통계스터디 3장 : 선형회귀

선형회귀

X가 변할 때 Y가 어떻게 변하는지 선형 관계로 설명하는 모델

선형회귀의 가정

1. 선형성(Linearity) : 예측변수(X)와 반응변수(Y) 사이의 관계가 **직선(선형)으로 표현될 수 있어야 한다.**
2. 등분산성(Homoscedasticity) : 모든 X 값에서 오차의 분산(흩어짐)이 일정해야 함
3. 오차의 독립성(Independence of errors) : 오차들끼리는 서로 **영향을 주지 않아야 함**
4. 오차의 정규성(Nomality of errors) : 오차가 정규분포를 따른다.

단순선형회귀

> $$y = \beta_0 + \beta_1 x$$

⇒ 계수 β_0, β_1를 추정하려면, RSS (오차 제곱합)을 최소화해야 함.

📐 회귀계수 공식

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

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

- $\bar{x}$, $\bar{y}$는 각각 x, y의 **표본 평균**
- 이 식은 **표본에서 구한 값**으로 **모집단 회귀계수**를 추정하는 것

- **왜 추정값을 사용하는가?**
  - 모집단 전체의 회귀선은 알 수 없으므로 표본 데이터로부터 추정함
  - 이때 구한 계수는 추정값 (estimate), 사용하는 공식은 추정량 (estimator)
- **비편향추정량 (Unbiased Estimator)**

    - 어떤 추정량이 '비편향'이라는 것은 여러 표본에서 추정값을 구했을 때, 그 평균이 **모집단의 실제 값**과 같다는 의미
    - 표본평균 $\bar{x}$는 모집단 평균 μ의 비편향추정량
- **단일 추정값의 한계**
    - 단 하나의 표본에서 얻은 추정값은 모집단의 실제 값보다 크거나 작을 수 있음 (과소추정 또는 과대추정)
    - 하지만 추정량이 비편향이라면, 많은 표본에서 얻은 추정값들의 평균은 정확함

표준오차(SE, Standard Error)

단 하나의 표본으로 계산한 추정값은 실제 모집단 값과 다를 수 있는데, 그 추정값이 얼마나 벗어났는지를 수치로 표현한 것 (=추정값이 얼마나 흔들릴 수 있는지 불확실성을 나타내는 값)

- 단순선형회귀에서 회귀계수 β^0, β^1의 표준오차는 다음과 같이 계산된다:

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

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

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

- *$\sigma 2$는 오차항 ε의 분산(즉, Var(ε))으로, 실제로는 알 수 없기 때문에 이를 잔차제곱합(RSS)을 이용해 추정*
    - 오차항은 모델이 설명하지 못하는 y의 변동성

신뢰구간(Confidence Interval)

표본에서 계산한 추정값이 모집단의 진짜 값(모수)을 포함할 것이라고 신뢰할 수 있는 범위

- 표준오차 역할: 이 구간의 폭을 결정하는 핵심 요소로 표준오차가 클수록 구간이 넓어짐
- 신뢰구간 :

$$\hat{\theta} \pm z \times SE(\hat{\theta})$$

여기서 $\hat{\theta}$는 추정값, SE는 표준오차, z는 정규분포 기준값이다.

가설검정(Hypothesis test)

추정값이 귀무가설에서 주장하는 값과 통계적으로 유의미하게 다른지를 판단하는 절차
= 내가 관측한 차이가 우연이라고 보기 어려운가를 수치로 따져보는 것

- 표준오차 역할 : 차이가 얼마나 큰가를 정량적으로 판단할 수 있게 도와주는 기준 단위
- 귀무가설 H_0 : X와 Y 사이에 관계가 없다 (↔ 대립가설 : X와 Y 사이에 관계가 있다.)
  - 결국 $H_0 : \beta_1 = 0$ 이고 $H_1 : \beta \neq 0$ 을 검정해야 하는 것이다.
- t : 이 차이가 표준오차 몇 배만큼 큰가? ($\hat{\beta}_1$와 귀무가설 0 사이에 표준오차가 몇 개나 들어가는가?)

$$t = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$$

  - $\hat{\theta}$: 관측된 추정값
  - θ_0: 귀무가설에서 주장하는 값 (예: 0)
  - $SE(\hat{\theta})$: 표준오차

  - t통계량 : 개별 독립변수 하나가 종속변수에 통계적으로 유의미한 영향을 미치는지 검정
- p-value : t값이 계산되면 그 값이 얼마나 자주 일어나는지를 계산함. (이 정도 차이가 우연히 나올 확률)
  - 작을수록 우연히 나왔을 가능성이 낮다 = 귀무가설을 기각한다.

