회귀분석팀

6팀

김형석

김준령

윤여원

김현우

이채은

INDEX

1

Introduction

표본 평균, 표본 분산

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)

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

표본 표준 편차, 편차 제곱 합

표본 표준편차(Sample Standard Deviation)

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

편차 제곱 합

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

표본 공분산

표본 공분산(Sample Covariance)

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

$\widehat{Cov(X,Y)} > 0$: 양의 선형 상관관계를 가짐

$\widehat{Cov(X,Y)} < 0$: 음의 선형 상관관계를 가짐

$\widehat{Cov(X,Y)} = 0$: 선형 상관관계를 갖지 않음

1 Introduction

표본 공분산

표본 공분산(Sample Covariance)

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

표본 공분산이 선형관계의 강도를 의미하지는 않음

두 변수 X, Y 간에 단위(Scale) 가 다를 수 있어

선형관계의 강도를 비교할 수는 없음

$\widehat{Cov(X,Y)} > 0$: 양의 선형 상관관계를 가짐

$\widehat{Cov(X,Y)} < 0$: 음의 선형 상관관계를 가짐

$\widehat{Cov(X,Y)} = 0$: 선형 상관관계를 갖지 않음

단위를 통일시킨 표본 상관계수 도입 !

표본 상관계수

표본 상관계수(Sample Correlation Coefficient)

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

① -1 ~ 1의 값

② 1이나 -1에 가까울수록 **강한 선형 상관관계**를,

0에 가까울수록 **약한 선형 상관관계**를 가짐

③ 만약 변수 X, Y 가 **독립**인 확률변수라면 $\widehat{r_{xy}} = 0$

(단, 역은 X, Y ~ Normal distribution인 경우에만 성립함)

표본 상관계수

표본 상관계수(Sample Correlation Coefficient)

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

$\widehat{r_{xy}} = 0$ 이라고 해서 변수 간에 관계가 아예 없는 것이 아님

선형관계를 갖지 않지만, **비선형관계는 존재할 수 있음**

선형관계를 갖지 않지만 비선형관계는 존재하는 예시

2

Concepts of Regression Analysis

2 Concepts of Regression Analysis

회귀분석의 정의

회귀분석(Regression Analysis)

독립변수(X)와 종속변수(Y) 간에 관계를 설명하고 모델링하는 통계적 기법
모델링 시 함수의 형태를 가정하는 **모수적 방법(Parametric Method)**이자
지도학습(Supervised Learning) 종류에 해당 됨

회귀분석을 통해 변수들 간의 상관관계를 파악하여
종속변수(Y)의 값을 독립변수(X)들을 이용하여 설명하고 예측 가능

2 Concepts of Regression Analysis

회귀분석의 정의

회귀분석(Regression Analysis)

독립변수(X)와 종속변수(Y) 간에 관계를 설명하고 모델링하는 통계적 기법
모델링 시 함수의 형태를 가정하는 **모수적 방법(Parametric Method)**이자
지도학습(Supervised Learning) 종류에 해당 됨

회귀분석을 통해 변수들 간의 상관관계를 파악하여
종속변수(Y)의 값을 독립변수(X)들을 이용하여 **설명**하고 **예측** 가능

2 Concepts of Regression Analysis

회귀식의 일반 구조

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

① X_p: 독립 or 설명 변수(Independent or Explanatory variable)

② Y: 종속 or 반응 변수(Dependent or Response variable)

③ f : 독립변수들과 종속변수 간에 관계를 나타내는 함수

④ ε : 변수를 측정할 때 발생할 수 있는 오차 (Error term)

IID 한 정규분포를 따르며,

독립변수와의 상관성은 없다($\mathrm{Cov}(X, \varepsilon) = 0$) 라고 가정함

회귀분석의 특징

상관분석의 한계를 회귀분석을 통해 해결 가능

상관분석은 선형적 상관관계를 표현할 수 있지만,
구체적인 예측과 해석은 불가능하다는 한계를 가짐

회귀분석은 인과관계가 아닌 상관관계에 기반한 분석 기법

회귀분석의 결과 자체는 인과관계라 보긴 어렵지만,
해당 결과를 통해 인과관계 해석을 시도해볼 순 있음

2 Concepts of Regression Analysis

회귀분석의 특징

인과관계와 상관관계의 차이

상관관계는 인과관계의 필요조건!

상관분석의 한계를 회귀분석을 통해 해결 가능

상관분석은 선형적 상관관계를 표현할 수 있지만,
구체적인 예측과 해석은 불가능하다는 한계를 가짐

상관관계

인과관계

회귀분석은 인과관계가 아닌 상관관계에 기반한 분석 기법

회귀분석의 결과 자체는 인과관계라 보긴 어렵지만,
해당 결과를 통해 인과관계 해석을 시도해볼 순 있음

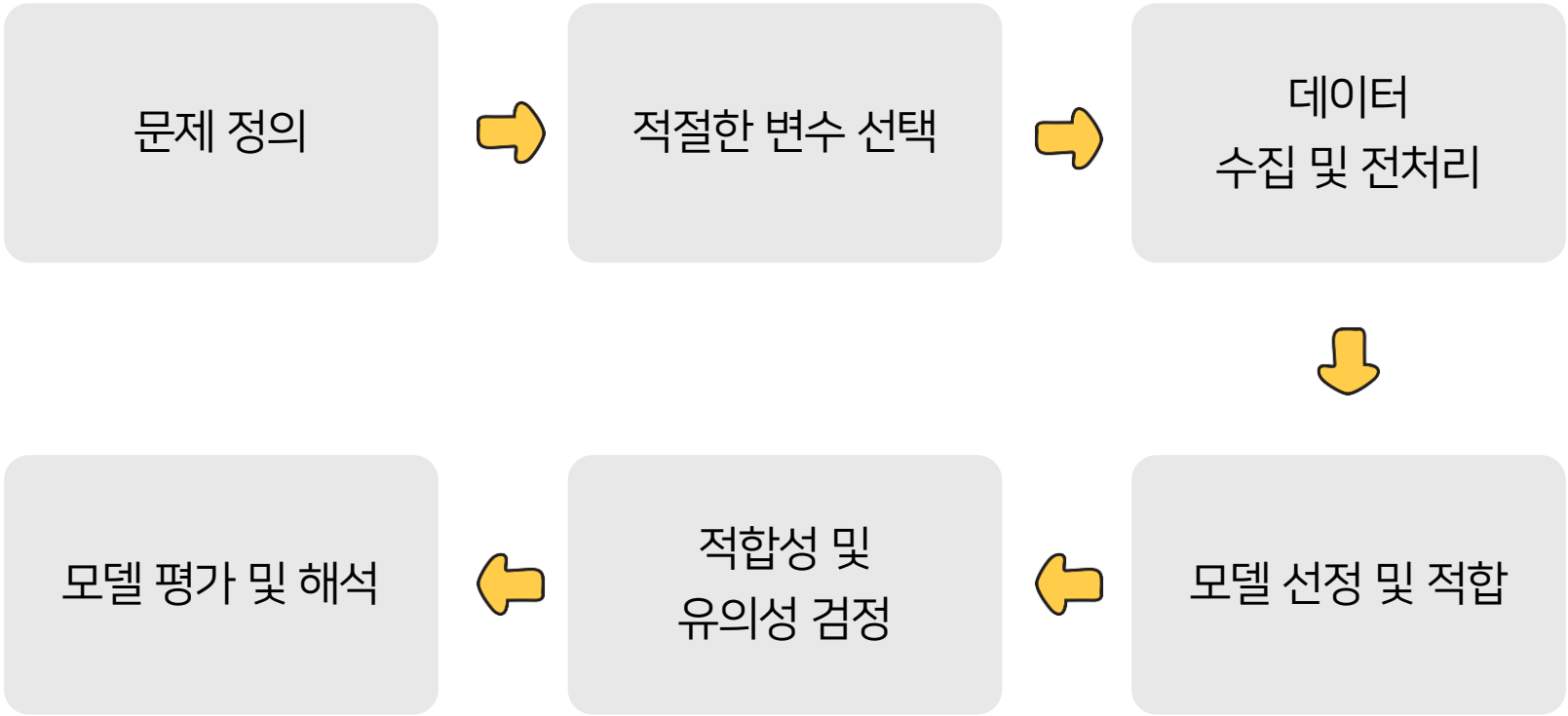
인과관계 성립 조건

① 두 사건은 통시적 관계에 놓임

② 선행 사건이 있어야 후행 사건이 존재

③ 두 시점의 사건 사이에
영향을 끼치는 요인은 더 이상 없음

회귀분석의 진행과정



회귀분석의 진행과정

문제 정의

적절한 변수 선택

데이터
수집 및 전처리

이번 주차에는 아래 두 단계에 대해 중점적으로 다룰 예정!

모델 평가 및 해석

적합성 및
유의성 검정

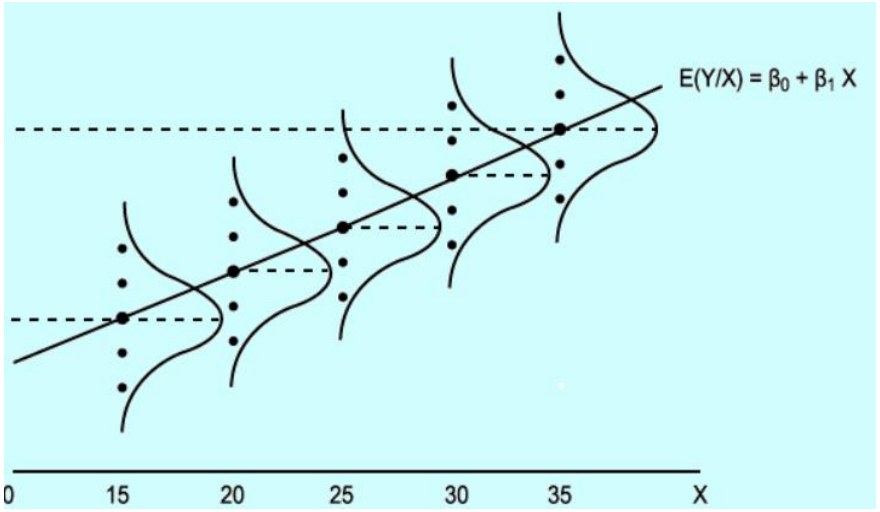
모델 선정 및 적합

3

Simple Linear Regression

단순 선형회귀분석 | 목적

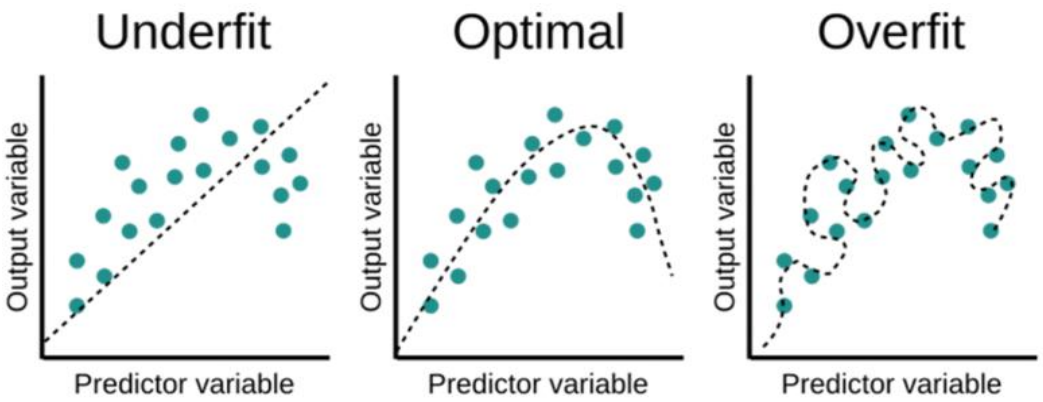
단일 반응변수(Y) 와 설명변수(X) 사이의
관계를 규명하는 것



단순 선형회귀분석 | 선형회귀식의 장점

① 변수의 영향력을 매우 간단하게 모델링할 수 있어, 모델의 해석력이 높음

② 모델이 과적합(Overfitting)에 빠질 위험성이 적음



3 Simple Linear Regression

단순 선형회귀식

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = E(Y_i|X_i) + \varepsilon_i$$
$$for\ i = 1, \dots, n$$

