HKU BUSINESS SCHOOL
港大經管學院

Business Statistics

Model Selection and Regularization

Zhanrui Cai

Assistant Professor in Analytics and Innovation

ISLR Chapter 5, 6

Linear Regression

- The model:
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$+ random error

- Estimate by the Least Square Method:
  - Suppose we have *n* subjects with data $(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i), i = 1, 2, \ldots n$.
  - The coefficient vector $\beta = (\beta_0, \beta_1, \beta_2, \ldots \beta_p)$ is estimated by

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Linear Regression: estimation

- The estimated coefficients are $(\widehat{\beta_1}, \widehat{\beta_2}, \ldots \widehat{\beta_p})$.
- The prediction for Y_i is: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_p} X_p$.
- The LS method is finding the smallest *RSS (Residual sum of squares)*:

$$RSS = \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2$$

Linear Regression: R square

- The total variability in Y *(Total sum of squares):*

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- The R^2 is defined as

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

- It measures how much variation in Y can be explained by the variation in *X*.
- Always between 0 and 1.

Linear Model Selection

- In linear model

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

- Model selection – select the relevant X features.
  - Number of X variables
  - Which X variables

- Prediction Accuracy: especially when $p > n$, to control the variance and enable model fitting.

- Model Interpretability: by removing irrelevant features through setting the corresponding coefficients to be zero.

Example – Credit Data Set

- From the ISLR book;
- The data contains information about credit card debt for 10,000 customers.
- Response variable: balance (average credit card debt for each individual)
- Covariates:
  - Income: in thousands of dollars
  - Limit: credit limit
  - Rating: credit rating
  - Cards: number of credit cards
  - Age: in years
  - Education: years of education
  - Own: house ownership

Example – Credit Data Set

- Covariates:
  - Student: student status
  - Married: martial status
  - Region: east, west, or south

```
> head(Credit)
   Income Limit Rating Cards Age Education Own Student Married Region Balance
1  14.891  3606    283     2  34        11  No      No     Yes  South     333
2 106.025  6645    483     3  82        15 Yes     Yes     Yes   West     903
3 104.593  7075    514     4  71        11  No      No      No   West     580
4 148.924  9504    681     3  36        11 Yes      No      No   West     964
5  55.882  4897    357     2  68        16  No      No     Yes  South     331
6  80.180  8047    569     4  77        10  No      No      No  South    1151
```

Three Model Selection Methods

- *Subset selection*: identify a subset of p predictors that we believe to be related to the response. Then fit a model on the reduced set of variables.

- *Shrinkage*: the estimated coefficients are shrunken towards zero relative to the original estimates. The shrinkage has the effect of reducing variance. Some shrinkage method may estimate the coefficients to be exactly zero, which is equivalent to variable selection.

- *Dimension reduction*: project the predictors into a smaller subspace. (will not be covered in this class)

