variable selection strategy only when it has a reasonably large coefficient in absolute value.

Problem of Ignorance of Model selection

- See another example of LASSO. Here, we look at the t statistics: $T_{j\cdot\widehat{M}} = \frac{\hat{\beta}_{j\cdot\widehat{M}} - \beta^0}{sd(\hat{\beta}_{j\cdot\widehat{M}})}$ and $T_{j\cdot\widehat{M}} = \frac{\hat{\beta}_{j\cdot\widehat{M}}}{sd(\hat{\beta}_{j\cdot\widehat{M}})}$.

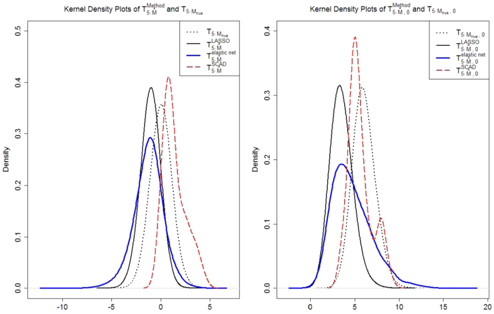

Figure: [Zhang et al., 2022]

Problem of Ignorance of Model selection

- Let us also check the coverage of the confidence interval.

Empirical coverage probabilities of 95% naïve confidence intervals for the true non-zero β_j^0's.

	Empirical coverage probabilities				
Model Selector + Estimation	β_1^0	β_3^0	β_4^0	β_5^0	β_7^0
LASSO + OLS	.691	.486	.695	.736	.560
Elastic net + OLS	.749	.493	.815	.867	.619
SCAD + OLS	.362	.736	.682	.727	.723

Figure: [Zhang et al., 2022]

Data Split

- Now, let us look at some methods that target the population parameter β_j.
- The key problem of post-selection inference is that we use the same data select model and do inference.
- A natural idea is to split the data so that select the model on one half and estimate the coefficient given the model on the second half.
- For the second half, the model is given; therefore, the traditional inference works here.
- For LASSO, we assume the sparsity true model. In the first half, we select non-zero predictors S. In the second half, if $n/2 > |S|$ we just run linear regression and obtain the traditional OLS estimator.
- For those $j \notin S$, we set $p = 1$ for test $H_{0,j} : \beta_{j \cdot M} = 0$

Data Split

- One issue of the single-sample-splitting method is its sensitivity with respect to the choice of splitting the entire sample: sample splits lead to wildly different p-values.
- Similarly to cross-validation K, we can do multiple sample splits.
- Sample splitting is invalid for dependent data. It inherently assumes independence of observations in the data
- Note: Data Split can also be used to conduct post-selection inference.

De-biasing LASSO

- Recall OLS estimator, we can write it as $\hat{\beta}_j^{ols} = \frac{Y' X_j^{\perp}}{X_j' X_j^{\perp}}$, where $X_j^{\perp}$ is the residual in the regression of X_j on X_{-j}. (Recall Frisch–Waugh–Lovell theorem).
- Put in Y, we get $\frac{Y' X_j^{\perp}}{X_j' X_j^{\perp}} = \beta_j + \frac{\epsilon' X_j^{\perp}}{X_j' X_j^{\perp}}$
- We obtain: $\sqrt{n}(\hat{\beta}_j^{ols} - \beta_j) = \frac{n^{-1/2} \epsilon' X_j^{\perp}}{n^{-1} X_j' X_j^{\perp}}$, which is asymptotically normal.
- In LASSO, we use a lasso regression with a regularization parameter λ to get $X_j^{\perp}$.
- Algebra shows that

$$\frac{Y' X_j^{\perp}}{X_j' X_j^{\perp}} = \beta_j + \sum_{k \neq j} P_{jk} \beta_j + \frac{\epsilon' X_j^{\perp}}{X_j' X_j^{\perp}}$$

  where $P_{jk} = \frac{X_k' X_j^{\perp}}{X_j' X_j^{\perp}}$
- The middle part is the bias. Naturally, we can correct for this term.

