회귀 분석(Regression Analysis)

류영표

목차

- 회귀분석이란?

회귀 분석이란

- 회귀 분석(Regression Analysis)은 통계학에서 관찰된 연속형 변수들에 대해 독립변수와 종속변수 사이의 인과관계에 따른 수학적 모델인 선형적 관계식을 구하여 어떤 독립변수가 주어졌을때 이에 따른 종속 변수를 예측한다. 또한 이 수학적 모델이 얼마나 잘 설명하고 있는지를 판별하기 위한 적합도를 측정하는 분석방법이다. -위키백과

선형 회귀(linear regression)

- 회귀는 현대 통계학을 이루는 큰 축
- 회귀 분석은 유전적 특성을 연구하던 영국의 통계학자 갈톤(Galton)이 수행한 연구에서 유래했다는 것이 일반론.

“부모의 키가 크더라도 자식의 키가 대를 이어 무한정 커지지 않으며, 부모의 키가 작더라도 대를 이어 자식의 키가 무한정 작아 지지 않는다.’

- 회귀 분석은 이처럼 데이터 값이 평균과 같은 일정한 값으로 돌아가려는 경향을 이용한 통계학 기법.

회귀(Regression) 개요

- 회귀는 여러 개의 독립변수와 한 개의 종속변수 간의 상관관계를 모델링 하는 기법을 통칭

아파트의 가격은 ?	방 개수	아파트 크기	주변 학군	근처 지하철 역 갯수

$$Y = W_1X_1 + W_2X_2 + W_3X_3 + \cdots + W_nX_n$$

Y는 종속변수, 아파트 가격

$X_1, X_2, X_3 \cdots, X_n$ 은 방 개수, 아파트 크기, 주변 학군등의 독립변수

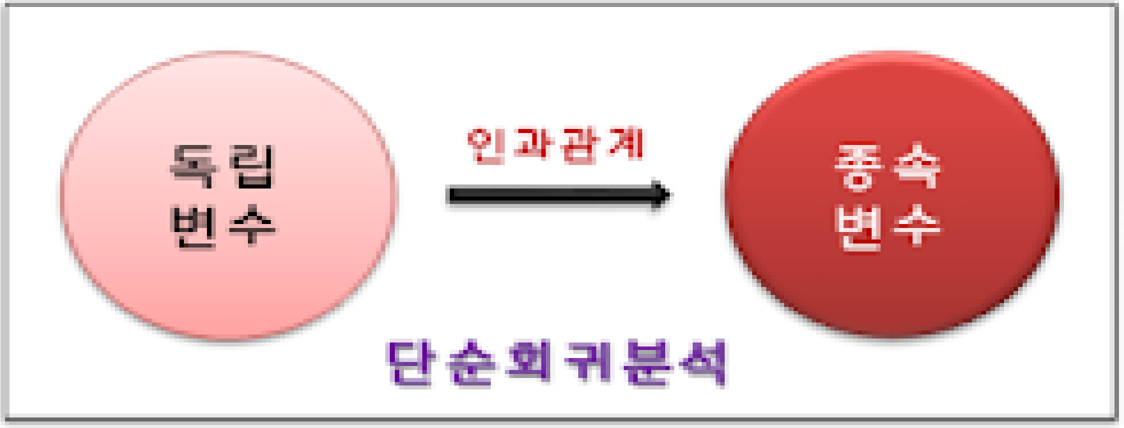
$W_1, W_2, W_3 \cdots, W_n$ 은 이 독립변수의 값에 영향을 미치는 회귀 계수(Regression coefficients)

