회귀분석(Regression Analysis)

목 차

1. 단순선형회귀
   (1) 기초개념
   (2) 선형회귀의 기본 가정
   (3) 단순선형회귀모형
   (4) 회귀계수 추정
   (5) 잔차분석

2. 중회귀분석
   (1) 기본 개념
   (2) 교호작용
   (3) 더미변수

3. 회귀모형의 수정
   (1) 표준화 회귀계수
   (2) 다중공선성

4. 결과해석
   (1) 잔차와 지렛값
   (2) 영향력 관측치
   (3) F 통계량, T 통계량 , P- value

5. 회귀모형의 선택
   (1) 변수 선택

1. 단순 선형 회귀

(1) 회귀분석 기초 개념

- 회귀분석(regression analysis) : 독립변수(x) 와 종속변수(y)를 통하여 회귀모형 가정

-> 가정된 회귀모형을 통하여 종속변수를 예측 또는 통계적 추론을 하는 분석 기법.

- 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링 등의 통계적 예측 등에 이용될 수 있다.
- 몸무게(Y) = 키(x1) + 성별(x2) + 인종(x3) : 선형회귀모형

1. 단순 선형 회귀

(2) 선형 회귀의 기본 가정

① 회귀모형은 다음과 같이 모수에 대해 선형(linear)인 모형이다

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, 2, \cdots, n$$

② 수집된 데이터의 확률분포는 정규분포를 따른다

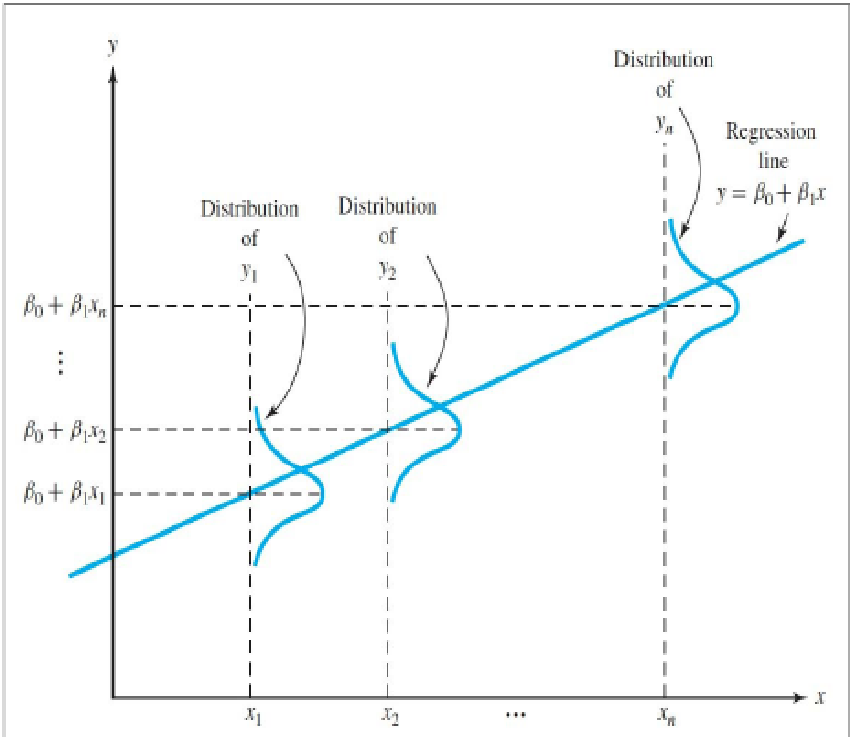
③ 오차항에 대한 가정

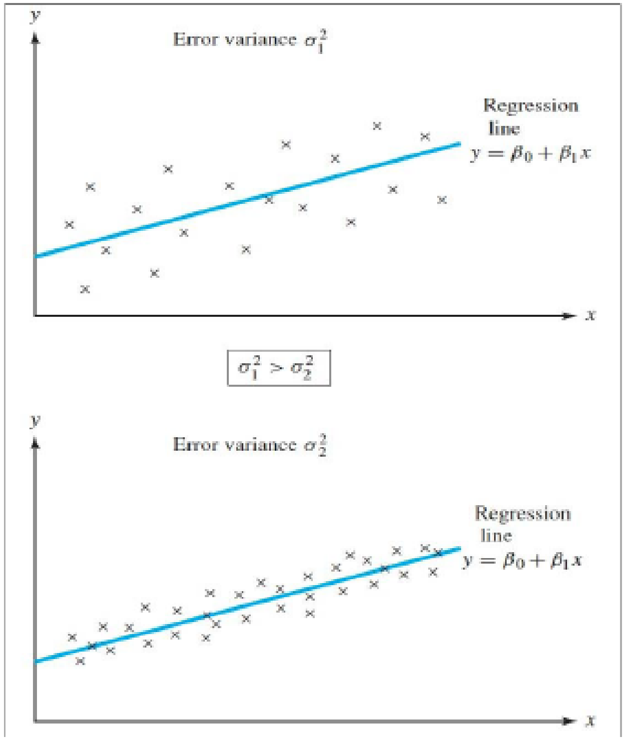
- 정규성 : 오차는 평균이 0, 분산이 σ^2 인 정규분포를 따름
- 등분산성 : 오차항은 모든 독립변수 값에 대하여 동일한 분산 σ^2 을 가짐
- 독립성 : 오차들을 서로 독립

④ 독립변수 상호간에는 상관관계가 없어야 한다

1. 단순 선형 회귀

