KCA University
Nairobi, Kenya

BSD 3101: PRINCIPLES OF DATA SCIENCE

Prepared By:

Linus Aloo

E-mail: linusaloo88@gmail.com

Phone: 0754188380

Recommended Reading

1. Sinan Ozdemir, “Principles of Data Science,” Packt Publishing,1st Ed. 2016.
2. Vijay Kotu and Bala Deshpande, “Data Science: Concepts and Practice,” Elsevier, Second Edition, 2019.
3. Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani, “An Introduction to Statistical Learning (with Applications in R), Springer (1st Ed.), 2014.

Additional Reading

1. Joel Grus, “Data Science from Scratch,” O’Reilly Media, 1st Ed., 2015.

CHAPTER 8: LINEAR MODEL SELECTION AND REGULARIZATION

8.0. Lesson Goals

- ❑ Understand best subset selection and stepwise selection methods for reducing the number of predictor variables in regression.
- ❑ Indirectly estimate test error by adjusting training error to account for bias due to overfitting (C_p, AIC, BIC, adjusted R^2).
- ❑ Directly estimate test error using validation set approach and cross-validation approach.
- ❑ Understand and know how to perform ridge regression and the lasso as shrinkage (regularization) methods.
- ❑ Understand and know how to perform principal components regression and partial least squares as dimension reduction methods.

8.1. Improving the Linear Model

- ❑ We may want to improve the simple linear model by replacing OLS (Ordinary Least Squares) estimation with some alternative fitting procedure.
- ❑ Why use an alternative fitting procedure?
  - ✓ Prediction Accuracy
  - ✓ Model Interpretability

Prediction Accuracy

- ❑ The OLS estimates have relatively **low bias** and **low variability** especially when the relationship between the response and predictors is linear and n >> p.

CHAPTER 8: LINEAR MODEL SELECTION AND REGULARIZATION

8.1. Improving the Linear Model

Prediction Accuracy

- ❑ If n is not much larger than p, then the OLS fit can have high variance and may result in over fitting and poor estimates on unseen observations.
- ❑ If $p > n$, then the variability of the OLS fit increases dramatically, and the variance of these estimates is infinite.

Model Interpretability

- ❑ When we have a large number of predictors in the model, there will generally be many that have little or no effect on the response.

