회귀분석팀

6팀

김민우

채희지

김다민

성준혁

천예원

INDEX

1

기본 수식

1 기본 수식

기초 수식

표본 평균
(Sample Mean)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

표본 분산
(Sample Variance)

$$S^2{}_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S^2{}_y = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

표본 표준편차
(Sample Standard Deviation)

$$S_x = \sqrt{S^2{}_x}$$

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

편차제곱합(변동)

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad S_{xy}$$

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

1 기본 수식

기초 수식

표본 평균
(Sample Mean)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

표본 분산
(Sample Variance)

$$S^2{}_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S^2{}_y = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

표본 표준편차
(Sample Standard Deviation)

$$S_x = \sqrt{S^2{}_x}$$

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

편차제곱합(변동)

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad S_{xy}$$

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

1 기본 수식

기초 수식

표본 평균 (Sample Mean)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

표본 분산 (Sample Variance)

$$S^2{}_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S^2{}_y = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

표본 표준편차 (Sample Standard Deviation)

$$S_x = \sqrt{S^2{}_x}$$

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

편차제곱합(변동)

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad S_{xy}$$

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

1 기본 수식

기초 수식

표본 평균 (Sample Mean)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

표본 분산 (Sample Variance)

$$S^2{}_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S^2{}_y = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

표본 표준편차 (Sample Standard Deviation)

$$S_x = \sqrt{S^2{}_x}$$

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

변동

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad S_{xy}$$

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

공분산 (Covariance)

공분산(Covariance)

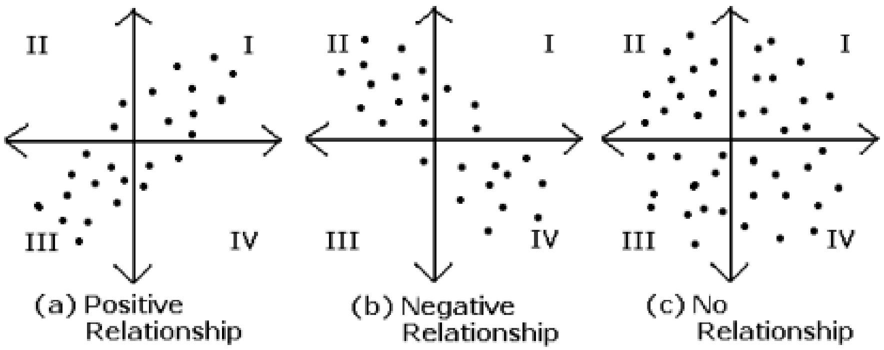
두 개의 확률변수의 선형 관계를 나타내는 값.

두 변수가 가지는 **선형관계의 방향성(양, 음)**만 나타낼 뿐이며,

어느 정도로 선형성을 갖는지는 표현하지 못함

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

공분산 (Covariance)



(a) 공분산>0 (b) 공분산<0 (c) 공분산≒0

$Cov(X, Y) > 0$: X가 증가할 때 Y도 증가

$Cov(X, Y) < 0$: X가 증가할 때 Y는 감소

$Cov(X, Y) = 0$: X, Y 두 변수 간 선형 상관관계 존재하지 않음

비선형 관계는 존재할 수도 있음!

공분산 (Covariance)

공분산은 확률 변수의 측정 단위에 영향을 많이 받으므로

상관성의 형태에 대해서는 나타낼 수 있지만, 상관성의 정도를 직접 나타낼 수는 없다!

Ex) A, B의 공분산과 C, D의 공분산 크기가 다르더라도,

각각 선형 관계를 나타내는 정도는 같을 수 있음

⋮

단순히 공분산이 더 크다고 해서 선형관계가 더 강하게 나타난다고 할 수 없음

1 기본 수식

공분산 (Covariance)

공분산은 확률 변수의 측정 단위에 영향을 많이 받으므로

상관성의 형태에 대해서는 나타낼 수 있지만, 상관성의 정도를 직접 나타낼 수는 없다!

이를 보완하기 위해 **상관계수** 사용!

1 기본 수식

상관계수(Correlation Coefficient)

상관계수(Correlation Coefficient)

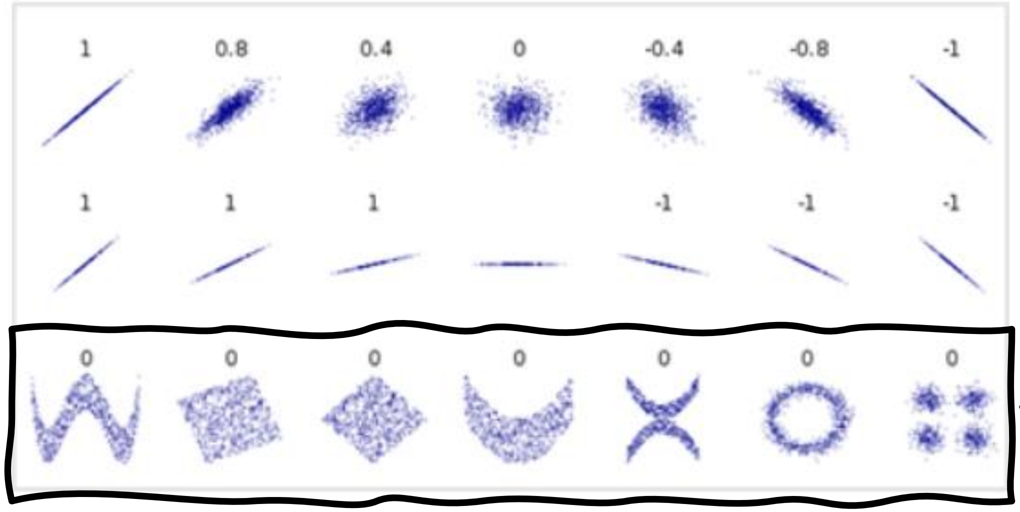
확률변수의 절대적 크기에 영향을 받지 않도록 단위화를 진행한

표준화된 공분산

$$r_{xy} = \frac{Cov(X,Y)}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

상관계수(Correlation Coefficient)

- 두 확률변수의 선형 상관관계의 여부와 선형적 상관성 크기까지 파악할 수 있는 지표
- -1부터 1까지의 값을 가짐
- 일반적으로 0.7 이상이면 강한 상관관계를 지닌다고 판단
- 확률변수 X, Y가 독립일 때 상관계수는 0이 됨



상관계수가 0일 때, 선형 관계가 없을 뿐
비선형 관계는 존재할 수 있음!

2

회귀분석이란?

2 회귀분석이란?

회귀분석의 정의

회귀분석

- 독립 변수와 종속 변수 간의 관계를 설명하고 모델링하는 통계적 기법
- 변수들 간의 상관관계를 파악하고, 특정 변수의 값을 다른 변수들을 이용해 설명하고 예측하는 방법
- 지도학습(Supervised Learning)의 한 종류

Train data로부터 하나의 함수를 유추해내기 위한 ML 방법. 유추된 함수 중 연속적인 값을 출력하는 회귀분석, 주어진 입력 벡터가 어느 집단인지 구분하는 분류 등이 속함.

2 회귀분석이란?

회귀분석의 정의

회귀분석의 종류

- **단순회귀분석**: 한 개의 종속변수와 한 개의 독립변수 사이의 관계 분석

- **다중회귀분석**: 한 개의 종속변수와 여러 개의 독립변수 사이의 관계 분석

회귀분석의 목적

- 변수들 간의 **관계**에 대한 표현

- 독립변수에 따른 **종속변수의 변화** 파악

- 미래 관측값에 대한 **예측**

2 회귀분석이란?

회귀식

회귀식

종속변수(Predictor, Feature) Y와 **독립변수**(Response) X의 관계를 함수식으로 표현한 것.

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

Y, **종속변수** : 독립변수에 의해 설명되는 변수. 반응변수라고도 불림.

X_k, **독립변수** : 종속변수를 설명하기 위한 변수. 설명변수, 예측변수라고도 불림.

ϵ, **오차항** : 변수를 측정할 때 발생할 수 있는 오차. 설명할 수 없는 무작위성을 지님.

2 회귀분석이란?

상관분석과의 차이

상관분석

- 두 변수의 관계만 표현 가능
- 변수 간 선형적 상관성 정도만 표현 가능하며, 구체적인 예측과 설명 불가능

회귀분석을 사용하는 이유!

