회귀분석팀

6팀

김보근

김민주

서유진

하희나

INDEX

1

회귀분석이란?

1 회귀분석이란?

회귀분석

- 독립변수와 종속변수 간의 관계를 설명하고 모델링하는 통계적 기법
- 변수들 간의 상관관계 파악

→ 특정 변수의 값을 다른 변수들을 이용하여 설명, 예측하는 방법

EXAMPLE) 범죄율 (독립변수)과 주택 가격 (종속변수)간의 관계

회귀분석의 목적

1) 변수들 간의 관계 표현

2) 독립변수에 따른 종속변수의 변화 파악

3) 미래 관측값 예측

1 회귀분석이란?

회귀식

독립변수 X와 종속변수 Y의 관계를 표현한 함수식

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

Y (종속변수) : 독립변수에 의해서 설명되는 변수

X_k (독립변수) : 종속변수를 설명하기 위한 변수

ε (오차항) : 변수를 측정할 때 발생할 수 있는 오차, 무작위성을 지님

1 회귀분석이란?

회귀분석 vs 상관분석

회귀분석

두 변수의 관계에 분명한 방향이 있을 때 사용 ex. 보석의 가격과 크기의 관계

상관분석

두 변수의 역할이 서로 대등할 때 사용 ex. 키와 몸무게의 관계

하지만, 두 변수의 관계, 선형적 상관성 정도만 표현 가능

→ 구체적인 예측과 설명 불가능

1 회귀분석이란?

회귀분석 vs 상관분석

회귀분석

두 변수의 관계에 분명한 방향이 있을 때 사용 ex. 보석의 가격과 크기의 관계

회귀분석을 사용하는 이유

독립변수가 한 단계 변화할 때마다

종속변수가 어떻게 변화할지를 안다면

더 유의미한 관계 파악 가능!

1 회귀분석이란?

회귀모델링 과정

1. 문제 정의

보근이의 학점을 가장 잘 표현할 수 있는 변수들은 무엇이 있을까?

2. 적절한 변수 선택

X_1, X_2, X_3 : 공부 시간, 통학 거리, 아침밥 식사 여부

3. 데이터 수집 및 전처리

보근이의 학점, 공부 시간, 집에서 학교까지의 거리, 아침식사 여부

1 회귀분석이란?

회귀모델링 과정

1. 문제 정의

보근이의 학점을 가장 잘 표현할 수 있는 변수들은 무엇이 있을까?

2. 적절한 변수 선택

X_1, X_2, X_3 : 공부 시간, 통학 거리, 아침밥 식사 여부

3. 데이터 수집 및 전처리

보근이의 학점, 공부 시간, 집에서 학교까지의 거리, 아침식사 여부

1 회귀분석이란?

회귀모델링 과정

1. 문제 정의

보근이의 학점을 가장 잘 표현할 수 있는 변수들은 무엇이 있을까?

2. 적절한 변수 선택

X_1, X_2, X_3 : 공부 시간, 통학 거리, 아침밥 식사 여부

3. 데이터 수집 및 전처리

보근이의 학점, 공부 시간, 집에서 학교까지의 거리, 아침식사 여부

1 회귀분석이란?

회귀모델링 과정

4. 모델 선정과 적합

적절한 회귀분석 모델 선택 및 적합

Ex. 선형/비선형, 단순/다중회귀, 모수/비모수, 일변량/다변량 등

5. 적합성 및 유의성 검정

✓ 회귀모델이 얼마나 데이터를 잘 설명하는가?

✓ 회귀계수는 통계적으로 유의한가?

6. 모형 평가

설정한 모델이 회귀 가정을 만족하는가?

→ 만족하지 않으면 처방! [회귀팀 2주차 클린업 예정]

1 회귀분석이란?

회귀모델링 과정

4. 모델 선정과 적합

적절한 회귀분석 모델 선택 및 적합

Ex. 선형/비선형, 단순/다중회귀, 모수/비모수, 일변량/다변량 등

5. 적합성 및 유의성 검정

- ✓ 회귀모델이 얼마나 데이터를 잘 설명하는가?
- ✓ 회귀계수는 통계적으로 유의한가?

6. 모형 평가

설정한 모델이 회귀 가정을 만족하는가?

→ 만족하지 않으면 처방! [회귀팀 2주차 클린업 예정]

1 회귀분석이란?

회귀모델링 과정

4. 모델 선정과 적합

적절한 회귀분석 모델 선택 및 적합

Ex. 선형/비선형, 단순/다중회귀, 모수/비모수, 일변량/다변량 등

5. 적합성 및 유의성 검정

- ✓ 회귀모델이 얼마나 데이터를 잘 설명하는가?
- ✓ 회귀계수는 통계적으로 유의한가?

6. 모형 평가

설정한 모델이 회귀 가정을 만족하는가?

→ 만족하지 않으면 처방! [회귀팀 2주차 클린업 예정]

1 회귀분석이란?

회귀모델링 과정

4. 모델 선정과 적합

적절한 회귀분석 모델 선택 및 적합

Ex. 선형/비선형, 단순/다중회귀, 모수/비모수, 일변량/다변량 등

7. 모형 해석

보근이가 현재보다 주당 2시간 더 공부하고,

자취방에서 통학을 하고, 아침밥을 매일 챙겨먹는다면

학점이 0.3만큼 오를 것이다!

6. 모형 평가

설정한 모델이 회귀 가정을 만족하는가?

→ 만족하지 않으면 처방! [회귀팀 2주차 클린업 예정]

2

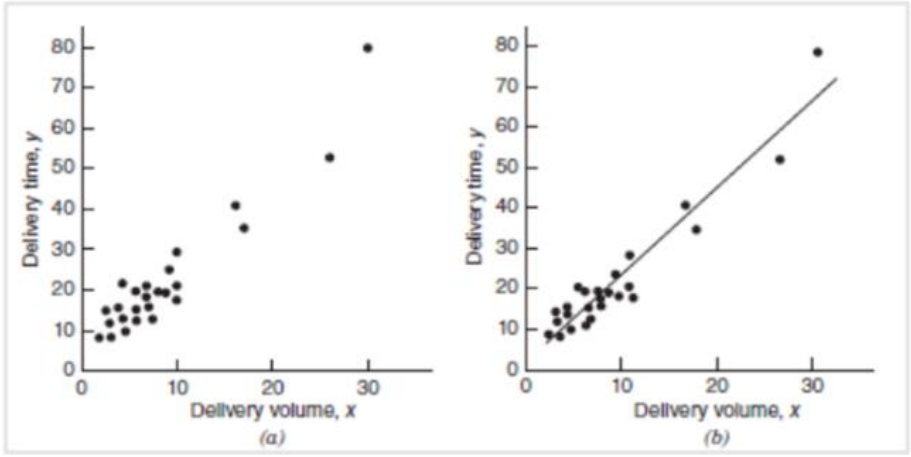
단순선형회귀분석

2 단순선형회귀분석

단순선형회귀 (Simple Linear Regression)

독립변수(X)와 종속변수(Y)의 관계를 가장 잘 설명하는 **직선**을 찾아 수식화!

