회귀분석팀

6팀

김민우

채희지

김다민

성준혁

천예원

INDEX

1

회귀 기본 가정

1 회귀 기본 가정

회귀 기본 가정이 가지는 의미

회귀분석은 **적은 수의 관측치**만으로도

모델을 구성할 수 있고, **좋은 예측과 추정**이 가능하다는 장점이 있음

그만큼 **많은 제약들**이 **예측력과 설명력을 뒷받침**하고 있기 때문

⋮

선형회귀모델이 만족해야 하는 제약들이 곧 **선형회귀의 기본 가정**

1 회귀 기본 가정

회귀 기본 가정이 가지는 의미

선형회귀의 기본 가정

- 모델의 선형성
- 오차의 정규성
- 오차의 등분산성
- 오차의 독립성

1 회귀 기본 가정

회귀 기본 가정이 가지는 의미

선형회귀 기본 가정

머신러닝 모델들도 이러한 가정들을 필요로 하는데,

모델의 가정은 모델이 만들어진 형태와 직접적으로 연관되어 있으므로

지켜지지 않으면 **모델의 성능이 급락하는 경우**가 많음

모델의 선형성

오차의 정규성

오차의 등분산성

오차의 독립성

1 회귀 기본 가정

회귀 기본 가정이 가지는 의미

잔차의 평균이 최대한 0에 가까워지는 정확한 회귀모델을

만드는 것이 궁극적인 목표!

모델의 선형성

오차의 정규성

현실에서는 추정한 모델과 실제 데이터의 차이가 발생하는데,

① 모델링을 할 때 미처 고려하지 못한 속성

②현실 세계의 여러 오차(잡음)

중 무엇 때문인지 확인해야 함!

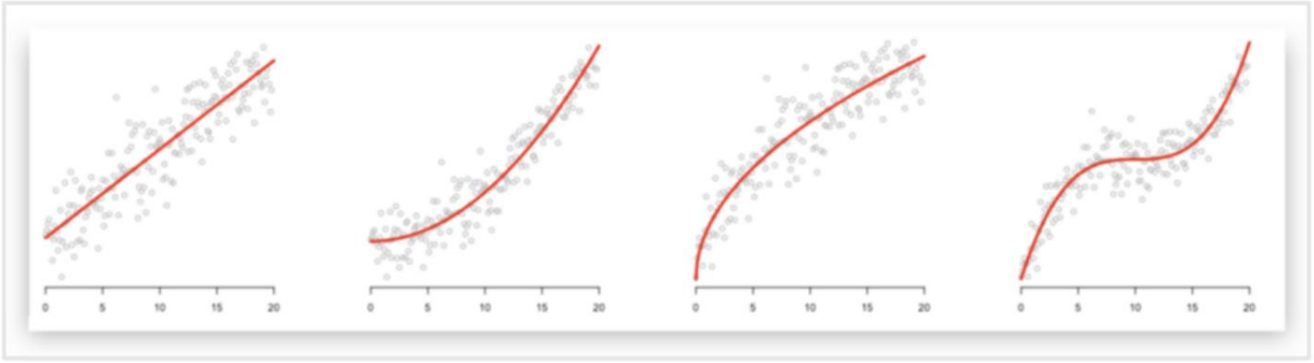
1 회귀 기본 가정

회귀분석의 기본 가정

1. 모델의 선형성 (Linearity)

설명변수와 반응변수의 관계는 **선형**이다

모델 자체가 선형성만 고려



치환의 과정을 통해 변화된 x를 새로운 x로 취급한다면,

위 결합들을 모두 **선형결합**으로 이해 가능

=선형성 만족!

1 회귀 기본 가정

회귀분석의 기본 가정

1. 모델의 선형성 (Linearity)

설명변수와 반응변수의 관계는 **선형**이다

모델 자체가 선형성만 고려

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

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

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

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

치환의 과정을 통해 변화된 x를 새로운 x로 취급한다면,

위 결합들을 모두 **선형결합**으로 이해 가능

=선형성 만족!

1 회귀 기본 가정

회귀분석의 기본 가정

1. 모델의 선형성 (Linearity)

설명변수와 반응변수의 관계는 **선형**이다

모델 자체가 선형성만 고려

$y = \frac{\beta_1 x}{\beta_0 + x}$ 와 같은 형태라면,

변환을 통해 선형으로 만들 수 없으므로 **비선형 모델!**

치환의 과정을 통해 변화된 x를 새로운 x로 취급한다면,

위 결합들을 모두 **선형결합**으로 이해 가능

=선형성 만족!

1 회귀 기본 가정

회귀분석의 기본 가정

2. 오차의 정규성 (Normality)

오차항은 **정규분포**를 따른다

오차의 평균은 0이다! (거의 위배되지 않음)

⋮

정규분포가 오차에 대한 확률분포이므로,

회귀식이 데이터를 잘 표현하고 있다면

오차들은 **단순 잡음(noise)**가 되어 **정규분포에 근사한 형태**가 나올 것!

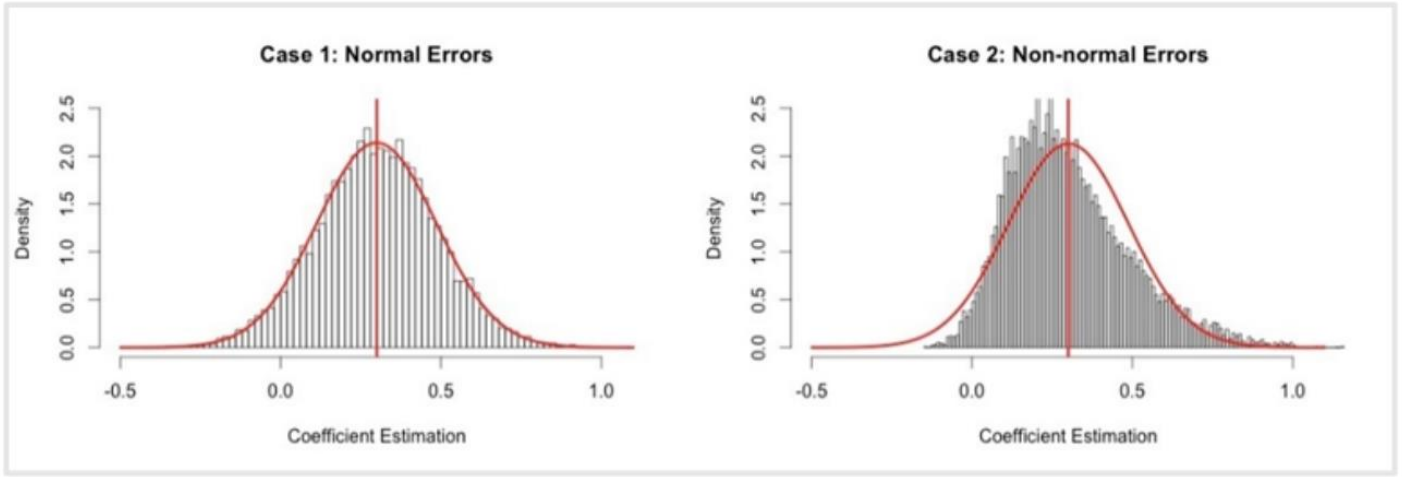
1 회귀 기본 가정

회귀분석의 기본 가정

2. 오차의 정규성 (Normality)

오차항은 **정규분포**를 따른다

오차의 평균은 0이다! (거의 위배되지 않음)



오차항의 정규성을 가정하기 때문에

회귀식과 개별 회귀 계수에 대한 **검정 시행 가능!**

1 회귀 기본 가정

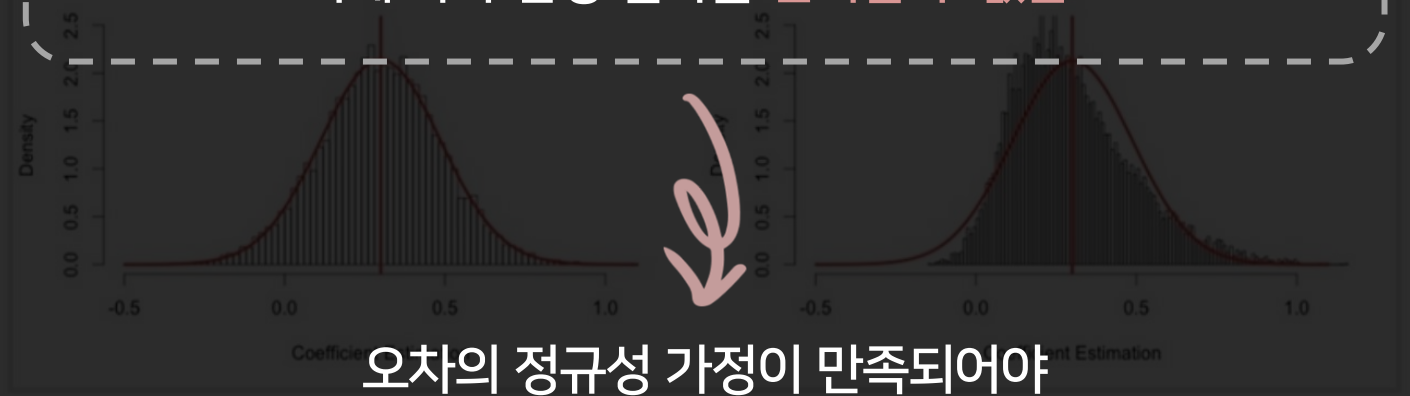
회귀분석의 기본 가정

2. 오차의 정규성 (Normality)

정규분포를 따르지 **않을** 경우

가설 검정에서 분포가 **왜곡**될 것이고,

이에 따라 검정 결과를 **신뢰할 수 없음**



오차의 정규성 가정이 만족되어야

회귀 모형의 해석이 유의미해짐

회귀식과 개별 회귀 계수에 대한 검정 시행 가능!

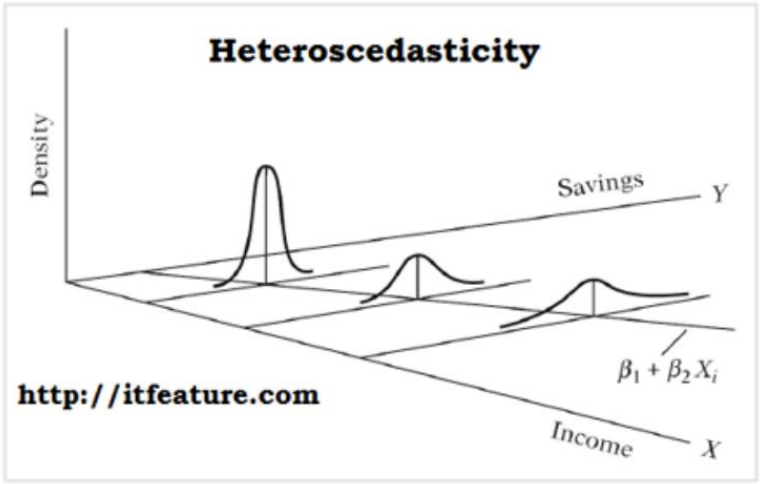
회귀분석의 기본 가정

3. 오차의 등분산성 (Homoscedasticity / Constant Variance)

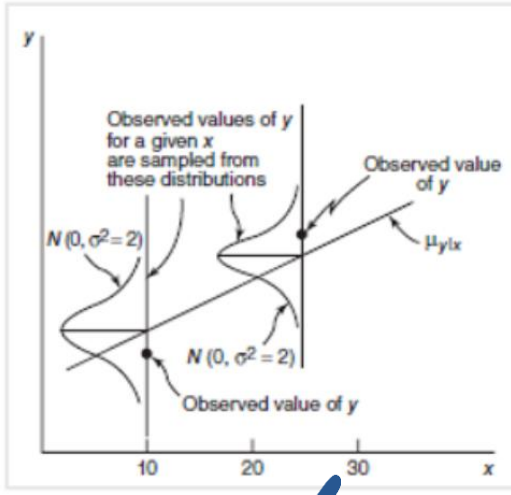
오차항의 분산은 **상수**다

분산은 σ^2으로 동일

↔ 이분산성 (Heteroscedasticity)



▲ 이분산성

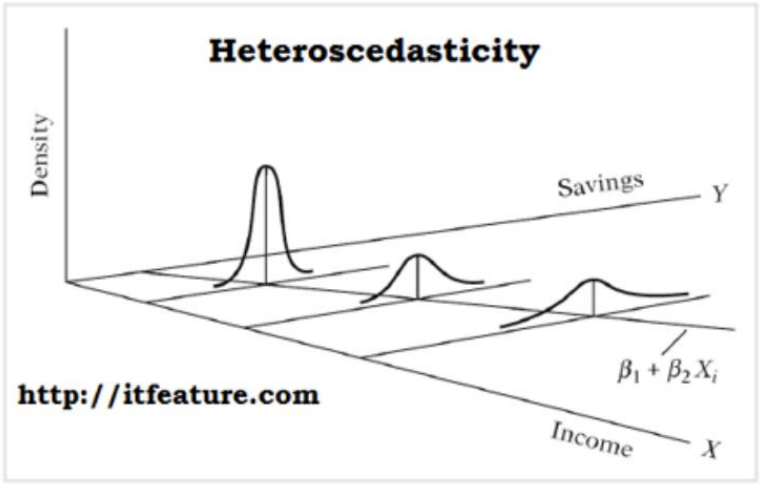


▲ 등분산성

회귀분석의 기본 가정

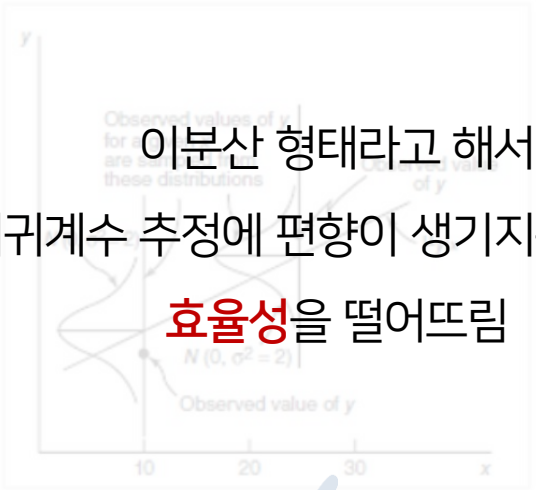
3. 오차의 등분산성 (Homoscedasticity / Constant Variance)

