# 회귀분석팀

6팀
유종석
윤경선
채소연
김진혁
안은선

# INDEX

# 1

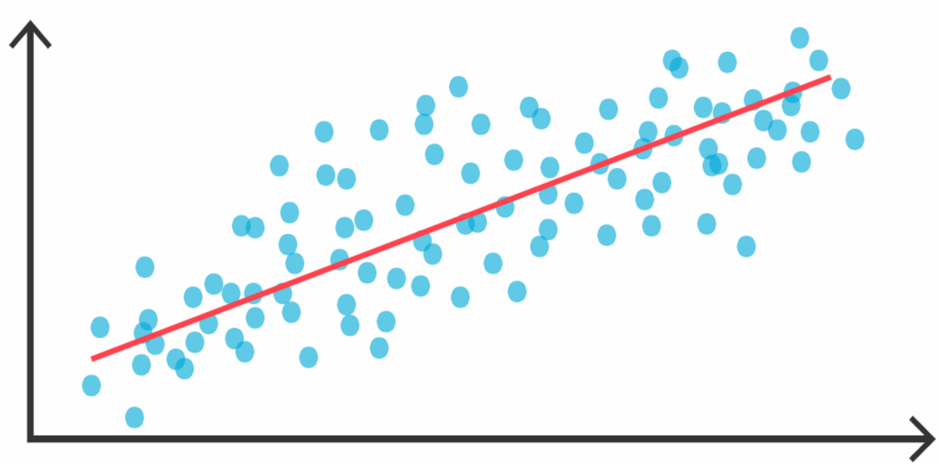
## 회귀분석이란?

# 1 회귀분석이란?

## 회귀분석의 정의

변수들 간의 상관관계를 파악하고,

특정 변수의 값을 다른 변수들을 이용해 설명하고 예측하는 기법



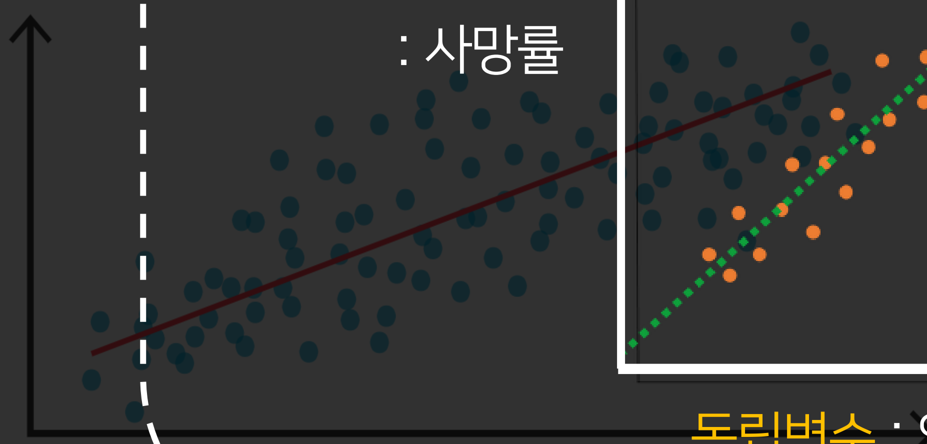
### 회귀분석의 목적

- ✔ 변수들 간의 관계에 대한 표현
- ✔ 독립변수에 따른 종속변수의 변화 파악
- ✔ 미래 관측값에 대한 예측

# 1 회귀분석이란?

## 회귀분석의 정의

변수들 간의 상관관계를 파악하고,
특정 변수의 값을 다른 변수들을 이용해 설명하고 예측하는 기법



Ex) 암 발병률과 사망률의 관계

종속변수
: 사망률


독립변수 : 암 발병률

회귀분석의 목적

- ✔ 변수들 간의 관계에 대한 표현
- ✔ 독립변수에 따른 종속변수의 변화 파악
- ✔ 미래 관측값에 대한 예측

## 회귀식

종속변수 Y와 독립변수 X의 관계를 함수식으로 표현

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

$Y$ 종속변수 : 독립변수에 의해 설명되는 변수

$X_i$ 독립변수 : 종속변수를 설명하기 위한 설명변수

$\varepsilon$ 오차항 : 변수 측정 시 발생할 수 있는 오차로 무작위성을 지님

## 상관분석과의 차이

### 상관분석

- ✔ 두 변수의 관계만 표현 가능
- ✔ 두 변수의 선형적 정도만 표현 가능

**변수에 대한 구체적인 예측과 설명이 불가능**

### 회귀분석

- ✔ 변수들의 상관관계에 기반한 모델
- ✔ 독립변수가 한 단계 변할 때마다 종속변수가 어떻게 변화하는지를 알 수 있음

# 1 회귀분석이란?

## 상관분석과의 차이

### 상관분석

✔ 두 변수의 관계만 표현 가능
✔ 두 변수의 선형적 정도만 표현 가능

변수에 대한 구체적인 예측과 설명이 불가능

### 회귀분석

✔ 변수들의 상관관계에 기반한 모델
✔ 독립변수가 한 단계 변할 때마다 종속변수가 어떻게 변화하는지를 알 수 있음

더 유의미한 관계 파악이 가능하기 때문에 회귀분석을 사용!

## 회귀 모델링 과정

① 문제 정의

나의 학점을 잘 표현할 수 있는 변수들은 무엇일까?

② 적절한 변수 선택

공부시간, 통학거리, 운동시간, 아침밥 여부 등등

③ 데이터 수집 및 전처리

학점, 공부시간, 집에서 학교까지의 거리, 주 당 운동 횟수,

아침밥 식사 여부 조사 후 전처리 작업 진행

# 1 회귀분석이란?

## 회귀 모델링 과정

### ① 문제 정의

나의 학점을 잘 표현할 수 있는 변수들은 무엇일까?

### ② 적절한 변수 선택

공부시간, 통학거리, 운동시간, 아침밥 여부 등등

### ③ 데이터 수집 및 전처리

학점, 공부시간, 집에서 학교까지의 거리, 주 당 운동 횟수,

아침밥 식사 여부 조사 후 전처리 작업 진행

# 1 회귀분석이란?

## 회귀 모델링 과정

### ① 문제 정의

나의 학점을 잘 표현할 수 있는 변수들은 무엇일까?

### ② 적절한 변수 선택

공부시간, 통학거리, 운동시간, 아침밥 여부 등등

### ③ 데이터 수집 및 전처리

학점, 공부시간, 집에서 학교까지의 거리, 주 당 운동 횟수,

아침밥 식사 여부 조사 후 전처리 작업 진행

## 회귀 모델링 과정

### ④ 모델 설정과 적합

적절한 회귀분석 모델 선택

(선형/비선형, 단순회귀/다중회귀, 모수/비모수, 일변량/다변량 등)

### ⑤ 모형 평가

설정한 모델이 회귀 가정을 만족하는지 확인

### ⑥ 모형 해석

지금보다 주 당 2시간 더 공부하고, 주 당 운동은 두 번하고,

서울에서 통학하고, 아침밥을 먹는다면 학점이 0.3만큼 오를 것!

# 1 회귀분석이란?

## 회귀 모델링 과정

### ④ 모델 설정과 적합

적절한 회귀분석 모델 선택

(선형/비선형, 단순회귀/다중회귀, 모수/비모수, 일변량/다변량 등)

### ⑤ 모형 평가

설정한 모델이 회귀 가정을 만족하는지 확인

### ⑥ 모형 해석

지금보다 주 당 2시간 더 공부하고, 주 당 운동은 두 번하고,

서울에서 통학하고, 아침밥을 먹는다면 학점이 0.3만큼 오를 것!

# 1 회귀분석이란?

## 회귀 모델링 과정

### ④ 모델 설정과 적합

적절한 회귀분석 모델 선택

(선형/비선형, 단순회귀/다중회귀, 모수/비모수, 일변량/다변량 등)

### ⑤ 모형 평가

설정한 모델이 회귀 가정을 만족하는지 확인

### ⑥ 모형 해석

지금보다 주 당 2시간 더 공부하고, 주 당 운동은 두 번하고, 서울에서 통학하고, 아침밥을 먹는다면 학점이 0.3만큼 오를 것!