'단순' 선형회귀이므로, 각 변수의 개수는 하나!



☑ '변수의 관계는 선형적이다'라는 가정 하에 직선 함수식을 가정

[회귀팀 2주차 클린업 예정]

2 단순선형회귀분석

단순선형회귀모델

$\varepsilon_i \sim N(0, \sigma^2)$라는 가정 하에서,

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

y_i : 종속변수 y의 i번째 관측값

x_i : 독립변수 x의 i번째 관측값

β_0, β_1 : 회귀계수 [추정해야 하는 모수!] [회귀식의 기울기와 절편]

ε_i : i번째 관측값에 의한 랜덤한 오차

2 단순선형회귀분석

단순선형회귀모델

단순선형회귀 모델의 **해석**

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

"x가 한 단위 증가할 때, y는 β_1**만큼 증가**한다."

$\Leftrightarrow$ Y의 평균값의 변화량이 β_1

y_i : 종속변수 y의 i번째 관측값

x_i : 독립변수 x의 i번째 관측값

β_0, β_1 : 회귀계수 [추정해야 하는 모수!] [회귀식의 기울기와 절편]

ε_i : i번째 관측값에 의한 랜덤한 오차

2 단순선형회귀분석

왜 직선인가?

선형 근사

- 변수의 영향력을 간단히 모형화 가능
- X의 변화에 따른 Y의 변화를 직관적으로 확인 가능

고차 근사를 하게 된다면?

모델의 복잡도가 높아져서 과적합 (overfitting)의 원인이 됨

2 단순선형회귀분석

왜 직선인가?

선형 근사

- 변수의 영향력을 간단히 모형화 가능
- X의 변화에 따른 Y의 변화를 직관적으로 확인 가능

고차 근사를 하게 된다면?

모델의 복잡도가 높아져서 **과적합 (overfitting)**의 원인이 됨

2 단순선형회귀분석

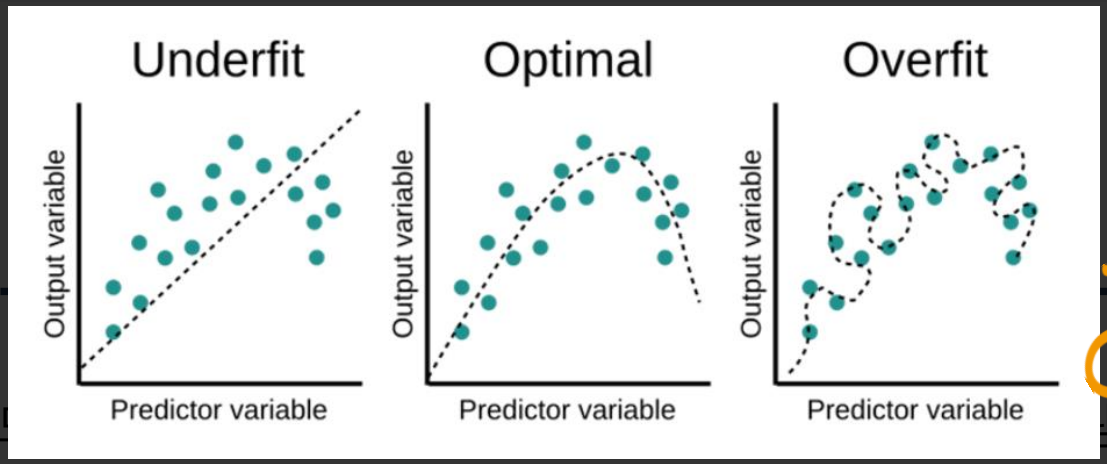
왜 직선인가?

선형 근사

과적합 (overfitting) 이란?

Training data에 대한 설명성은 높아도

Test data에 대한 설명성은 떨어지는 문제



모델의 분산 ⬆, 검증 데이터의 성능 ⬇

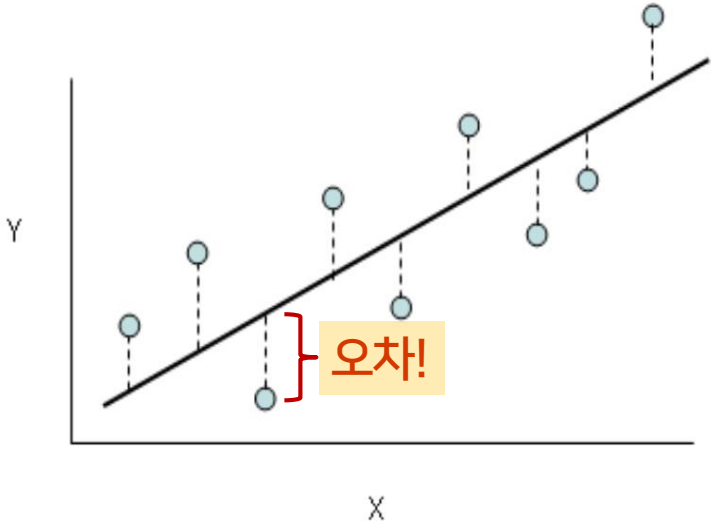
[데이터마이닝팀 1주차 클린업 참고]

2 단순선형회귀분석

모수의 추정 : 최소제곱법 (LSE : Least Square Estimation)

좋은 추정이란?

회귀직선과 관측치 사이의 오차가 작을수록 좋은 추정!



최소제곱법

오차의 제곱합을 최소화하는

모수를 추정하는 방법

2 단순선형회귀분석

모수의 추정 : 최소제곱법 (LSE : Least Square Estimation)

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

LSE의 목적!!

