

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

0

REVIEW

회귀분석



Regression Analysis

변수 사이의 관계를 모델링하는 통계적 기법

특정 변수들의 값을 이용하여 다른 변수를 설명하거나 예측

Ex) 스마트폰 이용시간(X변수)에 따른 기말고사 성적 변화(Y변수)

지도학습의 한 종류



지도학습이란?



결과의 예측이 목적인 학습 방법

결과변수와 특징변수가 모두 존재

다중선형회귀

단순선형회귀

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

다중선형회귀

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

+ 독립변수 + 독립변수
+ 독립변수
+ 독립변수
+ 독립변수 + 독립변수

단순선형회귀에 비해 **복잡한 관계** 설명이 용이

나머지 X변수들이 고정되어 있을 때,
 x_p 가 한 단위 증가하면 y 는 β_p 만큼 증가함을 의미

유의성 검정



전체 회귀계수에 대한 검정 : F-test

임계값

$$F_{(1-\alpha/2, p, n-p-1)}$$

- ✓ 귀무가설 **기각** if $F_0 \geq F_{(1-\alpha/2, p, n-p-1)}$
 - ▶ 적어도 한 개의 회귀계수는 0이 아님
- ✓ 귀무가설 **기각 안됨** if $F_0 < F_{(1-\alpha/2, p, n-p-1)}$
 - ▶ 모든 회귀계수는 0임
 - 모델 재설정 등 다른 조치 필요!

데이터 진단의 필요성

일반적인 경향에서 벗어나는 개별 데이터 존재

이상치, 지렛값, 영향점 등

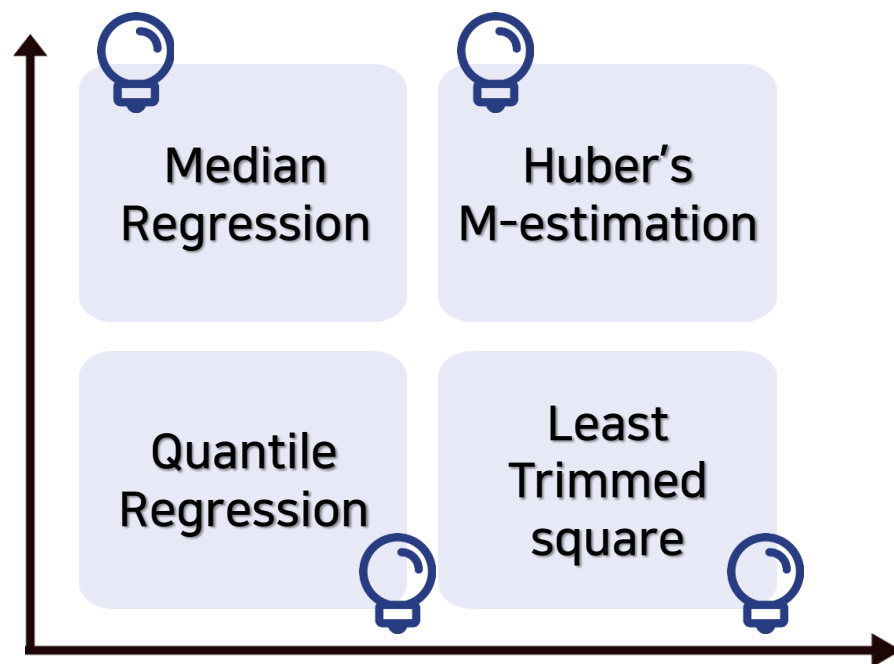
회귀 모형에 큰 영향을 미침

개별 데이터가 경향성에 벗어나는지 판단하여 처리 필요!

로버스트 회귀

로버스트 회귀

이상치의 영향력을 크게 받지 않는 회귀모형



1

회귀 기본가정

회귀 가정의 필요성



선형회귀가 여전히 사용되는 이유

적은 관측치로도 모델을 쉽게 구성 가능
여러 변수간 **복합적인 관계**를 설명할 수 있음

But, 강한 설명력에는 많은 제약 존재!



선형회귀분석의 기본 가정이 위배되면?

불안정한 모델 추정
설명력과 예측력을 잃음



회귀 가정의 필요성



선형회귀가 여전히 사용되는 이유

적은 관측치로도 모델을 쉽게 구성 가능

여러 변수간 복합적인 관계를 설명할 수 있음

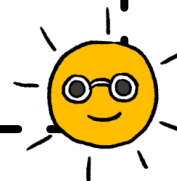
But, 강한 설명력에는 많은 제약 존재!



선형회귀분석의 기본 가정이 위배되면?

불안정한 모델 추정

설명력과 예측력을 잃음



선형회귀분석의 가정

변수에 대한 가정

오차항에 대한 가정

모델의 선형성

설명변수와 반응변수는
선형관계



설명변수의 독립성

설명변수들은 서로 독립

설명변수 \neq 확률변수

측정에 오차는
존재하지 않음



선형회귀분석의 가정

변수에 대한 가정

오차의 평균은 0



오차의 정규성

오차항은
정규분포를 따름

오차항에 대한 가정

오차의 등분산성

오차항의 분산은 상수



오차의 독립성

오차항은 서로 독립



선형회귀분석의 가정

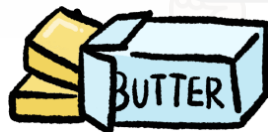
선형회귀분석의 기본 가정

선형성

정규성

독립성

등분산성



선형회귀분석의 가정



기본 가정 in 다중선형회귀모형

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon, \quad \varepsilon \sim MVN(0, \sigma^2 I)$$

선형성

정규성

▶ 독립변수는 선형결합을 이루며 종속변수 표현

▶ 오차항은 다변량정규분포를 따름

▶ 분산의 크기는 상수 σ^2 로 일정한 등분산

▶ 오차항은 서로 독립(I , Identity Matrix)



선형회귀분석의 가정



기본 가정 in 다중선형회귀모형

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon, \quad \varepsilon \sim MVN(0, \sigma^2 I)$$

선형성

정규성

▶ 독립변수는 선형결합을 이루며 종속변수 표현

▶ 오차항은 다변량정규분포를 따름

▶ 분산의 크기는 상수 σ^2 로 일정한 등분산

▶ 오차항은 서로 독립(I , Identity Matrix)



기본가정 진단법

시각적 방법



선형성 | 정규성 | 독립성 | 등분산성

4가지 기본 가정
진단 가능💡

가설 검정 방법



정규성 검정, 독립성 검정 등

2

잔차 플랫폼

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

잔차 플랏 출력

잔차 분포를 통해 **경험적 판단**에 근거한 **회귀진단** 가능
R에서 Plot() 함수를 통해 잔차의 분포를 표현



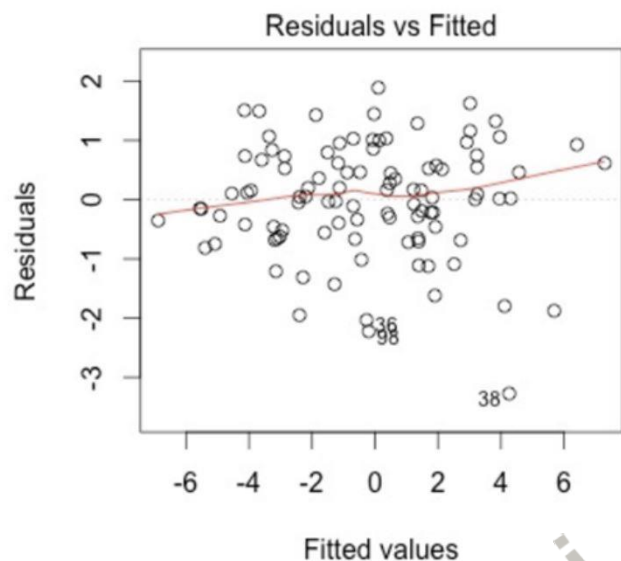
1) Residuals vs Fitted

2) Normal Q-Q plot

3) Scale - Location

4) Residuals vs Leverage

Residuals vs Fitted



기본 요소

선형성과 오차의 등분산성 확인

X축 : 예측값(\hat{y}), Y축 : 잔차($e = y - \hat{y}$)

빨간 실선 : 잔차추세선 (경향성)

▶ Local Regression으로 추정

해석

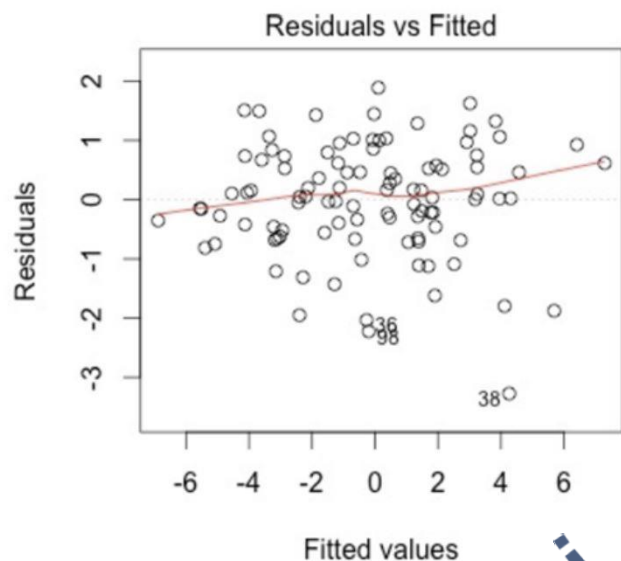
랜덤 패턴 + 수평적 잔차추세선

선형성과 오차의 등분산성 만족

2

잔차 플랏 (Residual Plot)

Residuals vs Fitted



기본 요소

선형성과 오차의 등분산성 확인

X축 : 예측값(\hat{y}), Y축 : 잔차($e = y - \hat{y}$)

빨간 실선 : 잔차추세선 (경향성)

▶ Local Regression으로 추정

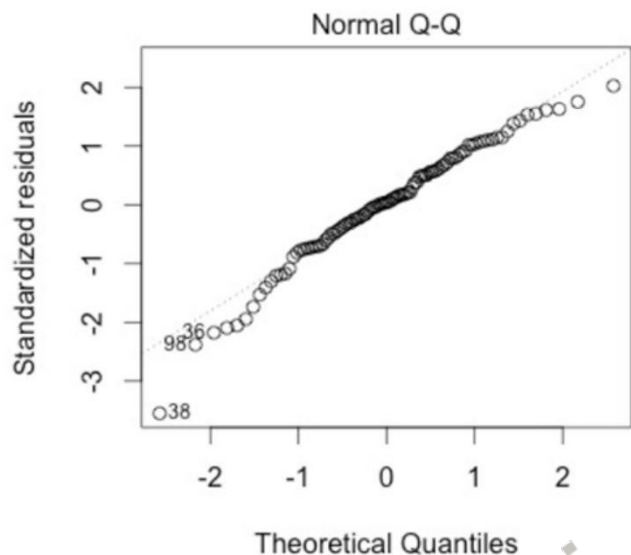
해석

랜덤 패턴 + 수평적 잔차추세선

선형성과 오차의 등분산성 만족

2 잔차 플랏 (Residual Plot)

Normal Q-Q plot



기본 요소

정규성 확인

X축 : 정규분포의 분위수

Y축 : 스튜던트 잔차(r_i)

