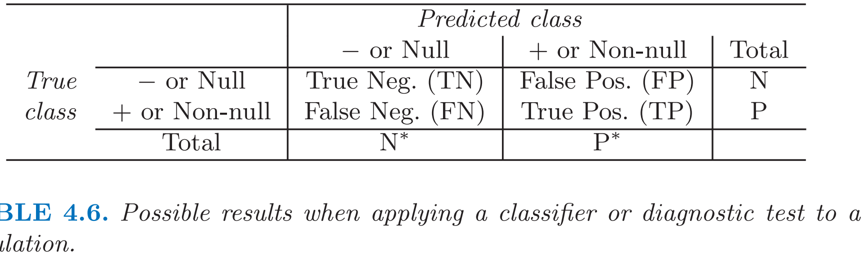
Week 5

Cogs 109: Data Analysis and Modeling

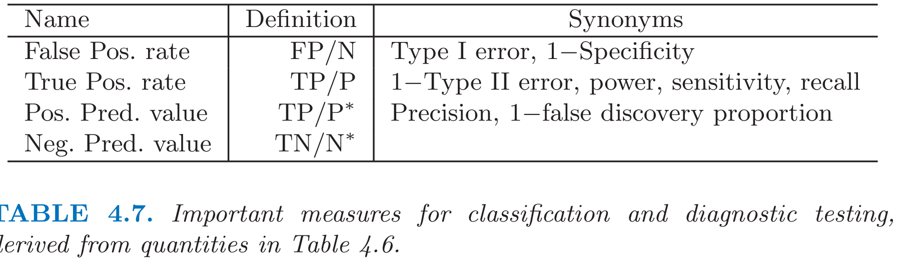
Fall 2017
Prof. Eran Mukamel

ROC plot concepts



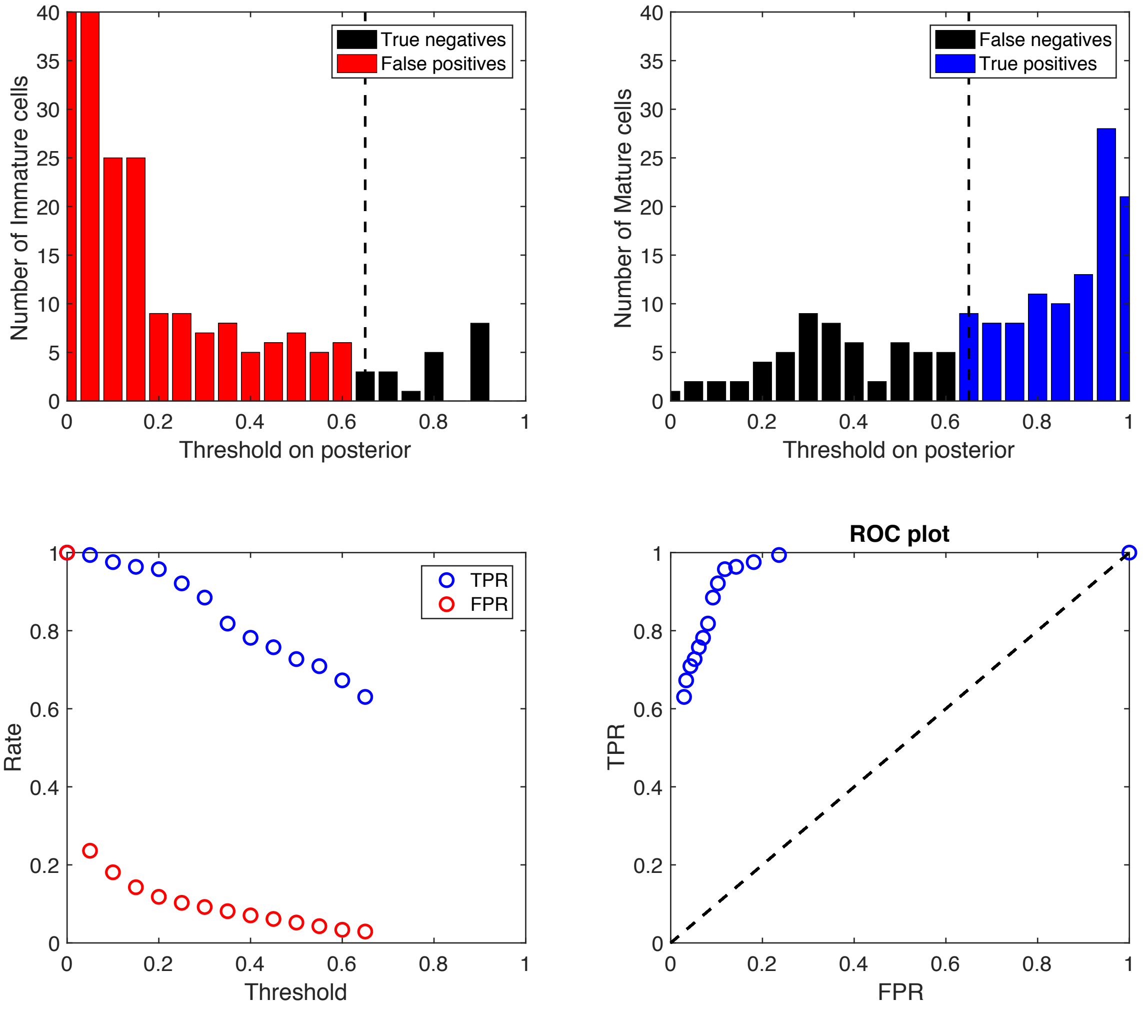
		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N^*	P^*	

ABLE 4.6. *Possible results when applying a classifier or diagnostic test to a pulation.*



Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P^*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N^*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*



A “model” is a whole class of possible predictions

- Linear, quadratic and cubic polynomials are different “models” or “model classes”