오차제곱합 $J(\beta_0, \beta_1)$을 최소화시키는 β_0, β_1를 찾자!!

아래로 볼록한 convex함수 → 각각의 모수에 대해 편미분

'미분값 = 0'을 만족시키는 β_0, β_1 찾기!

2 단순선형회귀분석

모수의 추정 : 최소제곱법 (LSE : Least Square Estimation)

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

LSE의 목적!!

오차제곱합 $J(\beta_0, \beta_1)$을 최소화시키는 β_0, β_1를 찾자!!

아래로 볼록한 convex함수 → 각각의 모수에 대해 편미분

'미분값 = 0'을 만족시키는 β_0, β_1 찾기!

2 단순선형회귀분석

모수의 추정 : 최소제곱법 (LSE : Least Square Estimation)

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

정규방정식

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

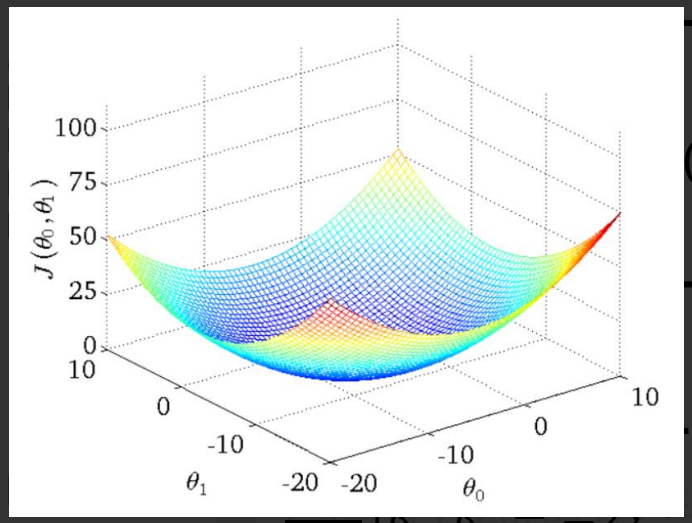
$$\frac{\partial J}{\partial \beta_1}\Big|_{\widehat{\beta_0},\widehat{\beta_1}} = -2\Sigma_{i=1}^{n}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) x_i = 0$$

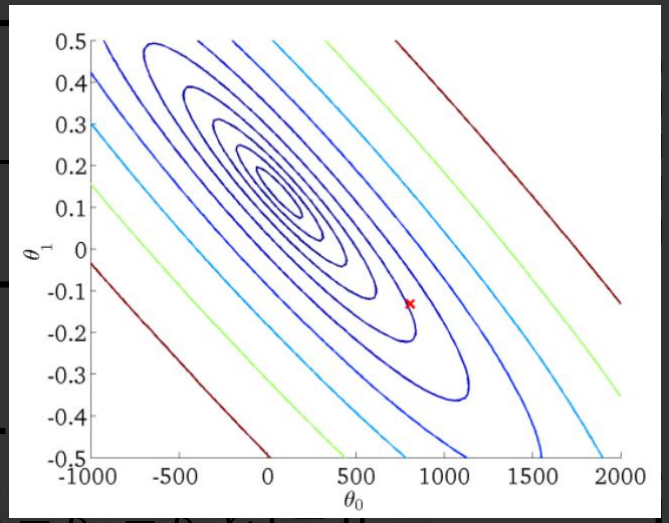
정규방정식을 β_0, β_1에 대해서 연립 → 회귀계수의 추정치

2 단순선형회귀분석

모수의 추정 : 최소제곱법 (LSE : Least Square Estimation)

오차제곱합 $J(\beta_0, \beta_1)$를 기하학적으로 나타낸 그래프





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

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

중심으로 갈수록 오차제곱합의 값이 작아짐!

= 우리가 구하고자 하는 **최적의 β_0, β_1**

정규방정식을 β_0, β_1에 대해서 연립 → 회귀계수의 추정치

2 단순선형회귀분석

모수의 추정 : 최소제곱법 (LSE : Least Square Estimation)

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

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

최소제곱법으로 추정한 $\widehat{\beta_0}$, $\widehat{\beta_1}$

최소제곱 추정치

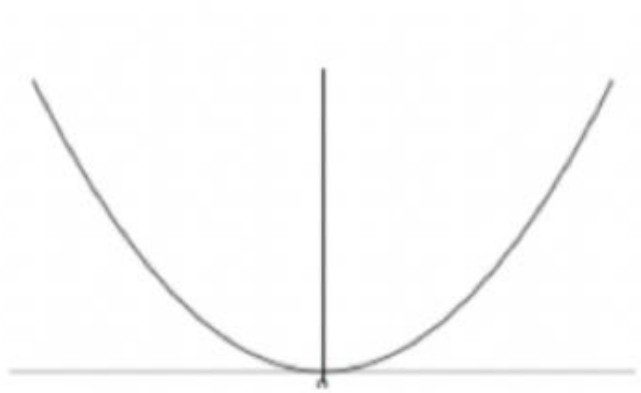
(LSE : Least Square Estimator)

왜 '오차의 제곱합'을 최소화하는가?

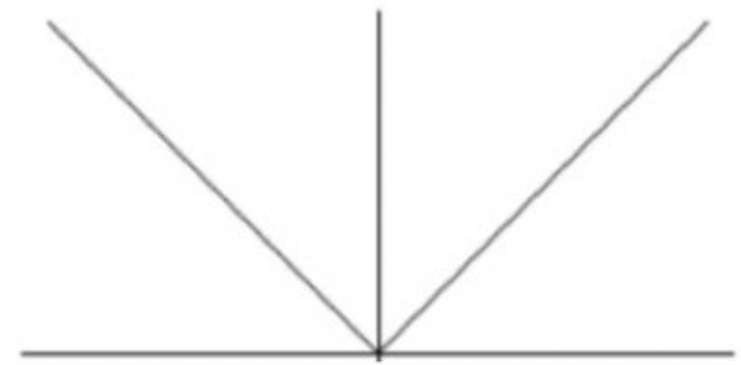
1) 미분하기에 편리 ('오차의 절댓값' 사용 시 미분 불가능)

2) 오차가 클수록 더 큰 페널티 부여 가능

3) 특정 조건 하에서 BLUE가 됨



오차제곱합



오차의 절대값

BLUE (Best Linear Unbiased Estimator)

분산이 제일 작은 선형 불편추정량

분산이 작음 = 추정량이 안정적!