$$\epsilon_i \sim N(0, \sigma^2) \quad \rightarrow \quad Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$



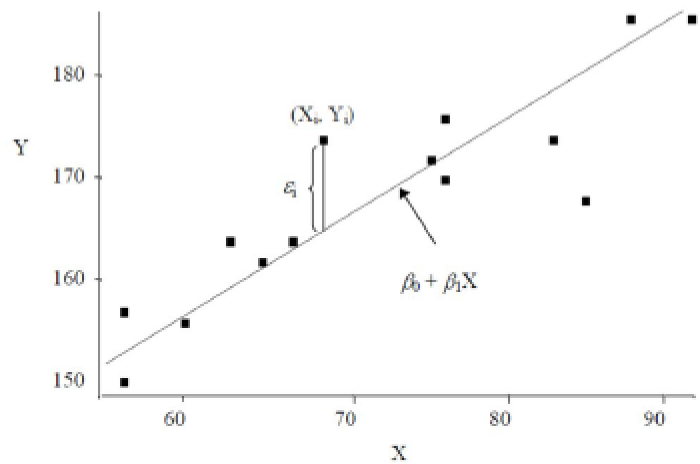


(3) 단순 선형 회귀 모형

- 단순 선형 회귀모델(simple linear regression model) : 독립변수 1개, 종속변수 1개

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, 2, \cdots, n$$

- Y_i : 종속변수 (dependent variable)
- X_i : 독립변수 (independent variable)
- β_0, β_1 : 회귀계수 (regression coefficient)
- ε_i : 오차항



BOAZ

(4) 회귀계수 추정

- 최소제곱법(method of least squares) : 오차항의 제곱합이 최소가 되는 회귀계수β_0, β_1추정

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

이를 최소화하는 계수 β_0, β_1을 찾기 위해 각각에 대하여 편미분

$$\frac{\partial Q}{\partial \beta_0} = -\sum_{i=1}^{n} 2[Y_i - (\beta_0 + \beta_1 X_i)] = 0, \qquad \frac{\partial Q}{\partial \beta_1} = -\sum_{i=1}^{n} 2X_i[Y_i - (\beta_0 + \beta_1 X_i)] = 0$$

(5) 잔차분석

- 잔차분석이란 오차의 추정치인 잔차를 이용하여 다음 정보를 얻어내는 과정을 의미
- 위의 회귀모델의 기본 가정들을 잘 따르는지에 대한 분석이다

① 설명변수와 종속변수 관계는 선형인가?
② 오차의 등분산성, 독립성, 정규성을 만족하는가?