① X_i: 독립변수 X 의 i 번째 관측 값

② Y: 종속변수 Y 의 i 번째 관측 값

③ β_0: $X = 0$ 일 때 Y 의 예측 값 (Unknown Intercept)

④ β_1 : X 가 한 단위 증가할 때 Y 의 평균 변화량 (Unknown Slope)

⑤ ε_i : 오차항

단순 선형회귀식

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = E(Y_i|X_i) + \varepsilon_i$$
$$for\ i = 1, \dots, n$$

Mean of the Model

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

평균적으로 회귀 직선에
데이터들이 존재함을 의미

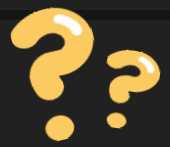
Variance of the Model

$$V(Y_i|X_i) = \sigma^2\ (\because\ \beta, X_i \text{ 는 상수})$$

데이터의 분산은 일정함을 의미

3 Simple Linear Regression

단순 선형회귀식



$E(Y_i|X_i)$ 와 $E(Y_i)$ 의 차이

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = E(Y_i|X_i) + \varepsilon_i \quad for\ i = 1, \dots, n$$

	$E(Y_i\|X_i)$	$E(Y_i)$
개념	X_i 값이 주어졌을 때, 개별 Y_i 값에 대한 Expectation	반응변수(Y) 전체에 대한 Expectation
수식 표현	$\beta_0 + \beta_1 X_i$	$\beta_0 + \beta_1 \mu_X$ ($\because \beta_1 X_i$ 는 상수)
특징	Number of i 개수만큼 $E(Y_i\|X_i)$ 값들이 존재	단일 값

모수의 추정 | 최소제곱법

최소제곱법(OLS: Ordinary Least Squares)

구하고자 하는 해와 실제 해 간의

오차 제곱합이 최소가 되는 해를 구하여

가장 적합한 회귀직선을 찾는 기법

⋮

$$\underset{\beta_0, \beta_1}{\arg min}\, J(\beta_0, \beta_1) = \underset{\beta_0, \beta_1}{\arg min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \underset{\beta_0, \beta_1}{\arg min} \sum_{i=1}^{n} {\varepsilon_i}^2$$

모수의 추정 | 최소제곱법

최소제곱법(OLS: Ordinary Least Squares)

구하고자 하는 해와 실제 해 간의

오차 제곱합이 최소가 되는 해를 구하여

가장 적합한 회귀직선을 찾는 기법

OLS Estimator or LSE (Least Squares Estimator)

모수 β, σ^2을 추정하는 것이 목적!

모수의 추정 | 최소제곱법

목적식에 대해 OLS 적용

$$\left.\frac{\partial J}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \cdots ①$$

$$\left.\frac{\partial J}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0 \cdots ②$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

모수의 추정 | 최소제곱법

$\hat{\beta}_1, \hat{\beta}_0$ 점 추정량 도출 과정

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{1}{S_{xx}}\left\{\sum_{i=1}^{n}\{(x_i - \bar{x})y_i - (x_i - \bar{x})\bar{y}\}\right\}$$

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{1}{n}\sum_{i=1}^{n}y_i - \frac{\bar{x}}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})y_i$$

$$= \sum_{i=1}^{n}\left\{\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}\right\}y_i$$

모수의 추정 | 최소제곱법

$\hat{\beta}_1, \hat{\beta}_0$ 점 추정량 도출 결과

$$\hat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \{ \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} \} y_i$$

모수의 추정 | 최소제곱법

$\hat{\beta}_0$의 분포 도출 과정

$$E(\hat{\beta}_0) = E(\bar{y}) - E(\hat{\beta}_1\bar{x}) = \bar{y} - \bar{x}\,E(\hat{\beta}_1) = \bar{y} - \beta_1\bar{x} = \beta_0$$

$$V(\hat{\beta}_0) = V\left[\sum_{i=1}^{n}\left\{\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{S_{xx}}\right\}y_i\right] = V(y_i|x_i)\sum_{i=1}^{n}\left\{\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{S_{xx}}\right\}^2$$

$$= \sigma^2\sum_{i=1}^{n}\left(\frac{1}{n^2} - \frac{2\,\bar{x}(x_i-\bar{x})}{n\,S_{xx}} + \frac{\bar{x}^2(x_i-\bar{x})^2}{{S_{xx}}^2}\right)$$

$$= \sigma^2\sum_{i=1}^{n}\left(\frac{1}{n^2} + \frac{\bar{x}^2(x_i-\bar{x})^2}{{S_{xx}}^2}\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2\,S_{xx}}{{S_{xx}}^2}\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

모수의 추정 | 최소제곱법

$\hat{\beta}_1$의 분포 도출 과정

$$E(\hat{\beta}_1) = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) E(y_i | x_i)$$

$$= \frac{1}{S_{xx}} \left\{ \beta_0 \sum_{i=1}^{n} (x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i \right\} = \frac{\beta_1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \beta_1$$

$$V(\hat{\beta}_1) = V\left[\frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i \right] = \frac{1}{{S_{xx}}^2} V\left[\sum_{i=1}^{n} (x_i - \bar{x}) y_i \right]$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{{S_{xx}}^2} V(y_i | x_i) = \frac{S_{xx}}{{S_{xx}}^2} \sigma^2 = \frac{\sigma^2}{S_{xx}}$$

모수의 추정 | 최소제곱법

$\hat{\beta}$ 의 분포 도출

오차항이 **IID 하게 정규분포를 따른다는 가정**을 고려하였을 때

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right))$$

모수의 추정 | 최소제곱법

회귀계수 분포 도출 결과를 자세히 보면...

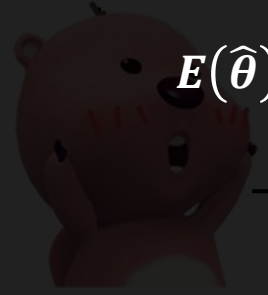
β 의 분포 도출

오차항이 IID 하게 정규분포를 따른다는 가정을 고려하였을 때

$$E(\hat{\beta}_0) = \beta_0, E(\hat{\beta}_1) = \beta_1$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right))$$



$E(\hat{\theta}) = \theta$ 를 만족할 때 $\hat{\theta}$ 은 Unbiased Estimator 라고 부르기 때문에,

$\hat{\beta}_0, \hat{\beta}_1$은 모두 Unbiased Estimator 임을 확인 가능!

모수의 추정 | 최대우도법

최대우도법(MLE: Maximum Likelihood Estimation)

표본(데이터)이 나올 가능도(Likelihood, $L(\theta)$)를
최대로 하는 모수 θ를 선택하는 방법

$$\hat{\theta} = \underset{\theta}{\arg max}\ L(\theta) = \underset{\theta}{\arg max}\ \prod_{i=1}^{n} f(x_i;\ \theta),\quad f(x_i;\ \theta)\ is\ the\ pdf\ of\ x$$

오차의 정규분포 가정이 있다면 LSE 와 MLE 의 결과는 동일

모수의 추정 | 최대우도법

최대우도법(MLE: Maximum Likelihood Estimation)

표본(데이터)이 나올 가능도(Likelihood, $L(\theta)$)를
최대로 하는 모수 θ를 선택하는 방법

$$\hat{\theta} = \underset{\theta}{\arg max}\ L(\theta) = \underset{\theta}{\arg max}\ \prod_{i=1}^{n} f(x_i;\ \theta),\quad f(x_i;\ \theta)\ is\ the\ pdf\ of\ x$$

오차의 정규분포 가정이 있다면 LSE 와 MLE 의 결과는 동일

BLUE

BLUE(Best Linear Unbiased Estimator)

분산이 가장 작은(Best) 선형(Linear) 불편(Unbiased) 추정량

분산이 가장 작다는 것은,
추정량이 안정적이라는 것을 의미!

BLUE 조건

① 오차항의 Expectation은 0

② 오차항의 Variation은 σ^2 로 일정

③ 오차항들 간에는 자기상관이 없음(Uncorrelated)

3 Simple Linear Regression

BLUE

BLUE(Best Linear Unbiased Estimator)

분산이 가장 작은(Best) 선형(Linear) 불편(Unbiased) 추정량

해당 3가지 조건들을 모두 만족 시,

Gauss Markov Theorem 에 따라 LSE 는 BLUE 가 됨 !

BLUE 조건

① 오차항의 Expectation은 0

만약 LSE 가 BLUE 가 되면, 해당 Estimator 는

불편성(Unbiasedness) 과 일치성(Consistency) 을 만족

회귀식의 적합도 | 잔차

잔차 (Residuals, e_i)

오차항(ε) 에 대한 추정치

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

LSE를 이용해 계산한 y_i에 대한 prediction

⋮

잔차의 합은 항상 0을 만족

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = 0$$

회귀식의 적합도 | 제곱합

잔차의 제곱합(RSS, Residual Sum of Squares)을 이용해

오차항의 분산(σ^2)에 대한 추정량 계산

⋮

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} {e_i}^2}{n-2} = \frac{RSS}{n-2}$$

해당 추정량이 불편성을 만족하도록

$n-2$ (= n − (설명변수 개수 + 1))로 나눔 (즉, $E(\hat{\sigma}^2) = \sigma^2$)

다중선형회귀에서도 마찬가지 적용!

회귀식의 적합도 | 제곱합

Sum of Squares

TSS (or SST)

Total Sum of Squares

y_i 값이 가지는 총 변동

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

ESS (or SSR)

Regression Sum of Squares

회귀직선이 설명하는 변동

$$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

RSS (or SSE)

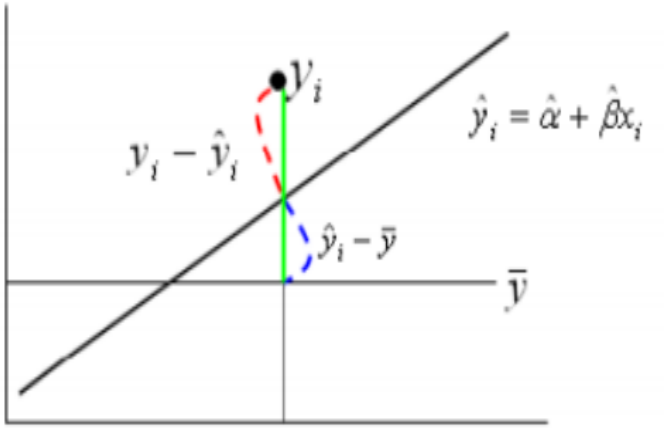
Residual Sum of Squares

회귀직선이 설명 못하는 변동

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

회귀식의 적합도 | 제곱합

y_i 값이 가지는 총 변동(TSS)은
회귀 직선이 설명하는 변동(ESS)과
회귀 직선이 설명하지 못하는 변동(RSS)으로 분리됨



$$TSS = ESS + RSS$$

회귀식의 적합도 | 제곱합

확률변수 Y 의 편차를,

변수 X 에 의해 설명되는 부분($= E(Y|X) - E(Y)$)과

그렇지 않은 부분($= Y - E(Y|X)$) 으로 분리 한 뒤,

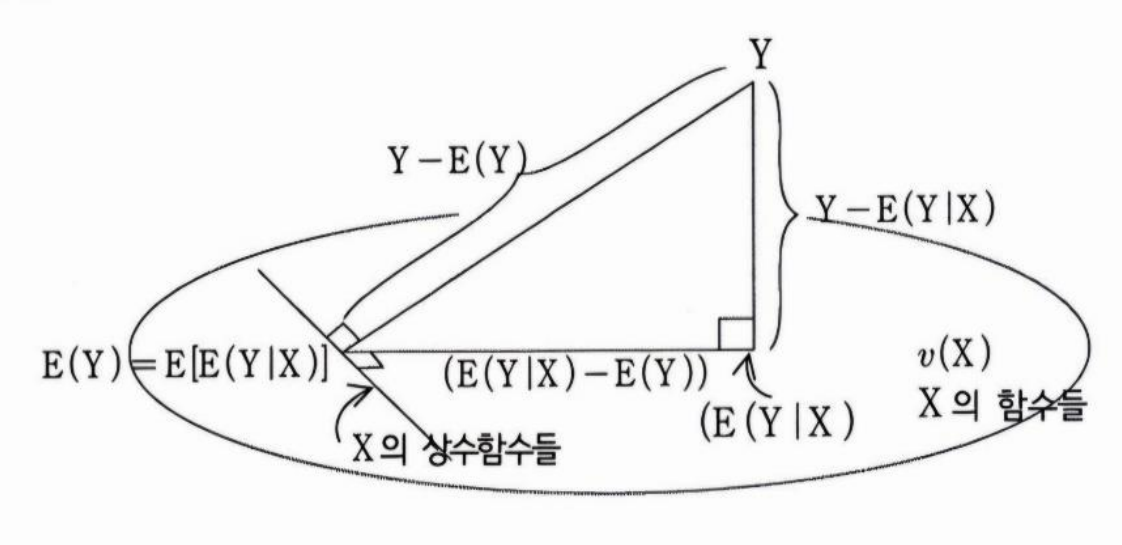
양변 제곱처리 후 Expectation 적용

$$Y - E(Y) = \{Y - E(Y|X)\} + \{E(Y|X) - E(Y)\}$$

$$E[\{Y - E(Y)\}^2] = E[\{Y - E(Y|X)\}^2] + E[\{E(Y|X) - E(Y)\}^2] \cdots ①$$

$$(\because \; E[\{Y - E(Y|X)\}\{E(Y|X) - E(Y)\}] = 0 \cdots ②)$$

회귀식의 적합도 | 제곱합



식 ①, ② 에 입각해 시각화를 해보면 위 그림과 같음

여기서 피타고라스의 정리를 통해 아래와 같은 결과를 도출

$$\{\boldsymbol{Y}-\boldsymbol{E}(\boldsymbol{Y})\}^2=\{\boldsymbol{Y}-\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X})\}^2+\{\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X})-\boldsymbol{E}(\boldsymbol{Y})\}^2 \cdots ③$$