De-biasing LASSO

- Consider the estimator $\hat{b}_j = \frac{Y' X_j^{\perp}}{X_j' X_j^{\perp}} - \sum_{k \neq j} P_{jk} \hat{\beta}_j$
- Similarly, we obtain

$$\sqrt{n}(\hat{b}_j - \beta_j) = \frac{n^{-1/2} \epsilon' X_j^{\perp}}{n^{-1} X_j' X_j^{\perp}} + \sum_{k \neq j} \sqrt{n} P_{jk} (\beta_k - \hat{\beta}_k)$$

- The first term converges normal. The second term is negligible under some regular conditions.
- This implies that we can conduct valid inference under normal distribution (asymptotically).
- Implementation: R package hdi, by [Dezeure et al., 2015]

PoSI

- Previous methods target population β_j in the structural model: $Y = X\beta + \epsilon$.
- [Berk et al., 2013] argues that we should focus on the submodel $\beta_{j\cdot\widehat{M}}$.
- Given a submodel $\widehat{M}$ selected by a generic model selection procedure, we consider the following CI s.t.

$$CI_{j\cdot\widehat{M}}(K) = (\hat{\beta}_{j,\widehat{M}} - K\hat{\sigma}\sqrt{[(X'_{\widehat{M}}X_{\widehat{M}})^{-1}]_{jj}}, \hat{\beta}_{j,\widehat{M}} + K\hat{\sigma}\sqrt{[(X'_{\widehat{M}}X_{\widehat{M}})^{-1}]_{jj}})$$

- If $K = t(n - |\widehat{M}|, 1 - \alpha/2)$, we obtain the naive confidence interval, which ignores the uncertainty from model selection; The CI is too short.
- Therefore, we hope to find a large K so that the CI is wider and the coverage be at least $1 - \alpha$.

PoSI

- To do this, we first construct simultaneous confidence intervals for all possible selected model (i.e.*regardless of model selection procedures and selected models.*):
- In other words, we hope to find CI s.t.

$$\mathbb{P}[\beta_{j\cdot M} \in CI_{j\cdot M}(K), \forall j \in M, M \in \mathcal{M}] \geq 1-\alpha$$

- If this is true, it implies that $\mathbb{P}[\beta_{j\cdot \widehat{M}} \in CI_{j\cdot \widehat{M}}(K), \forall j \in M] \geq 1-\alpha$ and $\mathbb{P}[\beta_{j\cdot \widehat{M}} \in CI_{j\cdot \widehat{M}}(K)] \geq 1-\alpha$.
- To obtain simultaneous control over all possible submodels, we need to find the largest value of K.
- Note that $\beta_{j\cdot \widehat{M}} \in CI_{j\cdot \widehat{M}}(K)$ is equivalent to $|\frac{\hat{\beta}_{j\cdot \widehat{M}}-\beta_{j,\widehat{M}}}{\hat{\sigma}\sqrt{[(X'_{\widehat{M}}X_{\widehat{M}})^{-1}]_{jj}}}| \leq K$. We use $t_{j\cdot \widehat{M}}$ to denote that ratio.
- [Berk et al., 2013] propose the following K_{PoSI} such that

$$K_{PoSI} = \min\{K \in \mathbb{R} | \mathbb{P}[\max_{M\in\mathcal{M}} \max_{j\in\widehat{M}} |t_{j\cdot \widehat{M}}| \leq K] \geq 1-\alpha\}$$

Advantages and Disadvantages

- Advantages: We can get the correct coverage regardless of the selection method.
- It can be applied to dependent data as well.
- Disadvantage:Because it is safeguarded against all possible selected submodels, it is necessarily conservative.

Conditional Selective Inference

- The previous PoSI method considers the simultaneous control.
- In some cases, we can derive the condition distribution of $\hat{\beta}_{\widehat{M}}$ giving $\widehat{M}$.
- For example, [Lee et al., 2016] shows that, under some conditions, the conditional distribution of the LASSO estimator is essentially a (univariate) truncated Gaussian.

Further Methods

- Given time constraints, we do not cover many other popular methods, for example, among others,
  1. Bayesian inference
  2. Bootstrap
  3. Methods from optimization theory
- We will cover inference on the treatment effects when using lasso in the later lectures.

References

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013).
Valid post-selection inference.
The Annals of Statistics, pages 802–837.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015).
High-dimensional inference: confidence intervals, p-values and r-software hdi.
Statistical science, pages 533–558.

Kuchibhotla, A. K., Kolassa, J. E., and Kuffner, T. A. (2022).
Post-selection inference.
Annual Review of Statistics and Its Application, 9:505–527.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016).
Exact post-selection inference, with application to the lasso.
The Annals of Statistics.

Zhang, D., Khalili, A., and Asgharian, M. (2022).
Post-model-selection inference in linear regression models: An integrated review.
Statistic Surveys, 16:86–136.