LSE가 BLUE가 될 3가지 조건

① 오차들의 평균은 0

② 오차들의 분산은 σ^2로 동일

③ 오차 간 자기 상관 X (uncorrelated)

최대가능도추정 (MLE : Maximum Likelihood Estimation)

확률적인 방법에 근거해서, 원하는 데이터가 나올
'가능도'를 최대로 하는 모수를 선택하는 방법

오차의 정규분포 가정이 있다면,
LSE와 MLE는 완전히 동일한 추정량 산출

적합성 (Goodness of fit) 검정

모형의 적합성에 대한 평가, 잔차를 이용해 검정 가능

잔차

- 추정한 회귀계수를 이용해 회귀직선을 만들었을 때, 오차의 추정량
- 실제 오차는 알 수 없기에, 추정된 직선을 통해 오차를 추정해야 함

$$e_i = y_i - \hat{y}_i = y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_i\right), \qquad \sum e_i = 0$$

적합성 (Goodness of fit) 검정

변동 분할

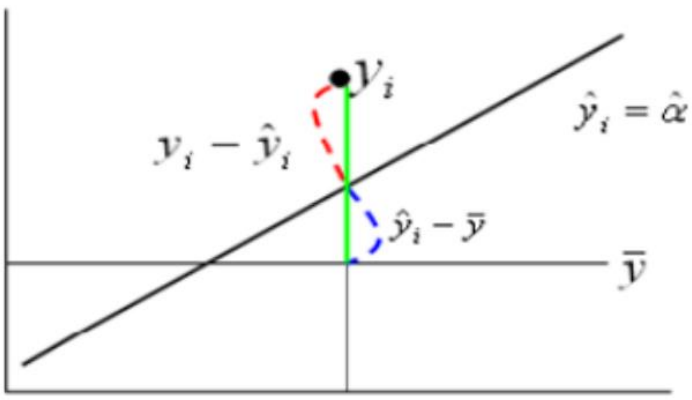
- SST (Total Sum of Squares, 총 변동) : $\sum(y_i - \bar{y})^2$
- SSR (Regression Sum of Square, 회귀선이 설명하는 변동) : $\sum(\hat{y}_i - \bar{y})^2$
- SSE (Residual Sum of Square, 회귀선이 설명하지 못하는 변동) : $\sum(y_i - \hat{y}_i)^2$

적합성 (Good

변동 분할

- SST (Total S
- SSR (Regres
- SSE (Residual Sum of Square, 회귀선이 설명하지 못하는 변동) : $\sum(y_i - \hat{y}_i)^2$

■변동 분할



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

$$SST = SSR + SSE$$

총 변동은

회귀선이 설명하는 변동 + 회귀선이 설명하지 못하는 변동

적합성 (Goodness of fit) 검정

결정계수 $\mathbf{R^2}$

- 총 변동(SST)에서 회귀식이 설명할 수 있는 비율(SSR)
- 회귀식이 얼마나 데이터를 잘 설명하는가를 판단

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

결정계수가 1에 가까울수록

모형이 데이터를 잘 설명한다고 판단!

유의성 검정

개별 모수의 추정량이 통계적으로 유의한지를 알아보는 검정

0) 오차의 정규분포 가정 : $\varepsilon_i \sim N(0, \sigma^2)$

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

β_0에 대해서도 동일한 방법으로 검정 가능

2 단순선형회귀분석

유의성 검정

개별 모수의 추정량이 통계적으로 유의한지를 알아보는 검정

귀무가설 기각

→ "X, Y 간 선형 관계가 **있다**"

0) 오차의 정규분포 가정 : $\varepsilon_i \sim N(0, \sigma^2)$

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

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

귀무가설을 기각 X

→ "X, Y 간 선형 관계가 **없다**"

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

4) 임계값 : $t_{(1-\alpha/2, n-2)}$

But **비선형적 관계** 있을 수도 있다!!

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

β_0에 대해서도 동일한 방법으로 검정 가능

3

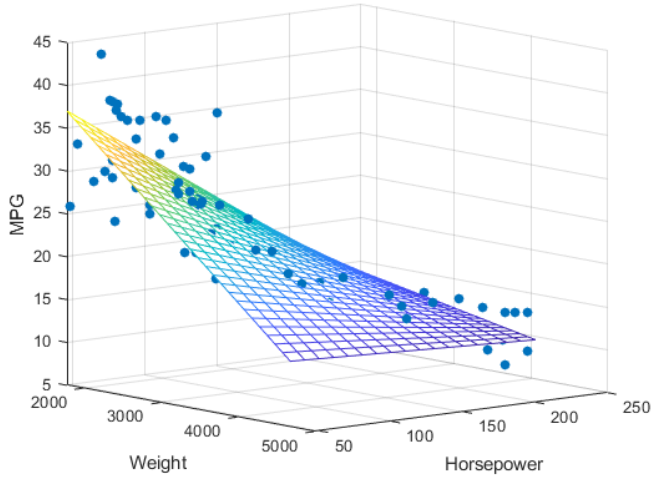
다중선형회귀

다중선형회귀

독립변수가 2개 이상인 경우의 회귀분석

단순선형회귀에 비해 **복잡한 관계 설명에 용이**

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



다중선형회귀

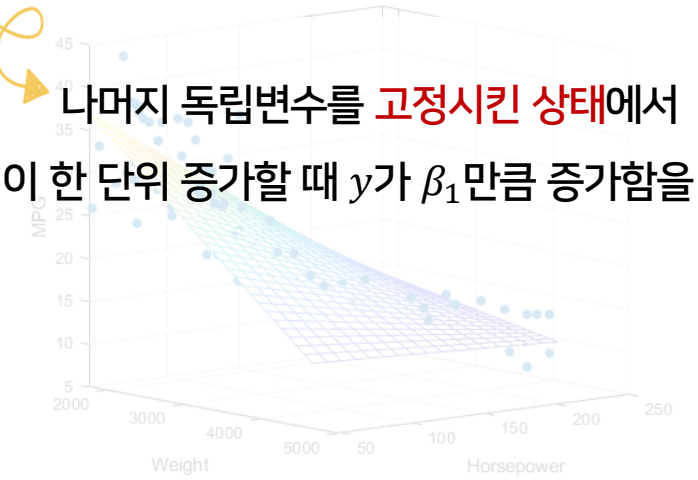
독립변수가 2개 이상인 경우의 회귀분석

단순선형회귀에 비해 **복잡한 관계 설명에 용이**

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

나머지 독립변수를 **고정시킨 상태**에서

x_1이 한 단위 증가할 때 y가 β_1만큼 증가함을 의미



모수의 추정 : 최소제곱법

단순선형회귀와 동일하게 최소제곱법을 활용해 모수 추정 가능

오차의 제곱합 : $\sum_i \epsilon_i^2 = \sum_i (y - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi})^2$

$$\frac{\partial J}{\partial \beta_0} = -2 \sum_i \left(y - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}\right) = 0$$

$$\vdots$$

$$\frac{\partial J}{\partial \beta_p} = -2 \sum_i \left(y - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}\right) x_{pi} = 0$$

