Week 2 : 회귀분석의 가정

1. 회귀 기본 가정

- 가정의 종류와 의미

2. 잔차 플랏

- 잔차 플랏의 종류와 해석

3.선형성 진단과 처방

- 선형성 가정이란?
- 진단과 위배되었을 경우 문제점
- 처방

4. 정규성 진단과 처방

- 정규성 가정이란?
- 진단과 위배되었을 경우 문제점
- 처방

5. 등분산성 진단과 처방

- 등분산성 가정이란?
- 진단과 위배되었을 경우 문제점
- 처방

6. 독립성 진단과 처방

- 독립성 가정이란?
- 진단과 위배되었을 경우 문제점
- 처방

0. 복습

1주차 클린업에서 회귀분석의 기본에 대해 배웠다. 선형회귀모델을 중심으로 모수를 추정하고, 적합성 검정, 유의성 검정, 이상치 탐지까지 알아보았다. 잠시 복습타임~!~!

- 회귀분석이란 독립변수와 종속변수 간의 관계를 설명하고 모델링하는 통계적 기법. 상관관계 기반의 모델링으로, 한 변수만을 고려하는 단순선형회귀부터 여러 X변수를 고려하는 다중선형회귀를 주로 다룸
- 회귀계수의 추정은 최소제곱법을 통해 진행하나, 오차의 정규성 가정이 있는 경우 MLE와 LSE는 완전히 동일한 추정량을 가짐
- 다중회귀에서 F 검정은 회귀식 자체에 대한 검정을 다루고, Partial F-test 검정은 일부 회귀 계수에 대한 검정이며, t 검정은 다른 변수를 고정시킨 상태에서 개별 변수의 유의성을 검정
- 회귀식의 Goodness of fit을 측정하는 지표로는 R^2과 ${R_{adj}}^2$가 있는데, 변수 개수가 늘어날 경우 ${R_{adj}}^2$로 모델 간의 비교를 진행할 수 있음.
- 하지만 회귀분석은 이상치에 민감한 경향을 가지기 때문에, 이를 Outlier, Leverage, Influential Point를 통해 각각 관측치를 확인해야 함
- 만약 이상치가 많을 경우, Median Regression, Huber’s M estimation, Least Trimmed Square 등의 Robust(이상치에 강건한) 모델을 고려할 수 있음
- 오늘은 회귀분석의 기본가정과 다중공선성에 대해 확인하면서 가정이 지켜졌는지 판단하는 방법과 가정이 위배되었을 때의 문제점, 처방법에 대해 배워보자.

---

1. 회귀 기본 가정

1) 모델의 가정이 지니는 의미

회귀는 결국 변수 간의 관계를 추정하는 과정이다. 우리는 분석을 통해 잔차가 평균인 0으로 회귀하는 정확한 회귀모델을 만드는 것이 목표! 하지만 여러 이유로 인해 우리가 추정한 모델과 실제 데이터 사이에 오차가 발생한다.

예를 들어, 어떤 사람의 체내 에너지(Y)와 전날 음주량(X)의 관계를 파악하고자 한다. 그러면 우리는 대략적인 선형관계를 추정하고 이에 대한 모수들을 구체적으로 추정하여 회귀식을 얻으려고 할 것이다. 하지만 이때 발생할 수 있는 모델과 실제 데이터 간의 오차를 두 가지 케이스로 나눠볼 수 있다.

- **Case 1**

기본가정인 선형성을 만족하지 못하거나, 혹은 전날 음주량 말고 전날 공부량, 전날 수면 시간 등 다른 변수들이 체내 에너지에 더 밀접한 관련이 있다면 이 회귀식을 통해 예측한 값은 실제 값과 큰 차이가 발생할 것이다.

- **Case 2**

  실제로 우리가 기본적으로 가정한 것이 맞다고 해도 현실 세계의 여러 오차에 의해 회귀식으로 예측한 값은 실제값과 약간의 차이를 가진다.

- 결과적으로 우리는 추정한 모델과 실제 데이터의 오차가 **Case 1) 모델링을 할 때 고려하지 못한 속성들** 때문인지, **Case 2) 현실의 어쩔 수 없는 잡음** 때문인지 확인해야 한다.
- 이를 위해 모델의 선형성, 오차의 등분산성, 오차의 정규성, 오차의 독립성이 지켜졌는지 진단해야 한다.

회귀분석은 많은 머신러닝 모델의 기본 토대가 되며, 적은 관측치만으로도 모델을 구성할 수 있고, 좋은 추정과 예측이 가능하다는 장점이 있다. 하지만 선형회귀모델이 이런 장점을 가지기 위해서는 '선형회귀분석의 4가지 기본가정' (모델의 선형성, 오차의 등분산성, 오차의 정규성, 오차의 독립성)이 지켜져야 한다. 만일 이 가정들이 지켜지지 않는다면 이런 장점들이 사라져서, 추정된 모델은 불안정할 것이며 설명력과 예측력을 잃고 말 것이다. 따라서 선형회귀의 기본가정이 필요하다.

회귀분석 이외의 머신러닝 모델들에도 여러 가정들이 존재하는데, 이러한 가정들이 지켜지지 않을 경우 모델의 성능이 급락하는 경우가 많다. 이번주차 클린업을 통해 회귀분석 가정의 진단과 처방 과정을 통해 모델의 가정이 어떤 의미를 갖는지 이해하고, 다른 모델을 사용할 때에도 모델을 근본적으로 이해한 후에 사용하는 사람들이 됩시다!

2) 선형회귀분석의 가정들

- 이 수식을 통해 회귀분석의 기본 가정들을 알 수 있다. 가정은 크게 변수에 대한 가정과 오차항에 대한 가정 두 가지로 구분 할 수 있다.

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \qquad \varepsilon \sim NID(0, \sigma^2)$$

1. **선형성** (Linearity) : 설명변수와 반응변수의 관계는 선형이다.

   식 자체가 x변수들의 '선형결합'으로 이루어져 있다. 모델 자체가 선형성만 고려하고 있다는 의미!

2. 오차의 **정규성** (Normality) : 오차항은 정규분포를 따른다.
3. 오차의 **독립성** (Independence / No autocorrelation) : 오차항은 서로 독립이다. 즉 오차항 간에 상관관계가 없다.

4. 오차의 **등분산성** (Homoscedasticity / Constant variance) : 오차항의 분산은 상수다.
   → 분산은 σ^2으로 동일하다.
5. 오차의 평균은 0이다.

여기서 '오차의 평균이 0' 이라는 가정은 거의 위반되지 않는다. 따라서 선형성과 오차의 정규성, 독립성, 등분산성에 초점을 맞춰 진단과 처방을 진행하면 된다!

---

2. 잔차 플랏 (Residual Plot)

4가지 가정을 진단하기 위해서 크게 두 가지가 동원된다.

1. **시각적(graphical) 방법** **2. 가설 검정을 이용한 방법**

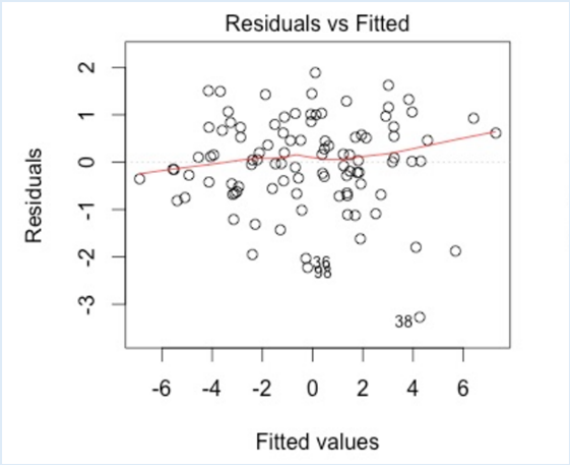
시각적 방법은 잔차 플랏을 많이 사용하며, 가설 검정은 각각의 가정마다 다른 검정 방법을 이용하여 이루어진다. 우선 각 가정의 진단 방법에 보편적으로 사용되는 잔차 플랏을 보는 법부터 배워보자.