CHAPTER 8: LINEAR MODEL SELECTION AND REGULARIZATION

8.1. Improving the Linear Model

Model Interpretability

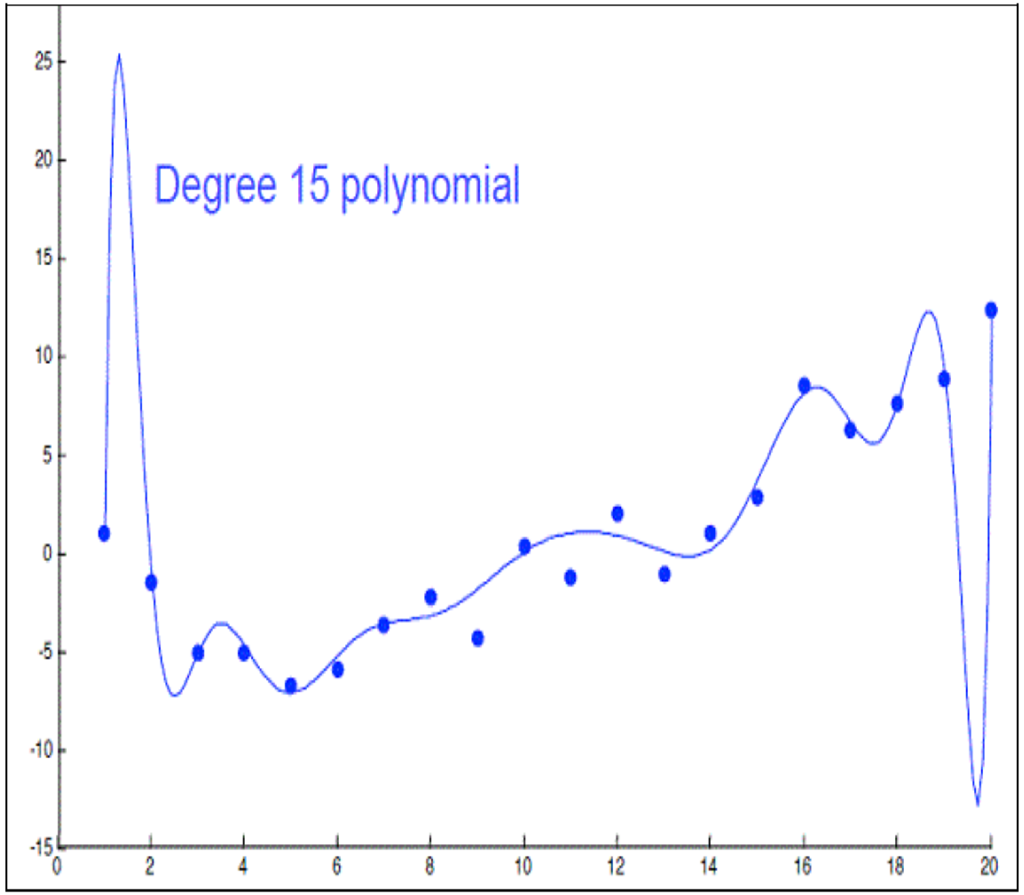
- By removing irrelevant features - that is, by setting the corresponding coefficient estimates to zero - we can obtain a model that is more easily interpreted.
- We will present some approaches for automatically performing ***feature selection***.

8.2. Feature/Variable Selection

- Carefully selected features can improve model accuracy, but adding too many can lead to overfitting.
  - Overfitted models describe random error or noise instead of any underlying relationship.
  - They generally have poor predictive performance on test data.

8.2. Feature/Variable Selection

- ❑ For instance, we can use a 15-degree polynomial function to fit the following data so that the fitted curve goes nicely through the data points.
- ❑ However, a brand new dataset collected from the same population may not fit this particular curve well at all.



8.2. Feature/Variable Selection

Three classes of methods:

1. Subset Selection

- Identify a subset of the *p* predictors that we believe to be related to the response; then, fit a model using OLS on the reduced set.
- Methods: best subset selection, stepwise selection

2. Shrinkage (Regularization)

- Involves shrinking the estimated coefficients toward zero relative to the OLS estimates; has the effect of reducing variance and performs variable selection.
- Methods: ridge regression, lasso

8.2. Feature/Variable Selection

Three classes of methods:

3. Dimension Reduction

- Involves projecting the *p* predictors into a *M*-dimensional subspace, where $M < p$, and fit the linear regression model using the *M* projections as predictors.
- Methods: principal components regression, partial least squares.