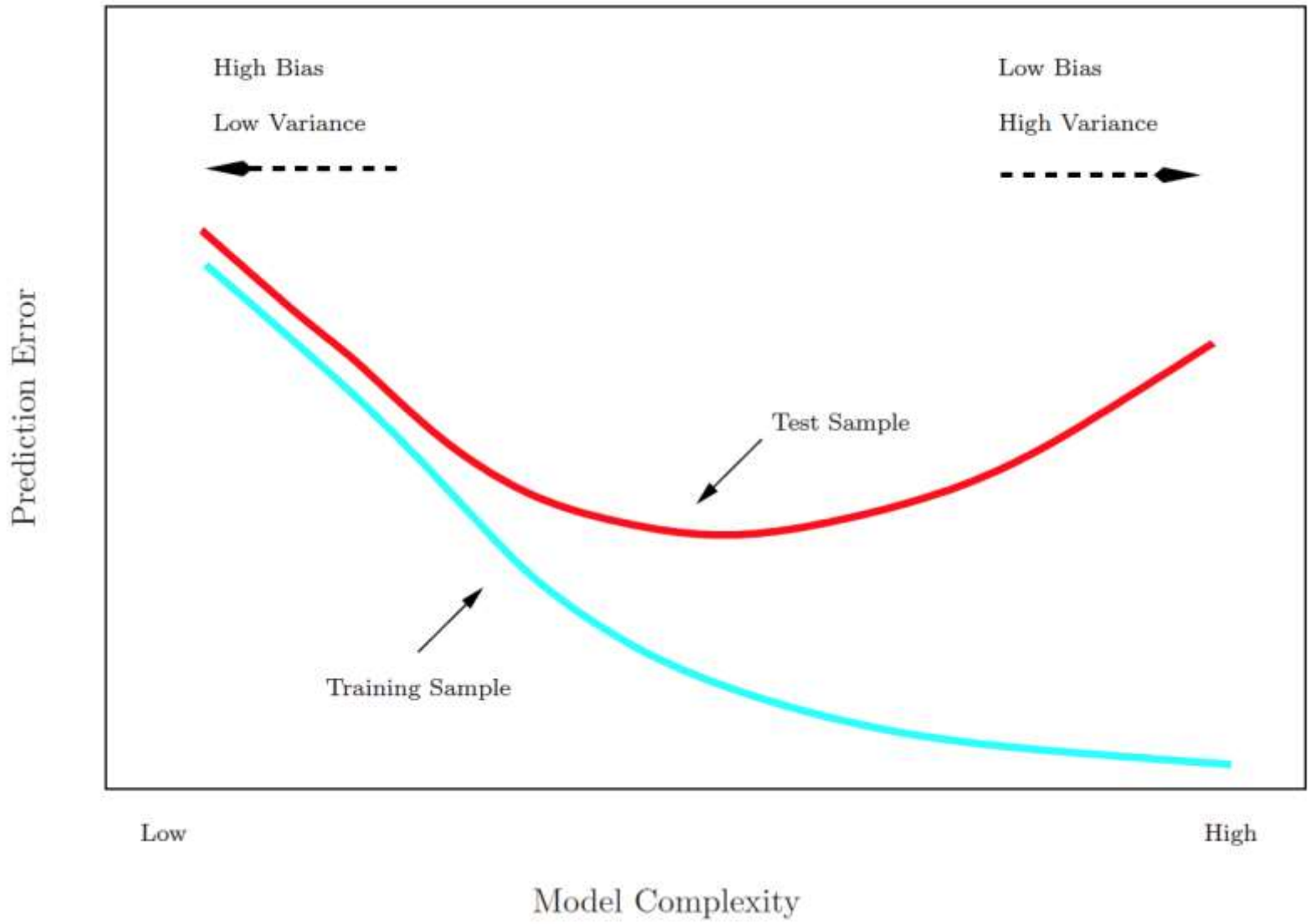
Can we just use R^2 to choose the best model?

- No!
- More predictors will always lead to smaller RSS and larger R^2, since these quantities are related to the training error.
- If use R^2, you will always include the most number of variables.
- We wish to choose a model with low test error, not low training error. Training error is usually a poor estimate of test error.
- Therefore, RSS and R^2 are not suitable for selecting the best model among models with different number of predictors.

Can we just use R^2 to choose the best model?

- Example: Fit a linear regression for Y with two predictors X_1 and X_2
- M1 only includes X_1
- M2 includes both X_1 and X_2
- RSS1 $\geq$ RSS2

Model Complexity & Prediction Error



Estimating Test Error: Two Approaches

1. Indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.

2. Directly estimate the test error, using either a validation set approach or a cross-validation approach.

Model Comparison Criteria

Adjustment to the Training Error

- Adjusted RSS and R^2 (proposed in 1920s)
- Mallow's C_p (proposed in 1973)
- Akaike Information Criterion (AIC) (proposed in 1973)
- Bayesian Information Criterion (BIC) (proposed in 1973)

Adjusted R^2

- For a least squares model with p variables, the Adjusted R^2 statistic is calculated as

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

  where TSS is the total sum of squares.

- A large Adjusted R^2 indicates a model with a small test error.

Adjusted R^2

- Maximize the Adjusted R^2 = Minimize $\frac{\text{RSS}}{n-p-1}$.
-
- Reason: Unlike R^2, Adjusted R^2 pays a price for the inclusion of unnecessary variables in the model.