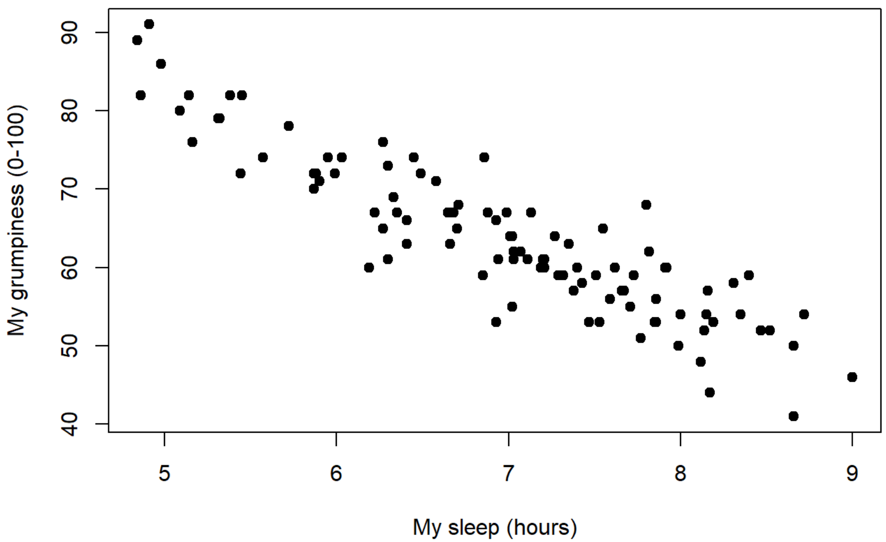
# 2

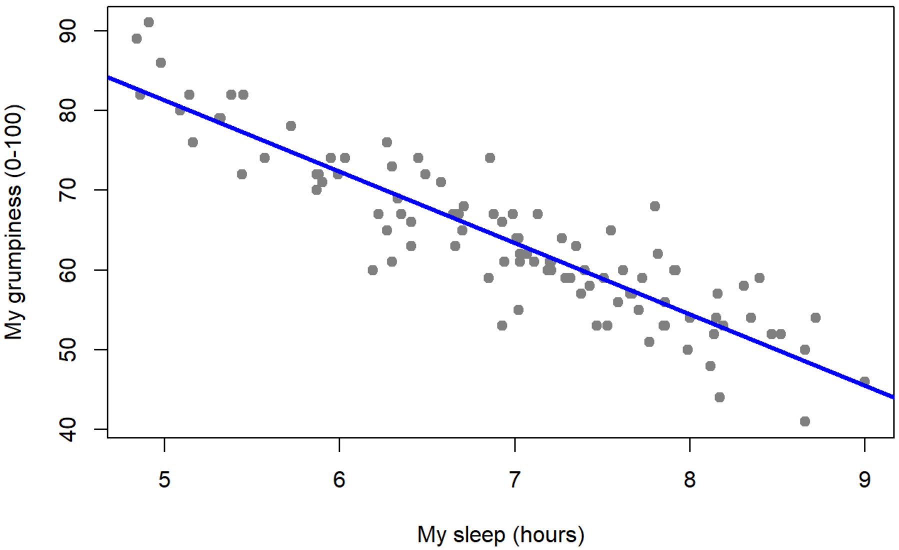
# 단순선형회귀

# 2 단순선형회귀

## 단순선형회귀식

X와 Y, 두 변수의 관계를 가장 잘 표현할 수 있는 직선을 찾아 수식화




두 변수의 관계가 선형적일 것이라는 가정 하에 직선 함수식 추정

## 단순선형회귀 모델

종속변수 Y와 독립변수 X의 관계를 잘 설명할 수 있는 직선

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

이때 $\epsilon_i \sim N(0, \sigma^2)$ 가정

$y_i$ 종속변수 y의 $i$번째 관측값

$x_i$ 독립변수 x의 $i$번째 관측값

$\beta_0, \beta_1$ 회귀계수 : 우리가 추정해야 할 모수

[ 좋은 모델을 만들기 위해서는 **회귀계수를 잘 추정**하는 것이 중요! ]

$\epsilon_i$ 오차항 : $i$번째 관측값에 의한 무작위적인 오차

## 단순선형회귀 해석

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

이때 $\epsilon_i \sim N(0, \sigma^2)$ 가정

**단순선형회귀 모델의 해석**

$X$가 한 단위 증가할 때, $Y$가 평균적으로 $\beta_1$만큼 증가

## 왜 직선인가?

### 직선을 이용할 경우

변수의 영향력을 간단하게 모형화 가능

독립변수의 변화에 따른 종속변수의 변화를 직관적으로 확인 가능

### 고차근사를 할 경우

과적합(overfitting) 문제의 원인이 됨

## 왜 직선인가?

### 직선을 이용할 경우

변수의 영향력을 간단하게 모형화 가능

독립변수의 변화에 따른 종속변수의 변화를 직관적으로 확인 가능

### 고차근사를 할 경우

과적합(overfitting) 문제의 원인이 됨

## 왜 직선인가?

직선을 이용할 경우

### 과적합(Overfitting)이란?

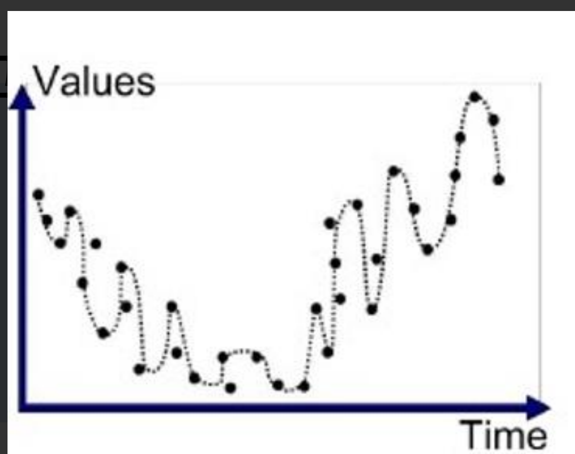
변수의 영향력을 간단하게 모형화 가능

: Train data에 대한 설명성은 높을 수 있으나

독립변수의 변화에 따른 종속변수의 변화를 직관적으로 확인 가능

Test data에 대한 설명성은 떨어진다는 의미

고




- 모델의 분산을 높임

ing) 문제의 원인이 됨

- 검증 데이터의 예측 성능 저하

데마팀 1주차 클린업 참고

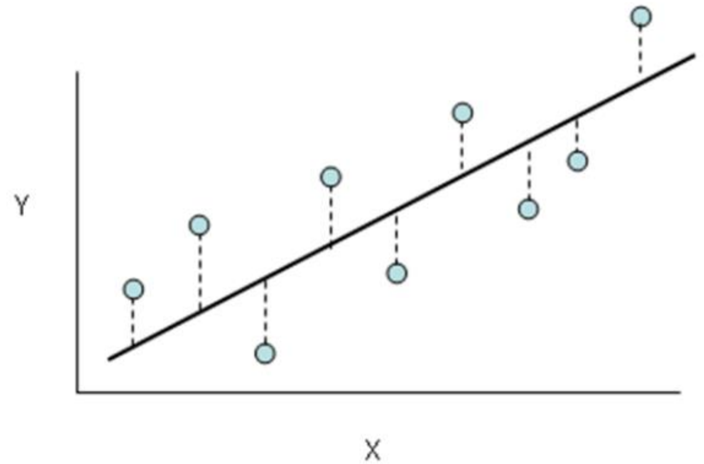
## 최소제곱법(LSE : Least Square Estimation Method)

**좋은 추정이란,**

우리가 만들어 낸 회귀직선과 관측치 사이의 오차가 최소화 되는 경우

**최소제곱법(LSE)**

오차의 제곱합을 최소화하여 모수를 추정하는 방법



## 최소제곱법을 이용한 모수 추정

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

오차 제곱합을 최소화 시키는 $\beta_0, \beta_1$를 찾는 것이 목적!

아래로 볼록한 Convex함수 → 각각의 모수를 편미분하여

'미분값=0'을 만족시키는 $\beta_0, \beta_1$을 구함

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

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

## 최소제곱법을 이용한 모수 추정

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}, \qquad \widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

$$where\ \bar{y} = \sum_{i=1}^{n} y_i/n,\ \bar{x} = \sum_{i=1}^{n} x_i/n,\ \ S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})\ and\ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

**최소제곱추정치**

(LSE : Least Square Estimator)

## 최소제곱법을 이용한 모수 추정

### 왜 오차의 '제곱합'을 최소화할까?

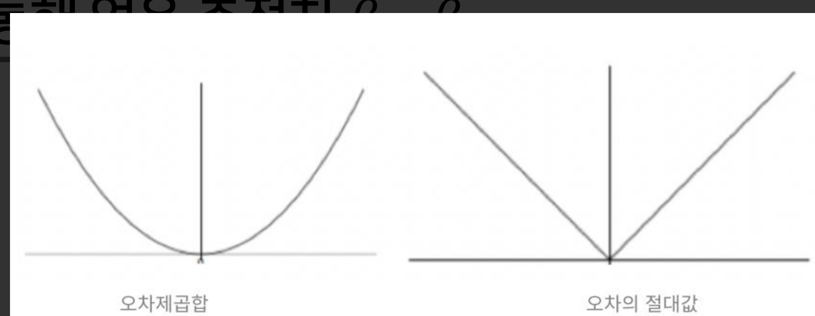
$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}, \qquad \widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

$$where\ \bar{y} = \sum_{i=1}^{n} y_i/n,\ \bar{x} = \sum_{i=1}^{n} x_i/n,\ S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})\ and\ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

① 미분의 편리성

② 오차가 클수록 더 큰 패널티 부여 가능

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




오차의 절대값 사용 시 미분 불가능한 점이 존재

## 최소제곱법의 가정과 특징

BLUE (Best Linear Unbiased Estimator)

분산이 제일 작은 선형 불편추정량

→ 분산이 작다는 것은 추정량이 안정적이라는 의미

① 오차들의 평균은 0

② 오차들의 분산은 $\sigma^2$으로 동일(등분산)

③ 오차 간에 자기상관 X (uncorrelated)

