Linear Regression

Estimation

Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon \ \text{ where } \ \epsilon \sim N\left(0, \sigma^2\right)$$

- RSS(Residual Sum of Squares)를 minimize하는 방향으로 각각의 β 추정

$$\begin{aligned} RSS &= \left(y_0 - \hat{y_0}\right)^2 + \cdots + \left(y_n - \hat{y_n}\right)^2 \\ &= \left(y_0 - \beta_0 - \beta_1 x_0\right)^2 + \cdots + \left(y_n - \beta_0 - \beta_1 x_n\right)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \end{aligned}$$

- 이러한 방법을 Least Squares 또는 OLS(Ordinary Least Squares)라고 부른다.
- 그리고 이때 각 β_0, β_1에 대한 First Order Condtion을 구하면 다음과 같은 결과를 얻게 된다.

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \end{aligned}$$

Multiple Linear Regression

$$\begin{aligned} &Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \\ &\text{where } \ \epsilon \sim N\left(0, \sigma^2\right) \end{aligned}$$

- 보통 matrix를 이용해 나타내며, 아래와 같다.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_{11} + \cdots + \beta_p X_{1p} \\ \vdots \\ \beta_0 + \beta_1 X_{n1} + \cdots + \beta_p X_{np} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$
$$\Leftrightarrow \ \boldsymbol{Y} = \boldsymbol{XB} + \epsilon$$

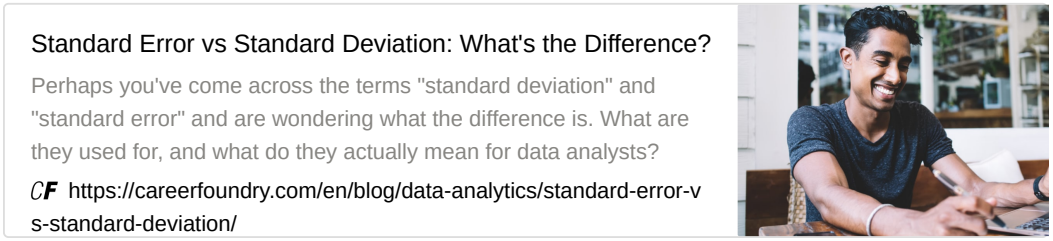
  - 이때 $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}$, and $\boldsymbol{B} \in \mathbb{R}^{p+1}$

- 이 경우 역시 마찬가지로 OLS를 통해 각각의 parameter(또는 coefficient)를 추정하며, 그 결과는 아래와 같다.

$$\hat{\boldsymbol{B}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

Assessment

- 추정된 parameter $\hat{\beta}$의 significance를 판단하는 방법
  - Standard Error의 계산
    - cf) Standard Deviation vs Standard Error



- Standard Deviation (표준 편차)
  - Descriptive statistic; 현재 가지고 있는 표본에 대해서, 그것들이 평균으로부터 대체로 얼마나 멀리 떨어져 있는지에 대한 측정

$$S.D. = \sqrt{Var} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

- Standard Error (표준 오차)
  - Inferential statistic; $i.i.d$를 가정하고 추출된 sample을 사용해 계산된 estimator $\hat{\theta}$가 그 기댓값으로부터 얼마나 떨어져 있는지에 관한 측정

$$S.E. = \sqrt{\mathbb{E}\left(\hat{\theta} - \mathbb{E}\hat{\theta}\right)^2}$$

- 즉, '얼마나 떨어져 있는지'에 관한 측정이라는 점에서 공통점이 있으나, standard deviation은 N개짜리 sample 하나 하나에 대해서 계산되는 값이라면, standard error는 N개짜리 sample을 여러 번 추출해 계산되는 estimator들에 대해서 계산되는 값이라고 할 수 있다.
  - 즉 100개짜리 sample을 100번 추출한다고 생각하면, 각각의 100개짜리 sample에 대해서 표준편차가 구해진다.
  - 그리고 이렇게 만들어진 100개짜리 sample 1세트를 가지고 계산한 표본평균 100개가 구해질 수 있고, 이러한 100개의 표본평균을 사용해 표준오차가 구해진다.

