회귀분석팀

6팀

조수미
김민지
손재민
박윤아
조웅빈

CONTENTS

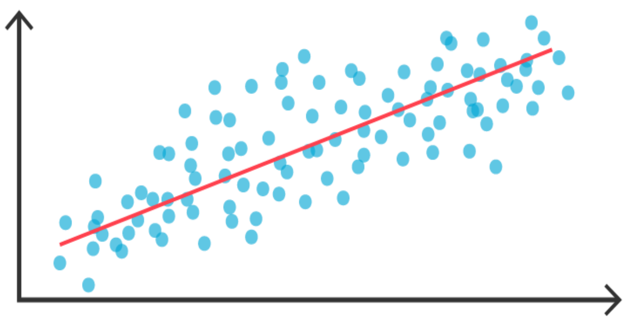
1

회귀분석이란?

회귀분석이란?

“ Regression + Analysis ”

회귀선을 찾아 관계를 설명!

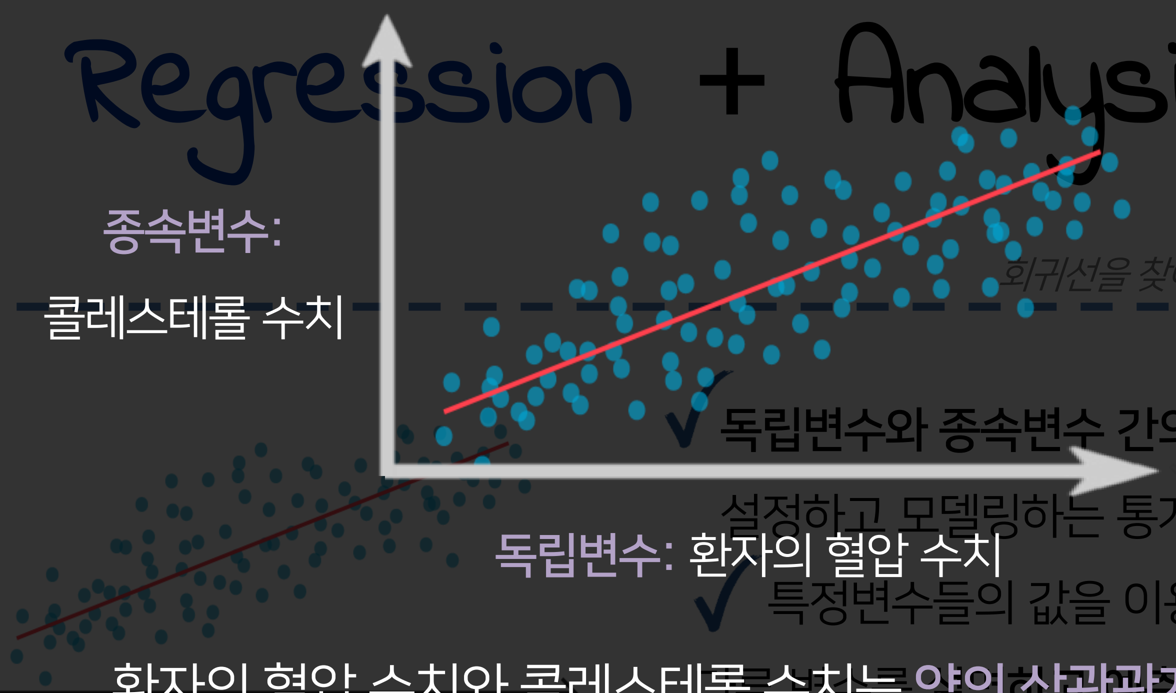


✓ **독립변수와 종속변수 간의 관계를** 설정하고 모델링하는 통계적 기법

✓ 특정변수들의 값을 이용하여 다른 변수를 **설명하고 예측**하는 방법

회귀분석이란?

"Regression + Analysis"



✓ 독립변수와 종속변수 간의 관계를 설정하고 모델링하는 통계적 기법

✓ 특정변수들의 값을 이용하여 … 하는 방법

환자의 혈압 수치와 콜레스테롤 수치는 양의 상관관계

회귀식

회귀식

종속변수 Y와 독립변수 X의 관계를 **함수식(f)**으로 표현한 모델

$$Y = f\left(X_1, X_2, \cdots, X_p\right) + \varepsilon$$

Y **종속변수** 독립변수에 의해서 설명되는 변수

X_k **독립변수** 종속변수를 설명하기 위한 변수

ε **오차항** 변수를 측정할 때 발생할 수 있는 오차

설명할 수 없는 무작위성을 가짐

회귀식

회귀식

종속변수 Y와 독립변수 X의 관계를 **함수식(f)**으로 표현한 모델

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

Y **종속변수** 독립변수에 의해서 설명되는 변수

X_k **독립변수** 종속변수를 설명하기 위한 변수

ε **오차항** 변수를 측정할 때 발생할 수 있는 오차

설명할 수 없는 무작위성을 가짐

회귀식

회귀식

종속변수 Y와 독립변수 X의 관계를 **함수식(f)**으로 표현한 모델

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

Y 종속변수 독립변수에 의해서 설명되는 변수

X_k **독립변수** 종속변수를 설명하기 위한 변수

ε 오차항 변수를 측정할 때 발생할 수 있는 오차

설명할 수 없는 무작위성을 가짐

회귀식

회귀식

종속변수 Y와 독립변수 X의 관계를 **함수식(f)**으로 표현한 모델

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

Y 종속변수 독립변수에 의해서 설명되는 변수

X_k 독립변수 종속변수를 설명하기 위한 변수

ε **오차항** 변수를 측정할 때 발생할 수 있는 오차

설명할 수 없는 무작위성을 가짐

1 회귀분석이란?

회귀 모델링 과정

① 문제정의

내 학점에 영향을 주는 요소에는 무엇이 있을까?

② 적절한 변수 선택

통학 거리, 공부 시간 등이 영향을 주지 않을까?

③ 데이터 수집 및 전처리

선정한 변수에 맞는 학생 데이터를 수집 및 전처리

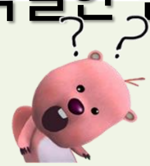
1 회귀분석이란?

회귀 모델링 과정

① 문제정의

내 학점에 영향을 주는 요소에는 무엇이 있을까?

② 적절한 변수 선택



통학 거리, 공부 시간 등이 영향을 주지 않을까?

③ 데이터 수집 및 전처리

선정한 변수에 맞는 학생 데이터를 수집 및 전처리

1 회귀분석이란?

회귀 모델링 과정

① 문제정의

내 학점에 영향을 주는 요소에는 무엇이 있을까?

② 적절한 변수 선택

통학 거리, 공부 시간 등이 영향을 주지 않을까?

③ 데이터 수집 및 전처리

선정한 변수에 맞는 학생 데이터를 수집 및 전처리

1 회귀분석이란?

회귀 모델링 과정

④ 모델설정과 적합

선형 vs 비선형 / 단순회귀 vs 다중회귀 등을 고려

⑤ 모형 평가

모델이 회귀분석의 가정을 만족하는지 평가

⑥ 모형 해석

현재상태에서 통학거리를 30분 줄이고,
공부시간을 한 시간 늘리면 학점이 0.5 정도 오를 것이다

회귀 모델링 과정

④ 모델설정과 적합

선형 vs 비선형 / 단순회귀 vs 다중회귀 등을 고려

⑤ 모형 평가

모델이 회귀분석의 가정을 만족하는지 평가

⑥ 모형 해석

현재상태에서 통학거리를 30분 줄이고,
공부시간을 한 시간 늘리면 학점이 0.5 정도 오를 것이다

회귀 모델링 과정

④ 모델설정과 적합

선형 vs 비선형 / 단순회귀 vs 다중회귀 등을 고려

⑤ 모형 평가

모델이 회귀분석의 가정을 만족하는지 평가

⑥ 모형 해석

현재상태에서 통학거리를 30분 줄이고,
공부시간을 한 시간 늘리면 학점이 0.5 정도 오를 것이다

2

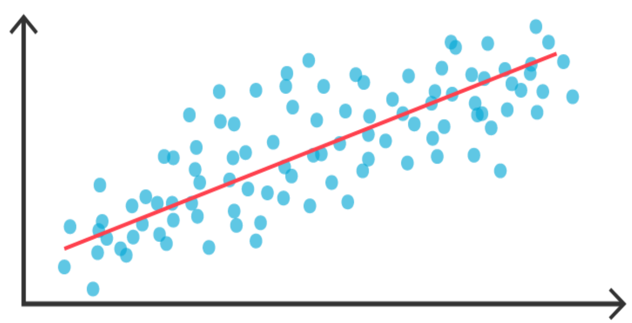
단순선형회귀

단순선형회귀

단순선형회귀 Simple Linear Regression

하나의 종속변수와 하나의 독립변수만을 가짐

두 변수의 관계를 가장 잘 표현하는 **직선을 추정**하는 것이 목적



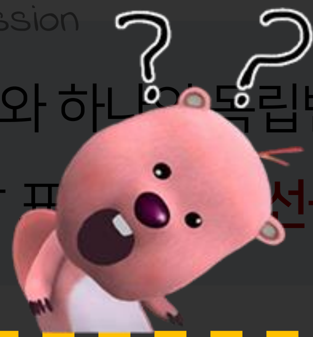
$$y_i = \beta_0 + \beta_1 x_1 + \epsilon, \ \ \epsilon_i \sim NID(0, \sigma^2)$$

