Linear Model Selection

Pham Mai Tam
15th June, 2022

Outline

1. Best subset selection

2. Stepwise selection

   - Forward stepwise selection
   - Backward stepwise selection

3. Choosing the Optimal Model

   - by using C_p, AIC, BIC, adjusted R^2
   - by using Validation Sets and Cross-Validation

Method

Alternatives to least squares

● **Prediction Accuracy:**

☞ n >> p: can use least squares to fit model

☞ n > p: variability in the least squares fit → over-fitting and consequently poor predictions on test set.

☞ especially when p > n: variance is infinite

→ using constrain or shrinkage methods 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 automatically performing feature selection.

Linear Model Selection

In multiple regression model building, there are 3 basic strategies:

1. Best subsets selection (all possible combinations)
2. Forward stepwise selection (addition)
3. Backward stepwise selection (deletion)

Example

A theme park analyst will use historical data to develop regression models for a “Guess Your Weight” game designed for n=236 children and adolescents. The Full Model is:

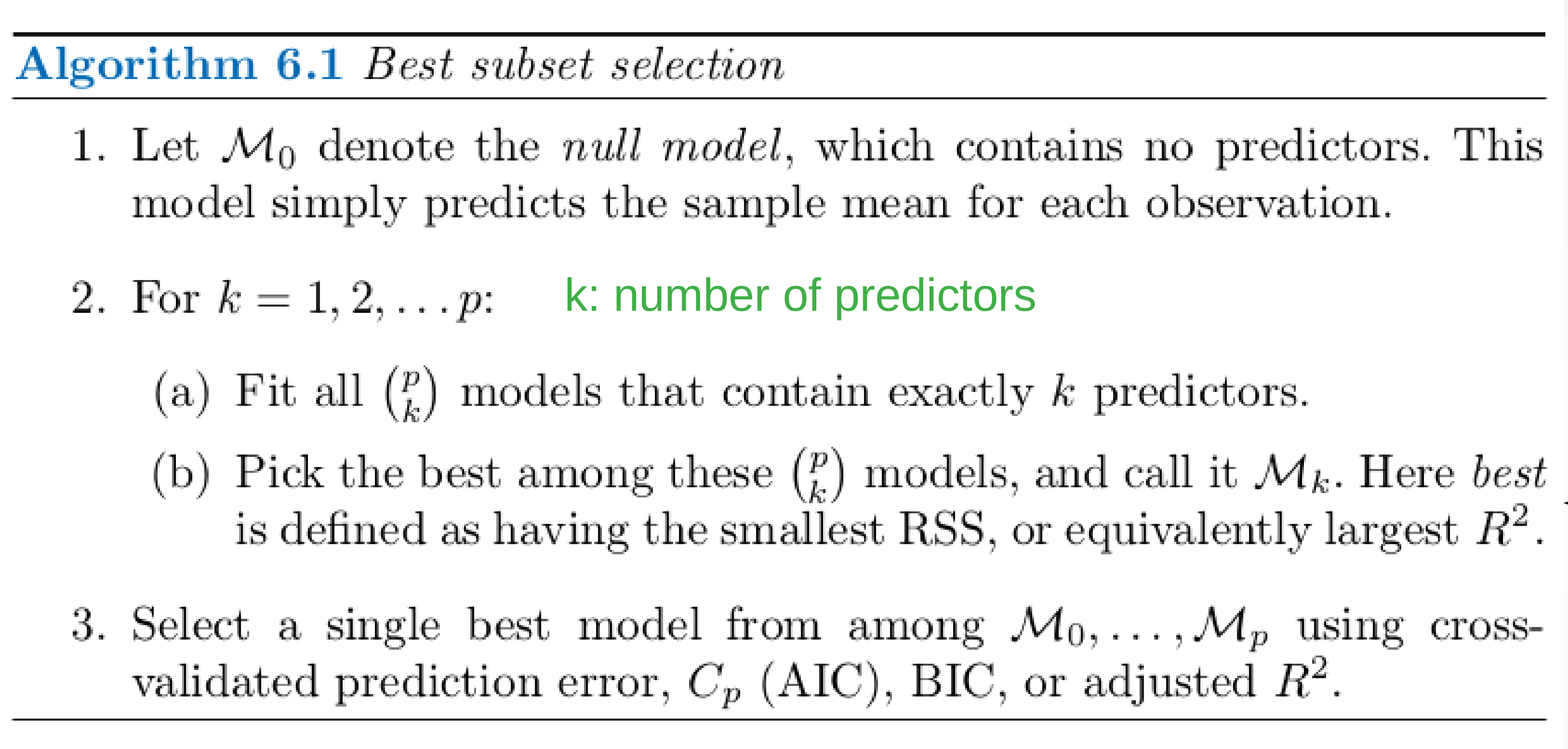
WEIGHT = SEX + AGE + HEIGHT

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}$$

Weight (y)
Sex (x_1)
Age (x_2)
Height (x_3)

