회귀분석팀

6팀

조수미
김민지
손재민
박윤아
조웅빈

CONTENTS

1

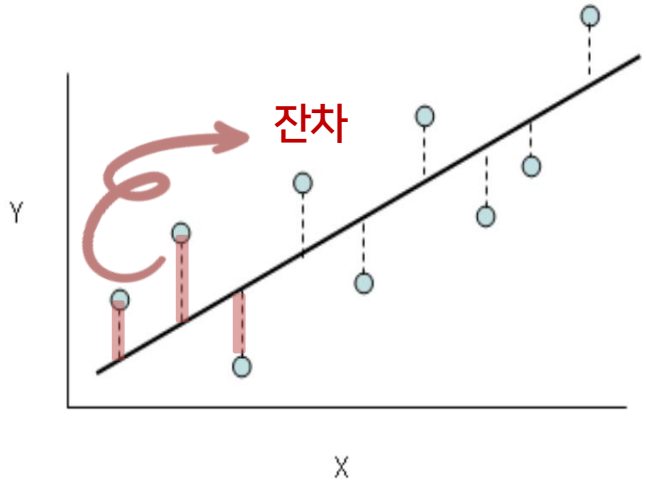
회귀 기본 가정

회귀 기본 가정

모델 가정의 목표

모형의 목표

잔차가 평균인 0으로 **회귀**하는 정확한 모델을 만드는 것



!!

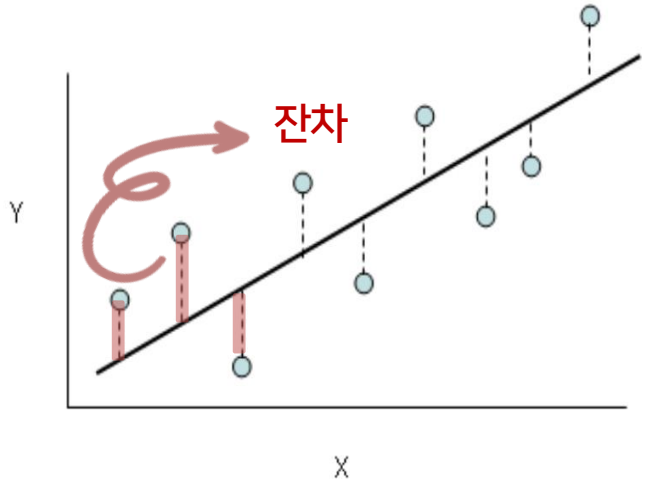
그러나, 추정한 모델과 실제 데이터 사이에는 **오차**가 발생

회귀 기본 가정

모델 가정의 목표

모형의 목표

잔차가 평균인 0으로 회귀하는 정확한 모델을 만드는 것



!!

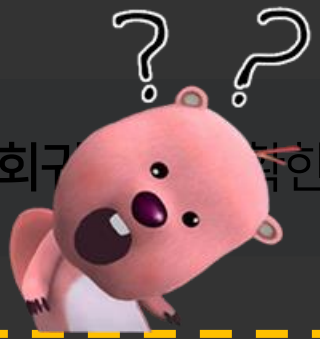
그러나, 추정한 모델과 실제 데이터 사이에는 **오차**가 발생

1 회귀 기본 가정

모델 가정의 목표

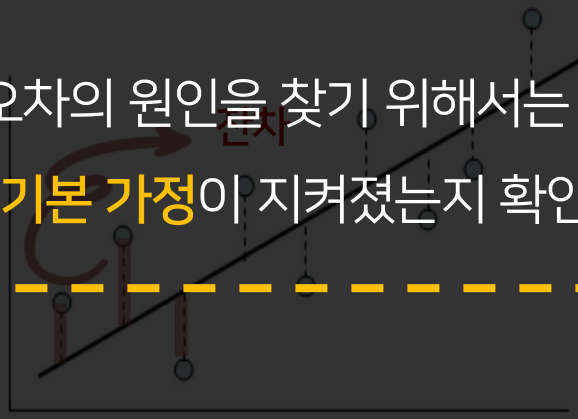
모형의 목표

잔차가 평균인 0으로 회귀[illegible]한 모델을 만드는 것



오차의 원인을 찾기 위해서는

회귀의 기본 가정이 지켜졌는지 확인 필요



!!

그러나, 추정한 모델과 실제 데이터 사이에는 **오차**가 발생

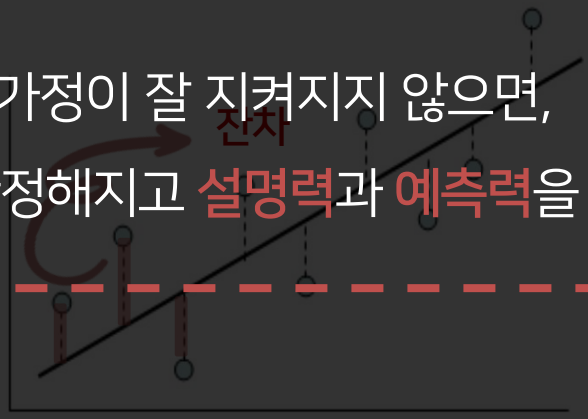
1 회귀 기본 가정

모델 가정의 목표

모형의 목표

잔차가 평균인 0으로 회 확한 모델을 만드는 것

가정이 잘 지켜지지 않으면,
모델이 불안정해지고 **설명력**과 **예측력**을 잃기 때문



!!

그러나, 추정한 모델과 실제 데이터 사이에는 **오차**가 발생

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

⑤ **오차의 평균은 0**

회귀 기본 가정

선형회귀분석 가정

회귀식

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

X와 Y의 관계가 선형으로 나타남

① **선형성** : 설명변수와 반응변수의 관계는 **선형**

② **오차의 정규성** : 오차항은 정규분포를 따름

③ **오차의 독립성** : 오차항은 서로 독립

④ **오차의 등분산성** : 오차항의 분산은 상수

⑤ **오차의 평균은 0**

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

⑤ 오차의 평균은 0

회귀 기본 가정

선형회귀분석 가정

회귀식

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

서로 다른 오차항 간 관계가 없어야 함

① **선형성** : 설명변수와 반응변수의 관계는 선형

② **오차의 정규성** : 오차항은 정규분포를 따름

③ **오차의 독립성** : 오차항은 **서로 독립**

④ **오차의 등분산성** : 오차항의 분산은 상수

⑤ **오차의 평균은 0**

회귀 기본 가정

선형회귀분석 가정

회귀식

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

서로 다른 오차항 간 분산이 일정해야 함

① **선형성** : 설명변수와 반응변수의 관계는 선형

② **오차의 정규성** : 오차항은 정규분포를 따름

③ **오차의 독립성** : 오차항은 서로 독립

④ **오차의 등분산성** : 오차항의 분산은 **상수**

⑤ 오차의 평균은 0

회귀 기본 가정

선형회귀분석 가정

회귀식

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

일반적으로 거의 위반되지 않는 가정

① **선형성** : 설명변수와 반응변수의 관계는 선형

② **오차의 정규성** : 오차항은 정규분포를 따름

③ **오차의 독립성** : 오차항은 서로 독립

④ **오차의 등분산성** : 오차항의 분산은 상수

⑤ **오차의 평균은 0**

1 회귀 기본 가정

선형회귀분석 가정

회귀식

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

일반적으로 거의 위반되지 않는 가정

오차의 평균이 0이라는 가정은 거의 위반되지 않으므로

선형성, 정규성, 독립성, 등분산성에 초점을 맞춰

진단과 처방을 살펴볼 예정

② 오차의 정규성 : 오차항은 정규분포를 따름

③ 오차의 독립성 : 오차항은 서로 독립

④ 오차의 등분산성 : 오차항의 분산은 상수

⑤ 오차의 평균은 0

2

잔차 플랏

잔차 플랏

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

잔차 플랏

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

잔차 플랏

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

잔차 플랏

시각적 방법

잔차 플랏 Residual Plot

잔차 분포를 통해 경험적 판단에 근거한 회귀진단이 가능

R의 plot() 함수를 통해 잔차의 분포를 쉽게 나타낼 수 있음

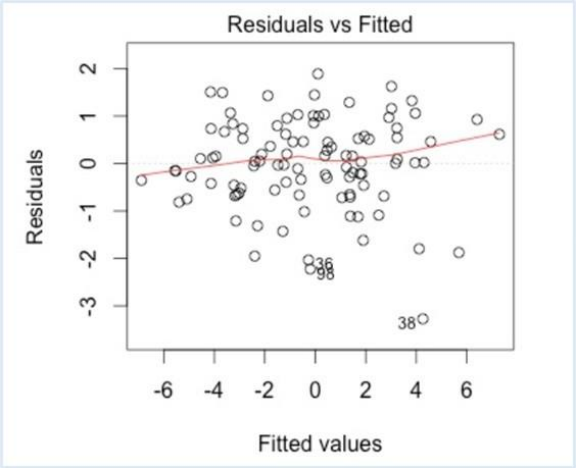
- Residuals vs Fitted
- Normal QQ Plot
- Scale-Location
- Residuals vs Leverage

잔차 플랏

잔차 플랏

Residuals vs Fitted

선형성과 오차의 등분산성 확인 가능



X축 : 예측값 $(\hat{y})$

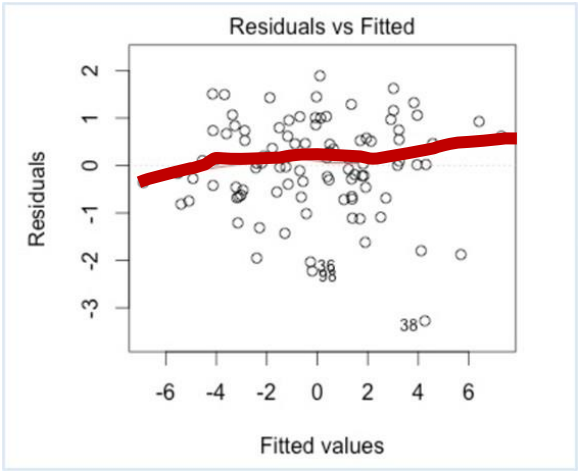
Y축 : 잔차 $(e = y - \hat{y})$

잔차 플랏

잔차 플랏

Residuals vs Fitted

선형성과 오차의 등분산성 확인 가능



!!

빨간 실선 : 전체적인 잔차들의 추세선으로,

잔차들의 분포를 Local Regression으로 추정한 직선

잔차 분포의 경향성을 나타내는 보조지표

잔차 플랏

잔차 플랏

Residuals vs Fitted

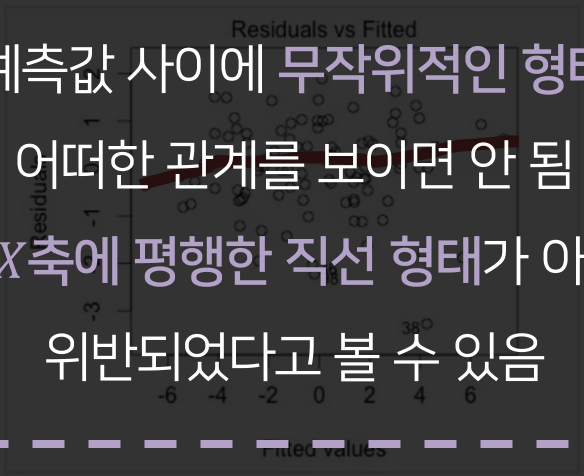
선형성과 오차의 **등분산성** 확인 가능

잔차와 예측값 사이에 **무작위적인 형태** 이외에

어떠한 관계를 보이면 안 됨

즉, 빨간 실선이 X**축에 평행한 직선 형태**가 아니라면 선형성이

위반되었다고 볼 수 있음



!!

빨간 실선 : 전체적인 잔차들의 추세선으로,

잔차들의 분포를 Local Regression으로 추정한 직선

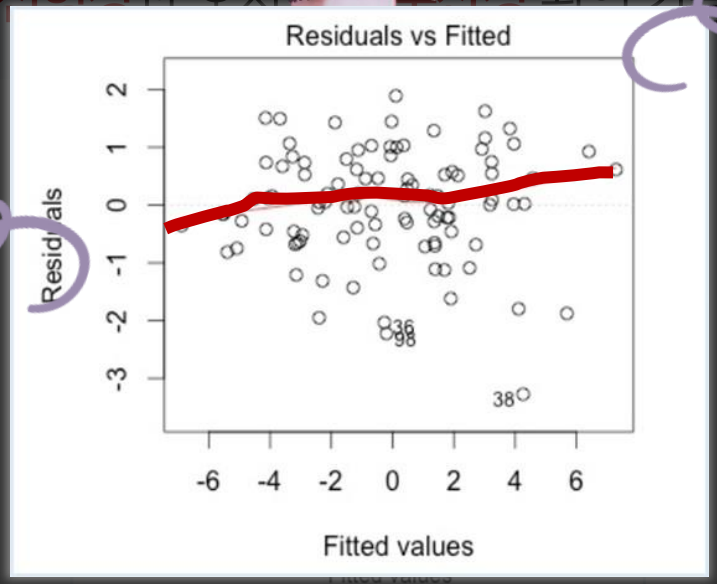
잔차 분포의 경향성을 나타내는 보조지표

잔차 플랏

잔차 플랏

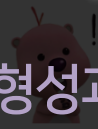
Residuals vs Fitted

선형성과 오차의 독립성 확인 가능



② 점들의 분포가 랜덤하게 퍼져있음

① 완만하게 수평을 이루고 있음



!!