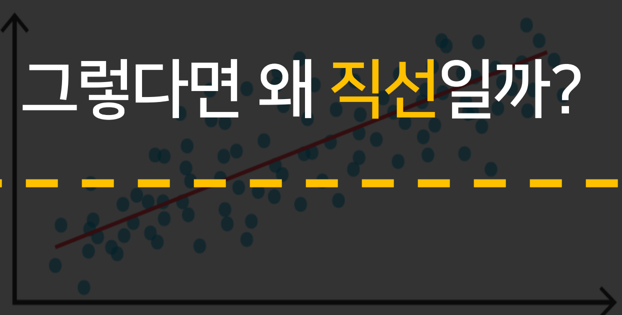
단순선형회귀

단순선형회귀 Simple Linear Regression

하나의 종속변수와 하나의 독립변수만을 가짐

두 변수의 관계를 가장 잘 표 … **선을 추정**하는 것이 목적

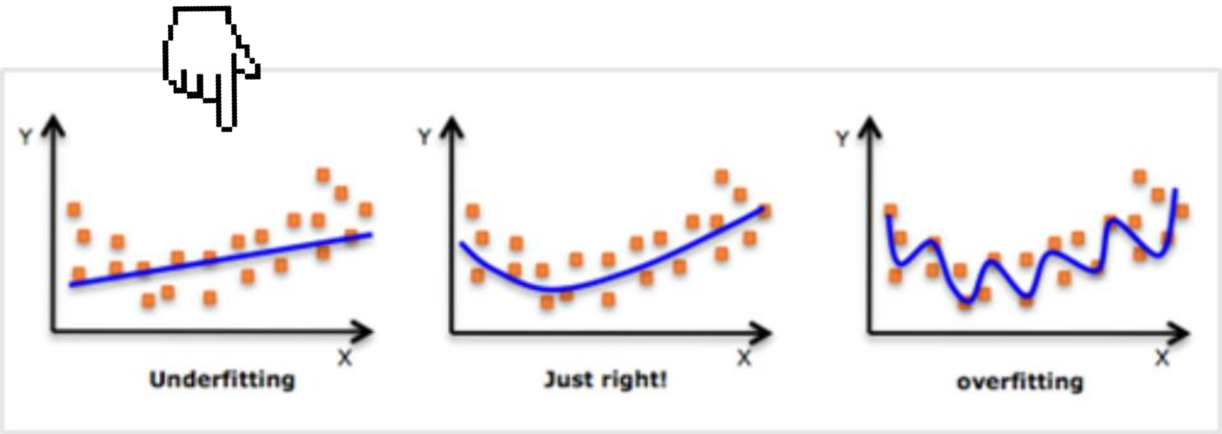




그렇다면 왜 **직선**일까?

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon,\ \ \epsilon_i \sim NID(0, \sigma^2)$$

단순선형회귀



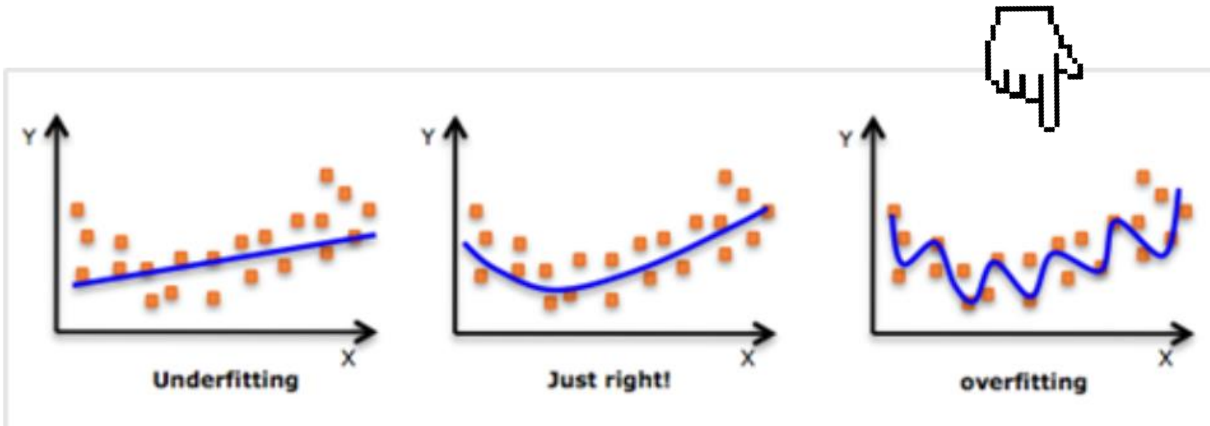
단순한 모델　　실제 모델　　복잡한 모델

변수의 영향력을 **간단하게 모형화** 할 수 있기 때문!

위의 그림처럼 2차원 평면 위의 직선은

X와 Y의 일대일대응 관계를 통해 **변화율을 직관적으로 이해** 가능

단순선형회귀



단순한 모델 실제 모델 복잡한 모델

고차함수로 추정할 경우

모델의 복잡도가 높아져 **과적합** 우려

단순선형회귀

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

y_i **종속변수** 종속변수 y의 i번째 관측값

x_i **독립변수** 독립변수 x의 i번째 관측값

ε_i **오차항** i번째 관측값에 의한 랜덤 오차

β_0, β_1 **회귀계수** 추정해야 할 모수

단순선형회귀

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

y_i **종속변수** 종속변수 y의 i번째 관측값

x_i **독립변수** 독립변수 x의 i번째 관측값

ε_i **오차항** i번째 관측값에 의한 랜덤 오차

β_0, β_1 **회귀계수** 추정해야 할 모수

단순선형회귀

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

y_i **종속변수** 종속변수 y의 i번째 관측값

x_i **독립변수** 독립변수 x의 i번째 관측값

ε_i **오차항** i번째 관측값에 의한 랜덤 오차

β_0, β_1 **회귀계수** 추정해야 할 모수

단순선형회귀

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

y_i 종속변수 종속변수 y의 i번째 관측값

x_i 독립변수 독립변수 x의 i번째 관측값

ε_i 오차항 i번째 관측값에 의한 랜덤 오차

$\boldsymbol{\varepsilon_i \sim NID(0, \sigma^2)}$

평균은 0, 분산은 σ^2를 가정

β_0, β_1 회귀계수 추정해야 할 모수

단순선형회귀

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

y_i **종속변수** 종속변수 y의 i번째 관측값

x_i **독립변수** 독립변수 x의 i번째 관측값

ε_i **오차항** i번째 관측값에 의한 랜덤 오차

$\beta_0,\ \beta_1$ **회귀계수** 추정해야 할 모수

회귀계수의 의미

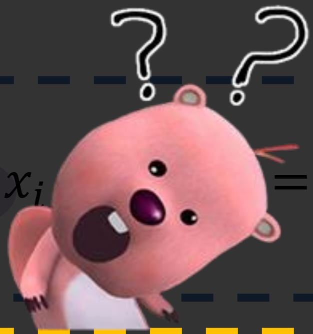
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

β_0: $x_i = 0$ 일때, **예측된 y값**

β_1: 설명변수가 한 단위 증가할 때 **종속변수가 변화하는 정도**

회귀계수의 의미

$y_i = \beta_0 + \beta_1 x_i \quad \cdots = 1, 2, \ldots, n$



그렇다면 **좋은 추정**이란 무엇일까?

!!

β_0: $x_1 = 0$ 일때, 예측된 y값

좋은 추정: 실제 데이터와 추정된 함수 사이의 오차가 최소화되는 경우

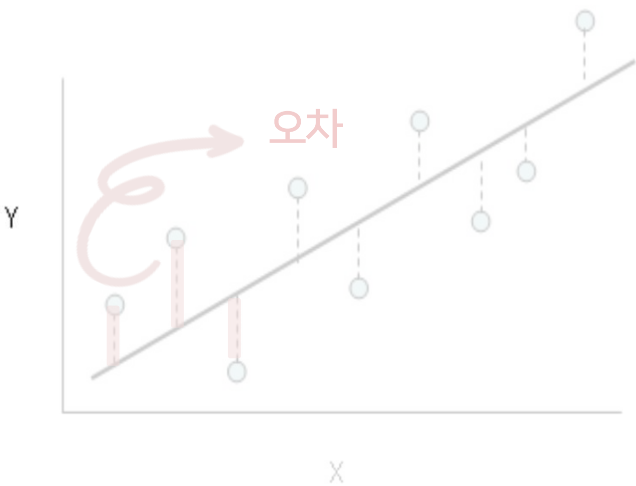
β_1: 설명변수가 한 단위 증가할 때 종속변수가 변화하는 정도

최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법



오차의 제곱합 최소화

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

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

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

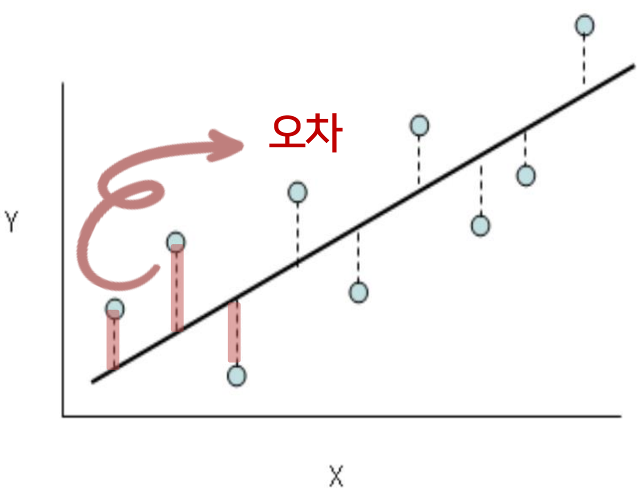
아래로 볼록한 이차함수 형태는
최소값을 가지므로 편미분!