오차항의 분산은 **상수**다



▲ 이분산성

이분산 형태라고 해서
회귀계수 추정에 편향이 생기지는 않지만
효율성을 떨어뜨림



▲ 등분산성

1 회귀 기본 가정

회귀분석의 기본 가정

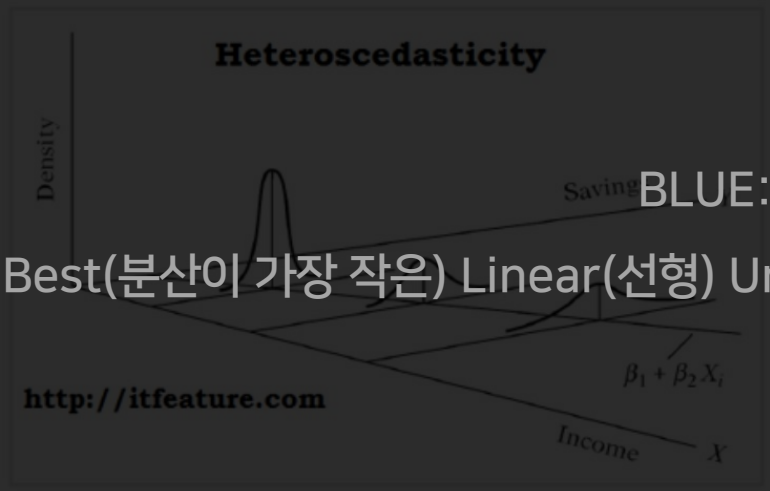
3. 오차의 등분산성 (Homoscedasticity / Constant Variance)

오차항의 분산은 상수다

분산이 일정하지 않고 **변화**한다면

최소제곱추정량이 BLUE의 조건을 만족하지 않아

최소 분산이 갖는 **효율성**을 지니지 못함



▲ 이분산성

BLUE:

Best(분산이 가장 작은) Linear(선형) Unbiased Estimator(불편추정량)

이분산 형태라고 해서

회귀계수 추정에 편향이 생기지는 않지만

효율성을 떨어뜨림

▲ 등분산성

자세한 내용은 1주차 클린업 참고!

1 회귀 기본 가정

회귀분석의 기본 가정

3. 오차의 등분산성 (Homoscedasticity / Constant Variance)

오차항의 분산은 **상수**다

분산은 σ^2으로 동일

↔ 이분산성 (Heteroscedasticity)

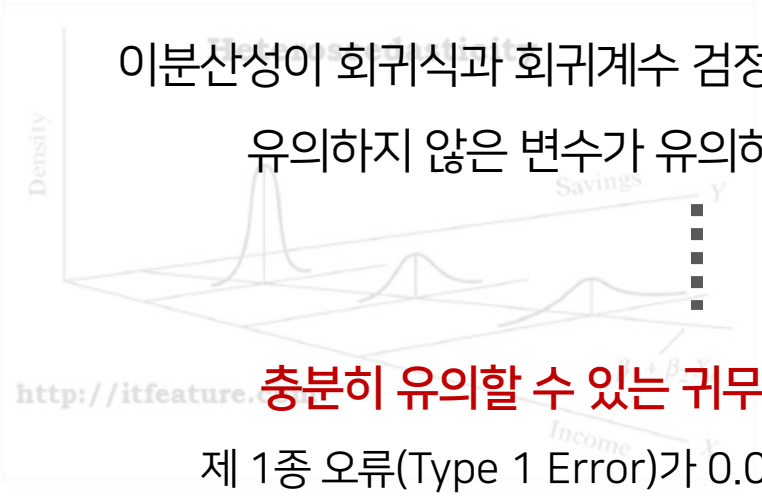
이분산성이 회귀식과 회귀계수 검정에 대한 신뢰도를 떨어뜨려,

유의하지 않은 변수가 유의하다고 나타날 수 있음

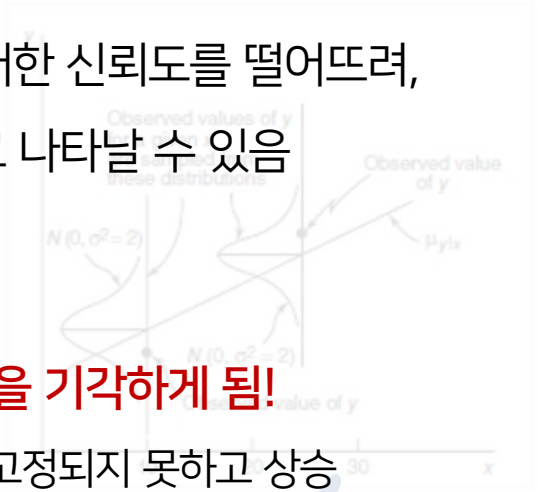
⋮

충분히 유의할 수 있는 귀무가설을 기각하게 됨!

제 1종 오류(Type 1 Error)가 0.05로 고정되지 못하고 상승



▲ 이분산성



▲ 등분산성

1 회귀 기본 가정

회귀분석의 기본 가정

4. 오차의 독립성 (Independence / No autocorrelation)

오차항은 **서로 독립**이다

↔ 자기상관성 (Autocorrelation)

오차의 독립성을 만족하지 않는다면,

최소제곱추정량이 **더 이상 BLUE가 아님**

1 회귀 기본 가정

회귀분석의 기본 가정

4. 오차의 독립성 (Independence / No autocorrelation)

오차항은 **서로 독립**이다

σ^2의 추정량과 회귀계수의 표준오차가

실제보다 심각하게 과소추정됨

오차의 독립성을 만족하지 않는다면,

최소제곱추정량이 **더 이상 BLUE가 아님**

유의성 검정의 결과를 **신뢰할 수 없음**

1 회귀 기본 가정

회귀분석의 기본 가정

4. 오차의 독립성 (Independence / No autocorrelation)

오차항은 **독립**이다

↔ 자기상관성 (Autocorrelation)

모델의 선형성과 **오차의 정규성, 등분산성, 독립성**에

초점을 맞춰 진단 및 처방 과정 진행!

오차의 독립성을 만족하지 않는다면,

최소제곱추정량이 **더 이상 BLUE가 아님**

2

잔차 플랏

2 잔차 플랏

회귀분석의 기본 가정 진단

회귀분석의 4가지 가정을 진단하려면?

① 가설 검정을 이용한 방법

② 시각적 (Graphical) 방법

판단에 대한 명확한 근거를 마련하기 위해 가설 검정의 과정은 필요

2 잔차 플랏

잔차 플랏

잔차 플랏

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

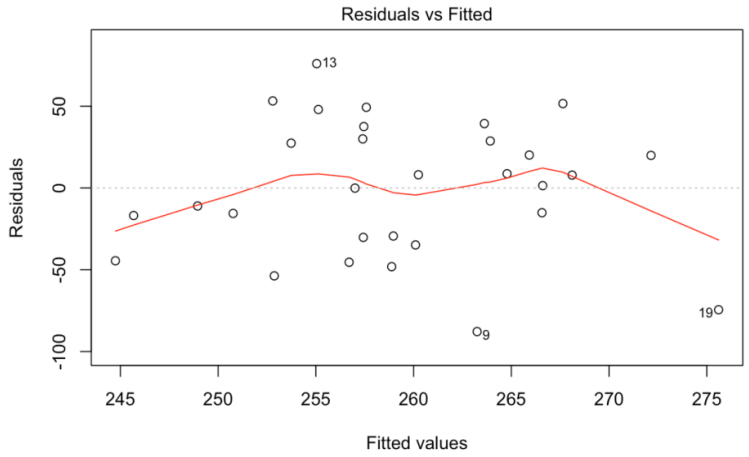
Residuals vs Fitted	Normal Q-Q (Quantile-Quantile)
Scale - Location	Residuals vs. Leverage

2 잔차 플랏

Residuals vs Fitted

① Residuals vs Fitted

독립변수와 종속변수 간의 **선형성**과 오차의 **등분산성**을 확인 가능



▲ x축 : 예측값($\hat{y}$)

▲ y축 : 잔차 ($e = y - \hat{y}$)

빨간 실선은 전체적인 잔차들의 추세선

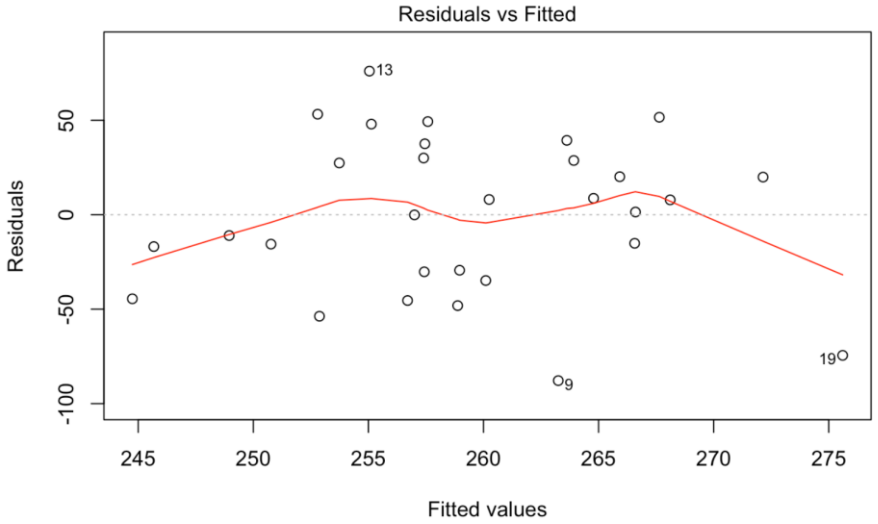
⋮

잔차들의 분포를 Local Regression으로 추정한 완만한 연결 직선이며,

잔차 분포의 패턴 및 경향성 표현

2 잔차 플랏

Residuals vs Fitted



선형성

빨간 실선이 X축에 평행한 직선이 아니라면 **선형성**이 위배된 것

등분산성

잔차와 예측값 사이에 무작위한 형태 이외의 어떠한 관계를 보이면 **등분산성**이 위배된 것

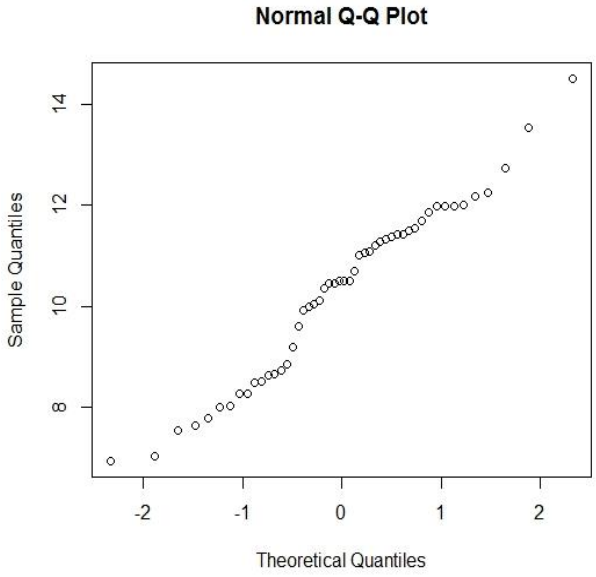
2 잔차 플랏

Normal Q-Q

② Normal Q-Q

오차의 정규성을 확인 가능

두 개 변수의 분포를 비교하기 위한 **비모수적 방법이자 시각적 방법**



▲ x축 : 정규분포의 분위수 값,

▲ y축 : 표준화 잔차

그래프가 $y = x$에 가까울수록

잔차가 정규성을 만족

⋮

잔차가 **정규분포 사분위수** 위에

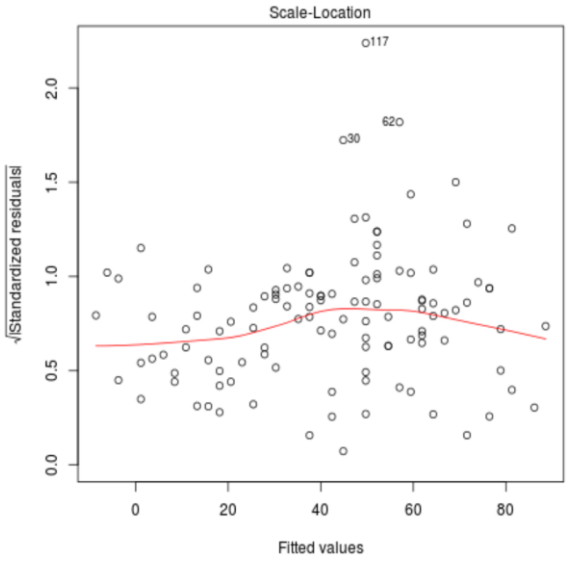
그대로 위치한다는 의미이므로!