선형성과 오차의 등분산성 가정이 위배되지 않음!

잔차들의 분포를 Local Regression으로 추정한 직선

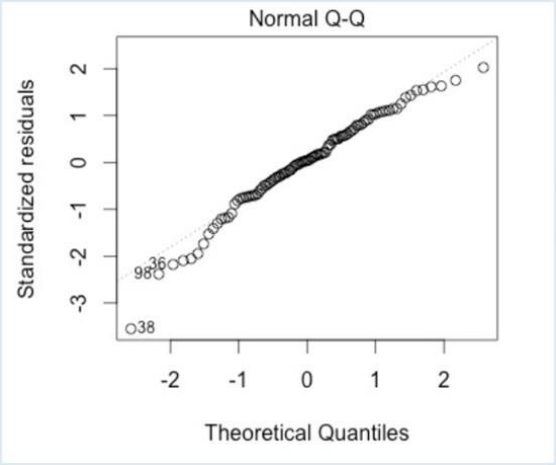
잔차 분포의 경향성을 나타내는 보조지표

잔차 플랏

잔차 플랏

Normal QQ Plot

정규성 확인 가능



X축 : 정규분포의 분위수 값

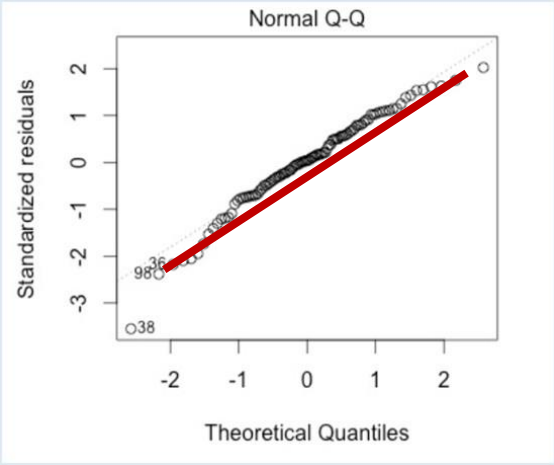
Y축 : 표준화 잔차 (r_i)

잔차 플랏

잔차 플랏

Normal QQ Plot

정규성 확인 가능



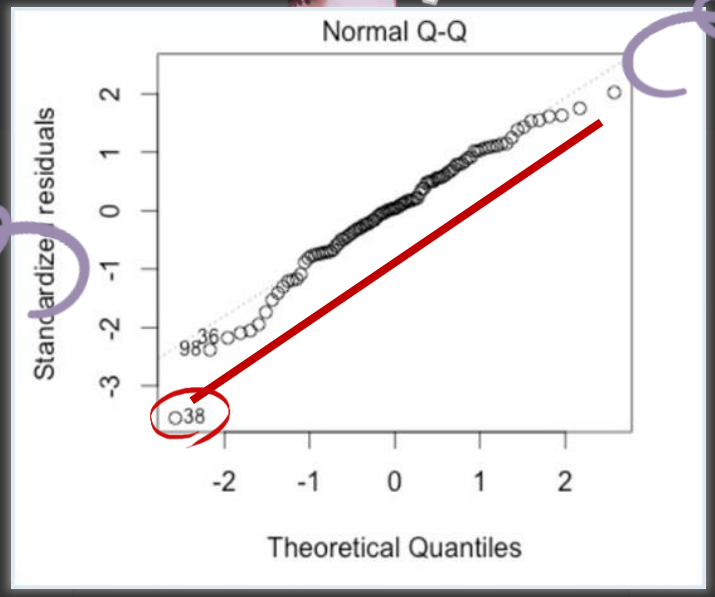
!! $y = x$에 가까울 수록, 잔차가 **정규성**을 만족함

직선이라는 것은 정규분포 사분위수 위에 그대로 위치한다는 의미

잔차 플랏

잔차 플랏

Normal QQ Plot



잔차 분포가 대부분
$y = x$를 따르고 있음

정규성을 만족하는
것처럼 보임

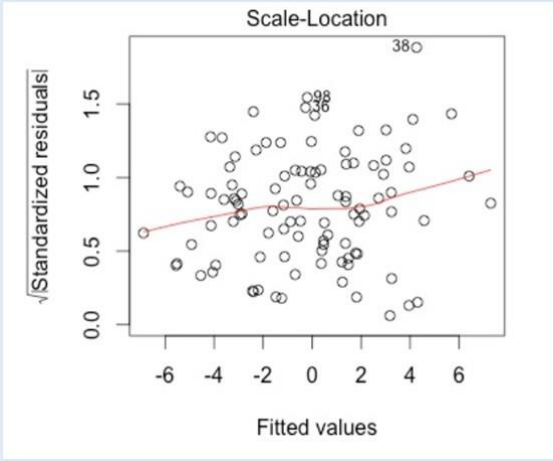
하지만 38번 관측치가 $y = x$에서 많이 벗어나므로
$y = x$에 가까울수록 잔차가 정규성을 만족함
추가적인 확인이 필요!
직선이라는 것은 정규분포 사분위수 위에 그대로 위치한다는 의미

잔차 플랏

잔차 플랏

Scale-Location

선형성과 오차의 등분산성 확인 가능, 보통 등분산성 고려



X축 : 예측값 $(\hat{y})$

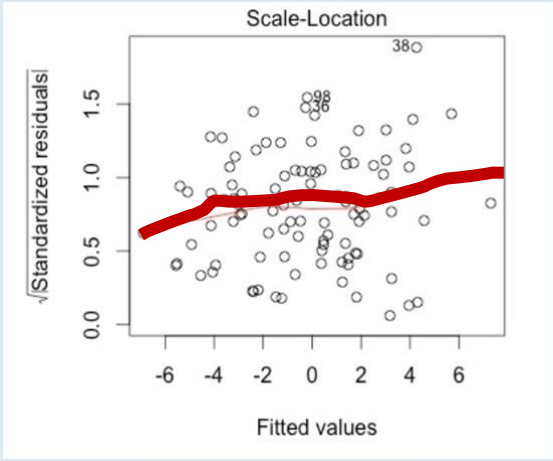
Y축 : 표준화잔차 절댓값 $(\sqrt{|e_i|/se(e_i)}\ )$

잔차 플랏

잔차 플랏

Scale-Location

선형성과 오차의 등분산성 확인 가능, **보통 등분산성 고려**



!!

빨간 실선 : Residuals vs Fitted와 비슷한 판단을 할 수 있음

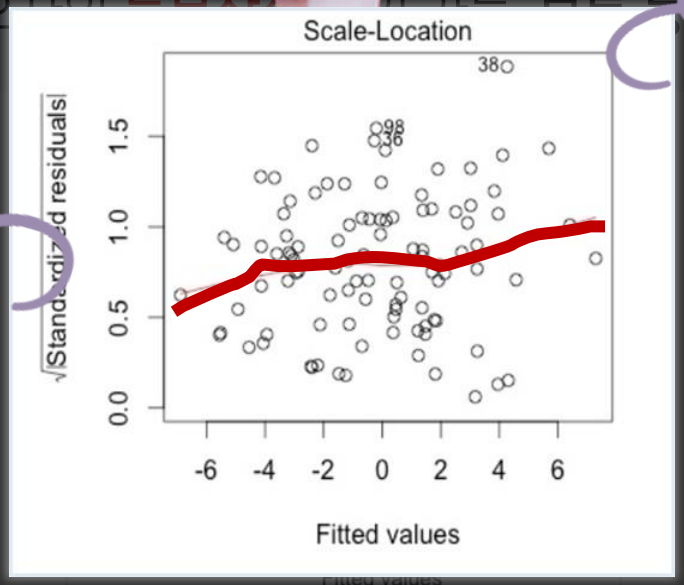
첫 번째 플랏과의 차이점은 **잔차에 절댓값이 씌워진 형태**

잔차 플랏

잔차 플랏

Scale-Location

선형성과 오차의 등분산성 가정 고려



② 점들의 분포가 랜덤하게 퍼져있음

① 완만하게 수평을 이루고 있음

!! 선형성과 오차의 등분산성 가정이 위배되지 않음!

빨간 실선 : Residuals vs Fitted와 비슷한 판단을 할 수 있음

첫 번째 플랏과의 차이점은 **잔차에 절댓값이 씌워진 형태**

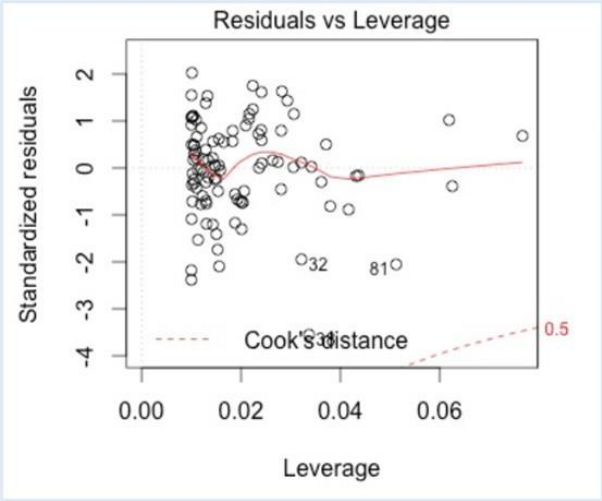
잔차 플랏

잔차 플랏

Residual vs Leverage

지난 주에 배운 영향점(Influential point)을 확인할 수 있음

선형성과 오차의 **등분산성** 확인 가능



X축 : 레버리지 (지렛값)

Y축 : 표준화 잔차 (r_i)

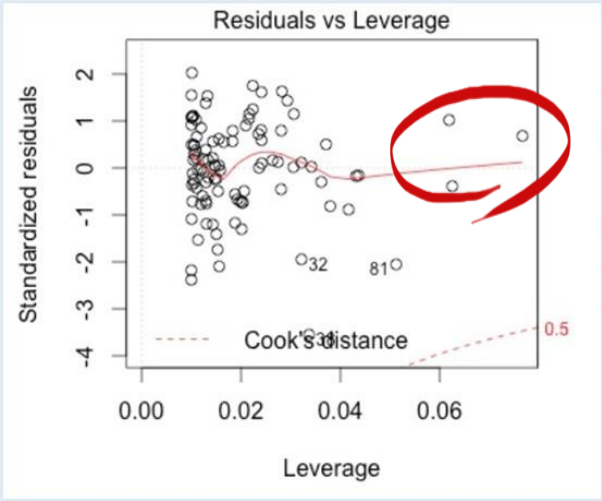
잔차 플랏

잔차 플랏

Residual vs Leverage

지난 주에 배운 영향점(Influential point)을 확인할 수 있음

선형성과 오차의 **등분산성** 확인 가능



!! 플랏의 **오른쪽의 위치**한 점들이 **leverage**가 큰 잔차

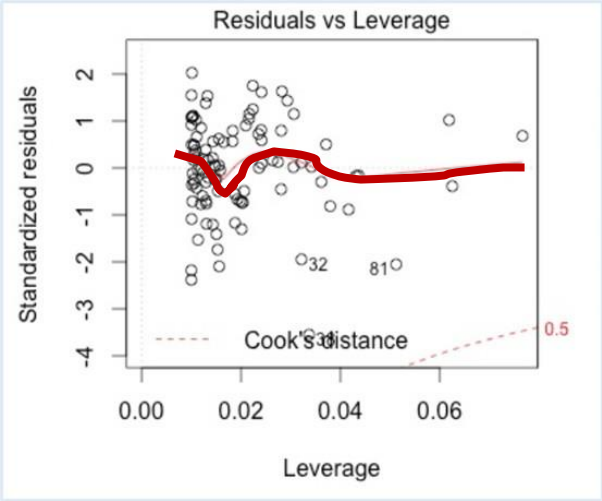
잔차 플랏

잔차 플랏

Residual vs Leverage

지난 주에 배운 영향점(Influential point)을 확인할 수 있음

선형성과 오차의 **등분산성** 확인 가능



!! **빨간 실선**으로부터 위아래로 멀리 떨어진 점들이 **outlier**

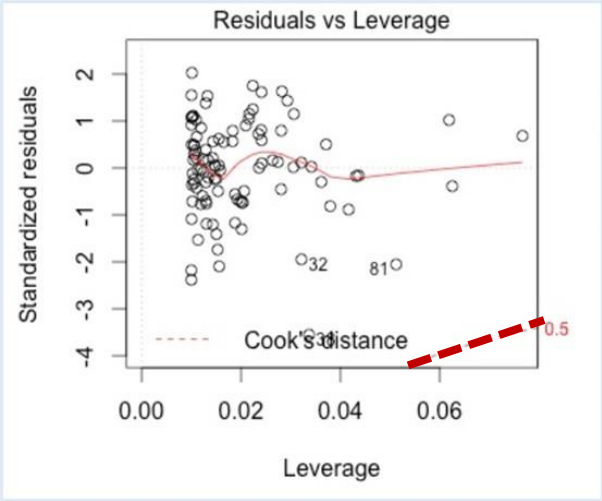
잔차 플랏

잔차 플랏

Residual vs Leverage

지난 주에 배운 영향점(Influential point)을 확인할 수 있음

선형성과 오차의 등분산성 확인 가능



!! 빨간 점선은 Cook's distance로 주로 0.5과 1에서의 경계가 표시됨

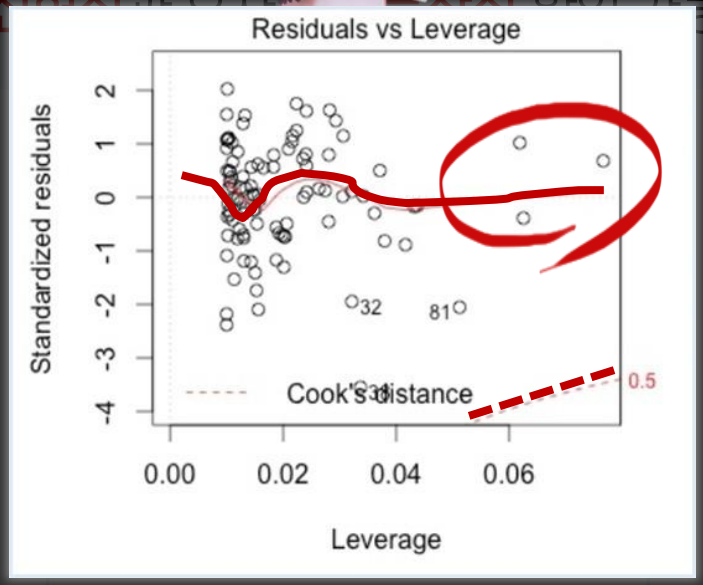
잔차 플랏

잔차 플랏

Residual vs Leverage

에운 영향점(Influential point)을 확인할 수 있음

선형성과 오차 [illegible] 확인 가능



Leverage

모든 관측치들이 0.5 경계 안에 있으므로 영향점은 딱히 없음!