2. 중선형 회귀

(1) 개념

다중회귀분석 :
설명변수(독립변수)가 2개 이상인 회귀모형을 분석대상으로 삼는 것

예) 아파트 가격=
a+b(평수)+c(연령)+d(단지규모) (a,b,c,d는 회귀계수)

*기본 가정:
독립변수는 2개 이상이며 각 독립변수는 종속변수와 선형 관계에 있다

다중 회귀 분석의 의의

① 분석내용을 향상시킬 수 있다

② 추가적인 독립변수를 도입함으로써 오차항의 값을 줄일 수 있다

③ 회귀계수 추정량의 편의를 제거할 수 있다
-> 단순회귀분석에서는 종속변수에 대한 중요한 설명변수를 누락함으로써 계수 추정량에 대해 편의를 야기할 수 있음. 다중회귀분석 통해 제거 가능

2. 중선형 회귀

다중 회귀 모형의 구조 -다중회귀모형의 일반형

- 종속변수 Yi가 상수항(α)과 k개의 독립변수 X1i, X2i, …., Xki 에 설명되는 모집단의 다중회귀모형

- 설명변수는 k개 존재하고 모수는 k+1개 존재하게 된다

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

$$\begin{cases} y_i & : \text{종속변수} \\ x_{1i}, x_{2i}, \cdots, x_{ki} & : \text{독립변수} \\ \beta_0, \beta_1, \cdots, \beta_k & : \text{회귀계수} \end{cases}$$

BOAZ

다중 회귀 모형의 구조 -다중회귀모형의 벡터표현

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\
&= (1,\ X_{i1},\ X_{i2}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon_i
\end{aligned}
$$

다중 회귀 모형의 구조 -다중회귀모형의 행렬표현

$$Y = X\beta + \varepsilon \ , \quad \varepsilon \sim N(0, I\sigma^2)$$

$$\begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdot \quad \cdot + \beta_n x_{1n} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdot \quad \cdot + \beta_n x_{2n} \\ \cdot \\ \cdot \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdot \quad \cdot + \beta_n x_{nn} \end{pmatrix}$$

2. 중선형 회귀

최소제곱법을 이용한 다중회귀계수의 추정

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})^2$$

$$\begin{aligned} S &= \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}' \boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} \end{aligned}$$

$\Rightarrow$ $\boldsymbol{\beta}$ 에 관해 미분 $\dfrac{\partial S}{\partial \boldsymbol{\beta}} = -2X'\boldsymbol{Y} + 2X'X\boldsymbol{\beta} = 0$

$===> \quad X'X\boldsymbol{b} = X'\boldsymbol{Y}$

$===> \quad \boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{Y}$

더미변수

- 범주형 설명변수(독립변수)를 의미. 명목척도(범주형)로 측정된 변수

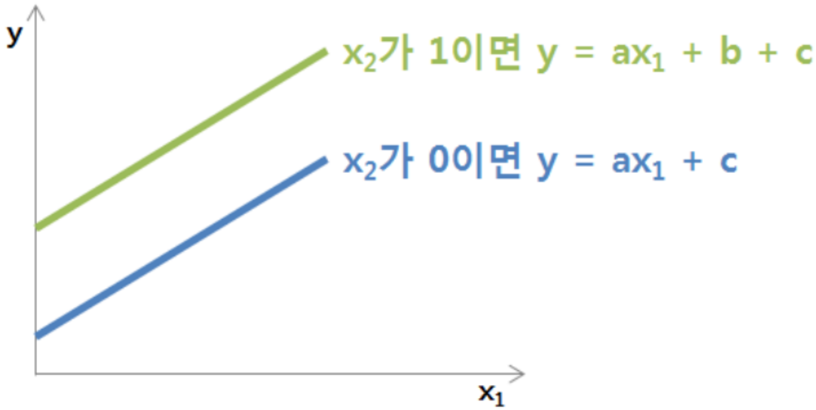
-> 범주형 변수를 연속형 변수로 만들어 회귀분석에서 범주형 변수를 회귀분석에서 사용할 수 있게끔 하는 것이 포인트