    - cf) Mean Squared Error

        - estimator $\hat{\theta}$가 true parameter에 얼마나 가깝게 추정되는지에 관한 측정
        - 표준오차의 제곱 + bias의 제곱 → unbiased estimator에 대해서는 표준오차의 제곱과 동일
    - Confidence Interval의 계산
        - 95% 신뢰구간은 $\hat{\beta} \pm 2 \cdot SE\left(\hat{\beta}\right)$으로 계산된다.
        - 그리고 이때 95% 신뢰구간의 의미는, N개짜리 sample 100세트를 사용해 $\hat{\beta} \pm 2 \cdot SE\left(\hat{\beta}\right)$의 길이를 갖는 구간을 만들었을 때, 평균적으로 95개의 구간이 true parameter β를 포함한다는 의미이다.
        - 즉 평균적으로 95개의 구간이 true parameter를 포함하도록 자른 구간이 95% 신뢰구간
    - p-value의 계산
        - p-value: H_0가 참인 분포 하에서 sample을 통해 계산된 estimator에 대한 test statistic보다 큰 값이 나올 확률
        - Linear regression에서 사용하는 test statistic은 t-statistic으로 $\frac{\hat{\beta}}{SE\left(\hat{\beta}\right)}$로 계산됨
        - 일반적으로 0.05보다 낮을 경우 significant하다고 한다.
    - Standard error - Confidence interval - p-value의 관계
        - Standard error가 낮을수록 confidence interval이 좁아진다.
            - 즉 추정된 $\hat{\beta}$가 움직일 수 있는 폭이 좁아지고, 따라서 추정된 값이 robust하다고 할 수 있다.
        - Standard error가 낮을수록 t-statistic은 커진다.
            - 따라서 주어진 H_0의 분포에서 계산된 t-statistic보다 큰 값이 나올 확률이 작아지며, p-value가 낮아지게 된다.
            - 그리고 이에 따라서 해당 변수가 유의미하다고 보일 가능성이 높아진다.
- R^2: 각 계수가 아닌 regression 모형 전체에 대한 정확도 측정
    - 개별 target value y_i에 대해서 전체 평균 $\bar{y}$와의 차이인 TSS는 다음과 같이 decompose됨

$$\begin{aligned}\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2 &= \sum_{i=1}^{N}\left(y_i-\hat{y}_i+\hat{y}_i-\bar{y}\right)^2 \\ &= \sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2+\sum_{i=1}^{N}\left(\hat{y}_i-\bar{y}\right)^2\end{aligned}$$

- 여기서 뒷부분은 regression으로 인해 발생한 차이로서, model을 통해 설명되는 부분을 의미
- 그리고 앞부분은 regression을 했음에도 불구하고 발생하는 차이로서, RSS에 해당
- **전체 TSS 가운데 뒷부분의 비중**이 얼마인지, 즉 model을 통해 설명되는 비중에 대한 측정이 R^2

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \\ &= 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \\ &= 1 - \frac{RSS}{TSS} \end{aligned}$$

- R^2가 높으면 높을수록 model의 설명력이 높다고 판단할 수 있음
  - BUT, **R^2는 feature의 개수가 늘어나면 무조건 높아지기 때문**에, 변수의 개수가 많은 경우 해석에 주의가 필요
  - 따라서 model size를 고려하는 Adjusted R^2도 많이 사용됨

- F-statistic
  - Model 자체적인 유의성을 파악할 수 있는 statistic

$$F = \frac{\frac{(TSS-RSS)}{p}}{\frac{RSS}{(n-p-1)}} = \frac{(TSS-RSS)\cdot(n-p-1)}{p\cdot RSS}$$

  - F-statistic이 유의미하면 적어도 하나의 variable은 유의미하다고 할 수 있음

Interpretation

- Linear regression 계수의 해석

  ❗ **Cetris paribus - 다른 모든 변수들이 고정되어 있을 때!!**

  - In general: predictor 1 unit이 증가했을 때 **평균적으로 변화하는 target의 unit**
    - Underlying assumption: 각각의 variable이 target에 미치는 영향이 모두 독립적이다
  - Dummy 변수의 계수: $D = 0$인 group과 비교했을 때 $D = 1$인 group에서 보이는 **target의 평균적인 차이**

$$Y = \beta_0 + \beta_1 \cdot \text{Price} + \beta_2 \cdot D_{\text{Urban}} + \beta_3 \cdot D_{\text{US}} + \epsilon$$

    - $D_{\text{Urban}} = 1, D_{\text{US}} = 1$

        - $y_i = \beta_0 + \beta_1 \cdot \text{Price} + \beta_2 + \beta_3$
    - $D_{\text{Urban}} = 0, D_{\text{US}} = 1$
        - $y_i = \beta_0 + \beta_1 \cdot \text{Price} + \beta_3$
    - $D_{\text{Urban}} = 1, D_{\text{US}} = 0$
        - $y_i = \beta_0 + \beta_1 \cdot \text{Price} + \beta_2$
    - $D_{\text{Urban}} = 0, D_{\text{US}} = 0$
        - $y_i = \beta_0 + \beta_1 \cdot \text{Price}$
    - β_2: Urban=1인 group과 Urban=0인 group에서 보이는 target value의 평균적인 차이
    - β_3: US=1인 group과 US=0인 group에서 보이는 target value의 평균적인 차이
- Interaction Term
    - 각 variable이 target에 미치는 영향이 독립적이지 않은 경우, 두 변수의 상호작용이 target에 미치는 영향을 통제(분리)하기 위함
    - Example: sales target에 대해서 TV와 라디오, 신문 광고의 영향을 파악하고자 할 때, 각각의 변수를 포함한 multiple regression을 사용할 수 있다
        - 그런데 이때 TV 광고가 송출되는 것이 라디오 광고 송출에 영향을 받는 상황을 가정하면, 각 변수 사이의 상호작용을 무시할 경우 해석에 bias가 있을 수 있음
            - 만약 라디오 광고가 TV 광고(효과)에 악영향을 미치는 경우 TV 광고에 붙은 계수는 TV 광고의 영향을 underestimate할 것이며,
            - 반대의 경우 TV 광고에 붙은 계수는 TV 광고의 영향을 overestimate할 것이다.
    - 따라서 두 변수 사이에 상호작용이 있다고 판단되는 경우, 두 변수를 곱한 교차항(interaction term)을 포함함으로써 더욱 정확한 해석이 가능함
    - Correlation vs Interaction term
        - **Correlation → 두 변수 사이에 직접적인 상관관계**
            - 위의 예시의 경우 TV 광고 송출 횟수와 라디오 광고 송출 횟수 사이의 상관관계가 될 것
            - TV 광고 송출을 늘림으로써 라디오 광고 송출 횟수가 줄어들었다면, 둘 사이에는 음의 상관관계가 있음
        - **Interaction → 두 변수가 제3의 변수에 미치는 영향에서 나타나는 관계**
            - 광고 - 매출 사이의 plot을 그린다고 가정하자. x축에는 광고 집행금액, y축에는 매출을 찍는다.
            - 만일 TV 광고 집행금액 - 매출의 그래프와 라디오 광고 집행금액 - 매출의 그래프가 서로 평행하다면, 두 그래프 사이의 차이는 매체의 차이에서 비롯될 뿐 상호작용

