WEEK 06

INSTRUCTOR: YANAN WU

TA: KHADIJA NISAR

SPRING 2025

5.1
MODEL SELECTION

STANDARD LINEAR MODEL

- $Y = \beta_0 + \beta_1 X_1 + \ \ldots + \beta_p X_P + \ \epsilon$

  1) Irrelevant variables lead to unnecessary complexity in the model
  2) Feature selection or variable selection: excluding irrelevant variables from a multiple regression model

SUBSET SELECTION

- Identifying a subset of the p predictors that related to the response
- Fitting a model using least squares on the reduced set of variables

BEST SUBSET SELECTION

1. Let 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, . . .*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 M_k. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $M_0, \ldots, M_p$ using cross validated prediction error, C_p ,AIC, BIC, or adjusted R^2.

BOSTON HOUSING PRICE DATA

Variable	Description
medv	**Median value of owner-occupied homes in $1000s**
crim	Per capita crime rate by town
zn	Proportion of residential land zoned for large lots
indus	Proportion of non-retail business acres per town
chas	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
nox	Nitrogen oxide concentration (parts per 10 million)
rm	Average number of rooms per dwelling
age	Proportion of owner-occupied units built before 1940
dis	Weighted mean of distances to employment centers
rad	Index of accessibility to radial highways
tax	Property tax rate per $10,000
ptratio	Pupil-teacher ratio by town
black	Proportion of Black residents per town
lstat	Lower status population percentage

$$medv = \beta_0 + \beta_1 zn + \beta_2 crim + \cdots \beta_{13} lstat$$

BEST SUBSET SELECTION FOR BOSTON DATA

1. $y \sim 1$ (Null Model)
2. $y = \beta_0 + \beta_1 x_1, y = \beta_0 + \beta_2 x_2 \ldots, y = \beta_{13} x_{13}$ (Models with k=1) (Best model: smallest RSS, or largest R^2)
3. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2, y = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \ldots y = \beta_0 + \beta_{p-1} x_{p-1} + \beta_p x_p$ (Models with k=2) (Best model: smallest RSS, or largest R^2)
4. Continue adding variables until the best model is selected.
5. Select a single best model from among $M_0, \ldots, M_p$ using cross validated prediction error, C_p ,AIC, BIC, or adjusted R^2.

