회귀분석팀

6팀
유종석
윤경선
채소연
김진혁
안은선

INDEX

1

회귀 기본 가정

1 회귀 기본 가정

모델의 가정이 지니는 의미

잔차의 평균이 최대한 0에 가까워지는 정확한 회귀모델을 만드는 것

현실에서는...

추정한 모델과 실제 데이터 사이에서 **오차 발생**

1 회귀 기본 가정

모델의 가정이 지니는 의미

잔차의 평균이 최대한 0에 가까워지는 정확한 회귀모델을 만드는 것

Check!

- ✓ 모델링을 할 때 미처 고려하지 못한 속성들
- ✓ 현실 세계의 여러 오차(잡음)

현실에서는...

추정한 모델과 실제 데이터 사이에서 오차 발생

1 회귀 기본 가정

회귀분석의 가정들

가정

- 선형성
- 등분산성
- 정규성
- 독립성

적은 수의 관측치로…

① **모델 구성** 가능

② **좋은 추정**과 **예측** 가능

1 회귀 기본 가정

회귀분석의 가정들

머신러닝 모델들에도 이러한 가정들이 들어가는데,

이는 모델이 만들어진 형태와 관련이 있기 때문에

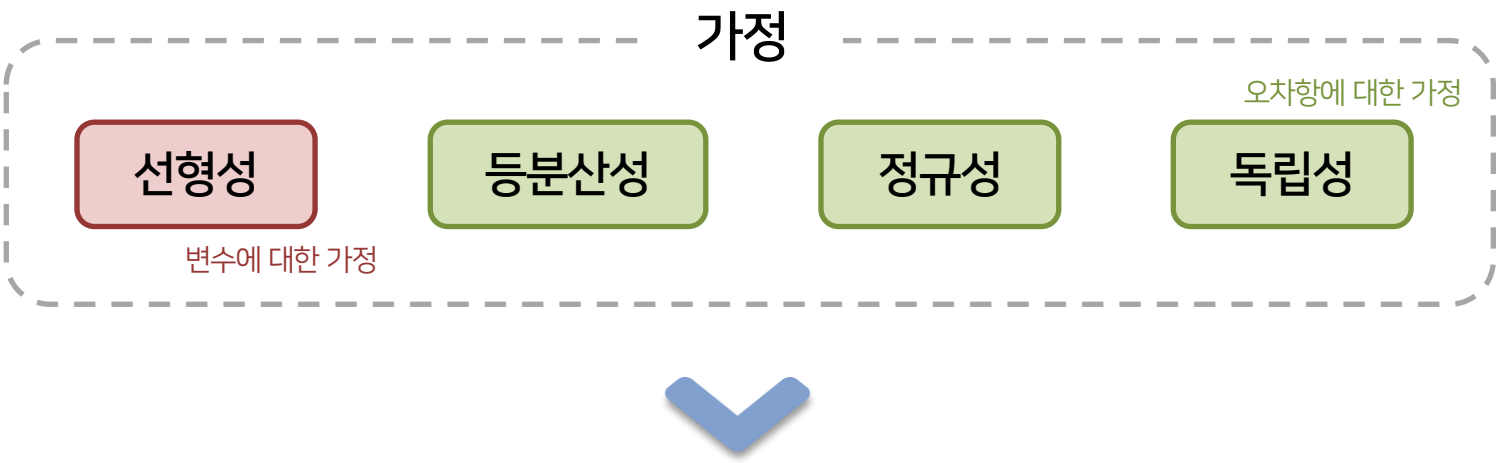
가정들이 지켜지지 않으면 **모델의 성능이 급락하는 경우**가 많음

모델 구성 가능

좋은 **추정**과 **예측** 가능

1 회귀 기본 가정

회귀분석의 가정들



회귀식

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

1 회귀 기본 가정

회귀분석의 가정들

선형성 *Linearity*

설명변수와 반응변수의 관계는 **선형**이다

모델 자체가 선형성만 고려하고 있음.

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

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

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

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

치환의 과정을 통해

변화된 x를 **새로운 x**로 취급할 수 있다면,

넓은 의미의 **선형결합**으로 이해 가능

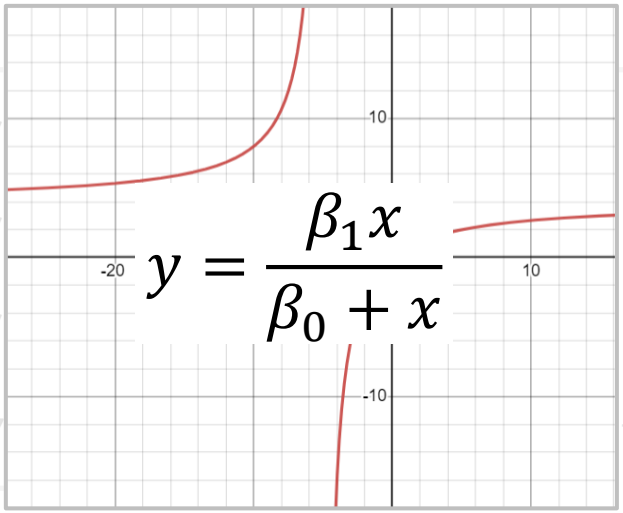
= 선형성 만족

회귀분석의 가정들

선형성 *Linearity*

설명변수와 반응변수의 관계는 **선형**이다

모델 자체가 선형성만 고려하고 있음.



치환의 과정을 통해

변화된 x를 새로운 x로 취급한다면,

위 결합들을 모두 선형결합으로 이해

= 선형성 만족

비선형모델

1 회귀 기본 가정

회귀분석의 가정들

오차의 정규성 *Normality*

오차항은 **정규분포**를 따른다

오차의 평균은 0이다(거의 위배되지 않는 가정)

정규분포가 오차에 대한 확률분포이기 때문에...

회귀식이 데이터를 잘 표현하고 있다면,

오차들은 **단순 잡음(noise)**이 되어 **정규분포에 근접**하는 형태가 나옴

1 회귀 기본 가정

회귀분석의 가정들

오차의 정규성 *Normality*

오차항은 **정규분포**를 따른다

오차의 평균은 0이다(거의 위배되지 않는 가정)

오차의 정규성을 가정하기 때문에

회귀식과 **개별 회귀계수**에 대한 **검정 시행 가능**

정규분포가 오차에 대한 확률분포이기 때문에...

회귀식이 데이터를 잘 표현하고 있다면,

오차들은 **단순 잡음(noise)**이 되어 **정규분포에 근접**하는 형태가 나옴

1 회귀 기본 가정

회귀분석의 가정들

오차의 정규성 *Normality*

오차항은 정규분포를 따른다

정규분포를 따르지 않을 경우

가설검정에서 분포가 왜곡될 것이고,

이에 따라 검정 결과를 신뢰할 수 없음

오차의 정규성 가정이 만족해야

회귀모형의 해석 가능성에도 의미를 부여할 수 있다.

정규분포가 오차에 대한 적절한 분포이기 때문에...

회귀식이 데이터를 잘 표현하고 있다면,

오차들은 단순 잡음(noise)이 되어 정규분포에 근접하는 형태가 나옴

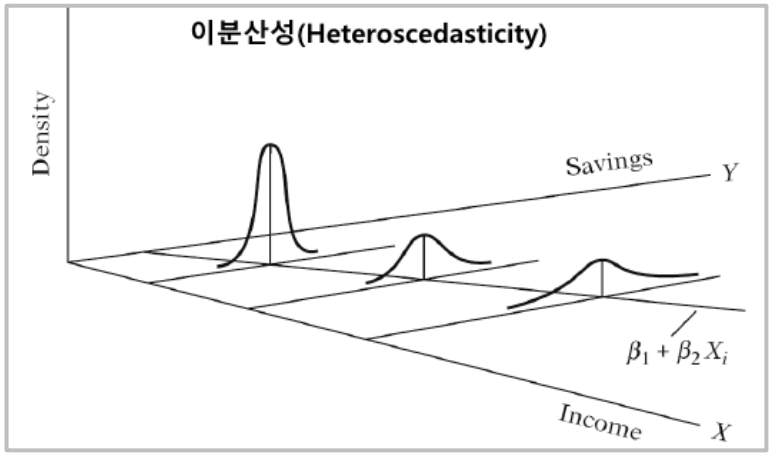
회귀분석의 가정들

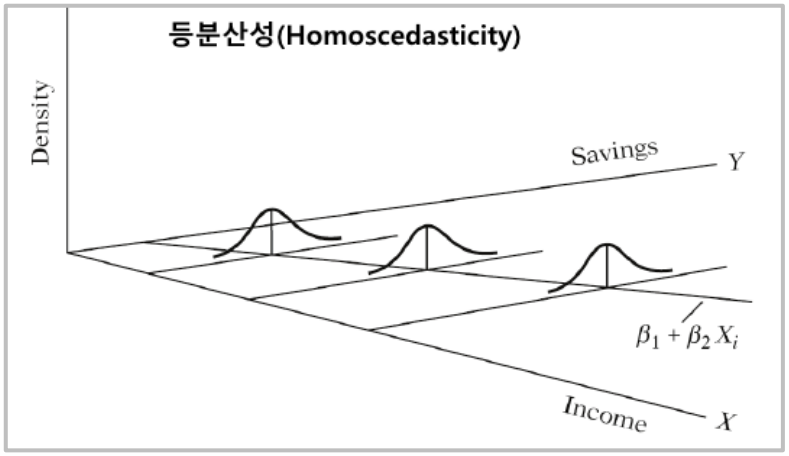
오차의 등분산성 *Homoscedasticity / Constant variance*

오차항의 **분산**은 **상수**다

분산은 σ^2으로 동일하다

↔ 이분산성(Heteroscedasticity)





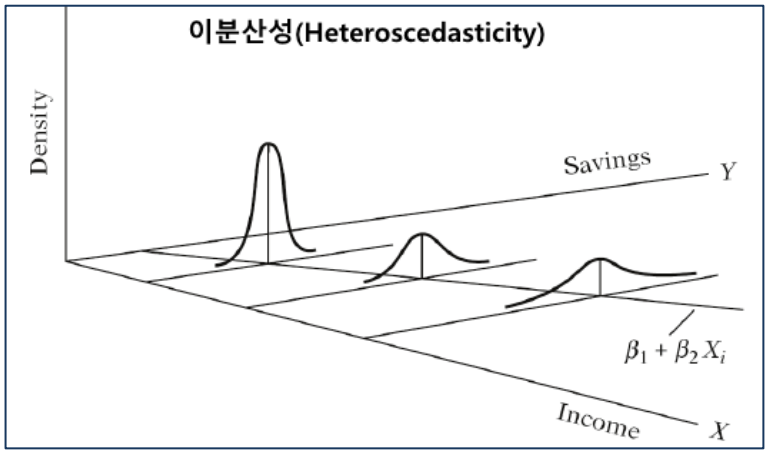
회귀분석의 가정들

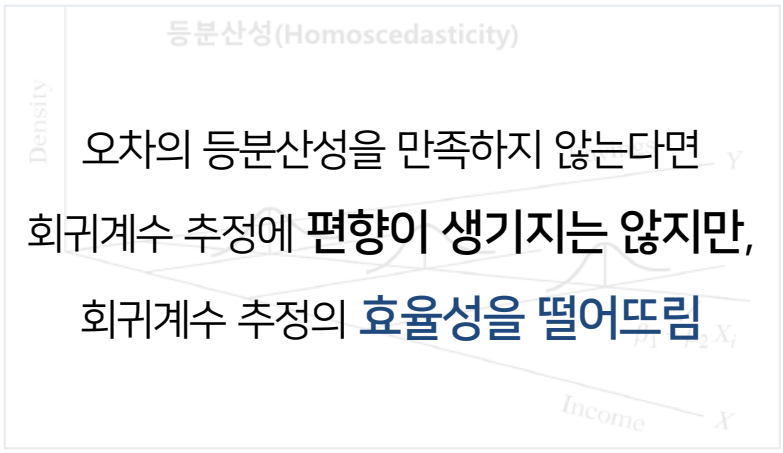
오차의 등분산성 *Homoscedasticity / Constant variance*

오차항의 **분산**은 **상수**다

분산은 σ^2으로 동일하다

↔ 이분산성(Heteroscedasticity)





오차의 등분산성을 만족하지 않는다면

회귀계수 추정에 **편향이 생기지는 않지만,**

회귀계수 추정의 **효율성을 떨어뜨림**

1 회귀 기본 가정

회귀분석의 가정들

오차의 등분산성

Homoscedasticity | Constant variance

오차항의 분산은 상수다

분산이 일정하지 않고 **변화**한다면

전체적인 회귀계수의 분산도 **커지게 되는데,**

그 결과 최소제곱추정량이 **BLUE의 조건**을 **만족하지 않아**

최소분산이 갖는 **효율성을 지니지 못함**

BLUE

: Best(분산이 제일 작은) Linear(선형) Unbiased Estimator(불편추정량)

[자세한 내용은 회귀 1주차 참고!]

