Model Selection: Introduction

Instructor: Haoran LEI

Hunan University

What we have covered:

- Linear (and additive) models:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Least squares (i.e., minimizing the MSE on the training dataset)

Roadmap

- Ch6: discuss some ways in which the linear model can be improved
  - The model is still linear, but we **replace least squares with alternative fitting procedures**
- Ch7: generalize the linear model in order to accommodate *non-linear*, but still *additive*, relationships.
- Ch8: more general *non-linear* models.
  - For example: trees, boosting

Beyond Least Squares

Ch6 sticks to linear model: Despite its simplicity, the linear model has advantages in terms of **interpretability** and often shows **good predictive performance**.

We want to improve on the Least Squares by

1. *selecting* features: improve on **interpretability**
2. *shrinking* the coefficients of features: improve on **predictive performance**

Why consider alternatives to least squares?

- **Prediction Accuracy:** especially when $p > n$, to control the variance.
- **Model Interpretability:** By removing irrelevant features — that is, by setting the corresponding coefficient estimates to zero — we can obtain a model that is more easily interpreted.
  - We will present some approaches for *automatical feature selection*.

Three *classes* of methods

1. **Subset Selection.** First, *identify a subset of the predictors* that are related to the response, then fit a model using LS.
    - Method of exhaustion, forward and backward stepwise methods

Three *classes* of methods

2. **Shrinkage.** Fit a model involving *all p predictors*, but the estimated coefficients are *shrunken towards zero* relative to the least squares estimates.
   - This **shrinkage** (also known as **regularization**) has the effect of reducing variance and can also perform variable selection.
   - We do not select the features explicitly, but rather *penalize the model* for the **number of coefficients** or the **size of coefficients** in various ways.
   - **Lasso** and **Ridge Regression** are two popular shrinkage methods.

3. **Dimension Reduction.**
   - Project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the variables.
   - Then these M projections are used as predictors to fit a linear regression model by least squares.

Final Remarks:

- These three classes of methods (or *ideas*) also apply to other models, while we focus on linear models here.

1. Subset Selection

Very simple idea:

- Our data contains p predictors, but we have a simpler model that involves only *a subset* of those predictors.
- The natural way is to consider every possible subset of p predictors (2^p in total), and then select the "best subset".

1. Subset Selection

Step 1. For $k = 1, 2, \ldots, p$:

