ESC

×

# Variable Selection

이규민 고정민

## ESC × CONTENTS

01. Review

Why regression analysis?

02. Linear Model Selection

Best Subset Selection / Stepwise Selection

03. Forward Stagewise Selection

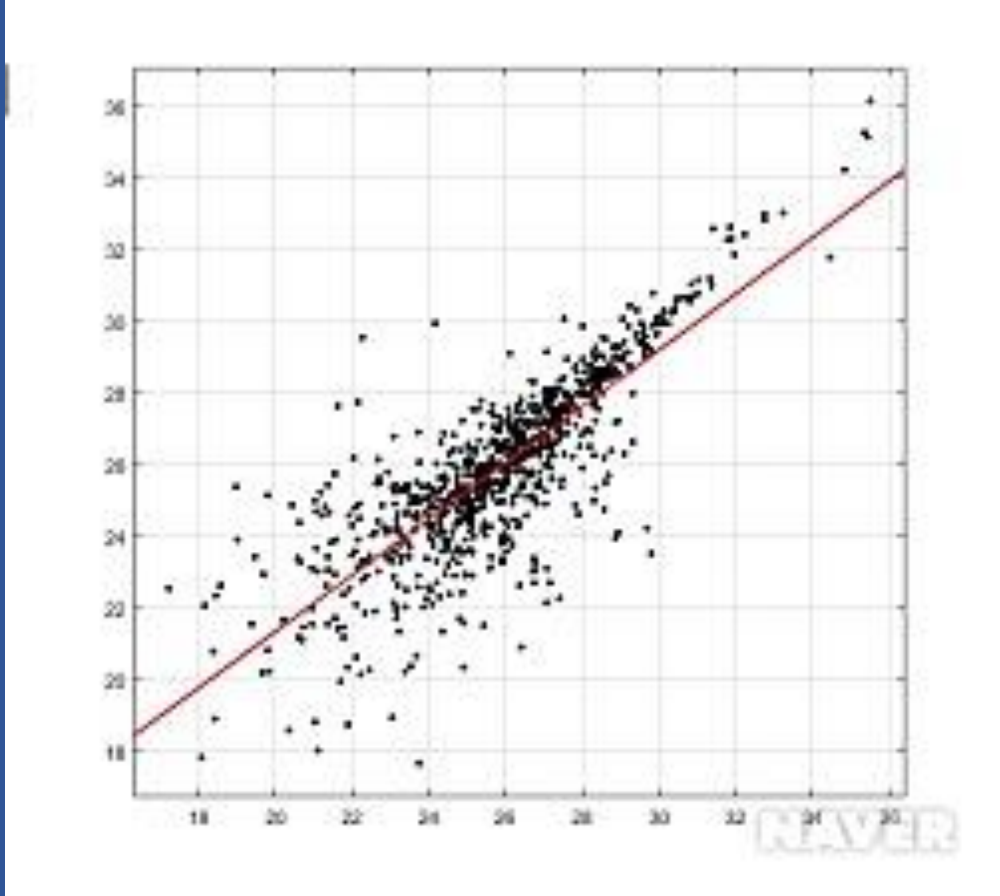
04. Criteria for Model selection

Loss, error, and various criteria

ESC

×

# 1. Review



출처 : https://terms.naver.com/entry.naver?docId=5141772&cid=60266&categoryId=60266

# Regression Analysis

## What is regression analysis?

변수들 사이의 함수적인 관련성을 규명하기 위해 어떤 수학적 모형을 가정하고 이 모형을 측정된 변수들의 자료로부터 추정하는 통계적 분석 방법

## What is regression model?

어떤 관계가 있을지에 대한 여러 가지 가설들을 회귀 '모형'이라 부른다. 100%까지는 아니라고 하더라도 간단한 가설을 통해 현실의 많은 부분을 설명해줄 수 있기 때문이다.

# Good Model

## What is good model?

$$Y = f\left(X_1, X_2, \ldots, X_p\right) + \epsilon$$

→ 회귀분석 변수들 사이의 관계를 정확하게 기술하거나 예측을 하려면 잔차는 최소가 되어야 한다.

## Guass-Markov Theorem

1. $X$의 각 값들에 대해 반응 변수 $Y$는 모수($\beta$)들의 선형결합이다. (linear in parameters)
2. 오차항 $\epsilon$는 주어진 모든 독립변수들의 값에서 0의 기댓값을 갖는다. (zero conditional mean)

   → 주어진 $X$에서 $Y$의 기댓값이 모수들에 대하여 선형이다.
3. 모형에 포함된 예측변수들이 어떤 값을 갖더라도 오차항 $\epsilon$의 조건부 분산은 일정하다. (homoskedasticity)

   → 주어진 $X$에서 $Y$의 분산이 독립변수들의 값에 의존하지 않는다.

MVUE (Minimum Variance Unbiased Estimator)