오차항의 추정량인 잔차의 분포를 통해 경험적 판단에 근거한 회귀 진단(diagnostic)이 가능해진다. R에서 제공하는 함수 `plot()`를 통해 잔차의 분포를 쉽게 나타낼 수 있으며, 이를 잔차 플랏이라고 부른다. 이 잔차 플랏들을 통해 선형 회귀 모델이 4가지 가정들을 만족하고 있는지 간단히 확인할 수 있다. 이 잔차 플랏을 출력하여 해석하는 방법에 대해 알아보자.

- **잔차 플랏 출력 (R)**

```
# 모델 정의 및 추정
model = lm(Y ~ X1 + X2 + ... + Xp, data = data) # 변수들의 선형 결합으로 표현되어 있다.
# plot display 화면 정의
par(mfrow = c(2,2)) # 이 코드를 실행한 이후 총 4개의 플랏을 한 화면에 (2,2) 그리드로 나타낼 것
# 잔차 플랏 출력
plot(model)
```

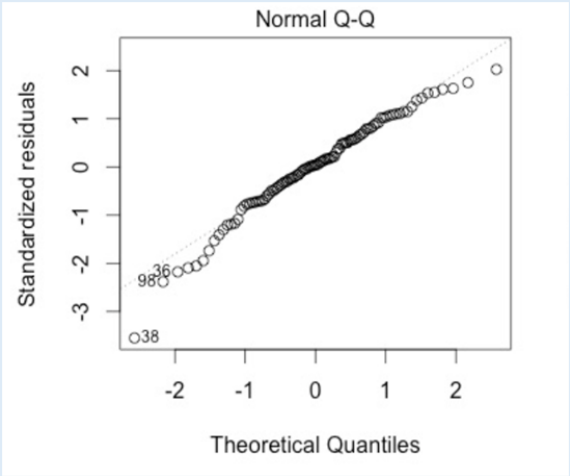
1) Residuals vs Fitted



선형성, 등분산성 가정 만족

- X축 : 예측값($\hat{y}$) fitted values, Y축 : 잔차($e = y - \hat{y}$) residuals
- 선형성과 오차의 등분산성 확인가능
- 빨간 실선 : 전체적인 잔차들의 추세선 → 잔차들의 분포를 Local Regression으로 추정한 직선이며 잔차 분포의 경향성을 나타내는 보조 지표
- 잔차와 예측값 사이에 무작위적인 형태 이외의에 어떠한 관계를 보이면 안 됨. 빨간 실선이 x축에 평행한 직선 형태가 아니라면 선형성이 위반되었다고 볼 수 있다.
- 해석 : 빨간 실선이 완만하게 수평을 이루고 있으며, 점들의 분포도 랜덤하게 퍼져있으므로, 선형성과 오차의 등분산성 가정은 위배되지 않은 것 같다.

2) Normal QQ plot

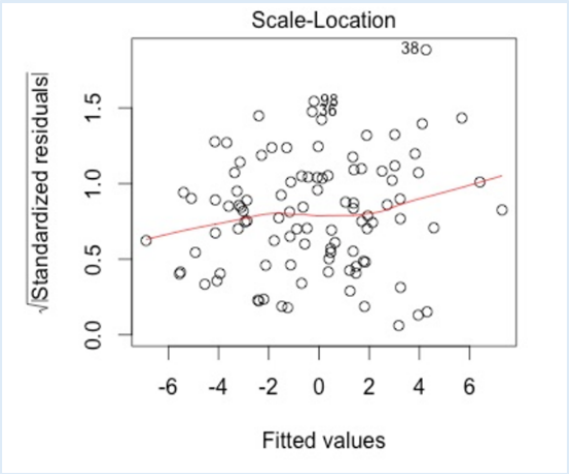


정규성 만족

- X축 : 정규분포의 분위수 값(Theoretical Quantile), Y축 : 표준화 잔차(Standardized residual)

- 정규성 확인 가능
- $y = x$ 그래프에 가까울 수록, 잔차가 정규성을 만족한다. 직선이라는 것은 정규분포 사분위수 위에 그대로 위치한다는 의미니까!
- 해석 : 잔차 분포가 대부분 점선 주변에서 벗어나고 있지 않아 정규성을 만족하는 것 같다. 하지만 38번 관측치가 점선에서 많이 벗어나는 걸 보았을 때 추가적인 확인이 필요하다.

3) Scale - Location

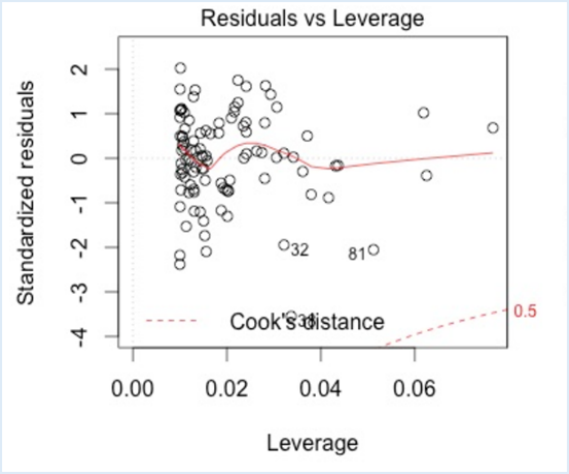


선형성, 등분산성 만족

- X축 : 예측값($\hat{y}$) fitted values, Y축 : 표준화잔차 ($\sqrt{|e_i|/se(e_i)}$)
- 선형성과 오차의 등분산성 확인가능. 보통 등분산성 고려
- 빨간 실선 : 전체적인 잔차들의 추세선 → 잔차들의 분포를 Local Regression으로 추정한 직선이며 잔차 분포의 경향성을 나타내는 보조 지표
- 해석 : 세 번째 플랏은 Scale-Location 플랏으로 이 플랏을 통해 첫 번째 플랏과 비슷한 판단을 할 수 있다. 첫 번째 플랏과의 차이점은 잔차에 절댓값이 씌워진 형태라는 것이다.

  빨간 실선이 완만하게 수평을 이루고 있으며, 점들의 분포도 랜덤하게 퍼져있으므로 선형성과 오차의 등분산성 가정은 위배되지 않은 것 같다.

4) Residuals vs Leverage



- X축 : 레버리지(지렛값), Y축 : 표준화 잔차
- 지난 주에 배운 영향점(influential point)을 확인할 수 있다. 플랏의 오른쪽의 위치한 점들이 leverage가 큰 잔차이며, 빨간 실선으로부터 위아래로 멀리 떨어진 점들이 outlier라고 생각할 수 있다.
- 빨간 점선 : Cook's distance로 주로 0.5과 1에서의 경계가 표시된다.
- 해석 : 이 플랏에서 모든 관측치들이 0.5 경계 안에 있으므로 영향점은 딱히 없는 것 같다.

이제 총 네 가지의 가정에 대해 자세하게 알아보자. 모든 가정은 똑같은 흐름으로 설명할 수 있다. 가정의 의미와 이를 어떻게 진단하는지, 만약 가정이 위배되었을 경우 어떤 문제가 발생하고 어떻게 대처하는지 순서대로 배울 것이다.

3. 선형성 진단과 처방

1) 선형성 가정

- 선형성 가정이란?

  반응 변수(Y)가 설명 변수(X)의 선형결합으로 이루어진다는 가정이다. 지난 주차에 배운 단순선형회귀와 다중선형회귀 모두 선형성 가정에서 출발한 모델이다. 만약 선형성이 위배되었다면, 변수 변환이나 비선형 모델을 추정함으로써 대처할 수 있다.