위 3가지 조건을 만족하면 LSE는 분산이 제일 작은 선형 불편추정량이 됨

## 최소제곱추정량(LSE) VS 최대가능도추정량(MLE)

### 최대가능도 추정 (MLE: Maximum Likelihood Estimator)

확률적인 방법에 근거해서, 데이터가 나올

'가능도'를 최대로 하는 모수를 선택하는 방법

① 추정 관측치가 항상 iid라는 가정 필수

② 오차의 정규분포 가정할 때

LSE와 MLE는 완전히 동일한 추정량을 산출

## 적합성(Goodness of fit) 검정

**모델의 적합성에 대한 평가 과정**

(회귀식이 얼마나 데이터를 잘 설명하는지 검정)

### 잔차(Residual)

✓ 추정한 회귀계수를 이용해 회귀직선을 만들었을 때 오차의 추정량을 의미

✓ 모집단(오차), 표본(잔차)라고 명칭만 변경

✓ $e_i = y_i - \widehat{y_i} = y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i), \ \sum e_i = 0$

## 적합성(Goodness of fit) 검정

### 변동 분할

- ✓ SST (Total Sum of Squares, 총 변동) : $\sum(y_i - \bar{y})^2$
- ✓ SSR (Regression Sum of Square, 회귀선이 설명하는 변동) : $\sum(\hat{y}_i - \bar{y})^2$
- ✓ SSE (Residual Sum of Square, 잔차제곱합,
  회귀선이 설명하지 못하는 변동) : $\sum(y_i - \hat{y}_i)^2$

$$SST = SSR + SSE$$

## 적합성(Goodness of fit) 검정

### 결정계수 $R^2$

총변동(SST)에서 회귀식이 설명할 수 있는 비율(SSR)

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

결정계수 $R^2$ 가 1에 가까울수록 좋음

## 유의성 검정

개별 모수의 추정량이 통계적으로 유의한지를 알아보는 과정

$\epsilon_i \sim N(0, \sigma^2)$라는 오차의 정규분포 가정 아래에서

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

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 : $t_{(1-\frac{\alpha}{2}, n-2)}$

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

# 2 단순선형회귀

## 유의성 검정


개별 모수의 추정량이 통계적으로 유의한지를 알아보는 과정

$\epsilon_i \sim N(0, \sigma^2)$라는 오차의 정규분포 가정 아래에서

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

② 추정량의 분포 : $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

③ 검정 통계량 : $t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{(n-2)}$

④ 임계값 : $t_{(1-\frac{\alpha}{2}, n-2)}$

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

귀무가설을 기각하지 못하더라도

X와 Y 사이에 선형적 관계가 없을 뿐,

비선형적인 관계가 있을 수도 있음

# 3

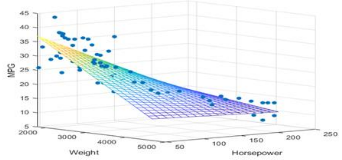
# 다중선형회귀

## 다중선형회귀

### 단순선형회귀

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

독립 변수 1개, 회귀 계수 2개




### 다중선형회귀

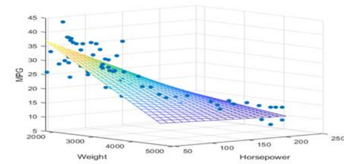
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

독립 변수 p개, 회귀 계수 p+1개

① 단순선형회귀에 비해 **복잡한 관계** 설명 가능

② 자연현상, 사회현상 파악에 유리

# 다중선형회귀




## 단순선형회귀

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

독립 변수 1개, 회귀 계수 2개

## 다중선형회귀

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

독립 변수 p개, 회귀 계수 p+1개

**다중선형회귀 모델의 해석**

나머지 X 변수들이 고정되었을 때,

$x_p$가 한 단위 증가하면 $y$는 $\beta_p$만큼 증가함

## 모수의 추정

단순선형회귀와 동일한 방식으로 모수의 추정치 산출 가능

$$S(\beta) = \sum_i (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_{pi})^2$$

$$(1)\ \frac{\partial S}{\partial \beta_0} = -2\sum_i \left(y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_{pi}\right) = 0$$

$$\vdots$$

$$(2)\ \frac{\partial S}{\partial \beta_p} = -2\sum_i \left(y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_{pi}\right) x_{pi} = 0$$

## 모수의 추정

단순선형회귀와 동일한 방식으로 모수의 추정치 산출 가능

$$S(\beta) = \sum_i (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_{pi})^2$$

다차원에 대한 표현을 해야 하므로

편미분을 통해 모수 추정 시 계산식이 매우 복잡해짐!

$$(1)\ \frac{\partial S}{\partial \beta_0} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_{pi}) = 0$$

$$(2)\ \frac{\partial S}{\partial \beta_p} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_{pi}) x_{pi} = 0$$

행렬을 이용하자!!

## 모수의 추정: 최소제곱법

$$y = X\beta + \epsilon \iff \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

✓ 다중선형회귀식을 **행렬**로 표현

✓ 단순선형회귀와 동일하게 **최소제곱법** 이용

## 최소제곱법

$$\min\ S(\beta) = \sum_{i=1}^{n} \epsilon^2 = \epsilon^T\epsilon = (y - X\beta)^T(y - X\beta)$$

목적함수 $S$를 $\beta$에 대해 미분하고

해당 미분식을 0으로 만들어주는 추정량 $\hat{\beta}$을 구함!

$$\frac{\partial S}{\partial \beta} = -2X'\left(Y - X\hat{\beta}\right) = 0$$

$$\rightarrow \hat{\beta} = (X'X)^{-1}X'y$$

**최소제곱법을 통해 얻은 추정된 회귀식**

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$$

($H$ = $X(X'X)^{-1}X'$ 는 투영행렬)

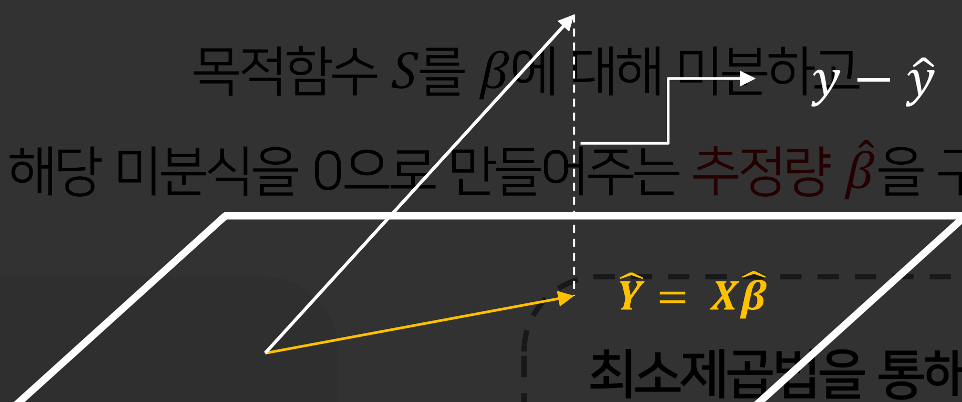
## 최소제곱법

$$\min\ S(\beta) = \sum_{i=1}^{n} e_i^2 = (y - X\beta)^T(y - X\beta)$$

### **투영 행렬**이란?

목적함수 $S$를 $\beta$에 대해 미분하고

해당 미분식을 0으로 만들어주는 추정량 $\hat{\beta}$을 구함!




$$\frac{\partial S}{\partial \beta} = -2X'\left(Y - X\hat{\beta}\right) = 0$$

$$\rightarrow \hat{\beta} = (X'X)^{-1}X'y$$

최소제곱법을 통해 얻은 추정된 회귀식

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$$

($H = X(X'X)^{-1}X'$ 는 투영행렬)

$y$ 를 $X$의 열공간에 가깝게 근사 시키기 위해 사용

→ $y$ 를 $X$의 열공간에 투영시킴으로써 근사해 $\hat{\beta}$ 를 찾음

선대팀 2주차 클린업 예정

## 유의성 검정

추정량이 통계적으로 유의한지 알아보는 검정

1. F-test
2. Partial F-test
3. T-test

## 유의성 검정

1. F-test: 전체 회귀계수에 대한 검정

### 가설 설정

$$\boldsymbol{H_0}: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$\boldsymbol{H_1}$: $\beta_1, \beta_2, \cdots, \beta_p$ 중 적어도 하나는 0이 아니다.

