Regression Problem

Data Scientist
안건이

목차

Loss Function

- Regression Loss Function

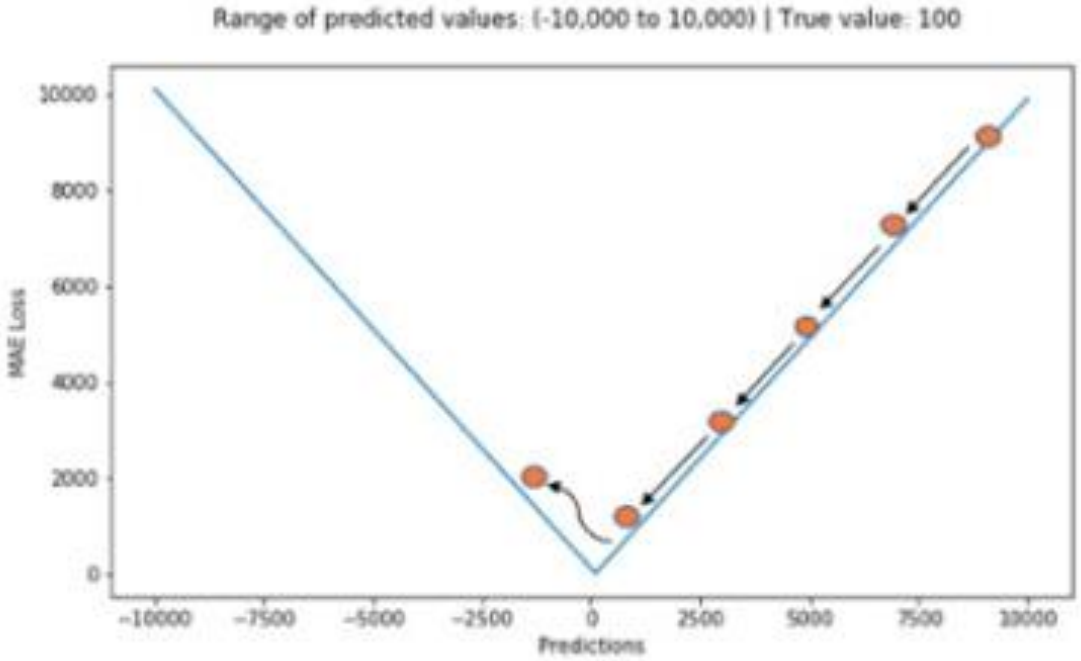
미분 불가능

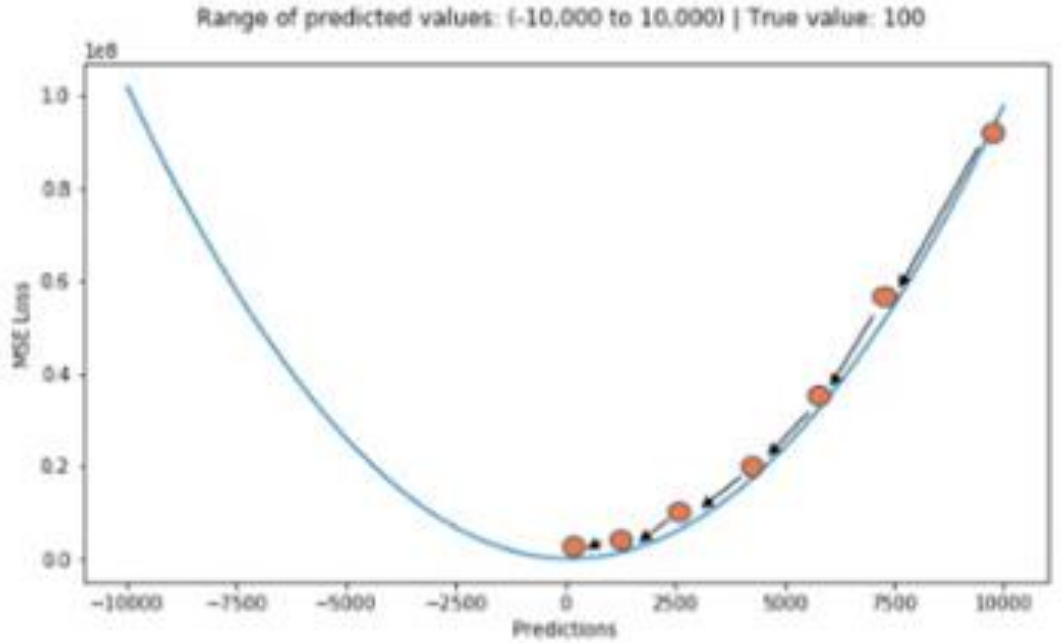
$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$$