독립변수가 한 단계 변할 때마다 **종속변수가 어떻게 변화할지** 알게 된다면,

더 유의미한 관계 파악이 가능해짐!

2 회귀분석이란?

상관분석과의 차이

상관분석

- 변수 간 선형적 상관성 정도만 표현 가능하며, 구체적인 예측과 설명 불가능

회귀분석을 사용하는 이유!

독립변수가 한 단계 변할 때마다 **종속변수가 어떻게 변화할지** 알게 된다면,

더 유의미한 관계 파악이 가능해짐!

회귀분석과 인과관계

회귀분석은 변수 간 상관관계를 기반으로 한 분석이며,

독립변수를 통해 종속변수를 예측하는 것이 목표.

독립변수와 종속변수를 가정해 분석하지만,

그 결과가 **인과 관계를 의미하지는 않음.**

독립변수가 종속변수를 잘 예측한다고 해서

인과 관계가 있다고 할 수 없음!

회귀 모델링 과정

① 문제 정의

나의 학점을 가장 잘 표현할 수 있는 변수들은 무엇이 있을까?

② 적절한 변수 선택

$X_1, X_2, \cdots, X_p$: 공부 시간, 통학 거리, 아침 식사 여부

③ 데이터 수집 및 전처리

나의 학점, 공부 시간, 집에서 학교까지의 거리, 아침 식사 여부 조사

회귀 모델링 과정

① 문제 정의

나의 학점을 가장 잘 표현할 수 있는 변수들은 무엇이 있을까?

② 적절한 변수 선택

$X_1, X_2, \cdots, X_p$: 공부 시간, 통학 거리, 아침 식사 여부

③ 데이터 수집 및 전처리

나의 학점, 공부 시간, 집에서 학교까지의 거리, 아침 식사 여부 조사

회귀 모델링 과정

① 문제 정의

나의 학점을 가장 잘 표현할 수 있는 변수들은 무엇이 있을까?

② 적절한 변수 선택

$X_1, X_2, \cdots, X_p$: 공부 시간, 통학 거리, 아침 식사 여부

③ 데이터 수집 및 전처리

나의 학점, 공부 시간, 집에서 학교까지의 거리, 아침 식사 여부 조사

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 만족하는가?

만족하지 않는다면, 처방 시도(회귀팀 클린업 2주차 예정)

⑥ 모형 해석

현재보다 주당 2시간 더 공부하고, 자취방에서 통학하고, 아침밥을 꼬박꼬박 챙겨

먹는다면 학점이 0.5만큼 오를 것이다! (제발..

2 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 만족하는가?

만족하지 않는다면, 처방 시도(회귀팀 클린업 2주차 예정)

⑥ 모형 해석

현재보다 주당 2시간 더 공부하고, 자취방에서 통학하고, 아침밥을 꼬박꼬박 챙겨

먹는다면 학점이 0.5만큼 오를 것이다! (제발..

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 만족하는가?

만족하지 않는다면, 처방 시도(회귀팀 클린업 2주차 예정)

⑥ 모형 해석

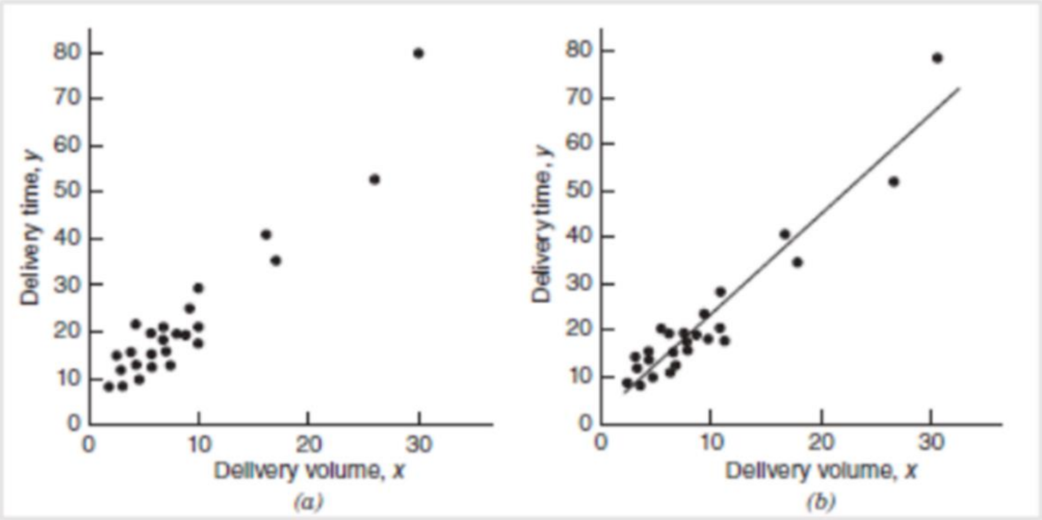
현재보다 주당 2시간 더 공부하고, 자취방에서 통학하고, 아침밥을 꼬박꼬박 챙겨 먹는다면 학점이 0.5만큼 오를 것이다! (제발..)

3

단순선형회귀

단순선형회귀

독립변수 X와 종속변수 Y의 **관계를 가장 잘 표현할 수 있는 직선**을 찾는 것



두 변수의 관계가 **선형적**일 것이라는 가정을 바탕으로 추정!

단순선형회귀 모델

선형회귀식

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

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

y_i : 종속변수 y의 i번째 관측값

x_i : 독립변수 x의 i번째 관측값

β_0 , β_1 : **회귀계수** = 우리가 추정해야 할 **모수**

ϵ_i : 오차항 = i번째 관측값에 의한 랜덤한 오차

단순선형회귀 모델

선형회귀식

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

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

단순선형회귀의 해석

X가 한 단위 증가할 때, Y가 **평균적으로 β_1 만큼 증가**한다.

왜 직선인가?

선형 근사를 할 경우

변수의 영향력을 **간단하게 모형화** 가능

X의 변화에 따른 Y의 변화를 직관적으로 확인 가능

고차 근사를 할 경우

모델의 복잡도가 높아져서,

과적합(overfitting) 문제의 원인이 됨

왜 직선인가?

선형 근사를 할 경우

변수의 영향력을 **간단하게 모형화** 가능

X의 변화에 따른 Y의 변화를 직관적으로 확인 가능

고차 근사를 할 경우

모델의 복잡도가 높아져서,

과적합(overfitting) 문제의 원인이 됨

단순선형회귀 모델

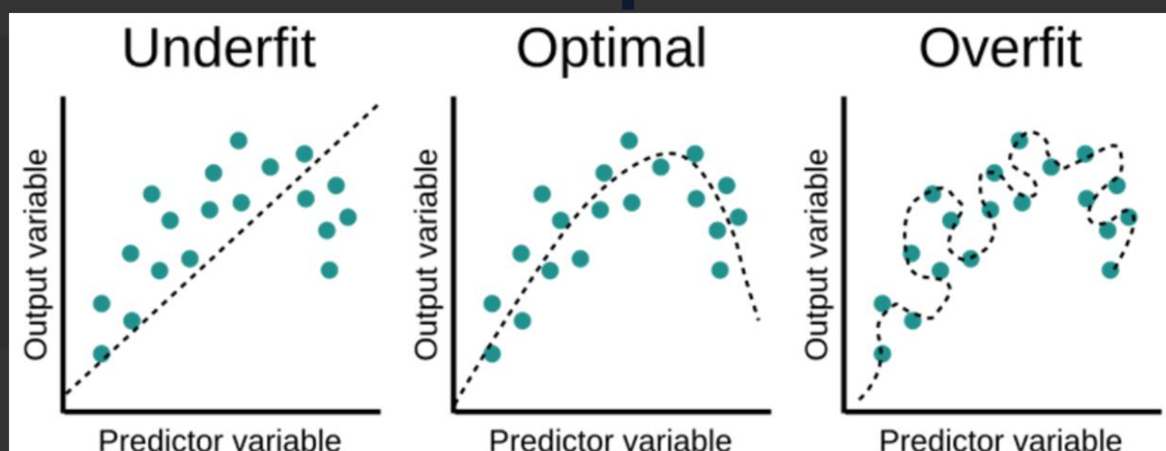
선형 근사를 할 경우

과적합(Overfitting)이란?

