회귀분석팀

6팀
고경현
박세령
박이현
박지성
심예진
이선민

INDEX

1

회귀분석이란?

1 회귀분석이란?

회귀분석

회귀팀 파이팅><
(feat. 학회장님)

Regression Analysis

변수 사이의 관계를 모델링하는 통계적 기법

특정 변수들의 값을 이용하여 다른 변수를 설명하거나 예측

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

지도학습의 한 종류

지도학습이란?

결과의 예측이 목적인 학습 방법

결과변수와 특징변수가 모두 존재

1 회귀분석이란?

회귀분석

회귀팀 파이팅><
(feat. 학회장님)

Regression Analysis

변수 사이의 관계를 모델링하는 통계적 기법

특정 변수들의 값을 이용하여 다른 변수를 설명하거나 예측

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

지도학습의 한 종류

지도학습이란?

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

ϵ (오차항, Error Term)

변수를 측정할 때 발생할 수 있는 오차

설명이 불가능한 무작위성

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

ϵ (오차항, Error Term)

변수를 측정할 때 발생할 수 있는 오차

설명이 불가능한 무작위성

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

ϵ (오차항, Error Term)

변수를 측정할 때 발생할 수 있는 오차

설명이 불가능한 무작위성

1 회귀분석이란?

회귀 모델링 과정

Ⅰ. 문제 정의

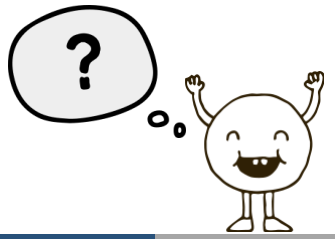
내 학점에 영향을 주는 요소에는 무엇이 있을까?

Ⅱ. 적절한 변수 선정

독립변수 X 선정 → 통학 거리, 공부 시간, 수강 학점이 영향을 주지 않을까?

Ⅲ. 데이터 수집 및 전처리

선정한 변수에 맞는 학생 데이터를 수집 및 전처리



1 회귀분석이란?

회귀 모델링 과정

Ⅳ. 모델 설정 및 적합

우리가 가진 데이터를 설명할 적절한 회귀분석 모델 선정

Ex) 다중회귀 모형 : $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_p} X_p$

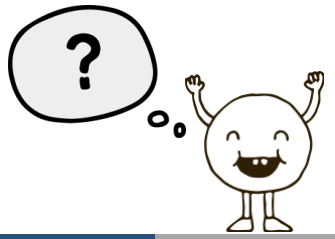
Ⅴ. 모델 평가

모델이 회귀분석의 가정을 만족하는지 평가 Ex) $\epsilon_i \sim NID(0, \sigma^2)$

Ⅵ. 모델 해석

분석을 바탕으로 독립변수와 종속변수간 관계 제시

3학점을 덜 듣는다면 추가적으로 GPA가 0.3정도 오르겠구나!