최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법



오차의 제곱합 최소화

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

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

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

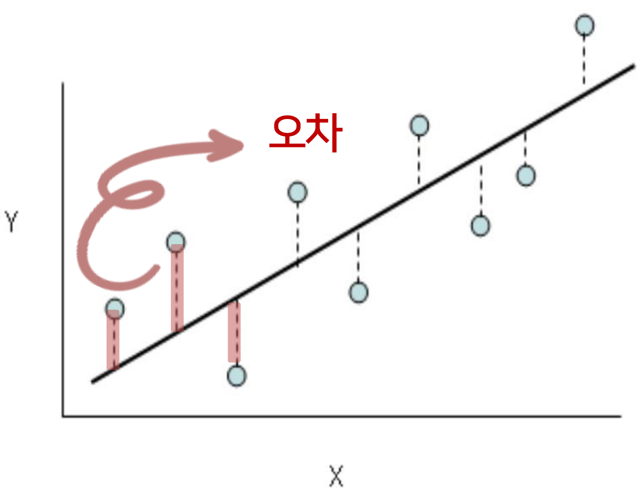
아래로 볼록한 이차함수 형태는
최소값을 가지므로 편미분!

최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법



오차의 제곱합 최소화

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

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

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

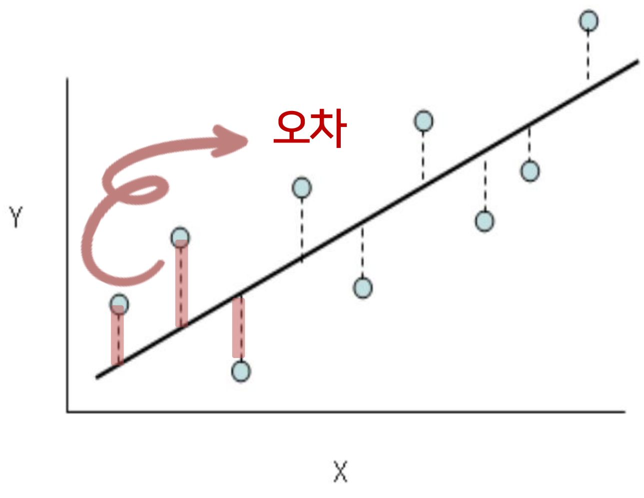
아래로 볼록한 이차함수 형태는
최소값을 가지므로 편미분!

최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법



오차의 제곱합 최소화

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

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

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

아래로 볼록한 이차함수 형태는
최소값을 가지므로 편미분!

최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법

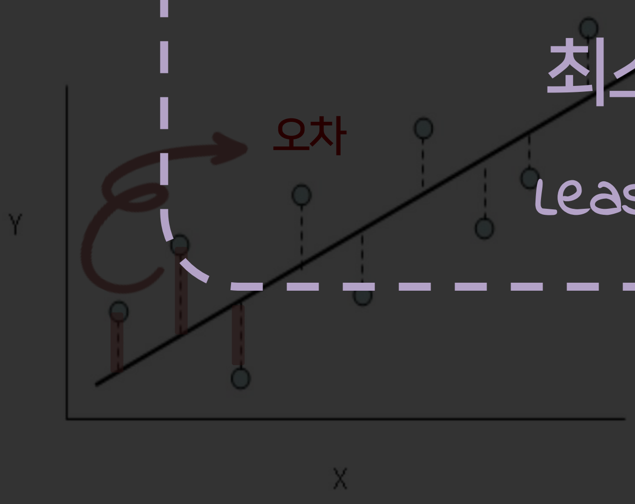
최소제곱법을 통해 얻은 추정치 $\widehat{\beta_0}$과 $\widehat{\beta_1}$

최소제곱추정치(LSE)

Least Sqaure Estimator

오차의 제곱합 최소화

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



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

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

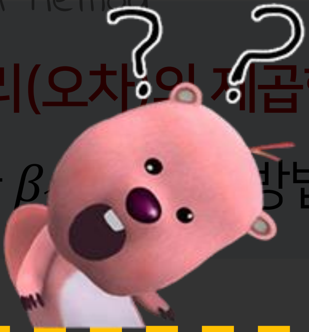
아래로 볼록한 이차함수 형태는 최소값을 가지므로 편미분!

최소제곱법

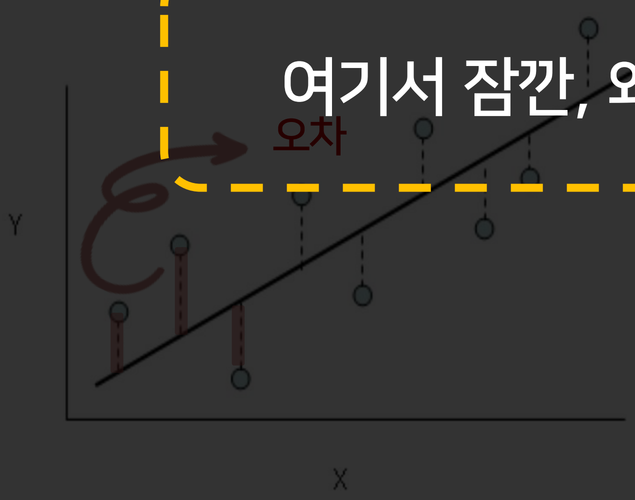
최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β 방법



여기서 잠깐, 왜 **오차의 제곱합**을 최소화할까?



오차의 제곱합 최소화

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

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

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

아래로 볼록한 이차함수 형태는
최소값을 가지므로 편미분!

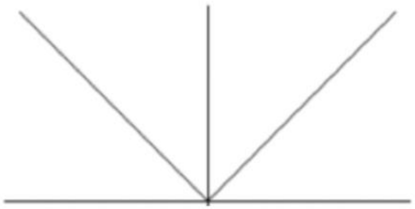
최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법

$y = x^2$



$y = |x|$

① 미분의 편리성

② 오차가 클수록 더 큰 패널티 부여 가능

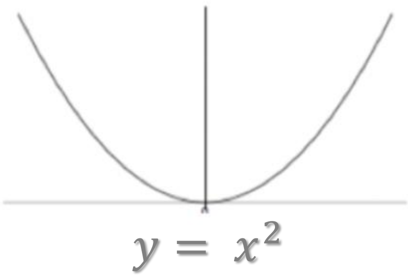
③ 오차의 절대값 사용 시 미분불가능한 점이 존재

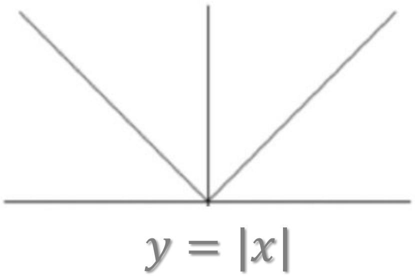
최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법





