6.5: Linear Model Selection and Regularization

Simone Collier

University of Toronto

Overview

Introduction

In this section we will look at alternative fitting procedures for linear models that may give better prediction accuracy of model interpretability compared to the standard least squares approach. The two classes of methods we will cover in this section are:

- **Subset Selection**
- **Shinkage**

Subset Selection

Subset Selection

Subset selection refers to the process of fitting a statistical model using only a subset of the predictors that are thought to be associated with the response. This can have the following results:

- Improved model interpretability by removing irrelevant variables, thereby reducing the complexity of the model.
- Improved prediction accuracy by a reduction in the risk of overfitting and the variance in the fitted coefficients.

We will cover two methods for subset selection:

- Best Subset Selection
- Stepwise Model Selection

Best Subset Selection

Best subset selection involves fitting a least squares regression for every possible combination of the p predictors and then choosing the best model among them all. It works as follows:

1. For $k = 1, 2, \ldots, p$:
   i. Fit all possible models that contain k predictors. (i.e. if $k = 2$ fit a separate model for each possible combination of two predictors)
   ii. Pick the best model from the set of models with k predictors and call it $\mathcal{M}_k$. The best model is the one with the smallest residual sum of squares (RSS) or largest R^2.
2. Select the best model from the set of $\mathcal{M}_0, \ldots, \mathcal{M}_p$ where $\mathcal{M}_0$ contains no predictors and simply predicts the sample mean for each observation.

The methods used to choose the best model in step 2 will be discussed later on.

Best Subset Selection

There are several difficulties associated with best subset selection.

- The number of models to be fit grows a lot with each increase in p.
- It can be very computationally expensive as a result.

Forward Stepwise Selection

Forward stepwise selection works by fitting models with progressively more predictors, adding the predictors into the model in the order of greatest model improvement. The process is as follows:

1. Let $\mathcal{M}_0$ denote the model that contains no predictors.
2. For $k = 0, 1, \ldots, p-1$:
   i. Fit all models that include all the predictors in $\mathcal{M}_k$ along with one additional predictor.
   ii. Choose the best model from the $p-k$ models above and call it $\mathcal{M}_{k+1}$. The best model is the one with the smallest RSS or highest R^2.
3. Select the best model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$.

Backward Stepwise Selection

Backward stepwise selection is simply the reverse of forward stepwise selection.

1. Let $\mathcal{M}_p$ be the model that contains all p predictors.
2. For $k = p,\ p-1, \ldots,\ 1$:
   i. Fit all models that contain all but one of the predictors in $\mathcal{M}_k$.
   ii. Choose the best from the k models and call it $\mathcal{M}_{k-1}$. The best model is the one with the smallest RSS or highest R^2.
3. Select the best model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$.

Exercises: Subset Selection

Open the Linear Model Selection and Regularization R Markdown file.

- Go over the "Getting Started" section together as a class.
- Go over the "Best Subset Selection" section together as a class.
- Go over the "Stepwise Selection" section together as a class.

Choosing the Optimal Model

Each of the subset selection methods require choosing the best model from a set of models. We cannot use RSS or R^2 for our final decision since they will always recommend the model containing all the predictors since it will have the lowest training error rate.

We need to estimate the test error. There are two approaches:

- Indirectly estimate the test error by adjusting the training error to include a measure of bias due to overfitting.
- Directly estimate the test error using a a validation set approach or cross-validation.

Indirect Test Error Estimation

We will consider four approaches used to adjust the training error.

- C_p
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Adjusted R^2

C_p

Given a fitted least squares model with d predictors, the C_p estimates the test MSE using

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

- $\hat{\sigma}^2$ is an estimate of the variance of the error for each response measurement.
- $2d\hat{\sigma}^2$ acts as a penalty term for the training RSS since it will underestimate the test error.
- The more predictors, the larger the penalty term.
- Choose the model with the lowest C_p

AIC

The AIC is usually used for models fit with maxiumum likelihood but in the case of a linear model with Gaussian errors, maximum likelihood and least squares are the same thing.

$$\text{AIC} = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

So for least squares models, C_p and AIC are the same. As a result, we choose the model with the lowest AIC value.

BIC

Given a fitted least squares model with d predictors, the BIC is

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n) d \hat{\sigma}^2\right)$$

- n is the number of observations
- The BIC places a heavier penalty on models with many variables.
- We choose the model with the lowest BIC.

Adjusted R^2

The adjusted R^2 statistic is the computed with

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

- Total sum of squares (TSS) $= \sum (y_i - \bar{y})^2$
- We choose the model with the largest adjusted R^2

Exercises: Indirect Error Estimation

Open the Linear Model Selection and Regularization R Markdown file.

