회귀분석팀

6팀

조수미
김민지
손재민
박윤아
조웅빈

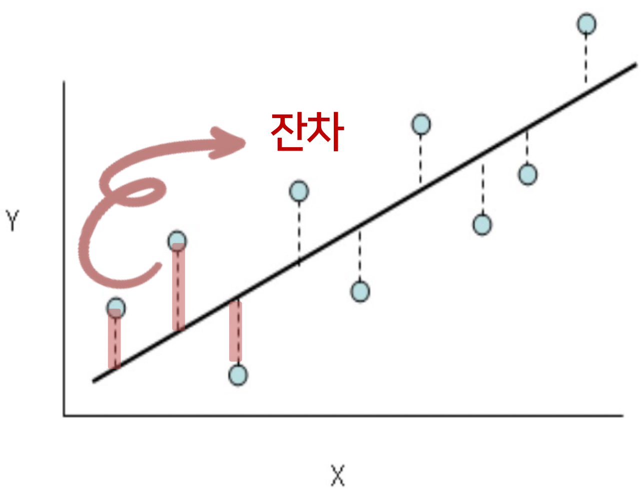
CONTENTS

1 회귀 기본 가정

모델 가정의 목표

모델의 목표

잔차가 평균인 0으로 회귀하는 정확한 모델을 만드는 것



그러나, 추정한 모델과 실제 데이터 사이에는 **오차**가 발생

가정이 잘 지켜지지 않으면,

모델이 불안정해지고 **설명력**과 **예측력**을 잃음

1 회귀 기본 가정

선형회귀분석 가정

회귀식

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

① **선형성** : 설명변수와 반응변수의 관계는 선형

② **오차의 정규성** : 오차항은 정규분포를 따름

③ **오차의 독립성** : 오차항은 서로 독립

④ **오차의 등분산성** : 오차항의 분산은 상수

기본 가정 진단

기본 가정 진단

회귀분석의 기본 가정을 진단하기 위해 크게 **두 가지 방법**을 동원

시각적 방법

Residual Plot

가설 검정

Statistical Hyphothesis Test

시각적 방법

잔차 플랏 Residual Plot

잔차 분포를 통해 경험적 판단에 근거한 회귀진단이 가능

R의 plot() 함수를 통해 잔차의 분포를 쉽게 나타낼 수 있음

- Residuals vs Fitted
- Normal QQ Plot
- Scale-Location
- Residuals vs Leverage

3 선형성 진단과 처방

선형성 가정

선형성 가정

반응변수 Y가 설명변수 X의 **선형결합**으로 이루어진다는 가정

단순선형회귀, 다중선형회귀 모두 선형성 가정에서 출발한 모델

↓

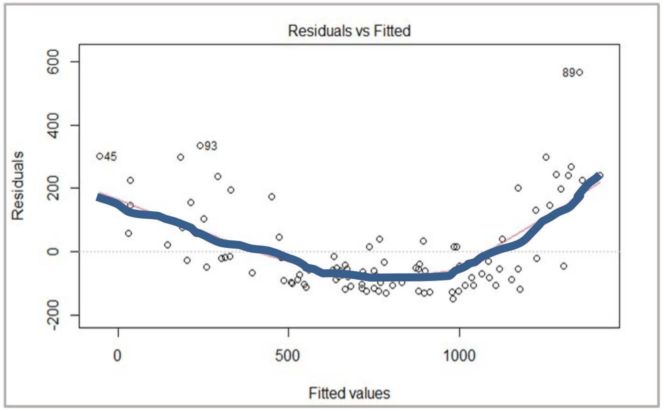
만약 **선형성 가정이 위배**되었다면?

↓

변수 변환이나 **비선형 모델**을 추정함으로써 대처할 수 있음

3 선형성 진단과 처방

진단 | 잔차 플랏



✓**선형성이 위배**되는 보통의 경우, **이차함수** 혹은 **삼차함수** 형태처럼 나타남

✓오른쪽 플랏은 **빨간 실선**이 **이차함수 꼴**을 보이므로 **선형성 위배**

선형회귀모델 자체가 성립하지 않으며

예측 성능도 현저히 떨어짐

처방 | 변수 변환

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon \rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

x_1의 관점에서는 선형결합이 아니므로

$x_1^2 = x_2$ 으로 변수를 변환하면, x_2와 y는 선형결합이다.