머신러닝 회귀 예측의 핵심은 주어진 피처와 결정 값 데이터 기반에서 학습을 통해 최적의 회귀 계수를 찾아내는 것

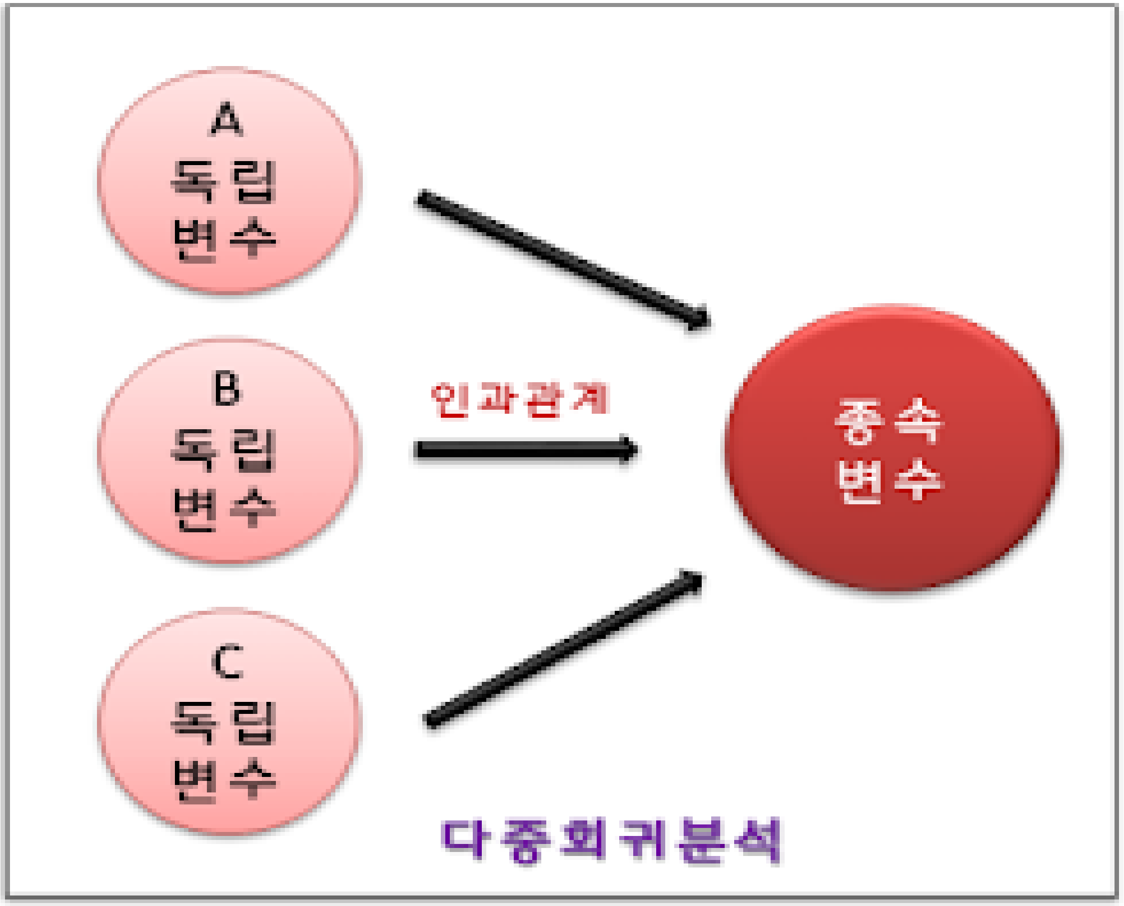
회귀의 유형

변수의 비교

A
변수
선형관계
B
변수
상관분석



독립
변수
인과관계
종속
변수
단순회귀분석



A
독립
변수
B
독립
변수
인과관계
종속
변수
C
독립
변수
다중회귀분석

ㅣ회귀분석이란

- 지도 학습(supervised learning)

  $\boldsymbol{Y} = \boldsymbol{f}\ (\boldsymbol{X})$ 에 대하여 입력 변수 (X)와 출력 변수 (Y) 의 관계에 대하여 모델링하는것

  (Y에 대하여 예측 또는 분류하는 문제)

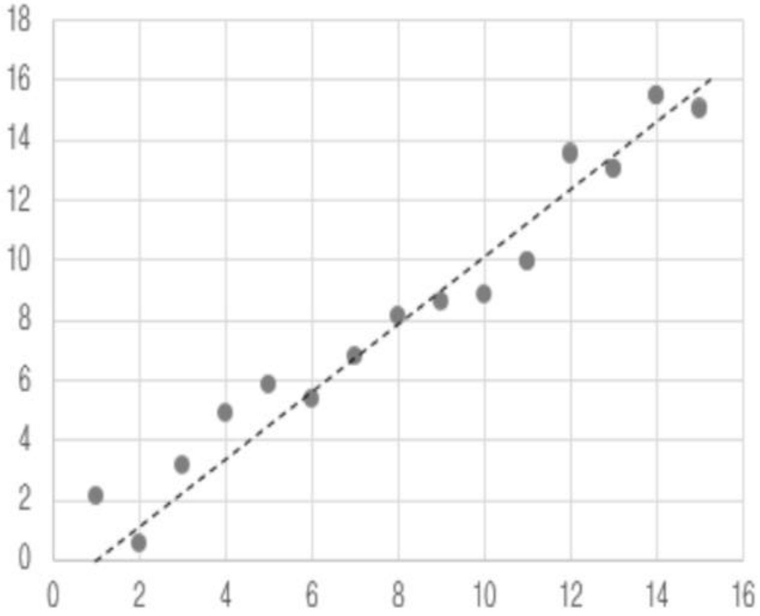
  - 회귀 (regression): 입력 변수 X에 대해서 연속형 출력 변수 Y를 예측
  - 분류 (classification): 입력 변수 X에 대해서 이산형 출력 변수 Y(class)를 예측

$$\underbrace{\boldsymbol{Y}}_{\text{연속형 변수}} = \overbrace{\boldsymbol{f}}^{\text{회귀 모형}}(\underbrace{\boldsymbol{X}}_{\text{연속/이산형 변수}})$$