빨간 점선은 Cook's distance로 주로 0.5과 1에서의 경계가 표시됨

잔차 플랏

잔차 플랏

Residual vs Leverage

주에 배운 영향점(Influential point)을 확인할 수 있음

선형성과 오차의 등분산성 확인 가능

이제 총 네 가지의 가정에 대해 자세하게 알아보자!

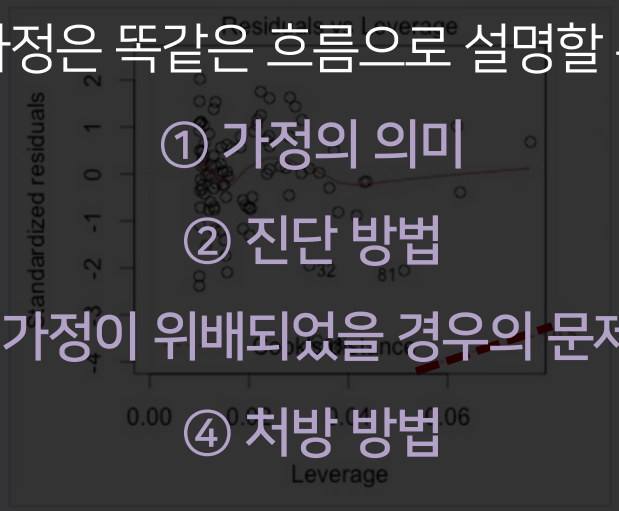
모든 가정은 똑같은 흐름으로 설명할 수 있음

① 가정의 의미

② 진단 방법

③ 가정이 위배되었을 경우의 문제점

④ 처방 방법



!!

빨간 점선은 Cook's distance로 주로 0.5과 1에서의 경계가 표시됨

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 위배되었다면?

↓

변수 변환이나 비선형 모델을 추정함으로써 대처할 수 있음

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 위배되었다면?

변수 변환이나 비선형 모델을 추정함으로써 대처할 수 있음

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 모델을 추정함으로써 대처할 수 있음

선형성 진단과 처방

선형성 가정

선형성 가정

반응변수 Y가 설명변수 X의 **선형결합**으로 이루어진다는 가정

단순선형회귀, 다중선형회귀 모두 선형성 가정에서 출발한 모델

↓

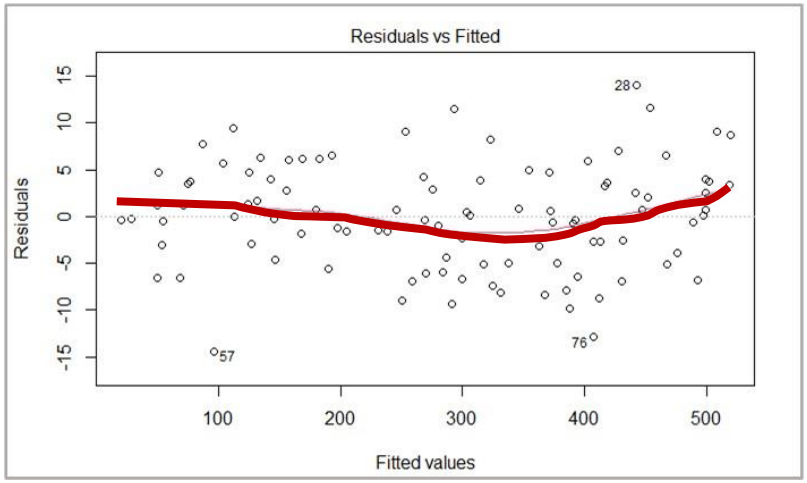
만약 **선형성 가정이 위배**되었다면?

↓

변수 변환이나 **비선형 모델**을 추정함으로써 대처할 수 있음

선형성 진단과 처방

진단 | 잔차 플랏



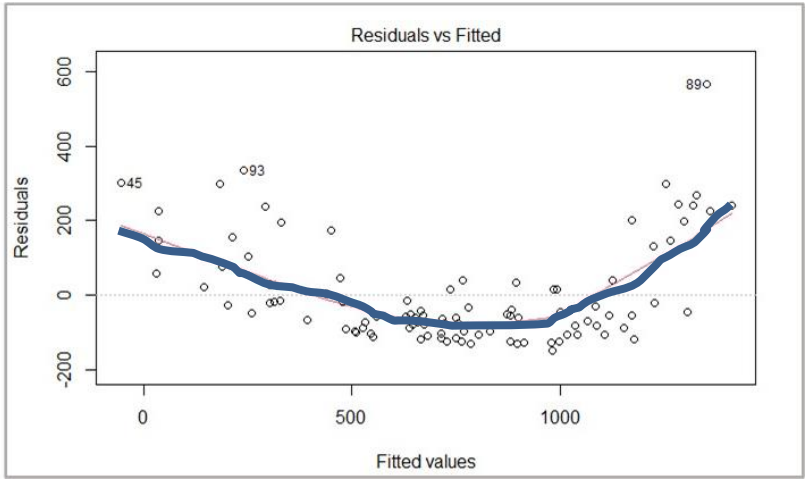
✓ **평균 0**을 중심으로 하는 ***X*축에 평행**한 직선 형태라면 **선형성을 만족**