## 유의성 검정

1. F-test: 전체 회귀계수에 대한 검정

### 검정통계량

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

MSR : 평균회귀제곱 MSE : 평균오차제곱

회귀식의 전반적인 계수가 얼마나 설명력을 갖는지를 보여줌

추정된 회귀식의 설명력이 높다는 것은 유의미한 $\hat{\beta}$이 있다는 것을 의미

## 유의성 검정

1. F-test: 전체 회귀계수에 대한 검정

**임계값**

$$F_{(1-a/2\ ,\ p,\ n-p-1)}$$

귀무가설 기각 if $F_0 \geq F_{(1-a/2\ ,\ p,\ n-p-1)}$

➔ 적어도 한 개의 회귀계수는 0이 아님

귀무가설 기각 안 됨 if $F_0 < F_{(1-a/2\ ,\ p,\ n-p-1)}$

➔ 모든 회귀계수는 0임

## 유의성 검정

1. F-test: 전체 회귀계수에 대한 검정

**임계값**

$$F_{(1-a/2\,,\,p,\,n-p-1)}$$

귀무가설 기각 if $F_0 \geq F_{(1-a/2\,,\,p,\,n-p-1)}$

➔ 적어도 한 개의 회귀계수는 0이 아님

귀무가설 기각 안 됨 if $F_0 < F_{(1-a/2\,,\,p,\,n-p-1)}$

➔ 모든 회귀계수는 0임

## 유의성 검정

1. F-test: 전체 회귀계수에 대한 검정

임계값 $F_{(1-a/2\,,\,p,\,n-p-1)}$

귀무가설 기각 if $F_0 > F_{(1-a/2\,,\,p,\,n-p-1)}$

➔ 적어도 한 개의 회귀계수는 0이 아님

귀무가설 기각 안 됨 if $F_0 < F_{(1-a/2\,,\,p,\,n-p-1)}$

➔ 모든 회귀계수는 0임

F-test의 귀무가설이 기각되지 않는다면,

$y = \beta_0 + \epsilon \ (\because \beta_1 = \beta_2 = \cdots = \beta_p = 0)$ 이므로

회귀식이 아무런 의미가 없음

↓

모델 재설정 등의 조치가 필요

## 유의성 검정

2. Partial F-test: 일부 회귀계수에 대한 검정

FM(Full Model)

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon_i$$

➔ 모든 변수를 사용한 다중회귀모형

RM(Reduced Model)

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_{j-1} x_{j-1} + \beta_{j+q} x_{j+q} + \cdots + \beta_p x_p + \epsilon_i$$

➔ 일부 계수를 특정 값으로 둔 회귀모형

*특정 상수값은 0으로 두는 것이 일반적*

## 유의성 검정

2. Partial F-test: 일부 회귀계수에 대한 검정

### 가설 설정

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

➔ RM이 맞음

$\boldsymbol{H_1}$: $\beta_j, \beta_{j+1}, \cdots, \beta_{j+q-1}$ 중 적어도 하나는 0이 아니다.

➔ FM이 맞음

## 유의성 검정

2. Partial F-test: 일부 회귀계수에 대한 검정

검정통계량

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

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

*변수를 제거하면 일반적으로 $SSE(RM) > SSE(FM)$*

## 유의성 검정

2. Partial F-test: 일부 회귀계수에 대한 검정

$q$개의 변수를 제거했을 때

모델이 설명하지 못하는 변동

검정통계량

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

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

*변수를 제거하면 일반적으로 $SSE(RM) > SSE(FM)$*

## 유의성 검정

2. Partial F-test: 일부 회귀계수에 대한 검정

검정통계량

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

모든 변수를 포함했을 때

모델이 설명하지 못하는 변동

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

*변수를 제거하면 일반적으로 $SSE(RM) > SSE(FM)$*

## 유의성 검정

2. Partial F-test: 일부 회귀계수에 대한 검정

검정통계량

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

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

변수를 제거하면 일반적으로 $SSE(RM) > SSE(FM)$

일반적으로 변수를 제거하면 $SSE(RM) > SSE(FM)$

이때, 제거된 변수가 모델에 유의미하다면

$SSE(RM)$는 월등히 커짐

검정통계량 $F_0$

검정통계량이 귀무가설을 기각시킬만큼 충분히 커짐

## 유의성 검정

2. Partial F-test: 일부 회귀계수에 대한 검정

**임계값**

$$F_{(1-a/2\,,\,p-q,\,n-p-1)}$$

귀무가설 기각 if $F_0 \geq F_{(1-a/2\,,\,p-q,\,n-p-1)}$

➔ $q$ 개의 회귀 계수 중 적어도 한 개의 회귀계수도 0이 아님

귀무가설 기각 안 됨 if $F_0 < F_{(1-a/2\,,\,p-q,\,n-p-1)}$

➔ $q$의 회귀계수는 0임

## 유의성 검정

2. Partial F-test: 일부 회귀계수에 대한 검정

### 임계값

$$F_{(1-a/2\,,\,p-q,\,n-p-1)}$$

귀무가설 기각 if $F_0 \geq F_{(1-a/2\,,\,p-q,\,n-p-1)}$

➔ $q$ 개의 회귀 계수 중 적어도 한 개의 회귀계수는 0이 아님

귀무가설 기각 안 됨 if $F_0 < F_{(1-a/2\,,\,p-q,\,n-p-1)}$

➔ $q$ 개의 회귀 계수 중 적어도 한 개의 회귀계수가 0임

## 유의성 검정

3. T-test: 개별 회귀계수에 대한 검정

### 가설 설정

$$\boldsymbol{H_0}: \beta_j = 0$$

➔ 다른 변수들이 다 적합된 상태에서 $x_j$ 는 통계적으로 유의하지 않음

$$\boldsymbol{H_1}: \beta_j \neq 0$$

➔ 다른 변수들이 다 적합된 상태에서 $x_j$ 는 통계적으로 유의함

## 유의성 검정

3. T-test: 개별 회귀계수에 대한 검정

### 검정통계량

$$\boldsymbol{t_j} = \frac{\widehat{\beta_j}}{s.e.(\widehat{\beta_j})}$$

T-test는 $x_j$ 변수 자체가 아니라 해당 변수를 추가적으로 적합했을 때 통계적 유의성을 확인하는 검정

## 유의성 검정

3. T-test: 개별 회귀계수에 대한 검정

**임계값**

$$\boldsymbol{t_{(a/2,\, n-p-1)}}$$

귀무가설 기각 if $|t_j| \geq t_{(a/2,\, n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져옴

귀무가설 기각 안 됨 if $|t_j| < t_{(a/2,\, n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져오지 않음

## 유의성 검정

3. T-test: 개별 회귀계수에 대한 검정

**임계값**

$$\boldsymbol{t_{(a/2,\, n-p-1)}}$$

귀무가설 기각 if $|t_j| \geq t_{(a/2,\, n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져옴

귀무가설 기각 안 됨 if $|t_j| < t_{(a/2,\, n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져오지 않음

## 유의성 검정

## T-test로 변수선택이 가능할까?

3. T-test: 개별 회귀계수에 대한 검정

T-test는 다른 변수들이 다 적합된 상태에서

해당 변수의 추가가 유의미한 설명력 증가를 가져오는지 판단하는 것

**임계값**

$t_{(a/2,\,n-p-1)}$

귀무가설 기각 if $|t_j| \geq t_{(a/2,\,n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져옴

귀무가설 기각 안 됨 if $|t_j| < t_{(a/2,\,n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져오지 않음

## 유의성 검정

3. T-test: 개별 회귀계수에 대한 검정

임계값

$t_{(a/2,n-p-1)}$

귀무가설 기각 if $|t_j| \geq t_{(a/2,n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져옴

귀무가설 기각 안 됨 if $|t_j| < t_{(a/2,n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져오지 않음

### T-test로 변수선택이 가능할까?

T-test는 다른 변수들이 다 적합된 상태에서
해당 변수의 추가가 유의미한 설명력 증가를 가져오는지 판단하는 것

따라서 다른 회귀식을 가정하면 해당 변수의 유의성도 바뀔 수 있음

## 유의성 검정

T-test로 변수선택이 가능할까?

3. T-test: 개별 회귀계수에 대한 검정

임계값

$t_{(a/2,\, n-p-1)}$

T-test로 변수를 선택하는 것은 매우 위험!

귀무가설 기각 if $|t_j| \geq t_{(a/2,\, n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져옴

귀무가설 기각 안 됨 if $|t_j| < t_{(a/2,\, n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져오지 않음

## 유의성 검정

3. T-test: 개별 회귀계수에 대한 검정

# F-test vs. T-test

<T-test 하기 전 F-test를 먼저 해야 하는 이유>

① 전체 회귀식에 대한 검정이 더 엄격함

② F-test를 기각 못해도 T-test는 기각하는 경우 발생 가능

임계값

귀무가설 기각 if $|t_j| \geq t_{(a/2,\, n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져옴

귀무가설 기각 안 됨 if $|t_j| < t_{(a/2,\, n-p-1)}$

➔ $x_j$의 추가는 유의미한 회귀식의 설명력 증가를 가져오지 않음

## F-test vs. T-test

<T-test하기 전 F-test를 먼저 해야 하는 이유>

① 전체 회귀식에 대한 검정이 더 엄격함

② F-test를 기각 못해도 T-test는 기각하는 경우 발생 가능

따라서 F-test를 먼저 시행해 봄으로써

모델 전체가 통계적으로 유의한지 확인해야 함

## 적합성(Goodness of fit) 검정

회귀모델이 데이터에 얼마나 잘 들어맞는지 확인하기 위한 적합성 검정

### 적합성 검정

#### 결정계수

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

#### 수정결정계수

$${R_a}^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

## 적합성(Goodness of fit) 검정의 종류

① 결정계수 (R square)

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

**변수↑ - 결정계수 값 ↑**

총 변동은 고정되어 있는데,

X 변수가 추가되면 회귀식으로 설명되는 변동이 조금이라도 증가

= $R^2$ 값도 증가

무의미한 변수 추가는 모델에 대한 해석도 어렵게 하고 예측에도 좋지 않은 영향을 끼칠 수 있음!

## 적합성(Goodness of fit) 검정의 종류

### ② 수정결정계수 (Adjusted R square)

$$R_a^{\,2} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

**변수의 개수가 다른 두 회귀식을 비교할 때 유용한 지표**

변수가 많은 쪽의 결정계수가 회귀식이 더 유의미한지 여부와 관련 없이 더 높을 수 있다.