스튜던트 잔차를 정규분포분위수와 비교

변수 분포가 직선형태이면 분포가 같음

모든 점이 점선에 있다면 잔차는 정규분포

해석

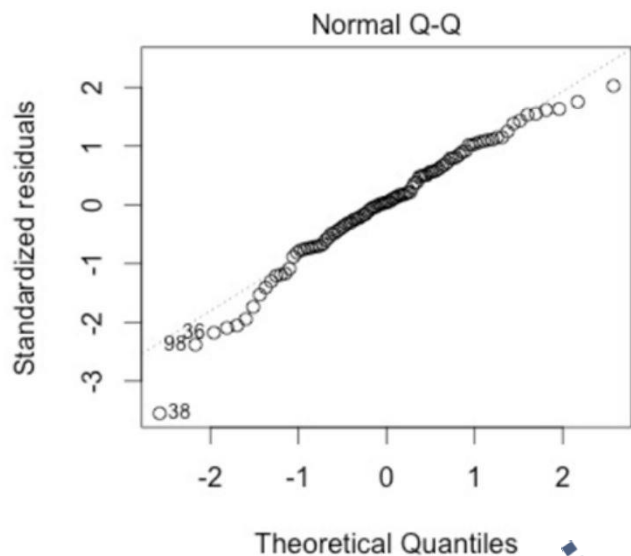
잔차가 대부분

점선 주변에 분포해 정규성 만족

38번 관측치가 벗어나므로 추가 확인 필요

2 잔차 플랏 (Residual Plot)

Normal Q-Q plot



기본 요소

정규성 확인

X축 : 정규분포의 분위수

Y축 : 스튜던트 잔차(r_i)

스튜던트 잔차를 정규분포분위수와 비교

변수 분포가 직선형태이면 분포가 같음

모든 점이 점선에 있다면 잔차는 정규분포

해석

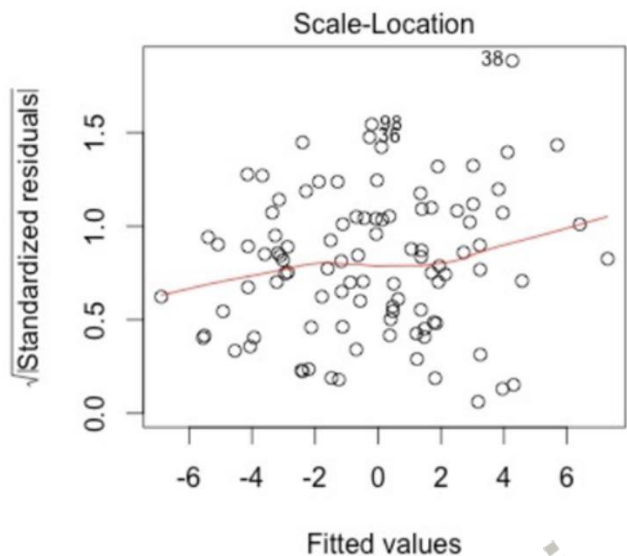
잔차가 대부분

점선 주변에 분포해 정규성 만족

38번 관측치가 벗어나므로 추가 확인 필요

2 잔차 플랏 (Residual Plot)

Scale - Location



기본 요소

선형성과 오차의 등분산성 확인(1번과 유사)

X축 : 예측값(\hat{y})

Y축 : 스튜던트 잔차 절대값($\sqrt{\frac{|e_i|}{se(e_i)}}$)

1번 플랏과 비슷한 판단 가능

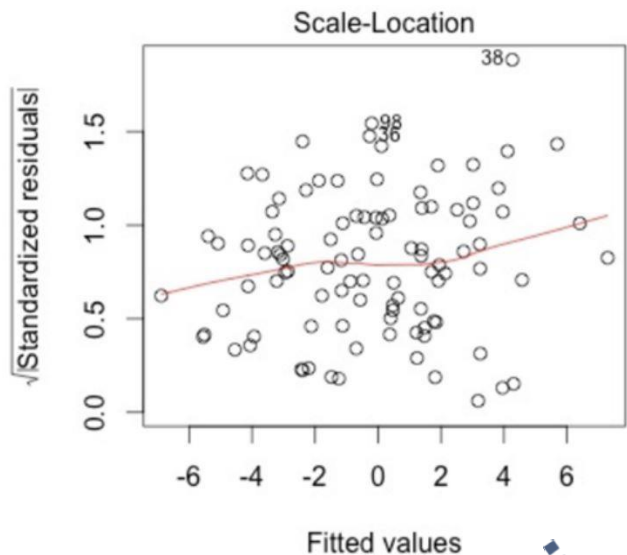
해석

랜덤 패턴 + 수평적 잔차추세선

선형성과 오차의 등분산성 만족

2 잔차 플랏 (Residual Plot)

Scale - Location



기본 요소

선형성과 오차의 등분산성 확인(1번과 유사)

X축 : 예측값(\hat{y})

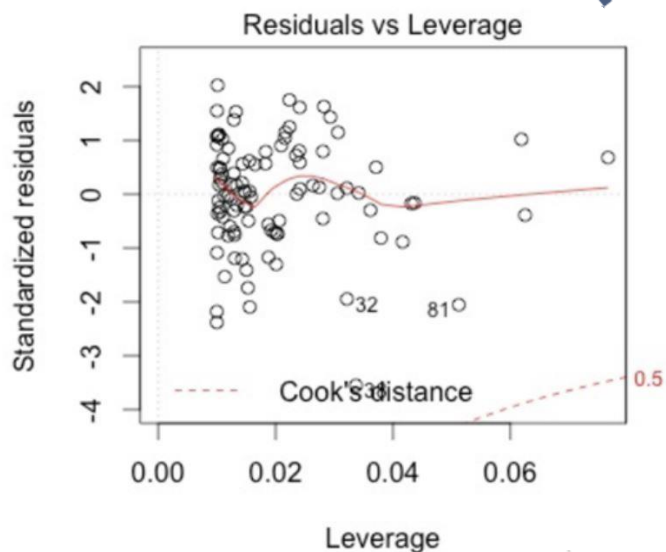
Y축 : 스튜던트 잔차 절대값($\sqrt{\frac{|e_i|}{se(e_i)}}$)

1번 플랏과 비슷한 판단 가능

해석

랜덤 패턴 + 수평적 잔차추세선
선형성과 오차의 등분산성 만족

Residuals vs Leverage



기본 요소

영향점(influential point) 확인

X축 : 레버리지값(\hat{y}), Y축 : 스튜던트 잔차(r_i)

플랏 우측에 위치한 점은 레버리지가 큼

빨간 실선으로부터 상하로 떨어지면 outlier

빨간 점선 : Cook's distance

해석

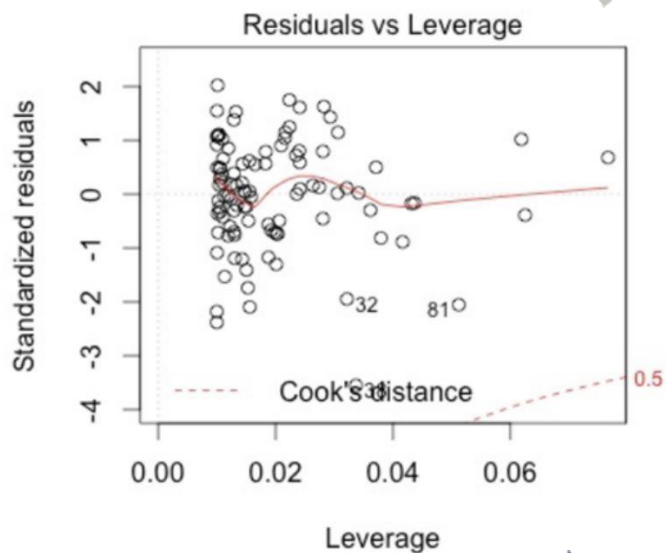
모든 관측치들이 0.5 경계 안에 위치

영향점은 존재하지 않는 것으로 보임

2

잔차 플랏 (Residual Plot)

Residuals vs Leverage



기본 요소

영향점(influential point) 확인

X축 : 레버리지값(\hat{x}), Y축 : 스튜던트 잔차(r_i)

플랏 우측에 위치한 점은 레버리지가 큼

빨간 실선으로부터 상하로 떨어지면 outlier

빨간 점선 : Cook's distance

해석

모든 관측치들이 0.5 경계 안에 위치
영향점은 존재하지 않는 것으로 보임

3

선형성 진단과 처방

선형성 가정

선형성 가정

반응변수(Y)가 설명변수(X)의 선형결합으로 이루어짐



선형성 가정



단순선형회귀 다중선형회귀



선형성 가정이 위배되었다면?

변수 변환이나 비선형 모델 추정으로 대처 가능!

선형성 가정

선형성 가정

반응변수(Y)가 설명변수(X)의 선형결합으로 이루어짐



선형성 가정



단순선형회귀 다중선형회귀

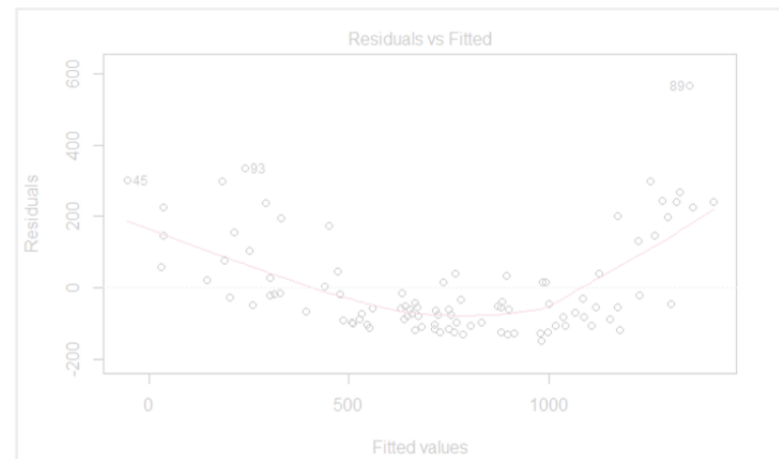
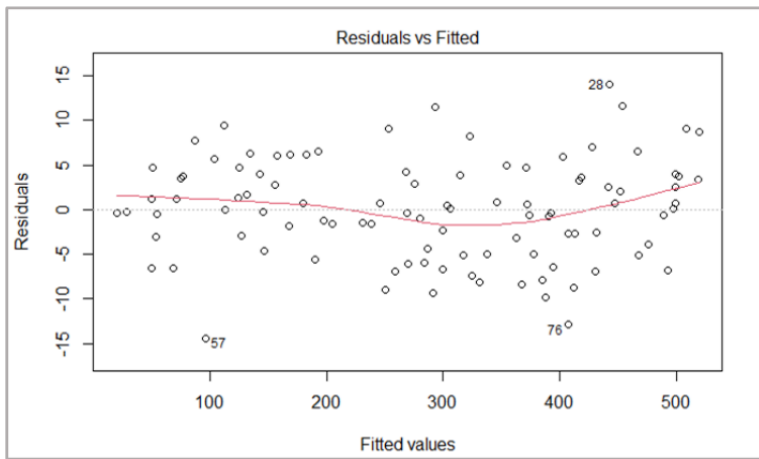


선형성 가정이 위배되었다면?

변수 변환이나 비선형 모델 추정으로 대처 가능!

