회 귀 분 석

ToBig's 10기 박규리

# Regression Analysis

## 선형회귀분석

# Contents

## 머신러닝이란?

"어떤 작업 T에 대한 컴퓨터 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상됐다면,
이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로 학습한 것이다."

## 머신러닝의 종류

지도학습 : 알고리즘에 주입하는 훈련 데이터에 레이블을 넣어 훈련시키는 것
비지도 학습 : 레이블 없이 데이터가 알아서 학습하는 것
강화 학습 : 환경을 관찰해서 행동을 실행하고 그 결과로 보상 또는 벌점을 받고, 시간이 지나면서 가장 큰 보상을 얻기 위해 정책이라는 최상의 전략을 스스로 학습

사례 기반 학습과 모델 기반 학습
: 머신러닝 시스템은 어떻게 일반화되는가

사례 기반 학습
: 이미 알고 있는 데이터와 새로운 데이터 간의 유사한 정도를 통해 예측하는 방법
Ex) knn

모델 기반 학습
: 학습 데이터로부터 일반화 할 수 있는 모델을 만들어 예측하는 방법
Ex) 회귀분석

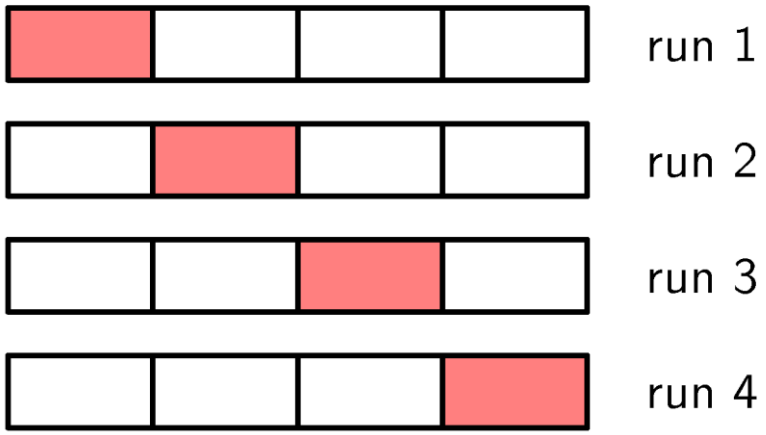
## 머신러닝의 주요 도전 과제

- 충분하지 않은 양의 훈련 데이터
- 대표성 없는 훈련 데이터 ex) 일부 집단 누락
- 낮은 품질의 데이터 ex) 에러, 이상치, 잡음 등
-
- 관련 없는 특성 ex) 필요한 특성이 없음
- 훈련 데이터 과대 적합(overfitting)

훈련 데이터 과대 적합
: 모델이 훈련 데이터에 최적화되어 있지만 새로운 데이터에 안맞아서 일반성이 떨어지는 것

해결책 중 하나
: 데이터에서 testset을 따로 떼내어 모델을 만든 후 test셋에 적용하여 비교하기
➔ 모델이 Test셋에 과적합될 우려
➔ K fold cross validation 사용(교차검증)

Ex) 4 fold cross validation
주어진 데이터를 4개로 나누어
3개는 모델을 훈련하고
1개는 모델을 테스트하며,
이것을 4번 반복한 후,
테스트 정확도를 평균하는 것
*보통 10 fold cross validation을 많이 씀



Machine learning flow

데이터 분석을 어떻게 해야 하는가?

- 직접 데이터를 뜯어보며 많은 고민을 해야합니다

- 막막한 분들은 아래 내용을 참고해보면서 시작해보았으면 좋겠습니다

https://github.com/ExcelsiorCJH/Hands-On-ML/blob/master/Chap02-End_to_End_ML_Project/Chap02-End_to_End_ML_Project.ipynb
(핸즈온 머신러닝 : 교재 코드라서 굉장히 깔끔)

https://github.com/KaggleBreak/walkingkaggle
(캐글뽀개기 스터디 자료 : 데이터 분석을 실제 해본 분들의 고민과 코드가 담겨있어 굉장히 유용한 자료)

# 두 가지 형태의 모형

## 두 가지 형태의 모형

### 1) 결정적 모형 : input 과 output의 관계가 오차 없이 명확

$$Y = f(X_1, \cdots, X_p)$$

예)
- ✓ 힘 = 질량 × 가속도
- ✓ 화씨 = 32 + 1.8 × 섭씨

### 2) 통계적 모형: output이 input에 의해 영향을 받는 경향을 보이며 언제나 오차를 수반

우리들의 관심사

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

(의문)
다른 변수도 가능한가? 다른 변수가 더 설명력이 높은가?
이 식은 어떻게 나온 건가? 이 식은 얼마나 믿을 수 있는가?

예)
- ✓ 매출액 = 100 + 0.1 × 광고비 + $\epsilon$
- ✓ 수축기 혈압 = 110 + 0.1 × 연령 + 0.15 × 몸무게 + $\epsilon$

# 두 가지 형태의 모형

## 정규분포

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad,\quad -\infty<x<+\infty\quad,\quad e=2.71828\cdots$$

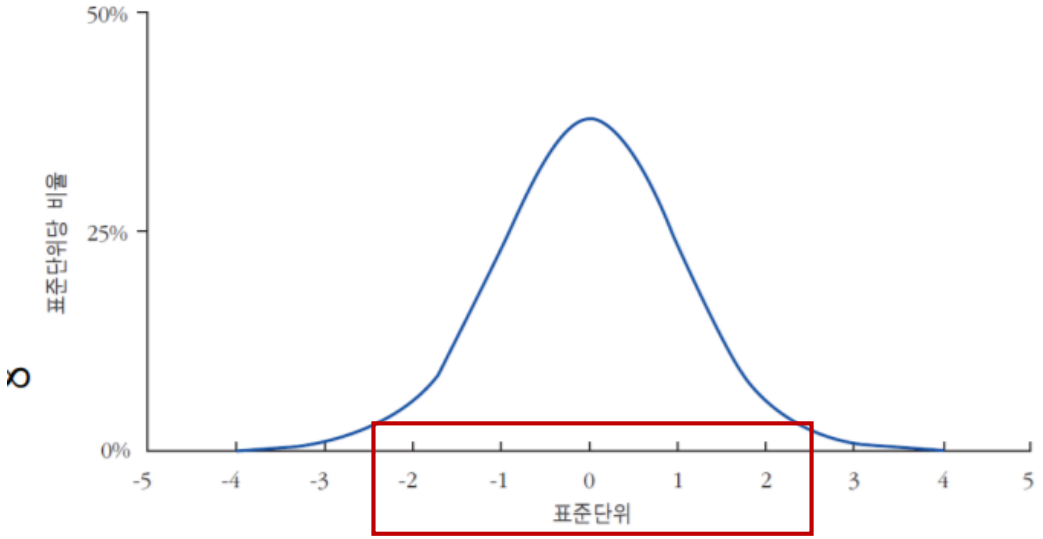
하나의 이상적인 히스토그램. 하나의 수학적 모형. 개념상 모집단의 분포.