변수가 추가됨에 따라 증가하는 결정계수에 변수 개수라는 패널티를 부과한 형태

## 적합성(Goodness of fit) 검정의 종류

**② 수정결정계수 (Adjusted R square)**

$$R_a^{\ 2} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

**변수의 개수가 다른 두 회귀식을 비교할 때 유용한 지표**

변수가 많은 쪽의 결정계수가 회귀식이 더 유의미한지 여부와 관련 없이 더 높을 수 있다.

변수가 추가됨에 따라 증가하는 결정계수에 변수 개수라는 패널티를 부과한 형태

## 적합성(Goodness of fit) 검정의 종류

### ② 수정결정계수 (Adjusted R square)

$$R_a{}^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

변수의 개수가 다른 두 회귀식을 비교할 때 유용한 지표

**활용**

AIC(Akaike Information Criterion)

BIC(Baysian Information Criterion)

변수의 개수가 다른 두 회귀식을

비교할 때 유용하게 사용 가능

변수가 많은 ... 높을 수 있다.

변수가 추 ... 과한 형태

회귀팀 3주차 클린업 예정

## 적합성(Goodness of fit) 검정의 종류

### ② 수정결정계수 (Adjusted R square)

$$R_a^{\,2} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

변수의 개수가 다른 두 회귀식을 비교할 때 유용한 지표

**한계**

결정계수처럼

'전체 변동 중에 회귀식이 설명하는 변동'으로

해석할 수는 없다!

변수가 많은 … 높을 수 있다.

변수가 추 … 무관한 형태

# 4

# 데이터 진단

## 데이터 진단이란?

데이터 중에 **일반적인 경향에서 벗어나는** 점들이 있음

최소제곱 회귀모형을 크게 **바꾸거나**

그에 따라 **성능을 저하**시키기도 함

## 표준화 잔차

### 스튜던트화 잔차

Y값의 단위에 따라 잔차의 차이가 많이 나는 상황을 방지하기 위해

일반화된 상황에서도 적용할 수 있도록 표준화한 것

$\sigma$는 모수이므로 알 수 없어 이에 대한 추정량인 $\hat{\sigma}$을 넣어줌

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

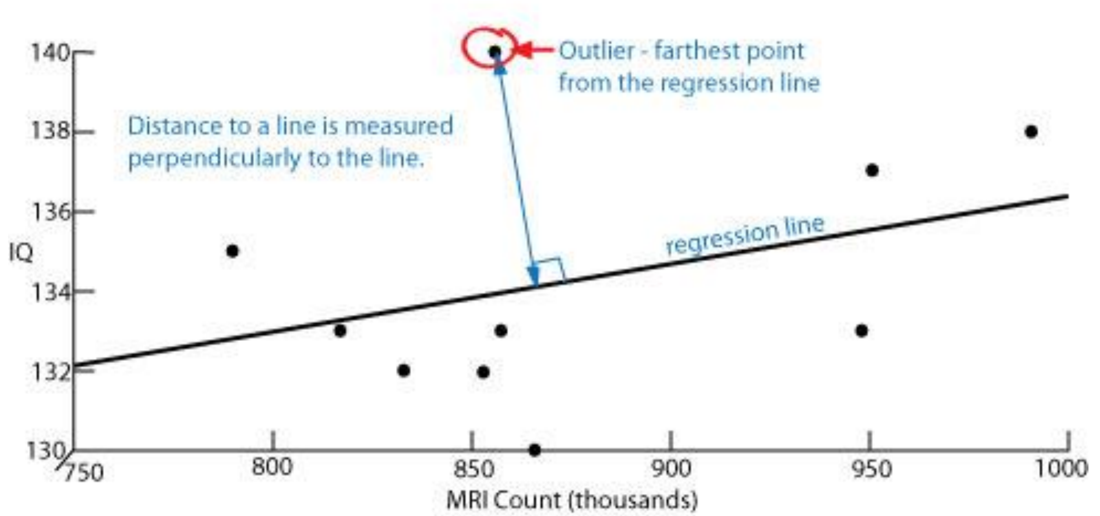
$$\hat{\sigma} = \sqrt{\frac{SSE}{n - p - 1}}$$

## 이상치(Outlier)

### 이상치

표준화 잔차가 매우 큰 값

y를 기준으로 절댓값이 큰 값




보통 $|r_i| > 3$ 이면 이상치라고 판단

## 지렛값(Leverage point)

**지렛값**

$x$ 의 평균 $\bar{x}$ 에서 멀리 떨어져 있어 기울기에 큰 영향을 주는 값

표준화했을 때 $x$의 기준에서 절댓값이 큰 값

$$H = X(X^tX)^{-1}X^t$$

$$h_{ii} = x_i^t\,(X^tX)^{-1}x_i$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

## 지렛값(Leverage point)

### 지렛값

$x$ 의 평균 $\bar{x}$ 에서 멀리 떨어져 있어 기울기에 큰 영향을 주는 값

표준화했을 때 $x$의 기준에서 절댓값이 큰 값

$H = X(X^tX)^{-1}X^t$

$h_{ii} = x_i^t(X^tX)^{-1}x_i$

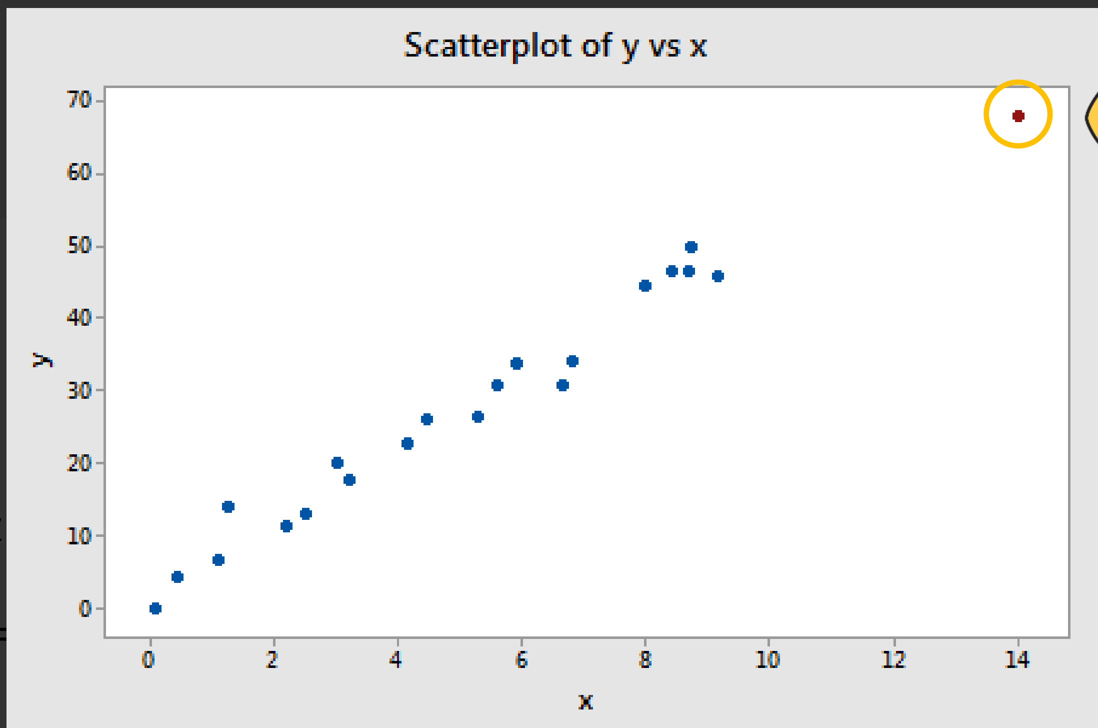
$x_i$값과 $\bar{x}$ 의 차이가 클수록 $h_{ii}$ 가 커진다

= $x$ 평균에서 멀수록 leverage 값이 상승한다

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

# 지렛값(Leverage point)

## 지렛값




Leverage point

$$\left( h_{ii} > \frac{2(p+1)}{n} \right)$$

이면 지렛값으로 판단

$x_i$

$$= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

# 영향점(Influential point)

## 영향점

회귀직선의 기울기에 상당한 영향을 주는 관측치

### Cook’s distance

영향점을 확인하는 표준적인 지표로,

특정 데이터를 지웠을 때 회귀선이 변하는 정도를 가리킴

## 영향점(Influential point)

Outlier와 Leverage point의 진단만으로 회귀직선 변화 파악 X

### Outlier

x평균 주위에 위치할 경우
기울기를 변화시키지 못함

### Leverage Point

회귀직선의
연장선에 있을 수 있음.