미분 가능

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$





Model

Regression VS **Classification**

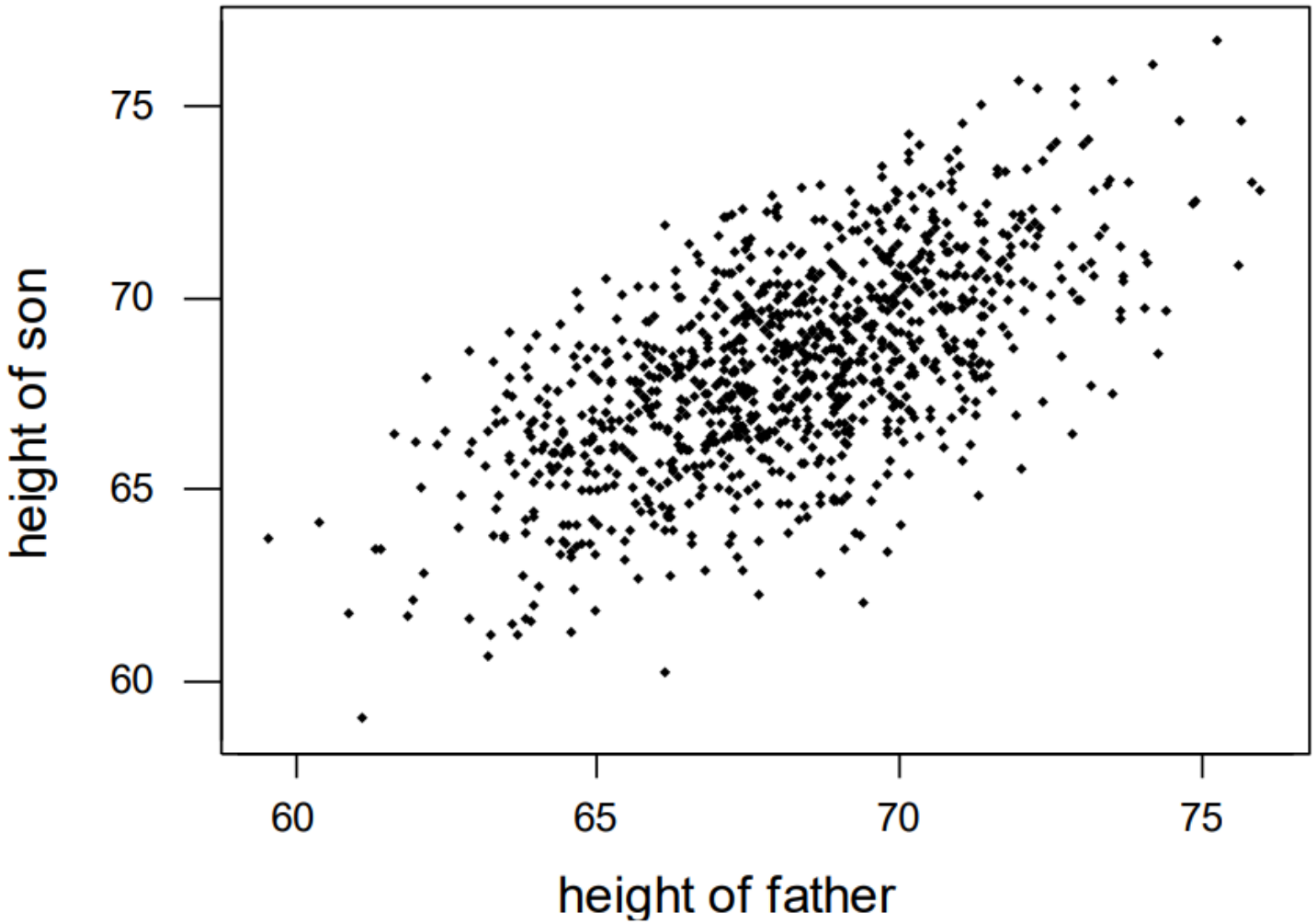
Y : Numeric　　Y : Categorical

더 어려운 문제는 ?

eXplainable 한 건 ?