✓ **빨간 실선**이 ***X*축과 완만하게 수평**을 이루고 있으며,

점들의 분포도 랜덤하게 퍼져있으므로 **선형성을 만족**

선형성 진단과 처방

진단 | 잔차 플랏

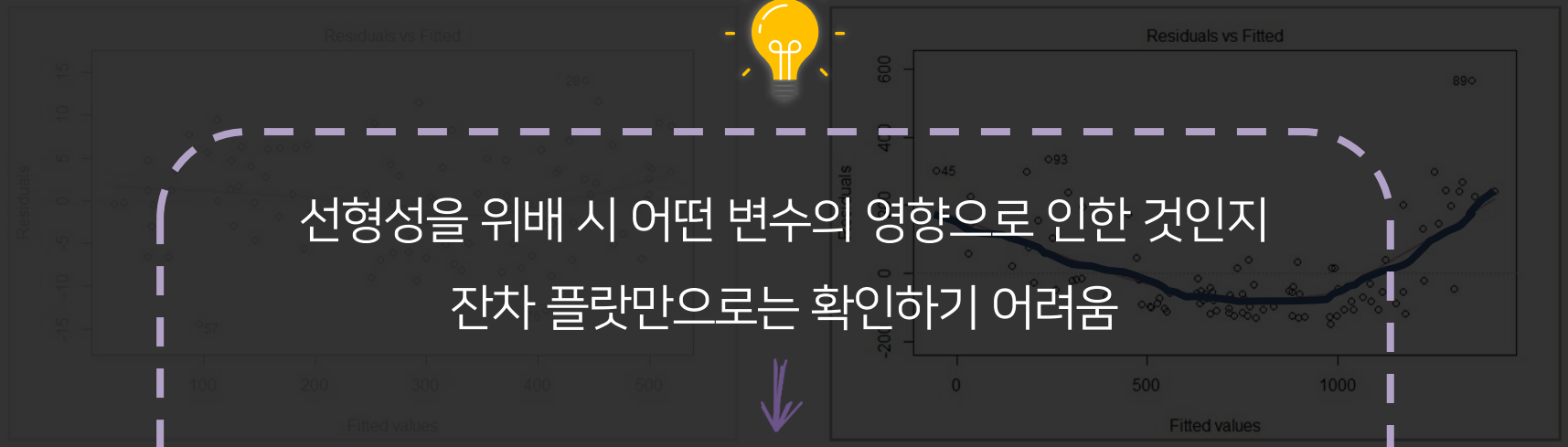


✓ **선형성이 위배**되는 보통의 경우, **이차함수** 혹은 **삼차함수** 형태처럼 나타남

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

3 선형성 진단과 처방

진단 | 잔차 플랏



선형성을 위배 시 어떤 변수의 영향으로 인한 것인지

잔차 플랏만으로는 확인하기 어려움

어떤 변수의 영향으로 인한 것인지

개별변수마다 선형성을 확인해야 함

✓ 선형성이 위배되는 보통의 경우, **이차함수** 혹은 **삼차함수** 형태처럼 나타남

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

선형성 진단과 처방

진단 | Partial Residual Plot

실선

잔차의 추세선

점선

최소제곱법을 통해

회귀선을 추정한 것

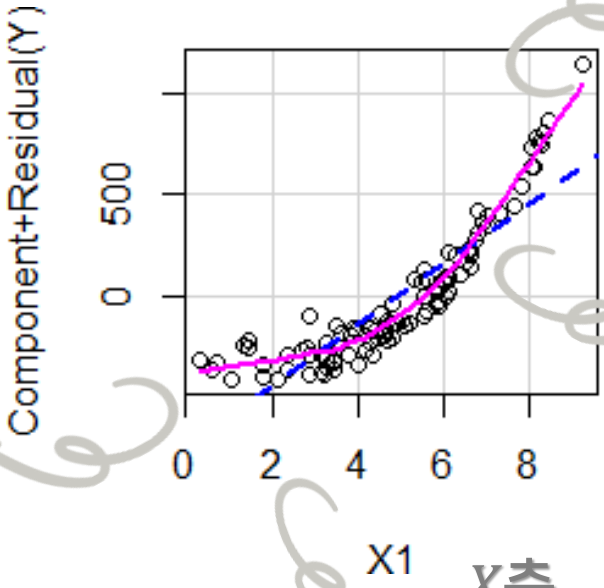
Y축

선형성을 보고싶은

변수를 제외한

나머지 변수로

회귀식을 적합한 잔차



이렇게 일부의 변수로 적합한

모델의 잔차를 이용하기 때문에

Partial residual plot이라고 불림

X축

선형성을 판단하기

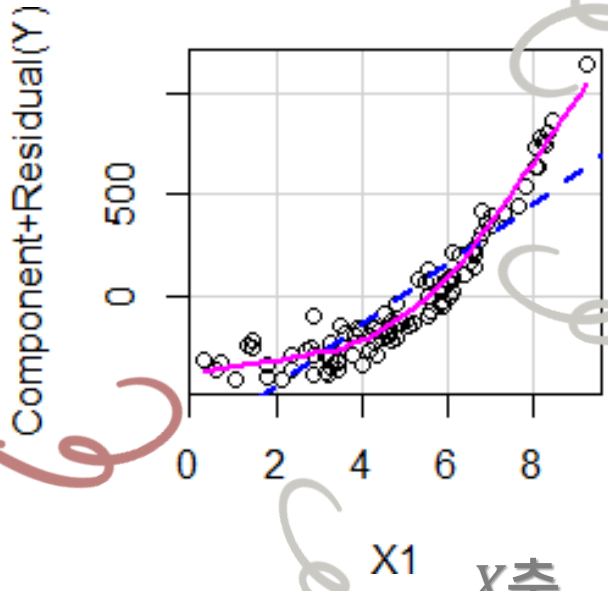
위한 변수 x_i

선형성 진단과 처방

진단 | Partial Residual Plot

Y축

선형성을 보고싶은
변수를 제외한
나머지 변수로
회귀식을 적합한 잔차



실선

잔차의 추세선

점선

최소제곱법을 통해
회귀선을 추정한 것

X축

선형성을 판단하기
위한 변수 x_i

이렇게 일부의 변수로 적합한
모델의 잔차를 이용하기 때문에
Partial residual plot이라고 불림

선형성 진단과 처방

진단 | Partial Residual Plot

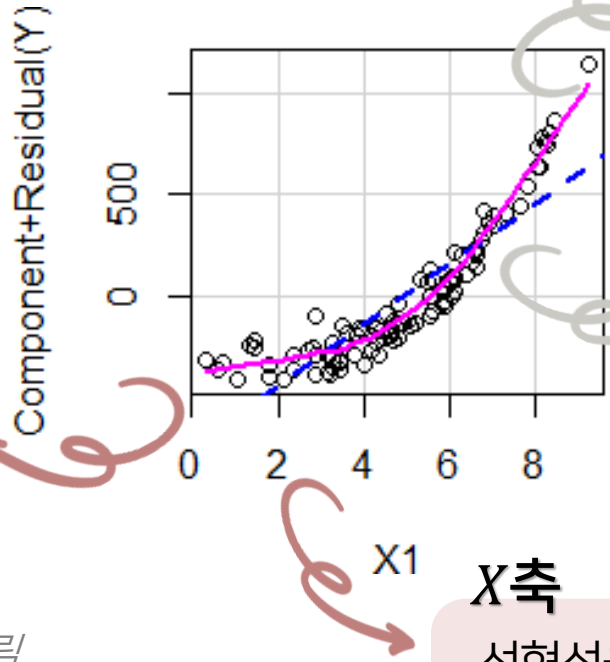
실선

잔차의 추세선

점선

최소제곱법을 통해

회귀선을 추정한 것



Y축

선형성을 보고싶은

변수를 제외한

나머지 변수로

회귀식을 적합한 잔차

이렇게 일부의 변수로 적합한

모델의 잔차를 이용하기 때문에

Partial residual plot이라고 불림

X축

선형성을 판단하기

위한 변수 x_i

선형성 진단과 처방

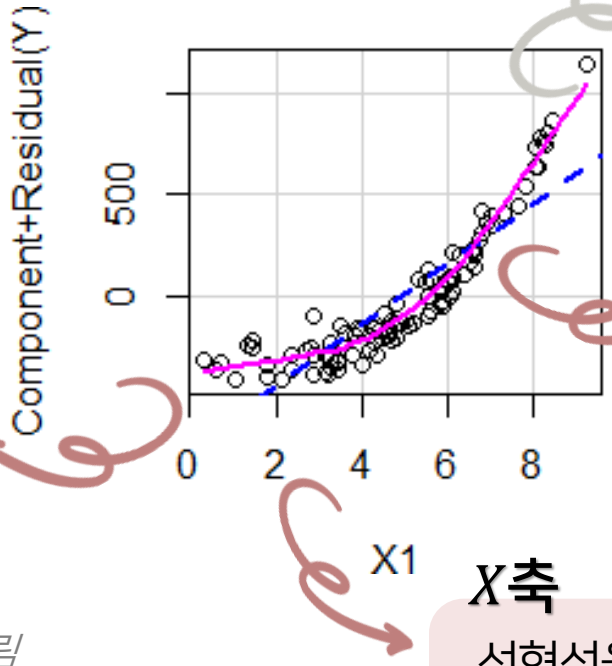
진단 | Partial Residual Plot

실선

잔차의 추세선

Y축

선형성을 보고싶은 변수를 제외한 나머지 변수로 회귀식을 적합한 잔차



점선

최소제곱법을 통해 회귀선을 추정한 것

이렇게 일부의 변수로 적합한 모델의 잔차를 이용하기 때문에 Partial residual plot이라고 불림

X축

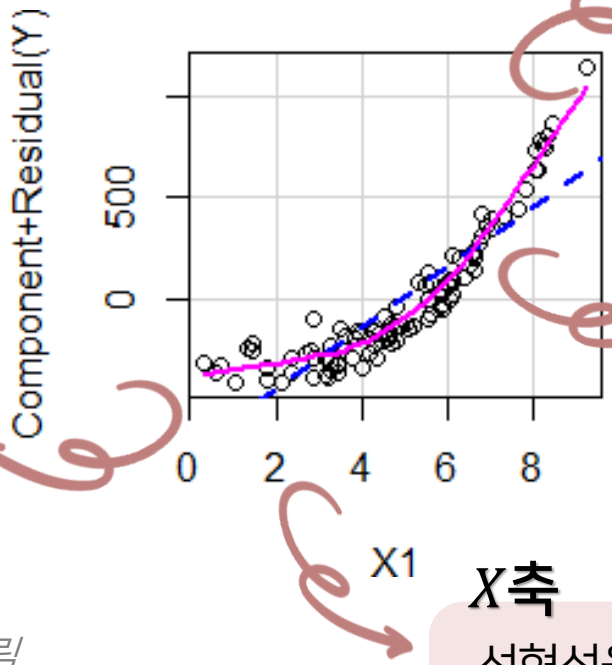
선형성을 판단하기 위한 변수 x_i

선형성 진단과 처방

진단 | Partial Residual Plot

Y축

선형성을 보고싶은 변수를 제외한 나머지 변수로 회귀식을 적합한 잔차



실선

잔차의 추세선

점선

최소제곱법을 통해 회귀선을 추정한 것

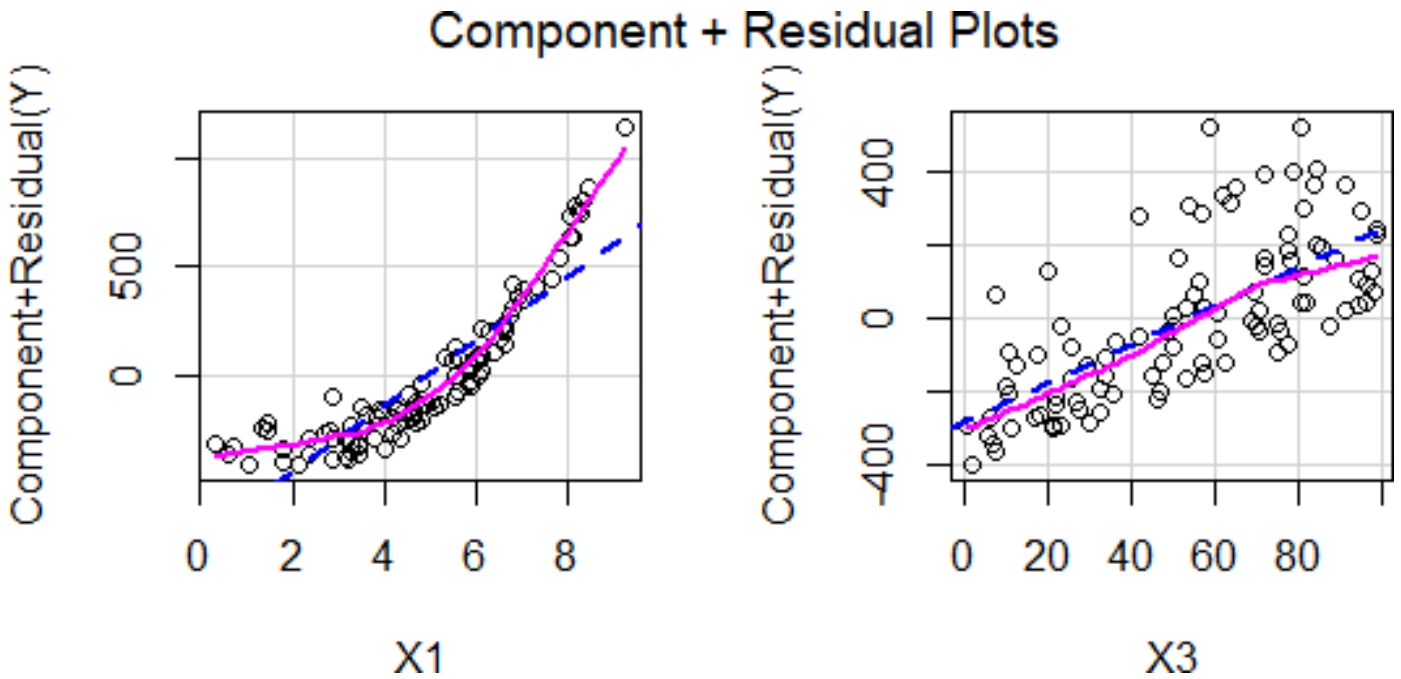
X축

선형성을 판단하기 위한 변수 x_i

이렇게 일부의 변수로 적합한 모델의 잔차를 이용하기 때문에 Partial residual plot이라고 불림

선형성 진단과 처방

진단 | Partial Residual Plot

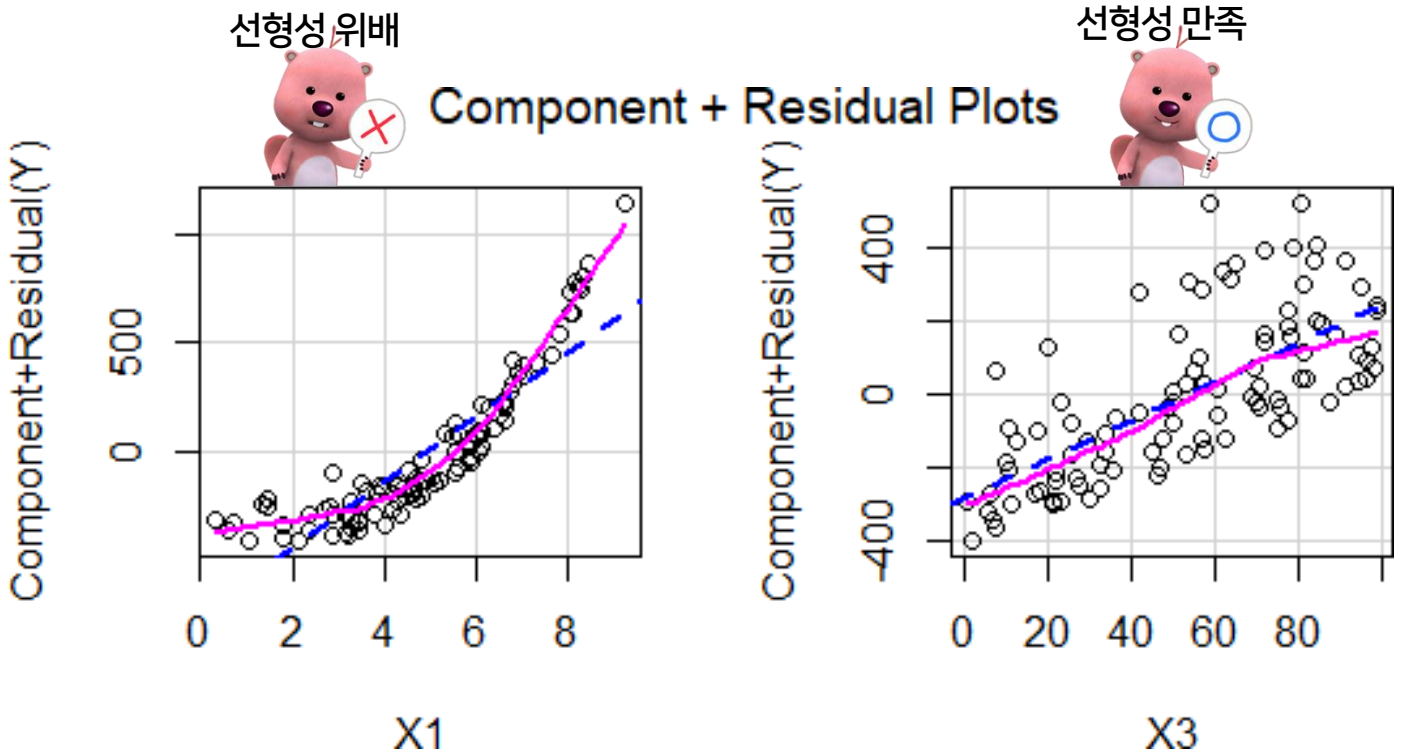


R car 패키지의 crPlots()

일반적으로 서로 다른 두 선이 **일치**하면 **선형성이 만족**되었다고 판단

선형성 진단과 처방

진단 | Partial Residual Plot

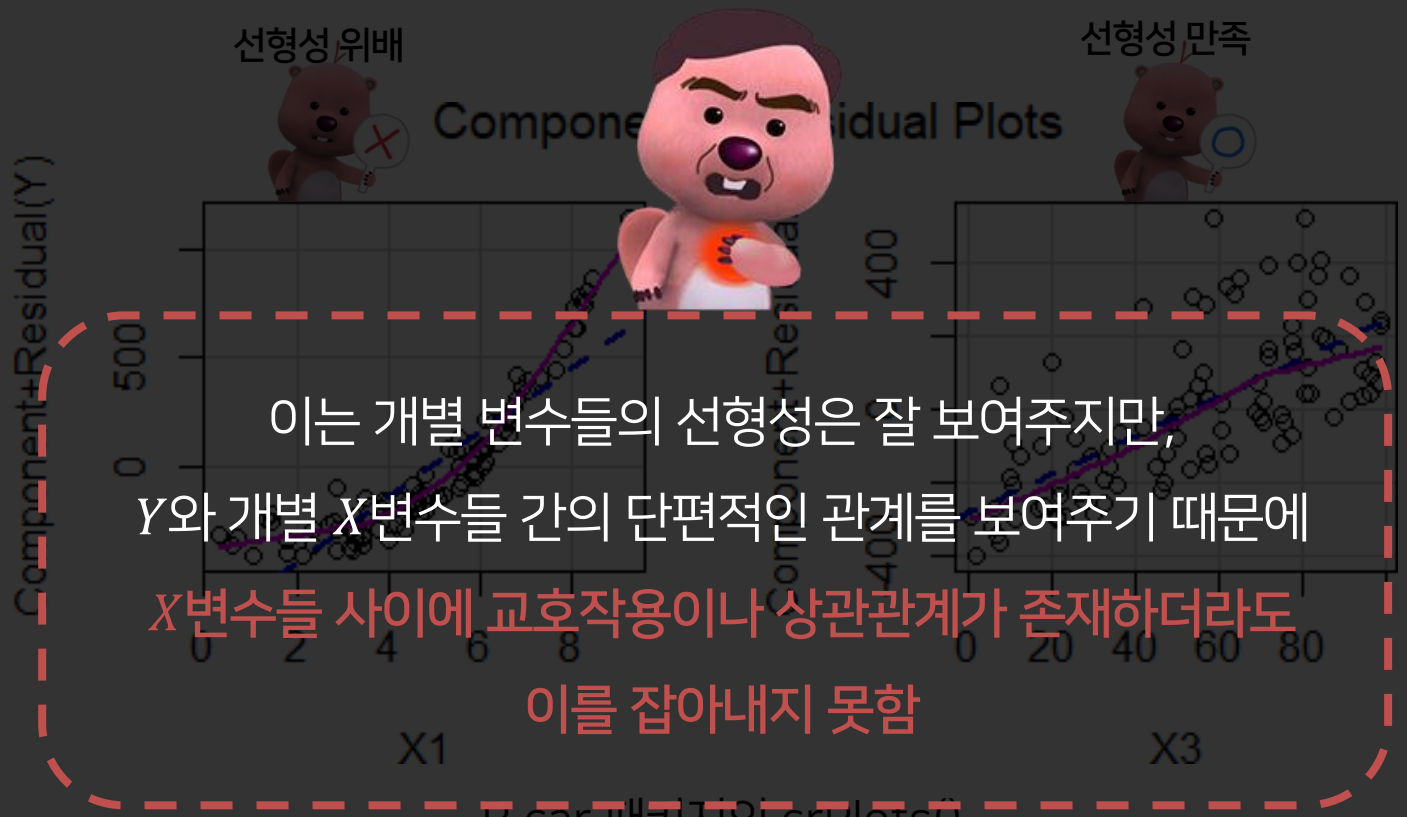


R car 패키지의 crPlots()

