Lecture 11: Model selection and regularization

Reading: Sections 7.4 - 7.7, 3.4

GU4241/GR5241 Statistical Machine Learning

Linxi Liu

February 23, 2018

What do we know so far

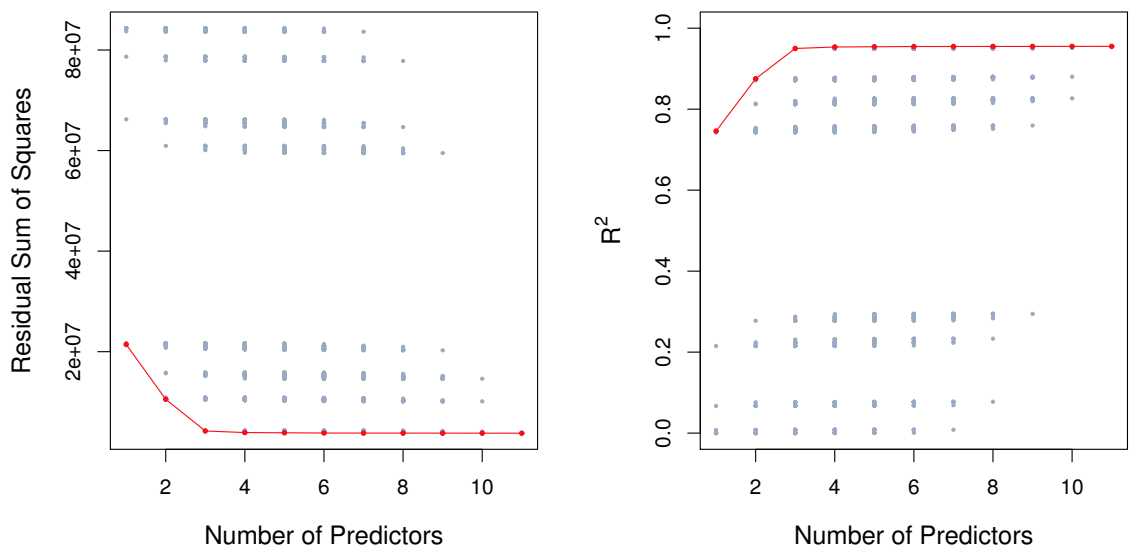
- In linear regression, adding predictors always decreases the training error or RSS.
- However, adding predictors does not necessarily improve the test error.
- Selecting significant predictors is hard when n is not much larger than p.
- When $n < p$, there is no least squares solution:

$$\hat{\beta} = \underbrace{(\mathbf{X}^T\mathbf{X})}_{\text{Singular}}^{-1}\mathbf{X}^T y.$$

So, we must find a way to select fewer predictors.

Best subset selection

- Simple idea: let's compare all models with k predictors.
- There are $\binom{p}{k} = p!/\left[k!(p-k)!\right]$ possible models.
- Choose the model with the smallest RSS. Do this for every possible k.



- Naturally, the RSS and R^2 improve as we increase k.

Best subset selection

To optimize k, we want to minimize the test error, not the training error.

We could use **cross-validation**, or alternative estimates of test error:

1. **Akaike Information Criterion (AIC):**

$$\frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2k\hat{\sigma}^2)$$

后面的部分是penlty

where $\hat{\sigma}^2$ is an estimate of the irreducible error.

Best subset selection

To optimize k, we want to minimize the test error, not the training error.

We could use **cross-validation**, or alternative estimates of test error:

1. **Akaike Information Criterion (AIC) or C_p:**

$$\frac{1}{n}(\text{RSS} + 2k\hat{\sigma}^2)$$

where $\hat{\sigma}^2$ is an estimate of the irreducible error.

for different model there will have different form of AIC and BIC

Best subset selection

To optimize k, we want to minimize the test error, not the training error.

We could use **cross-validation**, or alternative estimates of test error:

1. **Akaike Information Criterion (AIC) or C_p:**
2. **Bayesian Information Criterion (BIC):**

$$\frac{1}{n}(\text{RSS} + \log(n)k\hat{\sigma}^2)$$

BIC得到的值比较小，simper model

Best subset selection

To optimize k, we want to minimize the test error, not the training error.