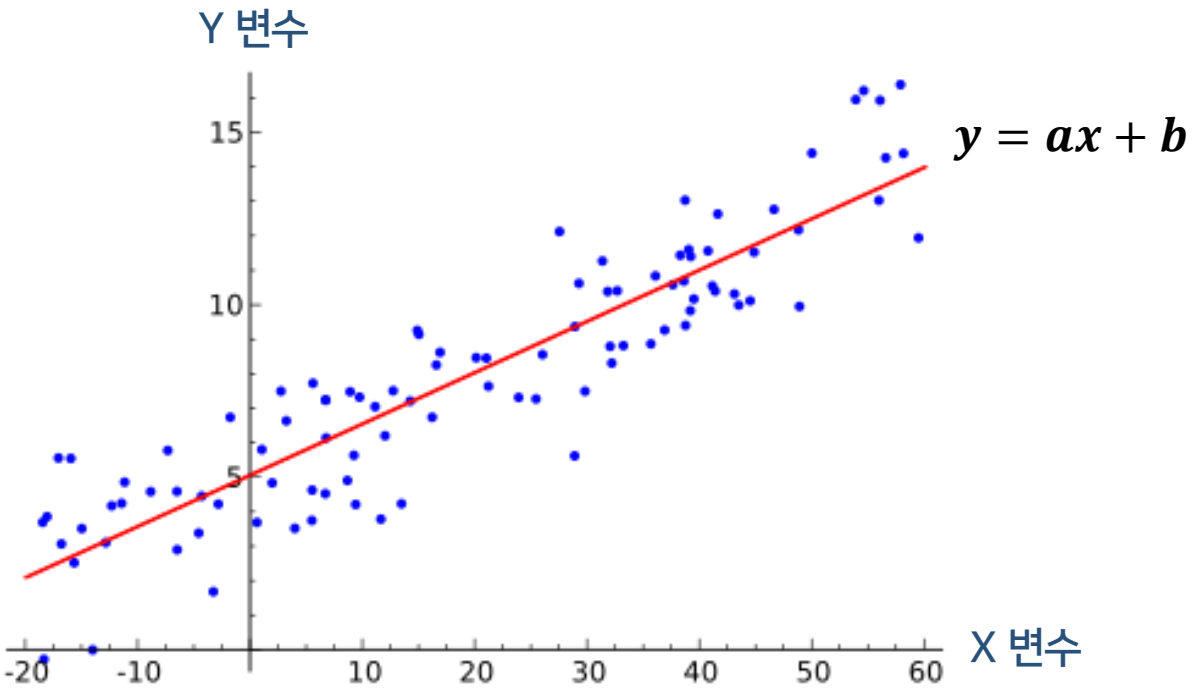
2

단순선형회귀

단순선형회귀

하나의 종속변수(Y)와 하나의 독립변수(X)만을 갖는 회귀모델

두 변수 간의 관계를 가장 잘 표현하는 직선 추정이 목적



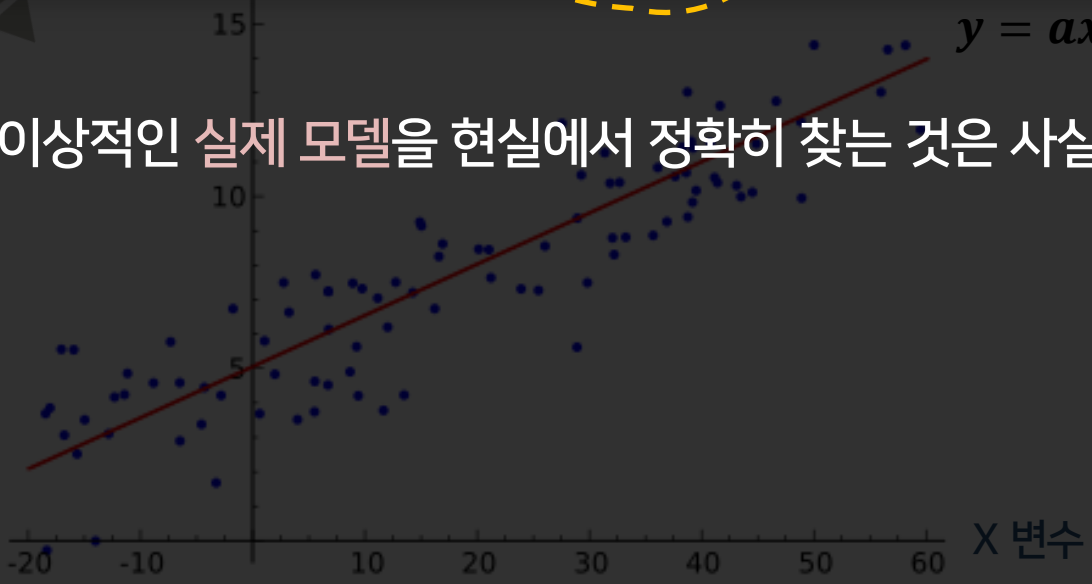
직선 추정의 목적

단순선형회귀

단순한 모델 | 실제 모델 | 복잡한 모델

$$y = ax + b$$

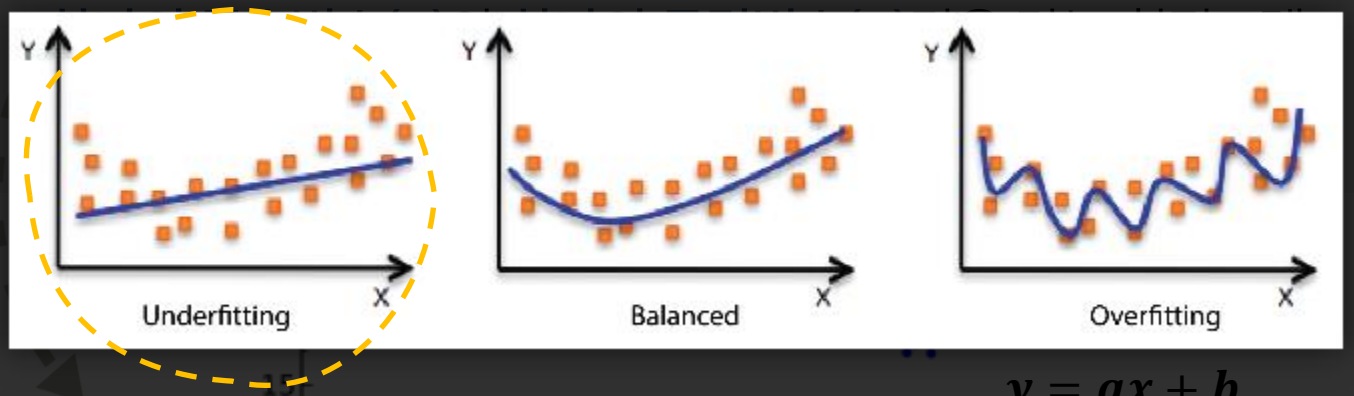
이상적인 실제 모델을 현실에서 정확히 찾는 것은 사실상 불가능



직선 추정의 목적

단순선형회귀

단순한 모델 simple



실제 모델

복잡한 모델

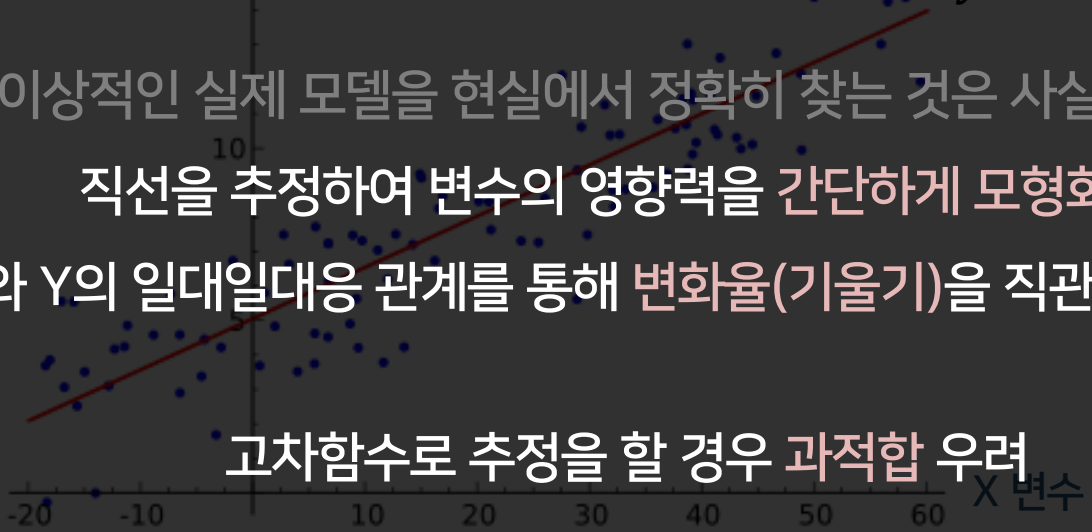
$$y = ax + b$$

이상적인 실제 모델을 현실에서 정확히 찾는 것은 사실상 불가능

직선을 추정하여 변수의 영향력을 간단하게 모형화 가능

X와 Y의 일대일대응 관계를 통해 변화율(기울기)을 직관적으로 이해

고차함수로 추정을 할 경우 과적합 우려



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

ϵ_i : i번째 관측값에 의한 랜덤 오차

평균은 0, 분산은 σ^2 를 가정

β_0, β_1 : 회귀계수, 추정해야 할 모수

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

ϵ_i : i번째 관측값에 의한 랜덤 오차

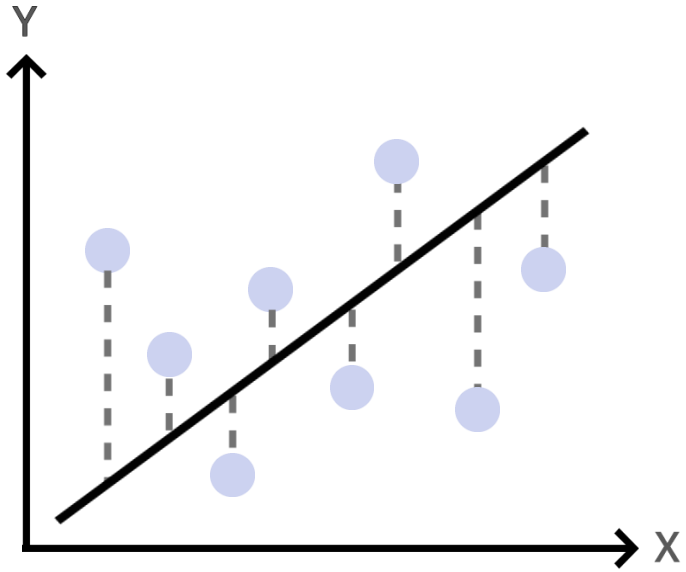
평균은 0, 분산은 σ^2 를 가정

β_0, β_1 : 회귀계수, 추정해야 할 모수

특정한 함수를 가정하는 모수적 방법

2 단순선형회귀

모수의 추정 : 최소제곱법



실제 데이터 사이의 관계를 함수식으로 표현하기 위해 모수 추정(β_0, β_1)

좋은 추정 : 실제 데이터와 추정된 함수 사이의 오차가 최소화 되는 경우

오차의 제곱합을 최소화하여 모수를 추정하는 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 β_0, β_1를 찾는 것이 목적

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

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

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

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 β_0, β_1를 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

모수의 추정 : 최소제곱법

최소제곱법을 사용하는 이유

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

미분의 편리성

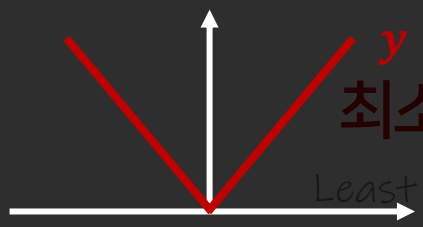
오차제곱합 $S(\beta_0, \beta_1)$를 최소화하는 β_0, β_1를 찾는 것이 목적

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

오차가 클수록 더 큰 패널티 부여 가능

오차의 절대값 사용 시 미분불가능한 점 존재

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



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 σ^2 로 동일

③ 오차간 자기상관이 없음(Independent)

Gauss-Markov Theorem

BLUE(Best Linear Unbiased 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 σ^2 로 동일

③ 오차간 자기상관이 없음(Independent)

