

## Chapter06

## 선행회귀

# I 선행회귀의 원리 II

FASTCAMPUS  
ONLINE

금융공학/퀀트 I

강사. 장순용

# I 키포인트

- 선형회귀 OLS 해.
- 선형회귀 학습과 예측.
- 더미변수.



# I 선형회귀 원리

- 회귀모형:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

# I 선형회귀 해

- 벡터와 행렬 사용 표기:

$$y_j = \beta_0 + \beta_1 x_{j,1} + \beta_2 x_{j,2} + \cdots + \beta_K x_{j,K} + \varepsilon_j$$



$$j \in [1, n]$$

# I 선형회귀 해


- 벡터와 행렬 사용 표기:

$$\overrightarrow{Y} = \tilde{X} \overrightarrow{\beta} + \overrightarrow{\varepsilon}$$

# I 선형회귀 해

- 벡터와 행렬 사용 표기:


$$\vec{Y} = \tilde{X} \vec{\beta} + \vec{\varepsilon}$$


$$\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

# I 선형회귀 해

- 벡터와 행렬 사용 표기:


$$\vec{Y} = \tilde{X} \vec{\beta} + \vec{\varepsilon}$$


$$\tilde{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,K} \\ 1 & x_{2,1} & \cdots & x_{2,K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,K} \end{pmatrix}$$

# I 선형회귀 해

- 벡터와 행렬 사용 표기:

$$\vec{Y} = \tilde{X} \vec{\beta} + \vec{\varepsilon}$$




$$\vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$$



# I 선형회귀 해

- 벡터와 행렬 사용 표기:

$$\vec{Y} = \tilde{X} \vec{\beta} + \vec{\varepsilon}$$


$$\vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

# I 선형회귀 해

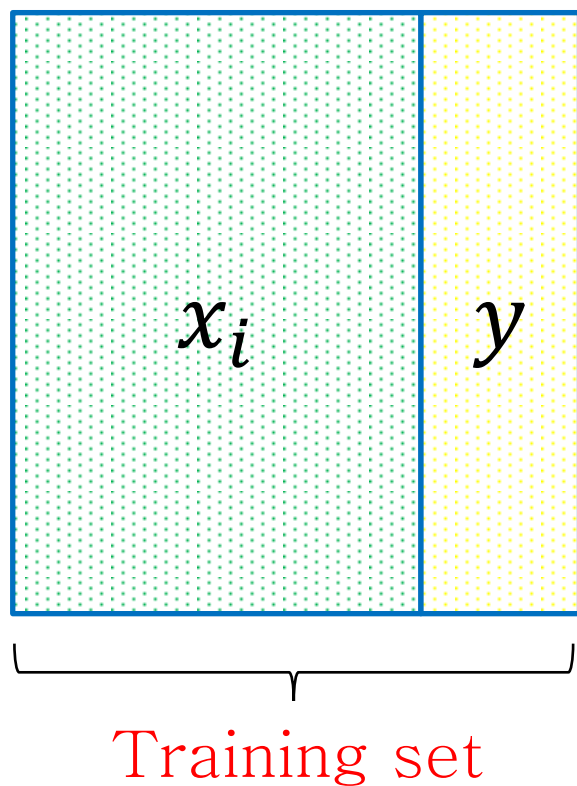
- 선형회귀의 OLS (Ordinary Least Squares solution) 해:

$$\vec{\beta} = \left[ (\tilde{X}^t \tilde{X})^{-1} \tilde{X}^t \right] \vec{Y}$$



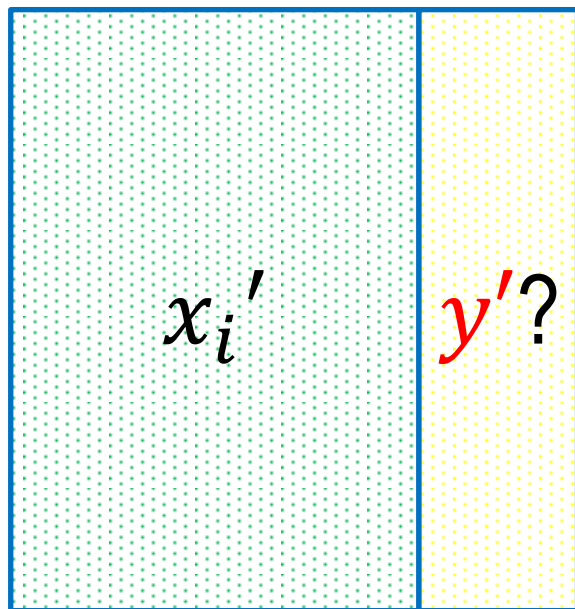
$\|\vec{\epsilon}\|^2$ 를 최소화 하는 계수벡터이다.