$y = \beta_0 + \beta_1 x + \varepsilon$ **Model 1: Linear**

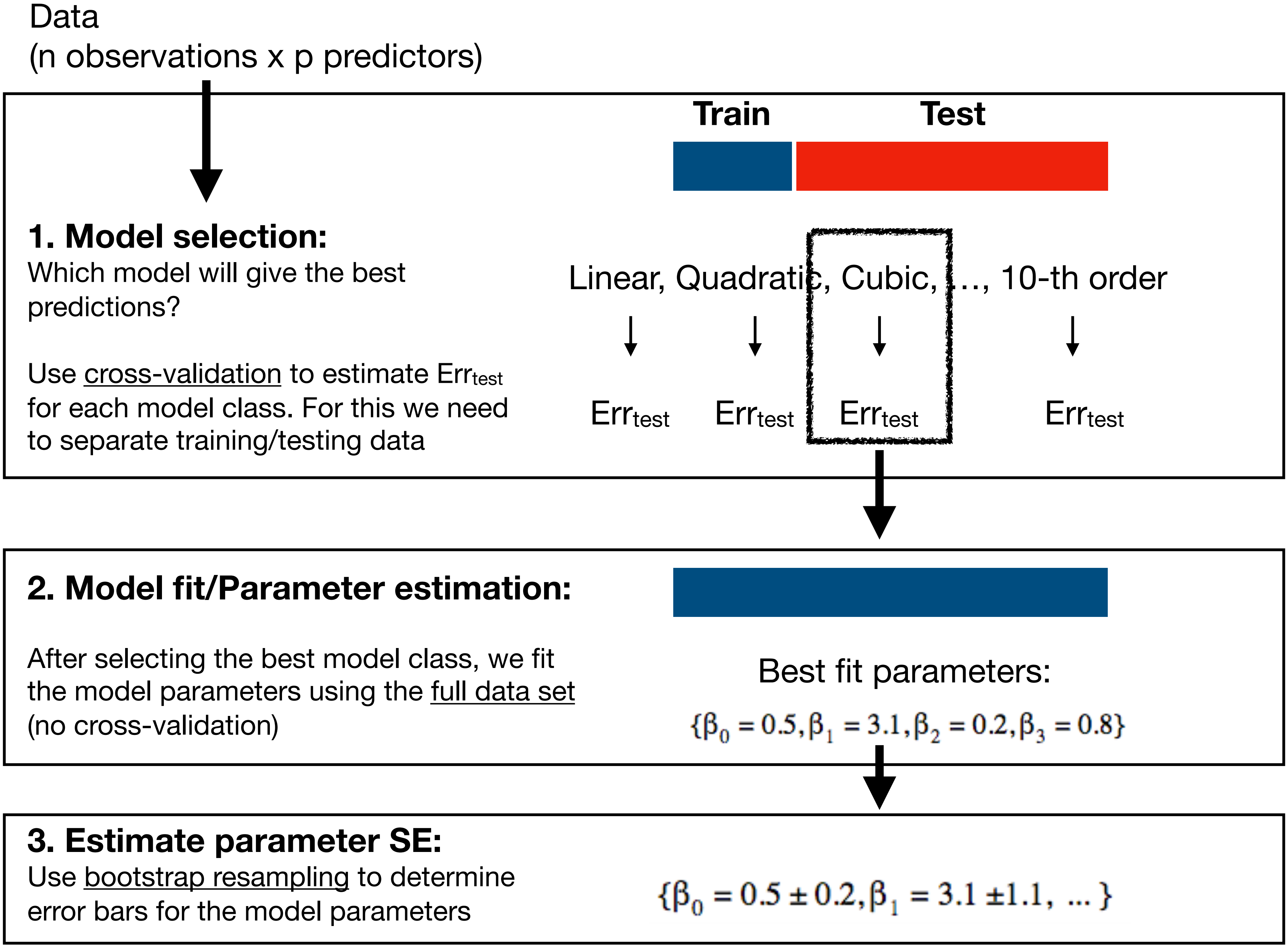
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ **Model 2: Quadratic**

- A particular “model fit” corresponds to a specific set of parameter values (and their corresponding SE, p-value etc.)

$$\{\beta_0 = 1, \beta_1 = 3\} : \; y = 1 + 3x + \varepsilon$$

$$\{\beta_0 = 0.3, \beta_1 = 5.1\} : \; y = 0.3 + 5.1x + \varepsilon$$

Predictive data modeling workflow



Data
(n observations x p predictors)

1. Model selection:
Which model will give the best predictions?

Use cross-validation to estimate $\mathrm{Err}_{\mathrm{test}}$ for each model class. For this we need to separate training/testing data

Train **Test**

Linear, Quadratic, Cubic, …, 10-th order

$\mathrm{Err}_{\mathrm{test}}$ $\mathrm{Err}_{\mathrm{test}}$ $\mathrm{Err}_{\mathrm{test}}$ $\mathrm{Err}_{\mathrm{test}}$

2. Model fit/Parameter estimation:

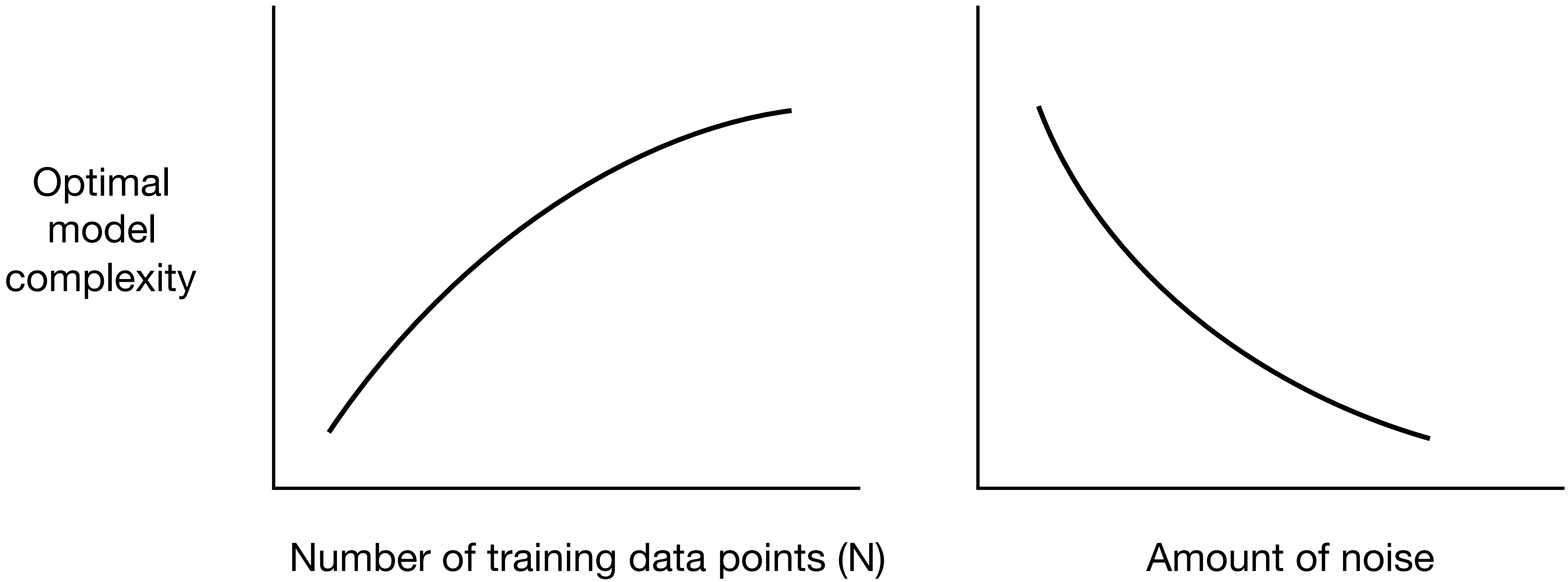
After selecting the best model class, we fit the model parameters using the full data set (no cross-validation)