1 회귀 기본 가정

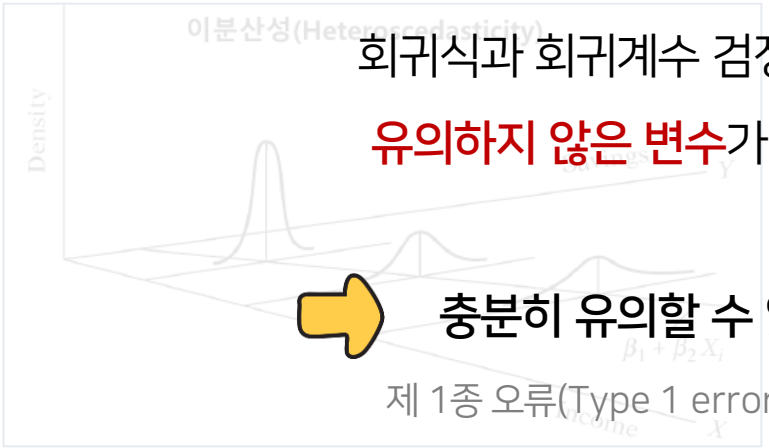
회귀분석의 가정들

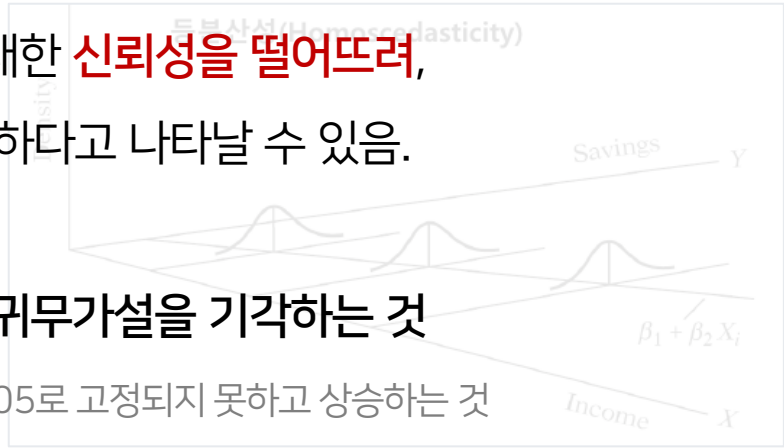
오차의 등분산성 *Homoscedasticity / Constant variance*

오차항의 **분산**은 **상수**다

분산은 σ^2으로 동일하다

↔ 이분산성(Heteroscedasticity)





회귀식과 회귀계수 검정에 대한 **신뢰성을 떨어뜨려**,

유의하지 않은 변수가 **유의**하다고 나타날 수 있음.

충분히 유의할 수 있는 귀무가설을 기각하는 것

제 1종 오류(Type 1 error)가 0.05로 고정되지 못하고 상승하는 것

분산이 일정하지 않고 변화한다면 전체적인 회귀계수의 분산도 커지게 되는데,
그 결과 최소제곱추정량이 BLUE의 조건을 만족하지 않아 최소분산이 갖는 효율성을 지니지 못한다.

1 회귀 기본 가정

회귀분석의 가정들

오차의 독립성 *Independence / No autocorrelation*

오차항은 **서로 독립**이다

↔ 자기상관성(Autocorrelation)

오차의 독립성이 **만족하지 않다면**, 최소제곱추정량이 **더 이상 BLUE가 아님**.

1 회귀 기본 가정

회귀분석의 가정들

오차의 독립성

Independence | No autocorrelation

오차항은 서로 독립이다

$\hat{\sigma}^2$의 추정량과 회귀계수의 표준오차가

실제보다 심각하게 과소추정됨

↔ 자기상관성(Autocorrelation)

오차의 독립성이 만족하지 않다면, 최소제곱추정량이 더 이상 BLUE가 아님.

유의성 검정의 결과를 신뢰할 수 없고,

Prediction Interval도 넓어지게 됨

$\hat{\sigma}^2$의 추정량과 회귀계수의 표준오차가 실제보다 심각하게 과소추정된다.
따라서 유의성 검정의 결과를 신뢰할 수 있고, Prediction Interval도 넓어지게 된다.

1 회귀 기본 가정

회귀분석의 가정들

오차의 독립성

Independence | No autocorrelation

오차항은 서로 독립이다

↔ 자기상관성(Autocorrelation)

선형성과 **오차의 정규성**, **등분산성**, **독립성**에

초점을 맞추어 **진단**과 **처방** 과정을 진행

오차의 독립성이 **만족하지 않다면**, 최소제곱추정량이 **더 이상 BLUE가 아님**.

σ^2의 추정량과 회귀계수의 표준오차가 실제보다 심각하게 과소추정된다.
따라서 유의성 검정의 결과를 신뢰할 수 없고, Prediction Interval도 넓어지게 된다.

2

잔차 플랏

기본 가정 진단

기본 가정 진단

회귀분석의 기본 가정을 진단하기 위해 두 가지의 방법이 사용됨

① 시각적 (graphical) 방법

② 가설 검정을 이용한 방법

물론 시각적 방법을 사용하더라도 판단에 대한
명확한 근거를 마련하기 위해 **가설 검정의 과정 필요**

2 잔차 플랏

시각적 방법

잔차 플랏 *Residual plot*

오차항의 추정량인 **잔차의 분포**를 통해 경험적 판단에 근거한 회귀 진단 가능

① Residuals vs Fitted

② Normal QQ
(Quantile-Quantile)

③ Scale - Location

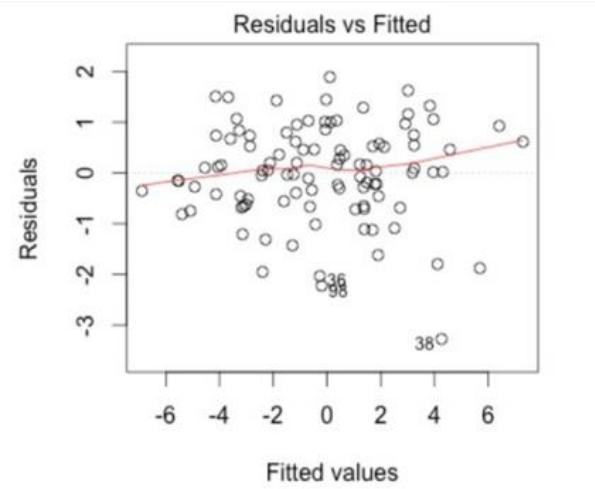
④ Residuals vs Leverage

2 잔차 플랏

잔차 플랏

① Residuals vs Fitted

선형성 & 오차의 등분산성 확인



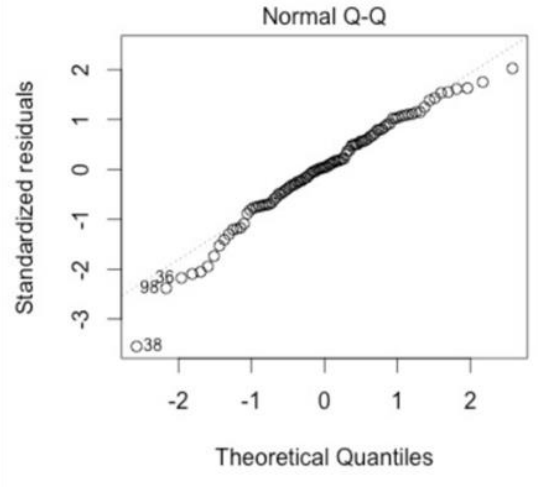
✓ 빨간 실선은 전체적인 잔차들의 추세선

잔차들의 분포를 Location Regression으로 추정한 완만한 연결 직선이며 **잔차 분포의 패턴, 경향성**을 나타내는 보조지표

잔차 플랏

② Normal QQ(Quantile-Quantile)

오차의 정규성 확인



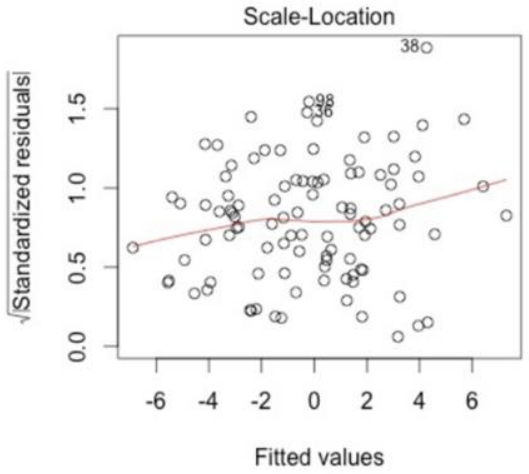
- ✓ 두 개 변수의 분포를 비교하기 위한 대표적인 **비모수적 방법**이자 **시각적 방법**
- ✓ 그래프가 $y = x$에 가까울 수록 잔차가 정규성을 만족

잔차 플랏

③ Scale - Location

선형성, 오차의 등분산성 확인

Mainly 등분산성



추세선 :

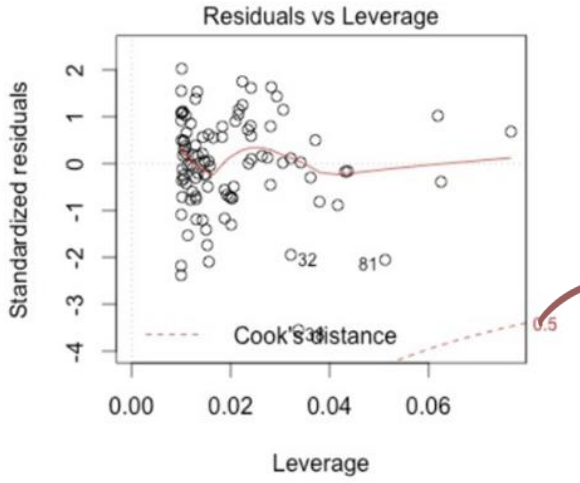
잔차들의 분포를 Local Regression으로 추정한 직선이며 **잔차 분포의 경향**을 나타내는 보조지표

✓ **잔차에 절댓값이 씌워진 형태**라는 점에서 Residual vs Fitted plot과 차이가 있음

잔차 플랏

④ Residuals vs Leverage

영향점(Influential point) 확인



- ✓ 플랏의 **오른쪽**에 위치한 점들이 **leverage가 큰 잔차**
- ✓ 빨간 실선으로부터 **위아래**로 멀리 떨어진 점들이 **outlier**
- Cook's distance로
  주로 **0.5와 1** 사이에서 경계가 표시됨

Cook's distance : 회귀직선의 모양에 영향을 미치는 점을 찾는 과정으로, 기울기, 절편에 크게 영향을 미치는 점을 찾는다
→ 이는 leverage와 잔차에 비례