- 정규분포의 확률밀도함수(probability density function)
- $\mu$ 를 모평균, $\sigma$를 모표준편차라고 부름.
- 모집단: 모평균과 모표준편차
- 표본: 표본평균과 표본표준편차

# 두 가지 형태의 모형

## 표준정규분포 (standard normal distribution)

$$f(z) = \frac{1}{2\pi} e^{\left\{-\frac{1}{2}z^2\right\}}, -\infty < z < \infty$$




평균이 0이고 표준편차가 1인 정규분포: $Z \sim N(0, 1)$

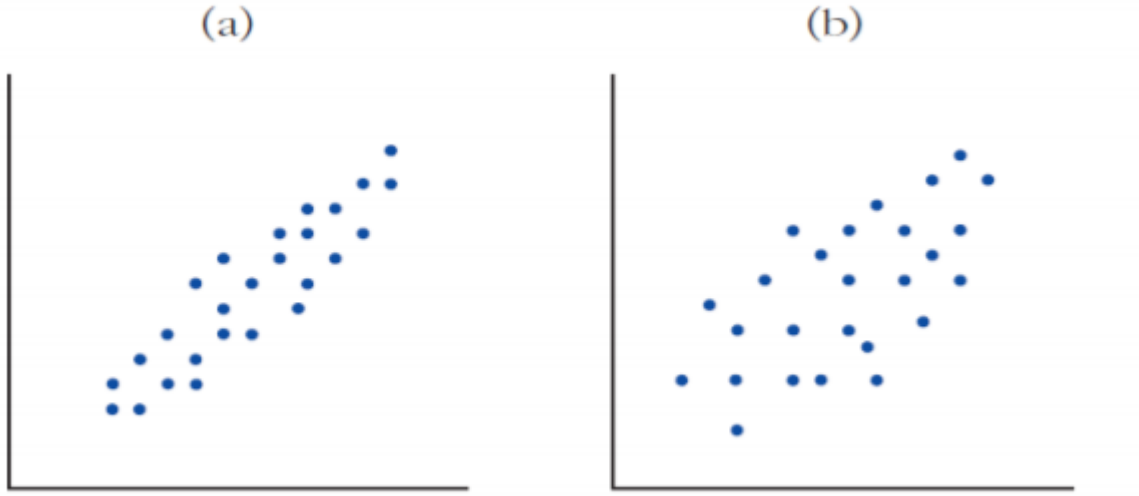
평균으로부터 2시그마 안에 95%의 데이터가들어갑니다

정규분포곡선의 68-95-99.7 규칙
- 표준단위로 -1부터 1까지 영역의 넓이 : 약 68%
- 표준단위로 -2부터 2까지 영역의 넓이 : 약 95%
- 표준단위로 -3부터 3까지 영역의 넓이 : 약 99.7%

- 정규분포곡선은 평균과 표준편차에 의해 그 모양이 완벽하게 묘사된다.
- 즉, 정규분포를 따르는 히스토그램은 중심과 중심 주위로 퍼진 정도 등 두 정 보만으로 100% 묘사된다.

# 상관계수

## 상관계수의 개념과 필요성




- 가로든 세로든 평균과 표준편차가 동일해도 두 변수의 관계는 상이
- 위의 두 산포도는 가로든 세로든 중심과 퍼진 정도가 동일하지만 (a)가 (b) 보다 더 강한 선형관계를 보임
- 두 변수간 선형관계의 방향과 강도가 얼마나 되는지 측정할 필요성 대두
- 상관계수는 두 변수간 선형관계의 방향과 강도 측정

# 상관계수

## 상관계수 공식과 특징

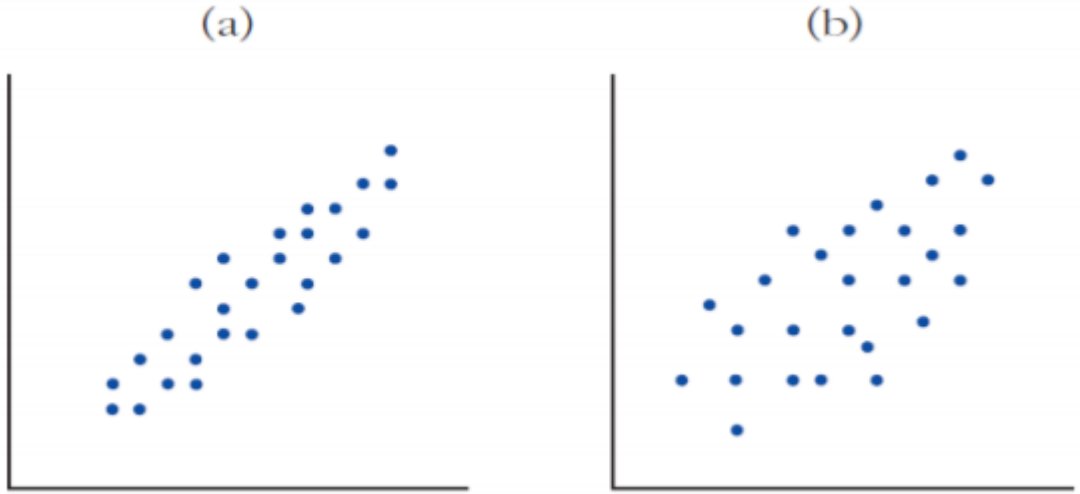
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

공분산을 각각(x와 y)의 표준편차로 나눈 것

- 범위: $-1 \leq r \leq 1$
- 상관계수 = 1 또는 −1 이면 완전상관(perfect correlation)
  - 모든 점들이 정확히 하나의 선 위에 위치
  - 양의 상관관계이면 점의 분포가 우상향
  - 음의 상관관계이면 점의 분포가 우하향
- 두 변수의 표준편차가 모두 0이면 상관계수를 정의할 수 없음
- 두 변수 중 어느 한 변수만의 표준편차가 0이면 상관계수는 0

# 상관계수

## 상관계수의 의미

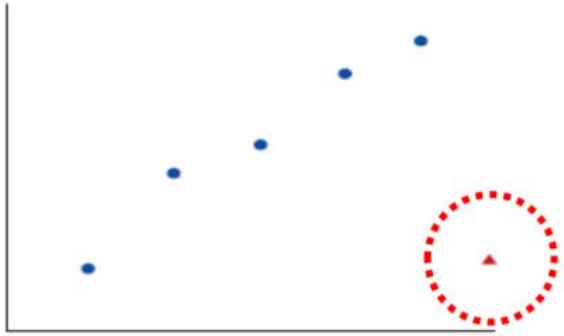



- '상관계수=0.8'은 산포도 상에서 80%의 점들이 하나의 선 주위에 빽빽하게 밀집해 있다는 것을 의미하지 않는다.
- '상관계수=0.8'은 상관계수가 0.4일 때보다 선형관계의 강도가 강하기는 하지만 정확히 두 배로 강하다는 것을 의미하지도 않는다.
  그저 상대적인 수치일 뿐

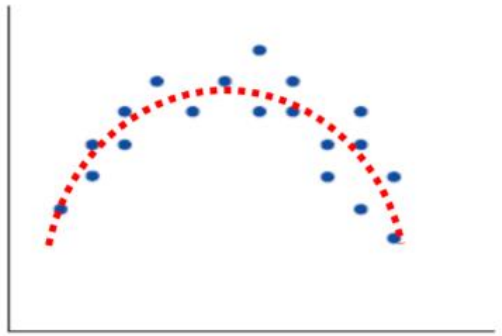
# 상관계수

## 상관계수가 유의미하지 않은 경우

(a) 이탈값



(b) 비선형관계



- 이탈값(outlier)이 존재하는 경우
- 두 변수간 관계가 비선형인 경우

# 회귀분석

**회귀분석**
**: 반응 변수와 설명 변수를 관측하여 회귀 함수를 추정하는 것**
**: 집단별 평균을 분석하는 통계적 방법**

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

$Y$ : 반응변수 (*response variable*), 종속변수 (*dependent variable*)