8.3. Subset Selection

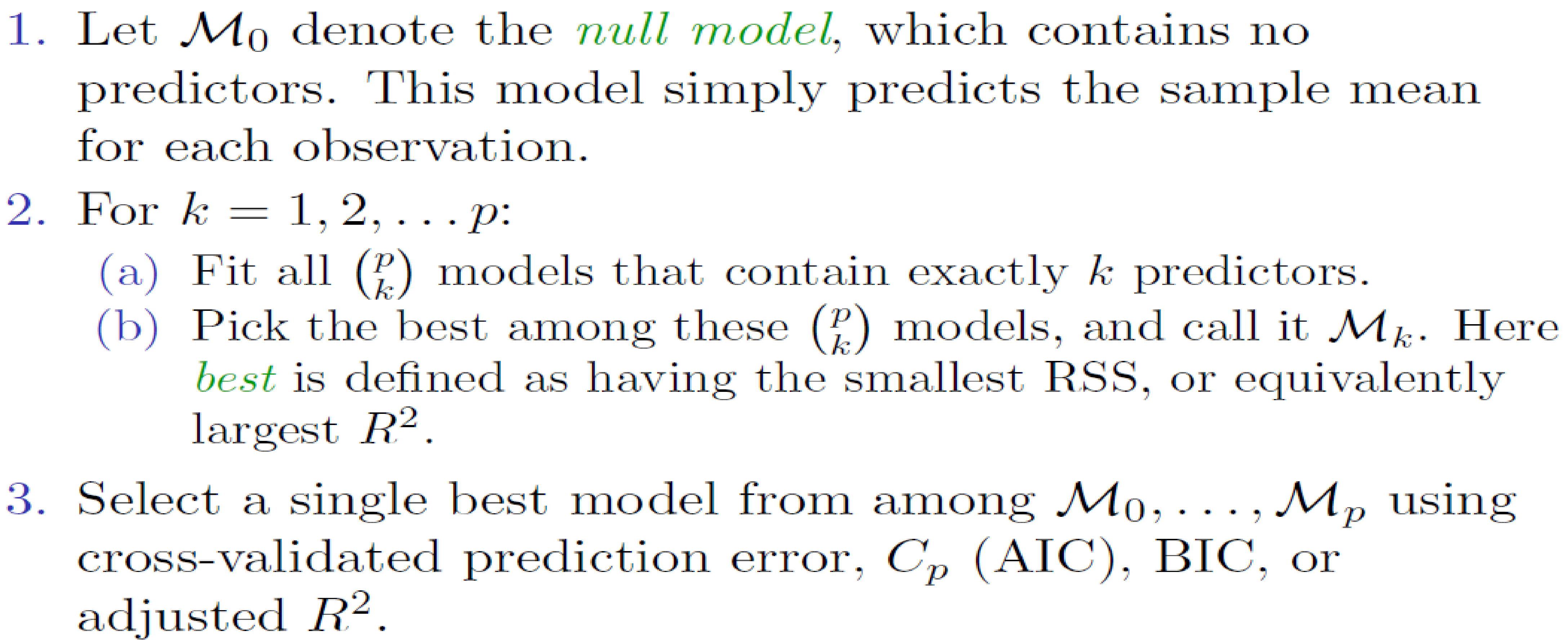
❑ Best subset and stepwise model selection procedures

8.3.1 Best Subset Selection

❑ We fit a separate OLS regression for each possible combination of the p predictors:

