Week 2 : 회귀분석의 가정

<목차>

---

0. 복습

1주차 클린업에서 회귀분석의 기본 개념에 대해 알아보는 시간을 가졌습니다.

잠시 복습을 해보자면,

- 회귀분석이란 변수들 간의 관계를 모델링하는 통계적 기법으로, 상관관계 기반의 모델링이다. 하나의 X변수만을 고려하는 단순선형회귀와 여러 X변수들을 고려하는 다중선형회귀가 있다.
- 회귀계수의 추정은 최소제곱법을 통해 진행한다. 오차의 정규성 가정이 있는 경우 LSE는 MLE와 동일한 추정량을 산출하며, 특정 조건이 만족되면 LSE는 BLUE가 된다.
- 다중회귀분석에서의 F 검정은 회귀식 자체에 대한 검정을, Partial F-test는 일부 회귀 계수에 대한 검정을, t-test는 다른 변수를 고정시킨 상태에서 개별 변수의 유의성을 검정한다.
- 회귀식의 적합성 검정은 R^2를 통해 확인할 수 있는데, 단순히 변수 개수에 따른 결정계수 증가를 방지하기 위해 ${R_{adj}}^2$로 변수 증가에 따른 페널티를 부과한다.
- 회귀분석은 이상치에 민감한 경향을 가지기 때문에, Outlier, Leverage, Influential Point를 확인하며 데이터를 진단하는 것이 좋다.
- 이상치가 많을 경우, Median Regression, M-estimation, Least Trimmed Square 등의 로버스트(이상치에 강건한) 모델을 고려할 수 있다.

2주차 클린업에서는 회귀분석의 가정에 대해 확인하면서, 만약 기본 가정이 위배되었을 경우 발생할 수 있는 문제점과 각각에 대한 처방법에 대해 알아보겠습니다. 이번 주도 양이 많지만,,, 조금만 힘내봅시다 !!

---

1. 회귀 기본 가정

1) 회귀 기본 가정이 가지는 의미

- 회귀분석은 적은 수의 관측치만으로도 모델을 구성할 수 있고, 좋은 예측과 추정이 가능하다는 장점이 있다. 회귀분석이 이런 장점을 가지는 것은 그만큼 많은 제약들이 예측력과 설명력을 뒷받침하고 있기 때문이다. 선형회귀모델이 만족해야 하는 제약을 '선형회귀의 기본가정'이라고 하며 다음과 같다.

  → 모델의 선형성, 오차의 등분산성, 오차의 정규성, 오차의 독립성

- 회귀분석은 많은 머신러닝 모델의 기반이 되며, 따라서 머신러닝 모델들에도 이러한 가정을 필요로 하는 경우가 있다. 모델의 가정은 모델이 만들어진 형태와 직접적으로 연관되어 있기 때문에 가정이 지켜지지 않으면 모델의 성능이 급락하는 경우가 많다. 오늘 회귀분석의 가정의 진단과 처방하는 과정을 배우며 모델에 대한 이해를 바탕으로 모델을 활용하기 위한 기초를 쌓아보자.
- 회귀는 결국 변수 간의 관계를 추정하는 과정이다. 우리는 분석을 통해 **오차의 평균이 최대한 0에 가까워지는** 정확한 회귀모델을 만드는 것이 궁극적인 목표이다.

  하지만 현실에서는 우리가 추정한 모델과 실제 데이터 사이에서 차이가 발생하게 되는데, 이러한 이유가 **Case 1) 모델링을 할 때 미처 고려하지 못한 속성들** 때문인지, 혹은 **Case 2) 현실 세계의 여러 오차(잡음)**에 의해서 인지 확인해야 한다.

---

2) 회귀분석의 기본가정

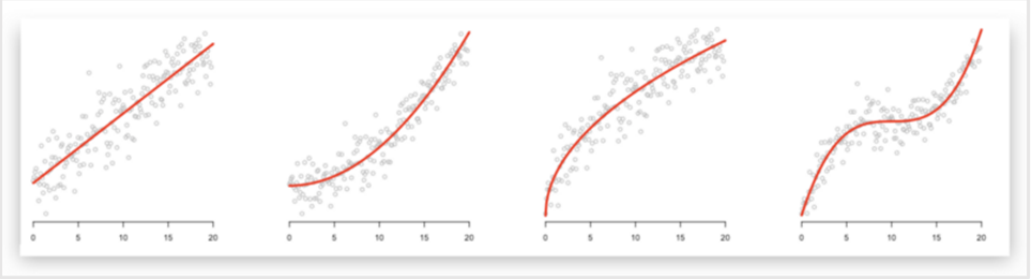
가정은 크게 변수에 대한 가정과 오차항에 대한 가정으로 구분할 수 있다.

- 변수에 대한 가정

  - **선형성(Linearity)** : 설명변수와 반응변수의 관계는 선형이다. (모델 자체가 선형성만 고려하고 있다)

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

  치환의 과정을 통해 변화된 x를 새로운 x로 취급한다면, 위 결합들을 모두 선형결합으로 이해할 수 있다. (선형성을 만족)

  그러나 $y = \frac{\beta_1 x}{\beta_0 + x}$와 같은 형태라면, 변환을 통해 선형을 만들 수 없어 비선형모델이다.

- **설명 변수의 독립성(No or little muliticollinearity)** : 설명 변수들은 서로 독립이다. (다중공선성과 관련, 뒤에서 더 자세히 살펴보자.)

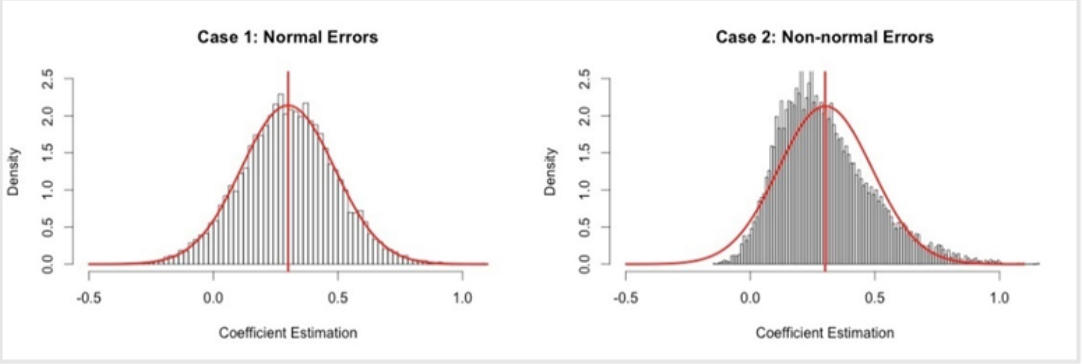
- 설명변수가 확률변수가 아니다 : 이는 측정에 오차가 존재하지 않는다는 가정이다. 이를 객관적으로 확인하는것은 어렵지만, 이 가정이 결과의 해석에는 도움을 준다. - Appendix에서 계속

- 오차항에 대한 가정

- 오차의 평균은 0이다. - Appendix에서 계속

- **오차의 정규성(Normality)** : 오차항은 정규분포를 따른다.

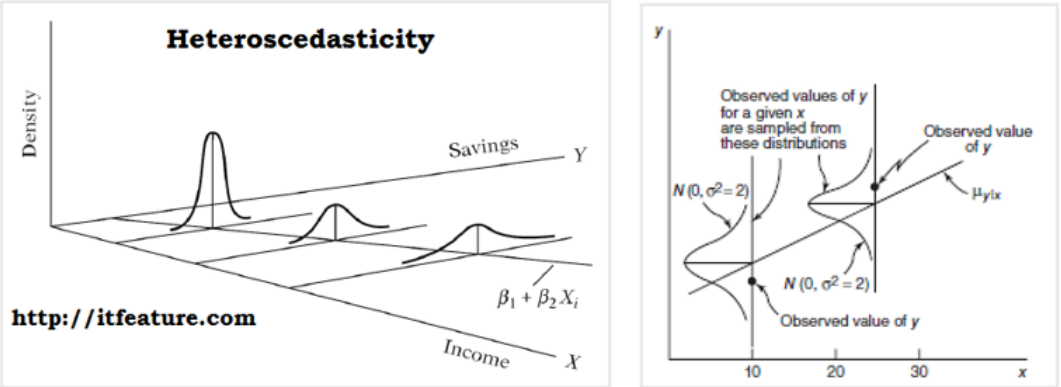
: 정규분포가 오차에 대한 확률분포이기 때문에, 회귀식이 데이터를 잘 표현하고 있다면 오차들은 단순 잡음(noise)가 되어 정규분포에 근접하는 형태가 나올 것이다.



: 오차의 정규성을 가정하기 때문에 회귀식과 개별 회귀계수에 대한 검정을 시행할 수 있다.

→ 정규분포를 따르지 않을 경우 가설검정에서 분포가 왜곡될 것이고, 이에 따라 검정결과를 신뢰할 수 없다. (Prediction의 경우 정규성에 민감하다고 알려져 있는데, 우리는 예측의 성능을 최우선으로 해서 데이터 분석을 진행하는 것이기 때문에, 오차의 정규성 가정이 만족해야 회귀모형의 해석 가능성에도 의미를 부여할 수 있다.)

- **오차의 등분산성(Homoscedasticity / Constant variance)** : 오차항의 분산은 상수다 (분산은 σ^2으로 동일하다) ↔ **이분산성(Heteroscedasticity)**



왼쪽이 이분산성(Heteroscedasticity), 오른쪽이 등분산성(Homoscedasticity)

이분산 형태라고 해서 회귀계수 추정에 편향이 생기지는 않지만, 회귀계수 추정의 효율성을 떨어뜨린다. 분산이 일정하지 않고 변화한다면 전체적인 회귀계수의 분산도 커지게 되는데, 그 결과 최소제곱추정량이 BLUE가 되도록 하는 조건(오차들의 평균이 0이고 분산이 σ^2으로 동일하며, 자기상관이 없음)을 만족하지 않아 최소분산이 갖는 효율성을 지니지 못한다.

→ 회귀식과 회귀계수 검정에 대한 신뢰성을 떨어뜨려, 유의하지 않은 변수가 유의하다고 나타날 수도 있다. 가설검정의 관점에서 말하면, 충분히 유의할 수 있는 귀무가설을 기각하는 것이고, 이는 제 1종 오류(Type 1 error)가 0.05로 고정되지 못하고 상승하는 것이다.

- **오차의 독립성(Independence / No autocorrelation)** : 오차항은 서로 독립이다 (오차항 간에 상관관계가 없다) ↔ **자기상관성(Autocorrelation)**

오차의 독립성이 만족하지 않다면, 역시 최소제곱추정량이 BLUE가 되지 못한다. 불편성은 만족하지만 효율성이 떨어진다.

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

이 가정들 중 선형성, 오차의 정규성, 등분산성, 독립성은 1주차의 다중선형회귀 모형의 함수식을 통해서도 확인해 볼 수 있다.

