회귀분석팀

6팀

김보근

김민주

서유진

하희나

INDEX

1

회귀 기본 가정

1 회귀 기본 가정

회귀 기본 가정이 갖는 의미

회귀분석은 **적은 수의 관측치**만으로도

모델을 구성할 수 있고, **좋은 예측과 추정**이 가능

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

선형회귀모델이 만족해야 하는 제약들 → **선형회귀의 기본 가정**

1 회귀 기본 가정

선형 회귀분석의 기본 가정

[선형회귀의 기본 가정]

모델의 선형성

설명변수와 반응변수의
관계가 **선형**이다

오차의 정규성

오차항은 **정규분포**를 따른다

오차의 등분산성

오차항의 분산은 **상수**이다

오차의 독립성

오차항은 서로 **독립**이다

2

선형성 진단과 처방

2 선형성 진단과 처방

선형성 가정이란?

반응 변수가 설명 변수의 **선형 결합**으로 이루어졌다는 가정

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

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

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

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

치환의 과정을 통해 변화된 x를 새로운 x로 취급한다면,

위 결합들을 모두 **선형 결합**으로 이해 가능

선형성 만족!

2 선형성 진단과 처방

선형성 가정 위배

선형성 가정이 위배되었다고 판단되는 경우,

모든 **선형회귀모델은 모델 자체가 성립하지 않음!**

⋮

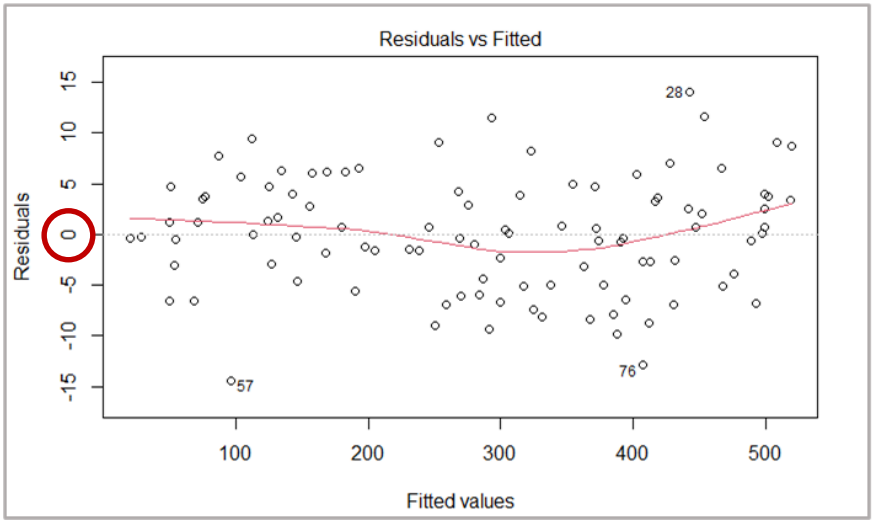
대부분 실제 모델보다 **과소추정**되어 예측 성능이 떨어질 것

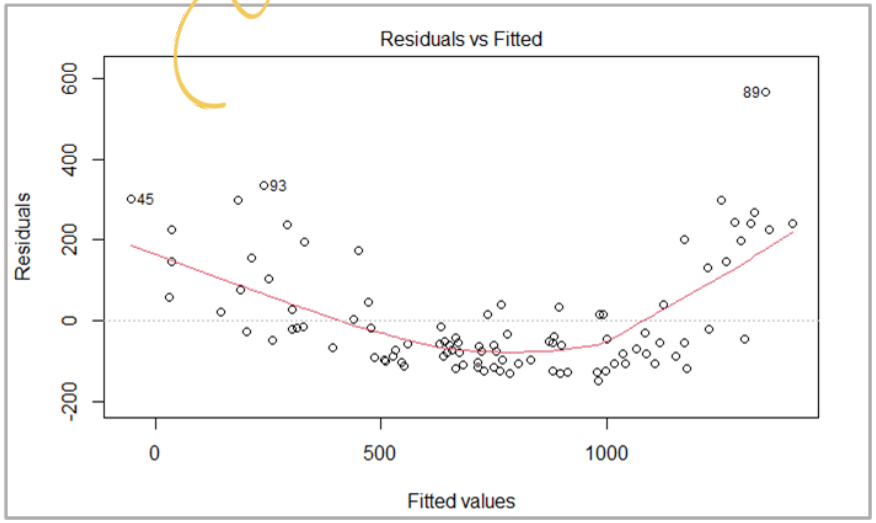
2 선형성 진단과 처방

진단 | ① Residuals vs Fitted Plot

빨간 실선은 전체적인

잔차들의 추세선!





▲ x축 : 예측값($\hat{y}$) , y축 : 잔차($e = y - \hat{y}$)

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

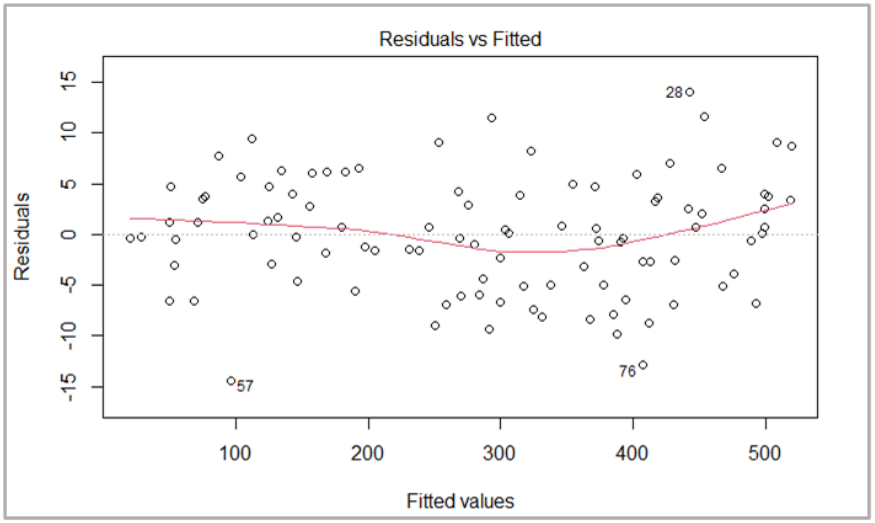
선형성이 위배된 경우, 일반적으로 추세선이 이차함수 혹은 삼차함수 꼴

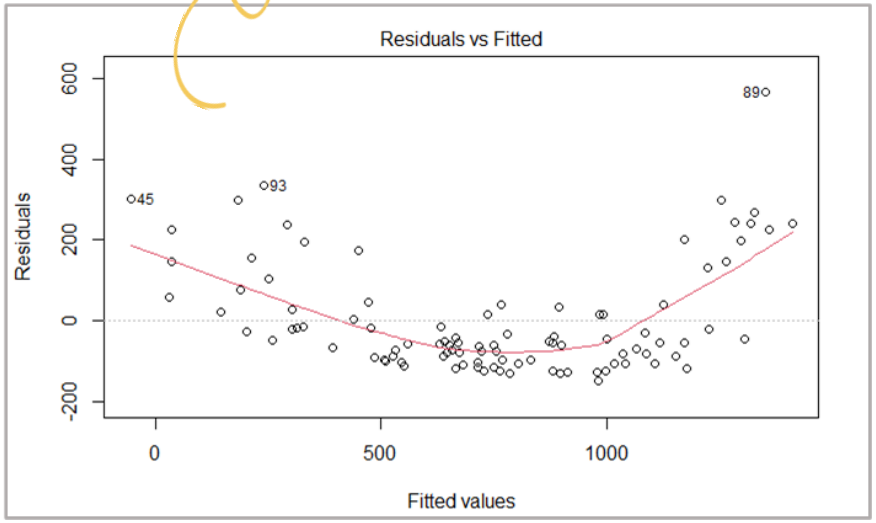
2 선형성 진단과 처방

진단 | ① Residuals vs Fitted Plot

빨간 실선은 전체적인

잔차들의 추세선!





▲ x축 : 예측값($\hat{y}$) , y축 : 잔차($e = y - \hat{y}$)

