회귀분석팀

6팀

김형석

김준령

윤여원

김현우

이채은

INDEX

INDEX

1

Standard Regression Assumptions

1 Standard Regression Assumptions

선형회귀모델의 기본 가정

변수에 대한 가정

선형성 (Linearity)

오차항에 대한 가정

정규성 (Normality)

등분산성 (Homoscedasticity)

독립성 (Independence)

1 Standard Regression Assumptions

변수에 대한 가정 | 선형성

선형성 (Linearity)

설명변수(X) 와 반응변수(Y) 가 서로 선형관계에 놓여있다는 가정

반응변수(Y) 의 평균 값은

회귀계수(β) 들의 Linear Combination 형태로 표현되어야 함

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 {X_i}^2 + \varepsilon_i$$

$$Y_i = \beta_0 e^{\beta_1 X_i} \leftrightarrow \ logY_i = {Y_i}^* = log\beta_0 + \beta_1 X_i$$

1 Standard Regression Assumptions

변수에 대한 가정 | 선형성

선형성 (Linearity)

설명변수(X) 와 반응변수(Y) 가 서로 선형관계에 놓여있다는 가정

반응변수(Y)의 평균 값은

선형성 가정이 위배된다면,

지금까지 배운 회귀 모형은 모두 선형회귀모델이기 때문에

모델자체가 성립하지 않게 됨

$$Y_i = \beta_0 e^{\beta_1 X_i} \leftrightarrow \log Y_i = Y_i^* = \log\beta_0 + \beta_1 X_i$$

오차항에 대한 가정 | 정규성

정규성 (Normality)

오차항이 $N(0,\ \sigma^2)$분포를 따라야 한다는 가정

정규성의 가정하에서 회귀계수에 대한
t-test, F-test 유의성 검정을 수행할 수 있음

오차항에 대한 가정 | 정규성

정규성 (Normality)

오차항이 $N(0,\ \sigma^2)$분포를 따라야 한다는 가정

정규성 가정이 위배된다면,

회귀분석 과정에서 수행한 **통계적 검정의 결과는 신뢰성을 잃게 됨**

오차항에 대한 가정 | 등분산성

등분산성 (Homoscedasticity)

오차항의 분산 값이 항상 σ^2 로 일정해야 한다는 가정

만약 오차항의 등분산성이 위배되면

아래와 같은 문제들이 발생할 수 있음

① LSE 는 더 이상 BLUE 가 아니게 됨

② 회귀계수 추정량의 분산 값이 커지게 돼 모델의 예측성능이 떨어지게 됨

오차항에 대한 가정 | 등분산성

등분산성 (Homoscedasticity)

오차항의 분산 값이 항상 σ^2 로 일정해야 한다는 가정

만약 오차항의 등분산성이 위배되면

아래와 같은 문제들이 발생할 수 있음

① LSE 는 더 이상 **BLUE 가 아니게 됨**

② 회귀계수 추정량의 분산 값이 커지게 돼 **모델의 예측성능이 떨어지게 됨**

오차항에 대한 가정 | 독립성

독립성 (Independence)

오차항들 간에 **자기상관성(Autocorrelation)** 이 없고,

서로 **독립**이어야 한다는 가정

만약 오차항의 독립성이 위배될 시 아래와 같은 문제들이 발생할 수 있음

① LSE 는 더 이상 BLUE 가 아니게 됨

② 오차항의 분산($= \sigma^2$) 및 회귀계수의 Standard Error($= se(\hat{\beta})$) 를 과소추정 하게 돼 통계적 검정의 정확도가 떨어짐

오차항에 대한 가정 | 독립성

독립성 (Independence)

오차항들 간에 **자기상관성(Autocorrelation)** 이 없고,

서로 **독립**이어야 한다는 가정

만약 오차항의 독립성이 위배될 시 아래와 같은 문제들이 발생할 수 있음

① LSE 는 더 이상 **BLUE 가 아니게 됨**

② **오차항의 분산($= \sigma^2$) 및 회귀계수의 Standard Error($= se(\hat{\beta})$) 를 과소추정** 하게 돼 통계적 검정의 정확도가 떨어짐

1 Standard Regression Assumptions

오차항에 대한 가정 | 독립성

독립성 (Independence)

오차항들 간에 자기상관성(Autocorrelation) 이 없고,

서로 **독립**이어야 한다는 가정

$$H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0\,, \qquad T_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

일반적으로 T_0의 값이 크면 귀무가설(H_0)을 기각하는데,

오차항의 독립성이 위배될 경우

만약 오차항의 독립성이 위배될 경우 아래와 같은 문제들이 발생할 수 있음

표준오차 $se(\hat{\beta})$의 값이 과소추정 되어 검정통계량 T_0값이 커짐에 따라

① LSE 는 더 이상 BLUE 가 아니게 됨

귀무가설을 더 자주 기각하게 됨

② 오차항의 분산($= \sigma^2$) 및 회귀계수의 Standard Error($= se(\hat{\beta})$) 를

과소추정 하게 돼 통계적 검정의 정확도가 떨어짐

과소추정이 문제가 되는 이유

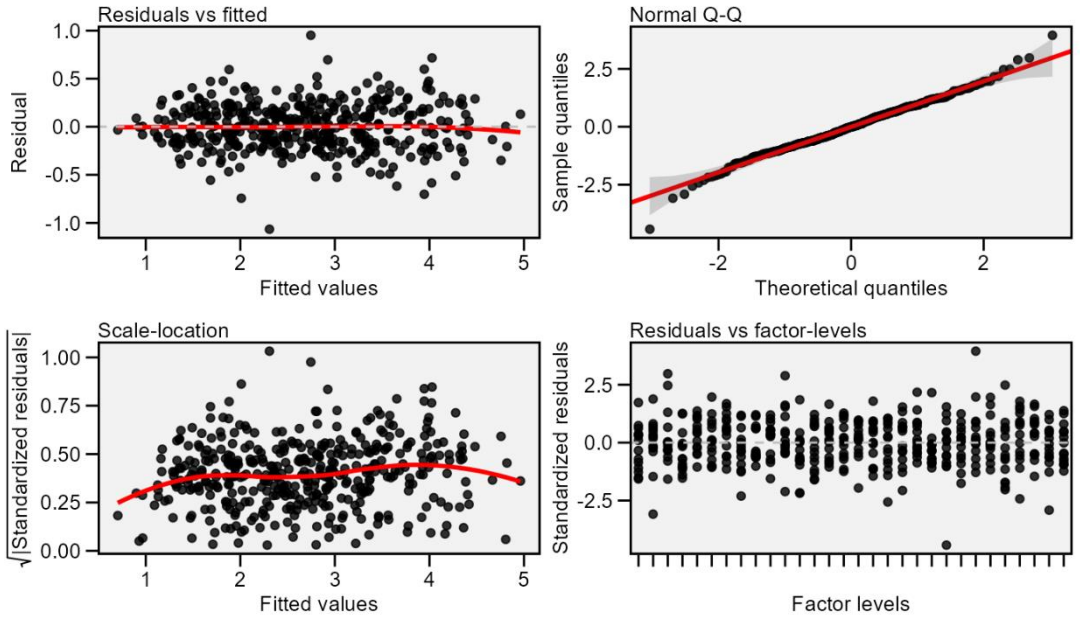
2

Residual Plot

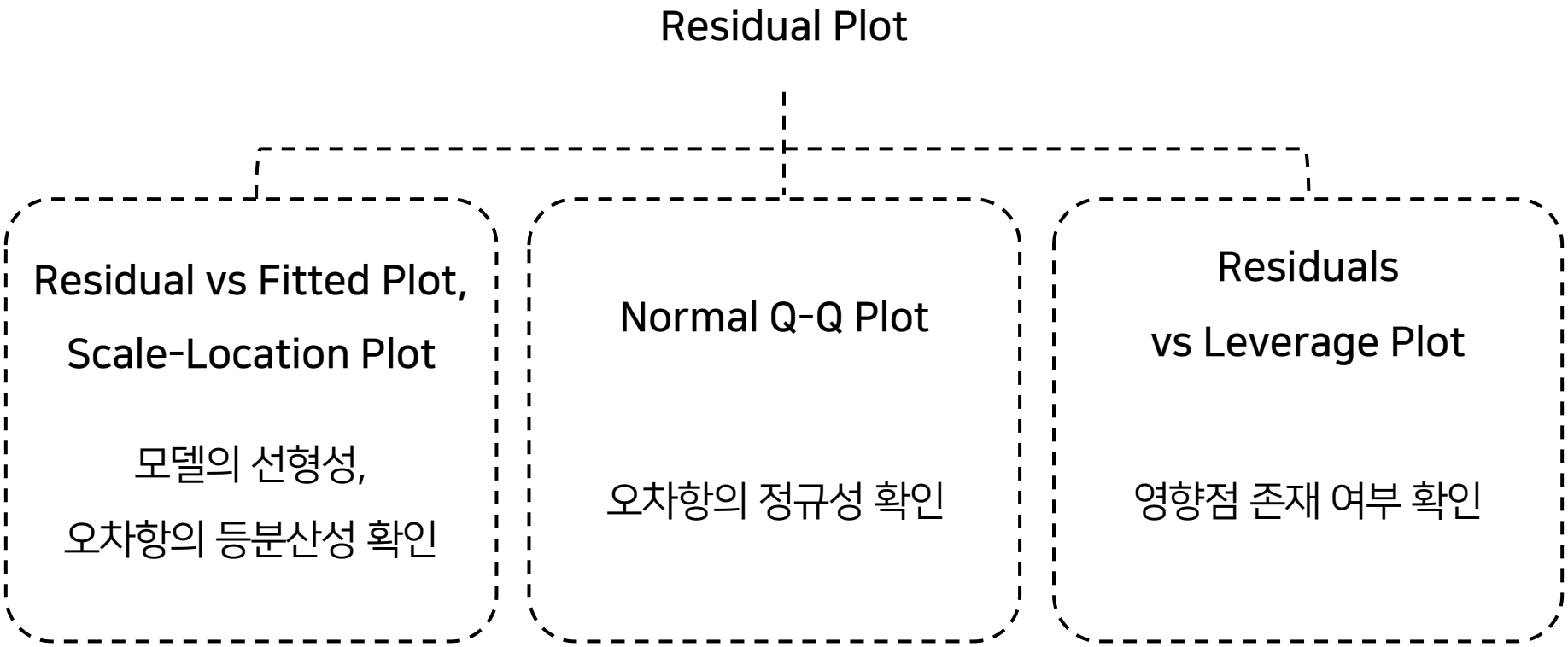
Residual Plot

잔차도 (Residual Plot)

회귀모델로부터 계산된 잔차(residual)를 이용하는 그래프들을 통칭하는 말



Residual Plot



2 Residual Plot

Residual vs Fitted Plot, Scale-Location Plot

Residual vs Fitted Plot

X 축에 Fitted Values(= $\hat{y}$) 를,

Y 축에 **잔차**를 놓고 그린 Scatter Plot

Scale-Location Plot

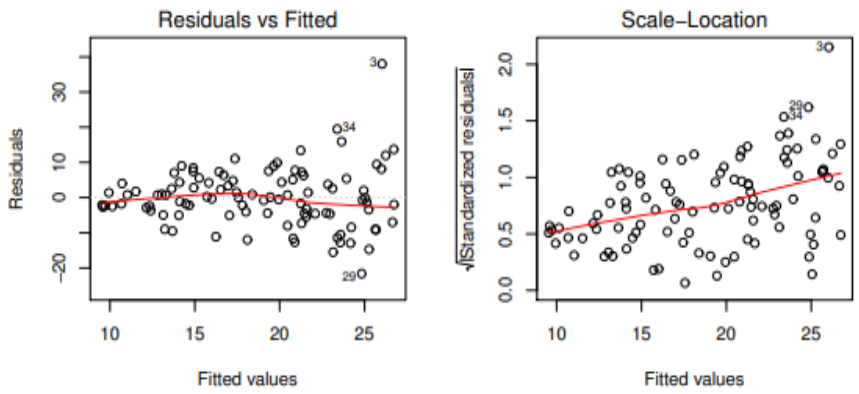
X 축에 Fitted Values(= $\hat{y}$) 를,

Y 축에 **내표준화 잔차**를 놓고 그린 Scatter Plot

2 Residual Plot

Residual vs Fitted Plot, Scale-Location Plot

설명변수와 반응변수 (Y)의 **선형성**과 **오차항의 등분산성**을 확인해볼 수 있음



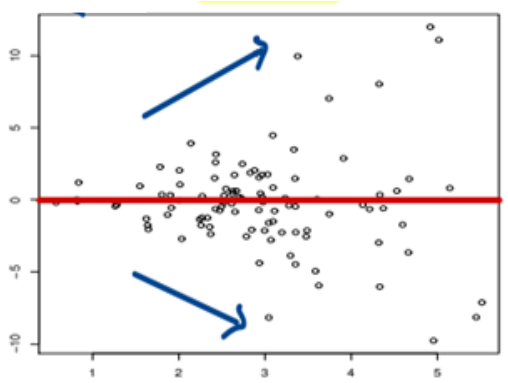
예측값과 잔차 사이의

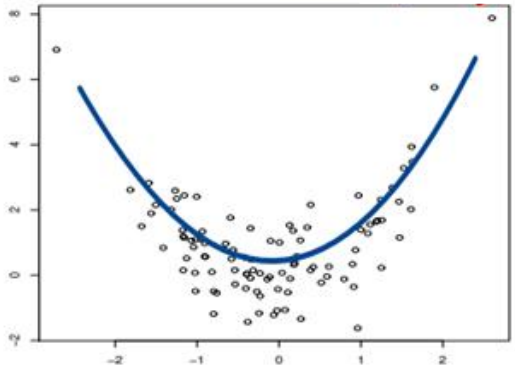
패턴이 존재하지 않고,

선형성과 등분산성 가정 만족

2 Residual Plot

Residual vs Fitted Plot, Scale-Location Plot





예측 값이 커짐에 따라 잔차들도
커지는 경향을 보이기 때문에
등분산성 가정 위배를
의심해 볼 수 있음

2차 함수 형태의 특정
패턴이 존재하기 때문에
선형성 가정 위배를
의심해 볼 수 있음

2 Residual Plot

Normal Q-Q Plot

Q-Q Plot (Quantile – Quantile plot)

두 확률분포를 서로 비교해보기 위해 그리는 Plot

오차항의 **정규분포**를 가정하기 때문에

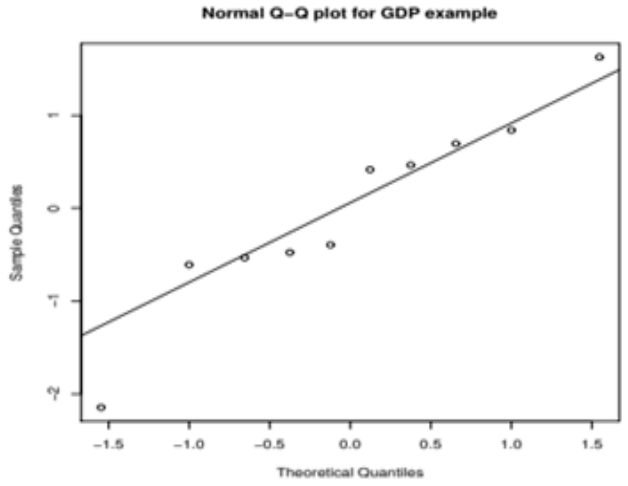
오차항의 분포와 정규분포를 서로 비교하는 Normal Q-Q Plot 을 사용