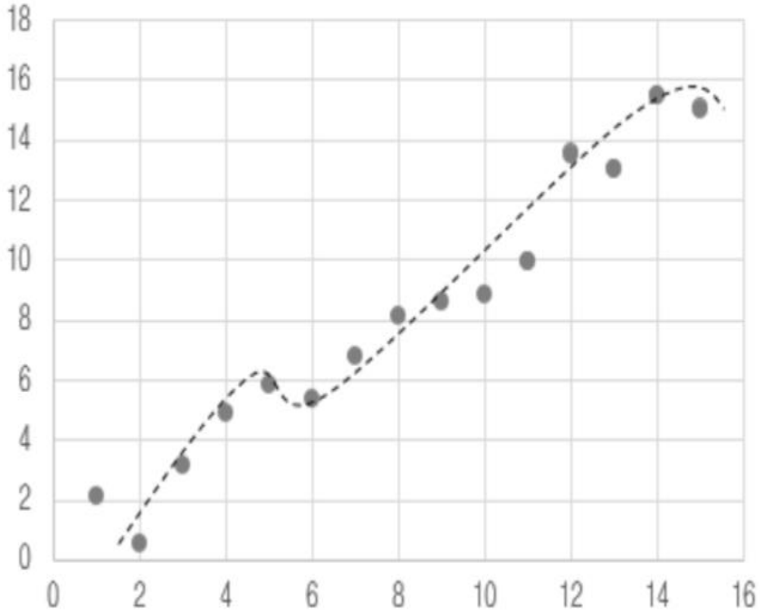
$$\underbrace{\boldsymbol{Y}}_{\substack{\text{이산형 변수} \\ \text{(클래스)}}} = \overbrace{\boldsymbol{f}}^{\text{분류 모형}}(\underbrace{\boldsymbol{X}}_{\text{연속/이산형 변수}})$$

▪ 회귀분석

- 입력 변수인 X의 정보를 활용하여 출력 변수인 Y를 예측하는 방법
- 회귀분석 중 간단한 방법으로는 선형회귀분석(좌측 그림)이 있으며, 이를 바탕으로 더 복잡한 회귀분석(우측 그림)이 개발



선형회귀



비선형회귀

대부분의 분류모델(SVM, Decision Tree 등)로도 회귀가 가능함.

선형회귀의 종류

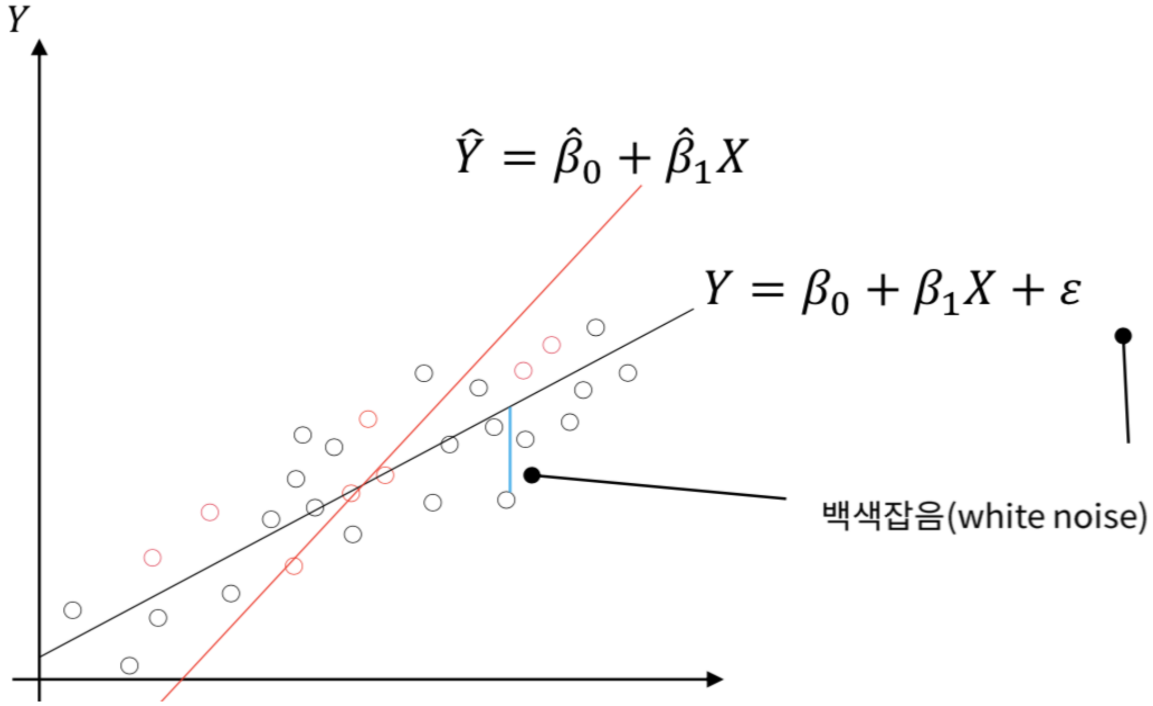
- 일반 선형 회귀 : 예측값과 실제 값의 RSS(Residual Sum of Squares)를 최소화할 수 있도록 회귀 계수를 최적화 하여, 규제(Regularization)를 적용하지 않은 모델
- 릿지(Ridge) : 릿지 회귀는 선형회귀에 L2 규제를 추가한 방식
- 라쏘(Lasso) : 라쏘 회귀는 선형 회귀에 L1 규제를 추가한 방식
- 엘라스틱넷(ElasticNet) : L2,L1 규제를 함께 결합한 모델
- 로지스틱 회귀(Logistic Regression): 로지스틱 회귀는 회귀라는 이름이 붙어 있지만, 사실은 분류에 사용되는 선형 모델