# I 선형회귀 학습



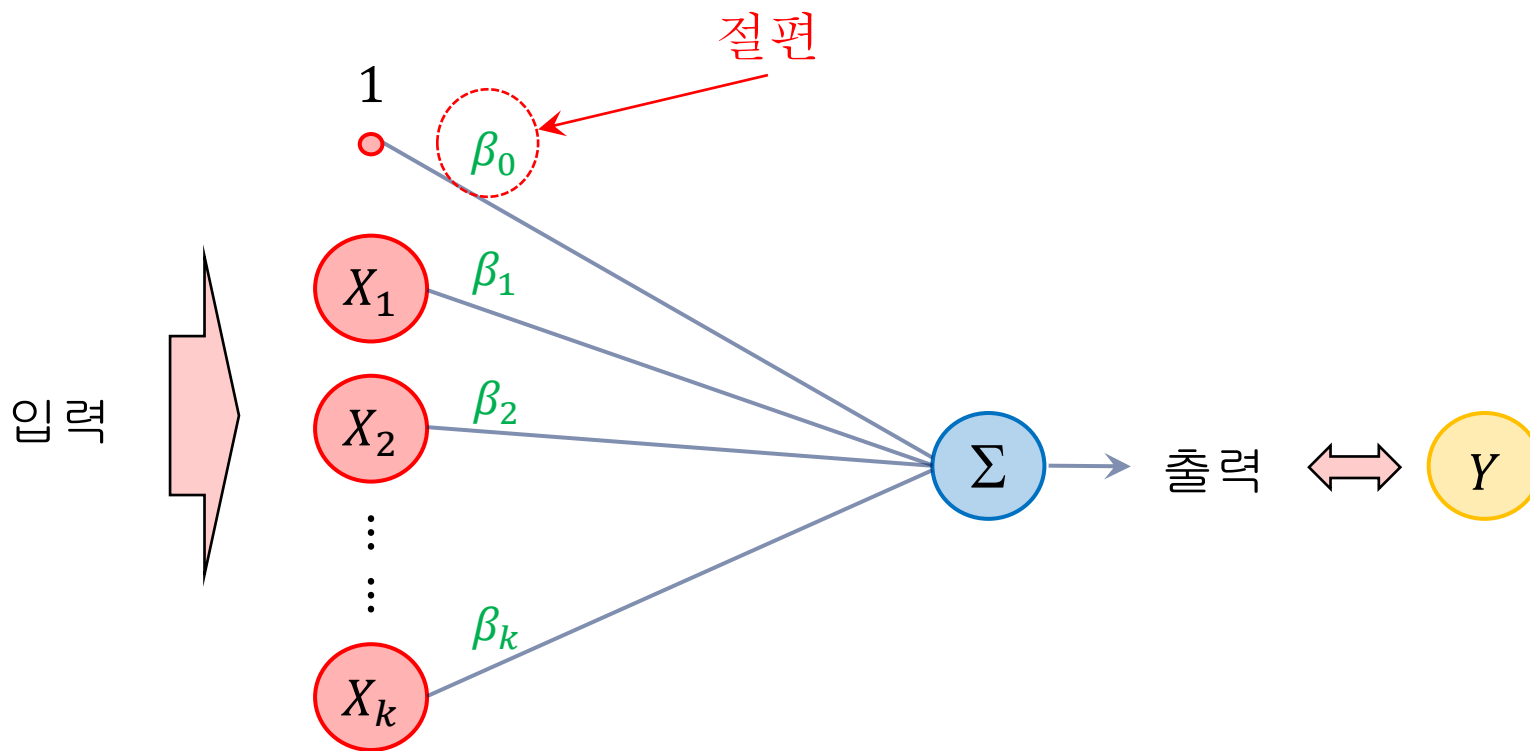
모델의 파라미터, 즉  $\{\beta_i\}$ 를 학습용 데이터를 사용하여 계산해 놓는다.