2 Residual Plot

Normal Q-Q Plot

X 축에는 표준정규분포의 분위 수,

Y 축에는 내표준화 잔차의 순서통계량을 표시함



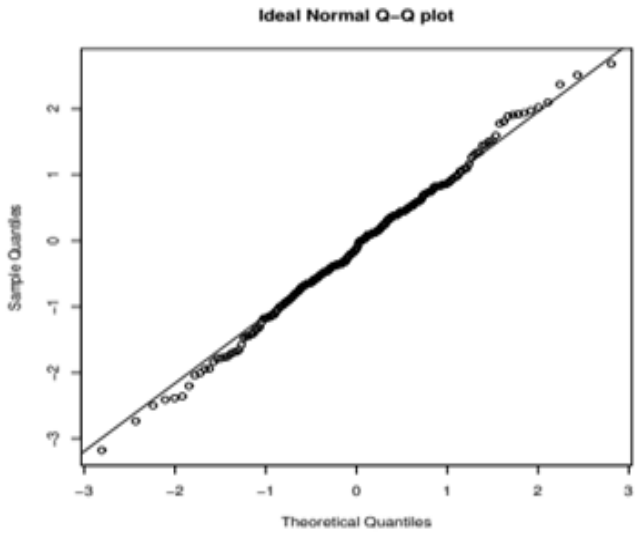
내표준화 잔차가 실제로 정규분포를 따른다면,

Normal Q-Q plot 은

$y=x$ 직선에 근사한 형태로 나타남

2 Residual Plot

Normal Q-Q Plot



좌측 Plot의 경우,

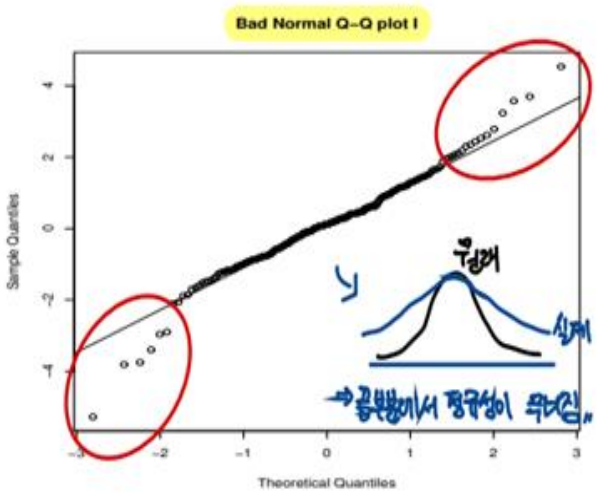
Plot 의 양상이 거의 직선 형태에 근접

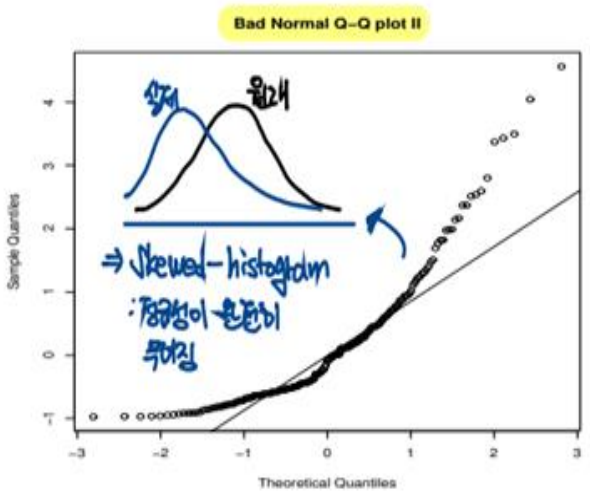
오차항의 정규성이 만족됐다고

판단할 수 있음

2 Residual Plot

Normal Q-Q Plot





반면에 위의 두 Plot 의 경우,

그래프의 형상이 직선에서 많이 벗어남

오차항의 정규성 가정 위배를 의심해볼 수 있음

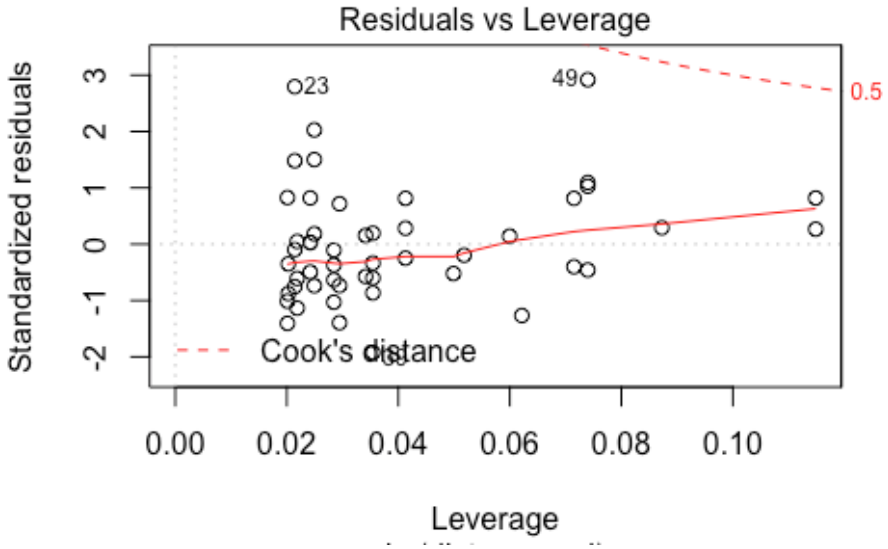
2 Residual Plot

Residuals vs Leverage Plot

Residuals vs Leverage Plot

이상치와 영향점을 확인하는 데 사용하는 진단 그래프

선형회귀모델의 기본 가정을 체크하는 게 목적이 아니라는 점에서
앞선 두 Plot 들과 차이점 존재

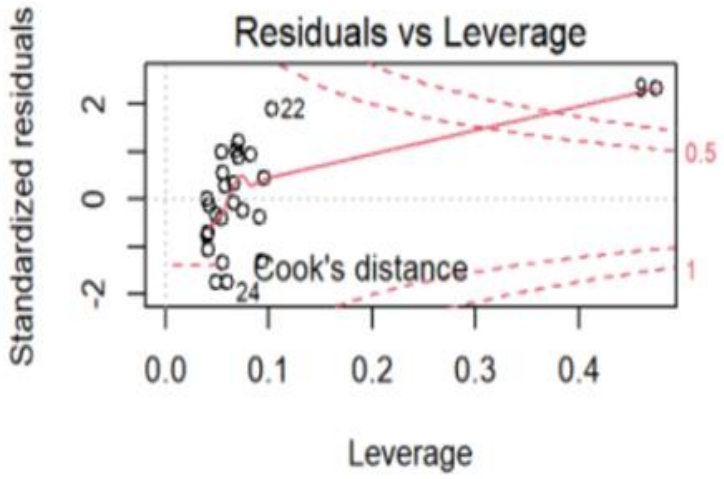


2 Residual Plot

Residuals vs Leverage Plot

X 축은 High leverage point를,

Y 축은 내표준화 잔차값(= Outlier 정도)을 표시함



빨간색 실선인

Cook's distance 경계선을 통해

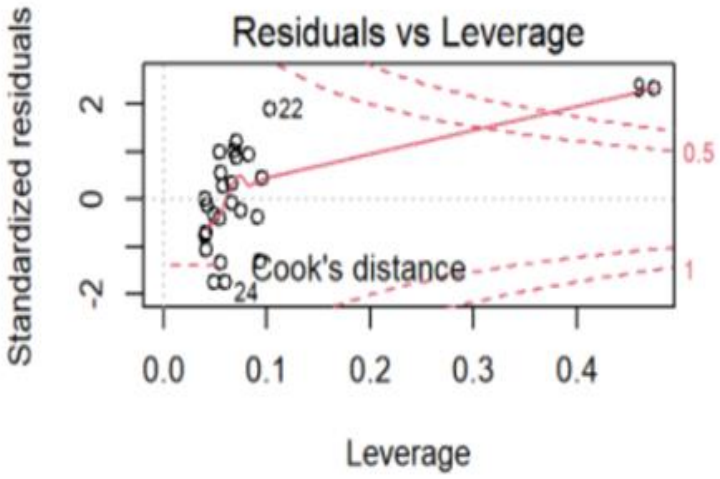
영향점의 존재 여부 확인

2 Residual Plot

Residuals vs Leverage Plot

X 축은 High leverage point를,

Y 축은 내표준화 잔차값(= Outlier 정도)을 표시함



좌측 Plot의 9번 관측치의 경우,

Cook's distance 값이 1을 넘어가기 때문에 **영향점으로 간주**할 수 있음

2 Residual Plot

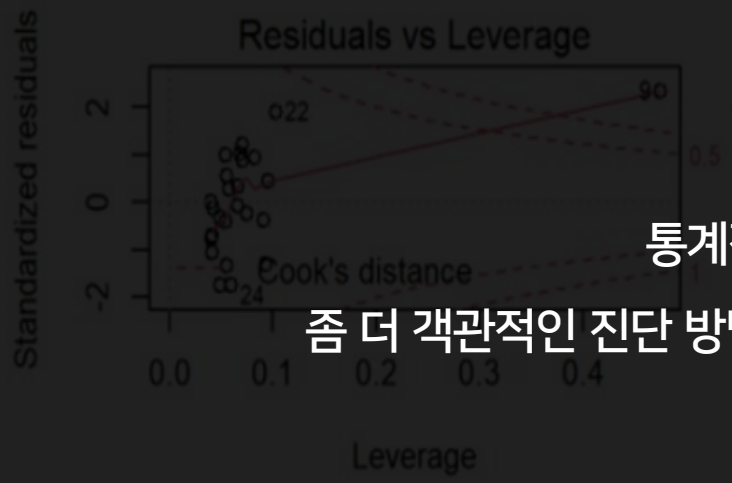
Residuals vs Leverage Plot

그래프를 통한 진단 방법의 한계

X 축은 High leverage point를,

Y 축은 내표준화 잔차값(= Outlier 정도)을 표시함

정확한 판단기준이 없어 분석자마다 상이한 결론을 도출될 수 있음



좌측 Plot의 9번 관측치의 경우,

Cook's distance 값이 1을 넘어가기 때문에 영향점으로 판단할 수 있음

통계적 검정 등

좀 더 객관적인 진단 방법을 적용해볼 필요성이 존재

3

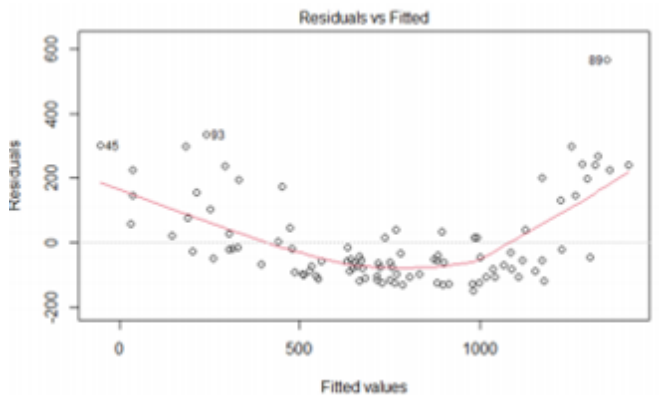
Linearity

3 Linearity

선형성의 진단

Residual vs Fitted Plot의 경우, 그래프의 추세 선이

x **축에 평행한 직선 형태**인지를 확인하는 과정을 통해 선형성을 판단함



① 판단이 다소 **주관적**일 수 있음

② **어떤 변수의 영향** 때문에 선형성이 위배됐는지 확인하기 어려움

3 Linearity

선형성의 진단

Residual vs Fitted Plot의 경우, 그래프의 추세 선이

x축에 평행한 직선 형태인지를 확인하는 과정을 통해 선형성을 판단함

Partial Residual Plot 을 통해

개별 설명 변수와 반응 변수 간의 선형성을 판단해보며

어떤 변수의 영향 때문에 선형성이 위배됐는지 확인 수 있음

② 어떤 변수의 영향 때문에 선형성이 위배됐는지 확인하기 어려움

3 Linearity

선형성의 진단 | Partial Residual Plot

부분 잔차도 (Partial Residual Plot)

특정 독립 변수가 종속 변수에 미치는 영향을

시각적으로 평가하기 위해 사용하는 Plot

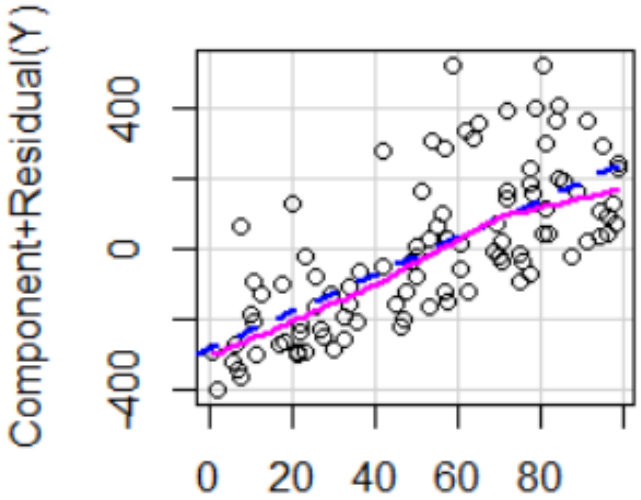
X 축은 단일 설명변수($\boldsymbol{X_{ki}}$) 를,

Y 축은 Partial Residual을 나타냄

Partial Residual = 회귀 모델의 잔차 값(e_i) + 단일 설명변수 기반 예측 값($\hat{\beta}_k X_{ki}$)

3 Linearity

선형성의 진단 | Partial Residual Plot



파란색 점선

: 단순선형회귀를 Fitting 시킨 결과

분홍색 실선

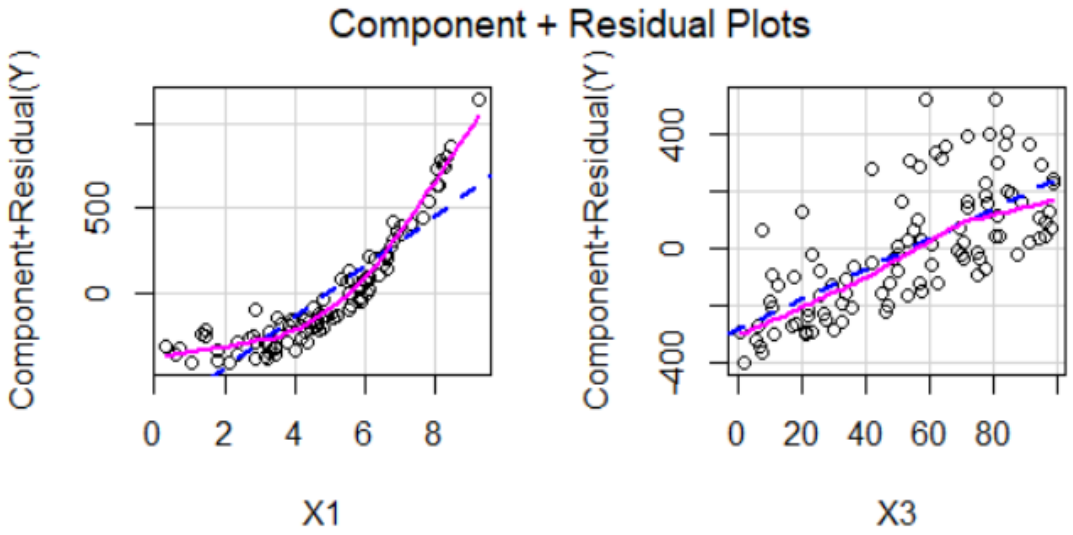
: Local Regression 을 Fitting 시킨 결과

두 가지 선의 양상이 서로 비슷하면,

해당 설명변수와 반응변수 간의 선형성은 만족됐다고 판단할 수 있음

3 Linearity

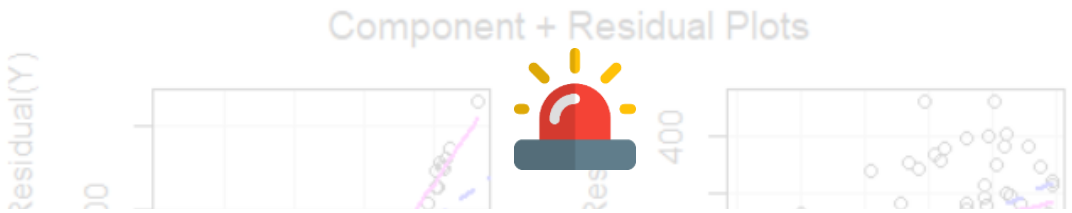
선형성의 진단 | Partial Residual Plot



따라서 위 Plot 의 결과를 해석하면

선형성이 위배되는 데 있어 X_1 의 영향이 컸을 거라고 판단할 수 있음

선형성의 진단 | Partial Residual Plot



그러나 설명변수들 간에 **교호작용**이나 **다중공선성**이 존재하는 경우

Partial Residual Plot 은 잘못된 정보를 제공할 수 있다는 한계점이 존재

설명변수들 간의 관계성을 반드시 먼저 확인해야 함

선형성이 위배되는 데 있어 X_3 의 영향이 컸을 거라고 판단할 수 있음

3 Linearity

선형성 위배 시 처방 | 변수 변환

변수 변환 (Transformation)

설명변수($\boldsymbol{X}$) 에 적절히 치환을 적용해서 비선형성 문제를 해결해주는 방법

변수변환을 통해 선형성을 확보할 수 있는 모델도,

넓은 의미에서 선형모델로 여김

Ex) log 변환

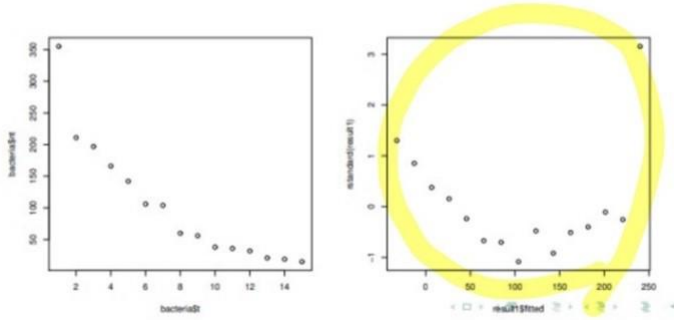
Linearizable Function : $y = \beta_0 x^{\beta_1}$

log 변환 후 Linear Form : $\log y = \log \beta_0 + \beta_1 x' \quad (x' = \log x)$

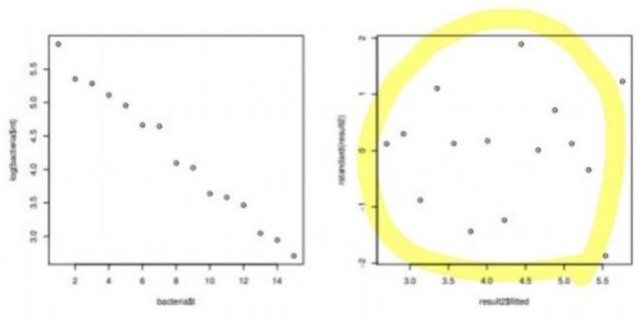
선형성 위배 시 처방 | 변수 변환

설명변수(X)-반응변수(Y) Scatter Plot 및 Residuals vs Fitted Plot

① 기존의 데이터



② log 변환을 취한 데이터



①과 같이 선형성 가정이 위배되는 경우,

②와 같이 변수 변환을 통해서 어느 정도 해결할 수 있음

선형성 위배 시 처방 | 다항 회귀

다항회귀 (Polynomial Regression)

설명변수(X)가 2차, 3차 등의 k차 다항식 형태로 표현된 회귀 모델

⋮

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 {X_i}^2 + \beta_3 {X_i}^3 + \cdots + \beta_k {X_i}^k$$

① 일반적으로 단순선형회귀 모델에 존재하는 설명변수에 대해,
거듭제곱 항들을 설명변수로 추가해주는 방식으로 모델을 작성

② 다중선형회귀에서의 분석 방법을 그대로 적용해볼 수 있음

3 Linearity

선형성 위배 시 처방 | 다항 회귀

다항회귀 (Polynomial Regressio

① 하나의 설명변수에 대해 거듭제곱 된 형태이기 때문에, 설명변수들 간의 상관관계 정도가 높아 **다중공선성** 문제가 발생할 수 있음

② 너무 높은 차수(k) 형태를 적용 시, 모델이 **Overfitting** 될 수 있음

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \cdots + \beta_k X_i^k$$

적절한 차수 값을 선택하고,

변수 생성 시 **설명변수들끼리 서로 직교하도록** 만들어주는 과정이 필요함

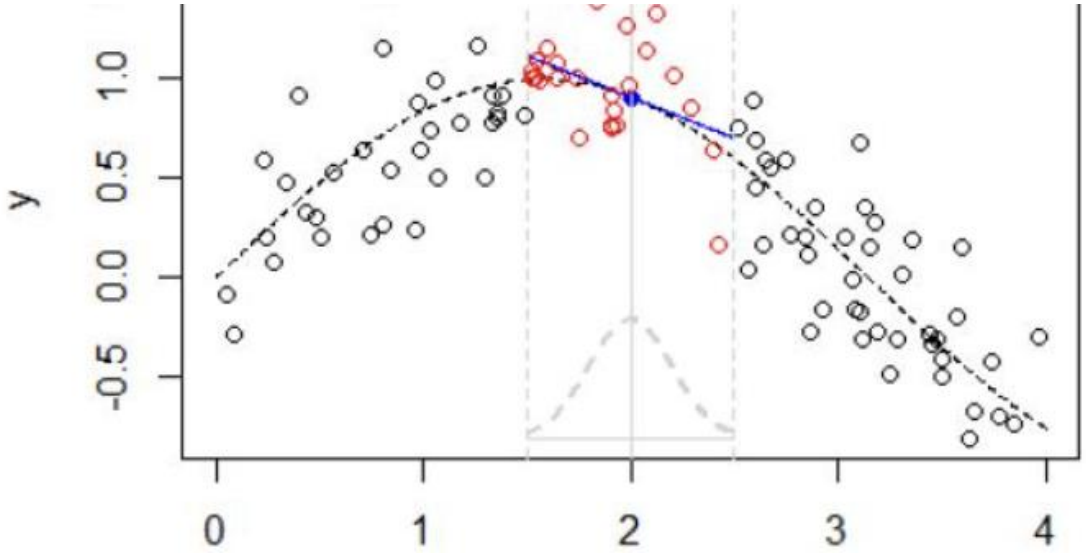
② 다중선형회귀에서의 분석 방법을 그대로 적용해볼 수 있음

선형성 위배 시 처방 | 비모수회귀

비모수회귀 (Non-parametric Regression)