Why Regression ?

Linear Regression – 프랜시스 골턴

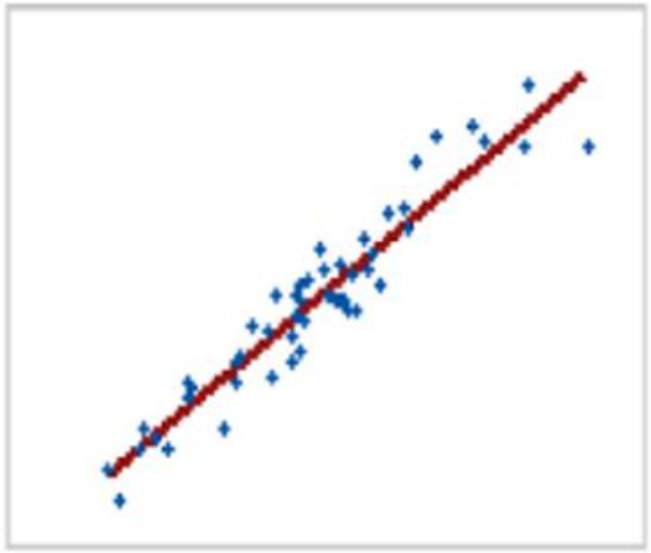


Francis Galton
(1822~1911)