3 다중선형회귀

모수의 추정 : 최소제곱법

단순선형회귀와 동일하게 최소제곱법을 활용해 모수 추정 가능

모수가 p+1개인 다차원 식이기 때문에
편미분을 통해 추정 시 계산식이 매우 복잡해짐!

오차의 제곱합 : $\sum e_i^2 = \sum (y - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi})^2$

$$\frac{\partial J}{\partial \beta_0} = -2 \sum_i (y - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}) = 0$$

⋮

$$\frac{\partial J}{\partial \beta_p} = -2 \sum_i (y - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}) x_{pi} = 0$$

행렬을 활용해 회귀계수 추정

3 다중선형회귀

모수의 추정 : 최소제곱법

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

⋮

목적함수

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

목적함수 J 를 β 에 대해 미분하고

미분식을 0으로 만들어주는 추정량 $\hat{\beta}$ 을 구할 수 있음

모수의 추정 : 최소제곱법

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

다중선형회귀식을 행렬로 표기한 것

목적함수

$$\min J(\beta) = \sum_{i}^{n} \epsilon_i^2 = \epsilon^T \epsilon = (Y - X\beta)^T (Y - X\beta)$$

행렬을 이용해 모수의 추정 가능

목적함수 J 를 β 에 대해 미분하고

미분식을 0으로 만들어주는 추정량 $\hat{\beta}$ 을 구할 수 있음

모수의 추정 : 최소제곱법

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

⋮

추정량

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

추정된 회귀식

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

($H = X(X^T X)^{-1} X^T$ 는 **투영행렬**)

3 다중선형회귀

모수의 추정 : 최소제곱법

투영 행렬이란?

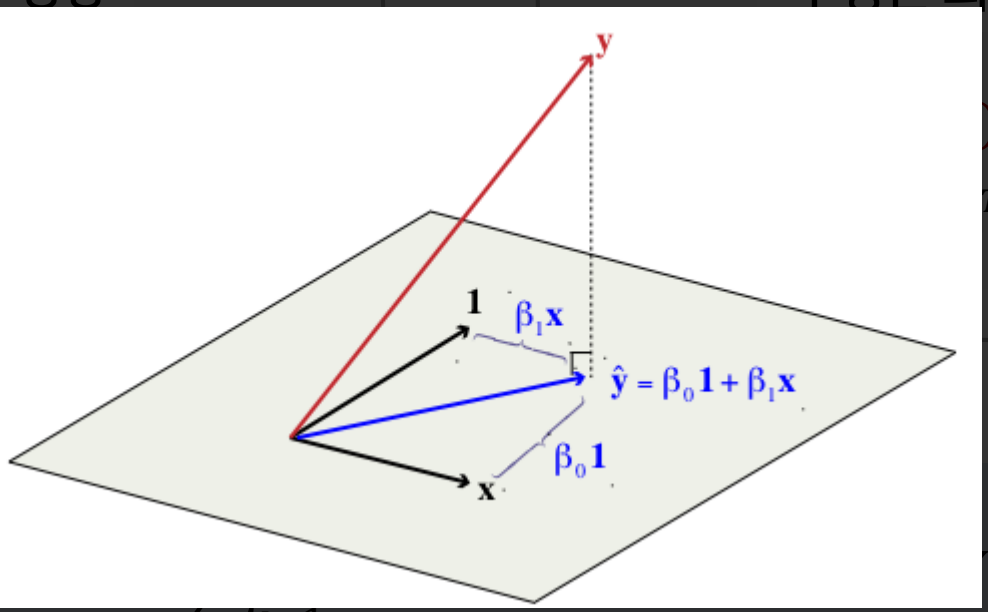
추정량

$\hat{\beta} = (X$

추정된 회귀식

$)^{-1}X^TY = HY$

^{T} 는 투영행렬)



목적함

mi

$\beta)$

Y 를 X 가 생성하는 공간에 가장 가깝게 근사(오차를 최소화)하기 위해 사용

→ Y 를 X 의 열공간에 투영함으로써 $\hat{\beta}$ 을 찾음

미분식을 0으로 만들어주는 추정량 $\hat{\beta}$ 을 구할 수 있음

선형대수학팀 클린업 2주차 참고!

적합성(Goodness of fit) 검정

회귀직선이 데이터에 얼마나 잘 들어맞는지 모형에 대한 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) 검정

결정계수(R square)

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

독립변수를 추가하면 반드시 R^2 값이 증가

아무 의미가 없는 변수라도 설명하는 변동이 생김

⋮

중요하지 않은 변수를 추가함에 따라 모델의 해석도 어려워지고,

예측에도 좋지 않은 영향

독립변수가 늘어날 때, 페널티를 줄 필요성이 생김

적합성(Goodness of fit) 검정

수정결정계수(Adjusted R square)