모델의 특정한 형태나 분포에 대해 가정하지 않고,

데이터를 통해 모형을 직접 추정하는 방법



선형성 위배 시 처방 | 비모수회귀

모수적 방법

고정된 개수의 파라미터들을 학습

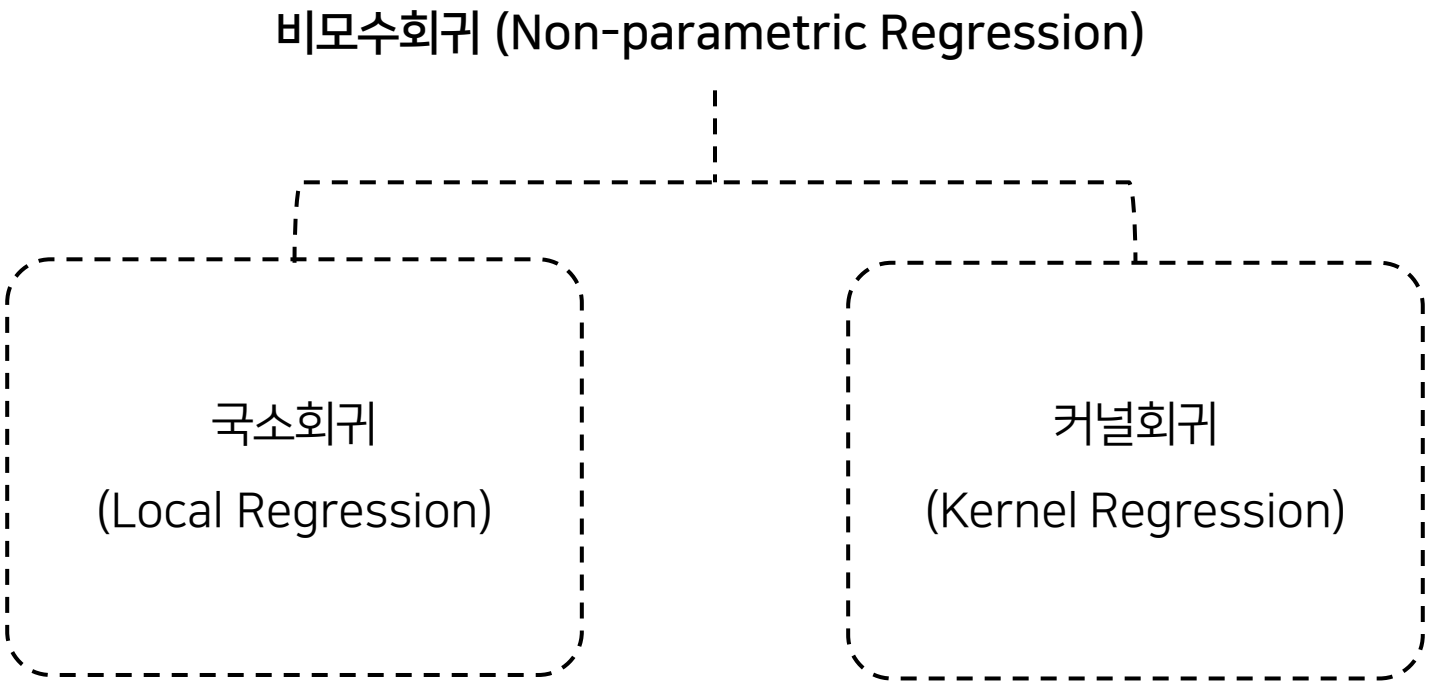
분포적 가정을 필요로 함

비모수적 방법

학습 데이터가 늘어남에 따라
파라미터의 개수도 늘어남

분포적 가정을 필요로 하지 않음

선형성 위배 시 처방 | 비모수회귀



3 Linearity

선형성 위배 시 처방 | 비모수회귀

국소회귀 (Local Regression)

데이터의 개수가 많은 반면, 차원 수는 매우 낮은 경우에 사용되는 모델

구체적으로 데이터 셋의 Feature 수가 2~3 개 이하인 경우에만 사용됨

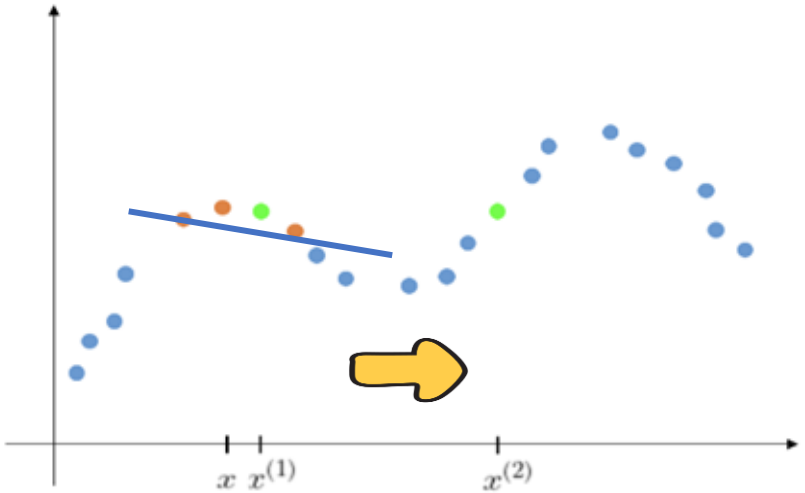
손실함수 : $J(\beta) = \sum_{i=1}^{n} w^{(i)}(y^{(i)} - \beta^T x^{(i)})^2$, $w^{(i)} = exp\left(-\frac{(x^{(i)} - x)^2}{2}\right)$

손실함수 특징 : Target x 에 가까운 관측치일수록 큰 가중치를 부여 받음

결과적으로 x 에 **가까운 관측치들을 위주**로 회귀 직선을 fitting 하게 됨

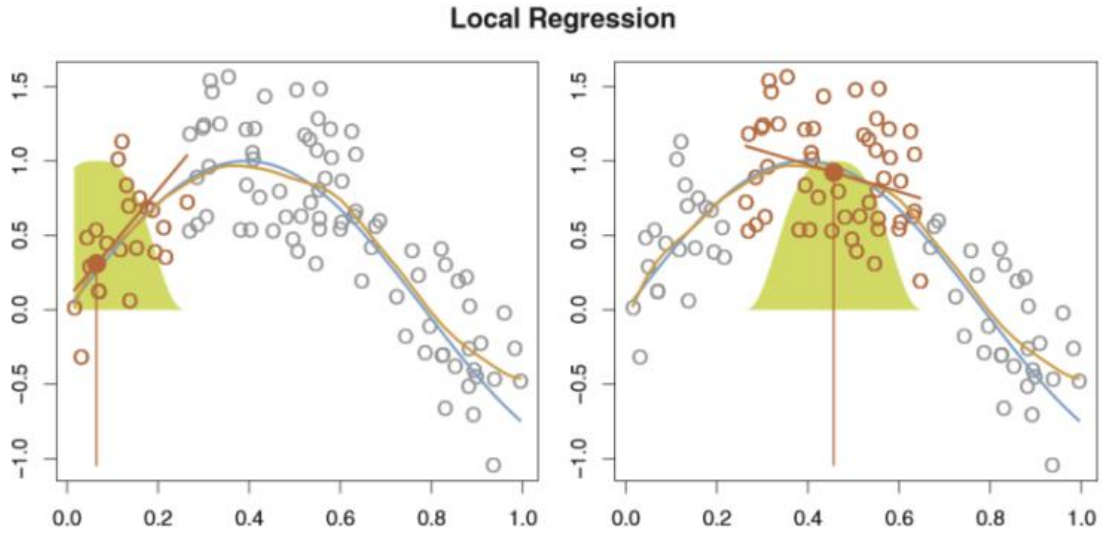
3 Linearity

선형성 위배 시 처방 | 비모수회귀



x 주변에 있는 데이터로만 회귀 직선을 fitting하는 과정을 반복 수행하면,
최종적인 추정선은 데이터의 **비선형 추세를 따라가는 형태**로 출력됨

선형성 위배 시 처방 | 비모수회귀

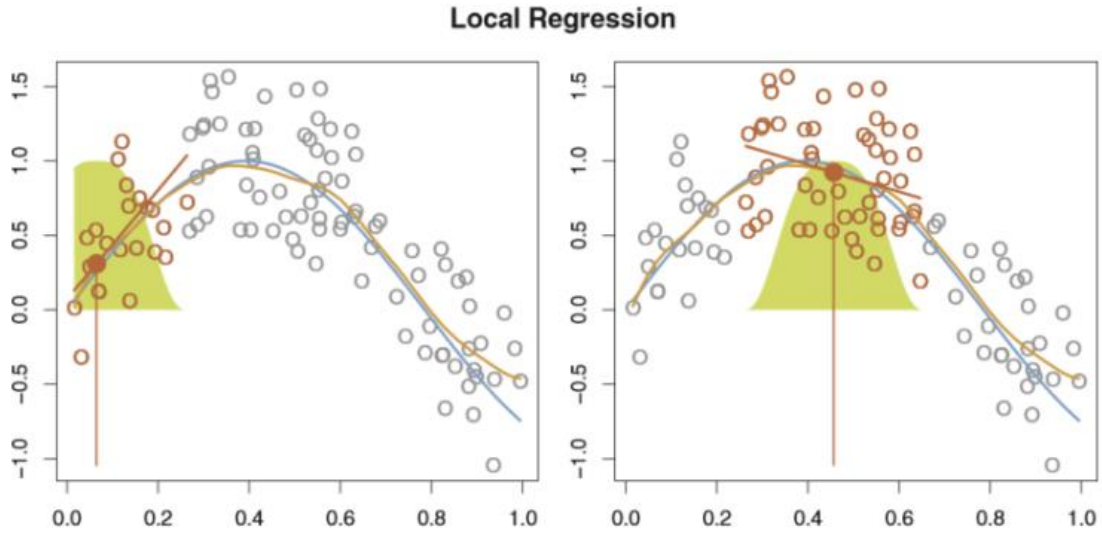


파란선 : 데이터를 생성한 실제 함수

주황선 : 국소회귀 모델로 추정된 선

초록색 종 모양의 면적에서 각 높이 : Target 관측치의 가중치 값(= $w^{(i)}$)

선형성 위배 시 처방 | 비모수회귀



파란선 : 데이터를 생성한 실제 함수

Local 범위에서의 빨간 선들이 결합되어 만들어진 주황색의 곡선이

데이터의 비선형 추세를 잘 나타냄을 확인할 수 있음

선형성 위배 시 처방 | 비모수회귀

가중치함수의 경우 **Target에 가까운 관측치일수록**

더 큰 가중치를 부여한다는 특성만 지닌다면 여러 함수들을 적용해볼 수 있음

Ex) Gaussian Kernel

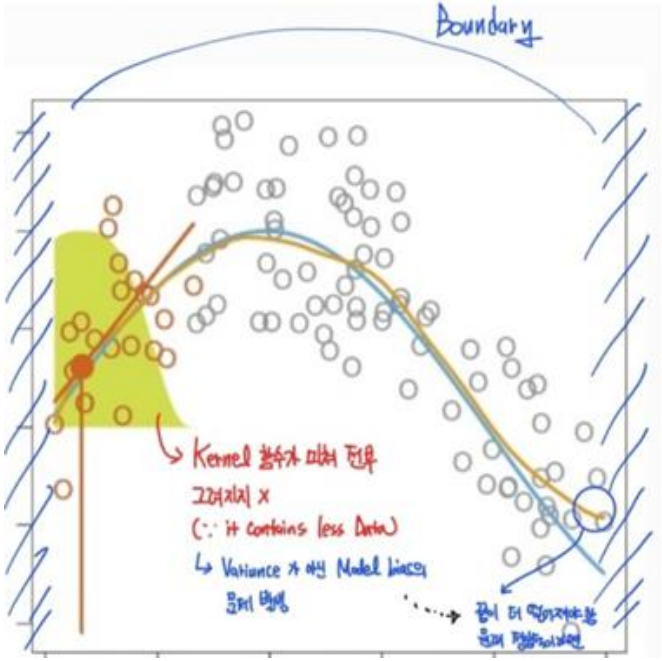
$$w^{(i)} = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x - x^{(i)}}{\lambda})^2]$$

대역폭 매개변수(Bandwidth parameter) λ 를 통해

관측치로부터 얼마나 가까운 데이터들만 이용할지 결정할 수 있음

3 Linearity

선형성 위배 시 처방 | 비모수회귀



Boundary Problem

경계선에 위치한 target 관측치에 대하여,

비교적 적은 양의 데이터로 인해

가중치의 불균형이 발생하는 문제

모델의 **Bias**에 영향을 미쳐, **예측 성능** 하락

(일차 회귀가 아닌 좀 더 Complex 한 모형을 Fitting 시키면

해당 문제를 어느 정도 해결 가능)

선형성 위배 시 처방 | 비모수회귀

커널회귀 (Kernel Regression)

커널 함수를 이용해 밀도함수를 추정한 뒤

해당 함수로부터 계산된 **평균값을 이용해 반응변수 값을 추정**하는 회귀 모델

국소회귀와는 달리 Feature 수가 많은 데이터 셋에서도 사용이 가능

커널 밀도함수의 추정 방식에 따라 크게 세 가지 종류로 분류 됨

① Nadaraya-Watson ② Priestley-chao ③ Gasse-Muller

선형성 위배 시 처방 | 비모수회귀

Nadaraya-Watson 회귀모델

추정식 : $Y = m(X) + \varepsilon,\ \ E(\varepsilon) = 0 \rightarrow\ \hat{Y} = \hat{m}(X)$

where $\hat{m}(X) = \frac{\sum_{i=1}^{n} K_\lambda(x-x_i)y_i}{\sum_{j=1}^{n} K_\lambda(x-x_j)}$ $(K_\lambda$: Kernel Function)

국소회귀와 유사하게,

Target 관측치 값마다 커널 함수를 적용해 예측 값을

계산하기 때문에 Boundary Problem 이 발생할 수 있음

선형성 위배 시 처방 | 비모수회귀

Nadaraya-Watson 회귀모델

커널함수(K_λ)는 일반적으로 Gaussian Kernel 함수를 많이 사용함

Gaussian Kernel 함수 : $w^{(i)} = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-x^{(i)}}{\lambda})^2]$

앞서 국소회귀에서 언급했던 Gaussian Kernel 함수식을 통해,

관측치가 target 에 가까워질수록 함수 값이 커지는 것을 알 수 있음

계산하기 때문에 Boundary Problem 이 발생할 수 있음

선형성 위배 시 처방 | 비모수회귀

Nadaraya-Watson 회귀모델

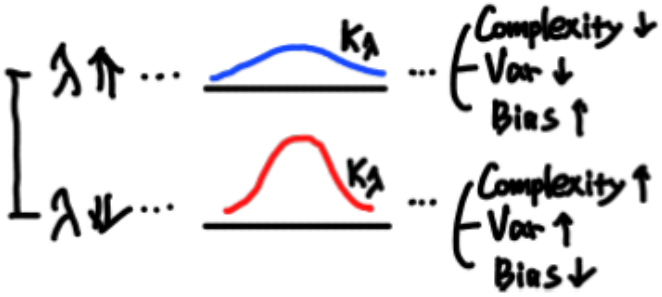
추정식 : $Y = m(X) + \varepsilon,\ \ E(\varepsilon) = 0 \rightarrow\ \hat{Y} = \hat{m}(X)$

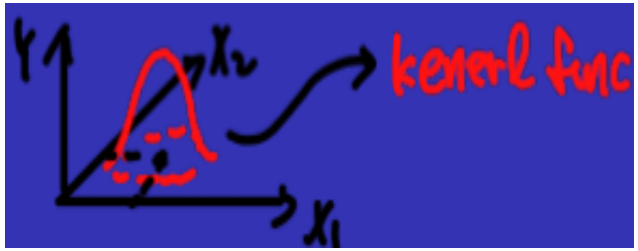
그렇다면 커널함수(K_λ) 에는 어떤 특징이 있을까?

커널함수의 특징에 대해 알아보자!