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

선형성, 오차의 정규성, 등분산성, 독립성 을 묶어 **회귀분석의 기본 4가지 가정**이라고 부른다. 2주차 클린업에서는 기본 가정들에 초점을 맞추어 진단과 처방 과정에 대해 알아보자.

2. 잔차 플랏 (Residual Plot)

회귀분석의 4가지 가정을 진단하기 위해서 크게 두 가지의 방법이 사용된다.

1) 시각적(graphical) 방법

2) 가설 검정을 이용한 방법

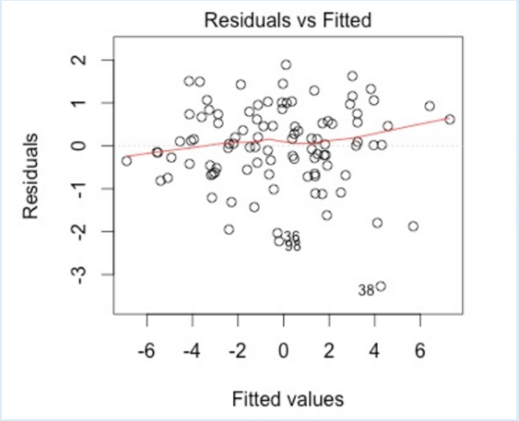
→ 물론 시각적 방법을 사용하더라도 판단에 대한 명확한 근거를 마련하기 위해 가설 검정의 과정은 필요하다.

오차항의 추정량인 잔차의 분포를 통해 경험적 판단에 근거한 회귀 진단이 가능해진다. 이는 R에서 제공하는 `plot()`을 통해 잔차의 분포를 쉽게 파악할 수 있다.