Mallow's C_p

- Mallow's C_p:

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

- $\hat{\sigma}^2$ is the estimated variance of ϵ, from the fitted model.
- p : # of predictors in the model.
- Select the model with the smallest C_p.
- Balance between residuals and model sizes.

Akaike Information Criterion

- The AIC is defined as:

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

- Select the model with the smallest AIC.

- For linear regression models
  - C_p and AIC: equivalent in linear model

Bayesian Information Criterion

- The BIC is defined as

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

- Select the model with the lowest BIC value.

- Notice that BIC replaces the $2p\hat{\sigma}^2$ used by C_p with a $\log(n)p\hat{\sigma}^2$ term, where n is the number of observations.

- Since $\log(n) > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

Subset Selection

Best subset and stepwise model selection procedures

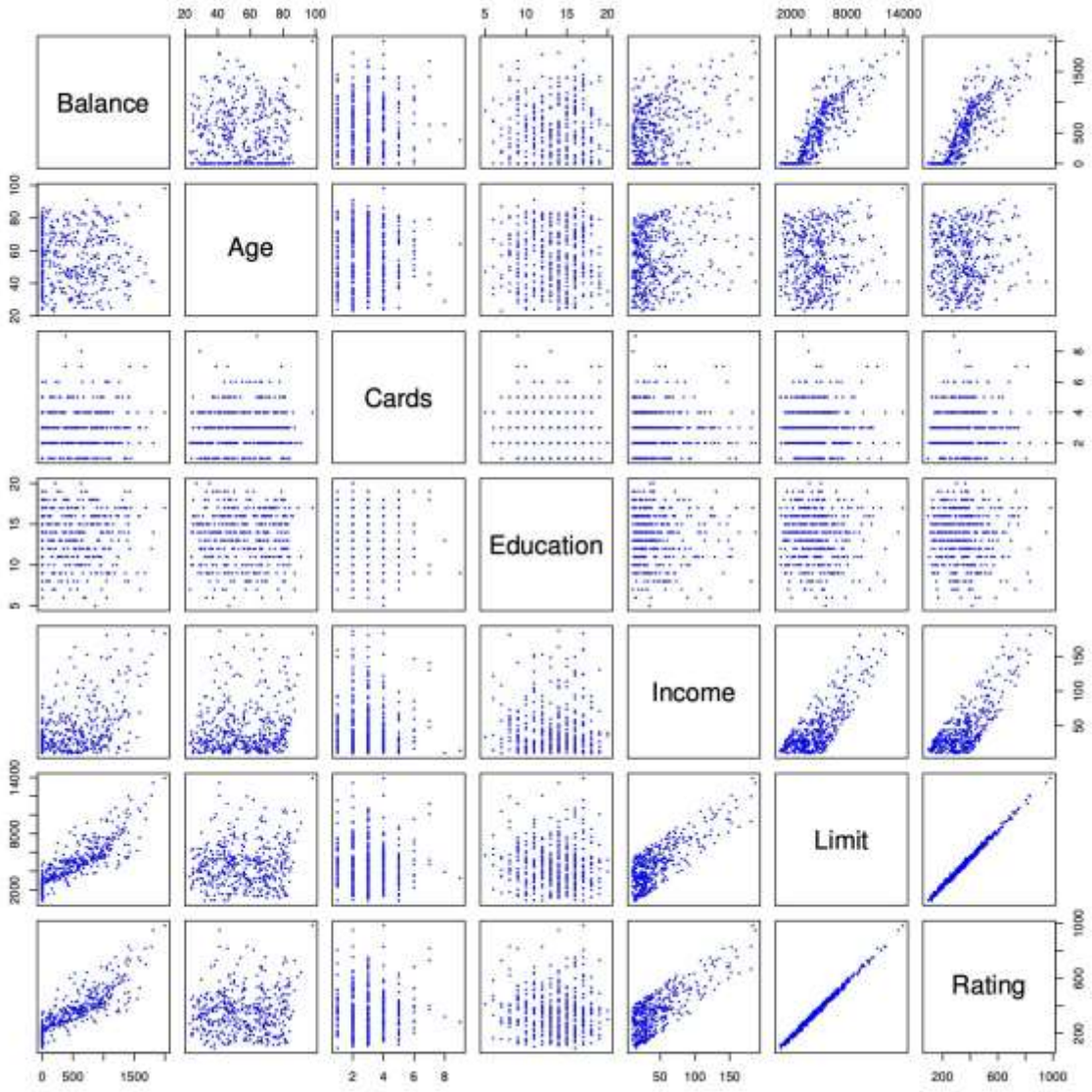
Best Subset Selection

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.

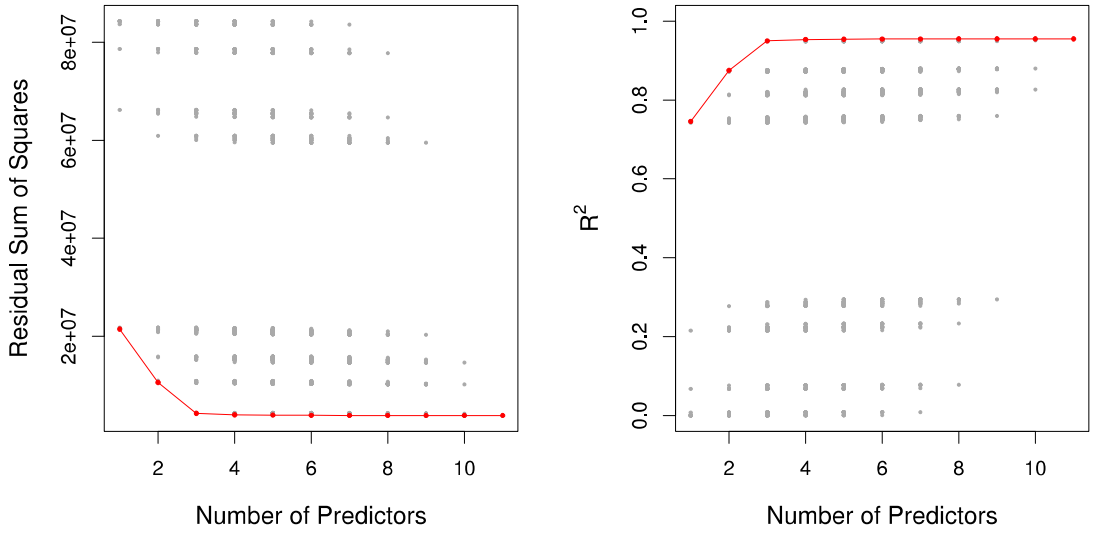
Example – Credit Data Set

- Response: balance (average credit card debt for each individual)
- Predictors:
  - Income: in thousands of doolars
  - Limit: credit limit
  - Rating: credit rating
  - Cards: number of credit cards
  - Age: in years
  - Education: years of education
  - Own: house ownership
  - Student: student status
  - Married: martial status
  - Region: east, west, or south