Train data에 대한 설명력은 높을 수 있지만

Test data에 대한 설명력은 떨어지는 문제

모델의 분산을 높이고, 검증 데이터의 예측 성능 저하시킴!

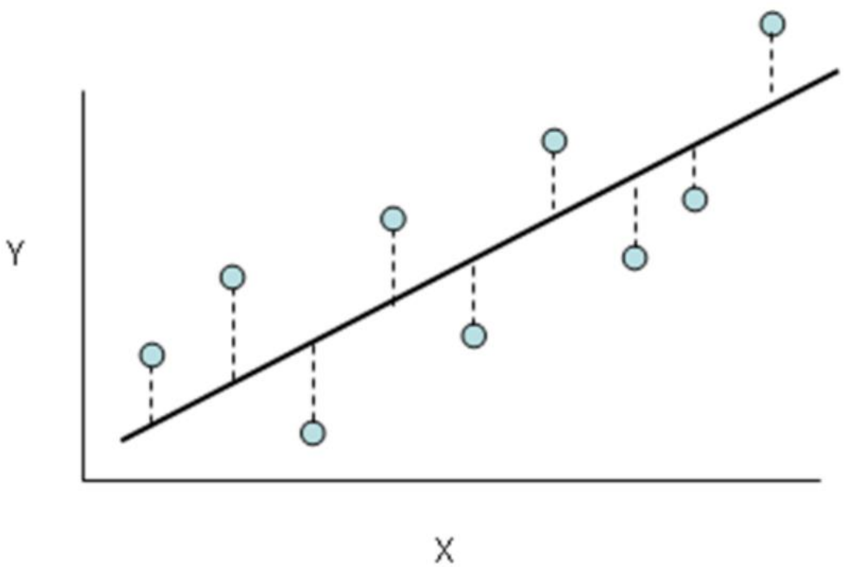


(데마팀 1주차 클린업 참고)

모수의 추정(LSE)

좋은 추정이란 ...

우리가 만들 회귀직선과 관측치 사이의 **오차가 작을수록** 좋은 추정!



최소제곱법 (LSE)

오차의 제곱합을 최소화하는

모수를 추정하는 방법

최소제곱법(LSE : Least Square Estimation Method)

오차의 제곱합

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

오차 제곱합을 최소화시키는 β_0, β_1 을 찾는 것이 목적!

아래로 볼록한 Convex 함수 → **'미분식=0'** 을 만족시키는 β_0, β_1을 구함

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

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

각각의 모수를 편미분하여 구함!

최소제곱법(LSE : Least Square Estimation Method)

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

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

최소제곱법으로 추정한 $\widehat{\beta_0}\,,\widehat{\beta_1}$ .... **최소제곱추정치** (LSE : Least Square Estimator)

최소제곱법 (LSE : Least Square Estimation Method)

왜 오차의 '제곱합'을 최소화할까?

① 미분이 **편리**하기 때문

② 오차가 클수록 **더 큰 페널티를 부여**할 수 있기 때문

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad , \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

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

오차의 '절댓값' 사용 시 미분 불가능 → 계산이 오래 걸리게 됨!

3 단순선형회귀

BLUE

BLUE(Best Linear Unbiased Estimator)

분산이 가장 작은 선형 불편추정량

(추정량이 안정적이기 때문에 더 유용한 성질)

LSE가 BLUE가 되는 3가지 조건

① 오차들의 평균은 0

② 오차들의 분산은 σ^2으로 동일 (등분산성)

③ 오차 간에 자기상관 없음 (uncorrelated)

최소제곱추정량(LSE) vs. 최대가능도추정량(MLE)

최대가능도 추정 (Maximum Likelihood Estimator)

확률적인 방법에 근거해서, 원하는 데이터가 나올

가능도를 최대로 하는 모수를 선택하는 방법

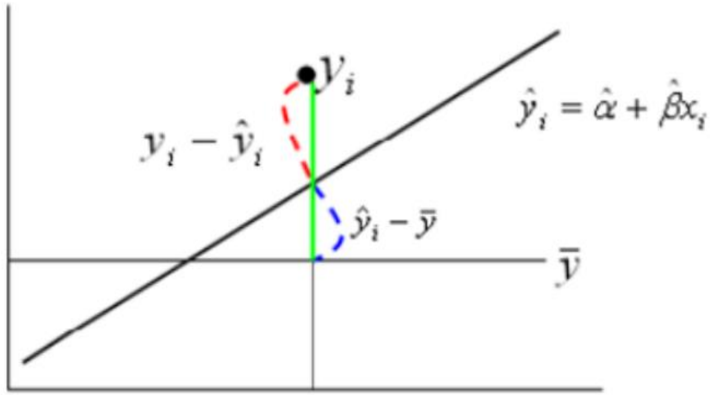
$\epsilon_i \sim N(0, \sigma^2)$ 라는 가정만 있다면 사용 가능

LSE와 MLE는 완전히 동일한 추정량을 산출!

적합성 검정(Goodness of Fit Test)

우리가 만든 회귀직선이 데이터를 얼마나 잘 설명하는가?

변동 분할을 통해 구한 **결정계수(R^2)**로 판단!



SST = SSR + SSE

SST: 총 변동

SSR: 회귀선이 설명하는 변동

SSE: 회귀선이 설명하지 못하는 변동

적합성 검정(Goodness of Fit Test)

결정계수 R^2

총변동(SST)에서 회귀직선이 설명하는 변동(SSR)의 비율

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

결정계수 R^2이 **1에 가까울수록**

회귀모형이 데이터를 **잘 설명한다는** 의미!

유의성 검정

개별 모수(회귀계수)의 추정량이 통계적으로 유의한가?

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

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

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

③ 검정 통계량을 분포에 적용 : $t_0 = \beta_1, \frac{\widehat{\beta_1}}{s.e.(\widehat{\beta_1})} \sim t_{(n-2)}$

④ 임계값 확인 : $t_{(1-\frac{a}{2}, n-2)}$

⑤ 통계적 검정(양측) : $|t_0| > t_{(1-\frac{a}{2}, n-2)}$ 이면, H_0 기각!

유의성 검정 해석

귀무가설을 기각했다면,

X와 Y 사이에 선형적 관계가 **있다**고 판단

귀무가설을 기각하지 못했다면,

X와 Y 사이에 선형적 관계가 **없다**고 판단

단, **비선형적 관계는 있을 수도** 있음!

유의성 검정 해석

귀무가설을 기각했다면,

X와 Y 사이에 선형적 관계가 **있다**고 판단

귀무가설을 기각하지 못했다면,

X와 Y 사이에 선형적 관계가 **없다**고 판단

단, **비선형적 관계는 있을 수도** 있음!

4

다중선형회귀

다중선형회귀

단순선형회귀

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

독립변수 1개 → 종속변수 1개

다중선형회귀

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

독립변수 p개 → 종속변수 1개

① 단순선형회귀보다 복잡한 관계를 더 잘 설명 가능

② 자연현상, 사회현상 파악에 유리

다중선형회귀

단순선형회귀

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

독립변수 1개 → 종속변수 1개

다중선형회귀

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

독립변수 p개 → 종속변수 1개

다중선형회귀 모델 해석

나머지 독립변수 X들이 **고정된 상태**에서

x_p 가 한 단위 증가할 때, y가 β_p만큼 증가함

모수의 추정 - 최소제곱법(LSE)

단순선형회귀와 동일하게 **최소제곱법(LSE)**을 활용해 모수 추정 가능

오차의 제곱합 : $\sum_i \left(y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_{pi}\right)^2$

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

$$\vdots$$

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

모수의 추정 - 최소제곱법(LSE)

단순선형회귀와 동일하게 **최소제곱법(LSE)**을 활용해 모수 추정 가능

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

모수가 (p+1)개인 **다차원** 식이기 때문에

편미분을 통해 추정 시 계산식이 매우 **복잡해짐!**

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

$$\vdots$$

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

행렬을 활용해 회귀계수를 추정!

모수의 추정

$$Y = X\beta + \epsilon \Leftrightarrow \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

✔ 다중선형회귀식을 행렬로 표현한 것!

✔ 행렬을 이용해 모수의 추정 가능!