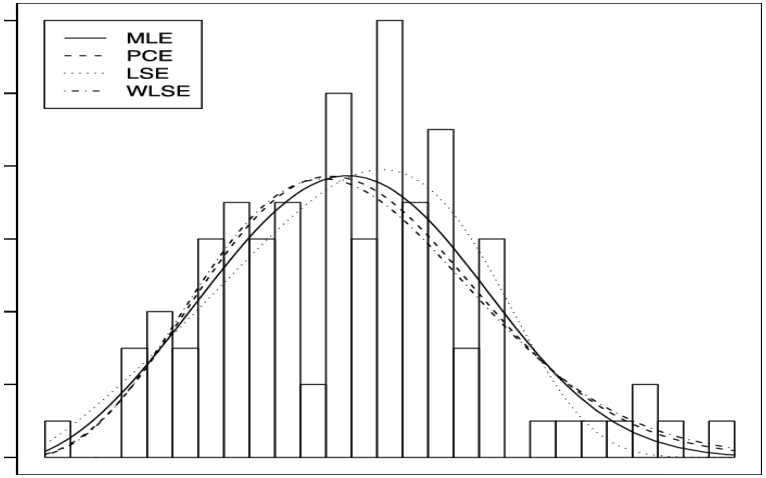
Gauss-Markov Theorem

BLUE(Best Linear Unbiased Estimator)

최대가능도추정량 vs 최소제곱추정량

최대가능도추정량 Maximal Likelihood Estimator

- 데이터가 나올 가능도(Likelihood)를 최대로 하는 모수 추정
- 관측치가 항상 iid라는 가정 필수
- 오차의 정규분포 가정 시 MLE = LSE



2 단순선형회귀

적합성 검정

01 최소제곱법으로 모수 추정

02 추정된 모수 = 회귀 계수

03 추정된 회귀 계수를 이용해 회귀직선

04 적합성 검정

실제 데이터에 잘 들어맞는지?

2 단순선형회귀

적합성 검정

01 최소제곱법으로 모수 추정

02 추정된 모수 = 회귀 계수

03 추정된 회귀 계수를 이용해 회귀직선

04 적합성 검정

실제 데이터에 잘 들어맞는지?

2 단순선형회귀

적합성 검정

01 최소제곱법으로 모수 추정

02 추정된 모수 = 회귀 계수

03 추정된 회귀 계수를 이용해 회귀직선

04 적합성 검정

실제 데이터에 잘 들어맞는지?

적합성 검정

잔차 Residual

오차의 추정량

실제값 y_i 와 추정값 $\hat{y}_i$의 차이

$$e_i = y_i - \hat{y}_i = y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_i\right)$$

오차와 잔차의 차이?

오차는 모집단의 측정치, 잔차는 표본의 측정치

적합성 검정

잔차 Residual

오차의 추정량

실제값 y_i 와 추정값 $\hat{y}_i$의 차이

$$e_i = y_i - \hat{y}_i = y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_i\right)$$

> 오차와 잔차의 차이?
>
> 오차는 모집단의 측정치, 잔차는 표본의 측정치

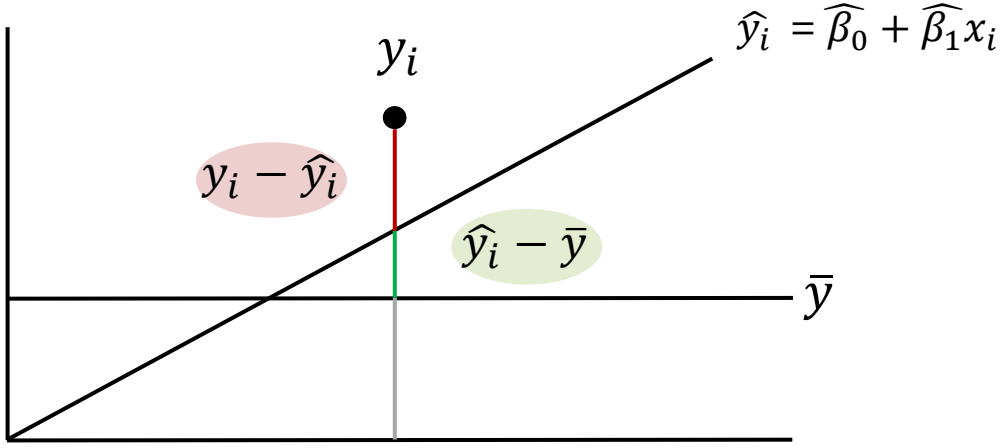
적합성 검정

변동 분할

총변동(Total Sum of Squares, SST) = $\sum(y_i - \bar{y})^2$

회귀변동(Regression Sum of Squares, SSR) = $\sum(\hat{y}_i - \bar{y})^2$

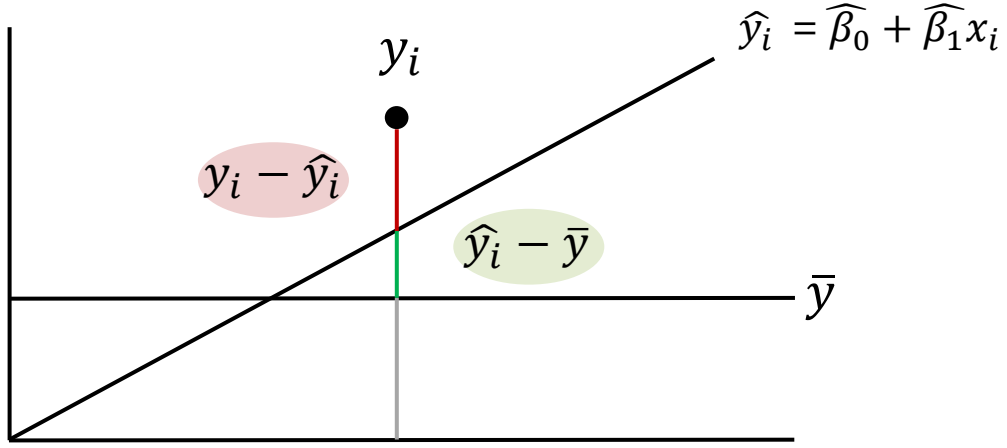
오차변동(Error Sum of Squares, SSE) = $\sum(y_i - \hat{y}_i)^2$



적합성 검정

변동 분할

$$SST = SSR + SSE$$



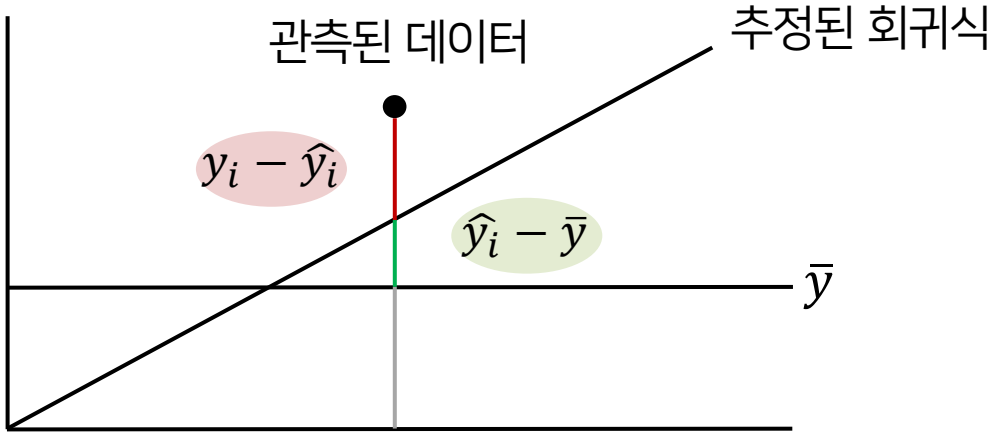
적합성 검정

변동 분할

총변동(Total Sum of Squares, SST) = $\sum(y_i - \bar{y})^2$

회귀변동(Regression Sum of Squares, SSR) = 설명 가능한 변동

오차변동(Error Sum of Squares, SSE) = 설명 불가능한 변동



적합성 검정

결정계수

총변동(SST)에서 회귀식이 설명할 수 있는 **비율**(SSR)