회귀식의 적합도 | 제곱합

Estimator 를 이용해 구한 $\widehat{E(Y|X)}$ 와 $\widehat{E(Y)}$ 의 값을 식 ③ 에 대입

$$\widehat{E(Y|X)} = \hat{\beta}_0 + \hat{\beta}_1 X, \quad \widehat{E(Y)} = \bar{Y} \ (Sample\ mean\ of\ Y)$$

$$\left\{Y - \widehat{E(Y)}\right\}^2 = \left\{Y - \widehat{E(Y|X)}\right\}^2 + \left\{\widehat{E(Y|X)} - \widehat{E(Y)}\right\}^2$$

$$\{\boldsymbol{Y} - \bar{\boldsymbol{Y}}\}^2 = \left\{Y - \hat{\beta}_0 - \hat{\beta}_1 X\right\}^2 + \left\{\hat{\beta}_0 + \hat{\beta}_1 X - \bar{Y}\right\}^2$$

$$= \left\{\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\}^2 + \left\{\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}\right\}^2$$

$$\{\boldsymbol{Y} - \bar{\boldsymbol{Y}}\}^2 = \left\{\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\}^2 + \left\{\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}\right\}^2$$

$$\therefore \ \boldsymbol{TSS} = \ \boldsymbol{RSS} + \boldsymbol{ESS}$$

회귀식의 적합도 | 결정계수

결정계수 (R^2)

반응변수(Y) 의 총 변동(TSS) 에서

회귀식이 설명하는 변동(ESS) 에 의해 설명되는 비율

⋮

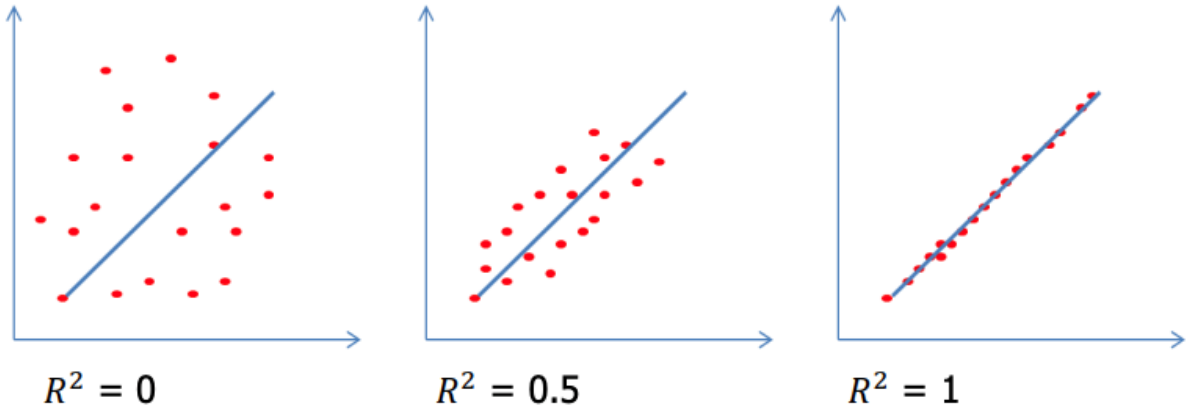
회귀모델을 이용해 반응변수(Y) 의 총 변동을

어느 정도만큼 설명할 수 있는지를 나타내는 지표

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

회귀식의 적합도 | 결정계수

모델이 설명할 수 있는 부분이 클수록(= ESS 의 값이 클수록),
해당 모델이 데이터를 잘 설명하고 있다고 볼 수 있음



결정계수 값의 범위는 **0 과 1 사이**이며,
1 에 가까울수록 회귀모델이 데이터를 잘 설명함

유의성 검정

유의성 검정

회귀계수들에 대한 추정량이 실제로 유의한 값인지 알아보는 검정

⋮

LSE($\hat{\beta}_0, \hat{\beta}_1$) 의 분포 형태

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)), \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

유의성 검정 | T-test

T-test

추정한 **회귀계수들의 유의성을 개별적으로 판단** 시 사용되는 검정방법

Null Hypothesis vs Alternative Hypothesis

$$H_0\colon \beta_0 = 0 \ \ vs \ \ H_1\colon \beta_0 \neq 0, \quad H_0\colon \beta_1 = 0 \ \ vs \ \ H_1\colon \beta_1 \neq 0$$

Test Statistic

$$t_{0,\beta_0} = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)} = \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}, \quad t_{0,\beta_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$$

유의성 검정 | T-test

T-test

추정한 회귀계수들의 [illegible] 사용되는 검정방법

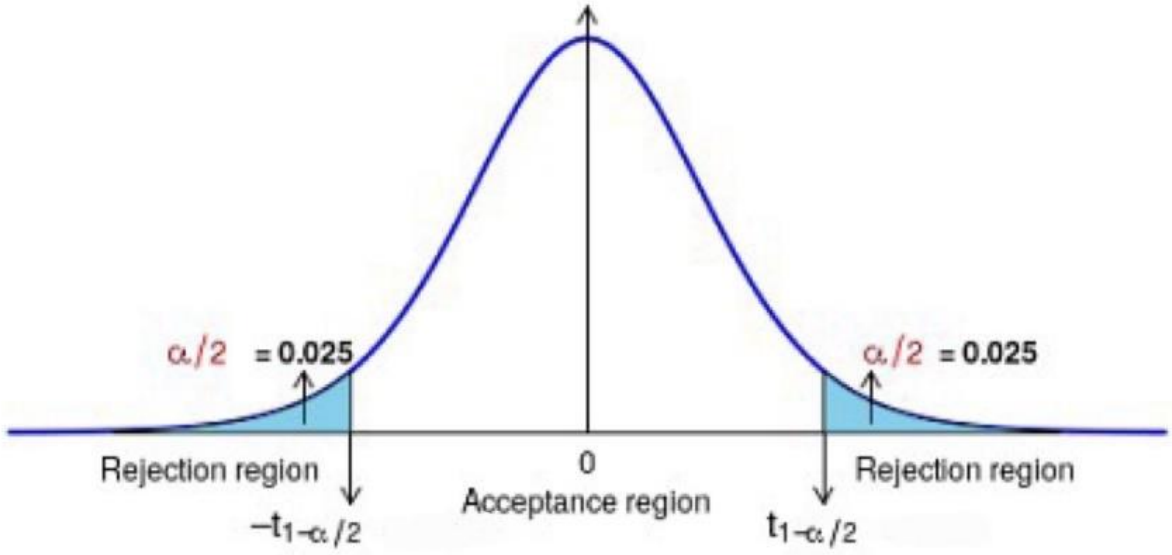
$\hat{\sigma}^2$을 이용하는 이유

σ^2은 모르는 모수이기에, $\hat{\sigma}^2$ 을 대신 이용
Null Hypothesis vs Alternative Hypothesis
이에 따라 Test Statistic 은 정규분포가 아닌
$H_0: \beta_0 = 0 \ vs \ H_1: \beta_0 \neq 0 \quad H_0: \beta_1 = 0 \ vs \ H_1: \beta_1 \neq 0$
t-분포를 따르며, 이때 t-분포의 자유도는 $n-2$ 가 됨

Test Statistic

$$t_{0,\beta_0} = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)} = \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}, \quad t_{0,\beta_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$$

유의성 검정 | T-test



귀무가설을 Reject 하면 해당 계수 값은 유의하다고 결론 내림

$$Reject\ H_0\ if\ |t_0| > t_{1-\alpha/2,\,n-2}\ (Significance\ Level:\ \alpha)$$

유의성 검정 | F-test

F-test

T-test 와 달리, **계수 전체를 하나로 묶어** 유의성을 판단 시 사용되는 검정방법

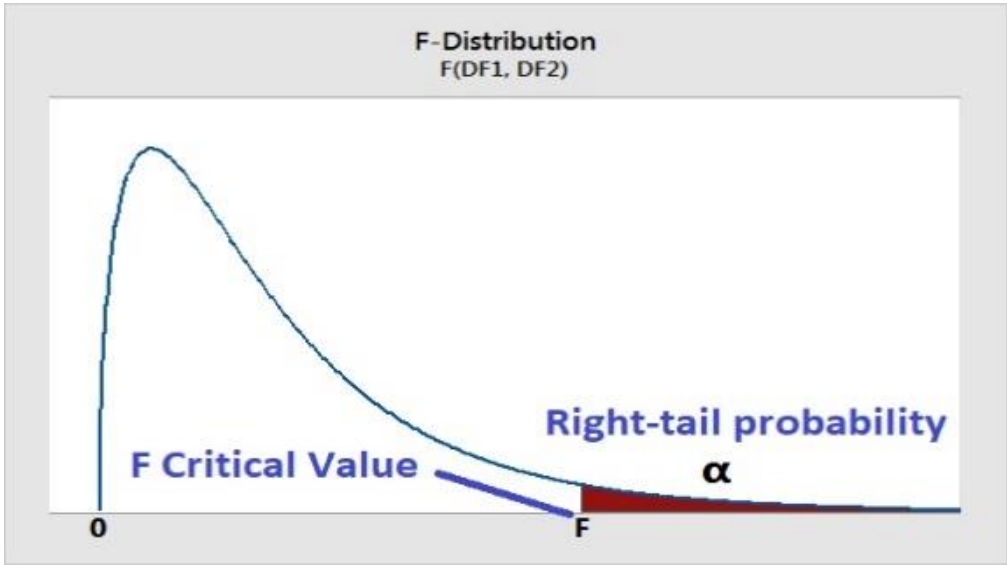
Null Hypothesis vs Alternative Hypothesis

$$H_0: \beta_1 = 0 \ \ vs \ \ H_1: O/W$$

Test Statistic

$$F_0 = \frac{MSR}{MSE} = \frac{ESS \ / \ df \ of \ ESS}{RSS \ / \ df \ of \ RSS} = \frac{ESS \ / 1}{RSS \ / \ n - (1+1)}$$

유의성 검정 | F-test



귀무가설을 Accept하면 모든 회귀계수들이 유의하지 않다고 결론 내림

$$Accept\ H_0\ if\ |F_0| < F_{1-\alpha/2,\,p,\,n-(p+1)}\ \ (Significance\ Level:\ \alpha)$$

단순선형회귀는 설명변수(X) 가 1개 뿐이기에, F-test 보다는 T-test 를 주로 이용 !