여러가지 변수변환 방법

Function	Transformations of x and/or y	Resulting model
$y = \beta_0 x^{\beta_1}$	$y' = log(y), x' = log(x)$	$y' = log(\beta_0) + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = ln(y)$	$y' = ln(\beta_0) + \beta_1 x$
$y = \beta_0 + \beta_1 log(x)$	$x' = log(x)$	$y = \beta_0 + \beta_1 x'$
$y = \dfrac{x}{\beta_0 x - \beta_1}$	$y' = \dfrac{1}{y}, x' = \dfrac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

변수 변환을 통해 선형성을 확보할 수 있는 모델도 넓은 의미에서 선형 모델!

정규성 가정

정규성 가정

반응변수 Y를 측정할 때 발생하는 오차는 **정규분포**를 따를 것이라는 가정

회귀식이 데이터를 잘 표현한다면

↓

잔차들은 단순 측정 오차인 Noise라 여겨짐

↓

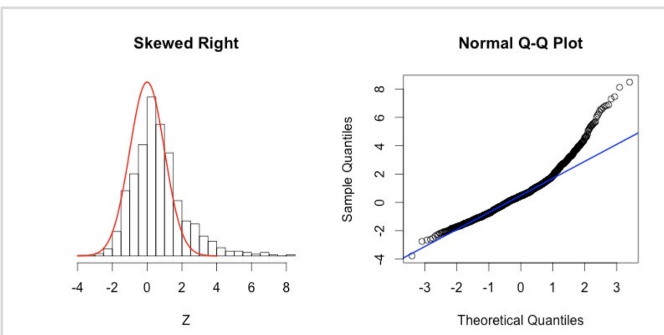
잔차들의 분포는 **정규분포**와 흡사한 형태

진단 | Normal QQ Plot

Normal QQ Plot

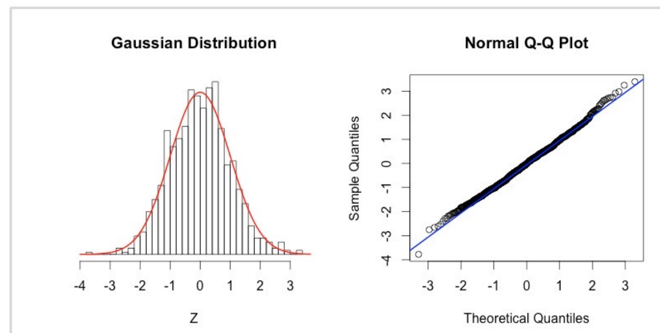
정규성을 파악하기 위한 **비모수적 방법**

$y = x$ 직선에 가까울수록 정규성을 만족



정규성을 만족하지 못하는 경우

왜도 (Skewed)가 양수인 경우
: 자료가 오른쪽으로 늘어져 있음



정규성을 만족하는 경우

자료의 분포가 정규분포에 가까움

진단 | 가설 검정

가설

H_0: 주어진 데이터는 **정규분포를 따른다**

H_1: 주어진 데이터는 **정규분포를 따르지 않는다**

Empirical CDF Test

관측치들을 **작은 순서대로 나열**한 후 **누적 분포 함수**를 그린 것

잔차의 ECDF와 **정규분포의** CDF를 비교함으로써 검정

자세한 방법은 다르지만 모두 잔차의 ECDF를 이용해 정규분포와 비교!

Anderson Darling Test

Kolmogorov Smirnov Test

진단 | 가설 검정

가설

H_0: 주어진 데이터는 **정규분포를 따른다**

H_1: 주어진 데이터는 **정규분포를 따르지 않는다**

Empirical CDF Test

관측치들을 **작은 순서대로 나열**한 후 **누적 분포 함수**를 그린 것

잔차의 ECDF와 정규분포의 CDF를 비교함으로써 검정

자세한 방법은 다르지만 모두 잔차의 ECDF를 이용해 정규분포와 비교!

Anderson Darling Test

Kolmogorov Smirnov Test

정규성이 위배됐을 경우 문제점

검정통계량이 t분포 또는 F분포를 따르지 않게 됨

(t분포, F분포는 정규분포를 전제하므로)

p-value에 의해 유의한

검정 결과와 예측 결과가 나와도,

신뢰할 수 없음

처방 | 변수 변환

Box-Cox Transformation

Y를 변환함으로써 **정규성**이나 **등분산성**을 해결해주는 방법

통계적 검정에 따라 변수 변환

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ log(y) & \text{if } \lambda = 0 \end{cases}$$