진단 | ① 잔차 플랏

측정값과 잔차 비교

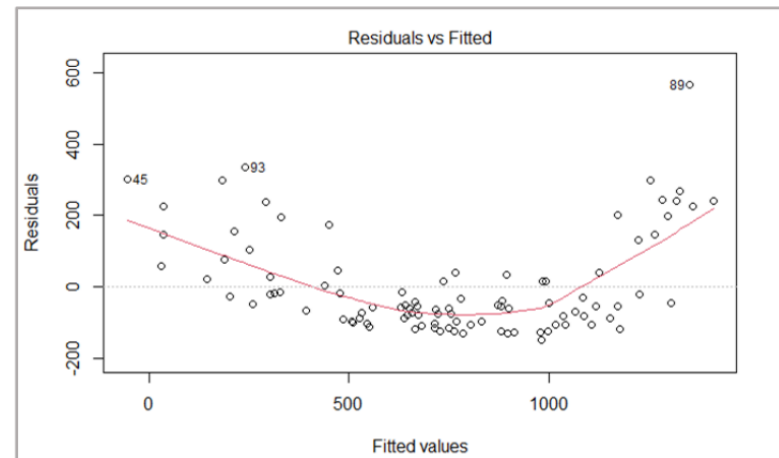
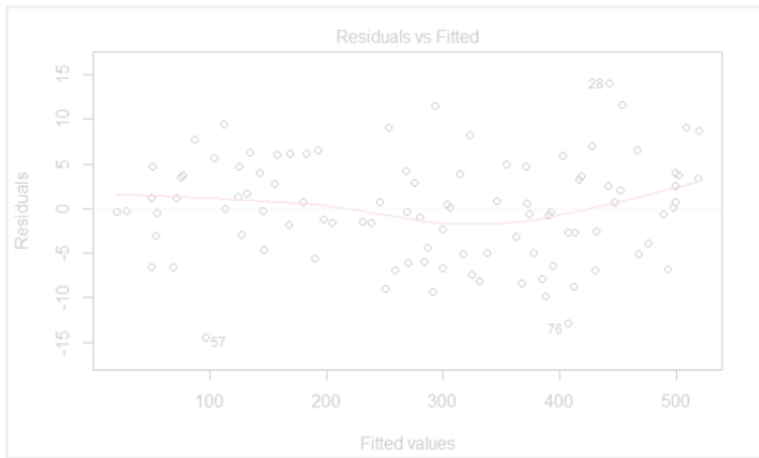


잔차의 추세가 X축과 비슷함 ▶ 선형성 만족



진단 | ① 잔차 플랏

측정값과 잔차 비교



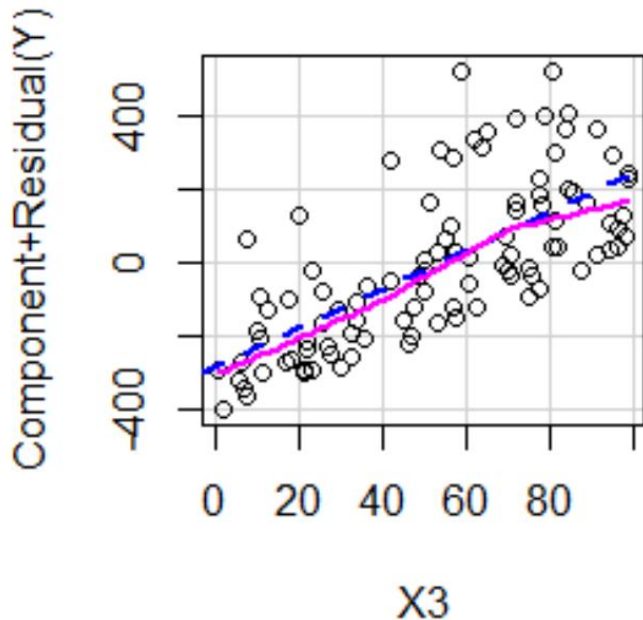
잔차의 추세가 이차함수 꼴 ▶ 선형성 위반



진단 | ② Partial residual plot

X와 잔차 비교

개별 독립 변수와 종속 변수 간의 선형성 판단 가능

**Y축**

전체 모형에서 선형성을 보고싶은
변수를 제외한 나머지 변수로 회귀식을
적합한 후의 잔차

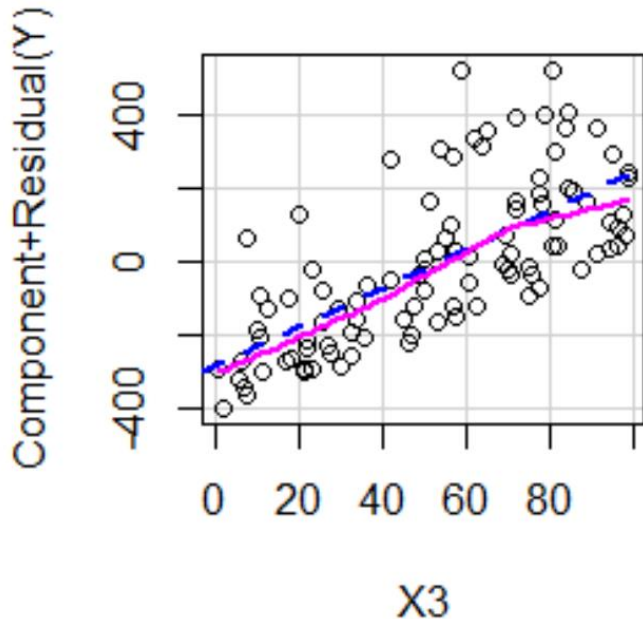
X축

선형성을 판단하기 위한 변수

진단 | ② Partial residual plot

X와 잔차 비교

개별 독립 변수와 종속 변수 간의 선형성 판단 가능

**점선**

점들의 분포를 최소제곱방법을 통해
추정한 회귀선

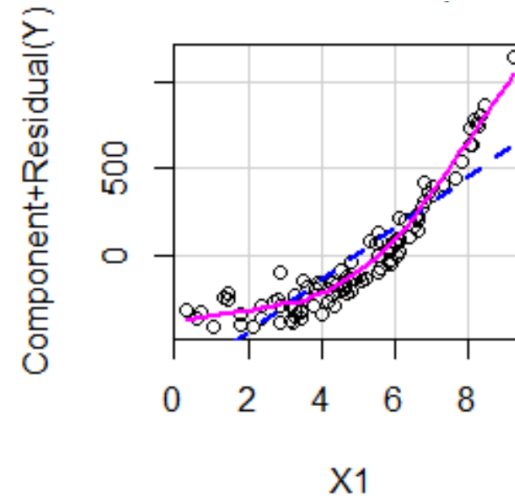
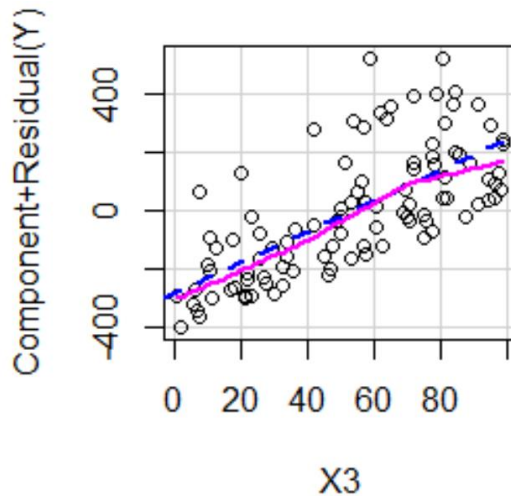
실선

점들의 분포를
Local Regression을 통해 추정한 선

진단 | ② Partial residual plot

X와 잔차 비교

개별 독립 변수와 종속 변수 간의 선형성 판단 가능



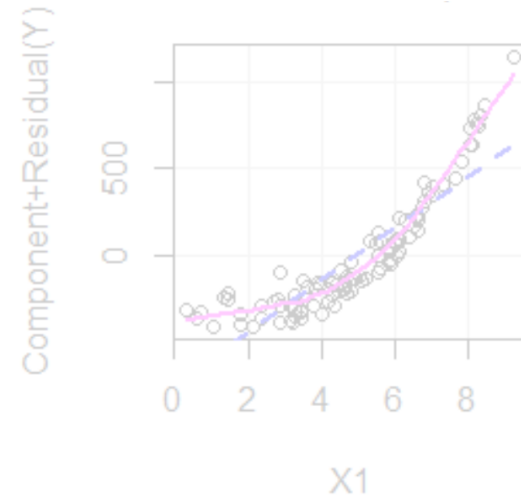
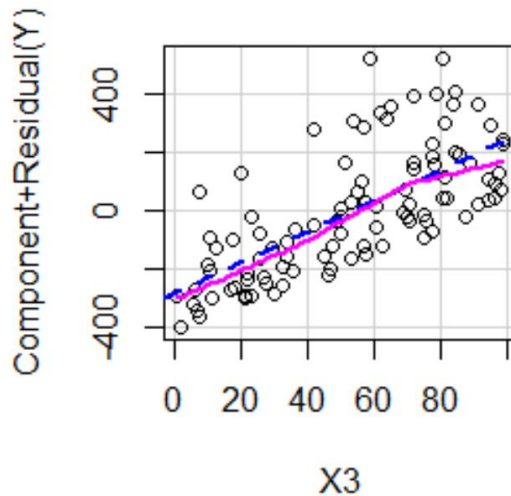
R에서는 car 패키지의 `crPlots()`를
이용하여 위와 같은 플랏을 출력



진단 | ② Partial residual plot

X와 잔차 비교

개별 독립 변수와 종속 변수 간의 선형성 판단 가능



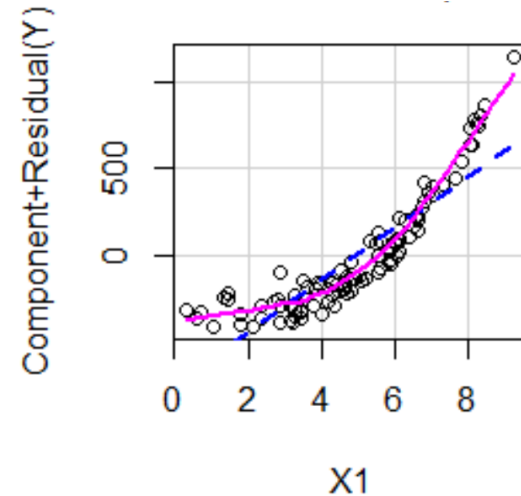
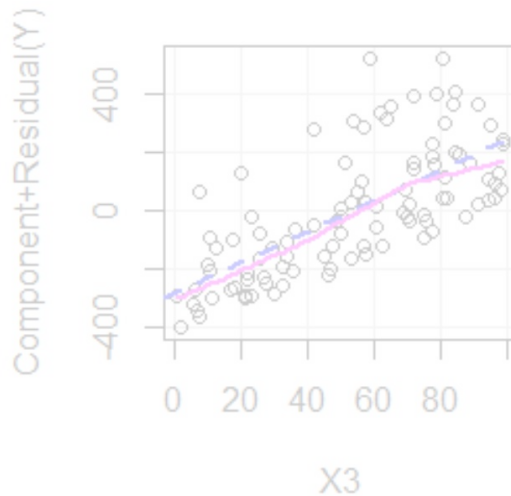
점선과 실선이 일치 ▶ X3과 Y는 선형



진단 | ② Partial residual plot

X와 잔차 비교

개별 독립 변수와 종속 변수 간의 선형성 판단 가능



점선과 실선이 불일치 ▶ X1과 Y는 비선형



진단 | ② Partial residual plot

X와 잔차 비교

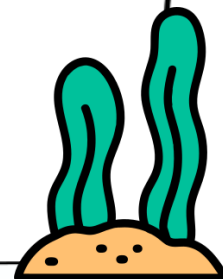
개별 독립 변수와 종속 변수 간의 선형성 판단 가능



개별 변수들의 **선형성**을 판단하기에는 좋은 방법

But, Y와 X변수들 간의 **단편적인 관계**만을 보여줌

∴ X변수들 사이의 교호작용이나 상관관계 파악은 어려움



점선과 실선이 불일치 ▶ X1과 Y는 비선형



진단 | ② Partial residual plot

X와 Y의 선형성 비교



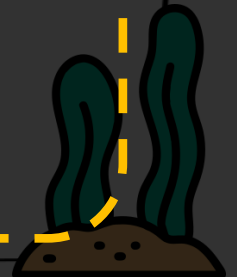
개별 독립 변수와 종속 변수의 선형성 판단 가능

선형성이 위반되었을 경우,

개별 변수들의 선형성을 판단하기에는 좋은 방법

선형회귀모델 자체가 성립하지 않으며But, Y와 X변수들 간의 **단편적인** 관계만을 보여줌**예측 성능도 현저히 떨어짐!**

∴ X변수들 사이의 교호작용이나 상관관계 파악은 어려움



점선과 실선이 불일치 ▶ X1과 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_2 와 y 의 선형 결합

여러가지 변수 변환 방법

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



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_2 와 y 의 선형 결합

여러가지 변수 변환 방법

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



처방 | ① 변수 변환

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

변수 변환

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

일부 변수가 비선형 결합

 x_2 와 y 의 선형 결합

여러가지 변수 변환 방법

변수 변환을 통해 선형성을 확보할 수 있는 모델 역시
넓은 의미에서 **선형 모형**!



