Lecture 13: Linear Model Selection and Regularization (Textbook 6.1)

Nayel Bettache

Department of Statistical Science, Cornell University

Improving LSE?

- Recall the linear model is

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon.$$

- How can we improve the model fitting ?

Prediction Accuracy

- Recall the linear model is

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon.$$

- How can we improve the model fitting ?
- *Prediction Accuracy*: If the true relationship is $\approx$ linear, LSE has low bias.
  - If $n \gg p$, LSE has low variance.
  - If $n \gtrsim p$, LSE has large variance, possibly resulting in overftting.
  - If $n < p$, there is no longer a unique LSE: infinitely many solutions. They show zero error on the training data, but very poor test set performance.
  - By **constraining** or **shrinking** the estimated coefficients, we can reduce the variance at the cost of a negligible increase in bias.

Model Interpretability

- Recall the linear model is

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon.$$

- How can we improve the model fitting ?
- *Model Interpretability*: What if the model considers irrelevant variables ?
  - Leads to unnecessary complexity
  - By removing these variables: we can obtain a model that is more easily interpreted.
  - Problem: LSE is unlikely to yield any coeffcient estimates that are exactly zero.
  - Solution: feature selection or variable selection.

Alternatives

- *Subset Selection*: Identifying a subset of the p predictors that we believe to be related to the response. Then fit LSE on these p predictors.
- *Shrinkage*: Fitting a model involving all predictors but shrinking towards zero the coefficients. This shrinkage (also known as regularization) has the effect of reducing variance.
- *Dimension Reduction*: This approach involves projecting the predictors into a smaller M-dimensional subspace. Then these M projections are used as predictors to fit a linear regression model by least squares.

Subset Selection

Two methods for selecting subsets of predictors: **best subset** and **stepwise model selection**.

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

For best subset selection, we need to compare 2^p models.

Best Subset Selection

- Step 2 identifies the best model (on the training data) for each subset size, in order to reduce the problem from one of 2^p possible models to one of $p + 1$ possible models.
- Warning: RSS of these $p + 1$ models decreases monotonically, and the R^2 increases monotonically, as the number of features included in the models increases.
- In Step 3, we should not use RSS or R^2, because we want a model with small **test** error not training error, and we can't compare RSS or R^2 on models with $\neq$ number of predictors.
- Best subset selection becomes computationally infeasible for large values of p.
- Best subset selection may also suffers from statistical problems when p is large. The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.

Forward Stepwise Selection

Computationally efficient alternative to best subset selection.

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

For forward stepwise selection, we need to compare 1 null model plus $p-k$ models in iteration k. So, in total $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models much fewer than 2^p models.

Forward Stepwise Selection

- It has computational advantage over best subset selection.
- It can be used in high-dimensional setting with $n < p$.
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

The Credit Card Data

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

Backward Stepwise Selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

- For backward stepwise selection, we also compare $1 + p(p+1)/2$ models much fewer than 2^p models.
- It only works when $n > p$.
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

Estimating the Test Error

There are two common approaches:

- We can directly estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in Chapter 5.
- We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting. This class of methods contain C_p, AIC, BIC, and adjusted R^2.

C_p, AIC, BIC, and adjusted R^2

These techniques adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables.

- **Mallow's** C_p:

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

d is the total # of parameters in the model and $\hat{\sigma}^2$ is an estimate of $var(\epsilon)$.

Essentially, the C_p adds a penalty $2d\hat{\sigma}^2$ to the training RSS to adjust for the fact that the training error tends to underestimate the test error.

C_p tends to take on a small value for models with a low test error, so when determining which of a set of models is best, we choose the model with the lowest C_p value.

C_p, AIC, BIC, and adjusted R^2

- **AIC**:

$$AIC = -2 \log L + 2d,$$

L is the maximized value of the likelihood function for the estimated model. In the linear model with Gaussian errors, $AIC = C_p/\hat{\sigma}^2$. In this case, AIC is proportional to C_p, which yields the same selected model.

- **BIC**:

$$BIC = -2 \log L + (\log n)d,$$

BIC places a heavier penalty $(\log n)d$ on models with many variables, and hence results in the selection of smaller models than AIC and C_p. For both AIC and BIC, we select model with lowest values.