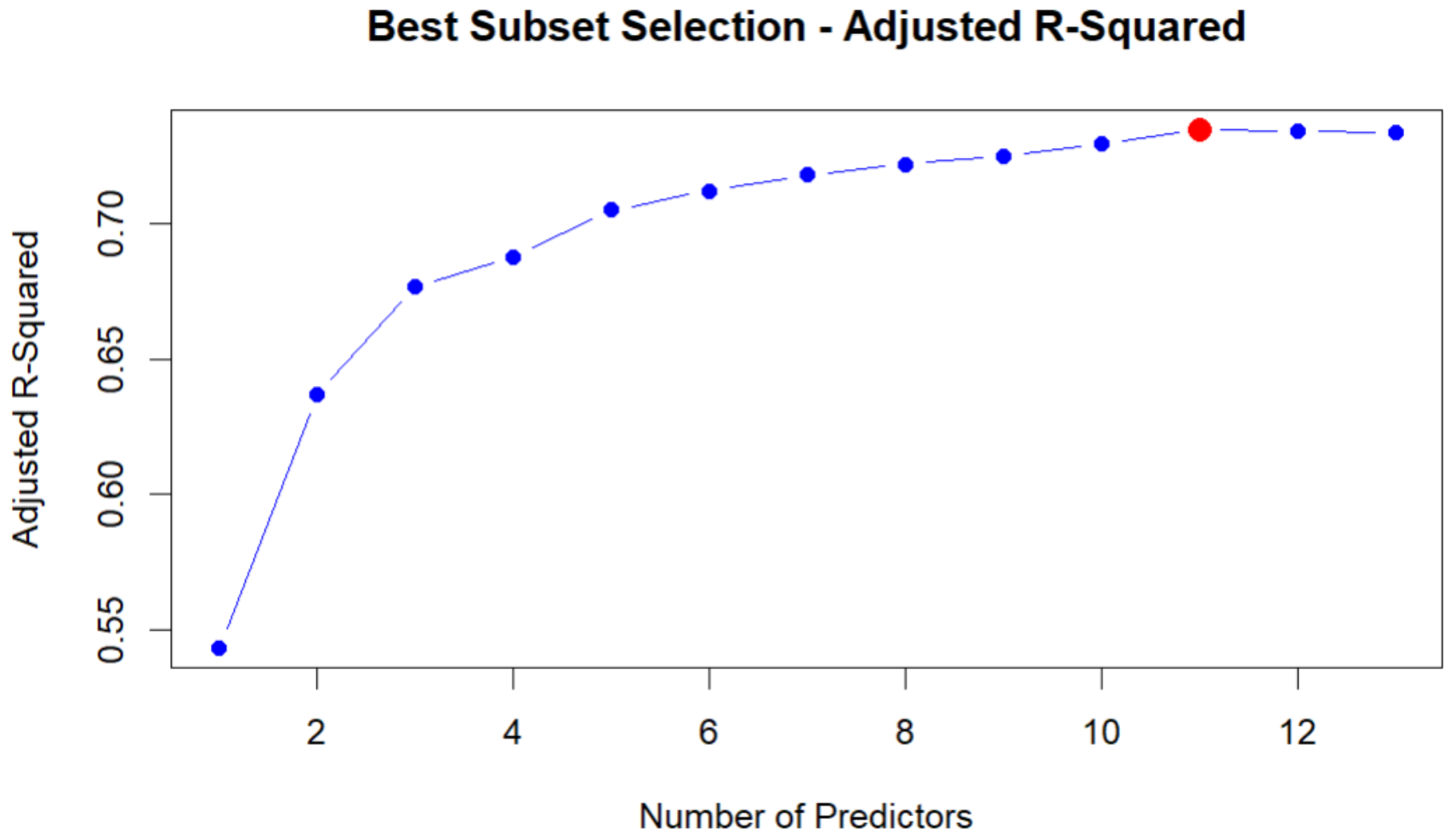
BEST SUBSET SELECTION

1. $medv \sim 1$ (Null Model)
2. $medv = \beta_0 + \beta_1 crime, medv = \beta_0 + \beta_1 rm$ …,$medv = \beta_0 + \beta_{13} tax$ (Single Predictor Models)
3. $medv = \beta_0 + \beta_1 crime + \beta_2 rm, medv = \beta_0 + \beta_1 crime + \beta_2 tax$,… $medv = \beta_0 + \beta_{p-1} black + \beta_p lstat$ (Two Predictor Models)
4. Continue adding variables until the best model is selected.