Example – Credit Data Set

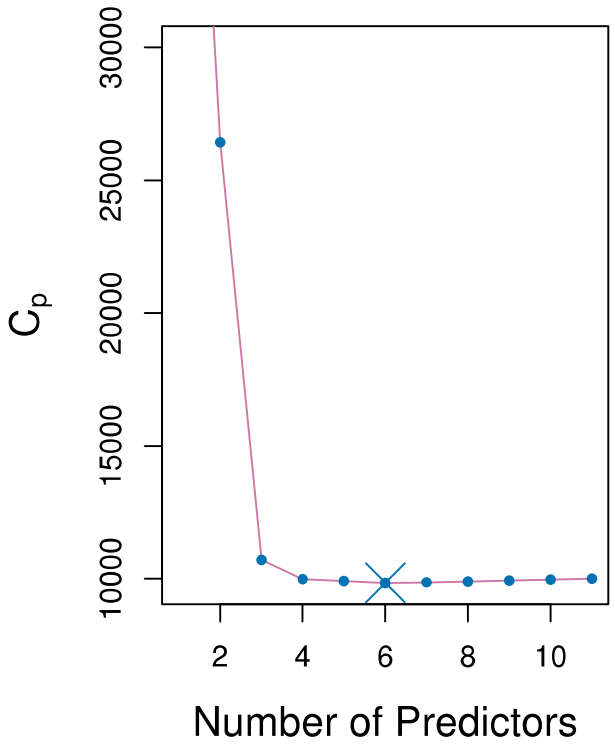


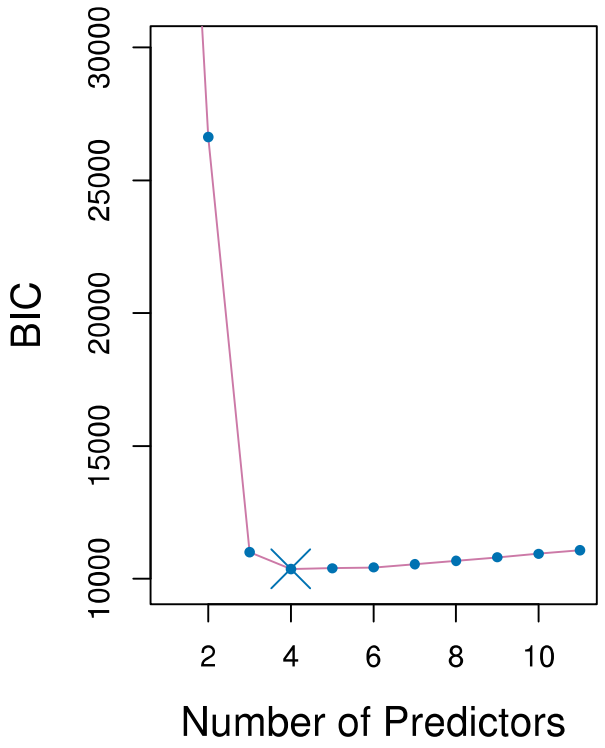
Example – Credit Data Set

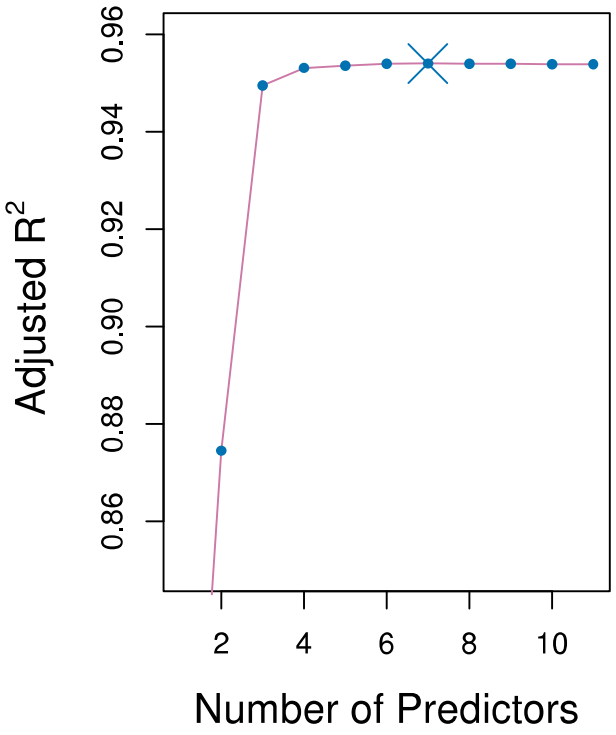


For each possible model containing a subset of the ten predictors in the `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the* x*-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables*

Credit Data Example: Best Subset







Credit Data Example: implementation

- Let's ask GPT: How to perform the best subset selection in R? And use the adjusted R square, Cp, AIC, and BIC to select the best model? Please demonstrate it using simulated dataset.

Extension to Other Models

- Although we have presented best subset selection here for least squares regression, the same ideas apply to other types of models, such as logistic regression (to be discussed).

Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large *p*.
- Best subset selection may also suffer 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 perform poorly on future data.
- An enormous search space can lead to overfitting and high variance of the coefficient estimates.
- For both reasons, stepwise methods, which explore a far more restricted set of models, are attractive alternatives.