## I 선형회귀 예측

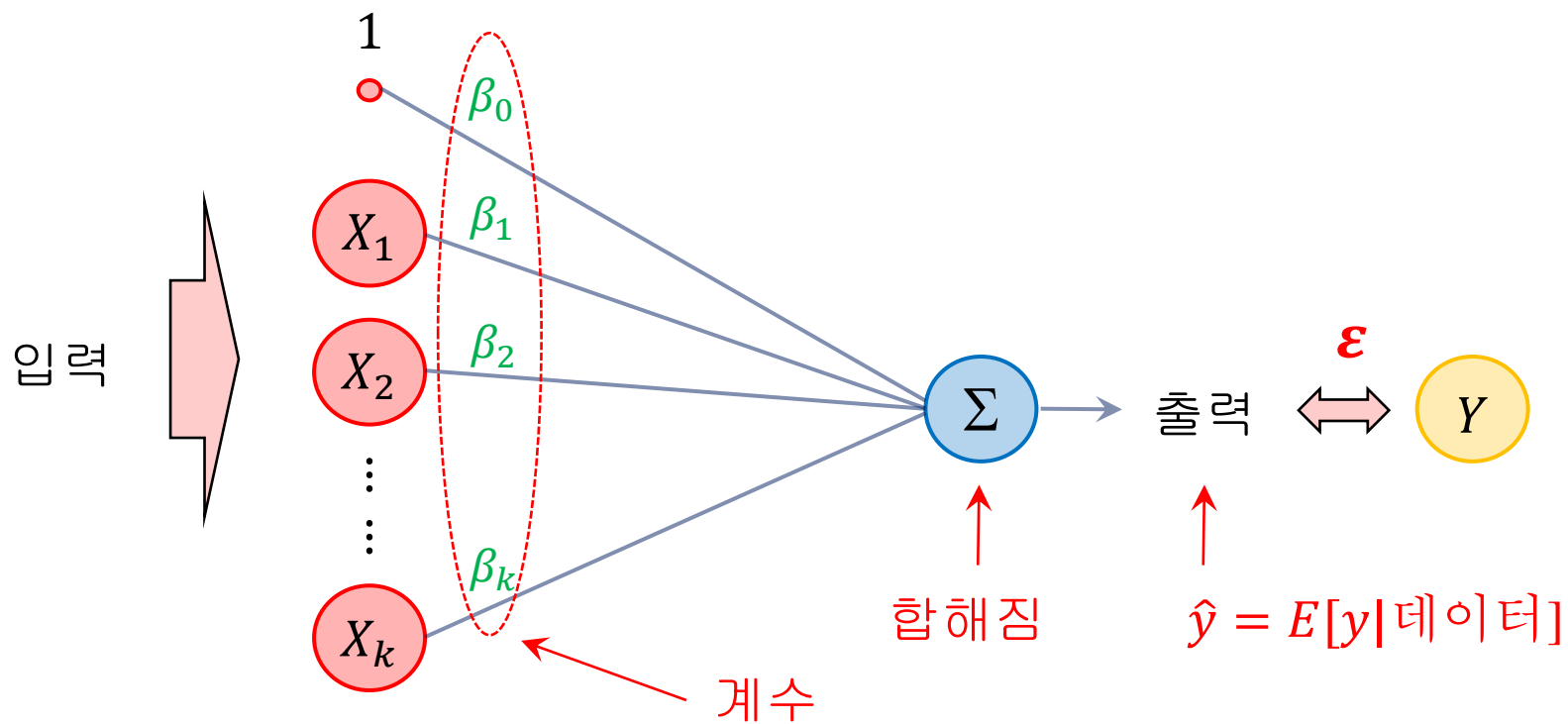


독립변수의 값이 새롭게 주어졌을 때  $\{x_i'\}$ , 모르는 상태인 종속변수의 값  $y'$  을 계산을 통해서 알아낸다.