CHAPTER 8: LINEAR MODEL SELECTION AND REGULARIZATION

8.3.1 Best Subset Selection

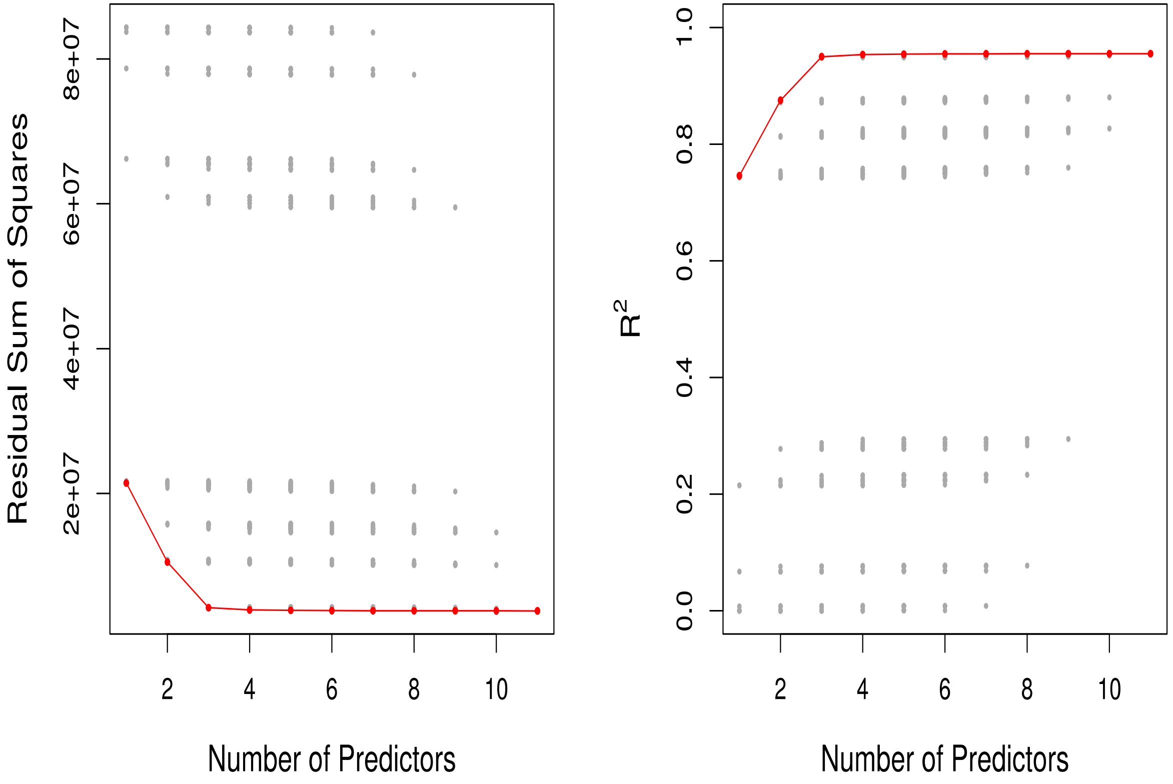


1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   - (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   - (b) Pick the best among these $\binom{p}{k}$ models, 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.

8.3.1 Best Subset Selection

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 RSS (R^2) will always decline (increase) as the number of predictors included in the model increases, so they are not very useful statistics for selecting the *best* model.
- ❑ The red line tracks the best model for a given number of predictors, according to RSS and R^2



8.3. Best Subset Selection

- While best subset selection is a simple and conceptually appealing approach, it suffers from computational limitations.
- The number of possible models that must be considered grows rapidly as p increases.
- Best subset selection becomes computationally infeasible for value of p greater than around 40.

8.3.2 Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large p.
- 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.
- An enormous search space can lead to overfitting and high variance of the coefficient estimates.

8.3.2 Stepwise Selection

- ❑ More attractive methods include:

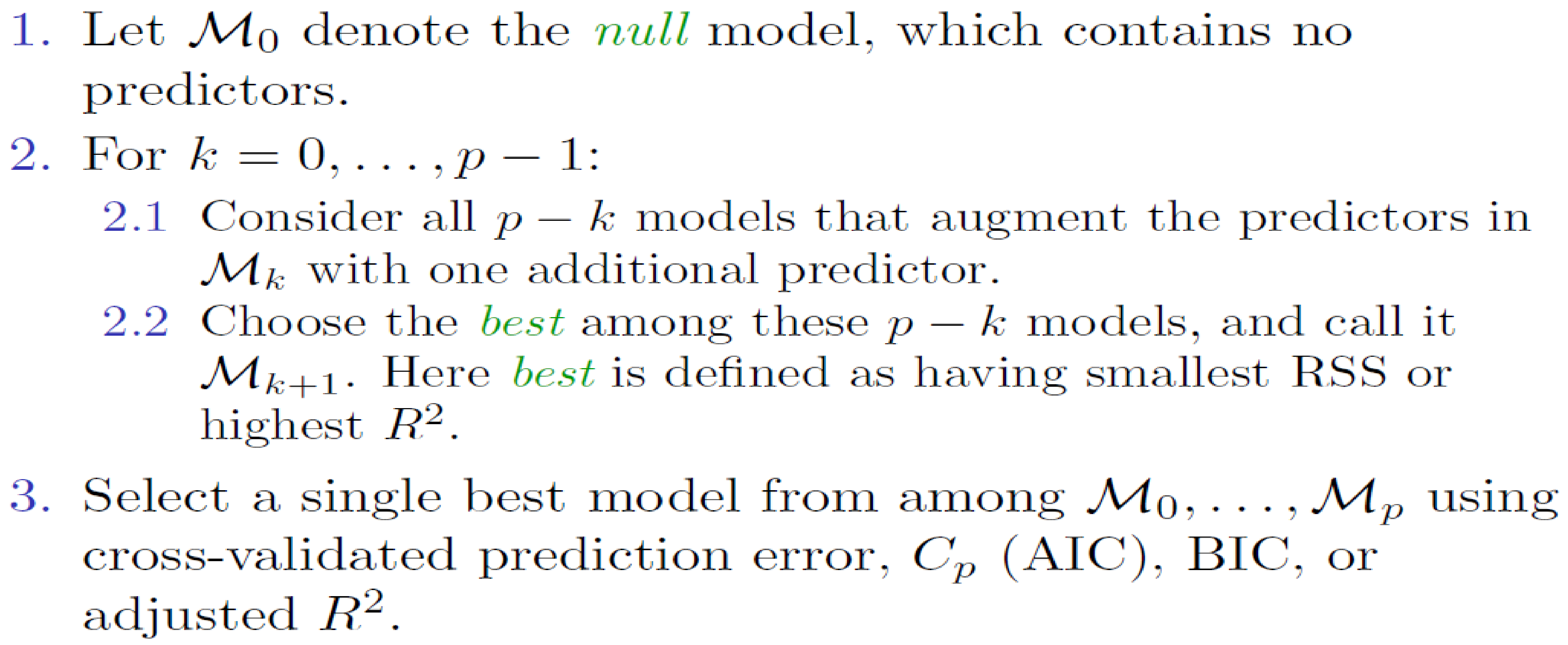
Forward Stepwise Selection

- ❑ Begins with a null OLS model containing no predictors, and then adds one predictor at a time that improves the model the most until no further improvement is possible.

Backward Stepwise Selection

- ❑ Begins with a full OLS model containing all predictors, and then deletes one predictor at a time that improves the model the most until no further improvement is possible.

