# 회귀분석팀

6팀
고경현
박세령
박이현
박지성
심예진
이선민

# INDEX

# 1 회귀분석이란?

## 회귀분석

*회귀팀 파이팅><*
*(feat. 학회장님)*

### Regression Analysis

**변수 사이의 관계**를 모델링하는 통계적 기법

특정 변수들의 값을 이용하여 다른 변수를 설명하거나 예측

Ex ) 스마트폰 이용시간(X변수)에 따른 기말고사 성적 변화(Y변수)

**지도학습**의 한 종류

### **지도학습**이란?

결과의 예측이 목적인 학습 방법

결과변수와 특징변수가 모두 존재

# 1 회귀분석이란?

## 회귀식 (회귀 모델)

회귀분석에서 변수들 간의 관계를 함수식으로 표현한 모델

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

$X$ 변수 (독립변수, Independent Variable)

종속변수를 설명하기 위한 변수

$Y$ 변수 (종속변수, Dependent Variable)

독립변수에 의해서 설명되는 변수

$\epsilon$ (오차항, Error Term)

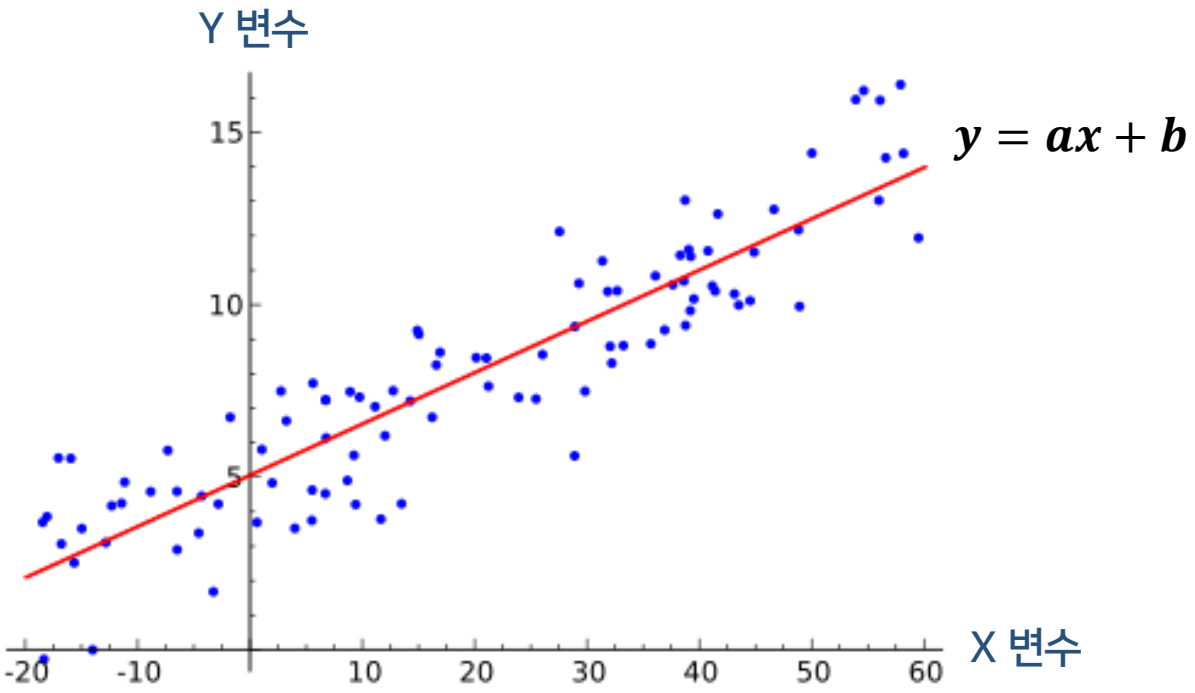
변수를 측정할 때 발생할 수 있는 오차

설명이 불가능한 무작위성

## 단순선형회귀

하나의 종속변수(Y)와 하나의 독립변수(X)만을 갖는 회귀모델

두 변수 간의 관계를 가장 잘 표현하는 직선 추정이 목적



# 2 단순선형회귀

## 회귀 모델

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

$$i = 1, \ldots, n \quad \epsilon_i \sim NID(0, \sigma^2)$$

$y_i$ : 종속변수 Y의 i번째 관측값

$x_i$ : 독립변수 X의 i번째 관측값

$\epsilon_i$ : i번째 관측값에 의한 랜덤 오차

평균은 0, 분산은 $\sigma^2$ 를 가정

$\beta_0, \beta_1$ : 회귀계수, 추정해야 할 모수

**특정한 함수를 가정하는 모수적 방법**

## 모수의 추정 : 최소제곱법

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

오차제곱합 $S(\beta_0, \beta_1)$를 최소화하는 $\beta_0, \beta_1$를 찾는 것이 목적

아래로 볼록한 이차함수 형태 → 최소값(= 극소값)을 가지므로 편미분!