## I 선형회귀 예측



## I 선형회귀 예측



# I 선형회귀 예측

- 관심점에 해당하는 독립변수의 값  $x_1', x_2', \dots, x_k'$ 가 주어졌을 때 다음 수식을 사용해서 종속변수의 예측값  $\hat{y} = E[y|\text{데이터}]$  를 구한다.

$$\hat{y} = \beta_0 + \beta_1 x_1' + \beta_2 x_2' + \dots + \beta_K x_K'$$



## I 선형회귀 예측

- 독립변수가 **하나 뿐인** 경우에는 예측값의 95% 신뢰구간은 다음과 같이 계산할 수 있다.

$$[\hat{y} - qt(0.975, n - 2) \sigma_{\hat{y}}, \hat{y} + qt(0.975, n - 2) \sigma_{\hat{y}}]$$

$$\sigma_{\hat{y}} = RMSE \times \sqrt{\left(\frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum (x - \bar{x})^2}\right)}$$



$qt()$ 은 R의 분위수 함수

## I 더미변수 (dummy variable)

- 더미변수는 0과 1만을 값으로 갖는 변수이다.  $\Rightarrow$  switch on/off의 역할.
- 명목형 변수를 모형에 추가하면 **유형의 가지수 - 1** 개의 더미변수 생성됨.

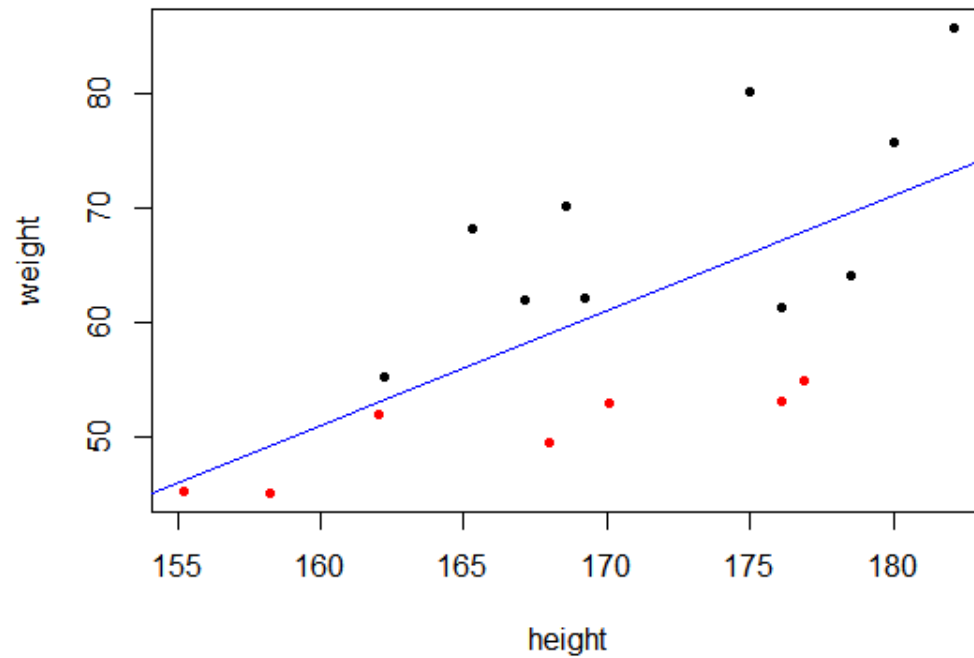
예). “남자”, “여자”와 같이 두 개의 유형을 값으로 갖는 “gender” 변수는 “gender여자”라는 한 개의 더미변수를 생성한다.

예). “setosa”, “versicolor”, “virginica”와 같이 세 개의 유형을 값으로 갖는 “Species” 변수는 “Speciesversicolor”, “Speciesvirginica” 라는 두 개의 더미변수를 생성한다.