Y가 X에 의해 설명되는 비율

1에 가까울수록 좋음

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

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 α

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 α

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

β_0 도 동일한 방법으로 검정할 수 있음

2 단순선형회귀

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 귀무 가설을 기각하지 못하는 경우,

이는 X와 Y의 선형적 관계가 없을 뿐

상관 관계가 없음을 의미하는 것은 아님

β_0 도 동일한 방법으로 검정할 수 있음

3

다중선형회귀

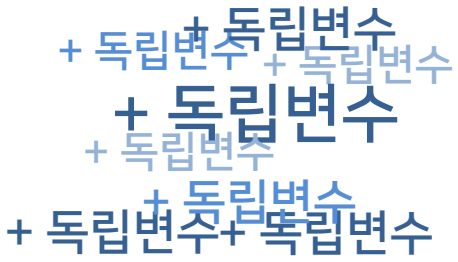
다중선형회귀

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_p가 한 단위 증가하면 y 는 β_p만큼 증가함을 의미

다중선형회귀

단순선형회귀

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

다중선형회귀

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

+ 독립변수
+ 독립변수 + 독립변수
+ 독립변수
+ 독립변수
+ 독립변수
+ 독립변수 + 독립변수

단순선형회귀에 비해 **복잡한 관계** 설명이 용이

나머지 X변수들이 고정되어 있을 때,
x_p가 한 단위 증가하면 y 는 β_p만큼 증가함을 의미

3 다중선형회귀

모수의 추정

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

- 다중선형회귀식을 **행렬**로 표현
- 단순선형회귀와 동일하게 최소제곱법(LSE)을 이용

모수의 추정 : 최소제곱법(LSE)

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

목적함수 $\boldsymbol{S}$ 를 $\boldsymbol{\beta}$에 대해 미분

해당 미분식 = 0

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

추정된 $\hat{\boldsymbol{\beta}}$을 이용하여 회귀식 추정하기

추정된 회귀식 $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$

3 다중선형회귀

모수의 추정 : 최소제곱법(LSE)

목적함수

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

목적함수 $\boldsymbol{S}$ 를 $\boldsymbol{\beta}$에 대해 미분

해당 미분식 = 0

추정량

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

추정된 $\hat{\boldsymbol{\beta}}$을 이용하여 회귀식 추정하기

추정된 회귀식

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

3 다중선형회귀

모수의 추정 : 최소제곱법(LSE)

목적함수

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

목적함수 $\boldsymbol{S}$ 를 $\boldsymbol{\beta}$에 대해 미분

해당 미분식 = 0

추정량

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

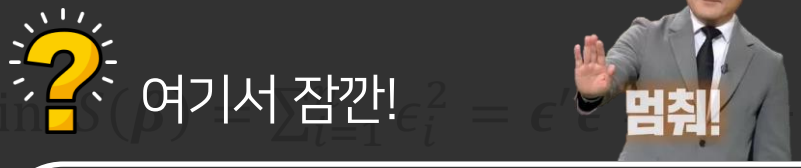
추정된 $\hat{\boldsymbol{\beta}}$을 이용하여 회귀식 추정하기

추정된 회귀식

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

3 다중선형회귀

모수의 추정 : 최소제곱법(LSE)



여기서 잠깐!

아래식에서 $\boldsymbol{H}$는 **투영 행렬**을 의미!

그렇다면, $\boldsymbol{H}$(투영행렬)가 가진

의미와 성질은 무엇일까?

추정된 회귀식

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

모수의 추정 : 최소제곱법(LSE)

투영 행렬이란?

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

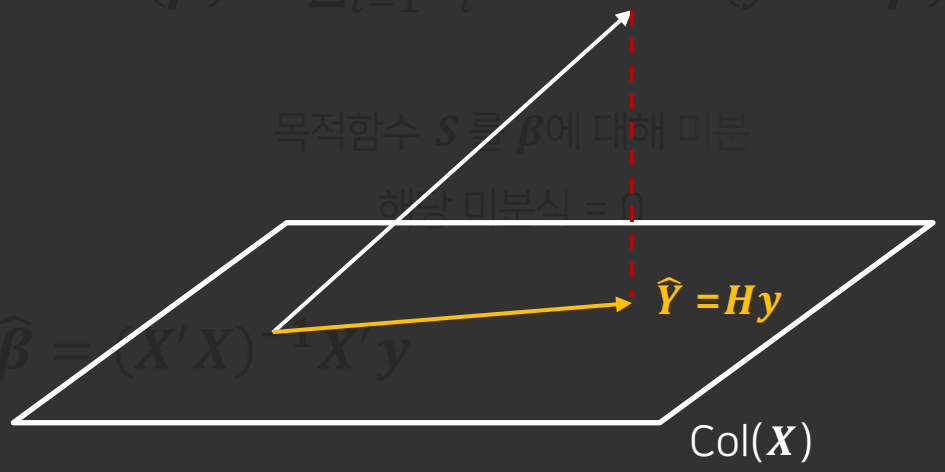
목적함수 S를 $\boldsymbol{\beta}$에 대해 미분

해당 미분식 = 0

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

추정된 $\hat{\boldsymbol{\beta}}$을 이용하여 회귀식 추정하기

추정된 회귀식 $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$

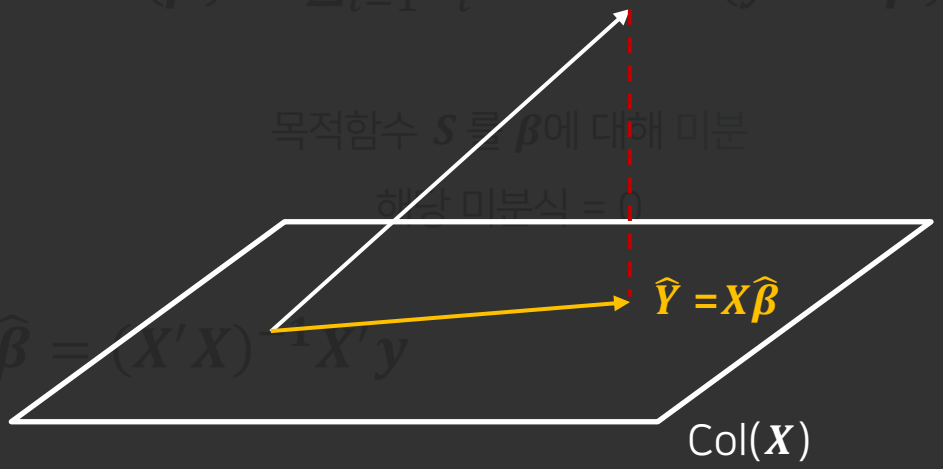


투영(projection)시키는 행렬

여기선, 투영 행렬 H가 X의 열공간으로 투영을 시킴

모수의 추정 : 최소제곱법(LSE)

투영 행렬이란?