- 잔차 플랏 출력

```
# 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, 오차의 등분산성 확인

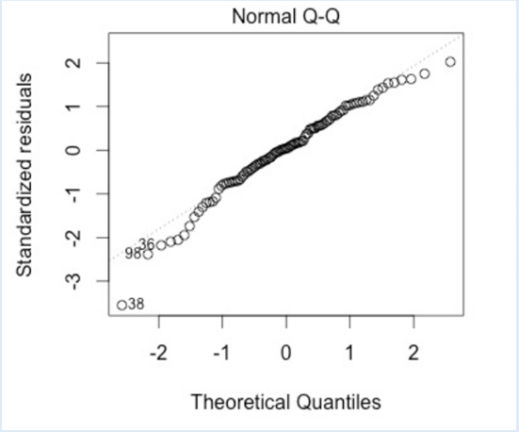


선형성, 등분산성 가정 만족

- X축 : 예측값($\hat{y}$) fitted values, Y축 : 잔차($e = y - \hat{y}$) residuals
- 빨간 실선 : 전체적인 잔차들의 추세선인데, 이는 잔차들의 분포를 Local Regression으로 추정한 완만한 연결 직선이며 잔차 분포의 패턴, 경향성을 나타내는 보조 지표이다. (Local Regression에 대해서는 뒤에서 다룰 예정!)
- 해석 방법 : 잔차와 예측값 사이에 무작위한 형태 이외의 어떠한 관계를 보이면 등분산성이 위배,

  빨간 실선이 x축에 평행한 직선이 아니라면 선형성이 위배되었다고 볼 수 있음.
- 해석 : **빨간 실선이 완만하게 수평**을 이루고 있고 **점들의 분포가 random하게 퍼져있어** 선형성과 오차의 등분산성 가정을 위배하지 않았다. (추가로 잔차의 평균이 0이라는 가정도 체크해볼 수 있다.

2) Normal QQ (Quantile-Quantile) → 오차의 정규성 확인

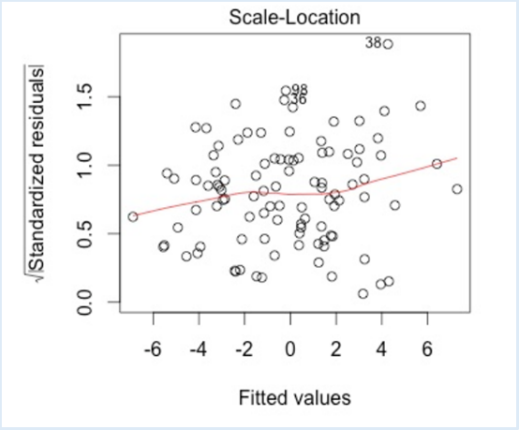


정규성 만족

- X축 : 정규분포의 분위수 값(Theoretical Quantile), Y축 : 표준화 잔차 (Standardized residual)
- 두 개 변수의 분포를 비교하기 위한 대표적인 비모수적 방법이자 시각적 방법
- 해석 방법 : **그래프가 $y = x$에 가까울 수록 잔차가 정규성을 만족한다.**

  → 잔차가 정규분포 사분위수 위에 그대로 위치한다는 의미이기 때문에, 잔차가 정규분포의 분포를 따른다는 것을 의미한다.
- 해석 : 해당 분포에서는 잔차 분포가 대부분 점선 주변을 벗어나고 있지 않아서 정규성을 만족하는 것 같지만, 38번 관측치의 경우 점선에서 많이 벗어나 있기 때문에 이에 대해서는 추가적인 확인이 필요하다.

3) Scale-Location → 오차의 등분산성 확인(주로 등분산성)

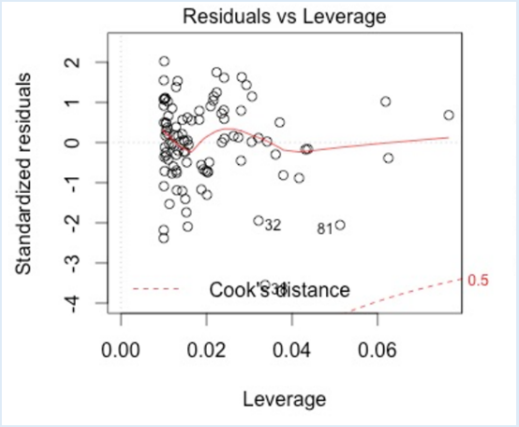


선형성, 등분산성 만족

- X축 : 예측값($\hat{y}$) fitted values, Y축 : 표준화잔차 ($\sqrt{|e_i|/se(e_i)}$)
- 빨간 실선은 전체적인 잔차들의 추세선이며, 잔차들의 분포를 Local Regression으로 추정한 직선
- Residual vs Fitted plot과 비슷하지만 잔차에 절댓값이 씌워진 형태라는 점에서 차이가 있다.
- 해석 방법 : 잔차와 예측값 사이에 무작위한 형태이외의 어떠한 관계를 보이면 등분산성이 위배
- 해석 : **점들의 분포가 random하게 퍼져있어** 선형성과 오차의 등분산성 가정을 위배하지 않았다.

  38번 처럼 표준화 잔차의 크기가 커서 0에서 멀리 떨어져 있을 경우, 회귀직선이 해당 Y를 잘 예측하지 못한다는 의미이기 때문에 Outlier일 가능성이 있다.

4) Residuals vs Leverage



Cook’s Distance 값이 전부 0.5미만이다

- X축 : 레버리지(지렛값), Y축 : 표준화 잔차
- 영향점(Influential point)을 확인할 수 있는 플랏으로, 플랏의 오른쪽의 위치한 점들이 leverage가 큰 잔차이며, 빨간 실선으로부터 위아래로 멀리 떨어진 점들이 outlier라고 생각할 수 있다.
- 빨간 점선 : Cook's distance(회귀직선의 모양에 영향을 미치는 점을 찾는 과정으로, 기울기, 절편에 크게 영향을 끼치는 점을 찾는다 → 이는 leverage와 잔차에 비례)로 주로 0.5과 1 사이에서 경계가 표시된다.
- 해석 방법 : Cook’s distance 0.5이상이면 influential point 후보, 1 보다 크면 매우 강력한 후보
- 해석 : 플랏에서 모든 관측치들이 한쪽에 몰려있고 0.5 경계 안에 있으므로 영향점은 없는 것 같다고 판단할 수 있다.(모형이 잘 적합했음을 의미)

---

3. 선형성 진단과 처방

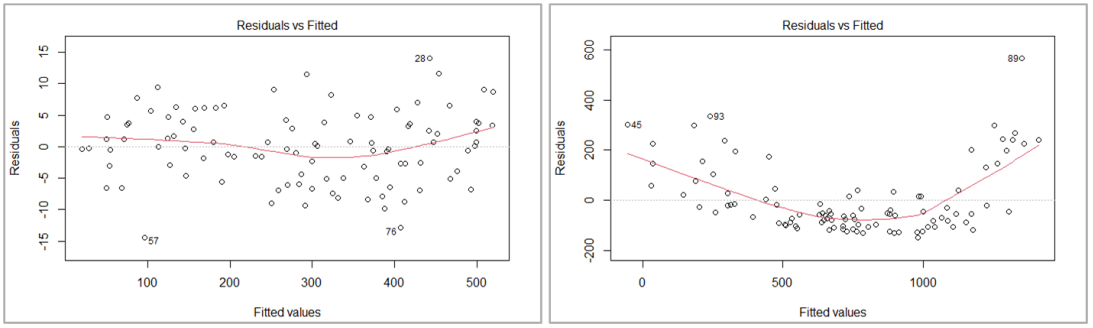
- 반응 변수가 설명 변수의 선형결합으로 이루어졌다는 가정

1) 진단

(1) 잔차 플랏을 통한 진단

- 평균 0을 중심으로 하는 x축에 평행한 직선 형태가 아니라면 선형성이 위반되었다고 볼 수 있다

  → 선형성이 위배되는 경우, 일반적으로 이차함수 혹은 삼차함수 형태로 표현된다.



잔차의 추세가 수평과 비슷하지 않고 오른쪽처럼 이차함수 꼴이라면 선형성이 위배된 것이다

(2) Partial residual plot을 통한 진단

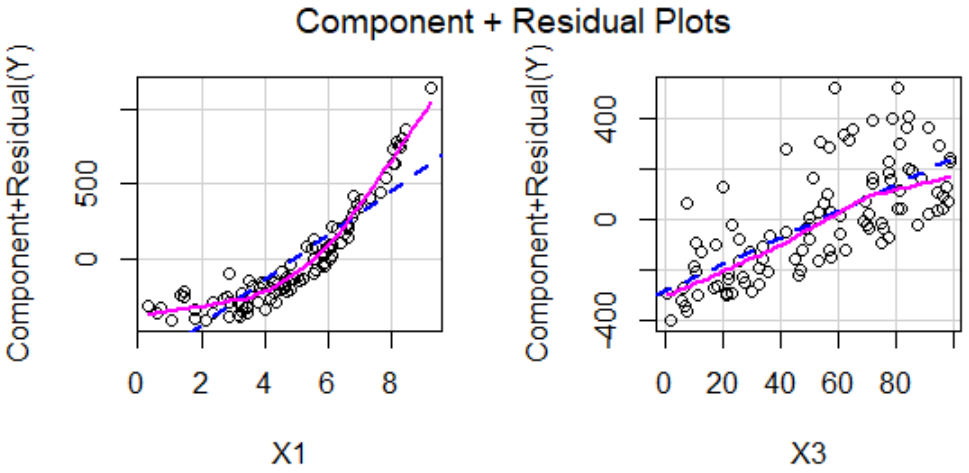
- 개별 독립 변수와 종속 변수 간의 선형성을 판단하기 좋은 플랏

  → 선형성을 만족하지 못할 때, 어떤 변수의 영향으로 선형성이 만족하지 않는 것인지 잔차 플랏으로는 확인하기 힘들기 때문에, **개별 변수의 영향을 확인**하는 과정이 필요하다.

- Car 패키지의 `crPlots()` 함수를 통해 개별 변수의 선형성을 파악 (Partial F-test처럼 이해하자!)

- X축 : x_i변수

  Y축 : Partial residual : $residual + \hat{\beta}_i X_i$ (단일 예측변수 기반 예측값+회귀식의 실제 잔차)



- 파란 점선: Partial residual과 x_i의 적합된 직선

  → 점들의 분포를 최소제곱방법을 통해 회귀선을 추정한 것

- 핑크색 실선: Partial residual의 추세선으로, 점들의 분포를 Local regression을 통해 추정한 선

- 해석 : 일반적으로 서로 다른 두 선이 일치하면 선형성이 만족되었다고 판단할 수 있다.
  - (왼쪽 플랏) X1과 Y는 선형적인 관계를 지니지 않는다.
  - (오른쪽 플랏) 그래프를 통해 판단해본다면 X3 변수가 Y를 선형적으로 잘 설명하고 있다

- 한계

  개별 변수의 선형성을 판단하기에는 좋은 방법이지만 단점이 존재
  - Y와 개별 X 변수들간의 단편적인 관계만을 보여주기 때문에, X 변수들 사이에 **교호작용이나 상관관계**가 존재하더라도 이를 파악하지는 못한다.
  - 심각한 다중공선성이 있는 경우 잘못된 정보를 제공할 수 있다.

  교호작용 : 한 요인의 효과가 다른 요인의 수준에 의존하는 경우로 변수 간의 시너지 효과를 의미 → ex) 운동과 식단관리를 동시에 진행하면 건강이 더 좋아지는 것

2) 선형성 가정이 위배될 경우 발생하는 문제점

위 진단에 따라 선형성 가정이 위배되었다고 판단되는 경우, 지금까지 배운 모형은 선형회귀모델이므로 모델자체가 성립하지 않는다. 대부분 실제 모델보다 과소추정이 된 경우(실제 데이터의 형태가 선형보다 복잡한 경우)로 예측성능이 떨어질 것이다.

3) 처방

(1) 변수변환

변수 변환을 통해 비선형 관계를 해결할 수 있다.

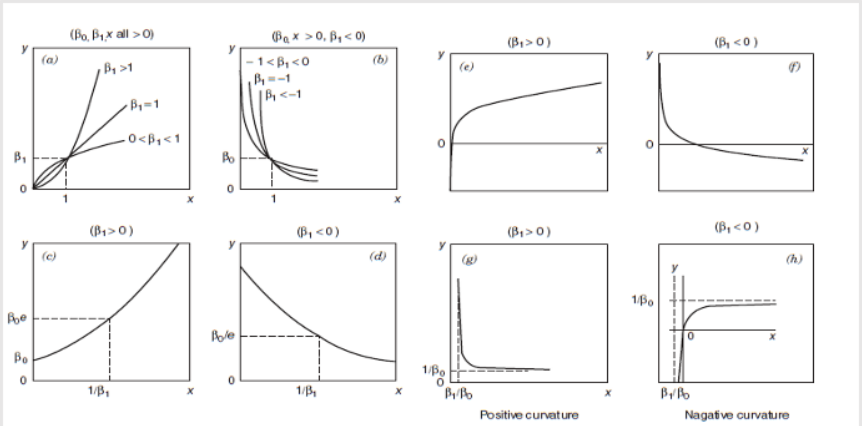


Figure 5.4 Linearizable functions. (From Daniel and Wood [1980], used with permission of the publisher.)

- 치환의 과정을 통해 x를 변화시켜 이를 새로운 x로 취급한다면, 선형결합을 만족하도록 만들 수 있음!

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

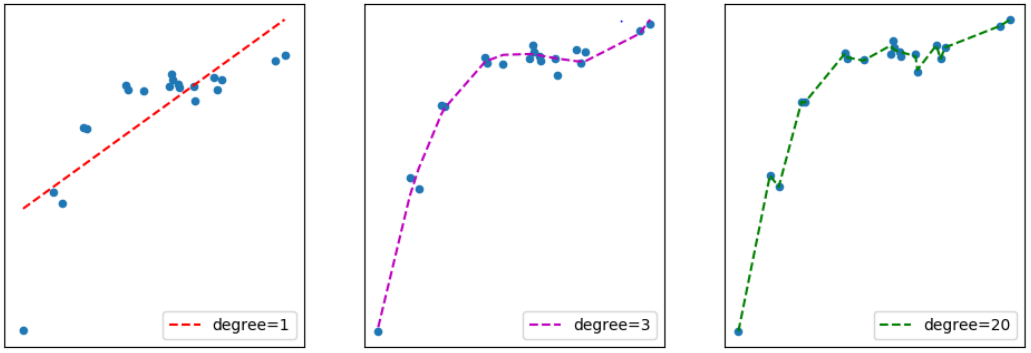
- 변수변환을 통해 선형성을 확보할 수 있는 모델도 넓은 의미에서 선형모델이라 부름

  → 포아송, 음이항, 로지스틱 회귀 모델을 일반화 선형모형(GLM)이라 부른다.

- 그런데 위에서 언급했듯 $y = \frac{\beta_1 x}{\beta_0 + x}$처럼 변환을 통해 선형결합 꼴을 만들 수 없는 비선형 모델도 있다.

(2) 다항 회귀(Polynomial Regression)

- 회귀식의 독립변수가 2차, 3차 방정식 같은 **다항식으로 표현**되는 것.



Underfit
High Bias
Low Variance

Correct Fit
Low Bias
Low Variance

Overfit
Low Bias
High Variance

회귀식이 하나의 설명변수의 k차 다항회귀식으로 주어지는 경우를 살펴보자.

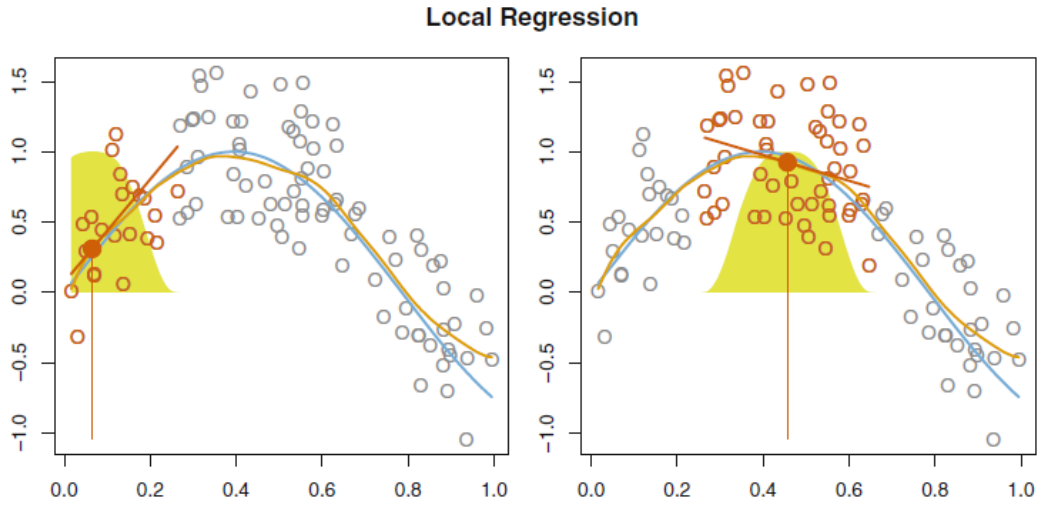
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \epsilon$$

X의 각 거듭제곱항을 하나의 새로운 설명변수로 생각하면, 다중회귀모형과 동일하게 생각할 수 있다. 그렇기에 추정 또한 다중선형회귀에서의 분석 방법을 그대로 적용할 수도 있다. 다만, 과적합 방지를 위해 적당한 차수 k를 설정해야한다는 과제가 생기며 설명변수들이 X의 거듭제곱꼴이기 때문에 설명변수들 간 상관관계가 높아져 후에 배울 **다중공선성**이 발생할 수 있다.

따라서 통계 패키지 등에서는 **직교다항식**을 이용한 알고리즘을 사용하여 계수들을 추정하게 된다. X의 거듭제곱을 직교하도록 유도하여 상관관계를 없애는 것 (직교다항식의 유도 등은 대학원 수준의 내용)

(3) 국소 회귀(Local regression)

- 비선형 문제여도 적은 범위(로컬, Local)에서 관찰하면 선형 문제라고 보는 접근 방식이다.



- 다항 회귀의 경우 새로운 변수를 추가하여 해결하는 모수적(parametric) 방법이고, 국소 회귀는 **비모수적(non-parametric) 방법**이다.
- Local Regression은 말 그대로 Local(지역적인)에 있는 데이터들로 회귀 모델링을 하는 방법이다. target data x_0를 중심으로 그 대역폭 내의 데이터 $x_i \in \mathcal{N}(x_0)$들만을 사용하여 부분적으로 선형 회귀 모델을 구성한다. 이 때 대역폭은 조정할 수 있다.

→ 가중치는 주로 정규분포와 비슷하게 생긴 Radius Basis Function(RBF) 혹은 tri-cubic function 에 기반하여 산정된다. (R에서 기본값은 tri-cubic function)

4. 정규성 진단과 처방

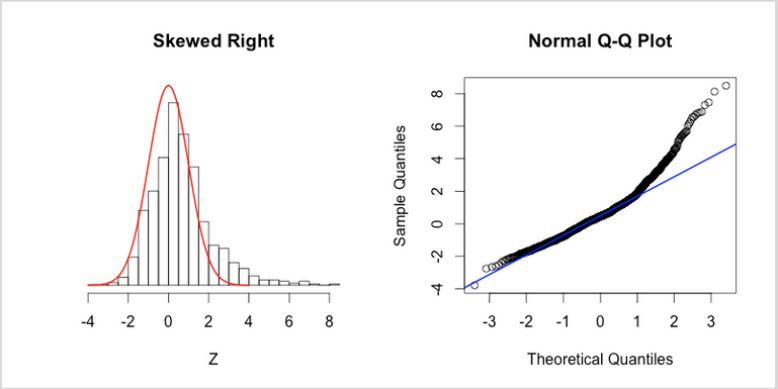
- **정규성 가정** : 반응 변수를 측정할 때 발생하는 오차는 정규분포를 따를 것이라는 가정

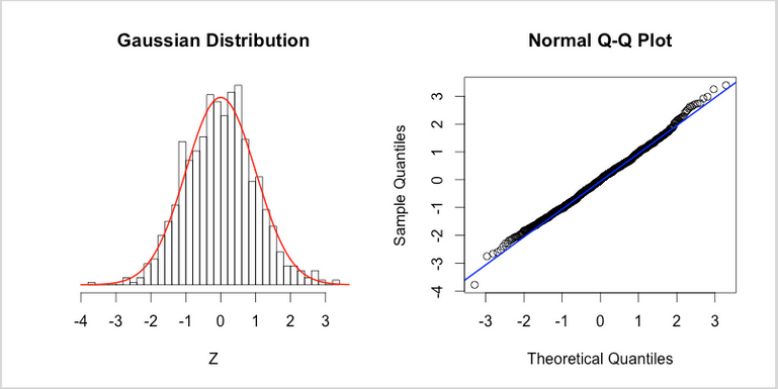
  회귀식이 데이터를 잘 표현한다면 잔차들은 단순한 측정 오차(noise)라 여겨지고, 이 잔차들의 분포는 정규분포와 흡사한 형태가 될 것이다! (사실 정규성 가정은 쉽게 위배되지 않는 가정임)

1) 진단

(1) Normal Q-Q plot

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





아래의 경우가 정규성을 만족하는 경우이다.

(2) 정규성 검정

: 플랏으로 확인하는 경우에 판단이 주관적일 수 있기 때문에, 정규성 만족 여부가 정말 명확하게 보이지 않는 경우를 제외하고는 **통계적 방법에 의한 가설검정**으로 확인해야 한다.

- 가설

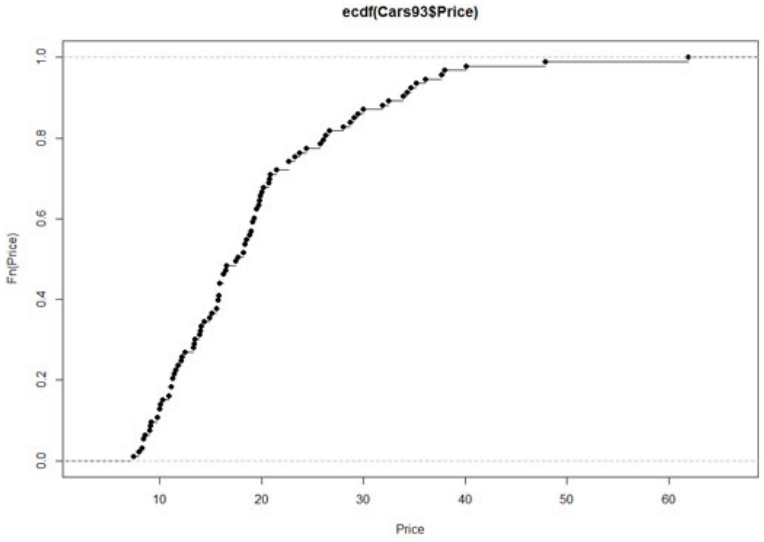
H_0 : 주어진 데이터는 정규분포를 따른다.

H_1 : 주어진 데이터는 정규분포를 따르지 않는다.

→ 귀무가설을 기각하지 못하는 것 = 정규분포를 따르는 상황을 원함

- 정규성 검정 1) - Empirical CDF를 이용한 Test

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

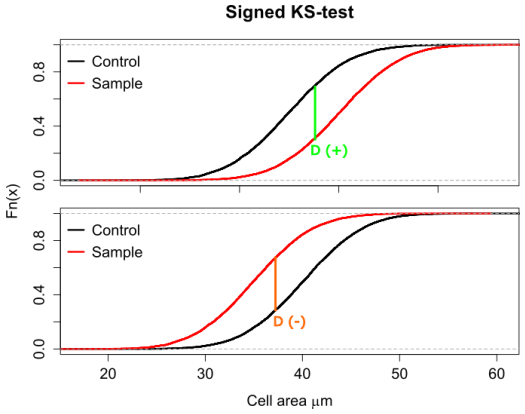


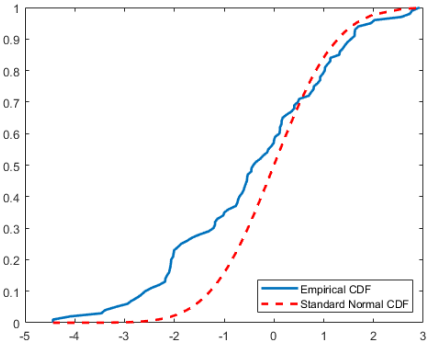
정규성 검정을 하기 위해 잔차의 ECDF와 정규분포의 CDF를 비교하여 검정한다.

→ 검정 방법 : Kolmogorov-Smirnov Test, Anderson Darling Test

a. Kolmogorov-Smirnov Test (K-S검정)

하나의 모집단이 어떤 특정한 분포함수를 갖는 지 알아보는 검정법으로, 귀무가설 하에서 표본분포함수(sample distribution function)가 어떤 이론적 분포함수(theoretical distribution)와 유사한지 검정하는 방법이다.





왼쪽에서는 빨간색 선이 표본분포함수(F_n)이며 비교 대상이 검정색 선, 이론분포함수(F_0)이다. 그림에서 보듯 검정통계량 D는 표본분포함수와 이론분포함수의 차이이며 이 차이가 크면 H_0을 기각하여 표본분포함수와 이론분포함수는 같지 않다고 판단할 수 있다.

오른쪽에서는 표본정규분포의 CDF와 비교하여 정규성 충족여부를 검정하고 있다.