Target 관측치 값마다 커널 함수를 적용해 예측 값을
계산하기 때문에 Boundary Problem 이 발생할 수 있음

선형성 위배 시 처방 | 비모수회귀



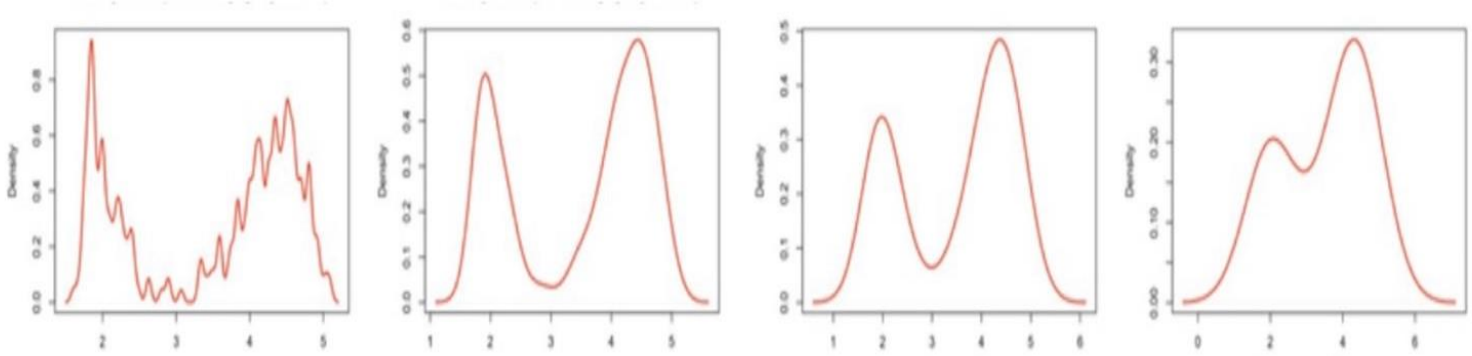


커널 함수의 대역폭(Bandwidth)
λ의 값이 커질수록
Complexity 가 감소하고,
커널 회귀 추정선은 Smooth한 형태가 됨

고차원 데이터 셋을 이용할 경우,
Gaussian Kernel 함수는 Multivariate
Normal 분포 식 형태로 취하게 됨

3 Linearity

선형성 위배 시 처방 | 비모수회귀

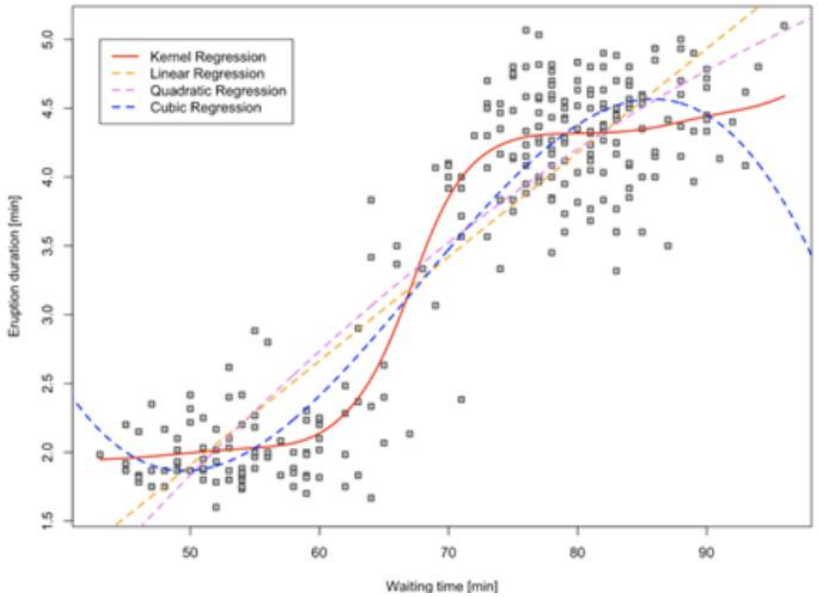


Bandwidth(=λ) 값이 커질수록 커널 함수의 Complexity 가 감소해

커널 밀도 함수의 Complexity 역시 점점 감소됨을 확인 가능

3 Linearity

선형성 위배 시 처방 | 비모수회귀



단순선형회귀나 다항회귀보다

데이터의 비선형 추세를

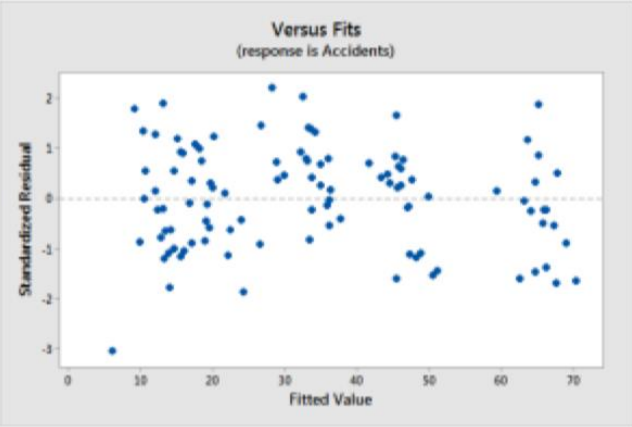
잘 따른 것 역시 확인 가능

4

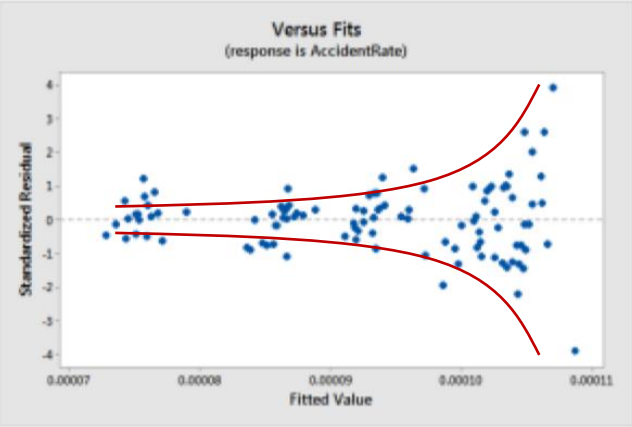
Homoscedasticity

4 Homoscedasticity

등분산성의 진단 | 잔차플롯 확인



등분산성을 만족하는 잔차플롯



등분산성을 만족하지 않는 잔차플롯

잔차플롯을 통해 산점도의 퍼진 정도가 일정한지 확인하여

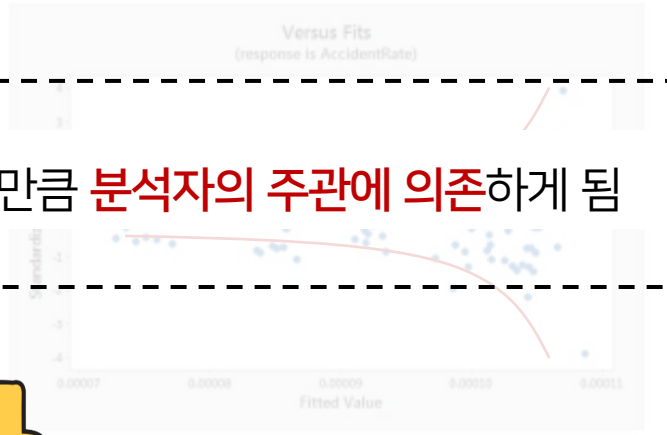
등분산성을 만족하는지 확인할 수 있음

4 Homoscedasticity

등분산성의 진단 | 잔차플롯 확인

시각적으로 등분산성을 진단하는 만큼 **분석자의 주관에 의존**하게 됨

등분산성을 만족하는 잔차플롯



등분산성을 만족하지 않는 잔차플롯

통계적 검정 **BP Test**를 통해 정량적으로 등분산성을 진단해보자!

등분산성을 만족하는지 확인할 수 있음

4 Homoscedasticity

등분산성의 진단 | BP Test

BP Test

추출한 **잔차의 제곱을 반응변수로** 두고 다시 선형 회귀모델을 적합하여

얻은 **결정계수(R^2) 를 통해 등분산성을 확인**하는 통계적 검정

BP Test의 원리

① 오차의 평균을 0으로 가정하므로

오차 제곱의 평균은 오차의 분산과 같음

$$Var(\varepsilon) = E(\varepsilon^2) - E^2(\varepsilon) = E(\varepsilon^2)$$

② 잔차 제곱을 반응변수로 두고 다음과 같은 선형 회귀모델을 적합

$$e_i{}^2 = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \cdots + \gamma_p X_{pi} \; (i = 1, \ldots, n)$$

4 Homoscedasticity

등분산성의 진단 | BP Test

BP Test

추출한 **잔차의 제곱을 반응변수로** 두고 다시 선형 회귀모델을 적합하여

얻은 **결정계수(R^2) 를 통해 등분산성을 확인**하는 통계적 검정

BP Test의 원리

① 오차의 평균을 0으로 가정하므로

오차 제곱의 평균은 오차의 분산과 같음

$$Var(\varepsilon) = E(\varepsilon^2) - E^2(\varepsilon) = E(\varepsilon^2)$$

② 잔차 제곱을 반응변수로 두고 다음과 같은 선형 회귀모델을 적합

$$e_i{}^2 = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \cdots + \gamma_p X_{pi} \ (i = 1, \ldots, n)$$

4 Homoscedasticity

등분산성의 진단 | BP Test

$$e_i{}^2 = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \cdots + \gamma_p X_{pi} \; (i = 1, \ldots, n)$$

모형에서 만약 $\boldsymbol{\gamma_i}$ 계수가 0이 아니라면

오차의 분산이 설명변수에 따라 증가하거나 감소한다고

판단할 수 있음

4 Homoscedasticity

등분산성의 진단 | BP Test

Null Hypothesis vs Alternative Hypothesis

$$H_0:\ \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0\ \ vs\ \ H_1:\ Not\ H_0$$

⋮

잔차는 등분산성을 따라야 하므로, **귀무가설이 기각되지 않기를 원함!**

Test Statistic

$$\chi^2{}_{stat} = n\,R^2 \sim \chi^2{}_{p-1}$$

$$Reject\ H_0\ \ if\ \chi^2{}_{stat} > \chi^2{}_{p-1,\alpha}\ \ (Significance\ Level:\ \alpha)$$

4 Homoscedasticity

등분산성의 진단 | BP Test

Test Statistic

$$\chi^2_{stat} = n R^2 \sim \chi^2_{p-1}$$

↓

잔차는 결정계수가 **클수록 이분산성**을,

작을수록 등분산성을 띤다고 볼 수 있음

4 Homoscedasticity

등분산성의 진단 | BP Test

정밀한 BP Test 결과를 위해서는 다음 가정들을 충족해야 함!

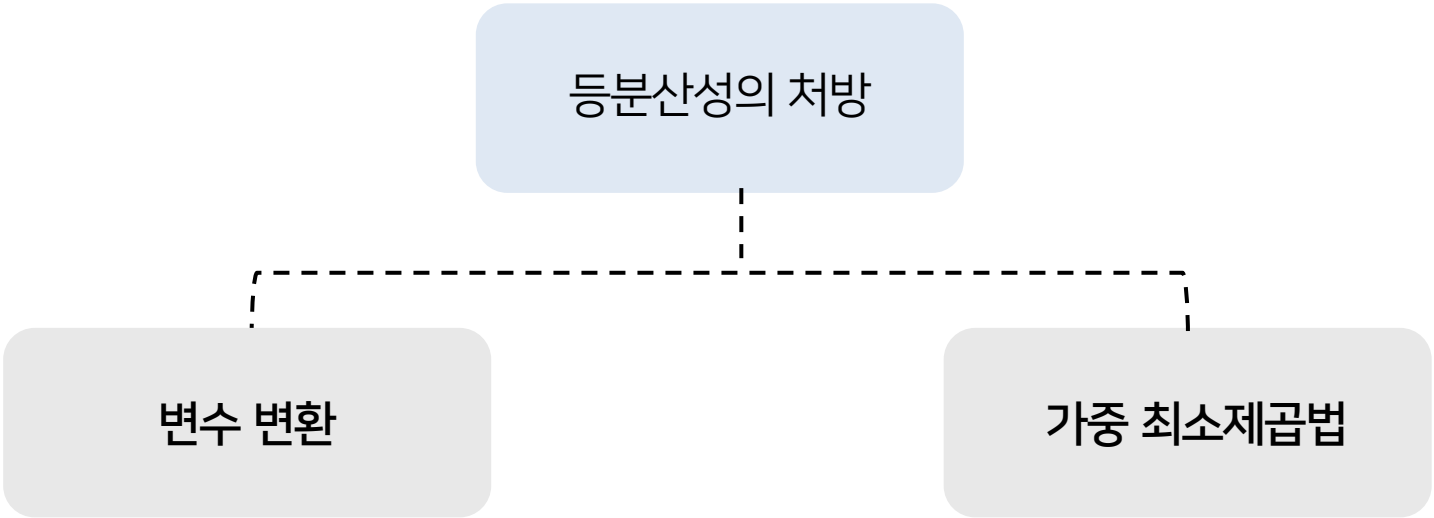
① 표본 수가 많을 것

검정통계량이 카이제곱 분포를 따르기 위한 가정

② 오차항은 서로 독립이며, 정규분포를 따를 것

③ 오차항의 분산은 설명변수와 연관이 있을 것

등분산성의 처방



등분산성의 처방 | 변수 변환

변수 변환

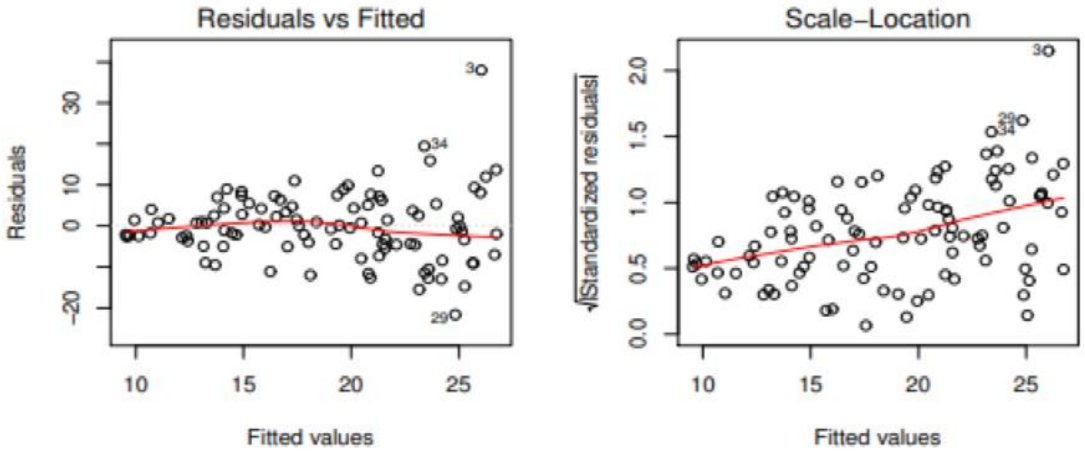
반응변수(Y) 에 적절한 변환을 적용해 이분산성 문제를 해결하려는 방법으로,

로그 변환 ($\log Y$), 제곱근 변환 ($\sqrt{Y}$) 등 변환을 시도

루피 얼굴 변환

4 Homoscedasticity

등분산성의 처방 | 변수 변환

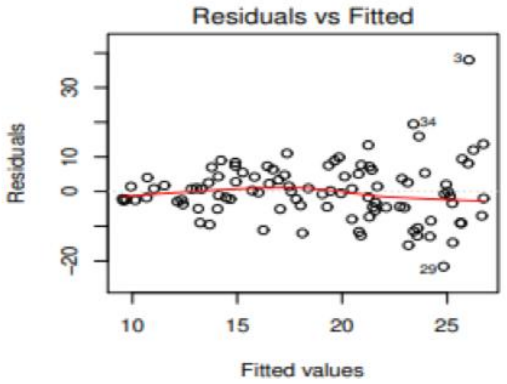


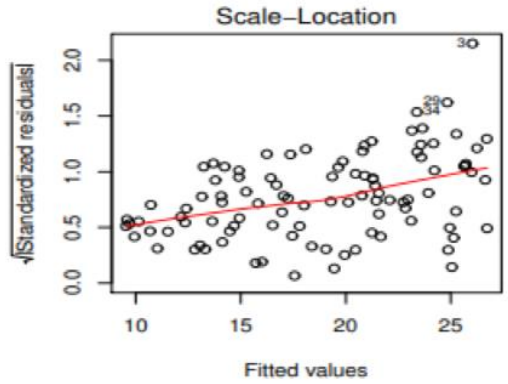
두 종류의 잔차 플롯에서 이분산성을 보이는 것을 확인

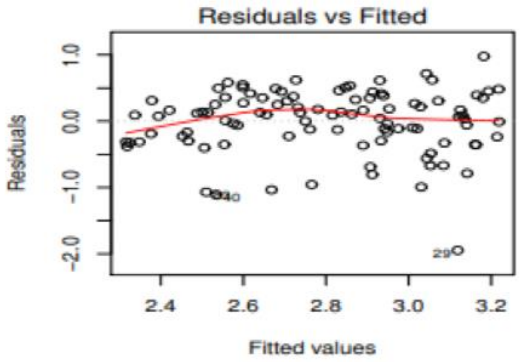
반응변수 Y에 대해 **로그변환**을 시도

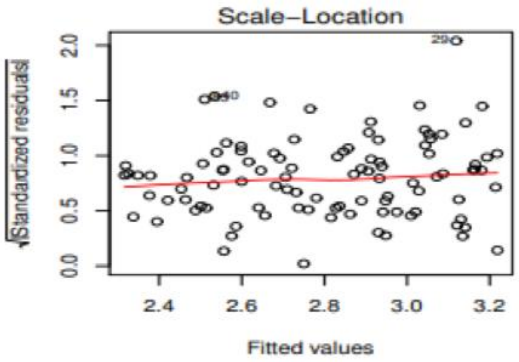
4 Homoscedasticity

등분산성의 처방 | 변수 변환



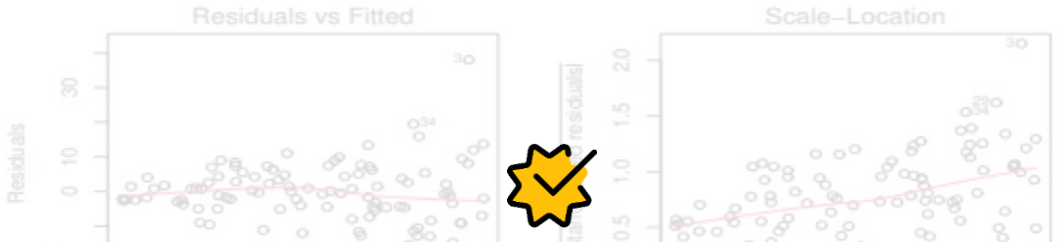




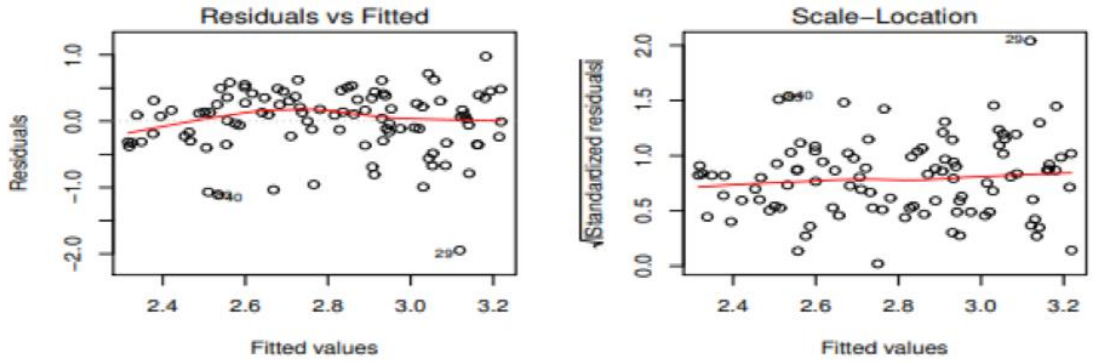


4 Homoscedasticity

등분산성의 처방 | 변수 변환



앞서 확인한 이분산성이 상당 부분 해소된 것을 확인!

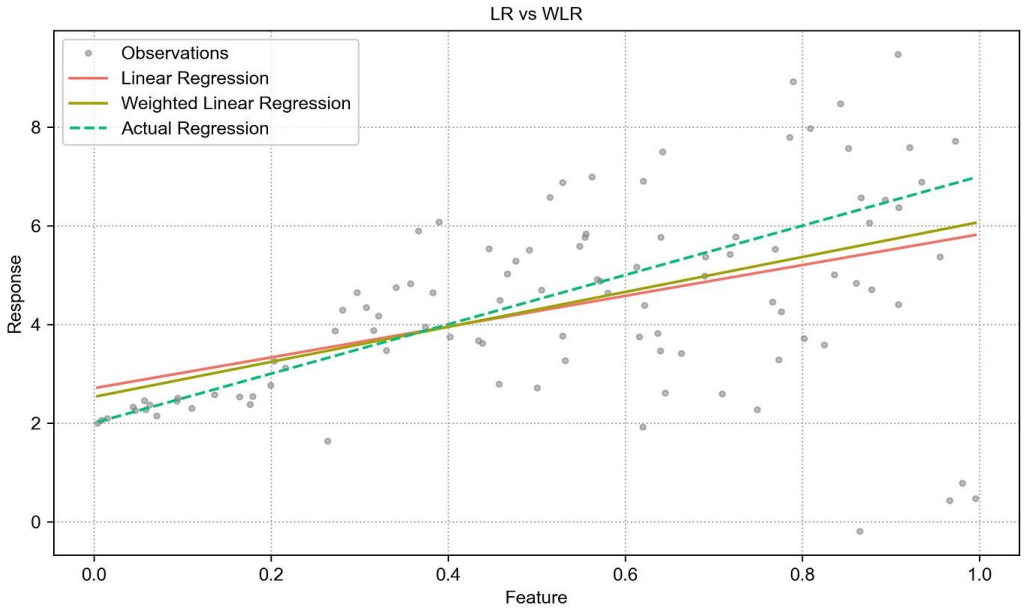


등분산성의 처방 | 가중 최소제곱법

가중 최소제곱법 (WLS, Weigted Least Squares)

가중치를 이용해 오차항의 분산을 이분산 형태로 설정해준 뒤,

가중오차의 제곱합을 최소로 하는 방식으로 회귀계수를 추정하는 방법



등분산성의 처방 | 가중 최소제곱법

$$Y = X\beta + \varepsilon,\ \varepsilon \sim N(0,\ \sigma^2 W^{-1})\ \ where\ W = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

$$Var(Y_i) = \frac{\sigma^2}{w_i} \Leftrightarrow Var(\sqrt{w_i}\, Y_i) = \sigma^2$$

위 관계를 고려하여 반응변수 $\sqrt{w_i}\, Y_i$ 에 대하여 회귀모델을 다시 설정

⋮

$$W^{1/2}\, Y = W^{1/2}\, X\beta + \varepsilon^*,\ \varepsilon^* \sim N(0,\ \sigma^2)$$

등분산성의 처방 | 가중 최소제곱법

$$Y = X\beta + \varepsilon,\ \varepsilon \sim N(0,\ \sigma^2 W^{-1})\ \ where\ W = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

$$Var(Y_i) = \frac{\sigma^2}{w_i} \Leftrightarrow Var(\sqrt{w_i}\, Y_i) = \sigma^2$$

변환 후 **가중오차 ε^* 는 등분산성을 만족하므로**

위 관계를 가중오차 제곱합을 최소로 하는 **OLS 적용 가능!** 과시 설정

$$W^{1/2} Y = W^{1/2} X\beta + \varepsilon^*,\ \varepsilon^* \sim N(0,\ \sigma^2)$$

4 Homoscedasticity

등분산성의 처방 | 가중 최소제곱법

어떤 관측치에 어느 정도의 가중치를 부여하게 될까?