① 미분의 편리성

② 오차가 클수록 더 큰 패널티 부여 가능

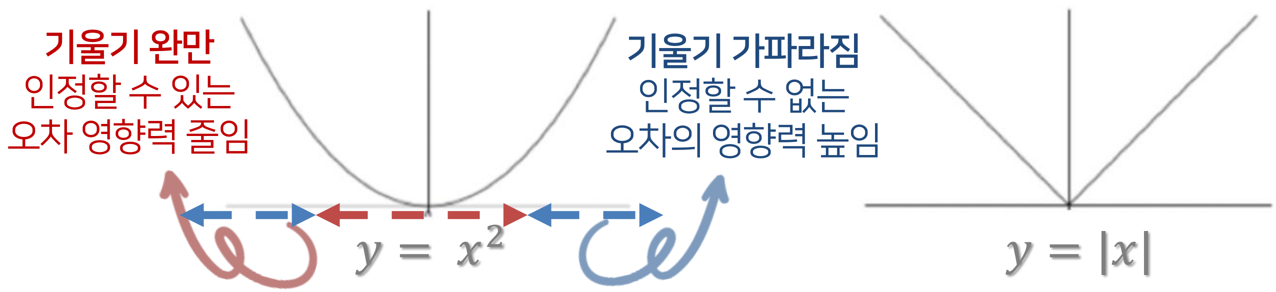
③ 오차의 절대값 사용 시 미분불가능한 점이 존재

최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법



① **미분의 편리성**

② 오차가 클수록 더 큰 **패널티 부여** 가능

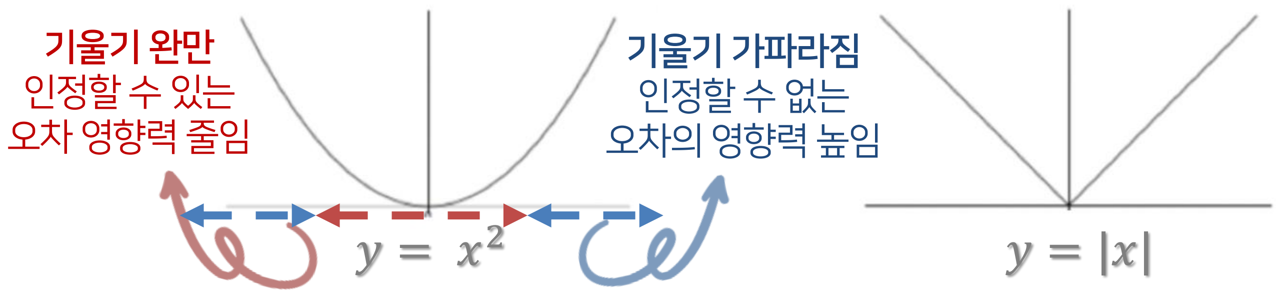
③ 오차의 절대값 사용 시 **미분불가능한 점**이 존재

최소제곱법

최소제곱법 Least Square Estimator Method

y_i와 회귀선 위의 y값의 거리(오차)의 제곱합이 최소가 되도록 하는

β_0과 β_1 을 찾는 방법



① **미분의 편리성**

② 오차가 클수록 더 큰 **패널티 부여** 가능

③ 오차의 절대값 사용 시 **미분불가능한 점**이 존재

최소제곱법의 가정과 특징

BLUE Best Linear Unbiased Estimator

분산이 제일 작은 선형 불편추정량

분산이 작다는 것은 **추정량이 안정적**이라는 의미

① 오차들의 **평균은 0**

② 오차들의 **분산은 σ^2 로 동일**

③ 오차간 **자기상관이 없음**
Independent

세 가지 조건이 만족되면, LSE는 선형불편추정량 중
분산이 가장 작은 안정적인 추정량이 됨

최소제곱법의 가정과 특징

BLUE Best Linear Unbiased Estimator

분산이 제일 작은 선형 불편추정량

분산이 작다는 것은 **추정량이 안정적**이라는 의미

① 오차들의 **평균은 0**

② 오차들의 **분산은** σ^2 로 동일

③ 오차간 **자기상관이 없음**

Independent

세 가지 조건이 만족되면, LSE는 선형불편추정량 중
분산이 가장 작은 안정적인 추정량이 됨

최소제곱법의 가정과 특징

BLUE Best Linear Unbiased Estimator

분산이 제일 작은 선형 불편추정량

분산이 작다는 것은 **추정량이 안정적**이라는 의미

① 오차들의 **평균은 0**

② 오차들의 **분산은 σ^2 로 동일**

③ 오차간 **자기상관이 없음**
Independent

세 가지 조건이 만족되면, LSE는 선형불편추정량 중
분산이 가장 작은 안정적인 추정량이 됨

최소제곱법의 가정과 특징

BLUE Best Linear Unbiased Estimator

분산이 제일 작은 선형 불편추정량

분산이 작다는 것은 **추정량이 안정적**이라는 의미

① 오차들의 **평균은 0**

② 오차들의 **분산은 σ^2 로 동일**

③ 오차간 **자기상관이 없음**
Independent

세 가지 조건이 만족되면, LSE는 선형불편추정량 중
분산이 가장 작은 안정적인 추정량이 됨

최소제곱법의 가정과 특징

BLUE Best Linear Unbiased Estimator

분산이 제일 작은 선형 불편추정량

분산이 작다는 것은 **추정량이 안정적**이라는 의미

① 오차들의 **평균은 0**

② 오차들의 **분산은** σ^2 **로 동일**

③ 오차간 **자기상관이 없음**
Independent

세 가지 조건이 만족되면, LSE는 선형불편추정량 중
분산이 가장 작은 안정적인 추정량이 됨

최대가능도추정법

최대가능도추정법 Maximum Likelihood Estimation

확률적인 방법에 근거해, 데이터가 나올
가능도(Likelihood)를 최대로 하는 모수를 추정

✓ 추정 관측치가 항상 iid라는 가정 필수
✓ **오차의 정규분포**를 가정할 때
MLE 와 LSE는 완전히 **동일한 추정량** 가짐

최대가능도추정법

최대가능도추정법 Maximum Likelihood Estimation

확률적인 방법에 근거해, 데이터가 나올

가능도(Likelihood)를 최대로 하는 모수를 추정

✓ 추정 관측치가 항상 iid라는 가정 필수

✓ 오차의 정규분포를 가정할 때

MLE 와 LSE는 완전히 **동일한 추정량** 가짐

최대가능도추정법

최대가능도추정법 Maximum Likelihood Estimation

확률적인 방법에 근거해, 데이터가 나올

가능도(Likelihood)를 최대로 하는 모수를 추정

✓ **추정 관측치가 항상 iid**라는 가정 필수

✓ **오차의 정규분포**를 가정할 때

MLE 와 LSE는 완전히 **동일한 추정량** 가짐

적합성 검정

적합성 검정 Goodness of Fit

회귀직선이 데이터에 **얼마나 잘 들어맞는지** 모형에 대한 적합성 검정

오차

모집단에서 얻은 회귀식에서의 실제값 y_i 와 추정값 $\hat{y}_i$의 차이

↓

오차는 모집단의 측정치

vs

잔차

표본에서 얻은 회귀식에서의 실제값 y_i 와 추정값 $\hat{y}_i$의 차이

↓

잔차는 표본의 측정치

적합성 검정

적합성 검정 Goodness of Fit

회귀직선이 데이터에 **얼마나 잘 들어맞는지** 모형에 대한 적합성 검정

오차

모집단에서 얻은 회귀식에서의

실제값 y_i와 추정값 $\hat{y}_i$의 차이

오차는 모집단의 측정치

vs

잔차

표본에서 얻은 회귀식에서의

실제값 y_i와 추정값 $\hat{y}_i$의 차이

잔차는 표본의 측정치

적합성 검정

적합성 검정 Goodness of Fit

회귀직선이 데이터에 **얼마나 잘 들어맞는지** 모형에 대한 적합성 검정

오차

모집단에서 얻은 회귀식에서의 실제값 y_i와 추정값 $\hat{y}_i$의 차이

오차는 모집단의 측정치

vs

잔차

표본에서 얻은 회귀식에서의 실제값 y_i와 추정값 $\hat{y}_i$의 차이

잔차는 **표본**의 측정치

적합성 검정

적합성 검정 Goodness of Fit

회귀직선이 데이터에 **얼마나 잘 들어맞는지** 모형에 대한 적합성 검정

모집단에서의 실제 오차값을 알 수 없기에 잔차 사용

오차

모집단에서 얻은 회귀식에서의 실제값 y_i와 추정값 $\hat{y}_i$의 차이

오차는 모집단의 측정치

vs

잔차

표본에서 얻은 회귀식에서의 실제값 y_i와 추정값 $\hat{y}_i$의 차이

잔차는 표본의 측정치

적합성 검정

적합성 검정 Goodness of Fit

회귀직선이 데이터에 **얼마나 잘 들어맞는지** 모형에 대한 적합성 검정

모집단에서의 실제 오차값을
알 수 없기에 잔차 사용

잔차의 공식

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad \sum e_i = 0$$

오차

모집단에서 얻은 회귀식에서의
실제값 y_i와 추정값 $\hat{y}_i$의 차이

오차는 모집단의 측정치

vs

잔차

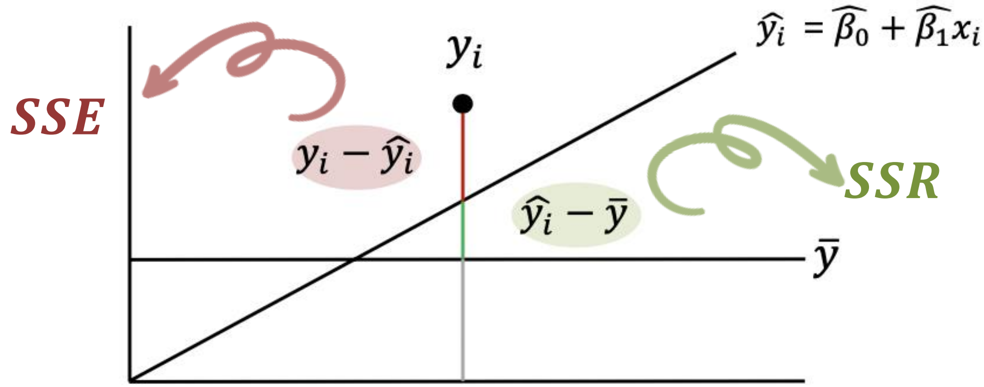
표본에서 얻은 회귀식에서의
실제값 y_i와 추정값 $\hat{y}_i$의 차이

잔차는 표본의 측정치

적합성 검정

변동분할

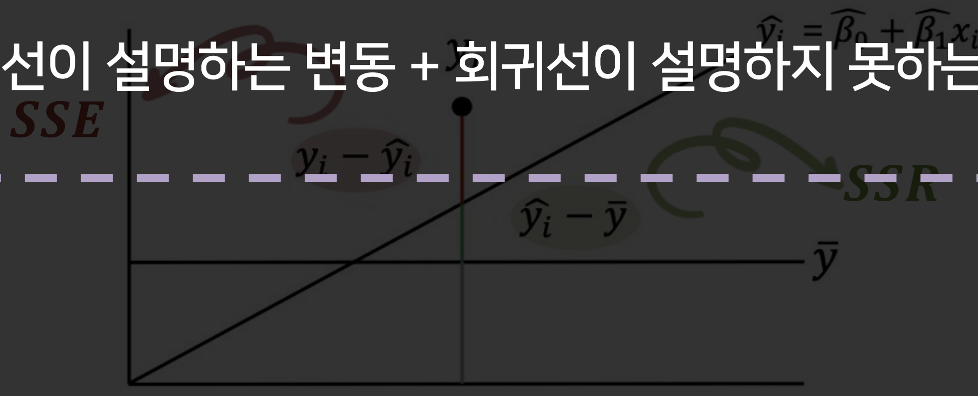
- 총변동(Total Sum of Squares, *SST*) : $\sum(y_i - \bar{y})^2$
- 회귀변동(Regression Sum of Square, *SSR*) : $\sum(\hat{y}_i - \bar{y})^2$
- 오차변동(Residual Sum of Square, *SSE*) : $\sum(y_i - \hat{y}_i)^2$



