◎ 회귀 분석 Linear Regression

- 변수 간에 선형적 인과 관계가 있는지 분석.
- 반응 변수가 연속형인 경우 Y
- 설명 변수는 연속형과 범주형 모두 가능

진행 회귀

◎ linear regression 의 유형

종속 변수, 독립 변수 ⇒ 이산형 or, 연속형

- 단순 선형 회귀 모형

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$ 오차

우리 관심사

평가-Y

반응변수 / 설명변수 ⇐ 반드시 연속형

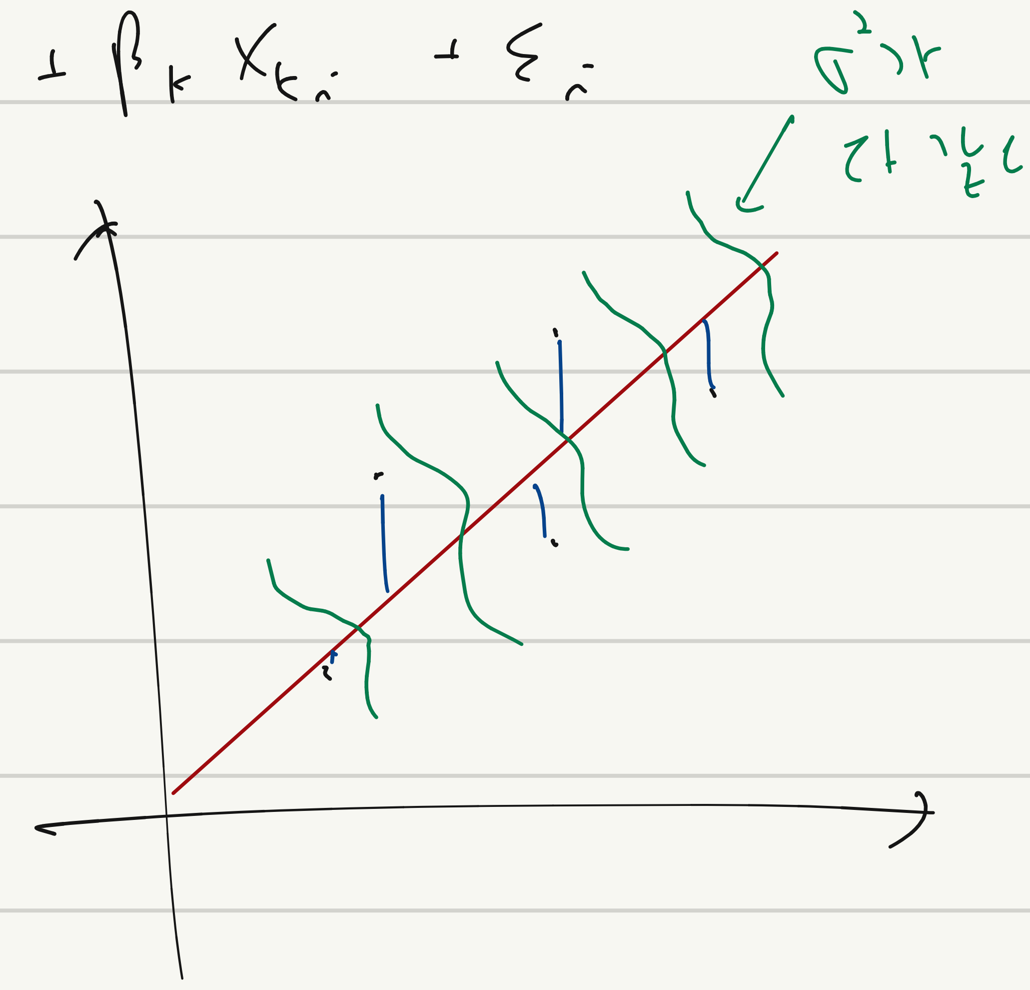
단순 회귀분석 Simple LR

다중 Multiple LR

- 다중 선형 회귀 모형

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

σ^2가 다 같다



- 선형 회귀 모형의 가정
  - $\varepsilon_i \overset{iid}{\sim} N(\beta^T x, \sigma^2), \ \forall i$