We could use **cross-validation**, or alternative estimates of test error:

1. **Akaike Information Criterion (AIC) or C_p:**
2. **Bayesian Information Criterion (BIC):**
3. **Adjusted R^2:**

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Best subset selection

To optimize k, we want to minimize the test error, not the training error.

We could use **cross-validation**, or alternative estimates of test error:

1. **Akaike Information Criterion (AIC) or C_p:**
2. **Bayesian Information Criterion (BIC):**
3. **Adjusted R^2:**

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$$

Best subset selection

To optimize k, we want to minimize the test error, not the training error.

We could use **cross-validation**, or alternative estimates of test error:

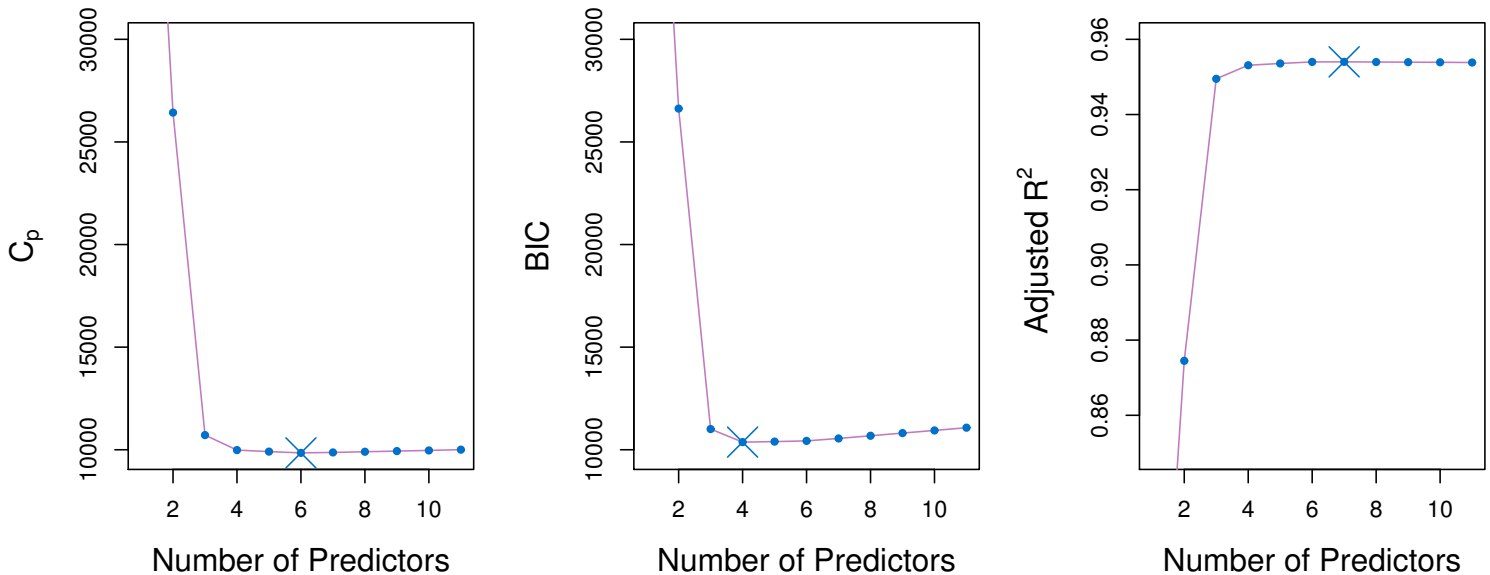
1. **Akaike Information Criterion (AIC) or C_p:**
2. **Bayesian Information Criterion (BIC):**
3. **Adjusted R^2:**

How do they compare to cross validation:

- They are much less expensive to compute.
- They are motivated by asymptotic arguments and rely on model assumptions (eg. normality of the errors).
- Equivalent concepts for other models (e.g. logistic regression).