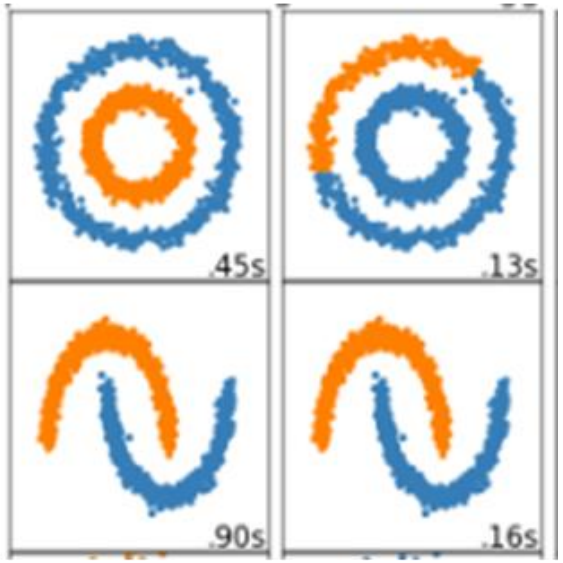
Linear Regression

- 장점
  - Model이 간단하기 때문에 모델 학습 시간이 짧음
  - 선형성 데이터에 적합함
  - 해석력이 가장 좋음

- 단점
  - 복잡한 데이터에 적합하지 않음
  - 비선형성 데이터에 취약함



선형



비선형

Mathematical Expression