모형의 정확도 평가

1. **잔차표준오차(RSE, residual standard error)**
   - 평균적으로 예측이 얼마나 빗나갔는지를 나타내는 값

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

- y_i: 실제값
- $\hat{y}_i$: 예측값
- $n-2$: 자유도 (단순선형회귀에서 회귀계수 두 개를 추정했기 때문)

2. R^2

- 회귀모형이 전체 변동성 중에서 얼마나 설명했는가(X를 사용해 설명할 수 있는 Y의 변동비율)를 나타냄. 전체 변동 중 모델이 설명한 비율

$$R^2 = 1 - \frac{RSS}{TSS}$$

- $RSS = \sum(y_i - \hat{y}_i)^2$: 예측 오차의 제곱합
- $TSS = \sum(y_i - \bar{y})^2$: 전체 변동성 (y의 분산 총합)

- 해석
  - 값 범위: $0 \leq R^2 \leq 1$ (1에 가까울수록 모델이 데이터를 잘 설명함)
  - ex. R^2이 0.85면, 전체 y의 변동 중 85%를 모델이 잘 설명했다는 것을 의미
  - RSS = 모델이 설명하지 못한 오차, TSS = 모델없이 데이터 자체의 전체 변동성
- 단순선형 회귀에서는 $R^2 = r^2$ 가 된다. (이 때, r^2 는 상관계수를 의미함)

다중선형회귀

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

- y: 종속변수 (예측하고자 하는 값)
- $x_1, x_2, \ldots, x_p$: 독립변수들
- β_0: 절편 (intercept)
- $\beta_1, \beta_2, \ldots, \beta_p$: 각 독립변수의 계수 (기울기)
- ε: 오차항 (예측되지 못한 부분)

가설 검정

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \leftrightarrow H_a$: 적어도하나의β_j는0이아니다.
- F-통계량 : 전체 회귀모형이 통계적으로 유의미한 가?

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

  - TSS(전체 변동) = ESS(회귀모형이 설명한 변동) + RSS(회귀모형이 설명하지 못한 오차)
  - 평균 제곱 개념 도입 - 단순히 합이 아니라 평균적인 크기를 비교하고 싶다!
    - 모형이 변수 하나당 평균적으로 설명하는 변동성
    - 한 관측값당 평균적으로 발생한 예측 오차
  - F통계량으로 예측변수와 반응변수 사이의 연관성을 검정하는 접근법은 p가 상대적으로 작고 n에 비해서는 확실히 작을 때 잘 작동한다.

변수 선택

반응변수와 연관되어 있는 예측변수만 선택하는 작업

- 전진선택(forward selection) : RSS가 가장작은 변수를 하나씩 추가
- 후진선택(backward selection) : 모든 변수가 포함된 상태에서 p-값이 가장 큰 변수를 제거

- 혼합선택(mixed selection) : 전진선택을 하다가 모형 내 변수 중 하나의 p-값이 특정 임계값보다 커지는 순간부터 후진선택

모형 적합도

1. R^2

- 값이 1에 가까울수록 모형이 데이터를 잘 설명하지만, 변수를 많이 넣으면 무조건 증가하게 된다. (설명력)
- 수정된 결정계수 Adjusted R^2 사용

$$\text{Adjusted } R^2 = 1 - \left(\frac{RSS/(n-p-1)}{TSS/(n-1)} \right)$$

  - **변수 수(p)**를 고려한 R^2
  - 변수 추가에 따른 과적합 가능성을 보완
  - → 모델 성능 평가할 때는 **일반적으로 Adjusted R^2**가 더 믿을 만함

2. **RSE**

$$RSE = \sqrt{\frac{RSS}{n-p-1}}$$

  - RSS: 잔차 제곱합 (모델이 설명하지 못한 오차의 합)
  - n: 관측값 수
  - p: 독립변수 개수
  - → RSE는 **잔차의 표준편차**, 즉 **예측 오차의 평균적인 크기**

3. **F통계량** : 그 설명이 유의한가? (모형의 설명력을 통계적으로 검정, p-value와 함께 사용)
4. **잔차분석** : 잔차의 분포와 패턴 분석해서 정규성, 등분산성, 선형성, 독립성 가정을 만족하는지 확인

질적 예측 변수

- 더미변수로 변환해서 모델링
- 이진 질적 변수 (2개의 범주, ex. 주택소유/미소유)
  - 주택 소유한 사람과 비소유자의 차이로 해석될 수 있음
- 다중 질적 변수
  - n개 범주 중 (n-1)개의 더미 변수 생성 (ex. 서울, 대구, 부산 … → 서울1/0, 대구 1/0)
- 결국 질적 변수의 회귀계수는 기준 그룹과의 차이를 나타냄