2 잔차 플랏

Scale-Location

③ Scale-Location

오차의 등분산성을 확인 가능



▲ x축 : 예측값($\hat{y}$),

▲ y축 : 표준화잔차($\sqrt{|e_i|/se(e_i)}$)

빨간 실선은 전체적인 **잔차들의 추세선**

⋮

잔차들의 분포를 Local Regression으로

추정한 완만한 연결 직선이며,

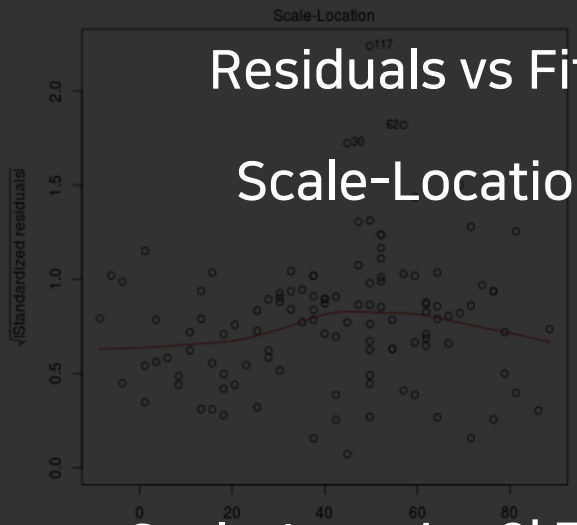
잔차 분포의 경향성 표현

2 잔차 플랏

Scale-Location

③ Scale-Location

Residuals vs. Fitted 와
오차의 등분산성을 확인 가능
Scale-Location의 차이점?



Residuals vs Fitted의 y축 : $e = y - \hat{y}$

빨간 실선은 전체적인 잔차들의 추세선

Scale-Location의 y축 : $\sqrt{|e_i|/se(e_i)}$

잔차들의 분포를 Local Regression으로
추정한 완만한 연결 직선이며,
표현

Scale-Location의 잔차는 절댓값이 씌워진 형태!

▲ x축 : 예측값($\hat{y}$),

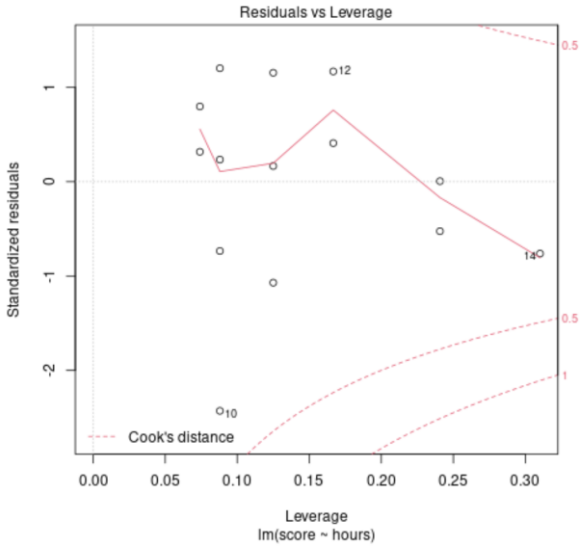
▲ y축 : 표준화잔차($\sqrt{|e_i|/se(e_i)}$)

2 잔차 플랏

Residuals vs Leverage

④ Residuals vs Leverage

영향점 (Influential Point) 확인 가능



▲ x축 : Leverage (지렛값),

▲ y축 : 표준화잔차($\sqrt{|e_i|/se(e_i)}$)

플랏의 오른쪽에 위치한 점들은

leverage가 큰 잔차

빨간 실선으로부터 위아래로 멀리

떨어진 점들은 **outlier**

빨간 점선은 cook's distance를 나타냄

0.5보다 크면 영향점 후보!

3

선형성 진단과 처방

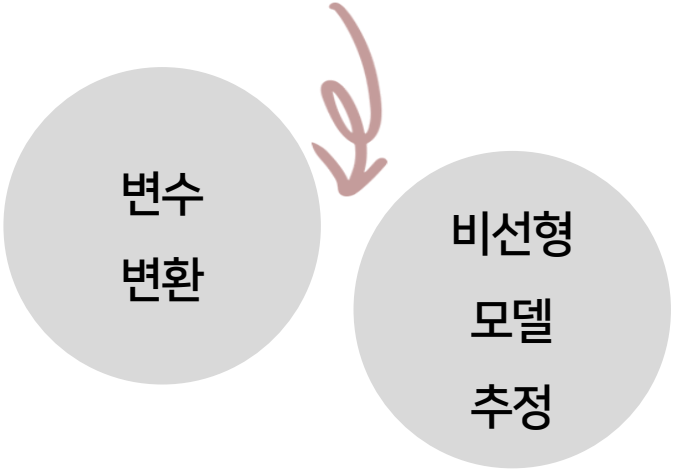
3 선형성 진단과 처방

선형성 가정이란?

선형성 가정

반응 변수가 설명 변수의 **선형결합**으로 이루어졌다는 가정

선형성 위배



을 통해 해결!

3 선형성 진단과 처방

선형성 가정이 위배될 경우 발생하는 문제점

선형성 가정이 위배되었다고 판단되는 경우,

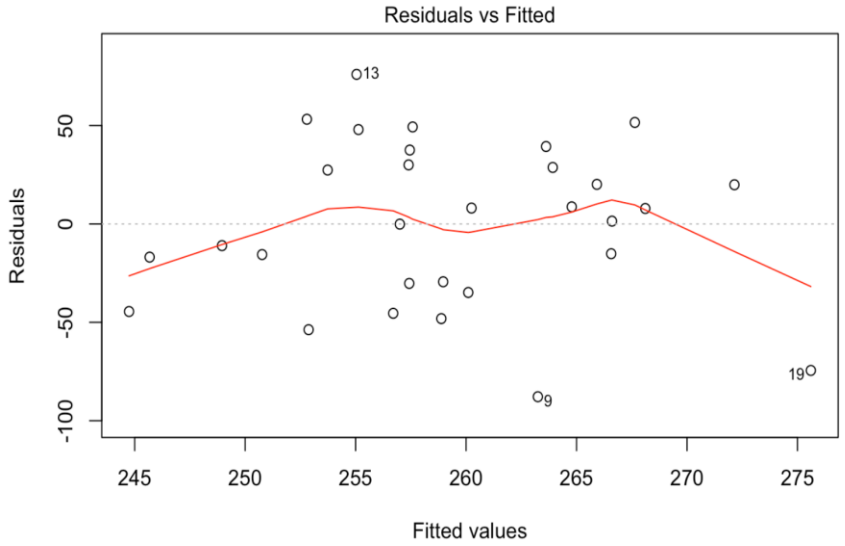
모든 **선형회귀모델은 모델 자체가 성립하지 않음**!

⋮

대부분 실제 모델보다 **과소추정**되어 예측 성능이 떨어지게 됨

3 선형성 진단과 처방

진단 | ① Residuals vs Fitted Plot



▲ x축 : 예측값($\hat{y}$)

▲ y축 : 잔차($e = y - \hat{y}$)

추세선이 평균 0을 중심으로 하는 **x축에 평행한 상태가 아니라면 선형성이 위반**된 것!

선형성이 위배되는 경우, 일반적으로 추세선이 이차함수 혹은 삼차함수 꼴

진단 | ② Partial Residual Plot

개별 독립 변수와 종속 변수 간의 **선형성**을 확인하기 좋은 플랏

⋮

Residual vs Fitted Plot을 이용하면

어떤 변수의 영향으로

선형성이 위배되었는지 알기 **어려움**

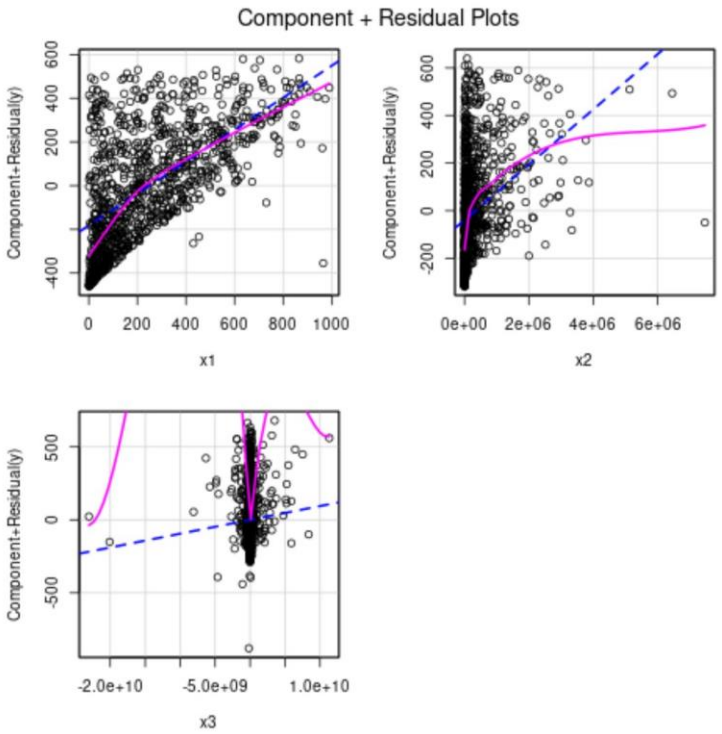
⋮

Partial Residual Plot은

개별 변수의 영향을 확인할 수 있음

3 선형성 진단과 처방

진단 | ② Partial Residual Plot



파란 점선

Partial Residual과 x_i가 적합된 직선.

OLS를 통해 점들의 분포를 추정한 회귀선

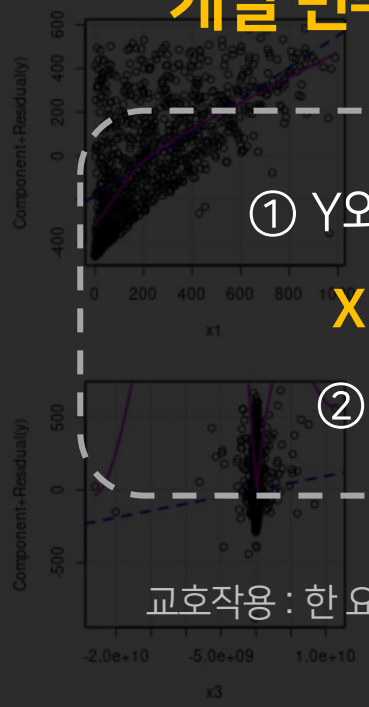
분홍 실선

점들의 분포를 Local Regression을 통해 추정한 잔차의 추세선

일반적으로 **두 선이 일치**하면 선형성이 만족되었다고 판단

3 선형성 진단과 처방

진단 | ② Partial Residual Plot



개별 변수의 선형성을 판단하기에는 좋은 방법이지만,

① Y와 개별 X 변수들 간의 단편적인 관계만 보여주므로,

X 변수 사이의 교호작용과 상관관계 파악 불가능

② 심각한 **다중공선성**이 있는 경우 **왜곡 가능성** ↑

교호작용 : 한 요인의 효과가 다른 요인의 수준에 의존하는 경우로 변수간의 시너지 효과를 의미

일반적으로 **두 선이 일치**하면 선형성이 만족되었다고 판단

3 선형성 진단과 처방

처방 | ① 변수변환

변수 변환

변수의 변환을 통해 비선형 관계를 해결할 수 있음

치환의 과정을 통해 x를 변화시켜 이를 새로운 x로 취급하면 선형결합 만족

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

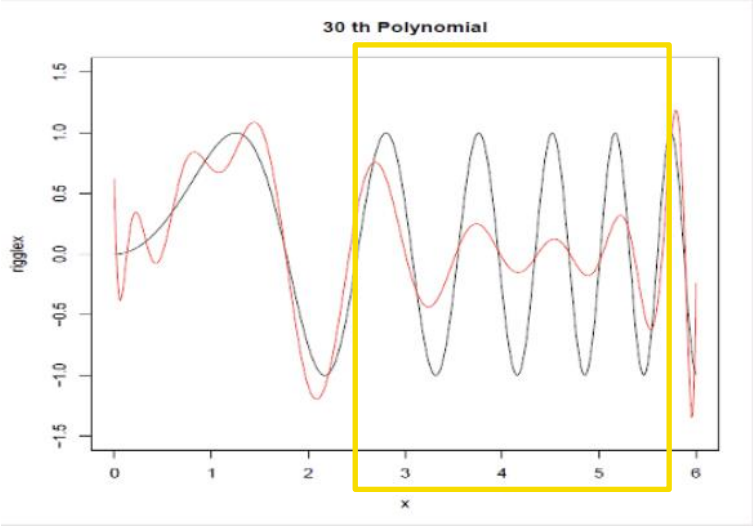
변수 변환을 통해 선형성을 확보할 수 있는 모델도,

넓은 의미에서 선형모델이라 부름

처방 | ② 비선형 회귀

Polynomial Regression

고차항을 고려하는 다항회귀



✓

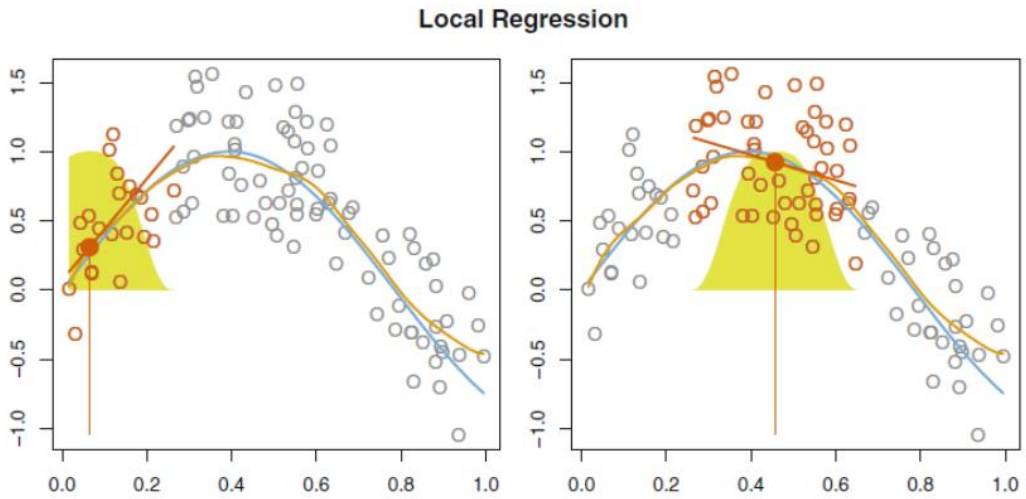
Partial Residual Plot에서
2차 이상의 곡선 형태가 나타나는 경우,
특정 변수에 대해서만 고차항 사용

그러나 초고차항을 추가하더라도 경향을 잡아내기가 힘들어 일반적으로는 3차까지만을 고려함

처방 | ② 비선형 회귀

Local Regression

지역적인(local) 데이터들로 회귀 모델링을 하는 방법.