- 0과 1로 나타내며 다른 회귀계수들을 추정하는데 영향을 미치지는 않는다

- K 개의 범주를 가지는 경우 더미변수는 k-1개

더미변수

회귀식 : $y = ax_1 + bx_2 + c$

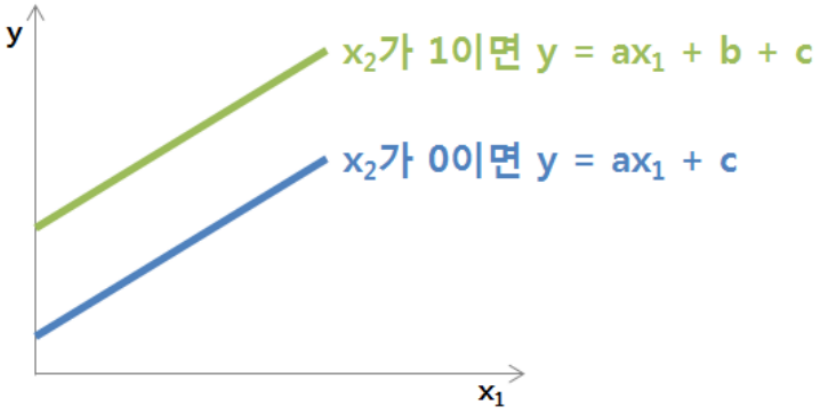


원래 회귀식에서 x2가 1이면 b만 남아서 y절편은 b+c가 됨

원래 회귀식에서 x2가 0이면 b도 0이되어서 y절편은 c가 됨

더미변수

회귀식 : $y = ax_1 + bx_2 + c$



- 더미변수는 회귀 기울기를 바꾸지 않고 절편만 바꾸어 평행 이동시키는 역할

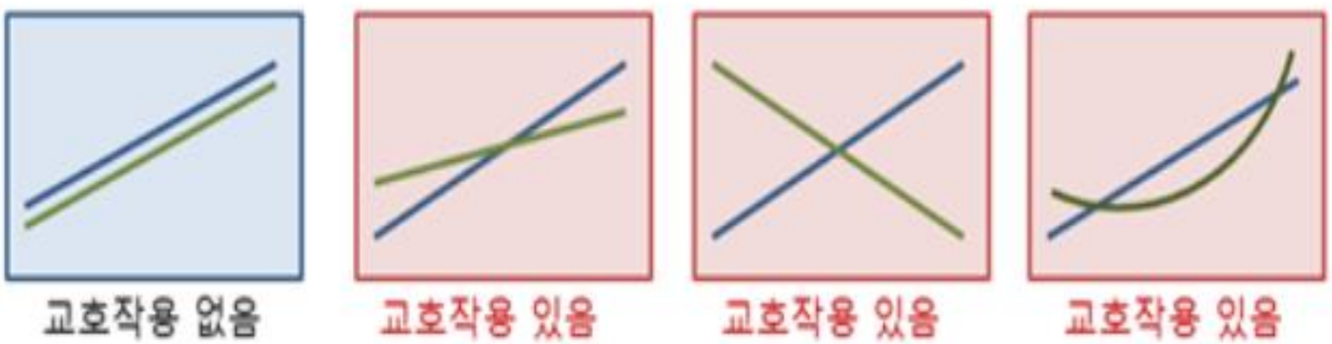
- 더미변수를 첨가 전후의 회귀모델이 평행하게 나타날 때 더미변수의 첨가적 효과(additive effect)가 있다고 한다

- 더미변수와 설명변수(독립변수)간에 교호작용이 있는 경우 기울기가 달라진다

교호작용

-독립변수 간의 조합으로 인해 나타나는 효과를 의미

Y=a+bX1+CX2+d(X1*X2) 로 표현하고, (X1*X2)항이 교호작용 항



표준화 회귀계수

- 회귀계수는 각각의 설명변수(독립변수)와 종속변수간의 관계를 표현하는데, 측정 단위에 따라 회귀계수가 달라진다

- 다중회귀의 경우 각 독립변수는 단위나 수치의 크기 범위가 서로 다른 경우가 많이 발생한다
-> 따라서 회귀계수의 크기 비교를 위해 회귀계수를 표준화

- 설명변수의 표준화한 회귀계수가 크다는 것은 이 설명변수에 의해 종속변수가 더 큰 영향을 받고 있다는 것을 의미한다

표준화 회귀계수를 계산하는 방법

- 방법 1) 변수의 표준화 이용

$$z_{ij} = \frac{x_{ij} - \bar{x}}{s(x_i)}, \ i = 1, \cdots, p; j = 1, \cdots, n$$

- 방법 2) 표준편차를 이용

$$b_k^* = b_k \times \frac{s_k}{s_y}, \ k = 1, \cdots, p$$

(bk*: k번째 독립변수의 표준화된 회귀계수, bk: k번째 독립변수의 추정된 회귀계수, Sk: k번째 독립변수의 표준편차, Sy: 종속변수의 표준편차

다중공선성

- 일반적으로 회귀모형에서 독립변수간에 정확한 선형관계(완전공선성)는 나타나지 않는다

- 그러나 독립변수들 간에 상관관계가 높게 나타나는 문제가 발생하는데 이를 다중공선성이라고 한다

- 즉, 회귀분석에서 종속변수와 독립변수 간에 선형성은 전제조건으로 존재하지만 독립변수들 간에는 선형관계가 없어야 한다!!