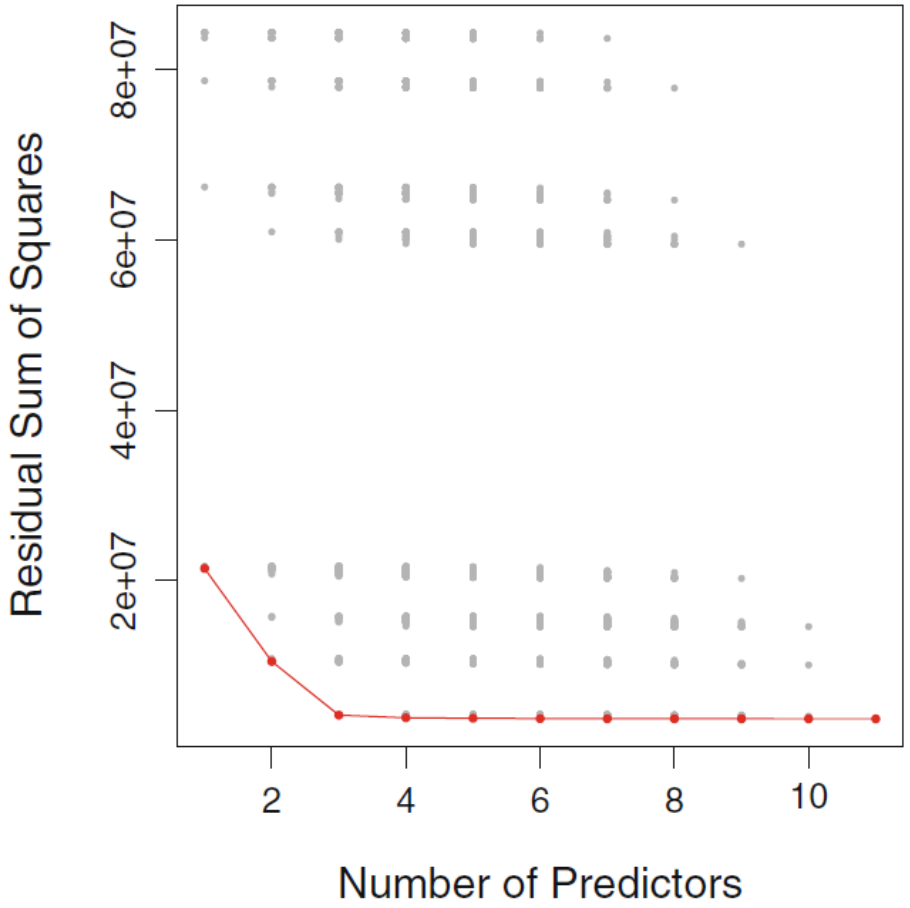
regsubsets() in R

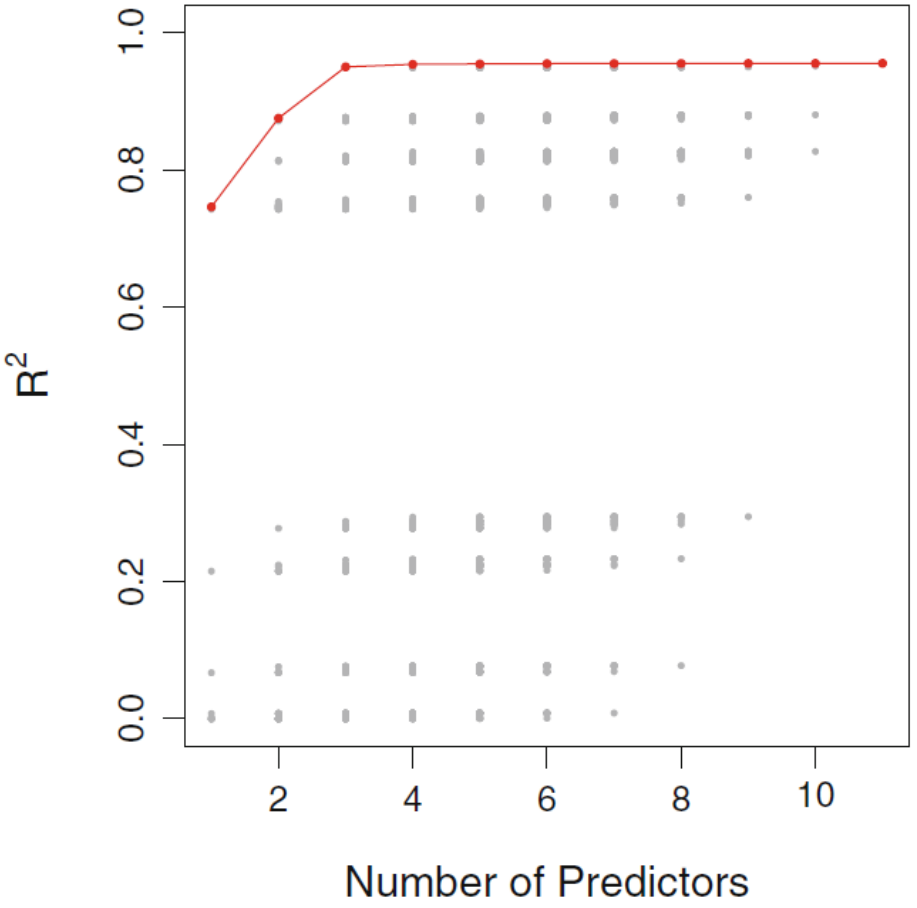
```
Subset selection object
Call: regsubsets.formula(medv ~ ., data = Boston, nbest = 1, nvmax = 13)
13 Variables  (and intercept)
        Forced in Forced out
crim        FALSE      FALSE
zn          FALSE      FALSE
indus       FALSE      FALSE
chas        FALSE      FALSE
nox         FALSE      FALSE
rm          FALSE      FALSE
age         FALSE      FALSE
dis         FALSE      FALSE
rad         FALSE      FALSE
tax         FALSE      FALSE
ptratio     FALSE      FALSE
black       FALSE      FALSE
lstat       FALSE      FALSE
1 subsets of each size up to 13
Selection Algorithm: exhaustive
          crim zn  indus chas nox rm  age dis rad tax ptratio black lstat
1  ( 1 )  " "  " " " "   " "  " " " " " " " " " " " "     " "   "*"
2  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " " "     " "   "*"
3  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " "*"     " "   "*"
4  ( 1 )  " "  " " " "   " "  " " "*" " " "*" " " " " "*"     " "   "*"
5  ( 1 )  " "  " " " "   " "  "*" "*" " " "*" " " " " "*"     " "   "*"
6  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " " " "*"     " "   "*"
7  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " " " "*"     "*"   "*"
8  ( 1 )  " "  "*" " "   "*"  "*" "*" " " "*" " " " " "*"     "*"   "*"
9  ( 1 )  "*"  " " " "   "*"  "*" "*" " " "*" "*" " " "*"     "*"   "*"
10  ( 1 ) "*"  "*" " "   " "  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
11  ( 1 ) "*"  "*" " "   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
12  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
13  ( 1 ) "*"  "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"
```