적합성 검정

변동분할

- 총변동(Total Sum of Squares, SST) : $\sum(y_i - \bar{y})^2$
- 회귀변동(Regression Sum of Square, SSR) : $\sum(\hat{y}_i - \bar{y})^2$
- 오차변동(Residual Sum of Square, SSE) : $\sum(y_i - \hat{y}_i)^2$



$$SST = SSR + SSE$$

총 변동은

회귀선이 설명하는 변동 + 회귀선이 설명하지 못하는 변동

적합성 검정

결정계수 Coefficient of Determinant

총 변동(SST)에서 회귀식이 설명할 수 있는 비율(SSR)

즉, Y가 X에 의해 설명되는 비율로, 1에 가까울수록 좋음

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

잔차와 연관지어 본다면,

잔차제곱합(SSE)은 회귀식이 설명할 수 없는 실제값과 추정값 사이의 오차이므로

총 변동 대비 잔차제곱합이 차지하는 비율이 작을수록 좋음

적합성 검정

결정계수 Coefficient of Determinant

총 변동(SST)에서 회귀식이 설명할 수 있는 비율(SSR)

즉, Y가 X에 의해 설명되는 비율로, 1에 가까울수록 좋음

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

잔차와 연관지어 본다면,

잔차제곱합(SSE)은 회귀식이 설명할 수 없는 실제값과 추정값 사이의 오차이므로

총 변동 대비 잔차제곱합이 차지하는 비율이 작을수록 좋음

유의성 검정

유의성 검정 Significance Test

전체 회귀식이 아닌 **개별 모수**의 추정량이 통계적으로 유의한지를 알아보는 과정

β_0도 동일한 방법으로 검정하면 됨

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

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

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

유의성 검정

유의성 검정 Significance Test

전체 회귀식이 아닌 **개별 모수**의 추정량이 통계적으로 유의한지를 알아보는 과정

β_0도 동일한 방법으로 검정하면 됨

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

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

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

유의성 검정

유의성 검정 Significance Test

전체 회귀식이 아닌 **개별 모수**의 추정량이 통계적으로 유의한지를 알아보는 과정

β_0도 동일한 방법으로 검정하면 됨

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

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

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

귀무가설을 기각하지 못해도,

X와 Y사이에 선형적 관계가 없을 뿐,

아무 의미가 없다는 게 아님!

3

다중선형회귀

다중선형회귀

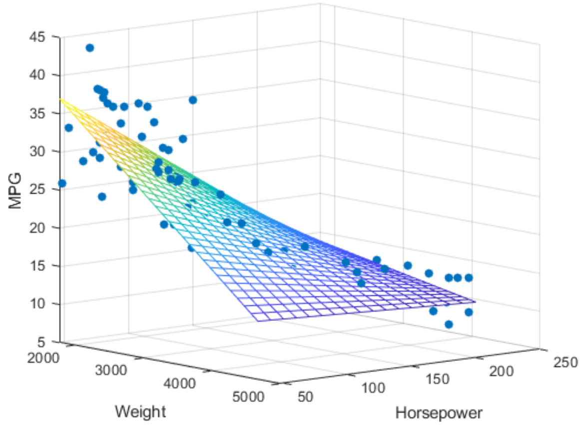
다중선형회귀 Multiple Linear Regression

2개 이상의 독립변수를 가짐

단순회귀분석에 비해 **복잡한 관계 설명에 용이**

설명변수가 p개로 확장

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,\ \epsilon \sim NID(0, \sigma^2)$$



다중선형회귀

다중선형회귀 Multiple Linear Regression

2개 이상의 독립변수를 가짐

단순회귀분석에 비해 **복잡한 관계 설명에 용이**

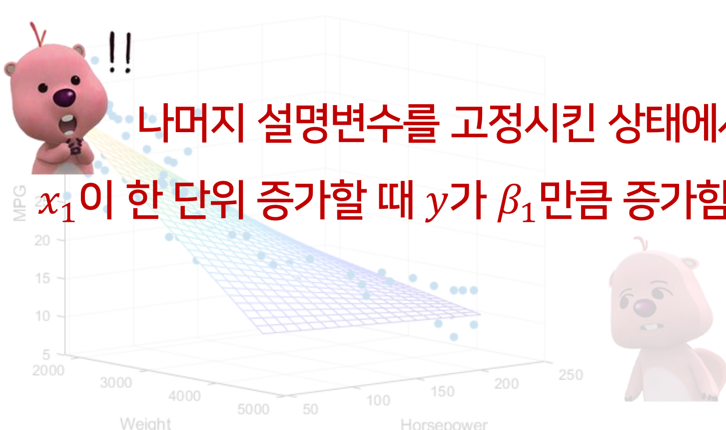
설명변수가 p개로 확장

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \;\; \epsilon \sim NID(0, \sigma^2)$$

!!

나머지 설명변수를 고정시킨 상태에서

x_1이 한 단위 증가할 때 y가 β_1만큼 증가함을 의미



최소제곱법

최소제곱법 Least Square Estimator Method

다중선형회귀에서는 **행렬**을 이용하여 계산

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

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

$$\Updownarrow$$

"$Y = X\beta + \epsilon$"

최소제곱법

최소제곱법 Least Square Estimator Method

다중선형회귀에서는 **행렬**을 이용하여 계산

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

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

$$\Updownarrow$$

$$\text{“}\boldsymbol{Y = X\beta + \epsilon}\text{”}$$

최소제곱법

목적함수

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

목적함수 S를 β에 대해 미분, 미분값 $= 0$

추정량

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

추정된 회귀식

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

최소제곱법

목적함수

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

추정량

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

추정된 회귀식

추정된 $\hat{\beta}$를 활용해 회귀식 추정

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

최소제곱법

목적함수

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

추정량

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

추정된 회귀식

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

이때, $H = X(X'X)^{-1}X'$ 는 투영행렬

적합성 검정

적합성 검정 Goodness of Fit

회귀직선이 데이터에 **얼마나 잘 들어맞는지** 모형에 대한 적합성 검정

결정계수

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

변수가 늘어나면

자연스럽게 **값이 증가**

vs

수정결정계수

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

변수가 추가됨에 따라

결정계수에 **패널티 부과**

적합성 검정

적합성 검정 Goodness of Fit

회귀직선이 데이터에 **얼마나 잘 들어맞는지** 모형에 대한 적합성 검정

결정계수

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

변수가 늘어나면

자연스럽게 **값이 증가**

vs

수정결정계수

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

변수가 추가됨에 따라

결정계수에 **패널티 부과**

적합성 검정

적합성 검정 Goodness of Fit

회귀직선이 데이터에 **얼마나 잘 들어맞는지** 모형에 대한 적합성 검정

결정계수

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

변수가 늘어나면

자연스럽게 **값이 증가**

vs

수정결정계수

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

변수가 추가됨에 따라

결정계수에 **패널티 부과**

R^2_{adj}가 더 높은 회귀식이 더 좋은 회귀식

유의성 검정

유의성 검정 Significance Test

추정량이 **통계적으로 유의**한지를 알아보는 과정

F-test

Partial F-test

T-test

유의성 검정

유의성 검정 Significance Test

추정량이 **통계적으로 유의**한지를 알아보는 과정

F-test

Partial F-test

T-test

유의성 검정

유의성 검정 Significance Test

추정량이 **통계적으로 유의**한지를 알아보는 과정

F-test

Partial F-test

T-test

유의성 검정

유의성 검정 Significance Test

추정량이 **통계적으로 유의**한지를 알아보는 과정

F-test

Partial F-test

T-test

유의성 검정

F-test

전체 회귀계수에 관한 검정

가설 설정

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

$\boldsymbol{H_1}: not\ \boldsymbol{H_0}$ ($\beta_0,\ \beta_1, \cdots, \beta_p$ 중 적어도 하나는 0이 아니다.)

유의성 검정

F-test

전체 회귀계수에 관한 검정

검정통계량

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

검정 통계량은 회귀계수가 얼마나 **설명력**을 갖는지를 의미

유의성 검정

F-test

전체 회귀계수에 관한 검정

임계값

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

검정통계량이 **임계값보다 크다**면 귀무가설을 **기각**

검정통계량이 크고, p-value 작은 경우

유의성 검정

F-test

전체 회귀계수에 관한 검정

임계값

$$F_0 \geq F_{\left(1-\frac{\alpha}{2},\, p,\, n-p-1\right)}$$

검정통계량이 **임계값보다 크다**면 귀무가설을 **기각**

검정통계량이 크고, p-value 작은 경우

F-test의 귀무가설이 기각되지 않을 경우,

$\boldsymbol{\beta_0 = \beta_1 = \cdots = \beta_p = 0}$ 이므로

회귀계수 값이 통계적으로 유의하지 않음을 의미

"모델을 다시 세우는 등의 조치가 필요"

유의성 검정

Partial F-test

일부 회귀계수에 관한 검정

가설 설정

Full model (FM) = 모든 변수를 사용한 회귀모형

Reduced Model (RM) = 일부 계수를 특정 값으로 둔 축소모형

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

$\boldsymbol{H_1}$: not $\boldsymbol{H_0}$ ($\beta_J, \beta_{J+1}, \cdots, \beta_{j+q-1}$ 중 적어도 하나는 0이 아니다)

특정 상수값은 0으로 두는 경우가 가장 일반적

유의성 검정

Partial F-test

일부 회귀계수에 관한 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 제거했으므로 일반적으로 SSE(RM) > SSE(FM)이 성립