## 영향점(Influential point)

영향점

회귀직선의 기울기에 상당한 영향을 주는 관측치

Outlier와 Leverage point의 진단만으로 회귀직선 변화 파악 X

Outlier와 Leverage point를 동시에 고려하는 지표

**Cook's Distance**

## 영향점(Influential point)

### 영향점

회귀직선의 기울기에 상당한 영향을 주는 관측치

### Cook’s distance

영향점을 확인하는 표준적인 지표로,

특정 데이터를 지웠을 때 회귀선이 변하는 정도를 가리킴

## 영향점(Influential point)

### 영향점

회귀직선의 기울기에 상당한 영향을 주는 관측치

### Cook's distance

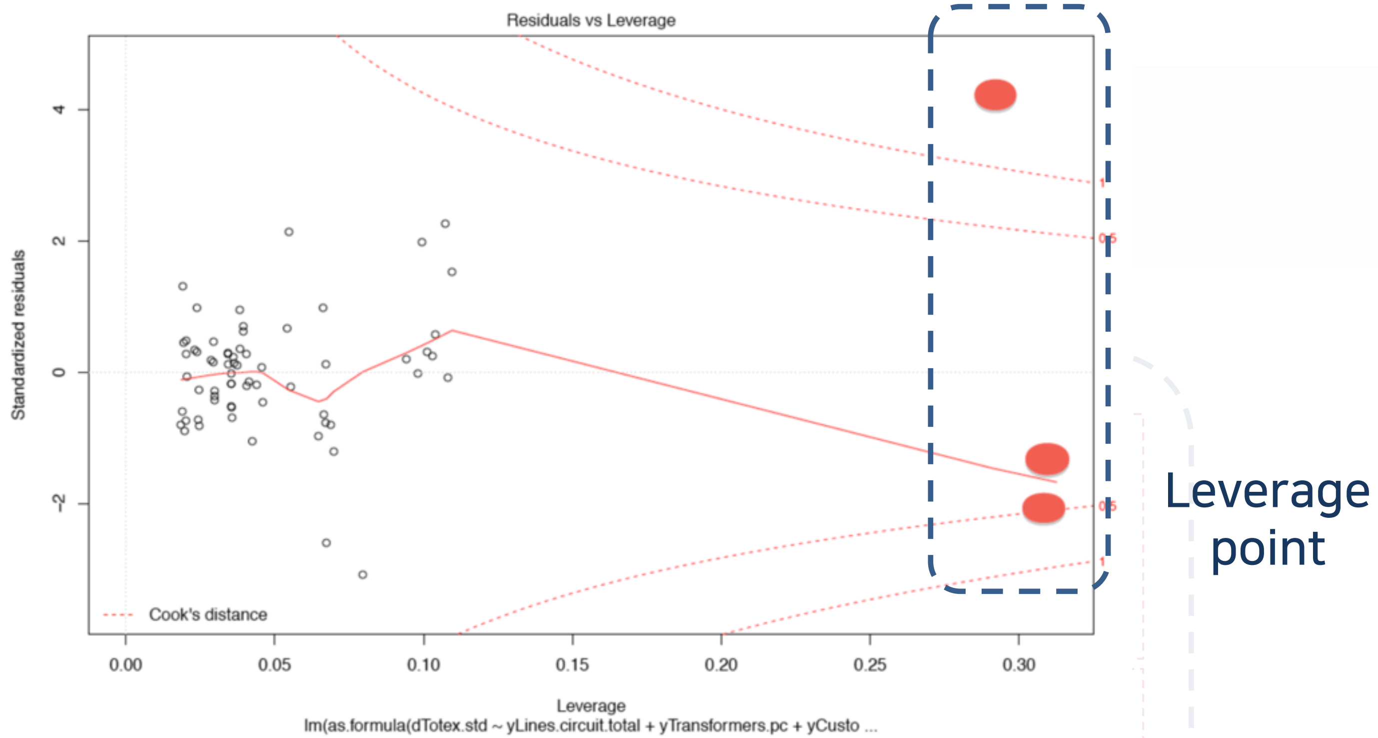
Outlier Leverage

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

✓ Outlier와 Leverage 각각이 커지면 커질수록 $C_i$가 커짐

✓ 보통 $C_i > 1$ 이면 영향점으로 판단

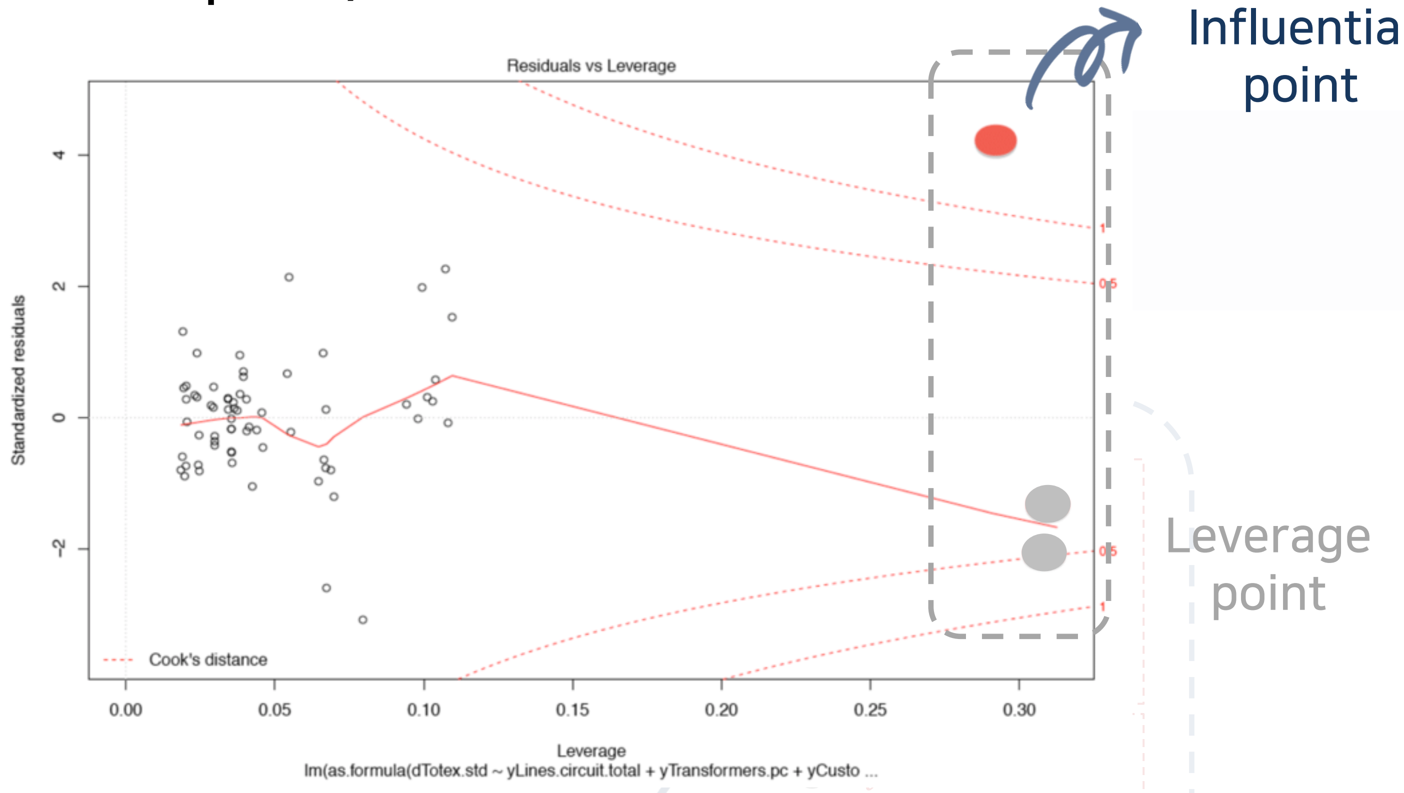
## 영향점(Influential point)