3

선형성 진단과 처방

3 선형성 진단과 처방

선형성 가정

선형성 *Linearity*

설명변수와 반응변수의 관계는 **선형**이다

모델 자체가 **선형성만 고려**하고 있으므로

선형성을 만족하지 않는다면 **모델 자체가 성립하지 않음.**

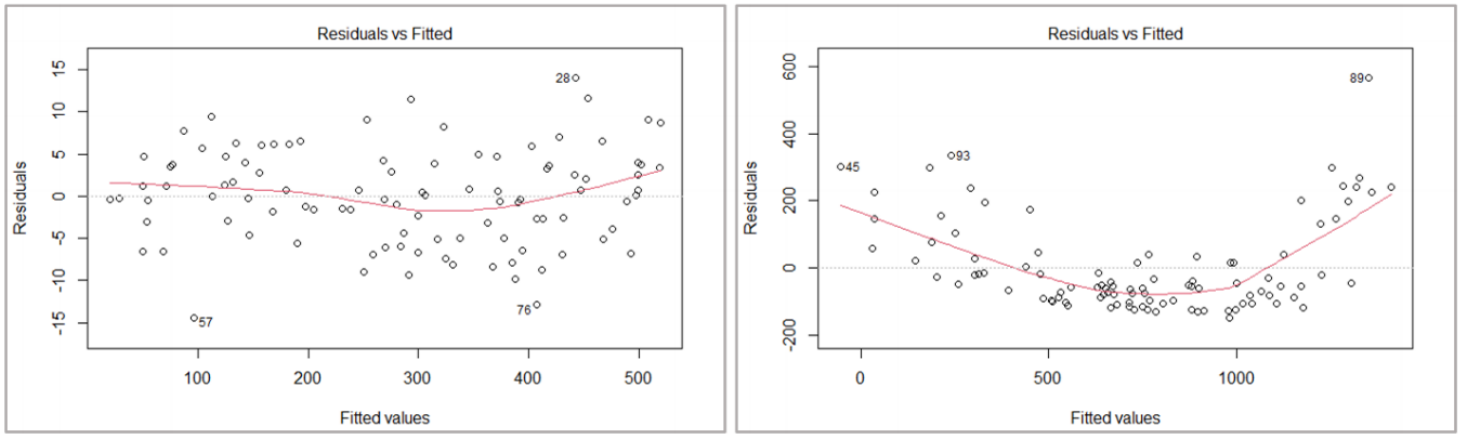
3 선형성 진단과 처방

진단 | ① 잔차 플랏을 통한 진단

잔차 플랏

평균 0을 중심으로 하는 x축에 평행한 **직선 형태가 아니라면**

선형성 위반되었다고 판단



잔차의 추세선이 x축과 비슷하지 않고 오른쪽처럼 이차함수 꼴이라면 선형성 위배

진단 | ② Partial residual plot을 통한 진단

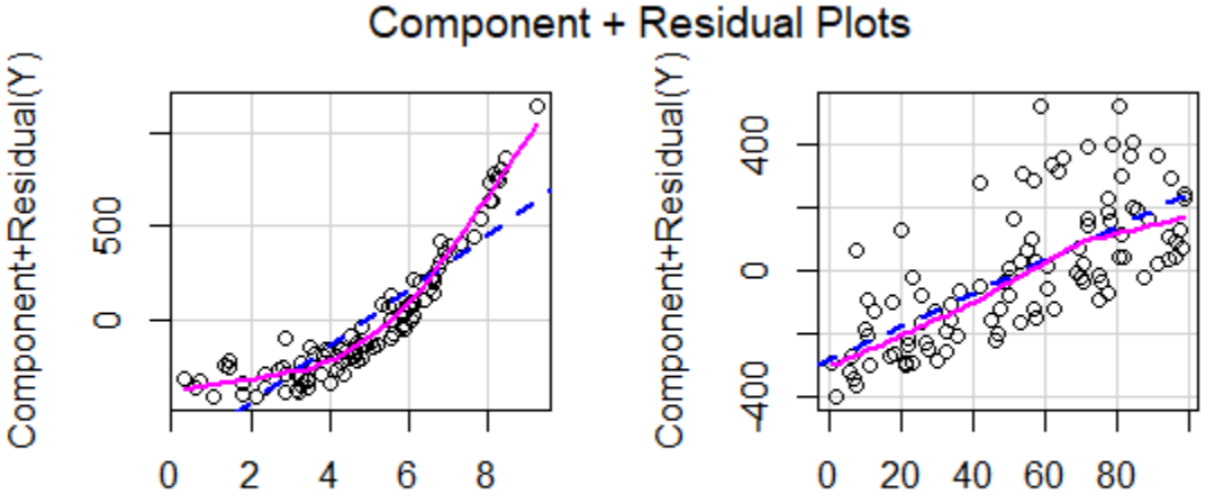
Partial Residual Plot

개별 독립 변수와 종속 변수 간의 선형성을 판단하기 좋은 플랏

선형성을 만족하지 못할 때, 어떤 변수의 영향으로
선형성이 만족되지 않는 것인지 **잔차 플랏으로는 확인이 어려움**

개별 변수의 영향을 확인하는 과정 필요

진단 | ② Partial residual plot을 통한 진단



✓ 파란 점선 : Partial residual과 x_i의 적합된 직선

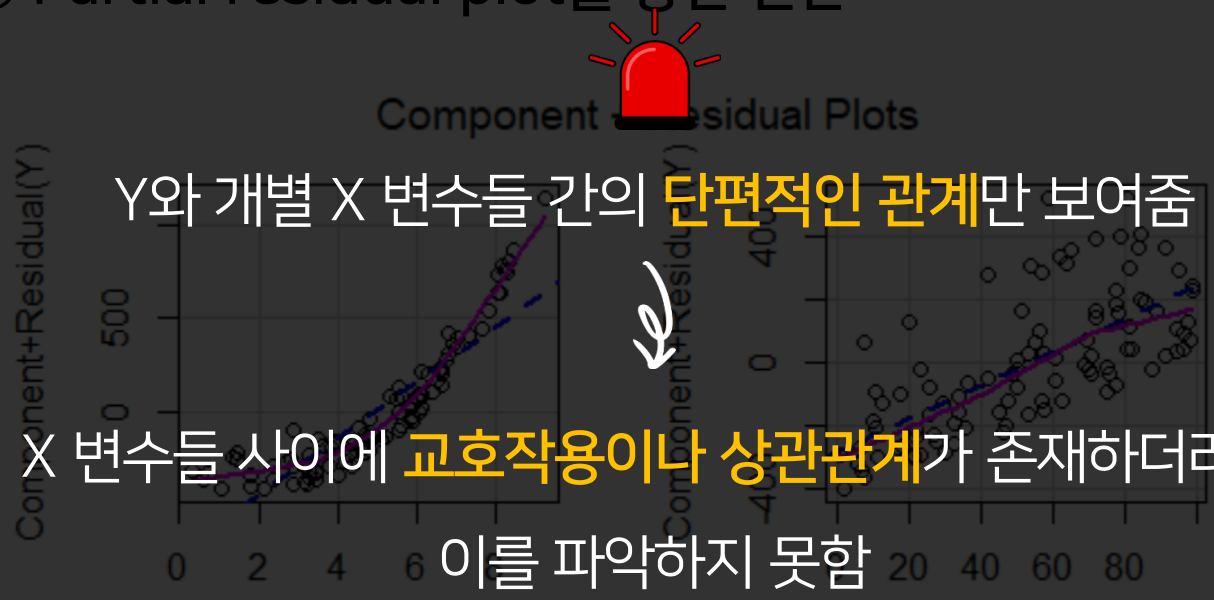
→ 점들의 분포를 최소제곱방법을 통해 회귀선을 추정한 것

✓ 핑크색 실선 : 잔차의 추세선

→ 새로운 변수에 의해 선형적으로 설명되어야 하는 부분

3 선형성 진단과 처방

진단 | ② Partial residual plot을 통한 진단



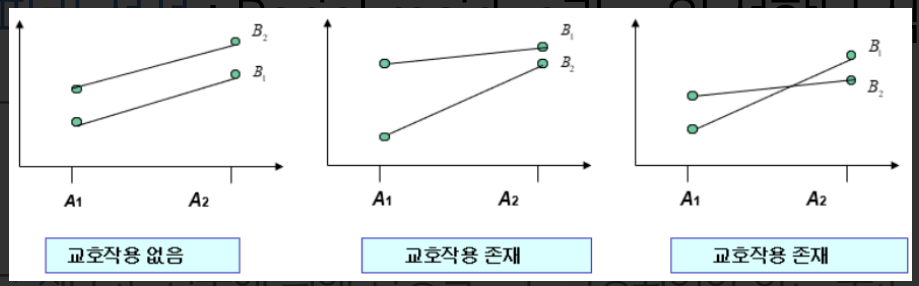
Y와 개별 X 변수들 간의 **단편적인 관계**만 보여줌

X 변수들 사이에 **교호작용이나 상관관계**가 존재하더라도

이를 파악하지 못함

교호작용 : 한 요인의 효과가 다른 요인의 수준에 의존하는 경우로 변수 간의 시너지 효과 의미

① [illegible] 선



처방 | ① 변수변환

변수 변환을 통해 **비선형 관계 해결**

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

✓ 변수변환을 통해 선형성을 확보할 수 있는 모델도 넓은 의미에서 선형모델이라 부름

처방 | ② 비선형 회귀

모델 자체를 **비선형 회귀 모델**에 적합

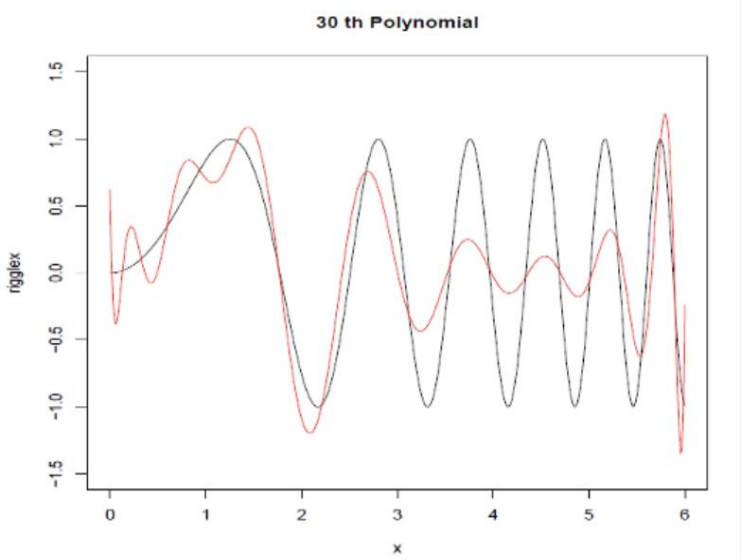
① Polynomial Regression

② Local Regression

처방 | ② 비선형 회귀

Polynomial Regression

고차항을 고려하는 다항회귀



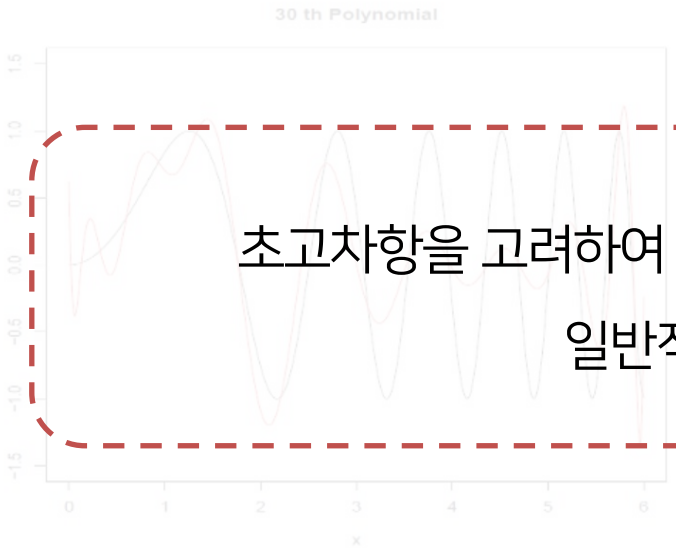
Residual plot과 Partial regression plot에서