Subset selection

Best subset model selection procedure



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 $R^2$.
```

k: number of predictors

M_o just contain intercept (β_o)

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

With p=3:

M_o: no predictors

M1: best models among C^1_3=3 models containing 1 predictors

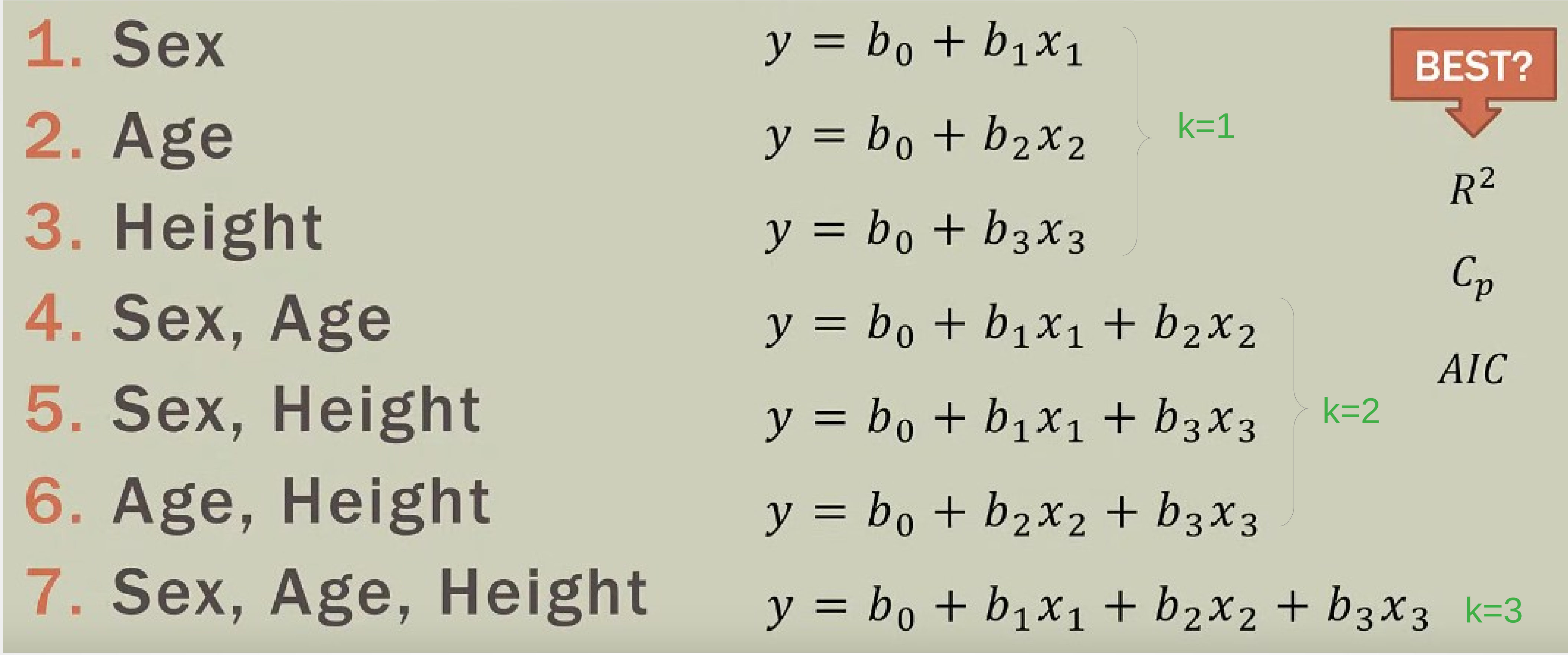
M2: best model among C^2_3=3 models containing 2 predictors

M3: best model among C^3_3=1 model containing 3 predictors

2^p (2^3=8) possible models containing subsets of p predictors

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^{n-k} b^k$$

BEST SUBSET SELECTION



1. Sex
2. Age
3. Height
4. Sex, Age
5. Sex, Height
6. Age, Height
7. Sex, Age, Height

$$y = b_0 + b_1 x_1$$
$$y = b_0 + b_2 x_2$$
$$y = b_0 + b_3 x_3$$

k=1

$$y = b_0 + b_1 x_1 + b_2 x_2$$
$$y = b_0 + b_1 x_1 + b_3 x_3$$
$$y = b_0 + b_2 x_2 + b_3 x_3$$

k=2

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

k=3

R^2

C_p

AIC

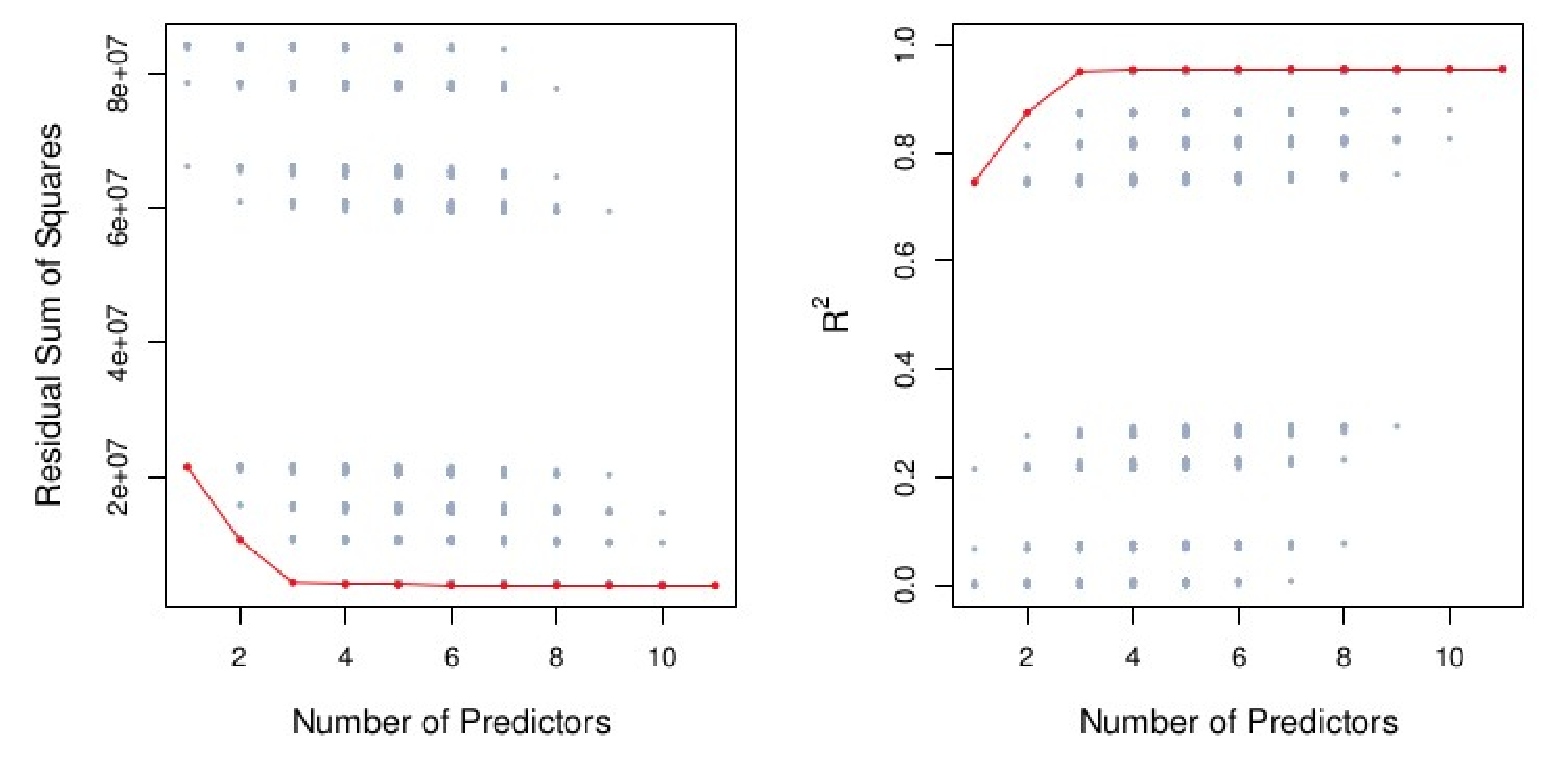
7 POSSIBLE MODELS

1. Sex
2. Age
3. Height
4. Sex, Age