```
#Kolmogorov Smirnov Test
ks.test(x=fit$residuals, y="pnorm")
# Kolmogorov Smirnov Test는 y의 입력값에 따라 다른 분포와의 비교도 가능하다.
# 도출된 p-value값으로 가설 기각여부 결정
```

b. Anderson-Darling Test (A-D 검정)

데이터가 특정 분포를 따르는 지 검정하는 적합도 검정의 하나이다. 이는 Kolmogorov Smirnov Test(K-S검정)를 수정한 적합도 검정으로, 특정 분포의 꼬리(tail)에 K-S 검정보다 가중치를 더 두어 수행된다.

```
#Anderson Darling Test
library(nortest)
ad.test(fit$residuals)
# 도출된 p-value값으로 가설 기각여부 결정
```

→ 두 방법 모두 잔차의 ECDF를 이용해 정규분포와 비교하는 방법이다.

해당 정규성 검정을 위해 A-D검정은 0.05로, K-S는 0.15로 유의수준을 설정한다. 이렇게 유의수준이 서로 다른 이유는 각 검정 방법별로 검정력에 차이가 있기 때문이며, 이를 통해 A-D 방법이 K-S 방법에 비해 엄격한 방법임을 알 수 있다.

- 정규성 검정 2) - 정규분포의 분포적 특성을 이용하는 Test

a. Shapiro Wilk Test : 표본이 정규분포로부터 추출된 것인지를 확인하기 위한 검정방법

- 정규분포 분위수 값과 표준화 잔차 사이의 선형관계를 확인 (QQ plot의 아이디어와 동일)

- 샘플 크기가 작을 때(5000개 이하인) 주로 사용하고, R 내장함수에서 `shapiro.test()` 함수 안에 residual 값을 넣으면 된다.

- 유의할 점은 여기서 귀무가설을 기각하지 못 했다는 것은 정규분포를 따르지 않는다고 말할 근거가 부족한 것일 뿐 100% 정규성이 만족된다는 뜻은 아니다.

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

b. Jarque-Bera Test

정규분포의 왜도가 0, 첨도가 3이라는 사실에서 기인한 방법. 잔차의 분포가 정규분포와 달라질수록 왜도나 첨도의 변화가 생기고, 통계량 값이 커져 유의수준을 넘어서면 귀무가설을 기각하는 방법

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

여기에서 n은 데이터의 개수, S는 표본의 왜도(Skewness), K는 표본의 첨도(Kurtosis)를 의미

데이터가 정규분포에서 발생했다면, 검정통계량은 자유도가 2인 카이제곱분포를 근사적으로 따름

- tseries 패키지 안에 있는 `jarque.bera.test()` 함수 안에 residual 값을 넣어준다.

- 이상치에 민감한 왜도를 사용하는 만큼, 이상치를 삭제했을 때 정규분포임이 드러나는 경우가 있을 수 있다. (회귀분석 보다는 시계열 분석에서 정규성 검정할 때 사용하는 경우가 많음)

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

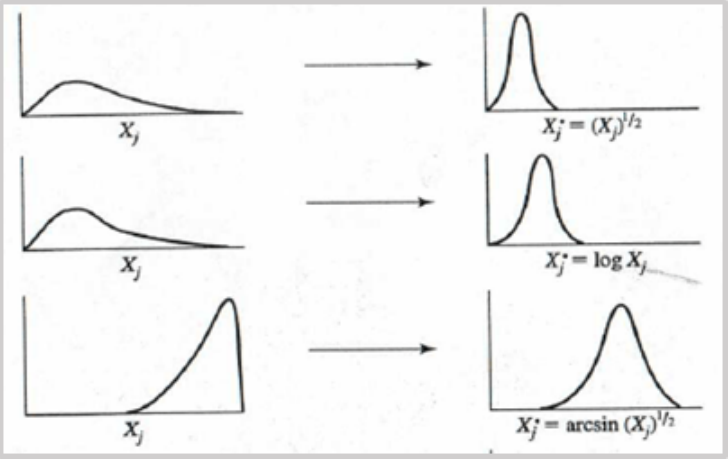
2) 정규성 가정이 위배될 경우 발생하는 문제점

선형회귀에서 오차의 정규성을 가정하고 있기 때문에 정규분포로부터 파생되는 표본분포에 대해 t분포, F분포를 사용할 수 있다. 회귀분석에 사용되는 F-test와 t-test는 모두 오차의 정규분포를 전제로 했었다.

만약 오차가 정규분포를 따르지 않아 정규성 가정이 위배되었다면, 검정통계량이 t분포 혹은 F분포를 따르지 않기 때문에, 가설 검정 결과가 p-value에 의해 유의하게 나오더라도 검정 결과와 예측의 결과를 신뢰할 수 없게 된다.

3) 처방

(1) 변수 변환



자의적으로 변수 변환을 진행할 수 있지만, 주관적 판단 하에 이루어지기 때문에 객관성을 확보하기는 힘들다.

a. Box-cox Transformation

반응변수(Y)를 변환함으로써 정규성과 추후에 나올 등분산성을 해결해주는 방법. 통계적인 검정에 따라 변수를 변환한다는 점에서 자의적인 변수변환보다 객관적이면서도 효율적이다. 반응변수가 **양수일 때** 적용 가능하다.

아래 식에서 λ를 변화시키면서 y가 정규성을 만족하도록 만드는데, 일반적으로 λ는 -5에서 5사이의 값을 사용하는데, λ가 0인 경우에는 log-transformation을 해준다.

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ log(y) & \text{if } \lambda = 0 \end{cases}$$ → 모든 실수 λ에 대해 연속임

ex) λ가 +2이면 이차함수, +0.5면 루트변환, +0면 로그변환, -1은 역수변환을 의미함

Lambda	Standard Transformation
-3	Inverse Cube
-2	Inverse Square
-1	Inverse
-0.5	Inverse Square Root
0	Logarithmic
0.5	Square Root
1	No Transformation
2	Square
3	Cube

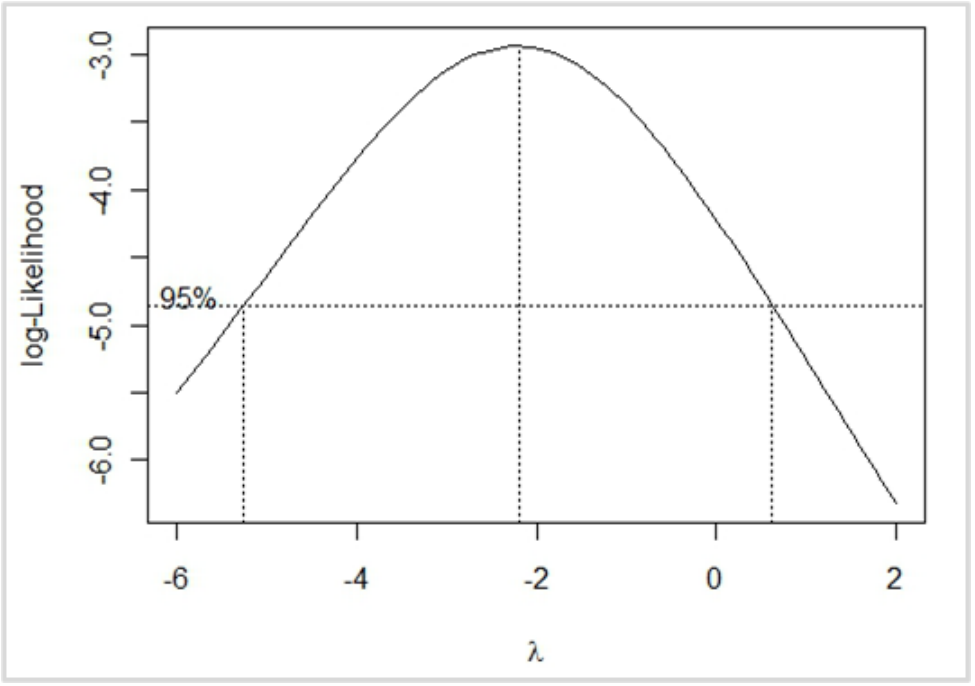
파라미터인 λ에 따라 달라지는 변환의 종류

이 때 최적의 λ는 최대우도함수(ML)을 통해 신뢰구간을 구한 후에 신뢰구간 내의 로그우도함수(ML)를 최대화하는 λ를 최적의 값으로 선택한다.

- car 패키지의 `powerTransform()`을 통해 구현할 수 있다.