$Var(Y_i) = \frac{\sigma^2}{w_i}$ 관계를 살펴본다면,

반응변수의 분산과 가중치 간 반비례 관계를 확인할 수 있음

변환 후 가중오차 ε^* 는 등분산성을 만족하므로

따라서 가중 최소제곱법 적용 시

가중오차 제곱합을 최소로 하는 OLS 적용 가능!

상대적으로 정확도가 높은(=분산이 작은) 관측치에

상대적으로 높은 가중치를 부여하게 됨

등분산성의 처방 | 가중 최소제곱법

가중 최소제곱법의 모수 추정

$$L^w = \left(W^{1/2}\,Y - W^{1/2}\,X\beta\right)^T \left(W^{1/2}\,Y - W^{1/2}\,X\beta\right)$$

최소로 하는 β 추정

Normal Equation

$$\frac{\partial L^w}{\partial \beta} = -2X^T W(Y - X\beta) = 0$$

WLSE (Weighted Least Square Estimator)

$$\hat{\beta}^w = (X^T W X)^{-1} X^T W Y$$

등분산성의 처방 | 가중 최소제곱법

가중 최소제곱법의 모수 추정

$$L^w = \left(W^{1/2}\,Y - W^{1/2}\,X\beta\right)^T \left(W^{1/2}\,Y - W^{1/2}\,X\beta\right)$$

최소로 하는 β 추정

Normal Equation

WLSE는 **선형회귀의 기본 가정하에서 도출된 Estimator**이므로

BLUE(Best Linear Unbiased Estimator)임!

WLSE (Weighted Least Square Estimator)

$$\hat{\beta}^w = (X^T W X)$$

BLUE의 정의와 조건에 대한 자세한 내용은

회귀분석팀 1주차 클린업 자료 확인!

4 Homoscedasticity

등분산성의 처방 | 가중 최소제곱법

일반적으로 가중치를 추정하기는 어렵지만,

다음의 경우에는 쉽게 가중치를 추정해볼 수 있음!

① **반응변수**가 **비율이나 확률**로 주어진 데이터

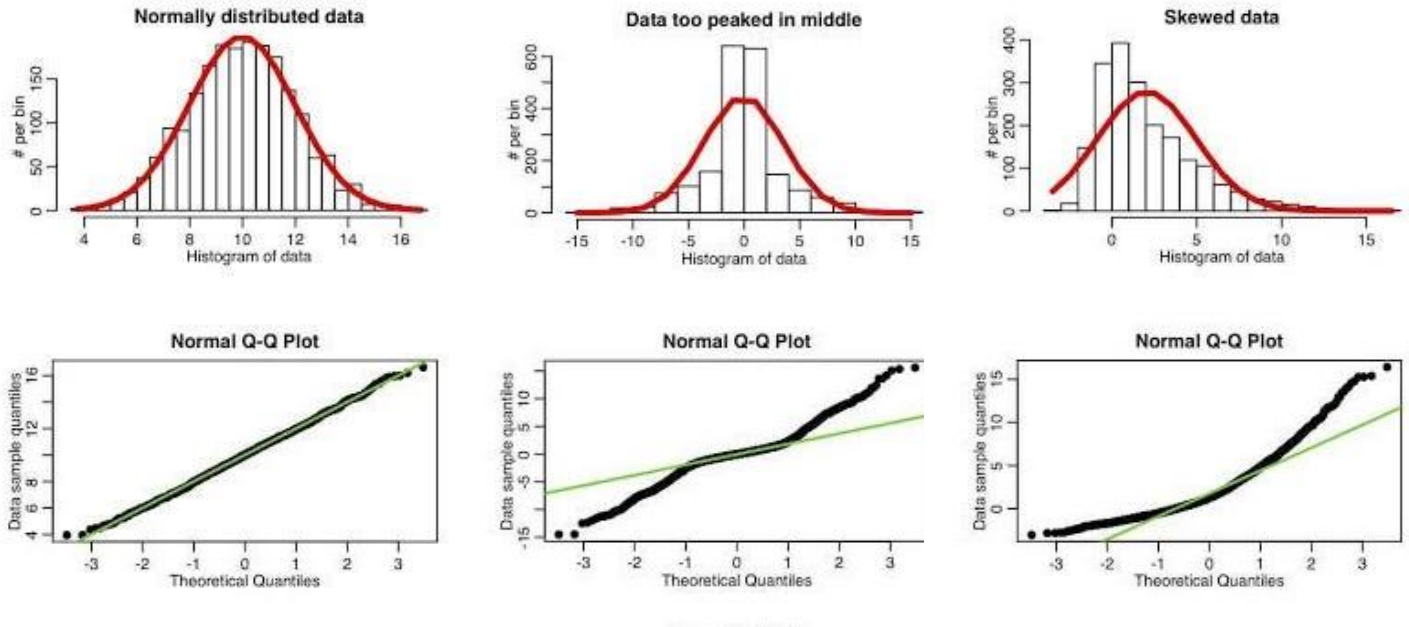
② **그룹의 평균**으로 주어져 있고, **그룹의 크기**를 알고 있는 데이터

5

Normality

5 Normality

정규성의 진단 | Normal Q-Q Plot 확인

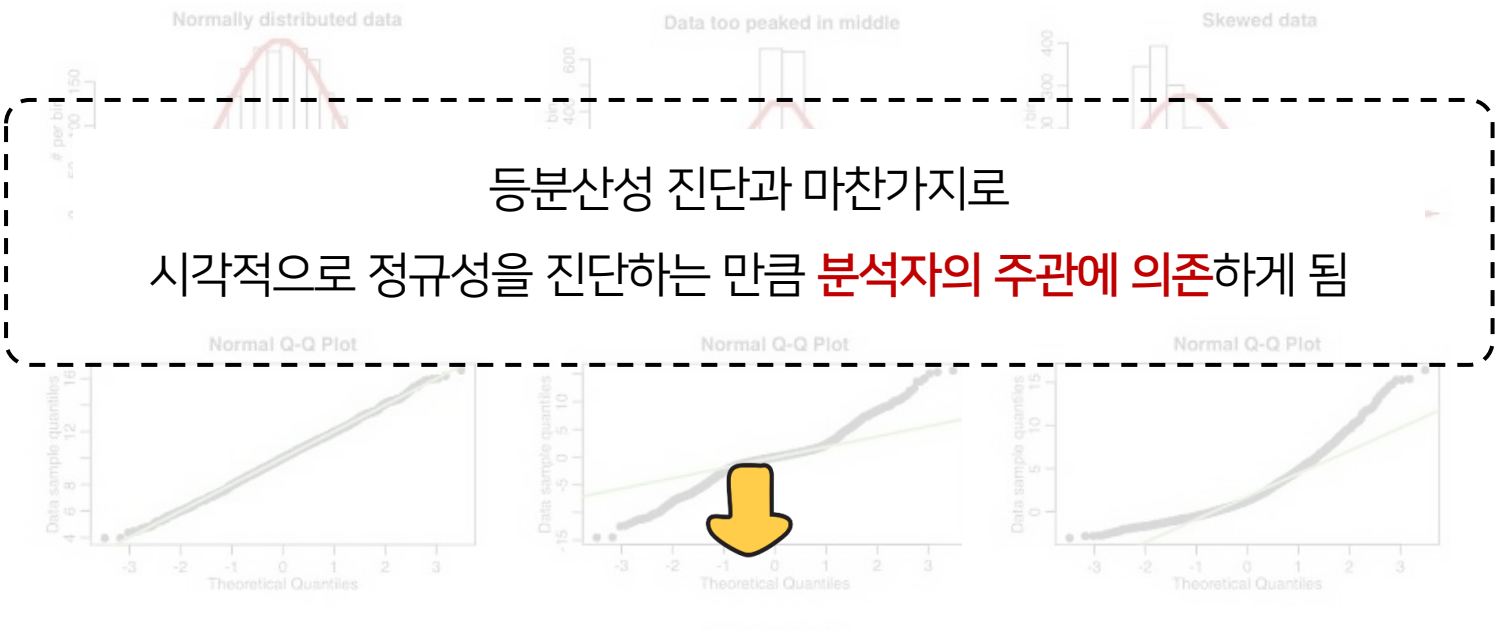


잔차의 sample Quantile(검은선)이 Theoretical Quantile(초록선)에

근접하는지 여부를 확인하여 정규성을 만족하는지 확인할 수 있음

5 Normality

정규성의 진단 | Normal Q-Q Plot 확인



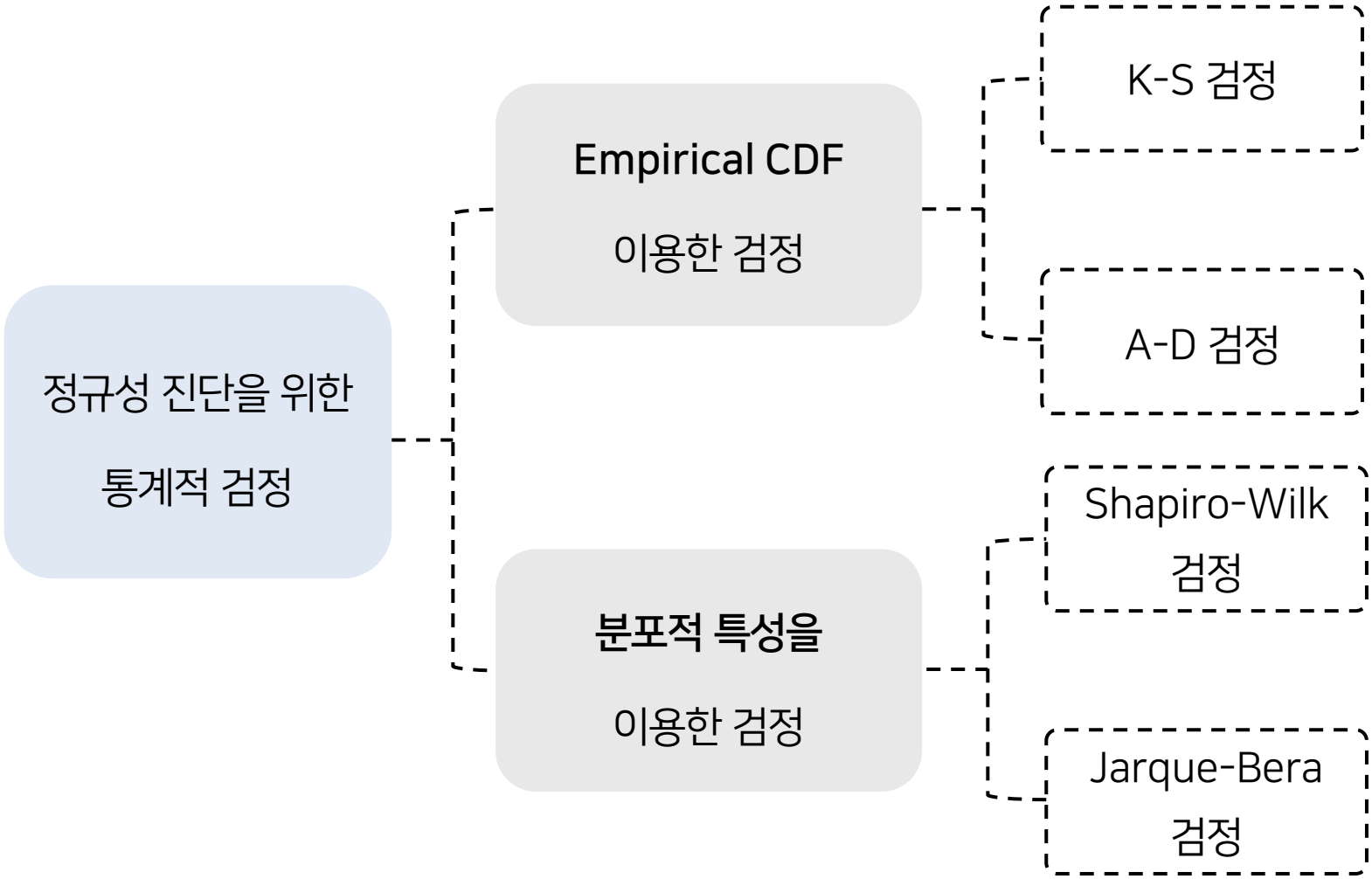
등분산성 진단과 마찬가지로

시각적으로 정규성을 진단하는 만큼 **분석자의 주관에 의존**하게 됨

여러 통계적 검정방법을 통해 정량적으로 정규성을 진단해보자!

근접하는지 여부를 확인하여 정규성을 만족하는지 확인할 수 있음

정규성의 진단 | 여러가지 통계적 검정



5 Normality

정규성의 진단 | 여러가지 통계적 검정

검정 통계량은 각기 다르지만,

귀무가설은 검정에 상관없이 동일하다는 것에 유의!

Null Hypothesis vs Alternative Hypothesis

H_0: 주어진 데이터는 정규분포를 따른다 vs H_1: $Not\ H_0$

잔차는 정규성을 따라야 하므로, **귀무가설이 기각되지 않기를 원함!**

정규성의 진단 | CDF, EDF

누적분포함수 (CDF, Cumulative Distribution Function)

주어진 확률변수가 특정 값보다 같거나 작을 확률을 나타내는 함수

적합도 검정에서는 이론적 누적분포함수(Theoretical CDF)라고도 쓰임

x_0 가 주어졌을 때

$$F(x_0) := P_X(X \leq x_0)$$

정규성의 진단 | CDF, EDF

경험적 누적분포함수 (EDF, Empirical Distribution Function)

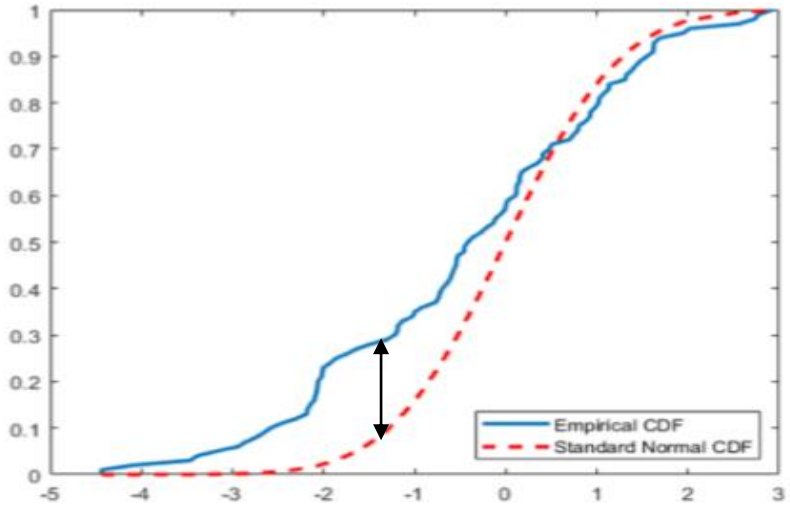
주어진 데이터의 관측값을 크기 순으로 정렬한 후

각 값의 누적빈도를 계산하여 구한 누적분포함수의 추정치

x_0 가 주어졌을 때

$$\widehat{F_n}(x_0) = \frac{number\ of\ X_i \le x_0}{total\ number\ of\ observations} = \frac{\sum_{i=1}^{n} I(X_i \le x_0)}{n}$$

정규성의 진단 | CDF, EDF



파란선 : 표본의 EDF

빨간선 : 검정하고자 하는 분포의 CDF

검은색 화살표 : EDF와 CDF 값의 차이

정규성의 진단 | K-S Test

K-S Test (Kolmogorov – Smirnov Test)

확률분포의 동일성을 비모수적으로 검정하는 방법

One Sample K-S Test

어떤 하나의 표본이

특정 분포의 CDF와 유사한지 검정

Two Sample K-S Test

두 집단의 표본이

서로 동일한 분포를 가지는지 검정

5 Normality

정규성의 진단 | K-S Test

K-S Test (Kolmogorov – Smirnov Test)

확률분포의 동일성을 비모수적으로 검정하는 방법

One Sample K-S Test

어떤 하나의 표본이

특정 분포의 CDF와 유사한지 검정

Two Sample K-S Test

잔차의 분포가 정규분포와 동일한지

확인하는 것이 목표이므로,

One-Sample K-S Test 에 집중해볼 것임!