처방 | ② 비선형 회귀

선형성 포기

모델 자체를 비선형 회귀 모델에 적합 시키는 방법

Polynomial Regression

$$Y \sim X_1 + X_1^2 + X_2 + X_2^2 + \epsilon$$

고차항을 고려하는 다항 회귀

변수의 차수를 다양하게 바꾸어 모델에 넣어주는 방법

삼차까지만 고려



처방 | ② 비선형 회귀

선형성 포기

모델 자체를 비선형 회귀 모델에 적합 시키는 방법

Local Regression

비선형 회귀 방법 & 비모수적 방법 사용

Local에 있는 데이터들로 회귀 모델링을 하는 방법

K개의 이웃 데이터들만 사용하여 부분적으로 회귀 모델 구성

모든 k개의 이웃에 각기 다른 가중치 부여

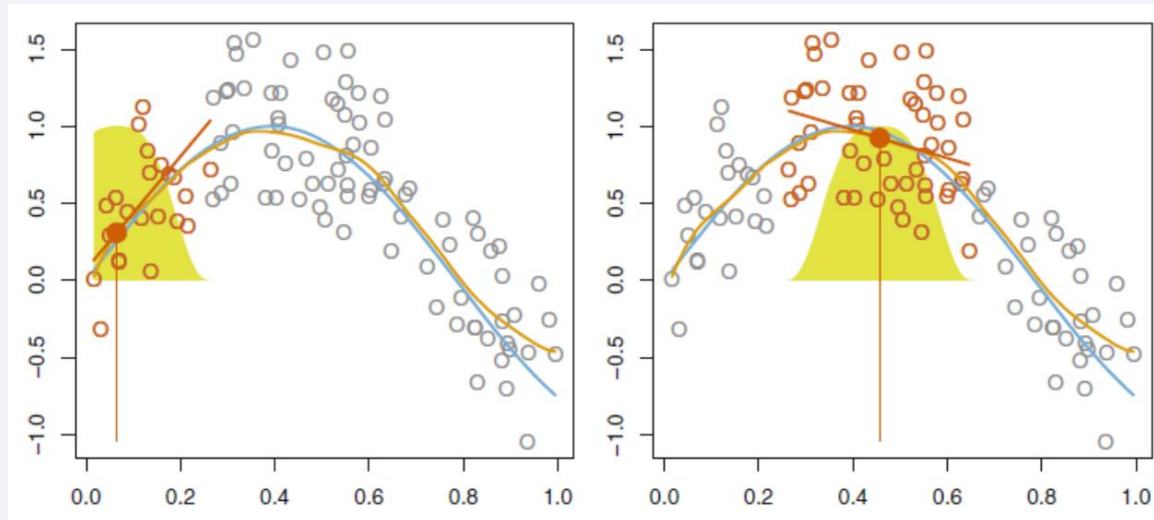


처방 | ② 비선형 회귀

선형성 포기

모델 자체를 비선형 회귀 모델에 적합 시키는 방법

Local Regression



4

정규성 진단과 처방

정규성 가정

정규성 가정

반응 변수 Y 를 측정할 때 발생하는 오차는
정규분포를 따를 것이라는 가정



회귀식이 데이터를 잘 표현한다면!

잔차들은 단순한 측정 오차, 즉 **Noise**라 여겨짐
잔차들의 분포는 **정규분포**와 흡사한 형태

정규성 가정

정규성 가정

반응 변수 Y 를 측정할 때 발생하는 오차는
정규분포를 따를 것이라는 가정



회귀식이 데이터를 잘 표현한다면!

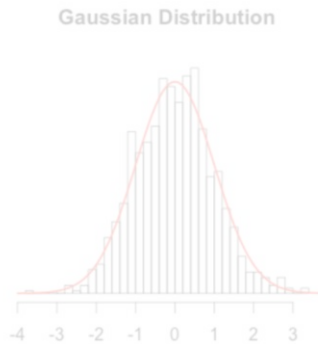
잔차들은 단순한 측정 오차, 즉 Noise라 여겨짐
잔차들의 분포는 정규분포와 흡사한 형태

진단 | ① Normal Q-Q plot

Normal Q-Q plot

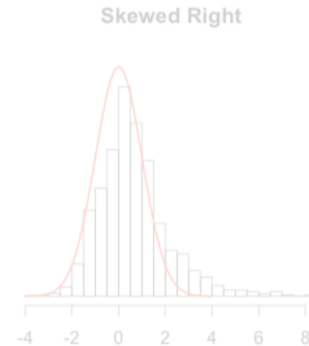
정규성을 파악하기 위한 대표적인 **비모수적** 방법

직선에 가까운 형태이면 **정규성** 만족



Gaussian Distributuion

정규성 만족



Right Skewed

정규성 불만족

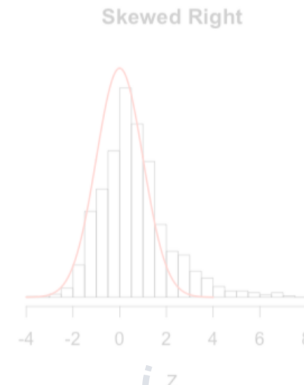
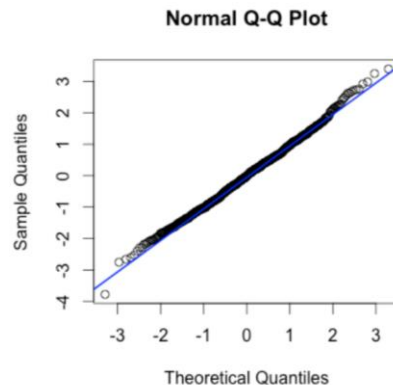
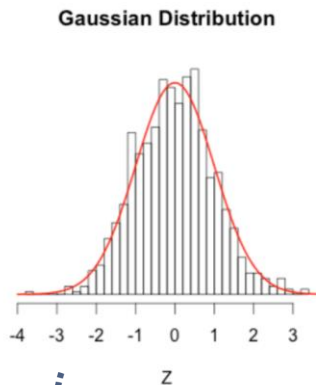


진단 | ① Normal Q-Q plot

Normal Q-Q plot

정규성을 파악하기 위한 대표적인 **비모수적** 방법

직선에 가까운 형태이면 **정규성** 만족



Gaussian Distributuion

정규성 만족



Right Skewed

정규성 불만족

4

정규성 진단과 처방

진단 | ① Normal Q-Q plot

Normal Q-Q plot

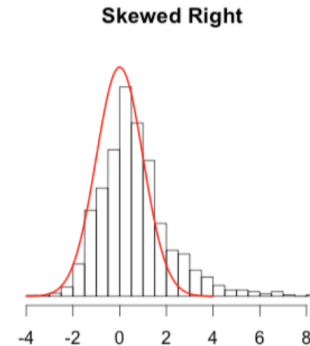
정규성을 파악하기 위한 대표적인 **비모수적** 방법

직선에 가까운 형태이면 **정규성** 만족



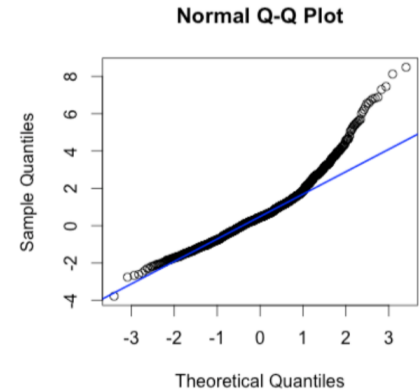
Gaussian Distributuion

정규성 만족



Right Skewed

정규성 불만족



4

정규성 진단과 처방

진단 | ① Normal Q-Q plot

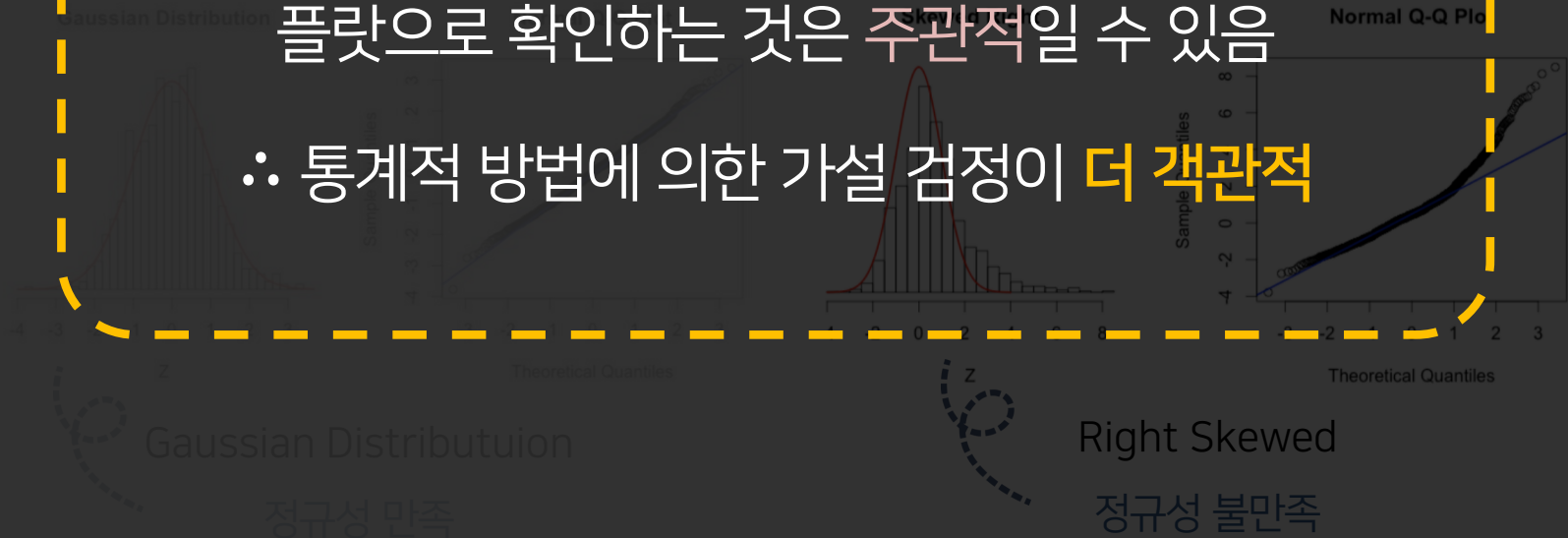
Normal Q-Q plot

정규성을 파악하기 위한 대표적인 비모수적 방법
 직선에 가까운 형태이면 정규성 만족



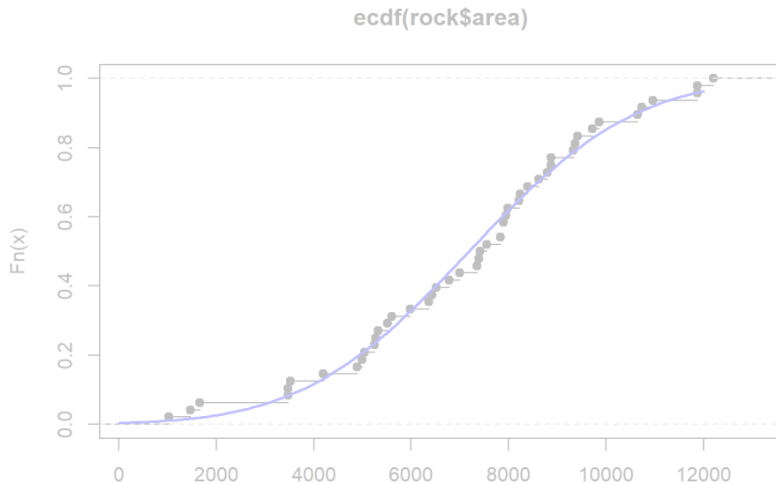
플랏으로 확인하는 것은 주관적일 수 있음

∴ 통계적 방법에 의한 가설 검정이 더 객관적



진단 | ② 정규성 검정

가설

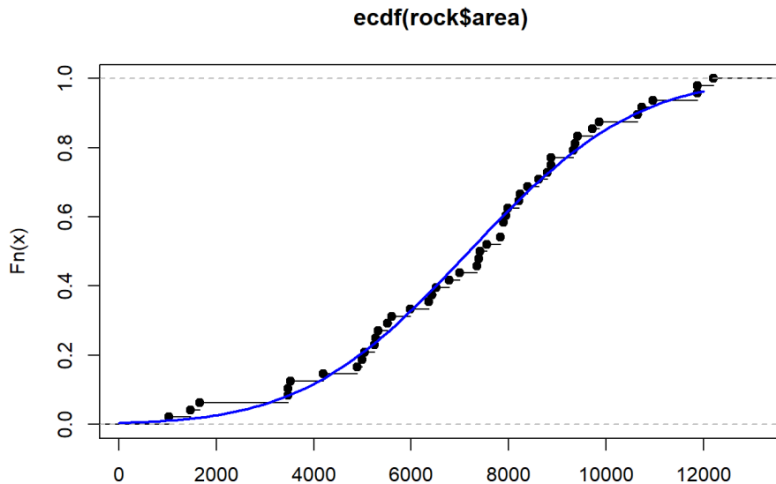
 H_0 : 주어진 데이터는 정규분포를 따른다. H_1 : 주어진 데이터는 정규분포를 따르지 않는다.