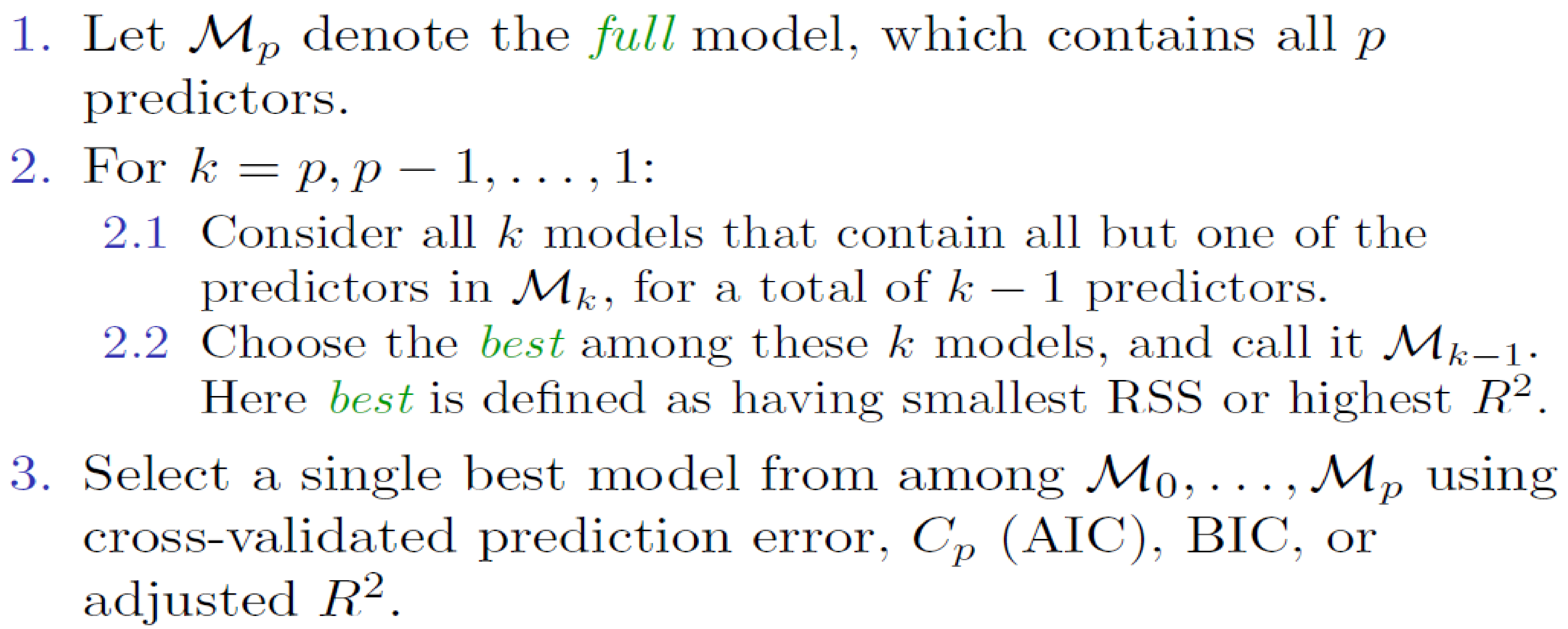
Forward Stepwise Selection



```
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 $R^2$.
```

CHAPTER 8: LINEAR MODEL SELECTION AND REGULARIZATION

Backward Stepwise Selection



1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   - 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.

Stepwise Selection (cont.)

- ❑ Both forward and backward stepwise selection approaches search through only $1 + p(p + 1)/2$ models, so they can be applied in settings where p is too large to apply best subset selection.
- ❑ Both of these stepwise selection methods are *not* guaranteed to yield the best model containing a subset of the p predictors.
- ❑ Forward stepwise selection can be used even when $n < p$, while backward stepwise selection requires that $n > p$.
- ❑ There is a *hybrid* version of these two stepwise selection methods.

8.4. Choosing the Optimal Model

- ❑ The model containing all 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.
- ❑ Thus, 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

1. We can indirectly estimate test error by making an *adjustment* to the training error to account for the bias due to overfitting.
2. We can *directly* estimate the test error, using either a validation set approach or a cross-validation approach.

8.5. Other Measures of Comparison

□ To compare different models, we can use other approaches:

- Adjusted R^2
- AIC (Akaike information criterion)
- BIC (Bayesian information criterion)
- Mallow’s C_p (equivalent to AIC for linear regression)