다중공선성으로 인한 문제

a) 계수 추정량의 분산 값이 커진다
-> t-검정 통계량이 작아져 귀무가설 (H0: $\beta_j = 0$)을 기각할 가능성이 희박해진다
-> 중요한 회귀계수임에도 검정결과 유의하지 않게 나타남

b) 계수 추정량이 데이터의 크기 변화, 혹은 설명변수의 누락 또는 부적절한 변수의 포함 등에 의해 민감하게 변화한다
->
하나의 데이터를 바꾸거나 제거할 때 계수 추정량에 큰 변화
하나 이상의 설명변수를 추가(제거)하면 계수 추정량에 큰 변화

다중공선성의 점검

a) 높은 R square 값과 낮은 t-검정치
-> R square값은 크고 개별 회귀계수에 대한 t-검정치가 낮은 경우, 모형 전체의 설명력은 높으나 각 계수의 추정치의 표준오차가 매우 크다는 것을 의미
이는 독립변수간에 다중공선성이 높을 때 나타나는 현상

b) 설명변수들 간의 높은 상관계수 값
-> 상관계수가 0.5~0.9 이상이면 독립변수들 간에 다중공선성이 존재한다고 간주.

c) VIF 지수 확인
-> 보통 10 이상일 경우 다중공선성이 존재한다고 간주.