$X_1, \cdots, X_p$ : 공변량, 설명변수 (*covariates*)

$\epsilon$ : 오차항 (*error term*)

$f(\cdot)$ : 회귀함수 (*regression function*)

# 회귀모형의 종류

## 1) 모수회귀모형(parametric regression model)

f(x1,x2,…xp)
: 이 함수를 찾기 위해 모든 형태의 함수를 고려할 수 없으므로 적절한 함수를 고르는 것은 불가능
그래서 함수의 형태를 고정시키는 제약을 둔다

- 단순선형회귀모형 (simple linear regression model)

$$f(X) = \beta_0 + \beta_1 X$$

- 다중선형회귀모형 (multiple linear regression model)

$$f(X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- 비선형회귀모형 (non-linear regression model)

$$f(X) = \frac{\beta_0 \cdot X}{\beta_1 + X}$$

함수는 이렇게 생겼다고 생각할거임~

- **k**-차 다항회귀모형 (**k**-th degree polynomial regression model)

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$$

- 로지스틱 회귀모형 (logistic regression model)

반응변수가 이항분포 (성공의 횟수)를 따를 때

- 로그선형모형 (log-linear regression model)

반응변수가 포아송 분포 (사건의 발생 건수)를 따를 때

함수는 이렇게 생겼다고 생각할거임~

# 회귀모형의 종류

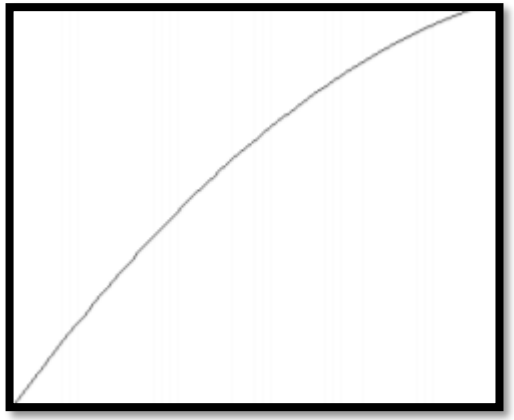
2) 비모수회귀모형(non parametric regression model)
: 회귀함수 f의 형태를 구체적으로 명시하지 않고 함수의 추정치를 계산

어떤 함수군에 속한다던지 등 아주 가벼운 가정만 한다고 한다

# 단순선형회귀모형

모형의 종류

참모형(true model)

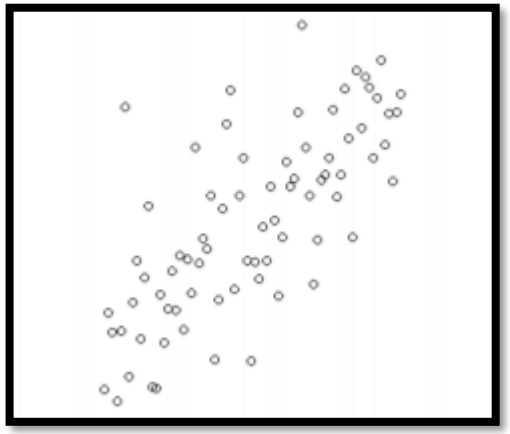


(only God knows)
실제 x와 y와의 관계

이 함수를 알 수 있다면
굳이 회귀분석 할 필요가 없다

그러나 알 수 없다

자료의 산점도



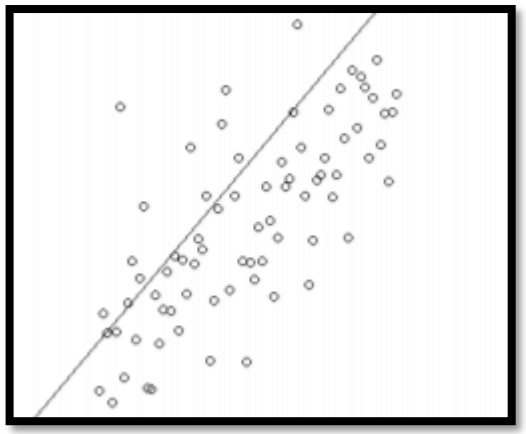
(Humans knows)
우리가 알 수 있는 것

실제 자료

# 단순선형회귀모형

모형의 종류

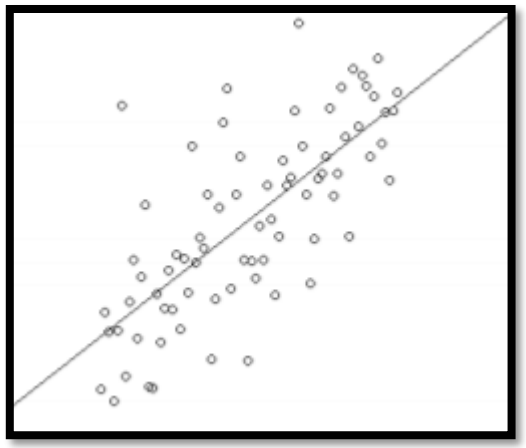
## 설정 모형(postulated model)



"x와 y는 직선의 관계를 가진다"
같은 모형설정을 한다

임의로 자료를 잘 판단해 직선이
잘 나타 내줄 것이라고 설정

## 적합모형(fitted model)

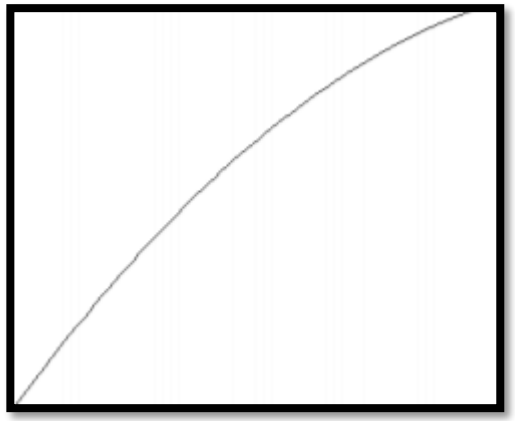


여러 개의 설정 모델 중 이 산포도를
가장 잘 표현해줄 것 같은 적합한 직선

# 단순선형회귀모형

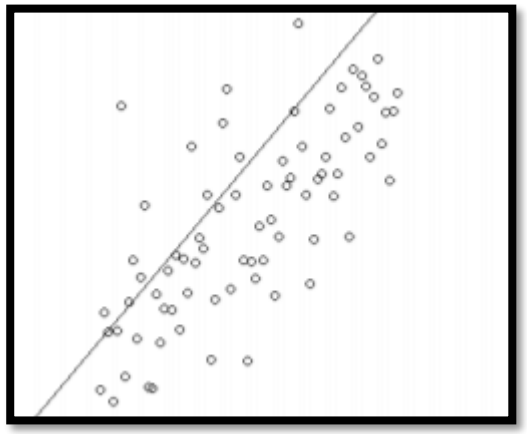
모형의 종류

참모형(true model)



≠

설정 모형(postulated model)