선형성 만족

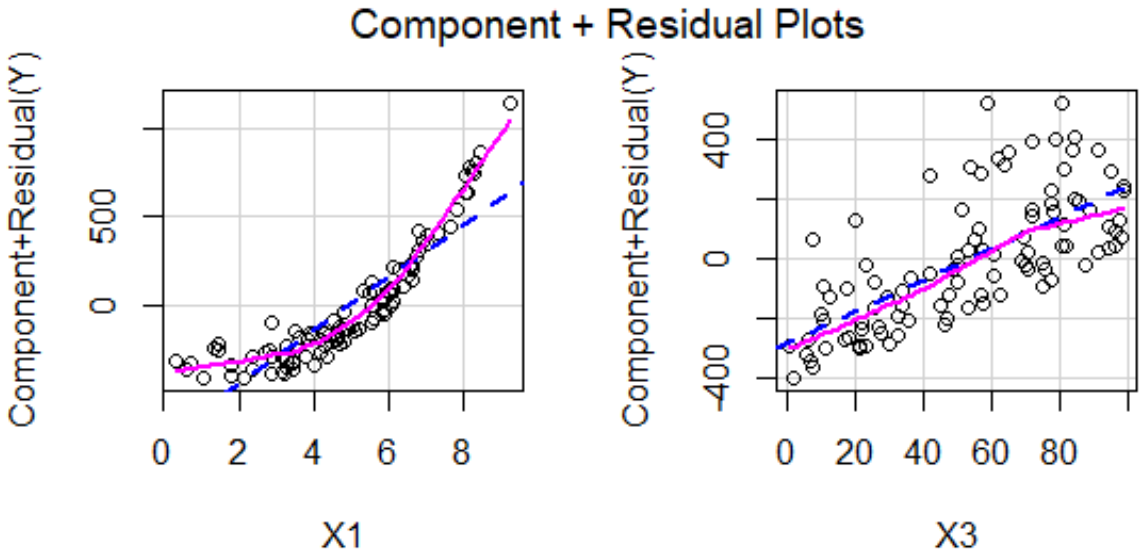
→ 추세선이 x축에 평행

선형성 위배

→ 추세선이 이차함수 형태

진단 | ② Partial Residual Plot

개별 독립 변수와 종속 변수 간의 **선형성** 확인 가능

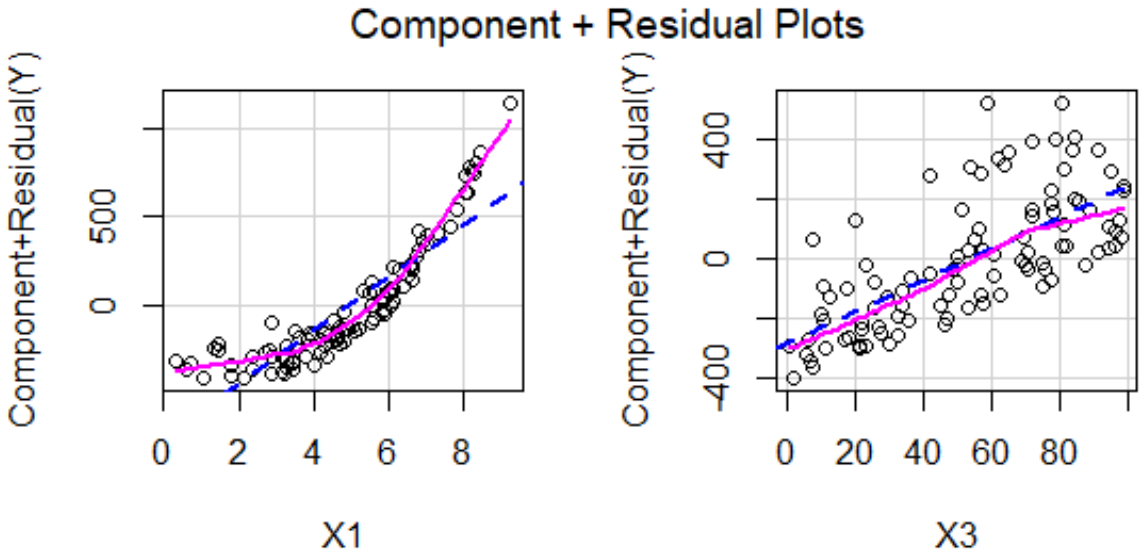


파란 점선과 분홍색 실선이 서로 비슷하면 선형성이 만족되었다고 판단

2 선형성 진단과 처방

진단 | ② Partial Residual Plot

개별 독립 변수와 종속 변수 간의 **선형성** 확인 가능



선형성 위배

→ X1, Y가 선형적이지 않음

선형성 만족

→ X3, Y가 선형적임

2 선형성 진단과 처방

처방 | ① 변수변환

변수의 변환을 통해 비선형 관계를 해결할 수 있음

치환의 과정을 통해 x를 변화시켜 이를 새로운 x로 취급하면 선형결합 만족

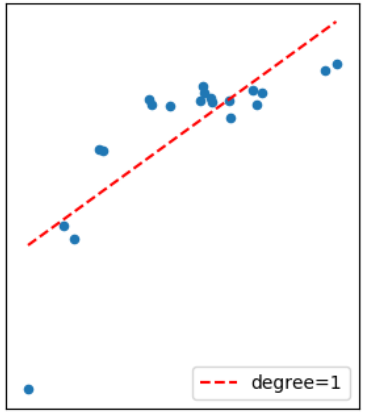
Function	Transformations of x and/or y	Resulting model
$y = \beta_0 x^{\beta_1}$	$y' = log(y), x' = log(x)$	$y' = log(\beta_0) + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = ln(y)$	$y' = ln(\beta_0) + \beta_1 x$
$y = \beta_0 + \beta_1 log(x)$	$x' = log(x)$	$y = \beta_0 + \beta_1 x'$
$y = \frac{x}{\beta_0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

변수 변환을 통해 선형성을 확보할 수 있는 모델도,

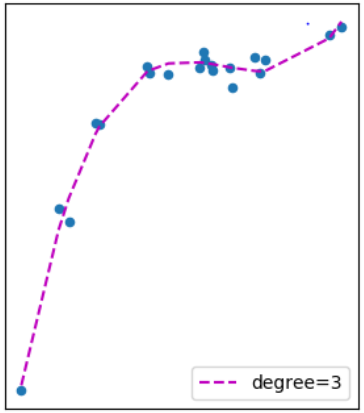
넓은 의미에서 선형모델이라 부름

처방 | ② 다항 회귀 (Polynomial Regression)

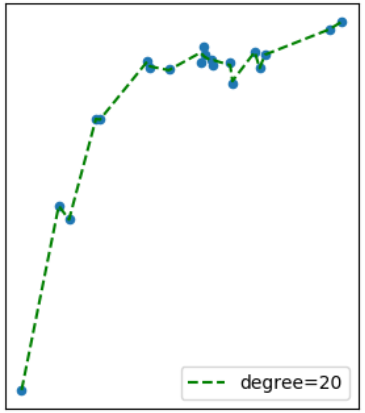
회귀식의 독립 변수가 2차, 3차와 같은 다항식으로 표현되는 것



Underfit
High Bias
Low Variance



Correct Fit
Low Bias
Low Variance



Overfit
Low Bias
High Variance

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \epsilon$$

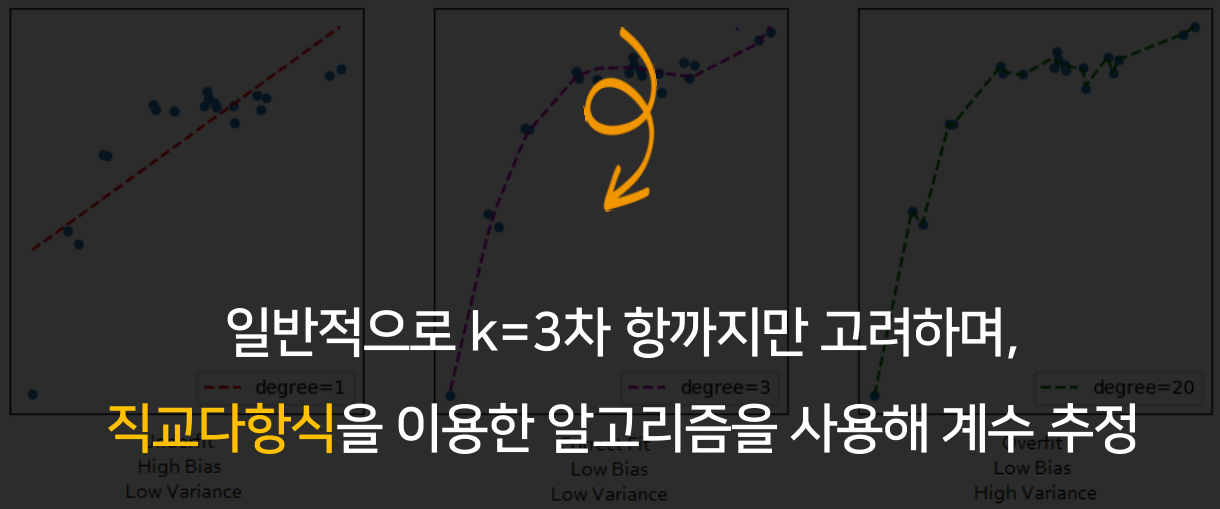
2 선형성 진단과 처방

처방 | ② 다항 회귀 (Polynomial Regression)

회귀식의 독립변수가 2차, 3차의 같은 다항식으로 표현되는 것

과적합 방지를 위해 **적절한 차수 k**를 선정해야 하며,

X의 거듭제곱 꼴이므로 변수들 간 **상관관계가 높아진다는 문제**



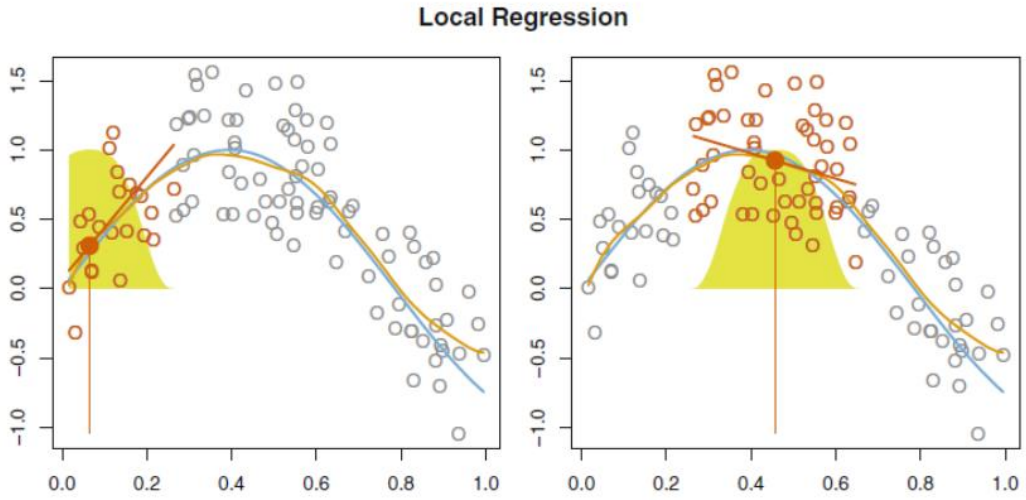
일반적으로 k=3차 항까지만 고려하며,

직교다항식을 이용한 알고리즘을 사용해 계수 추정

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \epsilon$$

처방 | ③ 국소 회귀 (Local Regression)

비선형 문제여도 작은 범위(Local)에서 관찰하면 선형 문제라고 보는 접근 방법



Target data x_0을 중심으로 그 주변 **대역폭 내의 데이터**

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

3

정규성 진단과 처방

3 정규성 진단과 처방

정규성 가정이란?

반응 변수를 측정할 때 발생하는 오차는 정규분포를 따를 것이라는 가정

회귀식이 데이터를 잘 표현한다면 

1) 잔차들은 단순한 측정 오차(noise)로 여겨짐

2) 잔차들의 분포는 정규분포와 흡사한 형태가 됨

3 정규성 진단과 처방

정규성 가정 위배

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

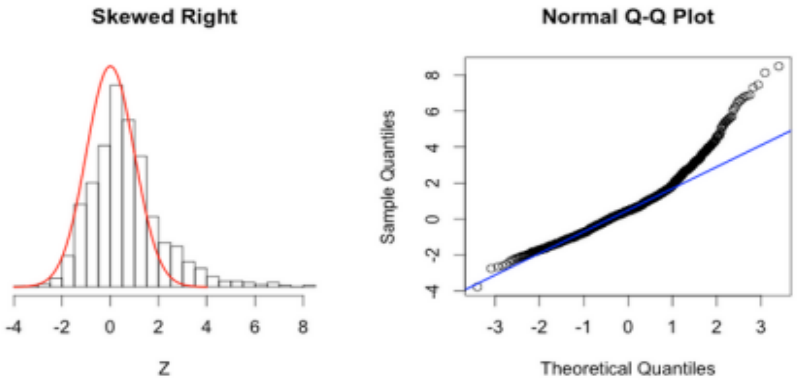
But 정규성 가정이 위배된다면?

가설 검정 결과가 p-value에 의해 유의하게 나오더라도,
검정 결과와 **예측 결과**를 신뢰할 수 없음

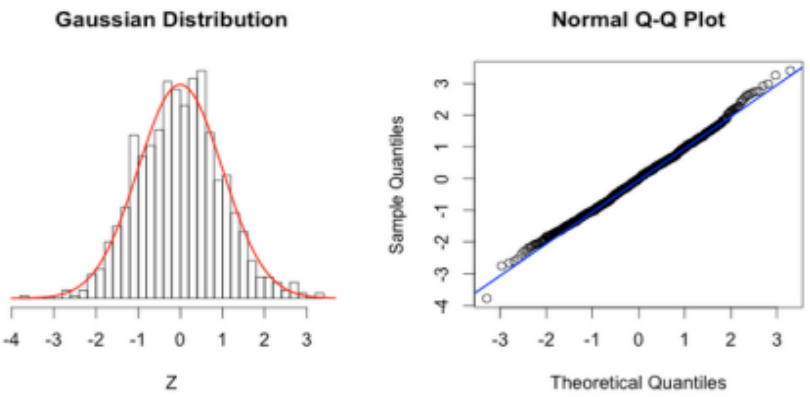
검정통계량이 F분포 혹은 t분포를 따르지 않게 되기 때문!