y가 정규성을 만족하도록 λ값을 조절

일반적으로 λ는 −5와 5 사이의 값을 사용

Yeo-Johnson Transformation

Box-cox transformation과 동일한 아이디어

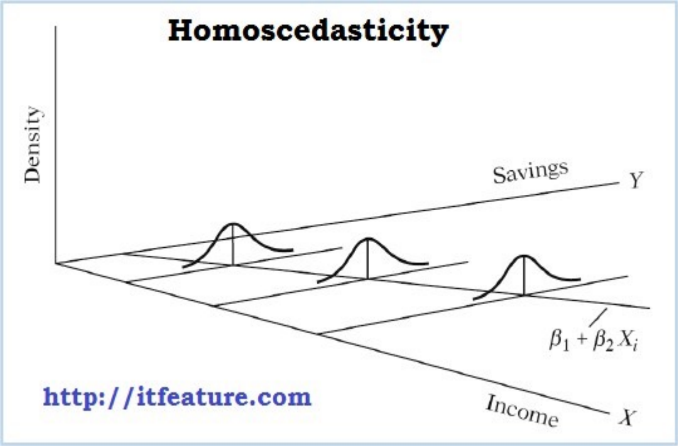
$$\psi(\lambda, y) = \begin{cases} ((y+1)^{\lambda} - 1)/\lambda\ , & \text{if } \lambda \neq 0,\ y \geq 0 \\ \log(y+1)\ , & \text{if } \lambda = 0,\ y \geq 0 \\ -[(-y+1)^{2-\lambda} - 1)]/(2-\lambda)\ , & \text{if } \lambda \neq 2,\ y < 0 \\ -\log(-y+1)\ , & \text{if } \lambda = 2,\ y < 0 \end{cases}$$