Best fit parameters:

$$\{\beta_0 = 0.5, \beta_1 = 3.1, \beta_2 = 0.2, \beta_3 = 0.8\}$$

3. Estimate parameter SE:
Use bootstrap resampling to determine error bars for the model parameters

$$\{\beta_0 = 0.5 \pm 0.2, \beta_1 = 3.1 \pm 1.1, \ \ldots\}$$

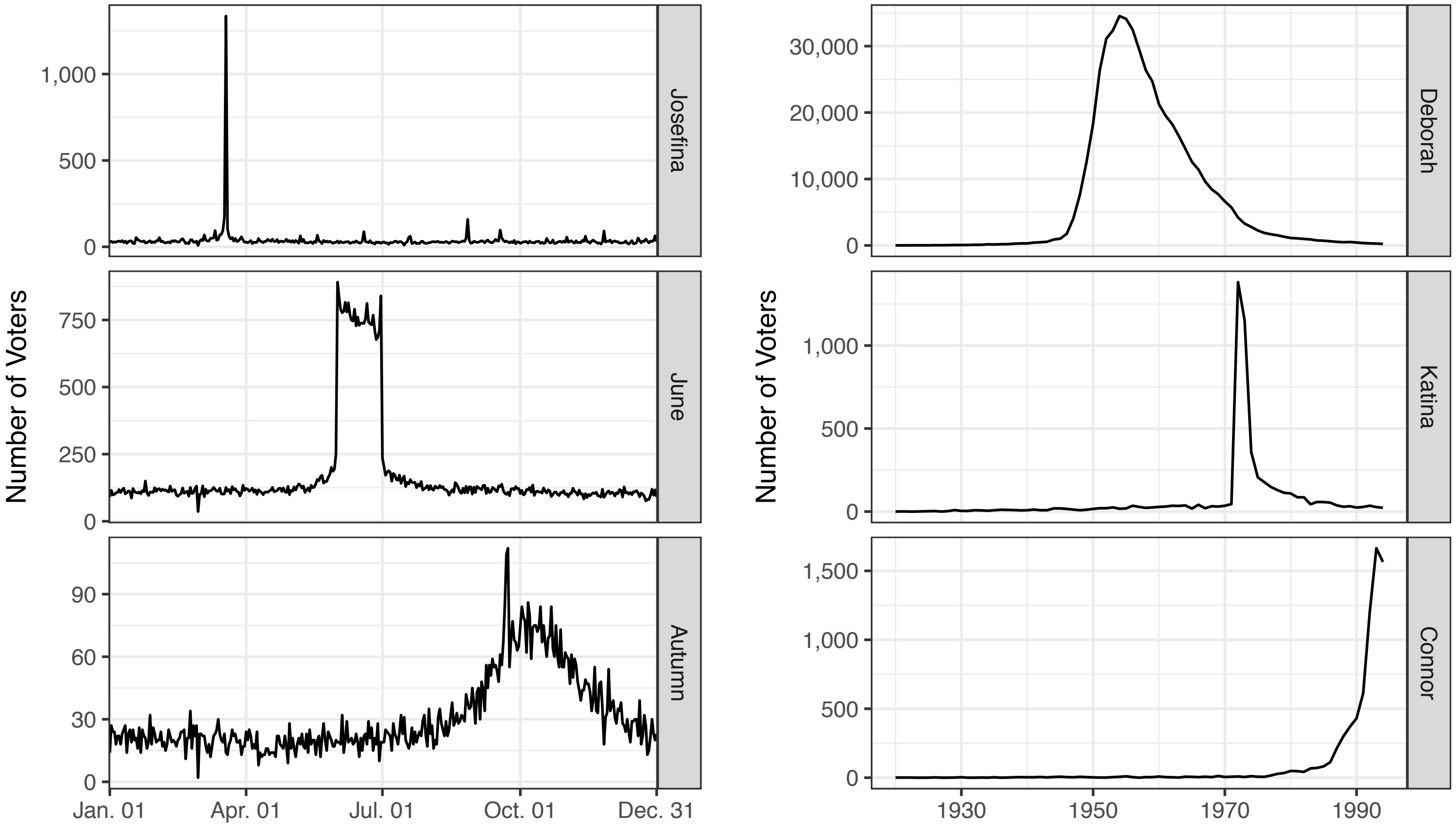

Optimal model complexity
Number of training data points (N)
Amount of noise

Cool data science application:

One Person, One Vote: Estimating the Prevalence of Double Voting in U.S. Presidential Elections*

Sharad Goel
Stanford University

Marc Meredith
University of Pennsylvania



(a)

(b)

Figure 3: Examples of names among 2012 voters with a non-uniform date of birth distribution, by day (a) or year (b) of birth.