최소제곱법

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

목적함수 S를 β에 대해 미분하고

미분식을 0으로 만들어주는 **추정량 $\hat{\beta}$** 를 구함!

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

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

최소 제곱법으로 추정된 회귀식 :

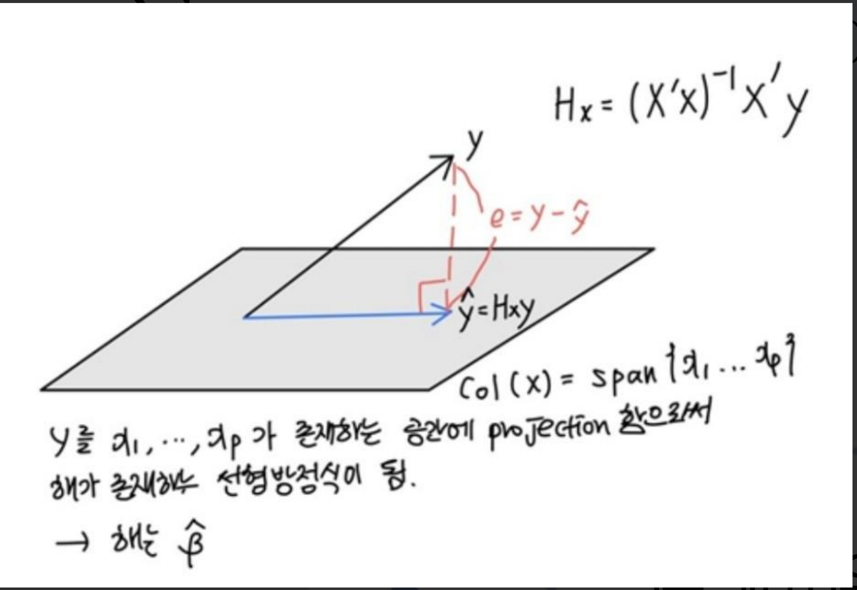
$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$$

($H = X(X'X)^{-1}X'$는 투영행렬)

최소제곱법

'투영행렬'이란?

$$S = \sum^{n} \ldots$$



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

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

…으로 추정된 회귀식 :

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

($H = X(X'X)^{-1}X'$는 투영행렬)

Y를 X의 열공간에 가깝게 근사시키기 위해 사용

→ Y를 X의 열공간에 투영시킴으로써 근사해 $\hat{\beta}$을 찾음

선대팀 2주차 클린업 예정!

유의성 검정

유의성 검정

추정량이 **통계적으로 유의**한지 알아보는 검정

다중선형회귀의 3가지 test

1. F-test

2. Partial F-test

3. T-Test

유의성 검정

1. F-test

전체 회귀계수에 대한 검정

가설 설정

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

$\boldsymbol{H_1}$: $\beta_0, \beta_1, \ldots, \beta_p$ 중 적어도 하나는 0이 아니다

귀무가설이 기각되어야 모형이 의미 있음

유의성 검정

1. F-test

전체 회귀계수에 대한 검정

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

MSR: 평균회귀제곱, MSE: 평균오차제곱

F_0 값이 임계치보다 충분히 크면 귀무가설을 기각할 수 있음

SSR과 SSE이 각각 분자, 분모에 위치하므로,

회귀식이 설명한 부분이 그렇지 않은 부분보다 충분히 크다는 것을 의미

유의성 검정

1. F-test

전체 회귀계수에 대한 검정

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

MSR: 평균회귀제곱, MSE: 평균오차제곱

F_0 값이 임계치보다 충분히 크면 귀무가설을 기각할 수 있음

→ SSR과 SSE이 각각 분자, 분모에 위치하므로,

회귀식이 설명한 부분이 그렇지 않은 부분보다 충분히 크다는 것을 의미

유의성 검정

1. F-test

전체 회귀계수에 대한 검정

임계값 : $\boldsymbol{F_{(1-\frac{a}{2},\, p,\, n-p-1)}}$

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

▶ 적어도 한 개의 회귀계수는 0이 아님

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

▶ 모든 회귀 계수는 0임

유의성 검정

1. F-test

전체 회귀계수에 대한 검정

임계값 : $\boldsymbol{F_{(1-\frac{a}{2},\, p,\, n-p-1)}}$

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

▶ 적어도 한 개의 회귀계수는 0이 아님

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

▶ 모든 회귀 계수는 0임

4 다중선형회귀

유의성 검정

1. F-test

F-test의 귀무가설이 기각되지 않는다면

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

회귀식이 아무 의미가 없음

임계값 : $F_{(1-\frac{\alpha}{2},\, p,\, n-p-1)}$

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

▶ 적어도 한 개의 회귀계수는 0이 아님

모델 재설정 등 조치가 필요

② 귀무가설 기각 안됨 if $F_0 < F_{(1-\frac{\alpha}{2},p,n-p-1)}$

▶ 모든 회귀 계수는 0임

유의성 검정

2. Partial F-test

일부 회귀계수에 대한 검정

가설 설정

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

H_1 : not H_0 (RM이 틀렸다, q개 중 적어도 한 개의 회귀계수가 0이 아니다)

유의성 검정

2. Partial F-test

일부 회귀계수에 대한 검정

검정 통계량

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

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

유의성 검정

2. Partial F-test

일부 회귀계수에 대한 검정

q개의 변수를 제거했을 때

모델이 설명하지 못하는 변동

검정 통계량

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

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

유의성 검정

2. Partial F-test

일부 회귀계수에 대한 검정

모든 변수를 포함했을 때

모델이 설명하지 못하는 변동

검정 통계량

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

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

유의성 검정

일반적으로 변수를 제거하면 $SSE(RM) > SSE(FM)$

2. Partial F-test

일부 회귀계수에 대한 검정

이때 **제거된 변수가 모델에 유의미** 하다면

$SSE(RM)$은 월등히 커짐

검정 통계량

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

검정통계량 F_0

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

검정통계량이 귀무가설을 기각시킬만큼 충분히 커짐

유의성 검정

2. Partial F-test

일부 회귀계수에 대한 검정

임계값 : $F_{\left(1-\frac{a}{2},p,n-p-1\right)}$

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

▶q개의 회귀계수 중 모든 회귀계수가 0이 아님

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

▶q개의 회귀계수 중 적어도 한 개의 회귀계수는 0임

유의성 검정

2. Partial F-test

일부 회귀계수에 대한 검정

임계값 : $F_{\left(1-\frac{a}{2},p,n-p-1\right)}$

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

▶ q개의 회귀계수 중 모든 회귀계수가 0이 아님

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

▶ q개의 회귀계수 중 적어도 한 개의 회귀계수는 0임

유의성 검정

3. T-test

개별 회귀계수에 대한 검정

가설 설정

$H_0 : \beta_j = 0$ (다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하지 않다)

$H_1 : \beta_j \neq 0$ (다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하다)

유의성 검정

3. T-test

개별 회귀계수에 대한 검정

검정통계량

$$t_J = \frac{\hat{\beta}_J}{s.e.(\hat{\beta}_J)}$$

t-test는 나머지 변수들이 다 적합된 상태에서

x_j를 추가적으로 적합했을 때 통계적 유의성을 검정

4 다중선형회귀

유의성 검정

3. T-test

개별 회귀계수에 대한 검정

임계값 : $\boldsymbol{t_{\left(\frac{a}{2},n-p-1\right)}}$

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

▶ x_j의 추가는 유의미한 회귀식의 설명력 증가를 가져옴

② 귀무가설 **기각** 안됨 if $|t_j| < t_{\left(\frac{a}{2},n-p-1\right)}$

▶ x_j의 추가는 유의미한 회귀식의 설명력 증가를 가져오지 않음

유의성 검정

3. T-test

개별 회귀계수에 대한 검정

임계값 : $\boldsymbol{t_{\left(\frac{a}{2}, n-p-1\right)}}$

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

▶ x_j의 추가는 유의미한 회귀식의 설명력 증가를 가져옴

② 귀무가설 기각 안됨 if $|t_j| < t_{\left(\frac{a}{2}, n-p-1\right)}$

▶ x_j의 추가는 유의미한 회귀식의 설명력 증가를 가져오지 않음

유의성 검정

3. T-test

개별 회귀계수에 대한 검정

그럼 F-test와 t-test는

뭐가 다른 거야?