5. Sex, Height
6. Age, Height
7. Sex, Age, Height

Model 7 is called the full model. Models 1-6 are called reduced models (or nested models)

Which model makes the BEST predictions while also being the simplest?
What do we mean by BEST?
How do we measure which is BEST?

Best subset selection



The red frontier tracks the best model for a given number of predictors, according to RSS and R^2

STEPWISE SELECTION

○ For *computational reasons*, best subset selection cannot be applied with very large p. *Why not? (p=40, 2^{40} ~ 1,000,000,000,000 possible models)*

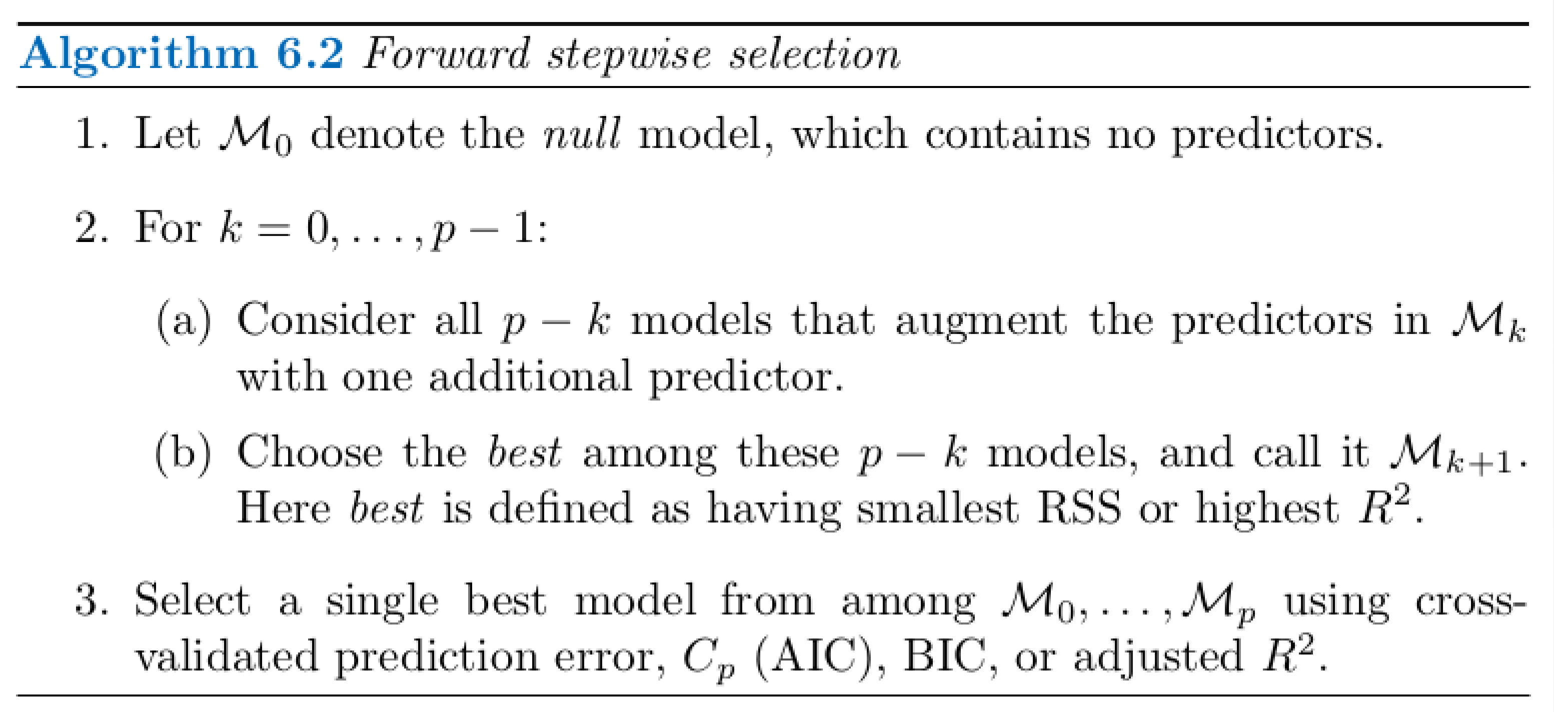
○ Best subset selection may also suffer from *statistical problems* when p is large: larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.

→ an enormous search space can lead to *over-fitting* and high variance of the coefficient estimates.

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

e.g p=3 → $2^3 = 8$ (7 possible models + M_0 just contain intercept)

Forward Stepwise Selection



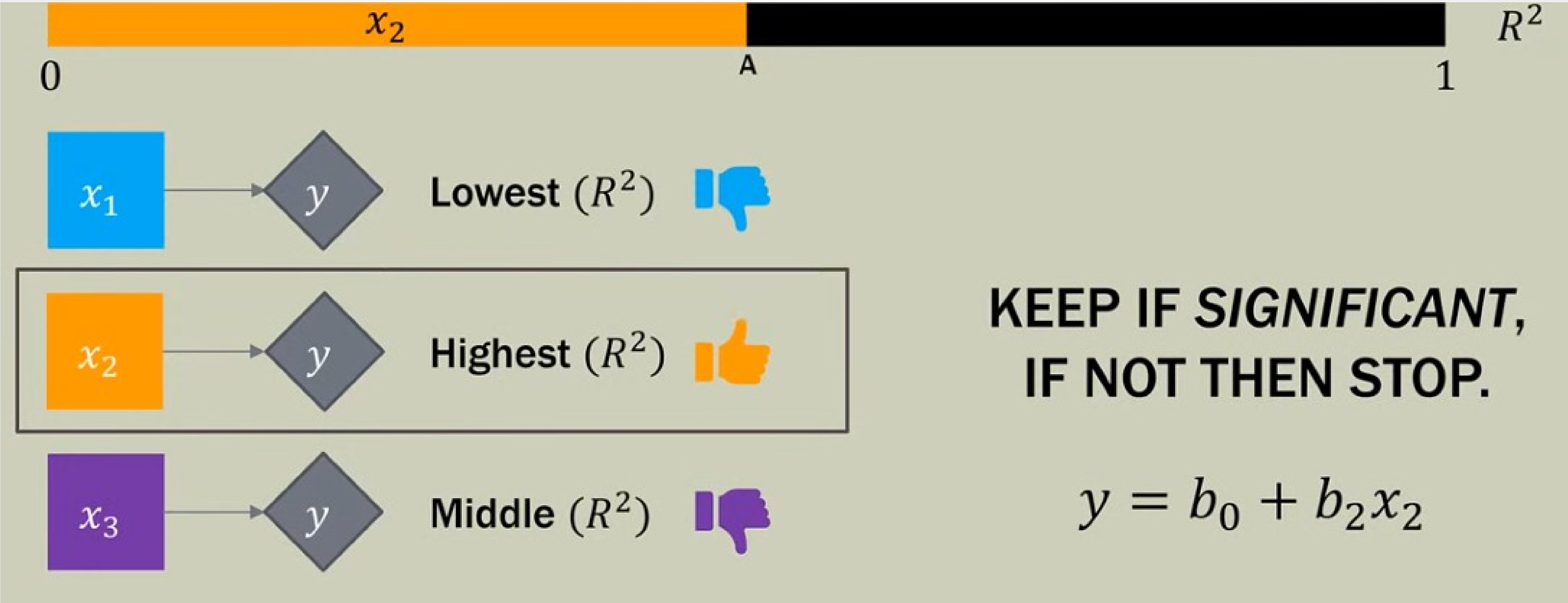
Algorithm 6.2 *Forward stepwise selection*

