Lecture 17: Feature Selection

Ailin Zhang

2023-06-21

Summary: How to address non-constant variance?

- Stabilize Variance
  - Variance-Stabilizing Transformation
  - Box-Cox Method
  - Weighted Least Square
- Bootstrap

The Bootstrap Method

- Bootstrap is a very useful method. It provides another remedy for the nonconstant variance problem.
- More generally, the Bootstrap provides a computationally intensive alternative method used primarily for computing standard errors, confidence intervals, and tests when either the assumptions needed for standard methods are questionable, or where standard methods are not readily available.

Idea of the Bootstrap

- Suppose we have a sample $y_1, \ldots, y_n$ from distribution G. What is a confidence interval for the population median?
- If we know G, this can be done by simulation from G as follows.
  1. Sample $y_1^*, \ldots, y_n^*$ from G.
  2. Compute and save the median of the sample in step 1.
  3. Repeat steps 1 and 2 for B times. (The larger the value of B, the more precise the ultimate answer.)
  4. To construct 95% confidence interval, take the 2.5 and 97.5 percentiles of sample medians.
- However, we usually do not know G and so directly simulation is not available. Efron (1979) pointed out that the observed data ($\hat{G}$) can be used to estimate G, and then we can sample from the estimate $\hat{G}$

Bootstrap: Sample with Replacement

1. Number the observations in the dataset from 1 to n. Take a random sample with replacement of size n from these case numbers
2. Create the resample from the original data using the case numbers. For repeating numbers, we repeat the corresponding sample by th same number of times. Compute the regression using this dataset and save the values of the coefficient estimates.
3. Repeat steps 1 and 2 for B times.
4. Estimate a 95% confidence interval for each estimates by the 2.5 and 97.5 percentiles of the B bootstrap samples.

Feature Selection

In real applications:

- Many predictors
- Several candidate models
    - all may pass the usual diagnostics and tests
- How do we pick the best model?

What is feature (model) selection?

- the process of choosing a “best” subset of available predictors.
- there might not be a single “best” subset.
- We do want a model we can interpret or justify with respect to the questions of interest.

Questions of Feature Selection

1. Which variables to include? $X_1, X_2, \ldots$
2. What transformation should these variables take? $X_1^2, log(X_2), 1/X_3, \ldots$
3. Should we include interaction terms, like $X_1 * X_2, X_1/(X_1 + X_2), \ldots$?

- Ideally, we answer these questions simultaneously.
- Instead, we will focus on the first question.
- We can then use variable transformations as needed.
- It would be impossible to compare all possible forms of all possible variables.

What Happens if We Miss Necessary Predictors or Include Unnecessary Predictors?

- The **Mean Squared Error (MSE)** of $\hat{\beta}$ is defined as

$$MSE(\hat{\beta}) = \mathrm{E}[(\hat{\beta} - \beta)^2]$$

- **Note**: This MSE is different from the MSE = SSE/dfE of a MLR model.
- One can show that MSE = Variance + (Bias)^2

Proof:

$$\begin{aligned}
\mathrm{E}[(\hat{\beta} - \beta)^2] &= \mathrm{E}[(\hat{\beta} - \beta - E[\hat{\beta}] + E[\hat{\beta}])^2] \\
&= \mathrm{E}[(\hat{\beta} - E[\hat{\beta}])^2 + (E[\hat{\beta}] - \beta)^2 + 2(E[\hat{\beta}] - \beta)(\hat{\beta} - E[\hat{\beta}])] \\
&= (\text{Variance of } \hat{\beta}) + (\text{Bias of } \hat{\beta})^2 + 0
\end{aligned}$$

What Happens if We Miss Necessary Predictors?

Suppose the “correct” model contains q predictors, but we only use p predictors and ignore the rest $(q - p)$ predictors.