Forward Stepwise Selection

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
- In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.
- Computational advantage over best subset selection is clear.
- It is not guaranteed to find the best possible model out of all models containing subsets of the p predictors.

Forward Stepwise Selection: Details

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   - 2.1 Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   - 2.2 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.

Example: Credit Data (Page 231)

# 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`

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 Stepwise Selection

- Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection.

- However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Backward Stepwise Selection: Details

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   - 2.1 Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   - 2.2 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.

More on Backward Stepwise Selection

- Like forward stepwise selection, the backward selection approach searches through only $1 + p(p + 1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection.

- Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors.

- Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

Extensions

- Hybrid-stepwise selection considers both forward and backward moves at each step, and selects the best of the two.
  - Pros: computationally efficient, error made at an earlier stage can be corrected later.
  - Need a criterion to decide whether to add or drop at each step. e.g. AIC takes proper account of both the number of parameters and how good the model fits.

Validation Set and Cross-Validation

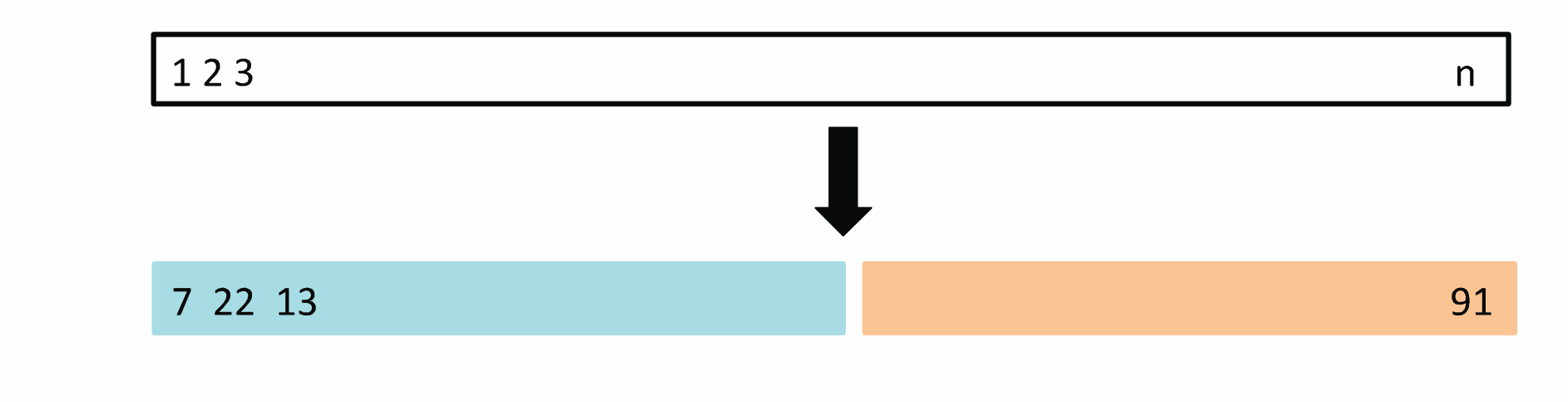
Validation and Cross-Validation

- Each of the procedures returns a sequence of models $\mathcal{M}_k$ indexed by model size $k = 0, 1, 2, \ldots$. Our job here is to select $\hat{k}$. Once selected, we will return model $\mathcal{M}_{\hat{k}}$
- We compute the validation set error or the cross-validation error for each model $\mathcal{M}_k$ 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* σ^2.
- 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 σ^2.

Validation-set Approach

- We would like to pick the model that has smallest testing error. But no test data is available at the training stage!
- Create an artificial testing data from training data!
- We randomly divide the available set of samples into two parts: a training set and a validation or hold-out set.
- The model is fit on the training set, and then used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error. Use MSE.

The Validation Process



A random splitting into two halves: left part is training set, right part is validation set

Drawbacks of the Validation Set Approach

- Only a subset of the observations are used to fit the model.
- The validation estimate of the test error can be highly variable, depending on the split of the raw data.
- The validation test error may tend to overestimate the test error for the model fit on the entire data set.
- Cross-validation (CV) to the rescue!