Target data x_0을 중심으로 그 주변의 **k개의 이웃 데이터**

$x_0 \in N(x_0)$ 들만을 사용하여 **부분적으로 회귀** 모델을 구성

3 선형성 진단과 처방

처방 | ② 비선형 회귀

Local Regression

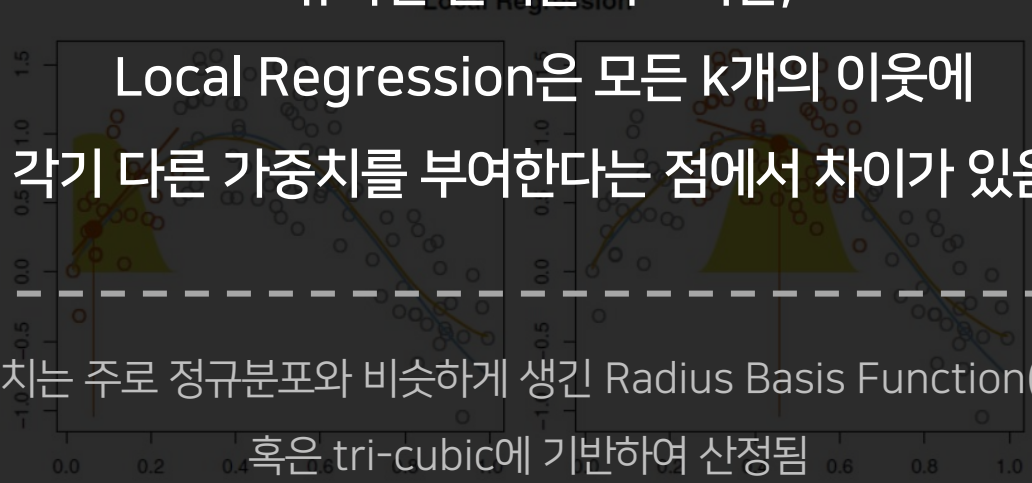
Local Regression vs. KNN 알고리즘

지역적인(local) 데이터들로 회귀 모델링을 하는 방법

유사한 논리를 따르지만,

Local Regression은 모든 k개의 이웃에

각기 다른 가중치를 부여한다는 점에서 차이가 있음!



가중치는 주로 정규분포와 비슷하게 생긴 Radius Basis Function(RBF)

혹은 tri-cubic에 기반하여 산정됨

Target data x_0을 중심으로 그 주변의 **k개의 이웃 데이터**

$x_0 \in N(x_0)$ 들만을 사용하여 **부분적으로 회귀** 모델을 구성

4

정규성 진단과 처방

4 정규성 진단과 처방

정규성 가정이란?

정규성 가정

모델 적합시 나타나는 **오차가 정규분포**를 따를 것이라는 가정

회귀식이 데이터를 잘 표현한다면?

① 잔차들은 단순한 측정 오차(noise)로 간주됨

② 잔차들의 분포는 정규분포와 흡사한 형태가 됨

4 정규성 진단과 처방

정규성 가정이 위배되었을 때 발생하는 문제

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

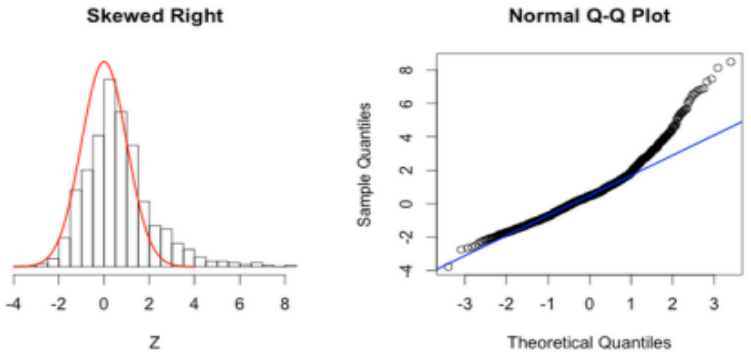
정규성 가정이 위배된다면?

가설 검정 결과가 p-value에 의해 유의하게 나오더라도,

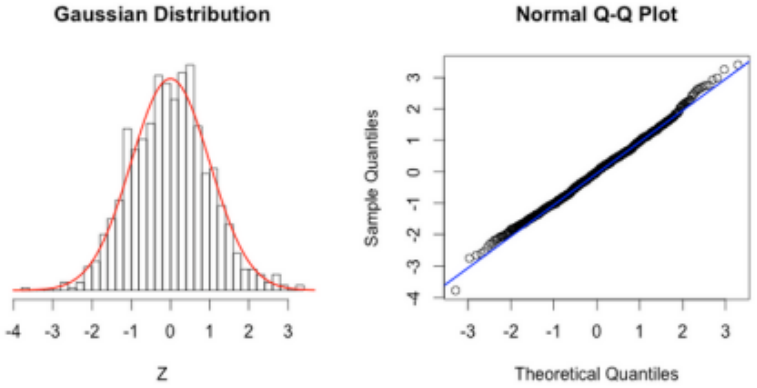
검정 결과와 **예측 결과**를 신뢰할 수 없음

4 정규성 진단과 처방

진단 | ① Normal Q-Q Plot을 통한 진단



▲ 정규성을 만족하지 못함



▲ 정규성을 만족함

Normal Q-Q Plot이 $y = x$에 가까우면 **정규성**을 만족

진단 | ② 통계적 검정

Plot으로 확인하는 경우 판단이 주관적일 수 있으므로,

Plot이 명확한 경우가 아니라면 **통계적 방법에 의한 가설검정**으로 확인해야 함

가설

H_0 : 주어진 데이터는 정규분포를 따른다.

H_1 : 주어진 데이터는 정규분포를 따르지 않는다.

⋮

귀무가설을 기각하지 못해 정규성을 만족하는(정규분포를 따르는) 상황을 원함

4 정규성 진단과 처방

진단 | ② 통계적 검정 : Empirical CDF를 이용한 검정

Empirical CDF

관측치들을 작은 순서대로 나열한 후 누적 분포 함수를 그린 것

잔차의 Empirical CDF와 정규분포의 CDF를 비교하여 검정

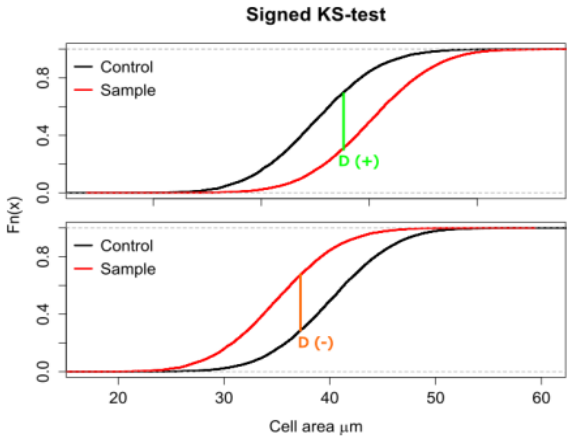
⋮

Kolmogorov Smirnov Test	Anderson Darling Test

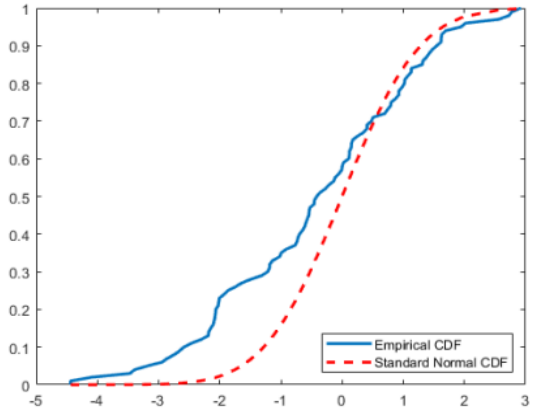
진단 | ② 통계적 검정 : Empirical CDF를 이용한 검정

1. Kolmogorov-Smirnov Test (K-S Test)

하나의 모집단이 어떤 특정한 분포함수를 갖는지 알아보는 방법



▲ 이론분포함수 ▲ 표본분포함수



▲ 이론분포함수 ▲ 표본분포함수

귀무가설 하에서 표본분포함수가 어떤 이론적 분포함수와 유사한지를 검정

이론분포함수와 표본분포함수의 차(D)가 크면 H_0을 기각

진단 | ② 통계적 검정 : Empirical CDF를 이용한 검정

2. Anderson-Darling Test (A-D Test)

K-S 검정을 수정한 방식으로, **특정 분포의 꼬리(tail)에 K-S보다 더 가중치**를 둠

꼬리 부분에 해당하는 데이터에
대해 민감하게 반응

정규성 검정을 위해 A-D 검정은 0.05로, K-S 검정은 0.15로 유의수준 설정

⋮

A-D 검정이 K-S 검정에 비해서 **더 엄격함**

4 정규성 진단과 처방

진단 | ② 통계적 검정 : 정규분포의 분포적 특성을 이용한 검정

1. Shapiro Wilk Test

표본이 정규분포로부터 추출된 것인지를 확인하기 위한 검정 방법

- ✓ 정규분포 **분위수 값**과 **표준화 잔차** 사이의 **선형관계** 확인 (Q-Q Plot 아이디어와 동일)
- ✓ **관측치가 5000개 이하**인 데이터에서만 가능한 검정 방법

귀무가설 H_0 를 기각하지 못했다는 것은 정규분포를 따르지 않는다고
말할 근거가 부족한 것일 뿐, **100% 정규성을 만족된다는 의미는 아님**

진단 | ② 통계적 검정 : 정규분포의 분포적 특성을 이용한 검정

2. Jarque-Bera Test

정규분포의 왜도가 0, 첨도가 3이라는 사실에서 기인한 검정 방법

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

n : 데이터의 개수

⋮

잔차의 분포가 **정규분포와 달라질수록 왜도/첨도에 변화**가 생기고

통계량 값이 커져 유의수준을 넘어서면 **귀무가설을 기각**하게 됨

진단 | ② 통계적 검정 : 정규분포의 분포적 특성을 이용한 검정

2. Jarque-Bera Test

정규분포의 왜도가 0, 첨도가 3이라는 사실에서 기인한 검정 방법

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

이상치에 민감한 왜도(skewness)를 이용하는 만큼,

이상치를 삭제했을 때 정규분포임이 드러나는 경우가 많음

n : 데이터의 개수

잔차의 분포가 **정규분포와 달라질수록 왜도/첨도에 변화**가 생기고

통계량 값이 커져 유의수준을 넘어서면 **귀무가설을 기각**하게 됨

진단 | ② 통계적 검정 : 정규분포의 분포적 특성을 이용한 검정

2. Jarque-Bera Test

정규분포의 왜도가 0, 첨도가 3이라는 사실에서 기인한 검정 방법

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

이상치에 민감한 왜도(skewness)를 이용하는 만큼,

이상치를 삭제했을 때 정규분포임이 드러나는 경우가 많음

⋮

정규성이 위반된 것을 확인했을 때,

어떻게 처방할 수 있을까?

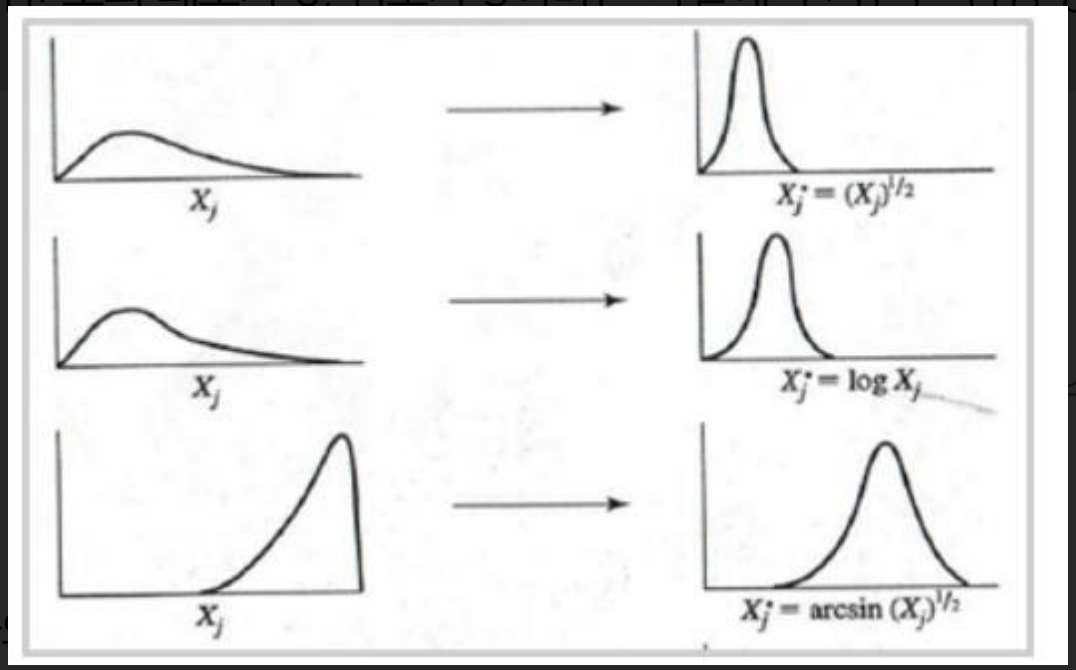
4 정규성 진단과 처방

진단 | ② 통계적 검정 : 정규분포의 분[illegible]특성을 이용한 검정

2. Jarqu[illegible]

변수변환을 통해 정규성을 처방해줄 수 있음!