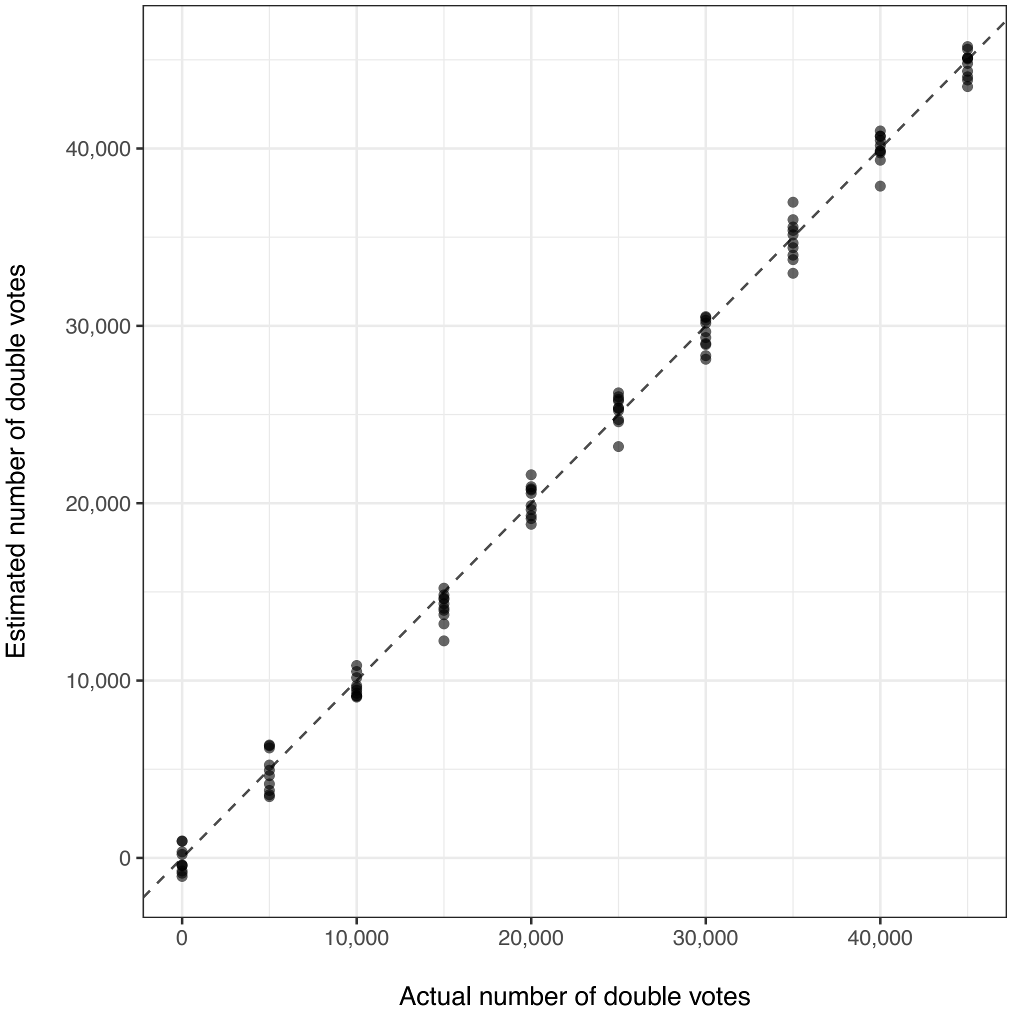
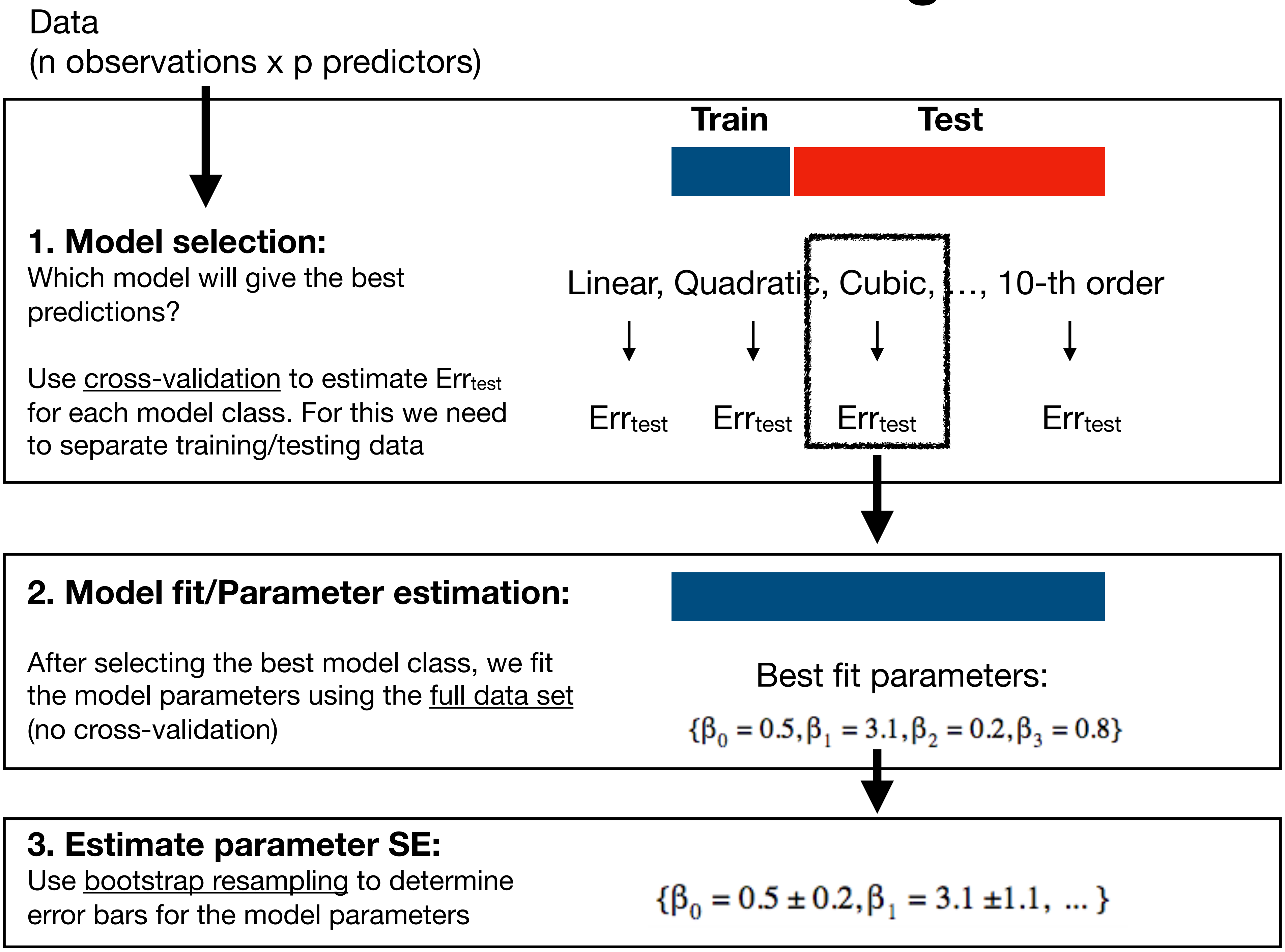


Figure A.2: Estimated number of duplicate records in a simulation compared to actual number of records duplicated.

Predictive data modeling workflow



Data
(n observations x p predictors)

Train

Test

1. Model selection:

Which model will give the best predictions?

Use cross-validation to estimate Err_{test} for each model class. For this we need to separate training/testing data

Linear, Quadratic, Cubic, …, 10-th order

Err_{test} Err_{test} Err_{test} Err_{test}

2. Model fit/Parameter estimation:

After selecting the best model class, we fit the model parameters using the full data set (no cross-validation)

Best fit parameters:

$$\{\beta_0 = 0.5, \beta_1 = 3.1, \beta_2 = 0.2, \beta_3 = 0.8\}$$

3. Estimate parameter SE:

Use bootstrap resampling to determine error bars for the model parameters

$$\{\beta_0 = 0.5 \pm 0.2, \beta_1 = 3.1 \pm 1.1, \; \ldots \}$$

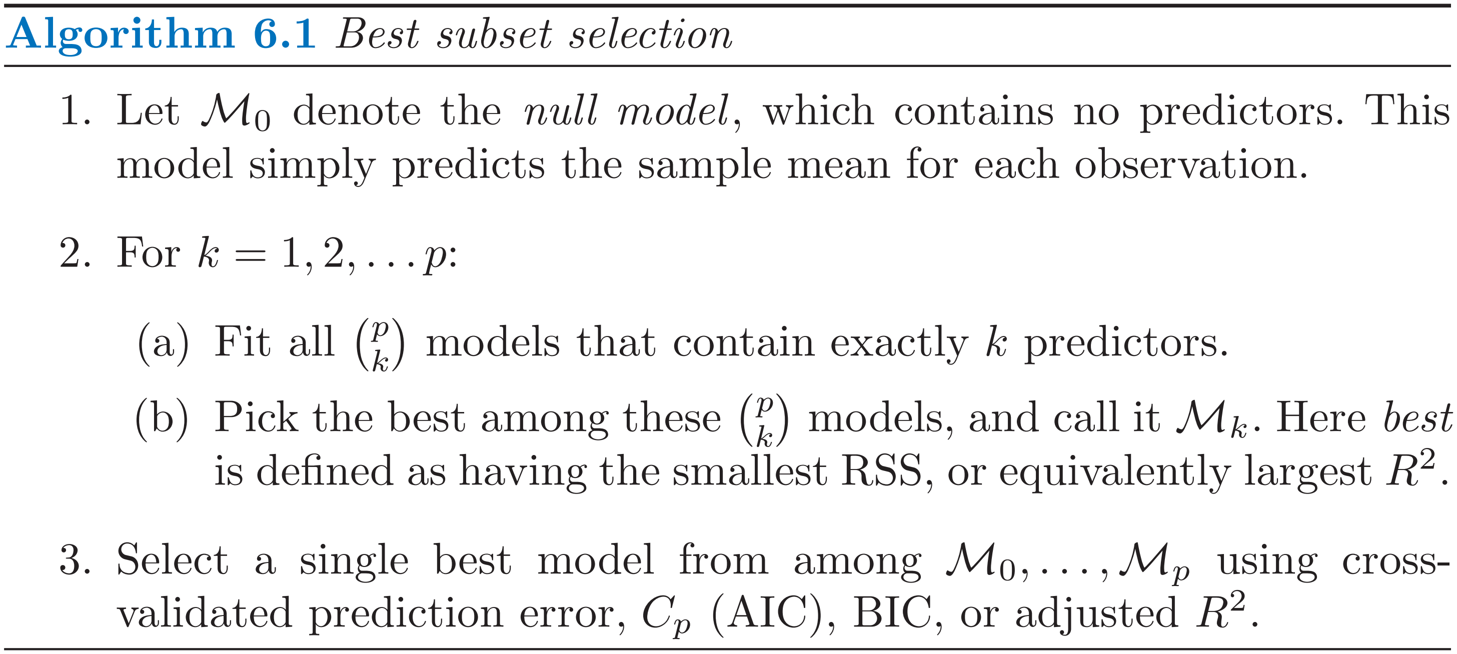
Model selection and regularization

- So far we have assumed that we can just try out several models and choose the best one (“brute force search”)

- In real data sets, there may be an *huge* number of possible models, making the brute force method impractical

- Need methods for more efficiently searching for good models

Method 1: Best subset



Algorithm 6.1 *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.

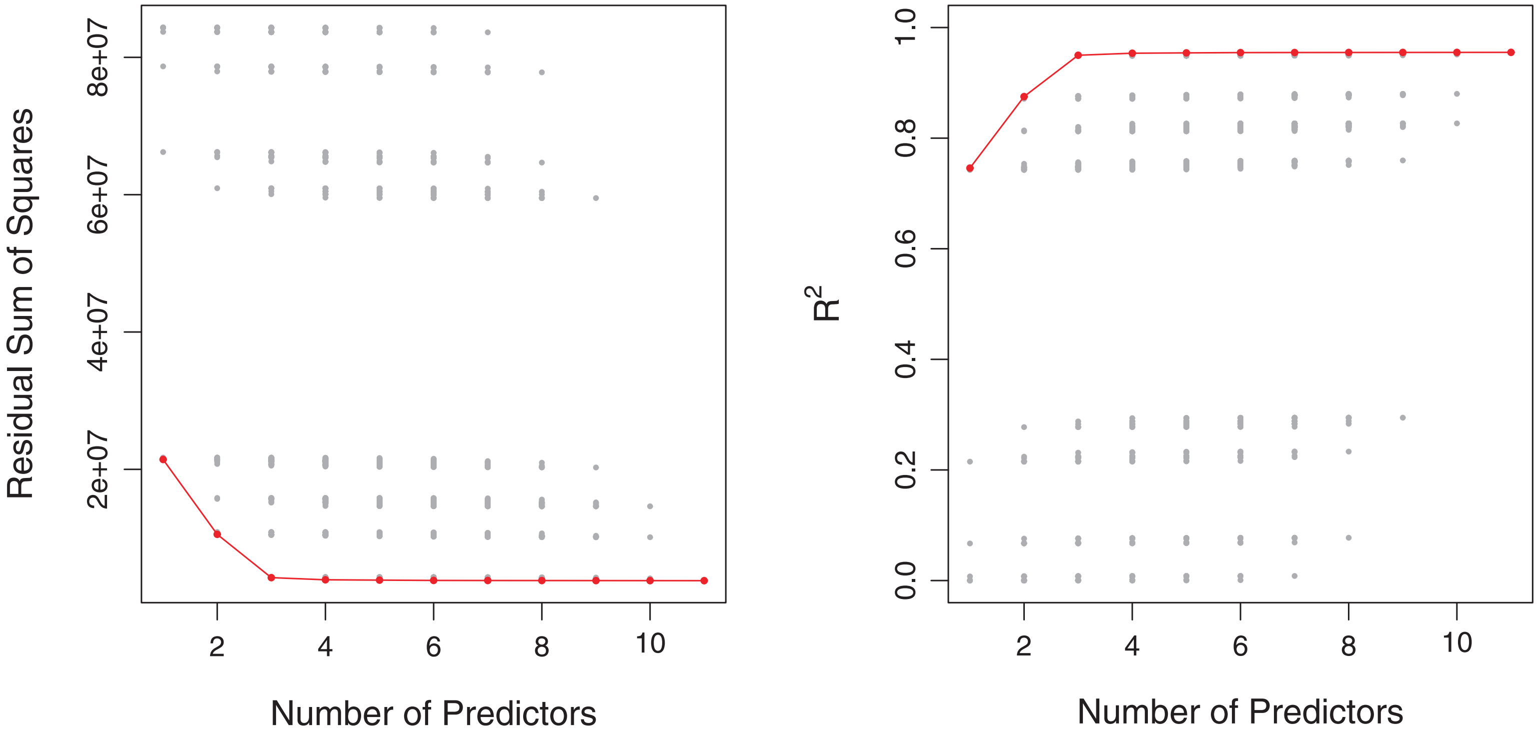
2. For $k = 1, 2, \ldots p$:

   (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.

- We will always use cross-validation in Step 3; you may ignore the other approaches (Cp, AIC, BIC, etc.)


Residual Sum of Squares
2e+07
4e+07
6e+07
8e+07
2
4
6
8
10
Number of Predictors
R^2
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
Number of Predictors

When to use cross-validation (MSEtest) and when to use training data (MSEtrain)

- In <u>step 2</u> of the selection, we are comparing many models with the same number of parameters:

  $k = 2:$

  - Model 1: $Blood\ pressure = \beta_0 + \beta_1 Sex + \beta_2 Age \quad \rightarrow \quad MSE_{train}(1)$
  - Model 2: $Blood\ pressure = \beta_0 + \beta_1 Sex + \beta_2 Exercise \quad \rightarrow \quad MSE_{train}(2)$
  - $\boxed{\text{Model 3: } Blood\ pressure = \beta_0 + \beta_1 Age + \beta_2 Diet \quad \rightarrow \quad MSE_{train}(3)}$