ㅣ회귀분석이란

- 단순 선형 회귀분석

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- 입력 변수가 X, 출력 변수가 Y일 때, 단순선형회귀의 회귀식은 검은 선으로 나타낼 수 있음
- β_0는 절편(intercept), β_1은 기울기(slope)이며 합쳐서 회귀계수(coefficients)로도 불림



- 검은 점: 모집단의 모든 데이터
- 빨간 점: 학습집합의 데이터
- 실제 β_0와 β_1은 구할 수 없는 계수로 데이터(학습집합)를 통해 이 둘을 추정해서 사용

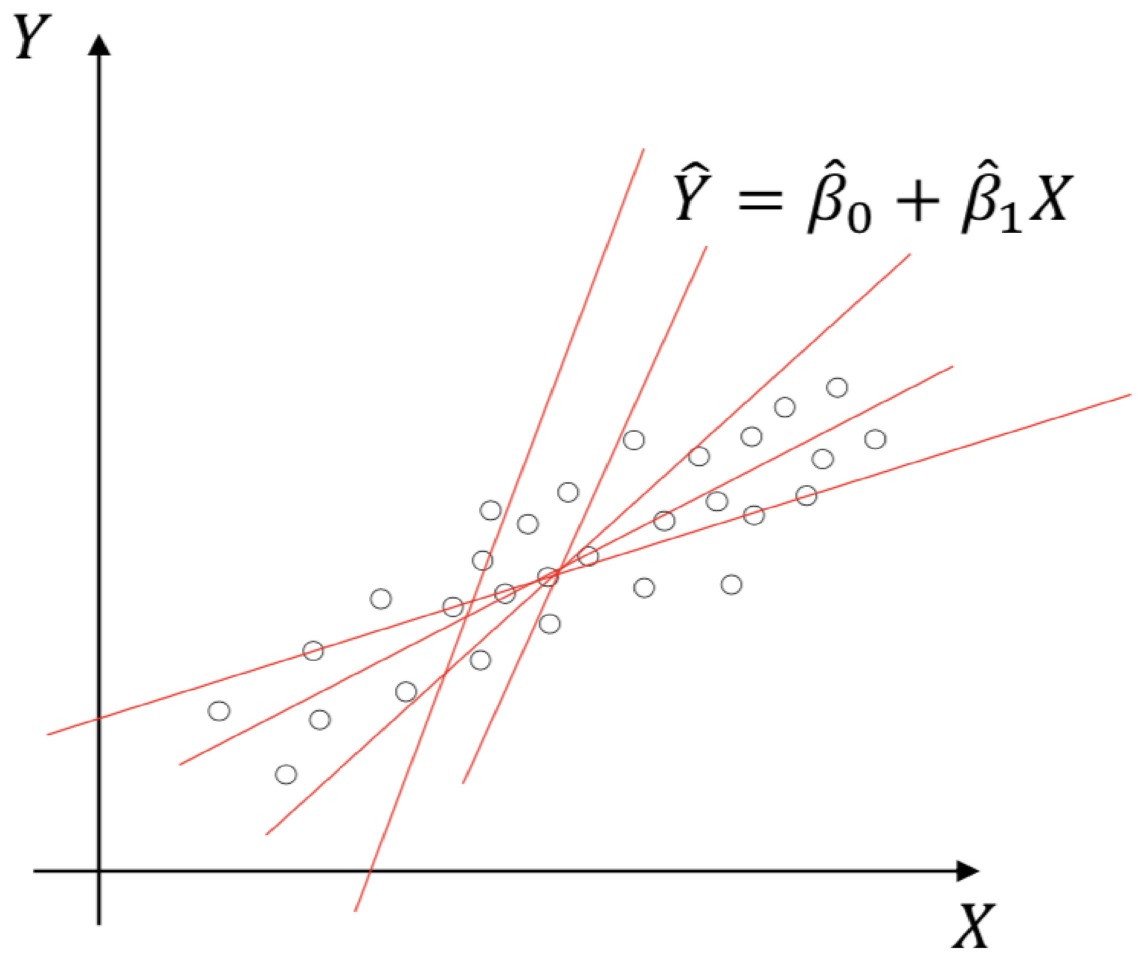
- 가정 :

  $\varepsilon_i \sim i.i.d\ N(0, \sigma^2)$, $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, X와 Y는 선형관계

최적의 회귀 모델을 만든다는 것은 바로 전체 데이터의 잔차(오류 값)) 합이 최소가 되는 모델을 만든다는 의미. 동시에 오류 값 합이 최소가 될 수 있는 최적의 회귀 계수를 찾는다는

- 단순 선형 회귀분석
  - 어떻게 추정 할까?