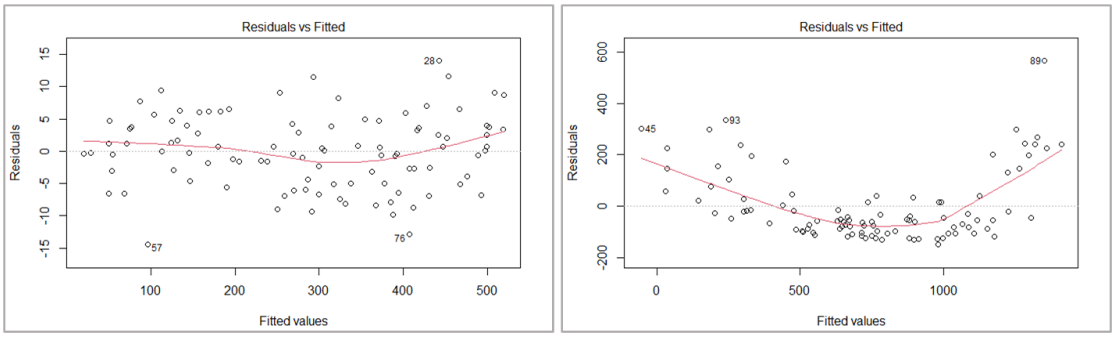
2) 진단

우리 모델의 선형성이 잘 지켜졌는지 어떻게 판단할 수 있을까?

2)-1. 진단 : 잔차 플랏

- 다음과 같이 평균 0을 중심으로 하는 x축에 평행한 직선 형태가 아니라면 선형성이 위반되었다고 볼 수 있다.
- 선형성이 위배되는 보통의 경우, 이차함수 혹은 삼차함수 형태처럼 나타난다.

"Residual vs Fitted" Plot



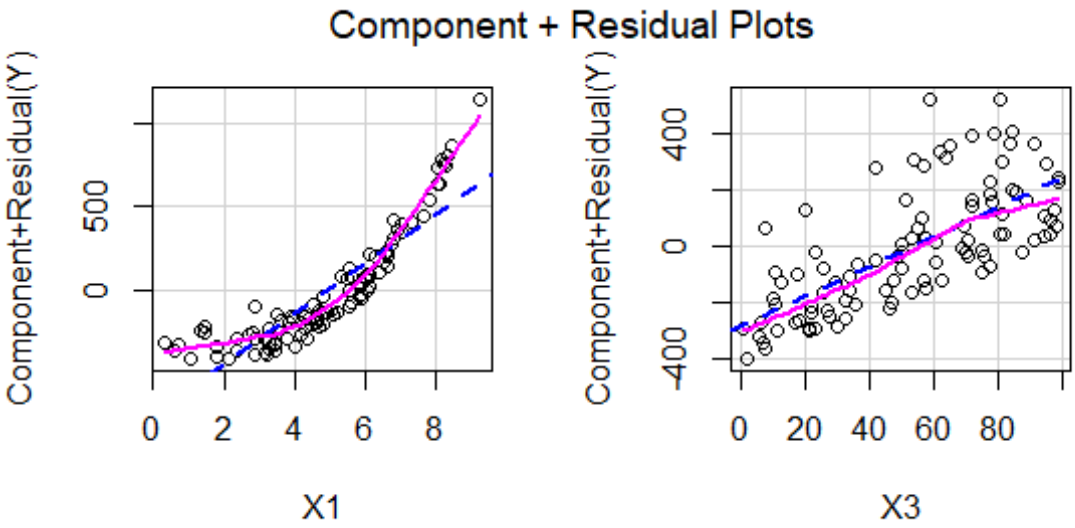
No Pattern Evident / Non-Linearity Evident

- 왼쪽 플랏은 빨간 실선이 X축과 완만하게 수평을 이루고 있으며, 점들의 분포도 랜덤하게 퍼져있으므로, 선형성을 만족한다.
- 오른쪽 플랏은 빨간 실선이 이차 함수 꼴을 보이므로 선형성이 위반되었다고 볼 수 있다.

---

2)-2. 진단 : Partial residual plot

- Car 패키지의 `crPlots()` 함수를 통해 개별 변수의 선형성을 파악할 수 있다.
- 선형성을 만족하지 못할 때, 어떤 변수의 영향으로 인한 것인지 잔차 플랏만으로는 확인하기 어렵다. 따라서 어떤 변수의 영향으로 인한 것인지 개별 변수마다 선형성을 확인해야 한다.
- crPlots에서 시각화 해주는 것은 Partial Regression Plot이다. 1주차 클린업에서 개별 회귀계수 검정 때 다른 변수를 고정시킨 상태에서 해당 변수의 영향력을 본다고 했죠? 그것과 비슷한 아이디어!



- X축 : x_i변수, Y축 : Partial residual ($y - \beta_i x_i$를 제외한 모든 회귀식 성분)

  X축은 선형성을 판단하기 위한 변수를 의미하며, Y축은 전체 모형에서 선형성을 보고싶은 변수를 제외한 나머지 변수로 회귀식을 적합한 후의 잔차이다. 이렇게 일부의 변수로 적합한 모델의 잔차를 이용하기 때문에 Partial residual plot이라고 불린다.

- 파란 점선: Partial residual과 x_i의 적합된 직선. 즉,점들의 분포를 최소제곱방법을 통해 회귀선을 추정한 것
- 보라색 실선: 잔차의 추세선으로, 점들의 분포를 Local regression을 통해 추정한 선이다. 즉, 새로운 변수에 의해 선형적으로 설명되어야 하는 부분
- 해석 : 일반적으로 서로 다른 두 선이 일치하면 선형성이 만족되었다고 판단할 수 있다.
  - (왼쪽 플랏) X1과 Y의 관계는 선형이 아니다.
  - (오른쪽 플랏) X3와 Y의 관계는 선형인 듯 하다. 즉, X3 변수가 Y를 선형적으로 잘 설명하고 있다.

- 한계

  개별 변수들의 선형성을 판단하기에는 아주 좋은 방법이지만 단점이 존재한다.

  - Y와 개별 X변수들 간의 단편적인 관계를 보여주기 때문에 X변수들 사이에 *교호작용이나 상관관계가 존재하더라도 이를 잡아내지 못한다.
  - 심각한 다중공선성이 존재할 경우 잘못된 정보를 제공할 수 있다.

  *교호작용 : 한 요인의 효과가 다른 요인의 수준에 의존하는 경우 즉, 변수간의 시너지 효과

X1과 X2는 Y에 영향을 끼치지 않지만, X1과 X2가 결합됨으로써 Y에 중요한 영향을 끼칠 수 있음
ex) 비타민과 다른영향제를 같이 먹으면 건강이 더 좋아지는 효과

3) 위배되었을 경우 문제점

위 진단에 따라 위배되었다고 판단하는 경우, 우리가 지금까지 배우고 있던 모델은 선형회귀 모델이므로 모델 자체가 성립하지 않는다. 대부분 실제 모델보다 과소추정이 된 경우이며 이 때 예측 성능도 현저히 떨어진다.

4) 처방

4)-1. 처방 : 변수 변환

비선형 관계는 변수 변환을 통해 해결할 수 있다.