- Go over the section 2.3 together as a class.
- 10 minutes for students to complete the questions from section 2.3.
- Questions should be completed at home if time does not allow.

Direct Test Error Estimation

We can compute the validation set error or the cross-validation error for each model we are attempting to decide between and choose the one with the smallest test error estimate.

If there are several models that have similar test errors, use the **one-standard-error rule**:

1. Calculate the standard error of the estimated test MSE for each model.
2. Select the smallest model that has an estimated test error within one SE of the smallest MSE of all the models.

Indirect vs Direct Test Error Estimation

- Advantages of direct test error estimation via the validation set approach or cross-validation:
  - provides a direct estimate of the test error
  - makes fewer assumptions about the true model

- Advantages of indirect test error estimation via AIC, BIC, C_p, and adjusted R^2:
  - less computation time for problems with large p or n
    (however, with the fast computers nowadays this is hardly ever an issue with cross-validation)

Exercises: Direct Error Estimation

Open the Linear Model Selection and Regularization R Markdown file.

- Go over the section 2.4 together as a class.
- 10 minutes for students to complete the questions from section 2.4.
- Questions should be completed at home if time does not allow.

Shrinkage Methods

Shrinkage Methods

Shrinkage methods fit a model using a technique that regularizes the coefficient estimates. This results in shrinking the coefficients towards zero and thereby reducing their variance.

The two shrinkage methods we will cover are:

- Ridge regression
- Lasso

Ridge Regression

Ridge regression works the same as the least squares method except instead of minimizing the RSS to find estimates of $\beta_0, \beta_1, \ldots, \beta_p$, we minimize

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- $\lambda \sum_j \beta_j^2$ is a shrinkage penalty that shrinks the estimates of β_j towards zero.
- $\lambda \geq 0$ is a tuning parameter that controls the impact of the penalty term
  - When $\lambda = 0 \Rightarrow$ ridge regression $=$ least squares.
  - When $\lambda \to \infty \Rightarrow$ ridge regression coefficients approach zero.

Ridge Regression

There are a few key considerations for ridge regression:

- Different resulting coefficient estimates $\hat{\beta}_\lambda^R$ for different λ.
- Shrinkage penalty is applied to $\beta_1, \ldots, \beta_p$ but not β_0.
- Ridge regression coefficient estimates are not scale equivalent.
  - Always standardize the predictors so that they are on the same scale before performing ride regression

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

The Lasso

The Lasso, like ridge regression, minimizes a different quantity than least squares regression in order to estimate the regression coefficients.

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

- β_j^2 in ridge regression is replaced by $|\beta_j|$ in the lasso.
- The penalty forces some coefficient estimates to be zero for sufficiently large λ.

Ridge Regression vs The Lasso

The advantages of ridge regression and the lasso over the standard least squares method are related to the bias-variance trade-off.

- As λ increases, the flexibility of the regression fit decreases
  - decrease in variance
  - increase in bias
  - The test MSE is closely related to variance + bias^2
- Ridge regression and the lasso work best when least squares estimates have high variance.
  - when p is almost as large or larger than n, ridge regression and the lasso can still perform well by trading off a small increase in bias with a large decrease in variance.
- Both method are substantially less computationally expensive compared to subset selection

Ridge Regression vs The Lasso

- Model interpretability
  - Ridge regression results in a model with all p predictors which can be difficult to interpret.
  - The lasso performs variable selection so the models are much easier to interpret.

- Prediction accuracy
  - The lasso will perform better when a small subset of the predictors have substantial coefficients and the remaining coefficients are negligible.
  - Ridge regression will perform better when many of the predictors are associated with the response.
  - These are not qualities known beforehand so cross-validation can be used to determine the best approach for a particular problem.

Selecting the Tuning Parameter

Cross-validation offers a simple way to choose the best tuning parameter for ridge regression and the lasso.

1. Take many λ values.
2. Compute the cross-validation error of the fitted model for each λ.
3. Choose the tuning parameter that gives the smallest cross-validation error.
4. Refit the model using the chosen tuning parameter and the complete set of observations.

Exercises: Ridge Regression & The Lasso

Open the Linear Model Selection and Regularization R Markdown file.

- Go over the "Ridge Regression" section together as a class.
- Go over the "The Lasso" section together as a class.
- 10 minutes for students to complete the questions.
- Questions should be completed at home if time does not allow.

References

Chapter 6 of the ISLR2 book:

James, Gareth, et al. "Linear Model Selection and Regularization." An Introduction to Statistical Learning: with Applications in R, 2nd ed., Springer, 2021.