일반적으로 서로 다른 두 선이 **일치**하면 **선형성이 만족**되었다고 판단

선형성 진단과 처방

진단 | Partial Residual Plot



이는 개별 변수들의 선형성은 잘 보여주지만,

Y와 개별 X변수들 간의 단편적인 관계를 보여주기 때문에

X변수들 사이에 교호작용이나 상관관계가 존재하더라도

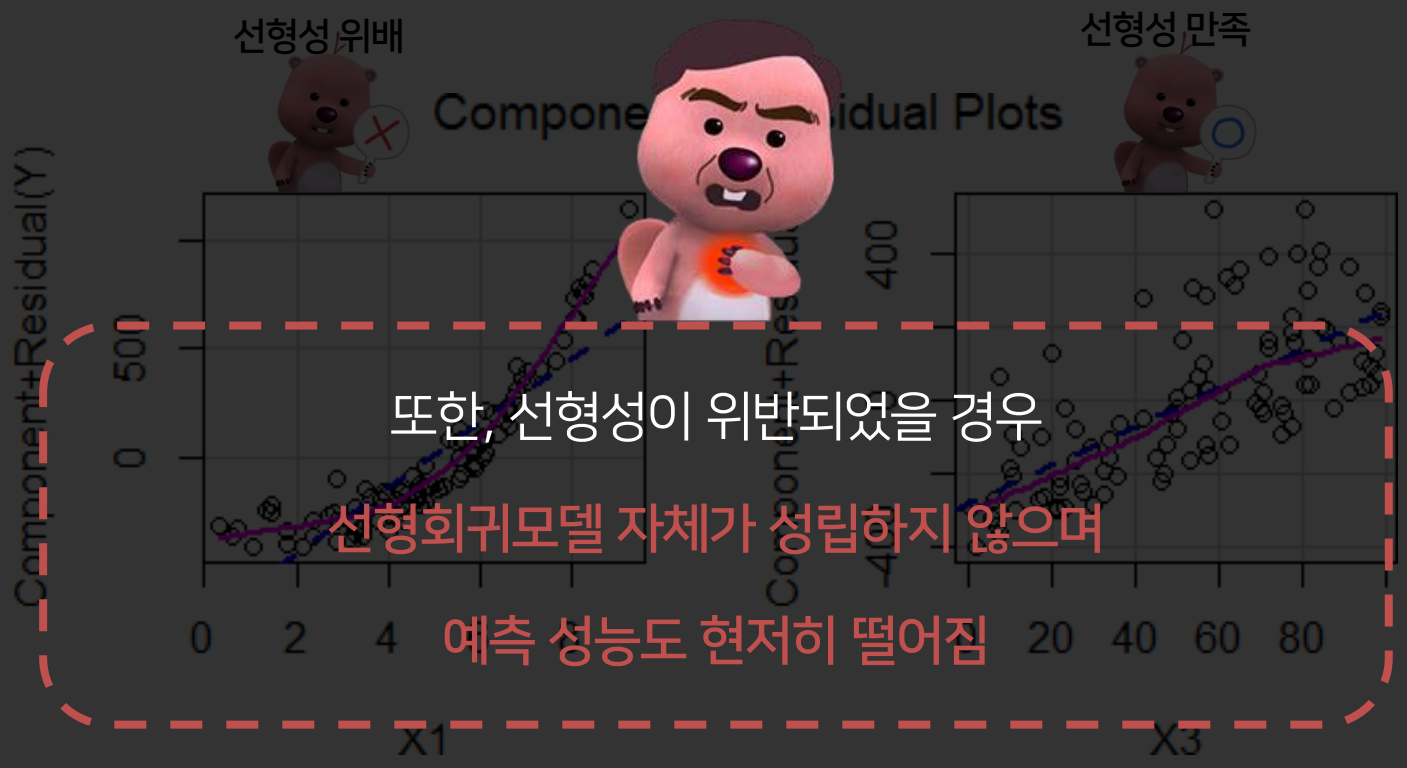
이를 잡아내지 못함

R car 패키지의 crPlots()

일반적으로 서로 다른 두 선이 **일치**하면 **선형성이 만족**되었다고 판단

선형성 진단과 처방

진단 | Partial Residual Plot



R car 패키지의 crPlots()

일반적으로 서로 다른 두 선이 **일치**하면 **선형성이 만족**되었다고 판단

처방 | 변수 변환

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

x_1의 관점에서는 선형결합이 아니므로

$x_1^2 = x_2$ 으로 **변수를 변환**하면, x_2와 y는 선형결합이다.

$$y = \beta_0 e^{\beta_1 x_1} \quad \rightarrow \quad y^* = logy = log\beta_0 + \beta_1 x_1 = \beta_0{}^* + \beta_1 x_1$$

$$y^* = logy = log\beta_0 + \beta_1 x_1 = \beta_0{}^* + \beta_1 x_1$$

양변에 로그를 취해주면, y^*와 x_1은 선형결합이 된다.

처방 | 변수 변환

여러가지 변수변환 방법

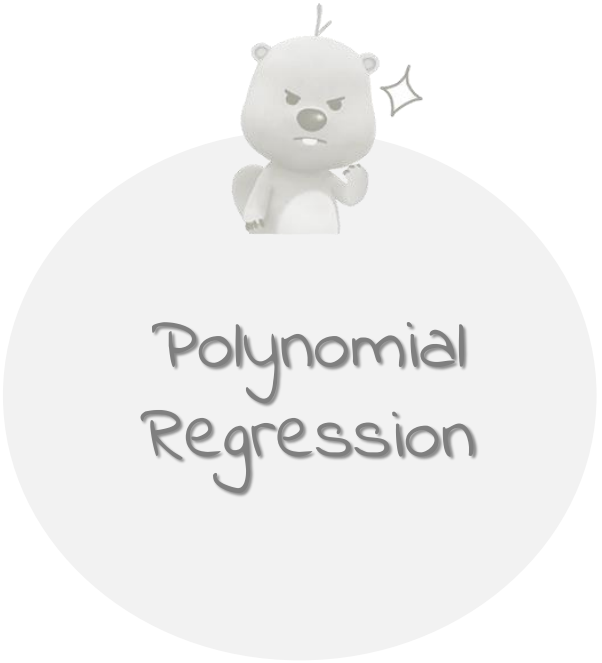
Function	Transformations of x and/or y	Resulting model
$y = \beta_0 x^{\beta_1}$	$y' = log(y), x' = log(x)$	$y' = log(\beta_0) + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = ln(y)$	$y' = ln(\beta_0) + \beta_1 x$
$y = \beta_0 + \beta_1 log(x)$	$x' = log(x)$	$y = \beta_0 + \beta_1 x'$
$y = \frac{x}{\beta_0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

!! **변수 변환을 통해 선형성을 확보**할 수 있는 모델도 넓은 의미에서 선형 모델!