- Linear Regression

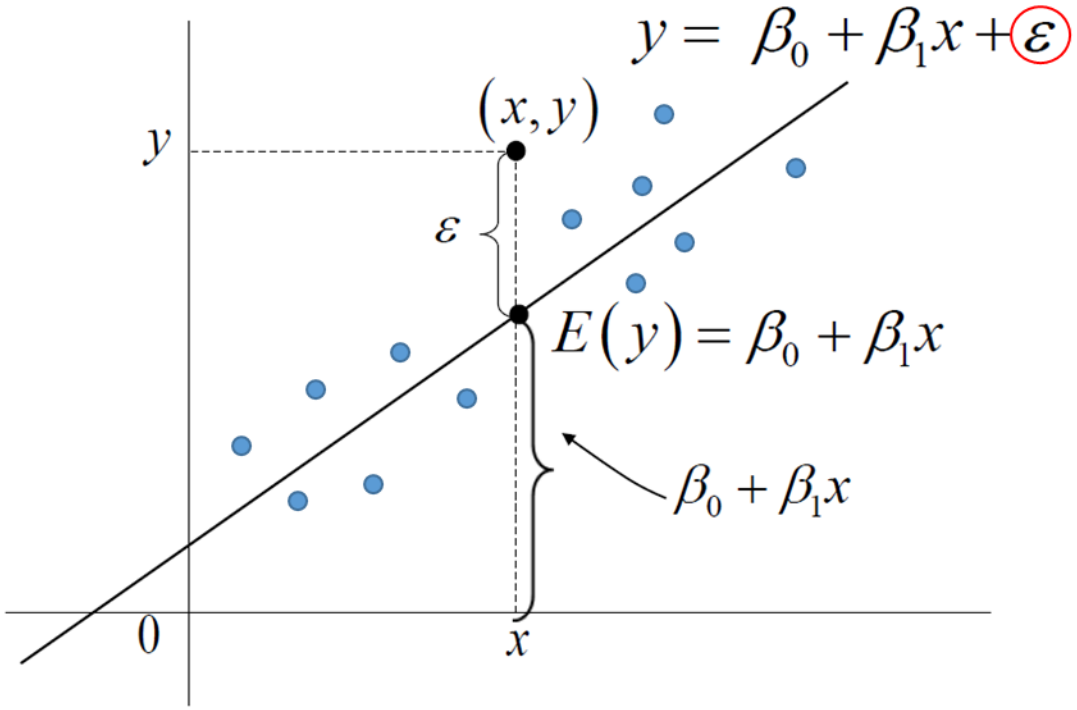
$$E(Y_i) = \beta_0 + \beta_1 x_i$$

Dependent Variable (DV)

Constant (상수)

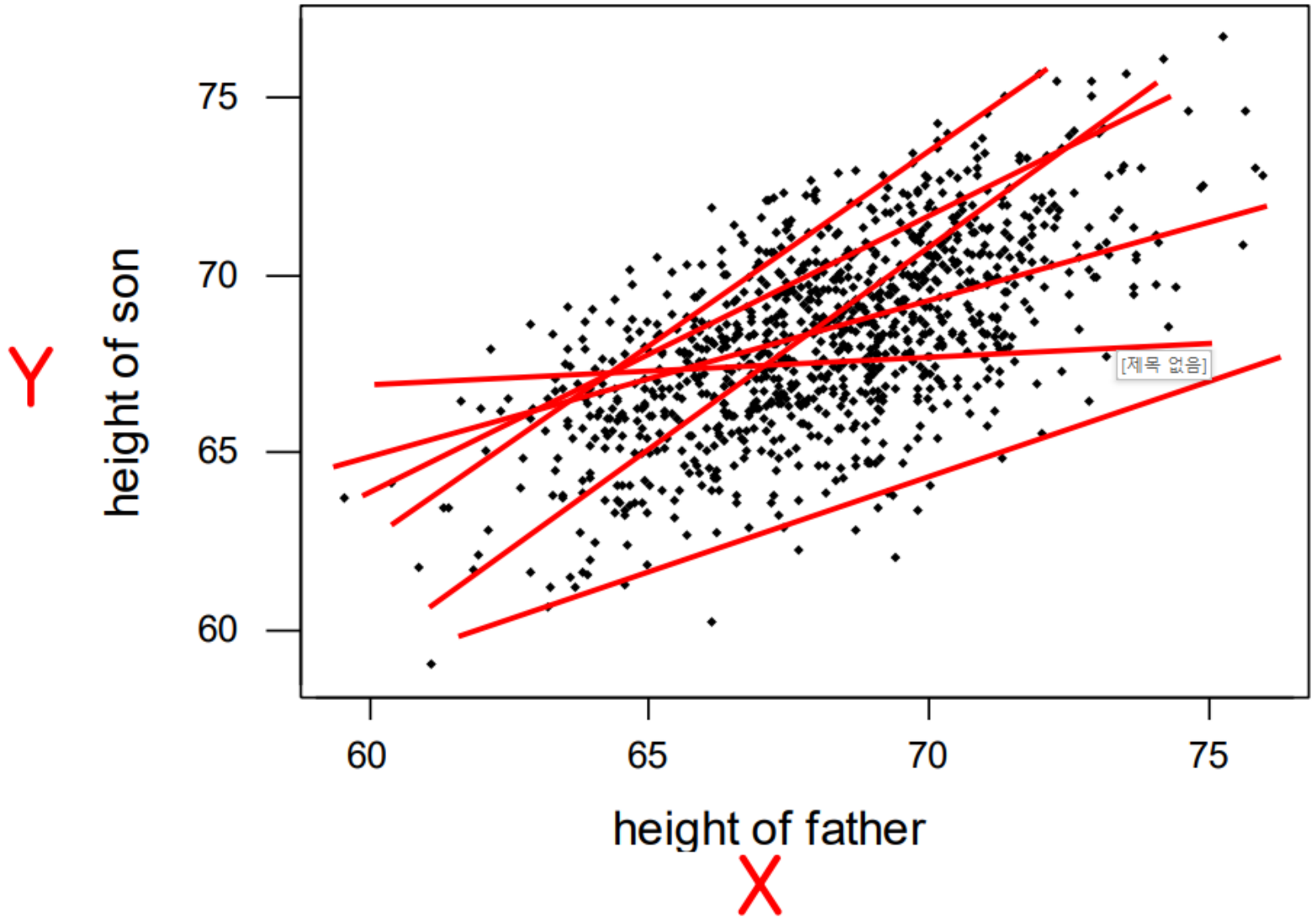
Coefficient (계수)