이차 이상의 곡선 형태가 나타날 경우 **사용**

처방 | ② 비선형 회귀

Polynomial Regression

고차항을 고려하는 다항회귀



초고차항을 고려하여 적합하더라도 경향을 잡아내기 힘들어

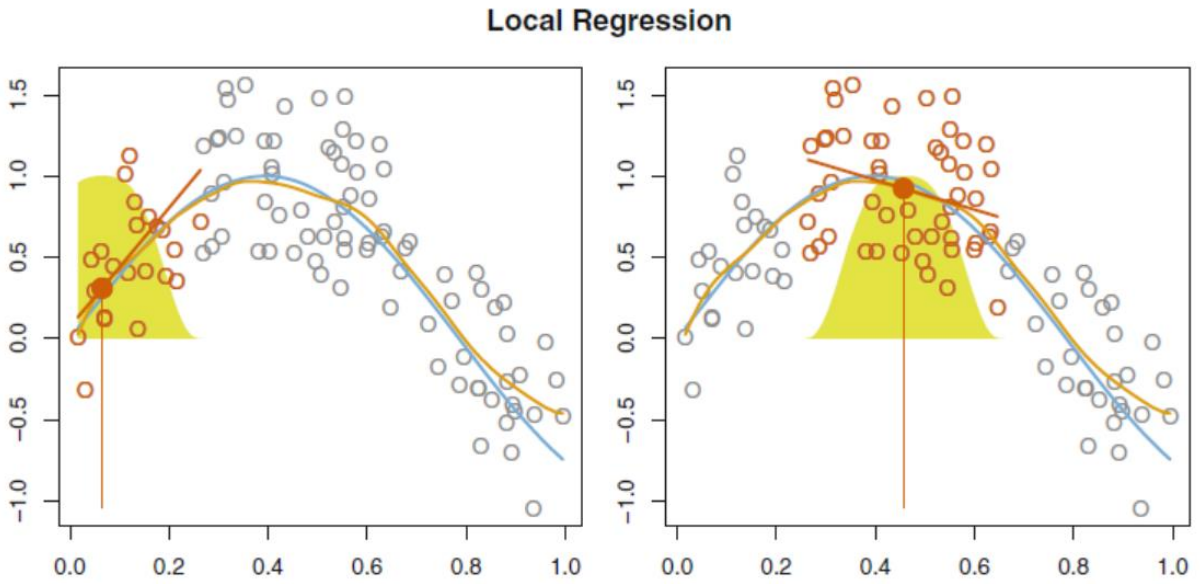
일반적으로 **3차까지만 고려**

처방 | ② 비선형 회귀

Local Regression

비선형 회귀 방법이자 **비모수적 방법**을 사용하는 회귀 모델

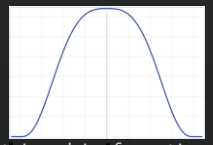
잔차 플랏의 추세선을 나타낼 때 쓰임



처방 | ② 비선형 회귀

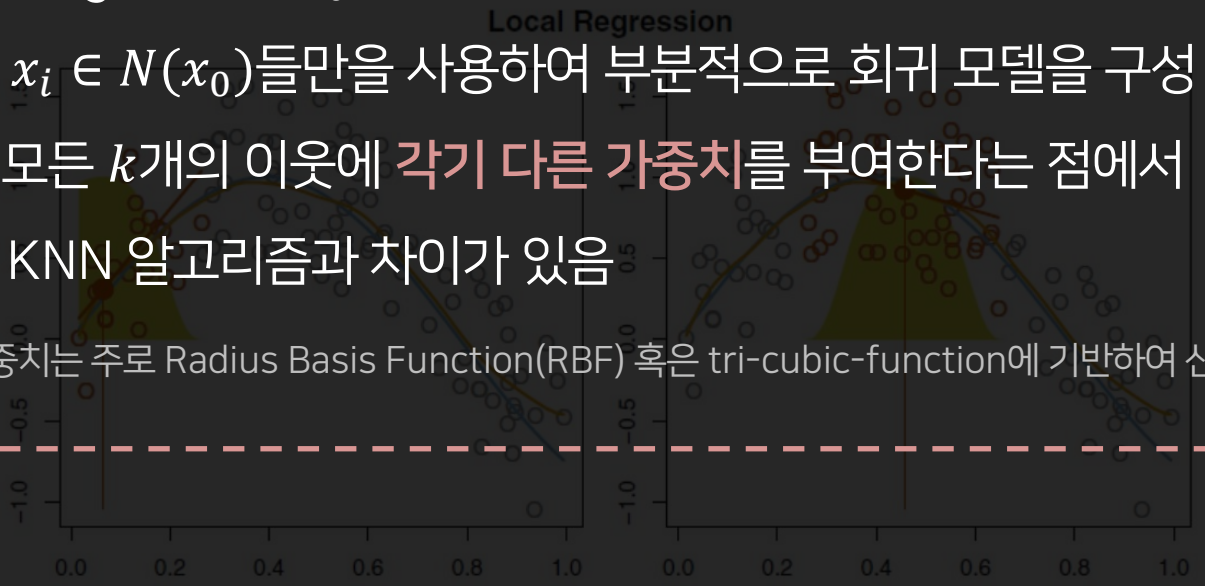
Local Regression

Local Regression



- Local(지역적인)에 있는 데이터들로 회귀 모델링 하는 방법
- Target data x_0를 중심으로 그 주변의 k개의 이웃 데이터 $x_i \in N(x_0)$들만을 사용하여 부분적으로 회귀 모델을 구성
- 모든 k개의 이웃에 각기 다른 가중치를 부여한다는 점에서 KNN 알고리즘과 차이가 있음

가중치는 주로 Radius Basis Function(RBF) 혹은 tri-cubic-function에 기반하여 산정



처방 | ② 비선형 회귀

Local Regression 추가 설명

비선형(non-linear), 비모수적(non-parametric) 방법을 이용한 회귀모델로,
물리적으로(주로 유클리드 거리) 가까운 값들은 종속변수에
비슷한 영향을 끼칠 것이라는 가정 하에 생겨난 방법

- ✓ Local(지역적인)에 있는 데이터들로 회귀 모델링 하는 방법
- ✓ Target data x_0를 중심으로 그 주변의 k개의 이웃 데이터 $x_i \in N(x_0)$들만을 사용하여 부분적으로 회귀 모델을 구성

특정한 target point x_0에 가까운 데이터들은 y의 예측에 중요하다고 판단하여
가중치를 많이 주고, 반대로 멀리 있는 데이터들은 가중치를 적게 주어 회귀모델 적합

4

정규성 진단과 처방

4 정규성 진단과 처방

정규성 가정

오차의 정규성 *Normality*

반응 변수를 측정할 때 발생하는 **오차**는 **정규분포**를 따를 것이라는 가정

회귀식이 데이터를 잘 표현한다면

잔차들은 단순한 측정 오차라 여겨지고,

잔차들의 분포는 정규분포와 흡사한 형태가 됨

4 정규성 진단과 처방

정규성 가정

오차의 정규성 *Normality*

반응 변수를 측정할 때 발생하는 **오차**는 **정규분포**를 따를 것이라는 가정

회귀식이 데이터를 잘 표현한다면

잔차들은 단순한 **측정 오차**라 여겨지고,

잔차들의 분포는 **정규분포**와 흡사한 형태가 됨

4 정규성 진단과 처방

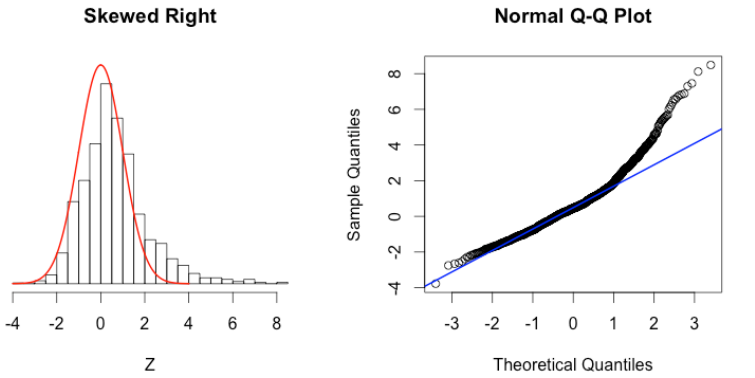
정규성 진단

Normal Q-Q Plot

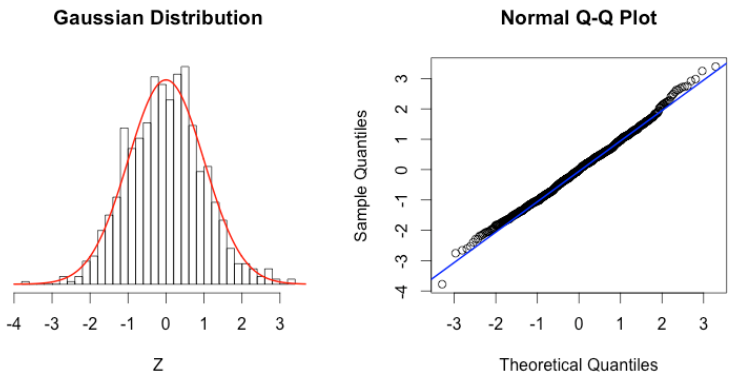
정규성을 파악하기 위한 **비모수적인 방법**

R에서 회귀식에 plot() 함수를 사용했을 때 두 번째로 나오는 plot

점들이 y=x 직선에 가까우면 정규성 만족



정규성을 만족하지 못한 경우



정규성을 만족한 경우

4 정규성 진단과 처방

정규성 진단

Normal Q-Q Plot

정규성을 파악하기 위한 비모수적인 방법

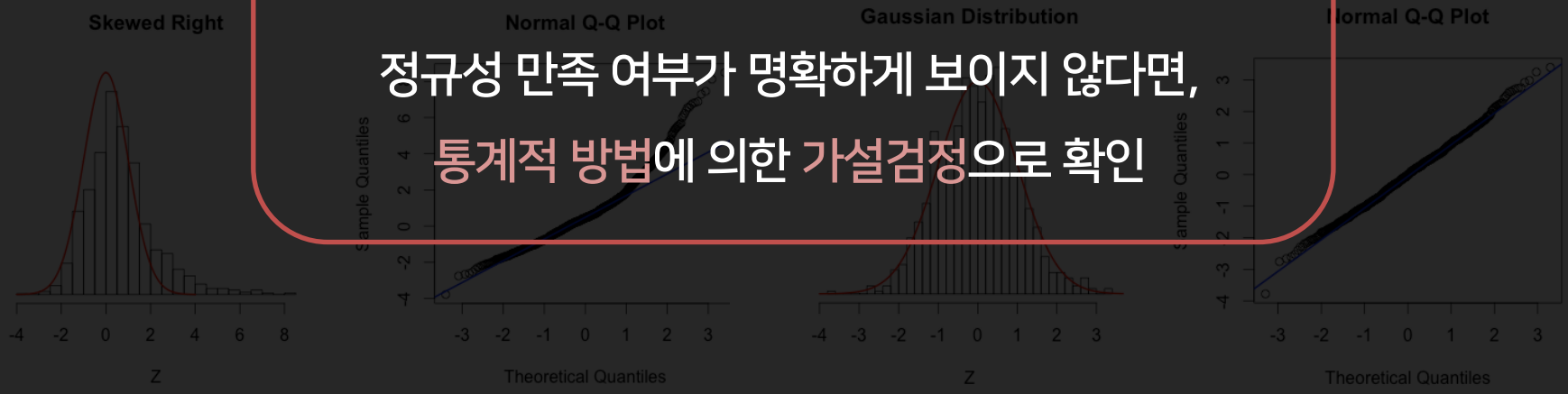
R에서 회귀식에 plot() 함수를 사용했을 때 두 번째로 나오는 plot

Plot으로 확인하는 경우, 판단이 주관적일 수도 있음

점들이 y=x 직선에 가까우면 정규성 만족

정규성 만족 여부가 명확하게 보이지 않다면,

통계적 방법에 의한 가설검정으로 확인



정규성을 만족하지 못한 경우

정규성을 만족한 경우

4 정규성 진단과 처방

정규성 진단

가설 설정

H_0: 주어진 데이터는 정규분포를 따른다.

H_1: 주어진 데이터는 정규분포를 따르지 않는다.

우리가 원하는 것은

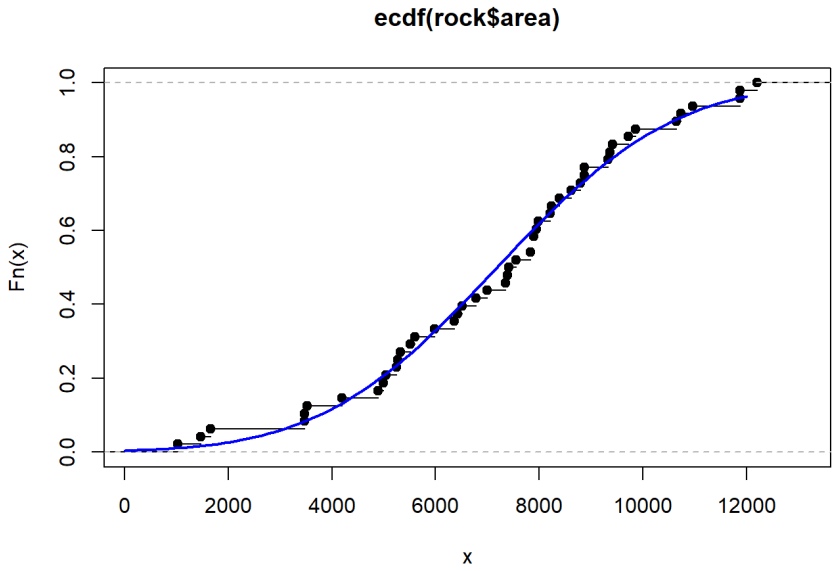
귀무가설을 **기각하지 못하는 것**!