BEST MODEL BASED ON ADJUSTED R^2



RSS and R^2 change with the number of predictors increase



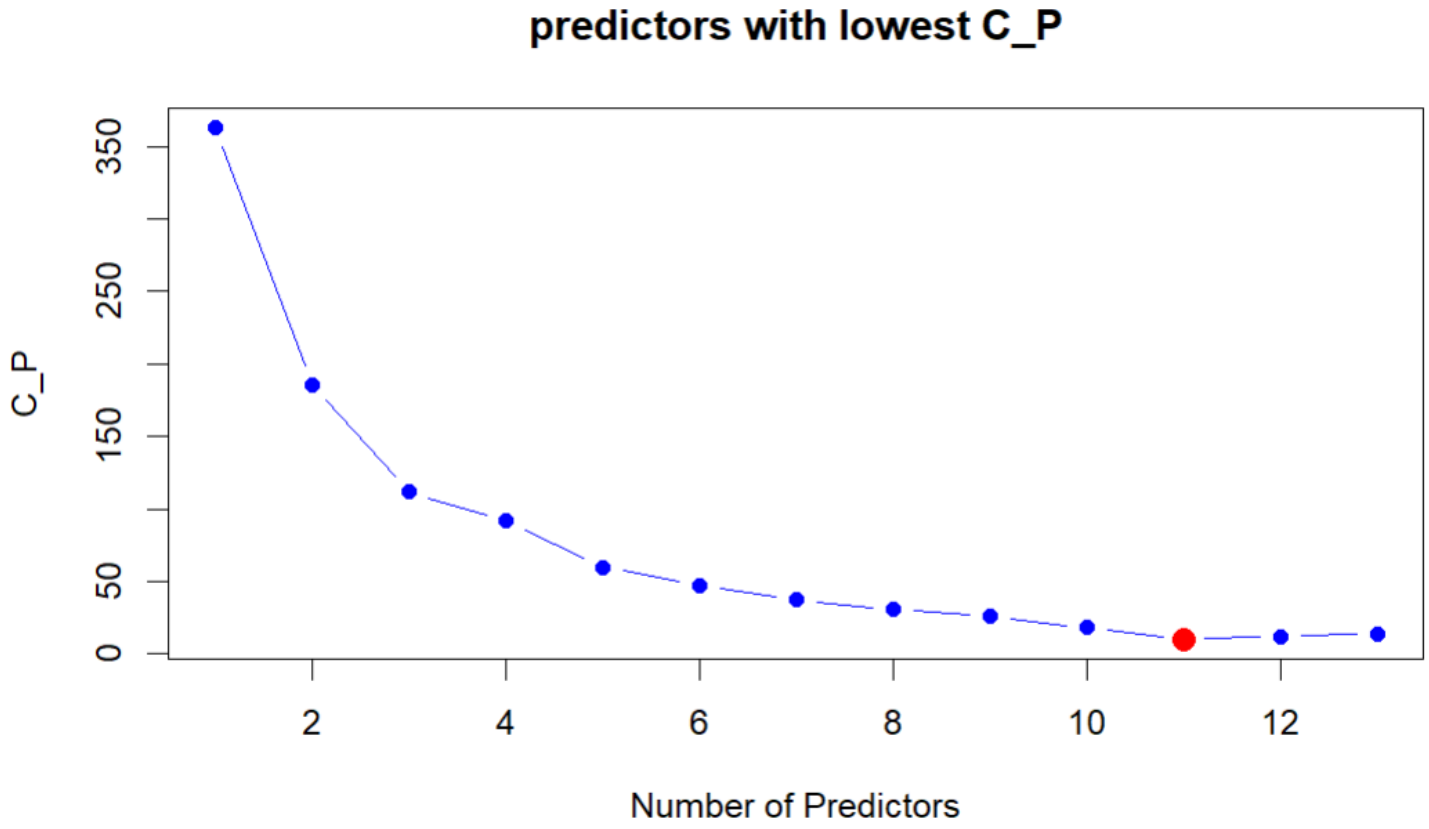


James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in python*. Springer Nature.