유의성 검정

Partial F-test

일부 회귀계수에 관한 검정

q개의 변수를 제거했을 때 모델이 설명하지 못하는 변동

모든 변수를 포함했을 때 모델이 설명하지 못하는 변동

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 제거했으므로 일반적으로 SSE(RM) > SSE(FM)이 성립

유의성 검정

Partial F-test

일부 회귀계수에 관한 검정

q개의 변수를 제거했을 때 모델이 설명하지 못하는 변동

모든 변수를 포함했을 때 모델이 설명하지 못하는 변동

검정통계량

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

이때, 귀무가설을 기각시키기 위해서는

***SSE(RM)*이 *SSE(FM)*보다 훨씬 커야함**

변수를 제거했으므로 일반적으로 SSE(RM) > SSE(FM)이 성립

유의성 검정

Partial F-test

일부 회귀계수에 관한 검정

q개의 변수를 제거했을 때 모델이 설명하지 못하는 변동

모든 변수를 포함했을 때 모델이 설명하지 못하는 변동

잔차제곱합

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

이때, 귀무가설을 기각시키기 위해서는

$SSE(RM)$이 $SSE(FM)$보다 훨씬 커야함

변수를 제거했으므로 일반적으로 SSE(RM) > SSE(FM)이 성립

제거된 변수가 의미있다면 $SSE(RM)$가 매우 커지므로,

검정통계량이 귀무가설을 기각시킬만큼

충분히 커진다는 의미!

유의성 검정

Partial F-test

일부 회귀계수에 관한 검정

임계값

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

F-test와 마찬가지로 검정통계량이 **임계값보다 크다**면 귀무가설을 **기각**

제거된 변수가 모델에 유의미하다면 검정통계량 $\boldsymbol{F_0}$가 커지고 p-value는 작아지므로

유의성 검정

Partial F-test

일부 회귀계수에 관한 검정

임계값

$$F_0 \ge F_{\left(1-\frac{\alpha}{2},\, q,\, n-p-1\right)}$$

F-test와 마찬가지로 검정통계량이 **임계값보다 크다**면 귀무가설을 **기각**

제거된 변수가 모델에 유의미하다면 검정통계량 F_0가 커지고 p-value는 작아지므로

회귀식 전체에 대한 F-test는 Partial F-test의 한 종류이므로

Partial F-test가 더 일반적인 검정이지만

보편적으로 F-test를 더 많이 사용

유의성 검정

T-test

개별 회귀계수에 대한 검정

회귀계수 추가의 유의성을 판단하기 위해 사용

가설 설정

$\mathbf{H_0}$: $\beta_j = 0$ 다른 변수들이 다 적합된 상태에서 설명변수 x_j는 유의하지 않음

$\mathbf{H_1}$: $\beta_j \neq 0$ 다른 변수들이 다 적합된 상태에서 설명변수 x_j는 유의함

유의성 검정

T-test

개별 회귀계수에 대한 검정

회귀계수 추가의 유의성을 판단하기 위해 사용

검정통계량

$$\boldsymbol{t_j} = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$$

T-test는 x_j 변수 자체가 아니라 **변수 추가의 유의성을 확인**

유의성 검정

T-test

개별 회귀계수에 대한 검정

회귀계수 추가의 유의성을 판단하기 위해 사용

임계값

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

F-test와 마찬가지로 검정통계량이 **임계값보다 크다**면 귀무가설을 **기각**

귀무가설을 기각시킨다면 변수의 추가는 회귀식의 설명력 증가에 기여

유의성 검정

T-test

개별 회귀... 한 검정

회귀계수 추가의 유... 하기 위해 사용

임계값

T-test를 활용해 변수를 선택하는 것에 주의!

$$|t_0| > t_{\left(\frac{\alpha}{2}, n-p-1\right)}$$

다른 회귀식을 가정하면

해당 변수의 유의성은 바뀔 수 있음

F-test와 마찬가지로 검정통계량이 **임계값보다 크다**면 귀무가설을 **기각**

변수선택법에 관한 내용은 3주차에서 만나요!

귀무가설을 기각시킨다면 변수의 추가는 회귀식의 설명력 증가에 기여

유의성 검정

T-test

개별 회귀계수에 대한 검정

회귀계수 추가의 유의성을 판단하기 위해 사용

① F-test가 전체 회귀식에 대한 가정이 더욱 엄격함

② F에서는 기각하지 못해도 T에서 기각하는 경우가 발생 가능

임계값

$$|t_j| \geq t_{\left(\frac{\alpha}{2},\, n-p-1\right)}$$

따라서 F-test를 먼저 시행해

전체 모델이 통계적으로 유의한지 확인하는 것이 중요

F-test와 마찬가지로 검정통계량이 **임계값보다 크다면** 귀무가설을 **기각**

귀무가설을 기각시킨다면 변수의 추가는 회귀식의 설명력 증가에 기여

4

데이터 진단

데이터 진단의 필요성

데이터 진단, 왜 필요해?

이상치, 지렛값, 영향점 등

일반적인 경향에서 벗어나는 데이터 존재

회귀 모형에 큰 영향을 미침

데이터가 일반적인 경향에서 벗어나는지 **1) 판단 2) 처리**

잔차를 이용해 데이터 진단을 할 수 있을까?

4 데이터 진단

데이터 진단의 필요성

데이터 진단, 왜 필요해?

이상치, 지렛값, 영향점 등

일반적인 경향에서 벗어나는 데이터 존재

회귀 모형에 큰 영향을 미침

데이터가 일반적인 경향에서 벗어나는지 **1) 판단 2) 처리**

잔차를 이용해 데이터 진단을 할 수 있을까?

데이터 진단의 필요성

데이터 진단, 왜 필요해?

이상치, 지렛값, 영향점 등

일반적인 경향에서 벗어나는 데이터 존재

회귀 모형에 큰 영향을 미침

데이터가 일반적인 경향에서 벗어나는지 1) **판단** 2) **처리**

잔차를 이용해 데이터 진단을 할 수 있을까?

4 데이터 진단

데이터 진단의 필요성

데이터 진단, 왜 필요해?

이상치, 지렛값, 영향점 등

일반적인 경향에서 벗어나는 데이터 존재

회귀 모형에 큰 영향을 미침

데이터가 일반적인 경향에서 벗어나는지 **1) 판단 2) 처리**

잔차를 이용해 데이터 진단을 할 수 있을까?

데이터 진단의 필요성

데이터 진단, 왜 필요해?

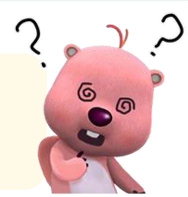
이상치, 지렛값, 영향점 등

일반적인 경향에서 벗어나는 데이터 존재

회귀 모형에 큰 영향을 미침

데이터가 일반적인 경향에서 벗어나는지 **1) 판단 2) 처리**

잔차를 이용해 데이터 진단을 할 수 있을까?



데이터 진단의 필요성

데이터 진단, 왜 필요해?

치, 지렛값, 영향점 등

일반적인 경향에서 벗어나는 데이터 존재

회귀 모형에 큰 영향을 미침

1) 판단 2) 처리

잔차를 이용해 데이터 진단을 할 수 있을까?

NO!

잔차는 y**값의 단위**에 영향을 많이 받기 때문

좀 더 **일반화된 상황**에서 적용하도록 **표준화** 필요!

스튜던트화 잔차

스튜던트화 잔차 Studentized Residual

Y값의 단위에 영향을 많이 받는 일반 잔차를 보완한 지표

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

σ는 모수이므로 알 수 없기 때문에 추정량 사용

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

관측값이 일반적인 경향에서 벗어나는지 **판단하는 기준!**

스튜던트화 잔차

스튜던트화 잔차 Studentized Residual

Y값의 단위에 영향을 많이 받는 일반 잔차를 보완한 지표

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

σ는 모수이므로 알 수 없기 때문에 추정량 사용

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

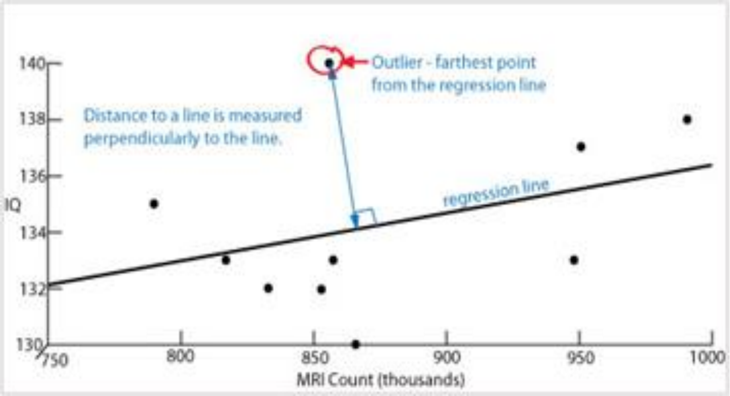
관측값이 일반적인 경향에서 벗어나는지 **판단하는 기준!**

이상치

이상치 outlier

스튜던트화 잔차가 매우 큰 값

표준화했을 때 y의 기준에서 **절대값이 큰 값**



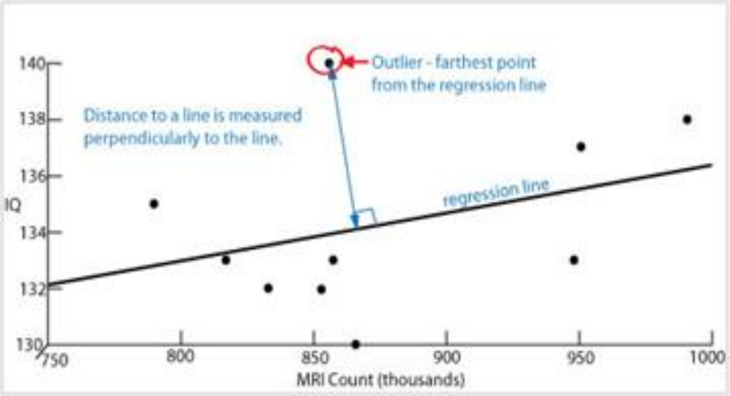
보통 $|r_i| > 3$ 이면 이상치라고 판단!