3 정규성 진단과 처방

진단 | ① Normal Q-Q Plot



▲ 정규성 **불만족**



▲ 정규성 **만족**

☑ X축 : 정규분포의 분위수 값, Y축 : 표준화 잔차

☑ Normal Q-Q Plot이 $y = x$에 가까울수록 잔차가 **정규성**을 만족

진단 | ② 통계적 검정

Plot으로 확인하는 경우 판단이 주관적일 수 있으므로,

명확한 경우가 아니라면 **통계적 방법에 의한 가설검정**으로 확인!

가설

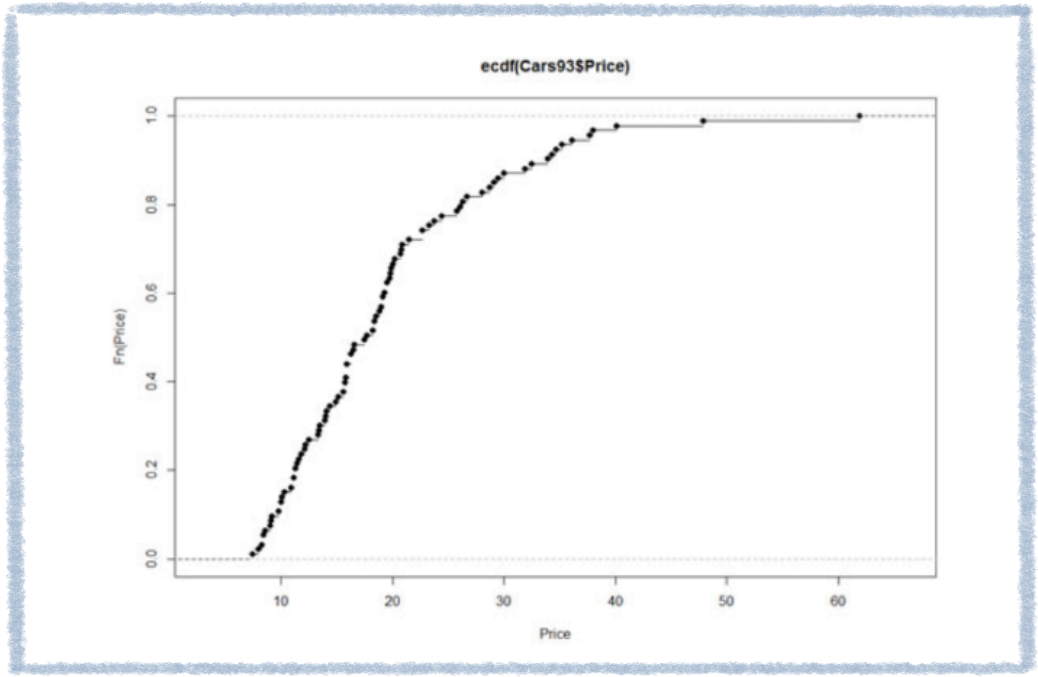
H_0 : 주어진 데이터는 정규분포를 따른다.

H_1 : 주어진 데이터는 정규분포를 따르지 않는다.

귀무가설을 기각하지 못해 정규성이 만족되는 것이 바람직함

진단 | ② 통계적 검정 : Empirical CDF

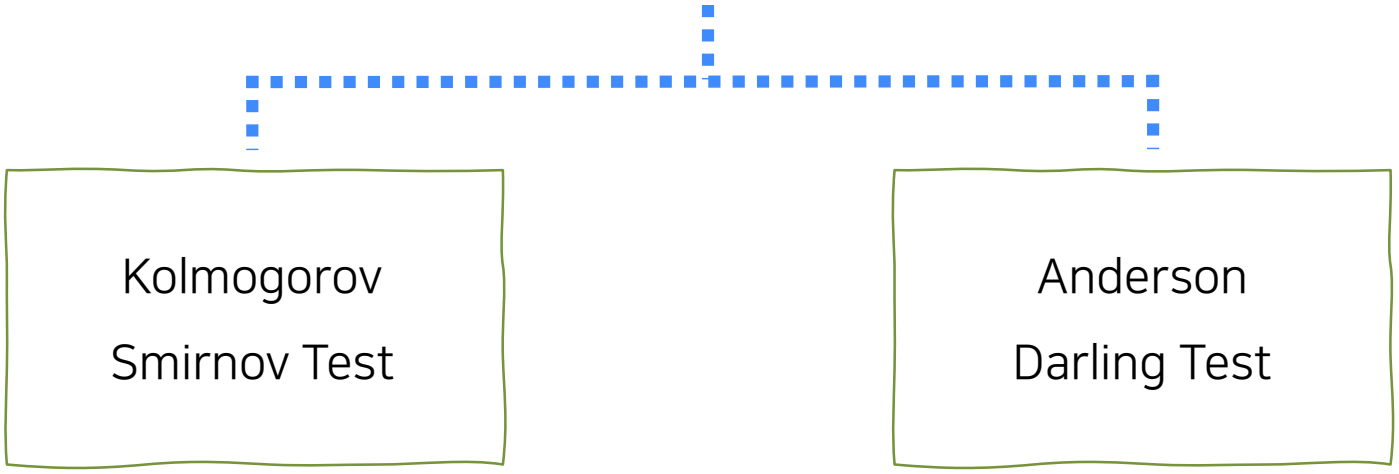
관측치들을 작은 순서대로 나열한 후 누적 분포 함수를 그린 것



진단 | ② 통계적 검정 : Empirical CDF

관측치들을 작은 순서대로 나열한 후 누적 분포 함수를 그린 것

잔차의 Empirical CDF와 정규분포의 CDF를 비교하여 검정



진단 | ② 통계적 검정 : Empirical CDF

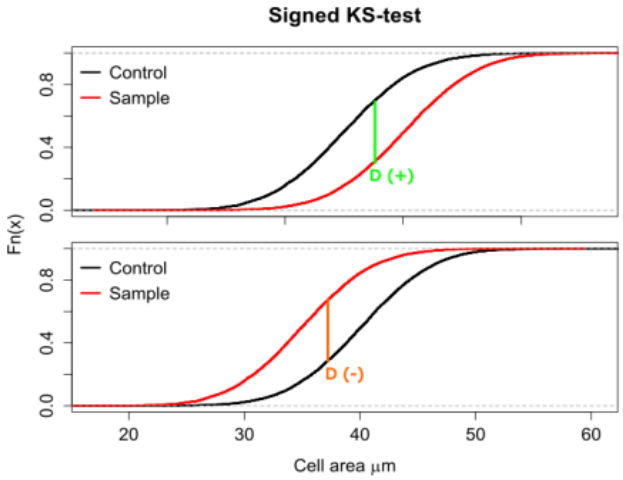
1. Kolmogorov-Smirnov Test (K-S 검정)

하나의 모집단이 어떤 **특정한 분포함수**를 갖는지 알아보는 검정법

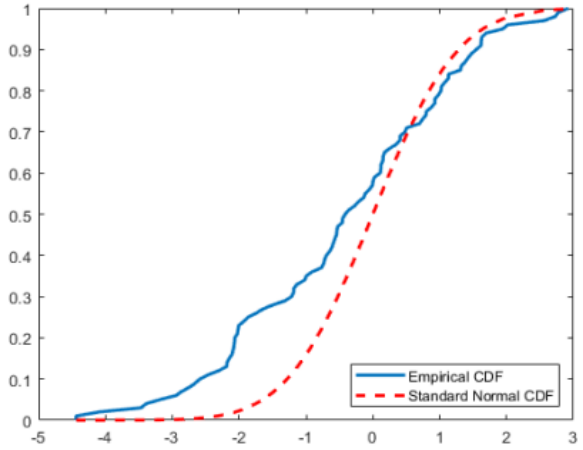
귀무가설 하에서 표본분포함수(sample distribution function)가
어떤 이론적 분포함수(theoretical distribution)와 유사한지 검정

진단 | ② 통계적 검정 : Empirical CDF

1. Kolmogorov-Smirnov Test (K-S 검정)



▲ 이론분포함수 ▲ 표본분포함수



▲ 이론분포함수 ▲ 표본분포함수

검정통계량 !!

이론분포함수와 표본분포함수의 차(D)가 크면 H_0 기각

[표본분포와 이론분포함수는 같지 않다]

진단 | ② 통계적 검정 : Empirical CDF

2 . Anderson-Darling Test (A-D 검정)

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

꼬리 부분의 데이터에 민감하게 반응

A-D 검정 vs. K-S 검정

A-D 검정은 0.05로, K-S 검정은 0.15로 유의수준 설정

A-D 검정이 K-S 검정에 비해서 **더 엄격함**

진단 | ② 통계적 검정 : 정규분포의 분포적 특성 이용

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 검정 가능

3 정규성 진단과 처방

진단 | ② 통계적 검정 : 정규분포의 분[illegible] 특성 이용

1. Shapiro Wilk Test

표본이 정규분포로부터 추출[illegible]지를 확인하기 위한 검정 방법

귀무가설 H_0 를 기각하지 못했다는 것은

정규분포를 따르지 않는다고 말할 근거가 부족하다는 의미

✓ 정규분포 **분위수 값**과 **표준화 잔차** 사이의 **선형관계** 확인

(Q-Q Plot 아이디어와 동일)

✓ [illegible]

100% 정규성이 만족된다는 의미는 아님 !!

진단 | ② 통계적 검정 : 정규분포의 분포적 특성 이용

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 : 데이터 개수

$skew$: 표본의 왜도

$kurt$: 표본의 첨도

잔차의 분포가 정규분포와 달라질수록 왜도/첨도 변화

→ 통계량 값이 커져 유의수준을 넘어서면 귀무가설 기각

3 정규성 진단과 처방

진단 | ② 통계적 검정 : 정규분포의 분포적 특성 이용

2. Jarque-Bera Test

변수변환을 통해 정규성을 처방해줄 수 있음!