C_p, AIC, BIC, and adjusted R^2

- **Adjusted** R^2:

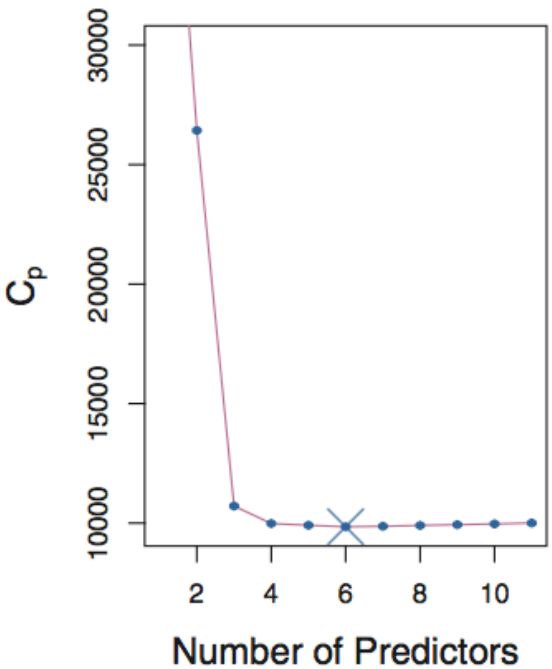
$$\textit{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}.$$

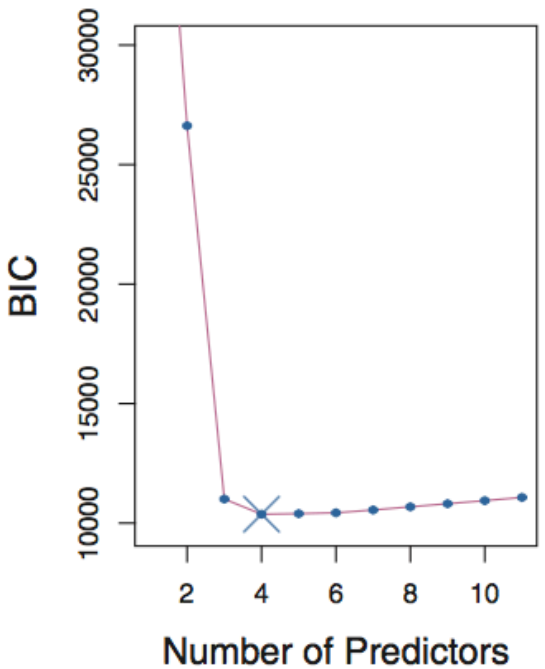
Unlike C_p, AIC, and BIC, for which a **small** value indicates a model with a low test error, a **large** value of adjusted R^2 indicates a model with a small test error.

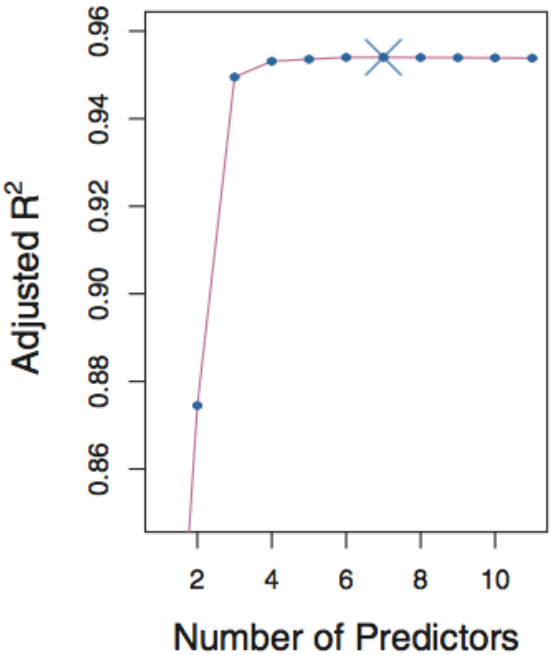
Maximizing the adjusted R^2 is equivalent to minimizing $RSS/(n-d-1)$. While RSS always decreases as the number of variables in the model increases, $RSS/(n-d-1)$ may increase or decrease, due to the presence of d in the denominator.

Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model.

The Credit Card Data







Validation and Cross-Validation

- We compute the validation set error or the cross-validation error for each model M_k under consideration, and then select the k for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance.
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance.

The Credit Card Data