Confidence Interval vs Prediction Interval

선형회귀모델에서의 Predicton 값

① 개별 반응변수 자체에 대한 Prediction = $\hat{y}$

② 개별 반응변수의 Expectation 에 대한 Prediction = $\widehat{E[y|x]} = \hat{\mu}$

Ex) 반응변수(Y) = 키, 설명변수(X) = 몸무게인 선형회귀모델

①

나의 몸무게(x_0) 로

나의 키(y_0) 를 예측

②

나의 몸무게(x_0) 로

나와 몸무게가 비슷한 사람들의

평균 키($\mu_0 = E(y_0|x_0)$) 를 예측

Confidence Interval vs Prediction Interval

구간 추정을 위해 **추정량과 실제 값 간의 차에 대한 분포**를 알아야 함

구할 수 있는 Prediction 값은 $\hat{y}_0, \hat{\mu}_0$ 2개이기 때문에,

결과적으로 $\hat{\boldsymbol{y}}_0 - \boldsymbol{y}_0$ 와 $\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0$ 의 분포를 알 필요가 있음

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0,\ \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \quad \rightarrow \quad \hat{\boldsymbol{y}}_0 - \boldsymbol{y}_0 \sim ?$$

$$\mu_0 = E(y_0|x_0) = \beta_0 + \beta_1 x_0,\ \hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \quad \rightarrow \quad \hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0 \sim ?$$

Confidence Interval vs Prediction Interval

$\hat{\mu}_0 - \mu_0$ 의 분포

$$\hat{\mu}_0 - \mu_0 \sim N(0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{{S_{xx}}^2} \right)$$

Confidence Interval

$\hat{y}_0 - y_0$ 의 분포

$$\hat{y}_0 - y_0 \sim N(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{{S_{xx}}^2} \right)$$

Prediction Interval

y_0 는 모수가 아닌 개별 값이기 때문에

Confidence Interval 이라는 말 대신

Prediction Interval 이라는 용어를 사용

Confidence Interval vs Prediction Interval

$(1-\alpha)$X100% Prediction Interval(= PI) for y_0

개별 반응변수 값에 대한 추정 Interval

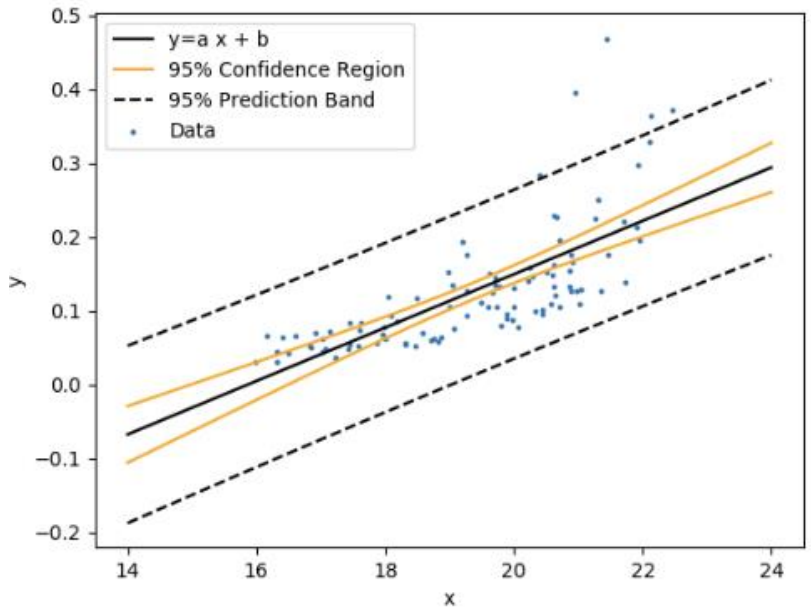
$$\hat{y}_0 - t_{\alpha/2,n-2}\, se(\hat{y}_0 - y_0) \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\, se(\hat{y}_0 - y_0)$$

$(1-\alpha)$X100% Confidence Interval(= CI) for μ_0

개별 반응변수 값의 평균에 대한 추정 Interval

$$\hat{y}_0 - t_{\alpha/2,n-2}\, se(\hat{y}_0 - y_0) \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\, se(\hat{y}_0 - y_0)$$

Confidence Interval vs Prediction Interval



일반적으로 CI보다

PI의 구간 폭이

더 넓은 형태를 보임

4

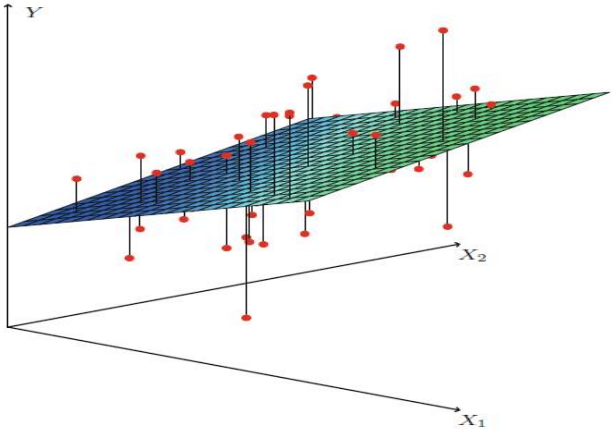
Multiple Linear Regression

4 Multiple Linear Regression

다중 선형회귀분석 | 정의

다중 선형회귀분석

하나의 반응변수(Y) 와 여러 개의 설명변수(X) 들 사이의
관계 규명을 위해 사용되는 선형회귀분석 방법



단순선형회귀모델보다 빈번히 사용되지만,
차원 수가 커짐에 따라
Overfitting 에 빠질 위험 존재

4 Multiple Linear Regression

다중 선형회귀식

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$
$$= E(Y_i|X) + \varepsilon_i \ (i = 1, \ldots, n)$$

- $\boldsymbol{X_{ji}}$: j 번째 독립변수 X의 i 번째 관측 값으로, 확률변수가 아닌 고정변수
- $\boldsymbol{Y_i}$: 종속변수 Y의 i 번째 관측 값으로, 확률변수
- $\boldsymbol{\beta_j}$: 우리가 추정해야 하는 모수이자 회귀계수 값 $(j = 1, \ldots, p)$
- $\boldsymbol{\varepsilon_i}$: i번째 관측 값에 의한 랜덤오차로, $iid\ N(0, \sigma^2)$ 의 분포를 따름

4 Multiple Linear Regression

다중 선형회귀식

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$
$$= E(Y_i|X) + \varepsilon_i \ (i = 1, \ldots, n)$$

주로 Matrix 를 이용해 Notation 을 작성

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

4 Multiple Linear Regression

다중 선형회귀식

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$
$$= E(Y_i|X) + \varepsilon_i \ (i = 1, \ldots, n)$$

Mean of the Model

$$\boldsymbol{E(Y_i|X)}$$
$$= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} = X\beta$$

평균적으로 해당 p 차원 평면에 데이터들이 존재함을 의미

Variance of the Model

$$\boldsymbol{V(Y_i|X)} = \sigma^2 \ (\because \ \beta_j, X_{ji} \text{ 는 상수})$$

데이터들의 분산은 일정함을 의미

다중 선형회귀모델의 해석

$\boldsymbol{\beta_0}$

$all\ X = 0$ 일 때 Y_j 의 예측 값

(즉, Intercept)

$\boldsymbol{\beta_j}$

다른 설명변수들($X_m,\ m \neq j$) 을 **고정**했을 때, X_j 값 변화에 따른 Y 의 평균 변화량

모수의 추정 | 최소제곱법

최소제곱법 (Ordinary Least Squares)

다중선형회귀에서도 OLS 방식을 이용해 모수(β) 를 추정

단순선형회귀 때와 마찬가지로 **회귀 기본가정 충족 시**,

Gauss-Markov Theorem 에 따라 **LSE 는 BLUE 가 되어**

불편성과 일치성을 만족하고 **MLE 와 같은 결과**를 가짐

4 Multiple Linear Regression

모수의 추정 | 최소제곱법

OLS 적용 시 풀어야 되는 **편미분식의 개수**가 $\boldsymbol{p+1}$ **개**로 늘어나고,

행렬 연산이 **이용**된다는 점에서 단순선형회귀와 차이점이 존재

[다중선형회귀에서 행렬연산을 통해 LSE 를 구하는 과정]

$$WTS)\ Minimize\ \ S(\beta) = (Y - X\beta)^T(Y - X\beta) = \varepsilon^T\varepsilon$$

$$S(\beta) = Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta$$

$$\frac{\partial}{\partial\beta}S(\beta) = -X^TY - X^TY + X^TX\hat{\beta} + X^TX\hat{\beta} = 0$$

$$2\,X^TX\hat{\beta} = 2\,X^TY$$

$$\therefore\ \ \hat{\beta} = (X^TX)^{-1}X^TY\,,\ \ if\ (X^TX)^{-1}\ exists$$

모수의 추정 | 최소제곱법

LSE 를 대입한 다중선형회귀 추정모델의 식

$$Y = X\hat{\beta} = HY$$

$\boldsymbol{H = X(X^TX)^{-1}X^T}$ 를 Hat Matrix라고 하며,

해당 행렬은 symmetric 하며 idempotent 하다는 성질을 가짐

(즉, $H^T = H \ \& \ HH = H$)

모수의 추정 | 최소제곱법

$\hat{\beta}$의 분포

$$E\left[\hat{\beta}\right] = E[(X^TX)^{-1}X^TY\ ] = \beta$$

$$V\left[\hat{\beta}\right] = V[(X^TX)^{-1}X^TY\ ] = \sigma^2\,(X^TX)^{-1}$$

$$\therefore\ \ \hat{\boldsymbol{\beta}}\ \sim\ \boldsymbol{N(\beta, \sigma^2(X^TX)^{-1})}$$

$Y|X$의 분포

$$E\left[\hat{\beta}\right] = E[\ X\beta + \varepsilon\ ] = X\beta$$

$$V[Y|X] = V[\ X\beta + \varepsilon\ ] = \sigma^2$$

$$\therefore\ \ \boldsymbol{Y|X\ \sim\ N(X\beta, \sigma^2)}$$

회귀식의 적합도 | 잔차

잔차 (Residuals, e_i)

오차항(ε) 에 대한 추정치

$$e = Y - \hat{Y} = Y - X\hat{\beta}$$

⋮

단순선형회귀에서와 마찬가지로

잔차의 제곱합을 $n -$ (변수개수(= p) + 1) 로 나눠

오차항의 분산(σ^2) 에 대한 추정치를 계산

$$\hat{\sigma}^2 = \frac{e^T e}{n-p-1} = \frac{RSS}{n-p-1}$$

회귀식의 적합도 | 제곱합

Sum of Squares

TSS (or SST)

Total Sum of Squares

y_i 값이 가지는 총 변동

$$TSS = (Y - \bar{Y})^T (Y - \bar{Y})$$

ESS (or SSR)

Regression Sum of Squares

회귀 직선이 설명하는 변동

$$ESS = (\hat{Y} - \bar{Y})^T (\hat{Y} - \bar{Y})$$

RSS (or SSE)

Residual Sum of Squares

회귀 직선이 설명 못하는 변동

$$RSS = (Y - \hat{Y})^T (Y - \hat{Y})$$

회귀식의 적합도 | 제곱합

다중선형회귀에서도 Y_i 값이 가지는 총 변동($\boldsymbol{TSS}$)은
회귀 직선이 설명하는 변동($\boldsymbol{ESS}$) 과
회귀 직선이 설명하지 못하는 변동($\boldsymbol{RSS}$)으로 분리 가능

$$\boldsymbol{TSS} = (Y-\bar{Y})^T(Y-\bar{Y}) = \left(Y-\hat{Y}+\hat{Y}-\bar{Y}\right)^T\left(Y-\hat{Y}+\hat{Y}-\bar{Y}\right)$$

$$= \left(Y-\hat{Y}\right)^T\left(Y-\hat{Y}\right) + \left(\hat{Y}-\bar{Y}\right)^T\left(\hat{Y}-\bar{Y}\right) + 2\left(Y-\hat{Y}\right)^T\left(\hat{Y}-\bar{Y}\right)$$

$$= ESS + RSS + 2\left(Y-\hat{Y}\right)^T\left(\hat{Y}-\bar{Y}\right) = \boldsymbol{ESS} + \boldsymbol{RSS}$$

$$\left(\because \left(\hat{Y}-\bar{Y}\right) = Y^T\left(H_J - H_J\right)Y = 0\right)$$

$$\therefore \quad \boldsymbol{TSS} = \boldsymbol{RSS} + \boldsymbol{ESS}$$

회귀식의 적합도 | 제곱합

반응변수(Y) 의 총 변동인 TSS 값은 모델의 종류와 관계없이 항상 일정

만약, 같은 Y 데이터에 대해 **2개의 설명변수(X_1, X_2) 를 이용한 회귀모델**과

3개의 설명변수(X_1, X_2, X_3) 를 이용한 회귀모델이 존재할 때,

다음과 같은 등식이 항상 성립하게 됨

$$TSS = ESS(X_1, X_2) + RSS(X_1, X_2) = ESS(X_1, X_2, X_3) + RSS(X_1, X_2, X_3)$$

회귀식의 적합도 | 수정된 결정계수

수정된 결정계수 (Adjusted R^2)