**최소제곱법을 통해 얻은 추정치 $\widehat{\beta_0}$, $\widehat{\beta_1}$**

**최소제곱추정치**

Least Square Estimator

## 최소제곱법의 가정과 특징

아무런 조건(가정) 없이 사용 가능

세 가지 조건이 만족되면,

LSE는 선형불편추정량 중 분산이 가장 작은 안정적인 추정량이 됨

① 오차들의 평균은 0

② 오차들의 분산은 $\sigma^2$ 로 동일

③ 오차간 자기상관이 없음(Independent)

Gauss-Markov Theorem

**BLUE(Best Linear Unbiased Estimator)**

## 적합성 검정

### 결정계수

**잔차제곱합**(SSE)은 회귀식이 설명할 수 없는

실제값과 추정값 사이의 오차

∴ 총 변동 대비 잔차제곱합이 차지하는 비율이 **작을수록** 좋음

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

## 유의성 검정

전체 회귀식이 아닌, 개별 모수의 추정량이

통계적으로 **유의한지**를 알아보는 과정

$\epsilon_i \sim NID(0, \sigma^2)$ 라는 오차의 정규분포 가정 하에 이루어짐

### 검정 과정

- 가설 검정 : $H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$
- 추정량의 분포 : $\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$
- 검정 통계량 : $t_0 = \frac{\widehat{\beta_1}}{se(\widehat{\beta_1})} \sim t_{(n-2)}$
- 임계값 : $t_{(1-\alpha/2, n-2)}$
- 검정(양측) : If $|t_0| > t_{(1-\alpha/2, n-2)}$ reject $H_0$ at $\alpha$

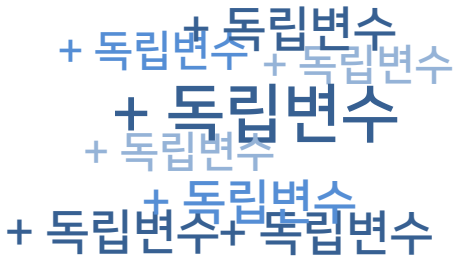
## 다중선형회귀

### 단순선형회귀

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

### 다중선형회귀

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$




단순선형회귀에 비해 **복잡한 관계** 설명이 용이

나머지 X변수들이 고정되어 있을 때,
$x_p$가 한 단위 증가하면 $y$ 는 $\beta_p$만큼 증가함을 의미

## 유의성 검정

전체 회귀계수에 대한 검정 : F-test

가설설정

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0$$

$$H_1 : not\ H_0$$

검정통계량

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

회귀식으로 설명하지 못하는 부분에 비해 얼마나 설명이 가능한가를 보여줌

→ 회귀식의 전반적인 계수가 얼마나 설명력을 갖는지를 보여줌

## 유의성 검정

일부 회귀계수에 대한 검정 : Partial F-test

**검정통계량**

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

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

검정통계량의 의미

$SSE(RM)$

$q$ 개의 변수를 제거했을 때

모델이 설명하지 못하는 변동

제거된 변수가
모델에 유의미하다면 월등히 커짐

$SSE(FM)$

모든 변수를 포함했을 때

모델이 설명하지 못하는 변동

# 3 다중선형회귀

## 유의성 검정

개별 회귀계수에 대한 검정 : t-t

임계값 $t_{(\alpha/2,\, n-p-1)}$

F-test를 t-test보다 먼저 검정할 것

- 귀무가설 기각 if $|t_j| \geq t_{(\alpha/2,\, n-p-1)}$
  - $x_j$를 새로 추가하는 것은 통계적으로 유의함
- 귀무가설 기각 안됨 if $|t_j| < t_{(\alpha/2,\, n-p-1)}$
  - $x_j$를 새로 추가하는 것은 통계적으로 유의하지 않음

전체 모델에 대한 F값을 확인해 봄으로써

모델 전체가 **통계적으로 유의**한지를 먼저 확인해야함

다른 변수들이 다 적합 된 상태

# 4 데이터 진단

## 데이터 진단의 필요성

일반적인 경향에서 벗어나는 개별 데이터 존재

이상치, 지렛값, 영향점 등

회귀 모형에 큰 영향을 미침

개별 데이터가 경향성에 벗어나는지 판단하여 처리 필요!

데이터 진단 과정

## 영향점

### 영향점 Influential point

회귀직선의 기울기에 상당한 영향을 주는 점

### Cook's Distance

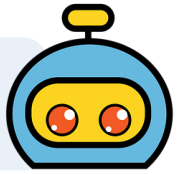
Outlier와 Leverage를 동시에 고려하는 지표로,

특정 데이터를 지웠을 때 회귀선이 변화하는 정도를 나타냄

## 영향점의 처리

영향점이 있으면 왜 안되나요?

추정량을 불안정하게 (분산을 크게) 만듦



잘못된 모델의 해석

예측 성능 저하

로봇이 채팅방을 나갔습니다.

---

영향점 제거

이상치에 강건한 모델링

## 로버스트 회귀

### 로버스트 회귀

이상치의 영향력을 크게 받지 않는 회귀모형