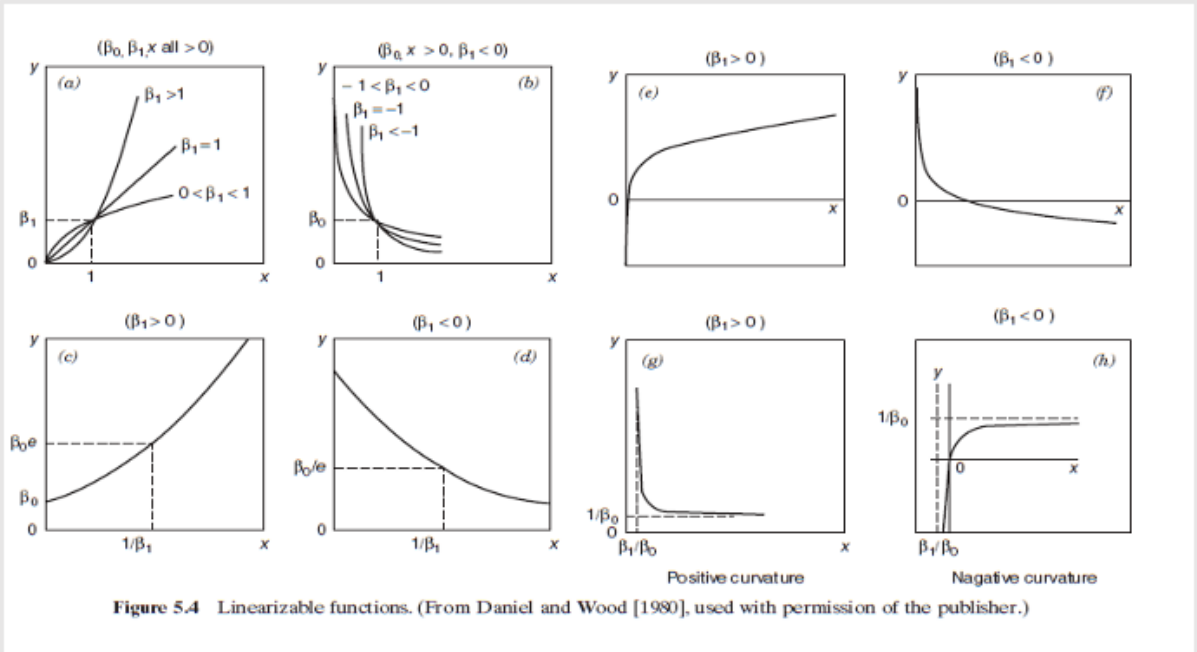


Figure 5.4 Linearizable functions. (From Daniel and Wood [1980], used with permission of the publisher.)

변수를 변환해서 우리가 흔히 생각하는 선형 모형이 만들어진다면, 그 모델은 선형성을 만족했다고 할 수 있다. 예를 들어,

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$
$$y = \beta_0 + \beta_1 log x_1 + \varepsilon$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \varepsilon$$
$$y = \beta_0 e^{\beta_1 x_1} \;\rightarrow\; y^* = log y = log\beta_0 + \beta_1 x_1 = {\beta_0}^* + \beta_1 x_1$$

- 변환된 x를 새로운 x로 취급한다면, 이 모든 결합들은 선형결합을 만족한다. 또한 승법 모형(곱) 또한 y에 log변환을 할 경우 가법모형으로 변환 가능하다.
- 단, $y = \frac{\beta_1 x}{\beta_0 + x}$ 와 같은 형태라면, 이는 변환을 통해 선형을 만들 수 없기 때문에 비선형 모델이다.

예시로 아래와 같이 변수 변환하여 비선형성을 해결할 수 있다.

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

이렇게 변수 변환을 통해 선형성을 확보할 수 있는 모델도 넓은 의미에서 선형 모형이라고 부른다. 그렇기 때문에 포아송, 음이항, 로지스틱 회귀 모델을 일반화 선형 모형이라고 부른다.

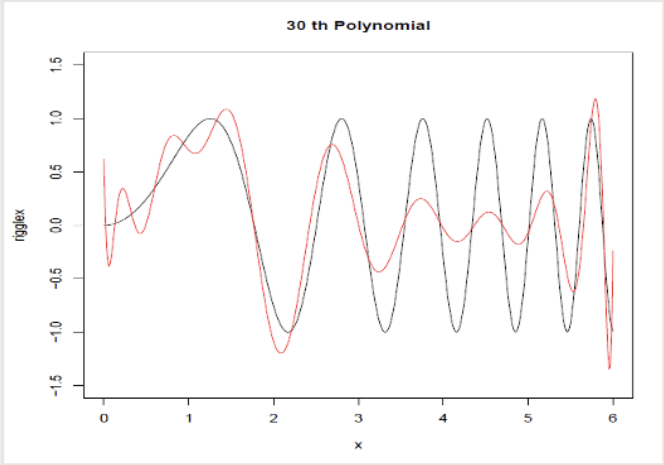
---

4)-2. 처방 : 비선형 회귀

모델 자체를 비선형 회귀 모델에 적합시키는 방법을 사용할 수 있다. 비선형 회귀 모델이란 선형 회귀 모델이 아닌 모든 모델을 일컫는 말로 그 종류는 매우 다양하다. 몇 가지만 소개하고자 한다.

(1) 비선형 회귀 : Polynomial Regression

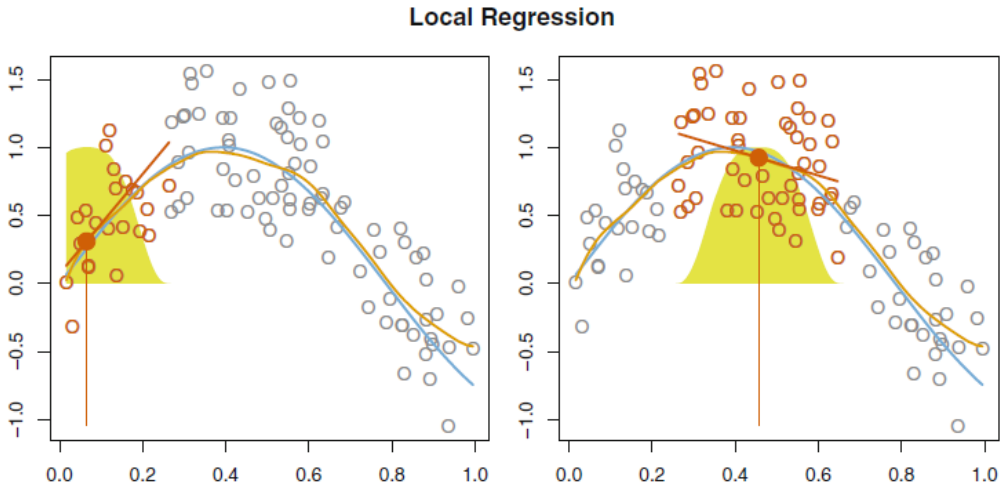
- 고차항을 고려하는 다항 회귀이다.
- 잔차 플랏이나 Partial residual plot을 보았을 때 이차 이상의 곡선 형태가 나타날 경우, 이를 설명하기 위해 변수의 차수를 다양하게 바꾸어 모델에 넣어주는 Polynomial Regression을 통해 해결 가능하다.
- 하지만 초고차항을 고려하여 모델을 적합하더라도 경향을 잘 잡아내기는 힘들어 삼차까지만 고려한다.



초고차항을 적합해도 경향을 못잡아내기 때문!

(2) 비선형 회귀 : Local Regression

- 비선형 회귀 방법이자 비모수적 방법을 사용하는 회귀 모델이다. 앞서 알아보았던 잔차 플랏의 추세선을 나타낼 때에도 쓰인다. 간단하게 아이디어만 알아보도록 하자.
- Local Regression은 말 그대로 Local(지역적인)에 있는 데이터들로 회귀 모델링을 하는 방법이다. 타겟 데이터 x_0를 중심으로 그 주변의 k개의 이웃 데이터 $x_i \in \mathcal{N}(x_0)$들만 사용하여 부분적으로 회귀 모델을 구성한다. KNN 알고리즘과 비슷하지만 모든 k개의 이웃에 각기 다른 가중치를 부여한다. 타겟 데이터 x_0와 가깝다면 큰 가중치를, 멀다면 작은 가중치를 부여한다. 이 가중치는 주로 정규분포와 비슷하게 생긴 Radius Basis Function(RBF) 혹은 tri-cubic function 에 기반하여 산정된다.