BEST MODEL BASED ON C_p, AIC

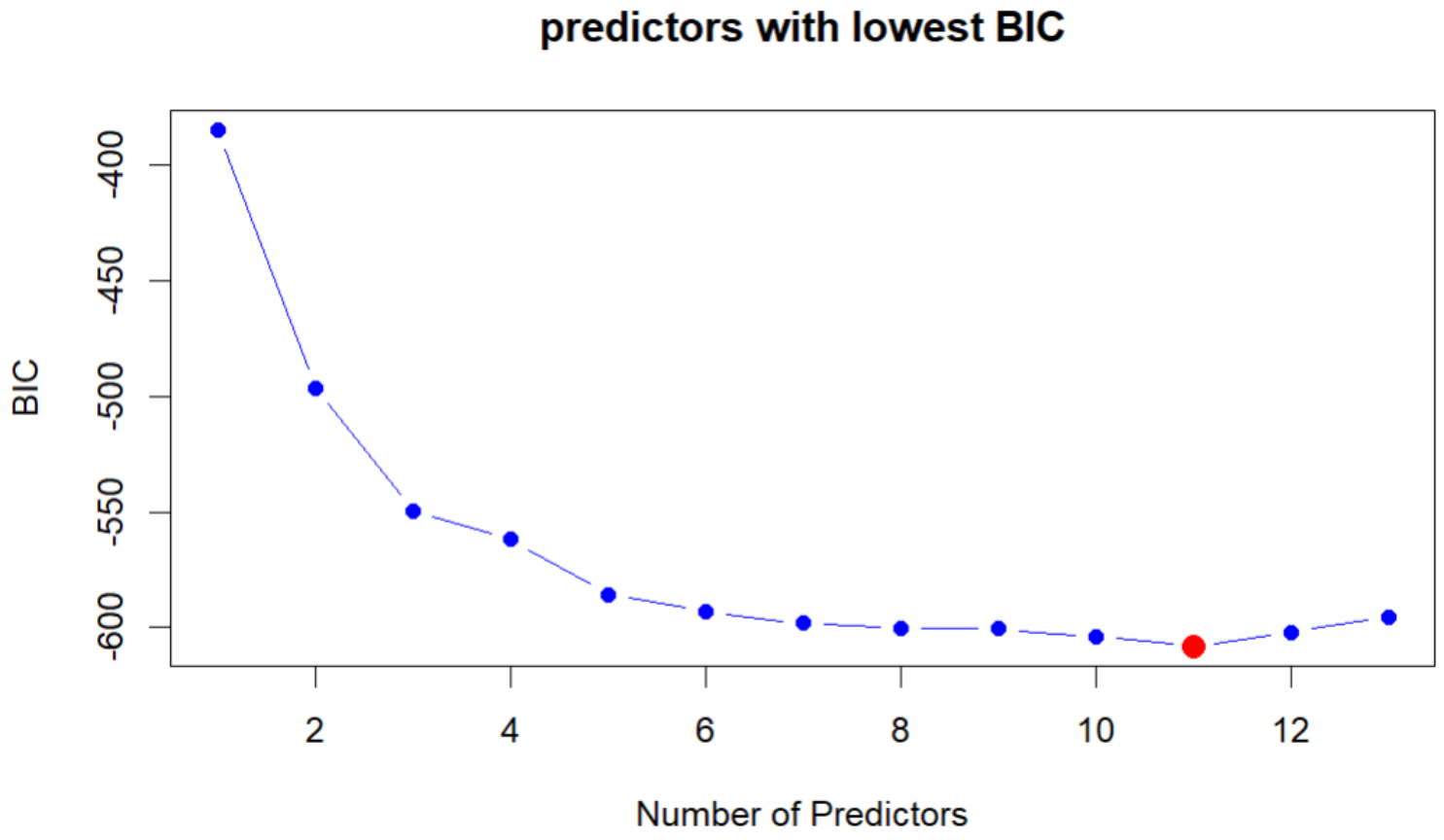
- $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$
- $AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$