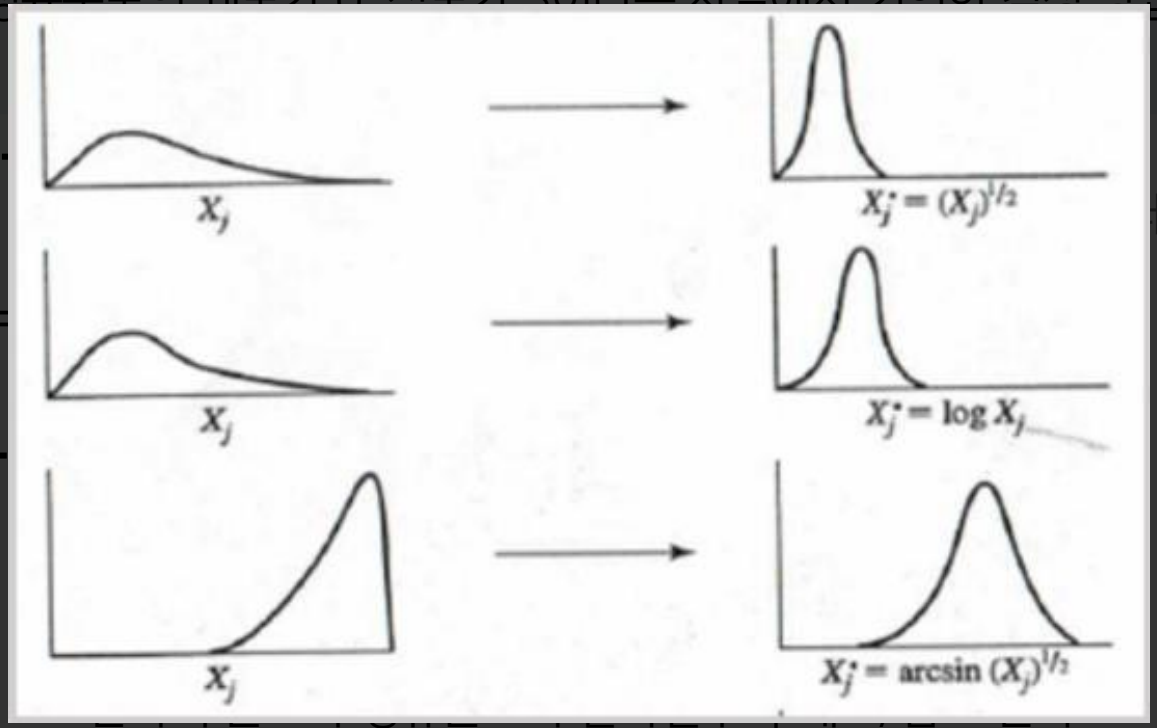
정규분포의 왜도가 0, 첨도가 3이라는 사실에서 기인한 검정 방법

$$JB =$$

데이터 개수

표본의 왜도

표본의 첨도



→ 통계량 값이 커져 유의수준을 넘어서면 귀무가설 기각

3 정규성 진단과 처방

진단 | ② 통계적 검정 : 정규분포의 분포적 특성 이용

2. Jarque-Bera Test

But 단순 변수 변환은 **주관적인** 판단 하에 이루어지므로

객관성을 확보하기 힘듦

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

n : 데이터 개수

$skew$: 표본의 왜도

$kurt$: 표본의 첨도

1) Box-cox Transformation

2) Yeo-Johnson Transformation

잔차의 분포가 정규분포와 달라질수록 왜도/첨도 변화

→ 통계량 값이 커져 유의수준을 넘어서면 귀무가설 기각

처방 | ① Box-cox Transformation

반응변수 (Y)를 변환함으로써 정규성과 등분산성을 해결해주는 방법

λ 를 변화시키면서 y 가 정규성, 등분산성을 만족하도록 함

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

- 일반적으로 λ는 -5에서 5 사이의 값
- 변환된 y가 모든 실수 λ에 대해 연속

처방 | ① Box-cox Transformation

최대우도함수(ML)을 통해 신뢰구간을 구한 후

신뢰구간 내의 로그우도함수를 최대화하는 λ를 최적의 값으로 선택

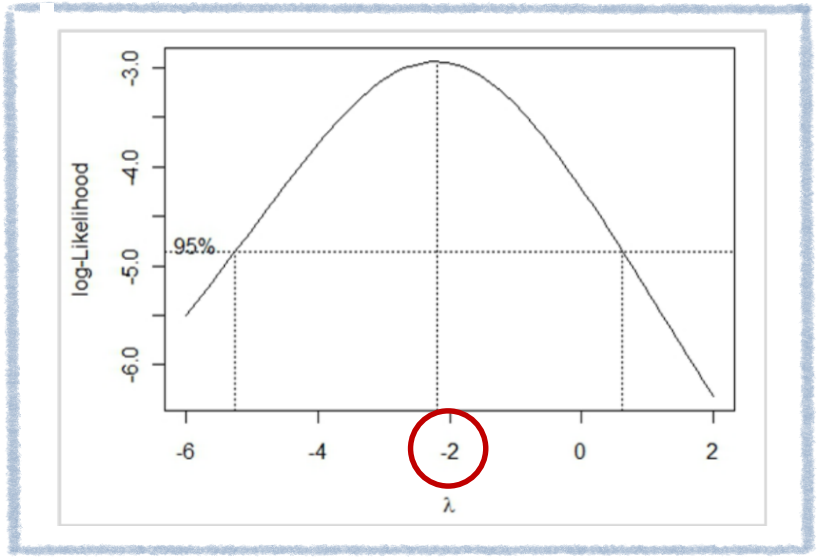
ex. λ=2이면 이차함수, λ=0.5이면 루트, λ=0이면 로그, λ=-1이면 역수변환을 의미

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

- 일반적으로 λ는 -5에서 5 사이의 값
- 변환된 y가 모든 실수 λ에 대해 연속

3 정규성 진단과 처방

처방 | ① Box-cox Transformation



95% 내의 λ값 중

가능도함수가 최대가 되는 λ를 선택

λ를 정수로 택하면 변수 변환 관계 쉽게 파악 가능!

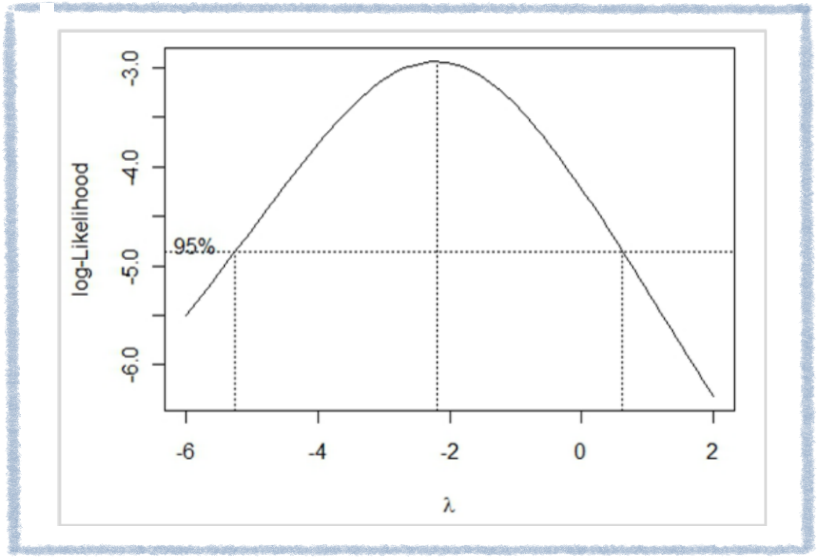
하지만 Box-cox Transformation은

Y가 0 이하일 경우 사용할 수 없음

λ=0이 되면 y가 $\log(y)$로 변환되기 때문!

3 정규성 진단과 처방

처방 | ① Box-cox Transformation



95% 내의 λ값 중

가능도함수가 최대가 되는

-2 근방의 λ를 선택

λ를 정수로 택하면 변수 변환 관계 쉽게 파악 가능!

하지만 Box-cox Transformation은

Y가 0 이하일 경우 사용할 수 없음

λ=0이 되면 y가 $\log(y)$로 변환되기 때문!

처방 | ② Yeo-Johnson Transformation

- Box-cox Transformation과 같은 아이디어

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

Box-cox와 달리 변수 범위에 대한 제약 없음

3 정규성 진단과 처방

처방 | ② Yeo-Johnson Transformation

- Box-cox Transformation과 같은 아이디어

Yeo-Johnson Transformation이 전체 범위에서 사용 가능한데

Box-cox Transformation도 사용하는 이유?

✓ Box-cox Transformation은 전체 범위에 대한 해석에 용이하기 때문!

✓ Yeo-Johnson Transformation은 제곱이 λ, 2-λ로 달라

전체 범위에 대한 해석이 모호해질 수 있음

4

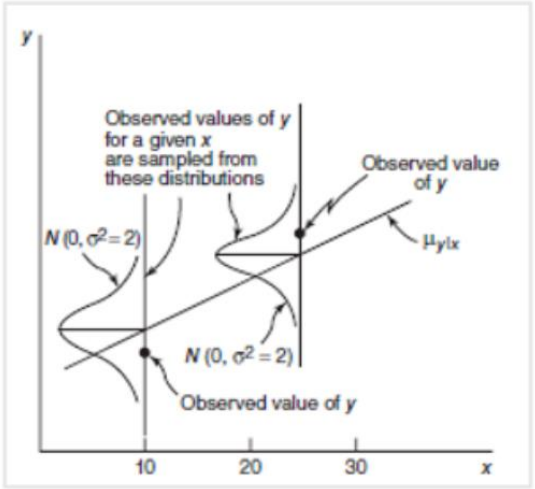
등분산성 진단과 처방

4 등분산성 진단과 처방

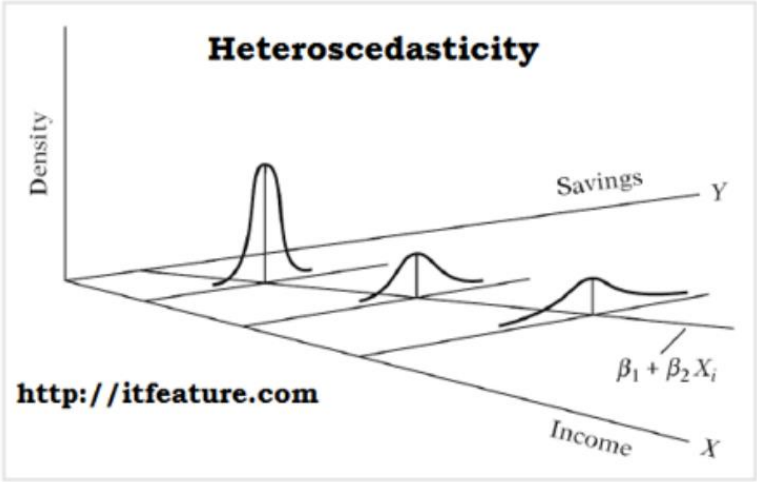
등분산성 가정이란?

오차의 모든 분산은 동일해야 한다는 가정

등분산성



이분산성

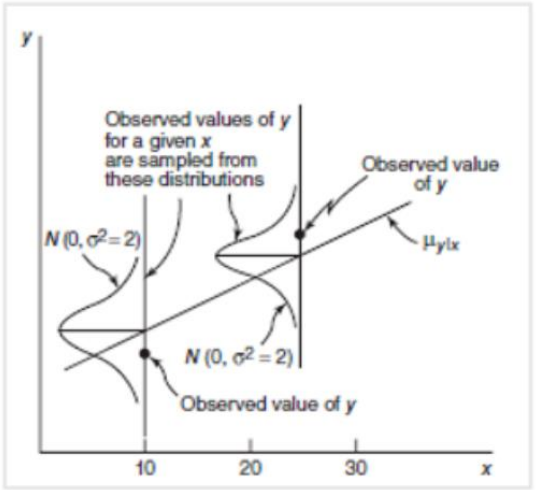


4 등분산성 진단과 처방

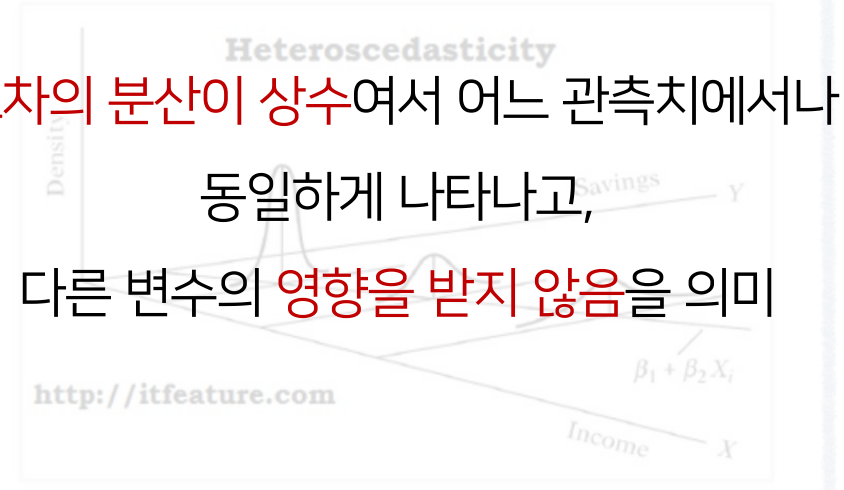
등분산성 가정이란?