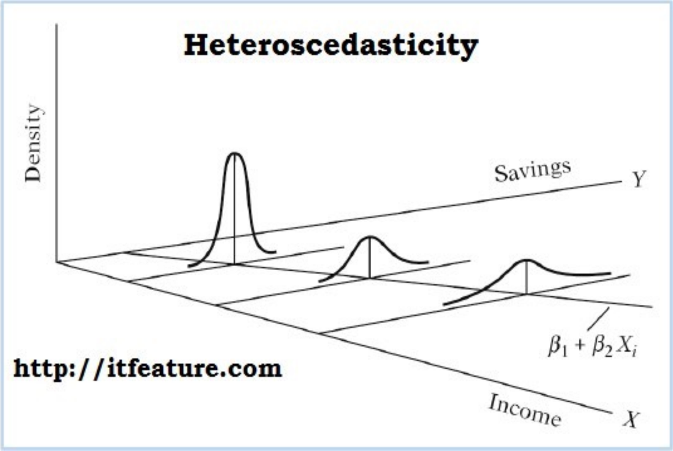
등분산성 가정

등분산성 가정

오차항의 분산은 어느 관측치에서나 **상수 σ^2**으로 동일하고
다른 변수의 영향을 받지 않는다는 가정

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



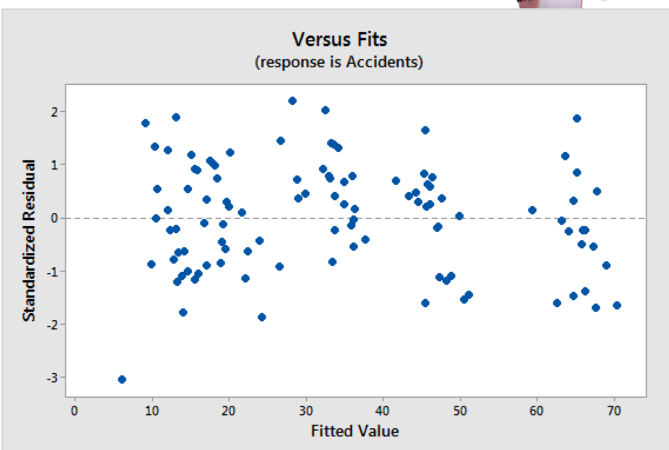


등분산성 진단과 처방

진단 | 잔차 플랏

등분선성 만족

잔차가 RANDOM하게 분포



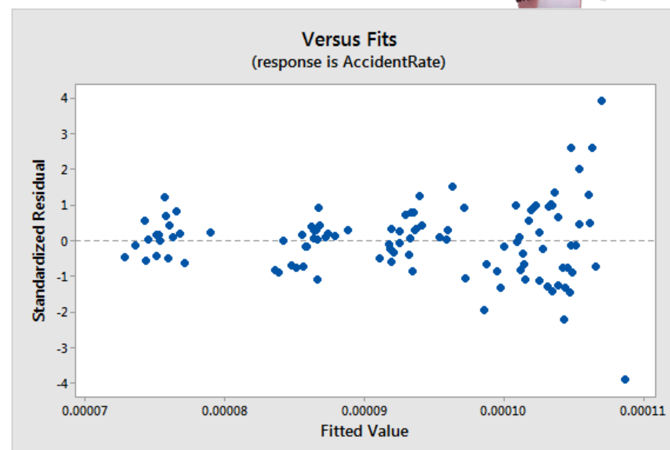
Fitted value $\hat{y}$값에

상관없이 잔차의

퍼짐의 정도가 일정

등분산성 위배

잔차가 특정 패턴을 가짐



Fitted value $\hat{y}$값이

커짐에 따라 잔차의

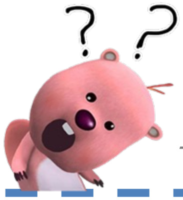
퍼짐의 정도가 변화

진단 | 가설 검정

가설

H_0: 주어진 데이터는 **등분산성을 지닌다**

H_1: 주어진 데이터는 **등분산성을 지니지 않는다**



우리가 원하는 것은?

귀무가설을 기각하지 못하는 것

즉 주어진 데이터가 **등분산성을 지니는 것**

이분산성의 문제점

이분산은 OLS 추정량의 분산을 **과소추정**

유의하지 않은 변수가 유의하다는 **잘못된 결과 도출 가능**

제 1종 오류(Type 1 error) 발생

가설검정의 신뢰성 하락

OLS 추정량이 **BLUE**가 되지 못함

Best Linear Unbiased Estimator

제1종오류 : 귀무가설이 실제로 참이지만, 이에 불구하고 귀무가설을 기각하는 오류

처방 | 변수 변환

가중 회귀 제곱 weighted Least Square

등분산이 아닌 형태의 데이터마다 **다른 가중치**를 주어서

등분산을 만족하게 해주는 **일반화된 최소제곱법**의 한 형태

$$\sum w_i(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2, \qquad w_i \propto \frac{1}{\sigma_i^2}$$

W_i는 **가중치이며, 분산에 반비례**

분산이 커 신뢰도가 낮은 부분의 관측치 → 적은 가중치 (분산을 작게 만듦)
분산이 작아 신뢰도가 높은 부분의 관측치 → 큰 가중치 (분산을 크게 만듦)

작은 가중치를 가지는 관찰값은

회귀계수 값을 결정하는데 **적은 영향을 미침**

독립성 가정

독립성 가정

공분산이 $0\ (Cov(e_i, e_j) = 0)$

오차항끼리는 **서로 독립**이라는 가정

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

독립성 가정 위배 시 오차항끼리의 **자기상관(autocorrelation)** 존재

시간적으로 자기상관

: **시계열 분석**을 이용

공간적으로 자기상관

: **공간회귀**를 통해 접근

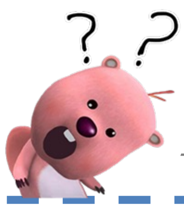
6 독립성 진단과 처방

진단 | 가설 검정

가설

H_0: 잔차들이 **서로 독립이다 (자기상관성이 없다)**

H_1: 잔차들이 **서로 독립이 아니다 (자기상관성이 있다)**



우리가 원하는 것은?

귀무가설을 기각하지 못하는 것

즉 주어진 데이터의 **잔차들이 서로 독립인 것**

독립성 진단

R lmtest 패키지의 dwtest() 함수 사용
car 패키지의 durbinWatsonTest() 함수 사용

Durbin Waston Test

앞 뒤 관측치의 **1차 자기상관성(first order autocorrelation)**을 확인

검정통계량 : $d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$

First order autocorrelation : $\widehat{\rho_1} = \frac{\widehat{Cov}(e_i, e_{i-1})}{\sqrt{V(e_i)} \cdot \sqrt{V(e_{i-1})}} \approx \frac{\sum_{i=1}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}$

$\therefore d \approx 2(1 - \widehat{\rho_1}) \rightarrow$ $\mathbf{[0, 4]}$ 값을 가짐

$\widehat{\rho_1}$은 표본 잔차 자기상관(sample autocorrelation of the residuals)

$[\mathbf{-1, 1}]$ 값을 갖는 e_i와 e_{i-1}의 상관계수의 꼴 !

6 독립성 진단과 처방

처방

가변수 생성

뚜렷한 **계절성**이 있다고 판단되면, **가변수** 생성
주기 함수인 **삼각함수** $\boldsymbol{cos(t), sin(t)}$**의 선형결합**으로 주기를 표현하는 방법

분석 모델 변경

☝️ **시간**에 따라 자기상관을 가지는 경우
시계열 모델 사용 (ex. $AR(p)$ 모형)

✌️ **공간**에 따라 자기상관을 가지는 경우
공간회귀모델 사용