▶ 관측치들을 작은 순서대로 나열한 후,
관측치들로 **누적 분포 함수**를 그린 것

▶ 정규성 검정을 위해 잔차의 ECDF와
정규분포의 CDF를 비교하여 검정

진단 | ② 정규성 검정 | ① Empirical CDF를 이용한 test

Empirical CDF를 이용한 test



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

진단 | ② 정규성 검정 | ① Empirical CDF를 이용한 test

Empirical CDF를 이용한 test



Anderson Darling Test

```
#Anderson Darling Test  
library(nortest)  
ad.test(fit$residuals)
```

▶ 관측치들을 작은 순서대로 나열한 후,
관측치들을 작은 순서대로 나열한 후,

진단 | ② 정규성 검정 | ① Empirical CDF를 이용한 test

Empirical CDF를 이용한 test



Kolmogorov Smirnov Test

```
#Kolmogorov Smirnov Test  
ks.test(x=fit$residuals, y="pnorm")
```

▶ 관측치들을 작은 순서대로 나열한 후,
관측치들보다 정규분포 함수를 그릴 것

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

정규분포의 분포적 특성을 이용하는 test



Shapiro Wilk test

```
#Shapiro Wilk Test  
shapiro.test(fit$residuals)
```

소표본 데이터에서만 가능

우리의 데이터가 정규분포를 따른다고 가정 후, **검정통계량** 계산

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

정규분포의 분포적 특성을 이용하는 test



Jarque - Bera test

```
#Jarque-Bera Test  
library(tseries)  
jarque.bera.test(fit$residuals)
```

정규분포의 왜도가 0, 첨도가 3이라는 점에서 기반하는 방법
잔차의 왜도와 첨도를 정규분포의 것과 비교하여 검정 통계량 계산

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

정규분포의 분포적 특성을 이용하는 test



Jarque - Bera test

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

정규분포의 왜도가 0, 첨도가 3이라는 점에서 기반하는 방법

잔차의 왜도와 첨도를 정규분포의 것과 비교하여 검정 통계량 계산

4

정규성 진단과 처방

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



정규분포의 부포적 특성을 이용하는 test

정규성이 위반되었을 경우

Jarque - Bera 검정통계량이 t분포 또는 F분포를 따르지 않게 됨

t분포, F분포는 정규분포에서 파생되므로!

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

가설 검정 결과가 p-value에 의해 유의하게 나와도,

결과 신뢰할 수 없음

예측 결과 또한 신뢰하기 어려움!

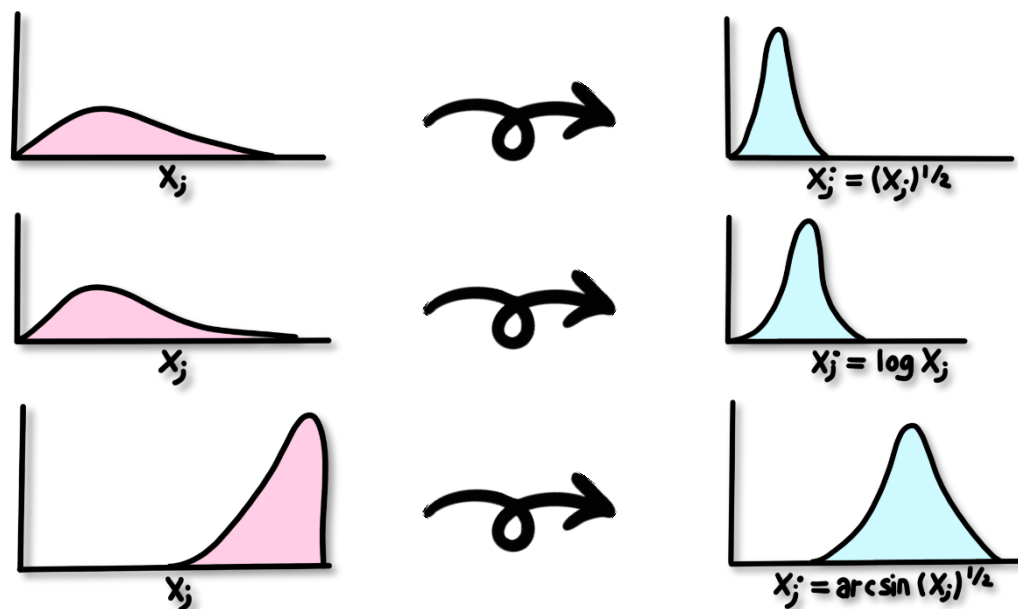
4

정규성 진단과 처방

처방 | ① 변수 변환

분포를 보고 주관적인 판단 하에 변수 변환

▶ 객관성 확보 어려움



처방 | ② Box-Cox Transformation

λ 를 변화시키면서 y 가 정규성을 만족하도록 만드는 방법

▶ 통계적인 방법에 의한 변수 변환



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

일반적으로 λ 는 -5에서 5 사이의 값을 사용
 λ 가 0인 경우, log-transformation을 적용



4

정규성 진단과 처방

처방 | ② Box-Cox Transformation

• 최적의 λ 를 구하는 방법 λ 를 변화시키면서 y 가 정규성을 만족하도록 만드는 방법

ML방법을 통해 신뢰구간을 구하고

신뢰구간 내의 로그우도함수를 최대화하는 λ 선택

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

$$L(\lambda) = -\frac{1}{2}n \cdot \ln(SSR(\lambda))$$

일반적으로 $L(\hat{\lambda}) - L(\lambda) \leq \frac{1}{2} \chi_{\alpha,1}^2$ 값을 사용
 λ 가 0인 경우, log-transformation을 적용

$$SSR(\lambda) \leq SSR(\hat{\lambda}) e^{\chi_{\alpha,1}^2/n}$$

4

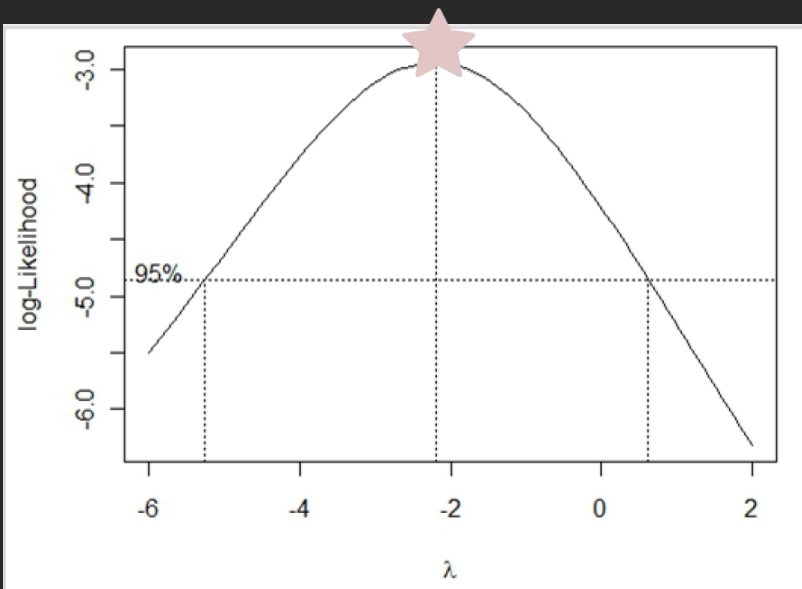
정규성 진단과 처방

처방 | ② Box-Cox Transformation

최적의 λ 를 구하는 방법

```
#load car library
library(car)

#take a value of lambda
trans = powerTransform(data$variable)
summary(trans)
```



성을 만족하도록 만드는 방법

에 의한 변수 변환

95% 내의 λ 값 중 가능도함수가

최대가 되는 λ 값 찾기

1 $if \lambda \neq 0$

(y) $if \lambda = 0$ -2근방의 λ 선택

-2처럼 정수로 λ 를 선택한다면

변수 변환 관계를 쉽게 파악 가능

서 5 사이의 값을 사용

ansformation을 적용

처방 | ② Box-Cox Transformation



Box-Cox Transformation은 y 가 $\log(y)$ 로 변환될 수 있으므로
 y 가 **항상 양수**여야 하는 단점 존재



해결법 ①

y 에 일정 값을 더해주기

▶ λ_2 를 넣어 양수가 되도록 함

$$y(\lambda_1, \lambda_2) = \begin{cases} \frac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ \log(y + \lambda_2) & \text{if } \lambda_1 = 0 \end{cases}$$

처방 | ② Box-Cox Transformation



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

y 가 항상 양수여야 하는 단점 존재



해결법 ①

y 에 일정 값을 더해주기

▶ λ_2 를 넣어 양수가 되도록 함

$$y(\lambda_1, \lambda_2) = \begin{cases} \frac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ \log(y + \lambda_2) & \text{if } \lambda_1 = 0 \end{cases}$$

처방 | ③ Yeo-Johnson Transformation



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

y 가 항상 양수여야 하는 단점 존재



해결법 ②

Yeo-Johnson Transformation

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

처방 | ③ Yeo-Johnson Transformation



Box-Cox Transformation은 y 가 $\log(y)$ 로 변환될 수 있으므로
 y 가 항상 양수여야 하는 단점 존재



해결법 ②

Yeo-Johnson Transformation

R코드

```
#take a value of lambda
trans = powerTransform(data$variable, family = "yjpower")
summary(trans)
```

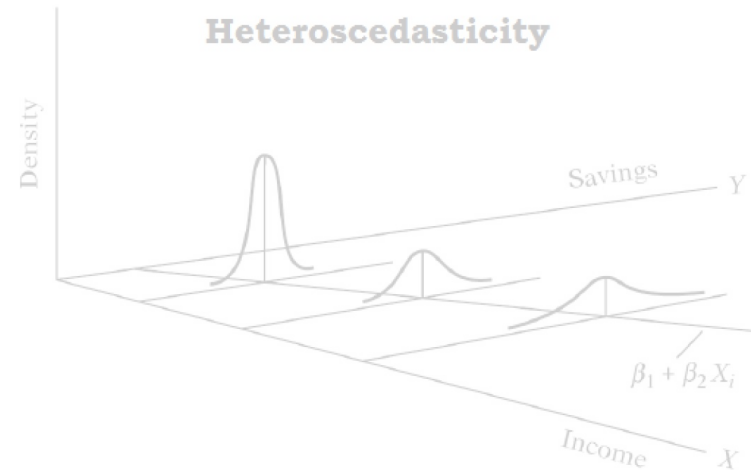
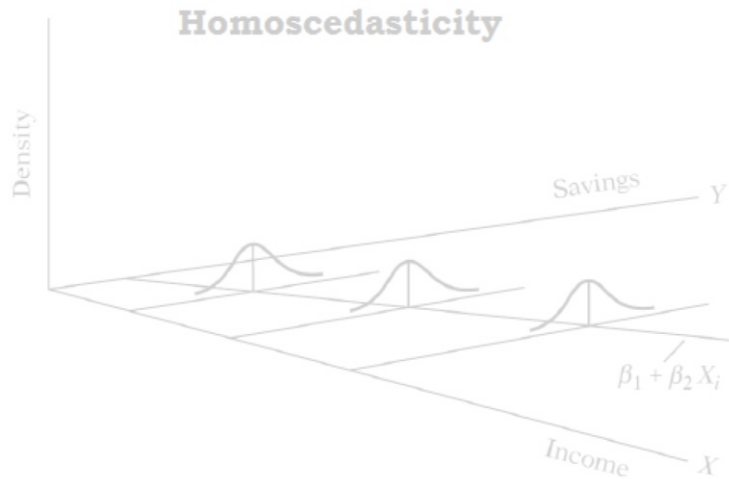
5

