회귀분석팀

6팀
고경현
박세령
박이현
박지성
심예진
이선민

INDEX

1. 회귀 기본가정

2. 잔차 플랏

3. 선형성 진단과 처방

4. 정규성 진단과 처방

5. 등분산성 진단과 처방

6. 독립성 진단과 처방

1 회귀 기본가정

회귀 가정의 필요성

선형회귀가 여전히 사용되는 이유

적은 관측치로도 모델을 쉽게 구성 가능

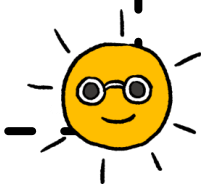
여러 변수간 복합적인 관계를 설명할 수 있음

But, 강한 설명력에는 많은 제약 존재!

선형회귀분석의 기본 가정이 위배되면?

불안정한 모델 추정

설명력과 예측력을 잃음



1 회귀 기본가정

선형회귀분석의 가정

기본가정 진단법

시각적 방법

잔차 플랏

선형성 | 정규성 | 독립성 | 등분산성

4가지 기본 가정 진단 가능

가설 검정 방법

가정마다 다른 검정 방법

정규성 검정, 독립성 검정 등

2 잔차 플랏 (Residual Plot)

잔차 플랏 출력

잔차 분포를 통해 **경험적 판단**에 근거한 **회귀진단** 가능

R에서 Plot() 함수를 통해 잔차의 분포를 표현

1) Residuals vs Fitted

2) Normal Q-Q plot

3) Scale – Location

4) Residuals vs Leverage

3 선형성 진단과 처방

선형성 가정

선형성 가정

반응변수(Y)가 설명변수(X)의 **선형결합**으로 이루어짐

선형성 가정 ⇒ 단순선형회귀 다중선형회귀

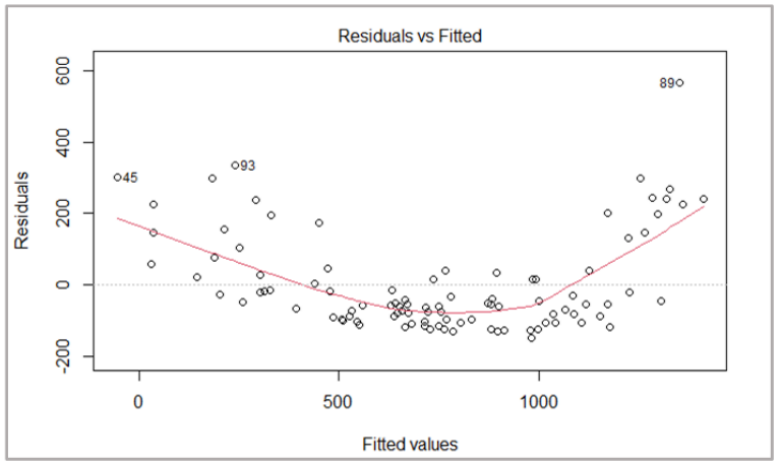
선형성 가정이 위배되었다면?

변수 변환이나 비선형 모델 추정으로 대처 가능!

3 선형성 진단과 처방

진단 | ① 잔차 플랏

측정값과 잔차 비교



잔차의 추세가 이차함수 꼴 ▸ 선형성 위반

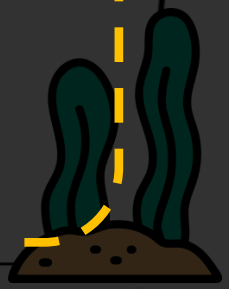
3 선형성 진단과 처방

진단 | ② Partial residual plot

X와 [illegible] 비교

개별 독립 변수와 종[illegible]의 선형성 판단 가능

선형성이 위반되었을 경우,
개별 변수들의 선형성을 판단하기에는 좋은 방법
선형회귀모델 자체가 성립하지 않으며
But, Y와 X변수들 간의 단편적인 관계만을 보여줌
예측 성능도 현저히 떨어짐!
∴ X변수들 사이의 교호작용이나 상관관계 파악은 어려움



점선과 실선이 불일치 → X와 Y는 비선형

처방 | ① 변수 변환

$$y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \epsilon$$

→ 일부 변수가 비선형 결합

변수 변환

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

→ x_2와 y의 선형 결합

여러가지 변수 변환 방법

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

정규성 가정

정규성 가정

반응 변수 Y를 측정할 때 발생하는 오차는

정규분포를 따를 것이라는 가정

회귀식이 데이터를 잘 표현한다면!

