BIOS 835: Post-Selective Inference

Alexander McLain

September 14, 2023

Movitation

- Setting: low-dimensional linear regression.
- Method: backward selection is used to trim 20 variables down to 5.
- Question: how do you construct confidence intervals (CIs) or perform hypothesis testing?

Post-selective inference

- Post-selective inference refers to the problem of making statistical inferences (like confidence intervals or hypothesis tests) after a model selection procedure has been applied.
- The challenge arises because traditional inferential procedures assume that the model structure (i.e., which predictors are included) is fixed in advance.
- The whole idea of model selection is to select the model structure based on the data.
- Problems:
  - Constructing CIs or performing hypothesis testing using standard methods after model selection leads to invalid inference (CI to narrow, increased type I errors).
  - Many procedures don't have "standard methods".

Approaches

- **Invariance-based methods** for post-selective inference leverage the properties of certain statistics that remain invariant under model selection procedures.
- **Data Splitting** split the data into two parts. Use one part for model selection (e.g., determining which variables are non-zero with Lasso) and the other part for inference.
- **Selective Inference Tools:** These methods condition on the selection event (i.e., which variables were selected) when performing inference.
- **Conformal Inference:** a very general tool for providing prediction intervals from any machine learning approach (any). It involves multiple refits of the model and can be computationally intensive.

Data Splitting

- We'll discuss data splitting for the standard LASSO model:

$$Y = X\beta + \epsilon$$

  where Y is the $n \times 1$ response vector and

$$\hat{\beta} = \arg\min_{\beta \in R^p} \left\{ ||Y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}$$

- Given a dataset of size n, randomly partition it into two subsets:
  - a **selection set** of size m and
  - an **inference set** of size $n - m$.

Data Splitting with the LASSO

- Designate matrices and vectors: X_{sel}, Y_{sel} for the selection set and X_{inf}, Y_{inf} for the inference set.
- Fit a LASSO regression model to X_{sel} and Y_{sel}:

$$\hat{\beta}_{sel} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2m} ||Y_{sel} - X_{sel}\beta||_2^2 + \lambda ||\beta||_1 \right\}$$

- The optimal λ value can be chosen using cross-validation within the selection set.
- Identify the set of predictors, S, where $\hat{\beta}_{sel}$ has non-zero coefficients.

Data Splitting with the LASSO

- Using the predictors in S, fit an ordinary least squares (OLS) regression model to X_{inf} and Y_{inf}.
- This will give unbiased estimates and valid statistical inferences, such as p-values and confidence intervals, since this dataset was not used in the model selection process.

Data Splitting

- Data splitting is a simple method that results in valid inference since there is no “double dipping” of the data.
- However, because the data is split, the methods can suffer from reduced statistical power, especially if the dataset isn’t large.
- Extensions:
  - consider multiple rounds of data splitting, the results from the different partitions can then be aggregated (e.g., by averaging) to obtain more robust estimates and inferences.
  - Advanced methods like cross-conformal prediction combine ideas from both data splitting and cross-validation.

Selective Inference Tools

- Selective inference refers to the practice of making statistical inferences that are valid not just conditionally on the selected model but also considering the entire model selection process.
- These methods require complex mathematical derivations of the distribution of model quantities.
- We'll go through the basic idea for lasso, just to give you the overall picture.

Selective Inference Concepts

- Suppose that $\mathrm{y} \sim \mathrm{N}(\mu, 1)$ and estimate μ only if:

$$|y_i| \geq c > 0$$

- If $0 < \mu < \mathrm{C}$ we will always overestimate μ if we use the standard MLE, y itself.

Selective Inference Tools

- The LASSO solution can be characterized by the Karush-Kuhn-Tucker (KKT) optimality conditions.
- The KKT conditions for LASSO can be written in the form:

  $$X_j^T(y - X\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_j)$$

  for all predictors j in the selected set S, and

  $$|X_j^T(y - X\hat{\beta})| \leq \lambda$$

  for all predictors j not in S.
- These conditions describe the relationships between the residuals, the predictors, and the LASSO estimates.