C_p and AIC are proportional to each other

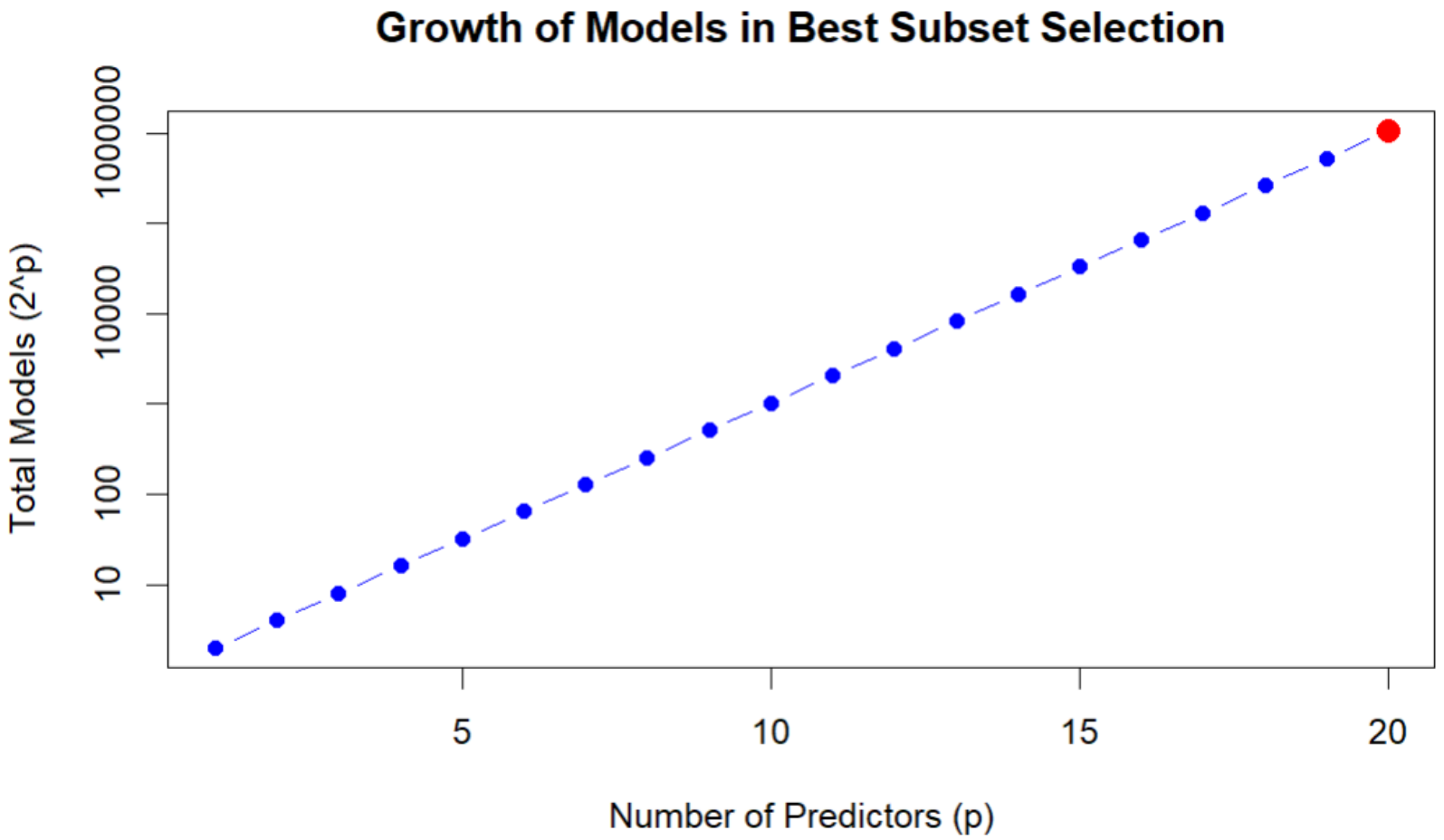


BEST MODEL BASED ON BIC

- $\mathrm{B}IC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$



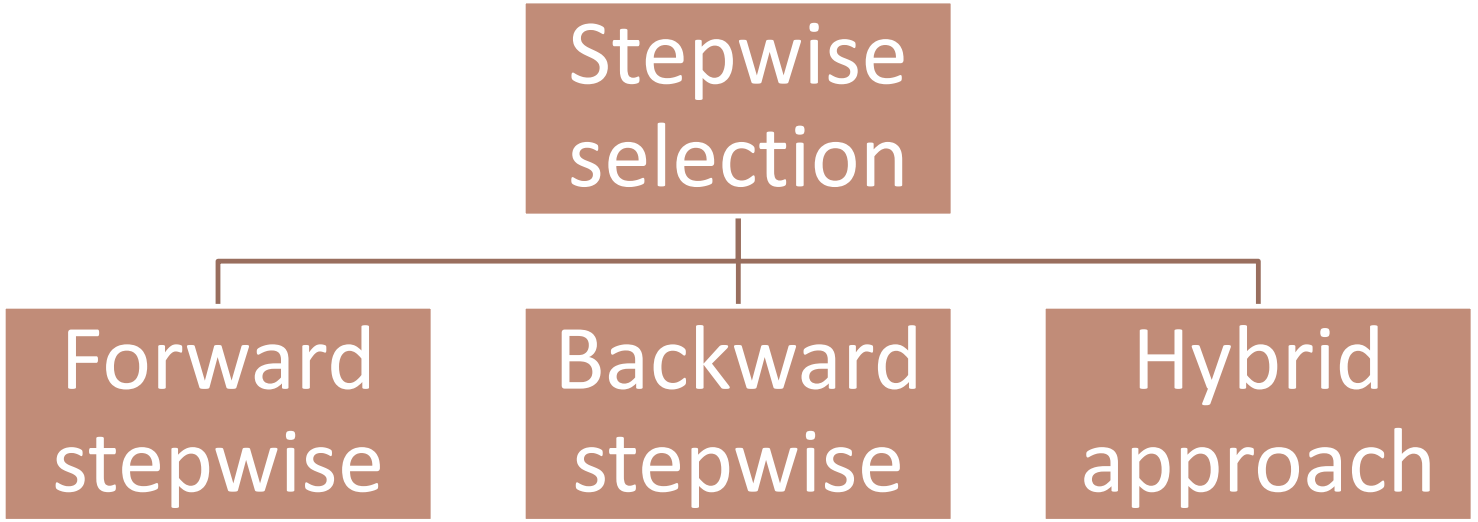
EXPONENTIAL COMPUTATIONAL COMPLEXITY



Issue: Becomes computationally infeasible for large p.

2^p models for p predictors

STEPWISE SELECTION



FORWARD STEPWISE SELECTION

1. $medv \sim 1$ (Null Model)

2. $medv = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 crime}, medv = \beta_0 + \beta_1 rm$ …,$medv = \beta_0 + \beta_{13} tax$ (Single Predictor Models)

3. $medv = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 crime} + \beta_2 rm, medv = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 crime} + \beta_2 tax$,… $medv = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 crime} + \beta_p lstat$