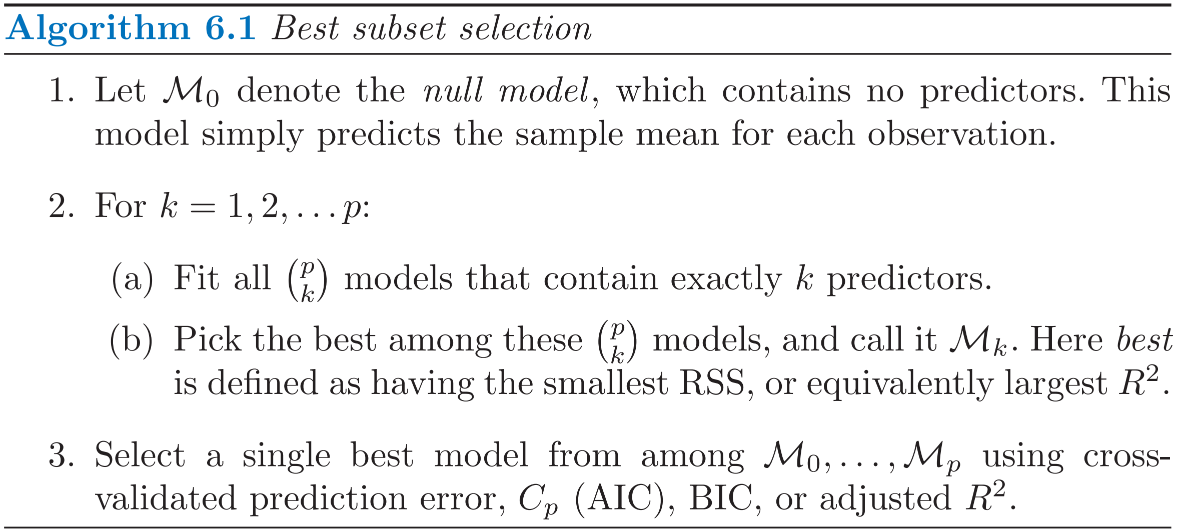
- These models are all equally flexible

- Although they may overfit the data (MSEtrain < MSEtest), they will all overfit by approximately the same amount. Thus we can use MSEtrain to compare them and select the best one (lowest MSEtrain)

- In <u>step 3</u> we are comparing models with different numbers of parameters. Therefore, cross-validation is essential so we can choose a model with low MSEtest.

- Recall that MSEtrain is always lower for more flexible models, so MSEtrain cannot be used to compare models with different levels of flexibility.

  - Model 1, $k = 1: \quad Blood\ pressure = \beta_0 + \beta_1 Sex \quad \rightarrow \quad MSE_{test}(1)$
  - Model 2, $k = 2: \quad Blood\ pressure = \beta_0 + \beta_1 Sex + \beta_2 Age \quad \rightarrow \quad MSE_{test}(2)$
  - $\boxed{\text{Model 3, } k = 3: \quad Blood\ pressure = \beta_0 + \beta_1 Sex + \beta_2 Age + \beta_3 Exercise \quad \rightarrow \quad MSE_{test}(3)}$



Algorithm 6.1 *Best subset selection*

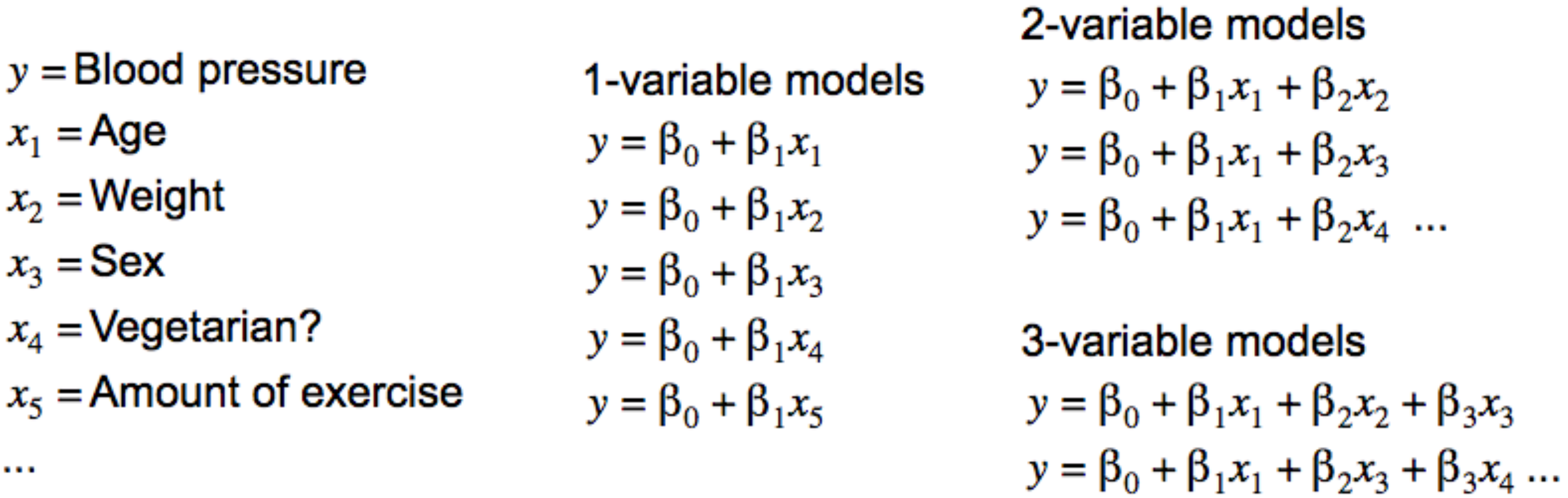
1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and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

No cross-validation; compare models using MSEtrain

Use cross-validation; compare models using MSEtest

The problem:
Too many models

- Complete model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$



$y =$ Blood pressure
$x_1 =$ Age
$x_2 =$ Weight
$x_3 =$ Sex
$x_4 =$ Vegetarian?
$x_5 =$ Amount of exercise
...

1-variable models

$y = \beta_0 + \beta_1 x_1$
$y = \beta_0 + \beta_1 x_2$
$y = \beta_0 + \beta_1 x_3$
$y = \beta_0 + \beta_1 x_4$
$y = \beta_0 + \beta_1 x_5$

2-variable models

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_4$ ...

3-variable models

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4$ ...

...

Number of possible subsets grows exponentially (2^P)