이 둘이 다르기 때문에 모형 설정에 대한 오차 존재 ( model error)

# 단순선형회귀모형

## 단순선형회귀모형

- **단순선형회귀모형 (simple linear regression model)**

cf .직선회귀모형 (straight line regression model)

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

Y : 반응변수 알고싶은 변수

X : 설명변수 알고 싶은 변수를 설명해 주는 변수

$\epsilon$ : 오차항. 흔히, $\epsilon \sim N(0, \sigma^2)$ 이라고 가정 (정규성 가정)
오차항은 평균이 0이고 분산이 $\sigma^2$를 따름
회귀모형으로 설명되지 않는 error

$\beta_0$, $\beta_1$ : 회귀계수 (regression coefficients), 추정해야 할 모수
회귀식을 도출하기 위해 필요한 추정할 모수

회귀식의 절편과 계수와 분산을 추정

⇒ **단순회귀모형의 목표 : 3개의 모수 $\beta_0$, $\beta_1$, $\sigma^2$ 을 추정**

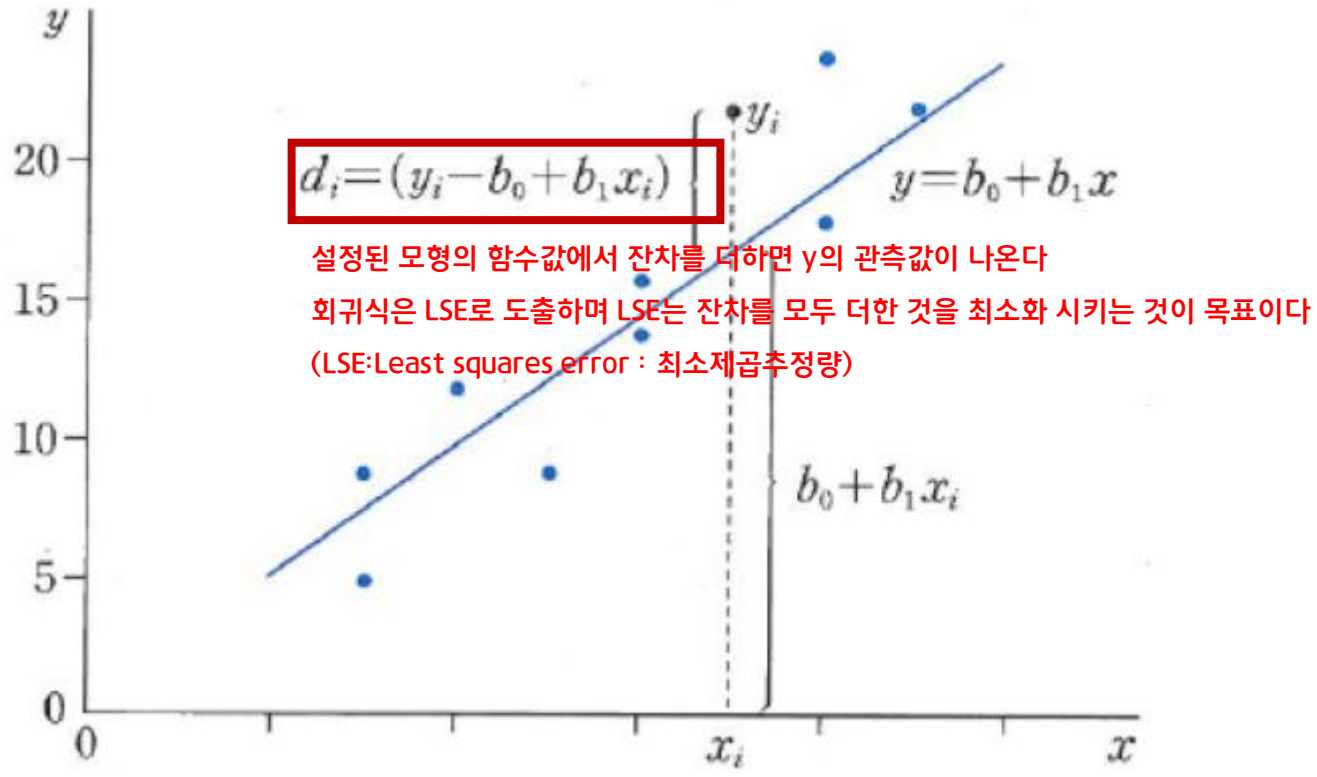
이를 위해 n개의 자료 $(X_i, Y_i)$, $i = 1,2,\cdots,n$을 관측.

즉, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ , $i = 1,2,\cdots,n$

# 최소제곱추정량(Least squares error)

**최소제곱법**

OLS(Ordinary Least Squares)는 가장 기본적인 결정론적 회귀 방법으로
Residual Sum of Squares(RSS)를 **최소화**하는 가중치 벡터 값을 미분을 통해 구한다.




설정된 모형의 함수값에서 잔차를 더하면 y의 관측값이 나온다

회귀식은 LSE로 도출하며 LSE는 잔차를 모두 더한 것을 최소화 시키는 것이 목표이다

(LSE:Least squares error : 최소제곱추정량)

# 최소제곱추정량(Least squares error)

OLS는 가장 기본적인 결정론적 회귀 방법으로 Residual Sum of Squares(RSS)를 **최소화하는 가중치 벡터 값을 미분**을 통해 구한다.

실제값

모형으로부터의 오류를 최소화하는 값 추정

$$D = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \text{ 를 최소로 하는 } \beta_0,\ \beta_1 \text{를 추정}$$

모형으로부터의 추정값

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

모형

이를 위해 편미분

$$\frac{\partial D}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial D}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

D를 베타제로와 베타 원으로 편미분한 값이 0이 되는 것이 최소값이에요

위의 식 풀면 밑에 식이 돼요

- $\beta_0$의 최소제곱추정량 : $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ 밑의 결과랑 나와있는 결과 대입
- $\beta_1$의 최소제곱추정량 : $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$, X의 y의 공분산을 x의 분산으로 나눈 값

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}),\quad S_{XX} = \sum (X_i - \bar{X})^2$$

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

# 최소제곱추정량(Least squares error)

## 추정치 신뢰구간

적합된 회귀식 (fitted regression line) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$X_i$에서의 적합치 (fitted value at $X_i$) $\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$i$번째 잔차 ($i$th residual) $e_i = Y_i - \widehat{Y_i}$

$\sigma^2$의 추정치 $s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$

1번째 값

모분산은 추정 해야 한다
잔차를 제곱하여 더한 것을 자유도(n-2)로 나눈 것으로 추정

$$\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{XX}}} \sim t(n-2)$$

추정 베타1과 실제 베타1의 차이는
T분포(자유도n-2)를 따르고 있다

$$\frac{\hat{\beta}_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}} \sim t(n-2)$$