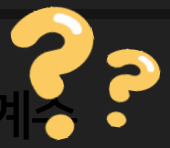
결정계수의 한계점을 보완한 것으로,

결정계수의 분자 분모의 값을 각각의 자유도로 나눠 정의함

$$Adjusted\_R^2 = \frac{ESS/(p-1)}{TSS/(n-1)} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

4 Multiple Linear Regression

회귀식의 적합도 | 수정된 결정계수



수정된 결정계수를 사용하는 이유

다중선형회귀에서는 **설명변수(X) 의 개수가 많아질수록**
반응변수의 변동을 더 많이 설명할 수 있어,

ESS 의 값은 계속 증가하고 RSS 값은 감소

설명변수 개수를 늘릴 수록 **결정계수(R^2) 의 값은 높아지지만,**
회귀모델이 복잡해져 변수들 간의 관계를 설명하기 어려워짐

유의성 검정

유의성 검정

다중선형회귀분석에서도 회귀계수들의 추정량에 대해 통계적 검정 가능

⋮

$\hat{\beta}$ 의 분포 형태

$$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1}), where\ \sigma^2(X^TX)^{-1} = \begin{bmatrix} V(\hat{\beta}_1) & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & V(\hat{\beta}_p) \end{bmatrix}$$

$\sigma^2(X^TX)^{-1}$ Matrix 의 diagonal element 들은

각각 순서대로 $\hat{\beta}_j \ \ (j = 1 \sim p)$ 의 분산 값에 해당

유의성 검정 | T-test

T-test

Null Hypothesis vs Alternative Hypothesis

$$H_0: \beta_j = 0 \ \ vs \ \ H_1: \beta_j \neq 0$$

Test Statistic

$$t_{\beta_j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

귀무가설을 Reject 하면 해당 계수 값은 유의하다고 결론 내림

$$Reject\ H_0\ if\ \left|t_{\beta_j}\right| > t_{1-\alpha/2,\, n-p-1}\ \ (Significance\ Level:\ \alpha)$$

유의성 검정 | F-test

F-test

다중선형회귀에서는 회귀계수에 대한 유의성 검정 진행 시,

항상 F-test 를 먼저 진행하고 그 **결과에 따라 t-test를 추가로 진행**할지 판단

Null Hypothesis vs Alternative Hypothesis

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \ vs \ H_1: O/W$$

Test Statistic

$$F_0 = \frac{MSR}{MSE} = \frac{ESS \ / \ df \ of \ ESS}{RSS \ / \ df \ of \ RSS} = \frac{ESS \ / p}{RSS \ / \ n-(p+1)}$$

귀무가설을 Accept하면 모든 회귀계수 값들이 유의하지 않다고 결론

$$Accept \ H_0 \ if \ |F_0| < F_{1-\alpha/2, p, n-(p+1)} \ (Significance \ Level: \ \alpha)$$

유의성 검정 | Partial F-test

Partial F-test

추가된 변수가 회귀식의 설명력을 유의미하게 증가시키는지 검정하는 방법

Null Hypothesis vs Alternative Hypothesis

$$H_0:\ \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0 \ \ vs \ \ H_1:\ O/W$$

Test Statistic

$$F_0 = \frac{ESS(FM) - ESS(RM)\ /\ p-q}{RSS(FM)\ /\ \ n-p-1} = \frac{\{R^2(FM) - R^2(RM)\}\ \ /\ p-q}{\{1-R^2(FM)\}\ /\ n-p-1}$$

$* \ FM = Full\ Regression\ model, RM = Reduced\ Regression\ model$

$* \ p = df\ of\ ESS(FM), q = df\ of\ ESS(RM), R^2 =$ 결정계수

유의성 검정 | Partial F-test

Partial F-test

추가된 변수가 회귀식의 설명력을 유의미하게 증가시키는지 검정하는 방법

Null Hypothesis vs Alternative Hypothesis

$$H_0: \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0 \ vs \ H_1: O/W$$

Test Statistic

계수들을 한번에 묶어서 유의성 검정을 진행하는 F-test 의 특성 상,

중요도 확인 대상인 설명변수들의 인덱스는 항상 연속적이어야 함

$* FM = Full\ Regression\ model, RM = Reduced\ Regression\ model$

$* p = df\ of\ ESS(FM), q = df\ of\ ESS(RM), R^2 =$ 결정계수

유의성 검정 | Partial F-test

Partial F-test

추가된 변수가 회귀식의 설명력을 유의미하게 증가시키는지 검정하는 방법

Null Hypothesis vs Alternative Hypothesis

$H_0: \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0 \ vs \ H_1: O/W$

그런데 Partial F-test 의 Test Statistic 에서

$F_0 = \frac{ESS}{\quad} \quad \frac{)\} \ / \ p-q}{-p-1}$

$ESS(FM) - ESS(RM)$ 식이 쓰인 이유는 무엇일까?

예시를 통해 알아보자 !

$* \ FM = Full \ Regression \ model, RM = Reduced \ Regression \ model$

$* \ p = df \ of \ ESS(FM), q = df \ of \ ESS(RM), R^2 =$ 결정계수

유의성 검정 | Partial F-test

일반적으로 선형회귀 모델에서 설명변수(X) 에 대한 변수 중요도를 확인할 때,
해당 설명변수의 $Type\ \mathrm{I}\ \ ESS$ 를 계산함

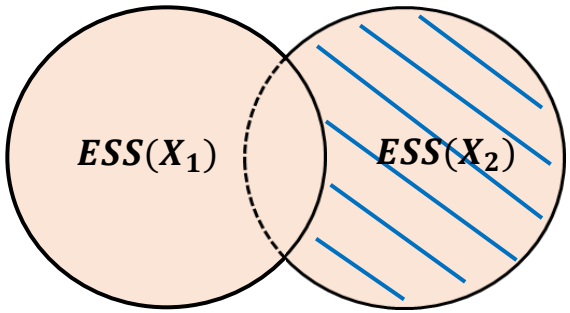
X_1, X_2 를 이용하여 다중선형회귀모델을 만든다고 가정 시,

$$Type\ \mathrm{I}\ \ ESS\ of\ X_2 = ESS(X_2|X_1) = ESS(X_1, X_2) - ESS(X_1)$$

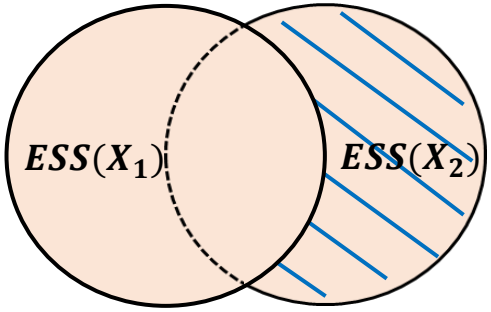
유의성 검정 | Partial F-test

$\mathrm{ESS}(X_2|X_1)$, $\mathrm{ESS}(X_1, X_2)$, $\mathrm{ESS}(X_1)$ 값들의 관계 확인

①



②



Blue Region 이 차지하는 면적인

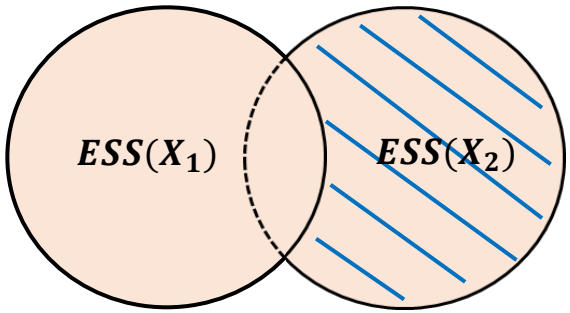
$\mathrm{ESS}(\boldsymbol{X_2}|\boldsymbol{X_1}) = \mathrm{ESS}(\boldsymbol{X_1}, \boldsymbol{X_2}) - \mathrm{ESS}(\boldsymbol{X_1})$ 의 값이 클수록

설명변수 $\boldsymbol{X_2}$ 는 회귀모델에서 중요한 역할을 함

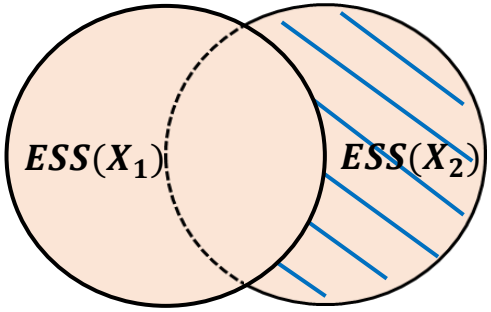
유의성 검정 | Partial F-test

$\mathrm{ESS}(X_2|X_1)$, $\mathrm{ESS}(X_1, X_2)$, $\mathrm{ESS}(X_1)$ 값들의 관계 확인

①



②



여기서 주의할 점은,

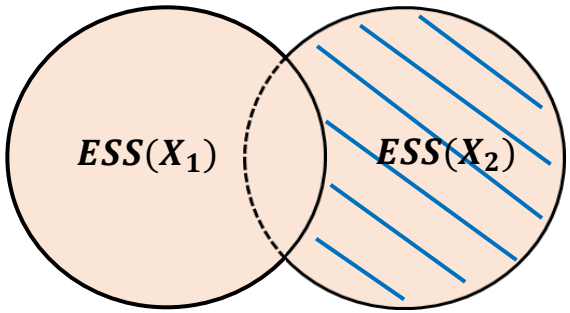
단독적인 것이 아닌 X_1 이 given 된 상황에서

X_2 변수가 중요한 역할을 수행한다고 봐야 함

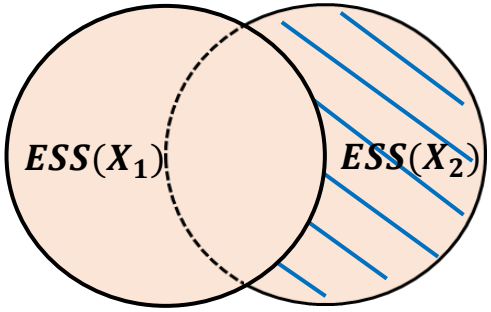
유의성 검정 | Partial F-test

$\mathrm{ESS}(X_2|X_1)$, $\mathrm{ESS}(X_1, X_2)$, $\mathrm{ESS}(X_1)$ 값들의 관계 확인

①



②



두 변수 간의 Correlation 정도가 작다면,

교집합 부분의 면적이 작아져 $\mathrm{ESS}(X_2|X_1)$ 의 값이 커지게 됨

① 에서가 ② 에서보다, 회귀모델 내의 X_2 변수 중요도가 더 높음

유의성 검정 | Partial F-test

반대로 ESS($X_2|X_1$) 의 값이 작다면

X_2 변수가 중요한 역할을 수행하지 못한다는 것을 의미

이 경우 $\boldsymbol{X_1, X_2}$ 가 모두 들어있는 **Full Model 을 이용할 필요 없이,**

$\boldsymbol{X_1}$ **하나로만 이루어진 Reduced Model** 로도

주어진 데이터를 충분히 설명 가능 !

4 Multiple Linear Regression

유의성 검정 | Partial F-test

반대로 ESS($X_2|X_1$) 의 값이 작다면

X_2 변수가 중요한 설명을 하지 못한다는 것을 의미

이러한 관계성 때문에,

Partial F-test 는 Test Statistic 의 분자 부분에

$\boldsymbol{ESS(FM|RM) = ESS(FM) - ESS(RM)}$ 식을 반영하여 검정을 수행

이 경우 $\boldsymbol{X_1, X_2}$ 가 모두 들어있는 Full Model 을 이용할 필요 없이,

$\boldsymbol{X_1}$ 하나로만 이루어진 Reduced Model 로도

주어진 데이터를 충분히 설명 가능 !