    - 여러 개의 직선 중 가장 좋은 직선은?



➢ <u>직선과 데이터의 차이가 평균적으로 가장 작아지는 직선</u>

- 어떻게 추정 할까?

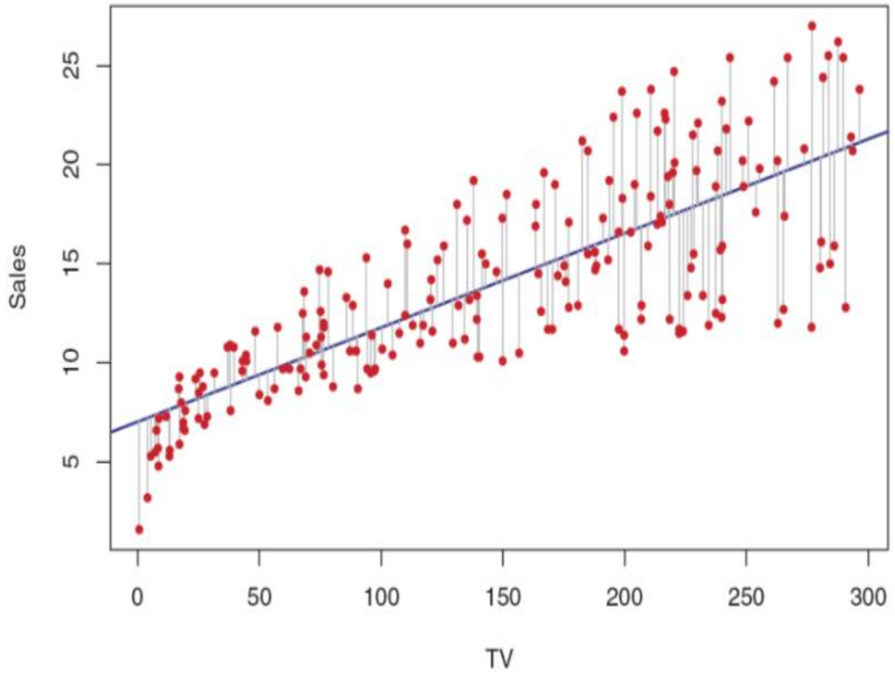
실제 값과 우리가 추정한 값의 차이가 적으면 적을 수록 좋을 것

$$e_i = y_i - \hat{y}_i$$

실제 값과 우리가 추정한 값의 차이를 잔차(residual)라고 하며 이를 최소화 하는 방향으로 추정

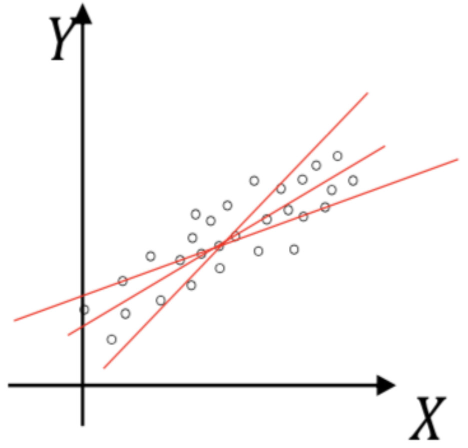
- 잔차를 그림으로 나타내면 오른쪽 그림과 같음
- 잔차의 제곱합(SSE; Error Sum of Squares)는 아래와 같이 표현 가능

$$SSE = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2$$



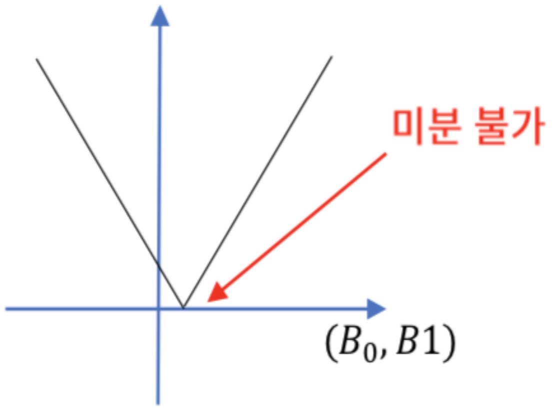
- 굳이 잔차의 제곱합을 최소화 시키는 이유
- 잔차의 합이 0이 되는 해는 무수히 많음 (유일한 해를 찾지 못함)

