Chapter 03. 선형회귀

3.1 단순선형회귀

- 하나의 설명변수 X에 기초하여 양적 반응변수 Y를 예측한다.
- X와 Y 사이에 선형적 상관관계가 있다고 가정한다. 수학적으로 다음과 같이 나타낸다.

$$Y \approx \beta_0 + \beta_1 X$$

- 기호 ≈는 "근사적으로 모델링된"이라는 뜻이다.
- β_0와 β_1는 계수(coefficient) 또는 파라미터(parameter)로, 각각 절편(intercept)과 기울기(slope)를 나타낸다.
- 훈련 데이터를 사용하여 모델 계수에 대한 추정치를 구하면, 다음과 같이 Y의 값을 예측할 수 있다.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

3.1.1 계수 추정

- 실제로 β_0와 β_1은 알려져 있지 않다. 따라서 데이터를 이용하여 계수를 추정해야 한다.
- 다음은 X와 Y 측정값으로 구성된 n개 관측치 쌍을 나타낸다고 하자.

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- 우리의 목적은 추정된 직선이 주어진 n개의 데이터 포인트에 가능한 가깝게 되는 $\hat{\beta}_0$와 $\hat{\beta}_1$을 찾는 것이다.
- 가장 흔하게 사용되는 기법은 **최소제곱법(Least Squares Approach)**이다.
- X의 i번째 값에 기초한 Y의 예측값을 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$라고 하자. 그러면 $e_i = y_i - \hat{y}_i$는 i번째 잔차(residual)을 나타내며, 이는 실제 관측된 Y값과 예측한 Y값의 차이를 의미한다.
- 그리하여 **잔차제곱합(RSS, Residual Sum of Squares)**은 다음과 같이 정의한다.

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

- RSS는 또한 아래와 같이 쓸 수 있다.

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

- 최소제곱법은 RSS를 최소화하는 $\hat{\beta}_0$와 $\hat{\beta}_1$을 선택한다. 이 둘은 아래와 같이 얻어진다. *(proof)*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \overline{x}$$