Confidence Interval vs Prediction Interval

$\hat{\mu}_0 - \mu_0$ 의 분포

$$\hat{\mu}_0 - \mu_0 \sim N(0, \sigma^2 x_0{}^T (X^T X)^{-1} x_0)$$

$$x_0 = (1, x_{01}, x_{02}, \dots, x_{0p})^T$$

Confidence Interval

$\hat{y}_0 - y_0$ 의 분포

$$\hat{\mu}_0 - \mu_0$$

$$\sim N(0, \sigma^2 (1 + x_0{}^T (X^T X)^{-1} x_0)$$

Prediction Interval

단순선형회귀 때와 마찬가지로

y_0 은 모수가 아닌 개별 값이기 때문에

Prediction Interval 이라는 용어를 사용

Confidence Interval vs Prediction Interval

$(1-\alpha)$X100% Prediction Interval(= PI) for y_0

개별 반응변수 값에 대한 추정 Interval

$$\hat{y}_0 - t_{\alpha/2,\,n-p-1}\, se(\hat{y}_0 - y_0) \;\le\; y_0 \;\le\; \hat{y}_0 + t_{\alpha/2,\,n-p-1}\, se(\hat{y}_0 - y_0)$$

$(1-\alpha)$X100% Confidence Interval(= CI) for μ_0

개별 반응변수 값의 평균에 대한 추정 Interval

$$\hat{\mu}_0 - t_{\alpha/2,\,n-p-1}\, se(\hat{\mu}_0 - \mu_0) \;\le\; \mu_0 \;\le\; \hat{\mu}_0 + t_{\alpha/2,\,n-p-1}\, se(\hat{\mu}_0 - \mu_0)$$

가변수의 정의

가변수 (Dummy Variable)

오직 두 가지 값 만을 갖는 변수

가변수의 값은 범주를 나타내기에 수치적으로 가지는 의미는 없음

일반적으로, k개의 범주가 존재하면 $k-1$개의 가변수가 필요함

가변수의 정의

가변수 (Dummy Variable)

오직 두 가지 값 만을 갖는 변수

가변수의 값은 범주를 나타내기에 수치적으로 가지는 의미는 없음

일반적 **회귀모형에 가변수를 많이 포함하게 될 때** 필요함

Design Matrix X가 Full rank가 되지 못하므로

OLS 과정에서 **회귀계수 추정이 불안정**해짐

가변수의 생성

Ex) 2개의 범주 값이 존재하는 경우

광고매체가 만약 '뉴스, 방송'이라는 **두 종류**의 범주를 가진다면,

아래와 같이 질적변수들을 생성해낼 수 있음

$$Z = \begin{cases} 1\ (\text{광고매체} = \text{뉴스}) \\ 0\ (\text{광고매체} = \text{방송}) \end{cases}$$

가변수의 생성

Ex) 2개의 범주가 존재하는 경우

1이 0보다 크다고 해서,

광고매체가 뉴스일 경우가 방송일 경우보다

더 의미가 있다고 볼 수는 없음

$$z = \begin{cases} 0\ (\text{광고매체} = \text{방송}) \end{cases}$$

가변수의 생성

Ex) 3개의 범주 값이 존재하는 경우

광고매체가 만약 '뉴스, 방송, 라디오' 라는 **세 종류**의 범주를 가진다면,

아래와 같이 질적변수들을 생성해낼 수 있음

$$Z_1 = \begin{cases} 1\ (\text{광고매체} = \text{뉴스}) \\ 0\ (\text{광고매체} \neq \text{뉴스}) \end{cases}, \quad Z_2 = \begin{cases} 1\ (\text{광고매체} = \text{방송}) \\ 0\ (\text{광고매체} \neq \text{방송}) \end{cases}$$

4 Multiple Linear Regression

가변수를 이용한 회귀분석

설명변수(X)들이 **질적변수**와 함께 주어지는 경우

가변수를 이용하여 회귀모델 작성

질적변수 : '성별, 학년' 등 자료가 속하는 범주(category)들을 나타내는 값을 갖는 변수

① 평균 등의 측도를 알아보는 것은 큰 의미가 없음

② 결과 해석에 주의가 필요함

③ 질적변수의 특성에 따른 해석이 요구됨

4 Multiple Linear Regression

가변수를 이용한 회귀분석

설명변수(X)들이 **질적변수**와 함께 주어지는 경우

가변수를 이용하여 회귀모델 작성

질적변수 : '성별, 학년' 등 자료가 속하는 범주(category)들을 나타내는 값을 갖는 변수

① 평균 등의 측도를 알아보는 것은 큰 의미가 없음

② 결과 해석에 주의가 필요함

③ 질적변수의 특성에 따른 해석이 요구됨

가변수를 이용한 회귀분석

Ex) 3개의 범주 값이 존재하는 경우

매출액과 광고비 간의 관계에 대한 회귀분석 + **광고매체의 효과** 분석 목적

광고매체는 '뉴스, 방송, 라디오' 라는 **세 종류**의 범주를 가짐

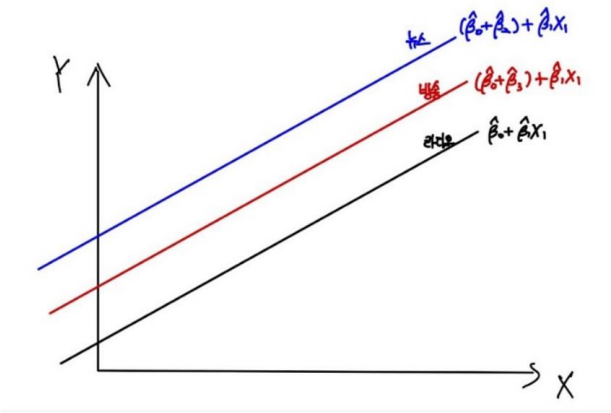
$$Z_1 = \begin{cases} 1\ (\text{광고매체} = \text{뉴스}) \\ 0\ (\text{광고매체} \neq \text{뉴스}) \end{cases}, \quad Z_2 = \begin{cases} 1\ (\text{광고매체} = \text{방송}) \\ 0\ (\text{광고매체} \neq \text{방송}) \end{cases}$$

$$X_1 = \text{광고비}$$

4 Multiple Linear Regression

가변수를 이용한 회귀분석

Ex) 3개의 범주 값이 존재하는 경우



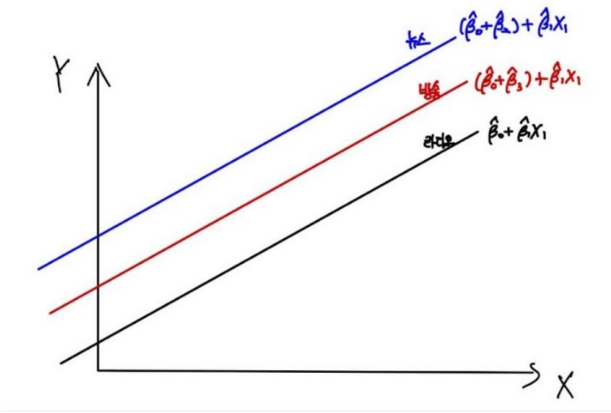
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 Z_1 + \beta_3 Z_2$$

$$Y = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1 \text{ (광고매체 = 뉴스)} \\ (\beta_0 + \beta_3) + \beta_1 X_1 \text{ (광고매체 = 방송)} \\ \beta_0 + \beta_1 X_1 \text{ (광고매체 = 라디오)} \end{cases}$$

4 Multiple Linear Regression

가변수를 이용한 회귀분석

Ex) 3개의 범주 값이 존재하는 경우



모델 적합 결과,

'뉴스 > 방송 > 라디오'의 대소 관계대로

매출액 평균값($E(Y|X)$) 분포의 차이를 확인

만약 회귀 계수에 대한 t-test 결과 $\hat{\beta}_2$ & $\hat{\beta}_3$ 가

모두 유의하지 않다는 결론이 나왔다면

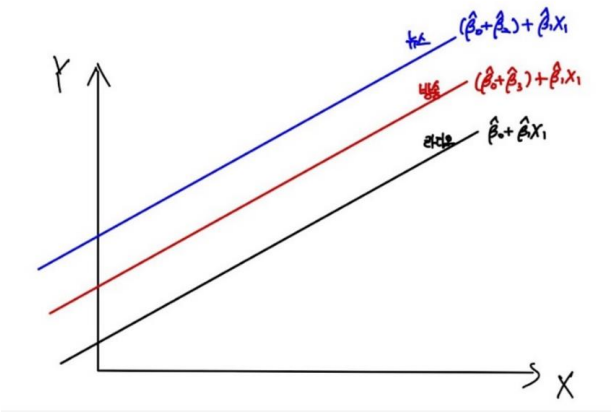
'광고매체의 효과'는 큰 영향을 미치지 못한다고

판단 내릴 수 있음

4 Multiple Linear Regression

가변수를 이용한 회귀분석

Ex) 3개의 범주 값이 존재하는 경우



'광고 매체별 광고비에 따른 매출액의 증가 정도가 **모두 동일**하다' 는 가정이 전제됨

교호작용을 고려한 선형회귀 모형을 통해 해당 가정을 검정!

교호작용을 고려한 선형회귀 모델

교호작용 (Interaction Effect)

한 요인의 효과가 다른 요인의 수준에 **의존**하는 것

교호작용을 고려한 선형회귀모델에는

양적변수와 질적변수 간의 곱 형태로 이루어진

Interaction term을 설명변수 부분에 추가해줘야 함

교호작용을 고려한 선형회귀 모델

Ex) 질적 설명변수와 교호작용을 고려할 때

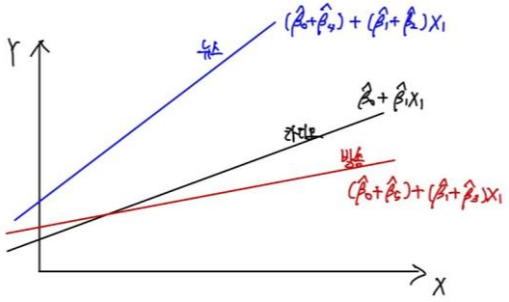
$$Z_1 = \begin{cases} 1\ (\text{광고매체} = \text{뉴스}) \\ 0\ (\text{광고매체} \neq \text{뉴스}) \end{cases}, Z_2 = \begin{cases} 1\ (\text{광고매체} = \text{방송}) \\ 0\ (\text{광고매체} \neq \text{방송}) \end{cases}$$

$$X_1 = \text{광고비},$$

$$X_2 = X_1 Z_1\ , X_3 = X_1 Z_2 \text{ (Interaction term)}$$

교호작용을 고려한 선형회귀 모델

Ex) 질적 설명변수와 교호작용을 고려할 때



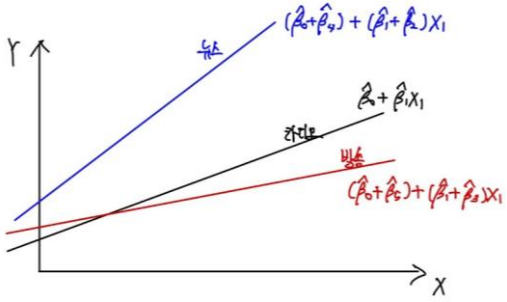
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 Z_1 + \beta_5 Z_2$$

$$Y = \begin{cases} (\beta_0 + \beta_4) + (\beta_1 + \beta_2) X_1 \text{ (광고매체 = 뉴스)} \\ (\beta_0 + \beta_5) + (\beta_1 + \beta_3) X_1 \text{ (광고매체 = 방송)} \\ \beta_0 + \beta_1 X_1 \text{ (광고매체 = 라디오)} \end{cases}$$

4 Multiple Linear Regression

교호작용을 고려한 선형회귀 모델

Ex) 질적 설명변수와 교호작용을 고려할 때



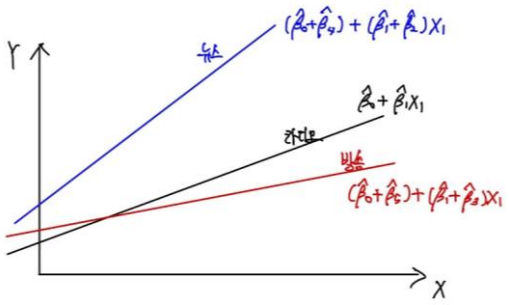
모델 적합 결과

$\hat{\beta}_2 > 0$ & $\hat{\beta}_3 < 0$ 라는 결과가 나왔다면,

'뉴스 > 라디오 > 방송' 순서대로 광고비(X_1) 증가에 따른 매출액(Y) 증가분이 높다고 해석할 수 있음

교호작용을 고려한 선형회귀 모델

Ex) 질적 설명변수와 교호작용을 고려할 때



만약 회귀 계수에 대한 t-test 결과 $\hat{\beta}_2$ & $\hat{\beta}_3$ 가
모두 유의하지 않다는 결론이 나왔다면,
해당 모델은 **교호작용이 없다**고 판단 내릴 수 있음