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.

(a) Forward selection begins with a model containing no predictors, then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.

→ *Computational advantage based on a reduced number of possible models*

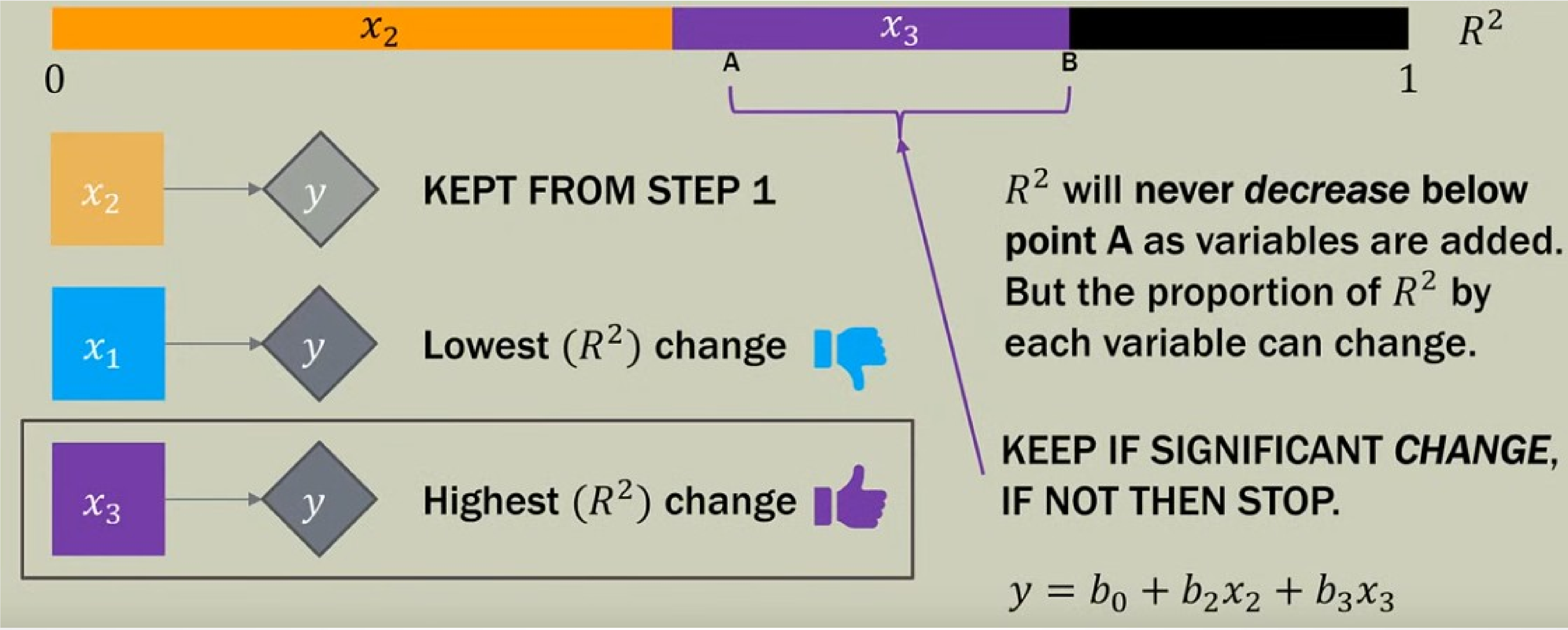
FORWARD SELECTION



KEEP IF *SIGNIFICANT*, IF NOT THEN STOP.

$$y = b_0 + b_2 x_2$$

FORWARD SELECTION



KEPT FROM STEP 1

Lowest (R^2) change 👎

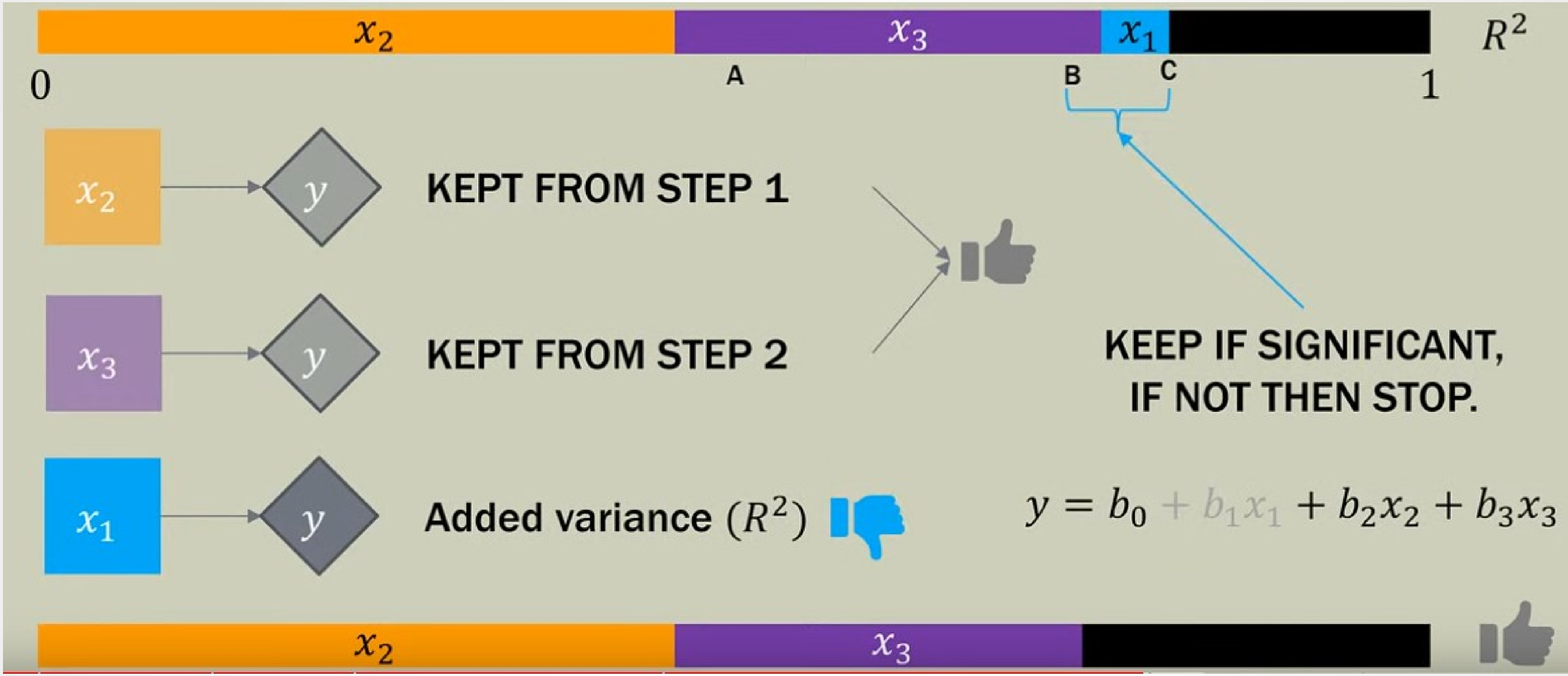
Highest (R^2) change 👍

R^2 **will never *decrease* below point A** as variables are added. But the proportion of R^2 by each variable can change.

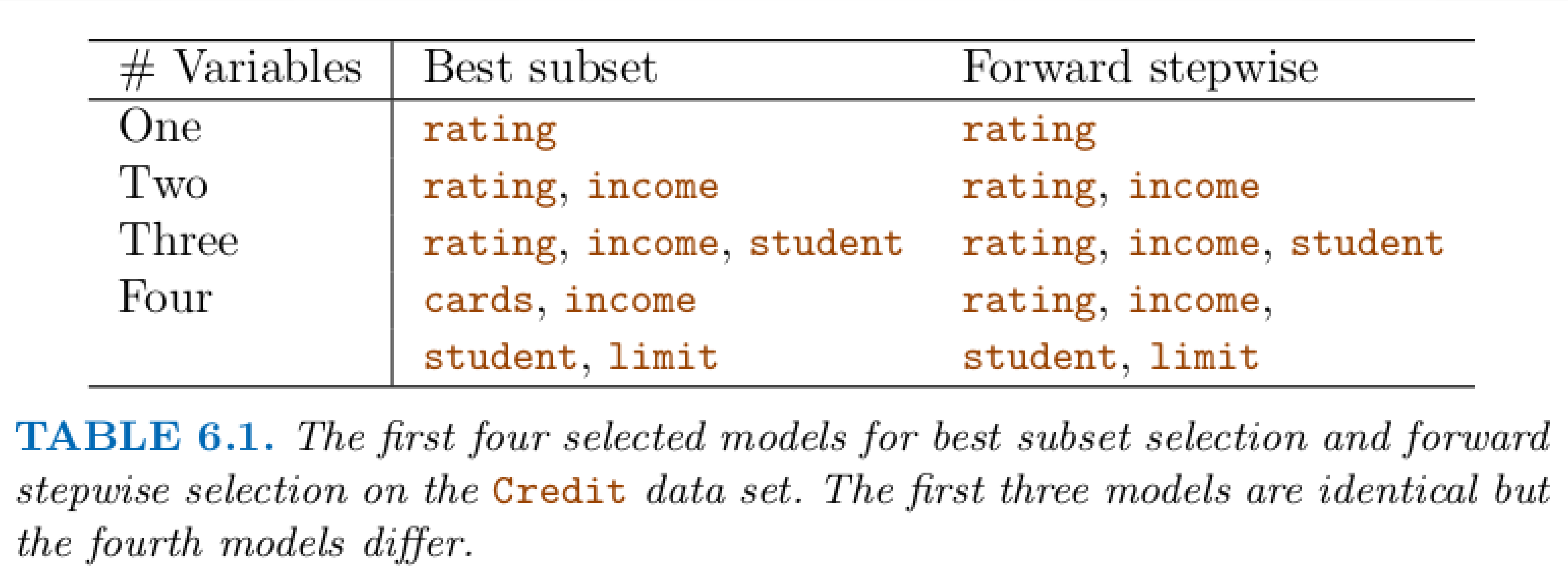
KEEP IF SIGNIFICANT *CHANGE*, IF NOT THEN STOP.

$$y = b_0 + b_2x_2 + b_3x_3$$

FORWARD SELECTION



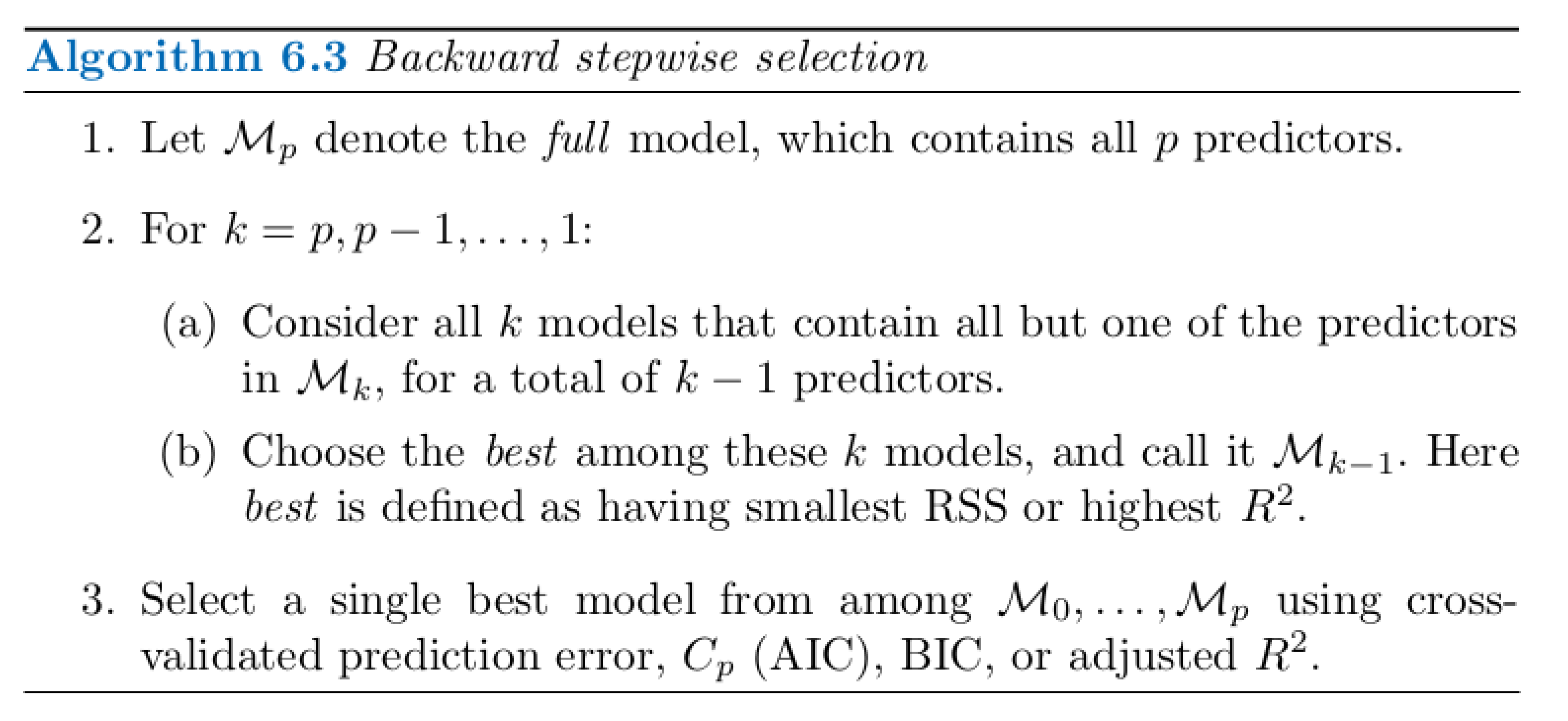
Different results b/w subset and stepwise selection



# 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

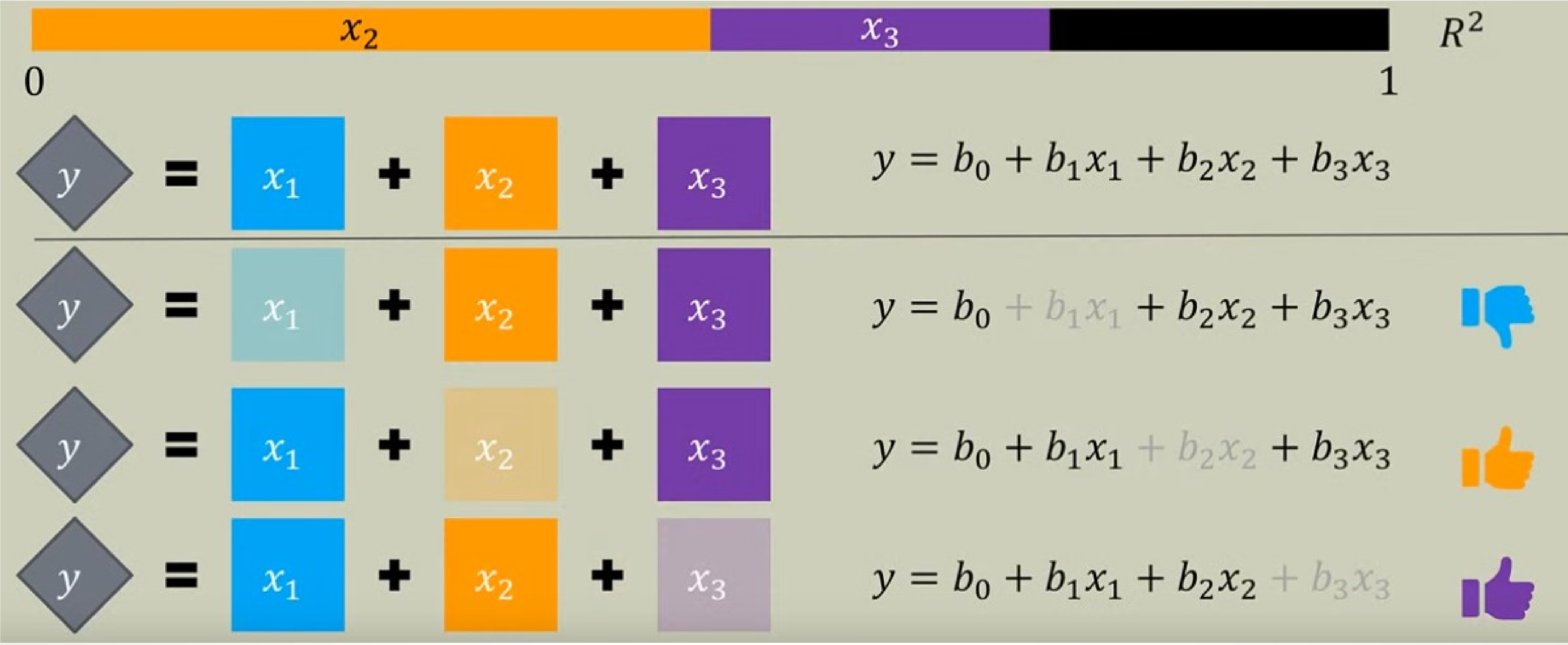


Algorithm 6.3 *Backward stepwise selection*

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.