Example

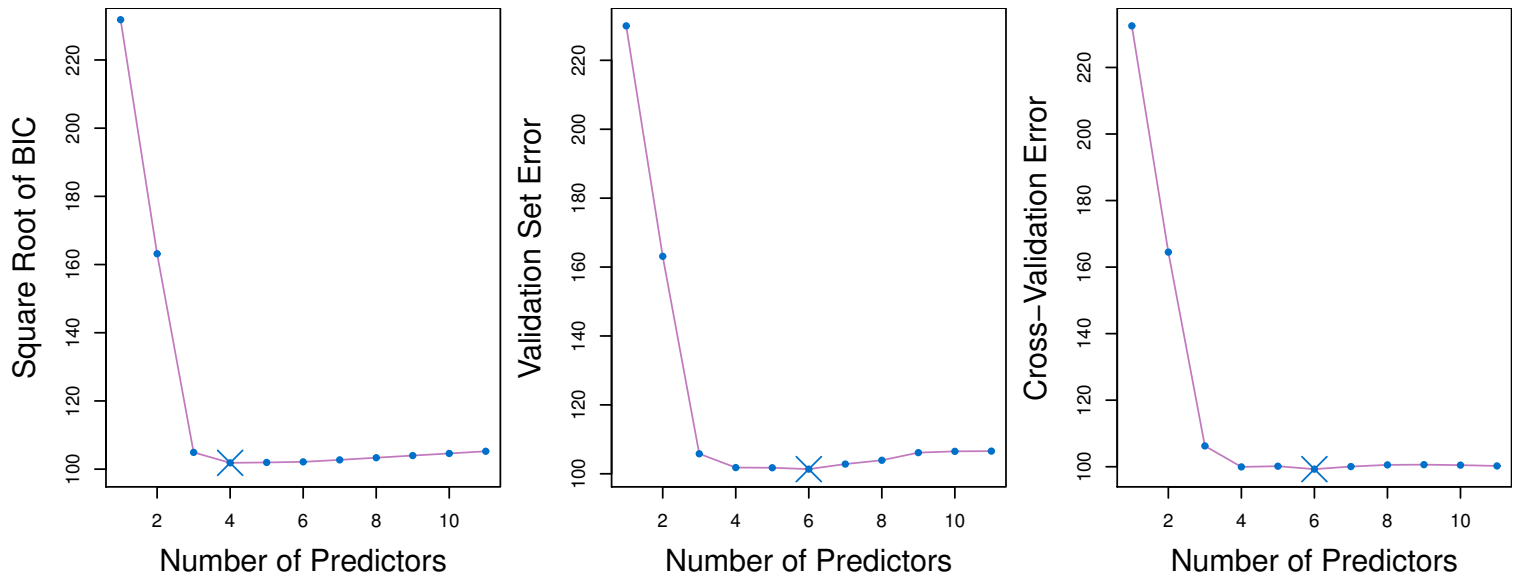
Best subset selection for the Credit dataset.



In linear adjusted R square is just equal to cp

Example

Cross-validation vs. the BIC.



Recall: In k-fold cross validation, we can estimate a standard error or accuracy for our test error estimate. Then, we apply the **one standard-error rule**.

Stepwise selection methods

Best subset selection has 2 problems:

1. It is often very expensive computationally. We have to fit 2^p models!
2. If for a fixed k, there are too many possibilities, we increase our chances of overfitting. The model selected has *high variance*.

In order to mitigate these problems, we can restrict our search space for the best model.

This reduces the variance of the selected model at the expense of an increase in bias.

Forward 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.

Forward selection vs. best subset

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

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* `Credit` *data set. The first three models are identical but the fourth models differ.*

Backward 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.

Forward vs. backward selection

- You cannot apply backward selection when $p > n$.
- Although it seems like they should, they need not produce the same sequence of models.

  Example. $X_1, X_2 \sim \mathcal{N}(0, \sigma)$ independent.

$$X_3 = X_1 + 3X_2$$
$$Y = X_1 + 2X_2 + \epsilon$$

  Regress Y onto X_1, X_2, X_3.

  - Forward: $\{X_3\} \to \{X_3, X_2\} \to \{X_3, X_2, X_1\}$
  - Backward: $\{X_1, X_2, X_3\} \to \{X_1, X_2\} \to \{X_2\}$

Other stepwise selection strategies

- **Mixed stepwise selection:** Do forward selection, but at every step, remove any variables that are no longer "necessary".
- **Forward stagewise selection:** Do forward selection, but after every step, modify the remaining predictors such that they are uncorrelated to the selected predictors.
- ...