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

R^2 에 변수 개수에 대한 페널티를 부과한 형태

⋮

R^2_{adj} 가 높은 모델이 적은 변수로 좋은 설명력을 가진다는 의미이므로

더 좋은 모델이라고 할 수 있음

그러나 R^2와 달리 그 자체로 해석은 어려움

3 다중선형회귀

유의성 검정

추정량이 **통계적으로 유의**한지를 알아보는 검정

다중선형회귀의 3가지 test

1. F-test : 전체 회귀계수에 대한 검정
2. Partial F-test : 일부 회귀계수 group에 대한 검정
3. T-test : 개별 회귀계수에 대한 검정

유의성 검정

F-test

전체 회귀계수에 대한 검정

1. 가설 설정

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

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

⋮

귀무가설이 기각되어야 모형이 의미 있다고 할 수 있음

유의성 검정

F-test

전체 회귀계수에 대한 검정

2. 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

검정통계량이 임계값보다 크다면 귀무가설 기각

→ SSR과 SSE가 각각 분자, 분모에 위치하므로,

회귀식이 설명하는 부분이 그렇지 않은 부분보다 충분히 크다는 것을 의미

유의성 검정

F-test

전체 회귀계수에 대한 검정

2. 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

검정통계량이 임계값보다 크다면 귀무가설 기각

→ SSR과 SSE가 각각 분자, 분모에 위치하므로,

회귀식이 설명하는 부분이 그렇지 않은 부분보다 충분히 크다는 것을 의미

유의성 검정

F-test

전체 회귀계수에 대한 검정

3. 임계값

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

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

▶ 적어도 한 개의 회귀계수는 0이 아님

② $F_0 < F_{\left(1-\frac{a}{2},\ p,\ n-p-1\right)}$ 이면 귀무가설 기각 X

▶ 모든 회귀계수가 0임, 모델 재설정 필요

유의성 검정

Partial F-test

일부 회귀계수 group에 대한 검정

1. 가설 설정

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

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

Full model (FM) = 모든 변수를 사용한 회귀모형

Reduced Model (RM) = 일부 계수($\beta_j \sim \beta_{j+q-1}$)를 특정 값으로 둔 축소 모형

유의성 검정

Partial F-test

일부 회귀계수 group에 대한 검정

2. 검정통계량

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

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

유의성 검정

Partial F-test

일부 회귀계수 group에 대한 검정

2. 검정통계량

q개의 변수를 제거했을 때

모델이 설명하지 못하는 변동

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

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

유의성 검정

Partial F-test

일부 회귀계수 group에 대한 검정

2. 검정통계량

모든 변수를 포함했을 때

모델이 설명하지 못하는 변동

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

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

3 다중선형회귀

유의성 검정

일반적으로 변수를 제거하면 $SSE(RM) > SSE(FM)$

Partial F-test

일부 회귀계수 group에 대한 검정

이 때 제거된 변수가 의미있는 변수라면,

2. 검정통계량

$SSE(RM)$이 매우 커질 것

모든 변수를 포함했을 때

모델이 설명하지 못하는 변동

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

검정통계량 F_0

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

검정통계량이 귀무가설을 기각시킬만큼 충분히 커짐

3 다중선형회귀

유의성 검정

Partial F-test

일부 회귀계수 group에 대한 검정

2. 검정통계량

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

Partial F-test가 더 일반적인 검정이지만

보편적으로 F-test를 더 많이 사용

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

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

유의성 검정

T-test

개별 회귀계수에 대한 검정

1. 가설 설정

$\mathbf{H_0}$: β_j=0 (다른 변수들이 적합된 상태에서 설명변수 x_j는 유의하지 않음)

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

⋮

귀무가설을 기각하면, 다른 변수들이 적합된 상태에서 x_j를 추가적으로 적합하는 것이 회귀식의 설명력을 유의미하게 증가시킨다고 할 수 있음

유의성 검정

T-test

개별 회귀계수에 대한 검정

2. 검정통계량

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

3. 임계값

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

- ▶ 현재 모델에서 x_j의 회귀 계수는 0이 아님
- ▶ x_j의 추가는 유의미한 회귀식 설명력의 증가를 가져옴

3 다중선형회귀

유의성 검정

T-test

t-test를 활용해 변수를 선택할 수 있을까?

개별 회귀계수에 대한 검정

t-test는 다른 변수들이 다 적합된 상태에서
해당 변수의 추가가 유의미한 설명력의 증가를 가져오는지 확인하는 것

2. 검정통계량

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

3. 임계값

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

▶ 현재 모델에서 x_j의 회귀 계수는 0이 아님

▶ x_j의 추가는 유의미한 회귀식 설명력의 증가를 가져옴

3 다중선형회귀

유의성 검정

T-test

t-test를 활용해 변수를 선택할 수 있을까?

개별 회귀계수에 대한 검정

t-test는 다른 변수들이 다 적합된 상태에서
해당 변수의 추가가 유의미한 설명력의 증가를 가져오는지 확인하는 것

2. 검정통계량

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

3. 임계값

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

다른 회귀식을 가정하면 해당 변수의 유의성도 바뀔 수 있음!

현재 모델에서 x_j의 회귀계수는 0이 아니

t-test로 변수를 선택하는 것은 매우 위험

▶ x_j의 추가는 유의미한 회귀식 설명력의 증가를 가져옴

3 다중선형회귀

T-test vs F-test

F-test를 먼저 시행해야 하는 이유

✓ 전체 회귀식에 대한 검정이 더 엄격

✓ F-test를 기각하지 못해도 T-test는 기각하는 경우 발생 가능

F-test를 먼저 시행하여

모델 전체가 통계적으로 유의한지 확인해야 함

4

데이터 진단

4 데이터 진단

잔차

잔차(e)

왜 데이터 진단을 해야 할까?

설명할 수 없는 오차(ϵ)의 추정치

관측된 종속변수와 추정된 종속변수의 차($y - \hat{y}$)로 구해짐

일반적인 경향을 벗어난 점들

$e = (I - H)y$

최소제곱 회귀모형을 크게 변화시키거나, **성능을 저하**시킬 수 있음

스튜던트화 잔차(Studentized residual)

단위의 영향을 받는 잔차를 일반화해서 적용하도록 표준화한 것

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

잔차

잔차(e)

설명할 수 없는 오차(ϵ)의 추정치

관측된 종속변수와 추정된 종속변수의 차($y - \hat{y}$)로 구해짐

$$e = (I - H)y$$

스튜던트화 잔차(Studentized residual)