F-test는 모든 설명변수의 유의성을,

T-test는 특정변수가 추가될 때의 유의성을 검정..

유의성 검정

3. T-test

개별 회귀계수에 대한 검정

그럼 F-test랑 T-test 중 대체
뭘 먼저 해야 하는데?

유의성 검정

F-test vs. T-test

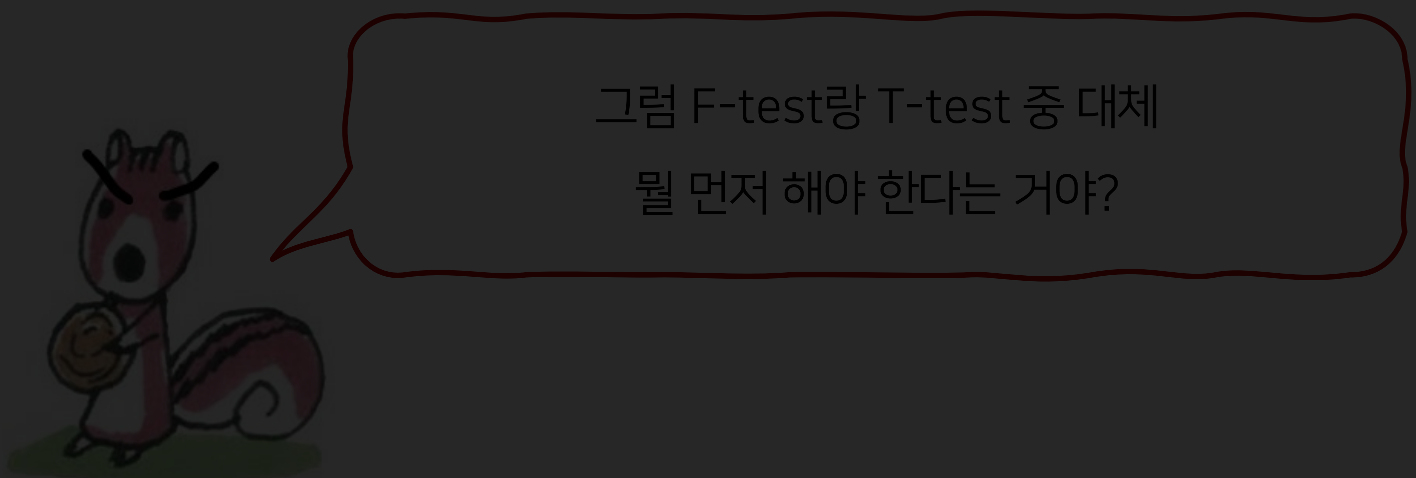
(F-Test를 먼저 수행해야 한다!)

3. T-test

개별 회귀계수에 대한 검정

① 전체 회귀식에 대한 검정이 더 엄격함

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



유의성 검정

3. T-test

개별 회귀계수에 대한 검정

F-test vs. T-test

(F-Test를 먼저 수행해야 한다!)

① 전체 회귀식에 대한 검정이 더 엄격함

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

F- test를 먼저 시행해 봄으로써

모델 전체가 **통계적으로 유의한지** 확인해야 함

4 다중선형회귀

유의성 검정

T-test로 변수 선택이 가능할까?

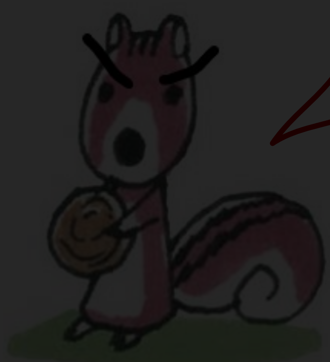
3. T-test

개별 회귀계수에 대한 검정

T-test는 다른 변수들이 다 **적합**된 상태에서

해당 변수의 추가가 유의미한 설명력 증가를 가져오는지 판단하는 것

그럼 F-test랑 T-test 중 대체
뭘 먼저 해야 한다는 거야?



유의성 검정

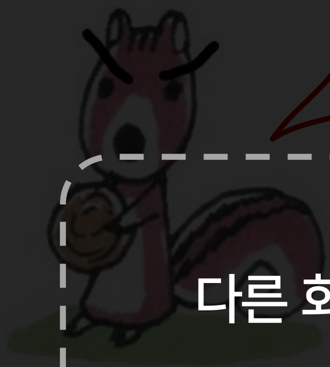
T-test로 변수 선택이 가능할까?

3. T-test

개별 회귀계수에 대한 검정

T-test는 다른 변수들이 다 **적합**된 상태에서

해당 변수의 추가가 유의미한 설명력 증가를 가져오는지 판단하는 것



그럼 F-test랑 T-test 중 대체

뭘 먼저 해야 한다는 거야?

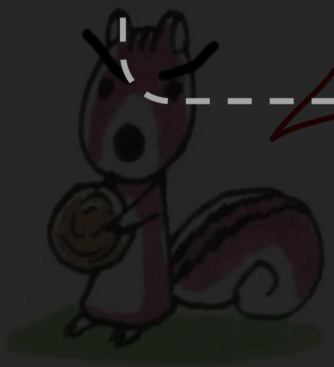
다른 회귀식을 가정하면 **해당 변수의 유의성**도 바뀔 수 있음

3. 유의성 검정

T-test로 변수 선택이 가능할까?

T-test

개별 회귀계수에 대한 검정



그럼 F-test랑 T-test 중 대체 뭘 먼저 해야 한다는 거야?

T-test로 변수를 선택하는 것은 **매우 위험 !**

적합성(Goodness of fit)

결정계수 (R square) R^2

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

관찰값의 **전체 변동** 대비 회귀 모델이 **설명한 변동**

모델을 잘 설명할수록 값이 1에 가까워짐

적합성(Goodness of fit)

결정계수 (R square) R^2

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

하지만 결정계수는 독립변수가 추가될 때 항상 값이 증가

아무 의미 없는 변수를 추가해도 값이 오름

중요하지 않은 변수를 추가함에 따라 모델의 해석도 어려워지고,

예측에도 좋지 않은 영향을 미침

독립변수가 늘어날 때, **페널티**를 줄 필요성이 생김

적합성(Goodness of fit)

수정결정계수 (Adjusted R square) R^2_{adj}

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

R^2 에 **변수 개수 증가**에 대한 페널티를 부과한 형태

✓ R^2_{adj} 가 높은 회귀 모델이 더 좋은 모델!

(그러나 R^2와 달리 그 자체로 해석이 어려움)

5

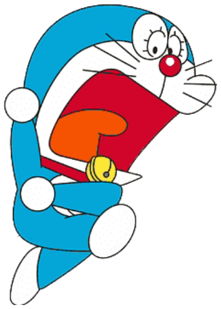
데이터 진단

왜 데이터를 진단해야 할까?

일반적인 경향을 벗어난 점들

왜곡

성능 저하



최소제곱 회귀모형에 **문제 발생!**

5 데이터 진단

잔차

잔차

설명할 수 없는 오차(ε)의 추정치.

관측된 종속변수(y)와 예측된 종속변수($\hat{y}$)의 차를 통해 구해짐

스튜던트화 잔차

Y값의 단위에 영향을 크게 받는 잔차의 한계를 해소하기 위해

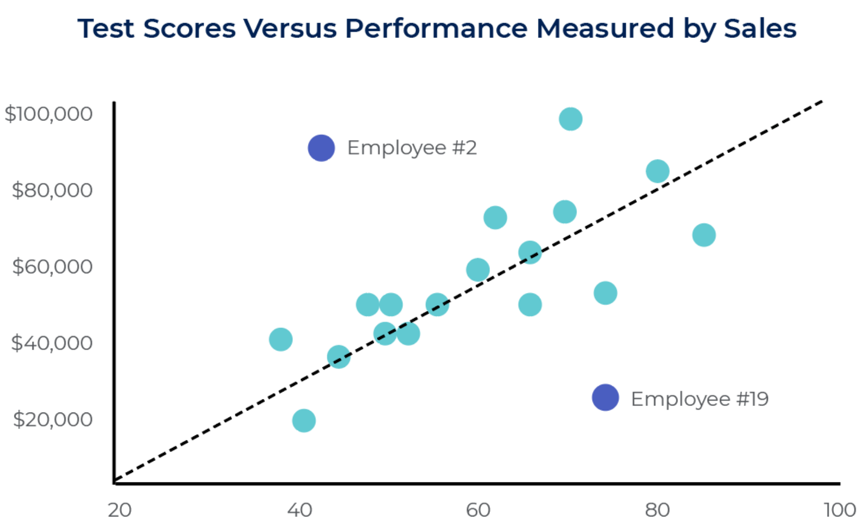
일반화해서 적용할 수 있도록 표준화한 것

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

이상치(Outlier)

이상치

스튜던트화 잔차가 매우 큰 값 (y를 기준으로 절댓값이 큰 값)



일반적으로 $|r_i| > 3$ 인 점을 이상치라고 판단

지렛값(Leverage point)

지렛값

x의 평균으로부터 멀리 떨어져 있어 기울기에 영향을 주는 값.