정규성의 진단 | K-S Test

Test Statistic

$$D = \sqrt{n} \sup_x |\hat{F}_n(x) - F(x)|$$

$$Reject\ H_0 \quad if\ p - value\ for\ D < \alpha \quad (Significance\ Level:\ \alpha)$$

* $\sup_x$: Supremum, Least upper bound

$|\hat{F}_n(x) - F(x)|$: K-S 통계량은 **표본의 EDF와 검정하고자 하는 분포의**

CDF 간의 **거리**를 측정하는 방식으로 계산

정규성의 진단 | A-D Test

A-D Test (Anderson-Darling Test)

데이터가 특정 분포를 따르는지 검정하는 방법

앞선 K-S Test를 수정한 검정으로,

분포의 꼬리(tail) 부분에 높은 가중치를 두고 검정을 수행

K-S Test 보다 엄격한 검정방법

정규성의 진단 | A-D Test

Test Statistic

$$A^2 = n \int_{-\infty}^{\infty} \frac{\{\widehat{F_n}(x_0) - F(x)\}^2}{F(x)\big(1 - F(x)\big)} dF(x)$$

$$\approx -n - \sum_{i=1}^{n} \frac{2i-1}{n} [\ \ln\big(F(Z_i)\big) + \ln\big(F(Z_{n+1-i})\big)\ ]$$

$$\vdots$$

$$Reject\ H_0 \quad if\ A^2 > critical\ value\ of \quad theoretical\ distribution$$

$$(Significance\ Level:\ \ \alpha)$$

정규성의 진단 | A-D Test

Test Statistic

$$A^2 = n\int_{-\infty}^{\infty} \frac{\{\widehat{F_n}(x_0) - F(x)\}^2}{F(x)\big(1-F(x)\big)}\, dF(x)$$

$$\approx -n - \sum_{i=1}^{n} \frac{2i-1}{n}\left[\ln\big(F(Z_i)\big) + \ln\big(F(Z_{n+1-i})\big)\right]$$

$F(x)$: 검정을 통해 확인하고자 하는 분포의 누적분포함수로,
정규성 진단 시 $F(x)$는 정규분포의 이론적 누적분포함수

정규성의 진단 | A-D Test

Test Statistic

$$A^2 = n\int_{-\infty}^{\infty} \frac{\{\widehat{F_n}(x_0) - F(x)\}^2}{F(x)\big(1-F(x)\big)}\, dF(x)$$

$$\approx -n - \sum_{i=1}^{n} \frac{2i-1}{n}\left[\ln\big(F(Z_i)\big) + \ln\big(F(Z_{n+1-i})\big)\right]$$

$\widehat{F_n}(x_0)$: 표본의 경험적 누적분포함수로,

정규성 진단 시 $\widehat{F_n}(x_0)$는 잔차의 경험적 누적분포함수

정규성의 진단 | A-D Test

Test Statistic

$$A^2 = n\int_{-\infty}^{\infty} \frac{\{\widehat{F_n}(x_0) - F(x)\}^2}{F(x)\big(1-F(x)\big)}\,dF(x)$$

$$\approx -n - \sum_{i=1}^{n} \frac{2i-1}{n}\left[\ln\big(F(Z_i)\big) + \ln\big(F(Z_{n+1-i})\big)\right]$$

$\{F_n(x) - F(x)\}^2$: 표본의 EDF와 특정 분포의 CDF와의 거리

즉, 해당 값의 차이가 클수록 검정통계량이 커짐

정규성의 진단 | A-D Test

Test Statistic

$$A^2 = n \int_{-\infty}^{\infty} \frac{\{\widehat{F_n}(x_0) - F(x)\}^2}{F(x)\big(1 - F(x)\big)} dF(x)$$

$$\approx -n - \sum_{i=1}^{n} \frac{2i-1}{n} [\ln\big(F(Z_i)\big) + \ln\big(F(Z_{n+1-i})\big)]$$

$F(x)\big(1 - F(x)\big)$: A-D Test에서 꼬리 부분에 더 가중치를 두게 함

정규성의 진단 | A-D Test

Test Statistic

$$A^2 = n\int_{-\infty}^{\infty} \frac{\{\widehat{F_n}(x_0) - F(x)\}^2}{F(x)\big(1-F(x)\big)}\, dF(x)$$

$$\approx -n - \sum_{i=1}^{n} \frac{2i-1}{n}\left[\ln\big(F(Z_i)\big) + \ln\big(F(Z_{n+1-i})\big)\right]$$

Z_i : x_i를 표준화시킨 뒤 오름차순 정렬하였을 때 i번째 값

정규성의 진단 | Shapiro-Wilk Test

Shapiro-Wilk Test

정규분포의 분위수와 내표준화 잔차 간의 선형관계를 확인해

표본이 정규분포로부터 추출된 것인지 확인하는 검정 방법

⋮

표본 수가 적을 때 (SAS 기준 2,000개, R 기준 5,000개 이하)

Shapiro-Wilk Test 를 수행할 것을 권장

정규성의 진단 | Shapiro-Wilk Test

Test Statistic

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$where \ \ (a_i, \dots, a_i) = \frac{m^T \, V^{-1}}{(m^T \, V^{-1} \, V^{-1} m)^{1/2}}$$

$$Reject \ H_0 \ \ if \ \ W > critical \ value \ of \ \ \alpha \ (Significance \ Level: \ \ \alpha)$$

정규성의 진단 | Shapiro-Wilk Test

Test Statistic

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$where\ \ (a_i, \dots, a_i) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

- $x_{(i)}$: 표본의 i번째 order statistics
- m : iid 표준정규분포의 Order statistic 에 대한 기댓값 벡터 집합
- V : iid 표준정규분포의 Order statistic 에 대한 공분산 행렬

정규성의 진단 | Jarque-Bera Test

Jarque-Bera Test

정규분포의 왜도와 첨도가 각각 0과 3이라는 특성을 이용하여

표본의 왜도와 첨도에 따라 정규성 만족여부를 판단하는 검정

5 Normality

정규성의 진단 | Jarque-Bera Test

Test Statistic

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

$$Reject\ H_0\ \ if$$

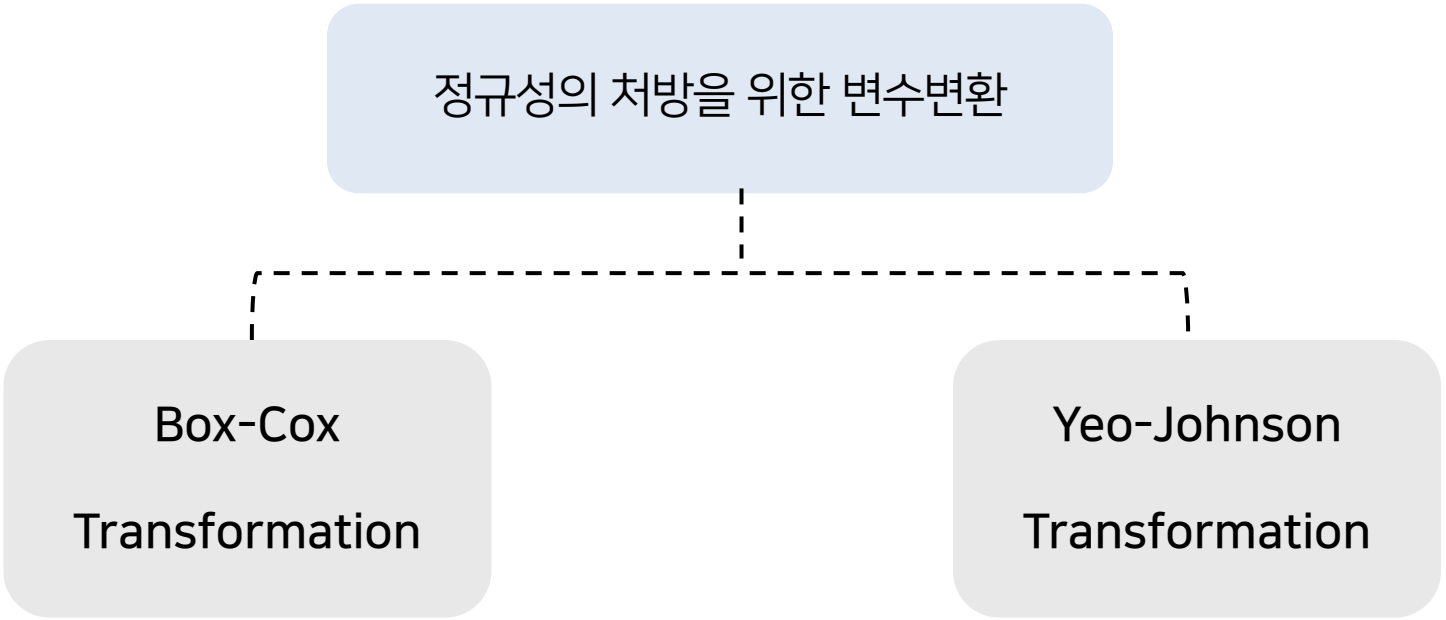
$$W > critical\ value\ of\ \ chi - square\ (df = 2, Significance\ Level:\ \ \alpha)$$

S : 관측치의 왜도 K : 관측치의 첨도

이상치에 민감한 왜도 값을 사용하는 만큼,

이상치를 제거했을 때 정규분포임이 드러나는 경우가 많음!

정규성의 처방



두 방법은 정규성의 처방은 물론 이분산성 문제해결에도 사용 가능함!

정규성의 처방 | Box-Cox Transformation

Box-Cox Transformation

파라미터 λ 값에 따라 정규분포에 가깝도록 반응변수 Y를 변환하는 방법

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

$\vdots$

반응변수 y 에 대해 위와 같은 변환을 수행하였을 때

$\mathbf{y(\lambda)}$가 근사적으로 정규분포에 가깝게 되도록 하는 $\boldsymbol{\lambda}$ 를 제안

정규성의 처방 | Box-Cox Transformation

Box-Cox Transformation

파라미터 λ 값에 따라 정규분포에 가깝도록 반응변수 Y를 변환하는 방법

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

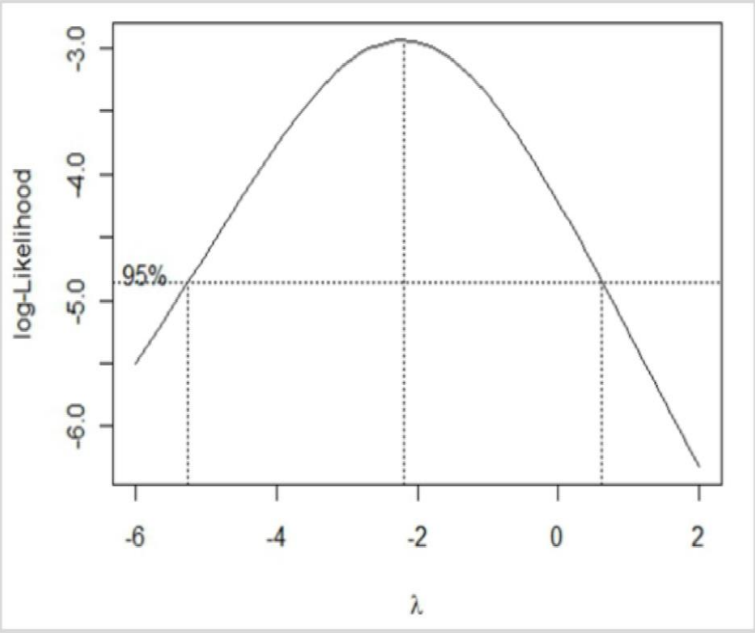
$\lambda = 0$ 일 때 로그변환을 수행하므로,

Box-Cox 변환은

반응변수가 양수인 값에 대해서만 적용할 수 있다는 점에 유의!

$y(\lambda)$가 근사적으로 정규분포에 가깝게 되도록 하는 λ를 제안

정규성의 처방 | Box-Cox Transformation



λ 값을 적절한 구간에서 조금씩 움직이며

로그우도함수가 최대가 되는 λ 값을 선택

정규성의 처방 | Yeo-Johnson Transformation

Yeo-Johnson Transformation

Box-Cox 변환 시 반응변수가 양수여야 한다는 제약을 극복한 변환으로,

파라미터 λ 값에 따라 정규분포에 가깝도록 반응변수 Y를 변환하는 방법

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

6

Independence

독립성의 진단 | Durbin-Watson Test

Durbin-Watson Test

관측치의 1차 자기상관성 (First order autocorrelation) 을 확인하여

독립성 여부를 판단하는 검정

* 1차 자기상관성 : (t – 1) 시점의 잔차와 t 시점의 잔차 간 상관관계

Null Hypothesis vs Alternative Hypothesis

H_0: 잔차들 간 1차 자기상관성이 존재하지 않는다 vs H_1: 그렇지 않다

Test Statistic

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

6 Independence

독립성의 진단 | Durbin-Watson Test

Durbin-Watson Test

관측치의 1차 자기상관성 (First order autocorrelation) 을 확인하여

독립성 여부를 판단하는 검정

* 1차 자기상관성 : (t – 1) 시점의 잔차와 t 시점의 잔차 간 상관관계

Null Hypothesis vs Alternative Hypothesis

H_0: 잔차들 간 1차 자기상관성이 존재하지 않는다 vs H_1: 그렇지 않다

Test Statistic

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

독립성의 진단 | Durbin-Watson Test

검정통계량의 전개

분자의 완전 제곱식을 전개 시

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

⋮

표본의 크기 n이 충분히 크다면

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

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

독립성의 진단 | Durbin-Watson Test

검정통계량의 근사 결과

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

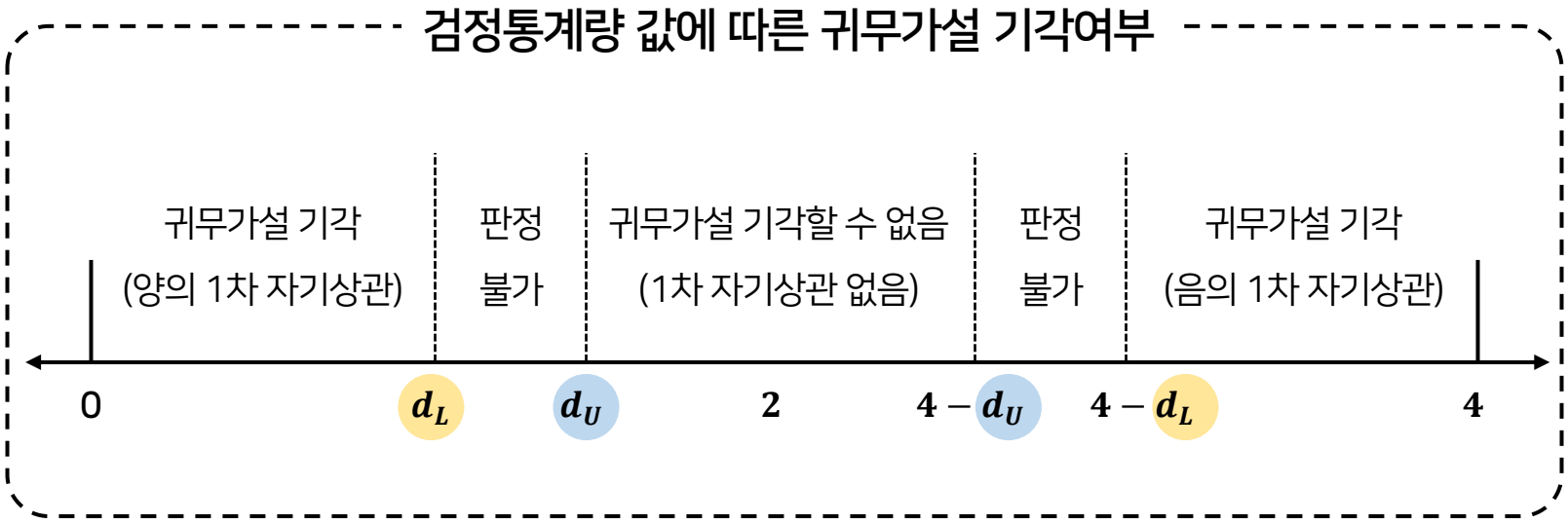
$$where\ \hat{\rho}_1 = \frac{\widehat{Cov}(e_i, e_{i-1})}{\sqrt{V(e_i)}\sqrt{V(e_{i-1})}} = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} {e_i}^2}$$

ⓐ 1차 표본자기상관계수($\hat{\rho}_1$)는 -1부터 1 사이의 값을 가짐

ⓑ $\hat{\rho}_1$과 검정통계량(d)는 선형관계라는 점을 고려할 때,

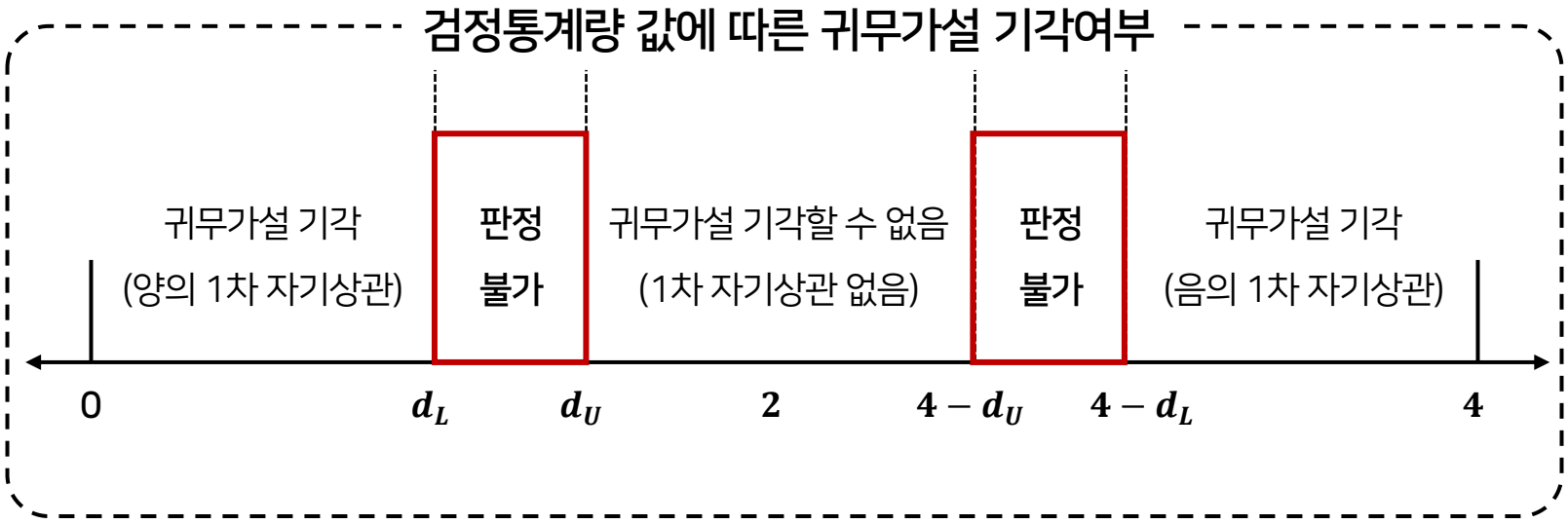
만약 데이터가 **독립적이라면 검정통계량 d는 2에 가까운 값을 가질 것**임

독립성의 진단 | Durbin-Watson Test



① Durbin-Watson Test 의 임계값(Critical Value) d_L, d_U 은
Durbin-Watson 검정표를 통해 구해볼 수 있음

독립성의 진단 | Durbin-Watson Test



② 검정통계량이 **불확실 영역(Inconclusive Region)**에 속하게 될 때,
Durbin-Watson Test 의 판정을 내리지 못하게 됨

독립성의 진단 | Durbin-Watson Test

검정통계량 값에 따른 귀무가설 기각여부

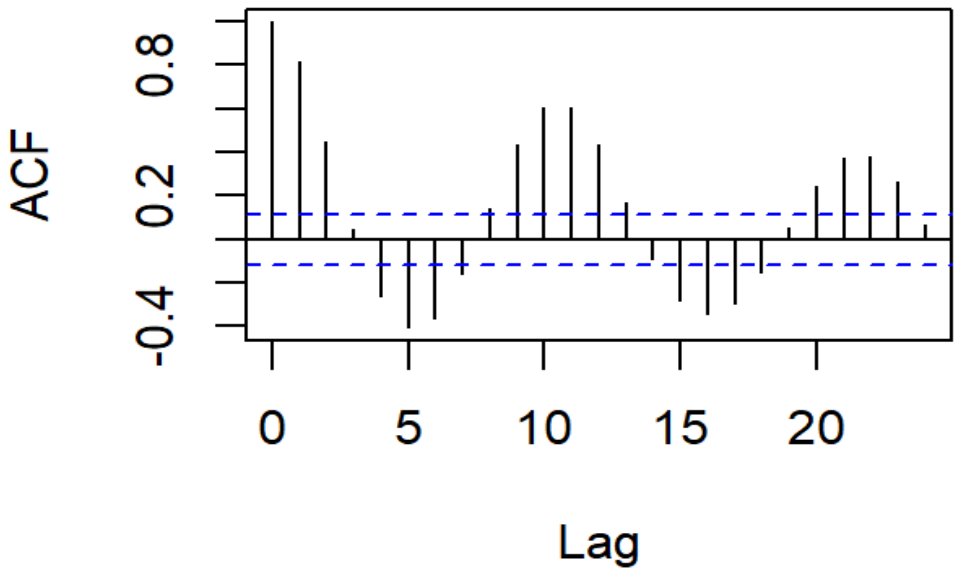
Durbin-Watson Test의 경우 1차 자기상관만을 고려하므로
2차 이상의 고차원 자기상관 혹은 계절성이 있는 경우
해당 검정을 적용할 수 없음

2차 이상의 자기상관을 확인할 수 있는 ACF Plot을 확인!