회귀분석에서는 y 를 X의 열공간에 가깝게 근사 시키기 위해 사용

→ y 를 X의 열공간에 투영시킴으로써 해 $\hat{\beta}$ 이 존재하는 선형방정식이 됨

모수의 추정 : 최소제곱법(LSE)

투영 행렬의 성질

★ **대칭** (Symmetric)

$$H' = \left(X(X'X)^{-1}X'\right)' = X(X'X)^{-1}X' = H$$

★ **멱등** (Idempotent)

$$H^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H$$

추정된 회귀식

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

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

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

3 다중선형회귀

유의성 검정

전체 회귀계수에 대한 검정 : F-test

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 안됨 if $F_0 < F_{(1-a/2,\ p,\ n-p-1)}$
  - 모든 회귀계수는 0임
    → 모델 재설정 등 다른 조치 필요!

유의성 검정

전체 회귀계수에 대한 검정 : F-test

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 안됨 if $F_0 < F_{(1-a/2,\, p,\, n-p-1)}$
  - 모든 회귀계수는 0임

    → 모델 재설정 등 다른 조치 필요!

유의성 검정

일부 회귀계수에 대한 검정 : Partial F-test

Full Model (FM) : 기존 모든 변수를 사용한 다중회귀모형

Reduced Model (RM) : 일부 회귀 계수를 특정한 값(보통 0)으로 두는 축소 모형

가설설정

$$H_0: \beta_j = \beta_{j+1} = \cdots = \beta_{j+q-1} = 0 \text{ (RM이 맞다)}$$

$$H_1: not\ H_0$$

유의성 검정

일부 회귀계수에 대한 검정 : Partial F-test

Full Model (FM) : 기존 모든 변수를 사용한 다중회귀모형

Reduced Model (RM) : 일부 회귀 계수를 특정한 값(보통 0)으로 두는 축소 모형

가설설정

$$H_0: \beta_j = \beta_{j+1} = \cdots = \beta_{j+q-1} = 0 \text{ (RM이 맞다)}$$

$$H_1: not\ H_0$$

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

변수를 제거하면 당연히 커짐

$SSE(FM)$

모든 변수를 포함했을 때

모델이 설명하지 못하는 변동

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

변수를 제거하면 당연히 커짐

$SSE(FM)$

모든 변수를 포함했을 때

모델이 설명하지 못하는 변동

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

하지만 제거된 변수가
모델에 유의미하다면 월등히 커짐

$SSE(FM)$

모든 변수를 포함했을 때

모델이 설명하지 못하는 변동

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

$SSE(RM)$ — 하지만 제거된 변수가 모델에 유의미하다면 월등히 커짐 — $SSE(FM)$

q 개의 변수를 제거했을 때 모델이 설명하지 못하는 변동

모든 변수를 포함했을 때 모델이 설명하지 못하는 변동

검정통계량 F_0 커짐 → P-value 작아짐

유의성 검정

일부 회귀계수에 대한 검정 : Partial F-test

임계값

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

▸ q개의 회귀 계수 중 적어도 한 개의 회귀 계수는 0이 아님

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

▸ q개의 회귀 계수는 0임

유의성 검정

일부 회귀계수에 대한 검정 : Partial F-test

임계값

- 귀무가설 기각 if $F_0 \geq F_{(1-a/2,\, p-q,\, n-p-1)}$
  - q개의 회귀 계수 중 적어도 한 개의 회귀 계수는 0이 아님

- 귀무가설 기각 안됨 if $F_0 < F_{(1-a/2,\, p-q,\, n-p-1)}$
  - q개의 회귀 계수는 0임

유의성 검정

개별 회귀계수에 대한 검정 : t-test

가설설정

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

다른 변수들은 모두 적합 된 상태라고 가정!

검정통계량

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

유의성 검정

개별 회귀계수에 대한 검정 : t-test

가설설정

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

다른 변수들은 모두 적합 된 상태라고 가정!

검정통계량

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

유의성 검정

개별 회귀계수에 대한 검정 : t-test

임계값 $t_{(\alpha/2,\, n-p-1)}$

- 귀무가설 기각 if $|t_j| \geq t_{(\alpha/2,\, n-p-1)}$
  - x_j를 새로 추가하는 것은 통계적으로 유의함

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

다른 변수들이
다 적합 된 상태

3 다중선형회귀

유의성 검정

개별 회귀계수에 대한 검정 : t-test

임계값 $t_{(\alpha/2,\, n-p-1)}$

- 귀무가설 기각 if $|t_j| \geq t_{(\alpha/2,\, n-p-1)}$
  - x_j를 새로 추가하는 것은 통계적으로 유의함

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

다른 변수들이
다 적합 된 상태

3 다중선형회귀

유의성 검정

개별 회귀계수에 대한 검정 : t-t

t-test를 통해 변수선택을 하는 것은 위험!

임계값 $t_{(\alpha/2,\, n-p-1)}$

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

t-test 는 다른 변수들이 **다 적합 된 상태**에서

x_j 를 **추가적으로 적합했을 때**

회귀식의 **설명력을 증가**시키는지 알아보고자 하는 것

다른 변수들이
다 적합 된 상태

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_j를 새로 추가하는 것은 통계적으로 유의함
- 귀무가설 기각 안됨 if $|t_j| < t_{(\alpha/2,\, n-p-1)}$
  - x_j를 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

적합성 검정

수정 결정계수

변수가 추가됨에 따라 증가하는 결정계수에 **변수개수**라는 **패널티를 부과**한 형태

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

변수개수 패널티

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

- **패널티가 왜 필요한가?**

의미 없는 변수임에도 변수가 증가함에 따라

SSR이 증가하여 R^2 이 불필요하게 증가되는 문제점을 보완하기 위함!

적합성 검정

수정 결정계수

변수가 추가됨에 따라 증가하는 결정계수에 **변수개수**라는 **패널티를 부과**한 형태

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

변수개수 패널티

$$R^2_{adj} = 1 - \frac{SS\,E/(n-p-1)}{SS\,T/(n-1)}$$

패널티가 왜 필요한가?

의미 없는 변수임에도 변수가 증가함에 따라

SSR이 증가하여 R^2 이 불필요하게 증가되는 문제점을 보완하기 위함!

4

데이터 진단

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!

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

스튜던트 잔차

관측값이 경향성에서 벗어나는지 판단하는 기준!

스튜던트 잔차

y값의 단위에 영향을 많이 받는 잔차의 특성을 고려하여,

좀 더 **일반화된 상황**에서 적용할 수 있도록 **표준화**한 것

$$\mathbf{r_i} = \frac{\mathbf{e_i}}{\hat{\sigma}\sqrt{\mathbf{1} - h_{ii}}}, \hat{\sigma} = \sqrt{\frac{\mathbf{SSE}}{\mathbf{n - p - 1}}}$$

4 데이터 진단

스튜던트 잔차

관측값이 경향성에서 벗어나는지 판단하는 기준!

스튜던트 잔차

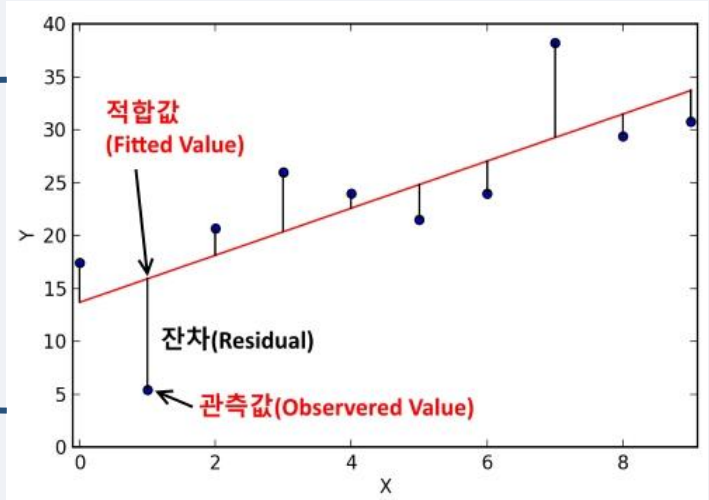
일반 잔차는 왜 안되나요?