이상치가 y의 관점이었다면, 지렛값은 x를 기준으로 관찰 !

투영행렬

$$\mathrm{H} = \mathrm{X}(X^T X)^{-1} X^T$$

대각원소

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

지렛값(Leverage point)

지렛값

x의 평균으로부터 멀리 떨어져 있어 기울기에 영향을 주는 값.

이상치가 y의 관점이었다면, 지렛값은 x를 기준으로 관찰 !

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

$(x_i - \bar{x})^2$의 크기가 클수록,

즉 x_i(관측값)와 $\bar{x}$(평균)의 차가 클수록 h_{ii}가 증가한다.

투영행렬

$$H = X(X^TX)^{-1}X^T$$

대각원소

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

지렛값(Leverage point)

지렛값

x의 평균으로부터 멀리 떨어져 있어 기울기에 영향을 주는 값.

이상치가 y의 관점이었다면, 지렛값은 x를 기준으로 관찰 !

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

$h_{ii} > \frac{2(p+1)}{n}$ 인 값을 **지렛값**으로 판단!

투영행렬

$$\mathrm{H} = \mathrm{X}(X^T X)^{-1} X^T$$

대각원소

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

영향점(Influential point)

영향점

회귀 직선의 기울기에 유의미한 영향을 미치는 관측치

이상치와 지렛값을 동시에 고려함

이상치

x 평균 주변에 위치

↓

기울기를 변화시킬 수 X

+

지렛값

$\bar{x}$ 에서 멀리 떨어져있을 뿐,

회귀직선의 연장선 위에

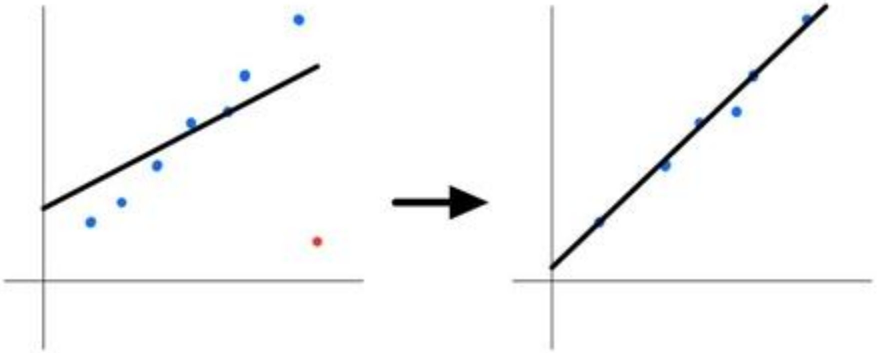
있을 수 있음

영향점(Influential point)

영향점

회귀 직선의 기울기에 유의미한 영향을 미치는 관측치

이상치와 지렛값을 동시에 고려함



▲ 빨간 점의 유무에 따라 회귀직선의 기울기가 크게 변화함.

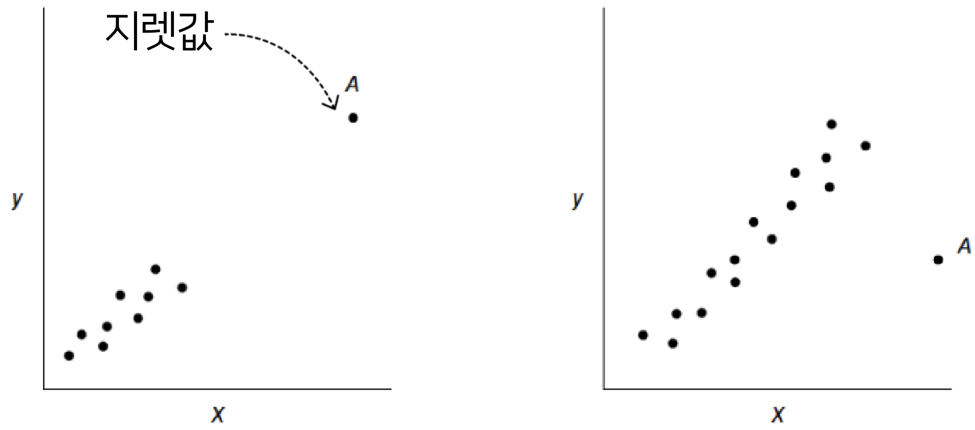
빨간 점은 **영향점**으로 간주될 수 있음!

영향점(Influential point)

영향점

회귀 직선의 기울기에 유의미한 영향을 미치는 관측치

이상치와 지렛값을 동시에 고려함



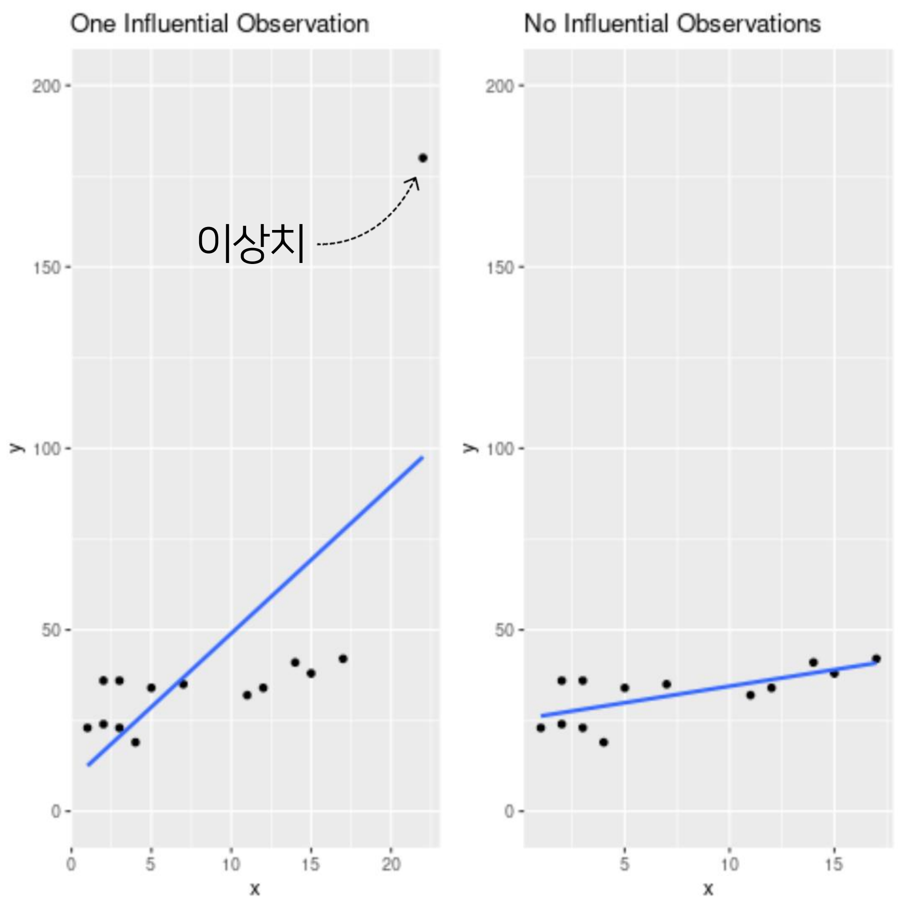
▲ 지렛값이라고 해서 모두 영향점인 것은 아님 !