이상치

이상치 outlier

스튜던트화 잔차가 매우 큰 값

표준화했을 때 y의 기준에서 **절대값이 큰 값**



보통 $|r_i| > 3$ 이면 이상치라고 판단!

지렛값

지렛값 Leverage Point

x의 평균 $\bar{x}$ 에서 멀리 떨어져 있어 **기울기에 큰 영향**을 주는 값

표준화했을 때 **x의 기준**에서 **절댓값이 큰 값**

앞에서 살펴본 투영행렬 H 의 대각 원소 기호 h_{ii} 와 혼동 주의!

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

x_i와 $\bar{x}$ 의 차이가 클수록 h_{ii} 가 커짐 → x 의 평균에서 멀수록 지렛값 상승

$h_{ii} > \frac{2(p+1)}{n}$ 이면 지렛값으로 판단!

지렛값

지렛값 Leverage Point

x의 평균 $\bar{x}$ 에서 멀리 떨어져 있어 **기울기에 큰 영향**을 주는 값

표준화했을 때 **x의 기준**에서 **절댓값이 큰 값**

앞에서 살펴본 투영행렬 H 의 대각 원소 기호 h_{ii} 와 혼동 주의!

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

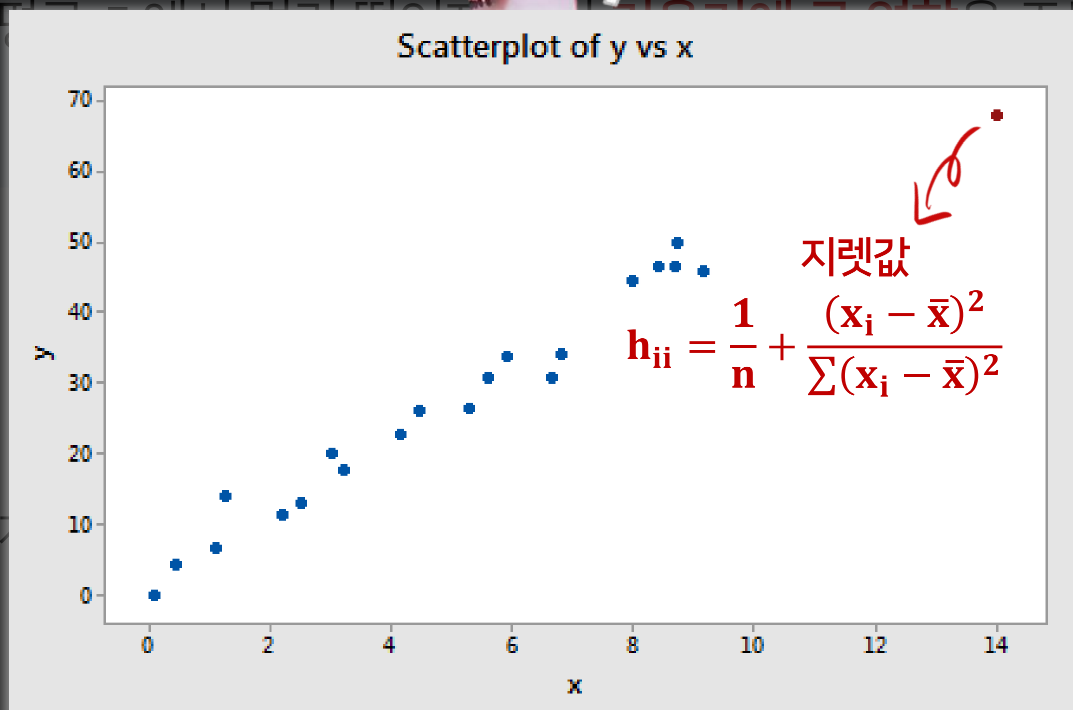
x_i와 $\bar{x}$ 의 차이가 클수록 h_{ii} 가 커짐 → x 의 평균에서 멀수록 지렛값 상승

$\boldsymbol{h_{ii} > \frac{2(p+1)}{n}}$ 이면 지렛값으로 판단!

지렛값

지렛값 Leverage Point

x의 ... 는 값



h_{ii} 와 혼동 주의!

x_i와 $\bar{x}$ 의 ... 렛값 상승

지렛값

지렛값 Leverage Point

x의 평균 $\bar{x}$ 에서 멀리 떨어져 있어 **기울기에 큰 영향**을 주는 값

표준화했을 때 x의 기준에서 **절댓값이 큰 값**

But,

이상치나 지렛값이라고 해서

회귀직선을 변화시킨다고 단정 지을 수 없음

...의 대각 원소 기호 h_{ii} 와 혼동 주의!

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

✓ **X의 평균 주변에 위치**한 이상치는 기울기를 변화시키지 못함

✓ 지렛값이라도 **회귀선의 연장선**에 존재할 수 있음

x_i와 $\bar{x}$ 의 차이가 클수록 ... 커지고 → x의 평균에서 멀수록 지렛값 상승

!!

$h_{ii} > \frac{2(p+1)}{n}$ 이면 지렛값으로 판단!

영향점

영향점 Influential Point

회귀직선의 기울기에 상당한 영향을 주는 점

이상치와 지렛값을 동시에 고려하는 지표

Cook's Distance

영향점을 확인하는 표준적인 지표

특정 데이터를 지웠을 때 **회귀선이 변하는 정도**를 나타냄

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

영향점

영향점 Influential Point

회귀직선의 기울기에 상당한 영향을 주는 점

이상치와 지렛값을 동시에 고려하는 지표

Cook's Distance

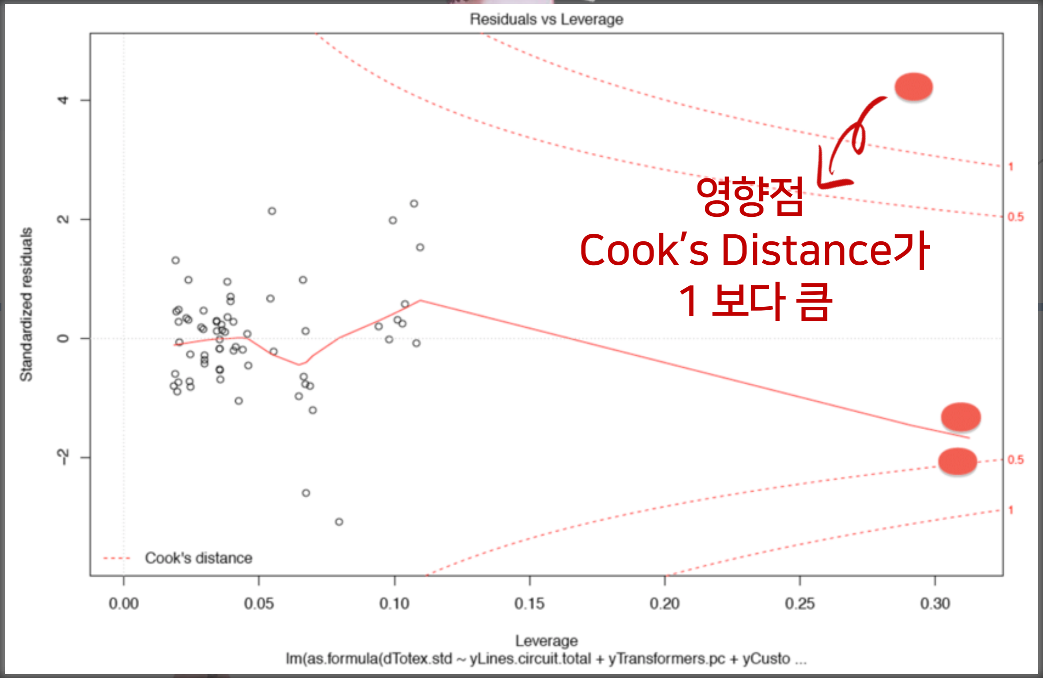
이상치 $$C_i = \frac{r_i^2}{p+1} \times \frac{h_{ii}}{1-h_{ii}}$$ 지렛값

$C_i > 1$ 이면 영향점으로 판단 !