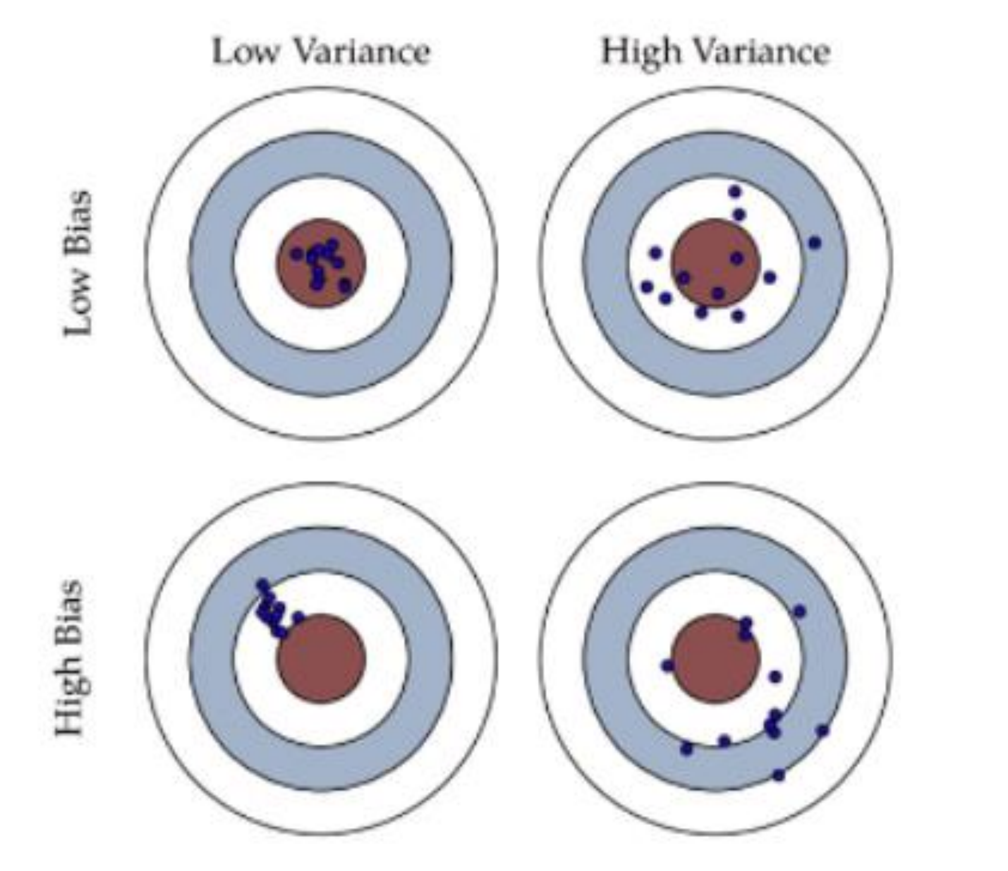
# Bias vs Variance

Good Explanatory Model 현재의 데이터를 잘 설명하는 모델

$$MSE_{(trainig)} = (Y - \hat{Y})^2$$

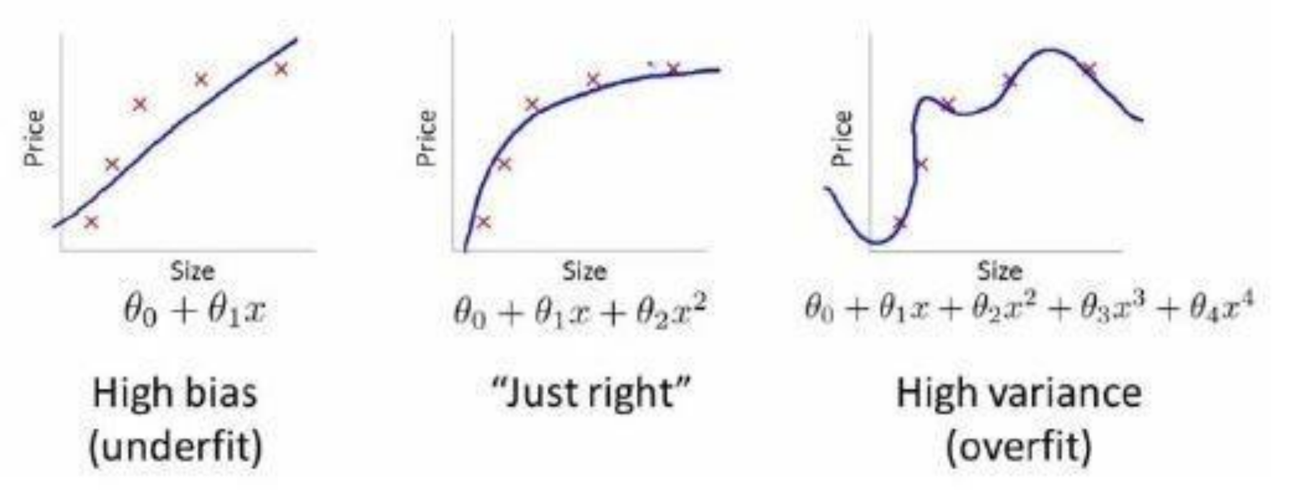
Good Predictive Model 미래의 데이터에 대한 예측 성능이 좋은 모델

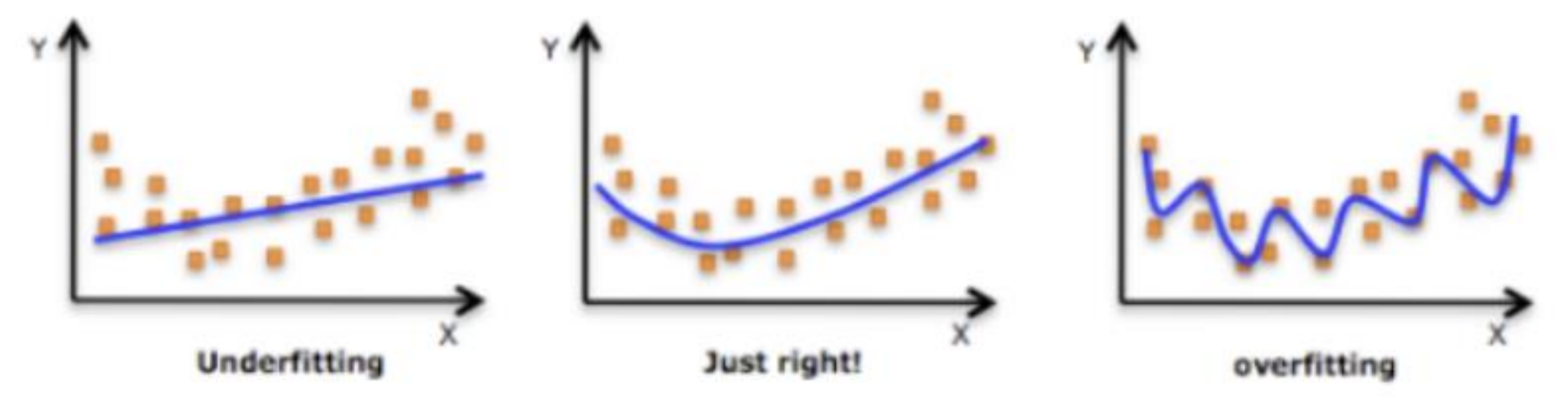
$$\begin{aligned}\text{Expected MSE} \quad &= E\left[(Y - \hat{Y})^2 | X\right] \\ &= \sigma^2 + (E[\hat{Y}] - \hat{Y})^2 + E[\hat{Y} - \mathrm{E}[\hat{Y}]]^2 \\ &= \sigma^2 + Bias^2(\hat{Y}) + \mathrm{Var}(\hat{Y}) \\ &= \text{Irreducible Error} + Bias^2 + \text{Variance}\end{aligned}$$



Bias가 증가되더라도 variance 감소폭이 더 크다면 expected MSE는 감소(예측 성능 증가)

# Overfitting vs Underfitting





# Overfitting vs Underfitting

1. Overfitting 과대적합

   샘플데이터에 너무 정확히 학습되어 샘플데이터에는 정확도가 높지만, 다른 데이터에서는 정확도가 떨어지는 현상

   앞으로 발생하기 어려운 ‘가짜 패턴’을 포착해 정확한 예측을 할 수 없게 된다.

2. Underfit 과소적합

   샘플데이터가 모자라거나 제대로 학습이 되지 않아, 학습데이터에 대해서 원하는 결과를 도출 못하는 현상

   패턴을 포착하지 못해서 정확한 예측을 할 수 없게 된다.

# Why Consider Variable Selection?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}$$

## Prediction Accuracy (Preventing Overfitting)

반응변수와 예측변수들 사이의 실제 관계가 근사적으로 선형일일 때 최소 제곱 추정치들은 편향이 적을 것이다. $n \gg p$이면 즉, n이 p보다 훨씬 크다면 최소 제곱 추정치는 적은 분산을 갖게 될 것이고 test 관측치에 대해서도 잘 작동할 것이다. 그러나 n이 p보다 훨씬 크지 않다면 최소 제곱 적합에 많은 변동이 존재할 수 있어 결과적으로 과적합과 모형을 훈련하는데 이용하지 않은 미래의 관측치에 대해 좋지 않은 예측을 하게 될 것이다. 그리고 $p>n$이면 더 이상 유일한 최소 제곱 계수 추정치가 없다.

## Model Interpretability

다중 회귀모형에서 사용되는 몇몇 또는 많은 변수들이 사실 반응변수와 연관되지 않은 경우가 다반사이다. 그런 관련 없는 변수들을 포함하는 것은 결과적인 모형을 복잡하게 만든다. 이러한 변수를 제거함으로써, 즉, 대응하는 계수의 추정치를 0으로 설정함으로써 우리는 훨씬 더 쉽게 해석되는 모형을 얻을 수 있다.

01

## Subset Selection

중요한 변수를 선정하고 중요하지 않은 변수를 버리는 작업

02

## Shrinkage ( = regularization)

중요하지 않은 변수에 해당하는 coefficient의 절대값을 낮추는 방법

03

## Dimension Reduction

$p$개의 예측 변수들을 $M$차원 부분 공간으로 $projecting$

ESC

×

# 2. Linear Model Selection

전체 $p$개의 예측변수(X) 중 일부 $k$개 만을 사용하여 회귀 계수 beta를 추정하는 방법

01

## Best Subset Selection

02

## Forward Stepwise Selection

03

## Backward Stepwise Selection

04

## Hybrid Approaches

# Best Subset Selection

1. Let $\mathcal{M}_0$ denote that the ***null model***, which contains no predictors. This model simply predicts the sample mean for each observation.

$$Y = \beta_0 + \epsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} \cdot \beta_0 + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}$$

# Best Subset Selection

2. For $k = 1, 2, \ldots, p$:
   - (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictiors.
   - (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 $R^2$.