처방 | 비선형회귀

비선형회귀

모델 자체를 **선형회귀모델이 아닌 다른 모델**에 적합시키는 방법



선형성 진단과 처방

처방 | 비선형회귀

비선형회귀

모델 자체를 **선형회귀모델이 아닌 다른 모델**에 적합시키는 방법

Polynomial Regression

Local Regression

선형성 진단과 처방

처방 | 비선형회귀

비선형회귀

모델 자체를 **선형회귀모델이 아닌 다른 모델**에 적합시키는 방법

Polynomial Regression

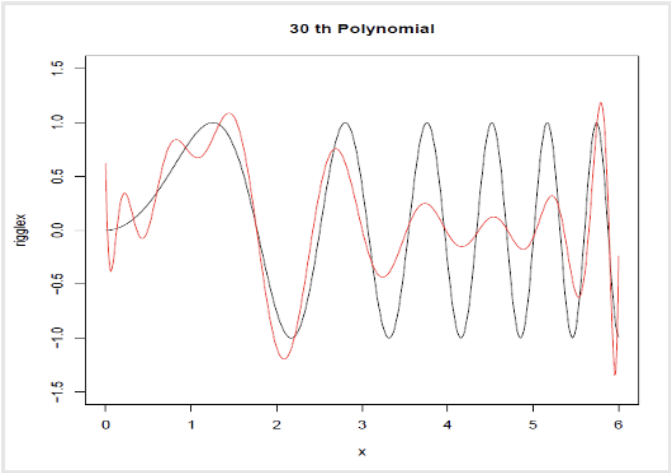
Local Regression

처방 | 비선형회귀

Polynomial Regression

고차항을 고려하는 다항회귀

변수의 **차수**를 바꾸어 모델에 넣어주는 방법



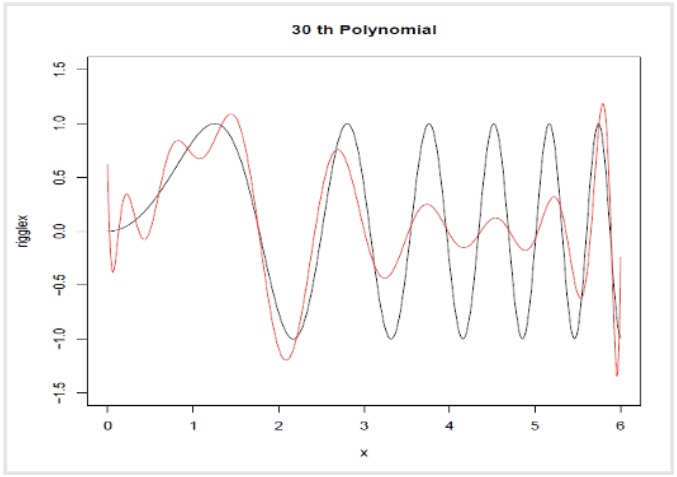
!! 초고항을 적합해도 경향을 못 잡아내기에 **3차까지만 고려**

처방 | 비선형회귀

Polynomial Regression

고차항을 고려하는 다항회귀

변수의 **차수**를 바꾸어 모델에 넣어주는 방법



!! 초고항을 적합해도 경향을 못 잡아내기에 **3차까지만 고려**

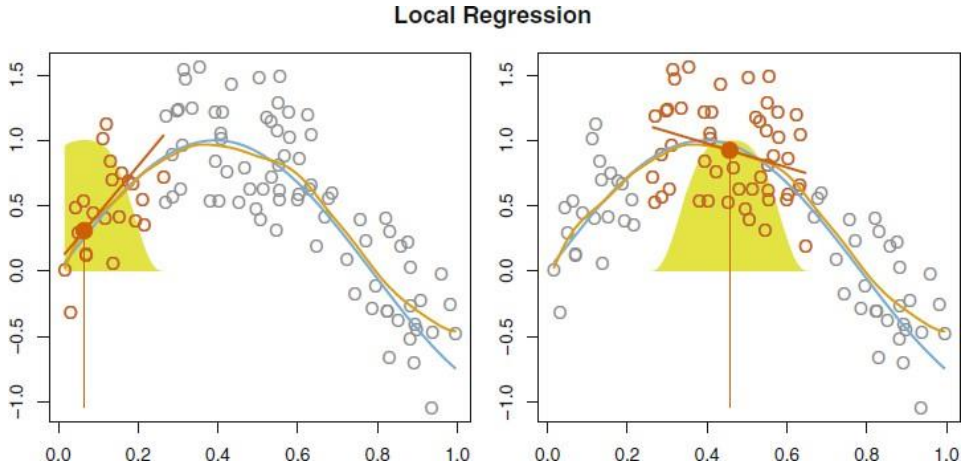
처방 | 비선형회귀

Local Regression

타겟 데이터 x_0 와 가깝다면 큰 가중치를, 멀다면 작은 가중치를 부여

비선형 회귀 방법 & **비모수적** 방법 사용

Local에 있는 데이터들로 회귀 모델링을 하는 방법



타겟 데이터 x_0 를 중심으로 그 주변의 k개의 이웃 데이터들만 사용하여 부분적으로 회귀 모델 구성

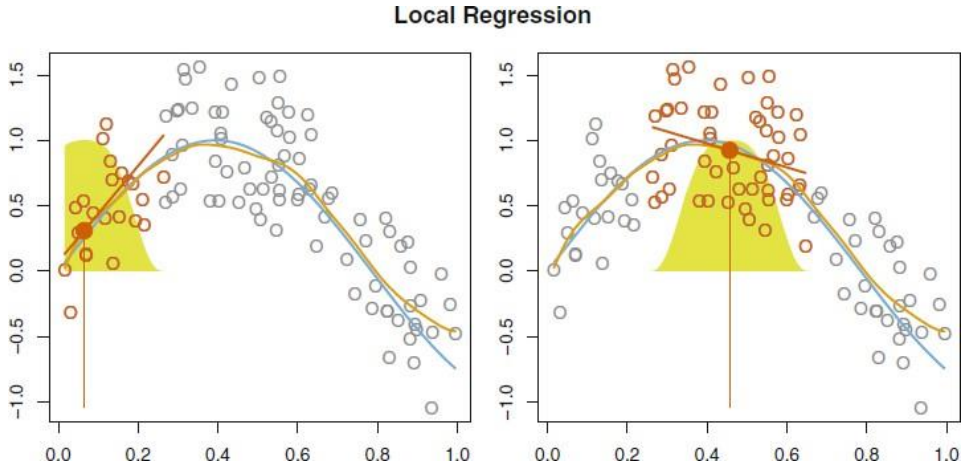
처방 | 비선형회귀

Local Regression

타겟 데이터 x_0 와 가깝다면 큰 가중치를, 멀다면 작은 가중치를 부여

비선형 회귀 방법 & **비모수적** 방법 사용

Local에 있는 데이터들로 회귀 모델링을 하는 방법



타겟 데이터 x_0 를 중심으로 그 주변의 ***k*개**의 이웃 데이터들만 사용하여 부분적으로 회귀 모델 구성

4

정규성 진단과 처방

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

정규성 진단과 처방

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

정규성 가정

정규성 가정

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

회귀식이 데이터를 잘 표현한다면

잔차들은 단순 측정 오차인 **Noise**라 여겨짐

잔차들의 분포는 **정규분포**와 흡사한 형태

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

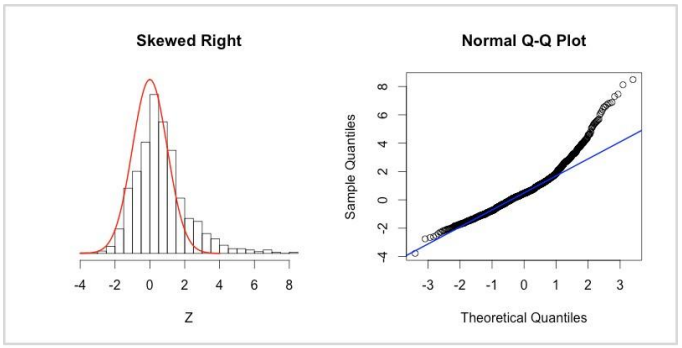
정규성 진단과 처방

진단 | Normal QQ Plot

Normal QQ Plot

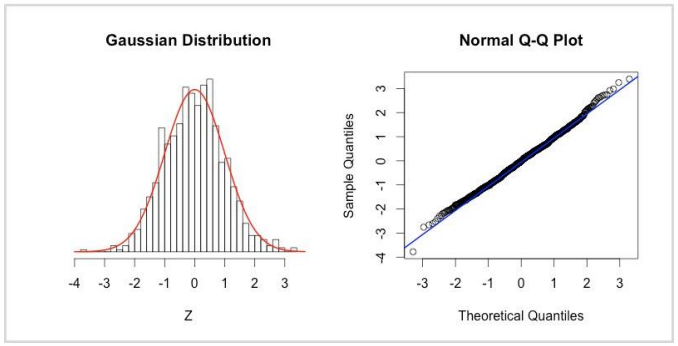
정규성을 파악하기 위한 **비모수적 방법**

$y = x$ 직선에 가까울수록 정규성을 만족



정규성을 만족하지 못하는 경우

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



정규성을 만족하는 경우

자료의 분포가 정규분포에 가까움

정규성 진단과 처방

진단 | Normal QQ Plot

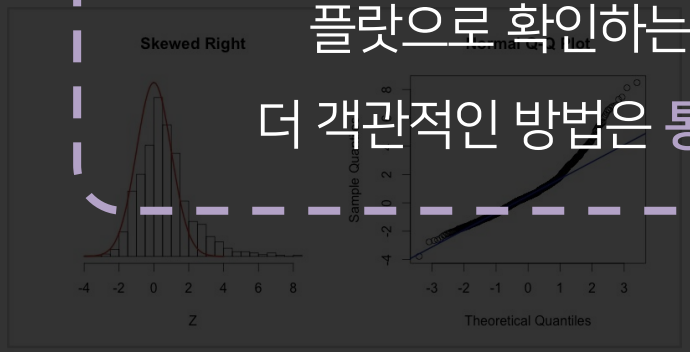
Normal QQ Plot

정규성을 파악하기 위한 **비모수적 방법**

$y = x$ 직선에 가까울수록 정규성을 만족

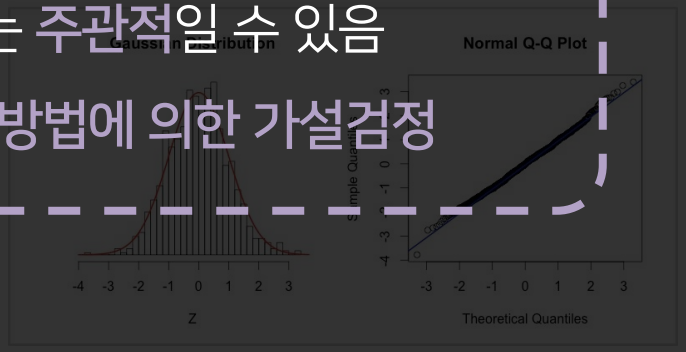
플랏으로 확인하는 경우는 주관적일 수 있음

더 객관적인 방법은 통계적 방법에 의한 가설검정



정규성을 만족하지 못하는 경우

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



정규성을 만족하는 경우

자료의 분포가 정규분포에 가까움

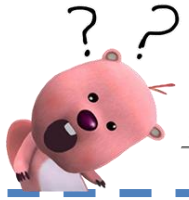
정규성 진단과 처방

진단 | 가설 검정

가설

H_0: 주어진 데이터는 **정규분포를 따른다**

H_1: 주어진 데이터는 **정규분포를 따르지 않는다**



우리가 원하는 것은?

귀무가설을 기각하지 못하는 것

즉 주어진 데이터가 **정규분포를 따르는 것**

정규성 진단과 처방

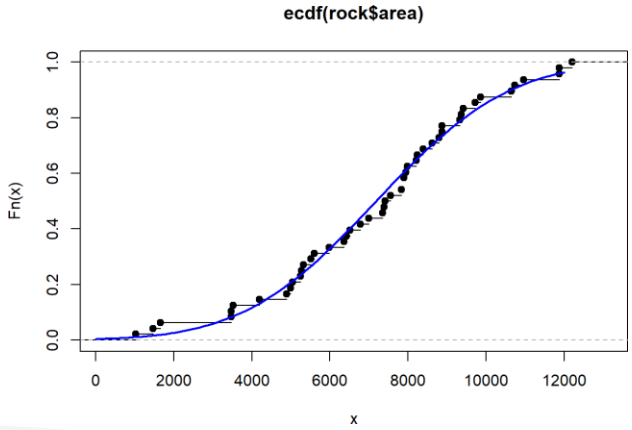
진단 | Empirical CDF Test

Empirical CDF Test

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

잔차의 ECDF와 **정규분포의 CDF**를 비교함으로써 검정

자세한 방법은 다르지만 모두 잔차의 ECDF를 이용해 정규분포와 비교!



Anderson Darling Test

Kolmogorov Smirnov Test

정규성 진단과 처방

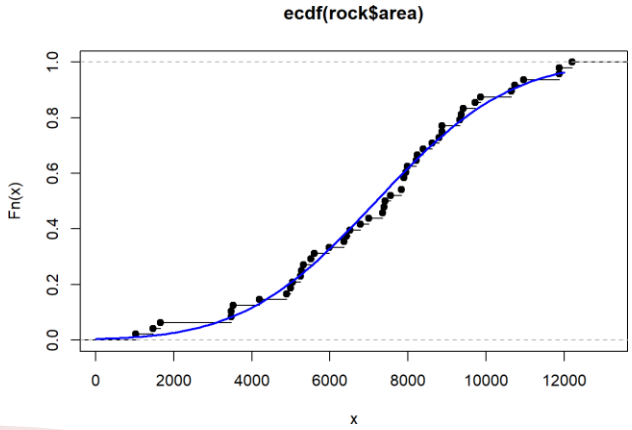
진단 | Empirical CDF Test

Empirical CDF Test

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

잔차의 ECDF와 **정규분포의 CDF**를 비교함으로써 검정

자세한 방법은 다르지만 모두 잔차의 ECDF를 이용해 정규분포와 비교!



Anderson Darling Test

Kolmogorov Smirnov Test

정규성 진단과 처방

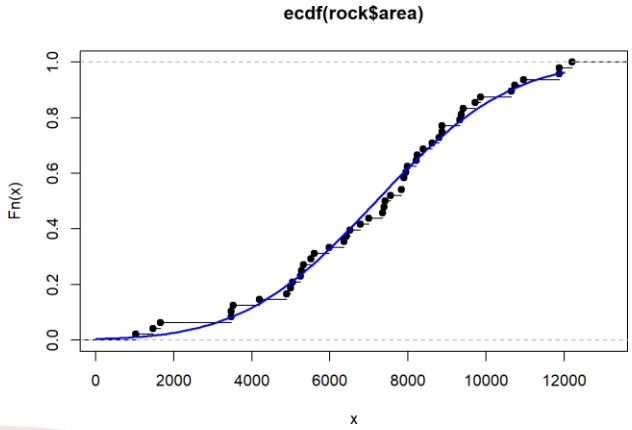
진단 | Empirical CDF Test

Empirical CDF Test

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

잔차의 ECDF와 **정규분포의 CDF**를 비교함으로써 검정

자세한 방법은 다르지만 모두 잔차의 ECDF를 이용해 정규분포와 비교!



Anderson Darling Test

Kolmogorov Smirnov Test

정규성 진단과 처방

진단 | 정규분포의 분포적 특성 이용 Test

Shapiro wilk Test

QQ plot의 아이디어와 동일
정규분포 분위수 값과
표준화 잔차 사이의
선형관계를 확인하는 검정

관측치가 5000개 이하의 데이터에서만 가능
R shapiro.test() 함수로 사용가능

Jarque-Bera Test

정규분포의 왜도가 0,
첨도가 3이라는 점에
기반하는 검정 방법

잔차의 분포가 정규분포와 달라질수록 왜도, 첨도의 변화
통계량 값이 커져 유의수준을 넘어서면 귀무가설을 기각
R Tseries의 jarque.bera.test() 함수로 사용 가능

정규성 진단과 처방

진단 | 정규분포의 분포적 특성 이용 Test

Shapiro wilk Test

QQ plot의 아이디어와 동일

정규분포 분위수 값과

표준화 잔차 사이의

선형관계를 확인하는 검정

관측치가 5000개 이하의 데이터에서만 가능

R shapiro.test() 함수로 사용가능

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

표준정규분포의 왜도(0), 첨도(3)와 비교해

차이가 커질수록 검정통계량 값이 커짐

Jarque-Bera Test

정규분포의 왜도가 0,

첨도가 3이라는 점에

기반하는 검정 방법

잔차의 분포가 정규분포와 달라질수록 왜도, 첨도의 변화

통계량 값이 커져 유의수준을 넘어서면 귀무가설을 기각

R Tseries의 jarque.bera.test() 함수로 사용 가능

정규성 진단과 처방

진단 | 분포적 특성 이용 Test

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

표준정규분포의 왜도(0), 첨도(3)와 비교해
차이가 커질수록 검정통계량 값이 커짐

Shapiro wilk Test

QQ plot의 아이디어와 동일

정규분포 분위수 값과

표준화 잔차 사이의

선형관계를 확인하는 검정

관측치가 5000개 이하의 데이터

R shapiro.test() 함수로 사용가능

Jarque-Bera Test

정규분포의 왜도가 0,

첨도가 3이라는 점에

기반하는 검정 방법

잔차의 분포가 정규분포와 달라질수록 왜도, 첨도의 변화
통계량 값이 커져 유의수준을 넘어서면 귀무가설을 기각

tseries의 jarque.bera.test() 함수로 사용 가능

정규성이 위배됐을 경우 문제점

검정통계량이 t 분포 또는 F분포를 따르지 않게 됨

(t분포, F분포는 정규분포를 하므로)

↓

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

!!

최적의 λ는 ML 방법을 통해 신뢰구간을 구한 후

신뢰구간 내 로그우도함수(ML)를 최대화하는 값을 선택

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

!!