4. 정규성 진단과 처방

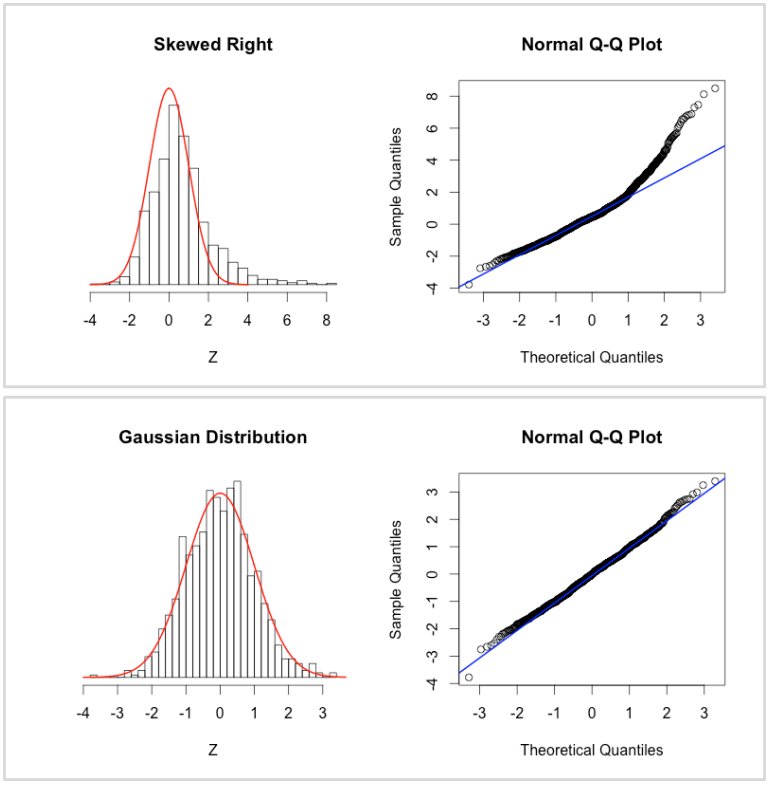
1) 정규성 가정

반응 변수 Y를 측정할 때 발생하는 오차는 정규분포를 따를 것이라는 가정이다. 우리의 회귀식이 데이터를 잘 표현한다면 잔차들은 단순한 측정 오차, Noise라 여겨지고, 이 잔차들의 분포는 정규분포와 흡사한 형태가 될 것이다. → 정규성 가정은 쉽게 위반되지 않음

2) 진단

2)-1. 진단 : Normal Q-Q plot

정규성을 파악하기 위한 대표적인 비모수적인 방법이다. R에서는 회귀식에 `plot()` 함수를 쓰면 두 번째로 나오는 플랏으로, 점들이 $y = x$ 직선에 가까우면 정규성을 만족한다.



(위) 분포가 정규성을 만족하지 않는 경우
(아래) 분포가 정규성을 만족하는 경우

2)-2. 진단 : 정규성 검정

하지만 플랏으로 확인하는 경우에는 판단이 주관적일 수밖에 없다. 그래서 너무 명확하게 정규성을 만족하거나, 만족하지 못하는 경우를 제외하고는 통계적 방법에 의한 가설검정을 통

해 확인하는 것이 더 객관적이다.

- 가설

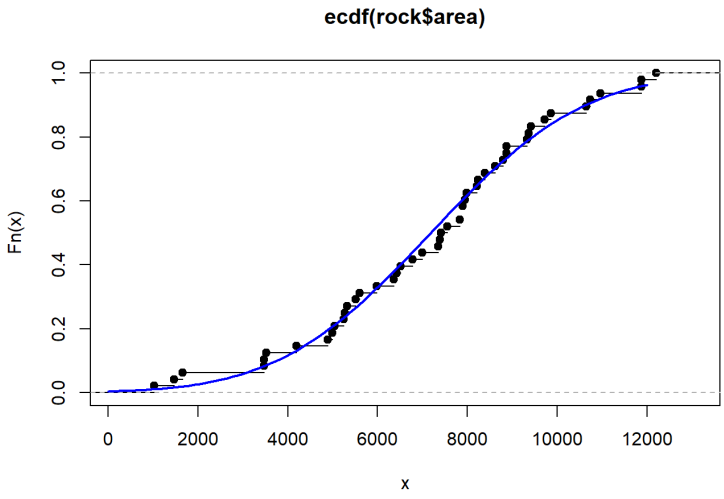
  H_0 : 주어진 데이터는 정규분포를 따른다.

  H_1 : 주어진 데이터는 정규분포를 따르지 않는다.

  → 우리가 원하는 것은 귀무가설을 기각하지 못하는 것 = 정규분포를 따르는 것이겠죠?

(1) 정규성 검정 : Empirical CDF를 이용한 Test

Empirical CDF(ECDF, 경험적 누적밀도함수) 란 아래 그림처럼 관측치들을 작은 순서대로 나열한 후 관측치들로 누적 분포 함수를 그린 것이다.



정규성 검정을 하기 위해 잔차의 ECDF와 정규분포의 CDF를 비교하여 검정한다. 검정 방법은 크게 두 가지로, Anderson Darling Test와 Kolmogorov Smirnov Test가 있다.

- ***Anderson Darling Test***

```
#Anderson Darling Test
library(nortest)
ad.test(fit$residuals)
```

- ***Kolmogorov Smirnov Test***

```
#Kolmogorov Smirnov Test
ks.test(x=fit$residuals, y="pnorm") # Kolmogorov Smirnov Test는 y의 입력값에 따라
 다른 분포와의 비교도 가능하다.
```

자세한 방법은 다르지만 모두 잔차의 ECDF를 이용해 정규분포와 비교한다는 점에서 같다.

(2) 정규성 검정 : 정규분포의 분포적 특성을 이용하는 Test

- ***Shapiro Wilk Test***
  - QQ plot의 아이디어와 동일하게, 정규분포 분위수 값과 표준화 잔차 사이의 선형관계를 확인하는 검정이다.
  - 관측치가 5000개 이하인 데이터에서만 가능하다.
  - R 기본함수로 내장되어 있으며, `shapiro.test()` 함수 안에 residual 값을 넣으면 된다.

```
#Shapiro Wilk Testt
shapiro.test(fit$residuals)
```

- ***Jarque-Bera Test***
  - 정규분포의 왜도가 0, 첨도가 3이라는 점에 기반하는 방법이다.
  - 잔차의 분포가 정규분포와 달라질수록 왜도나 첨도의 변화가 생기고, 통계량 값이 커져 유의수준을 넘어서면 귀무가설을 기각하게 된다.
  - 수식 : $JB = n(\frac{(\sqrt{skew})^2}{6} + \frac{(kurt-3)^2}{24})$
  - tseries 패키지 안에 있는 `jarque.bera.test()` 함수를 사용해 잔차를 넣어주면 된다.

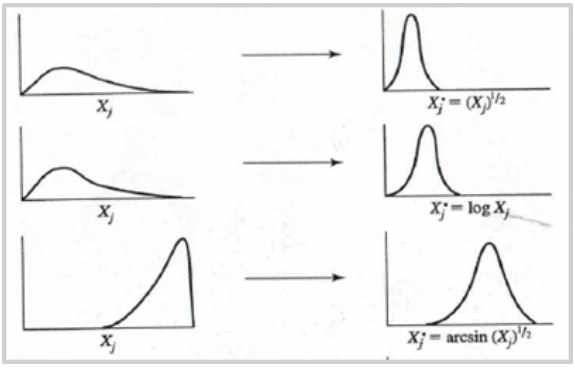
```
#Jarque-Bera Test
library(tseries)
jarque.bera.test(fit$residuals
```

3) 위배되었을 경우 문제점

선형 회귀에서 오차의 정규성을 가정하기 때문에 정규분포로부터 파생되는 표본분포 t분포, F분포를 사용할 수 있다. 즉, 회귀분석에 사용되는 F-test와 t-test 검정은 정규분포를 전제한다. 하지만 위와 같은 진단 결과에 따라 오차가 정규분포를 따르지 않아 정규성 가정이 위배되었다면, 검정통계량이 t분포 또는 F분포를 따르지 않게 되므로 가설 검정 결과가 p-value에 의해 유의하게 나오더라도 검정 결과를 신뢰할 수 없게 된다. 또한 예측의 결과를 신뢰하기도 어렵다.

4) 처방

4)-1. 처방 : 변수 변환



분포를 보고 자의적으로 위와 같이 여러 방법으로 변수 변환을 해줄 수 있다. 하지만 어디까지나 주관적인 판단 하에 이루어지므로 객관성을 확보하기는 어렵다.