3. 회귀 모형의 수정

다중공선성에 대한 대책

a) 설명 변수의 제외
-> 상관관계 높은 변수 중 하나 또는 일부를 모형에서 제거

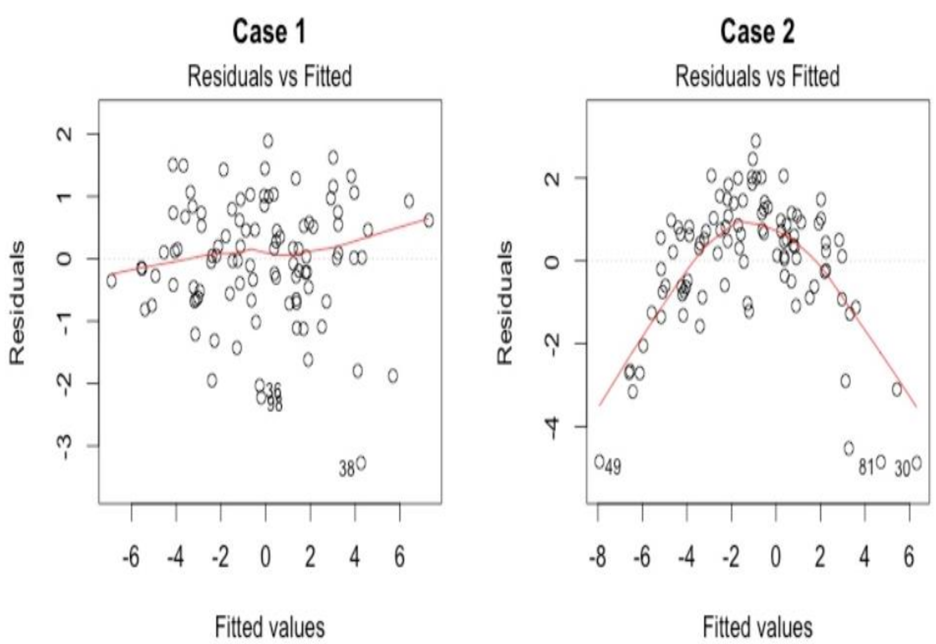
b) 새로운 관측치를 추가

c) 능형회귀(ridge regression) 또는 주성분회귀(principal components regression) 등의 추정법 사용

4. 결과 해석

(1) 잔차와 지렛값
- 그래프 분석 : 그래프의 경향성 확인

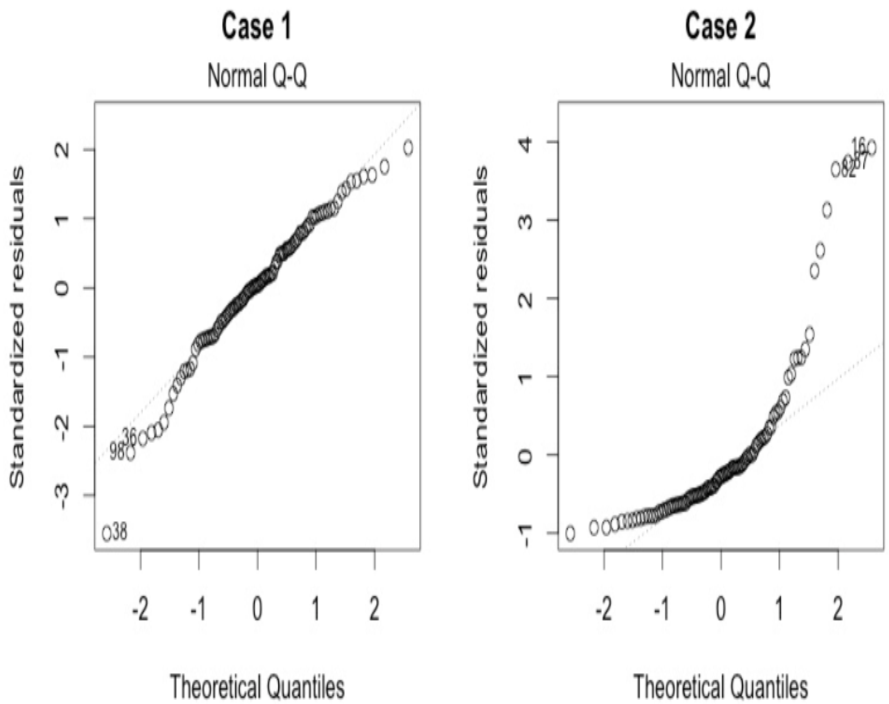
(1) residuals vs fitted



- x축 : 선형회귀로 예측된 y값
- y축 : 잔차
- 선형 회귀에서 오차는 평균이 0이고 분산이 일정한 정규분포 가정
- 따라서 예측된 y값과 무관하게 잔차의 평균은 0이고 분산은 일정해야 함
- 직선을 기준으로 데이터들의 분포가 특정한 경향을 보이지 않으면 등분산성을 만족

4. 결과 해석

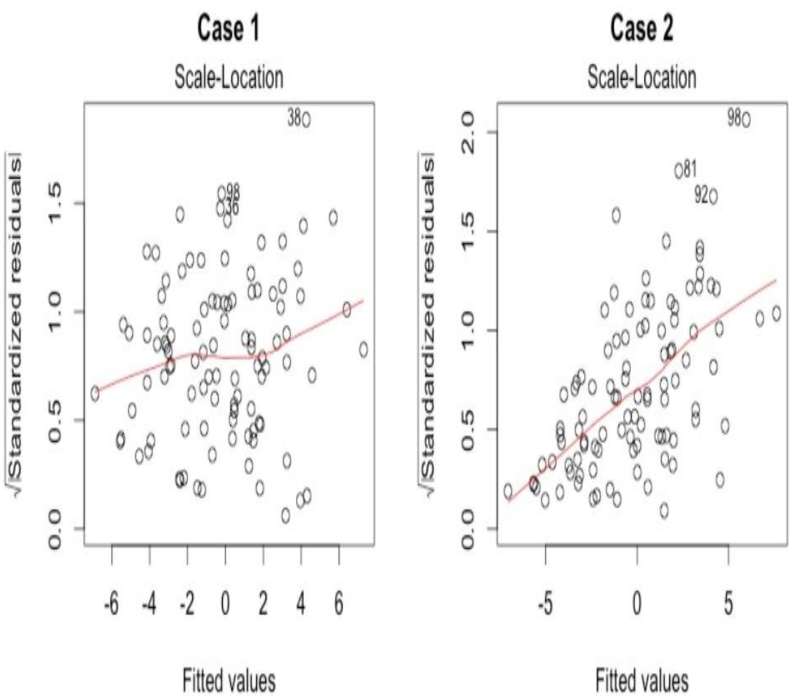
(2) Normal Q-Q



- 잔차가 정규분포를 따르는지 확인하기 위한 Q-Q도
- x축 : 통계적 모집단의 분위수(Quantile) , 회귀분석 가정에 따라 정규분포 따름
- y축 : 데이터 샘플의 표준화 잔차(회귀모델)
- 선형(y=x)에 가까울수록 회귀모델이 잘 추론되었음을 의미