최적의 λ는 ML 방법을 통해 신뢰구간을 구한 후

신뢰구간 내 로그우도함수(ML)를 최대화하는 λ 값을 선택

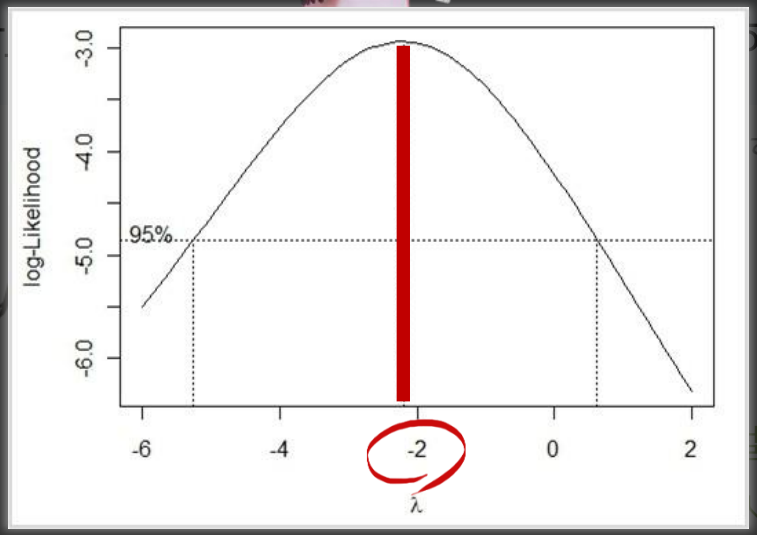
정규성 진단과 처방

처방 | 변수 변환

Box-Cox Transformation

Y를 변환함 해주는 방법

통계적 검정에 따라 변수 변환



사용

Tip: 정수로 λ 값을 선택한다면 변수변환 관계를 쉽게 파악 가능

최적 한 후

신뢰구간 값을 선택

95% 내의 λ 값 중 가능도 함수가
최대가 되는 −2 근방의 λ를 선택

처방 | 변수 변환

Box-Cox Transformation

Y를 변환함으로써 **정규성** **등분산성**을 해결해주는 방법

통계적 검정에 따라 변수 변환

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

Box-Cox Transformation은 y가 **log(y)로 변환**될 수 있으므로

y 가 0이하일 때는 사용할 수 없는 단점이 존재

일반적으로 λ는 −5와 5 사이의 값을 사용

!!

최적의 λ는 ML 방법을 통해 신뢰구간을 구한 후

신뢰구간 내 로그우도함수(ML)를 최대화하는 값을 선택

처방 | 변수 변환

Yeo-Johnson Transformation

Box-cox transformation과 동일한 아이디어

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

!!

범위를 잘게 쪼개어 y값이 0보다 작더라도 사용 가능

처방 | 변수 변환

Yeo-Johnson Transformation

Box-cox transformation과 동일한 아이디어

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

!!

범위를 잘게 쪼개어 y값이 0보다 작더라도 사용 가능

5

등분산성 진단과 처방

등분산성 진단과 처방

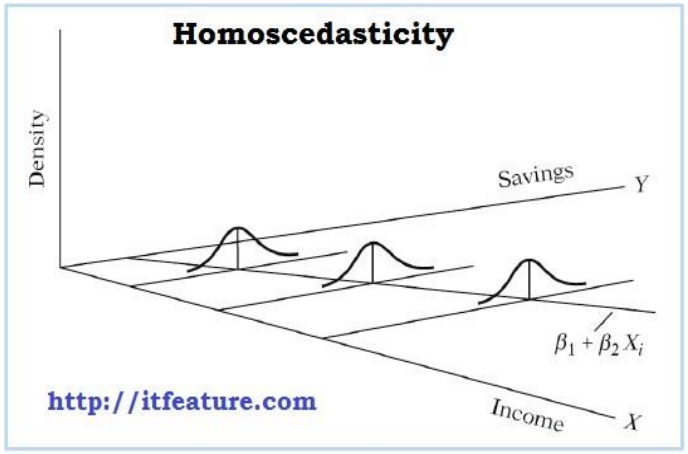
등분산성 가정

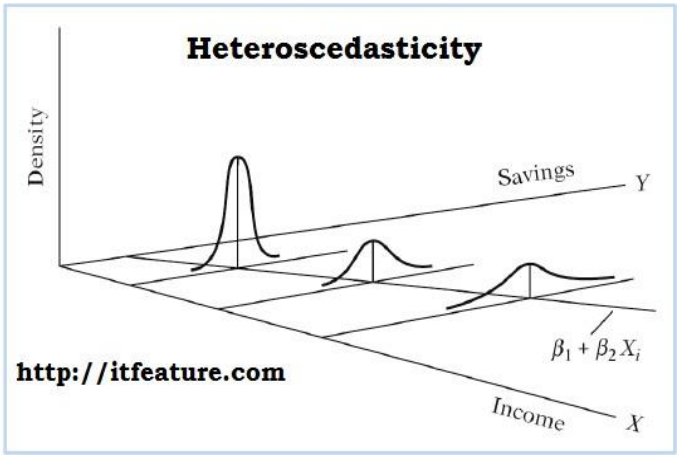
등분산성 가정

오차항의 분산은 어느 관측치에서나 **상수 σ^2**으로 동일하고

다른 변수의 영향을 받지 않는다는 가정

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

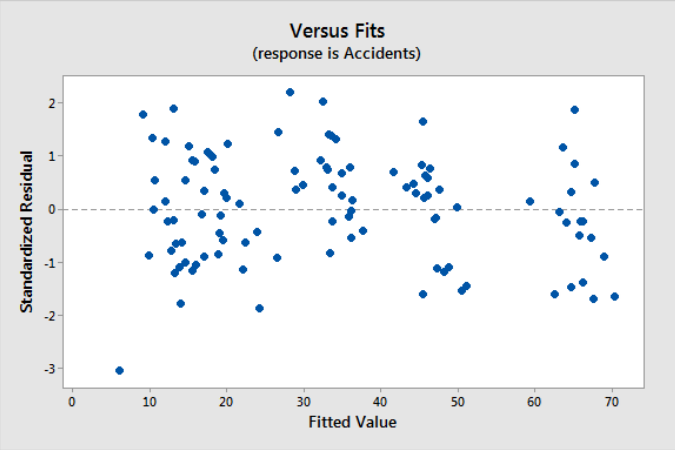




등분산성 진단과 처방

진단 | 잔차 플랏

잔차가 RANDOM하게 분포

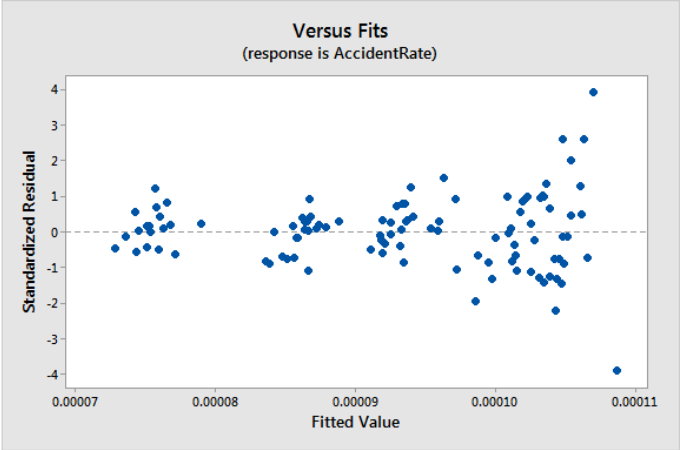


Fitted value $\hat{y}$값에

상관없이 잔차의

퍼짐의 정도가 일정

잔차가 특정 패턴을 가짐



Fitted value $\hat{y}$값이

커짐에 따라 잔차의

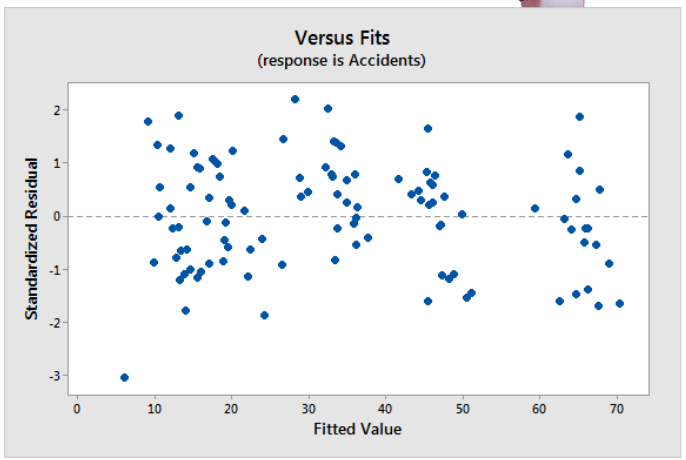
퍼짐의 정도가 변화

등분산성 진단과 처방

진단 | 잔차 플랏

등분선성 만족

잔차가 RANDOM하게 분포



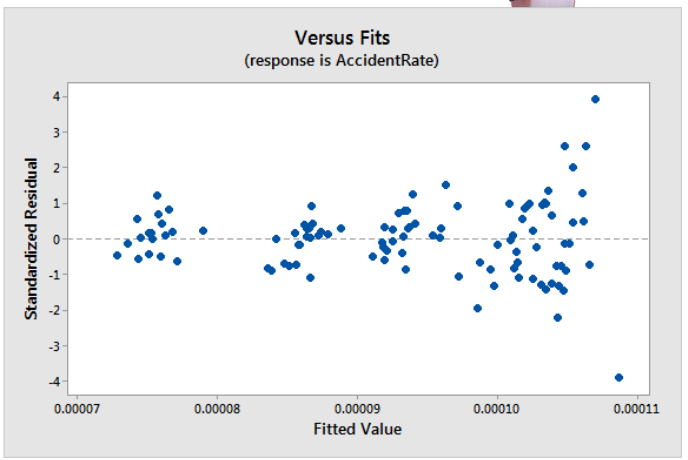
Fitted value $\hat{y}$값에

상관없이 잔차의

퍼짐의 정도가 일정

등분산성 위배

잔차가 특정 패턴을 가짐



Fitted value $\hat{y}$값이

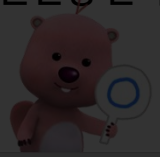
커짐에 따라 잔차의

퍼짐의 정도가 변화

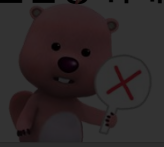
5 등분산성 진단과 처방

진단 | 잔차 플랏

등분선성 만족

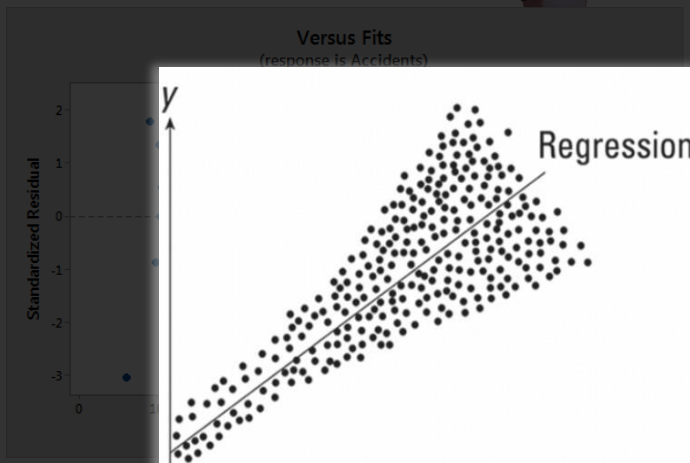


등분산성 위배



잔차가 RANDOM하게 분포

잔차가 특정 패턴을 가짐



상관없이 잔차의

퍼짐의 정도가 일정

커짐에 따라 잔차의

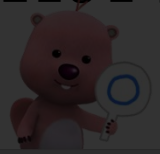
퍼짐의 정도가 변화

산점도를 통해서도 이분산성 확인 가능!

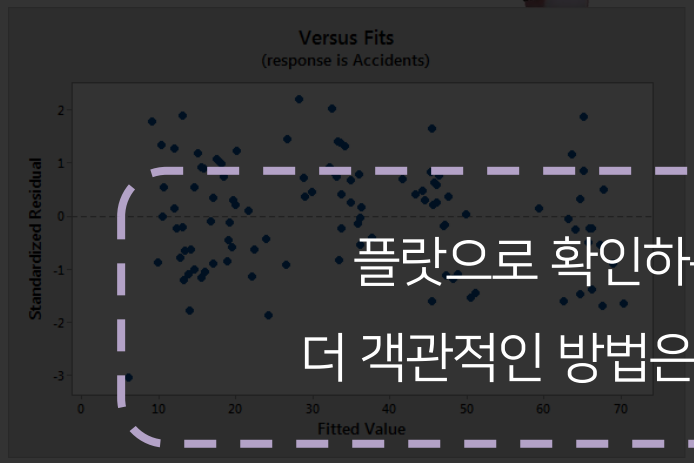
5 등분산성 진단과 처방

진단 | 잔차 플랏

등분선성 만족



잔차가 RANDOM하게 분포

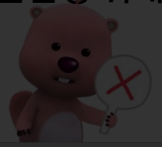


Fitted value $\hat{y}$값에

상관없이 잔차의

퍼짐의 정도가 일정

등분산성 위배



잔차가 특정 패턴을 가짐

Fitted value $\hat{y}$값이

커짐에 따라 잔차의

퍼짐의 정도가 변화

플랏으로 확인하는 경우는 주관적일 수 있음

더 객관적인 방법은 통계적 방법에 의한 가설검정

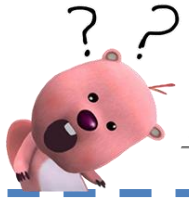
등분산성 진단과 처방

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**

등분산성 진단과 처방

진단 | Breusch-Pagan Test

Breusch-Pagan Test (BP Test)

오차의 분산이 **등분산**인지 아닌지 판단

잔차가 **독립변수들의 선형결합**으로 표현되는지 검정

R lmstat 패키지의 bptest() 함수 사용/car 패키지의 ncvTest() 함수 사용

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

분산이 **설명변수 X에 대한 선형결합**으로 되어있다는 가정

분산과 설명변수 X 간 **결정계수 값(R^2)**이 높으면 등분산성 위배

5 등분산성 진단과 처방

진단 | Breusch-Pagan Test

Breusch-Pagan Test (BP Te

오차의 분산이 등 지 아닌지 판단

오차의 추정량인 잔차가 **독립변** **선형결합**으로 표현되는지 검정

R lmstat 패키지의 bptest() 함수 사용/car 패키지의 ncvTest() 함수 사용

한계점

비선형적 결합으로 이루어진 이분산성 파악할 수 없음

Sample의 **크기가 커야** (**대표본이어야**) 사용 가능

분산이 **설명변수 X에 대한 선형결합**으로 되어있다는 가정

분산과 설명변수 X 간 **결정계수 값(R^2)**이 높으면 등분산성 위배

등분산성 진단과 처방

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 BLUE가 되지 못함

Best Linear Unbiased Estimator

제1종오류 : 귀무가설이 실제로 참이지만, 이에 불구하고 귀무가설을 기각하는 오류

BLUE가 궁금하다면 회귀분석 1주차 클린업 내용 참고!