추정 베타0과 실제 베타0의 차이는
T분포(자유도n-2)를 따르고 있다

$\Rightarrow \beta_1$에 대한 $100 \times (1-\alpha)\%$ 신뢰구간

$$\hat{\beta}_1 \pm t_{\alpha/2}\frac{s}{\sqrt{S_{XX}}}$$

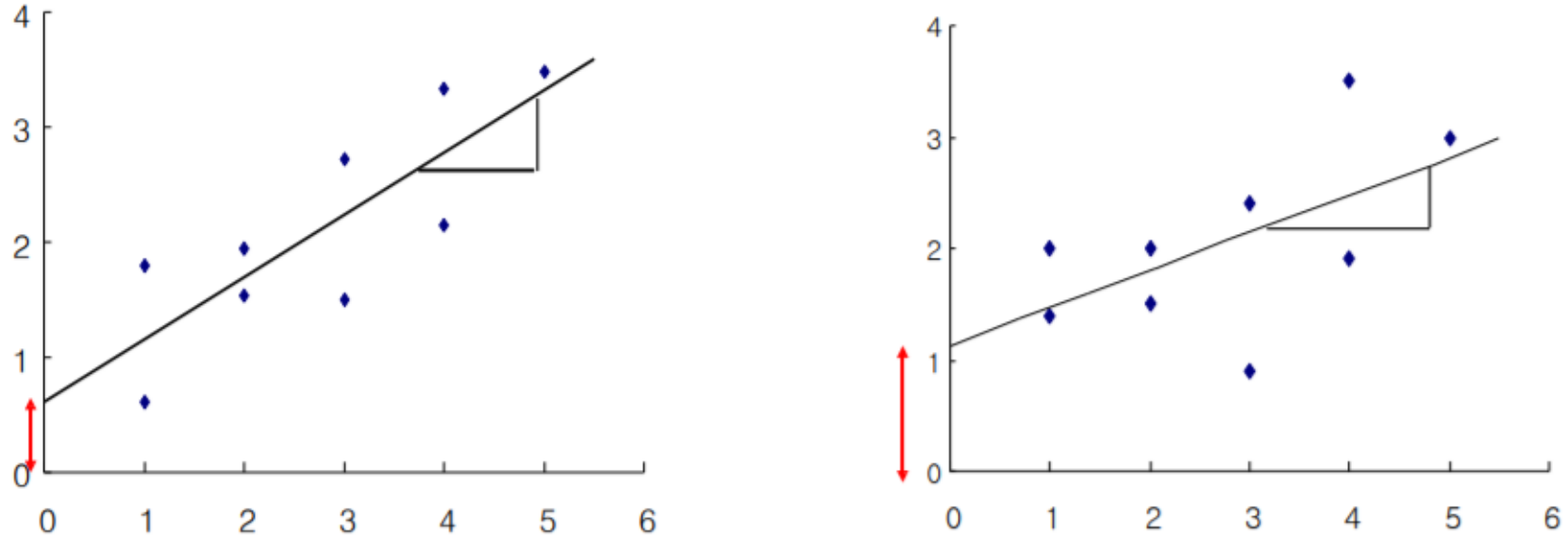
$\Rightarrow \beta_0$에 대한 $100 \times (1-\alpha)\%$ 신뢰구간

$$\hat{\beta}_0 \pm t_{\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$$

옆의 값으로 신뢰구간 구할 수 있음

# 회귀분석의 유의성

## 회귀분석 추정량의 표준오차

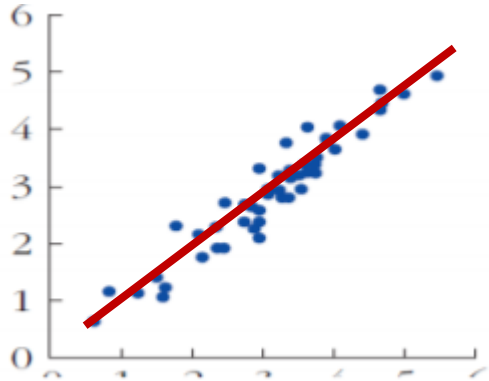


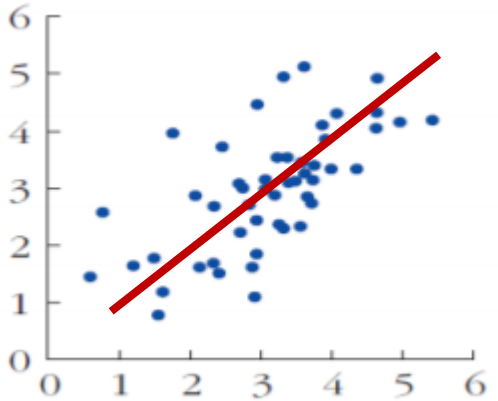

자료가 변하면 절편과 기울기가 모두 변한다.
이 때 얼마나 변화할지는 절편과 기울기 추정량의 표준오차를 보고 짐작 가능

표본이 바뀌면 절편과 기울기가 달라지겠지만 이것이 대략적으로 얼마나 바뀔 가능성이 있는가?
표준오차를 보면 된다!

# 회귀모형의 적합도

## 회귀모형의 적합도 intro




두 산포도에 대한 회귀식의 표본 수와 기울기가 같으며 회귀식이 같다고 가정

그러나 두 식은 실제 데이터에 적합된 정도가 다르다(=설명력이 다르다)

# 회귀분석의 유의성

## 회귀분석 추정량의 표준오차(변동성)

단순회귀분석에서 절편과 기울기 추정량의 표준오차

$$SE(a)=\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \quad SE(b)=\frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

평균이 y축에서 떨어져있으면 절편 근처에 자료가 별로 없으니 변동성 높아짐

샘플 사이즈가 커지면 절편의 불확실성이 감소

회귀직선으로부터 관측치 하나하나가 엄청 떨어져있으면 불확실성이 커짐

$$\hat{SE}(a)=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \quad \hat{SE}(b)=\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

절편의 표준오차

계수의 표준오차

평균에 X가 몰려 있지 않고 퍼져 있어야 불확실성 감소

평균에 X가 몰려 있지 않고 퍼져 있어야 불확실성 감소

$\sigma$: 회귀분석 오차의 표준편차

$\hat{\sigma}$: 회귀분석 오차의 표준편차에 대한 추정치
앞에서는 이를 RMSE로 표기했었음

$$\hat{\sigma}=\sqrt{\sum_{i=1}^{n}(y_i-a-bx_i)^2/(n-2)}$$

잔차들을 다 제곱해서 합치고 자유도로 나누고 루트를 씌운 시그마 햇

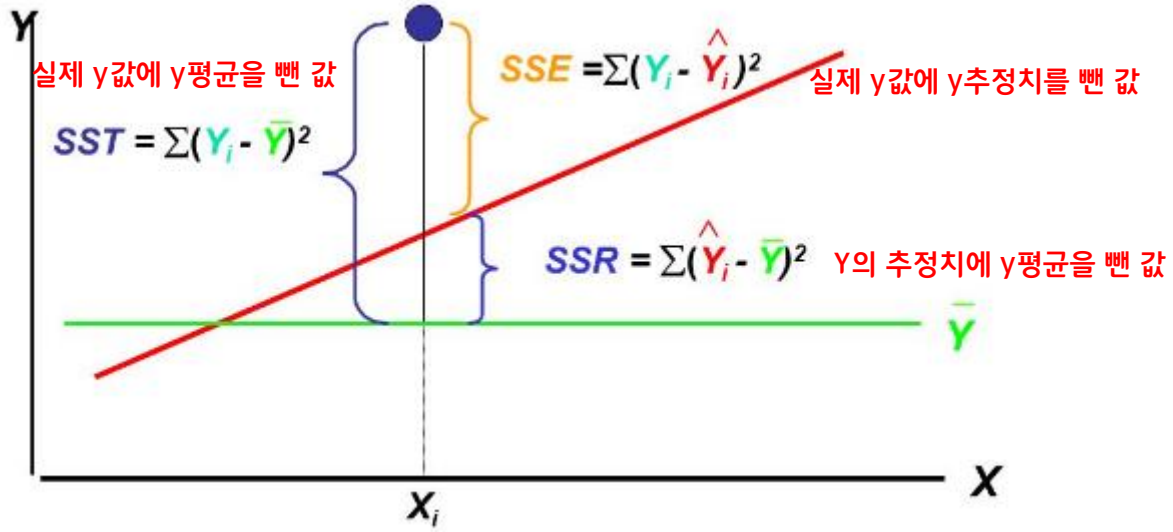
스케터 플랫 상에 점 하나는 회귀직선으로부터 평균적으로 얼마나 떨어져 있느냐?