4. 결과 해석

(3) Scale – Location



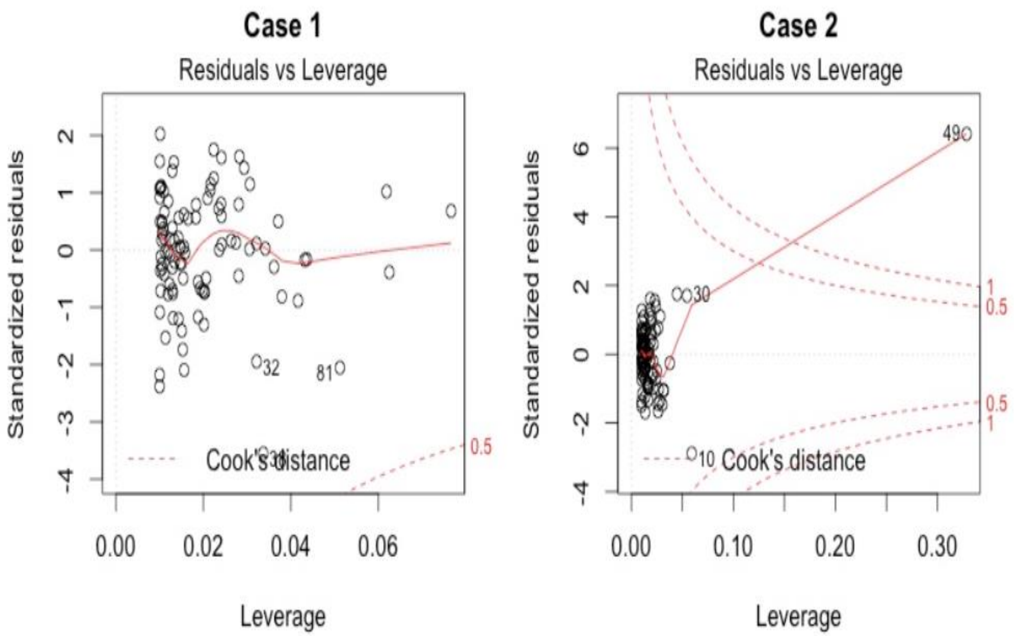
- x축 : 회귀식에서 예측값
- y축 : 표준화 잔차에 루트를 씌운 값

- 가능한 값들이 한 곳에 몰리지 않고 골고루 분포(오차의 등분산성)하며 잔차의 값이 작을수록 좋은 회귀 모형임

- 특정 위치에서 0에서 멀리 떨어진 값이 관찰된다 = 표준화 잔차가 크다, 이상치이다.

4. 결과 해석

(4) Residuals vs Leverage

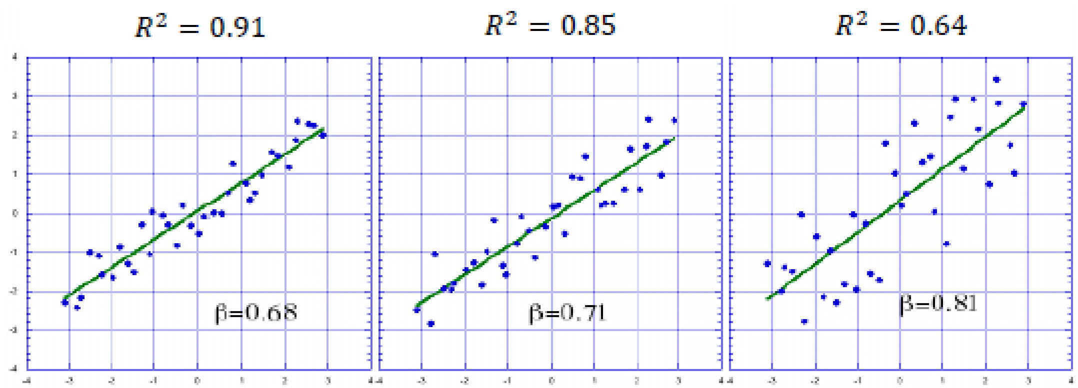


- x축 : leverage(지렛값), 하나의 관측치가 예측 모형에서 많이 벗어나 전체 모형의 회귀 계수에 큰 영향을 주는 값(outlier)
- y축 : 표준화 잔차
- 쿡의 거리(Cook's Distance) : 지렛값의 크기를 나타내는 정도