(1) 변수변환 : Box-Cox Transformation

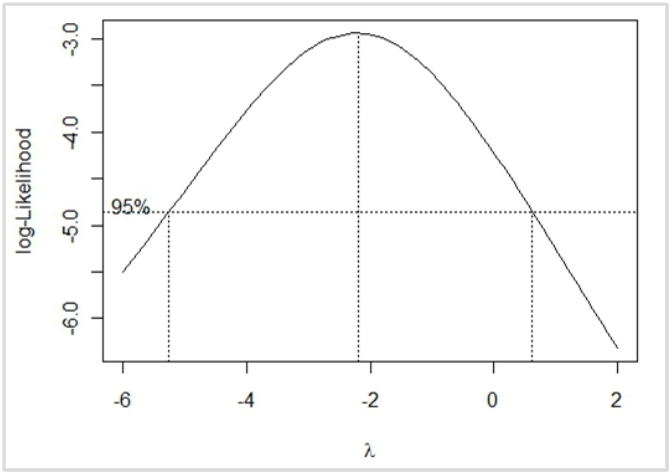
Y를 변환함으로써 정규성 혹은 뒤에 나올 등분산성을 해결해주는 방법이다. 위 변수변환의 단점을 보완하여 통계적인 검정에 따라 변수를 변환한다는 점에서 효율적이다!

아래 식에서 λ를 변화시키면서 y가 정규성을 만족하도록 만드는 것이다. 일반적으로 λ는 -5에서 5사이의 값을 사용한다. λ가 0인 경우에는 log-transformation을 해준다.

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

이 때 최적의 λ (람다)는 아래와 같이 ML방법을 통해 신뢰구간을 구한 후에 신뢰구간 내의 로그우도함수(ML)를 최대화하는 λ를 최적의 값으로 선택한다.

R 코드를 통해 아래와 같은 결과를 시각화 할 수 있다. 아래의 예시에서는 95% 내의 λ값 중 가능도함수가 최대가 되게하는 -2 근방의 λ를 선택하면 될 것 같다. 또는 -2를 λ로 선택해도 된다. 이렇게 정수로 λ를 선택한다면 제곱, 역수, 제곱근 등 변수 변환 관계를 쉽게 비교적 알 수 있다는 장점이 있다.



<R 코드>

```
#load car library
library(car)

#take a value of lambda
trans = powerTransform(data$variable)
summary(trans)
```

Box-Cox Transformation은 y가 $log(y)$로 변환 될 수 있으므로, Y가 0 이하일 때에는 사용할 수 없다. 그럴 때에는 다음 방법을 사용하여 y를 변환해줄 수 있다.

(2) 변수변환 : Yeo-Johnson Transformation

Box-Cox transformation과 동일한 아이디어.

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

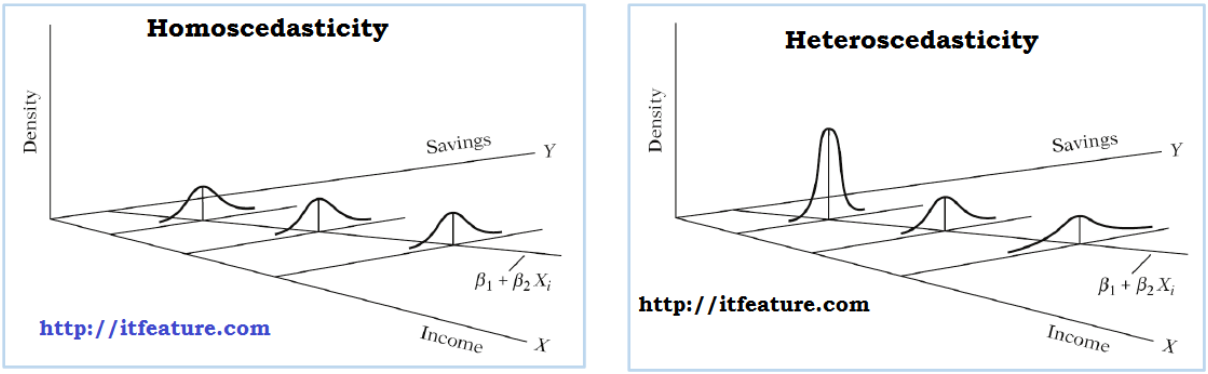
<R 코드>

```
#take a value of lambda
trans = powerTransform(data$variable, family = "yjpower")
summary(trans)
```

5. 등분산성 진단과 처방

1) 등분산성 가정

- 오차의 분산은 σ^2로 동일하다. 이를 등분산(Homoskedasticity)라고 하고, 이 가정이 깨지면 이분산(Heteroskedasticity)라고 한다.



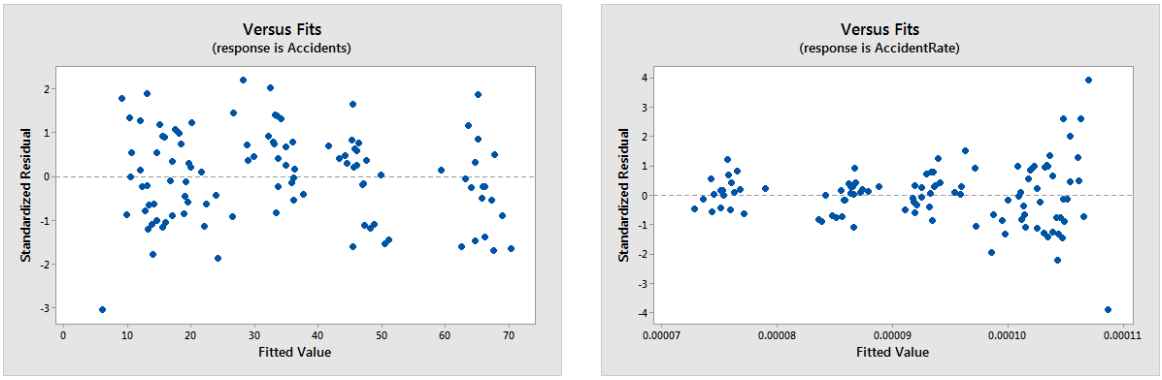
(왼)등분산 / (오)이분산

오차항은 확률변수이므로 기댓값과 분산을 가진다. 이때 오차항의 분산이 상수(어느 관측치에서나 동일하며, 다른 변수의 영향을 받지 않음)라는 가정이 바로 등분산성 가정이다. 영어로 Homoscedasticity라고 한다. 위 그림을 보면 오차의 분산이 등분산인 경우는 왼쪽 그림처럼 데이터의 어느 곳에서나 y의 조건부 분포의 모양이 같을 테지만, 등분산이 아니라면 오른쪽 그림처럼 지점에 따라 y값이 가지는 조건부 분포가 들쭉날쭉할 수 있다.

2) 진단

2)-1. 진단 : 잔차 플랏

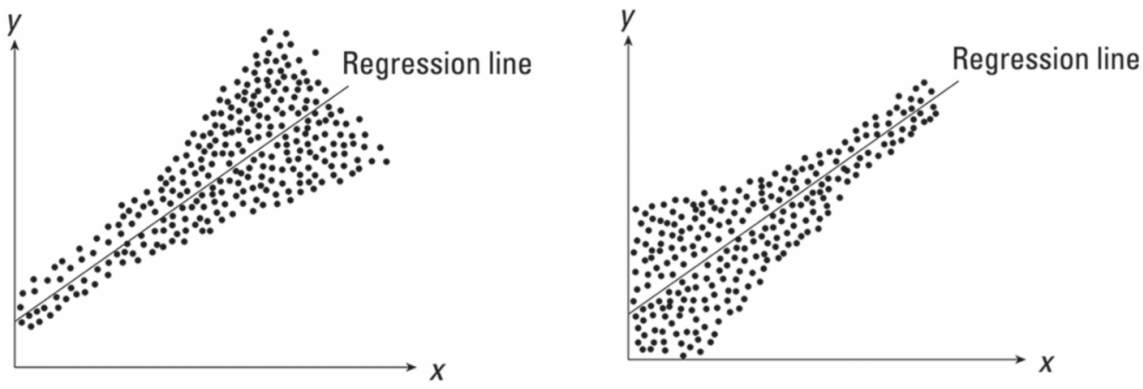
대표적으로 잔차 플랏의 첫 번째 플랏인 'residual vs fitted' 플랏과 세 번째 플랏인 'scale-location' 플랏을 종합적으로 보고 판단하여 확인할 수 있다.