5

Discussion of Influential Observations

5 Discussion of Influential Observations

잔차의 문제점

회귀식 기본가정에 따라 오차항(ε) 은 iid 한 정규분포를 따르지만,
오차의 추정량인 **잔차(e) 는 등분산성 및 독립성을 만족하지 않음**

표준화된 잔차 개념을 도입해야 할 필요성 인식

내표준화 잔차의 정의

내표준화 잔차 (Internally Studentized Residual)

잔차의 비등분산성 문제를 해결하기 위한 표준화된 잔차

⋮

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

내표준화 잔차의 특징

① 등분산성을 만족 (Unit Variance)하여 데이터에서 **추세와 거리가 먼 관측치**를 식별하는 데 용이함

$$V(r_i) = \frac{1}{\left(\hat{\sigma}\sqrt{1-h_{ii}}\right)^2} V(e_i) = 1$$

② 오차항에 대한 가정을 만족 시 **근사적으로 표준정규분포**를 따름

$$r_i \approx \mathrm{N}(0,1)$$

5 Discussion of Influential Observations

내표준화 잔차의 특징

근사적으로 표준정규분포를 따르지 않는다면,

실제 오차항의 가정은 위배됐다고 판단해볼 수 있음

따라서, 일반적으로 **오차항의 가정에 대해 진단**할 때에는

내표준화 잔차를 활용하게 됨

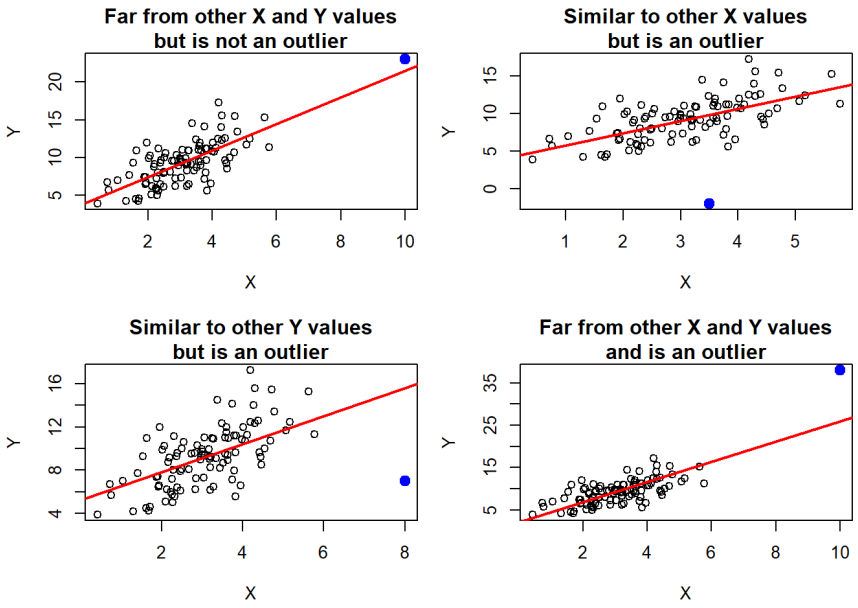
5 Discussion of Influential Observations

이상치

이상치 (Outlier)

데이터의 추세와는 거리가 먼 관측치

내표준화 **잔차 절댓값($= |r_i|$) 이 2 or 3 보다 클 때**, 이를 이상치로 간주



5 Discussion of Influential Observations

지렛값

지렛값 (Leverage Value)

Hat Matrix($\boldsymbol{H}$)의 $\boldsymbol{i}$번째 대각원소($\boldsymbol{h_{ii}}$)값

$$H = X(X^TX)^{-1}X^T = \begin{bmatrix} h_{11} & \cdots & h_{1p} \\ \vdots & \ddots & \vdots \\ h_{p1} & \cdots & h_{pp} \end{bmatrix}$$

Rule of thumb 를 적용해,
특정 $\boldsymbol{k}$번째 관측치의 지렛값($= \boldsymbol{h_{kk}}$) 이
평균 지렛값의 2배 ($= \boldsymbol{2\,tr(H)/n}$) 보다 크면,
해당 관측치를 High Leverage Point로 간주

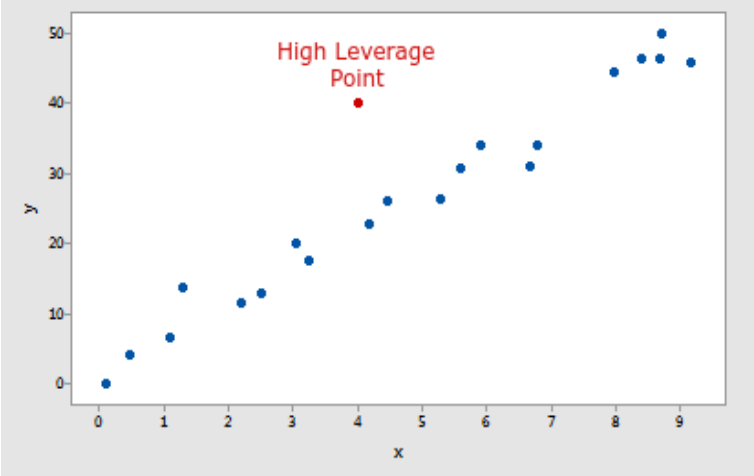
5 Discussion of Influential Observations

지렛값

지렛값 (Leverage Value)

Hat Matrix($\boldsymbol{H}$)의 $\boldsymbol{i}$번째 대각원소($\boldsymbol{h_{ii}}$)값

$$H = X(X^TX)^{-1}X^T = \begin{bmatrix} h_{11} & \cdots & h_{1p} \\ \vdots & \ddots & \vdots \\ h_{p1} & \cdots & h_{pp} \end{bmatrix}$$



Rule of thumb 를 적용해,

특정 $\boldsymbol{k}$번째 관측치의 지렛값($= \boldsymbol{h_{kk}}$) 이

평균 지렛값의 2배 ($= \boldsymbol{2\,tr(H)/n}$) 보다 크면,

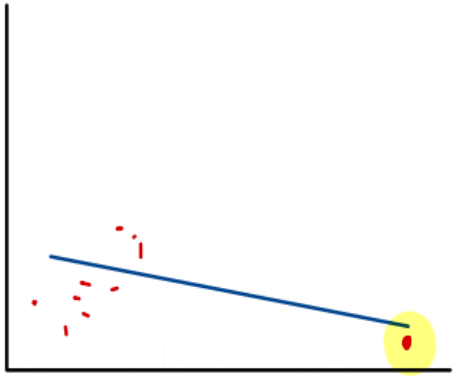
해당 관측치를 High Leverage Point로 간주

5 Discussion of Influential Observations

영향점

영향점 (Influential Observations)

회귀 직선 모양에 **큰 변화를 야기시키는 관측치로,**
일반적으로 이상치이거나 높은 지렛값을 가짐



영향점은 회귀 직선을 영향점 쪽으로 끌어오고,
나머지 데이터로부터 멀어지게 하는 경향을 지님

영향점

Cook's distance

영향점을 판단하는 한 측도로

표준화 잔차(r_i)와 지렛값(h_{ii})의 조합으로 표현되며,

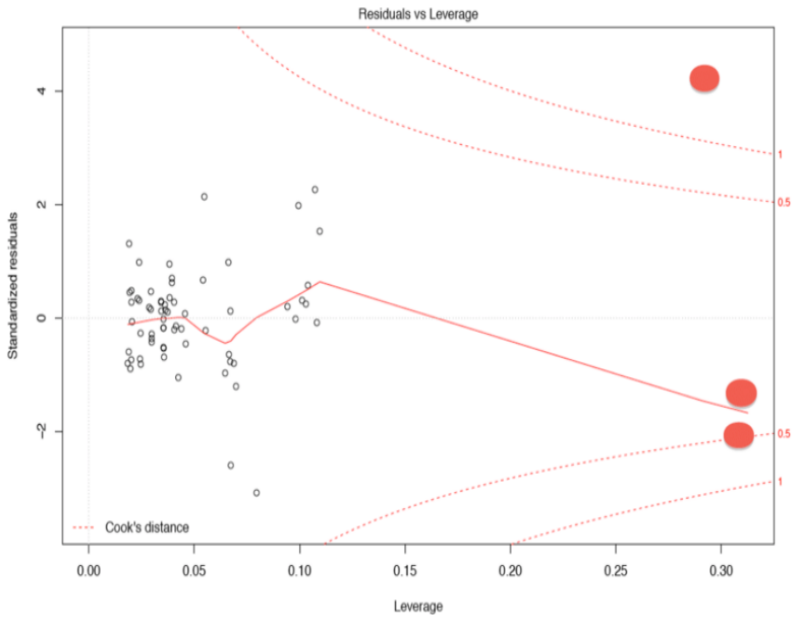
표준화 잔차 값이 커질수록 혹은 지렛값이 커질 수록 C_i값이 커짐

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

보통 C_i**의 값이 1보다 크면** 해당 관측치를 영향점이라고 판단

5 Discussion of Influential Observations

영향점



Cook's distance 계산 결과

① 우측 상단 Red point

Outlier임과 동시에 High Leverage Point

→ 영향점으로 간주

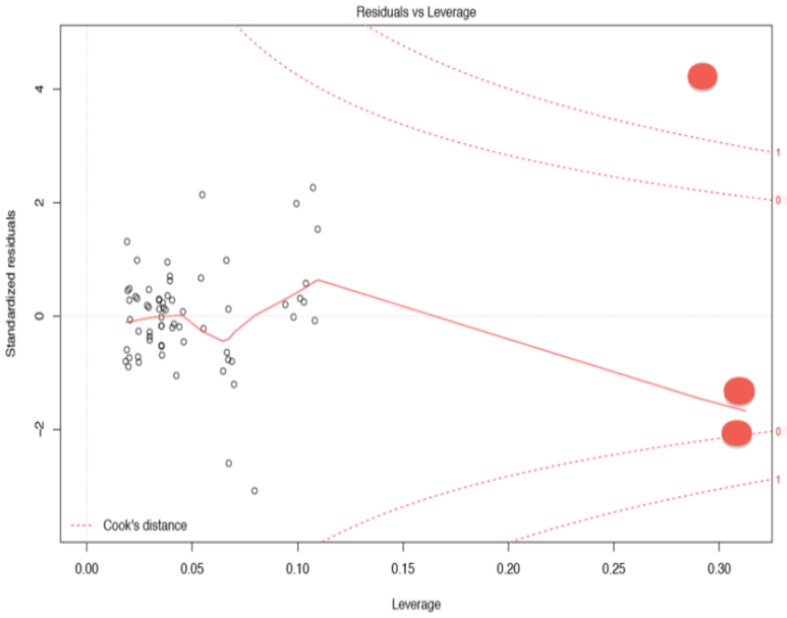
② 우측 하단 Red point

High Leverage Point이지만 Outlier가 아님

→ 영향점이 아니라고 간주

5 Discussion of Influential Observations

영향점



Cook's distance 계산 결과

영향점을 제거하는 것도 좋은 방법이지만,

영향점도 유의미한 데이터일 가능성도 있기에

이상치에 강건한(Robust) 회귀 모델을

적용해보는 것도 하나의 방법!