잔차들은 단순한 측정 오차, 즉 Noise라 여겨짐

잔차들의 분포는 정규분포와 흡사한 형태

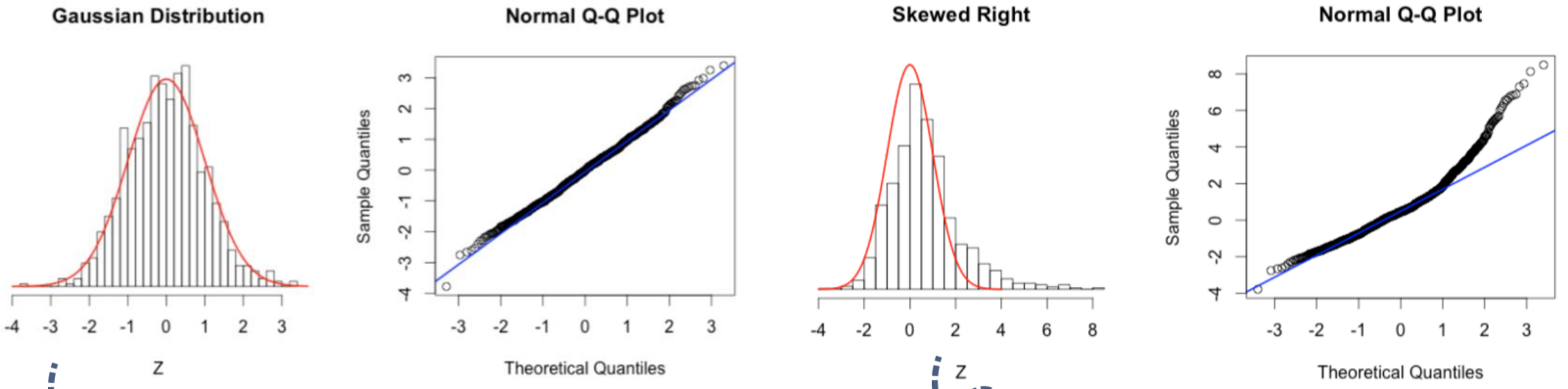
4 정규성 진단과 처방

진단 | ① Normal Q-Q plot

Normal Q-Q plot

정규성을 파악하기 위한 대표적인 **비모수적** 방법

직선에 가까운 형태이면 **정규성** 만족



Gaussian Distributuion

정규성 만족

Right Skewed

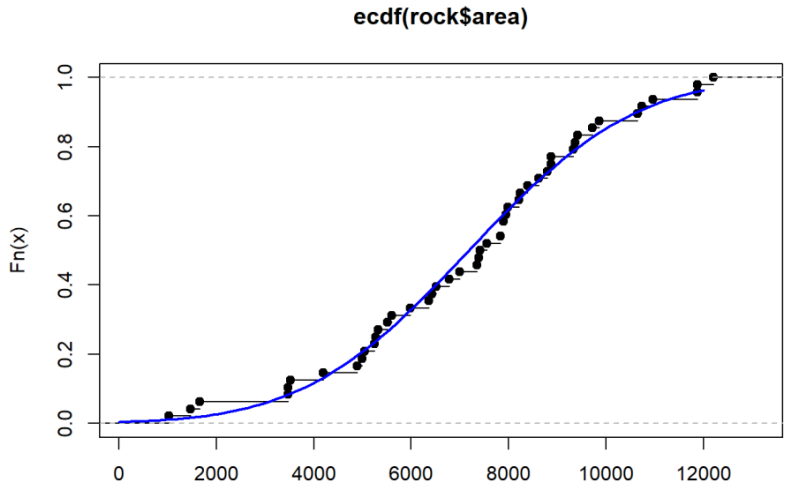
정규성 불만족

진단 | ② 정규성 검정

가설

H_0 : 주어진 데이터는 정규분포를 따른다.

H_1 : 주어진 데이터는 정규분포를 따르지 않는다.



- 관측치들을 작은 순서대로 나열한 후, 관측치들로 **누적 분포 함수**를 그린 것
- 정규성 검정을 위해 잔차의 ECDF와 정규분포의 CDF를 비교하여 검정

4 정규성 진단과 처방

진단 | ② 정규성 검정 | ② 정규분포의 분포적 특성을 이용하는 test

정규분포의 분포적 특성을 이용하는 test

정규성이 위반되었을 경우

검정통계량이 t분포 또는 F분포를 따르지 않게 됨

t분포, F분포는 정규분포에서 파생되므로!

$$JB = n\left(\frac{(\sqrt{skew})^2}{6} + \frac{(kurt-3)^2}{24}\right)$$

가설 검정 결과가 p-value에 의해 유의하게 나와도,

결과 신뢰할 수 없음

예측 결과 또한 신뢰하기 어려움!