등분산성 진단과 처방

등분산성 가정

등분산성 가정

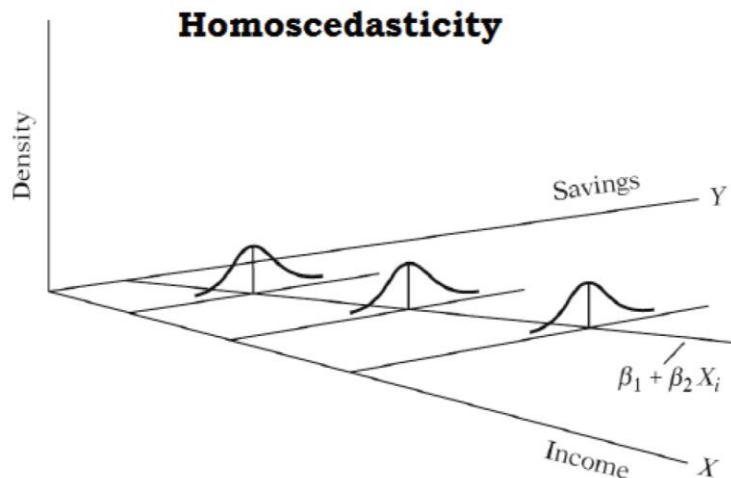
오차항의 분산이 어느 관측치에서나 동일하며
다른변수의 영향을 받지 않는 즉, **상수**라는 가정



등분산성 가정

등분산성 가정

오차항의 분산이 어느 관측치에서나 동일하며
다른변수의 영향을 받지 않는 즉, **상수**라는 가정



데이터 어느 곳에서도
 y 의 조건부 분포의 모양이 같음



오차의 분산은 **등분산**을 이룸

등분산성 가정

등분산성 가정

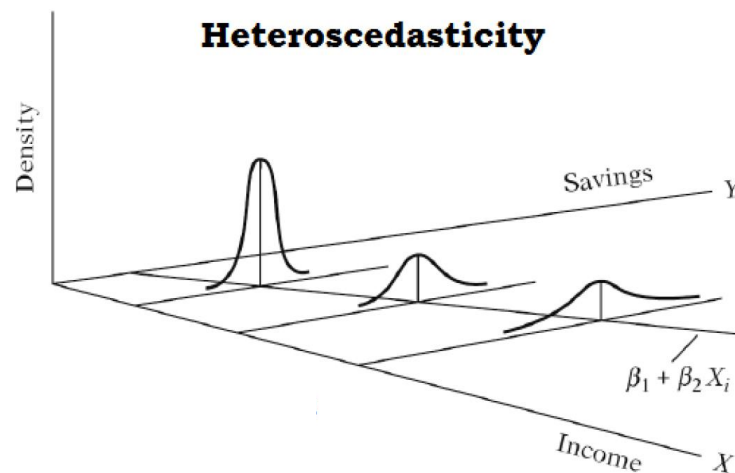
오차항의 분산이 어느 관측치에서나 동일하며
다른변수의 영향을 받지 않는 즉, **상수**라는 가정



지점에 따라
 y 의 조건부 분포의 모양이
같지 않음



등분산이 **위배**됨



등분산성 가정

등분산성 가정



오차항의 분산이 어느 관측치에서나 동일하며
다른 변수의 영향을 받지 않는 즉 상수라는 가정

이분산 (Heteroscedasticity)

오차항의 분산이 상수가 아니다!

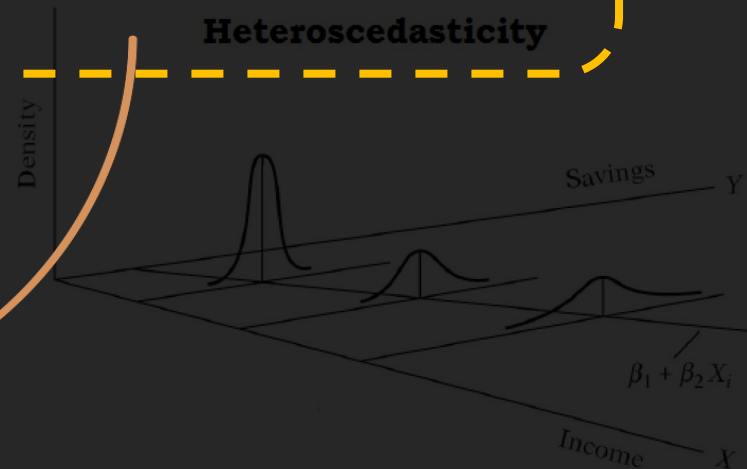
지점에 따라

y 의 조건부 분포의 모양이

같지 않음



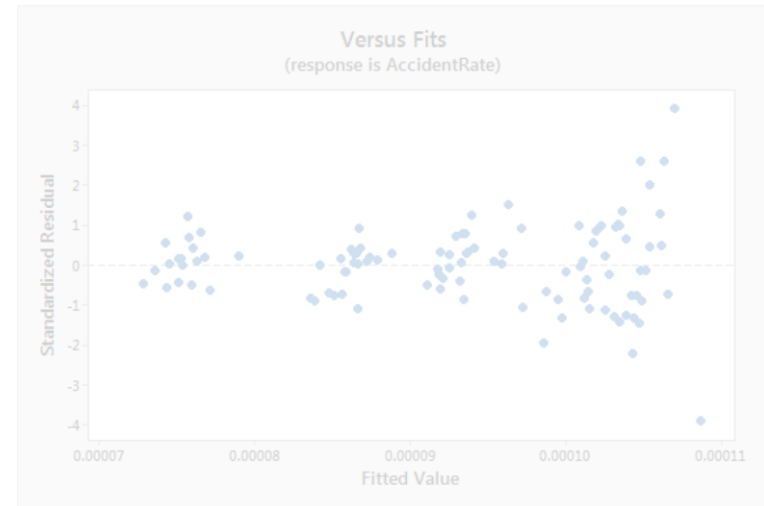
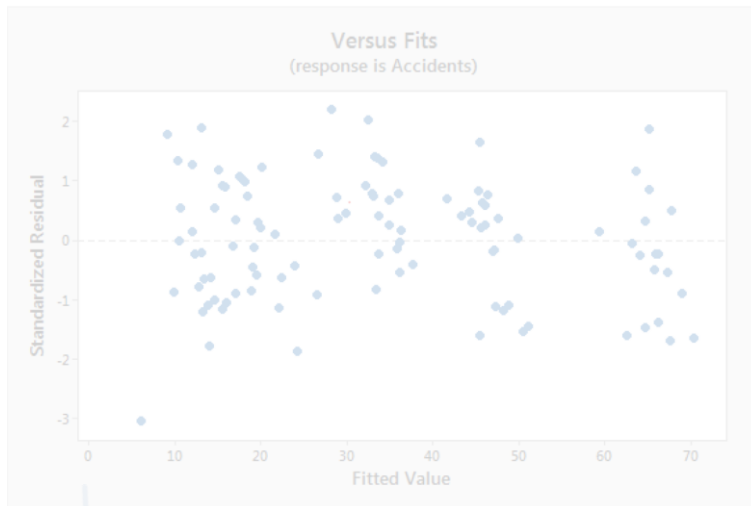
등분산이 **위배**됨



진단 | ① 잔차 플랏



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

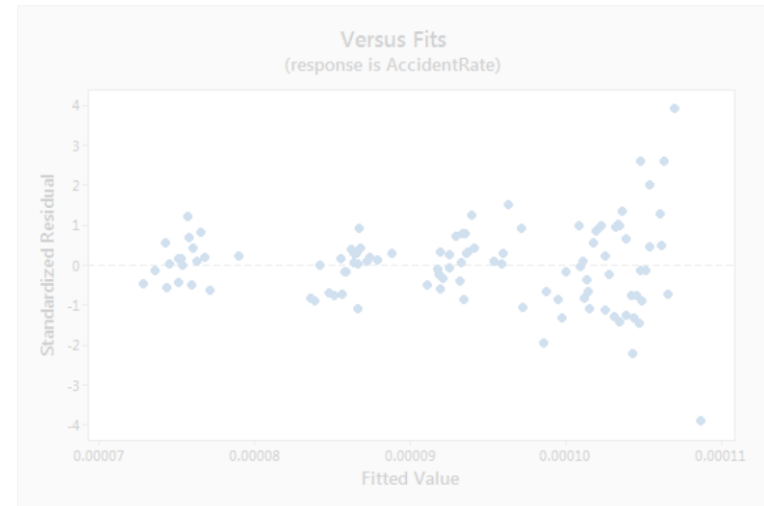
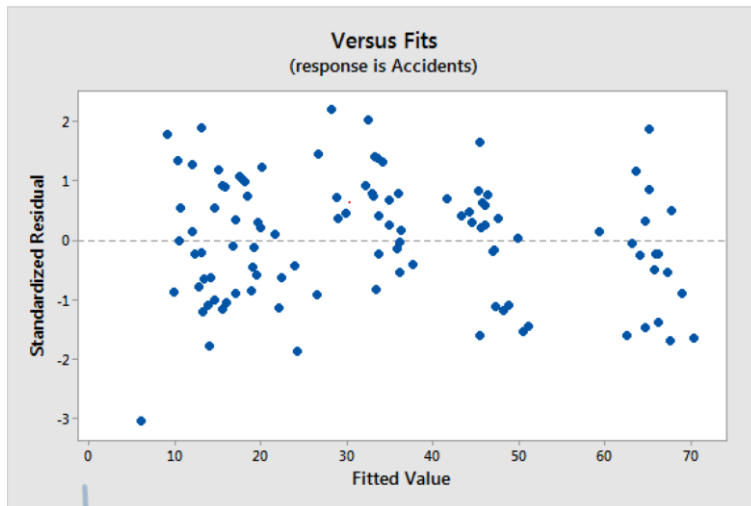


