회귀분석팀

6팀

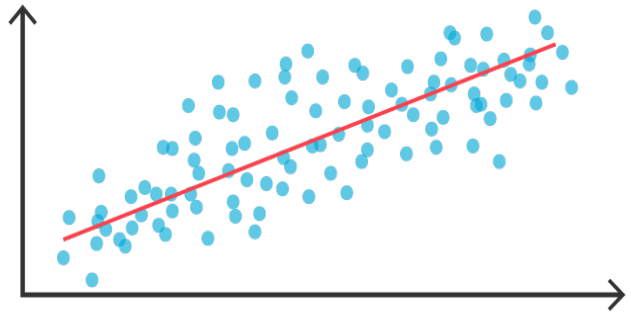
조수미
김민지
손재민
박윤아
조웅빈

CONTENTS

회귀분석이란?

"Regression + Analysis"

회귀선을 찾아 관계를 설명!



✓ **독립변수와 종속변수 간의 관계**를 설정하고 모델링하는 통계적 기법

✓ 특정변수들의 값을 이용하여 다른 변수를 **설명하고 예측**하는 방법

회귀분석이란?

회귀식

회귀식

종속변수 Y와 독립변수 X의 관계를 **함수식(f)**으로 표현한 모델

$$Y = f(X_1, X_2, \cdots, X_p) + \varepsilon$$

- Y **종속변수** 독립변수에 의해서 설명되는 변수
- X_k **독립변수** 종속변수를 설명하기 위한 변수
- ε **오차항** 변수를 측정할 때 발생할 수 있는 오차

설명할 수 없는 무작위성을 가짐

단순선형회귀

단순선형회귀 Simple Linear Regression

하나의 종속변수와 하나의 독립변수만을 가짐

두 변수의 관계를 가장 잘 표현하는 **직선을 추정**하는 것이 목적

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

y_i **종속변수** 종속변수 y의 i번째 관측값

x_i **독립변수** 독립변수 x의 i번째 관측값

ε_i **오차항** i번째 관측값에 의한 랜덤 오차

β_0, β_1 **회귀계수** 추정해야 할 모수

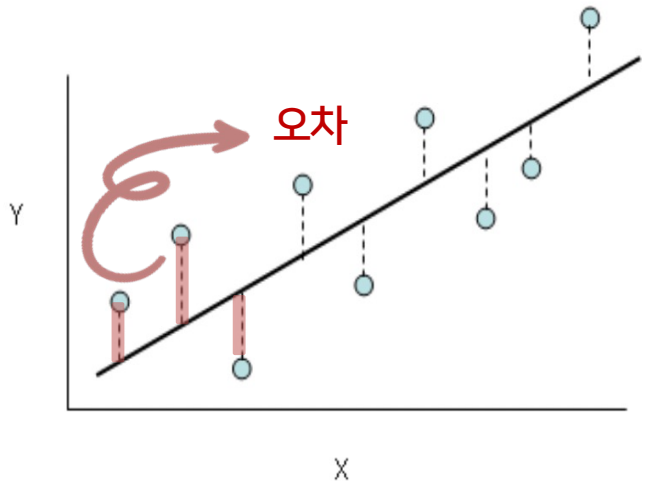
$\varepsilon_i \sim NID(0, \sigma^2)$
평균은 0, 분산은 σ^2를 가정

최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법



오차의 제곱합 최소화

$$\text{argmin } S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

$$\frac{\partial S}{\partial \beta_0} | \widehat{\beta_0}, \widehat{\beta_1} = -2\Sigma_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} | \widehat{\beta_0}, \widehat{\beta_1} = -2\Sigma_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i) x_i = 0$$

아래로 볼록한 이차함수 형태는
최소값을 가지므로 편미분!

최소제곱법의 가정과 특징

BLUE Best Linear Unbiased Estimator

분산이 제일 작은 선형 불편추정량

분산이 작다는 것은 **추정량이 안정적**이라는 의미

① 오차들의 **평균은 0**

② 오차들의 **분산은 σ^2 로 동일**

③ 오차간 **자기상관이 없음**

Independent

세 가지 조건이 만족되면, LSE는 선형불편추정량 중
분산이 가장 작은 안정적인 추정량이 됨

적합성 검정

적합성 검정 Goodness of Fit

회귀직선이 데이터에 **얼마나 잘 들어맞는지** 모형에 대한 적합성 검정

모집단에서의 실제 오차값을 알 수 없기에 잔차 사용

오차

모집단에서 얻은 회귀식에서의 실제값 y_i와 추정값 $\hat{y}_i$의 차이

↓

오차는 모집단의 측정치

vs

잔차

표본에서 얻은 회귀식에서의 실제값 y_i와 추정값 $\hat{y}_i$의 차이

↓

잔차는 표본의 측정치

적합성 검정

결정계수 Coefficient of Determinant

총 변동(SST)에서 회귀식이 설명할 수 있는 비율(SSR)

즉, Y가 X에 의해 설명되는 비율로, 1에 가까울수록 좋음

$$\uparrow R^2 = \frac{\uparrow SSR}{SST} = 1 - \frac{\downarrow SSE}{SST}$$

!!