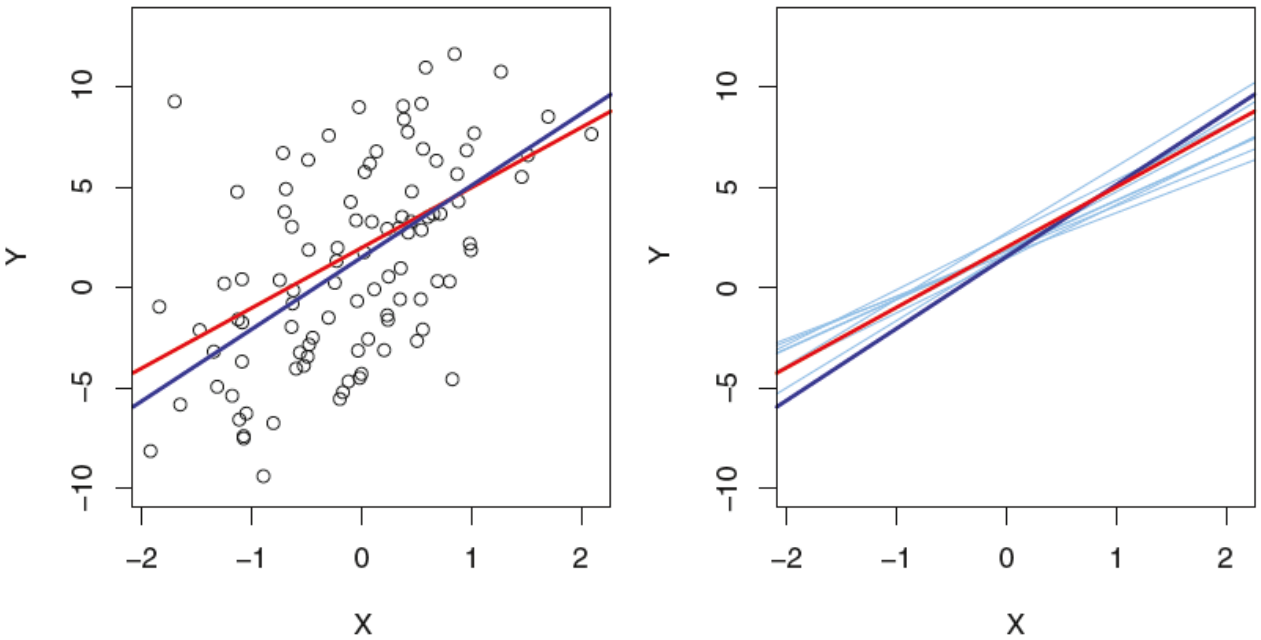
3.1.2 계수 추정값의 정확도 평가

- X와 Y의 실제 상관관계는 $Y = f(X) + \epsilon$의 형태를 가지며, ϵ은 평균이 0인 랜덤오차항이다.
- 오차항 ϵ은 위의 단순한 모델이 수반하는 한계를 보완하기 위한 것으로, X와 독립적이다.

    - X와 Y의 실제 관계는 아마도 선형적이지 않을 수도 있다.
    - Y 값의 변화를 초래하는 다른 변수들이 있을 수 있다.
    - 측정 오차가 있을 수 있다.
- 만약 f가 선형함수로 근사된다면 그 관계는 다음과 같다.

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

- 위 모델은 모회귀선을 정의하며, X와 Y의 상관관계에 가장 잘 맞는 선형근사이다. 하지만 모회귀선은 관측되지 않기에 대신 최소제곱직선을 추정한다. 다시 말해, β_0와 β_1을 각각 $\hat{\beta}_0$와 $\hat{\beta}_1$를 사용하여 추정하는 것이다.
- 이러한 추정은 최소제곱계수추정에 대해 **비편향 성질(Unbiasedness)**이 성립하기 때문에 적절하다.
- 즉, 특정 데이터셋에 대해 $\hat{\beta}_0$ 또는 $\hat{\beta}_1$은 β_0 또는 β_1과 정확히 일치하지는 않을 것이다. 그러나 아주 많은 데이터셋에 대해 얻은 추정값들의 평균을 낼 수 있으면 그와 정확하게 일치할 것이다. *(proof)*
    - 왼쪽 그림: red line = 모회귀선으로 알려진 실제 상관관계 $f(X) = 2 + 3X$, blue line = 최소제곱선
    - 오른쪽 그림: red line = 모회귀선, deep blue line = 최소제곱선, light blue lines = $Y = \beta_0 + \beta_1 X + \epsilon$ 모델을 이용하여 생성한 10개의 서로 다른 데이터셋에 대응하는 10개의 최소제곱선



- 반대로 많은 수의 데이터셋에 대해서는 모회귀선에 근접하지만, 하나의 추정값은 그와 얼마나 다를 것인가?
- **표준오차(SE, Standard Error)**는 추정값 $\hat{\beta}_0$와 $\hat{\beta}_1$이 실제 β_0와 β_1와 어느 정도로 다른지 나타낸다.

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

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

- $\sigma^2 = Var(\epsilon)$이고, σ의 추정치는 잔차표준오차인 $RSE = \sqrt{RSS/(n-2)}$로 구해진다.
- 표준오차는 β_1(또는 β_0)에 대한 **신뢰구간**을 구하는 데 사용될 수 있다. 이러한 범위는 데이터 표본으로부터 계산된 하한값과 상한값으로 정의된다. 즉, 아래의 구간은 대략 95%의 확률로 β_1의 실제값을 포함할 것이다.