정규성 진단

Empirical CDF (경험적 누적밀도함수)

관측치들을 **작은 순서대로** 나열한 후 관측치들로 그린 **누적 분포 함수**

잔차의 ECDF와 정규분포의 CDF와 비교하여 검정



4 정규성 진단과 처방

정규성 진단

Empirical CDF (경험적 누적밀도함수)

관측치들을 **작은 순서대로** 나열한 후 관측치들로 그린 **누적 분포 함수**

👉 잔차의 ECDF와 정규분포의 CDF와 비교하여 검정

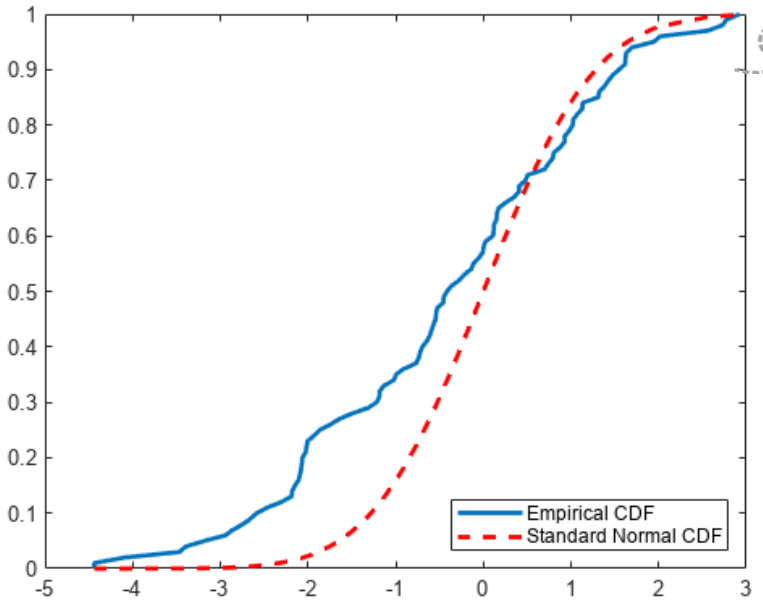
1. Kolmogorov Smirnov Test
2. Anderson Darling Test

정규성 진단

Kolmogorov-Smirnov Test (K-S검정)

귀무가설 하에서 표본분포함수가

어떤 이론적 **분포함수**와 **유사**한지 검정하는 방법



빨간색 선이 비교대상인 이론분포함수를
파란색 선이 표본분포함수를 의미함

둘의 차이가 크면 H_0 을 기각

잔차가 정규분포를 따르지 않음

4 정규성 진단과 처방

정규성 진단

Anderson-Darling Test (A-D검정)

K-S 검정을 수정한 적합도 검정으로,

분포의 **꼬리**에 K-S 검정보다 **가중치**를 더 두어 수행

데이터가 특정 분포를 따르는지 검정하는 적합도 검정 중 하나

정규성 검정을 위해 …

K-S 검정은 0.15로, A-D 검정은 0.05로 유의수준 설정

유의수준이 다른 이유는 각 검정 방법별로 검정력에 차이가 있기 때문

A-D 방법이 K-S 방법보다 더 엄격한 검정 방법

4 정규성 진단과 처방

정규성 진단

Anderson-Darling Test (A-D검정)

K-S 검정을 수정한 적합도 검정으로,

분포의 **꼬리**에 K-S 검정보다 **가중치**를 더 두어 수행

데이터가 특정 분포를 따르는지 검정하는 적합도 검정 중 하나

정규성 검정을 위해 …

K-S 검정은 **0.15**로, A-D 검정은 **0.05**로 유의수준 설정

유의수준이 다른 이유는 각 검정 방법별로 검정력에 차이가 있기 때문

A-D 방법이 K-S 방법보다 더 엄격한 검정 방법

정규성 진단

정규분포의 **분포적 특성**을 이용한 Test

1. Shapiro Wilk Test
2. Jarque Bera Test

4 정규성 진단과 처방

정규성 진단

Shapiro Wilk Test

정규분포 분위수 값과 **표준화 잔차** 사이의 **선형관계** 확인

관측치가 5000개 이하인 데이터에서만 가능

R에서 Shapiro.test() 함수 안에 residual 값을 넣어 확인 가능

H_0 을 기각하지 못 했다는 것은

정규분포를 따르지 않는다고 말할 근거가 부족할 뿐!

100% 정규성을 만족한다는 뜻이 아닐 수도 있음을 주의

4 정규성 진단과 처방

정규성 진단

Jarque-Bera Test

정규분포의 **왜도가 0**, **첨도가 3**이라는 사실에서 기인한 방법

Tseries 패키지의 jarque.bera.test() 함수 안에 residual 값을 넣어 확인

n은 데이터의 계수, s는 표본의 왜도, k는 표본의 첨도를 의미

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

이상치에 민감한 왜도를 사용하기 때문에

이상치를 삭제했을 때

정규분포임이 드러나는 경우가 많음

잔차의 분포가 정규분포와 다를수록 왜도나 첨도의 변화가 생김

결국, 통계량 값이 커져 유의수준을 넘어서면 H_0 을 기각

4 정규성 진단과 처방

정규성 진단

Jarque-Bera Test

정규분포의 **왜도가 0**, **첨도가 3**이라는 사실에서 기인한 방법

Tseries 패키지의 jarque.bera.test() 함수 안에 residual 값을 넣어 확인

n은 데이터의 계수, s는 표본의 왜도, k는 표본의 첨도를 의미

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

이상치에 민감한 왜도를 사용하기 때문에

이상치를 삭제했을 때

정규분포임이 드러나는 경우가 많음

잔차의 분포가 정규분포와 다를수록 왜도나 첨도의 변화가 생김

결국, 통계량 값이 커져 유의수준을 넘어서면 H_0 을 기각

4 정규성 진단과 처방

정규성 가정이 위배될 경우 발생하는 문제점

회귀 분석에 사용되는 **F-test, T-test**는 모두 **정규분포**를 전제

정규성 가정이 위배된다면...

가설 검정 결과가 p-value에 의해 유의하게 나오더라도

검정 결과와 예측 결과를 신뢰할 수 없음

검정통계량이 t분포 혹은 F분포를 따르지 않기 때문

4 정규성 진단과 처방

정규성 가정이 위배될 경우 발생하는 문제점

회귀 분석에 사용되는 **F-test, T-test**는 모두 **정규분포**를 전제

정규성 가정이 위배된다면...

가설 검정 결과가 p-value에 의해 유의하게 나오더라도

검정 결과와 **예측 결과**를 **신뢰할 수 없음**

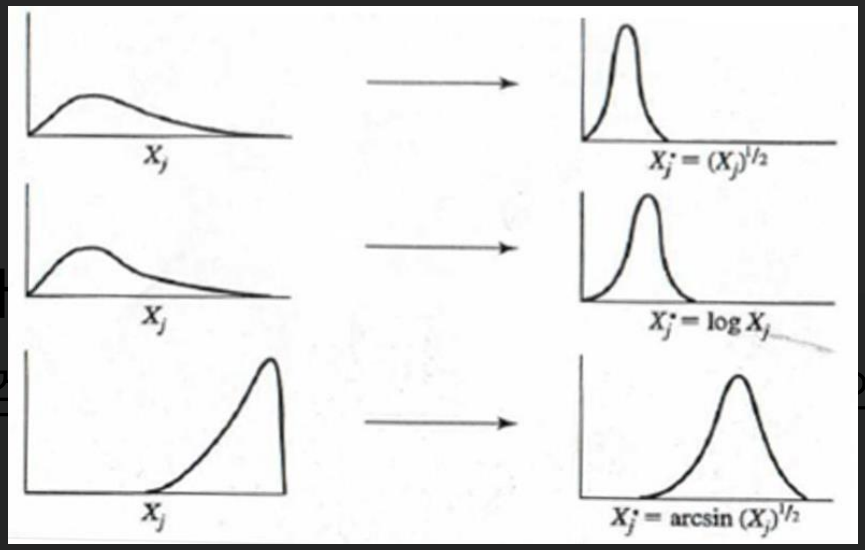
검정통계량이 t분포 혹은 F분포를 따르지 않기 때문

4 정규성 진단과 처방

정규성 가정이 위배될 경우 발생하는 문제점

회귀 분석에 사용되는 F-test, T-test는 모두 정규분포를 전제

변수 변환을 통해서 **정규성**을 **처방**해줄 수 있음



정규성 가

가설 검

오더라도

검정통계량이 t분포 혹은 F분포를 따르지 않기 때문

4 정규성 진단과 처방

정규성 가정이 위배될 경우 발생하는 문제점

회귀분석에 사용되는 F-test, T-test는 모두 정규분포를 전제

그러나 ...

변수 변환이 주관적 판단 하에 이루어지기 때문에

객관성을 확보하기 어려움

정규성 가정이 위배된다면...

가설 검정 결과가 p-value에 의해 유의하게 나오더라도

검정 결과에 대한 결론을 내리기 어려움

Box-cox Transformation

Yeo-Johnson Transformation

검정통계량이 t분포 혹은 F분포를 따르지 않기 때문

정규성 처방

Box-cox Transformation

통계적인 검정에 따라 변수 변환(비선형 변환)을 진행해주는 방법

Car 패키지의 powerTransform을 통해 구현 가능

일반적으로 λ는 -5에서 5 사이의 값
λ=0일 때 log transformation을 해줌

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

λ를 **변화**시키면서

y가 **정규성, 등분산성**을 만족하도록!

최적의 λ은 최대우도함수를 통해 구한 후

신뢰구간 내 로그우도함수를 최대화하는 λ를 최적의 값으로 선택

4 정규성 진단과 처방

정규성 처방

Box-cox Transformation

통계적인 검정에 따라 변수 변환(비선형 변환)을 진행해주는 방법

Car 패키지의 powerTransform을 통해 구현 가능

일반적으로 λ는 -5에서 5 사이의 값
λ=0일 때 log transformation을 해줌

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

λ를 **변화**시키면서

y가 **정규성, 등분산성**을 만족하도록!

최적의 λ은 최대우도함수를 통해 구한 후

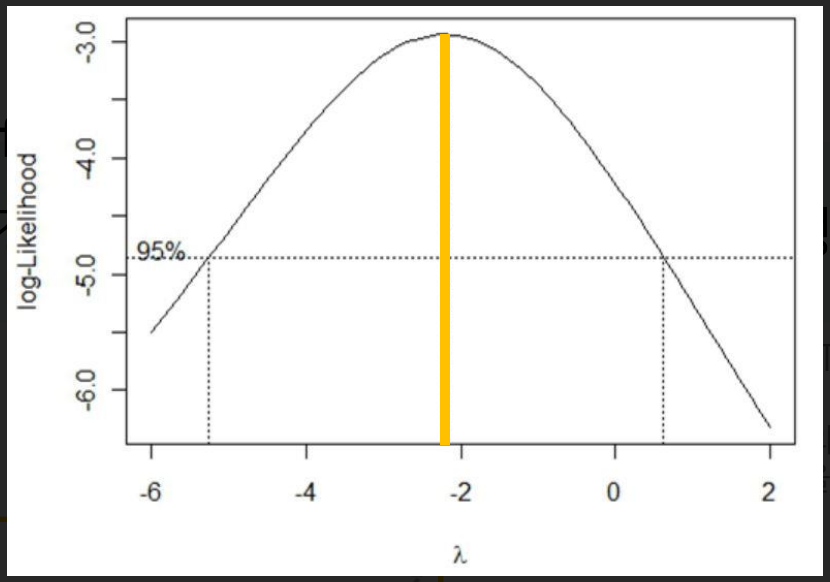
신뢰구간 내 로그우도함수를 최대화하는 λ를 최적의 값으로 선택

4 정규성 진단과 처방

정규성 처방

Box-cox Trans

통계적인 검 해주는 방법



Transform을 통해 구현 가능

값
해줌

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

λ를 **변화**시키면서

95% 내의 λ값 중 **가능도함수**가 **정규성, 등분산성**을 만족하도록!

최대가 되게 하는 **-2 근방**의 λ를 선택

최적의 λ은 최대우도함수를 통해 구한 후

λ를 -2로 선택하는 것도 가능!

신뢰구간 내 로그우도 함수를 최대화하는 λ를 최적의 값으로 선택

λ를 정수로 선택했을 때 변수 변환 관계 파악이 쉽다는 장점이 있음

4 정규성 진단과 처방

정규성 처방

Box-cox Transformation

통계적인 검정에 따라 변수 변환(비선형 변환)을 진행해주는 방법

그러나

Box-cox Transformation은

y가 0 이하일 경우에는 사용할 수 없음

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

λ = 0이 되면 y가 log(y)로 변환될 수 있기 때문

λ를 **변화**시키면서

y가 **정규성, 등분산성**을 만족하도록!