(좌) 등분산 / (우) 이분산, fitted value에 quadratic하게 비례하는 모습

위의 예시의 오른쪽은 $\hat{y}$값이 커짐에 따라 절대값이 커지는 잔차를 갖는 이분산성을 띄는 데이터이다. 수많은 예시 중 하나일 뿐이며, 퍼짐의 정도가 일정하지 않고 퍼짐이 증가하거나 감소하거나, 혹은 X 평균 부분의 퍼짐이 큰 형태 등 다양한 이분산의 형태가 존재한다.

이분산성(Heteroscedasticity)을 띄는 데이터를 단순선형회귀의 산점도에서 확인해볼 수 있으며 다음 그림과 같다.



이분산의 형태

- 이렇듯 퍼짐의 정도가 일정하지 않고, 퍼짐이 증가하거나 감소하거나, 혹은 x평균 부분의 퍼짐이 큰 형태 등 다양한 이분산의 형태가 존재한다.
- 이처럼 그래프 상으로 명확하게 나타나는 이분산 형태도 있지만, 육안으로 판단하기 어려운 이분산의 형태도 있기 때문에 이를 판단하기 위한 테스트 방법들을 존재한다. 이어서 배워보자!

2)-1. 진단 : Breusch-Pagan test (BP test)

오차의 추정량인 잔차가 독립변수들의 선형결합으로 표현되는지 아닌지를 검정하기 위한 아이디어에서 출발한 방법이다. 즉 우리가 원하는 오차의 분산이 등분산인지 아닌지 판단하는 검정 방법이다.

- 분산이 설명변수(X)에 대한 선형결합으로 되어있다는 가정을 바탕으로 한다.
- 분산과 설명변수(X) 간에 세운 회귀식의 결정계수 값(R^2)이 높으면 등분산이 아니게 된다. 즉, 이분산성을 지닌다.
- 단점은 분산과 X변수가 선형결합으로 이뤄졌다는 가정을 바탕으로 하기 때문에, 비선형결합으로 만들어지는 이분산성을 잡아낼 수 없다.
- 가설
  - H_0 : 주어진 데이터는 등분산성을 지닌다.
  - H_1 : 주어진 데이터는 등분산성을 지니지 않는다. (이분산성이다.)

잔차의 제곱이 독립변수의 선형결합으로 표현되는지, 그때의 설명력은 어느 정도인지 결정 계수를 통해 파악한다. 만약 오차가 독립변수에 의해서 충분히 표현된다면 결정계수는 커질 것이고, 검정통계량 또한 커질 것이다.

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

위 식에서와 같이 X 변수를 선형결합 시킨 식에서 R^2을 구한 뒤

(검정통계량) $\chi^2_{stat} = nR^2 \sim \chi^2_{p-1}$
$Reject \quad H_0 \quad if \quad \chi^2_{stat} > \chi^2_{p-1,\alpha}$ (임계값)

카이스퀘어 통계량을 구하고, 이 통계량이 임계값보다 크면 귀무가설을 기각한다. → 즉, 이분산성 존재한다는 의미

- 검정 통계량

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

$\chi^2_{p-1,\alpha}$
- 단점

  비선형적인 결합으로 이루어진 이분산성은 파악할 수 없으며, 샘플이 대표본이어야 사용 가능하다.

R에서는 lmstat 패키지의 `bptest()` 함수에 적합된 회귀식을 넣으면 된다.

```
#load lmtest library
library(lmtest)

#perform Breusch-Pagan Test
bptest(model)
```

혹은 car 패키지의 `ncvTest()` 를 사용하면 된다.

```
#load car library
library(car)

#perform NCV Test
ncvTest(model)
```

3) 위배되었을 경우 문제점

위배되었을 경우 이분산성(Heteroscedasticity)을 가진다고 하고 다음과 같은 문제가 발생한다.

- 이분산은 추정량의 분산을 증가시키지만, OLS 추정량은 이를 잡아내지 못한다. 따라서 이분산이 있는 경우 OLS 추정량의 분산은 실제 분산보다 작게 추정된다. → 이렇게 과소추정된 분산은 검정통계량을 크게 만들고(검정통계량에서 추정량의 분산이 분모에 들어가니까!), p-value를 작게 만들어 실제로는 유의하지 않은 변수를 유의하다고 나타낼 수 있게 된다.
- 이를 가설검정의 관점에서 말하면, 충분히 유의할 수 있는 귀무가설을 기각하는 과소추정 즉, 제 1종 오류(Type 1 error)가 범하게 되며 가설검정의 신뢰성을 떨어뜨린다.
- 지난 주에 배운 내용으로 설명하면 몇몇의 가정이 존재하기 때문에 가우스 마코프 정리에 의해 OLS 추정량은 BLUE가 되지만, 등분산성이 위배된 경우 더이상 BLUE가 아니게 된다.
-

4) 처방

4)-1 처방 : 변수 변환 (Box-Cox Transformations)

정규성을 만족시키기 위해 사용했던 각종 변수 변환 방법 등을 똑같이 적용할 수 있다.

4)-2. 처방: 가중 회귀 제곱 (WLS; Weighted Least Square)

등분산이 아닌 형태의 데이터마다 다른 가중치를 주어서 등분산을 만족하게 해주는 '일반화된 최소제곱법'의 한 형태이다. 분산이 커 신뢰도가 낮은 부분의 관측치에는 가중치를 적게 주어 분산을 작게 만들고, 분산이 작아 신뢰도가 높은 부분의 관측치에는 가중치를 크게 주어 분산을 크게 맞춰주는 방식이다.

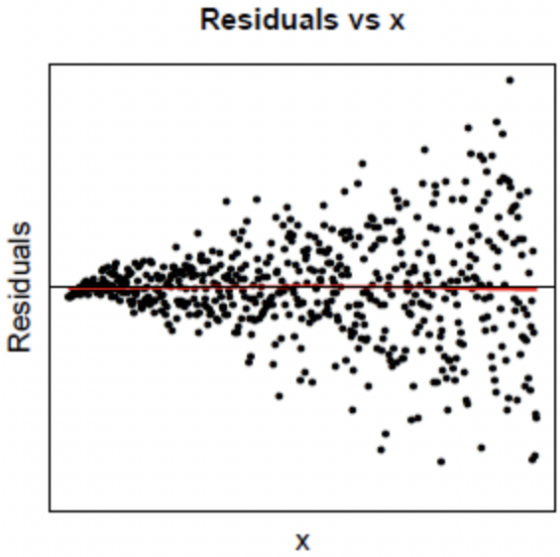
- **공식**

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

  여기서 가중치 w_i는 분산에 반비례하는 가중치이다. 따라서 WLS 방법을 사용할 경우 작은 가중치를 가지는 관찰값은 회귀계수 $\beta_0, \beta_1, \beta_2, ..., \beta_p$의 값을 결정하는 데 적은 영향을 미치게 된다. 이 때 우리는 총 n개의 잔차들의 분산을 알기 어렵기 때문에 경험적으로 선정해야 한다.

- **가중치 선정 방식 : 잔차플랏**

  명확한 패턴이 있는 경우 주로 잔차 플랏을 이용하여 판단하고, 그렇지 않다면 경험적으로 판단하여 가중치를 산정한다.



  이렇게 residual plot에서 분산이 점점 커지는 이분산성을 띄는 경우에는, $w_i \propto \frac{1}{x_i^2}$ 와 같은 방식으로 가중치 사용!

- **장점**

WLS를 통해 구한 추정량은 회귀 기본 가정 하에 다시 BLUE임!

---

6. 독립성 진단과 처방

1) 독립성 가정

- 독립성 가정이란 오차들은 서로 독립이라는 가정이다. 즉 개별 관측치에서 i번째 오차와 j번째 오차가 발생하는 것에 서로 영향을 미치지 않는다는 가정이다. 이를 식으로 표현하면 $Cov(e_i,\ e_j) = 0$인데 공분산이 0이라고 해서 반드시 독립인 것은 아님을 주의하자. 독립성 가정이 위배되었을 경우를 오차들 간의 자기상관(autocorrelation)이 있다고 한다. 즉 일종의 패턴을 지니는 것이다. 우리의 모델이 데이터를 잘 설명한다면, 설명하고 남은 잔차가 특정 패턴을 지니지 않는다. 그런데 시간적, 공간적으로 인접한 관측치들은 유사한 경향을 가지기 때문에 회귀식만으로 설명되지 않는 패턴이 남아 있을 수 있다.