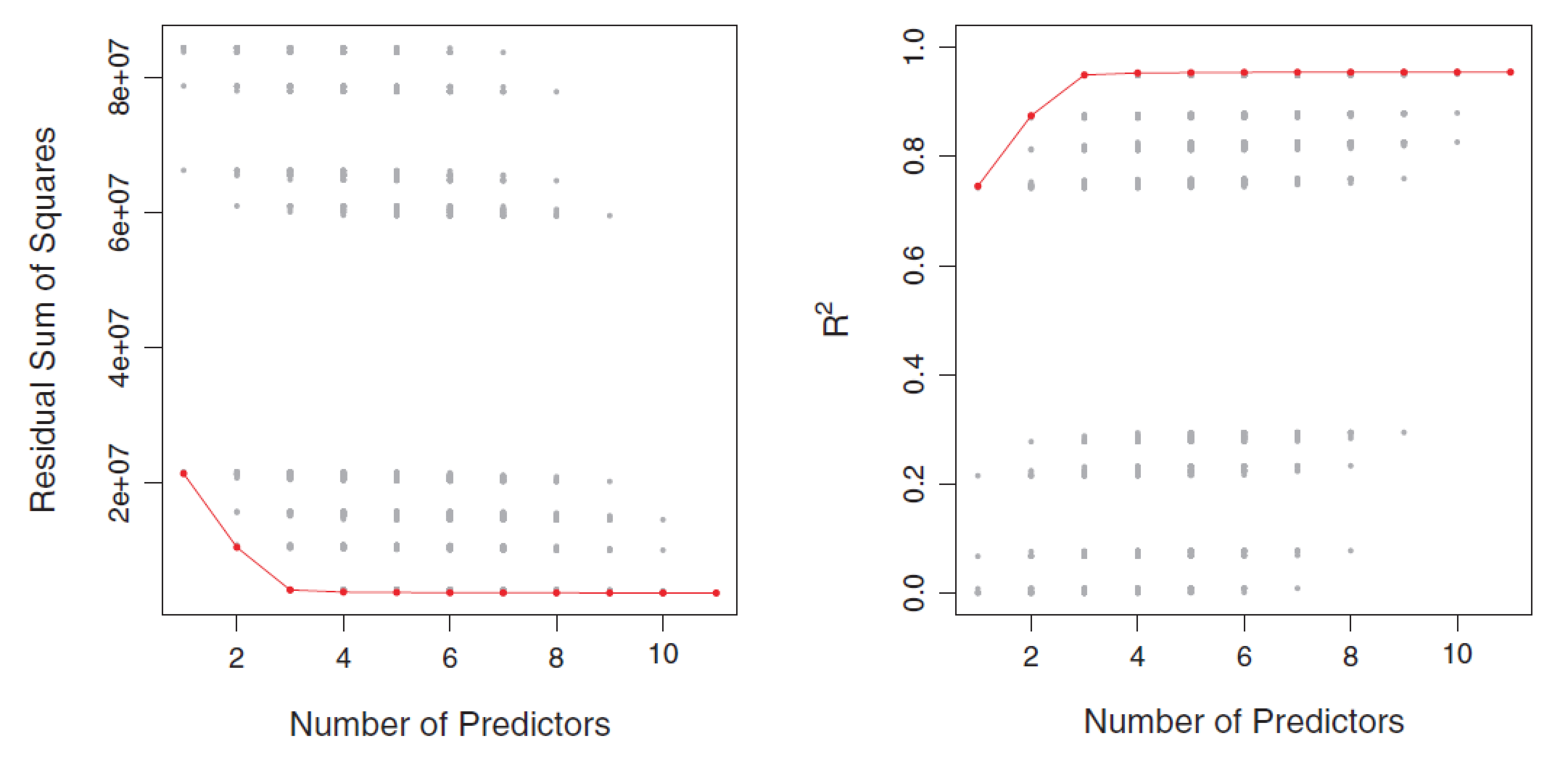
$$\left.\begin{array}{ll}
Y = \beta_0 + \epsilon & \rightarrow \binom{p}{0}\text{개} \\
Y = \beta_0 + \beta_1{}'X_1{}' + \epsilon & \rightarrow \binom{p}{1}\text{개} \\
Y = \beta_0 + \beta_1{}''X_1{}'' + \beta_2{}''X_2{}'' + \epsilon & \rightarrow \binom{p}{2}\text{개} \\
\quad\vdots & \\
\quad\vdots & \\
Y = \beta_0 + \beta_1{}'''X_1{}''' + \beta_2{}'''X_2{}''' + \cdots + \beta_p{}'''X_p{}''' + \epsilon & \rightarrow \binom{p}{p}\text{개}
\end{array}\right\} 2^p \text{ models}$$

# Best Subset Selection

3. Select a single best model from among $\mathcal{M}_0$, . . . , $\mathcal{M}_p$ using cross-validated predicton error, $C_p$ (AIC), BIC, or adjusted $R^2$.

$$\left.\begin{array}{ll}
Y = \beta_0 + \epsilon & \rightarrow \mathcal{M}_0 \\
Y = \beta_0 + \beta_1{}'X_1{}' + \epsilon & \rightarrow \mathcal{M}_1 \\
Y = \beta_0 + \beta_1{}''X_1{}'' + \beta_2{}''X_2{}'' + \epsilon & \rightarrow \mathcal{M}_2 \\
\quad\vdots & \\
\quad\vdots & \\
Y = \beta_0 + \beta_1{}'''X_1{}''' + \beta_2{}'''X_2{}''' + \cdots + \beta_p{}'''X_p{}''' + \epsilon & \rightarrow \mathcal{M}_p
\end{array}\right\} p + 1 \text{ models}$$

# Best Subset Selection



## Best Subset Selection

$$p = 10 \quad \longrightarrow \quad 2^p = 1024$$

$$p = 20 \quad \longrightarrow \quad 2^p = 1{,}048{,}576$$

$$\vdots$$

$$p = 40 \quad \longrightarrow \quad 2^p =$$

# Best Subset Selection

$$p = 10 \quad \longrightarrow \quad 2^p = 1024$$

$$p = 20 \quad \longrightarrow \quad 2^p = 1{,}048{,}576$$

$$\vdots$$

$$p = 40 \quad \longrightarrow \quad 2^p =$$

# Forward Stepwise Selection

1. Let $\mathcal{M}_0$ denote that the ***null model***, which contains no predictors.

$$Y = \beta_0 + \epsilon \qquad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} \cdot \beta_0 + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}$$

# Forward Stepwise Selection

2. For $k = 0, 1, \ldots, p - 1$ :
   (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 the smallest $RSS$, or highest $R^2$.

$$\left.\begin{array}{ll}
Y = \beta_0 + \epsilon & \rightarrow p - 0 \text{ 개} \\
Y = \beta_0 + \beta_1{}'X_1{}' + \epsilon & \rightarrow p - 1 \text{ 개} \\
Y = \beta_0 + \beta_1{}'X_1 + \beta_2{}'X_2{}' + \epsilon & \rightarrow p - 2 \text{ 개} \\
\quad \vdots & \\
\quad \vdots & \\
Y = \beta_0 + \beta_1{}'X_1{}' + \beta_2{}'X_2{}' + \cdots + \beta_p{}'X_p{}' + \epsilon & \rightarrow p \text{ - } (p - 1) \text{ 개}
\end{array}\right\}
\begin{array}{l}
\displaystyle\sum_{k=0}^{p-1} p - k \\
\displaystyle = 1 + \frac{p\,(\,p+1\,)}{2}
\end{array}$$

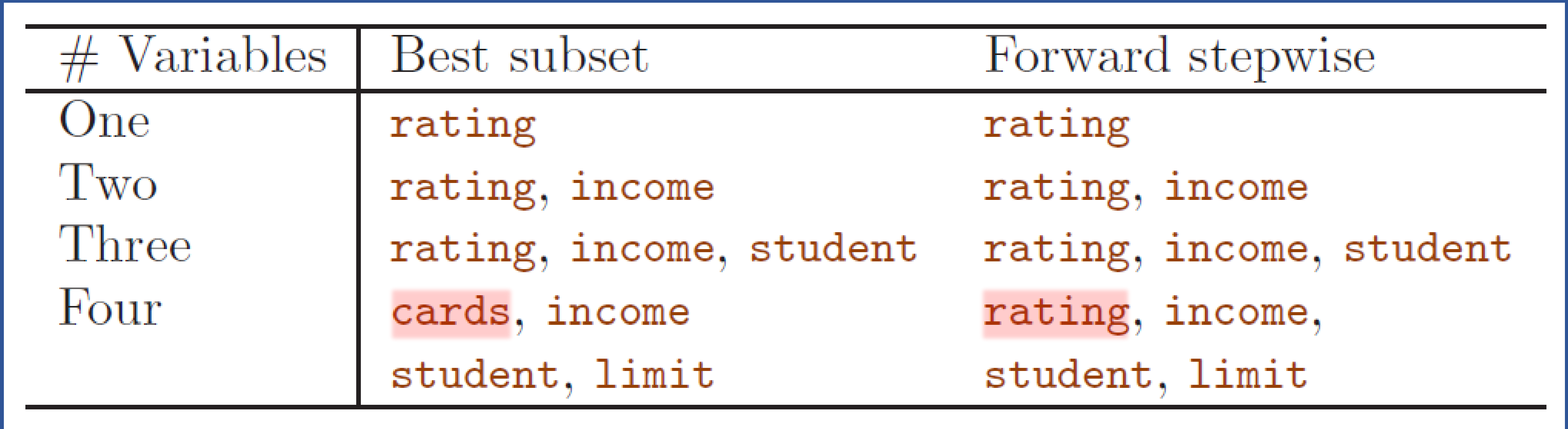
# Forward Stepwise Selection

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using across-validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.

$$p = 10 \quad \longrightarrow \sum_{k=0}^{p-1} p \; - k = 56$$

$$p = 20 \quad \longrightarrow \sum_{k=0}^{p-1} p \; - k = 211$$

# Forward 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<br>student, limit | rating, income,<br>student, limit |

**Underfitting Issue (모든 데이터를 보지 않아 문제 발생)**

모델이 데이터의 중요한 패턴을 학습하지 못해서 학습 데이터에서도 성능이 떨어진다.