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

이상치와 지렛값 각각이 커질수록 C_i (Cook’s Distance) 증가

보통 $C_i > 1$ 이면 영향점으로 판단

영향점(Influential Point)의 처리

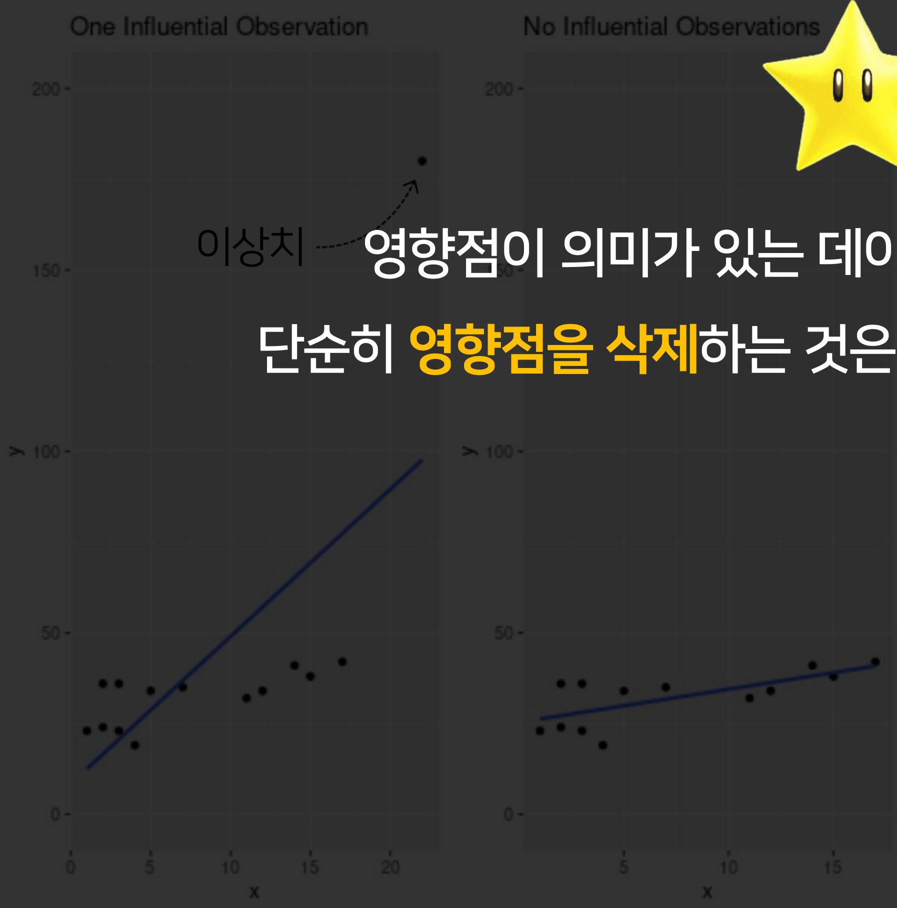


이상치는 **추정량을 불안정**하게 만듦

잘못된 모델의 해석, 예측 성능 저하

적절한 처리 必

영향점(Influential Point)의 처리



이상치는 **추정량을 불안정**하게 만듦

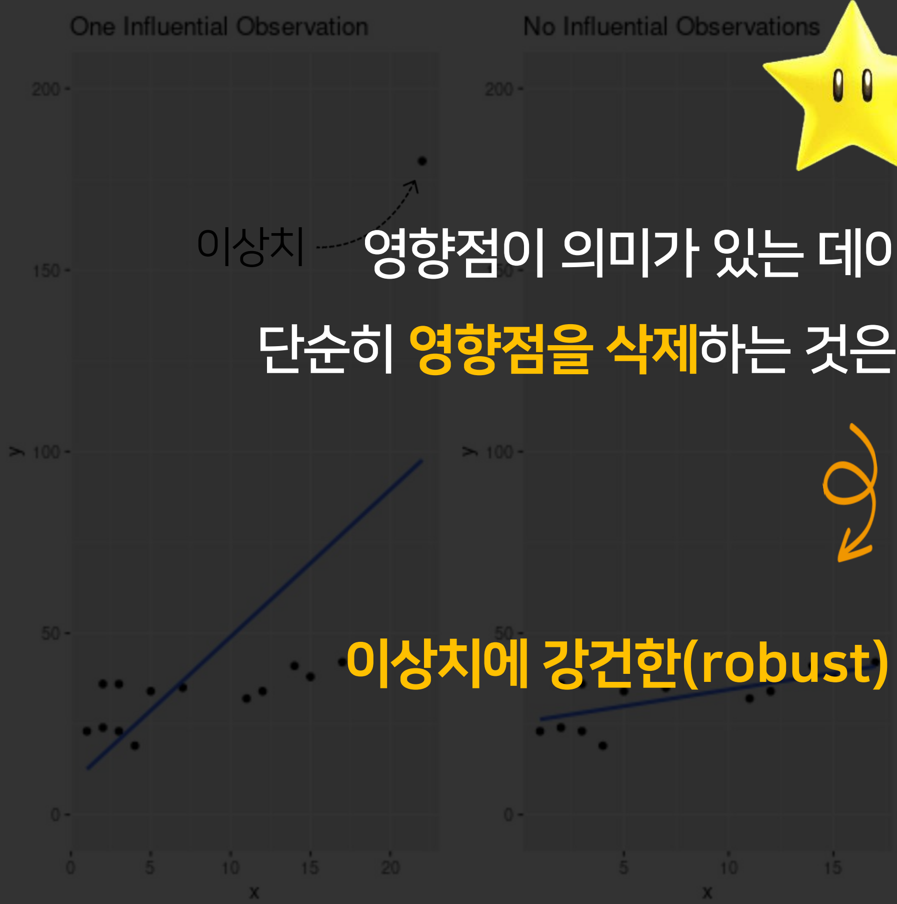
영향점이 의미가 있는 데이터일 수 있기 때문에,

단순히 **영향점을 삭제**하는 것은 적절한 처리로 볼 수 없음!

잘못된 모델의 해석, 예측 성능 저하

적절한 처리 必

영향점(Influential Point)의 처리



이상치는 **추정량을 불안정**하게 만듦

잘못된 모델의 해석, 예측 성능 저하

적절한 처리 必

영향점이 의미가 있는 데이터일 수 있기 때문에,

단순히 **영향점을 삭제**하는 것은 적절한 처리로 볼 수 없음!

이상치에 강건한(robust) 모델링이 필요한 이유!

6

로버스트 회귀

로버스트 회귀

로버스트 회귀

이상치의 영향을 줄이는 회귀분석 방법

Median Regression

Huber's M-estimation

Least Trimmed Square

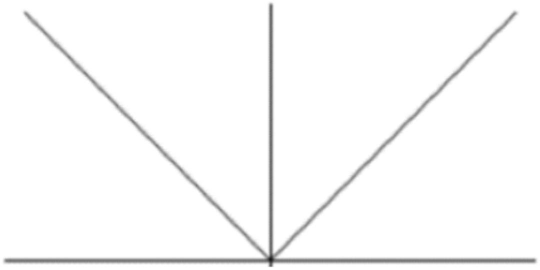
그리고… Support Vector Regression

Median Regression

Median Regression

이상치에 대해 너무 큰 가중치를 주는 최소제곱회귀의 단점을 극복하는 회귀분석 방법

▲ LSE



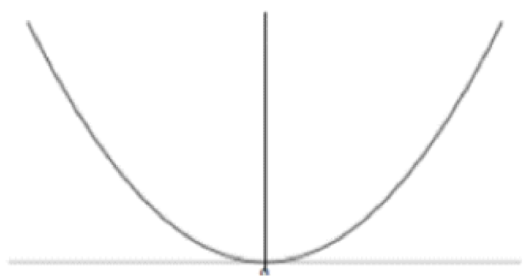
▲ Median Regression

모든 경우에 대해 **동일한 가중치**를 부여하는 방식으로 극복!