영향점

영향점 Influential Point

시에 고려하는 지표



영향점 처리의 필요성

영향점은 추정량의 **분산을 크게** 만듦

잘못된 모델의 해석과 **예측 성능 저하**

영향점 처리를 통해 **이상치에 강건한(robust) 모델링**

영향점 처리의 필요성

영향점은 추정량의 **분산을 크게** 만듦

잘못된 모델의 해석과 **예측 성능 저하**

영향점 처리를 통해 **이상치에 강건한(robust) 모델링**

영향점 처리의 필요성

영향점은 추정량의 **분산을 크게** 만듦

잘못된 모델의 해석과 **예측 성능 저하**

영향점 처리를 통해 **이상치에 강건한(robust) 모델링**

5

로버스트 회귀

로버스트 회귀

로버스트 회귀 모형 Robust Regression

이상치의 영향을 크게 받지 않는 회귀모형

Median Regression

Huber's M-estimation

Least Trimmed Square

로버스트 회귀

로버스트 회귀 모형 Robust Regression

이상치의 영향을 크게 받지 않는 회귀모형

Median Regression

Huber's M-estimation

Least Trimmed Square

로버스트 회귀

로버스트 회귀 모형 Robust Regression

이상치의 영향을 크게 받지 않는 회귀모형

Median Regression

Huber's M-estimation

Least Trimmed Square

로버스트 회귀

로버스트 회귀 모형 Robust Regression

이상치의 영향을 크게 받지 않는 회귀모형

Median Regression

Huber's M-estimation

Least Trimmed Square

Median Regression

Median Regression

평균보다 **중앙값이 이상치에 덜 민감**하다는 아이디어 착안

독립변수 X의 변화에 따른 종속변수 Y의 **조건부 중앙값**을 추정하는 방법

최소제곱회귀

오차의 제곱합 최소화

$$(\,argmin \sum \varepsilon_i^2\,)$$

이상치에 너무 큰 가중치

vs

Median Regression

오차의 절대값의 합 최소화

$$(\,argmin \sum |\varepsilon_i|\,)$$

항상 동일한 가중치

Median Regression

Median Regression

평균보다 **중앙값이 이상치에 덜 민감**하다는 아이디어 착안

독립변수 X의 변화에 따른 종속변수 Y의 **조건부 중앙값**을 추정하는 방법

최소제곱회귀

오차의 제곱합 최소화

$$(\,argmin \sum \varepsilon_i^2)$$

이상치에 너무 큰 가중치

vs

Median Regression

오차의 절대값의 합 최소화

$$(\,argmin \sum |\varepsilon_i|)$$

항상 동일한 가중치

Median Regression

Median Regression

평균보다 **중앙값이 이상치에 덜 민감**하다는 아이디어 착안

독립변수 X의 변화에 따른 종속변수 Y의 **조건부 중앙값**을 추정하는 방법

최소제곱회귀

오차의 제곱합 최소화

$$(argmin \sum \varepsilon_i^2)$$

이상치에 너무 큰 가중치

vs

Median Regression

오차의 절대값의 합 최소화

$$(argmin \sum |\varepsilon_i|)$$

항상 동일한 가중치

Median Regression

Median Regression

평균보다 **중앙값이 이상치에 덜 민감**하다는 아이디어 착안

독립변수 X의 변화에 따른 종속변수 Y의 **조건부 중앙값**을 추정하는 방법

최소제곱회귀

오차의 제곱합 최소화

$$(\, argmin \sum \varepsilon_i^2)$$

이상치에 너무 큰 가중치

vs

Median Regression

오차의 절대값의 합 최소화

$$(\, argmin \sum |\varepsilon_i|)$$

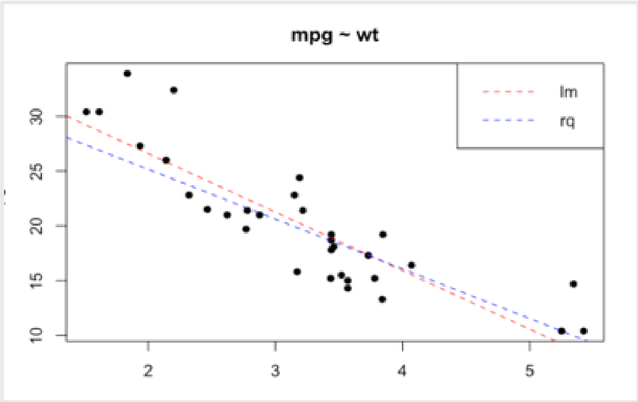
항상 동일한 가중치

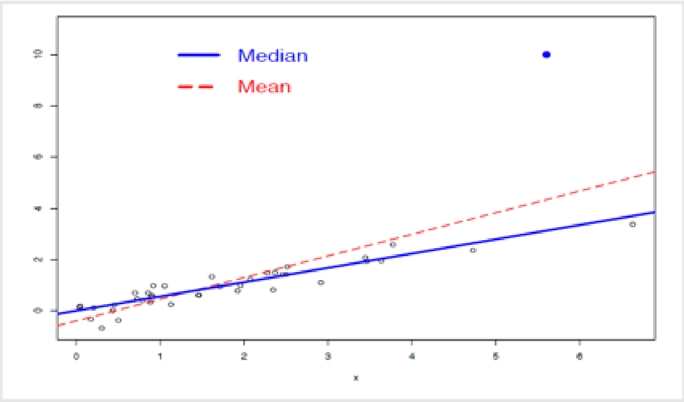
Median Regression

Median Regression

평균보다 **중앙값이 이상치에 덜 민감**하다는 아이디어 착안

독립변수 X의 변화에 따른 종속변수 Y의 **조건부 중앙값**을 추정하는 방법





분포 가정과 등분산 가정이 없는 모델

R에서 **quantreg 패키지**의 rq() 함수 사용

Huber's M-estimation

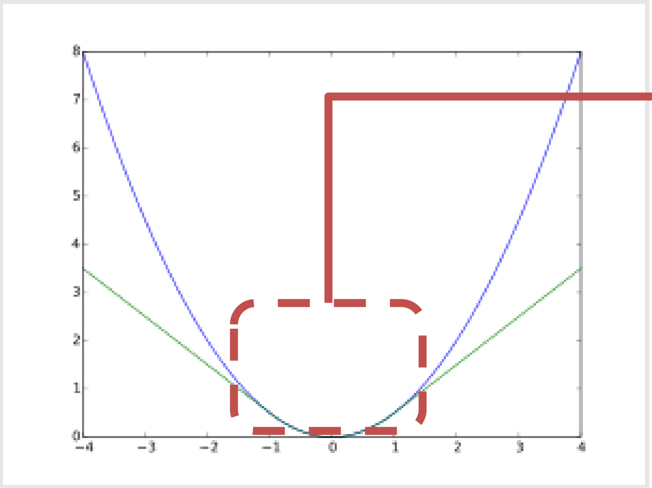
Huber's M-estimation

이상치에 대한 **지나친 패널티 부여를 없애는 방법**

잔차가 **특정 상수값보다 크면**, 잔차의 '제곱'이 아닌 **1차식**으로 바꾸어

이상치에 강건한 회귀계수를 추정하는 방법

c는 특정 상수값



$$\rho(e) = \frac{1}{2}e^2 \quad if\ |e| \le c,$$

$$\rho(e) = c|e| - \frac{1}{2}c^2 \quad otherwise$$

R에서 MASS 패키지의 rlm() 함수 사용

Huber's M-estimation

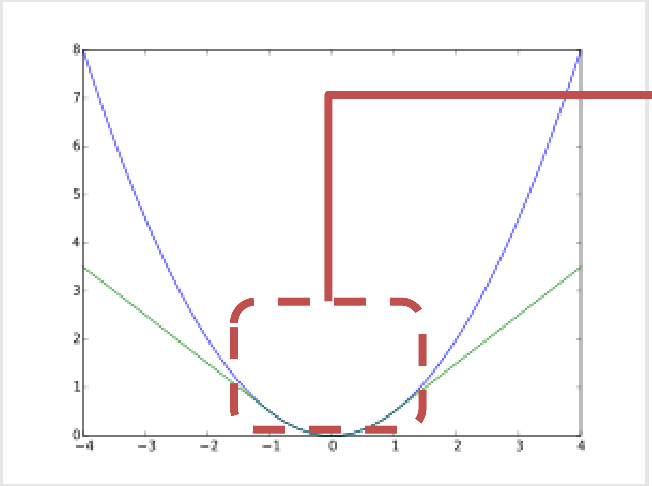
Huber's M-estimation

이상치에 대한 **지나친 패널티 부여를 없애는 방법**

잔차가 **특정 상수값보다 크면**, 잔차의 '제곱'이 아닌 **1차식**으로 바꾸어

이상치에 강건한 회귀계수를 추정하는 방법

c는 특정 상수값



$$\rho(e) = \frac{1}{2}e^2 \quad if\ |e| \leq c,$$

$$\rho(e) = c|e| - \frac{1}{2}c^2 \quad otherwise$$

R에서 MASS 패키지의 rlm() 함수 사용

Least Trimmed Square

Least Trimmed Square

통계적 기준에 따라 **잔차가 너무 큰 관측치를 제거**하고
회귀계수를 추정하는 방법

$r_{(j)}$ 는 오름차순으로 나열한 잔차

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

n개의 obs. 중 h개만 사용하여 회귀식을 만드는데,

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

obs가 별로 없는 경우나 영향점이 존재하지 않는 경우 주의해서 사용

Least Trimmed Square

Least Trimmed Square

통계적 기준에 따라 **잔차가 너무 큰 관측치를 제거**하고

회귀계수를 추정하는 방법

$r_{(j)}$ 는 오름차순으로 나열한 잔차

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

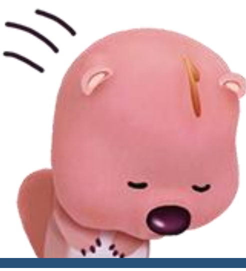
n개의 obs. 중 h개만 사용하여 회귀식을 만드는데,

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

obs가 별로 없는 경우나 영향점이 존재하지 않는 경우 주의해서 사용

다음주 예고

1. 회귀분석의 4가지 기본 가정
2. 잔차 Plot
3. 선형성 가정
4. 정규성 진단과 처방
5. 등분산성 진단과 처방
6. 독립성 진단과 처방



감사합니다