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}_{retained} + \underbrace{\beta_{p+1} x_{i,p+1} + \cdots + \beta_q x_{iq}}_{omitted}$$

Matrix Notation

$$Y = X\beta + \epsilon = X_p\beta_p + X_r\beta_r + \epsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \left[\begin{array}{ccccc|ccc} 1 & x_{11} & x_{12} & \dots & x_{1p} & x_{1(p+1)} & \dots & x_{1q} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} & x_{2(p+1)} & \dots & x_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n1} & \dots & x_{np} & x_{n(p+1)} & \dots & x_{nq} \end{array}\right], \beta = \left[\begin{array}{c} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ \hline \beta_{p+1} \\ \vdots \\ \beta_q \end{array}\right]$$

- The full linear model with q variables: $Y = X_p\beta_p + X_r\beta_r + \epsilon$

$$\hat{\beta}^* = (X^TX)^{-1}X^TY$$

- The linear model with only p predictors: $Y = X_p\beta_p + \epsilon$

$$\hat{\beta}_p = (X_p^TX_p)^{-1}X_p^TY$$

What Happens if We Miss Necessary Predictors? (Effect on β)

- $E(\hat{\beta}_p) = \beta_p + A\beta_r$, where $A = (X_p^T X_p)^{-1} X_p^T X_r$
- $Var(\hat{\beta}_p) = \sigma^2 (X_p^T X_p)^{-1}$
- $Var(\hat{\beta}^*) = \sigma^2 (X^T X)^{-1}$
- $MSE(\hat{\beta}_p) = \sigma^2 (X_p^T X_p)^{-1} + A\beta_r \beta_r^T A^T$
- $\hat{\sigma}_p^2 = \dfrac{(Y - X_p\hat{\beta}_p)^T (Y - X_p\hat{\beta}_p)}{n-p-1}$ is generally biased upward.

What Happens if We Miss Necessary Predictors? (Effect on β)

- $\hat{\beta}_p$ is a **biased** estimate of β_p unless (1) $\beta_r = 0$ or (2) $X_p^T X_r = 0$

  $E(\hat{\beta}_p) = \beta_p + A\beta_r$, where $A = (X_p^T X_p)^{-1} X_p^T X_r$

- The matrix $\mathrm{Var}(\hat{\beta}^*) - \mathrm{Var}(\hat{\beta}_p)$ is positive semidefinite.

  Variances of the least squares estimates of regression coefficients obtained from the full model are larger than the corresponding variances of the estimates obtained from the subset model.

- Variance-Bias Trade off: The smaller model might have smaller MSE if the increment in the (Bias)^2 is less than the reduction in variance

What Happens if We Miss Necessary Predictors? (Effect on Y)

Effect of deleting predictors on the prediction of Y is similar

- the smaller model has smaller variance but greater bias
- MSE will decrease for the smaller model if the increment in the (Bias)^2 is less than the reduction in variance
- The best models will keep the important variables: predictors with high $|\beta_j|$

What Happens if We Include Unnecessary Predictors?

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}_{necessary} + \underbrace{\beta_{p+1} x_{i,p+1} + \cdots + \beta_q x_{iq}}_{=0, redundant}$$

If some X_j's have coefficients β_j's equal to 0, but we include them in the model,

- We gain nothing in the precision in estimating β's and predicting y
- The variance in estimation and prediction will increase

Model Selection Criteria (Model Evaluation)

The way we evaluate a model depends on what we hope to achieved with our model:

- Description
- Prediction
- Control

In many cases, these uses overlap.

There might not be a single best model

Goal 1: Description

- To describe a given process or understand and model the variation in a complex interacting system.
- Interpretability: Lots of thinking required about which variables are important
- Two contradicting requirements:
  1. Account for as much of the variation as possible;
  2. Understanding and interpretation are easier with fewer variables.
- Strategy: Choose the smallest set of variables that accounts for the largest percentage of variation in the response.

Goal 2: Prediction

- To use the patterns in our current data set to estimate the mean of future response or to predict a future value.
- This is also called forecasting. The focus is on applying knowledge to observations not in the current data.
- Strategy: Minimize the MSE of the estimation or prediction

$$MSE(\hat{y}) = \mathrm{E}[(\hat{y} - y)^2]$$

Goal 3: Control

- To manipulate the response by altering some of the predictor variables.
- Minimize the MSE of $\hat{\beta}_j$

$$MSE(\hat{\beta}) = \mathrm{E}[(\hat{\beta} - \beta)^2]$$

Summary of Model Comparison Methods

- Nested models
    - F-test (Come along with P-values)
- Ant two models with the same response (no P-values)
    - MSE (Mean Squared Error = SSE/(n-p-1))
    - AIC (Akaike Information Criterion)
    - BIC (Bayesian Information Criterion)
    - Mallow's C_p (has fallen out of favor)
- Any two models with the same response (but possibly differently transformed)
    - Adjusted R^2