처방 | ② Box-Cox Transformation

λ를 **변화**시키면서 y가 **정규성**을 만족하도록 만드는 방법

▸ **통계적인** 방법에 의한 변수 변환

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

일반적으로 λ는 -5에서 5 사이의 값을 사용

λ가 0인 경우, log-transformation을 적용

처방 | ③ Yeo-Johnson Transformation

Box-Cox Transformation은 y가 $\log(y)$로 변환될 수 있으므로

y가 **항상 양수**여야 하는 단점 존재

해결법 ②

Yeo-Johnson Transformation

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

등분산성 가정

등분산성 가정

오차항의 분산이 어느 관측치에서나 동일하며

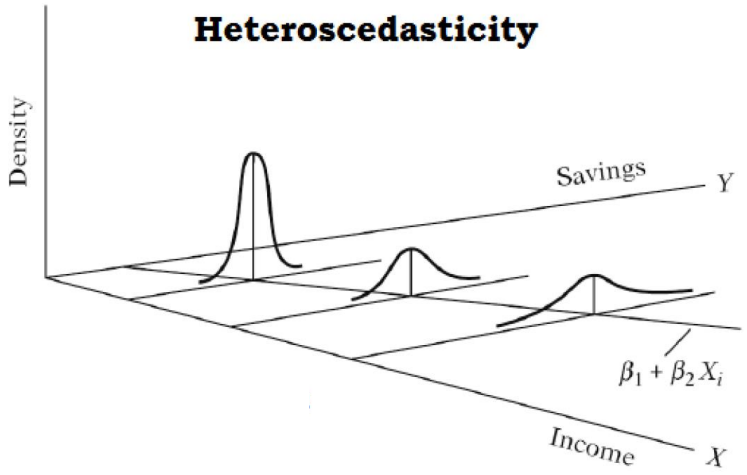
다른변수의 영향을 받지 않는 즉, 상수라는 가정

지점에 따라

y의 조건부 분포의 모양이

같지 않음

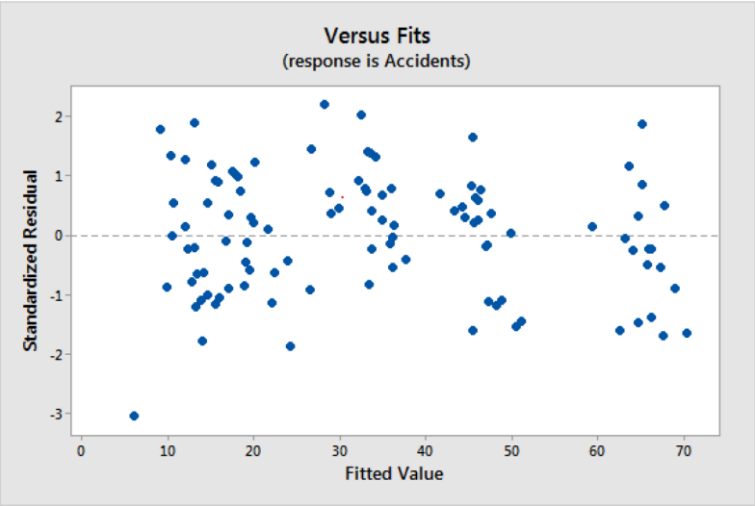
등분산이 위배됨

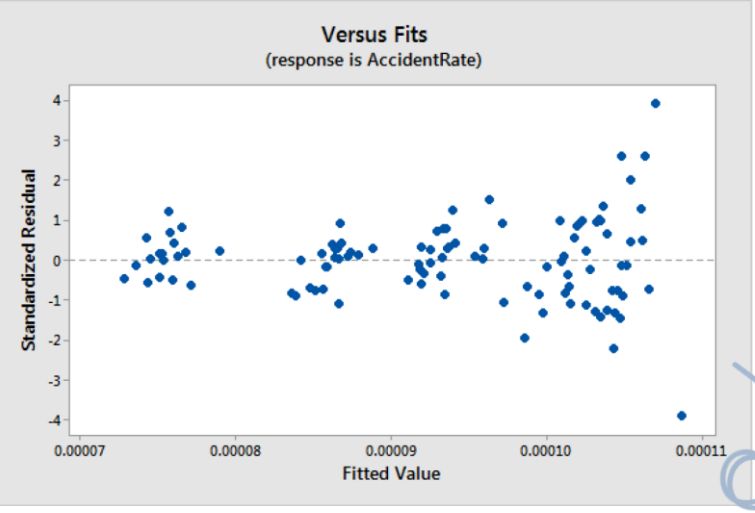


5 등분산성 진단과 처방

진단 | ① 잔차 플랏

'residual vs fitted' plot과 'scale-location' plot을 보고 종합적으로 판단





$\hat{y}$값이 커짐에 따라 잔차의 절대값이 커지는 형태 → 이분산성을 띔

5 등분산성 진단과 처방

진단 | ② Breusch-Pagan test (BP test)

등분산성이 위반되었을 경우

OLS 추정량의 분산은 실제 분산보다 작게 추정됨

이분산은 추정량의 분산을 증가시키지만, OLS 추정량은 이를 잡아내지 못함

한계점

검정통계량 증가 & P-value 감소

비선형적 결합으로 이루어진 이분산성은 파악 불가

유의하지 않은 회귀 계수조차 유의해짐

Sample의 크기가 커야 (대표본) 사용가능

5 등분산성 진단과 처방

처방

변수 변환

정규성을 만족시키기 위해 사용했던 각종 변수 변환 방법과 동일

가중 회귀