# Backward Stepwise Selection

1. Let $\mathcal{M}_p$ denote that the ***full model***, which contains all $p$ predictors.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}$$

# Backward Stepwise Selection

2. For $k = p, p \; - 1, \dots, 1$ :
   (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 the smallest $RSS$, or highest $R^2$.

$$\left.\begin{array}{ll} Y = \beta_0 + \beta_1'X_1' + \beta_2'X_2' + \cdots + \beta_p'X_p' + \epsilon & \rightarrow p \text{ 개} \\ Y = \beta_0 + \beta_1'X_1' + \beta_2'X_2' + \cdots + \beta_{p-1}'X_{p-1}' + \epsilon & \rightarrow p - 1 \text{ 개} \\ \quad \vdots & \\ \quad \vdots & \\ Y = \beta_0 + \beta_1'X_1' + \beta_2'X_2' + \epsilon & \rightarrow 2 \text{ 개} \\ Y = \beta_0 + \beta_1'X_1' + \epsilon & \rightarrow 1 \text{ 개} \end{array}\right\} \begin{array}{l} \displaystyle\sum_{k=0}^{p-1} p - k \\ \displaystyle = 1 + \frac{p\,(p+1)}{2} \end{array}$$

# Backward Stepwise Selection

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using across-validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.

$$n < p$$

**Overfitting Issue** (모델이 복잡해 문제 발생)

데이터의 양(또는 sample의 수)이 예측변수보다 많도록 해야 한다. **(so that the full model can be fit)**

# Hybrid Approaches

Forward Stepwise Selection + Backward Stepwise Selection

1. Forward Selection을 통해 변수를 선택
2. Backward Selection을 통해 선택된 변수들 중 중요하지 않은 변수 제거
3. 추가하거나 제거할 변수가 없을 때 종료

# Application in R

## ISL 6.5 Subset Selection Methods p.244

1. ISLR 패키지의 Hitters 데이터 불러오기 & 결측치 제거

2. leaps 패키지의 regsubsets 함수를 사용하여 Subset Selection 실행
   ( regsubsets(종속변수 ~., data = 데이터 셋 이름) )

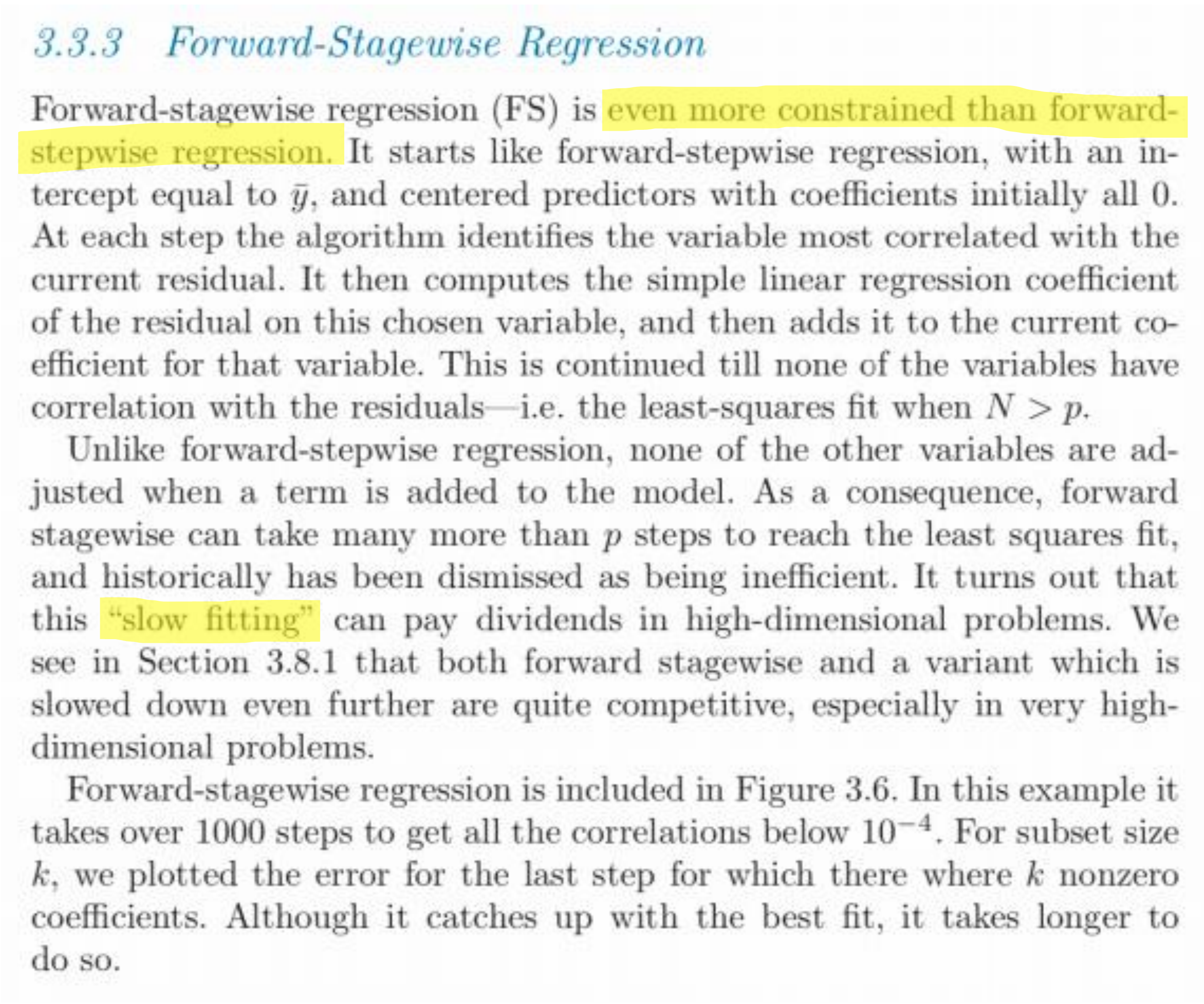
3. summary 함수로 단계별 선택된 변수들 확인

4. coef 함수로 단계별 beta의 추정치 확인
   ( coef( 회귀분석 모델, 선택된 변수의 개수) )

ESC

# 3. Forward Stagewise Selection

# Forward Stagewise Selection



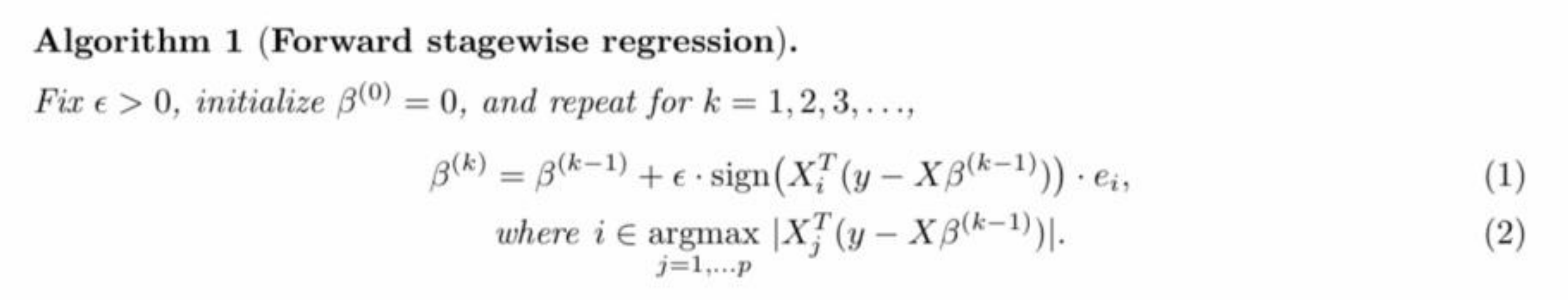
### 3.3.3 Forward-Stagewise Regression

Forward-stagewise regression (FS) is even more constrained than forward-stepwise regression. It starts like forward-stepwise regression, with an intercept equal to $\bar{y}$, and centered predictors with coefficients initially all 0. At each step the algorithm identifies the variable most correlated with the current residual. It then computes the simple linear regression coefficient of the residual on this chosen variable, and then adds it to the current coefficient for that variable. This is continued till none of the variables have correlation with the residuals—i.e. the least-squares fit when $N > p$.