스튜던트 잔차

관측값이 경향성에서 벗어나는지 판단하는 기준!

스튜던트 잔차

Y의 단위에 영향을 많이 받기 때문!



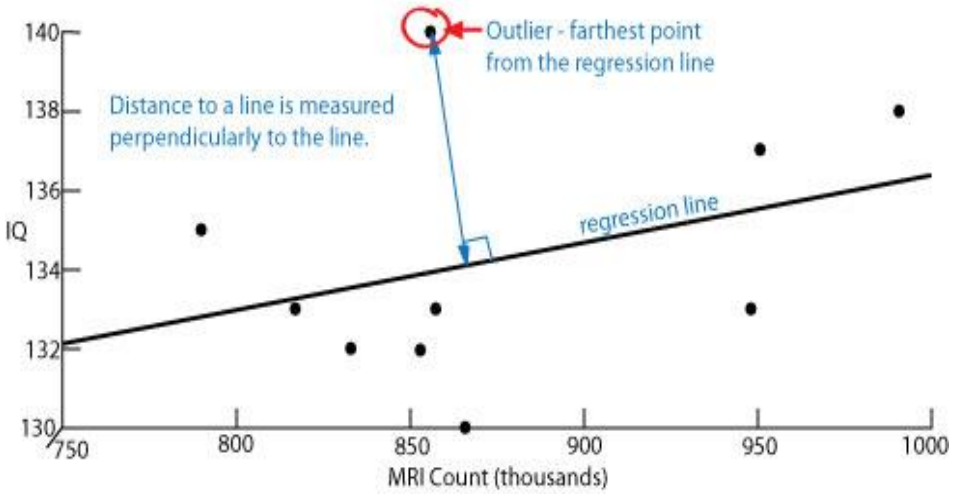
이상치

이상치 Outlier

스튜던트 잔차가 매우 큰 값

표준화 했을 때 Y의 기준에서 절대값이 큰 값

보통 $|r_i| > 3$ 이면 이상치라고 판단

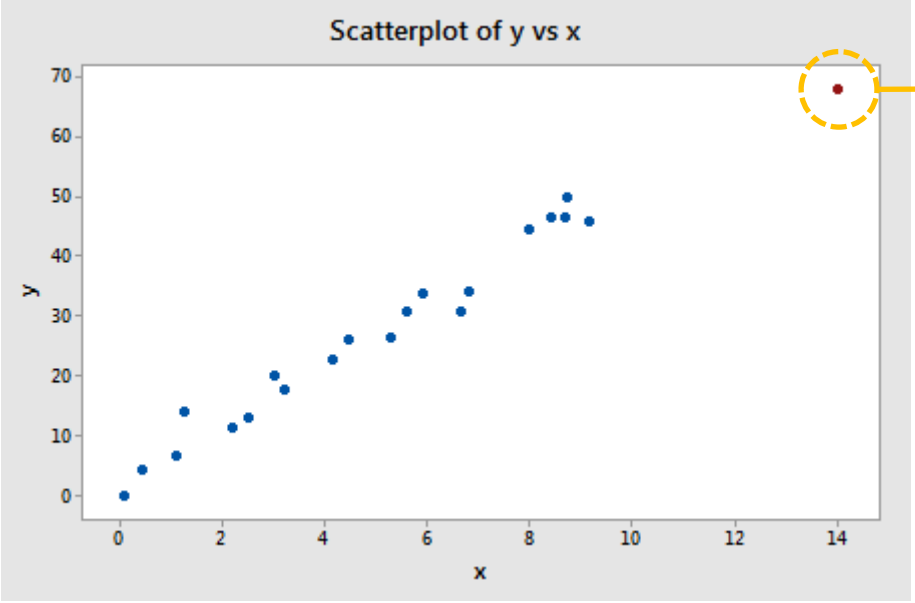


팀장님이 넣으래요,,

지렛값

지렛값 Leverage point

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



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

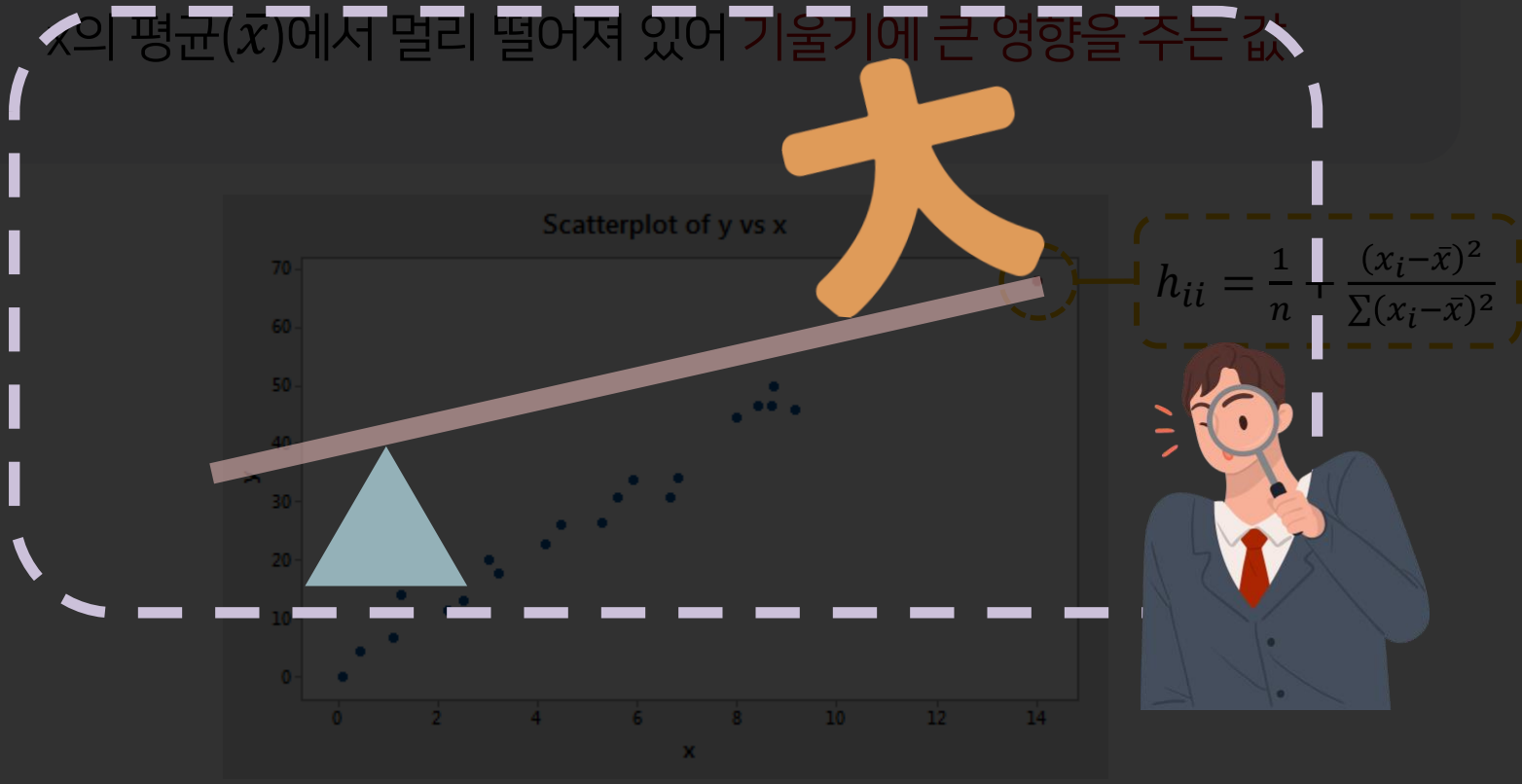
4 데이터 진단

지렛값

중심점에서 먼 점일수록 작은 힘만으로도

지렛값 Leverage Point 기울기를 쉽게 변화시키는 지렛대의 원리를 떠올려보자!