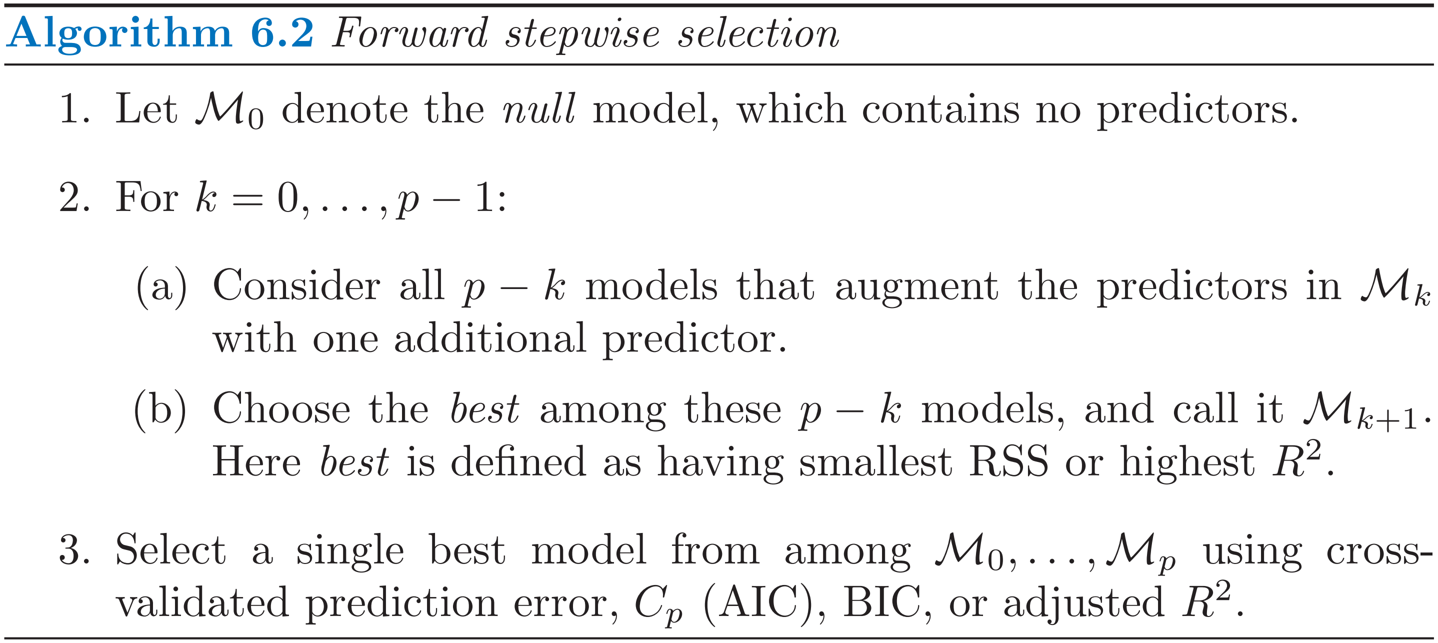
$$\sum_{k=0}^{P} \binom{P}{k} = 2^P$$

P=5, k=0, 1 models
P=5, k=1, 5 models
P=5, k=2, 10 models
P=5, k=3, 10 models
P=5, k=4, 5 models
P=5, k=5, 1 models
Total: 2^5 = 32 models

P=10, k=0, 1 models
P=10, k=1, 10 models
P=10, k=2, 45 models
P=10, k=3, 120 models
P=10, k=4, 210 models
P=10, k=5, 252 models
P=10, k=6, 210 models
P=10, k=7, 120 models
P=10, k=8, 45 models
P=10, k=9, 10 models
P=10, k=10, 1 models
Total: 2^10 = 1024 models

P=20, k=0, 1 models
P=20, k=1, 20 models
P=20, k=2, 190 models
P=20, k=3, 1140 models
P=20, k=4, 4845 models
P=20, k=5, 15504 models
P=20, k=6, 38760 models
P=20, k=7, 77520 models
P=20, k=8, 125970 models
P=20, k=9, 167960 models
P=20, k=10, 184756 models
…
Total: 2^20 = 1048576 models