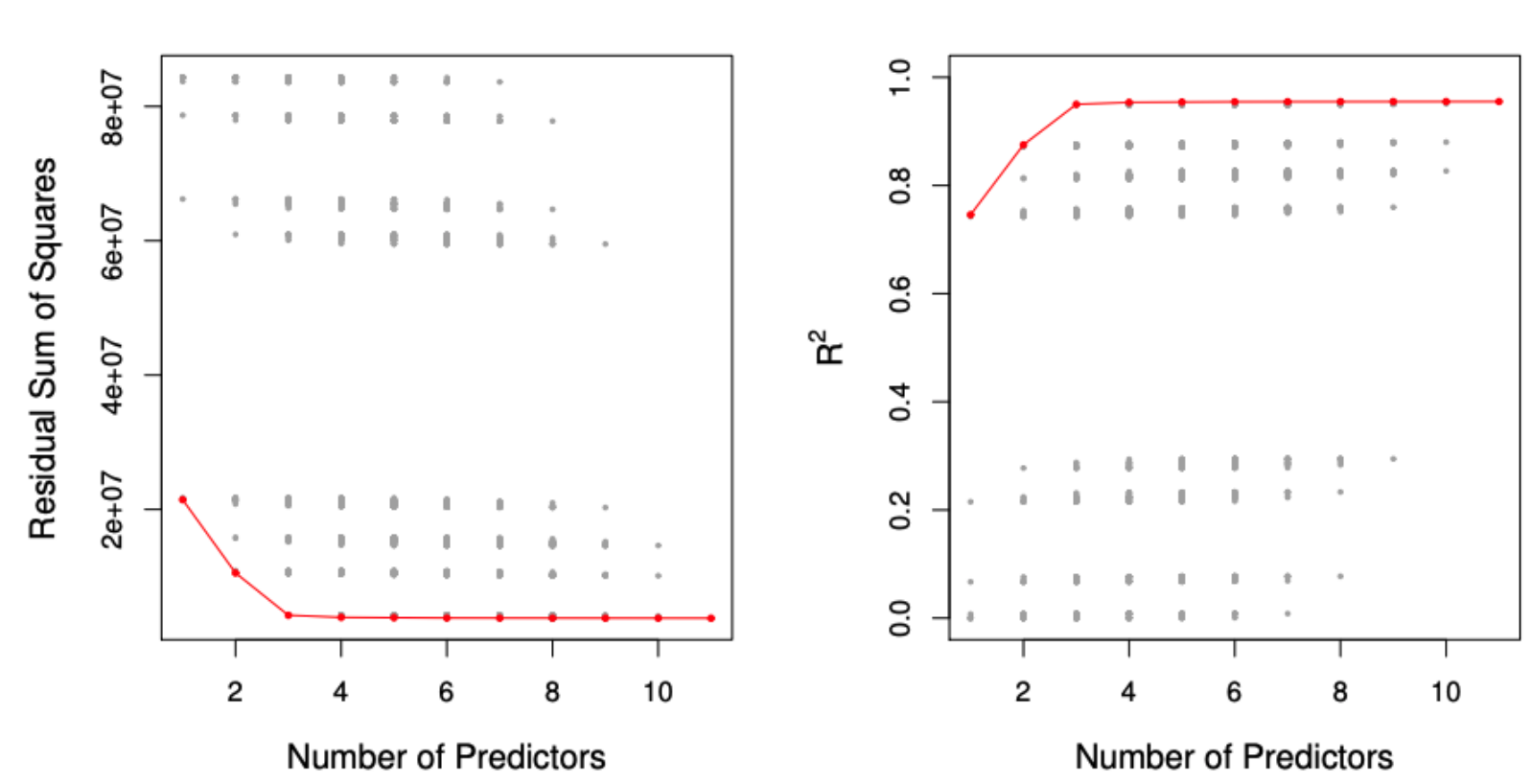
1. Fit all $\binom{p}{k}$ models that contain exactly k predictors.
2. Pick the "best" (i.e., having the smallest RSS/MSE) among these models. Call it M_k.

Step 2. Select a single "best" model from the p candidates, M_1, ..., M_p, based on:

- adjusted R^2, C_p (AIC), BIC, or cross-validated prediction error.

Example: Credit data set

Ten predictors ($p = 10$), including credit limit, credit range, # of cards, and so on. The response variable Y is card balance.



Stepwise Selection

When p is *not (very) small*, **best subset selection** method fails for two reasons:

1. *the computational cost*
2. *overfitting*

For both of these reasons, **stepwise methods**, which explore a far more restricted set of models, are attractive alternatives to best subset selection