정규분포의 왜도가 0, 첨도가 3이라는 사실에서 기인한 검정 방법



수

잔차[illegible]고

통계량 값이 커져 유의수준을 넘어서면 **귀무가설을 기각**하게 됨

4 정규성 진단과 처방

진단 | ② 통계적 검정 : 정규분포의 분포 특성을 이용한 검정

2. Ja

정규분포의 왜도 기인한 검정 방법

단순 변수 변환은 **주관적인** 판단 하에 이루어지므로

객관성을 확보하기 힘듦

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

1) 1) Box-cox Transformation

2) 2) Yeo-Johnson Transformation

잔차의 분포가 **정규분포와 달라질수록 왜도/첨도에 변화**가 생기고

통계량 값이 커져 유의수준을 넘어서면 **귀무가설을 기각**하게 됨

처방 | ① Box-cox Transformation

Box-cox Transformation

통계적인 검정에 따라 변수 변환(비선형 변환)을 진행해주는 방법

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

λ 를 **변화**시키면서

y 가 **정규성, 등분산성**을 만족하도록

일반적으로 λ는 -5에서 5 사이의 값을 사용

변환된 y가 모든 실수 λ에 대해 연속임

처방 | ① Box-cox Transformation

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

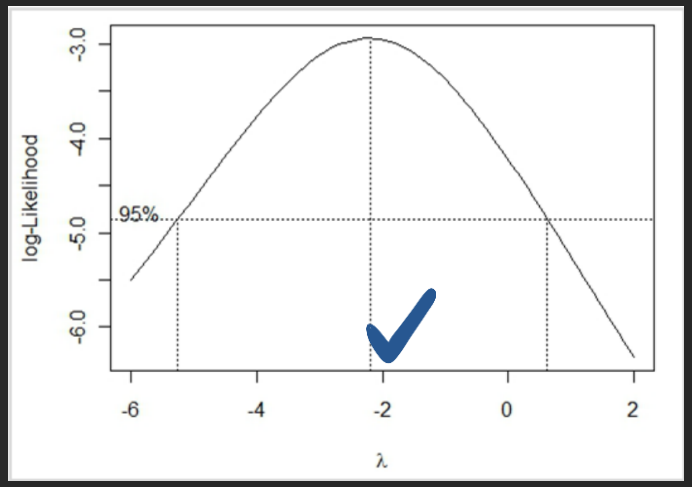
⋮

이 때 최적의 λ는 최대우도함수(ML)을 통해 신뢰구간을 구한 후

신뢰구간 내의 **로그우도함수를 최대화**하는 λ를 **최적의 값**으로 택함

4 정규성 진단과 처방

처방 | ① Box-cox Transformation



이 때 최적의 λ는 최대우도함수(ML)을 통해 신뢰구간을 구한 후
신뢰구간 내의 로그우도함수를 최대화하는 λ를 최적의 값으로 택함

95% 내의 λ값 중 **가능도함수가 최대**가
되게끔 하는 -2 근방의 λ를 선택

λ를 -2로 선택하는 것도 좋음!

λ를 **정수**로 택하면 **변수 변환 관계 파악**이 쉽기 때문

4 정규성 진단과 처방

처방 | ① Box-cox Transformation

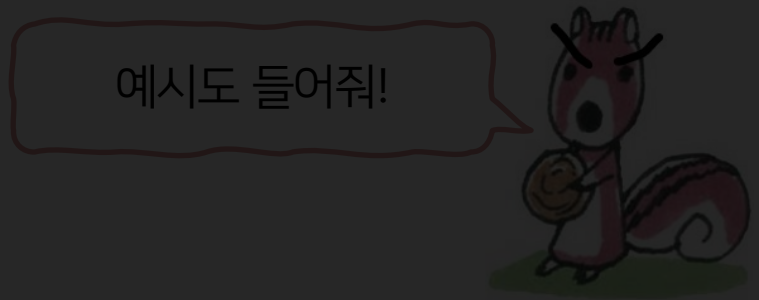
$$y(\lambda) \begin{cases} \frac{(y^{\lambda}-1)}{\lambda} & if\ \lambda \neq 0 \\ \log(y) & if\ \lambda = 0 \end{cases}$$

하지만 Box-cox Transformation은

y가 0 이하일 경우 사용할 수 없음

이 때 최적의 λ는 λ=0이 되면 y가 log(y)로 변환될 수 있기 때문

신뢰구간 내의 **로그우도함수를 최대화**하는 λ를 **최적의 값**으로 택함



처방 | ② Yeo-Johnson Transformation

Yeo-Johnson Transformation

Box-cox Transformation과 같은 아이디어

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

단, Box-cox와 달리 변수 범위에 대한 제약이 없음

4 정규성 진단과 처방

처방 | ② Yeo-Johnson Transformation

Yeo-Johnson Transformation이

전체 범위에서 사용 가능한데

Box-cox Transformation과 같은 아이디어

Box-cox Transformation도 사용하는 이유?

Box-cox Transformation은 해석에 장점이 있음

Yeo-Johnson Transformation처럼

제곱이 다르게 부여되면 전체 범위에 대한 해석이 모호해질 수 있음

단, Box-cox의 독립변수 범위에 대한 제약이 없음

Yeo-Johnson Transformation은 제곱이 λ, 2-λ로 다름

5

등분산성 진단과 처방

5 등분산성 진단과 처방

등분산성 가정이란?

등분산성 가정

오차의 **모든 분산**은 **동일**해야 한다는 가정!

가정이 충족되었다는 건,

분산이 상수여서 어느 관측치에서나 동일하게 나타나고,

다른 변수의 **영향을 받지 않음**을 의미

진단 방법

(1) 잔차 플랏(Residual Plot)

(2) BP(Breusch-Pagan) test

5 등분산성 진단과 처방

등분산성 가정이 위배되었을 때 발생하는 문제

① OLS 추정량의 분산이 **실제 분산보다 작게** 추정됨

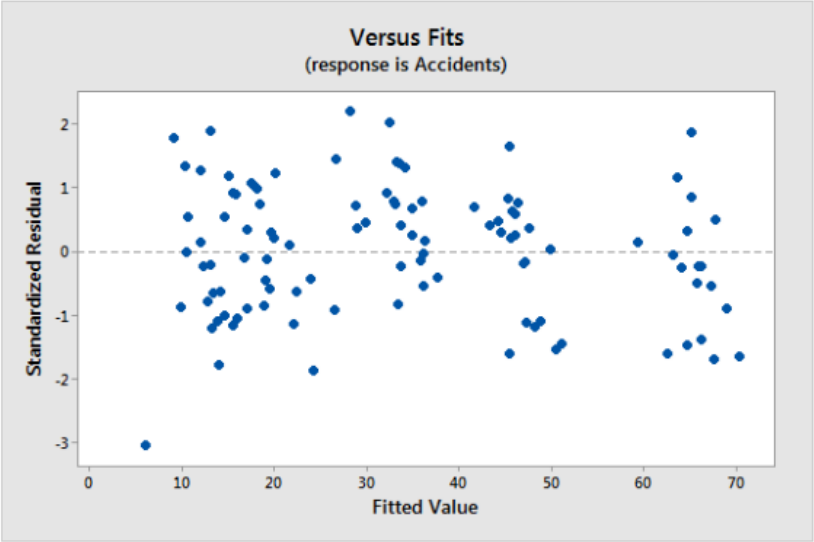
▶ 검정 통계량 증가 & P-value 감소

▶ 유의하지 않은 회귀계수를 유의하다고 판단(**제 1종 오류**)

② 조건을 만족하지 않으므로, OLS 추정량이 BLUE가 되지 못함

5 등분산성 진단과 처방

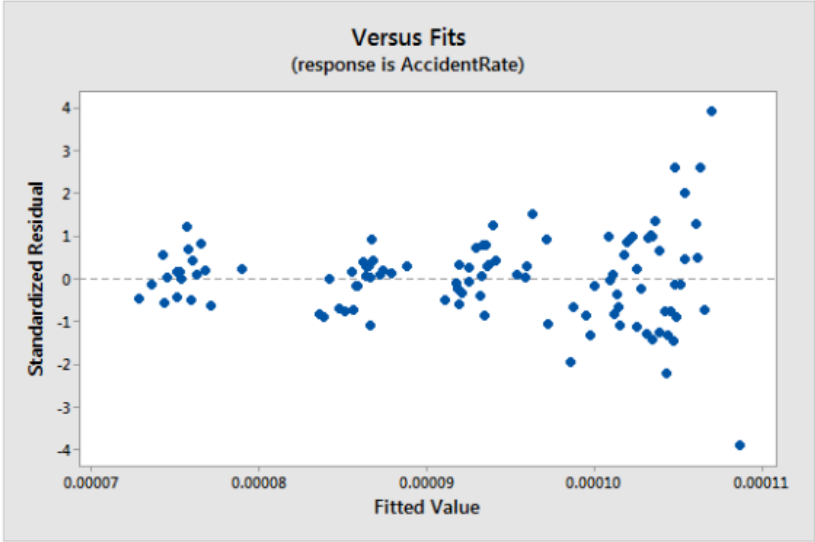
진단 | ① 잔차 플랏



▲ 등분산성을 **만족**한 경우

$\hat{y}$값에 상관없이

잔차 퍼짐의 정도가 **일정**함



▲ 등분산성을 **불만족**한 경우

$\hat{y}$값이 커지면서

잔차 퍼짐의 정도가 **일정하지 않음**

5 등분산성 진단과 처방

진단 | ① 잔차 플랏

Plot을 보아도 육안으로 **명확히 판단하기 어려운 경우**,

통계적 검정을 통해 등분산성을 확인할 수 있음!

▲ 등분산성을 **만족**한 경우

$\hat{y}$값에 상관없이

잔차 퍼짐의 정도가 **일정**함

▲ 등분산성을 **불만족**한 경우

$\hat{y}$값이 커지면서

잔차 퍼짐의 정도가 **일정하지 않음**

5 등분산성 진단과 처방

진단 | ② BP Test

BP(Breusch-Pagan) Test

잔차가 **독립변수**들의 **선형결합**으로 표현되는지 검정

설명변수의 증감에 따른 **오차의 분산 변화**를 통해

등분산성 지니는지 판단 가능

BP Test 기본 가정

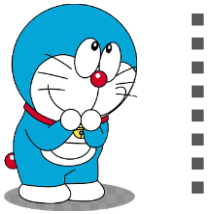
- ✓ 샘플 수가 많아야 함
- ✓ 오차항은 독립이고 정규분포를 따름
- ✓ 오차의 분산은 설명변수와 연관이 있음

진단 | ② BP Test

가설 설정

H_0 : 주어진 데이터는 등분산성을 지닌다.

H_1 : 주어진 데이터는 등분산성을 지니지 않는다 (이분산이다).



우리가 원하는 것은

귀무 가설을 **기각하지 못하는 것**!

5 등분산성 진단과 처방

진단 | ② BP Test

$$e^2 = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_p x_p + \epsilon'$$

잔차를 종속변수로 한 회귀 모형에서 결정계수(R^2)를 구함

오차가 독립변수에 의해 충분히 표현된다면

결정계수와 검정 통계량이 커질 것!

⋮

결정계수를 통해 잔차의 제곱이

독립변수의 **선형결합**으로 표현되는지와 그 때의 **설명력**을 파악

5 등분산성 진단과 처방

진단 | ② BP Test

검정 통계량

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

임계값

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

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

즉, 등분산성을 **만족하지 않음**을 의미

5 등분산성 진단과 처방

진단 | ② BP Test

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

임계값 $\chi^2_{P-1,\alpha}$

BP Test의 단점

- **비선형결합**으로 이루어진 이분산성은 파악할 수 없음
- 샘플이 **대표본**이어야 사용할 수 있음
- 오차의 정규성에 민감하게 반응하므로 **정규성**이 지켜진 상태인지 확인해야 함

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

즉, 등분산성을 **만족하지 않음**을 의미

5 등분산성 진단과 처방

처방 | ① 변수 변환(Box-cox Transformation)

Box-cox Transformation

정규성 만족을 위한 처방과 똑같이 **변수변환 방법**을 적용할 수 있음

가중 회귀 제곱 (WLS: Weighted Least Square)

등분산이 아닌 형태의 데이터마다 다른 가중치를 주어서

등분산을 만족하게 해주는 '**일반화된 최소제곱법**'의 형태 중 하나

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

w_i 는 가중치이며, 분산에 반비례

5 등분산성 진단과 처방

처방 | ① 변수 변환(Box-cox Transformation)

Box-cox Transformation

정규성 만족을 위한 처방과 똑같이 **변수변환 방법**을 적용할 수 있음

가중 회귀 제곱 (WLS: Weighted Least Square)

관측치마다 다른 가중치를 주어서 등분산을 만족하게 해주는

'**일반화된 최소제곱법**'의 형태 중 하나

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

w_i 는 가중치이며, 분산에 반비례

5 등분산성 진단과 처방

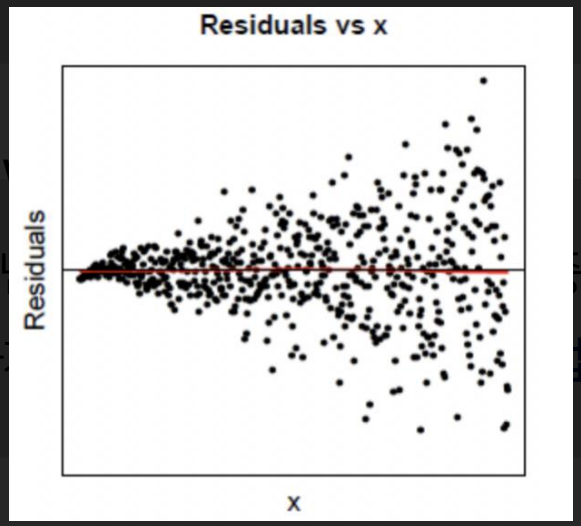
처방 | ① 변수 변환(Box-cox Transformation)