최적의 λ은 최대우도함수를 통해 구한 후

신뢰구간 내 로그우도함수를 최대화하는 λ를 최적의 값으로 선택

정규성 처방

Yeo-Johnson Transformation

Box-cox Transformation과 같은 아이디어

단, Box-cox와 달리 변수 범위에 대한 제약이 없음

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

4 정규성 진단과 처방

정규성 처방

Yeo-Johnson Transformation

Box-cox Transformation과 같은 아이디어!

단, Box-cox와 달리 변수 범위에 대한 제약이 없음

Yeo-Johnson Transformation이 전체 범위에서 가능함에도 불구하고

Box-cox Transformation도 사용하는 이유?

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

Box-cox Transformation은 해석에 장점이 있기 때문!

Yeo-Johnson Transformation처럼

제곱이 다르게 부여되면 전체 범위에 대한 해석이 모호해질 수 있음

Yeo-Johnson Transformation은 제곱이 $\lambda, 2-\lambda$로 다름

4 정규성 진단과 처방

정규성 처방

Yeo-Johnson Transformation

Yeo-Johnson Transformation이 전체 범위에서 가능함에도 불구하고

Box-cox Transformation도 사용하는 이유?

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

Box-cox Transformation은 해석에 장점이 있기 때문!

Yeo-Johnson Transformation처럼

제곱이 다르게 부여되면 전체 범위에 대한 해석이 모호해질 수 있음

Yeo-Johnson Transformation은 제곱이 $\lambda, 2-\lambda$로 다름

5

등분산성 진단과 처방

5 등분산성 진단과 처방

등분산성 가정

오차의 등분산성 *Homoscedasticity / Constant variance*

오차의 **모든 분산**은 **동일**함

회귀식이 데이터를 잘 표현한다면...

분산이 **상수**여서 어느 관측에서나 동일하게 나타남

다른 변수의 **영향을 받지 않음**

5 등분산성 진단과 처방

등분산성 가정

오차의 등분산성 *Homoscedasticity / Constant variance*

오차의 **모든 분산**은 **동일**함

회귀식이 데이터를 잘 표현한다면...

분산이 **상수**여서 어느 관측에서나 동일하게 나타남

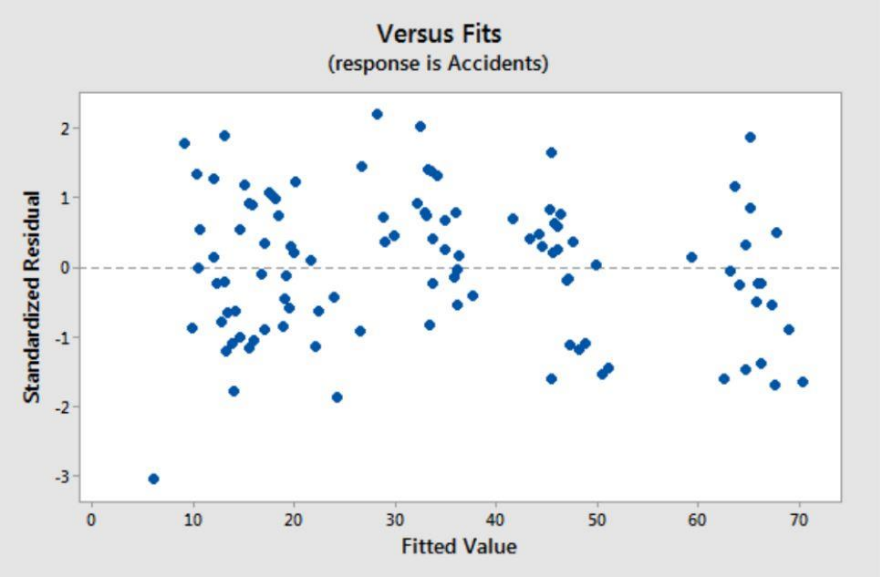
다른 변수의 **영향을 받지 않음**

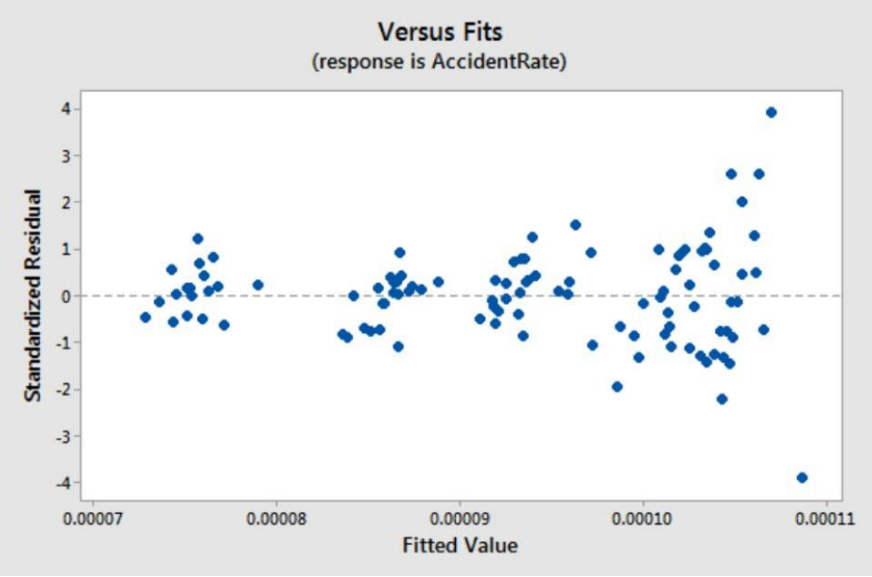
5 등분산성 진단과 처방

등분산성 진단

잔차 플랏

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





등분산성을 만족한 경우

$\hat{y}$값에 상관없이

잔차의 퍼짐의 정도가 일정함

등분산성을 만족하지 못한 경우

$\hat{y}$값이 커지면서

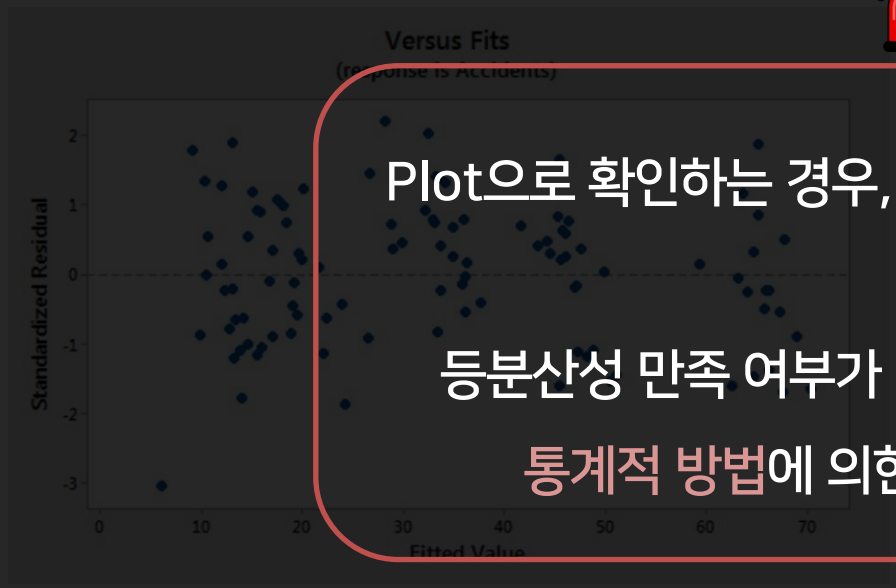
잔차의 퍼짐의 정도가 일정하지 않음

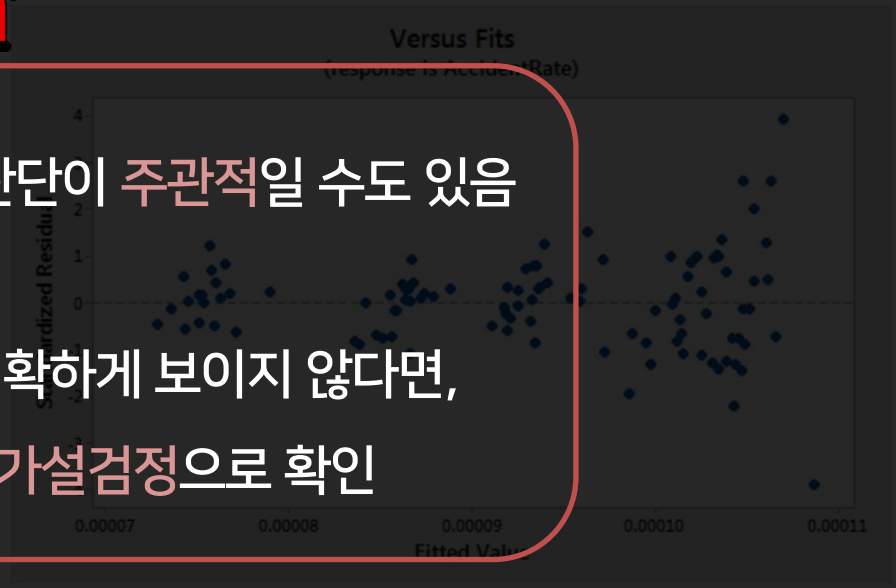
5 등분산성 진단과 처방

등분산성 진단

잔차 플랏

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





Plot으로 확인하는 경우, 판단이 주관적일 수도 있음

등분산성 만족 여부가 명확하게 보이지 않다면,

통계적 방법에 의한 가설검정으로 확인

등분산성을 만족한 경우

$\hat{y}$값에 상관없이

잔차의 퍼짐의 정도가 일정함

등분산성을 만족하지 못한 경우

$\hat{y}$값이 커지면서

잔차의 퍼짐의 정도가 일정하지 않음

5 등분산성 진단과 처방

등분산성 진단

BP(Breusch-Pagan) Test

잔차가 **독립변수**들의 **선형결합**으로 표현되는지 검정

R에서 lmstat 패키지의 bptest() 함수를 이용

👉 설명변수의 증감에 따른 **오차의 분산 변화**를 통해 **등분산성** 지니는지 판단 가능

BP Test 기본 가정

- ✔ 샘플 수가 많아야 함
- ✔ 오차항은 독립이고 정규분포를 따름
- ✔ 오차의 분산은 설명변수와 연관이 있음

5 등분산성 진단과 처방

등분산성 진단

BP(Breusch-Pagan) Test

잔차가 **독립변수**들의 **선형결합**으로 표현되는지 검정

R에서 lmstat 패키지의 bptest() 함수를 이용

👉 설명변수의 증감에 따른 **오차의 분산 변화**를 통해 **등분산성** 지니는지 판단 가능

BP Test 기본 가정

- ✔ 샘플 수가 많아야 함
- ✔ 오차항은 독립이고 정규분포를 따름
- ✔ 오차의 분산은 설명변수와 연관이 있음

5 등분산성 진단과 처방

등분산성 진단

가설 설정

H_0: 주어진 데이터는 등분산성을 지닌다.

H_1: 주어진 데이터는 등분산성을 지니지 않는다.

우리가 원하는 것은

귀무가설을 기각하지 못 하는 것!