(a) Unlike forward selection, backward selection 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


x_2
x_3
R^2
0
1
y = x_1 + x_2 + x_3
y = x_1 + x_2 + x_3
y = x_1 + x_2 + x_3
y = x_1 + x_2 + x_3

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

If the change is not significant, we leave it out. → Take out x_1

More on Stepwise Selection

○ Like forward stepwise selection, the backward selection approach searches through only 1 + p(p+1)/2 models (e.g. 1 + 40x(40+1)/2=821 <<< 2^{40} models in best subset selection) → advantage in computational resources

○ Like forward stepwise selection, backward 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.

GENERAL RULES AND CONCEPTS

- When adding variables to a model, R2 will never decrease
- At >1 variable, variables potentially start influencing each other
- Different techniques may result in different models
- R^2 alone is usually not sufficient to determine best model
- Aim for a model that is simple, but fits best
- Including more variables risks overfitting the model

Choosing the Optimal 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.

○ Training errors decreases when more variables are added, but testing errors do not

→ Training set RSS and training set R^2 cannot be used to select best models from different numbers of variables.

→ Instead, we have 2 approaches to estimate test error

Estimating test error: 2 approaches

○ We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to over-fitting (using C_p, BIC, AIC, adjusted R^2).

○ We can directly estimate the test error, using either a validation-set approach or a cross-validation approach.

→ We illustrate both approaches next

C_p, AIC, BIC, adjusted R^2

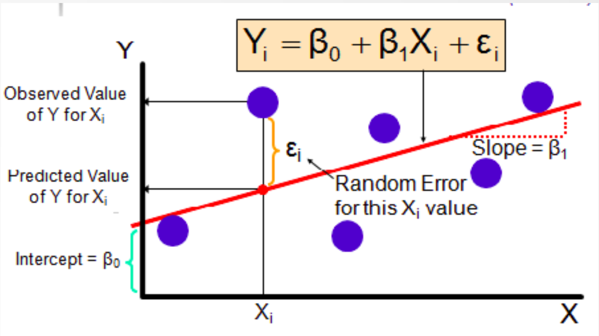
○ These techniques adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables.

○ The next figure displays how C_p, BIC, and adjusted R^2 work to select best model of each size produced by best subset selection.

C_p, AIC, BIC, adjusted R^2

○ Mallow's C_p:

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


Y
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Observed Value of Y for X_i