Unlike forward-stepwise regression, none of the other variables are adjusted when a term is added to the model. As a consequence, forward stagewise can take many more than $p$ steps to reach the least squares fit, and historically has been dismissed as being inefficient. It turns out that this "slow fitting" can pay dividends in high-dimensional problems. We see in Section 3.8.1 that both forward stagewise and a variant which is slowed down even further are quite competitive, especially in very high-dimensional problems.

Forward-stagewise regression is included in Figure 3.6. In this example it takes over 1000 steps to get all the correlations below $10^{-4}$. For subset size $k$, we plotted the error for the last step for which there where $k$ nonzero coefficients. Although it catches up with the best fit, it takes longer to do so.

# Forward Stagewise Selection



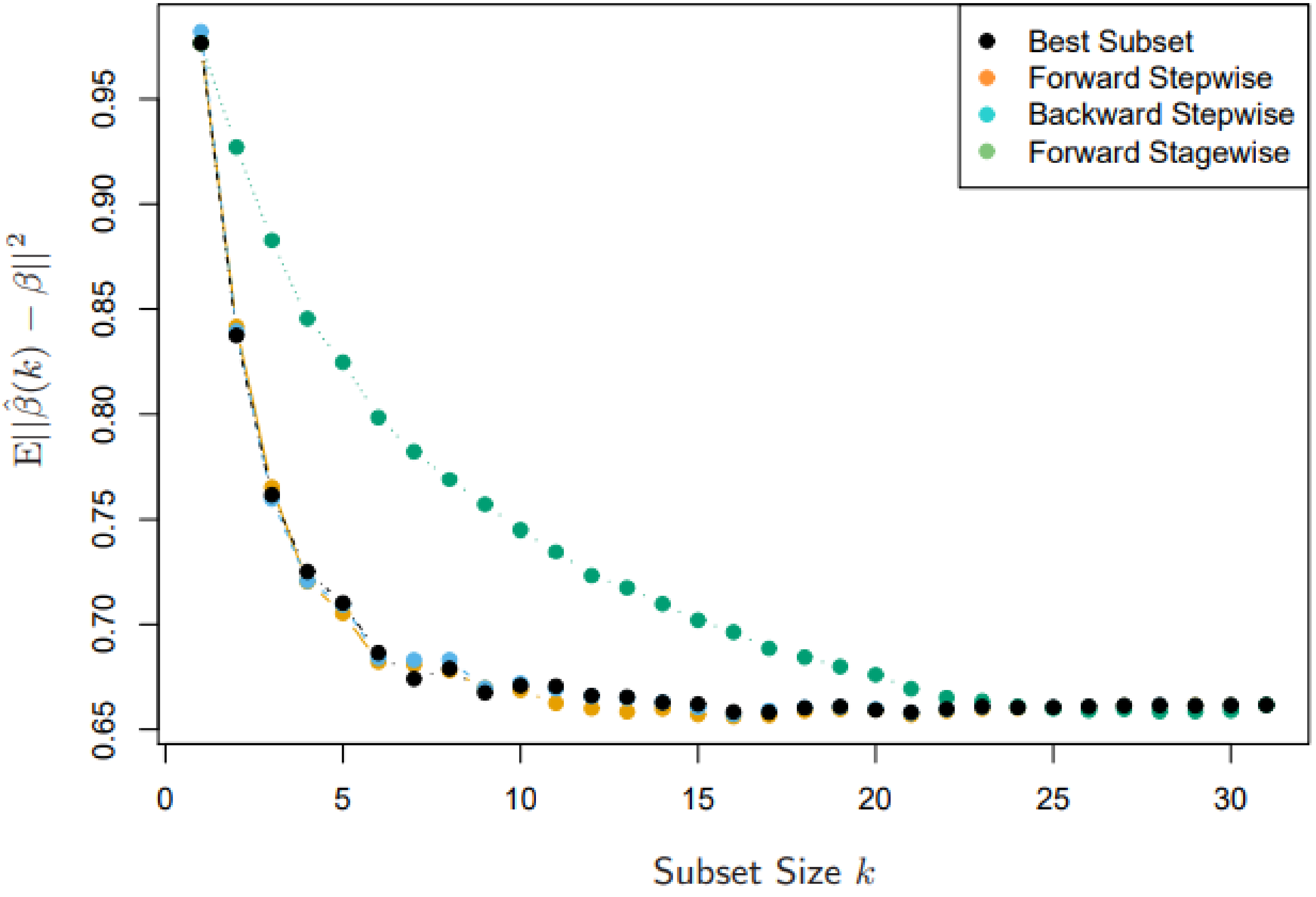
**Algorithm 1 (Forward stagewise regression).**

*Fix* $\epsilon > 0$, *initialize* $\beta^{(0)} = 0$, *and repeat for* $k = 1, 2, 3, \ldots,$

$$\beta^{(k)} = \beta^{(k-1)} + \epsilon \cdot \operatorname{sign}\big(X_i^T(y - X\beta^{(k-1)})\big) \cdot e_i, \quad (1)$$

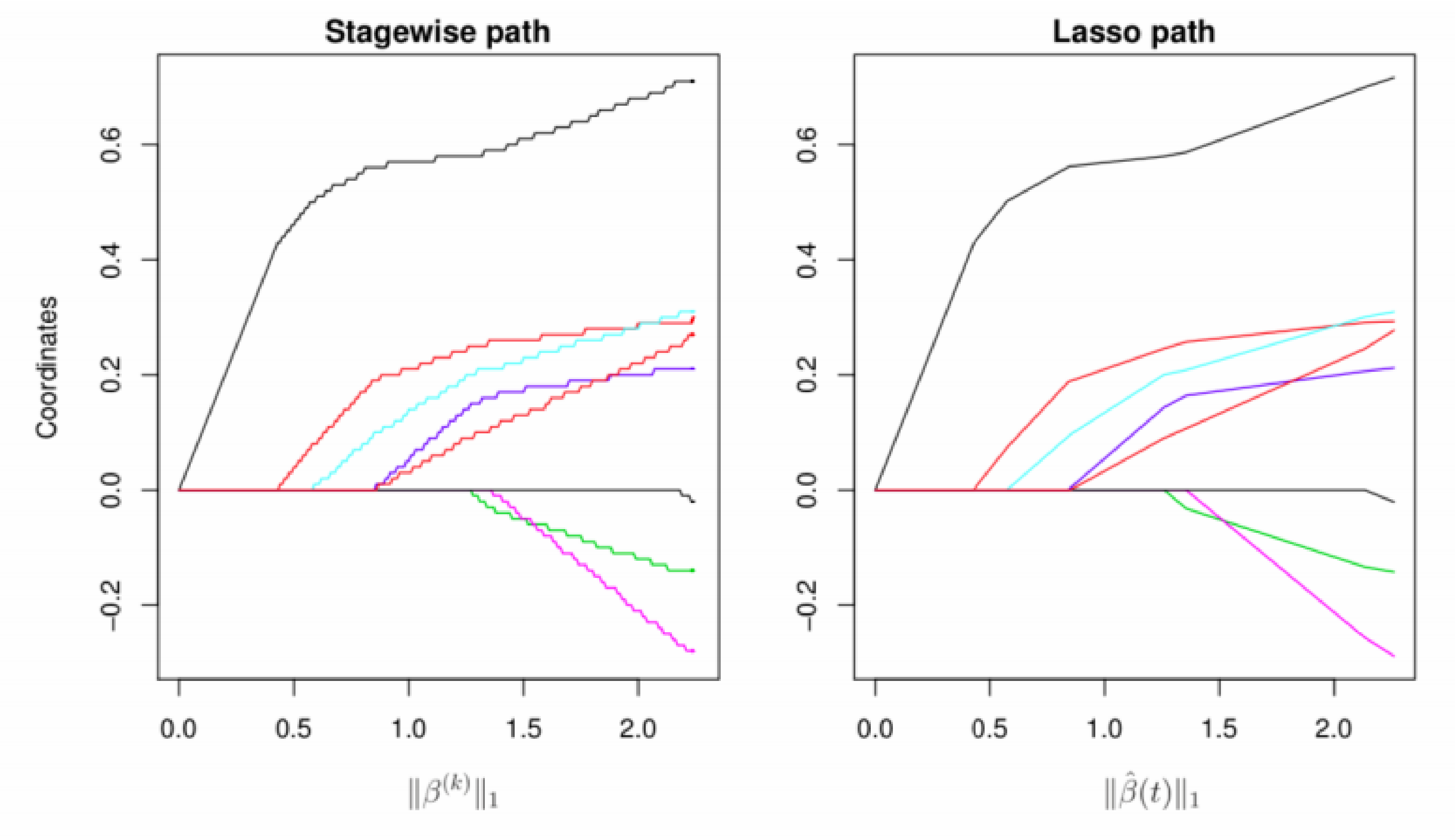
$$\text{where } i \in \underset{j=1,\ldots p}{\operatorname{argmax}} \; |X_j^T(y - X\beta^{(k-1)})|. \quad (2)$$