Leverage point라고 하더라도

모두 Influential point인 것은 아님을 확인할 수 있음

## 영향점(Influential point)

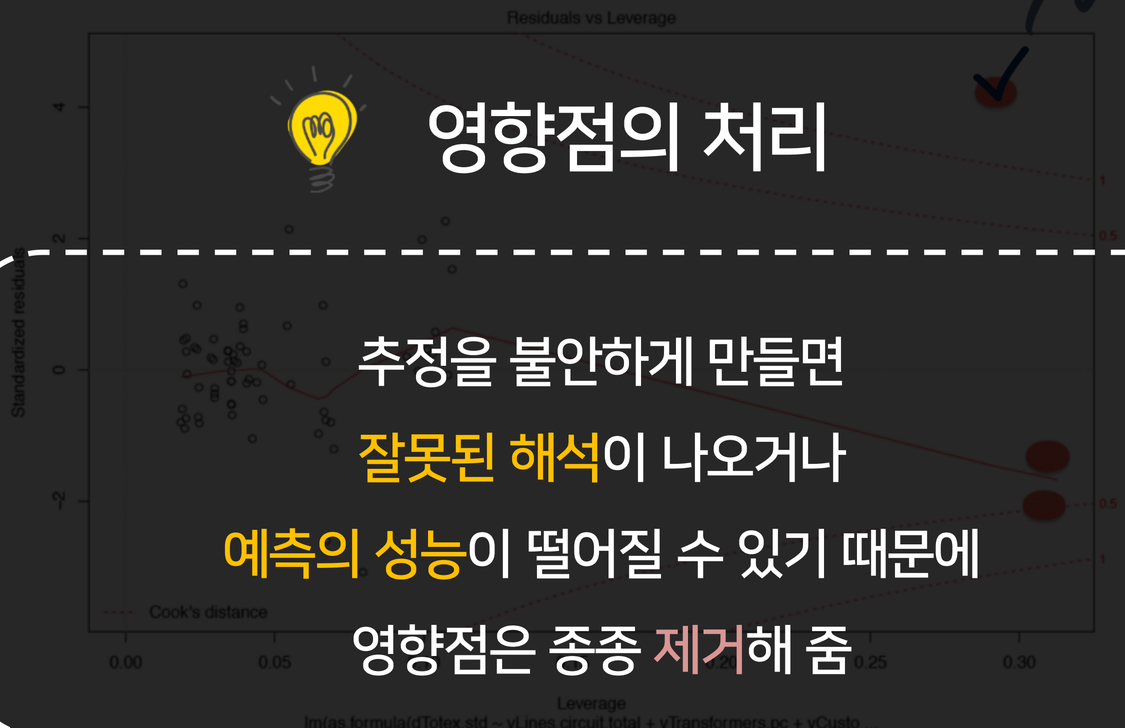



Leverage point라고 하더라도

모두 Influential point인 것은 아님을 확인할 수 있음

## 영향점(Influential point)

Influential point




### 영향점의 처리

추정을 불안하게 만들면

잘못된 해석이 나오거나

예측의 성능이 떨어질 수 있기 때문에

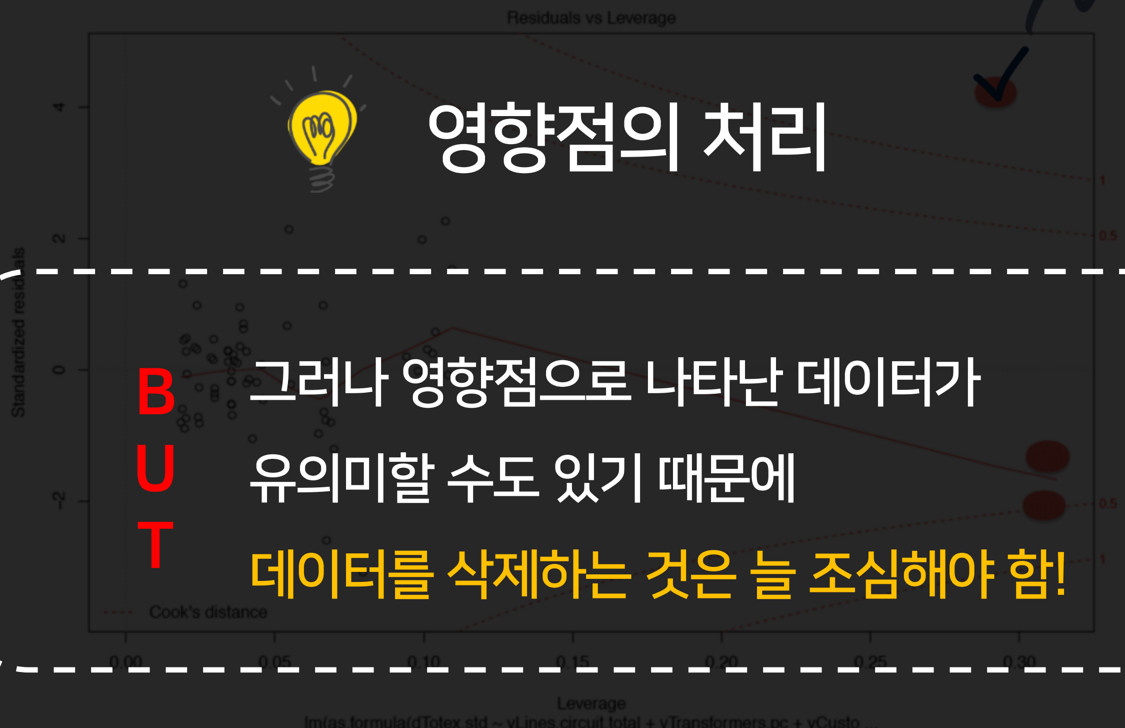
영향점은 종종 제거해 줌

Leverage point라고 하더라도

모두 Influential point인 것은 아님

## 영향점(Influential point)

Influential point




### 영향점의 처리

**BUT**

그러나 영향점으로 나타난 데이터가

유의미할 수도 있기 때문에

**데이터를 삭제하는 것은 늘 조심해야 함!**

Leverage point라고 하더라도

모두 Influential point인 것은 아님

## 영향점(Influential point)

Influential point


### 영향점의 처리

**B U T**

그러나 영향점으로 나타난 데이터가

유의미할 수도 있기 때문에

**데이터를 삭제하는 것은 늘 조심해야 함!**

Leverage point라고 하더라도

**이상치에 강건한(robust) 모델링이 필요!**

# 5

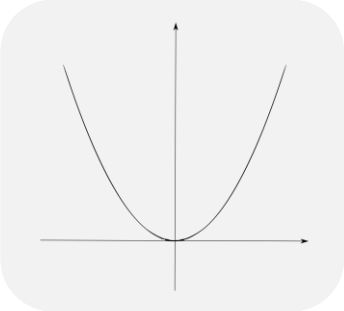
# 로버스트 회귀

## 로버스트 회귀란?

이상치의 영향을 줄이는 회귀분석 방법

1. Median Regression
2. Huber's M-estimation
3. Least Trimmed Square

## Median Regression



**최소제곱회귀**

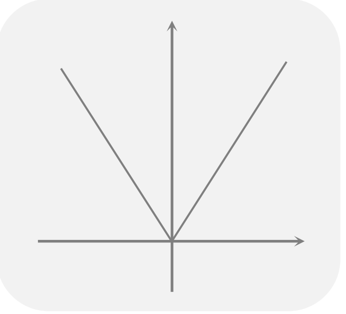
오차의 제곱합을 최소화하는 $\beta$를 찾는 방법

$$\sum \varepsilon_i^2 = (y - X\beta)^t (y - X\beta)$$

미분이 편리함!

이상치에 대해 너무 큰 가중치를 주는 경향이 있음

## Median Regression

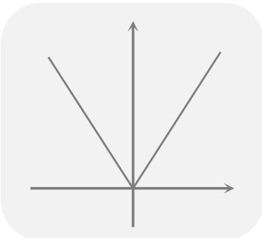


**Median Regression**

오차의 절댓값을 최소화하는 $\beta$를 찾는 방법

$\sum |\varepsilon_i|$

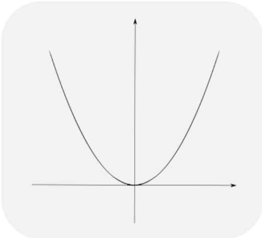
## Median Regression



Median Regression

- X에 따른 Y의 **중앙값** 반환
- 독립변수 X의 변화에 따른 종속변수 Y의 **조건부 중간값** 추정

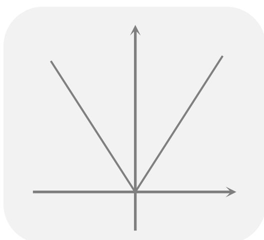
VS



최소제곱회귀

- X에 따른 **평균적인 Y** 반환
- 독립변수 X의 변화에 따른 종속변수 Y의 **조건부 평균**$(E(Y|X))$ 추정

## Median Regression



Median Regression

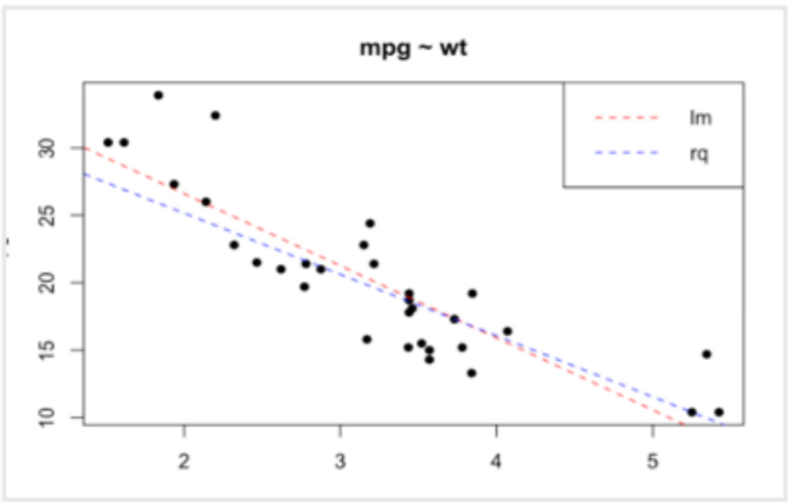
- X에 따른 Y의 **중앙값** 반환
- 독립변수 X의 변화에 따른 종속변수 Y의 **조건부 중간값** 추정

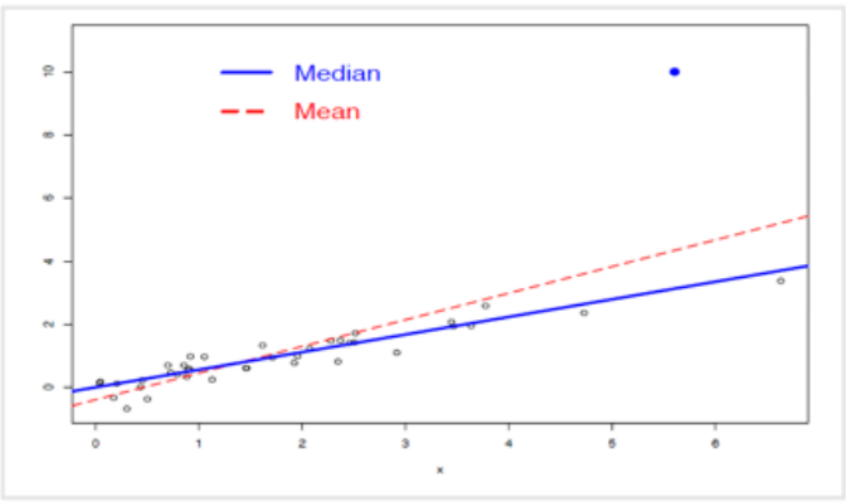
중심에서 **멀리 떨어진 이상치**에 **덜 민감한** 추정량을 가질 수 있음

대표값인 평균, 중앙값, 최빈값에 대해 배울 때,
중앙값은 이상치의 영향을 덜 받는다고 배웠죠?