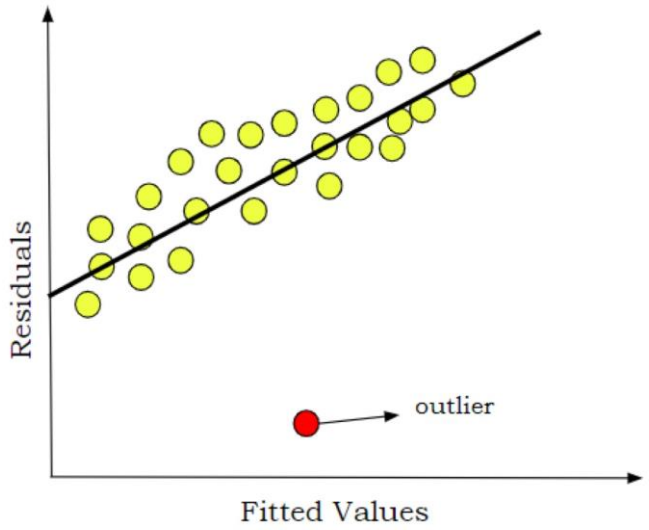
단위의 영향을 받는 잔차를 일반화해서 적용하도록 **표준화**한 것

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

4 데이터 진단

이상치(Outlier)

표준화 잔차가 매우 큰 값 (y를 기준으로 절댓값이 큰 값)



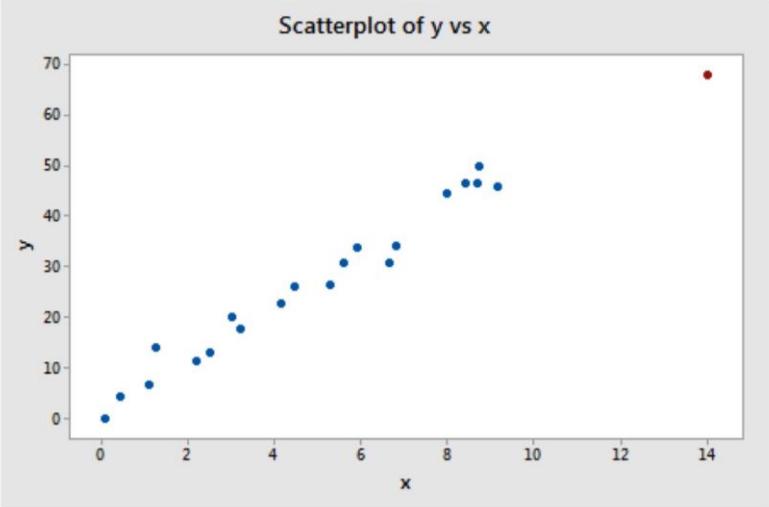
보통 $|r_i| > 3$ 인 점을
이상치로 판단

4 데이터 진단

지렛값(Leverage point)

x의 **평균에서 멀리 떨어져** 있어 **기울기에 큰 영향**을 주는 값

이상치가 y의 관점이었다면, 지렛값은 x의 관점



$h_{ii} > \frac{2(p+1)}{n}$ 인 점을

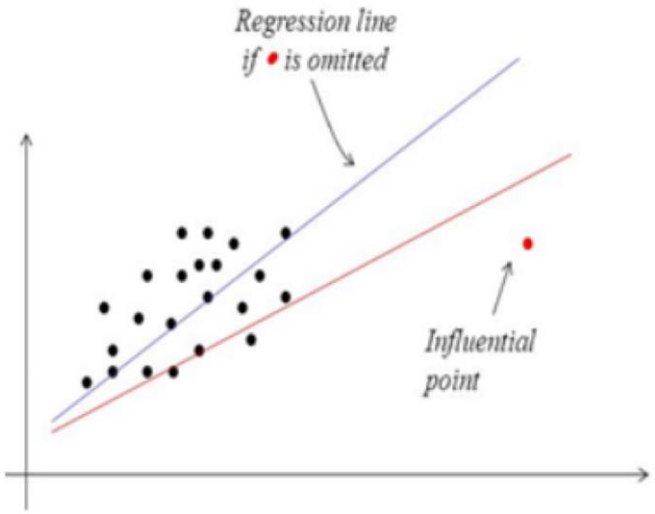
지렛값으로 판단

4 데이터 진단

영향점(Influential point)

회귀직선의 기울기에 상당한 영향을 주는 관측치

이상치와 지렛값을 동시에 고려



왼쪽그림의 경우

빨간 점의 유무에 따라

회귀직선의 기울기가 크게 변화

영향점!

4 데이터 진단

영향점(Influential point)

회귀직선의 기울기에 상당한 영향을 주는 관측치

이상치와 지렛값을 동시에 고려

이상치

x평균 주위에 위치하면

↓

기울기를 변화시키지 못함

지렛값

회귀직선의 연장선에 있다면

↓

기울기를 변화시키지 못함

앞 페이지 그림 참고

4 데이터 진단

영향점(Influential point)

회귀직선의 기울기에 상당한 영향을 주는 관측치

따라서 이상치와 지렛값을 동시에 고려하는 지표 필요!

Cook's distance!

이상치

x평균 주위에 위치하여
기울기를 변화시키지 못함

지렛값

회귀직선의 연장선에 있어
기울기를 변화시키지 못함

앞 페이지 그림 참고

4 데이터 진단

영향점(Influential point)

Cook's distance

회귀직선의 기울기에 상당한 영향을 주는 관측치

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

- ✓ 이상치와 지렛값 각각이 커질수록 C_i 커짐
- ✓ $C_i > 1$이면 영향점으로 판단

이상치

x평균 주위에 위치하여 기울기를 변화시키지 못함

지렛값

회귀직선의 연장선에 있어 기울기를 변화시키지 못함

앞 페이지 그림 참고

영향점의 처리

영향점은 추정량을 불안정하게 하고, 예측 성능 저하를 일으킬 수 있음

→ **적절한 처리** 필요

그러나 영향점이 의미 있는 데이터일 수 있기 때문에

단순히 **영향점을 삭제하는 것은 적절하지 않음**

영향점의 처리

영향점은 추정량을 불안정하게 하고, 예측 성능 저하를 일으킬 수 있음

→ **적절한 처리** 필요

그러나 영향점이 의미 있는 데이터일 수 있기 때문에

단순히 **영향점을 삭제하는 것은 적절하지 않음**

이상치에 강건한(robust) 모델링이 필요한 이유!