$$\sum_{i=1}^{n} e_i = e_1 + e_2 + \cdots + e_n = 0$$



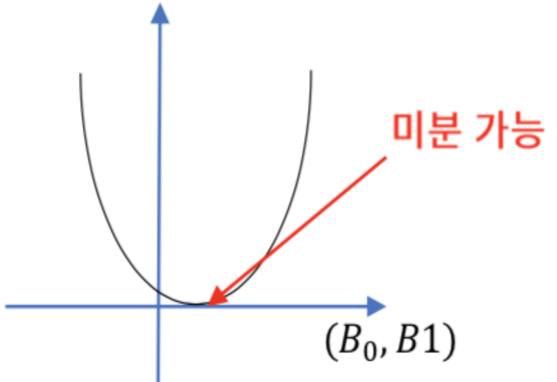
- 잔차의 절대값의 합은 미분이 불가능한 형태

$$\sum_{i=1}^{n} |e_i| = |e_1| + |e_2| + \cdots + |e_n|$$



- 잔차의 제곱 합은 미분이 가능한 형태로 유일한 해를 찾을 수 있음

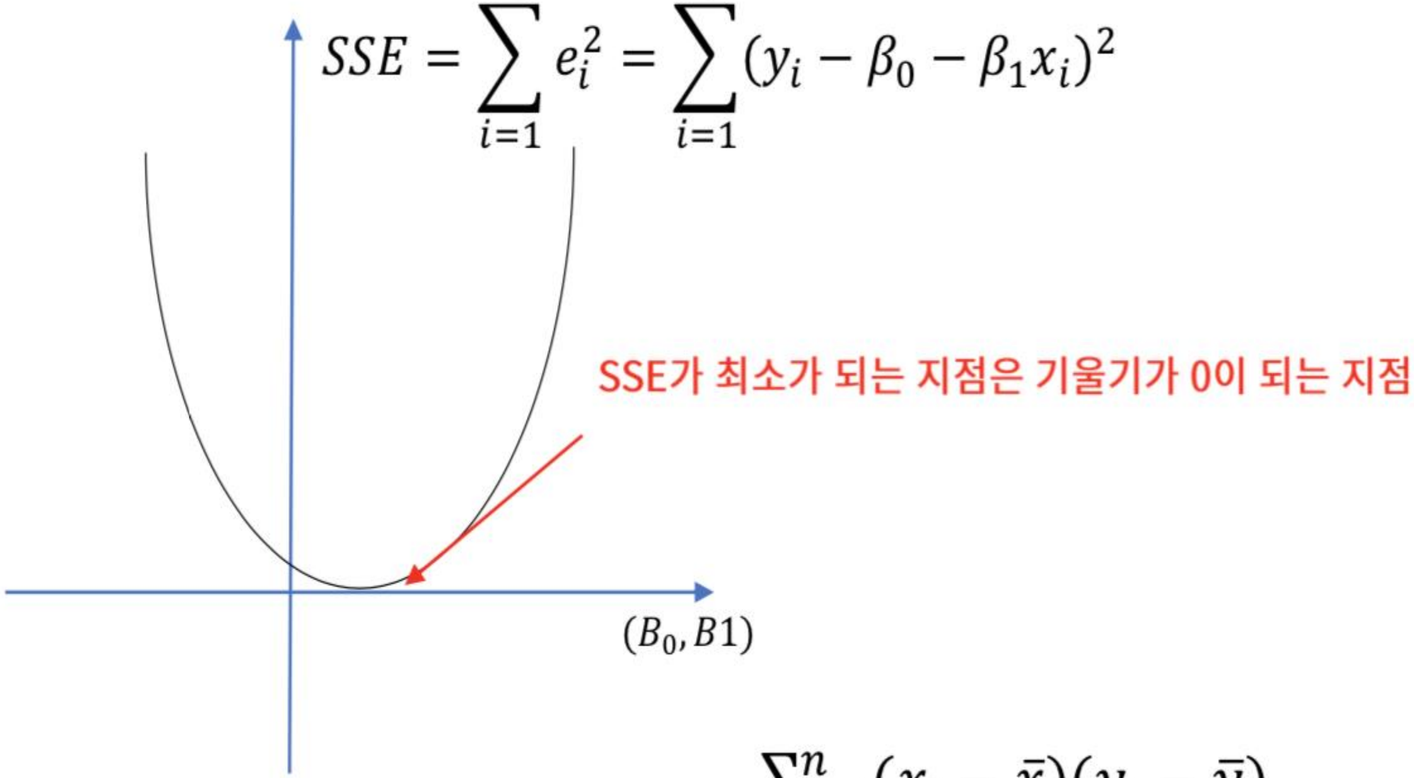
$$SSE = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2$$



회귀 계수의 추정

- SSE $\hat{\beta}_0$과 $\hat{\beta}_1$로 편미분하여 연립방정식을 푸는 방법(Least Square Method)