오차의 모든 분산은 동일해야 한다는 가정

등분산성



이분산성

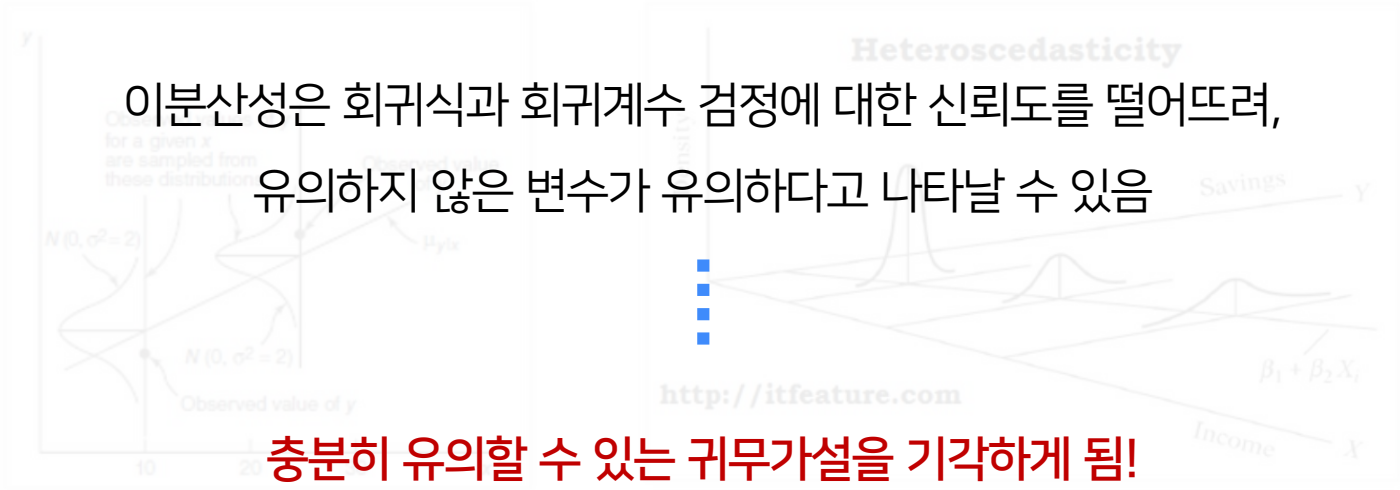


오차의 분산이 상수여서 어느 관측치에서나
동일하게 나타나고,
다른 변수의 영향을 받지 않음을 의미

4 등분산성 진단과 처방

등분산성 가정이란?

오차의 모든 분산은 동일해야 한다는 가정



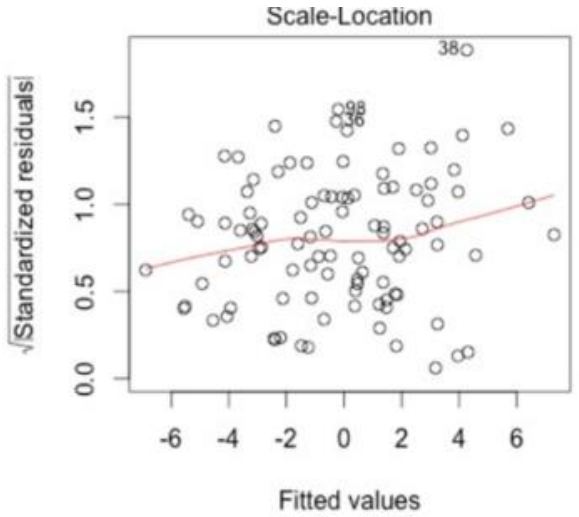
이분산성은 회귀식과 회귀계수 검정에 대한 신뢰도를 떨어뜨려,
유의하지 않은 변수가 유의하다고 나타날 수 있음

충분히 유의할 수 있는 귀무가설을 기각하게 됨!

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

4 등분산성 진단과 처방

진단 | ① 잔차 플랏 : Scale-Location



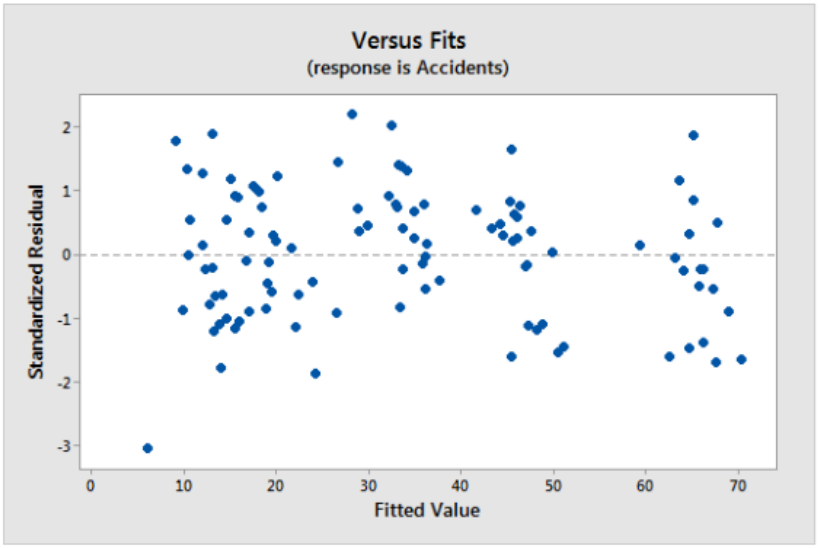
▲ X축 : 예측값($\hat{y}$), Y축 : 표준화 잔차

- 잔차에 절댓값이 씌워진 형태!
- 등분산성이 위배된 경우,

잔차와 예측값 사이에 random한 형태이외의 어떠한 관계가 나타남

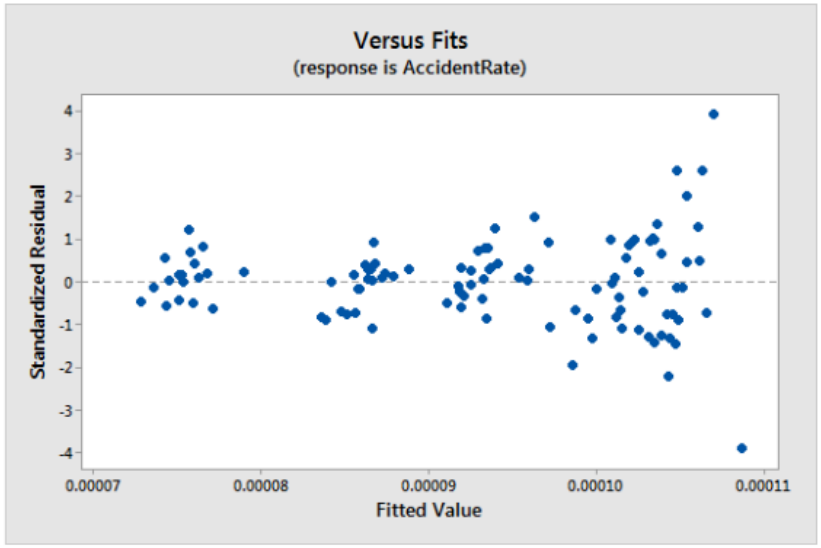
4 등분산성 진단과 처방

진단 | ① 잔차 플랏 : Residual vs Fitted



▲ 등분산성 **만족**

$\hat{y}$값에 상관없이
잔차 퍼짐의 정도가 일정

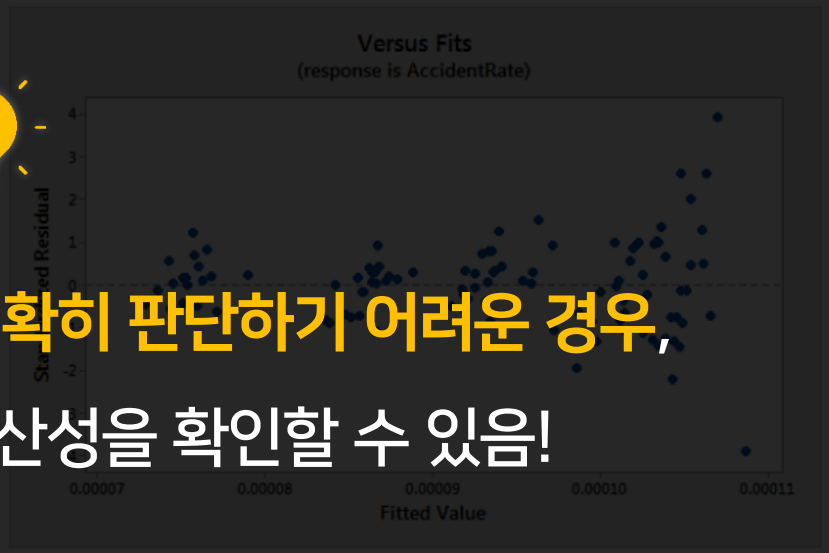


▲ 등분산성 **위배**

$\hat{y}$값이 커지면서
잔차 퍼짐의 정도가 일정하지 않음

4 등분산성 진단과 처방

진단 | ① 잔차 플랏 : Residual vs Fitted



Plot을 확인해도 육안으로 **명확히 판단하기 어려운 경우,**

통계적 검정을 통해 등분산성을 확인할 수 있음!

▲ 등분산성 **만족**

$\hat{y}$값에 상관없이
잔차 퍼짐의 정도가 일정

▲ 등분산성 **위배**

$\hat{y}$값이 커지면서
잔차 퍼짐의 정도가 일정하지 않음

진단 | ② 통계적 검정 : BP(Breusch-Pagan) Test

잔차가 **독립변수**들의 **선형결합**으로 표현되는지를 검정

⋮

설명변수의 증감에 따른 **오차의 분산 변화**를 통해

등분산성 판단 가능!

BP test의 기본 가정

① sample 수가 많아야 함

② 오차항은 독립이고 정규분포를 따라야 함

③ 오차의 분산은 설명변수와 연관이 있어야 함

4 등분산성 진단과 처방

진단 | ② 통계적 검정 : BP(Breusch-Pagan) Test

잔차가 **독립변수**들의 **선형결합**으로 표현되는지를 검정

⋮

설명변수의 증감에 따른 **오차의 분산 변화**를 통해

등분산성 판단 가능!

BP test의 기본 가정

① sample 수가 많아야 함

② 오차항은 독립이고 정규분포를 따라야 함

③ 오차의 분산은 설명변수와 연관이 있어야 함

4 등분산성 진단과 처방

진단 | ② 통계적 검정 : BP(Breusch-Pagan) Test

가설

H_0 : 주어진 데이터는 등분산성을 지닌다.

H_1 : 주어진 데이터는 등분산성을 지니지 않는다 (이분산이다).