# 회귀모형의 유의성

SST = SSR + SSE

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

SST = SSR + SSE



# 회귀모형의 유의성

SST = SSR + SSE

자유도

$$d.f. : \quad (n-1) = 1 + (n-2)$$

SST = SSR + SSE

실제 y값에 y평균을 뺀 값 — Y의 추정치에 y평균을 뺀 값

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

실제 y값에 y추정치를 뺀 값

SST(총제곱합)에 x에 대한 정보가 없다
관측치에 설명변수의 도움 없어도 계산이 가능한 식이다

$\beta_0 + \beta_1 X$

Y추정치는 원래 이 식
: 설명변수 x가 있어야 계산되는 값

결론 : 설명변수 x가 반응변수 y를 잘 설명해 주어야 SSE가 작아지며 SSE가 작아지면 SSR과 SST의 차이가 줄어든다

$$SST = SSR + SSE$$

SST (Total Sum of Squares) 총제곱합
SSR (Regression Sum of Squares) 회귀제곱합
SSE (Error Sum of Squares) 잔차제곱합

# 회귀모형의 유의성

$$F_0 = MSR/MSE$$

$$\frac{\text{회귀에 의한 평균 제곱합}}{\text{잔차에 의한 평균 제곱 합}} = \frac{\text{회귀에 의해 설명되는 부분}}{\text{잔차에 의해 설명되는 부분}}$$

회귀의 분산분석표

제곱합을 자유도로 나누면 평균제곱

| 요인 | 제곱합 | 자유도 | 평균제곱 | $F$비 |
|---|---|---|---|---|
| 회귀 | $SSR$ 회귀 제곱합 | 1 | $MSR = SSR/1$ 회귀에 의한 평균제곱합 | $F_0 = MSR/MSE$ |
| 오차 | $SSE$ 잔차 제곱합 | $(n-2)$ | $MSE = SSE/(n-2)$ 잔차에 의한 평균 제곱 합 | |
| 전체 | $SST$ 총 제곱합 | $(n-1)$ | | |

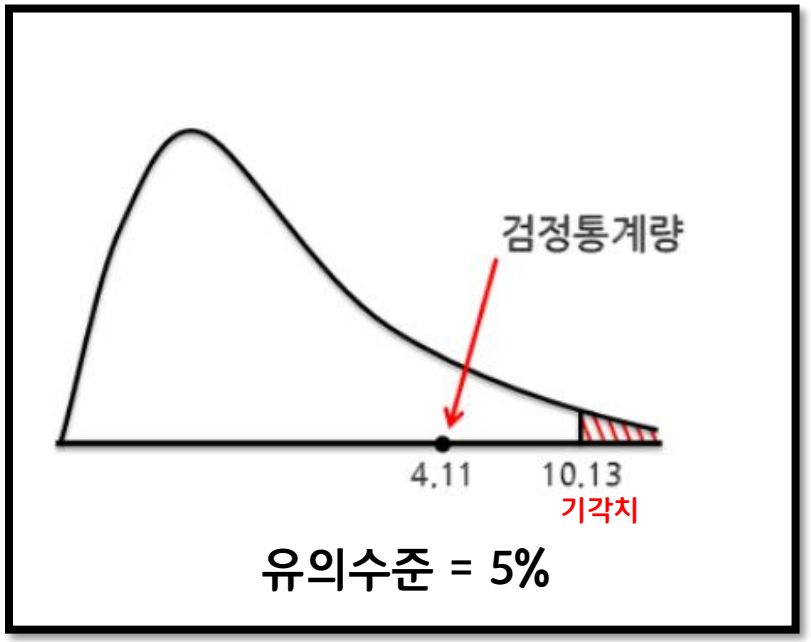
# 회귀모형의 유의성

## 모델의 유의성

$$F_0 = MSR/MSE$$

$$\frac{\text{회귀에 의한 평균 제곱합}}{\text{잔차에 의한 평균 제곱 합}} = \frac{\text{회귀에 의해 설명되는 부분}}{\text{잔차에 의해 설명되는 부분}}$$

귀무가설 하에서 F비 값은 자유도가 p-1, n-p인 F분포를 따릅니다.
F비가 F_a 보다 크면 귀무가설 기각하게 됩니다.
즉 F값이 이런 기각치보다 크면 귀무가설 기각하게 됩니다.
F_a라는 것은 F분포의 오른쪽 꼬리 끝의 면적이 alpha되는 값입니다.
이 값보다 크면 귀무가설을 기각하게 되는데, 이는 회귀관계가 없다는 사실
(설명변수가 반응변수를 제대로 설명 못한다는 사실)을 기각하는 것입니다.
따라서, 귀무가설을 기각하면 "회귀관계가 있다". "설명을 제대로 하고 있다"라는 것입니다.
그러기 위해서 F값이 이 값보다 커야 합니다.
F값이 크다는 얘기는 분모 값에 비해서 분자 값이 크다는 것입니다.
분자 값이 크다는 것은 회귀제곱합의 값이 커야 된다는 것입니다.
즉 설명을 하지 못하는 오차제곱합의 값이 작아야 됩니다.
SSR이 매우 커야 귀무가설을 기각하게 되고, 결과적으로 F값이 커지기 때문입니다.




이 상황에서는 귀무가설을 기각함으로써
모델이 유의하지 않다고 볼 수 있겠죠?

P값은 회귀계수가 유의한지(종속변수와 회귀관계가 있는지) 보고 F값은 모델 자체가 유의한지, 즉, 가설검정을 할 때 쓰인다

# 회귀모형의 적합도

$$F_0 = MSR/MSE$$

$$\frac{\text{회귀에 의한 평균 제곱합}}{\text{잔차에 의한 평균 제곱 합}} = \frac{\text{회귀에 의해 설명되는 부분}}{\text{잔차에 의해 설명되는 부분}}$$

결정계수 (coefficient of determination)

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

$$\frac{\text{회귀 제곱합}}{\text{총 제곱합}}$$

전체 변동 중에서 회귀에 의해 설명되는 부분의 비

적합이 잘 될수록(설명변수가 잘 설명할 수록) 결정계수가 1에 가까워진다(0과 1사이)

# 다중선형회귀모형

## 다중회귀모형

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

단순회귀와 가장 큰 다른 점
: 행렬로 표현

$$y = X\beta + \varepsilon$$

행렬로 표현 가능

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_1' \\ x_2' \\ \cdot \\ \cdot \\ \cdot \\ x_n' \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \cdot & \cdots & \cdots & \cdots & \cdots \\ \cdot & \cdots & \cdots & \cdots & \cdots \\ \cdot & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

