Ordinary Least Squares

임재민

1. 단순 선형 회귀

단순 선형 회귀 모델은 다음과 같은 식으로 표현됩니다:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

여기서 y_i는 종속 변수, x_i는 독립 변수, β_0는 상수항, β_1는 추정할 계수, ϵ_i는 오차 항이며 해당 모델의 목표는 잔차제곱의 합 즉, SSE를 최소화하는 것입니다:

$$\mathrm{SSE} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 .$$

SSE가 최소가 되는 지점을 찾기 위해 β_0와 β_1에 대해 미분합니다:

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

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

위 두 방정식을 동시에 풀면 다음과 같은 결과를 얻을 수 있습니다:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x},$$

이 때, $\bar{x}$와 $\bar{y}$는 각각 x_i와 y_i의 평균을 의미합니다.

2. 다중 선형 회귀

다중 선형 회귀 모델은 다음과 같은 식으로 표현됩니다:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

이를 행렬을 사용해 나타내면 다음과 같이 표현됩니다:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

여기서:

- $\mathbf{y}$는 $n \times 1$ 크기의 종속 변수 벡터,

- $\mathbf{X}$는 $n \times p$ 크기의 독립 변수 행렬 (절편을 위한 1의 열 포함),
- $\boldsymbol{\beta}$는 $p \times 1$ 크기의 계수 벡터,
- $\boldsymbol{\epsilon}$는 $n \times 1$ 크기의 오차 벡터.

목표는 단순 선형 회귀와 동일하게 SSE를 최소화하는 것입니다:

$$J(\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

SSE가 최소가 되는 지점을 찾기 위해 $\boldsymbol{\beta}$에 대해 미분하고 0으로 설정합니다:

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

이를 정리하면 Normal Equation을 얻을 수 있습니다:

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}.$$

이 후, $\mathbf{X}^T\mathbf{X}$가 역행렬이 가능하다는 전제 하에, 식을 $\boldsymbol{\beta}$에 대하여 풀면 다음과 같습니다:

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$