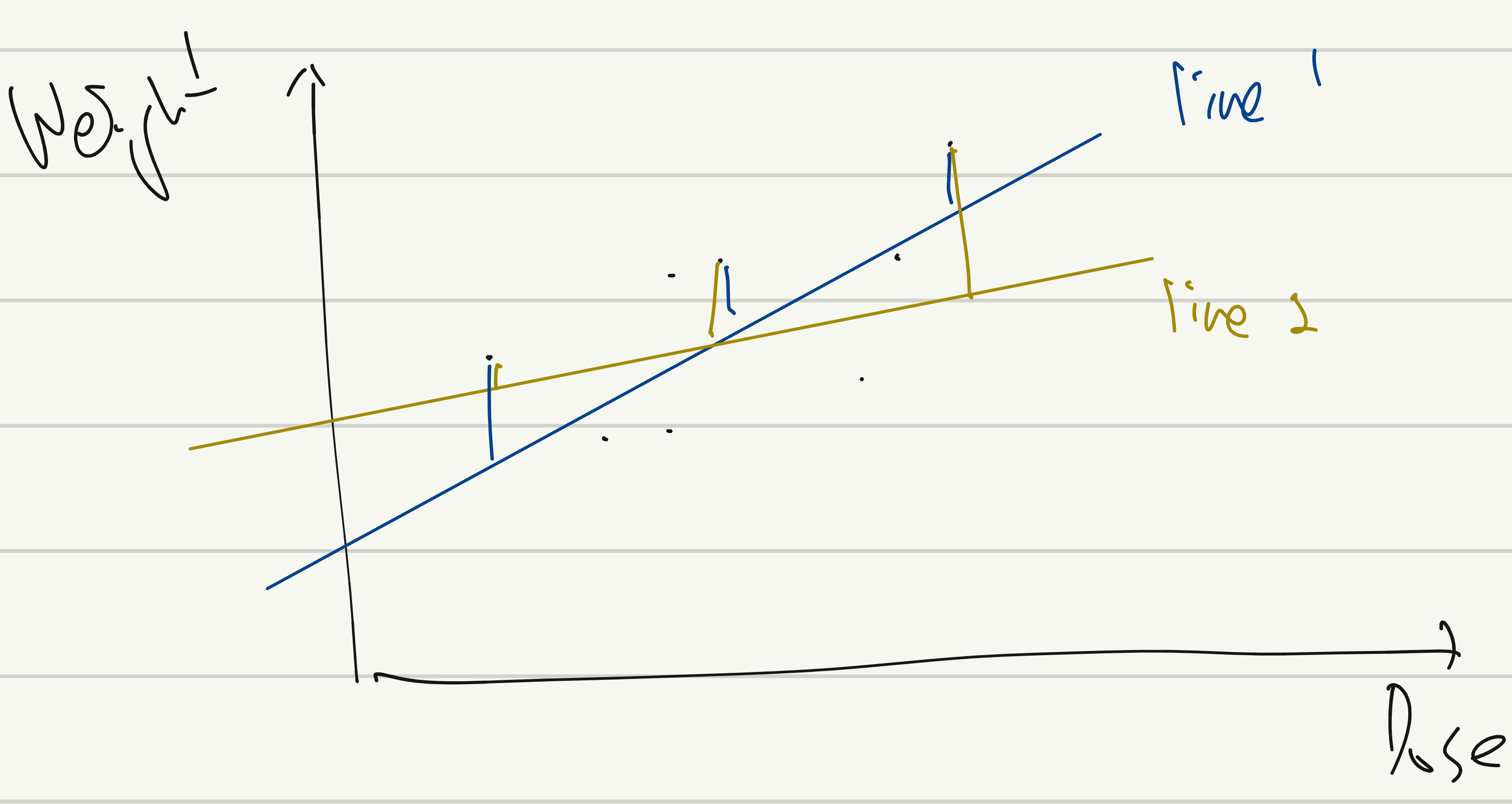
    where, $\beta^T x = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$

  * 대부분의 data가 선 근처에 존재한다.