독립성의 진단 | ACF Plot

ACF Plot (Autocorrelation Function Plot)

시차(lag)별 표본자기상관계수 값($\hat{\rho}_{lag}$)을 선으로 표현한 그래프



6 Independence

독립성의 진단 | ACF Plot

① 첫 번째 자기상관계수는 항상 1임

② 신뢰구간의 경계 값(파란선)을 넘었을 경우

해당 lag에서 데이터 간 자기상관이 존재한다고 판단

③ 전체적으로 경계선을 넘어가는 막대가 5개 미만일 때

데이터의 독립성 가정이 충족되었다고 판단

독립성의 진단 | ACF Plot

① 첫 번째 자기상관계수는 항상 1임

시각적으로 등분산성을 진단하는 만큼

여전히 **분석자의 주관에 의존**하게 되는 한계가 있음!

ACF Plot을 통한 계절성 확인, 정상화 등에 대해서는
시계열자료팀 1주차 클린업 참고!

③ 전 만일 때

데이터의 독립성 가정이 충족되었다고 판단

독립성의 처방

데이터가 독립성을 갖도록 변형해주는 것이 아님!

데이터의 자기상관 정도를 반영하여

적합한 회귀식을 모델링하는 방법 이용

- 시간적 자기상관
  - **자기회귀 오차모형**
  - 시계열자료팀 클린업 참고!
- 공간적 자기상관
  - **공간자기 회귀모형**
  - 3주차에 자세히 다뤄볼 예정!

7

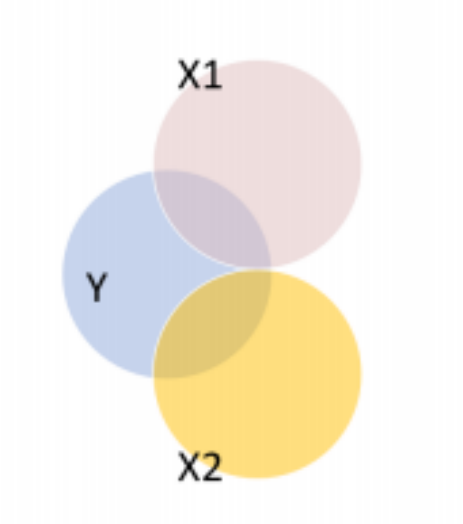
Multicollinearity

다중공선성

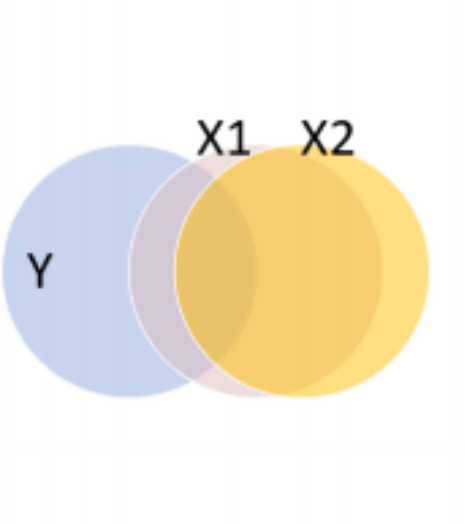
다중공선성 (Multicollinearity)

설명변수(X) 들 간에 서로 상관관계가 높거나

완전히 선형종속(Linearly dependent) 관계에 놓여있는 경우



[다중공선성이 없는 경우]



[다중공선성이 있는 경우]

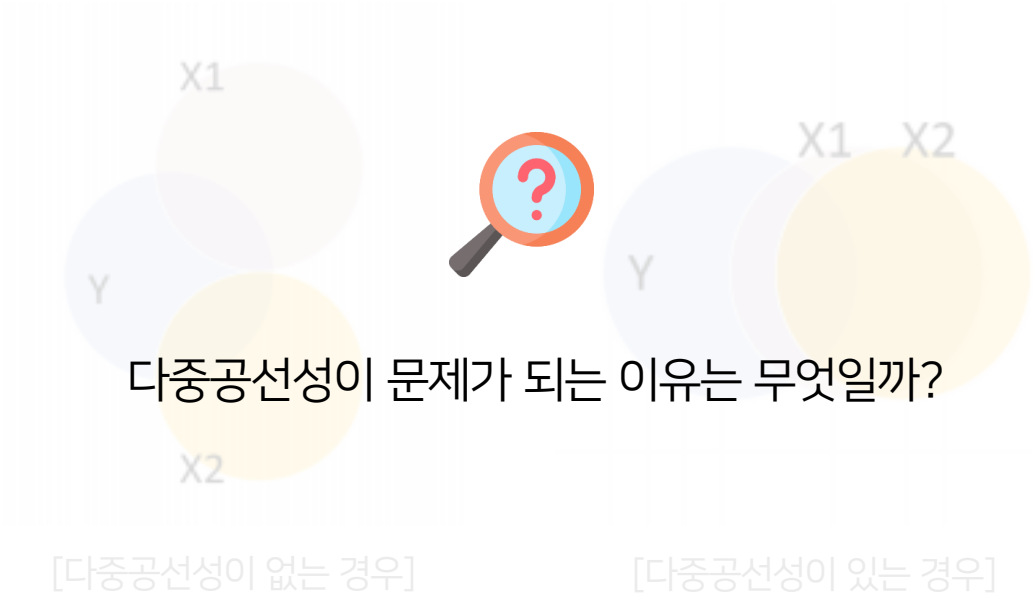
7 Multicollinearity

다중공선성

다중공선성 (Multicollinearity)

설명변수(X) 들 간에 서로 상관관계가 높거나

완전히 선형종속(Linearly dependent) 관계에 놓여있는 경우



[다중공선성이 없는 경우] [다중공선성이 있는 경우]

다중공선성이 문제가 되는 이유는 무엇일까?

다중공선성 | 문제점

선형종속 관계에 놓인 경우

설명변수들이 선형종속 관계에 놓이게 되면

X^TX 행렬의 역행렬이 존재하지 않게 되므로

OLS 과정을 통해 $\hat{\beta} = (X^TX)^{-1}X^TY$ 유일해 도출 불가능

회귀계수 추정이 불가능해짐

다중공선성 | 문제점

상관관계가 높은 경우

r : 설명변수들 간의 상관계수

$$X^TX = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}, (X^TX)^{-1} = \frac{1}{1-r^2}\begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}$$

$$|r| \to 1 \longrightarrow det[\,(X^TX)^{-1}\,] \uparrow \longrightarrow V(\hat{\beta}) \uparrow$$

$|r| \to 1$ 이면 **회귀계수의 분산이 매우 커지게 되어**

계수 추정의 정확도가 낮아지고 **회귀모형의 예측 정확도 감소**

7 Multicollinearity

다중공선성 | 진단 방법

① 중요하다고 생각한 설명변수의 회귀계수 추정치 **분산 값이 매우 큰 경우**

즉, 연관이 클 것이라 예상한 설명변수의 회귀계수에 대한 t-test 결과가 유의하지 않다는 결과가 나올 때

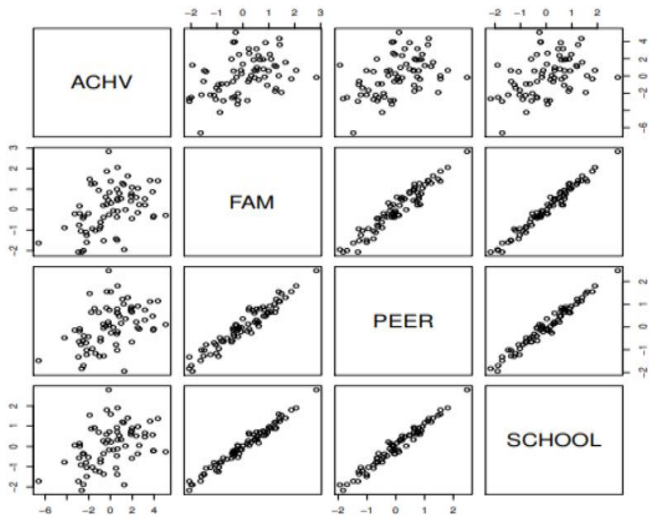
② 회귀계수 추정치의 부호가 사전적으로 기대하는 **부호와 일치하지 않는 경우**

Ex) 설명변수(X) 중 하나가 '공부시간', 반응변수(Y)가 '학점'일 때,
공부시간 설명변수의 계수 추정치 부호가 음수로 나온 경우

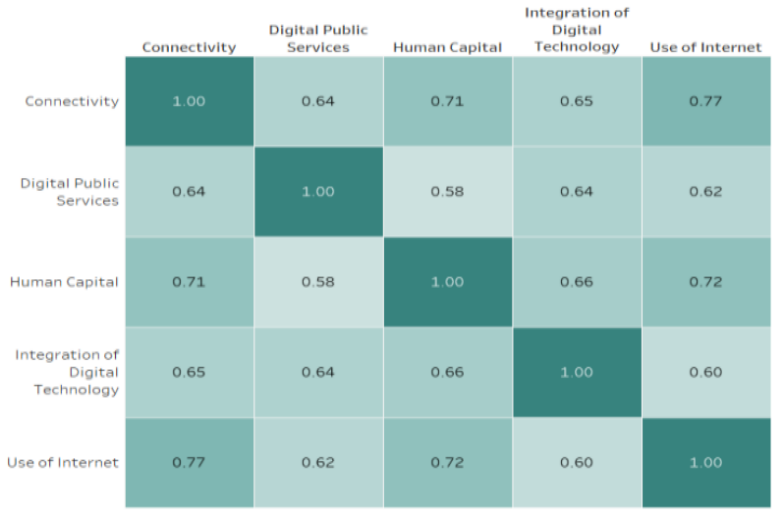
③ 특정 변수를 **추가**하거나 **삭제**했을 때 회귀계수 **추정 값이 크게 변하는 경우**

다중공선성 | 진단 방법

④ Scatter plot 및 Correlation matrix에서 높은 상관관계가 나타나는 경우



[Scatter plot]



[Correlation plot]

다중공선성 | 진단 방법

⑤ VIF(Variance Inflation Factor) 값이 10보다 크게 나오는 경우

VIF가 무엇인지 알아보도록 하자 !

다중공선성 | 진단 방법

VIF (Variance Inflation Factor)

설명변수(X) 들 중 특정 하나를 반응변수(Y) 로 둔 뒤,

남은 X 변수들을 설명변수로 하는 선형회귀분석모형을 모델링하여

변수 간의 관계성을 측정하는 척도

j 번째 설명변수를 반응변수 자리에 대입하면

$$X_j = \beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \cdots$$

$$\vdots$$

fitting 식 $X_j = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_{j-1} X_{j-1} + \hat{\beta}_{j+1} X_{j+1} + \cdots$ 의 $R^2 = {R_j}^2$

X_j 가 $(X_1, X_2, \cdots, X_{j-1}, X_{j+1}, \cdots)$ 과 높은 상관관계를 가지면 R_j 가 커짐

다중공선성 | 진단 방법

VIF (Variance Inflation Factor)

설명변수(X) 들 중 특정 하나를 반응변수(Y) 로 둔 뒤,

남은 X 변수들을 설명변수로 하는 선형회귀분석모형을 모델링하여

변수 간의 관계성을 측정하는 척도

$$VIF_j = \frac{1}{1 - {R_j}^2}$$

⋮

${R_j}^2$: j 번째 설명변수(X_j) 를 반응변수로 두고 나머지 설명변수에 대해

회귀분석을 수행했을 때, 얻어지는 결정계수의 값

다중공선성 | 진단 방법

VIF (Variance Inflation Factor)

설명변수(X) 들 중 특정 하나를 반응변수(Y) 로 둔 뒤,

남은 X 변수들을 설명변수로 하는 선형회귀분석모형을 모델링하여

변수 간의 관계성을 측정하는 척도

R_j^2 값이 1에 가까워 VIF_j 값이 매우 크다면,

X_j 가 남은 설명변수들의 선형결합 형태로 충분히 표현될 수 있어

다중공선성이 존재할 가능성이 매우 높을 것

다중공선성 | 진단 방법

VIF (Variance Inflation Factor)

설명변수(X) 들 중 특정 하나를 반응변수(Y) 로 둔 뒤,

남은 X 변수들을 설명변수로 하는 선형회귀분석모형을 모델링하여

변수 간의 관계성을 측정하는 척도

R_j^2 값이 1에 가까워 VIF_j 값이 매우 크다면,

X_j 가 남은 설명변수들의 선형결합 형태로 충분히 표현될 수 있어

다중공선성이 존재할 가능성이 매우 높을 것

VIF 값이 10보다 큰 경우

다중공선성이 존재한다고 판단 !

다중공선성 | 진단 방법

⑥ $\boldsymbol{X^T X}$ 의 **고유값 조사** 결과, $\boldsymbol{(\lambda_1/\lambda_p)^{1/2}}$ **값이 40보다 크게** 나오는 경우

$X^T X$ 에 대해 고유값 분해를 진행했을 때

0 에 가까운 고유값이 존재한다면 다중공선성을 의심해볼 수 있음

$$V^T(X^T X)V = diag(\lambda_1, \ldots, \lambda_p)$$

다중공선성 | 진단 방법

⑥ X^TX 의 고유값 조사 결과, $(\lambda_1/\lambda_p)^{1/2}$ 값이 40보다 크게 나오는 경우

0 에 가까운 고유값들이 존재하면

왜 다중공선성을 의심해볼 수 있는 걸까?

$$V^T(X^TX)V = diag(\lambda_1, \ldots, \lambda_p)$$

다중공선성 | 진단 방법

$diag(\lambda_1, \dots, \lambda_p)$ 의 k 번째 값인 λ_k 가 0 에 가까운 값을 가진다고 가정해본다면,

$$\boldsymbol{v_k}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{v_k} = (\boldsymbol{X}\boldsymbol{v_k})^T\,\boldsymbol{X}\boldsymbol{v_k} = \boldsymbol{\lambda_k} \approx \boldsymbol{0} \leftrightarrow \sum_{i=1}^{p} \boldsymbol{v_{ik}}\boldsymbol{X_i} \approx \boldsymbol{0}$$ 이 성립하게 됨

$\sum_{i=1}^{p} \boldsymbol{v_{ik}}\boldsymbol{X_i}$ 식에서 $\boldsymbol{v_{kk}}\boldsymbol{X_k}$ 부분만 좌항으로 이항 시켜주면 아래 식이 도출됨

$$-\boldsymbol{v_{kk}}\boldsymbol{X_k} = \boldsymbol{v_{1k}}\boldsymbol{X_1} + \cdots + \boldsymbol{v_{pk}}\boldsymbol{X_p}$$

0 에 가까운 고유값이 존재한다면 다중공선성을 의심해볼 수 있음

$V^T(X^TX)V = diag(\lambda_1, \dots, \lambda_p)$

즉 X_k 가 $X_1, \dots, X_p$ 의 **선형 결합**으로 표현돼

설명변수들이 서로 **선형종속관계**에 놓이게 되므로 다중공선성을 의심해봐야 함!

다중공선성 | 진단 방법

⑥ $\boldsymbol{X^T X}$ 의 **고유값 조사** 결과, $(\lambda_1/\lambda_p)^{1/2}$ **값이 40보다 크게** 나오는 경우

임의의 고유값 λ_p 에 대하여 $\boldsymbol{\lambda_p \to 0}$ **일수록** $(\lambda_1/\lambda_p)^{1/2}$ **의 값은 커지게** 됨

통상적으로 $(\lambda_1/\lambda_p)^{1/2}$ 값이 40을 넘으면

다중공선성이 존재한다고 판단

8

Endogeneity

내생성 | 원인

측정오차

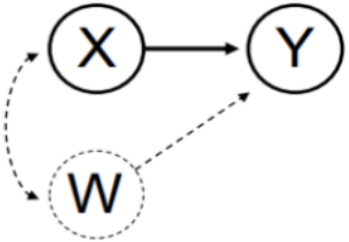
정확하지 않은 측정값을 사용하는 경우

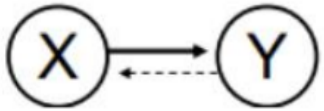
변수 누락

설명변수와 반응변수 모두 연관되는 생략변수가 존재하는 경우

동시성

설명변수와 반응변수 간의 인과관계가 불분명한 경우





8 Endogeneity

내생성 | 원인

측정오차

정확하지 않은 측정값을 사용하는 경우

변수 누락

설명변수와 반응변수

동시성

설명변수와 반응변수 간의

이번 회귀팀 클린업에서는

측정오차에 대해 중점적으로 알아볼 예정!

8 Endogeneity

측정오차

측정오차 (Measurement Error)

응답자가 잘못 응답하거나 설문 집단을 잘못 선정한 상황에서

회귀모델의 변수에 대해 정확하지 않은 측정값을 사용하여 발생하는 Error

군별	공군
복무기간	2019.08.05 ~ 2024.05.21
면제 기타 사유	

Ex) 전역 년도를 잘못 기입한 회귀 팀장의 실제 이력서 제출본

8 Endogeneity

내생성

내생성 (Endogeneity)

Measurement Error 등의 이유로 인해

설명변수(X) 와 오차항 간에 상관관계가 발생하는 현상

①

반응변수(Y) 에 측정오차가

발생하는 경우

②

설명변수(X) 에 측정오차가

발생하는 경우

8 Endogeneity

내생성

내생성 (Endogeneity)

Measurement Error 등의 이유로 인해

설명변수(X) 와 오차항 간에 상관관계가 발생하는 현상

측정오차로 인해 **내생성이 발생**하게 되면

회귀계수의 **추정량은 더 이상 BLUE 가 되지 못함**

발생하는 경우 발생하는 경우

회귀모델의 예측, 추정, 상관관계 분석에 있어 문제를 발생시킴

8 Endogeneity

내생성

내생성 (Endogeneity)

Measurement Error 등의 이유로 인해

설명변수(X) 와 오차항 간에 상관관계가 발생하는 현상

① ②

측정오차의 경우 설명변수(X) 에서 발생됐을 때, 내생성이 발생함

발생하는 경우 발생하는 경우

반응변수(Y) 와 설명변수(X) 각각에 대해

측정오차가 발생한 경우를 확인해보며 그 이유를 알아보자!