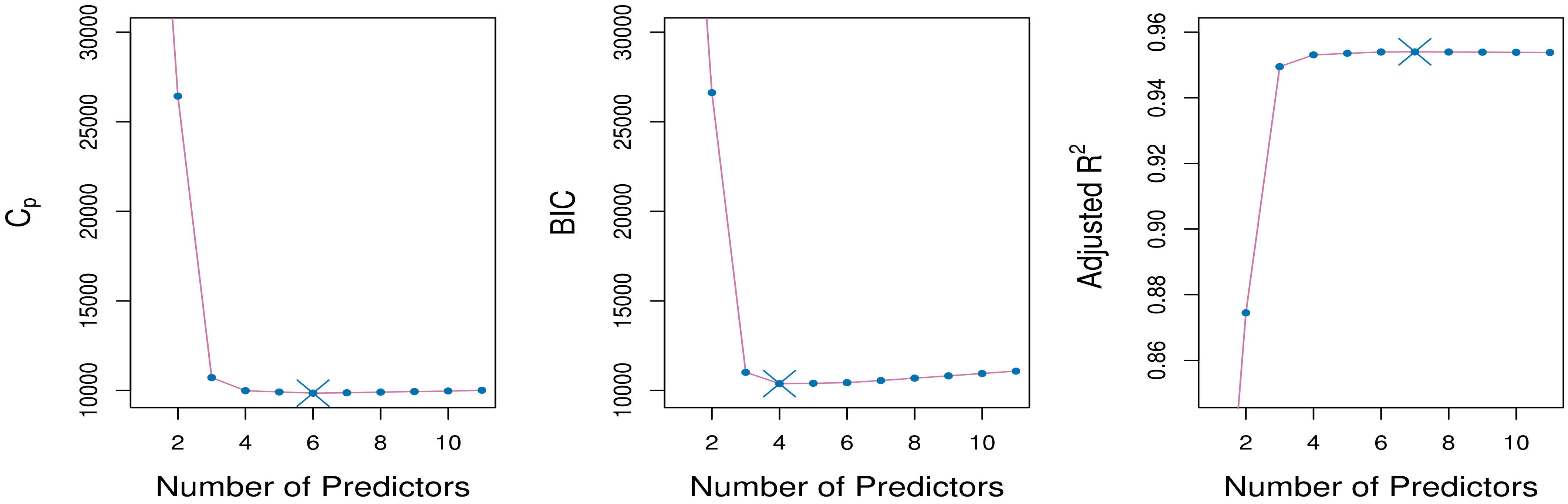
□ These techniques adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables.

□ These methods add penalty to RSS for the number of predictors in the model.

8.5. Other Measures of Comparison

Credit Data: C_p, BIC, and Adjusted R²

- ❑ A small value of C_p and BIC indicates a low error, and thus a better model.
- ❑ A large value for the Adjusted R² indicates a better model.



8.5. Other Measures of Comparison

8.5.1. Mallow's C_p

❑ For a fitted OLS model containing *d* predictors, the C_p estimate of test MSE:

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

where $\hat{\sigma}^2$ is an estimate of the variance of the error ε associated with each response measurement.

❑ Here, a penalty is added to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error.

8.5. Other Measures of Comparison

8.5.2. Akaike Information Criterion (AIC)

- Defined for a large class of models fit by maximum likelihood.

$$\text{AIC} = -2\log L + 2 \cdot d$$

where *L* is the maximized value of the likelihood function for the estimated model.

- In the case of the linear model with Gaussian errors, MLE and OLS are the same things; thus, C_p and AIC are equivalent.

8.5. Other Measures of Comparison

8.5.3. Bayesian Information Criterion (BIC)

- BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value.

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

- Since log *n* > 2 for an *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 $\mathrm{C_p}$.
- Notice that BIC replaces the $2d\hat{\sigma}^2$ used by $\mathrm{C_p}$ with a $\log(n)d\hat{\sigma}^2$ term, where *n* is the number of observations.

8.5. Other Measures of Comparison

8.5.4. Adjusted R^2

- For an OLS model with *d* variables, the adjusted R^2 is calculated:

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

where TSS is the total sum of squares.

- Unlike the other statistics, a large value of adjusted R^2 indicates a model with a small test error.
- The adjusted R^2 statistics *pays a price* for the inclusion of unnecessary variables in the model.

8.6. Validation and Cross-Validation

8.5.4. Adjusted R^2

- Each of the procedures returns a sequence of models indexed by model size k = 0, 1, 2, … Our job here is to select $\hat{k}$.
- We compute the validation set error or the CV error for each model under consideration, and then select the k for which the resulting estimated test error is smallest.
- This procedure provides a direct estimate of the test error, and it can also be used in a wider range of model selection tasks.
- We can also select a model using the *one-standard-error rule*.