가중치 선정 방식

Box-cox Transformation

1. 잔차 플랏 이용

정규성 만족을 위한 처방과 독같이 변수변환 방법을 적용할 수 있음



가중 회귀 제곱 (WLS:

등분산이 아 치를 주어서

등분산을 만족하 의 형태 중 하나

$$\sum w_i \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_n x_{in}\right)^2 \quad w_i \propto \frac{1}{\sigma_i^2}$$

Residual Plot에서 분산이 점점 커질 경우,

w_i 는 가중치이며, 분산에 반비례

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

5 등분산성 진단과 처방

처방 | ① 변수 변환(Box-cox Transformation)

2. 모델 기반 선정

① OLS로 다중선형회귀모형을 적합

② 다중선형회귀 모델 추정
(종속변수: 잔차의 제곱, 독립변수: 오차 분산에 영향을 주는 변수)

③ 이를 통해 n개의 적합값을 구하기

5 등분산성 진단과 처방

처방 | ① 변수 변환(Box-cox Transformation)

2. 모델 기반 선정

Box-cox Transformation

정규성 만족을 위한 처방으로 적절한 변수 변환 방법을 적용할 수 있음

④ 가중치 설정

적합값 제곱의 역수를 가중치로 설정, 기존 데이터에 가중회귀모델 적용

가중 회귀 제곱 (WLS: Weighted Least Square)

등분산이 아닌 형태의 데이터마다 다른 가중치를 주어서

등분산을 만족하는 형태로 변환하는 방법

⑤ 처음 모델과 가중치 적용 모델의 회귀계수 비교

▼

회귀계수의 차이가 작은 최적의 모형 선택

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

w_i 는 가중치이며, 분산에 반비례

5 등분산성 진단과 처방

처방 | ① 변수 변환(Box-cox Transformation)

Box-cox Transformation

정규성 만족을 위한 처방과 **변환 방법**을 적용할 수 있음

WLS의 장점은 구한 추정량이

회귀식의 기본 가정 하에 **BLUE**를 만족한다는 것

가중 회귀 제곱 (WLS: Weighted Least Square)

등분산을 만족하게 해주는 '**일반화된 최소제곱법**'의 형태 중 하나

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

w_i 는 가중치이며, 분산에 반비례

6

독립성 진단과 처방

6 독립성 진단과 처방

독립성 가정이란?

독립성 가정

오차항끼리 **서로 독립**이라는 가정

개별 관측치에서 i번째 오차와 j번째 오차가 발생하는 것에
서로 영향을 미치지 않는다

⋮

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

→ 오차들 간 상관성의 패턴이 있다는 것

6 독립성 진단과 처방

독립성 가정이란?

독립성 가정

자기상관(Autocorrelation)

우리 모델이 데이터를 잘 설명한다면, 설명하고 남은 잔차가

특정 패턴을 지니지 않아야 함!

개별 관측치에서 i번째 오차와 j번째 오차가 발생하는 것에

서로 영향을 미치지 않는다

하지만 **시간적/공간적으로 인접**한 관측치들은

유사한 경향을 가지고 있어 **설명하지 못한 패턴**이 남아있을 수 있음

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정

더빈-왓슨 검정

앞 뒤 관측치의 **1차 자기상관성**을 확인하는 검정

1차 자기상관성: 연이어 등장하는 오차들이 상관성을 지니는 것

귀무가설 H_0: 잔차들 간에 1차 자기상관이 없다(독립이다).

대립가설 H_1: 잔차들 간에 1차 자기상관이 있다(독립이 아니다).

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정

검정통계량

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

First Order Autocorrelation

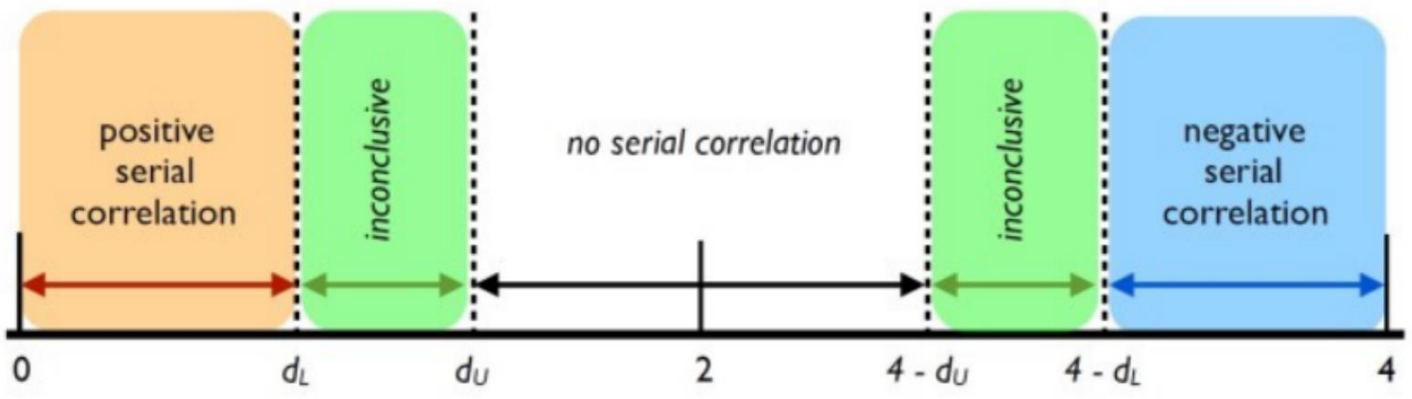
$$\widehat{\rho_1} = \frac{\widehat{Cov}(e_i, e_{i-1})}{\sqrt{V(e_i)} \cdot \sqrt{V(e_{i-1})}} \approx \frac{\sum_{i=1}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} {e_i}^2}$$

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

$\widehat{\rho_1}$ 표본 잔차 자기상관

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정



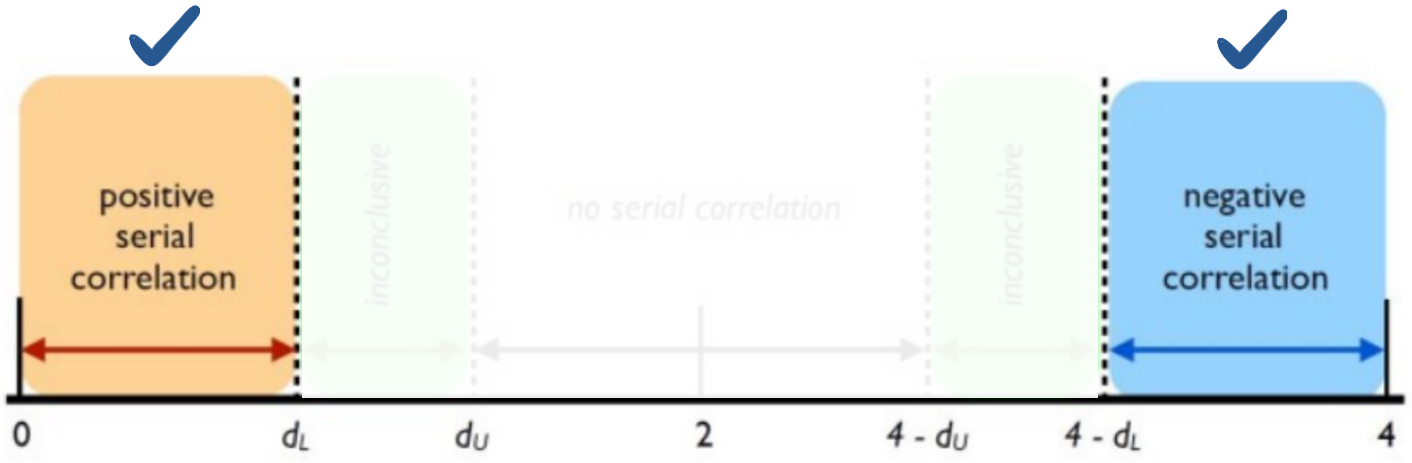
더빈 왓슨 검정표에서 데이터 개수 n과 변수의 개수 p에 따라

귀무가설 기각 여부를 판단하는 **cut-off** 값을 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



$d < d_L$이거나 $d > (4 - d_L)$일 경우 귀무가설 기각

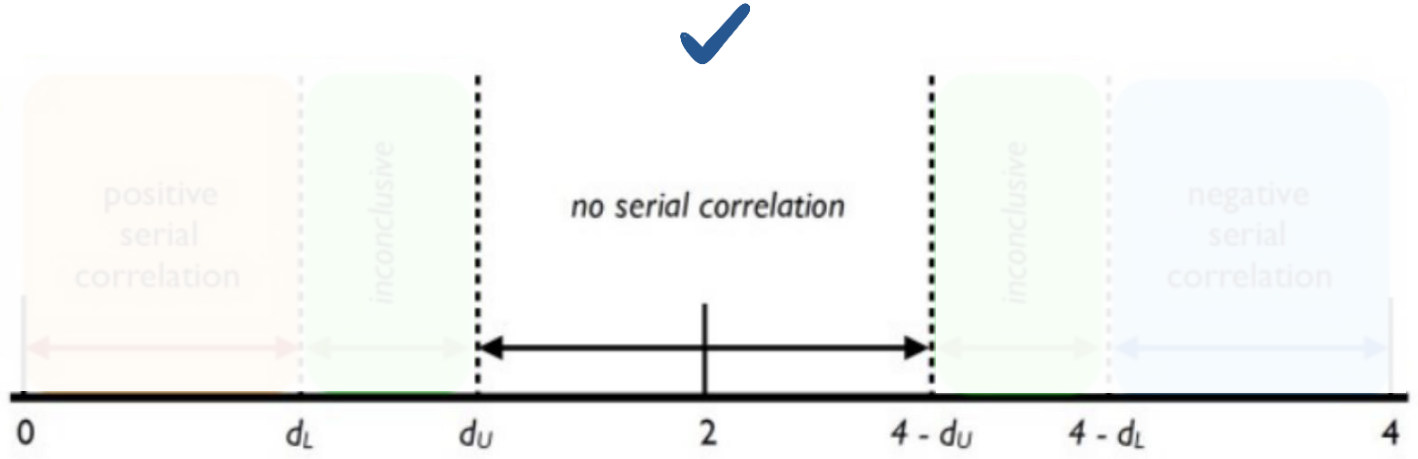
⋮

잔차들 간에 1차 자기상관이 있음

하한(d_L)보다 작으면 양의 자기상관, $(4 - d_L)$보다 크면 음의 자기상관 존재

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정



$d > d_L$이고 $d < (4 - d_L)$일 경우,

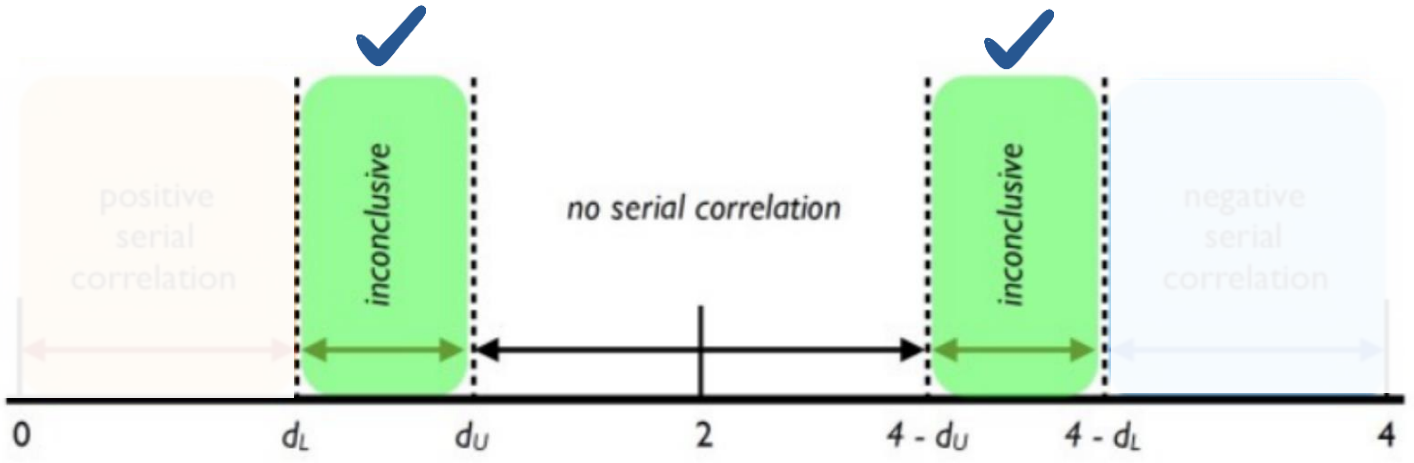
즉 2에 가까울 경우에는 귀무가설 기각 안 됨

⋮

잔차들 간에 1차 자기상관이 없음

6 독립성 진단과 처방

진단 | ① 더빈-왓슨 검정



더빈-왓슨 검정의 한계

① d가 상한과 하한 사이에 위치하게 된다면 자기상관성을 판단할 수 없음

② 바로 인접한 오차와의 1차 자기상관만 고려함

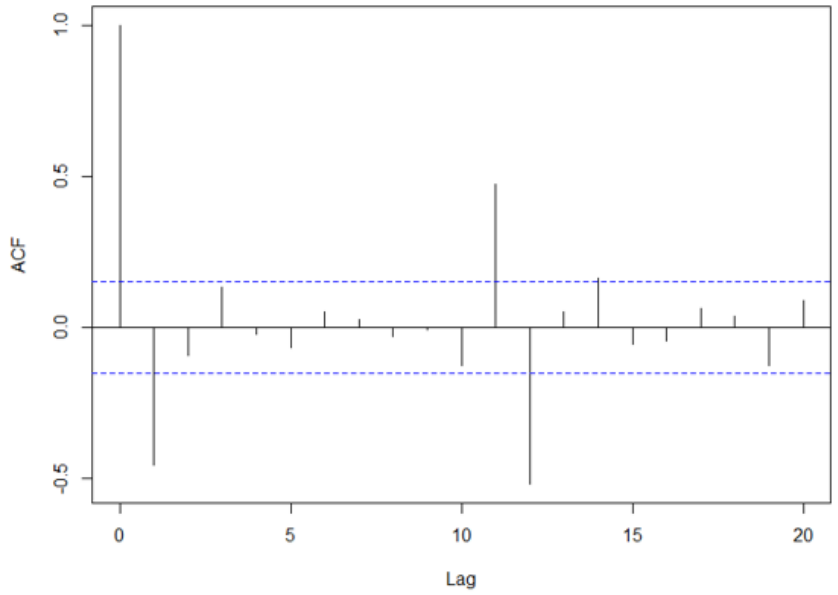
▶ 장기간 지속되는 자기상관이나, 계절성이 있는 경우 확인이 힘듦

진단 | ② Autocorrelation Function plot (ACF plot)