8 Endogeneity

측정오차 | 반응변수 (Y)

$E(\varepsilon) = 0,\ Cov(X, \varepsilon) = 0,\ Cov(v, \varepsilon) = 0$ 가정 하에서

측정오차가 발생한 부분을 제외한 Y 데이터를 이용해 회귀식을 작성 시

Gauss Markov Theorem 에 따라 추정량($\hat{\beta}$) 은 **불편성**과 **일치성**을 만족

$$\boldsymbol{y_i = y^*_i + v_i}$$

$$\boldsymbol{y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon} \cdots ①$$

y_i : 측정오차가 발생한 Y 데이터

y^*_i : 측정오차가 발생한 부분을 제외한 Y 데이터

8 Endogeneity

측정오차 | 반응변수 (Y)

그러나 식 ① 을 y_i 값을 이용해서 다시 표현 했을 때,

해당 회귀식의 추정량($\hat{\beta}$) 은 **불편성**과 **일치성**을 만족한다는 보장을 못함

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \cdots ①$$

$$\leftrightarrow y = y^* + v = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon + v \cdots ②$$

$\hat{\boldsymbol{\beta}}$ 은 BLUE 가 되지 못함

측정오차 | 반응변수 (Y)

그러나 식 ① 을 y_i 값을 이용해서 다시 표현 했을 때,

해당 회귀식의 추정량($\hat{\beta}$)은 불편성과 일치성을 만족한다는 보장을 못함

따라서 식 ② 로 부터 도출된 추정량($\hat{\beta}$) 이

BLUE 가 되기 위해선 **추가적인 가정들의 도입이 필요함**

$$\leftrightarrow \; y = y^* + v = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon + v \cdots ②$$

$\hat{\beta}$ 은 BLUE 가 되지 못함

8 Endogeneity

측정오차 | 반응변수 (Y)

식 ② 의 BLUE 만족을 위한 가정

ⓐ

$E(v) = 0$

Measurement Error 의

기댓값은 항상 0

⋮

추정량($\hat{\beta}$) 의 **불편성을 위한 가정**

ⓑ

$Cov(X, v) = 0$

Measurement Error 와

설명변수들 간에는 상관관계 X

⋮

추정량($\hat{\beta}$) 의 **일치성을 위한 가정**

8 Endogeneity

측정오차 | 반응변수 (Y)

식 ② 의 BLUE 만족을 위한 가정

ⓐ

$E(v) = 0$

v 는 반응변수(Y) 에서 발생된 측정 오차로

설명변수(X) 와는 서로 무관하다고 볼 수 있음

이 경우 **가정 ⓑ 는 항상 만족한다고 간주**

추정량($\hat{\beta}$) 의 불편성을 위한 가정

ⓑ

$\boldsymbol{Cov(X, v) = 0}$

Measurement Error 와

설명변수들 간에는 상관관계 X

추정량($\hat{\beta}$) 의 **일치성을 위한 가정**

8 Endogeneity

측정오차 | 반응변수 (Y)

식 ② 의 BLUE 만족을 위한 가정

ⓐ

$E(v) = 0$

Measurement Error 의

기댓값은 항상 0

추정량($\hat{\beta}$) 의 **불편성을 위한 가정**

ⓑ

$Cov(X, v) = 0$

따라서 **가정 ⓐ 의 만족 여부를**

중점적으로 따져볼 필요가 있음

추정량($\hat{\beta}$) 의 **일치성을 위한 가정**

8 Endogeneity

측정오차 | 반응변수 (Y)

식 ② 의 BLUE 만족을 위한 가정

ⓐ

$\boldsymbol{E(v) = 0}$

Measurement Error 의

기댓값은 항상 0

추정량($\hat{\beta}$) 의 **불편성을 위한 가정**

$E(v) = \delta$ 라고 가정해보면

절편 항의 추정량에는 bias 발생하게 됨

$$E(Y|X) = (\hat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \boldsymbol{\delta}) + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

추정량이 완전히 불편성을 가지려면

측정오차의 평균값은 0 이 돼야 함

8 Endogeneity

측정오차 | 반응변수 (Y)

식 ② 의 BLUE 만족을 위한 가정

ⓐ

$E(v) = 0$

Measurement Error 의

기댓

$E(v) = \delta$ 라고 가정해보면

절편 항의 추정량에는 bias 발생하게 됨

$E(Y|X) = (\hat{\beta}_0 + \delta) + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$

하지만 절편을 제외한 계수 추정량들을 다시 봐 보면 ...

추정량이 완전히 불편성을 가지려면

측정오차의 평균값은 0 이 돼야 함

추정량($\hat{\beta}$) 의 불편성을 위한 가정

측정오차 | 반응변수 (Y)

$$E(Y|X) = (\hat{\beta}_0 + \delta) + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

절편을 제외한 계수 추정량들에는 bias 가 발생하지 않음

관계분석 측면에서 반응변수(Y) 의 측정오차 발생은

Critical 한 영향을 주지 않는다는 사실을 확인 가능

8 Endogeneity

측정오차 | 반응변수 (Y)

앞서 살펴본 가정들의 만족 여부와는 관계 없이
계수 추정치의 분산 정도는 커지게 됨

$$V(y) = V(\varepsilon + v) = V(\varepsilon) + V(v) \ (\because Cov(v, \varepsilon) = 0) > V(\varepsilon)$$

$$\therefore V(\hat{\beta}) = V(\varepsilon + v)(X^T X)^{-1} > V(\varepsilon)(X^T X)^{-1}$$

관계분석 측면에서 반응변수(Y) 의 측정오차 발생은
BLUE 가 되더라도 '측정오차가 발생하지 않았을 때' 와 비교해보면
구간 추정 및 예측의 정확도는 감소할 수밖에 없음

8 Endogeneity

측정오차 | 반응변수 (Y)

정리

아래 가정들을 만족 시 OLS 추정량은 BLUE 가 됨

ⓐ $E(v) = 0$, ⓑ $Cov(X, v) = 0$

가정 ⓐ 가 위배돼도 절편에만 bias 가 발생하기 때문에

변수들 간의 관계 정도 파악에는 큰 문제가 되진 않음

8 Endogeneity

측정오차 | 설명변수 (X)

$E(\varepsilon) = 0,\ Cov(x_1{}^*, \varepsilon) = 0,\ Cov(\mu, \varepsilon) = 0$ 가정 하에서

측정오차가 발생한 부분을 제외한 X 데이터를 이용해 회귀식을 작성 시

Gauss Markov Theorem 에 따라 추정량($\hat{\beta}$) 은 **불편성**과 **일치성**을 만족

$$\boldsymbol{x_1 = x^*{}_1 + \mu}$$

$$\boldsymbol{y = \beta_0 + \beta_1 x_1{}^* + \varepsilon} \cdots ①$$

x_i : 측정오차가 발생한 X 데이터

$x^*{}_i$: 측정오차가 발생한 부분을 제외한 X 데이터

8 Endogeneity

측정오차 | 설명변수 (X)

그러나 식 ① 을 $x_1{}^*$ 값을 이용해서 다시 표현 했을 때,

해당 회귀식의 추정량($\hat{\beta}$) 은 **불편성**과 **일치성**을 만족한다는 보장을 못함

$$y = \beta_0 + \beta_1 x_1{}^* + \varepsilon \cdots ①$$

$$\leftrightarrow y = \beta_0 + \beta_1(x_1 - \mu) + \varepsilon = \beta_0 + \beta_1 x_1 + (\varepsilon - \beta_1\mu) \cdots ②$$

$\hat{\boldsymbol{\beta}}$ 은 BLUE 가 되지 못함

측정오차 | 설명변수 (X)

그러나 식 ① 을 ${x_1}^*$ 값을 이용해서 다시 표현 했을 때,

해당 회귀식의 추정량($\hat{\beta}$) 은 불편성과 일치성을 만족한다는 보장을 못함

반응변수(Y) 때와 마찬가지로 식 ② 로 부터 도출된 추정량($\hat{\beta}$) 이

BLUE 가 되기 위해선 **추가적인 가정들의 도입이 필요함**

$$\leftrightarrow \; y = \beta_0 + \beta_1(x_1 - \mu) + \varepsilon = \beta_0 + \beta_1 x_1 + (\varepsilon - \beta_1\mu) \cdots ②$$

$\hat{\beta}$ 은 BLUE 가 되지 못함

8 Endogeneity

측정오차 | 설명변수 (X)

식 ② 의 BLUE 만족을 위한 가정

ⓐ

$E(\mu) = 0$

Measurement Error 의 기댓값은 항상 0

ⓑ

$Cov(x_1, \mu) = 0$

Measurement Error 와 오류값 간에 상관관계 X

ⓒ

$Cov(x_1^*, \mu) \neq 0$

Measurement Error 와 참값 간에는 상관관계 존재

⋮

추정량($\hat{\beta}$) 의

불편성을 위한 가정

⋮

추정량($\hat{\beta}$) 의

일치성을 위한 가정

8 Endogeneity

측정오차 | 설명변수 (X)

식 ② 의 BLUE 만족을 위한 가정

ⓐ

$E(\mu) = 0$

ⓑ

$Cov(x_1, \mu) = 0$

ⓒ

$Cov(x_1^*, \mu) \neq 0$

Measurement Error 와

참값 간에는 상관관계 존재

반응변수(Y) 때와 달리

ⓒ 가정이 추가로 필요한 이유는 무엇일까?

추정량($\hat{\beta}$) 의

불편성을 위한 가정

추정량($\hat{\beta}$) 의

일치성을 위한 가정

8 Endogeneity

측정오차 | 설명변수 (X)

식 ② 의 BLUE 만족을 위한 가정

만약 가정 ⓒ 가 위배되어 $Cov(x_1{}^*, \mu) = 0$ 인 경우

$$Cov(x_1, \mu) = E[x_1\mu] - E[x_1]E[\mu] = E[x_1\mu]$$

$$= E[(x_1{}^* + \mu)\mu] = E[x_1{}^*\,\mu] + E[\mu^2] \neq 0$$

ⓒ

$\boldsymbol{Cov(x_1{}^*, \mu) \neq 0}$

Measurement Error 와 참값 간에는 상관관계 존재

결과적으로 가정 ⓑ $\boldsymbol{Cov(x_1, \mu) = 0}$ 가 위배돼

계수 추정량($\hat{\boldsymbol{\beta}}$) 은 일치성을 만족하지 않게 됨

불편성을 위한 가정

일치성을 위한 가정

측정오차 | 설명변수 (X)

따라서 가정 ⓒ 는, 가정 ⓑ 의 만족을 위한 충분조건이므로

만약 **가정 ⓒ 가 위배될 경우 추정량($\hat{\beta}$) 은 BLUE 가 될 수 없음**을 확인 가능

$$Cov(x_1, \mu) = E[x_1\mu] - E[x_1]E[\mu] = E[x_1\mu]$$

⋮

추가로 **설명변수(x_1) 와 오차항($\varepsilon - \beta_1\mu$) 사이 역시**

상관관계가 존재하게 되므로 **내생성이 생김을 확인 가능**

$$\boldsymbol{Cov(x_1, \varepsilon - \beta_1\mu)} = Cov(x_1, \varepsilon) - \beta_1 Cov(x_1, \mu)$$

$$= Cov(x_1^*, \varepsilon) + Cov(\mu, \varepsilon) - \beta_1 V(\mu) = \boldsymbol{-\beta_1 V(\mu) \neq 0}$$

측정오차 | 설명변수 (X)

설명변수(X) 에서의 측정오차로 인해 내생성이 발생하게 되면

계수 추정량($\hat{\beta}$) 은 정확히 아래와 같은 bias 를 가지게 됨

$$plim\ \hat{\beta} = \beta \left(\frac{V(\varepsilon)}{V(\varepsilon) + V(\mu)} \right) < \beta$$

OLS 추정량에 1보다 작은 값을 곱하는 꼴로써

Attenuation bias 라고 부름

8 Endogeneity

측정오차 | 설명변수 (X)

설명변수(X) 에서의 측정오차로 인해 내생성이 발생하게 되면

이러한 계수추정량의 bias 는 회귀분석 과정의

예측, 구간추정, 상관관계 파악에 있어 문제를 발생시키게 됨

$$plim\ \beta = \beta \left(\frac{}{V(\varepsilon) + V(\mu)} \right) < \beta$$

OLS 추 꼴로써

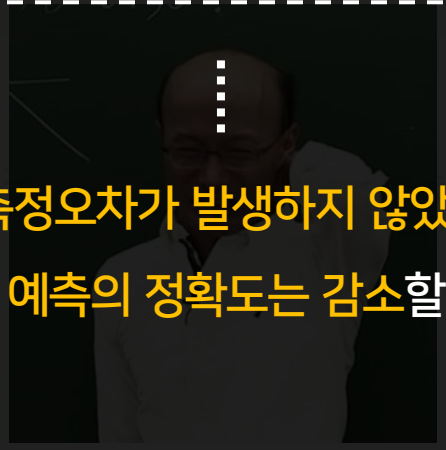
8 Endogeneity

측정오차 | 설명변수 (X)

앞서 살펴본 가정들의 만족 여부와는 관계 없이

계수 추정치의 분산 정도는 커지게 됨

$$V(\hat{\beta}) = V(\varepsilon - \beta_1\mu)(X^TX)^{-1} = \{V(\varepsilon) + \beta_1{}^2V(\mu)\}(X^TX)^{-1}$$
$$> V(\varepsilon)(X^TX)^{-1}$$



BLUE 가 되더라도 '측정오차가 발생하지 않았을 때' 와 비교했을 땐

구간 추정 및 예측의 정확도는 감소할 수밖에 없음

8 Endogeneity

측정오차 | 설명변수 (X)

정리

아래 가정들을 만족 시 OLS 추정량은 BLUE 가 됨

ⓐ $E(\mu) = 0$, ⓑ $Cov(x_1, \mu) = 0$, ⓒ $Cov({x_1}^*, \mu) \neq 0$

특히 가정 ⓒ 가 위배되면 OLS 추정량은 BLUE 가 될 수 없어

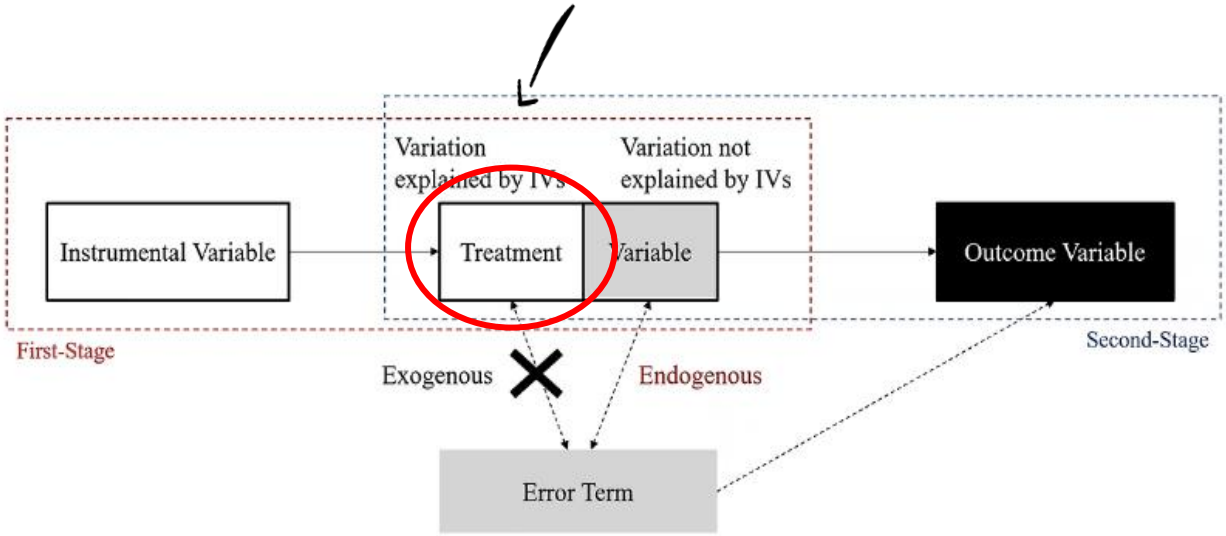
회귀모델의 예측, 구간추정, 상관관계 파악 과정에 있어 문제가 발생

내생성의 처방 | 도구 변수

도구 변수 (Instrumental Variables)

외생적인 설명변수와 내생적인 설명변수를 분리하는

도구적인 역할을 수행하는 변수



내생성의 처방 | 도구 변수

2SLS (2 Stage Least Squares)

도구 변수를 이용해 **설명변수의 외생적인 부분을 예측**한 뒤
이를 원래 회귀식에 다시 적용하여 **반응변수를 예측**하는 방법

0. Original Equation

$$y = \alpha_0 + \alpha_1 x + u$$

$$Cov(x, u) \neq 0$$

..... 내성성이 발생한 (즉 $Cov(x, u) \neq 0$) 원래 회귀식을 작성

내생성의 처방 | 도구 변수

2SLS (2 Stage Least Squares)

도구 변수를 이용해 **설명변수의 외생적인 부분을 예측**한 뒤

이를 원래 회귀식에 다시 적용하여 **반응변수를 예측**하는 방법

1. First-Stage

$$\boldsymbol{x = \beta_0 + \beta_1 z + v}$$

$$\boldsymbol{\hat{x} = \beta_0 + \beta_1 z}$$

$$\beta_1 \neq 0$$

$$Cov(z, u) = 0$$

$$Cov(z, v) = 0$$

.....

도구 변수(z) 를 이용해

설명변수의 외생적인 부분만을 예측

내생성의 처방 | 도구 변수

2SLS (2 Stage Least Squares)

도구 변수를 이용해 **설명변수의 외생적인 부분을 예측**한 뒤
이를 원래 회귀식에 다시 적용하여 **반응변수를 예측**하는 방법

2. Second-Stage

$$y = \gamma_0 + \gamma_1 \hat{x} + u$$

$$\therefore \ Cov(\hat{x}, u) = 0$$

.....

도구 변수(z) 로 예측한
설명변수의 **외생적 부분($\hat{x}$) 을,**
원래 회귀 식의 설명변수 자리에
다시 넣어 반응변수($\hat{y}$) 를 예측

8 Endogeneity

내생성 처방 | 도구 변수

2SLS (2 Stage Least Squares)

도구 변수를 이용해 설명변수의 외생적인 부분을 예측하고

이를 원래 회귀식에 다시 적용하여 반응변수를 예측하는 방법

$\boldsymbol{\beta_1 \neq 0}$: Relevance condition

$\boldsymbol{Cov(z, u) = 0}$: Exclusion condition

$\boldsymbol{Cov(z, v) = 0}$: Exogeneity condition

2. Second-Stage

$y = \gamma_0 + \gamma_1 \hat{x} + u$

$\therefore \ Cov(\hat{x}, u) = 0$

도구 변수(z) 로 예측한
설명변수의 외생적 부분($\hat{x}$) 을,
원래 회귀 식의 설명변수 자리에
다시 넣어 반응변수(y) 를 예측

First-Stage 의 세 조건들을 만족하기가 현실적으로 어렵고
도출된 추정량($\hat{\boldsymbol{\beta}}$) 의 분산은 이전 대비 커진다는 한계가 존재

다음 주 예고

1. Introduction(Prediction Error)

2. Subset Selection

3. Dimension Reduction

4. Shrinkage

5. Regression for Spatial data

감사합니다