X가 1단위 증가했을 때, Y에 미치는 영향도



Mathematical Expression

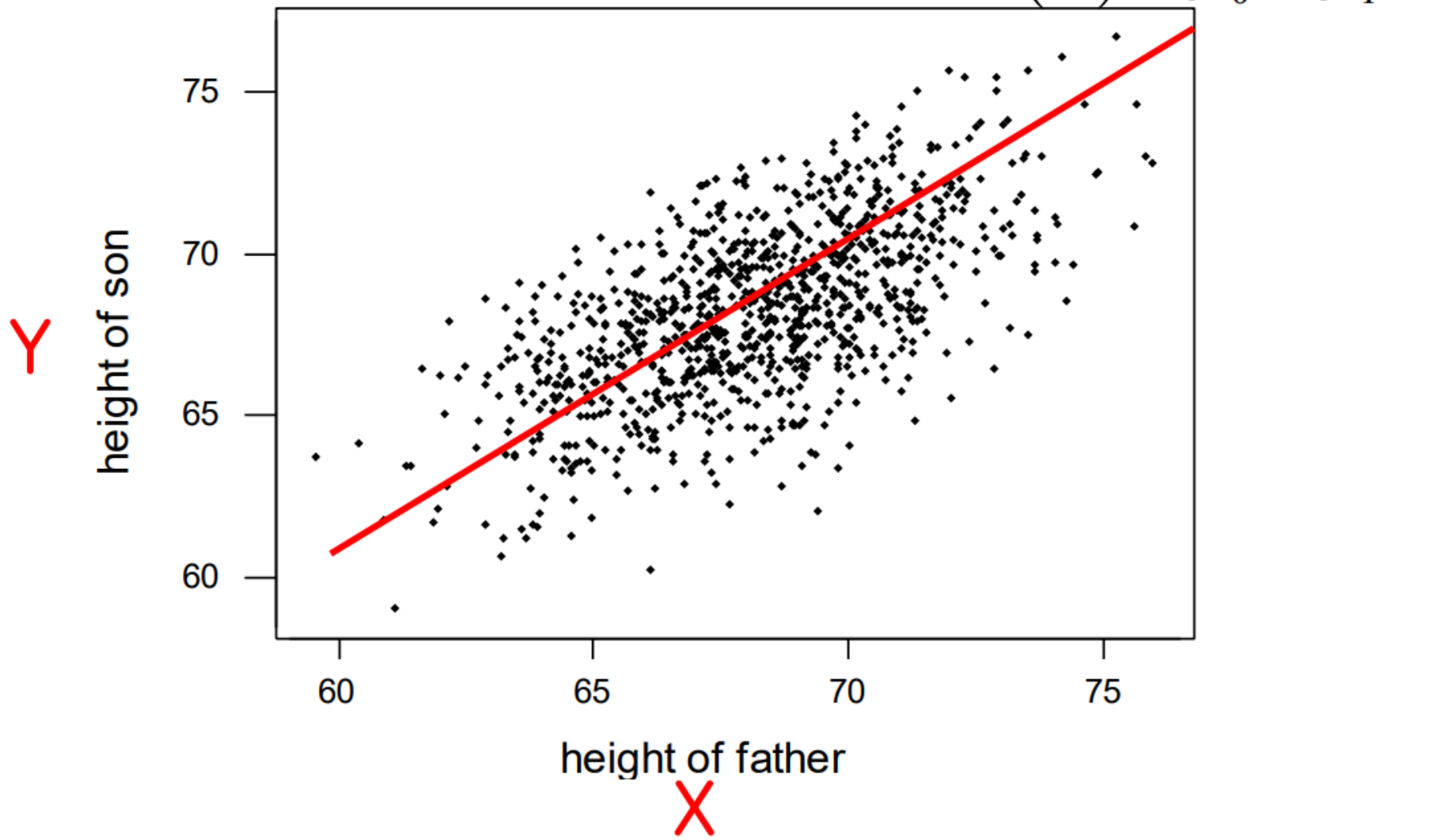
- Linear Regression



Mathematical Expression

- Linear Regression

$$E(Y) = \beta_0 + \beta_1 X$$



Mathematical Expression

- Multi-Linear Regression

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

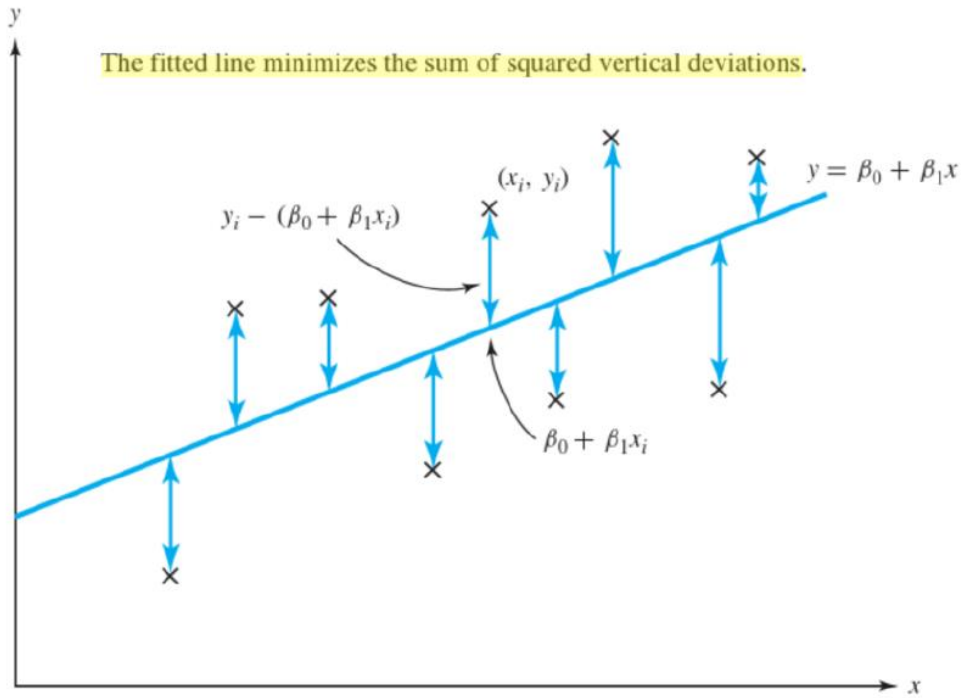
$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