잔차와 연관지어 본다면,

잔차제곱합(SSE)은 회귀식이 설명할 수 없는 실제값과 추정값 사이의 오차이므로

총 변동 대비 잔차제곱합이 차지하는 비율이 작을수록 좋음

유의성 검정

유의성 검정 Significance Test

전체 회귀식이 아닌 **개별 모수**의 추정량이 통계적으로 유의한지를 알아보는 과정

β_0도 동일한 방법으로 검정하면 됨

① 가설 설정 : $H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$

② 추정량의 분포 : $\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

③ 검정 통계량 : $t_0 = \frac{\widehat{\beta_1}}{se(\widehat{\beta_1})} \sim t_{(n-2)}$

④ 임계값 : $t_{(1-\alpha/2,n-2)}$

⑤ 검정(양측) : If$|t_0| > t_{(1-\alpha/2,n-2)}$, reject H_0 at α

다중선형회귀

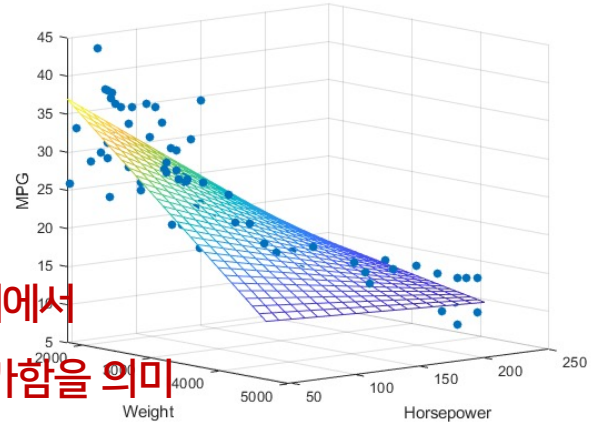
다중선형회귀 Multiple Linear Regression

2개 이상의 독립변수를 가짐

단순회귀분석에 비해 **복잡한 관계 설명에 용이**

설명변수가 p개로 확장

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,\ \epsilon \sim NID(0, \sigma^2)$$



나머지 설명변수를 고정시킨 상태에서

x_1이 한 단위 증가할 때 y가 β_1만큼 증가함을 의미

유의성 검정

F-test

전체 회귀계수에 관한 검정

가설 설정

$\boldsymbol{H_0}: \beta_0 = \beta_1 = \cdots = \beta_p = 0$

$\boldsymbol{H_1}: not\ \boldsymbol{H_0}$ ($\beta_0,\ \beta_1, \cdots, \beta_p$ 중 적어도 하나는 0이 아니다.)

검정통계량

$$\boldsymbol{F_0} = \frac{(SST-SSE)/p}{SSE/(n-p-1)} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE}$$

유의성 검정

Partial F-test

일부 회귀계수에 관한 검정

가설 설정

Full model (FM) = 모든 변수를 사용한 회귀모형

Reduced Model (RM) = 일부 계수를 특정 값으로 둔 축소모형

$\boldsymbol{H_0}$: $\beta_j = \beta_{j+1} = \cdots = \beta_{j+q-1} = 0$

$\boldsymbol{H_1}$: not $\boldsymbol{H_0}$ ($\beta_J, \beta_{J+1}, \cdots, \beta_{j+q-1}$ 중 적어도 하나는 0이 아니다)

검정통계량

$$\boldsymbol{F_0} = \frac{(SSE(RM) - SSE(FM))/(p-q)}{SSE(FM)/(n-p-1)}$$

$$= \frac{(SSR(FM) - SSR(RM))/(p-q)}{SSE(FM)/(n-p-1)} \sim \boldsymbol{F_{p-q,n-p-1}}$$

유의성 검정

T-test

개별 회귀계수에 대한 검정

회귀계수 추가의 유의성을 판단하기 위해 사용

가설 설정

$\mathbf{H_0}$: $\beta_j = 0$ 다른 변수들이 다 적합된 상태에서 설명변수 x_j는 유의하지 않음

$\mathbf{H_1}$: $\beta_j \neq 0$ 다른 변수들이 다 적합된 상태에서 설명변수 x_j는 유의함

검정통계량

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$$

데이터 진단

데이터 진단의 필요성

데이터 진단, 왜 필요해?

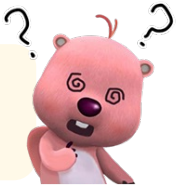
이상치, 지렛값, 영향점 등

일반적인 경향에서 벗어나는 데이터 존재

회귀 모형에 큰 영향을 미침

데이터가 일반적인 경향에서 벗어나는지 **1) 판단 2) 처리**

잔차를 이용해 데이터 진단을 할 수 있을까?



영향점

영향점 Influential Point

회귀직선의 기울기에 상당한 영향을 주는 점

이상치와 지렛값을 동시에 고려하는 지표

Cook's Distance

영향점을 확인하는 표준적인 지표

특정 데이터를 지웠을 때 **회귀선이 변하는 정도**를 나타냄

이상치 $C_i = \frac{r_i^2}{p+1} \times \frac{h_{ii}}{1-h_{ii}}$ 지렛값

!!

$C_i > 1$ 이면 영향점으로 판단 !

영향점 처리의 필요성

영향점은 추정량의 **분산을 크게** 만듦

잘못된 모델의 해석과 **예측 성능 저하**

영향점 처리를 통해 **이상치에 강건한(robust) 모델링**

로버스트 회귀

로버스트 회귀 모형 Robust Regression

이상치의 영향을 크게 받지 않는 회귀모형

Median Regression

Huber's M-estimation

Least Trimmed Square