5 등분산성 진단과 처방

등분산성 진단

결정계수를 통해 잔차의 제곱이

독립변수의 **선형결합**으로 표현되는지와 그때의 **설명력**을 파악

오차가 독립변수에 의해 충분히 표현된다면 결정계수와 검정통계량이 커질 것!

$$e^2 = \gamma_0 + \gamma_1 x_1 + \cdots \gamma_P x_P + \epsilon'$$

X 변수를 선형결합한 식에서 수정계수를 구함

5 등분산성 진단과 처방

등분산성 진단

결정계수를 통해 잔차의 제곱이

독립변수의 **선형결합**으로 표현되는지와 그때의 **설명력**을 파악

오차가 독립변수에 의해 충분히 표현된다면 결정계수와 검정통계량이 커질 것!

$$e^2 = \gamma_0 + \gamma_1 x_1 + \cdots \gamma_P x_P + \epsilon'$$

X 변수를 선형결합한 식에서 수정계수를 구함

5 등분산성 진단과 처방

등분산성 진단

검정 통계량

$$\chi^2_{stat} = nR^2 \sim \chi^2_{P-1}$$

임계값

$$\chi^2_{P-1,\alpha}$$

귀무가설 기각 if $\chi^2_{stat} > \chi^2_{P-1,\alpha}$

➔ 등분산성을 만족하지 않음

5 등분산성 진단과 처방

등분산성이 위배되었을 경우

① OLS 추정량의 분산이 실제 분산보다 작게 추정됨

- 검정통계량 증가 & P-value 감소
- 유의하지 않은 회귀 계수를 유의하다고 판단 (제 1종 오류)

② 조건을 만족하지 않으므로, OLS 추정량이 BLUE가 되지 못함

5 등분산성 진단과 처방

처방 | ① 변수 변환

변수 변환 *Variable Transformation*

정규성을 만족하기 위해서 사용한 **각종 변수 변환 방법** 똑같이 적용

가중 회귀 제곱 *Weighted Least Square*

등분산이 아닌 형태의 데이터마다 다른 가중치를 주어서
등분산을 만족하게 해주는 '일반화된 최소제곱법'의 한 형태

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

w_i는 가중치이며, 분산에 반비례

처방 | ② 가중 회귀 제곱

변수 변환 *Variable Transformation*

정규성을 만족하기 위해서 사용한 각종 변수 변환 방법 똑같이 적용

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

처방 | ② 가중 회귀 제곱

변수 변환 *Variable Transformation*

정규성을 만족하기 위해서 사용한 각종 변수 변환 방법 똑같이 적용

가중 회귀 제곱 *Weighted Least Square*

등분산이 아닌 형태의 데이터마다 **다른 가중치**를 주어서

등분산을 만족하게 해주는 '**일반화된 최소제곱법**'의 한 형태

가중 회귀 제곱을 통해 구한 추정량은

회귀 기본 가정 하에 **BLUE** 만족

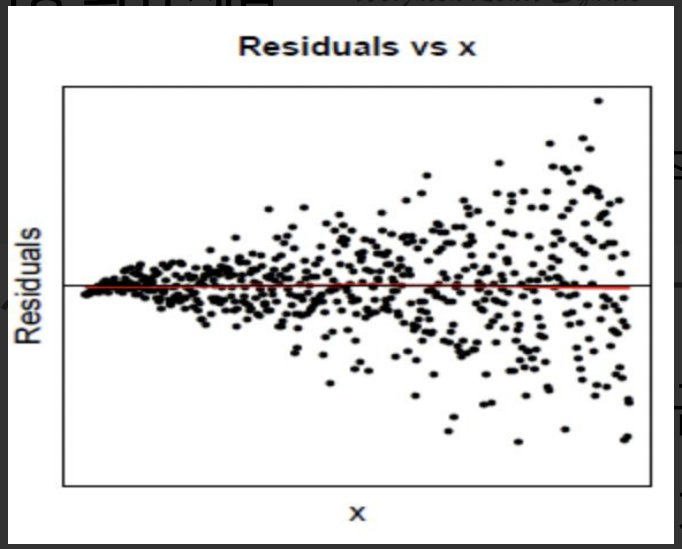
5 등분산성 진단과 처방

처방 | ② 가중 회귀 제곱

가중치 선정 방식

1. 잔차플랏 이용

가중 회귀 제곱 *Weighted Least Square*



데이터마다 다른 가중치를 주어서

residual plot에서

분산이 점점 커질 경우,

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

5 등분산성 진단과 처방

처방 | ② 가중 회귀 제곱

2. 모델 기반 선정

1. OLS로 다중선형회귀 모형 적합

가중 회귀 제곱 *Weighted Least Square*

등분산이 아닌 형태의 데이터마다 다른 가중치를 주어서
2. 다중선형회귀 모델 추정
등분산을 만족하게 해주는 '일반화된 최소제곱법'의 한 형태
*종속변수: 잔차의 제곱 / 독립변수: 오차 분산에 영향을 주는 변수

3. n개의 적합값 구하기

가중 회귀 제곱을 통해 구한 추정량은
회귀 기본 가정 하에 BLUE 만족

5 등분산성 진단과 처방

처방 | ② 가중 회귀 제곱

2. 모델 기반 선정

4. 가중치 설정

적합값 제곱의 역수를 가중치로 설정, 기존 데이터에 가중회귀모델 적용

5. 비교

처음 모델과 가중치 적용 모델의 회귀계수 비교

회귀계수의 차이가 작은 최적의 모형 선택

6

독립성 진단과 처방

6 독립성 진단과 처방

독립성 가정

독립성 가정 *Independence / No autocorrelation*

오차항끼리 **서로 독립**이라는 가정

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

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

오차들 간 상관성의 pattern이 있다는 것!

6 독립성 진단과 처방

독립성 가정

독립성 가정 *Independence / No autocorrelation*

오차항끼리 서로 독립이라는 가정

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

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

오차들 간 상관성의 pattern이 있다는 것!

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정

더빈-왓슨 검정 *Durbin-Watson Test*

앞 뒤 관측치의 **1차 자기상관성**을 확인하는 검정

1차 자기상관성: 연이어서 등장하는 오차들이 상관성을 지니는 것

귀무가설	H_0 : 잔차들 간에 1차 자기상관이 없다. = 잔차들이 서로 독립이다.
대립가설	H_1 : 잔차들 간에 1차 자기상관이 있다. = 잔차들이 서로 독립이 아니다.

진단 | ① 더빈-왓슨 검정

더빈-왓슨 검정 *Durbin-Watson Test*

앞 뒤 관측치의 **1차 자기상관성**을 확인하는 검정

1차 자기상관성: 연이어서 등장하는 오차들이 상관성을 지니는 것

검정통계량

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

First order autocorrelation

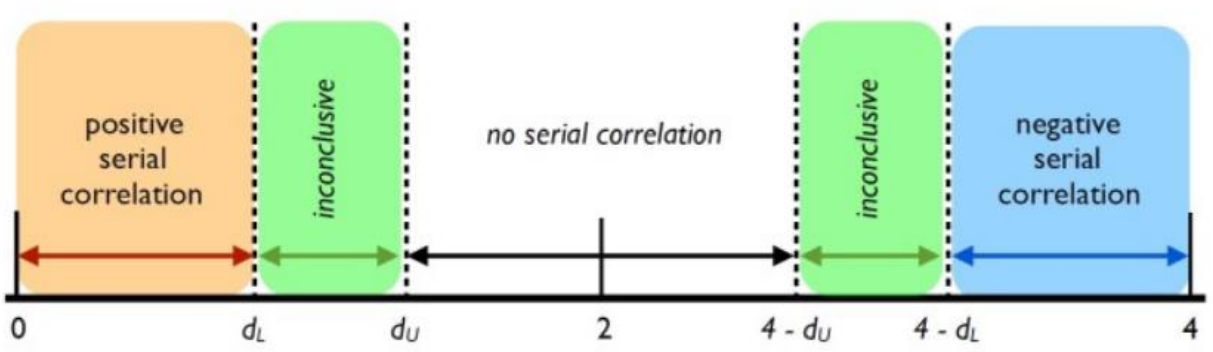
$$\widehat{\rho_1} = \frac{\widehat{Cov}(e_i, e_{i-1})}{\sqrt{V(e_i)} \cdot \sqrt{V(e_{i-1})}} \approx \frac{\sum_{i=1}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^{\,2}}$$

$$\therefore d \approx 2(1 - \widehat{\rho_1}) \quad \text{(범위: } [0,4]\text{)}$$

$\widehat{\rho_1}$: 표본 잔차 자기상관

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정



더빈 왓슨 검정표에서 데이터 개수 n과 변수의 개수 p에 따라

귀무가설 기각 여부를 판단하는 **컷 오프 값**을 알려줌

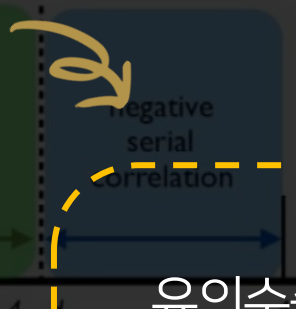
상한(d_U)과 하한(d_L)은 유의수준, 관측치 수, 설명변수 개수에 따라 달라짐

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정

더빈 왓슨 검정표

	k'*=1		k'=2		k'=3		k'=4		k'=5	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.390	1.142	-----	-----	-----	-----	-----	-----	-----	-----
7	0.435	1.036	0.294	1.676	-----	-----	-----	-----	-----	-----
8	0.497	1.003	0.345	1.489	0.229	2.102	-----	-----	-----	-----
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	-----	-----
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049
15	0.811	1.070	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736



유의수준: 0.01

n=15, p=3 일 때,

dL : 0.591

dU : 1.465

의 개수 p에 따라

값을 알려줌

관측치 수, 설명변수 개수에 따라 달라짐

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정

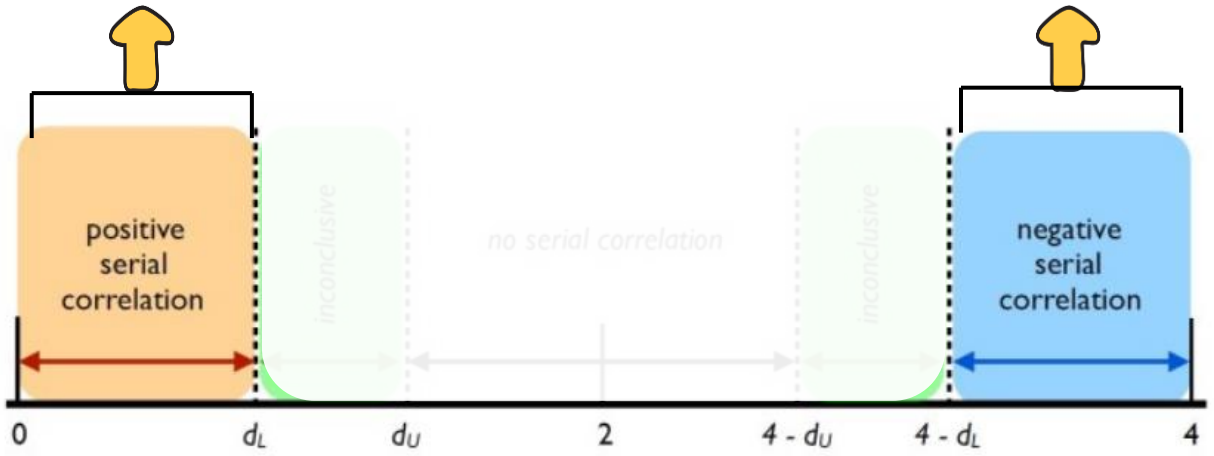
귀무가설 기각 = 잔차들 간에 1차 자기상관이 있음

if $d <$ 하한(d_L)