```
#load car library
library(car)

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



- 95% 내의 λ값 중 가능도함수가 최대가 되게하는 -2 근방의 λ를 선택하거나 혹은 정수중에서 가능도함수가 최대가 되는 지점 -2를 λ로 선택하면 된다.. 이렇게 정수로 λ를 선택한다면 제곱, 역수, 제곱근 등 변수 변환 관계를 쉽게 파악할 수 있다는 장점이 있다.
- 주의할 점은, Box-cox에서는 y가 log(y)로 변환될 수 있으므로, Y가 0 이하일 경우에는 사용할 수 없다. 그래서 뒤에 나오는 Yeo-Johnson을 살펴보자!

b. Yeo-Johnson Transformation

Box-cox와 아이디어는 같다. 그러나 Box-cox 변환과 달리 변수 범위에 대한 제약이 없다는 점에서 장점이 있다.

.

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

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

λ 경계값에 대한 설명은 2022-1학기 회귀분석팀 팀장님이 질의응답으로 남겨주셨던 필기로 대체합니다.

https://cafe.naver.com/powersat?iframe_url_utf8=/ArticleRead.nhn%3Fclubid=28935080%26articleid=4761

간단히 요약하자면, order를 지키고 continuity 속성을 가져가기 위해서이다.

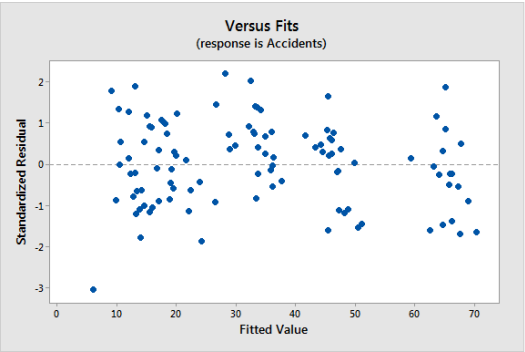
5. 등분산성 진단과 처방

: 오차의 모든 분산은 동일해야 한다는 가정(분산이 상수여서 어느 관측치에서나 동일하게 나타나고, 다른 변수의 영향을 받지 않는다)

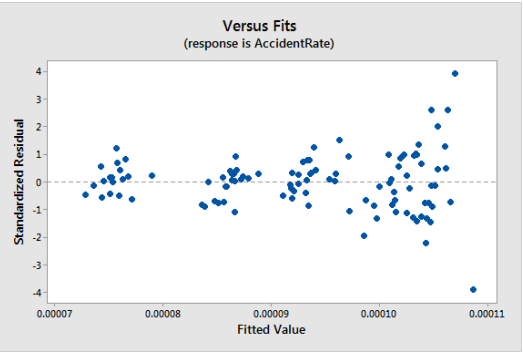
1) 진단

(1) 잔차 플랏

잔차 플랏의 ‘residual vs fitted’ plot과 ‘scale-location’ plot을 종합적으로 보고 판단한다



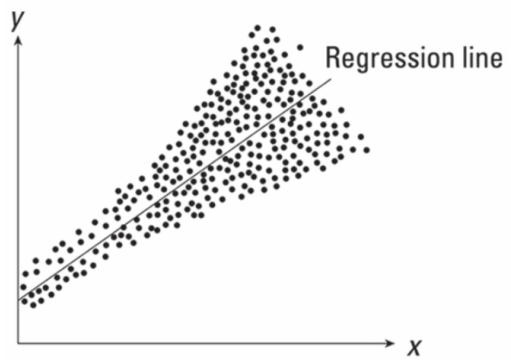
(왼) 등분산

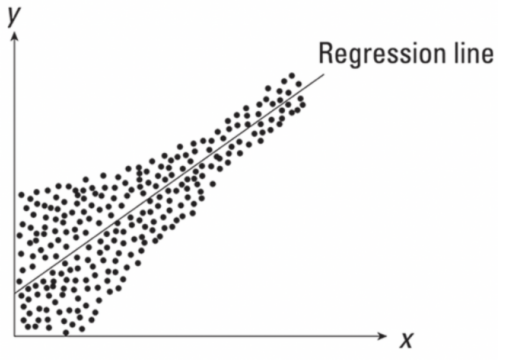


(오) 이분산

오른쪽 플랏은 $\hat{y}$값이 커짐에 따라 절대값이 커지는 잔차를 갖는, 이분산성을 띄는 데이터이다.

→ 이분산의 형태는 퍼짐이 증가하거나 감소하여 퍼짐의 정도가 일정하지 않은 것으로, X 평균 부분의 퍼짐이 큰 형태 등 다양한 이분산의 형태가 존재할 수 있다.





→ 육안으로 명확하게 판단하기 어려운 경우, 역시 검정의 방법을 통해 등분산성을 확인할 수 있다.

(2) BP(Breusch-Pagan) test

오차의 평균이 0이라면, 오차의 제곱은 오차의 분산이라고도 할 수 있을 것이다. 오차를 직접 구할 수는 없기 때문에, 잔차의 제곱 e_i^2을 반응변수로 두고 설명변수를 x_i로 하여 선형 모형을 적합하는 것을 생각해보자.

$$e_i^2 = \gamma_0 + \gamma_1 x_i, \ \ i = 1, 2, \cdots, n$$

여기서 귀무가설 $H_0 : \gamma_1 = 0$ 를 test하여 기각한다면 오차의 분산이 설명변수에 따라 증가하거나 감소한다고 말할 수 있을 것이다. 이것이 BP test의 기본적인 아이디어이며 검정을 위해 다음을 가정한다.

- BP test의 기본 가정

- 샘플 수가 많아야 한다

- 오차항은 독립이고 정규분포를 따라야 한다

- 오차의 분산은 설명변수와 연관이 있어야 한다

- 가설

$H_0 : \gamma_1 = 0$, 주어진 데이터는 등분산성을 지닌다.

$H_1 : \gamma_1 \neq 0$, 주어진 데이터는 등분산성을 지니지 않는다. (이분산이다.)

→ 우리는 데이터가 등분산성을 따르길 원하고, 귀무가설이 기각되지 않길 원한다.

- 검정통계량

  잔차의 제곱이 독립변수의 선형 결합으로 표현되는지, 그리고 그 때의 설명력은 어느 정도인지를 결정계수를 통해 파악한다. 만약 오차가 독립변수에 의해 충분히 표현된다면 결정계수는 커질 것이고 검정통계량 또한 커질 것이다.

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

  → X 변수를 선형결합한 식에서 R^2를 구하고,

  (검정통계량) $\chi^2_{stat} = nR^2 \sim \chi^2_{p-1}$

  (임계값) $\chi^2_{p-1,\alpha}$ 을 설정해서

  $Reject\ H_0 \ \ if \quad \chi^2_{stat} > \chi^2_{p-1,\alpha}$

  → 카이스퀘어 통계량이 임계값보다 크면 귀무가설을 기각한다 (이분산성이 존재한다는 의미)

- 구현

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.

- 단점

분산과 X변수가 선형 결합으로 이루어졌다는 가정을 바탕으로 하기 때문에, 비선형결합으로 이루어진 이분산성은 파악할 수 없으며, 샘플이 대표본이어야 사용할 수 있다.

또한, 오차의 정규성에 민감하게 반응하기 때문에 정규성이 지켜진 상태인지 확인해야 한다.

2) 위배되었을 때 발생하는 문제

이분산은 추정량의 분산을 증가시키지만, OLS 추정량은 이를 잡아내지 못하기 때문에, 이분산이 있는 경우 OLS 추정량의 분산은 실제 분산보다 작게 측정된다.

→ 과소추정된 분산은 검정통계량을 크게 만들고(검정통계량의 분모에 추정량의 분산이 들어가니까!), p-value를 작게 만들어 실제로는 유의하지 않은 변수를 유의하다고 나타낼 수 있게 된다.