  (시간적으로 자기상관O → 시계열분석을 이용/ 공간적으로 자기상관O → 공간회귀를 통해 접근)

2) 진단

2)-1. 진단 : Durbin Waston Test

더빈-왓슨 검정은 바로 앞 뒤 관측치의 1차 자기상관성(first order autocorrelation)을 확인하는 검정 방법이다.

- 가설

  H_0 : 잔차들이 서로 독립이다. (자기상관성이 없다)

  H_1 : 잔차들이 서로 독립이 아니다. (자기상관성이 있다)

- 검정 통계량

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

  $\text{First order autocorrleation } : \hat{\rho}_1 = \frac{\hat{Cov}(e_i, e_{i-1})}{\sqrt{V(e_i)} \cdot \sqrt{V(e_{i-1})}} \approx \frac{\sum_{i=1}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}$

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

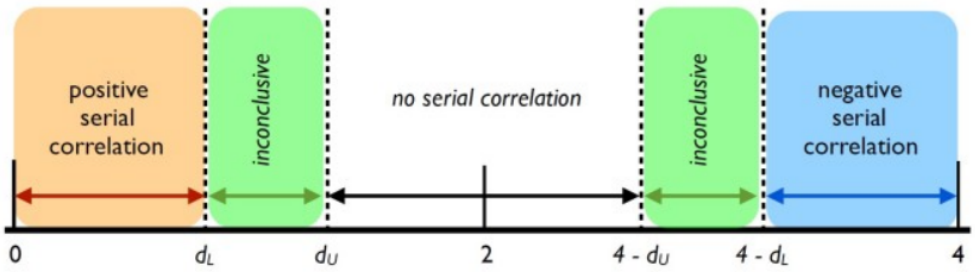
  이때 $\hat{\rho}_1$는 표본 잔차 자기상관(sample autocorrelation of the residuals)을 나타내는데, 이는 -1부터 1사이의 값을 갖는 e_i와 e_{i-1}의 상관계수의 꼴로 생각할 수 있다.

  따라서 더빈왓슨 통계량 $d \approx 2(1 - \hat{\rho}_1)$ 는 0~4 까지의 범위를 갖는다.

```
#Durbin Watson test
library(car)
durbinWatsonTest(fit) # fit은 적합된 모형
```

- 해석

  이 더빈 왓슨 검정의 귀무 가설은 '1차 자기상관이 없다'이다. 다시 말해서, $\hat{\rho}_1$ = 0 이다. 1차 자기상관계수는 더빈 왓슨 통계량과 근사한 선형관계가 존재하므로, $\hat{\rho}_1$이 0에 가까워진다면 d는 자연스레 0이 가까워질 것이다. 하지만 애매한 경우에는 어떻게 판단할까? 더빈 왓슨 검정 표가 존재한다. 이 표는 데이터의 개수 n과 변수의 개수 p에 따라 이 귀무가설을 기각할 수 있는지 없는지 판단하는 컷오프 값을 알려준다. 흔히 이 경계의 하한은 d_L 상한은 d_U 라고 표현한다.



  위 그림을 통해 편의상 한 방향만 설명하자면, 우리가 구한 검정 통계량 d가 하한보다 작다면 귀무가설을 기각할 수 있다. 이때 양의 자기상관(e_2가 e_1에 비해 커졌고, e_3도 e_2에 비해 커지는 경우, 앞 오차에 영향을 받는 경우라고 생각하면 된다)이 있다고 판단할 수 있다. 반대로 검정 통계량 d가 상한보다 크다면, 귀무가설을 기각시키지 못하게 된다.

  d가 0에 가까울수록 양의 상관관계를, 4에 가까울수록 음의 상관관계를 나타낸다. 2에 가까운 값이어야 귀무가설을 기각하지 못한다.

- 한계

  - d가 상한과 하한 사이에 위치하게 된다면 우리는 판단할 수 없다.

  - 바로 인접한 오차와의 1차 자기상관 즉, 첫 번째 순서의 자기상관성만 고려한다.
  →AR(1)구조만 파악할 수 있다는 것! 만약 자기상관이 오래 지속되거나, 계절성이 있는 경우 이를 확인하는 데에 한계가 있다.

    → AR모형에 대한 자세한 내용은 시계열팀 클린업 참고!

- R 코드

```
#train a linear model
fit<-lm(y~x1+x2+x3)

#perform Durbin Watson test 1
library(lmtest)
dwtest(fit) # fit은 적합된 모형

#perform Durbin Watson test 2
library(car)
durbinWatsonTest(fit$residuals)
```

3) 위배되었을 경우 문제

진단 결과 오차의 독립성이 위배되었을 경우, 즉 자기상관이 발생했을 경우 다음과 같은 결과가 발생한다.

- LSE의 가정 세가지를 만족하지 못하므로, 최소제곱추정량이 더 이상 BLUE가 아니게 된다.
- $\hat{\sigma}^2$의 추정량과 회귀계수의 표준오차가 실제보다 심각하게 과소추정된다. 따라서 유의성 검정의 결과를 신뢰할 수 없고, Prediction Interval도 넓어지게 된다.

4) 처방

4)-1. 가변수 만들기

- 뚜렷한 계절성이 있다고 판단되면, 이를 위해 가변수를 만든다. 계절성이 주기를 가진다는 점을 이용하여 주기 함수인 삼각함수 $cos(t), sin(t)$의 선형결합으로 주기를 표현하는 방법이다. 변수에 주기를 표현하는 가변수를 만듦으로써 대처가 가능하다.

4)-2. 분석 모델 변경

- 시간에 따라 자기상관을 가질 경우 자기상관을 고려하는 AR(p) 같은 시계열 모델을 사용한다. (시계열분석팀 클린업 참고~)
- 공간에 따라 자기상관을 가질 경우 공간의 인접도를 고려하는 공간회귀모델을 사용하기도 한다.

🌟gvlma package

(Global Validation of Linear Model Assumption)

선형성, 정규성, 등분산성을 한 번에 체크해주는 함수!

1. **Global Stat** : Are the relationships between your X predictors and Y roughly linear? Rejection of the null (p < .05) indicates a non-linear relationship between one or more of your X's and Y. → 선형성
2. **Skewness** : Is your distribution skewed positively or negatively, necessitating a transformation to meet the assumption of normality? Rejection of the null (p < .05) indicates that you should likely transform your data. → 정규성
3. **Kurtosis** : Is your distribution kurtotic (highly peaked or very shallowly peaked), necessitating a transformation to meet the assumption of normality? Rejection of the null (p < .05) indicates that you should likely transform your data. → 정규성
4. **Link function** : Is your dependent variable truly continuous, or categorical? Rejection of the null (p < .05) indicates that you should use an alternative form of the generalized linear model (e.g. logistic or binomial regression). → 선형성
5. **Heteroscedasticity** : Is the variance of your model residuals constant across the range of X (assumption of homoscedasticity)? Rejection of the null (p < .05) indicates that your residuals are heteroscedastic, and thus non-constant across the range of X. Your model is better/worse at predicting for certain ranges of your X scales. → 등분산성

- R 코드

```
assumptionTest<- gvlma(carsModel)
summary(assumptionTest)
```

- 한계

  gvlma를 이용하면 간편하기는 하지만, statistical testing 기법이 갖는 한계점처럼 유의수준 0.05에서 [가정 충족 || 가정 충족하지 않음]의 경계를 잘라 버리다 보니 융통성이 부족하다는 점이 있다. 선형회귀는 이런 가정 충족에 대해서 비교적 robust 한 편이다 보니 이 결과만 보고 비선형적 모델로 바로 넘어가는 등의 속단은 위험할 수 있다.