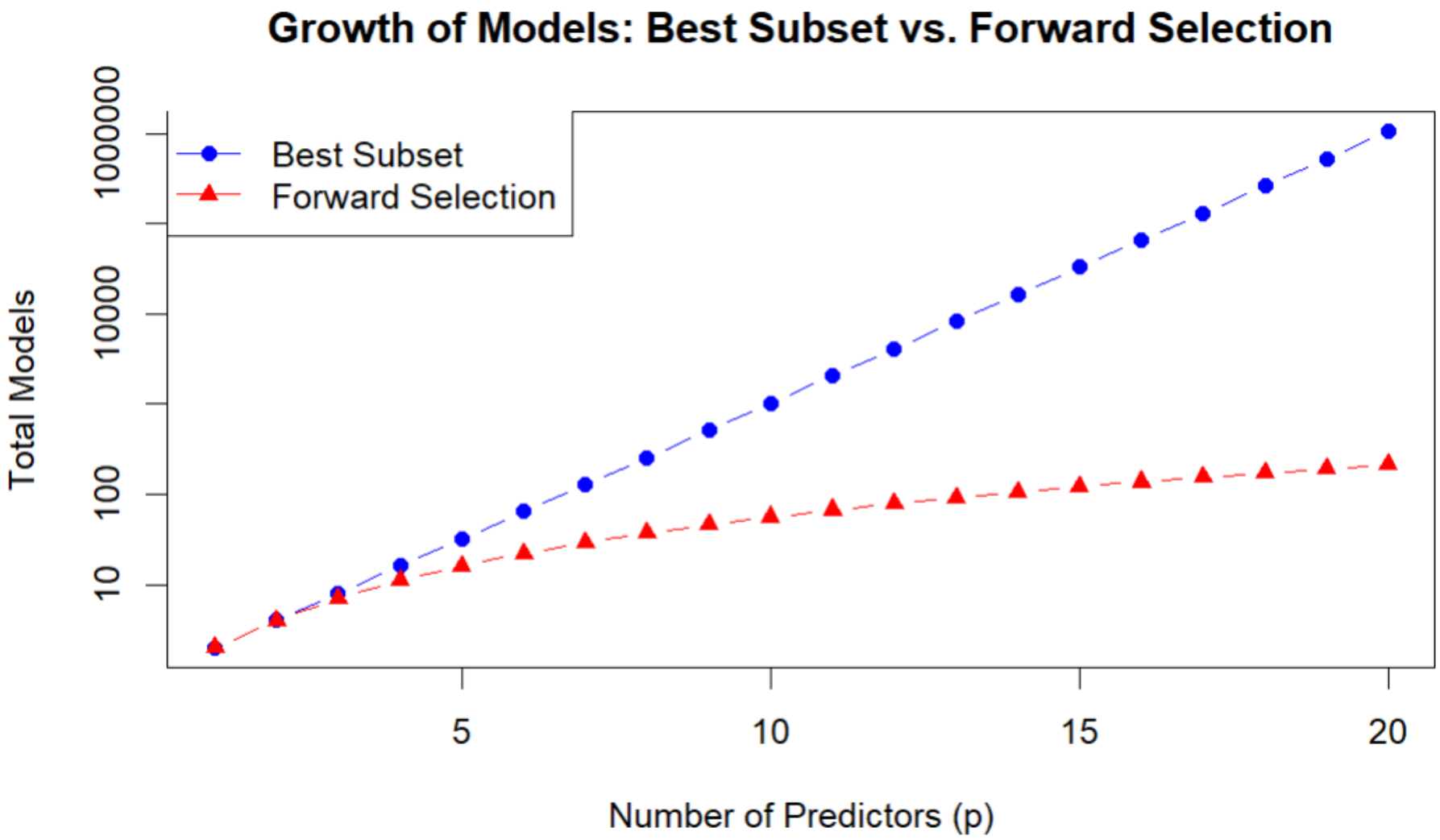
   (Two Predictor Models)

4. $medv = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 crime} + \boldsymbol{\beta_2 tax} + \beta_3 medv = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 crime} + \boldsymbol{\beta_2 tax} + \beta_3 chase = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 crime} +$

   $\boldsymbol{\beta_2 tax} + \beta_p lstat$ (Three Predictor Models)