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

비용 최소화 하기- 경사 하강법(Gradient Descent) 소개

- W 파라미터의 개수가 적다면 고차원 방정식으로 비용 함수가 최소가 되는 W 변수값을 도출할 수 있겠지만, W 파라미터가 많으면 고차원 방정식을 동원하더라도 해결하기가 어려움. 경사 하강법은 이러한 고차원 방정식에 대한 문제를 해결해주면서 비용 함수 RSS를 최소화하는 방법을 직관적으로 제공하는 뛰어난 방식

많은 W 파라미터가 있는 경우에 경사 하강법은 보다 간단하고 직관적인 비용함수 최소화 솔루션 제공

회귀 계수의 추정

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

$$\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - n\beta_0 = 0 \longrightarrow \beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \longrightarrow \therefore \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i\right) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + \beta_1 \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\Rightarrow \beta_1 \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - \beta_1 \sum_{i=1}^{n} {x_i}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i$$

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 회귀 계수의 해석

- $\hat{\beta}_1$의 해석 - X1이 1단위 증가할 때마다 y가 $\hat{\beta}_1$만큼 증가한다.

- 예시) radio광고 예산과 매출 간의 관계
- Radio광고 예산이 1증가 할 때 마다 매출은 0.2단위 만큼 증가한다. 그때의 유의성은 매우 높다.
- Radio광고 예산이 35 단위일 때 예상 매출액은 9.312+0.203*35=16.42단위이다

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \qquad \hat{Y} = 9.312 + 0.203X$$

회귀 계수의 의성을
판단하는 통계치

회귀 계수의 표준 오차

회귀 계수

Simple regression of sales on radio

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	$\boxed{0.203}$	0.020	9.92	< 0.0001

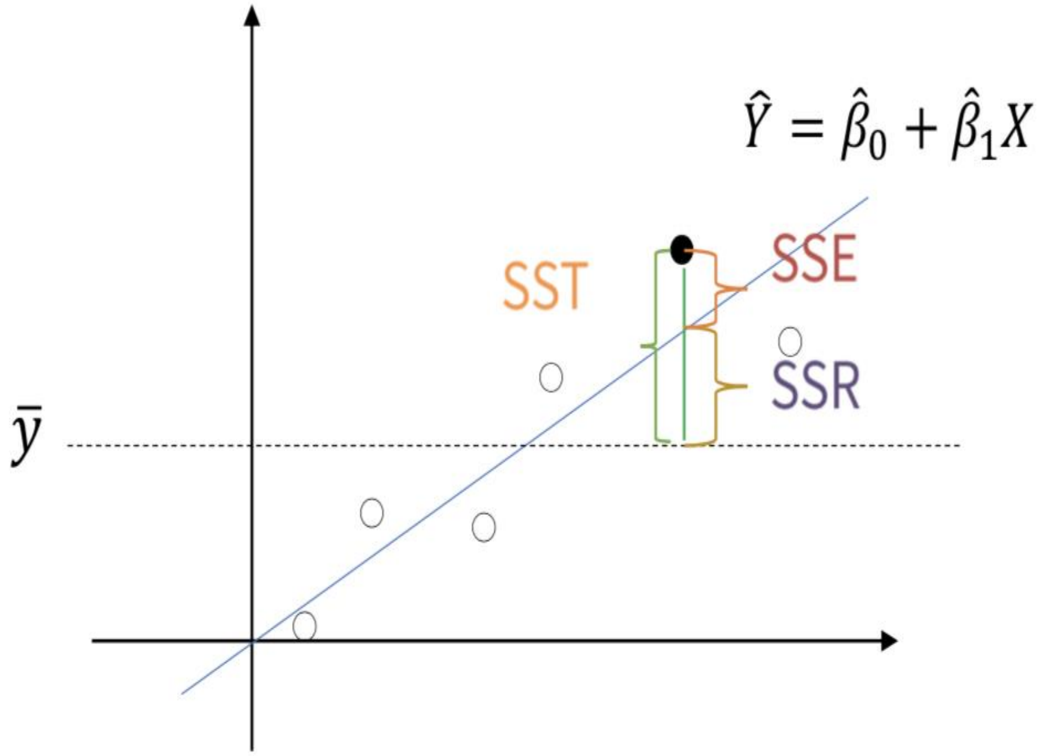
- 선형 회귀의 정확도 평가

  - 선형회귀는 잔차의 제곱합(SSE : Error sum of squares)를 최소화 하는 방법으로 회귀 계수를 추정
  - 즉, SSE가 작으면 작을수록 좋은 모델이라고 볼 수 있음
  - MSE(Mean Squared Error)는 SSE를 표준화한 개념

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

$$MSE = \frac{1}{n-2}SSE$$

▪ 선형 회귀의 정확도 평가



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

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

SST (Total sum of squares) / SSE (Error sum of squares) / SSR (Regression sum of squares)

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

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

※ 회귀 계수 추정

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

Source of Variation	Sum of Squares	Degree of freedom	Mean Square
Regression	SSR	1	SSR
Error	SSE	N-2	MSE
Total	SST	N-1	