High Leverage Point이지만 Outlier가 아님

→ 영향점이 아니라고 간주

L1 Loss & L2 Loss

L1 Loss & L2 Loss

손실함수의 종류들로서, 목적함수 식의 구조에 따라 그 종류가 구분됨

L1 Loss

Least Absolute Error(LAE)

$$L1\ Loss = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

절댓값 차를 구한 뒤

이를 합하는 방식을 적용

L2 Loss

Least Squares Error(LSE)

$$L2\ Loss = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

거리 제곱 차의 합에

루트를 씌우는 방식을 적용

5 Discussion of Influential Observations

L1 Loss & L2 Loss | 알고리즘 적용 시 특징 비교

분류	L1 Loss	L2 Loss
강건함 정도	**높음**	낮음
초기 조건에 대한 민감도	높음	**낮음**
해의 개수	하나 이상의 해를 가질 수 있음	항상 유일한 해를 가짐

영향점 발생 시 이용하는 회귀모델 들은 회귀계수 추정에 있어
OLS 의 목적함수(= L2 Loss) 대신 **L1 Loss 를 이용**하기도 하고,
L1 Loss 와 L2 Loss 를 조합 시킨 방법을 적용하기도 한다

저항 회귀

저항 통계 (Resistant Statistics)

특정 통계량이 비정상적 관측치에 얼마나 덜 영향을 받는지를 수치화한 개념

중앙값(Median), 분위 수(Quantile) 등

평균, 분산 등은 이상치에 따라 값이 변동되기 때문에 저항 통계에 속한다고 볼 수 없음

저항 회귀

저항 회귀 (Resistant Regression)

저항 통계량들을 회귀계수 추정 과정에 이용하는 회귀분석

① 상대적으로 영향점의 영향을 덜 받음

② 계수 추정과정에 있어 영향점과 같은 비정상적인 데이터를 완전히 배제

저항 회귀

계수 추정 방법은 흔히 아래와 같이 2가지로 분류해볼 수 있음

방법 ① : 특정 Quantile에 해당되는 잔차의 제곱을 최소화 하는 경우

방법 ② : 오름차순 정렬된 잔차들에 대해, 특정 순위까지 합을 최소화 하는 경우

⋮

위 방법들은 Closed form 형태의 Solution 이 도출되지 않음

따라서 대부분 'Iterated Reweighted Least Squares(IRLS)' 등과 같은

최적화 알고리즘을 적용하여 Solution 을 도출함

저항 회귀

Regression with Least Trimmed Sqaures (LTS)

잔차가 큰 데이터 포인트를 일부 제거한 후 나머지 데이터에 대해
최소제곱법으로 회귀 모델을 피팅하는 방법

① Loss 종류 : L2 Loss, 제약 조건이 존재

② 목적 : 일부 데이터 포인트를 제외하여 이상치의 영향을 감소

③ 추정량($= \hat{\beta}_{LTS}$): $\underset{\beta}{\arg min} \sum_{i=1}^{h} (y_i - {x_i}^T \beta)^2$

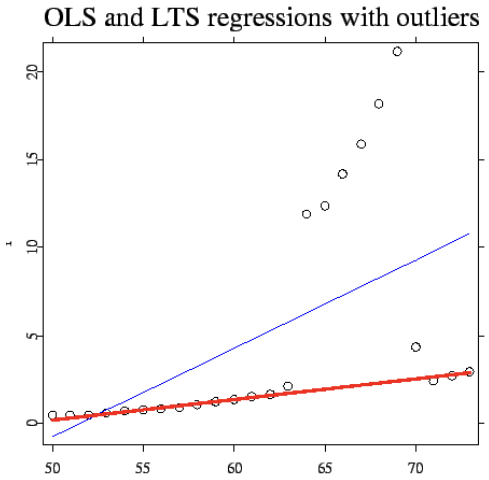
$(where\ h = \#\ of\ samples\ used\ to\ calculating\ the\ loss)$

5 Discussion of Influential Observations

저항 회귀

Regression with Least Trimmed Sqaures (LTS)

잔차가 큰 데이터 포인트를 일부 제거한 후 나머지 데이터에 대해
최소제곱법으로 회귀 모델을 피팅하는 방법



Plot 상단에 위치하는 Outlier 들에 대해,

LTS Regression 이 OLS Regression 보다

영향을 덜 받음을 확인 가능

5 Discussion of Influential Observations

강건 회귀

강건 통계 (Robust Statistics)

영향점과 같은 비정상적인 데이터 분포에도 불구하고,
일관되며 신뢰할 수 있는 결과를 제공하기 위해 설계된 통계적 방법론

영향점 및 노이즈의 영향을 최소화하도록 고안된 다양한 통계 기법과 절차를 포함하는 개념

강건 회귀

강건 회귀 (Robust Regression)

위와 같은 방법론(강건 통계)들을 회귀계수 추정 과정에 이용하는 회귀

계수 추정 과정에 있어 영향점과 같은 비정상적인 데이터의 영향을
감소시키는 것이지 **완전히 배제하는 것은 아니라는 점**

강건 회귀

강건 회귀의 방법론에 대한 대표적인 예시

① Quantile Loss ② M-estimator

⋮

위 방법들 또한 Closed form 형태의 Solution 이 도출되지 않음

따라서 최적화 알고리즘을 적용하여 Solution 을 도출함

강건 회귀 | 분위수 회귀

분위수 회귀(Quantile Regression)

분위수(= v) 에 따른 Quantile Loss 를 최소화하여, 회귀계수를 추정하는 모델

분위수(= v) = Percentile (= α) x 샘플 수

① Loss 종류: L1 Loss

② 목적: Y 변수의 평균이 아닌, **분위수(= y_v) 를 예측**

③ 추정량(= $\hat{\beta}_Q$): $\underset{\beta}{\arg min} \sum_{i=1}^{n} \rho_\alpha(y_i - {x_i}^T\beta)$

Where $\rho_\alpha(y_i - {x_i}^T\beta) = \begin{cases} (\alpha - 1)(y_i - {x_i}^T\beta) & when\ \ y_i - {x_i}^T\beta < 0 \\ \alpha(y_i - {x_i}^T\beta) & when\ \ y_i - {x_i}^T\beta \geq 0 \end{cases}$

이를 Quantile Loss 라고 부름!

강건 회귀 | 분위수 회귀

분위수 회귀(Quantile Regression)

분위수(= v) 에 따른 Quantile Loss 를 최소화하여, 회귀계수를 추정하는 모델

분위수(= v) = Percentile (= α) x 샘플 수

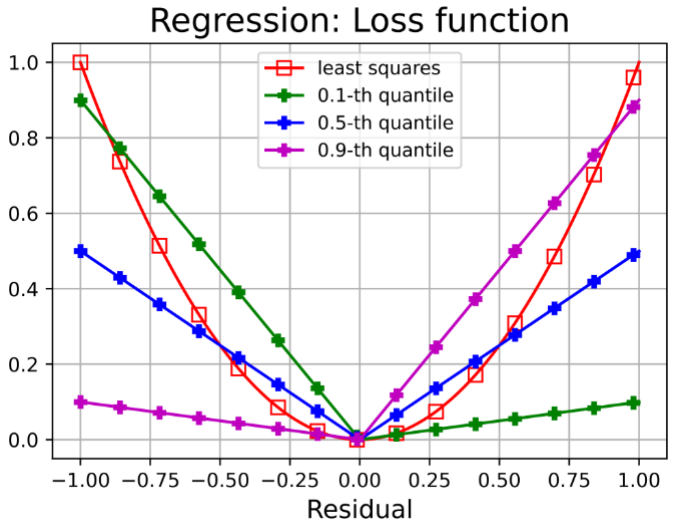
$$\rho_\alpha(y_i - x_i{}^T\beta) = \begin{cases} (\alpha - 1)(y_i - x_i{}^T\beta) & when\ \ y_i - x_i{}^T\beta < 0 \\ \alpha(y_i - x_i{}^T\beta) & when\ \ y_i - x_i{}^T\beta \geq 0 \end{cases}$$

Quantile Loss 는 지정한 Percentile(α) 값에 기반해,

$y_i - x_i{}^T\beta$ 값 부호에 따라 서로 다른 가중치를 부여하여

y_i 의 실제분위수(= y_{iv}) 와 추정분위수(= $x_i{}^T\hat{\beta}_v$) 간의 차이를 최소화

강건 회귀 | 분위수 회귀



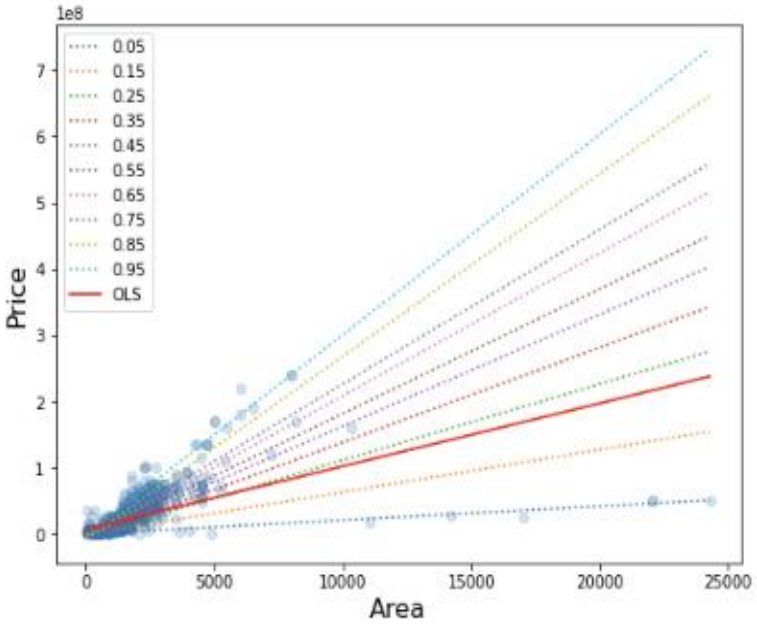
보편적으로 0.5 분위수(α = 0.5) 를 많이 이용하고,

이 경우 Quantile Loss 는 Least Absolute Error(LAE) 와 같게 됨

$$\arg\min_{\beta} \sum_{i=1}^{n} 0.5\,|y_i - {x_i}^T\beta| \cong \arg\min_{\beta} \sum_{i=1}^{n} |y_i - {x_i}^T\beta|$$

5 Discussion of Influential Observations

강건 회귀 | 분위수 회귀



OLS 를 적용한 회귀 직선보다,
Quantile Regression 들이 Outlier 의
영향을 덜 받고 데이터의 전반적인 추세를
잘 따라 감을 확인 가능!

강건 회귀 | Huber Loss를 이용한 회귀분석

Regression with Huber Loss

Huber Loss 를 이용하여 이상치의 영향을 줄이는 회귀모델

① Loss 종류 : L1 Loss 와 L2 Loss 를 혼합한 형태

② 추정량 $(= \hat{\beta}_M)$: $\underset{\beta}{\arg min} \sum_{i=1}^{n} \rho_\delta(\epsilon_i) = \underset{\beta}{\arg min} \sum_{i=1}^{n} \rho_\delta(y_i - {x_i}^T \beta)$

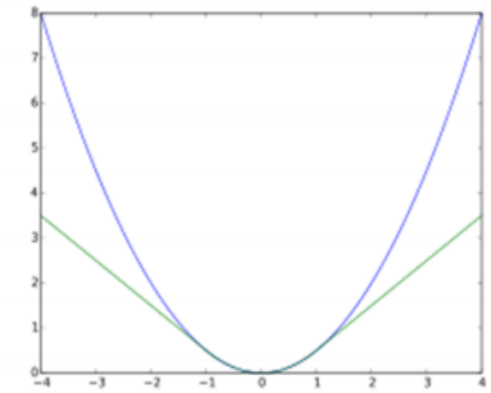
$$\rho_\delta(a) = \begin{cases} \frac{1}{2}a^2 & (|a| \le \delta) \\ \delta\left(|a| - \frac{1}{2}\delta\right) & (o/w) \end{cases}, \quad a\text{: } Residual,\ \delta\text{: } Hyper\ parameter$$

5 Discussion of Influential Observations

강건 회귀 | Huber Loss를 이용한 회귀분석

Regression with Huber Loss

Huber Loss 를 이용하여 이상치의 영향을 줄이는 회귀모델



목적함수 비교

초록색 : Huber's M-estimation

파란색 : OLS

잔차의 절대값이 **특정 상수 값(δ) 이하**면

OLS 의 목적함수와 동일하고,

그 **이상**이 되면 **일차식의 형태의 Penalty**를

적용하여 회귀 계수를 추정

다음 주 예고

1.Standard Regression Analysis

2. Residual Plot

3. Linearity

4. Homoscedasticity

다음 주 예고

5.Normality

6. Independence

7. Multicolinearity

8. Endogeity

감사합니다