Information Criteria

- $AIC = n\log(SSE_p/n) + 2p$
- $BIC = n\log(SSE_p/n) + p\log n$
- $C_p = 2p - n + \dfrac{SSE_p}{MSE_{all}}$
- Both Akaike and Bayesian Information Criteria reward small variance (SSE_p/n) and penalize larger models (p large).
- CAUTION: For AIC, BIC, and C_p, p = number of parameters (including the intercept). Different from our usual meaning of p (= # of predictors).
- Smaller AIC/BIC is better
- Models with AIC differ < 2 should be considered equally adequate.
- Similarly, models with BIC differ ≤ 2 are considered equally good
- BIC penalty for larger models is more severe: $p\log n > 2p$ whenever $n > 8$

Feature Selection Procedure

- The most direct and ideal approach is to examine all possible subsets of potential predictors.

  If it's possible to search over all possible subsets, then a good strategy is

  1. Choose the best models in each class of p-term models, for p = 0,1,...,q
  2. Analyze these best models more closely, including diagnostic plots, possible transformations, etc.
  3. Select the best model(s) (including transformations) by comparing both diagnostics and scores.

- In practice, if q=20. there are 2^{20} ($>$ 1,000,000) candidate models
- If q=30. there are 2^{30} ($>$ 1 billion) candidate models

Not practical to examine all of them unless q is quite small.

Feature Selection Procedure

Commonly used algorithms aimed to find the “best” model without look at all possible subsets:

- Forward selection (FS)
- Backward elimination (BE)
- Stepwise selection (SW)

Forward Selection (FS)

The Forward Selection algorithm consider all candidate subsets consisting of one additional term beyond the current subset

1. It begins with an intercept-only model.
2. At each iteration, add the predictor with the smallest P-value and then fit the new model
3. Stop if:
   - The new selected variable is not significant, or
   - All variables have been selected (all variables added).
4. Otherwise, fit the new model with the new added variable, and go to Step 2

The FS algorithm considers at most $q + (q-1) + \cdots + 2 + 1 = q(q+1)/2$ subsets, not all 2^q possible subsets.

Backward Elimination (BE)

1. The Backward Elimination (BE) algorithm begins with the full set of variables.
2. Eliminate the least significant variable (the one with the largest P-value)
3. Stop if:
    - All variables are significant, or
    - All variables have been eliminated (intercept-only model).
4. Otherwise, eliminate the least significant variable and go to Step 2.

The BE algorithm considers at most

$$q + (q-1) + \cdots + 2 + 1 = q(q+1)/2$$

subsets, not all 2^q possible subsets.

Stepwise Selection (SW)

1. At each iteration, the Stepwise Selection (SW) algorithm consider all models obtained by either adding or deleting one term to or from the current model.
2. At each iteration, choose the model with the lowest AIC or BIC
3. Stop if the model with the lowest AIC/BIC is the current model
4. Otherwise, let the model with the lowest AIC/BIC be the new current model and then go back to Step 1

Using the SW algorithm, a term added to a model might be removed at a later step

FS, BE Algorithms with AIC and BIC

- Instead of using P-values, we can use scoring methods like AIC and BIC.
- At any iteration, compare models based on the chosen scoring method.
- Among the models we are considering, we choose the model with the lowest AIC (or BIC).
- We stop the procedure when no candidates reduce the score.
- The major difference is that we are not judging variables based on significance levels, but only on the basis of how they affect the score.

Comments about the FS, BE, SW Algorithms

- The indicator variables of a categorical predictor should be included or removed altogether
  - better using AIC/BIC rather than P-values since adding/removing a categorical predictor may involve more than 1 parameter but a numerical predictor involves just 1 parameter
  - May simplify a categorical predictor by merging some categories
- If the possible pool of model terms include interactions, note
  - an interaction is never added unless all the lower order effects in the interaction are already included.
  - if an interaction is in the current model, none of its component variables or lower order interaction should be removed