# Forward Stagewise Selection



# Forward Stagewise Selection

+ Relationship between LASSO



ESC

×

# 4. Selection Criteria

# Loss and error

**Loss function** : $L(Y, \hat{f}(X))$ where

target variable $Y$

input $X$

prediction model $\hat{f}(X)$ estimated from training set $\mathcal{T}$

- Various types of loss functions

Squared error: $(Y - \hat{f}(X))^2$

Absolute error: $|Y - \hat{f}(X)|$

Deviance(= -2*log likelihood): $L(G, \hat{p}(X)) = -2\sum_{k=1}^{K} I(G = k) log_{\hat{p}_k}(X)$

0-1 loss: $L(G, \hat{G}(X)) = I(G \neq \hat{G}(X))$

# Loss and error

**Test error(generalization error)**

$$Err_{\mathcal{T}} = E[L(Y, \hat{f}(X))|\mathcal{T}]$$

: drawn randomly from joint distribution (random data)
: hard to predict so we use **expected test error**

**expected test error**

$$Err = E[L(Y, \hat{f}(X))] = E[Err_{\mathcal{T}}]$$

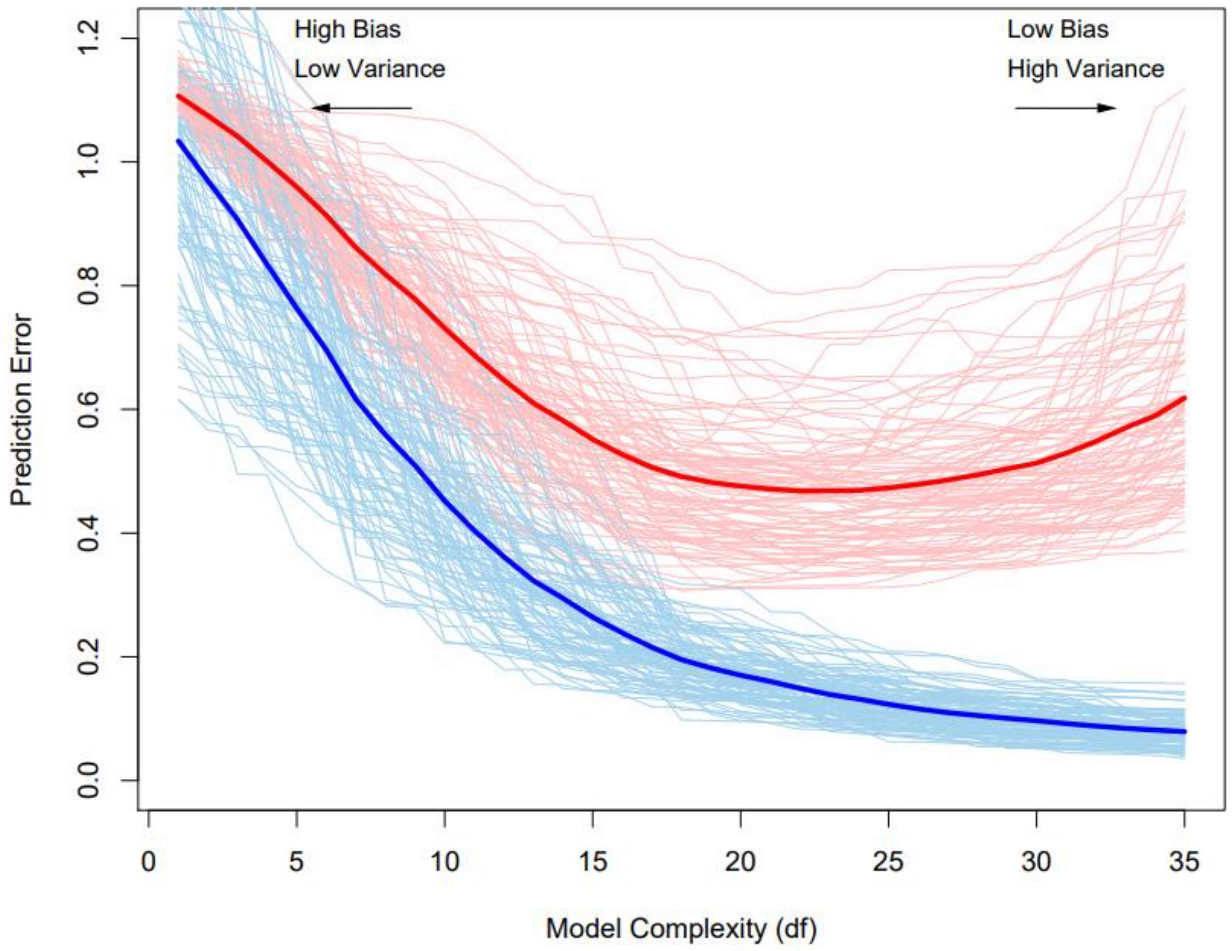
: most methods estimate this!

**Training error**

$$e\bar{r}r = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

: not good estimate for test error

# Loss and error



Thus, we estimate the **expected test error** to achieve goals of
1.model selection
2.model assessment

# Bias – variance decomposition

Our model was …… $Y = f(X) + \epsilon$ where $E(\epsilon) = 0$

$Var(\epsilon) = \sigma_\epsilon^2$

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance.}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}_k(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + \left[ f(x_0) - \frac{1}{k} \sum_{\ell=1}^{k} f(x_{(\ell)}) \right]^2 + \frac{\sigma_\varepsilon^2}{k}
\end{aligned}$$

# Bias – variance decomposition

**Linear model** : $\hat{f}_p(x) = x^T\beta$

$$Err(x_0) = E[(Y - \hat{f}_p(x_0))^2 | X = x_0]$$

$$= \sigma_\epsilon^2 + [f(x_0) - E\hat{f}_p(x_0)]^2 + \|h(x_0)\|^2\sigma_\epsilon^2$$

where

$$\hat{f}_p(x_0) = x_0^T\beta = x_0^T(X^TX)^{-1}X^Ty = h(x_0)^Ty$$

$$h(x_0) = X(X^TX)^{-1}x_0$$

+ in-sample error

$$\frac{1}{N}\sum Err(x_i) = \sigma_\epsilon^2 + \frac{1}{N}\sum[f(x_0) - E\hat{f}_p(x_0)]^2 + \frac{p}{N}\sigma_\epsilon^2$$