# 다중선형회귀모형

## 다중회귀모형

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

## 다중회귀모형의 LSE(최소제곱추정치) 구하는 법

잔차 제곱 다 더한 것

$$D = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon$$

똑같이 잔차 최소화 해주어야 한다

$$y = X\beta + \varepsilon$$

$$\| \quad \|$$

$$(y - X\beta)^T (y - X\beta)$$

잔차 제곱 나타낸 것

$$y^T y - 2\beta^T X y + X^T \beta^T \beta X$$

잔차 제곱 풀어쓴 것

베타로 편미분

$$\frac{\partial D}{\partial \beta} = -2X^T y + 2X^T X\beta$$

결과

$$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

위의 식 정리

# 다중선형회귀모형

## 다중회귀모형

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

SS : $SST = SSR + SSE$

d.f. : $(n-1) = (p-1) + (n-p)$

중회귀에서의 분산분석표

| 요인 | 제곱합 | 자유도 | 평균제곱 | $F$비 |
|---|---|---|---|---|
| 회귀 | $SSR$ | $p-1$ | $MSR = SSR/(p-1)$ | $F_0 = MSR/MSE$ |
| 오차 | $SSE$ | $n-p$ | $MSE = SSE/(n-p)$ | |
| 전체 | $SST$ | $n-1$ | | |

단순 회귀 모형과 같은 방식입니다

# 회귀분석의 보고

## 회귀분석 결과의 보고

단순회귀분석 결과의 보고

$$\hat{y} = a + b\,x$$
$$(SE(a)) \quad (SE(b))$$

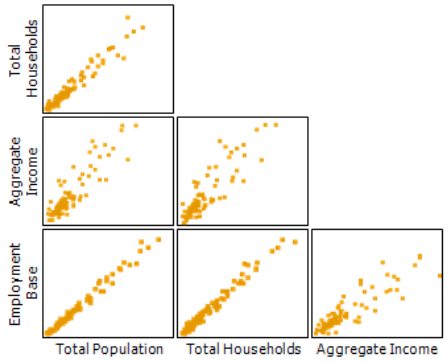
(단, 괄호 안은 표준오차)

관측치수=n, 결정계수=$R^2$, 추정의 표준오차=RMSE

회귀분석 결과 y는 a+bx로 추정되며, 95% 신뢰구간은 표준오차를 고려하여
절편 a는 a + -2SE(a) 이고 회귀계수 b는 b+-2SE(b) 으로 판명됩니다
또한 F값이 기각치를 넘었느냐에 따라 모델이 유의한지 판명할 수 있을 것 같습니다
신뢰구간 b+-2SE(b)에 0이 포함되지 않으므로, x라는 변수는 이 식에서 유의하다고 볼 수 있습니다 또한 관측치는 n개이고, 이 회귀식의 설명력은 $R^2$이며
관측치 하나가 회귀직선에서 평균적으로 떨어져 있는 거리는 RMSE라고 할 수 있습니다

# 기타

## 다중공선성




다중공선성(multicollinearity)란 독립 변수의 일부가 다른 독립 변수의 조합으로 표현될 수 있는 경우이다. 독립 변수들이 서로 독립이 아니라 상호상관관계가 강한 경우에 발생한다.

독립 변수가 서로 의존하게 되면 이른바 과최적화(over-fitting) 문제가 발생하여 회귀 결과의 안정성을 해치게 된다. 이를 방지하는 방법들은 다음과 같다.

해결책
변수 선택법으로 의존적인 변수 삭제
PCA(principal component analysis) 방법으로 의존적인 성분 삭제
정규화(regularized) 방법 사용

# 기타

VIF(Variance inflation Factor)

$$\mathrm{VIF}_i = \frac{\sigma^2}{(n-1)\mathrm{Var}[X_i]} \cdot \frac{1}{1-R_i^2}$$

독립변수를 다른 독립변수로 선형회귀한 성능을 나타낸 것

$R^2$가 크다는 것은 한 독립변수가 다른 독립변수에 대해 설명력이 크다는 것을 뜻한다

다중 공선성을 없애는 가장 기본적인 방법은 다른 독립변수에 의존하는 변수를 없애는 것이다. 가장 의존적인 독립변수를 선택하는 방법으로는 VIF(Variance Inflation Factor)를 사용할 수 있다. VIF는 독립변수를 다른 독립변수로 선형회귀한 성능을 나타낸 것이다

보통 10이상이면 다중공선성이 있다고 판단한다

# 기타

## 조정 결정 계수

$$R^2_{adj} = 1 - \frac{n-1}{n-K}(1-R^2) = \frac{(n-1)R^2 + 1 - K}{n-K}$$

K(독립 변수의 개수)가 늘어날 수록 결정계수에 패널티(감소)를 주는 것

선형 회귀 모형에서 독립 변수가 추가되면 결정 계수의 값은 항상 증가

그러나 독립 변수가 한없이 많은 모델은 좋은 모델이 아님.

독립 변수 추가 효과를 상쇄시키기 위한 다양한 기준들이 제시.

그 중 하나가 다음과 같이 독립 변수의 갯수 K에 따라 결정 계수의 값을 조정하는 조정 결정 계수

# 기타

## 레버리지(leverage)

$$\hat{y} = Hy$$

영향도 행렬(influence matrix) 또는 hat 행렬(hat matrix)

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{iN}y_N$$

H의 i번째 행, j번째 열 성분을 hij라고 하면 실제 결과값 yi과 예측값 y^i은 다음과 같은 관계를 가짐

레버리지(leverage)는 실제 종속변수값 y가 예측치(predicted target) y^에 미치는 영향을 나타낸 값이다

레버리지는 실제의 결과값 $y_i$이 예측값 $\hat{y}_i$에 미치는 영향, 즉 예측점을 자기 자신의 위치로 끌어 당기는 정도를 나타낸 것이다.

레버리지는 수학적으로 영향도 행렬의 대각성분 $h_{ii}$으로 정의된다.

레버리지가 크고 오차가 작은 표본은 오류가 아님

레버리지가 크고 오차가 큰 표본은 잘못된 오류를 만들어 내므로 제거하는 것이 좋다

# 기타

information criterion

$$\mathrm{AIC} = -2\log L + 2K$$

$$\mathrm{BIC} = -2\log L + K\log n$$

AIC, BIC를 최소화하는 것이 최적의 모델인데, K(변수의 개수)를 늘릴때 패널티(증가)를 해준다

BIC가 AIC보다 변수의 개수에 더 많은 패널티를 준다

조정 결정 계수와 함께 많이 쓰이는 모형 비교 기준은 최대 우도에 독립 변수의 갯수에 대한 손실(penalty)분을 반영하는 방법이다. 이를 정보량 규준(information criterion)이라고 하며 손실 가중치의 계산 법에 따라 AIC (Akaike Information Criterion)와 BIC (Bayesian Information Criterion) 두 가지를 사용한다.

AIC와 BIC 둘 다 값이 작을 수록 올바른 모형에 가깝다.