## I 더미변수 (dummy variable)

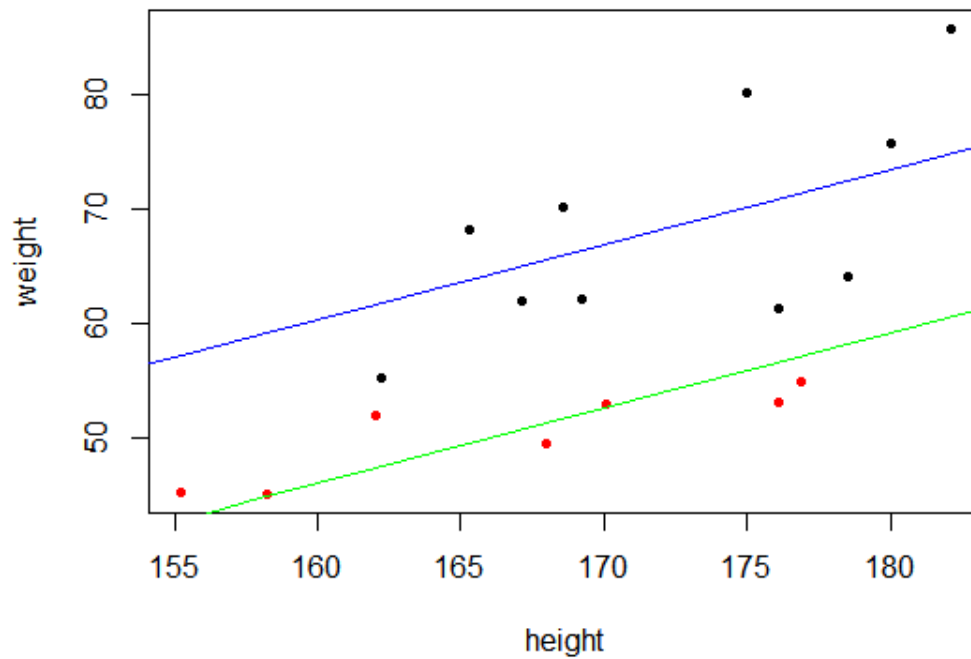
- 더미변수가 독립적으로 추가되면 해당 유형의 절편을 올리거나 내려주는 역할을 한다.
- 상호작용하는 더미변수는 해당 유형의 기울기를 조절해 주는 역할을 한다.

# I 더미변수 (dummy variable)



더미변수가 없음:  $\text{weight} \sim \text{height}$

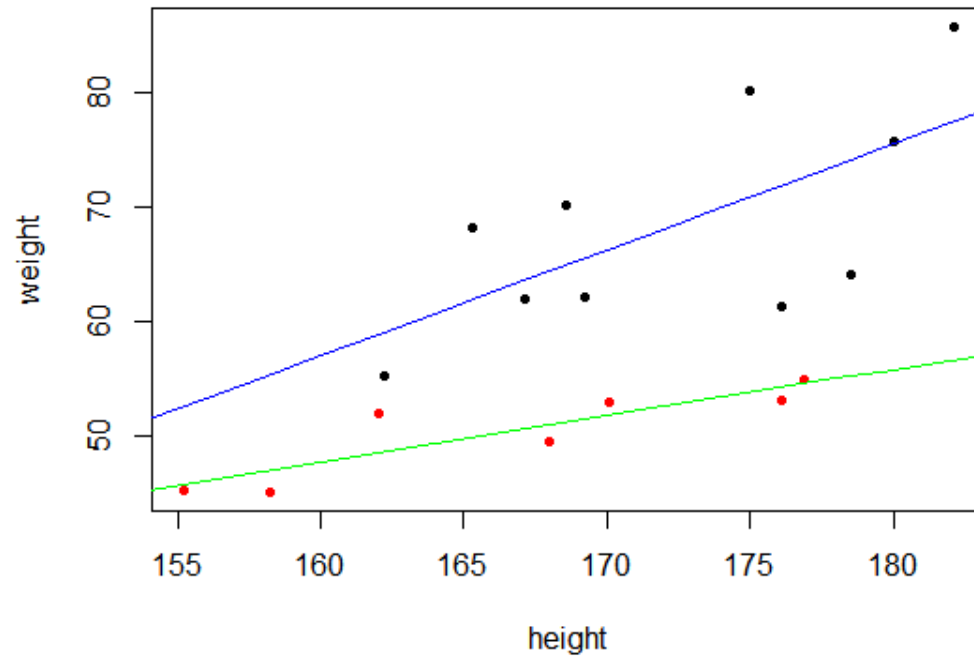
# I 더미변수 (dummy variable)



더미변수가 있음:  $\text{weight} \sim \text{height} + \text{gender}$

$R^2$  증가, MSE 감소

# I 더미변수 (dummy variable)



상호작용하는 더미변수가 있음:  $\text{weight} \sim \text{height} * \text{gender}$

$R^2$  증가, MSE 감소

I 끝.

감사합니다.