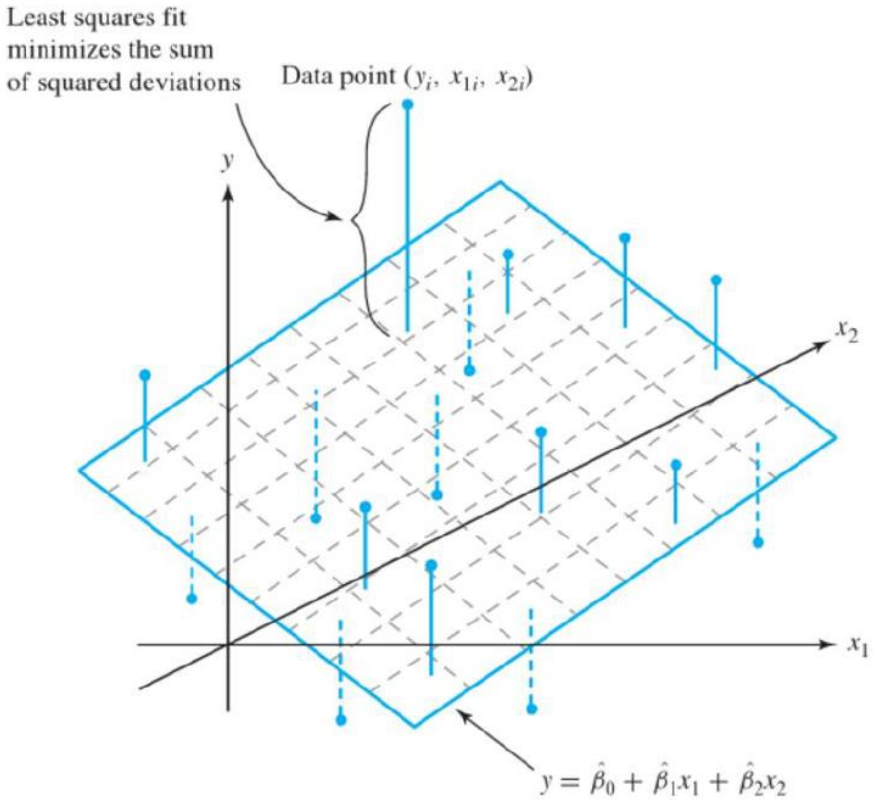
Mathematical Expression

- Simple & Multi-Linear Regression

$$\varepsilon_i = y_i - E(y_i) = y_i - (\beta_0 + \beta_1 x_i) = y_i - \beta_0 - \beta_1 x_i$$

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$





Mathematical Expression

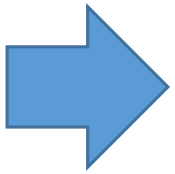
- Simple & Multi-Linear Regression Estimation

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$$

$$Minimize\,Q(\beta_0,\beta_1)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$$

$$\left.\frac{\partial Q(\beta_0,\beta_1)}{\partial\beta_0}\right|=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0$$

$$\left.\frac{\partial Q(\beta_0,\beta_1)}{\partial\beta_1}\right|=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)x_i=0$$



$$\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X},$$

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

$$\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$$

Mathematical Expression

- Simple & Multi-Linear Regression Estimation
  - β 검증
    - β에 대한 P-Value가 낮으면 기울기가 0이 아닌것으로 됨
      - 통상적으로 P-value가 0.05이하면 의미 있다고 판단함
    - 즉 H_0 은 기각 되며 H_1 이 채택 됨

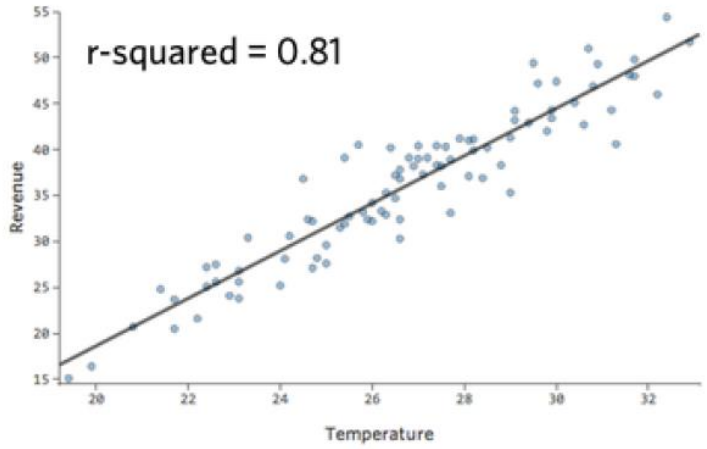
$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0$$

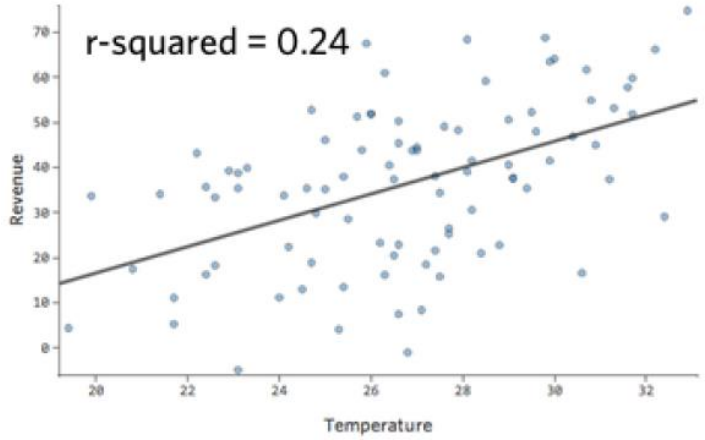
$$t^* = \frac{\hat{\beta}_1 - 0}{sd\{\hat{\beta}_1\}}$$