  ⇒ data가 선형적 pattern을 가지고 있다.

  - 독립성, 정규성, 등분산성.

◎ 회귀 계수의 추정 (How to fit regression line)



$e_1^2 + e_2^2 + e_3^2$

$e_1^2 + e_2^2 + e_3^2$

$\sum e_i^2 < \sum e_i^2$

$$\min\left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right)$$ ⇒ 최소 제곱 추정법.

$$\min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \longrightarrow \beta_0 \text{ 와 } \beta_1 \text{ 을 찾자}$$

⊚ 회귀계수 추정량.

$$\hat{\beta}_0, \hat{\beta}_1 = \min_{\beta_0, \beta_1} \sum e_i^2 = \min \sum (y - \hat{y})^2$$

$$= \min \sum (y_i - \beta_0 - \beta_1 x)^2$$

· parameter estimation

or

optimization.

미분해서

0이되는값

$$\frac{\partial}{\partial \beta_0} \sum (y_i - \beta_0 - \beta_1 x)^2$$

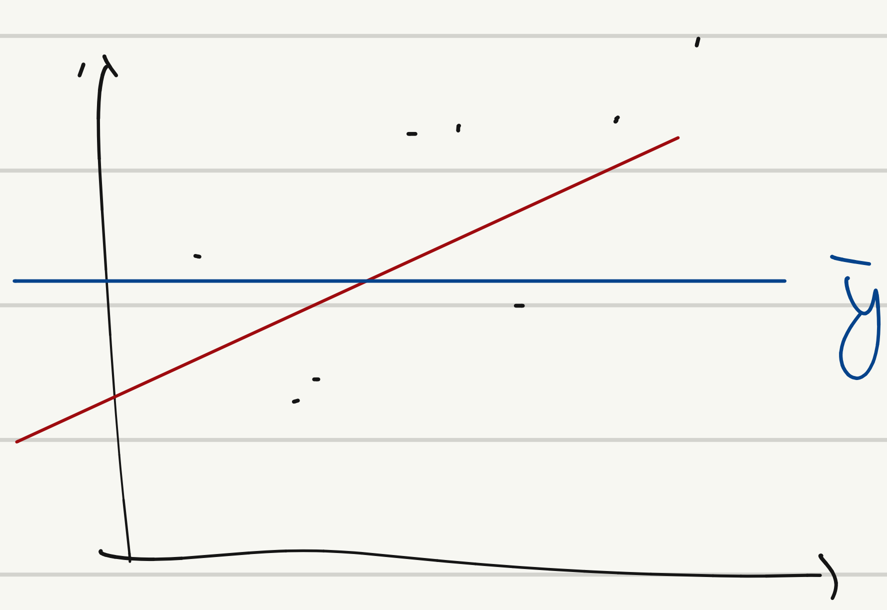
$$\frac{\partial}{\partial \beta_1} \sum (y_i - \beta_0 - \beta_1 x)^2$$

· SSE (Sum of squares of errors) $= \sum e_i^2$

$$\frac{\partial SSE}{\partial \beta_0} = 0, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial SSE}{\partial \beta_1} = 0, \quad \hat{\beta}_1 = \frac{\overbrace{\sum (x_i - \bar{x})(y_i - \bar{y})}^{S_{xy}}}{\underbrace{\sum (x_i - \bar{x})^2}_{S_{xx}}}$$

⊚ 총 변동의 분할.



$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2$$

$$+ 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$H_1: \beta_1 \neq 0$

$H_0: \beta_1 = 0$

Z value 자체 기각시켜야

X의 의미가 생긴다.

회귀직선 Regression.

$$\frac{\sum (\hat{y}_i - \bar{y})^2 / 1}{\sum (y_i - \hat{y}_i)^2 / n-2} \sim F_{(1, n-2)} = \underbrace{\frac{MSR}{MSE}}$$

이 값이 크면, data가

regression 직선에 밀집해 있다.