▪ 선형 회귀의 정확도 평가

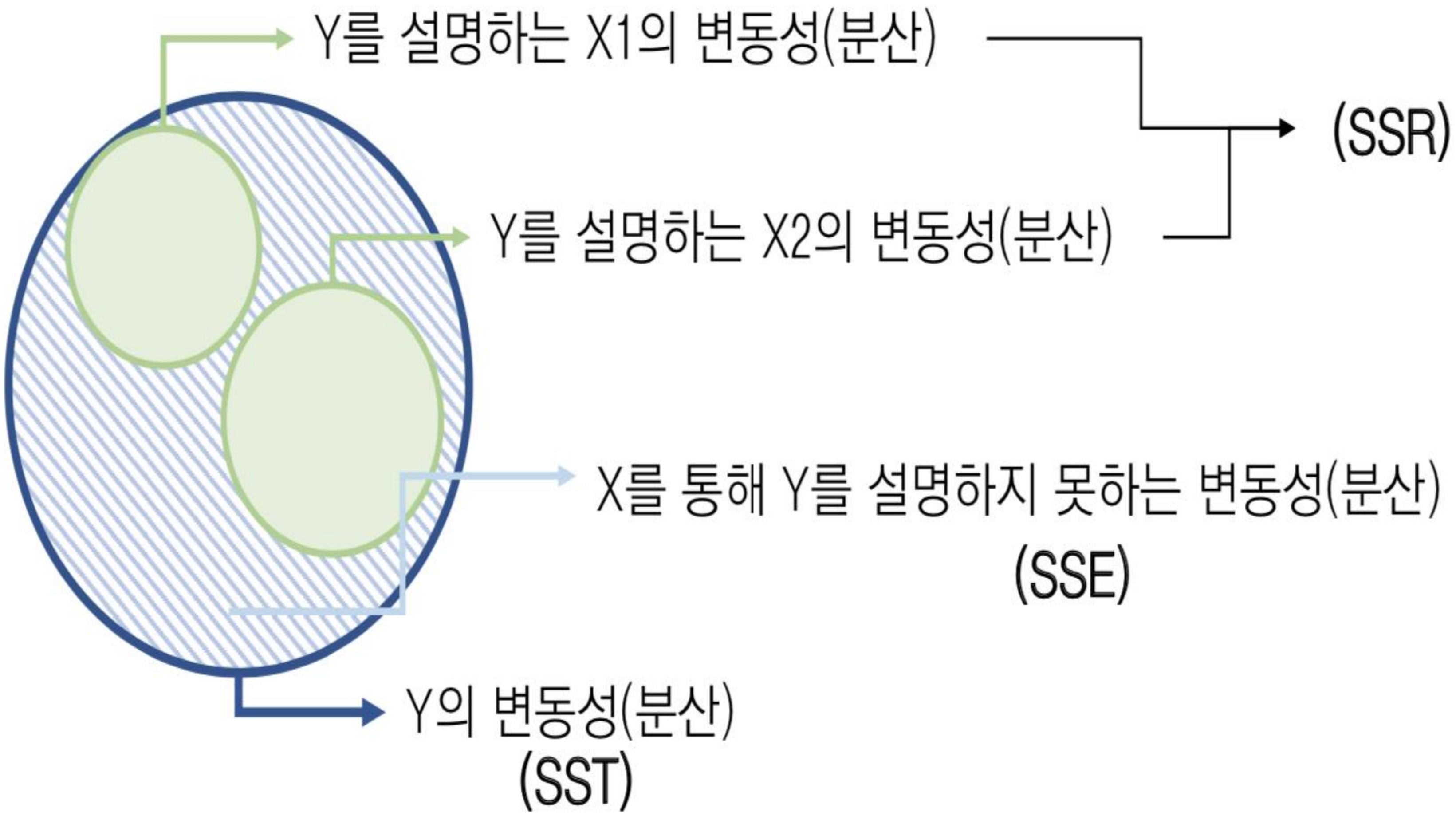
$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSR}$$

- Y의 총 변동은 회귀직선으로 설명 불가능 한 변동과 회귀직선으로 설명 가능한 변동으로 이루어져 있음
- R^2는 RSE의 단점을 보완한 평가지표로 0~1의 범위를 가짐
- R^2은 설명력으로 입력 변수인 X로 설명할 수 있는 Y의 변동을 의미
- R^2이 1에 가까울 수록 선형회귀 모형의 설명력이 높다는 것을 뜻함

입력 변수로 설명할 수 없는 변동 비율

$$R^2 = \frac{SST - SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \quad where\ SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- 선형 회귀의 정확도 평가
  - 회귀 분석은 결국 Y의 변동성을 얼마나 독립변수가 잘 설명하느냐가 중요
  - 변수가 여러 개일 때 각각 Y를 설명하는 변동성이 크면 좋은 변수 -> p-value자연스레 낮아짐



$$R^2 = \frac{SSR}{SST}$$