Autocorrelation function

1차 자기상관부터 p차 자기상관까지 고려하고,

신뢰구간을 반환하므로 통계적인 기준에 따라 판단 가능



진단 | ② Autocorrelation Function plot (ACF plot)

Autocorrelation function

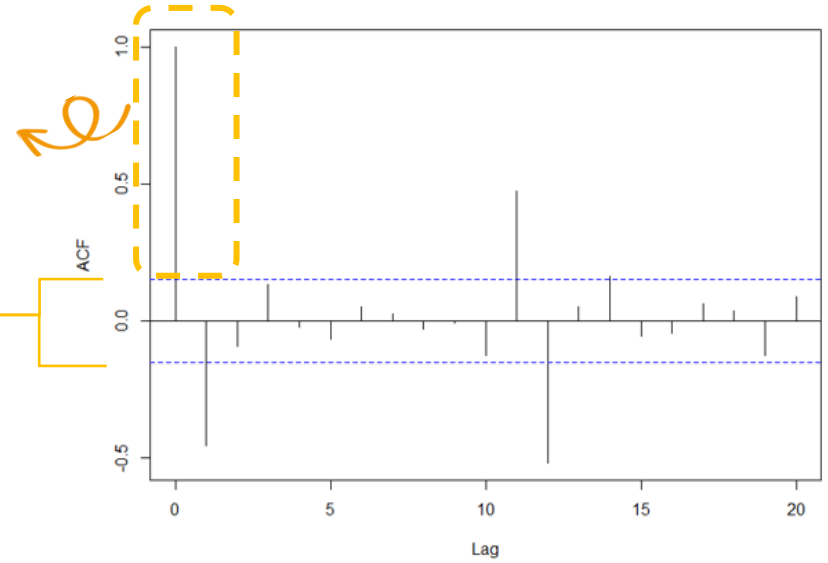
1차 자기상관부터 p차 자기상관까지 고려하고,

신뢰구간을 반환하므로 통계적인 기준에 따라 판단 가능

신뢰구간을 벗어나는 선

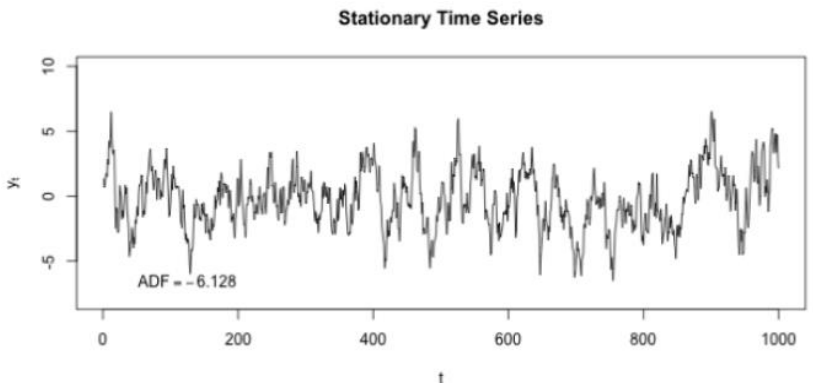
: p**차 자기상관**이 있다고 간주

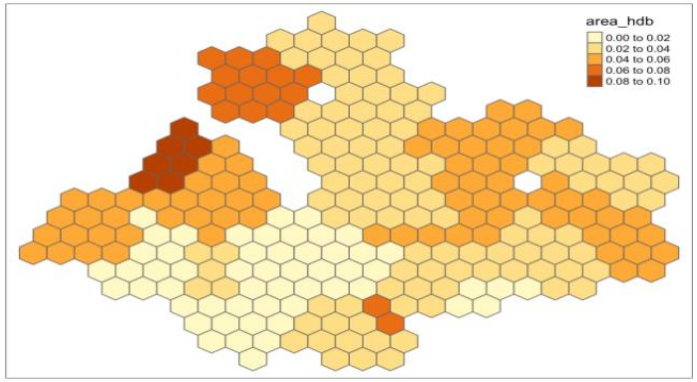
신뢰 구간



6 독립성 진단과 처방

처방 | ① 분석 모델 변경





분석 모델 변경

1) **시간**에 따른 자기상관

▶ 자기 상관을 고려하는 AR(p) 같은 **시계열 모델** 사용

2) **공간**에 따른 자기상관

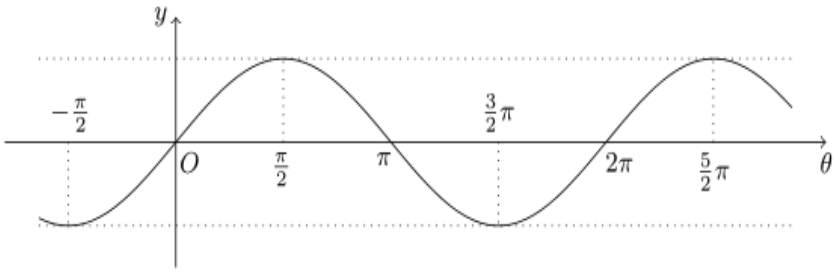
▶ 공간의 인접도를 고려하는 **공간회귀모델** (SEM, SLM 등) 사용

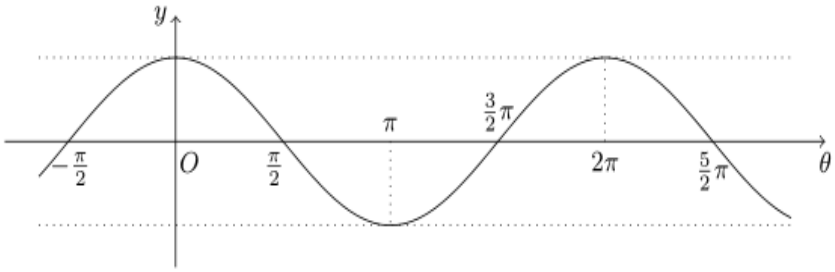
6 독립성 진단과 처방

처방 | ② 가변수 만들기

가변수 생성

뚜렷한 **계절성**이 있다고 판단되면, **가변수** 생성!





주기함수인 **삼각함수 cos(t), sin(t)**의 선형결합으로 주기를 표현!

6 독립성 진단과 처방

처방 | ② 가변수 만들기

gvlma package

가변수 만들기

(Global Validation of Linear Model Assumption)

뚜렷한 **계절성**이 있다고 판단되면, **가변수** 생성!