→ 즉, 충분히 유의할 수 있는 귀무가설을 기각하는 과소추정, 즉 제 1종 오류(Type 1 error)가 발생하게 되며, 이는 가설검정의 신뢰성을 떨어뜨린다.

또한 등분산성 조건이 만족하지 않는다면 OLS 추정량이 더 이상 BLUE일 수 없게 된다.
(BLUE의 조건: 오차의 평균 0, 오차가 등분산, 오차 간 자기상관성 없음)

3) 처방

(1) Box-cox Transformation

(2) 가중회귀제곱 (WLS: Weighted Least Square)

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

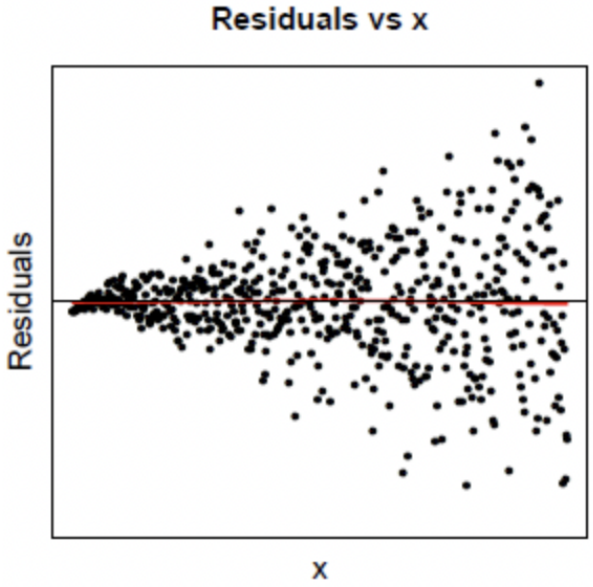
ex) 분산이 큰 부분의 관측치에는 가중치를 적게 주어 전체적인 분산을 비슷하게 맞춰주는 방식

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

→ 가중치는 보통 분산의 역수로 넣어주지만, 분산을 알기 어렵기 때문에 이 부분은 경험적으로 산정해야 한다 (사전 지식 or 잔차플랏을 통해 결정!)

- 가중치 선정 방식

  - 잔차플랏 이용



→ 이렇게 residual plot에서 분산이 점점 커질 경우에는 $w_i \propto \frac{1}{\sigma_i^2}$ 와 같은 방식으로 가중치를 준다!

- 모델 기반 선정

1. OLS로 다중선형회귀 모형을 적합시키고, 이 모형의 잔차를 구한다.
2. 잔차의 절대값을 종속변수로 두고, 기존 데이터의 변수들 중 오차의 분산에 영향을 주는 변수를 독립변수로 하는 다중선형회귀 모델을 OLS로 추정한다.
3. 2에서 구한 회귀계수를 $\hat{\eta}_0, \ldots, \hat{\eta}_p$라고 하면 x_i에 대응하는 i번째 적합 값 $s_i = \hat{\eta}_0 + \hat{\eta}_1 x_{i1} + \cdots + \hat{\eta}_p x_{ip}$ 를 구한다.

4. 적합값 제곱의 역수 $w_i = 1/s_i^2$ 를 가중치로 설정하여 기존 데이터에 가중 회귀모델을 적용한다.
5. 1에서 구한 회귀계수와 4에서 구한 회귀계수를 비교해서 차이가 크지 않다면 4의 모형을 최종 모형으로 하고, 그렇지 않다면 4에서 구한 모형으로 잔차를 계산한 뒤 다시 2~5의 과정을 거친다.

- 장점

  : WLS를 통해 구한 추정량은 회귀식의 기본 가정 하에 BLUE를 만족한다!

---

6. 독립성 진단과 처방

오차들은 서로 독립이라는 가정으로, 개별 관측치에서 i번째 오차와 j번째 오차가 발생하는 것에 서로 영향을 미치지 않는다는 가정

오차들이 서로 독립이 아니라서 독립성 가정이 위배되었다면, 이때 오차들 간 자기상관(autocorrelation)이 있다고 한다. ⇒ 오차들 간 상관성의 pattern이 있다는 것!

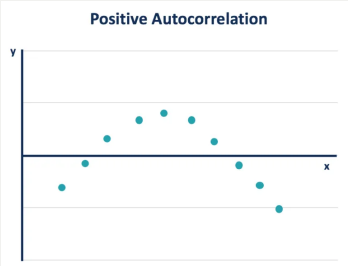
→ 우리 모델이 데이터를 잘 설명한다면 설명하고 남은 잔차가 특정 패턴을 지니지 않아야 하지만, 시간적/공간적으로 인접한 관측치들은 유사한 경향을 가지고 있어서 회귀식으로만으로 설명할 수 없는 패턴이 남아있을 수 있다.

⇒ 시간적 자기상관 : 시계열 분석
공간적 자기상관 : 공간회귀를 통한 접근(3주차 예정)

- 시계열모형 : 회귀분석에서 ‘시간적 자기상관’이라는 특별한 조건이 추가된 형태로 이해할 수 있다. (자세한 내용은 시계열 클린업 참고)

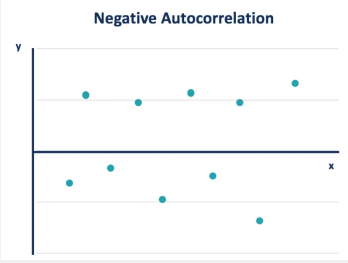
✅ 시계열 관점에서 Autocorrelation 살펴보기
: 연속적인 시간의 관점에 있는 변수들끼리의 상관관계 정도를 측정한 것이다. 과거 데이터를 활용하여 시간에 따른 pattern 혹은 trend를 찾는 것이기 때문에 활용성이 매우 높다. (stock market 등)
→ ARMA(autoregressive-moving-average model), ARIMA(autoregressive-integrated-moving-average model) 모델 등에서 활용됨

- Positive Autocorrelation



양의 자기상관을 가질 때 그래프는 다음과 같이 smooth curve를 가진다. 이를 통해 positive error는 positive error끼리, negative error는 negative error끼리 영향을 주고 받는 모습을 볼 수 있다.(경향성 읽기!)

- Negative Autocorrelation



음의 자기상관을 가질 때, 양과 음의 값이 번갈아서 나오는 모습을 볼 수 있다.

1) 진단

(1) Durbin Watson Test

더빈-왓슨 검정은 바로 앞 뒤 관측치의 1차 자기상관성(first order autocorrelation)을 확인하는 검정 방법이다. (1차 자기상관성이란, 연이어서 등장하는 오차들이 상관성을 지니는 것을 의미하며, $t-1$시점의 잔차와 t시점의 잔차 간의 상관관계를 측정한 것이다)

AR(1) 모형 : 1차 자기 회귀 모형 ⇒ $X_t = \ \square X_{t-1} + Z_t$

$\rho = Corr(X_t, X_{t-1}) = \frac{Cov(X_t, X_{t-1})}{\sqrt{Var(X_t)}\sqrt{Var(X_{t-1})}}$= $\square$

- 가설

  H_0 : 잔차들 간에 1차 자기상관이 없다. (잔차들이 서로 독립이다)

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

- 검정 통계량

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

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

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

  → $\hat{\rho}_1$는 표본 잔차 자기상관(sample autocorrelation of the residuals)을 나타내는데,

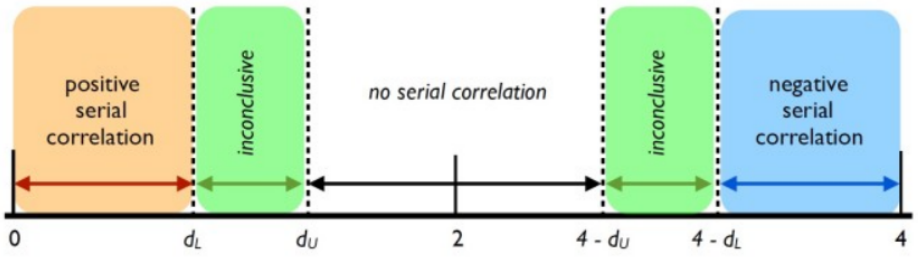
  이는 -1부터 1사이의 값을 갖는 e_i와 e_{i-1}의 상관계수의 꼴로 생각할 수 있다.

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

- 해석

  더빈 왓슨 검정의 귀무 가설은 '1차 자기상관이 없다' 즉, $\hat{\rho}_1$ = 0 이다. 1차 자기상관계수는 더빈 왓슨 통계량과 근사한 선형관계가 존재하므로, $\hat{\rho}_1$이 0에 가까워진다면 d는 자연스레 2에 가까워질 것이다.

  하지만 판단하기 애매한 경우에는 더빈 왓슨 검정 표를 참고한다! 이 표는 데이터의 개수 n과 변수의 개수 p에 따라 이 귀무가설을 기각할 수 있는지 없는지 판단하는 cut-off 값을 알려준다. 흔히 이 경계의 하한은 d_L 상한은 d_U 라고 표현한다.



위 그림에서 한 방향에 대해서만 살펴보자. 우리가 구한 검정 통계량 d가 하한(d_L)보다 작다면 귀무가설을 기각할 수 있다. 이때 양의 자기상관(positive serial correlation: 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 상한(d_U)보다 크다면, 귀무가설을 기각시키지 못하게 된다.

→ d가 0에 가까울수록 양의 상관관계를, 4에 가까울수록 음의 상관관계를 나타낸다.

(-1 < ρ < 0 :음의 상관관계, 0 < ρ < 1 : 양의 상관관계)

(2에 가까운 값이면 귀무가설을 기각하지 못한다! : $\hat{\rho}_1$이 0에 근사하고 있다는 의미이기 때문)

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

https://www.slideshare.net/Njirrming/durbin-watson-tables

- 한계

  - d가 상한(d_U)과 하한(d_L) 사이에 위치하게 된다면 우리는 판단할 수 없다. (inconclusive)

  - 바로 인접한 오차와의 1차 자기상관(첫번째 순서의 자기상관성)만을 고려하기 때문에 자기상관이 오래 지속되거나 계절성(seasonality)이 있는 등 주기가 있다면 확인하는 데 한계가 있다.