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



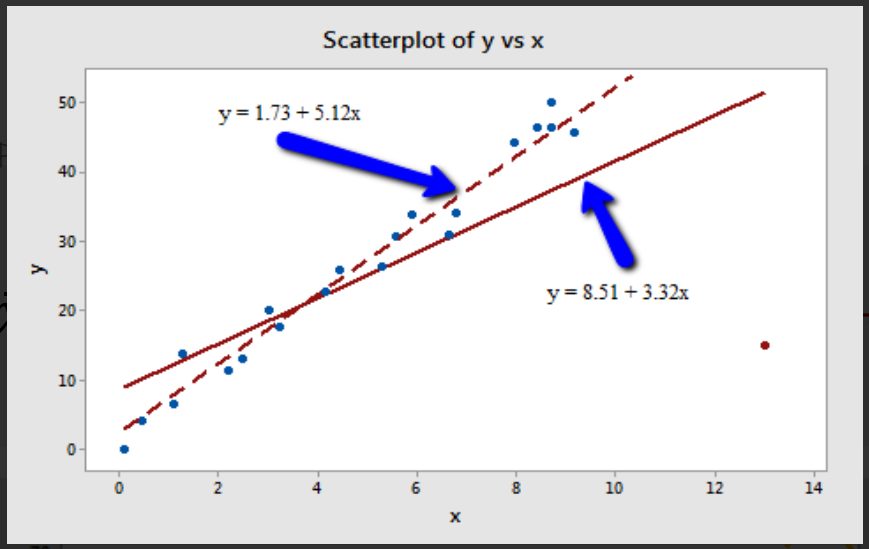
4 데이터 진단

지렛값

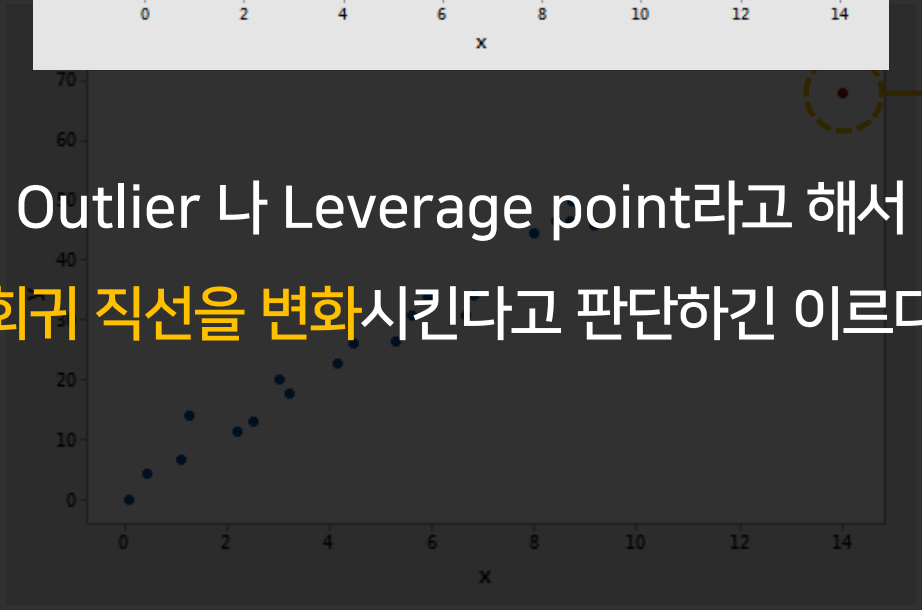
지렛값 Leverage

x의 평균(

을 주는 값



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



Outlier 나 Leverage point라고 해서

회귀 직선을 변화시킨다고 판단하긴 이르다

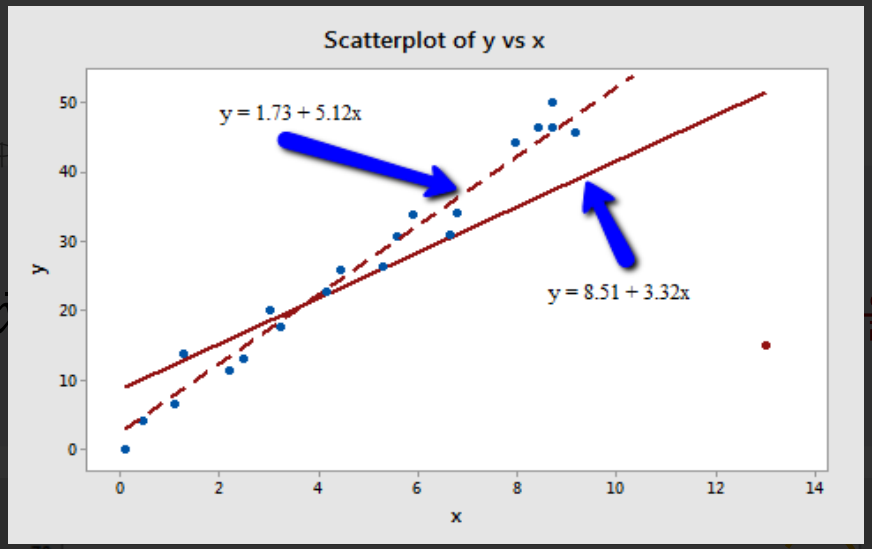
4 데이터 진단

지렛값

지렛값 Leverage

x의 평균(

을 주는 값



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

Outlier 나 Leverage point라고 해서

회귀 직선을 변화시킨다고 판단하긴 이르다

- 평균 주위에 있는 Outlier는 기울기 변화시키지 못함!
- Leverage point라도 회귀선의 연장선에 있을 수 있음!

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

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

영향점

영향점 Influential point

회귀직선의 기울기에 상당한 영향을 주는 점

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

이상치 고려
지렛값 고려

$C_i > 1$ 이면 영향점으로 판단

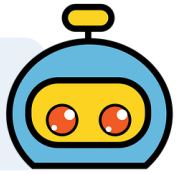
영향점의 처리

영향점이 있으면 왜 안되나요?

영향점의 처리

영향점이 있으면 왜 안되나요?

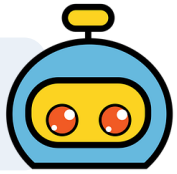
추정량을 불안정하게 (분산을 크게) 만듦



영향점의 처리

영향점이 있으면 왜 안되나요?