귀무가설을 기각하지 못해 등분산성이 만족되는 것이 바람직함

진단 | ② 통계적 검정 : BP(Breusch-Pagan) 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 오차의 제곱(오차의 분산)이

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

4 등분산성 진단과 처방

진단 | ② 통계적 검정 : BP(Breusch-Pagan) Test

검정통계량

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

임계값

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

if $\chi^2_{stat} > \chi^2_{P-1,\alpha}$, $reject\ H_0$

즉, 등분산성을 **만족하지 않음**을 의미

4 등분산성 진단과 처방

진단 | ② 통계적 검정 : BP(Breusch-Pagan) Test

검정통계량 임계값

BP Test의 단점

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

1) **비선형결합**으로 이루어진 이분산성은 파악 불가능

2) 샘플이 **대표본**이어야 함

3) 오차의 **정규성**에 민감하게 반응

➡ 정규성이 지켜진 상태인지 확인해야 함

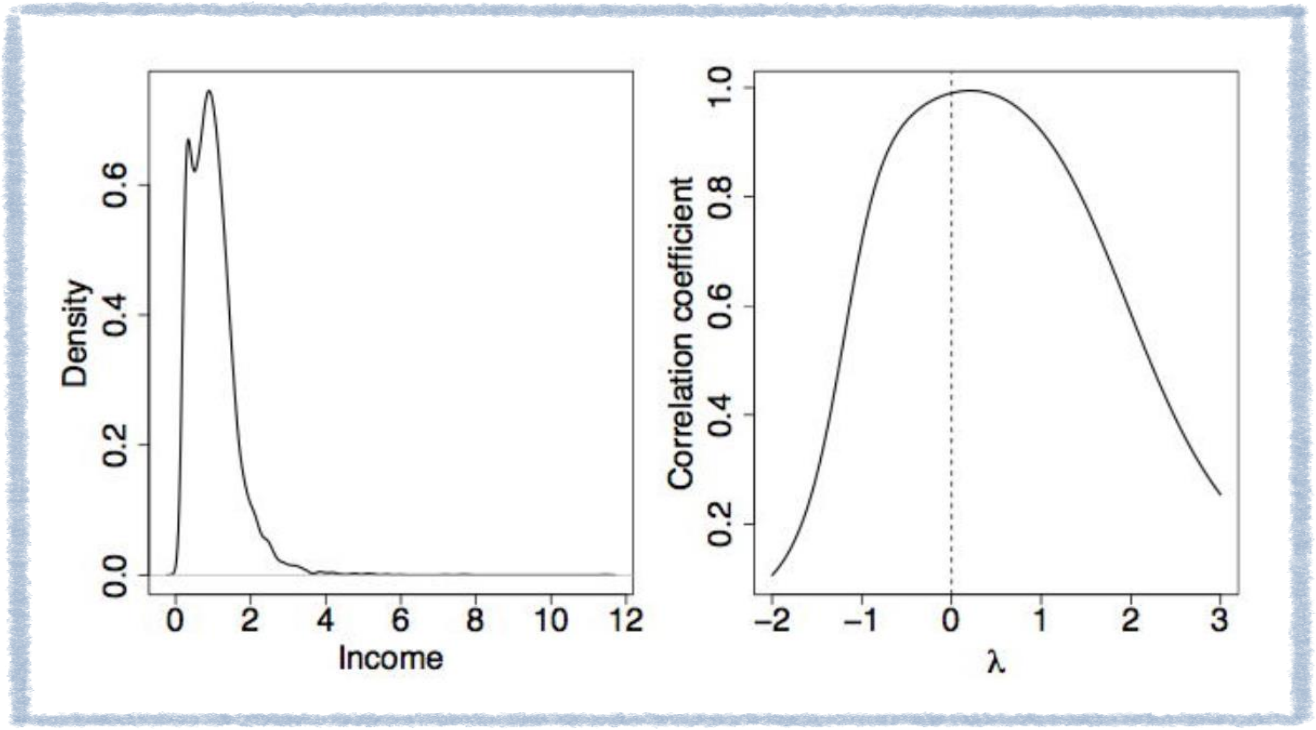
if $\chi^2_{stat} > \chi^2_{P-1,\alpha}$, $reject\ H_0$

즉, 등분산성을 **만족하지 않음**을 의미

4 등분산성 진단과 처방

처방 | ① 변수 변환(Box-cox Transformation)

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



처방 | ② 가중회귀제곱 (WLS: Weighted Least Square)

관측치마다 다른 가중치를 주어서 등분산을 만족하게 해주는
'일반화된 최소제곱법'의 형태 중 하나

분산이 큰 부분의 관측치에는 가중치를 적게 주어
전체적인 분산을 비슷하게 맞춰줌!

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

w_i 는 가중치이며, 분산에 반비례

4 등분산성 진단과 처방

처방 | ② 가중회귀제곱 (WLS: Weighted Least Square)

[가중치 선정 방식]

1. 잔차 플랏 이용

Residual Plot에서 분산이 점점 커질 경우,

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

2. 모델 기반 선정

다중선형회귀 모델을 OLS로 추정 후

(종속변수: 잔차의 절댓값, 독립변수: 오차 분산에 영향을 주는 변수들)

적합값을 이용해 가중치 부여

적합값 제곱의 역수

4 등분산성 진단과 처방

처방 | ② 가중회귀제곱 (WLS: Weighted Least Square)

[가중치 선정 방식]

가중회귀제곱(WLS)의 장점

1. 잔차 플랏 이용

Residual Plot에서 분산이 점점 커질 경우,

$\frac{1}{\sigma_i^2}$

회귀식의 기본 가정 하에 구한

가중회귀제곱(WLS) 추정량은 BLUE를 만족!

2. 모델 기반 선정

다중선형회귀 모델을 OLS로 추정 후

(종속변수: 잔차의 절댓값, 독립변수: 오차 분산에 영향을 주는 변수들)

적합값을 이용해 가중치 부여

적합값 제곱의 역수

5

독립성 진단과 처방

5 독립성 진단과 처방

독립성 가정

오차항들은 서로 독립이라는 가정

개별 관측치에서 i번째 오차와 j번째 오차가 발생하는 것에

서로 영향을 미치지 않는다

⋮

독립성 가정이 위배되었다면, 오차들 간 자기상관(Autocorrelation) 존재

→ 오차들 간 상관성의 pattern이 있다는 것

5 독립성 진단과 처방

독립성 가정

자기상관(Autocorrelation)이란?

모델이 데이터를 잘 설명한다면,
설명하고 남은 잔차가 특정 패턴을 지니지 않아야 함

시각적/공간적으로 인접한 관측치들은
회귀식만으로 설명할 수 없는 패턴이 남아있을 수 있음

5 독립성 진단과 처방

독립성 가정

오차항들은 서로 독립이라는 가정

개별 관측치에서 i번째 오차와 j번째 오차가 발생하는 것에

서로 영향을 미치지 않는다

오차 간 자기상관이 존재하면, σ^2의 추정량과 회귀계수의 표준오차가

실제보다 **과소추정**됨

독립성 가정이 위배되었다면, 오차들 간 자기상관(Autocorrelation) 존재

유의성 검정의 결과를 신뢰할 수 없고,

→ 오차들 간 상관성의 pattern이 있다는 것

Prediction Interval도 넓어짐

5 독립성 진단과 처방

진단 | 더빈-왓슨 검정(Durbin Watson Test)

앞 뒤 관측치의 1차 자기상관성을 확인하는 검정 방법

1차 자기상관성: 연이어 등장하는 오차들이 상관성을 지니는 것

귀무가설 H_0: 잔차들 간에 1차 자기상관이 없다. ($\hat{\rho}_1 = 0$)

대립가설 H_1: 잔차들 간에 1차 자기상관이 있다.

진단 | 더빈-왓슨 검정(Durbin Watson Test)

검정 통계량

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

First order autocorrelation

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

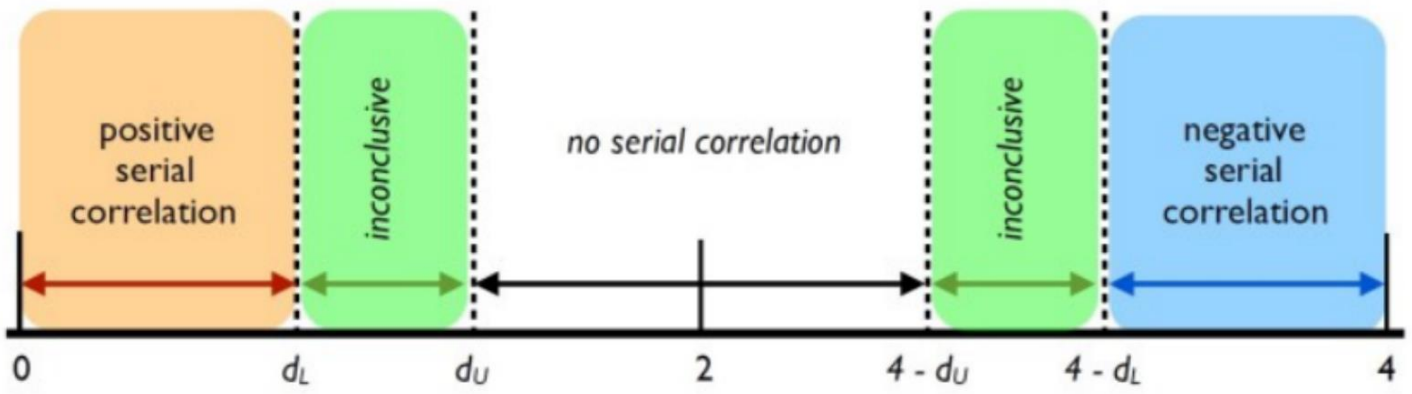
$$\therefore d \approx 2(1 - \hat{\rho}_1), 0 < d < 4$$

$\hat{\rho}_1$: 표본 잔차 자기상관

5 독립성 진단과 처방

진단 | 더빈-왓슨 검정(Durbin Watson Test)