# Bias – variance decomposition

+ bias can be decomposed also …

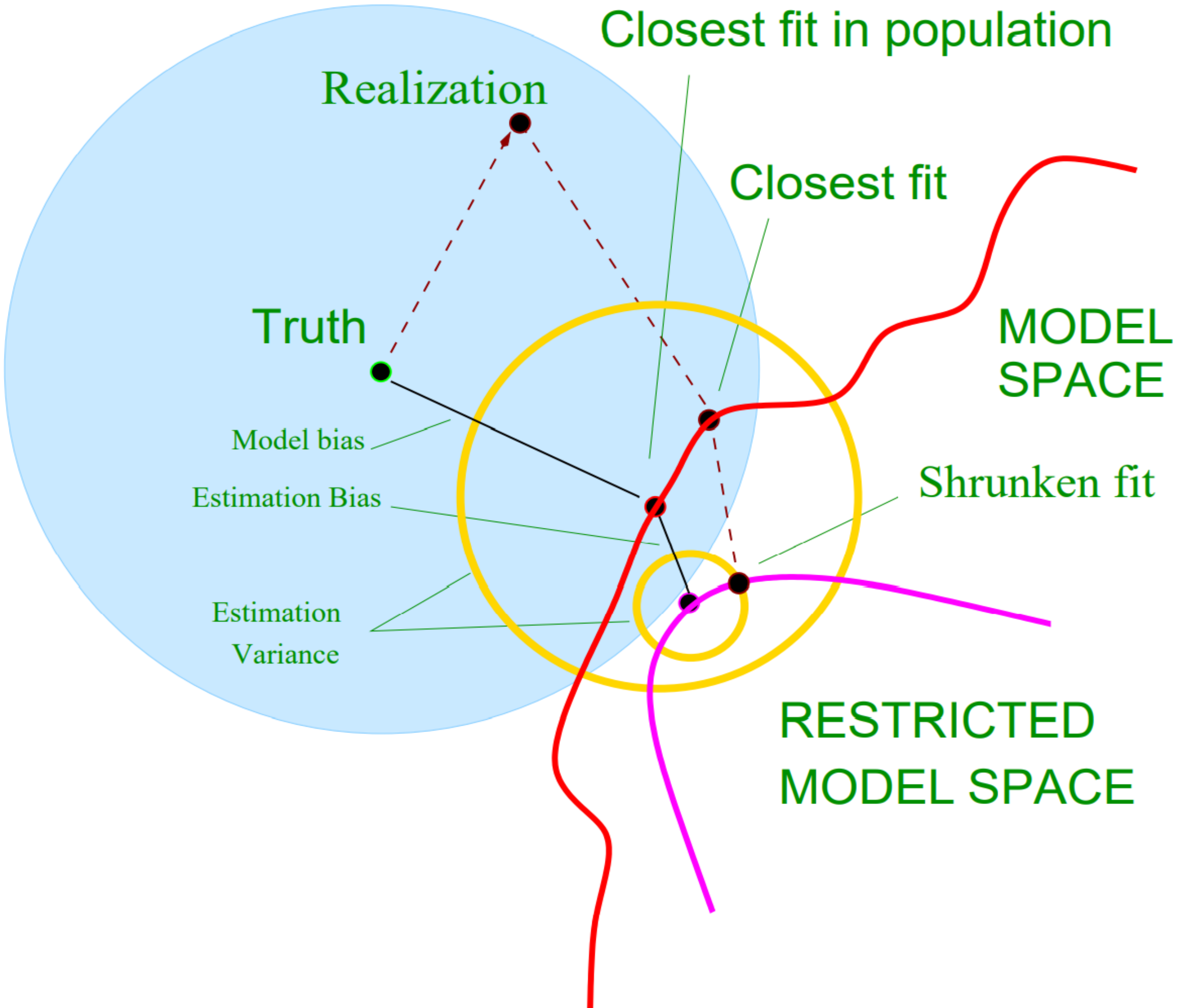
For $Err(x_0) = E[(Y - \hat{f}_p(x_0))^2|X = x_0]$

$$= \sigma_\epsilon^2 + [f(x_0) - E\hat{f}_p(x_0)]^2 + \|h(x_0)\|^2\sigma_\epsilon^2$$

and $\beta_* = \arg\min E(f(X) - X^T\beta)^2$

$$\begin{aligned} \mathrm{E}_{x_0}\left[f(x_0) - \mathrm{E}\hat{f}_\alpha(x_0)\right]^2 &= \mathrm{E}_{x_0}\left[f(x_0) - x_0^T\beta_*\right]^2 + \mathrm{E}_{x_0}\left[x_0^T\beta_* - \mathrm{E}x_0^T\hat{\beta}_\alpha\right] \\ &= \mathrm{Ave}[\mathrm{Model\ Bias}]^2 + \mathrm{Ave}[\mathrm{Estimation\ Bias}]^2 \end{aligned}$$

# Bias – variance decomposition



# Optimism of the training error

**In – sample** error

$$Err_{\mathcal{T}} = E_{X^0,Y^0}[L(Y^0,\hat{f}(X^0))|\mathcal{T}]$$

$$Err = E_{\mathcal{T}}E_{X^0,Y^0}[L(Y^0,\hat{f}(X^0))|\mathcal{T}]$$

$$\boxed{e\bar{r}r = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} L(y_i,\hat{f}(x_i))}$$

**vs**

$$\boxed{Err_{in} = \tfrac{1}{N}\textstyle\sum E_{Y^0}[L(Y_i^0,\hat{f}(x_i))|\mathcal{T}]}$$

**Optimism**

$$op \equiv Err_{in} - e\bar{r}r$$

$$w \equiv E_y(op)$$

$$w = \tfrac{2}{N}\textstyle\sum Cov(\hat{y}_i, y_i)$$

$$\boxed{\mathrm{E}_{\mathbf{y}}(\mathrm{Err}_{\mathrm{in}}) = \mathrm{E}_{\mathbf{y}}(\overline{\mathrm{err}}) + \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i)}$$

# Selection Criteria

So we use

$$\widehat{\mathrm{Err}}_{\mathrm{in}} = \overline{\mathrm{err}} + \hat{\omega}$$

and the way “optimism” is estimated varies

**Cp**

$$C_p = \overline{\mathrm{err}} + 2 \cdot \frac{d}{N} \hat{\sigma_\varepsilon}^2$$ using $$\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i) = d\sigma_\varepsilon^2$$

**AIC**(Akaike Information Criteria)

Using log likelihood, $-2 \cdot \mathrm{E}[\log \mathrm{Pr}_{\hat{\theta}}(Y)] \approx -\frac{2}{N} \cdot \mathrm{E}[\mathrm{loglik}] + 2 \cdot \frac{d}{N}$

$$AIC = e\bar{r}r + 2\frac{d}{N}\hat{\sigma}_\epsilon^2$$ same as Cp when Gaussian !

# Selection Criteria

**BIC**(Bayesian Information Criteria)

$$\mathrm{BIC} = -2 \cdot \mathrm{loglik} + (\log N) \cdot d.$$

**vs**

$$AIC = -2 \cdot loglik + 2 \cdot d$$

When we assume Gaussian with squared error loss,

$$\mathrm{BIC} = \frac{N}{\sigma_\varepsilon^2}\left[\overline{\mathrm{err}} + (\log N) \cdot \frac{d}{N}\sigma_\varepsilon^2\right].$$

From posterior of model

$$\begin{aligned}
\Pr(\mathcal{M}_m|\mathbf{Z}) &\propto \Pr(\mathcal{M}_m) \cdot \Pr(\mathbf{Z}|\mathcal{M}_m) \\
&\propto \Pr(\mathcal{M}_m) \cdot \int \Pr(\mathbf{Z}|\theta_m, \mathcal{M}_m)\Pr(\theta_m|\mathcal{M}_m)d\theta_m
\end{aligned}$$

we get posterior odds

$$\frac{\Pr(\mathcal{M}_m|\mathbf{Z})}{\Pr(\mathcal{M}_\ell|\mathbf{Z})} = \frac{\Pr(\mathcal{M}_m)}{\Pr(\mathcal{M}_\ell)} \cdot \frac{\Pr(\mathbf{Z}|\mathcal{M}_m)}{\Pr(\mathbf{Z}|\mathcal{M}_\ell)}$$