Method 2: Stepwise 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 $R^2$.
```

$$k = 1: \quad Blood\ pressure = \beta_0 + \beta_1 Sex$$

$$k = 2: \quad Blood\ pressure = \beta_0 + \beta_1 Sex + \beta_2 Age$$

$$k = 3: \quad Blood\ pressure = \beta_0 + \beta_1 Sex + \beta_2 Age + \beta_3 Exercise$$

Stepwise selection is much more efficient than best subset selection

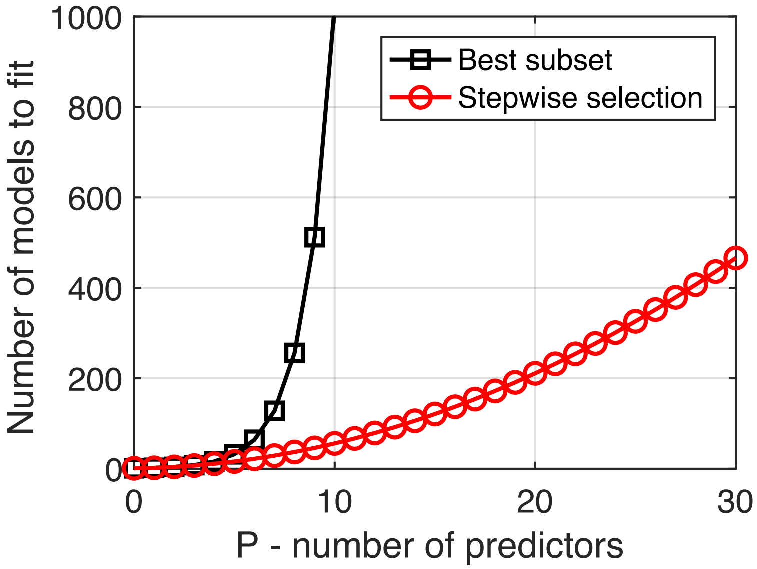
- Instead of exhaustively trying all 2^P possible subsets of parameters, at each stage we just try P subsets.

- The total number of models we end up fitting is:

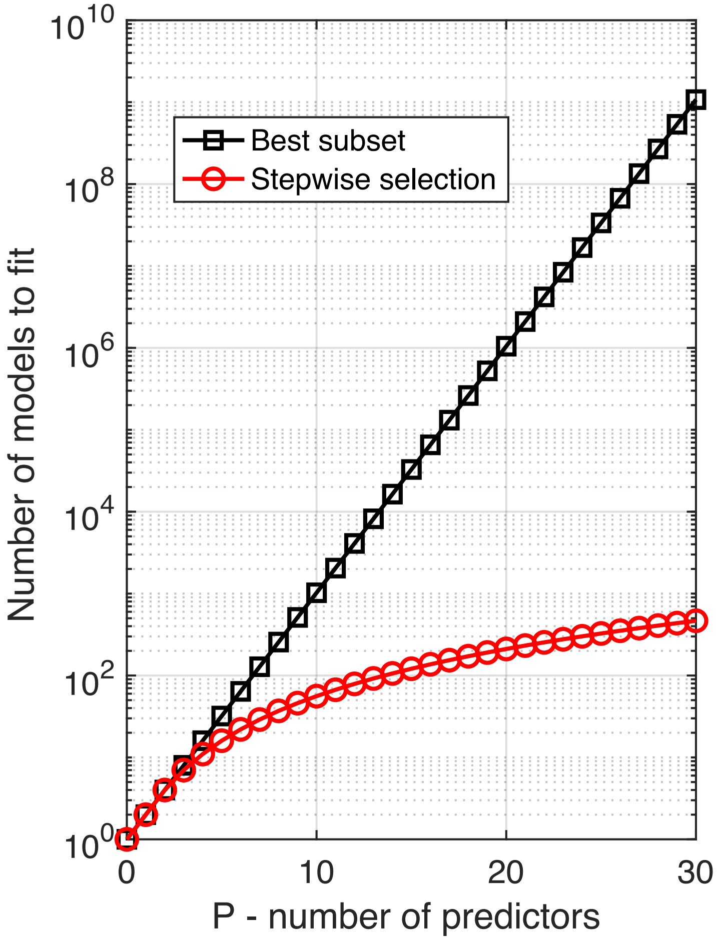
$$1 + \sum_{k=0}^{p-1}(p-k) = 1+p(p+1)/2$$

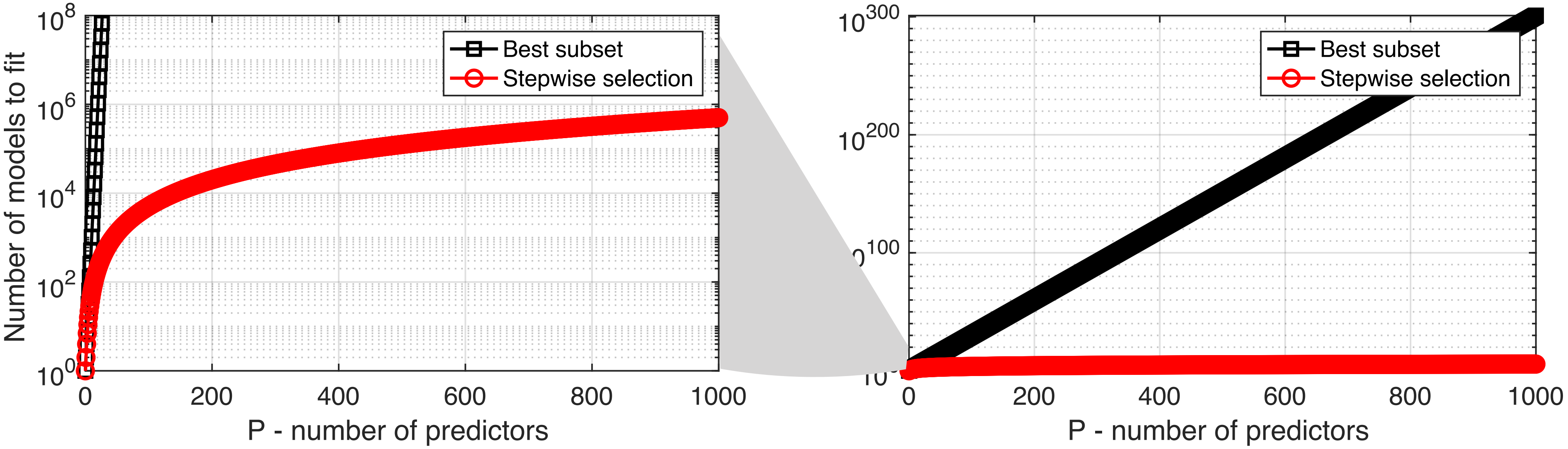
P=20, k=0, 20 models
P=20, k=1, 19 models
P=20, k=2, 18 models
P=20, k=3, 17 models
P=20, k=4, 16 models
P=20, k=5, 15 models
P=20, k=6, 14 models
P=20, k=7, 13 models
P=20, k=8, 12 models
P=20, k=9, 11 models
….

Total: 1+P(P+1)/2 = 211 models

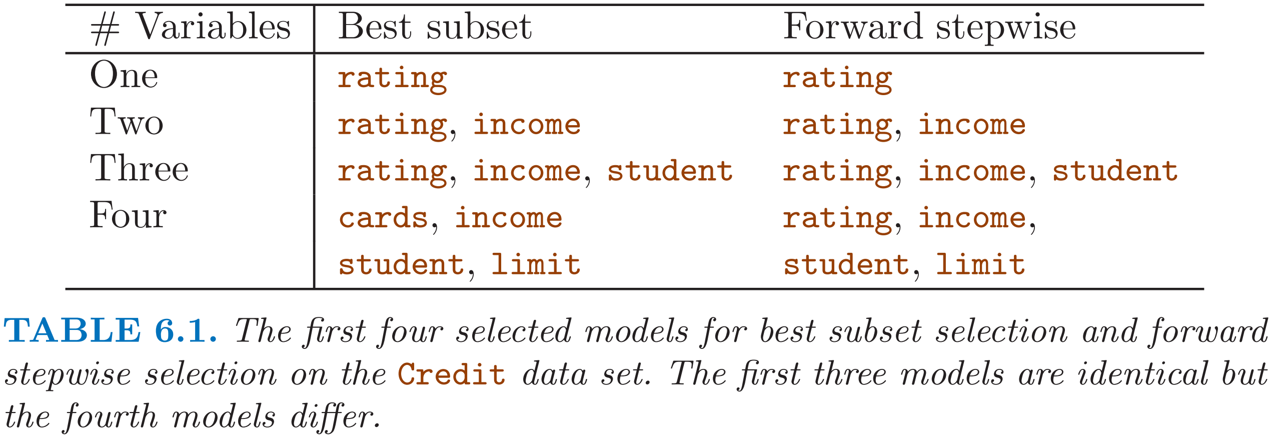


Same data on a logarithmic scale for the y-axis




Number of models to fit
10^{8}
10^{6}
10^{4}
10^{2}
10^{0}
0
200
400
600
800
1000
P - number of predictors
Best subset
Stepwise selection
10^{300}
10^{200}
10^{100}
10^{0}
0
200
400
600
800
1000
P - number of predictors
Best subset
Stepwise selection

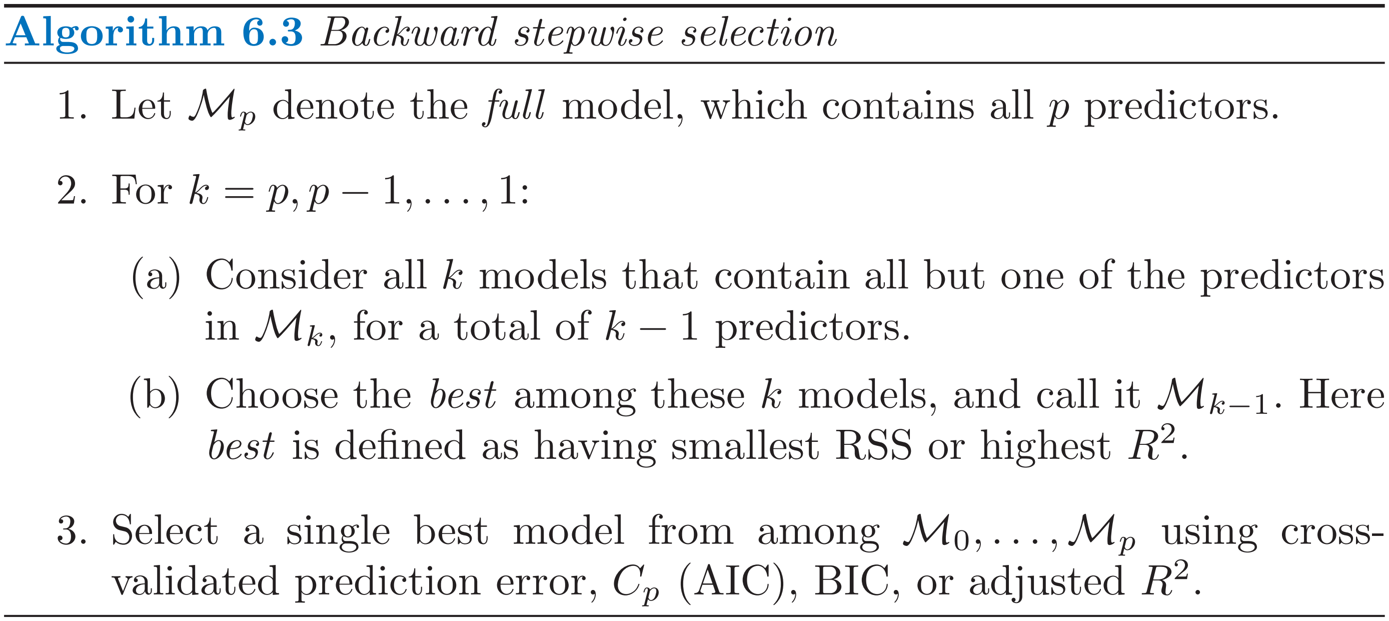
Stepwise selection may not find the absolute best model



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