선형모형의 확장

- 선형회귀모형의 가정 (현실에서는 자주 깨짐)
  - 가법성(additivity) : 예측변수 X_j 와 반응 Y 사이의 연관성이 다른 예측변수 값에 의존하지 않음
  - 선형성(linearity) : X_j 한 단위 변화와 연관된 반응변수 Y의 변화가 X_j 값에 관계없이 일정함
- 가법성 제거하기
  - 상호작용항 추가

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

- 비선형 관계
  - 다항회귀 : 독립변수 X의 고차항(X_2, X_3 등) 추가해서 비선형적인 패턴도 예측할 수 있게 만드는 모델

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

  - 변수는 여전히 **선형결합 형태로 쓰였기 때문에 선형회귀**임
  - 하지만 함수 형태는 **곡선**, 즉 **비선형 관계**를 표현할 수 있어

잠재적인 문제들

1. **데이터의 비선형성** → 잔차 그래프(residual plot)
    - 잔차에 U 패턴이 보이면 선형가정 부적합 (잔차는 랜덤하게 흩어져야 한다.)
2. **오차항 사이의 상관관계**
    - 선형회귀는 오차항은 무상관이라는 가정을 갖고 있음. But 시계열 데이터에서는 상관관계를 보이기도 함
    - 잔차 시계열 플롯, Durbin-Waston 검정(오차 간 자기상관 존재 여부 확인) 등으로 확인 가능
3. **이분산 오차항**
    - 오차항의 분산이 일정하지 않고 X의 값의 크기에 영향을 받는 경우
    - 표준오차가 잘못 계산됨 : p-value, 신뢰구간, t-test 등 추론 결과를 믿을 수 없게 됨
    - 잔차-예측값 그래프, Breusch-Pagan Test, White's Test 등으로 확인 가능
    - 변수 변환(로그, 루트 등), 가중 최소제곱법(WLS), Robust SE 등으로 해결
4. **이상치(outlier)** - y가 특이한 관측
5. **높은 지렛점** - x가 특이한 관측
    - 지렛값 통계량을 통해 한 관측의 지렛값을 계산할 수 있다.
6. **공선성(collinearity)**
    - 두 개 이상의 예측변수가 서로 밀접하게 관련되어 있는 상태
    - **다중공선성(multicollinearity)**
        - 두 개 이상의 예측변수가 서로 강하게 선형 결합될 수 있는 경우
    - 문제
        - 회귀계수 β를 정확하게 추정하게 어려움
        - 회귀계수의 표준오차가 커짐 → p-value가 커져서 유의하지 않은 것처럼 보임
        - 각 변수의 해석이 불안정하다.
    - 진단
        - 상관계수 행렬 확인 : 독립변수끼리 상관관계가 0.8 이상일 경우 공선성 의심

- VIF(Variance Infation Factor)

$$VIF_j = \frac{1}{1 - R_j^2}$$

  - X_j를 나머지 변수로 회귀했을 때 결정계수 R_j^2 기반
  - **VIF > 5 또는 10** → 다중공선성 심각

선형회귀 vs KNN

모수적 방법

- 데이터를 설명하는 고정된 형태 f(x)를 가정하고 그 모델의 파라미터(모수)를 추정하는 방법
- 모델의 형태를 미리 정하고, 그 모델 내에서 가장 잘 맞는 파라미터를 찾음

비모수적 방법

- 모델의 형태를 미리 가정하지 않고 데이터를 기반으로 모델의 구조 자체를 유연하게 학습하는 방법
- 데이터가 많을수록 더 정확하게 학습되고 복잡한 관계도 잘 포착 가능

KNN (K-Nearest Neighbors, k-최근접 이웃 알고리즘)

- 새로운 데이터가 주어졌을 때, 가장 가까운 K개의 이웃을 찾아 그 이웃들의 라벨을 기반으로 예측하는 알고리즘
- 작동 방식
  - 새로운 데이터 포인트가 주어짐
  - 학습데이터 중에서 이 포인트와 가장 가까운 K개의 점을 찾음 (보통 유클리디안 거리 사용)
  - 그 이웃들의 레이블 확인 (분류 - 다수결, 회귀 - 평균)
  - 최종 예측 결과 출력
- K의 최적값은 어떻게 찾아야 할까?
  - k값이 작으면 가장 유연한 적합을 제공 → 편향은 낮고 분산은 높아짐
  - k값이 크면 더 매끄럽고 변동이 덜한 적합을 제공

🔥 **언제 선형회귀(모수적 접근)이 KNN(비모수적 접근)보다 우수할까?**

- 모수 형태가 f의 참 형태와 가까운 경우
- 예측 변수당 관측 수가 적을 때
- 차원이 낮을 때