5. Continue adding variables until the best model is selected (using $C_p, AIC, BIC, or\ adjusted\ R^2$).

COMPARISON ON GROWTH OF MODELS

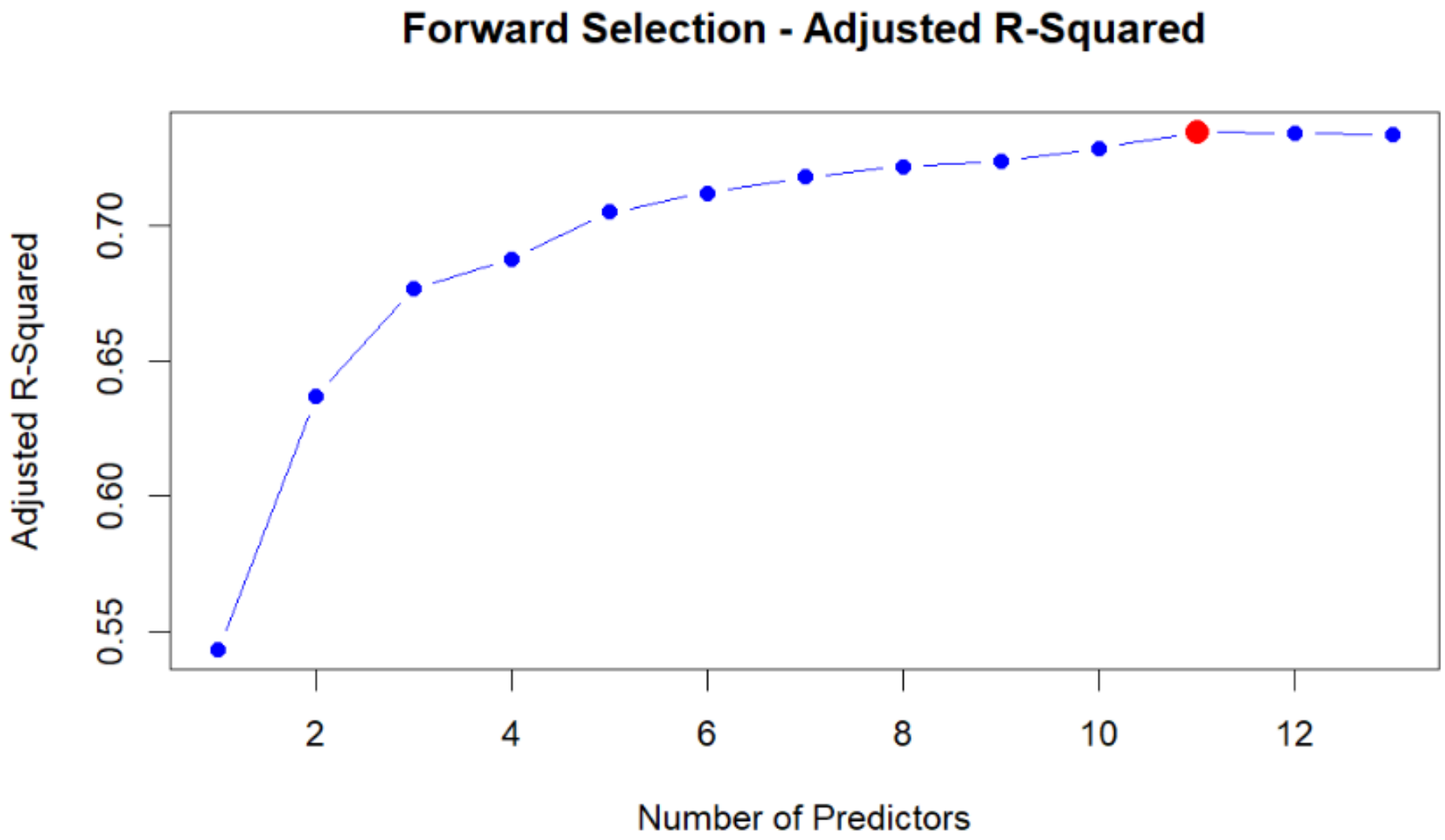


$1 + p(p + 1)$ models for p predictors

BEST MODEL BASED ON FORWARD SELECTION

```
Subset selection object
Call: regsubsets.formula(medv ~ ., data = Boston, nvmax = 13, method = "forward")
13 Variables  (and intercept)
        Forced in Forced out
crim        FALSE      FALSE
zn          FALSE      FALSE
indus       FALSE      FALSE
chas        FALSE      FALSE
nox         FALSE      FALSE
rm          FALSE      FALSE
age         FALSE      FALSE
dis         FALSE      FALSE
rad         FALSE      FALSE
tax         FALSE      FALSE
ptratio     FALSE      FALSE
black       FALSE      FALSE
lstat       FALSE      FALSE
1 subsets of each size up to 13
Selection Algorithm: forward
          crim zn  indus chas nox rm  age dis rad tax ptratio black lstat
1  ( 1 )  " "  " " " "   " "  " " " " " " " " " " " "     " "   "*"
2  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " "     " "   "*"
3  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " "*"     " "   "*"
4  ( 1 )  " "  " " " "   " "  " " "*" " " "*" " " " " "*"     " "   "*"
5  ( 1 )  " "  " " " "   " "  "*" "*" " " "*" " " " " "*"     " "   "*"
6  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " " " "*"     " "   "*"
7  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " " " "*"     "*"   "*"
8  ( 1 )  " "  "*" " "   "*"  "*" "*" " " "*" " " " " "*"     "*"   "*"
9  ( 1 )  "*"  "*" " "   "*"  "*" "*" " " "*" " " " " "*"     "*"   "*"
10  ( 1 ) "*"  "*" " "   "*"  "*" "*" " " "*" "*" " " "*"     "*"   "*"
11  ( 1 ) "*"  "*" " "   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
12  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
13  ( 1 ) "*"  "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"
```

ADJUSTED R^2


Forward Selection - Adjusted R-Squared
Adjusted R-Squared
0.55
0.60
0.65
0.70
2
4
6
8
10
12
Number of Predictors

POTENTIAL ISSUE IN FORWARD SELECTION

- **Forward Selection** follows a **greedy approach**, adding one variable at a time, which can lead to suboptimal models.
- Once a variable is included, it **cannot be removed later**, even if adding another variable would result in a better model.
- If two predictors are **highly correlated**, Forward Selection may pick **one too early**, preventing the other from being added—even if the other would lead to a better overall model.

BACKWARD STEPWISE SELECTION

1. $medv \sim \beta_0 + \beta_1 crime + \ldots + \beta_{13} tax$ (**Full Model**)
2. Compute the regression in the previous step while considering removing exactly one of the X variables, which regressions with **12** predictors.
3. Compute the regression in the previous step while considering removing exactly one of the remaining X variables, which creates regressions with **11** predictors.
4. Continue to perform steps similar to step 3 until all remaining X variables do have strong association.

BACKWARD STEPWISE SELECTION IN R

full_model <- lm(medv ~ ., data = Boston

AICbackward_model <- step(full_model, direction = "backward

summarysummary(backward_model)

By default, step() selects the model that minimizes AIC.

UPDATED

- Task 2.d and 2.b in Lab 03

WEEK 03

CODE DEMO SESSION

Instructor: Yanan Wu
TA: Khadija Nisar

Spring 2025