# 5 로버스트 회귀

## Median Regression




중심에서 멀리 떨어진 이상치

분포가정과 등분산가정이 없는 모델

대표값인 평균, 중앙값, 최빈값에 대해 배울 때, 중앙값은 이상치의 영향을 덜 받는다는 거 다 아시죠?

R에서는 'quantreg' 패키지의 rq() 함수 사용

# 5 로버스트 회귀

## Huber’s M-estimation

**최소제곱회귀는…**

이상치에 지나치게 큰 패널티를 부여하지만,

동시에 적정 수준 안에서는 패널티를 완화시켜줌


적정수준의 패널티를 완화시켜주는 형태는 유지하되,

**이상치에 대해 지나친 패널티를 부여하는 것을 막고자 함.**

## Huber's M-estimation

잔차가 특정 상수값보다 크면 잔차의 '제곱'이 아닌 1차식으로 바꾸어서

이상치에 강건한 회귀계수를 추정하는 방법

$\rho$ 함수

$$\rho(e) = \begin{cases} \frac{1}{2}e^2 & if\ |e| \le c \\ c|e| - \frac{1}{2}c^2, & otherwise \end{cases}$$

기존 **최소제곱추정법**의 목적함수와 동일

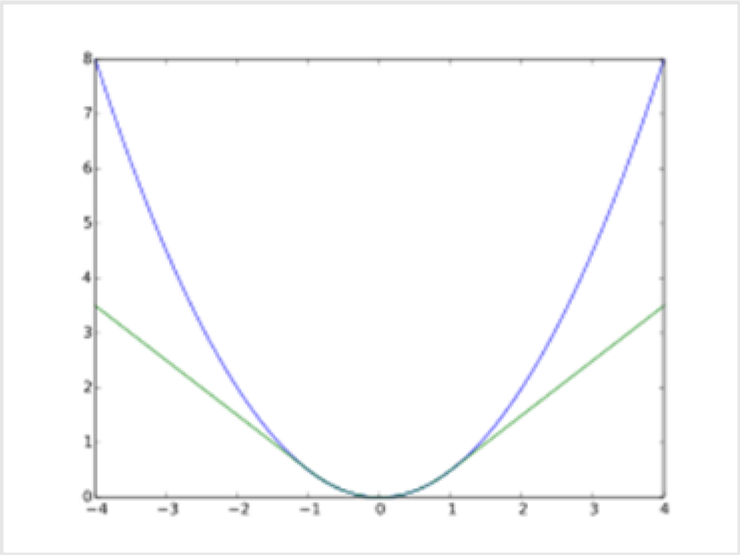
최적화해야하는 목적함수는

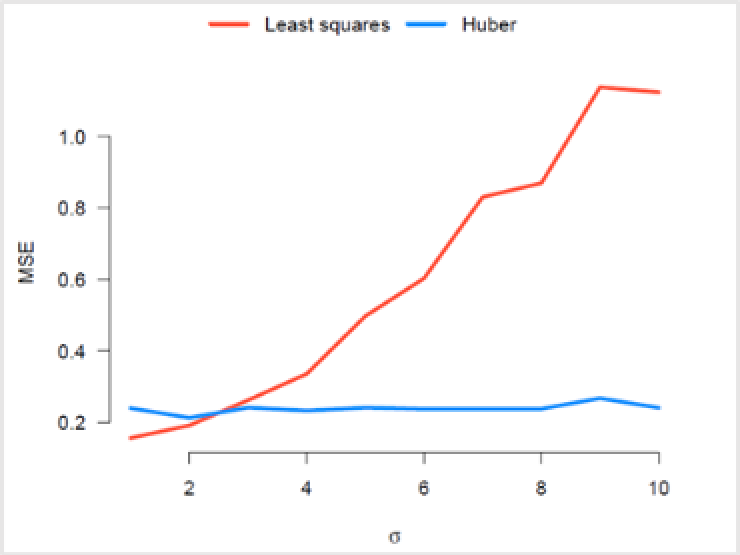
$\sum \rho(e)$

이상치에 대한 큰 패널티를 적용하지 않도록

**일차식의 형태**를 적용

## Huber's M-estimation




최소제곱회귀

Huber's M-estimation

이상치에 대한 패널티 완화로

MSE값이 작다는 것을 확인할 수 있음

이상치에 대한 큰 패널티를 적용하지 않도록

최적화해야하는 목적함수는

$\sum \rho(e)$

일차식의 형태를 적용

R에서는 'MASS' 패키지의 rlm() 함수 사용

## Least Trimmed Square

통계적 기준에 따라 잔차가 너무 큰 관측치를 제거하고 회귀계수를 추정하는 방식

$r_{(j)}$ 작은 순서부터 오름차순으로 나열한 잔차

$$\hat{\beta} = min \sum_{j=0}^{h} r_{(j)}^2 \begin{cases} r_1 \le r_2 \le \cdots \le r_h \\ \dfrac{n}{2} + 1 \le h \end{cases}$$

Obs가 별로 없는 경우나
영향점이 존재하지 않는 경우
주의해서 사용해야 함

n개의 obs 중 h개만 사용하여 회귀식을 만드는데,

$\binom{n}{h}$개의 회귀식 중 가장 잔차제곱합이 작은 회귀식을 사용

## Least Trimmed Square

통계적 기준에 따라 잔차가 너무 큰 관측치를 제거하고 회귀계수를 추정하는 방식

$r_{(j)}$ 작은 순서부터 오름차순으로 나열한 잔차

$$\hat{\beta} = min \sum_{j=0}^{h} r_{(j)}^2 \begin{cases} r_1 \le r_2 \le \cdots \le r_h \\ \dfrac{n}{2} + 1 \le h \end{cases}$$

Obs가 별로 없는 경우나

영향점이 존재하지 않는 경우

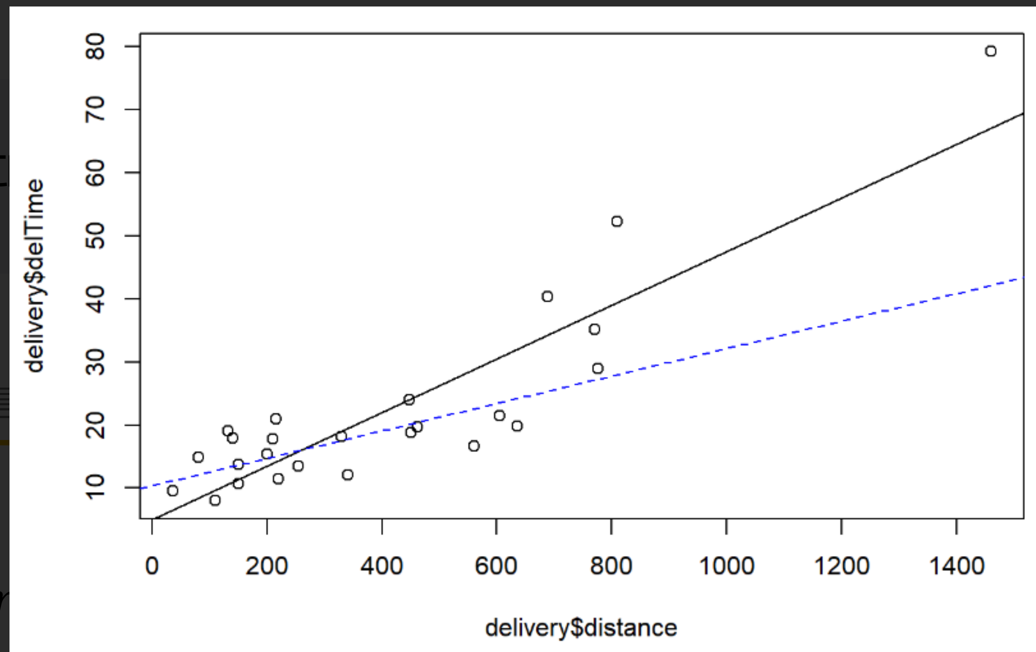
주의해서 사용해야 함

n개의 obs 중 h개만 사용하여 회귀식을 만드는데,

$\binom{n}{h}$개의 회귀식 중 가장 잔차제곱합이 작은 회귀식을 사용

# Least Trimmed Square

통계적 기준에 ㄷ … 추정하는 방식




이상치에 대한 패널티 완화로
기울기가 줄어들었음을 확인할 수 있음

$r_{(j)}$ 작은 순서부터 오름

$$\hat{\beta} = min \sum_{j=0}^{h} r$$

…는 경우나

…지 않는 경우

주의해서 사용해야 함

n개의 obs 중 h개만 사용하여 회귀식을 만드는데,

$\binom{n}{h}$개의 회귀식 중 가

R에서는 ‘robustbase’ 패키지의 lstReg() 함수 사용

# 다음주 예고

회귀분석의 기본 가정

잔차 플랏

선형성 진단과 처방

등분산성 진단과 처방

정규성 진단과 처방

독립성 진단과 처방

# 감사합니다