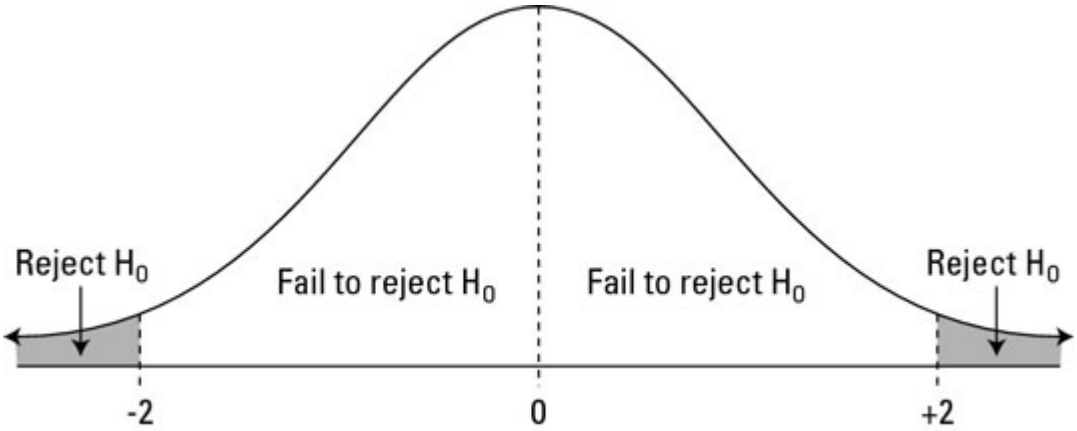
$$[\hat{\beta}_1 - 2 * SE(\hat{\beta}_1), \hat{\beta}_1 + 2 * SE(\hat{\beta}_1)]$$

- 표준오차는 또한 계수들에 대한 **가설검정**을 하는 데 사용될 수 있다.

    - H_0 : X와 Y 사이에 상관관계가 없다. ($\beta_1 = 0$)
    - H_1 : X와 Y 사이에 상관관계가 있다. ($\beta_1 \neq 0$)
- 다음과 같이 t-통계량을 구한다. 이는 귀무가설 하에서 자유도가 n-2인 t-분포를 따른다.

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- p-값은 $\beta_1 = 0$이라고 가정했을 때 어떤 값이 $|t|$와 같거나 큰 경우를 관측할 확률이다. 만약 p-값이 충분히 작으면, 귀무가설을 기각하고 X와 Y 사이에 상관관계가 있다고 결론을 내린다.
- p-값이 작으면 설명변수와 반응변수 사이에 어떠한 실질적인 상관성이 없는데도 우연에 의해 의미있는 상관성이 관측될 가능성이 거의 없음을 나타낸다. 그러므로, p-값이 작으면 상관성이 있다고 유추할 수 있다.
- 아래 그림으로 신뢰구간과 t-분포 및 가설검정 시의 p-값을 확인할 수 있다.



3.1.3 모델의 정확도 평가

- 귀무가설을 기각하고 대립가설을 채택했다면, 모델이 데이터에 적합한 정도를 수량화하고자 할 것이다.
- 선형회귀의 적합성은 잔차표준오차(RSE)와 R^2 통계량을 사용하여 평가한다.

> 잔차표준오차(RSE)

- 잔차표준오차(RSE, Residual Sum of Errors)는 오차항의 표준편차에 대한 추정값으로, 대략 반응변수 값이 실제 회귀선으로부터 벗어나게 될 평균값을 의미한다.

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

- 모델을 사용하여 얻은 예측값이 실제 결과값에 가까우면 RSE는 작을 것이고, 크게 다르면 상당히 클 것이다.

> R^2 통계량

- RSE는 Y의 단위로 측정되므로 적정한 RSE가 무엇인지 항상 명확한 것은 아니다.
- R^2 통계량은 비율의 형태를 취하므로 항상 0과 1 사이의 값을 가지며 Y의 크기와는 무관하다.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$
$$TSS = \sum (y_i - \bar{y})^2$$

- 여기서 TSS(Total Sum of Squares)는 총제곱합으로, 회귀가 수행되기 전 반응변수 Y에 내재하는 변동량으로 생각할 수 있다. 이에 반해 RSS는 회귀가 수행된 후에 설명되지 않고 남아 있는 변동량을 측정한다.

- 그러므로 TSS - RSS는 회귀를 수행함으로써 설명된 반응변수의 변동량을 측정하고, R^2는 X를 사용하여 설명될 수 있는 Y의 변동비율을 측정한다.
- R^2가 1에 가까우면 많은 부분이 회귀에 의해 설명되었다는 것, 0에 가까우면 거의 설명되지 않았다는 것이다.
- 같은 선형상관관계의 측도인 **상관계수(correlation coefficient)**를 사용할 수도 있다.

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

- 단순선형회귀에서 $R^2 = r^2$이다. *(proof)*