    → 즉, AR(1) 구조만 파악할 수 있다는 한계가 있다. (시계열 2주차 클린업 참고)

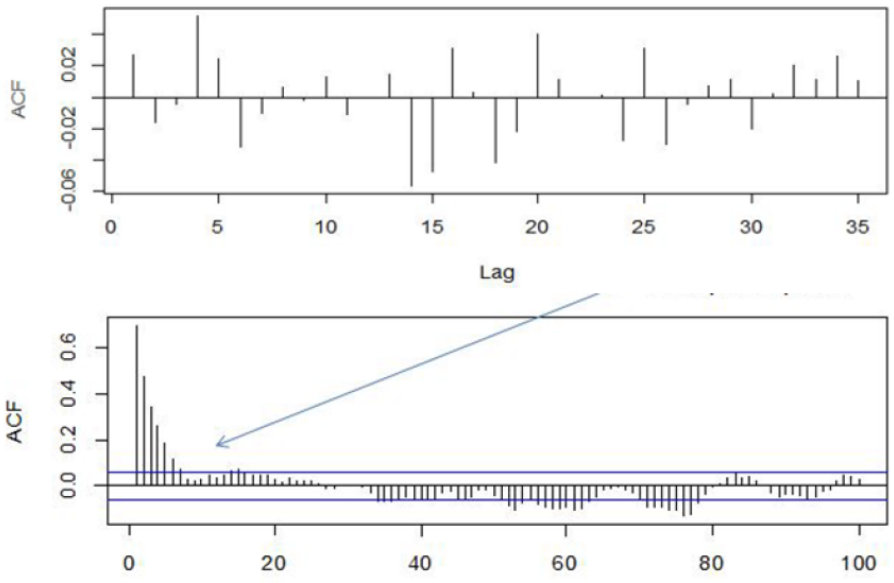
- R 코드

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

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

(2) Autocorrelation function plot (ACF plot)

: 더빈 왓슨 검정에서 판단할 수 없는 독립성의 경우가 존재하고, 1차 자기상관만을 고려하는 한계를 보완하는 오차 독립성 진단 방법이다. 1차 자기상관부터 p차 자기상관까지 고려하고 이에 따라 신뢰구간을 반환하기 때문에 통계적인 절차에 따라 판단할 수 있는 힘을 가지고 있다. 아래의 plot을 살펴보자. X축은 p, Y축은 p차 자기상관을 나타낸 것이다.



위(자기상관 없음), 아래(자기상관 있음)

위의 플랏은 자기상관이 없는 플랏이며 아래 플랏은 자기상관이 있는 플랏이다. 아래 플랏을 먼저 보고 위의 플랏과 비교해보자. 아래 플랏의 파란색 실선은 신뢰구간으로 신뢰구간을 벗어나는 선들

은 p차 자기상관이 있다고 간주할 수 있는 것들이다. 예를 들면, 아래 플랏의 초반 (1,2,3,4,5) 차수는 모두 양의 자기 상관이 꽤 강하게 있는 것으로 보인다. 이에 반해 위의 플랏은 신뢰구간이 보이지 않을 정도로 자기상관 수치가 모든 차수에서 작다. 따라서 자기상관이 없다고 할 수 있다.

(시계열팀 클린업 1주차 참고!)

- R 코드

```
# ACF plot 그리기
x<-sample(1:9,10,replace=TRUE)
acf(x) #plot으로 보여준다
```

2) 위배되었을 때 발생하는 문제

회귀계수의 추정치가 불편성은 만족하나 최소분산성을 만족하지 못하게 되며, 독립성 가정 하에 수행되는 신뢰구간의 계산, 가설검정 등 여러 가지 추론의 결과를 신뢰할 수 없게 된다.

특히, 오차항들이 양의 상관관계를 갖는 경우 오차항의 분산이나 회귀계수 추정치의 표준오차를 과소추정하는 경향이 있다.

3) 처방

(1) 설명변수의 추가 - 자기상관을 유발하는 변수를 설명변수로 모형에 추가

회사에서 몇 년에 걸친 판매액과 광고비의 관계를 분석한다고 하자. 이때 회사 자체의 크기(자본금, 종업원 수 등)를 모델에 포함시키지 않으면 양의 자기상관관계가 존재할 수 있다. 또한 경제 관련 문제에서는 시간과 관련된 변수나 시간 변수 자체를 추가할 수 있다.

(2) 분석 모델 변경

- 시간에 따른 자기상관 : 자기상관을 고려하는 AR(p)와 같은 시계열 모델 사용
- 공간에 따른 자기상관 : 공간의 인접도를 고려하는 공간회귀모델 사용

ex) SEM모형(Spatial Error Model, 공간오차모형), SLM모형(Spatial Lagged Model, 공간시차모형) 등

🤦 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)
```

```
> summary(gvlma(car_model))

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = car_model)