# Selection Criteria

**BIC**(Bayesian Information Criteria)

$$\begin{aligned}\Pr(\mathcal{M}_m|\mathbf{Z}) &\propto \Pr(\mathcal{M}_m)\cdot\Pr(\mathbf{Z}|\mathcal{M}_m)\\ &\propto \Pr(\mathcal{M}_m)\cdot\int \Pr(\mathbf{Z}|\theta_m,\mathcal{M}_m)\Pr(\theta_m|\mathcal{M}_m)d\theta_m\end{aligned}$$

Using Laplace approximation,

$$\log\Pr(\mathbf{Z}|\mathcal{M}_m) = \log\Pr(\mathbf{Z}|\hat{\theta}_m,\mathcal{M}_m) - \frac{d_m}{2}\cdot\log N + O(1).$$

So we get BIC as

$$-2\log\Pr(\mathbf{Z}|\hat{\theta}_m,\mathcal{M}_m)$$

+ comparing posteriors using

$$\frac{e^{-\frac{1}{2}\cdot\mathrm{BIC}_m}}{\sum_{\ell=1}^{M} e^{-\frac{1}{2}\cdot\mathrm{BIC}_\ell}}.$$

**+ Adjusted R squared**

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

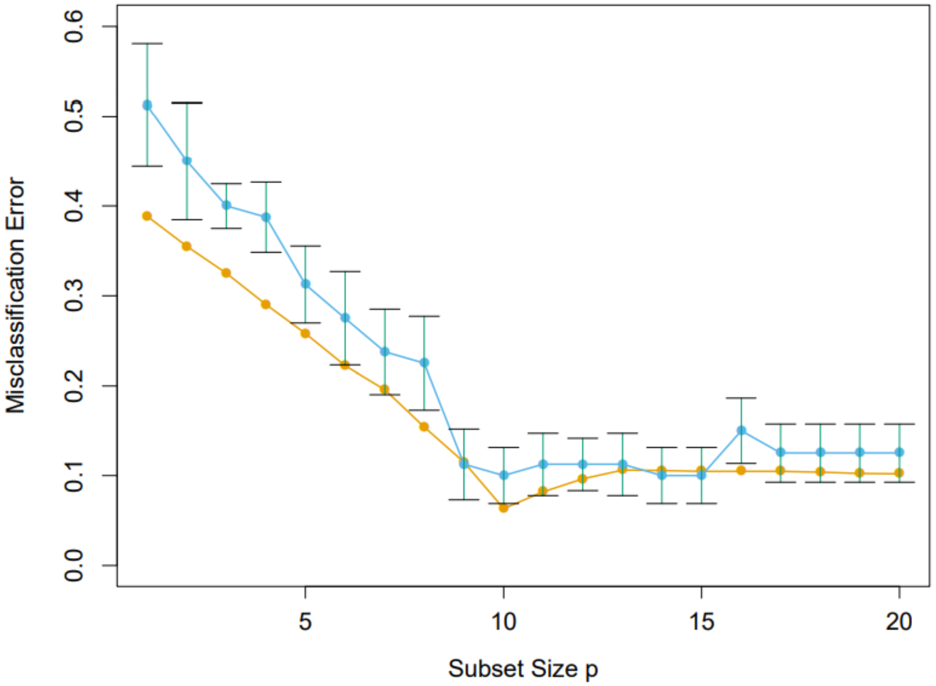
# Cross validation

Finally, **Cross validation**

$$\mathrm{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Train | Train | Validation | Train | Train |

Directly estimate expected test error



ESC

×

# Homework

# Homework

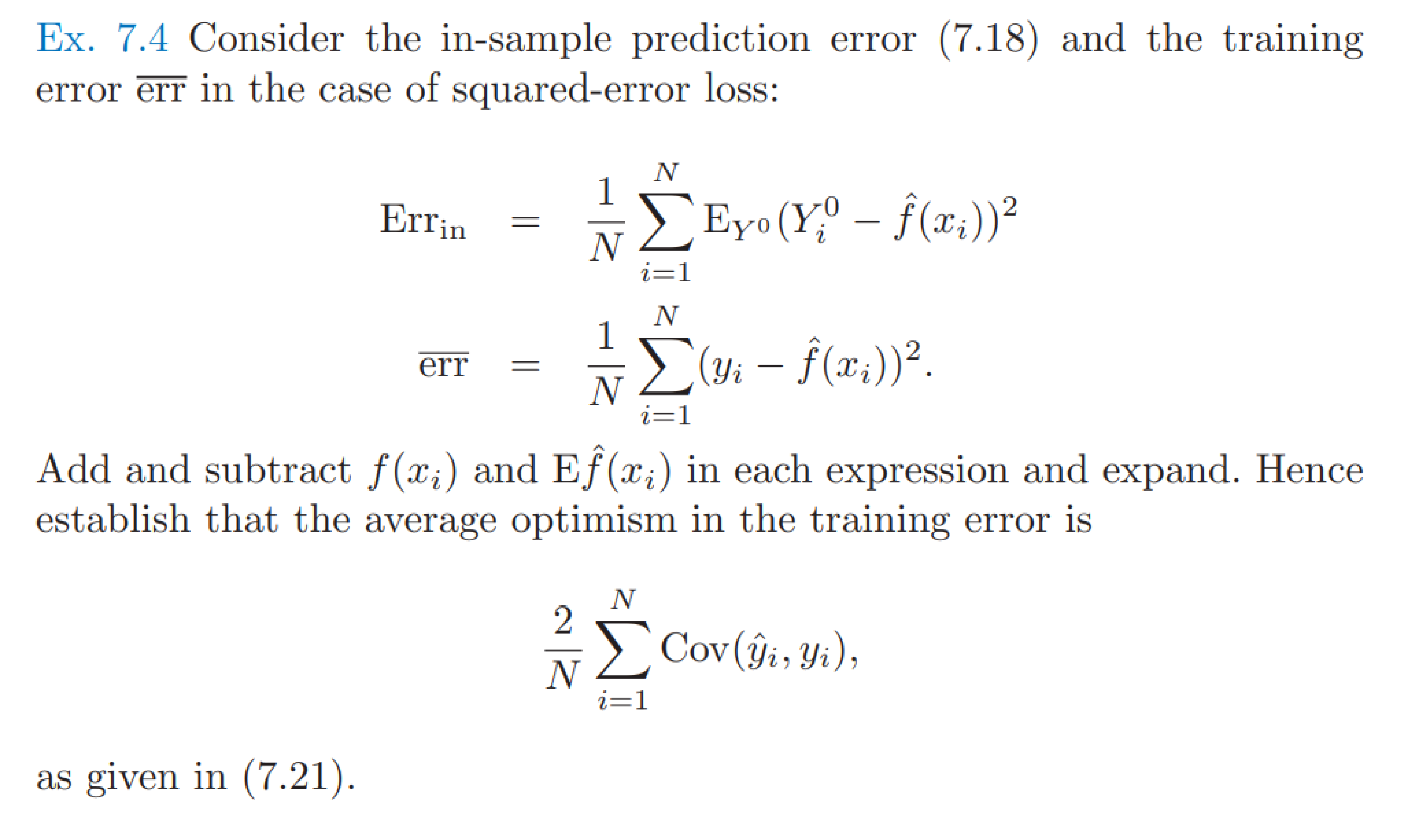
mlbench 패키지의 BostonHousing 데이터에 대하여
Forward Stepwise Selection 을 해봅시다.

종속변수는 집 가격을 나타내는 ‘medv’ 변수입니다. 나머지 13개의 변수는 독립변수로 설정합니다.
Hint : ISL 6.5.2를 참고하세요!

1. 단계별로 생성된 모델들 중 예측변수(또는 설명변수)가 4개인 모델에서, 새롭게 선택된 변수의 이름은 무엇인가요?

2. 그 변수의 추정된 계수의 값과 해당 모델의 결정계수 $R^2$는 무엇인가요?

# Homework



Ex. 7.4 Consider the in-sample prediction error (7.18) and the training error $\overline{\text{err}}$ in the case of squared-error loss:

$$\begin{aligned}\text{Err}_{\text{in}} &= \frac{1}{N}\sum_{i=1}^{N}\text{E}_{Y^0}(Y_i^0 - \hat{f}(x_i))^2\\ \overline{\text{err}} &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2.\end{aligned}$$

Add and subtract $f(x_i)$ and $\text{E}\hat{f}(x_i)$ in each expression and expand. Hence establish that the average optimism in the training error is

$$\frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i),$$

as given in (7.21).

**ESC**

×

Thank you