추정량을 불안정하게 (분산을 크게) 만듦

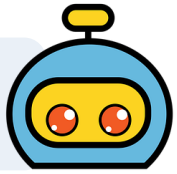


잘못된 모델의 해석

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

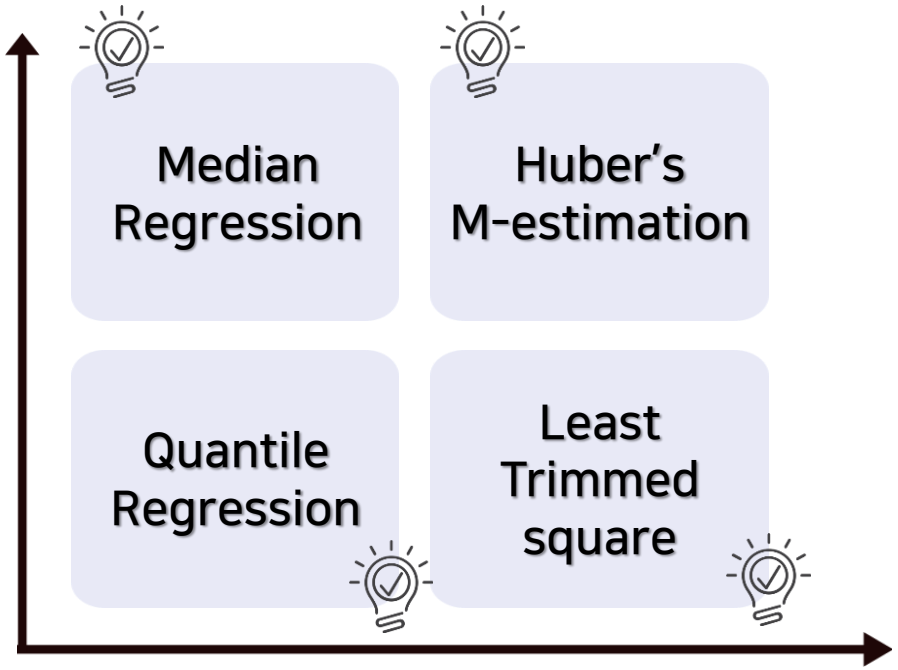
5

로버스트 회귀

로버스트 회귀

로버스트 회귀

이상치의 영향력을 크게 받지 않는 회귀모형

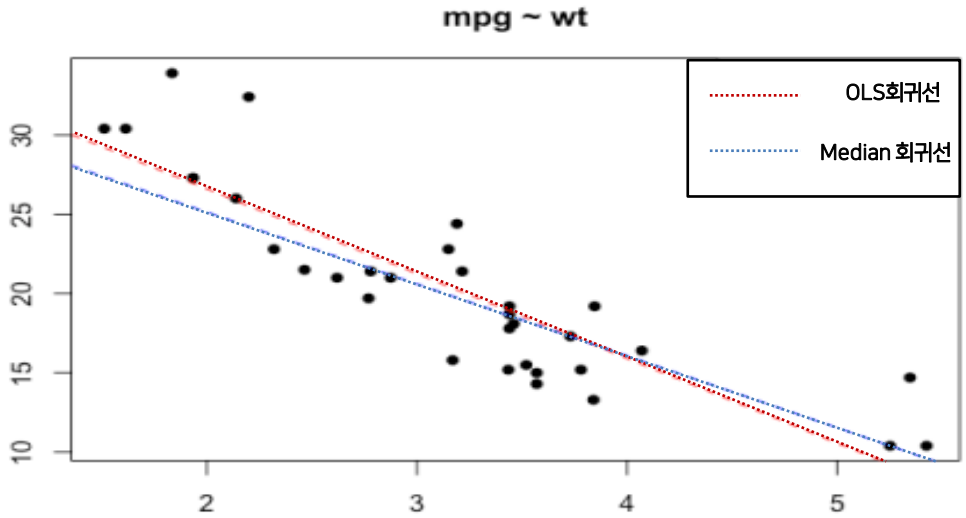


Median Regression

Median Regression

평균보다 중앙값이 이상치의 영향을 덜 받는다는 생각에 기초하여

독립변수 X의 변화에 따른 종속변수 Y의 조건부 중앙값을 추정하는 방법



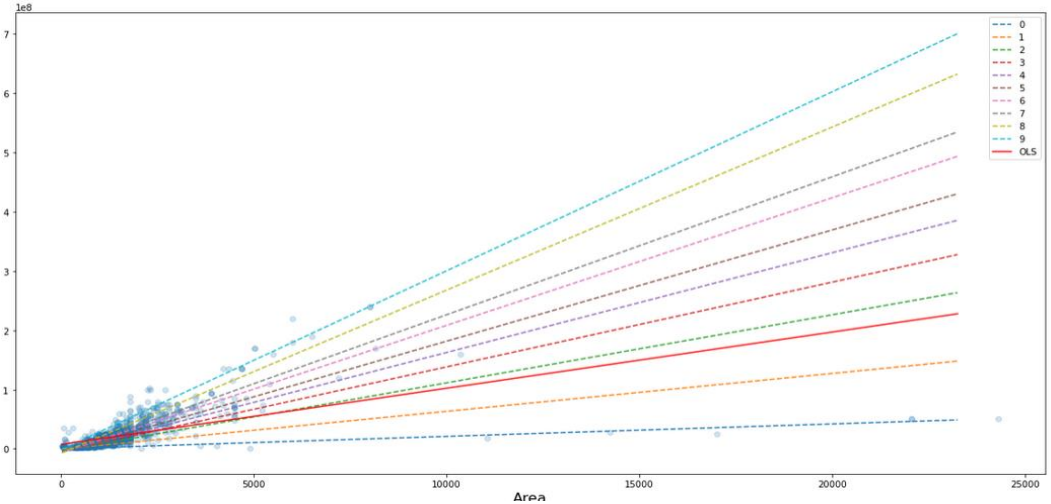
오차의 절대값의 합을 최소화

$$argmin \sum|\epsilon_i|$$

Quantile Regression

Quantile Regression

독립변수 x가 주어졌을 때, 종속변수 y의 분위수 값을 추정하는 방법



중간값이 백분위수에서 50번째 값

Median regression을 일반화한 방법

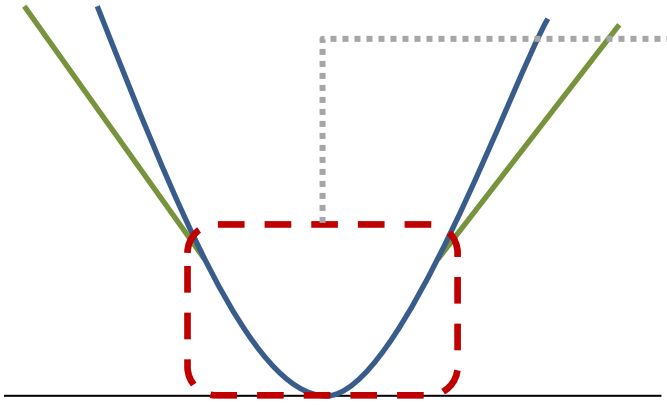
Huber's M-estimation

Huber's M-estimation

잔차가 특정 상수값보다 크면,

잔차의 '제곱'이 아닌 일차식으로 바꾸어 이상치에

강건한 회귀계수를 추정하는 방법



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

Least Trimmed Square

Least Trimmed Square

통계적 기준에 따라 잔차가 너무 큰 관측치를
제거하고 회귀계수를 추정하는 방법

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

- $r_{(j)}$는 작은 순서부터 오름차순으로 나열한 잔차
- N개의 관찰값 중 h개만 사용하여 회귀식을 만드는데,
  그 중 잔차제곱합이 가장 작은 회귀식을 사용
- 관찰값이 별로 없는 경우나 영향점이 존재하지 않는 경우 주의해서 사용해야 함

Least Trimmed Square

Least Trimmed Square

통계적 기준에 따라 잔차가 너무 큰 관측치를
제거하고 회귀계수를 추정하는 방법

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

- $r_{(j)}$는 작은 순서부터 오름차순으로 나열한 잔차
- N개의 관찰값 중 h개만 사용하여 회귀식을 만드는데,
  그 중 잔차제곱합이 가장 작은 회귀식을 사용
- 관찰값이 별로 없는 경우나 영향점이 존재하지 않는 경우 주의해서 사용해야 함

다음 주 예고

회귀분석의 4가지 기본 가정

가정 진단

가정 위배 시 문제점

가정 위배 시 처방법