Predicted Value of Y for X_i
ε_i
Slope = β_1
Random Error for this X_i value
Intercept = β_0
X_i
X

Where:
d: the total # of parameters used (# of variables + intercept);
$\hat{\sigma}^2$:an estimate of the variance of the error ϵ from each response measurement;
estimates $\hat{\sigma}^2$ using full model → be constant → d increases, penalty increases
n: is # of observation.

○ AIC: defined for a large class of models fit by maximum likelihood:

$$\mathrm{AIC} = -2\log L + 2\cdot d \xrightarrow{\text{For linear model}} \mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right),$$

where L is the maximized value of the likelihood function for the estimated model.
• In the case of the linear model with Gaussian errors, maximum likelihood and least squares are proportional, and C_p and AIC are proportional (In fact, minimize cost means maximum likelihood)

RSS

Likelihood= $\sigma^2 2\pi^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i-\theta^T x_i)^2}$

→ -2logL ~ 1/ $\hat{\sigma}^2$ x RSS

How to prove? → see slide of regularization

Details on BIC

Like C_p, the 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.

- Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$ 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. See Figure on slide 19.

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

Details on adjusted R^2

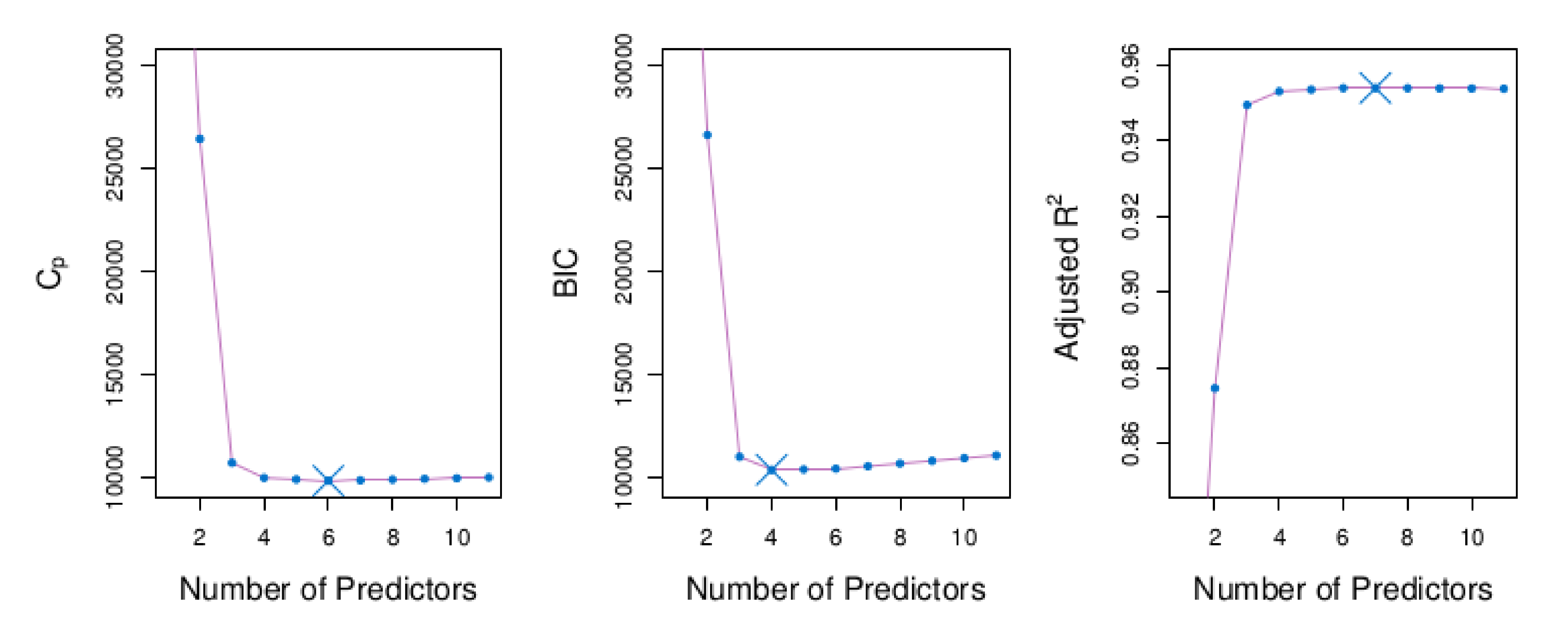
For a least squares model with d variables, the adjusted R^2 statistic is calculated as

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

where TSS is the total sum of squares.

- Unlike C_p, AIC, and BIC, for which a small value indicates a model with a low test error, a large value of adjusted R^2 indicates a model with a small test error.
- Maximizing the adjusted R^2 is equivalent to minimizing RSS/(n−d−1). While RSS always decreases as the number of variables in the model increases, RSS/(n−d−1) may increase or decrease, due to the presence of d in the denominator.
- Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model. See Figure on slide 19

C_p, BIC, adjusted R^2


C_p
Number of Predictors
BIC
Number of Predictors
Adjusted R^2
Number of Predictors