K-fold Cross-Validation (CV)

- *Widely used approach* for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other $K-1$ parts (combined), and then obtain predictions for the left-out kth part.
- This is done in turn for each part $k = 1, 2, \ldots K$, and then the results are combined.

K-fold CV Illustration

Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Validation	Train	Train	Train	Train

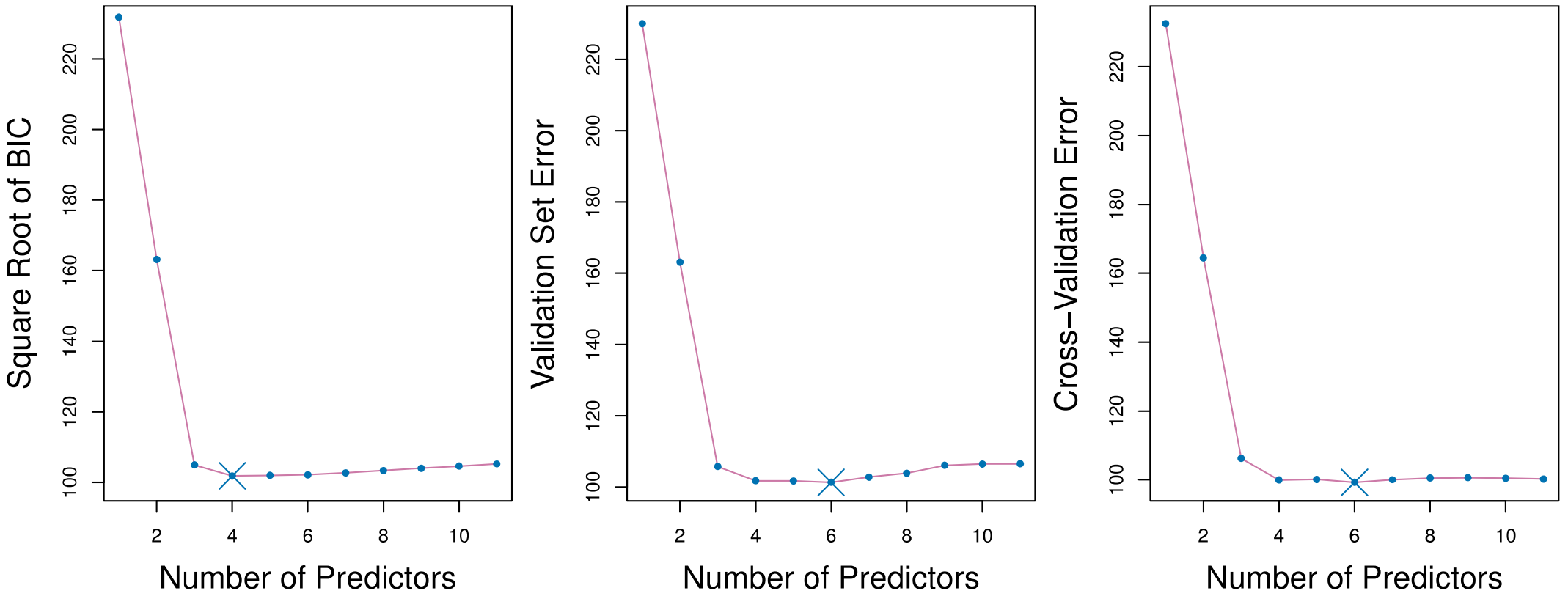
The Computing Details

- Let the K parts be $C_1, C_2, \ldots C_K$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k: if N is a multiple of K, then $n_k = n/K$.
- Compute

$$\text{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k$$

  where $\text{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.
- Setting $K = n$ yields n-fold or *leave-one out cross-validation* (LOOCV).

Credit Data Example: Validation



Details of Previous Figure

- The validation errors were calculated by randomly selecting three-quarters of the observations as the training set, and the remainder as the validation set.
- The cross-validation errors were computed using $k = 10$ folds.
- In this case, the validation and cross-validation methods both result in a six-variable model.
- However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.