더빈 왓슨 검정표



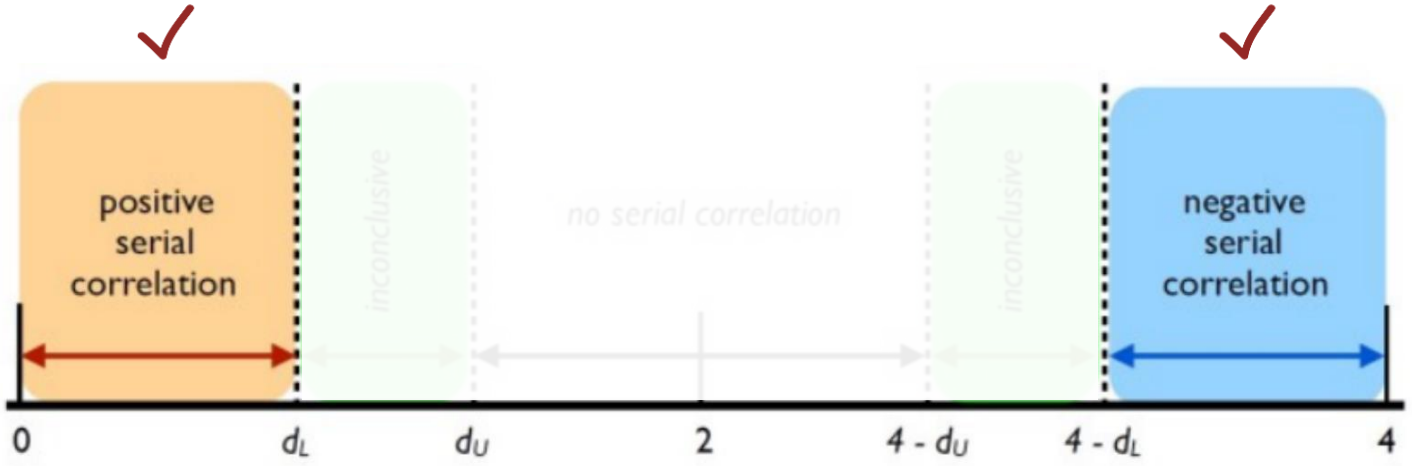
d_L : 경계의 하한, d_U : 경계의 상한

데이터의 개수 n과 변수의 개수 p에 따라

귀무가설 기각 여부를 판단하는 cut-off 값을 알려줌

5 독립성 진단과 처방

진단 | 더빈-왓슨 검정(Durbin Watson Test)



d_L : 경계의 하한, d_U : 경계의 상한

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

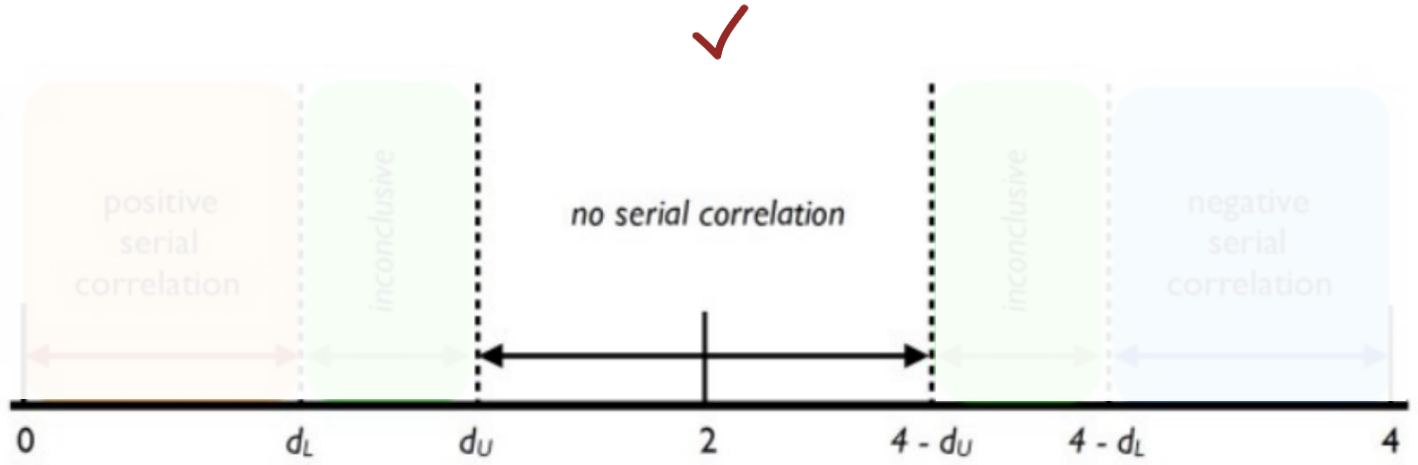
⋮

잔차들 간 1차 자기상관이 있음

(= 앞 오차에 영향을 받음)

5 독립성 진단과 처방

진단 | 더빈-왓슨 검정(Durbin Watson Test)



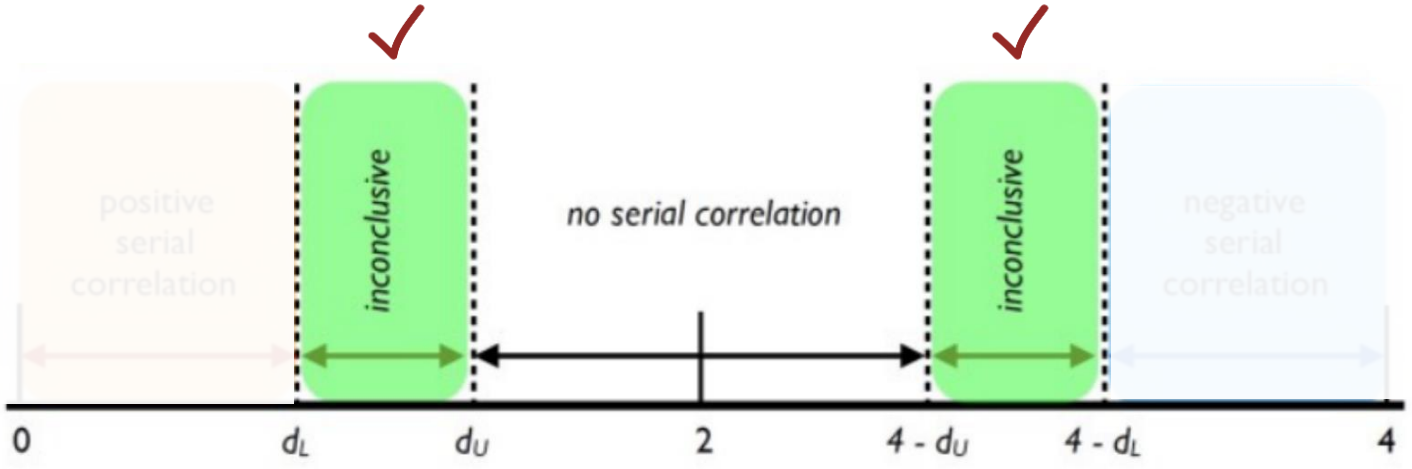
d_L : 경계의 하한, d_U : 경계의 상한

$d_L < d < (4 - d_L)$일 경우 귀무가설 기각 불가

잔차들 간 1차 자기상관 없음

5 독립성 진단과 처방

진단 | 더빈-왓슨 검정(Durbin Watson Test)



더빈-왓슨 검정의 한계

① d가 상한과 하한 사이에 위치하면 자기상관성 판단 불가

② 바로 인접한 오차와의 1차 자기상관만 고려

→ 자기상관이 오래 지속되거나, 계절성이 있는 경우 확인이 어려움

5 독립성 진단과 처방

처방

1) 설명변수 추가

자기상관을 유발하는 변수를 설명변수로 모형에 추가

2) 분석 모델 변경

✓ **시간**에 따른 자기상관

→ 자기 상관을 고려하는 AR(p) 같은 **시계열 모델** 사용

✓ **공간**에 따른 자기상관

→ 공간의 인접도를 고려하는 **공간회귀모델** 사용

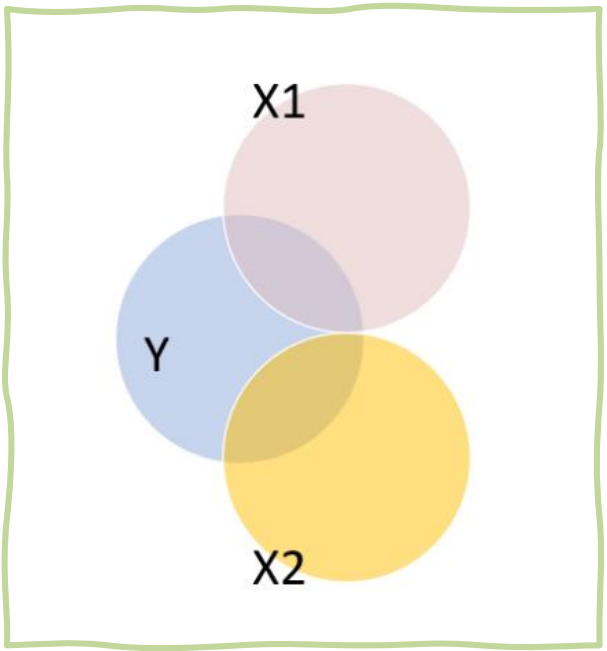
6

다중공선성

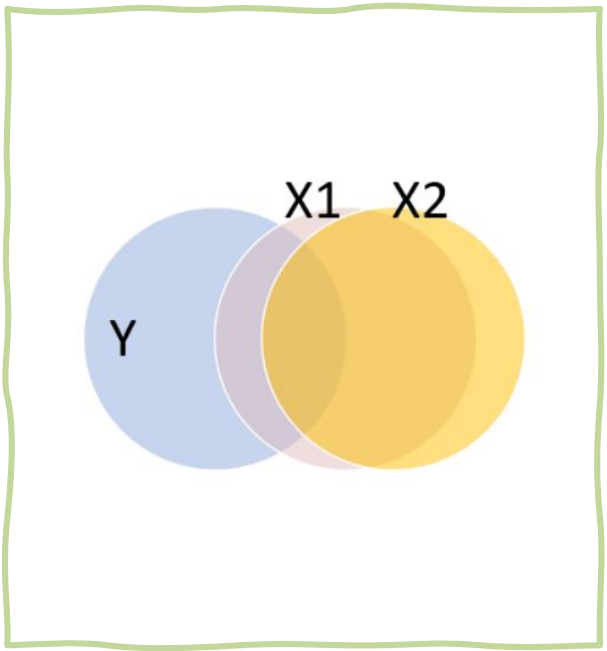
6 다중공선성

다중공선성(Multicollinearity)

설명변수 X_j들 사이에 **선형적인 상관관계**가 존재하는 것



다중공선성이 **없는** 경우



다중공선성이 **있는** 경우

6 다중공선성

다중공선성(Multicollinearity)

예시

Y : 학기별 성적

X_1 : 집에서 학교까지의 거리, X_2 : 통학 시간, X_3 : 공부 시간

⋮

X_2 변수는 $X_2 = aX_1$과 같은 식으로 X_1에 의해 완벽히 설명됨

X_2의 정보는 완전히 **필요하지 않은** 정보

6 다중공선성

다중공선성의 문제 | ① 추정량의 문제

1) OLS method를 이용한 모수 추정이 어려워진다

최소제곱법(OLS)을 통한 LSE로 적합된 모형

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'y$$

다중공선성이 존재하면?

선형종속이 되어 X는 full rank를 만족하지 않음 → $X'X$ 역시 full rank를 만족하지 않음 → $(X'X)^{-1}$ 존재하지 않아 정규방정식의 유일해 계산 불가능

다중공선성의 문제 | ① 추정량의 문제

2) 추정량을 불안정하게 만든다

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

$$(X'X)^{-1} = \frac{1}{\det(X'X)} adj(X'X)$$

이때, 다중공선성이 존재한다면 $\det(X'X) \approx 0$

추정량의 분산($Var(\hat{\beta})$) 또한 매우 커져 **계수의 추정이 불안정**해지는 문제 발생

→ Prediction Accuracy 심각하게 감소

다중공선성의 문제 | ② 모델의 문제

1) 모델의 검정 결과를 신뢰할 수 없다

다중공선성이 존재할 때 F-test는 통과하고, 적합성 검정을 위한 R^2 값도 괜찮은 수준

그러나 **유의한 개별 계수가 하나도 존재하지 않는 상황** 발생

개별 변수의 유의성 검정에서!