5

로버스트 회귀

로버스트 회귀

이상치의 영향을 줄이는 회귀분석 방법

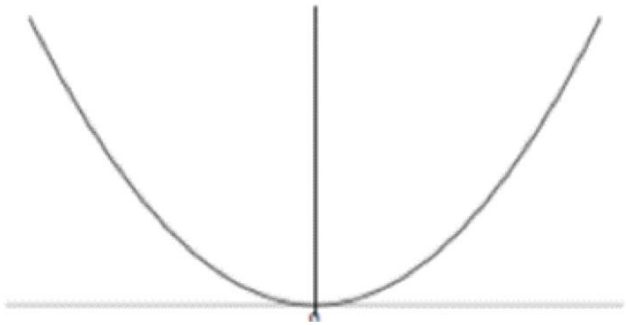
로버스트 회귀

이상치의 영향을 줄이는 회귀분석 방법

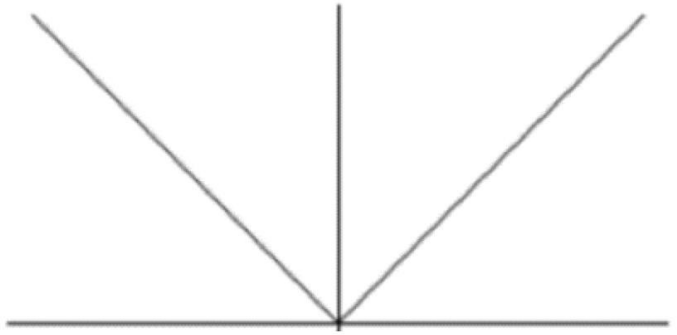
5 로버스트 회귀

Median Regression

모든 경우에 **동일한 가중치**를 주어 최소제곱회귀의 단점을 극복하는 회귀분석 방법



LSE

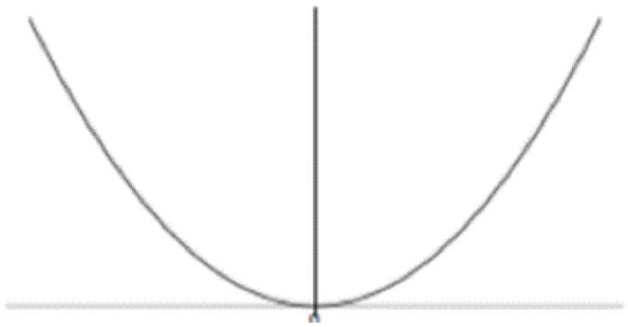


Median Regression

5 로버스트 회귀

Median Regression

모든 경우에 **동일한 가중치**를 주어 최소제곱회귀의 단점을 극복하는 회귀분석 방법



LSE

LSE

- ✓ **오차의 제곱합을 최소**로 하는 추정량
- ✓ X에 따른 평균적인 Y를 반환

5 로버스트 회귀

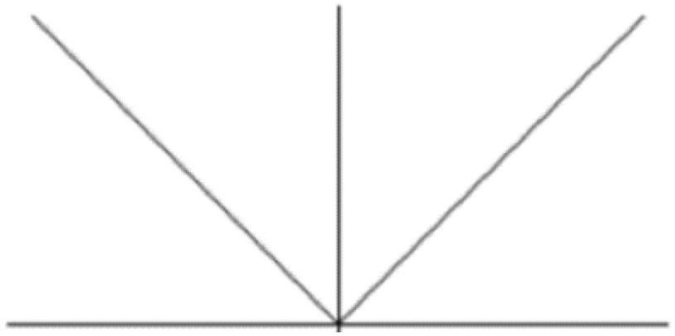
Median Regression

모든 경우에 **동일한 가중치**를 주어 최소제곱회귀의 단점을 극복하는 회귀분석 방법

Median Regression

- ✓ **오차 절대값의 합**을 최소화
- ✓ **조건부 중앙값** 추정

이상치에 덜 민감한 추정량 가짐



Median Regression

5 로버스트 회귀

Huber's M-estimation

이상치에 큰 가중치를 주는 최소제곱회귀의 단점을 극복하면서

적정 수준 내에서 페널티를 부여하는 회귀분석 방법

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

잔차의 절댓값이 c 이하

기존 최소제곱추정법의

목적함수와 동일

5 로버스트 회귀

Huber's M-estimation

이상치에 큰 가중치를 주는 최소제곱회귀의 단점을 극복하면서

적정 수준 내에서 페널티를 부여하는 회귀분석 방법

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

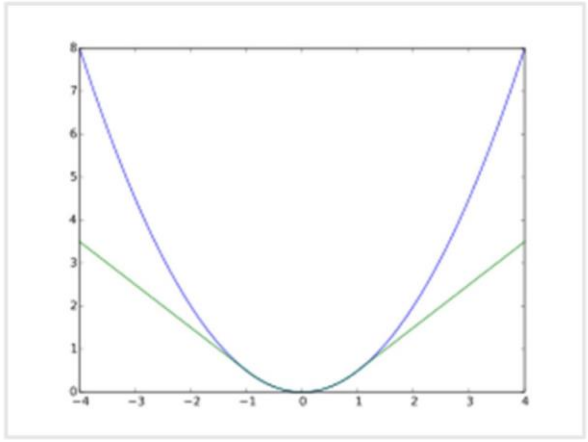
잔차의 절댓값이 c 이상

이상치에 큰 페널티를 주지 않는

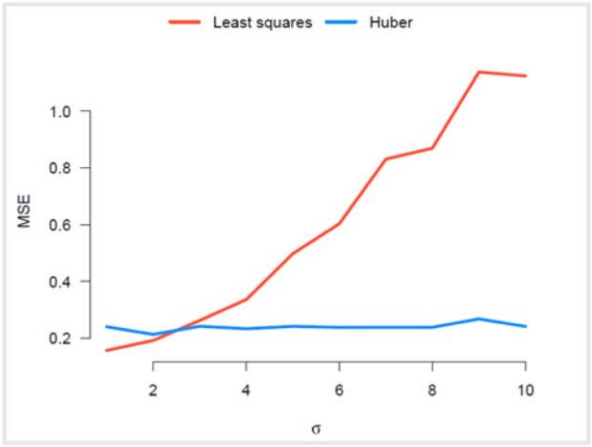
일차식 형태

5 로버스트 회귀

Huber's M-estimation



파란색 : 최소제곱회귀



초록색 : Huber's M-estimation

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

이상치에 대한 페널티를 완화하여 MSE값을 줄임

일차식 형태

5 로버스트 회귀

Least Trimmed Squre

통계적 기준에 따라 잔차가 너무 큰 관측치를 제거하고
회귀계수를 추정하는 방식

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

$r_{(j)}$는 작은 순서부터 오름차순으로 나열한 잔차

관찰값이 적거나 영향점이 없는 경우 주의하여 사용

감사합니다