➔ 양의 자기상관이 있음

if $d > (4 - d_L)$

➔ 음의 자기상관이 있음

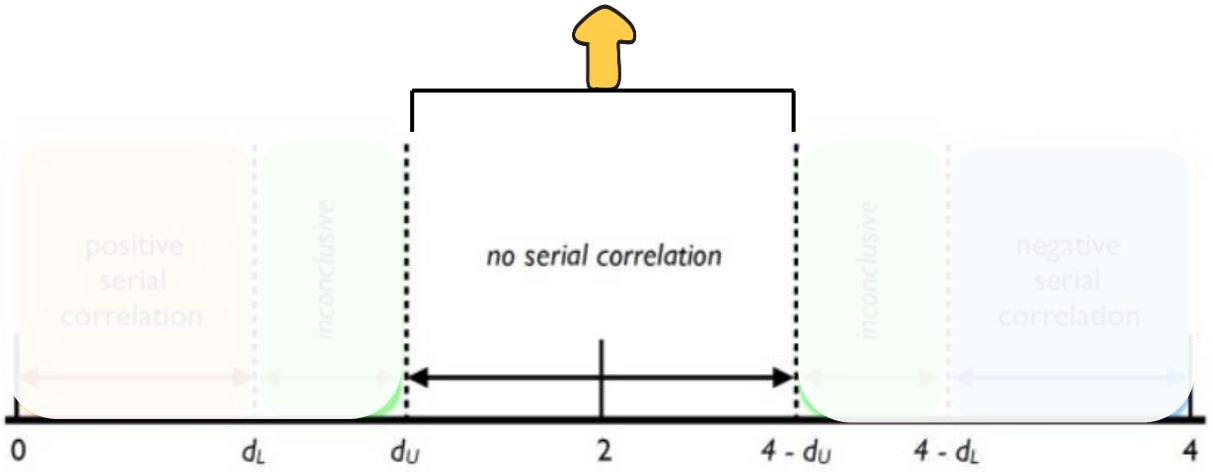


6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정

귀무가설 기각 안 됨 = 잔차들 간에 1차 자기상관이 없음

if $d >$ 상한(d_U) or $d < (4 - d_U)$



6 독립성 진단과 처방

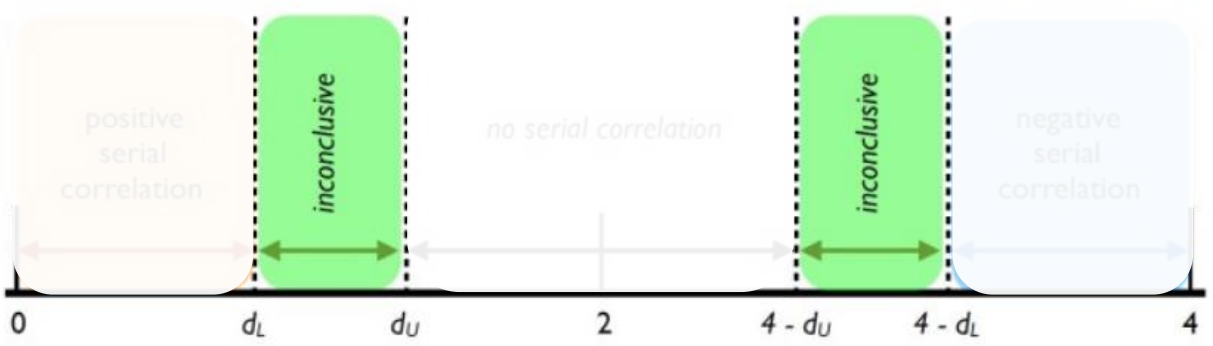
진단 | ① 더빈-왓슨 검정

한계

① d가 **상한과 하한 사이에 위치**한다면 판단할 수 없음

② 바로 인접한 오차와의 1차 자기상관만 고려함

➔ **자기상관이 오래 지속**되거나, **계절성**이 있는 경우 확인이 힘듦

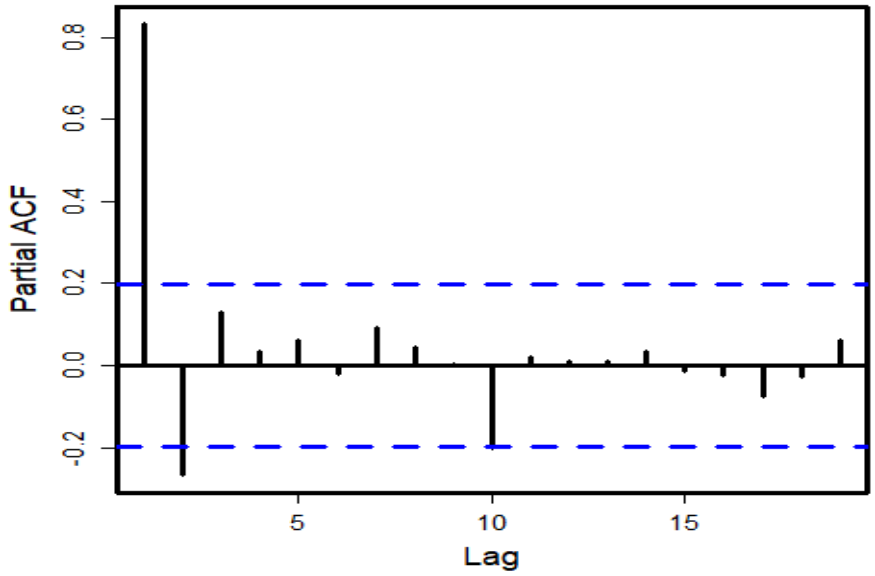


진단 | ② Autocorrelation function plot (ACF plot)

Autocorrelation function

1차 자기상관부터 p차 자기 상관까지 고려하고,

신뢰구간을 반환하므로 통계적인 절차에 따라 판단 가능



진단 | ② Autocorrelation function plot (ACF plot)

Autocorrelation function

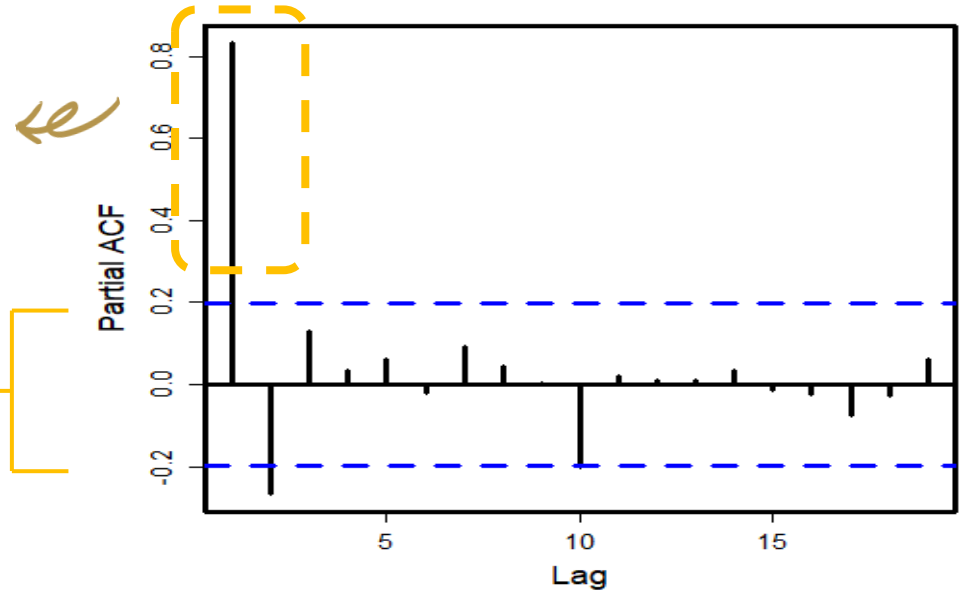
1차 자기상관부터 p차 자기 상관까지 고려하고,

신뢰구간을 반환하므로 통계적인 절차에 따라 판단 가능

신뢰구간을 벗어나는 선

: P차 자기 상관이 있다고 간주

신뢰구간



6 독립성 진단과 처방

진단 | ② Autocorrelation function plot(ACF plot)

Autocorrelation function

독립성이 위배되었을 경우

1차 자기 상관이 있다고 가정하고,

신뢰구간을 반환하므로 통계적인 절차에 따라 판단 가능

① LSE의 가정 세 가지를 만족하지 못함

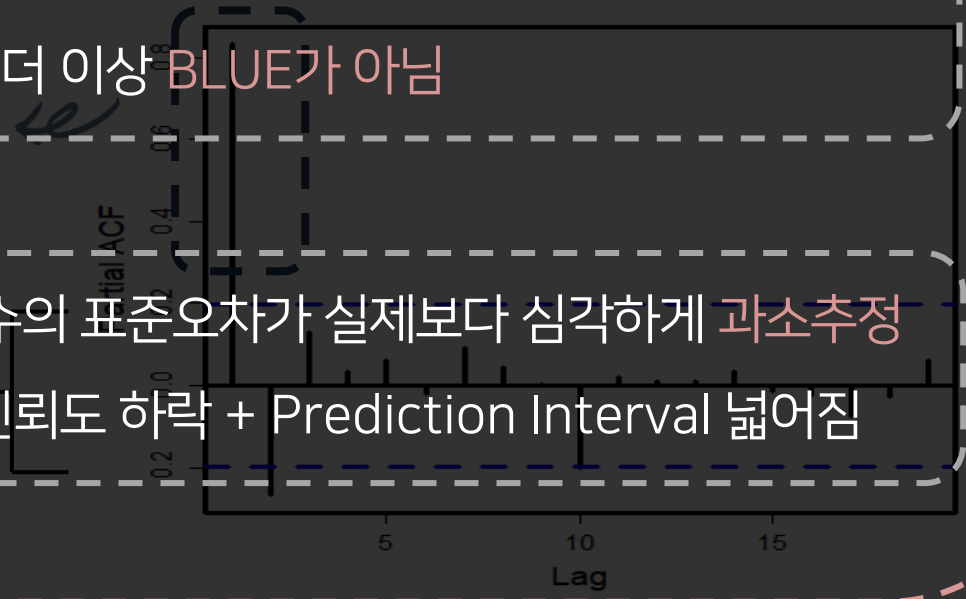
👉 최소제곱추정량이 더 이상 BLUE가 아님

신뢰구간을 벗어나는 점

: P차 자기 상관이 있다고 간주

② $\hat{\sigma}^2$의 추정량과 회귀계수의 표준오차가 실제보다 심각하게 과소추정

👉 유의성 검정 결과 신뢰도 하락 + Prediction Interval 넓어짐



6 독립성 진단과 처방

처방

가변수 만들기

뚜렷한 **계절성**이 있다고 판단되면, **가변수** 생성

분석 모델 변경

① 시간에 따라 자기상관을 가질 경우

➔ 자기상관을 고려하는 $AR(p)$ 같은 시계열 모델 사용

② 공간에 따라 자기상관을 가질 경우

➔ 공간의 인접도를 고려하는 공간회귀모델 사용

6 독립성 진단과 처방

처방

가변수 만들기

계절성이 주기를 가진다는 점을 이용

주기함수인 **삼각함수** $\cos(t)$, $\sin(t)$**의 선형결합**으로 주기 표현하는 방법

분석 모델 변경

① 시간에 따라 자기상관을 가질 경우

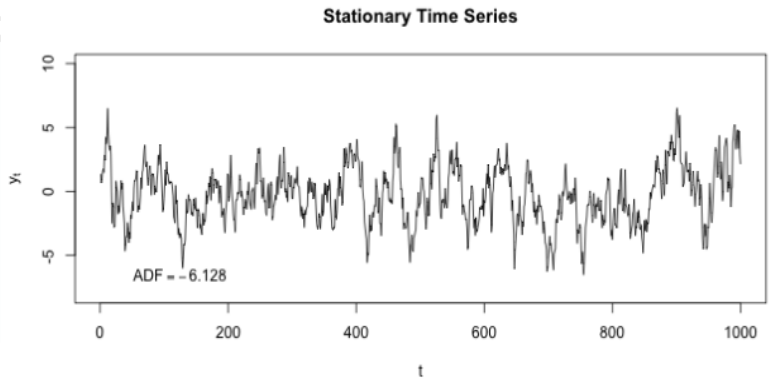
➔ 자기상관을 고려하는 $AR(p)$ 같은 시계열 모델 사용

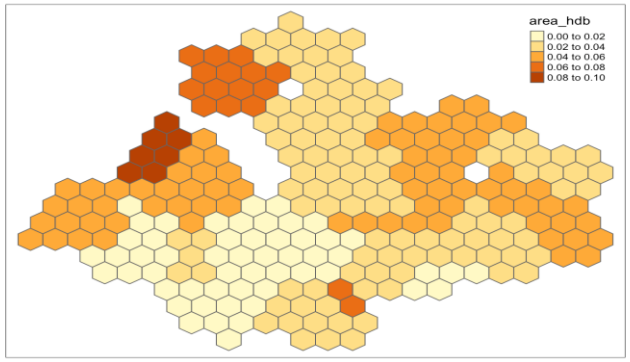
② 공간에 따라 자기상관을 가질 경우

➔ 공간의 인접도를 고려하는 공간회귀모델 사용

6 독립성 진단과 처방

처방





분석 모델 변경

① 시간에 따라 자기상관을 가질 경우

➔ 자기상관을 고려하는 $AR(p)$ 같은 시계열 모델 사용

② 공간에 따라 자기상관을 가질 경우

➔ 공간의 인접도를 고려하는 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        11.73816 0.019408 Assumptions NOT satisfied!
Skewness            2.37864 0.123004    Assumptions acceptable.
Kurtosis            0.02033 0.886622    Assumptions acceptable.
Link Function       8.57441 0.003409 Assumptions NOT satisfied!
Heteroscedasticity  0.76478 0.381838    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, 등분산성을 한 번에 체크해주는 함수

gvlma를 이용하면 간편하지만, 유의수준 0.05에서
경계를 잘라 버리다 보니 융통성이 부족함

선형회귀는 이런 가정 충족에 대해
비교적 robust하기 때문에 gvlma 결과만으로
비선형 모델을 선택하는 등의 판단은 위험할 수 있음

Global Stat : 선형성 / Skewness : 정규성 / Kurtosis : 정규성

Link Function : 선형성 / Heteroscedasticity : 등분산성

다음주 예고

다중공선성

변수선택법

정규화

감사합니다