이 없다고 판단할 수 있다.

- 그런데 만일 TV 광고 집행금액 - 매출의 그래프와 라디오 광고 집행금액 - 매출의 그래프의 차이가 광고 집행금액이 늘어남에 따라서 서로 **달라진다면**, 둘 사이의 차이는 매체의 차이 이상의 효과가 반영되었다고 볼 수 있다.
  - 그 차이가 증가한다면 TV 광고가 매출에 미치는 영향이 라디오 광고 집행금액이 늘어남에 따라서 늘어난다는 의미(순영향)이며,
  - 그 차이가 감소한다면 TV 광고가 매출에 미치는 영향이 라디오 광고 집행금액이 늘어남에 따라서 줄어든다는 의미(악영향)이다.
- 이때 라디오 광고 집행금액과와 TV 광고 집행금액 사이에는 어떠한 상관관계도 없을 수 있다.
  - 심지어 상호 악영향이 있더라도 직접적인 상관관계는 강한 양의 상관관계를 띨 수 있다.

- 교차항 계수의 해석
  - Interaction between continuous variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

    - X_2가 고정되어 있을 때 X_1의 효과는 이제 $\beta_1 + \beta_3 X_2$로 변화한다. 즉 β_3에 현재 주어진 X_2만큼을 곱한 양이 X_1의 효과에 추가된다는 의미이다.
    - 마찬가지로 X_2의 효과는 $\beta_2 + \beta_3 X_1$이 된다.
    - 따라서 β_3은 계수 절댓값 자체의 의미보다는 그 유의성을 가장 먼저 판단하며, 그 부호(순영향인지 악영향인지)를 그 다음으로 본다.
    - 한편 이제 β_1, β_2의 의미는 “서로가 없을 때(즉, $X_2 = 0, X_1 = 0$, respectively)
      각 X_1, X_2가 Y에 미치는 순수한 효과”가 된다.
  - Interaction between continuous and dummy variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \beta_3 X_1 D + \epsilon$$

    - X_1이 고정되어 있다고 가정하면
      - $D = 1$일 때 Y의 추정값은 $\beta_0 + \left(\beta_1 + \beta_3\right) X_1 + \beta_2$
      - $D = 0$일 때 Y의 추정값은 $\beta_0 + \beta_1 X_1$
    - 따라서 이 둘의 차이 $\beta_2 + \beta_3 X_1$은 $D = 0$인 group에 대비한 $D = 1$인 group의 평균적인 Y의 차이를 의미한다.

- Hierarchy principle in interaction: 두 변수 사이의 interaction을 포함하고 싶은 경우 개별 변수를 모두 포함해야 한다.

- High-order term
  - 만일 Y와 feature X의 관계로부터 비선형성(polynomial term)이 파악되거나 의심되는 경우 해당 variable을 제곱한 항을 모델에 포함시킬 수 있다.
  - Multiple regression에서도 개별 변수 X_j에 대해 Y를 plot했을 때 비선형성이 나타난다면, 해당 변수 X_j의 제곱항을 포함시킬 수 있다.
  - High-order term은 self-interaction term이라고도 볼 수 있으며, 따라서 hierarchy principle에 따라 high-order term을 포함시킬 경우 lower-order term을 모두 포함시키는 것이 일반적이다.
  - 또한, 비선형성이 포함되었음에도 불구하고 mulitple linear regression model이다.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

    - 이때 $\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$로 나타난다. 즉, X가 1 unit 증가할 때 Y의 평균적인 변화분에는 **그때 당시에 주어진 X의 값도 영향을 미친다**는 의미이다.
  - 다만 너무 고차항의 변수를 포함시킬 경우 overfitting이 발생할 수 있다는 점을 주의해야 한다.