선형성, 정규성, 등분산성을 한 번에 체크해주는 유용한 패키지

```
                     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

주기함수인 **삼각함수** $\cos(t)$, $\sin(t)$의 선형결합으로 주기를 표현!

Link Function : 선형성 / Heteroscedasticity : 등분산성

6 독립성 진단과 처방

처방 | ② 가변수 만들기

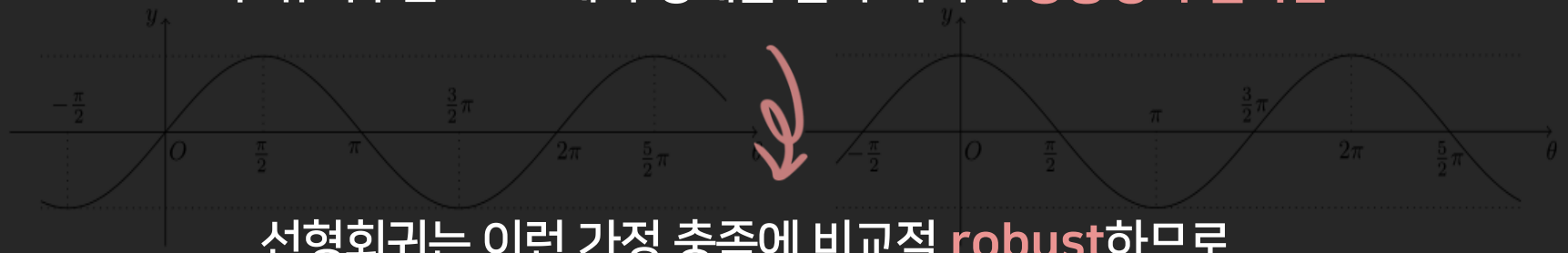
가변수 만들기

gvlma package의 한계

뚜렷한 **계절성**이 있다고 판단되면, **가변수** 생성!

선형성, 정규성, 등분산성을 한 번에 볼 수 있어 **간편하지만,**

모두 유의수준 0.05에서 경계를 잘라 버려서 **융통성이 떨어짐**



선형회귀는 이런 가정 충족에 비교적 **robust**하므로,

gvlma의 결과만 보고

주기함수인 삼각함수 cos(t), sin(t)의 선형결합으로 주기를 표현!

비선형 모델로 변경하는 등의 판단은 **위험**할 수 있음!

다음 주 예고

1. 다중공선성
2. 변수선택법
3. 정규화

Thank you!

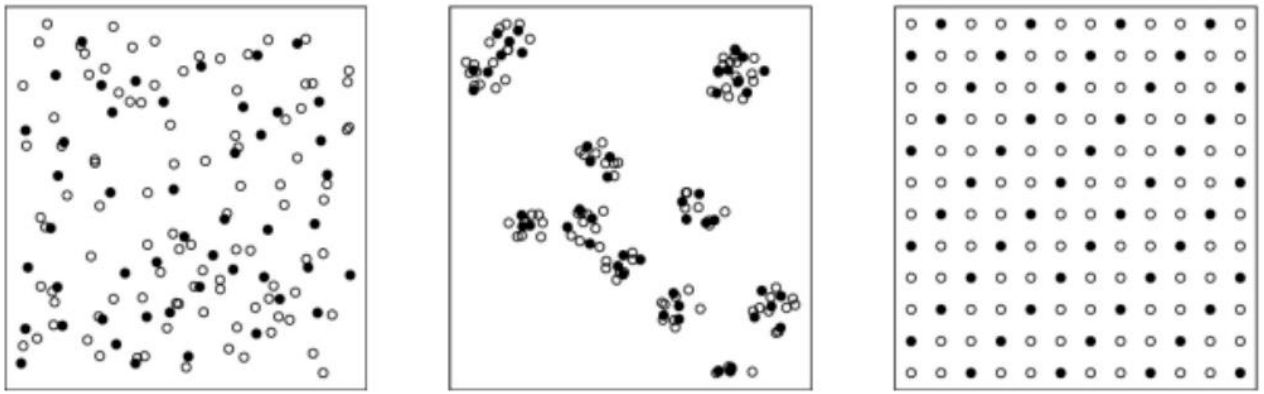
7

공간회귀분석

공간 데이터

공간 데이터

공간 상의 **위치** 또는 **좌표**와 관련된 속성의 집합



공간회귀

공간패턴을 형성하는 데 영향을 미친 **공간과정**을 파악

7 공간회귀분석

공간데이터의 특성 ① : 공간자기상관

Tobler 지리학 제 1 법칙

Everything is related to everything else,

but near things are more related than distant things

가까이 있을수록 **유사성**을 띄는 공간데이터의 특성을 **공간자기상관**이라 함

강남구는 노원구보다

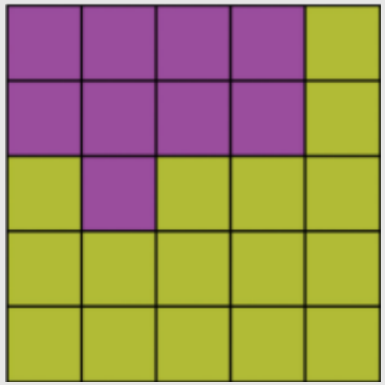
지리적으로 **가까운** 서초구와

아파트 가격이 **비슷함**

공간데이터의 특성 ① : 공간자기상관

positive spatial autocorrelation

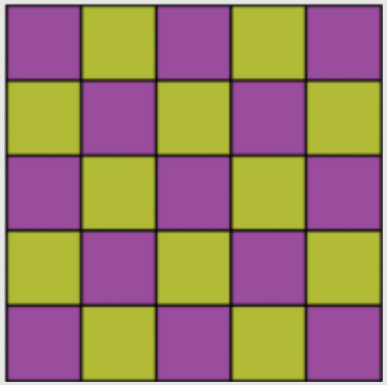
양의 공간자기상관



근처의 관측치들과 유사한 형태

negative spatial autocorrelation

음의 공간자기상관



근처의 관측치들과 상반된 형태

▲ 상관 방향에 따른 구분

공간데이터의 특성 ① : 공간자기상관

global spatial autocorrelation

전역적 공간자기상관

전체 구역이 가지는
하나의 공간자기상관 정도

ex) 서울시에서 나타나는
집값의 공간적인 패턴

local spatial autocorrelation

국지적 공간자기상관

특정 지점이 가지는
개별적인 공간자기상관 정도

ex) 혜화동에서 나타나는
집값의 공간적인 패턴

▲ 공간의 크기에 따른 구분

공간데이터의 특성 ② : 공간적 이질성

공간적 이질성

넓은 지역에서 나타나는 불규칙한 분포를 의미하며,

한 지역 내에 **서로 다른** 성격의 **하위 집단**이 존재하는 것을 말함.

→ 특정 사건이 **전 지역**에서 **동일한 강도**로 나타나는가?

example)

살인 사건이 집값에 미치는 **영향력의 크기**가 모든 지역에서 같은가?

→ 영향을 많이 받는 지역, 영향을 적게 받는 지역 등 **여러 유형**이 존재 가능

7 공간회귀분석

공간자기상관 진단

먼저, 알고자 하는 지역들이 **공간적으로 인접한지**부터 확인해야 함!

공간가중행렬 *spatial weights matrix*

지역 내의 지점들이 서로 공간적으로 인접하고 있는지의 여부를
파악할 수 있도록 **행렬**로 나타낸 것

→ 지역 간의 **잠재적 상호작용의 강도**를 말해줌

$$w_{ij} = \begin{cases} 1 & if\ i,j\ is\ neighbor \\ 0 & otherwise \end{cases}$$

공간가중행렬의 이웃 결정 기준

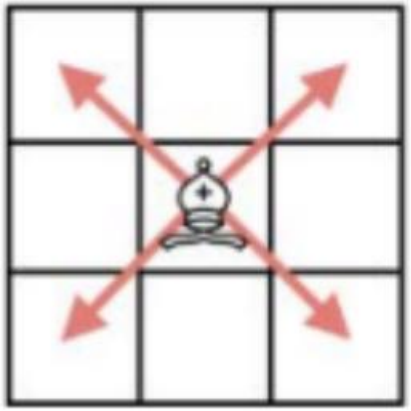
Data에 적합하게 스스로 정의도 가능!

<table>
<tr><td rowspan="3">Binary Contiguity Weights</td><td>Bishop Contiguity</td></tr>
<tr><td>Rook Contiguity</td></tr>
<tr><td>Queen Contiguity</td></tr>
<tr><td colspan="2">Distance-based Weights</td></tr>
<tr><td colspan="2">K-Nearest Neighbors Weights</td></tr>
</table>

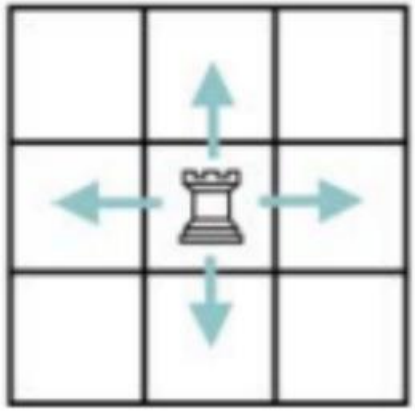
공간가중행렬의 이웃 결정 기준

① Binary Contiguity Weights

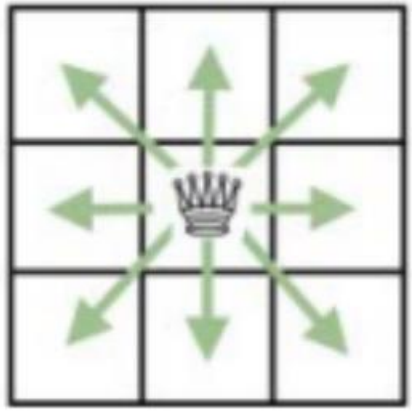
근접하고 있는 경우를 이웃으로 보는 방법



▲ Bishop Contiguity



▲ Rook Contiguity



▲ Queen Contiguity

가장 보편적으로 사용!

Rook Contiguity를 사용하여 직접 **공간가중행렬**을 만들어보자!

① Binary Contiguity Weights

A	B	C
D	E	F
G	H	I

E를 기준으로 면이 붙어 있는 B, D, F, H를 이웃으로 간주

▲ Bishop Contiguity

$$W = \begin{pmatrix} A & B & C & D & E & F & G & H & I & \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & A \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & B \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & C \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & D \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & E \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & F \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & G \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & H \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & I \end{pmatrix}$$

공간가중행렬의 이웃 결정 기준

② Distance-based Weights

특정 거리보다 가까우면 이웃으로 보는 방법

$$w_{ij} = \begin{cases} 1 & if\ d_{ij} < d \\ 0 & otherwise \end{cases}, \text{where } d = minimum\ distance$$

▲ 기준 거리(d)를 너무 작게 설정하면 이웃이 없는 고립된 점이 생길 수 있으므로,

d를 각 관측치별 최단거리보다는 크게 설정해야 함!

공간가중행렬의 이웃 결정 기준

③ K-Nearest Neighbors Weights

머신러닝의 KNN 알고리즘과 비슷한 방식으로

가장 근접한 K개의 점을 이웃으로 보는 방법

이렇게 만들어진 공간가중행렬은

그대로 쓰이지 않고, 정규화하여 사용됩니다!

공간가중행렬의 이웃 결정 기준

③ K-Nearest Neighbors Weights

머신러닝의 KNN 알고리즘과 비슷한 방식으로

가장 근접한 K개의 점을 이웃으로 보는 방법

이렇게 만들어진 공간가중행렬은

그대로 쓰이지 않고, **정규화**하여 사용됩니다!

공간가중행렬의 정규화

① Row Standardized Weights

행 단위로 정규화하는 방법으로, 가중치를 **각 행의 합**으로 나눠줌.

$$w^*_{ij} = \frac{w_{ij}}{\sum_{all\ j} w_{ij}}$$

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

공간가중행렬의 정규화

② Stochastic Weights

전체 행렬을 정규화하는 방법으로, 가중치를 **행렬 전체의 합**으로 나눠줌.

$$w^*_{ij} = \frac{w_{ij}}{\sum_{all\ i,j} w_{ij}}$$

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1/7 & 1/7 & 1/7 \\ 1/7 & 0 & 1/7 \\ 1/7 & 1/7 & 0 \end{pmatrix}$$

공간가중행렬의 정규화

② Stochastic Weights

이렇게 정규화까지 마친 후, 공간자기상관성에 대한 검정 수행!

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1/7 & 1/7 & 1/7 \\ 1/7 & 0 & 1/7 \\ 1/7 & 1/7 & 0 \end{pmatrix}$$

공간자기상관 진단

Moran's I 지수

지역 간 인접성을 나타내는 **공간가중행렬**과

실제 인접 지역들 간의 **속성 데이터**의 **유사성**을 측정하는 방법

$$I = \frac{N \sum_i^N \sum_j^N w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_i^N \sum_j^N w_{ij}) \sum_i^N (Y_i - \bar{Y})^2}$$

$$Z_I = \frac{I - E(I)}{\sqrt{Var(I)}} \quad where \;\; E(I) = -\frac{1}{N-1}$$

N : 지역 단위 수, Y_i : i지역의 속성, Y_j : j지역의 속성, $\bar{Y}$: 평균값, w_{ij} : 가중치

7 공간회귀분석

공간자기상관 진단

Moran's I 지수

▶ I 값의 범위 : -1 ~ 1

▶ 검정통계량 : Z_I → **Z검정**을 통해 전역적 공간자기상관의 유의성 판단

⋮

한계

전체 공간에서 패턴이 있는지 없는지만 알 수 있을 뿐,

핫스팟이나 **콜드스팟**의 위치는 **알 수 없음**!

* 핫스팟 : 자기상관이 비교적 강하게 나타나는 지역

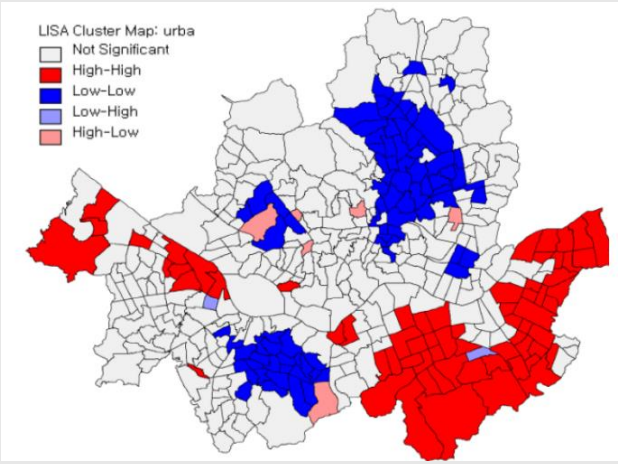
* 콜드스팟 : 자기상관이 비교적 약하게 나타나는 지역

7 공간회귀분석

공간자기상관 진단

LISA 지표 *Local Indicator of Spatial Association*

특정 개별 지역들이 전체 지역의 공간자기상관성에
얼마나 영향을 미치는지 파악하는 **국지적** 측정 방법



공간 자기상관이 **세부적**으로
어느 지역에서 나타나는 것인지 알 수 있다!

HH(high-high), LL(low-low): 공간적 군집지역
HL(high-low), LH(low-high): 공간적 이례지역

공간자기상관 진단

라그랑지 승수검정 *Lagrange Multiplier*

OLS 회귀모델의 **종속변수** 또는 **오차**에서

공간자기상관이 실재하지 않는다는 귀무가설에 대해 검정하는 것

공간자기상관이 **종속변수**(LM-Lag)에서 나타났는지,

오차(LM-Error)에서 나타났는지에 따라

사용해야 하는 공간회귀모델이 달라집니다!

7 공간회귀분석

공간회귀모델을 어떻게 선택해야 할까?

공간자기상관 진단

Moran I 지수, LISA로 공간자기상관성 확인

라그랑지 승수검정 *Lagrange Multiplier*

OLS 회귀모델의 종속변수 또는 오차에서

공간자기상관이 실재하지 않는다는 귀무가설에 대해 검정하는 것

라그랑지 승수검정으로 모델 선택

공간자기상관이 종속변수(LM-Lag)에서 나타났는지,

오차(LM-Error)에서 나타났는지에 따라

사용해야 하는 공간회귀모델이 달라짐

둘 다 유의 X	LM-Lag 유의	LM-Error 유의	둘 다 유의 O
OLS 회귀모델	공간시차모델	공간오차모델	Robust LM

공간자기상관 처방

앞서 말한 공간데이터의 특성(문제 유형)에 따라 해결방법이 달라짐

공간 자기상관성	공간시차모델(SLM)
	공간오차모델(SEM)
공간적 이질성	지리가중회귀모형(GWR)

공간 자기상관성 → 인접지역의 영향력을 **변수에 포함**시켜 통제

공간 이질성 → 각 지역마다 **다른 추정계수**로 영향력을 추정

공간자기상관 처방

① 공간시차모델 (SLM, Spatial Lag Model)

인접지역의 공간적 의존성을 변수로 투입시켜서

공간시차변수를 하나의 **설명변수**로 두는 모델

$$Y = \rho WY + X\beta + \varepsilon = (1 - \rho W)^{-1}(X\beta + \varepsilon)$$

공간시차변수

example)

주택가격 = 주택면적 + 건축년도 + 가구주의 소득 + 오차

공간시차변수 투입

주택가격 = **W*주택가격** + 주택면적 + 건축년도 + 가구주의 소득 + 오차

공간자기상관 처방

② 공간오차모델 (SEM, Spatial Error Model)

오차의 공간자기상관성은 주로 **숨겨진 설명변수**를 고려하지 못했기 때문

→ 오차를 **공간오차변수**로 변형시킨 모델

$$Y = X\beta + \mu = X\beta + (I - \lambda W)^{-1}\varepsilon$$

공간오차변수

example)

주택가격 = 주택면적 + 건축년도 + 가구주의 소득 + 오차

공간오차변수로 변형

주택가격 = 주택면적 + 건축년도 + 가구주의 소득 + **오차(W*오차 + ε)**

공간자기상관 처방

③ 지리가중회귀모델 (GWR, Geographically Weighted Regression)

변수들 간의 관계에 대한 **회귀계수가 지역마다 서로 다르다는 전제** 하에
지역 별로 회귀모델을 추정하는 국지적 방법

$$W_i^{1/2}Y = W_i^{1/2}X\beta_i + W_i^{1/2}X\varepsilon_i$$

$$\beta(u_i, v_i) = [X'W(u_i, v_i)X]^{-1}X'W(u_i, v_i)XY$$

주의

회귀분석이 분석단위(지역) 별로 이루어졌기 때문에
추정된 **회귀계수** 값은 **해당 격자에서만** 의미가 있음!

공간자기상관 처방

③ 지리가중회귀모델 (GWR, Geographically Weighted Regression)

변수들 간의 관계에 대한 회귀계수가 지역마다 서로 다르다는 전제 하에
지역 별로 회귀모델을 추정하는 국지적 방법

지리가중회귀모델을 사용했을 때, 같은 변수에 대한 **지역별 회귀계수의 차이가 크다면**

역으로 공간적 이질성이 존재한다고 볼 수 있음.

$$W_i^{1/2}Y = W_i^{1/2}X\beta_i + W_i^{1/2}X\varepsilon_i$$

$$\beta(u_i, v_i) = [X'W(u_i, v_i)X]^{-1}X'W(u_i, v_i)XY$$

단, 지리가중회귀모델은 OLS와 비교하는 과정이 없었기 때문에
대안모형(GWR)이 기준모형(OLS)을 개선했는지 점검해야 함!

주의

회귀분석이 분석단위(지역) 별로 이루어졌기 때문에
추정된 **회귀계수** 값은 **해당 격자에서만** 의미가 있음!

MSE, AIC, BIC 등의 방법을 사용하여 점검 (3주차 클린업 참고)

다음 주 예고

1. 다중공선성
2. 변수선택법
3. 정규화