잔차가 random하게 분포 ➡ 등분산성을 땀

진단 | ① 잔차 플랏



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

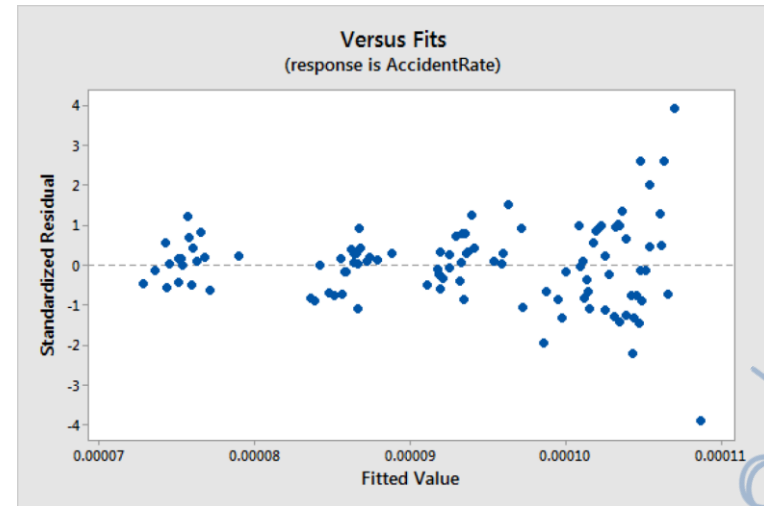
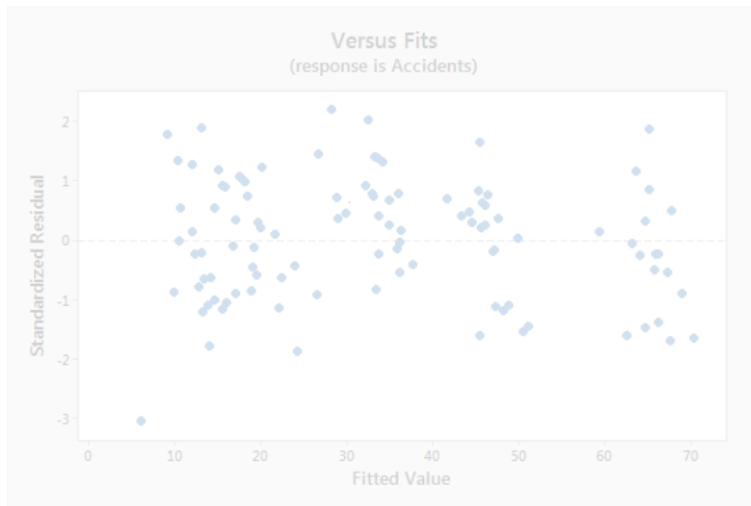


잔차가 random하게 분포 ➡ 등분산성을 땀

진단 | ① 잔차 플랏



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



\hat{y} 값이 커짐에 따라 잔차의 절대값이 커지는 형태 ➡ 이분산성을 땀

진단 | ② Breusch-Pagan test (BP test)

Breusch-Pagan test (BP test)

오차의 분산이 등분산인지 아닌지 판단하는 검정 방법



잔차가 독립변수들의 선형결합으로 표현되는지를
검정하기 위한 아이디어에서 출발

가설설정

H_0 : 주어진 데이터는 등분산성을 지님

H_1 : 주어진 데이터는 등분산성을 지니지 않음

임계값


$$\chi_{p-1, \alpha}^2$$

진단 | ② Breusch-Pagan test (BP test)

Breusch-Pagan test (BP test)

오차의 분산이 등분산인지 아닌지 판단하는 검정 방법



잔차가 독립변수들의 선형결합으로 표현되는지를
검정하기 위한 아이디어에서 출발

가설설정

H_0 : 주어진 데이터는 등분산성을 지님

H_1 : 주어진 데이터는 등분산성을 지니지 않음

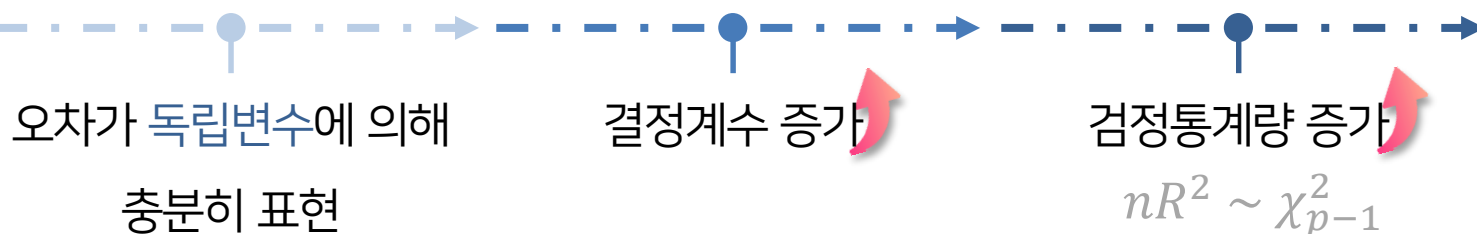
임계값


$$\chi_{p-1, \alpha}^2$$

진단 | ② Breusch-Pagan test (BP test)

검정통계량

$$e^2 = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_P X_P + \epsilon'$$



한계점

비선형적 결합으로 이루어진 이분산성은 파악 불가

Sample의 크기가 커야 (대표본) 사용가능



진단 | ② Breusch-Pagan test (BP test)

검정통계량

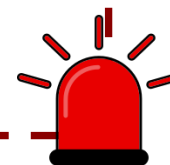
$$e^2 = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p + \epsilon'$$



한계점

비선형적 결합으로 이루어진 이분산성은 파악 불가

Sample의 크기가 커야 (대표본) 사용가능



진단 | ② Breusch-Pagan test (BP test)



검정통계량 $e^2 = v_0 + v_1 X_1 + \dots + v_p X_p + \epsilon'$

등분산성이 위반되었을 경우

오차가 독립변수에 의해

결정계수 증가

검정통계량 증가

충분히 OLS 추정량의 분산은 실제 분산보다 작게 추정됨 χ^2_{p-1}

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



진단 | ② Breusch-Pagan test (BP test)



검정통계량 $e^2 = \nu_0 + \nu_1 X_1 + \dots + \nu_p X_p + \epsilon'$



오차가 독립변수에 의해

결정계수 증가

검정통계량 증가

충분히 표현

1종 오류를 범하게 되며

$$nR^2 \sim \chi_{p-1}^2$$

한계점

가설 검정의 신뢰성이 떨어짐

비선형적 결합으로 이루어진 이분산성은 파악 불가

Sample의 크기가 커야 (대표본) 사용가능



진단 | Wait a minute !



오차들의 평균 = 0

오차들의 분산은 σ^2 으로 동일

오차 간의 자기 상관성이 없음



Gauss-Markov Theorem에 의해

OLS 추정량은 BLUE(Best Linear Unbiased Estimator)

회귀분석 1주차 참고

진단 | Wait a minute !

?

등분산성 가정이 위배된다면?

오차들의 평균 = 0

오차들의 분산은 σ^2 으로 동일

오차 간의 자기 상관성이 없음

Gauss-Markov Theorem에 의해

OLS 추정량은 BLUE(Best Linear Unbiased Estimator)

회귀분석 1주차 참고



진단 | Wait a minute !



오차들의 평균 = 0

오차들의 분산은 σ^2 으로 동일 X

오차 간의 자기 상관성이 없음



OLS 추정량은

BLUE(Best Linear Unbiased Estimator)가 아니다!



처방 | ① 변수 변환

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



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

$$Y = WX\beta + \epsilon$$

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



처방 | ② 가중 회귀

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



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

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

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



처방 | ② 가중 회귀

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



Advantages

WLS를 통해 구한 추정량은
회귀 기본 가정 하에 다시 BLUE임!



처방 | ② 가중 회귀

변수 변환

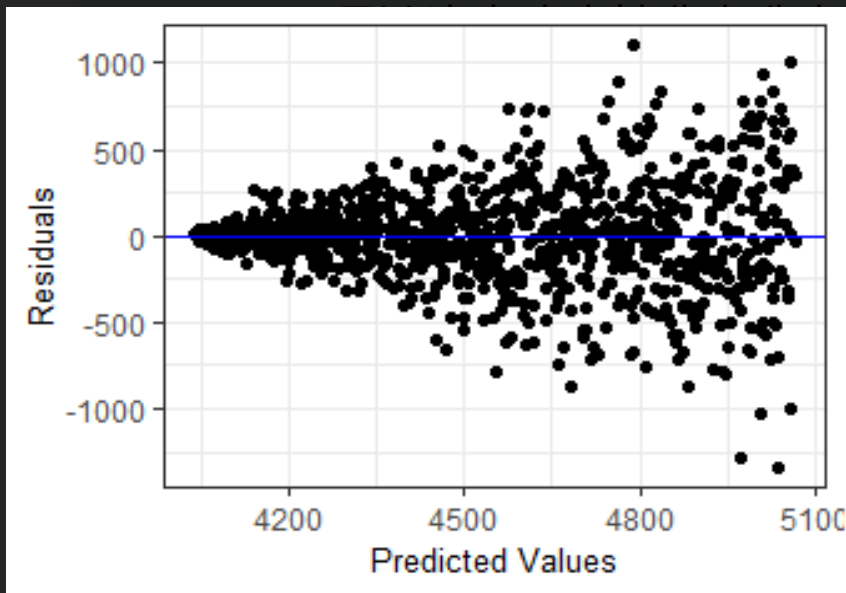


총 n개의 잔차들의 분산을 알기 어려움

∴ 잔차 플랏을 이용하여 판단

가중 회귀

(이 외 방법도 있음)



마다 다른 가중치를 주어서

변환된 '최소제곱법'의 한 형태

이렇게 residual plot에서

분산이 점점 커질 경우에는,

 $w_i \propto \frac{1}{x_i^2}$ 와 같은 방식으로 가중치 사용!


다시 BLUE임!

6

독립성 진단과 처방

독립성 가정

독립성 가정


$$\text{Cov}(e_i, e_j) = 0$$



오차들은 서로 독립이며, 개별 관측치에서 i 번째 오차와 j 번째 오차가 발생하는 것에 서로 영향을 미치지 않는다는 가정



독립성 가정이 **위배**된다면!

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

독립성 가정

독립성 가정

$$\text{Cov}(e_i, e_j) = 0$$

오차들은 서로 독립이며, 개별 관측치에서 i 번째 오차와 j 번째 오차가 발생하는 것에 서로 영향을 미치지 않는다는 가정



독립성 가정이 **위배**된다면!

- ▶ 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

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

진단 | ① 더빈-왓슨 검정

더빈-왓슨 검정

바로 앞 뒤 관측치의 1차 자기 상관성을 확인하는 검정 방법



가설설정

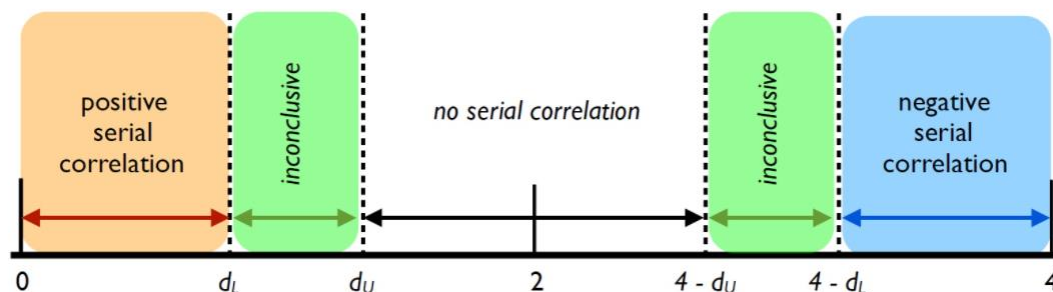
$$\hat{\rho}_1 = \frac{\widehat{Cov}(e_i, e_{i-1})}{\sqrt{V(e_i)} \cdot \sqrt{V(e_{i-1})}} \approx \frac{\sum_{i=2}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2}$$

$$\therefore d \approx 2(1 - \hat{\rho}_1)$$

검정통계량

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

진단 | ① 더빈-왓슨 검정



해석

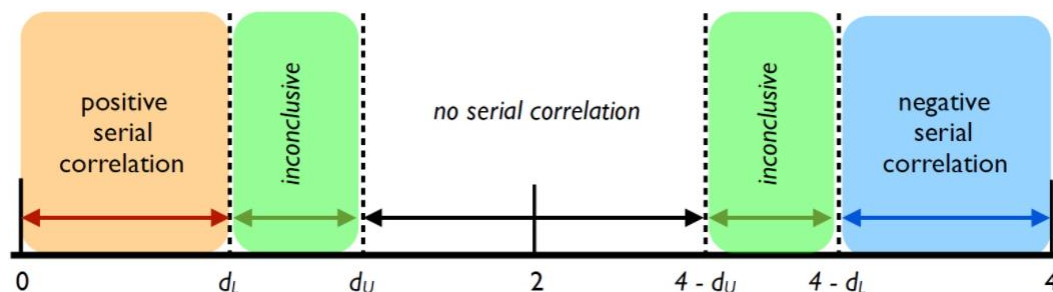
더빈 왓슨 검정 표에서 데이터 개수 n 과 변수의 개수 p 에 따라 귀무가설을 기각할 수 있는지 없는지 판단하는 **컷 오프 값**을 알려줌