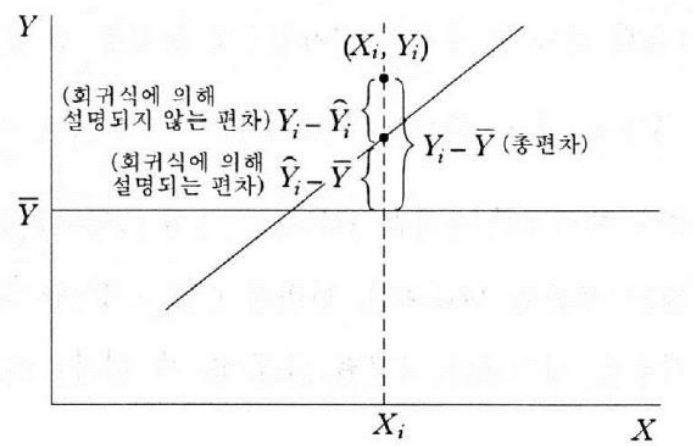
$$\text{If } |t^*| > t_{\alpha/2,n-2}, \text{ we reject } H_0$$

Mathematical Expression

- Simple & Multi-Linear Regression Estimation
  - 회귀모델의 정성적 적합도 판단
    - R^2 은 평균으로 예측한 것에 대비 분산을 얼마나 축소 시켰는지에 대한 판단
      - 보통은 아래의 수식과 달리 Correlation $(y, \hat{y})^2$ 으로 표현함
  - 과연 R^2 가 어느 정도 수치일 때 쓸만한 모델일까?







$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \quad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$\boldsymbol{SST = SSR + SSE}$$

Mathematical Expression

- Simple & Multi-Linear Regression Estimation
  - 회귀모델의 정량적 적합도 판단

성능지표 1: 평균오차

- 실제 값에 비해 과대/과소 추정 여부를 판단
- 부호로 인해 잘못된 결론을 내릴 위험이 있음

$$Average\ error = \frac{1}{n}\sum_{i=1}^{n}(y - y') = 0.342$$

Age	Actual Weight(y)	Predicted Weight(y')
1	5.6	6.0
2	6.9	6.4
3	10.4	10.9
4	13.7	12.4
5	17.4	15.6
6	20.7	21.5
7	23.5	23.0

Mathematical Expression

- Simple & Multi-Linear Regression Estimation
  - 회귀모델의 정량적 적합도 판단

성능지표 2: 평균 절대 오차(Mean absolute error; MAE)

- 실제 값과 예측 값 사이의 절대적인 오차의 평균을 이용

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y - y'|$$
$$= 0.829$$

Age	Actual Weight(y)	Predicted Weight(y')
1	5.6	6.0
2	6.9	6.4
3	10.4	10.9
4	13.7	12.4
5	17.4	15.6
6	20.7	21.5
7	23.5	23.0

Mathematical Expression

- Simple & Multi-Linear Regression Estimation
  - 회귀모델의 정량적 적합도 판단

성능지표 3: Mean absolute percentage error (MAPE)

- 실제값 대비 얼마나 예측 값이 차이가 있는지를 %로 표현
- 상대적인 오차를 추정하는데 주로 사용

$$MAPE = 100\% \times \frac{1}{n} \sum_{i=1}^{n} \frac{|y - y'|}{|y|}$$

$$= 6.43\%$$

Age	Actual Weight(y)	Predicted Weight(y')
1	5.6	6.0
2	6.9	6.4
3	10.4	10.9
4	13.7	12.4
5	17.4	15.6
6	20.7	21.5
7	23.5	23.0

Mathematical Expression

- Simple & Multi-Linear Regression Estimation
  - 회귀모델의 정량적 적합도 판단

성능지표 4 & 5: (Root) Mean squared error ((R)MSE)

- 부호의 영향을 제거하기 위해 절대값이 아닌 제곱을 취한 지표

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - y')^2$$

$$= 0.926$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - y')^2}$$

$$= 0.962$$

Age	Actual Weight(y)	Predicted Weight(y')
1	5.6	6.0
2	6.9	6.4
3	10.4	10.9
4	13.7	12.4
5	17.4	15.6
6	20.7	21.5
7	23.5	23.0

Q & A