- **Forward Stepwise Selection** and **Backward Stepwise Selection**

Choosing the best Model

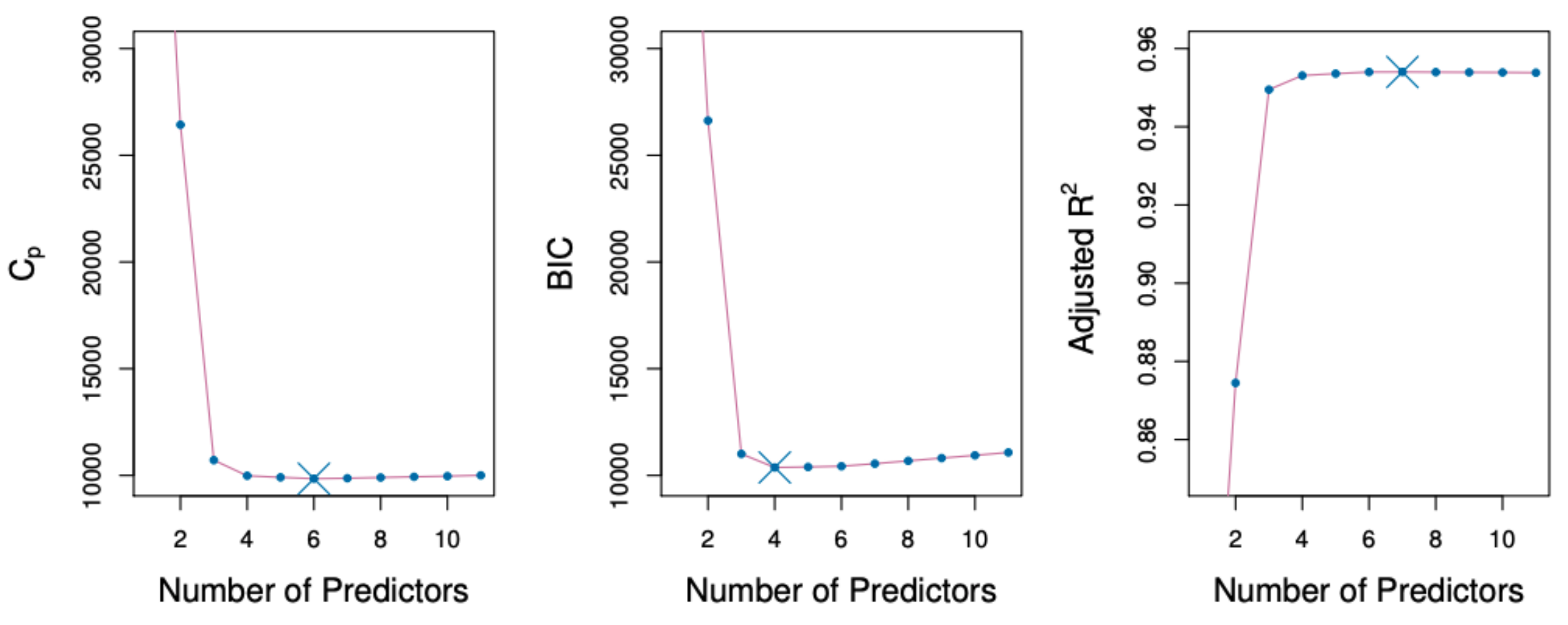
- The model containing all of the predictors will always have the **smallest RSS** and the **largest** R^2, since these quantities are related to the training error.
- We wish to choose a model with low **test error**, not a model with low **training error**. *Recall that training error is usually a poor estimate of test error.*
- Therefore, RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

Estimating test error: two approaches

- We choose the best model based on the **test error**, not the training error.
- We can *indirectly* estimate **test error** by making an adjustment to the training error to *account for the bias due to overfitting.*
- We can *directly* estimate the **test error**, using either a *validation set approach* or a *cross-validation approach*.
  - Also known as data-driven model selection.
- We illustrate both approaches next.

C_p, AIC, BIC, and Adjusted R^2

- These techniques can be viewed as indirect estimates of test error.
  - They adjust the *training error* for the *model size*, and can be used to select among a set of models with different numbers of variables.
- The next figure displays C_p, BIC, and Adjusted R^2 for the best model of each size produced by best subset selection on the **credit dataset**.



Credit data example

Details of these criterion: C_p and AIC

- **Mallow's C_p** defined as below, where d is the total # of parameters used and $\hat{\sigma}$ is an estimate of the variance of ϵ.

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

- The **AIC criterion** is defined for a large class of models fit by maximum likelihood:

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

- where L is the maximized value of the likelihood function.

Details of these criterion: C_p and AIC

- In the case of the **linear model with Gaussian errors**, maximum likelihood and least squares are the same thing, and C_p and AIC are *equivalent*.

Details of these criterion: BIC

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

- We select the model that has the lowest BIC value. Like Mallow's C_p, the BIC will penalize a model for having too many predictors (ie, a higher d).
- Compared to C_p, BIC replaces the $2d\hat{\sigma}^2$ in C_p by $\log(n) d \hat{\sigma}^2$.
- Since $\log n > 2$ for any $n > 7$, BIC generally places a heavier penalty on models with many variables. So the selected "best model" is smaller than C_p. (See the credit example above)

Details of these criterion: adjusted R^2

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

- Unlike C_p, AIC and BIC, a better model tends to have a higher adjusted R^2.
- Maximizing adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-d-1}$.
- An advantage of adjusted R^2 over C_p/AIC/BIC is that it does not require computing an estimate of σ^2.

From selection by criteria to data-driven selection

- The first three criteria (C_p, AIC and BIC) are developed by statisticians, each having its own strength in different setups.
  - Read this article if you are interested in the statistical theories behind these criteria.
- Adjusted R^2 has the advantages of being easier to compute and "understand."

With the rapid growth of the machine learning literature, more researchers start to adopt the *data-driven selection* methods: **validation** and **cross-validation**.