Median Regression

Median Regression

이상치에 대해 너무 큰 가중치를 주는 최소제곱회귀의 단점을 극복하는 회귀분석 방법



▲ LSE

오차의 제곱합을 최소로 하는 추정량 :

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

X에 따른 **평균적인** Y를 반환:

조건부 평균 E(Y|X)

Median Regression

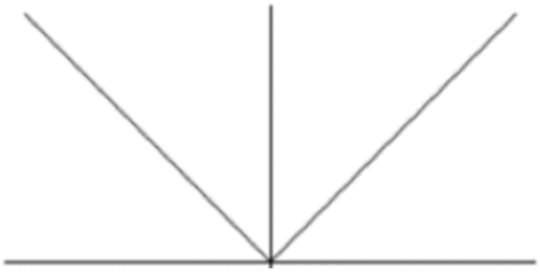
Median Regression

이상치에 대해 너무 큰 가중치를 주는 최소제곱회귀의 단점을 극복하는 회귀분석 방법

오차의 절댓값의 합을 최소로 하는 추정량 :

$\sum|\varepsilon_i|$

X에 따른 Y의 **조건부 중앙값**을 반환



▲ Median Regression

Median Regression

Median Regression

이상치에 대해 너무 큰 가중치를 … 최소제곱회귀의 단점을 극복하는
회귀 … 방법

중앙값이 이상치의 영향으로부터 비교적 자유롭다는 점을 이용한 것!

오차의 절댓값의 합을 최소로 하는 추정량 :
$\sum|\varepsilon_i|$

X에 따른 Y의 **조건부 중앙값**을 반환

▲ Median Regression

Huber's M-estimation

Huber's M-estimation

이상치에 너무 큰 가중치를 주는 최소제곱회귀의 단점을 극복하면서

적정 수준 내에서 페널티를 완화시키는 최소제곱회귀의 장점을 이용한 분석 방법

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

잔차의 절댓값이 c 이하

→ 최소제곱추정법의 목적함수와 동일

Huber's M-estimation

Huber's M-estimation

이상치에 너무 큰 가중치를 주는 최소제곱회귀의 단점을 극복하면서

적정 수준 내에서 페널티를 완화시키는 최소제곱회귀의 장점을 이용한 분석 방법

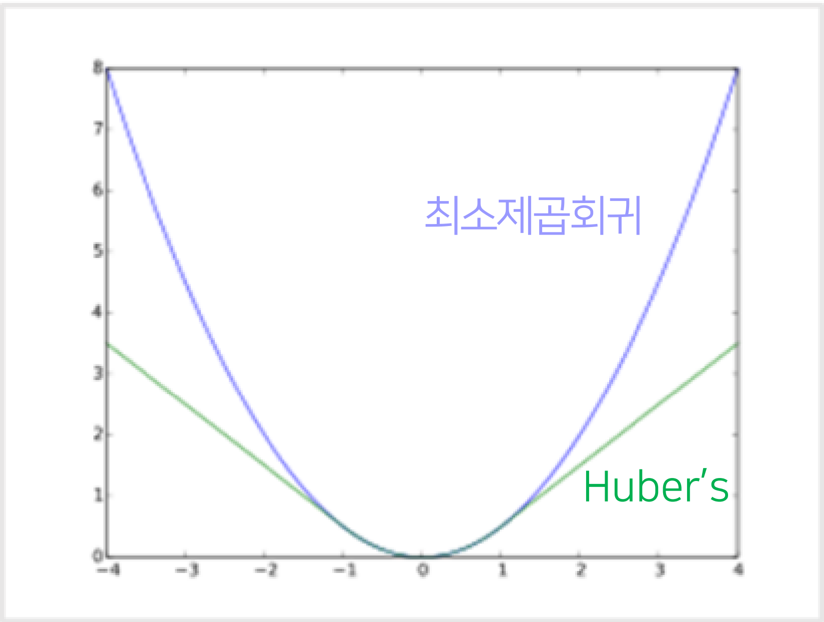
$$\mathrm{p(e)} = \begin{cases} \frac{1}{2}e^2, & if\ |e| \le c \\ c|e| - \frac{1}{2}c^2, & otherwise \end{cases}$$

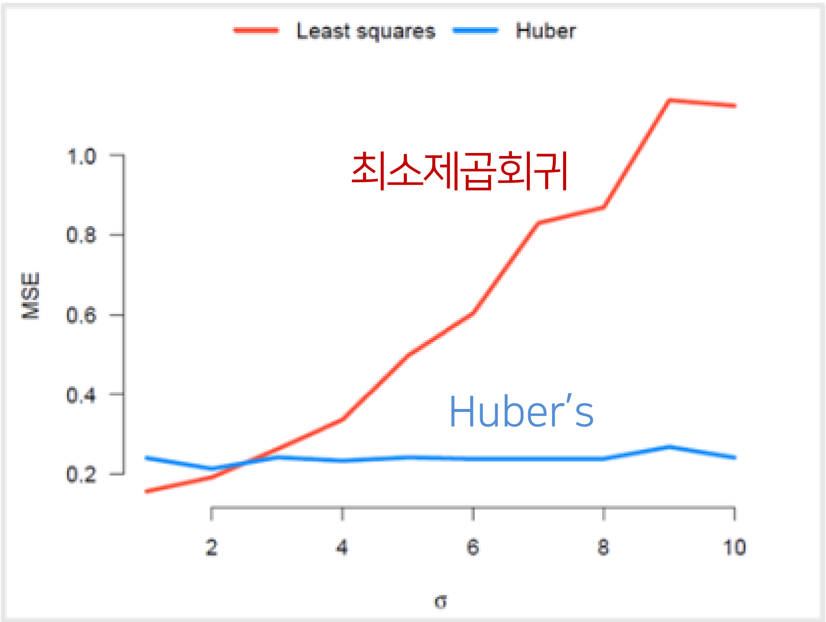
잔차의 절댓값이 c 이상

→ 이상치에 큰 페널티를 주지 않는

일차식의 형태

Huber’s M-estimation





이상치에 대한 페널티를 완화하여 MSE 값을 줄임

Least Trimmed Square

Least Trimmed Square

통계적 기준에 따라 **잔차가 너무 큰 관측치를 제거**하고

회귀계수를 추정하는 회귀분석 방법

$r_1 \sim r_h$ 의 제곱합

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

※ $r_{(j)}$ = j번째로 작은 residual

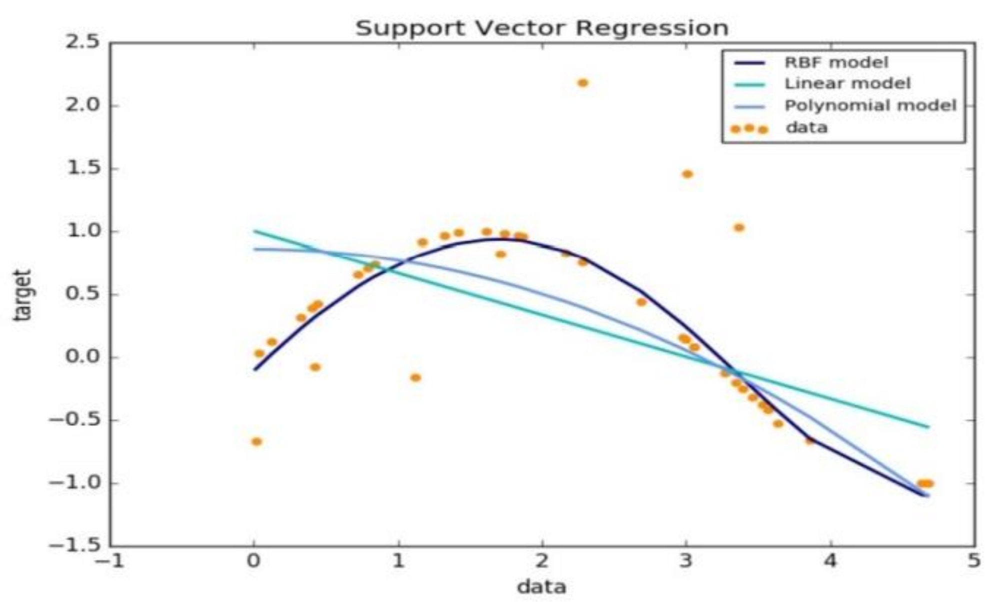
관측값의 개수가 적거나

영향점이 없는 경우 주의해서 사용!

Support Vector Regression

Support Vector Regression

Robust하면서 **비선형**적인 모델링이 가능한 회귀분석 방법



다음 주 예고

1. 회귀 기본 가정
2. 잔차 플랏
3. 선형성 진단과 처방
4. 정규성 진단과 처방
5. 등분산성 진단과 처방
6. 독립성 진단과 처방