등분산성 진단과 처방

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 BLUE가 되지 못함

Best Linear Unbiased Estimator

제1종오류 : 귀무가설이 실제로 참이지만, 이에 불구하고 귀무가설을 기각하는 오류

BLUE가 궁금하다면 회귀분석 1주차 클린업 내용 참고!

등분산성 진단과 처방

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

BLUE가 궁금하다면 회귀분석 1주차 클린업 내용 참고!

5 등분산성 진단과 처방

이분산성의 문제점 회귀분석팀 1주차 클린업 참고!

BLUE Best Linear Unbiased Estimator

분산이 제일 작은 선형 불편추정량

분산이 작다는 것은 **추정량이 안정적**이라는 의미

① 오차들의 **평균은 0**

② 오차들의 **분산은 σ^2 로 동일**

③ 오차간 **자기상관이 없음**

Independent

세 가지 조건이 만족되면, LSE는 선형불편추정량 중

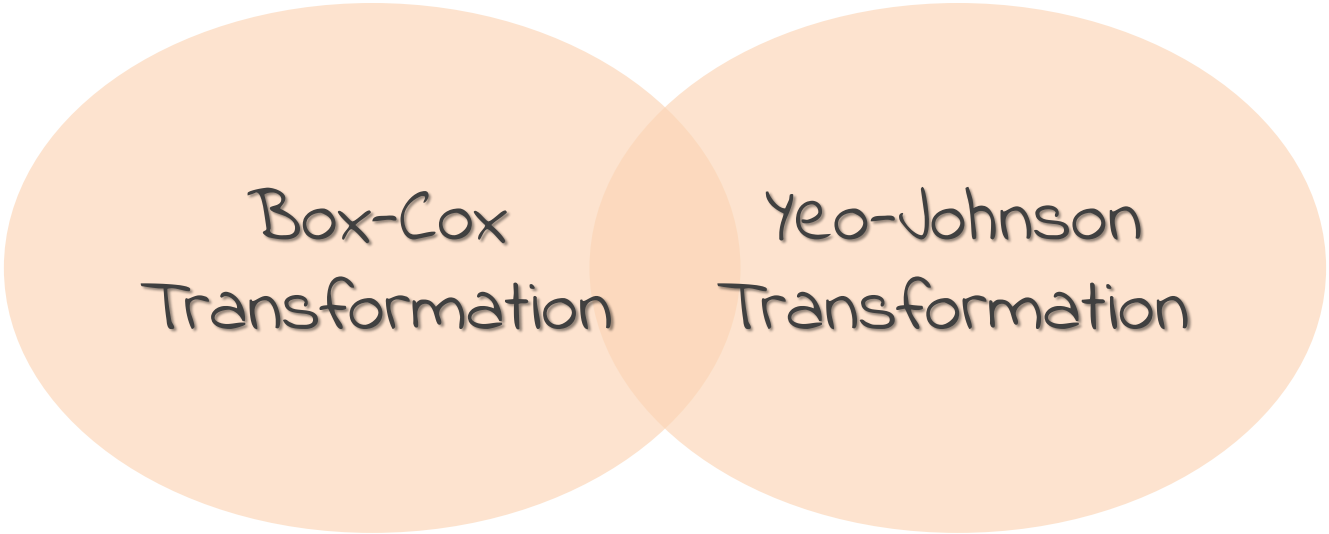
분산이 가장 작은 안정적인 추정량이 됨

제1종오류 : 귀무가설이 실제로 참이지만, 이에 불구하고 귀무가설을 기각하는 오류

BLUE가 궁금하다면 회귀분석 1주차 클린업 내용 참고!

등분산성 진단과 처방

처방 | 변수 변환



!!

정규성을 만족시키기 위해 사용했던

각종 변수 변환 방법 등을 똑같이 적용 가능

등분산성 진단과 처방

처방 | 변수 변환

가중 회귀 제곱 weighted least Square

W_i는 가중치이며, 분산에 반비례

등분산이 아닌 형태의 데이터마다 **다른 가중치**를 주어서

등분산을 만족하게 해주는 **일반화된 최소제곱법**의 한 형태

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

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**

등분산성 진단과 처방

처방 | 변수 변환

가중 회귀 제곱 weighted least Square

W_i는 가중치이며, 분산에 반비례

등분산이 아닌 형태의 데이터마다 **다른 가중치**를 주어서

등분산을 만족하게 해주는 **일반화된 최소제곱법**의 한 형태

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

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**

등분산성 진단과 처방

처방 | 변수 변환

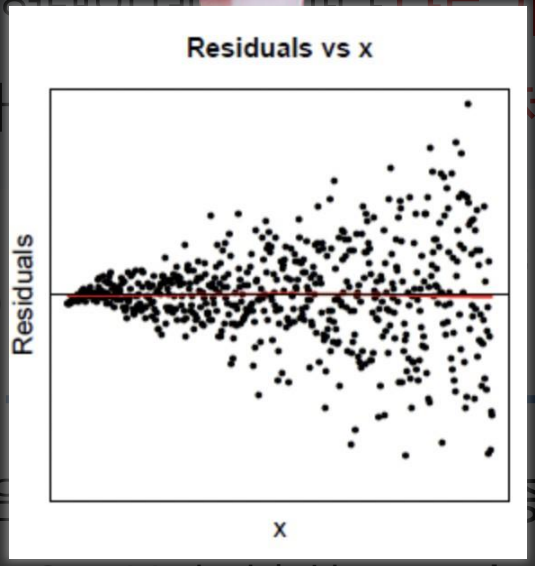
가중 회귀 제곱 weighted least Squar

w_i는 가중치이며, 분산에 반비례

등분산이 아닌 형태의 데 가중치를 주어서

등분산을 만족하 제곱법의 한 형태

$$\sum w_i(y_i - \beta_0 \quad , \qquad w_i \propto \frac{1}{\sigma_i^2}$$



분산이 커 신뢰도가 낮은 증치 (분산을 작게 만듦)

분산이 작아 신뢰도가 높은 부분의 관측치 → **큰 가중치** (분산을 크게 만듦)

잔차 플랏을 통해 **명확한 패턴**을 발견한다면

이를 토대로 **가중치 선정** ex) $w_i \propto \frac{1}{x_i^2}$

작은 가중치를 가지는 관찰값은

회귀계수 값을 결정하는데 **적은 영향을 미침**

6

독립성 진단과 처방

독립성 진단과 처방

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

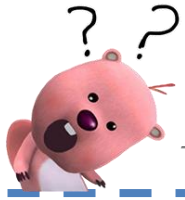
독립성 진단과 처방

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 진단과 처방

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

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

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

[-1, 1] 값을 갖는 e_i와 e_{i-1}의 상관계수의 꼴 !

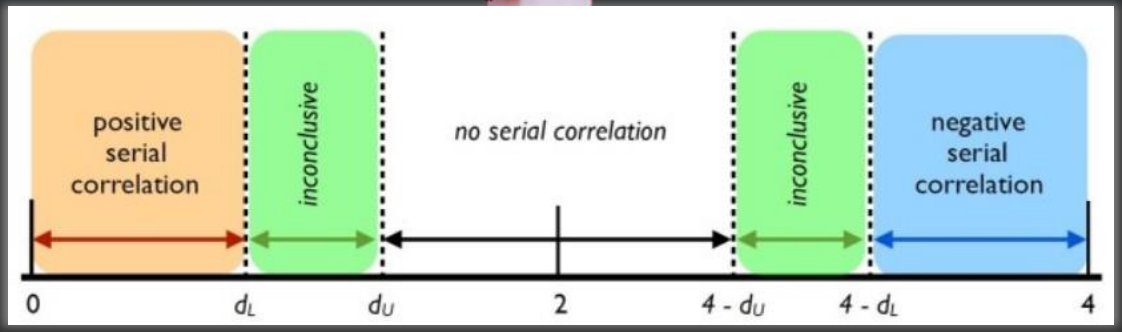
독립성 진단과 처방

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



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

더빈 왓슨 검정표에 따라 귀무가설 기각

d_L과 d_U는 귀무가설을 기각할 수 있는지

없는지를 판단하는 하한, 상한값

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

$\widehat{\rho_1}$은 표본 잔차 간 상관계수 (sample correlation of the residuals)

[-1, 1] 값을 갖는 e_i와 e_{i-1}의 상관계수의 꼴 !

6 독립성 진단과 처방

독립성 진단

Durbin Waston Test

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

앞 뒤 관측)을 확인

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

d가 **0** 에 가까울 수록 양의 상관관계

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

d가 **4** 에 가까울 수록 음의 상관관계

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

d가 **2** 에 가깝다면 귀무가설을 기각하지 못함!

(= 1차 자기상관이 없다)

6 독립성 진단과 처방

독립성 진단

Durbin Waston Test

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

앞 뒤 관측)을 확인

d가 상한과 하한 사이에 위치한다면 판단할 수 없음

바로 인접한 오차와의 1차 자기상관만을 고려함

($AR(1)$ 구조만을 파악할 수 있음)

자기상관이 오래 지속되거나, 계절성이 있는 경우 한계

AR모형에 대한 자세한 내용은 시계열팀 클린업 참고!

$\hat{\rho}_1$은 표본 ... residuals)

[-1, 1] 값을 갖는 e_i와 e_{i-1}의 상관계수의 꼴 !

독립성 진단과 처방

독립성 위배의 문제점

LSE의 가정 세가지를 만족하지 못하므로,
최소제곱추정량이 더 이상 **BLUE**가 되지 않음

σ^2 **의 추정량과 회귀계수의 표준오차**가 실제보다 심각하게 **과소추정**
유의성 검정 결과 신뢰도 하락, Prediction Interval 넓어짐

처방 | 가변수 생성

가변수 생성

뚜렷한 **계절성**이 있다고 판단되면, **가변수** 생성

계절성이 **주기**를 가진다는 점을 이용

주기 함수인 **삼각함수 $cos(t), sin(t)$의 선형결합**으로 주기를 표현하는 방법

변수에 주기를 표현하는 가변수를 만듦으로써 대처하는 방법

독립성 진단과 처방

처방 | 분석 모델 변경

분석 모델 변경

현 모델이 **시간, 공간에 따른 자기상관**을 가질 때 알맞은 모델로 변경

✌ **시간**에 따라 자기상관을 가지는 경우

시계열 모델 사용 (ex. $AR(p)$ 모형)

✌ **공간**에 따라 자기상관을 가지는 경우

공간회귀모델 사용

유용한 진단 패키지

gvlma package

Global Validation of Linear Model Assumption

선형성, 정규성, 등분산성을 한 번에 체크해주는 함수

	Value	p-value	Decision
Global Stat	113.688	0.000e+00	Assumptions NOT satisfied!
Skewness	37.022	1.168e-09	Assumptions NOT satisfied!
Kurtosis	50.181	1.402e-12	Assumptions NOT satisfied!
Link Function	25.760	3.867e-07	Assumptions NOT satisfied!
Heteroscedasticity	0.725	3.945e-01	Assumptions acceptable.

Global Stat : **선형성** / Skewness : **정규성**

Kurtosis : **정규성** / Link function : **선형성** / Heteroscedasticity : **등분산성**

유용한 진단 패키지

gvlma package

Global Validation of Linear Model Assumption

선형성, 정규성, 등분산 [illegible]에 체크해주는 함수

	Value	p-value	Decision
Global Stat	113.688	0.000e+00	Assumptions NOT satisfied!
Skewness	[illegible]	[illegible]	[illegible]
Kurtosis	50.181	1.402e-12	Assumptions NOT satisfied!
[illegible]	[illegible]	[illegible]	[illegible]
Heteroscedasticity	0.725	3.945e-01	Assumptions acceptable.

유의수준 0.05에서 [가정 충족|가정 충족하지 않음]의 경계를 잘라 버림

융통성 부족, 비선형적 모델로 바로 넘어가는 등의 속단은 위험

Global Stat : **선형성** / Skewness : **정규성**

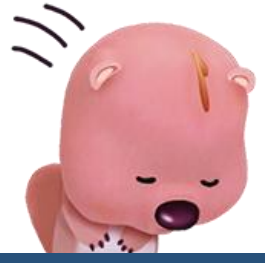
Kurtosis : **정규성** / Link function : **선형성** / Heteroscedasticity : **등분산성**

다음주 예고

1. 다중공선성

2. 변수선택법

3. 정규화



감사합니다