등분산이 아닌 형태의 데이터마다 다른 가중치를 주어서

등분산을 만족하게 해주는 '**일반화된 최소제곱법**'의 한 형태

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

$$\mathbf{Y} = \mathbf{WX}\beta + \epsilon$$

$$\hat{\beta}^{WLS} = (\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{WY}$$

독립성 가정

독립성 가정

$$Cov(e_i, e_j) = 0$$

오차들은 서로 독립이며, 개별 관측치에서 i번째 오차와 j번째 오차가 발생하는 것에 서로 영향을 미치지 않는다는 가정

독립성 가정이 위배된다면!

- 오차들간의 자기상관(autocorrelation)이 있다고 함
- 시공간 상의 데이터일 경우 오차들에 일종의 패턴이 존재할 수 있음

진단 | ① 더빈-왓슨 검정

더빈-왓슨 검정

바로 앞 뒤 관측치의 1차 자기 상관성을 확인하는 검정 방법

가설설정

H_0: 1차 자기 상관이 없다

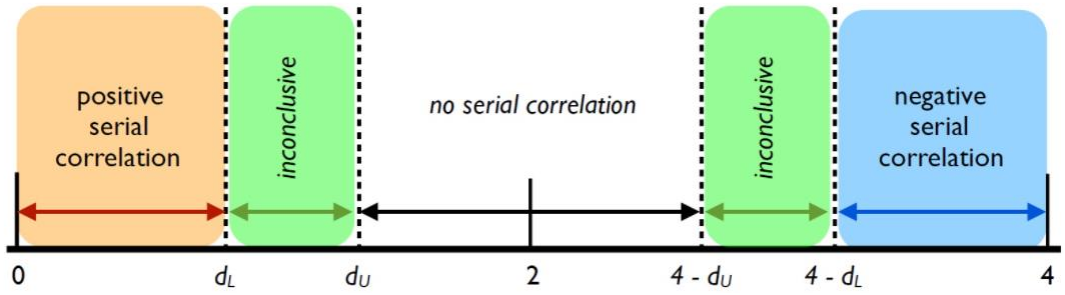
H_1: 1차 자기 상관이 있다(잔차들이 서로 독립이 아니다)

검정통계량

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

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정



해석

더빈 왓슨 검정 표에서 데이터 개수 n과 변수의 개수 p에 따라 귀무가설을 기각할 수 있는지 없는지 판단하는 **컷 오프 값**을 알려줌

- 귀무가설 **기각** if 검정통계량 d < 하한
  - 양의 자기상관이 있다고 판단
- 귀무가설 **기각 안됨** if 검정통계량 d > 상한

처방 | ② 분석 모델 변경

가변수 만들기

계절성이 주기를 가진다는 점을 이용하여 주기 함수인 삼각함수 cos(t), sin(t)의 선형결합으로 주기를 표현하는 방법

분석 모델 변경

시간에 따라 자기상관을 가질 경우 자기상관을 고려하는 AR(p) 같은 시계열 모델을 사용

공간에 따라 자기상관을 가질 경우 공간의 인접도를 고려하는 공간회귀모델을 사용