회귀계수들의 분산 커짐 → t 검정통계량 감소 → 귀무가설 기각 못함

다중공선성의 문제 | ② 모델의 문제

2) 모델 해석에 영향을 준다

개별 베타 계수 β_j의 해석

변수 x_j를 제외한 **나머지 변수가 고정**되어 있을 때,

x_j가 한 단계 증가했을 때의 증가량

다중공선성이 있다면?

나머지 변수가 고정되어 있다는 가정이 불가능해짐!

다중공선성의 문제 | ② 모델의 문제

2) 모델 해석에 영향을 준다

개별 베타 계수 β_j의 해석

변수 x_j를 제외한 **나머지 변수가 고정**되어 있을 때,

x_j가 한 단계 증가했을 때의 증가량

다중공선성이 있다면?

나머지 변수가 고정되어 있다는 가정이 불가능해짐!

6 다중공선성

다중공선성의 진단 | ① 직관적인 판단

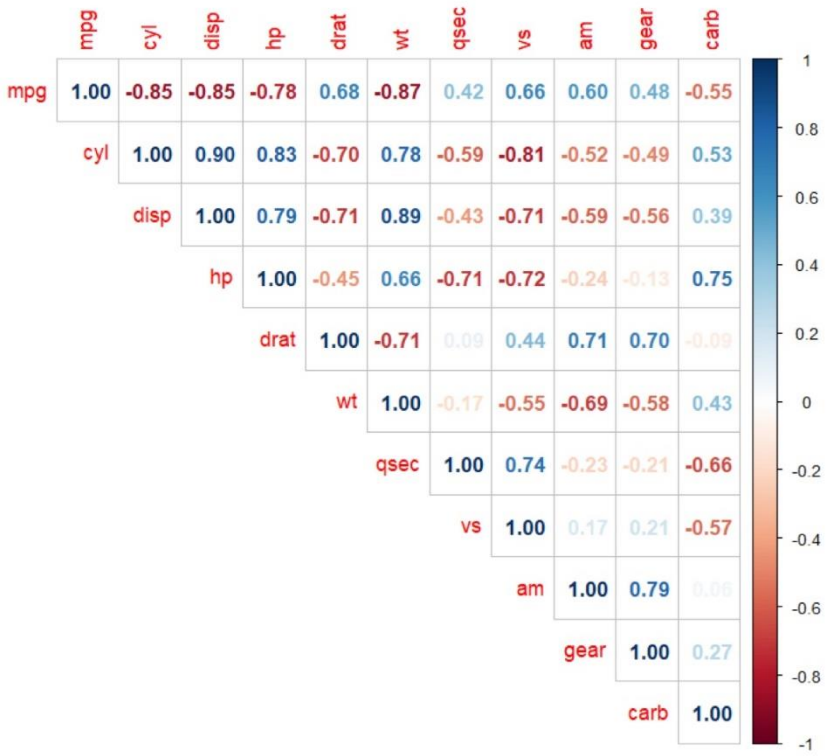
✓ F-test는 유의했지만, 개별 회귀계수들에 대한 T-test가
귀무가설을 대부분 기각하지 못하는 경우

✓ 상식적으로 유의한 회귀계수가, 유의하지 않은 경우
→ 이미 비슷한 설명변수가 모델에 포함되어 있을 것임!

✓ 추정된 회귀계수의 부호가 상식과 다를 경우
→ 이미 비슷한 설명변수가 모델에 포함되어 있을 것임!

6 다중공선성

다중공선성의 진단 | ② 상관계수 plot



✓ 변수들 사이의
선형관계 파악 가능

✓ 상관계수 **절댓값이 0.7 이상**일때
다중공선성 의심!

다중공선성의 진단 | ③ VIF (분산팽장인자)

$$VIF_j = \frac{1}{1-R_j^2} \quad , \qquad j = 1, \dots, p$$

R_j^2 : 선형회귀식 $x_j = \gamma_1 x_1 + \cdots + \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots + \gamma_p x_p$ 을 적합했을 때의 결정계수로, **나머지 변수가 x_j를 설명하는 정도**를 의미

R_j^2가 크다 → x_j가 **나머지 변수들의 선형결합**으로 표현될 수 있다 → **다중공선성** 존재

6 다중공선성

다중공선성의 진단 | ③ VIF (분산팽창인자)

$$VIF_j = \frac{1}{1-R_j^2}, \quad j = 1, \cdots, p$$

☑ 일반적으로 VIF가 10 이상일 경우

심각한 다중공선성이 존재한다고 판단

R_j^2 : 선형회귀식 $x_j = \gamma_1 x_1 + \cdots + \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots + \gamma_p x_p$ 을 적합했을 때의 결정계수로, 회귀식이 데이터를 설명하는 정도를 의미

☑ 다중공선성이 존재하지 않는다면

VIF 값 = 1

x_j가 나머지 변수들의

R_j^2가 크다 → 선형결합으로 표현될 수 있다 → 다중공선성 존재

다중공선성의 진단 | ④ $(X'X)$의 고유값 조사

설명변수들 간 선형종속 관계가 있는 경우,

X의 rank가 줄어들고 $(X'X)$의 rank도 p보다 줄어들게 됨

Rank는 **0보다 큰 고유값의 개수**로도 볼 수 있으므로

만약 **0에 가까운 고유값**이 있다면 다중공선성 의심 필요!

이 방법은 고유벡터를 통해 선형종속의 관계 형태를 대략적으로 알 수 있음

6 다중공선성

다중공선성의 해결방법

변수선택법, 차원축소, 정규화 등과 같은 방법 존재

회귀팀 3주차 클린업 예정

7

내생성과 측정 오차

7 내생성과 측정 오차

내생성(Endogeneity)

LSE가 BLUE가 될 조건

① 오차들의 평균은 0

② 오차들의 분산은 σ^2로 동일

③ 오차 간 자기 상관 X (No autocorrelation)

④ 설명변수와 오차의 상관관계 X $(Cov(X, e) = 0)$

위배될 경우, **내생성** 문제 발생

→ LSE 추정량에 attenuation bias를 일으킴

7 내생성과 측정 오차

내생성(Endogeneity)

LSE가 BLUE가 될 조건

① 오차들의 평균은 0

② 오차들의 분산은 σ^2로 동일

③ 오차 간 자기 상관 X (No autocorrelation)

④ 설명변수와 오차의 상관관계 X ($Cov(X, e) = 0$)

위배될 경우, **내생성 문제 발생**

→ LSE 추정량에 편향(bias)을 일으킴

7 내생성과 측정 오차

내생성이 발생하는 원인

내생성(Endogeneity)

설명변수와 오차 사이의 상관관계가 존재하는 것

내생성이 발생하는 원인

① 모델에 반영하지 않은 제3의 요인이 X와 Y에 영향을 미치는 경우

② X, Y가 서로 영향을 미치는 경우

③ X에 Measurement Error가 존재하는 경우

7 내생성과 측정 오차

내생성이 발생하는 원인

내생성(Endogeneity)

설명변수와 오차 사이의 상관관계가 존재하는 것

⋮

내생성이 발생하는 원인

① 모델에 반영하지 않은 제3의 요인이 X와 Y에 영향을 미치는 경우

② X, Y가 서로 영향을 미치는 경우

③ X에 Measurement Error가 존재하는 경우

7 내생성과 측정 오차

Measurement Error

내생성이 발생하는 원인

내생성(Endogeneity)

설문조사 등에서 응답자가 잘못 응답하거나, 설문 집단을 잘못 선정하여

회귀 모델에서 **변수에 대해 정확하지 않은 측정 값**을 사용할 때 발생하는 Error

Classical Error라고 가정

내생성이 발생하는 원인

Classical Error

① 모델에 반영하지 않은 제3의 요인이 X와 Y에 영향을 미치는 경우

② X, Y가 서로 영향을 미치는 경우

③ X에 Measurement Error가 존재하는 경우

X를 잘못 측정된 값, X^*를 X의 참 값, μ를 측정 오차라고 했을 때,

① $E(\mu) = 0$, ② $Cov(X^*, \mu) = 0$

를 만족하는 오차

내생성과 측정 오차

모델 설정

$$y = \beta_0 + \beta_1 X^* + e$$
$$= \beta_0 + \beta_1 (X - \mu) + e$$
$$= \beta_0 + \beta_1 X - \beta_1 \mu + e$$

$-\beta_1 \mu + e$ 를
새로운 오차항 e'로 가정

참값(X^*)은 모르는 값이므로 측정한 X 로 위의 모델을 세우게 되고,

$Cov(X, e') \neq 0$ 이면 내생성 문제 발생

내생성과 측정 오차

$Cov(X, \mu)$

$= \mathrm{E}(\mathrm{X}\mu) - \mathrm{E}(\mathrm{X})\mathrm{E}(\mu)$

$= \mathrm{E}(\mathrm{X}\mu) = \mathrm{E}[(X^* + \mu)\mu)]$

$= \mathrm{E}(X^*\mu) + \mathrm{E}(\mu\text{^}2) = \sigma_\mu^2$

$Cov(X, e')$

$Cov(X, e - \beta\mu) = E(Xe) - E(X\beta\mu)$

$= -\beta E(X\mu)$

$= -\beta Cov(X, \mu)$

$= -\beta\sigma_\mu^2 \neq 0$

$$Cov(X, e') \neq 0$$

7 내생성과 측정 오차

내생성과 측정 오차

$Cov(X, \mu)$

$= E(X\mu) - E(X)E(\mu)$

$= E(X\mu) = E[(X^* + \mu)\mu)]$

$= E(X^*\mu) + E(\mu^2) = \sigma_\mu^2$

$Cov(X, e') = Cov(X, e - \beta\mu) = E(Xe) - E(X\beta\mu)$

$= -\beta E(X\mu)$

$= -\beta Cov(X, \mu)$

$= -\beta\sigma_\mu^2 \neq 0$

X에 발생하는 Measurement Error는

Classical Error 가정 하에 **내생성 문제**를 일으킴

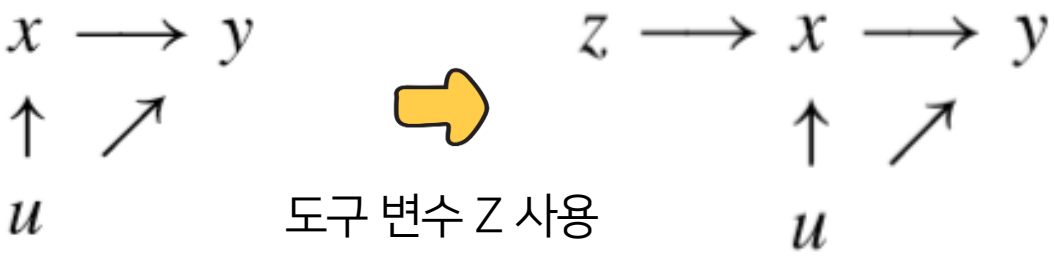
추정량이 감쇠 편향(Attenuation Biased)되는 결과

$$Cov(X, e') \neq 0$$

내생성 문제의 처방 | 도구 변수

도구 변수

X와는 상관관계가 있지만, 오차와는 상관관계가 없는 변수를 찾아 모델에 넣는 방법



제거하고자 하는 요인 u 와 관계가 없는 변수 z 를 찾아

z 와 x 로 회귀 모델을 세워 $\hat{x}$ 을 구하고, x 대신 $\hat{x}$ 을 회귀 모델에 사용

감사합니다!