Selective Inference Tools

- Leveraging the KKT conditions, one can derive that, conditionally on the selection event E, the centered vector $X_j^T(y - X\hat{\beta})$ follows a truncated Gaussian distribution.
- The truncation intervals depend on the selection event E and the signs of the active predictors in S.
- Given the truncated Gaussian framework, one can compute the selective p-values for testing:

  $$H_0 : \beta_j = 0$$

  against a two-sided alternative for any predictor j in S, using the tail probabilities of the truncated Gaussian distribution.

Selective Inference Tools

- The key challenge is computing these p-values efficiently, especially in high dimensions.
- Several algorithms have been proposed that provide approximations to the exact selective p-values, leveraging properties of Gaussian random vectors and convex optimization.
- `selectiveInference`, `selectiveMLE`, and `PSAT` are packages in `R` with various selective inference tools.

Selective Inference Tools

Pros:

- Valid inferences that account for Model Selection Uncertainty
- Flexibility: The concept behind selective inference can be applied beyond the LASSO to other variable selection procedures.

Cons:

- Computational Complexity
- Not Always Intuitive
- Methodological Limitations

Conformal inference

- Why prediction inference?
- Conformal inference is a general non-parametric approach to statistical inference that aims to provide valid coverage in finite samples, regardless of the underlying data-generating distribution.
- When combined with penalized linear regression, like LASSO, the goal is to build prediction intervals or hypothesis tests that have valid coverage properties, even when the model selection procedure is taken into account.

Conformal inference

- Calibration Set and Exchangeability:
    - Split the data randomly into two parts: a calibration set and a training set.
    - Under the assumption of exchangeability (which is a weaker assumption than i.i.d.), any permutation of the labels of the samples would have the same joint distribution.
- Nonconformity Score:
    - A nonconformity score is a measure of how "strange" or "atypical" an observation is, compared to the rest.
    - For regression, a typical nonconformity score might be the absolute residual:

      $$\alpha_i = |y_i - \hat{y}_i|$$

      where $\hat{y}_i$ is the prediction of the LASSO or other penalized regression model for the i-th sample.

Conformal inference

- Compute Nonconformity Scores for the Calibration Set:
  - Fit the penalized regression model to the training set.
  - Use this model to predict the outcomes in the calibration set.
  - Compute the nonconformity scores for each observation in the calibration set.
- This results in $\alpha_1, \alpha_2, \ldots, \alpha_{n.cal}$. Let's assume

$$\alpha_1 < \alpha_2 < \ldots < \alpha_n$$

where n is the size of the calibration set.

Conformal inference

Prediction Intervals:

- For a new observation, fit the model to the combined training and calibration set $\rightarrow \hat{y}_{new}$
- Predict the outcome for the new observation and compute its nonconformity score $\rightarrow \alpha_{new}$.
- Note that
$$\Pr(\alpha_{new} \leq \alpha_{\lceil (n+1)(1-\alpha) \rceil}) \geq 1 - \alpha$$
if $\lceil (n+1)(1-\alpha) \rceil \leq n$.

Conformal inference

Prediction Intervals:

- Let

$$\pi_{new}(y) = \frac{1}{n+1}\left\{1 + \sum_{i=1}^{n} I(|y - \hat{y}_{new}| \leq \alpha_i)\right\}$$

  be a function that finds the proportion of α's that are less than what α_{new} if $y_{new} = y$
- This is used to construct the conformal prediction interval is essentially

$$C_{\text{conf}}\ (X_{new}) = \{y : \pi_{new}(y) \leq (1 - \alpha)\}$$

  or the prediction interval is

$$[\hat{y}_{new} - \alpha_{\lceil (n+1)(1-\alpha)\rceil}, \hat{y}_{new} + \alpha_{\lceil (n+1)(1-\alpha)\rceil}]$$

Conformal inference

- If the exchangeability assumption holds, then conformal prediction intervals have valid coverage in finite samples.
- Notice we didn't use the model here. Conformal inference can be used with almost any approach.
- A challenge of conformal inference with penalized regression is that the penalty term and the variable selection process can influence the nonconformity scores.
- Care must be taken when doing CV.
- Conformal inference is computationally more demanding due to the need to repeatedly fit the model on various subsets of the data.