✓ 귀무가설 기각 if 검정통계량 $d <$ 하한

▶ 양의 자기상관이 있다고 판단

✓ 귀무가설 기각 안됨 if 검정통계량 $d >$ 상한

진단 | ① 더빈-왓슨 검정

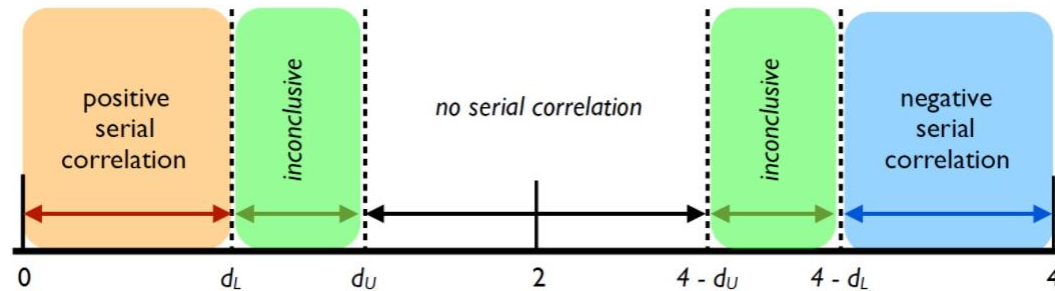


해석

더빈 왓슨 검정 표에서 데이터 개수 n 과 변수의 개수 p 에 따라 귀무가설을 기각할 수 있는지 없는지 판단하는 **컷 오프 값**을 알려줌

- ✓ 귀무가설 **기각** if 검정통계량 $d < \text{하한}$
 - ▶ 양의 자기상관이 있다고 판단
- ✓ 귀무가설 **기각 안됨** if 검정통계량 $d > \text{상한}$

진단 | ① 더빈-왓슨 검정



한계

D가 상한과 하한 사이에 위치할 때 **판단 할 수 없음**

바로 인접한 오차와의 **자기 상관만 고려**한다는 점에서 한계를 지님

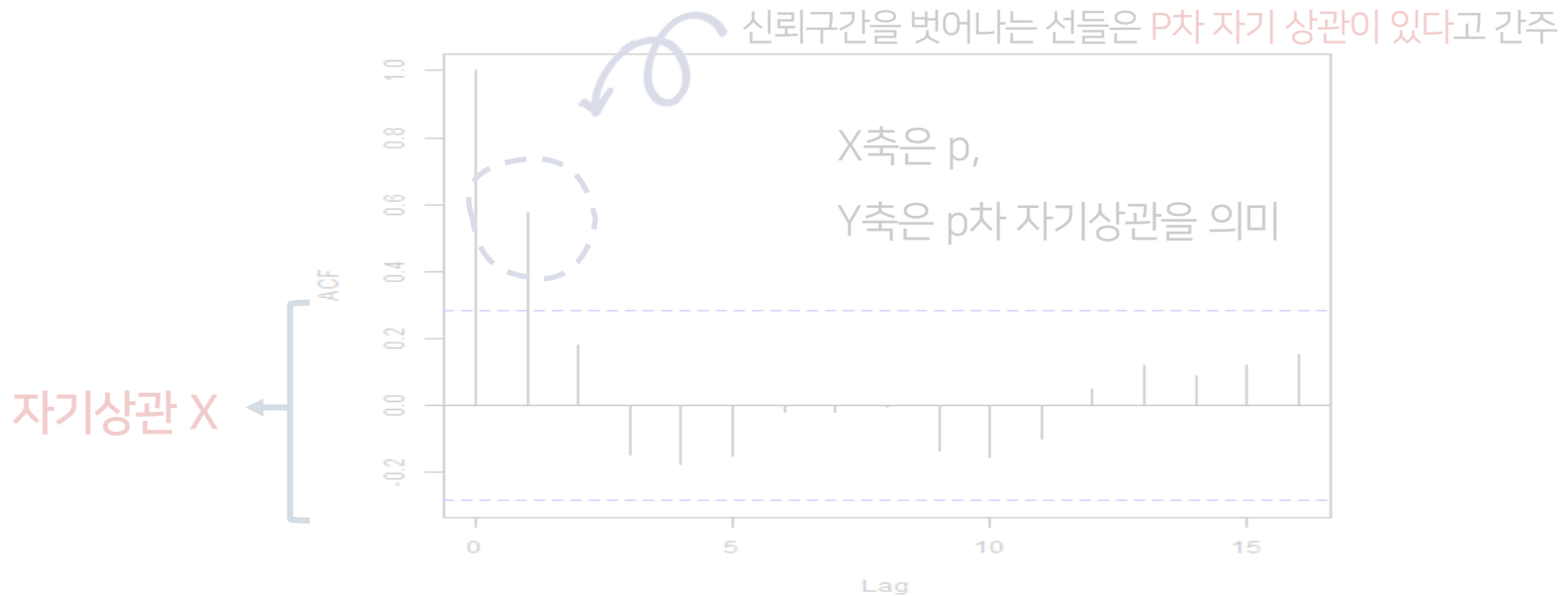
```
#Durbin Watson test
library(lmtest)
dwtest(fit) # fit은 적합한 모형
```



진단 | ② Autocorrelation function plot(ACF Plot)

Auto Correlation function

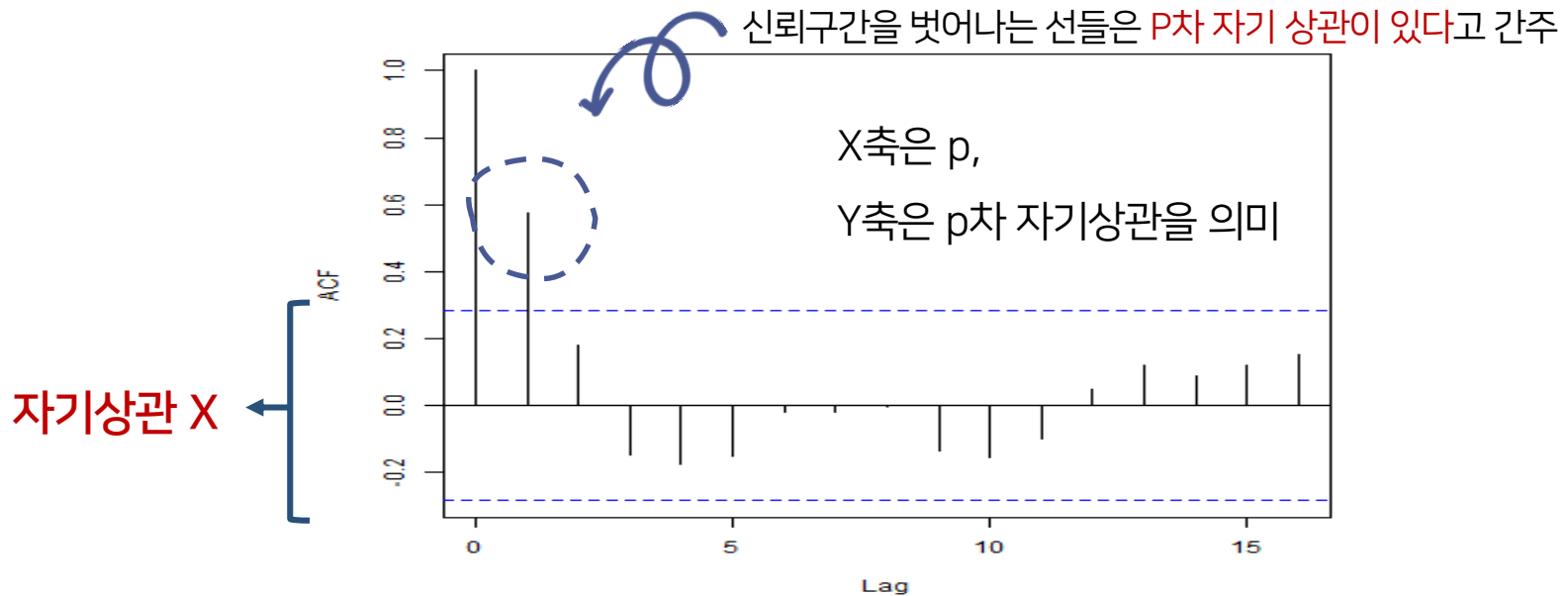
1차 자기상관부터 p차 자기 상관까지 고려하고,
신뢰구간을 반환하므로 통계적인 절차에 따라 판단 가능



진단 | ② Autocorrelation function plot(ACF Plot)

Auto Correlation function

1차 자기상관부터 p차 자기 상관까지 고려하고,
신뢰구간을 반환하므로 통계적인 절차에 따라 판단 가능



진단 | ② Autocorrelation function plot (ACF Plot)



Auto Correlation function

1차 자기상관부터 p차 자기 상관까지 고려하고,
신뢰구간을 반환하므로 통계적인 절차에 따라 판단 가능
독립성이 위반되었을 경우

신뢰구간을 벗어나는 선들은 p차 자기 상관이 있다고 간주

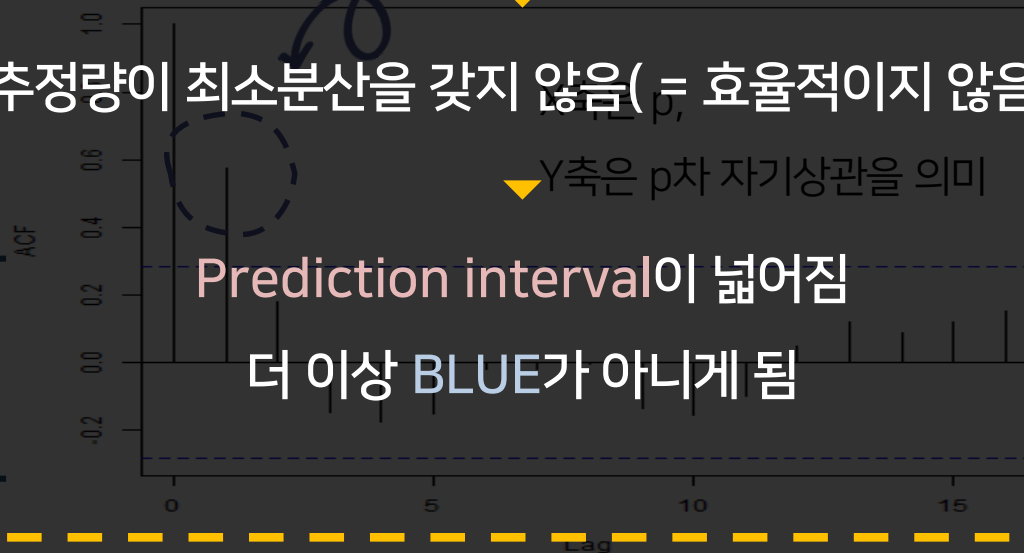
추정량이 최소분산을 갖지 않음 (= 효율적이지 않음)

Y축은 p차 자기상관을 의미

Prediction interval이 넓어짐

더 이상 BLUE가 아니게 됨

자기상관 X



처방 | ① 가변수 만들기

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



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



TIP our Team Is Perfect



선형성, 정규성, 등분산성을 한 번에 체크해 주는 함수!

gvlma package(Global Validation of Linear Model Assumption)



Global Stat : 선형성 판단



Skewness : 정규성 판단



Kurtosis : 정규성 판단



Heteroscedasticity: 등분산성 판단

다음 주 예고

다중공선성 문제

변수선택법

축소추정 방법