                     Value  p-value                   Decision
Global Stat        11.73816 0.019408 Assumptions NOT satisfied!
Skewness            2.37864 0.123004    Assumptions acceptable.
Kurtosis            0.02033 0.886622    Assumptions acceptable.
Link Function       8.57441 0.003409 Assumptions NOT satisfied!
Heteroscedasticity  0.76478 0.381838    Assumptions acceptable.
```

- 한계

  gvlma를 이용하면 간편하기는 하지만, statistical testing기법이 갖는 한계점처럼 유의수준 0.05에서 [가정 충족 || 가정 충족하지 않음]의 경계를 잘라 버리다 보니 융통성이 부족하다.

  선형회귀는 이런 가정 충족에 대해서 비교적 robust 한 편이다 보니 이 결과만 보고 비선형적 모델로 바로 넘어가는 등의 빠른 판단은 위험할 수 있다.

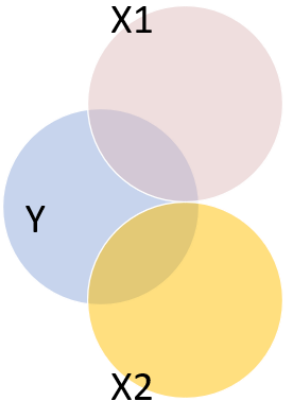
7. 다중공선성

1) 다중공선성(Multicollinearity)이란?

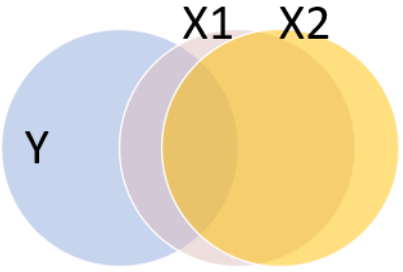
우리는 지금까지 회귀분석의 가정: 선형성, 등분산성, 정규성, 독립성에 대해 알아보았다. 그러나 지금부터 알아볼 다중공선성은 기본 가정 못지 않게 심각한 문제를 일으켜 회귀분석 자체를 위태롭게 하기 때문에 매우 주의해야 한다.

- **다중공선성** : 모델에서 설명변수 X_j들 사이에 서로 **선형적인 상관관계**가 존재하는 것

  아래 그림에서 확인해본다면, 첫 번째 그림에서는 $X1$과 $X2$가 각각 Y를 설명하고 있지만, 두 번째 그림에서는 $X1$과 $X2$가 서로 많이 겹친 채 Y를 설명하고 있다. 두 번째 그림처럼 $X1$과 $X2$가 겹쳐 있는 경우 다중공선성이 있다고 한다.



다중공선성이 없는 경우



다중공선성이 있는 경우

예시를 통해 살펴보자.
Y: 유진이의 학기별 성적, X_1: 집에서 학교까지의 거리, X_2: 통학 시간, X_3: 혈중 알코올 농도

위와 같은 변수들로 회귀분석을 시행한다고 해보자. 속도가 일정하다고 가정하면, $X_2 = aX_1$ 과 같은 식으로 X_2 변수가 X_1에 의해 완벽하게 설명될 수 있을 것이다. 즉, X_2의 정보는 완전히 필요하지 않은 정보라고 할 수 있다.

2) 다중공선성의 문제점

다중공선성이 있을 때 어떠한 문제들이 발생할까?

- 추정량의 문제

이전의 예시를 행렬의 관점에서 살펴보게 되면, linear dependent한 열이 생기는 것이므로 X가 full rank가 아니다. 따라서 $(\mathbf{X'X})$ 역시 full rank가 아니게 된다. 1주차 클린업에서 우리는 다중선형회귀모형을 적합시키기 위해 최소제곱법(OLS method)을 이용하여 LSE(Least Squared Estimator)를 구했다. 적합된 식은 다음과 같았다.

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

위 식에서 $\mathbf{X'X}$의 역행렬 $(\mathbf{X'X})^{-1}$을 구해야 하는데, 이것이 존재하지 않게 되어 정규방정식의 유일해를 구할 수 없게 된다. 즉 다중공선성 문제는 지금까지 우리가 모수를 추정하기 위해 써왔던 OLS method를 사용할 수 없게 만든다. 이렇게 완전하게 선형종속이 발생하는 경우를 Complete multi-colinearity라고 한다. 하지만 실생활에서 이렇게 완전한 선형 종속이 발생하는 경우는 드물다.

그렇다 하더라도 상관관계가 높은 경우 추정량을 불안정하게 만들기도 한다. 이해를 위해 두 개의 설명변수가 있는 모형에서의 $(\mathbf{X'X})$을 살펴보자.

$$(\mathbf{X'X}) = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

이 행렬의 행렬식 값은 $|1 - r_{12}^2|$ 이며, 변수 간 상관관계가 높다면 $det(X'X) \approx 0$이 될 것이다.

이제 추정량의 분산을 살펴보자. $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$ 이며, $(\mathbf{X'X})^{-1}$ = $\frac{1}{det(X'X)}adj(X'X)$ 이다. 즉, 다중공선성이 존재할 경우 추정량의 분산이 매우 커져 **계수의 추정이 불안정**해진다는 문제가 발생한다. 이에 따라 Prediction Accuracy도 심각하게 감소한다.

- 해석의 문제

1. 모델의 검정 결과를 신뢰할 수 없다.

다중공선성이 존재할 때 다중선형회귀모델은 전체 검정인 F-test (전체 회귀계수가 0인지, 적어도 하나는 0이 아닌지 검정)는 통과하고, 적합성 검정을 위한 R^2값도 괜찮은 수준이지만 유의한 개별 계수가 하나도 존재하지 않는 상황이 발생한다. 회귀계수들의 분산이 커짐에 따라 개별 변수의 유의성 검정에서 t 검정통계량은 작아지게 되고, β = 0 이라는 귀무가설을 기각하지 못하기 때문.

→ 전체 회귀식은 유의한데, 개별 회귀계수 중에는 유의한 것이 없는 결과가 발생한다.
(F-test가 t-test 보다 엄격한 검정인데, t-test에서 기각을 하지 못하는 것!)

2. 모델 해석에 영향을 준다.

다중선형회귀모델에서 개별 베타 계수 β_j의 해석은 '변수 x_j를 제외한 나머지 변수가 고정되어 있을 때, x_j가 한 단계 증가하면 증가하게 되는 증가량'이었다. 하지만 두 변수가 서로에게 영향을 주고 있다면, '나머지 변수가 고정되어 있을 때'라는 가정 상황이 불가능해질 것이다.

3) 다중공선성의 진단

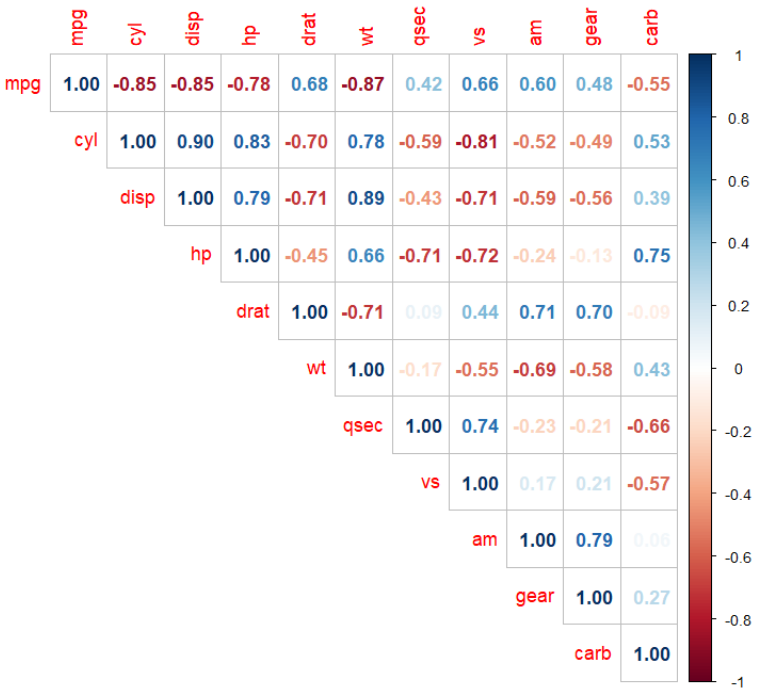
- 직관적인 판단
  1. F-test는 유의했지만 개별 회귀계수들에 대한 검정에서 귀무가설을 대부분 기각하지 못할 때
  2. 상식적으로 유의한 회귀계수가 유의하지 않다고 나올 경우

     → 이미 비슷한 설명변수가 모델에 포함되어 종속변수를 설명하고 있을 것이다
  3. 추정된 회귀계수의 부호가 상식과 다를 경우

     → 이미 비슷한 설명변수가 모델에 포함되어 종속변수를 설명하고 있어 부호가 반대로 발생했을 것이다

- 상관계수 plot

  변수들 사이의 선형관계 여부를 파악할 수 있다.



보통 절댓값을 기준으로 **상관계수가 0.7 이상**일 경우, 다중공선성을 의심할 수 있다.

그러나 3개 이상의 변수에 대해 선형종속 관계가 있는 경우, 상관계수의 값이 높지 않게 나올 수 있기 때문에 참고적으로만 활용해야 한다.

- **VIF(Variance Inflation Factor, 분산팽창인자)**

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

  - R_j^2 : 다중선형회귀모델 $x_j = \gamma_1 x_1 + \cdots + \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots + \gamma_p x_p$을 적합했을 때의 결정계수

    R_j^2는 회귀식이 데이터를 설명하는 정도를 의미한다. 즉, R_j^2가 크면 x_j가 나머지 변수들의 선형결합으로 충분히 표현될 수 있다는 것이므로 다중공선성이 있다는 것이다.

  - 일반적으로 VIF가 10 이상일 경우(결정계수가 0.9 이상) 심각한 다중공선성이 존재한다고 판단한다.
  - 다중공선성이 전혀 존재하지 않는다면 VIF 값은 1이 나온다. ($R_j^2 = 0$)

- **$(\mathbf{X}'\mathbf{X})$ 의 고유값 조사**

설명변수들 간 선형 종속 관계가 있는 경우, X의 rank가 줄어들고 $(\mathbf{X'X})$ 의 rank도 p보다 줄어들게 된다. rank는 **0보다 큰 고유값의 개수**로도 볼 수 있기 때문에, 만약 0에 가까운 고유값이 있으면 다중공선성을 의심할 필요가 있다.

고유값을 이용하는 방법의 장점은 VIF와 달리 고유벡터를 통해 선형종속의 관계의 형태를 대략적으로 알 수 있다는 것이다.

Spectral decomposition에 의해

$$V'(X'X)V = diag(\lambda_1, \lambda_2, ..., \lambda_{p-1})$$

의 관계식이 성립한다. 만약 다중공선성이 성립하는 경우, λ_k가 0이거나 0에 가까운 값을 갖게 되고,

$$v_k'(X'X)v_k = (Xv_k)'(Xv_k) \approx 0$$

이며, 또한 이는

$$\sum_{i=1}^{p-1} v_{ik}X_i = 0$$

이 성립함을 의미한다. (벡터 X_i는 X의 i번째 열벡터) 위 식은 설명변수들 간 성립하는 선형종속관계를 구체적으로 보여주는 것으로, 고유벡터의 각 원소 v_{ik}가 선형관계식에서 i번째 설명변수의 계수가 된다고 해석할 수 있다.

다중공선성의 존재여부는 $(\lambda_1/\lambda_p)^{1/2}$ 를 이용해 판별하는데, λ_1과 λ_p는 각각 $X'X$의 가장 큰 고유값과 가장 작은 고유값을 나타낸다. 40이 넘으면 다중공선성이 존재한다고 판단할 수 있다.

다중공선성을 해결하는 방법에는 **변수선택법**(Variable Selection), **차원축소**(Dimension Reduction), **정규화**(Regularization) 등이 있다.

이 방법들에 대해서는 다음 주에 깊게 배워보도록 합시다!!

8. Appendix

패키지에 뜬 거 보고 열심히 공부해서 썼어요... 조금만 힘내서 한번 봐봅시다!!!

1) Measurement Error, 그리고 내생성(endogeneity)

챕터 1에서, 오차항의 기대값이 0인 가정이 있었다는 것을 기억할 것이다. 그리고 이는 OLS 추정량이 BLUE가 되기 위한 가정 중 하나이기도 했다.

그렇다면 이 가정이 위배되는 경우는 어떻게 될까? 종속변수 Y에 Measurement Error가 발생하는 경우가 대표적인 예이다.

> 📌 **Measurement Error**
>
> 설문조사 등에서 응답자가 잘못 응답하거나, 설문 집단을 잘못 선정하여 회귀 모델에서 변수에 대해 정확하지 않은 측정 값을 사용할 때 발생하는 Error

먼저 Y(종속 변수)에 측정 오차가 발생하는 경우를 살펴보자. $Y^\star$를 참값, Y를 잘못 측정된 값이라고 정의하겠다. 그리고 v는 측정 시 발생하는 오차(measurement error)로, random variable이다.

그렇다면 $Y_i = Y_i^\star + \mu_i$ 과 같이 정의해보자. 이 식을 바탕으로 모델을 세워 본다면

$$y^\star = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e$$
$$\Leftrightarrow y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e + v$$

로, e 대신 $e+v$라는 새로운 오차항이 생겼다고도 볼 수 있을 것이다. y의 참값($y^\star$)으로 세운 위의 모델은 ①오차의 기댓값이 0이고, ②오차는 X와 공분산이 0이다 라는 가우스-마코프 정리의 가정을 만족한다고 하겠다. 밑의 식에서도, v 역시 위의 2가지 가정을 만족한다면 문제가 발생하지 않을 것이다.

그러나 만약 Measurement Error가 심각해서 v의 기대값이 0이 아닌 α가 되었다고 하면,

$$E(y|x) = (\hat{\beta}_0 + \alpha) + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

와 같이 절편 상수항이 왜곡(편향)되는 결과가 나타난다. 그렇지만 이는 관계 파악, 즉 기울기 추정량이 중요한 경우에는 사실 큰 문제가 되지는 않는다.

설명변수 X에 Measurement Error가 존재하는 경우에는 어떨까? 상황이 조금 복잡해진다. 2023-2 클린업 2주차에서 등장한 패키지과제가 이를 다루고 있다.

마찬가지로 $X^\star$를 참값, X를 잘못 측정된 값이라고 정의하자. 그렇다면 $X_i = X_i^\star + \mu_i$ 과 같이 정의될 수 있다.

그리고 이번엔 X의 Measurement Error μ에 몇 가지 가정을 해보겠다. 이는 Classical Error라고 하며, measurement error는 측정 장치에 의해 발생하며 오차의 크기는 측정되는 값에 의존하지 않는다는 살짝은 강력한 가정이다. 수식으로 표현하면 다음과 같다.

① $E(\mu) = 0$, ② $Cov(X^\star, \mu) = 0$

이제 모델을 세워보자.

$$\begin{aligned} y &= \beta_0 + \beta_1 X^\star + e \\ &= \beta_0 + \beta_1 (X - \mu) + e \\ &= \beta_0 + \beta_1 X - \beta_1 \mu + e \end{aligned}$$

로, 이번엔 e 대신 $-\beta_1\mu + e$ 라는 새로운 오차항이 생겼다고도 볼 수 있을 것이다.

그리고 Classical Error 가정 하에서,

$$\begin{aligned} Cov(X, \mu) &= E(X\mu) - E(X)E(\mu) \\ &= E(X\mu) = E[(X^\star + \mu)\mu)] \\ &= E(X^\star \mu) + E(\mu^2) = \sigma_\mu^2 \end{aligned}$$

이며, 이는 곧 우리가 세운 모델에서

$$
\begin{aligned}
Cov(X, e - \beta\mu) &= E(Xe) - E(X\beta\mu) \\
&= -\beta E(X\mu) = -\beta Cov(X, \mu) \\
&= -\beta\sigma_\mu^2 \neq 0
\end{aligned}
$$

X와 오차항 간에 상관관계가 존재하는 것이고, 이를 **내생성 문제**가 발생하였다고 한다. 이는 필연적으로 OLS추정량에 bias 를 일으킨다.

내생성 문제가 발생하는 원인은 다양하다. Measurement Error는 그 원인 중 하나인 것

과정은 생략하겠지만, 내생성 문제가 있는 모델의 OLS추정량에 극한을 취하게 되면

$$
plim(\hat{\beta}^{OLS}) = \beta(\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_{x^\star}^2})
$$

으로, OLS추정량에 1보다 작은 값을 곱하는 꼴로 표현된다. 그리고 이러한 bias를 Attenuation bias라고 한다.

내생성 문제를 처방하기 위해서는 **도구 변수**를 사용할 수 있다.
간단히 설명하자면, X와는 상관관계가 있지만 오차와는 상관관계가 없는 변수를 찾는 것이다. X가 오차항에 의해 **휘둘리고 있으니** 그러한 오차항과 관련이 없는 변수 Z를 찾고, 그 변수가 X를 움직이는 것이 얼마나 Y를 움직이는지를 보겠다는 것.

패키지과제를 예로 들자면, 강의실에서 집까지의 거리가 출석률과 음의 상관관계는 있지만, measurement error가 존재하지 않으며, 오차항과 무관할 것이다. 그러므로 dist(거리)를 atndre(출석률)을 위한 도구 변수로 사용할 수 있을 것이다.

X와는 상관관계가 높지만 Y와는 또 무관하다는 강력한 가정이 필요하며, 도구변수를 사용하게 되면 분산이 커지기 때문에 잘 사용하지는 않습니다!!! 더 깊게 들어가려면 고려해야 할 것이 많고, 계산과 수식 또한 복잡해지므로 이 정도만 하겠습니다~!