# 기타

## 더미 변수

**더미 변수는 범주형 변수를 0 또는 1만으로 표현되는 값으로**
**어떤 특징(feature)이 존재하는가 존재하지 않는가를 표시하는 독립 변수이다**

모형 : $\hat{y} \sim 1 + x_1 + x_2 + \cdots + x_D$ (모든 데이터에 대해)

$x =$" 남자 " $\rightarrow d = (1,0)$
$x =$" 여자 " $\rightarrow d = (0,1)$

모형 : $\hat{y} = w_{d1}d_1 + w_{d2}d_2 + w_2x_2 + \cdots + w_Dx_D$

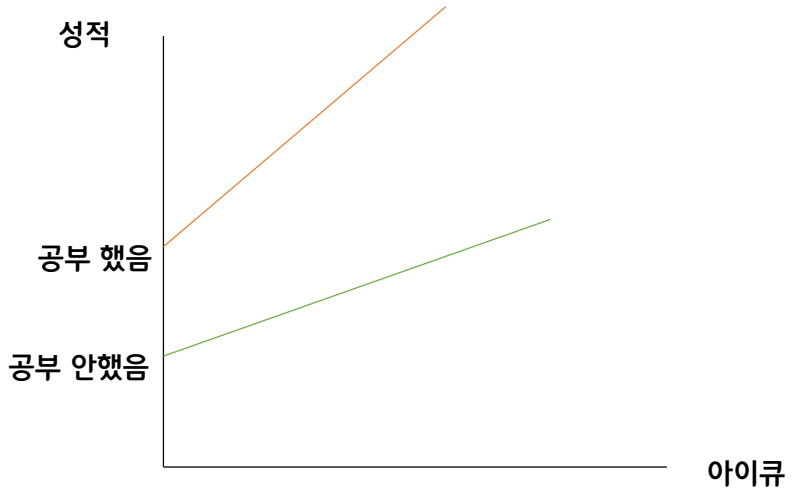
남자 $\hat{y} = w_{d1} + w_2x_2 + \cdots + w_Dx_D$

여자 $\hat{y} = w_{d2} + w_2x_2 + \cdots + w_Dx_D$

# 기타

## 상호작용이 있는 경우

만약 범주형 변수의 값이 달라질 때 상수항만 달라지는 것이 아니라 다른 독립 변수들이 미치는 영향도 달라지는 모형을 원한다면 상호작용(interaction)을 쓰면 된다. 예를 들어 범주형 입력 변수 x1과 연속값 입력 변수 x2를 가지는 회귀모형에서 연속값 입력 변수 x2가 미치는 영향 즉 가중치가 범주형 입력 변수 x1의 값에 따라 달라진다면 범주형 입력 변수를 더미 변수 d1으로 인코딩하고 연속값 입력 변수 x2는 d1과의 상호작용 항을 추가하여 사용한다. *그러나 상호작용 항은 반드시 다중공선성 발생




모형 : $\hat{y} = w_{d1}d_1 + w_{d2}d_2 + w_{d1,2}x_2 + w_{d2,2}d_2x_2$

공부 안했음 공부 했음 아이큐 공부 하면서 아이큐 높음

공부 안했음 $\hat{y} = w_{d1} + w_{d1,2}x_2$

공부 했음 $\hat{y} = w_{d2} + (w_{d1,2} + w_{d2,2})x_2$

# 과제

## <서술형 문제>

1. 회귀분석 결과에서 모형 진단할 때 쓰이는 5가지 선형회귀분석의 전제조건 (선형성, 정규성, 등분산성, 독립성, 비상관성) 에 대해서 간략하게 서술해주세요.

*9기 강의자료 참고 가능

*예측에 큰 고려사항이 아니지만 추론에 중요한 사항이므로 그 부분 생각하면서 공부해보기

2. 다음의 회귀분석 결과를 말로 풀어서 서술해주세요.
(모형 유의성, 적합성, 회귀계수 유의성 및 신뢰수준 ,
설명력 등 포함)

예시) 이 ~~ 수치로 보아 이 모델의 설명력은 ~~인 것 같으며…

제출은 워드, 한글, 파워포인트 등으로 첨부 부탁드립니다

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | target | R-squared: | 0.518 |
| Model: | OLS | Adj. R-squared: | 0.507 |
| Method: | Least Squares | F-statistic: | 46.27 |
| Date: | Wed, 16 Jan 2019 | Prob (F-statistic): | 3.83e-62 |
| Time: | 02:42:35 | Log-Likelihood: | -2386.0 |
| No. Observations: | 442 | AIC: | 4794. |
| Df Residuals: | 431 | BIC: | 4839. |
| Df Model: | 10 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 152.1335 | 2.576 | 59.061 | 0.000 | 147.071 | 157.196 |
| X1 | -10.0122 | 59.749 | -0.168 | 0.867 | -127.448 | 107.424 |
| X2 | -239.8191 | 61.222 | -3.917 | 0.000 | -360.151 | -119.488 |
| X3 | 519.8398 | 66.534 | 7.813 | 0.000 | 389.069 | 650.610 |
| X4 | 324.3904 | 65.422 | 4.958 | 0.000 | 195.805 | 452.976 |
| X5 | -792.1842 | 416.684 | -1.901 | 0.058 | -1611.169 | 26.801 |
| X6 | 476.7458 | 339.035 | 1.406 | 0.160 | -189.621 | 1143.113 |
| X7 | 101.0446 | 212.533 | 0.475 | 0.635 | -316.685 | 518.774 |
| X8 | 177.0642 | 161.476 | 1.097 | 0.273 | -140.313 | 494.442 |
| X9 | 751.2793 | 171.902 | 4.370 | 0.000 | 413.409 | 1089.150 |
| X10 | 67.6254 | 65.984 | 1.025 | 0.306 | -62.065 | 197.316 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.506 | Durbin-Watson: | 2.029 |
| Prob(Omnibus): | 0.471 | Jarque-Bera (JB): | 1.404 |
| Skew: | 0.017 | Prob(JB): | 0.496 |
| Kurtosis: | 2.726 | Cond. No. | 227. |

# 과제

## <데이터 분석>

1. 아파트 낙찰가 예측 문제, 데이터 전처리 및 feature selection, 시각화를 하고 RMSE를 k-fold validation으로 검증해주세요.

1) 파생변수는 3개 이상 만들어주세요
2) 마지막 최종 모델에 사용한 변수는 모두 레이블과의 분포를 시각화를 하셔야 합니다
3) 모든 전처리와 feature selection은 명확한 근거가 있어야 합니다(선택된 변수와 선택되지 않은 변수에 대한 명확한 설명 必)
4) Auction_master_train.csv는 사용해야 하며 나머지 데이터 사용 유무는 자유입니다
5) 모델은 회귀분석(정규화 된 모델 포함)이어야 하며, 예측 외에 필요로 쓰는 모델은 자유입니다
6) 언어는 자유입니다(R, python)

# Reference

1) 류근관 교수님의 경제통계학(k-mooc)

http://www.kmooc.kr/courses/course-v1:SNUk+SNU212_204_1k+2018_T2/about

http://www.kmooc.kr/courses/course-v1:SNUk+SNU212_204_2k+2018_T1/about

http://www.kmooc.kr/courses/course-v1:SNUk+SNU212_204_3k+2018_T2/about

2) 김충락 교수님의 R을 활용한 통계학개론(k-mooc)

http://www.kmooc.kr/courses/course-v1:SNUk+SNU212_204_2k+2018_T1/about

3) 데이터사이언스스쿨

https://datascienceschool.net/view-notebook/661128713b654edc928ecb455a826b1d/


들어주셔서 감사합니다.