4. 결과 해석

영향력 관측치

(1) 결정계수 R^2 (coefficient of determination) :
- 회귀모형에 의해 설명되어지는 변동의 비율
– R^2값이 클 수록 변동에 대한 설명력이 커지는 것이므로 좋은 모델이다.
- 일반적으로 0.8 이상이면 좋은 모델이라고 판단



4. 결과 해석

(2) Adjusted R square : 일반적으로 R^2는 독립변수의 수가 많아질수록 설명력이 높아지므로 함께 커진다. 이를 방지하기 위해 독립변수의 수에 벌점을 부여하여 adjusted R square값을 구하면 다음과 같음

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2), \qquad 0 \leq R^2_{adj} \leq 1$$

4. 결과 해석

F통계량, T통계량, P-value

- 회귀 모형의 타당성을 평가할 때, 즉 구축한 회귀모형이 원래의 데이터를 잘 설명하는지 를 확인하기 위해 볼 수 있는 지표

(1) F 통계량 - 모형이 통계적으로 유의미한지를 확인하는 지표; 값이 클수록 통계적으로 유의하다고 판단
- p-value와 함께 확인 (p-value는 작을수록 유의미)

(2) T 통계량 - 회귀 모형에서 각각의 계수(회귀계수)의 타당성을 검증하는 지표; t분포에서 회귀계수가 유의미하다는 기각역을 계산한 면적의 넓이가 p-value ; p-value가 일정수준 이하로 작다는 것 -> 기각역에 들어갈 확률이 커짐 -> 회귀계수의 설명력 증가

4. 결과 해석

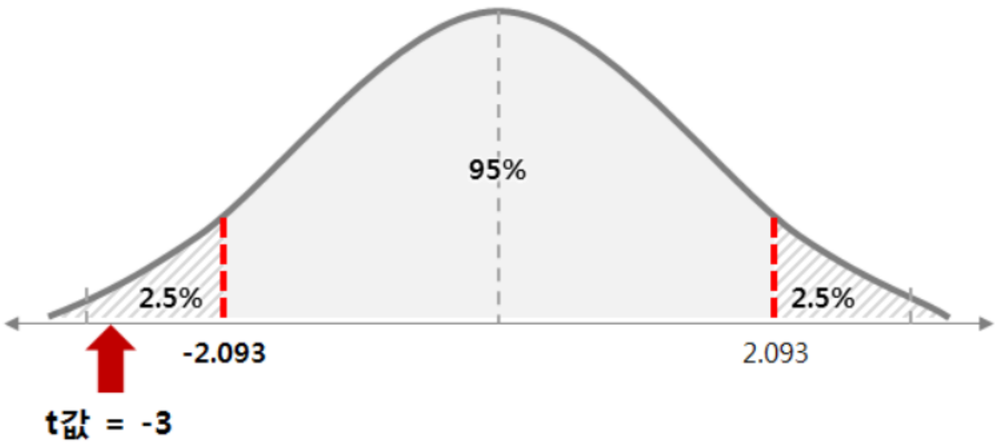
- β_i 의 타당성 검증

(1) 가설

$$H_0 : \beta_i = 0 \qquad vs. \qquad H_1 : \beta_i \neq 0, \qquad i = 0, 1, \cdots, k$$

(2) 기각역

$$t < -t_{\alpha/2, n-k-1} \qquad or \qquad t > t_{\alpha/2, n-k-1}$$



4. 결과 해석

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11169 -0.06756 -0.04627  0.04689  0.19610

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.9924243  0.2941729   3.374 0.008210 **
xproduction          0.3493581  0.0643149   5.432 0.000415 ***
xcolling degree days 0.0011870  0.0003109   3.818 0.004102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1126 on 9 degrees of freedom
Multiple R-squared:  0.9245,    Adjusted R-squared:  0.9077
F-statistic: 55.07 on 2 and 9 DF,  p-value: 8.949e-06
```

5. 회귀 모형의 선택

(1) 변수 선택

회귀모형을 구축하기 위해서는 어떤 변수를 선택해야 할지 고려해야 한다.

- 변수 선택법(variable selection) :
종속변수(Y)에 영향을 주는 독립변수(x) 선택, 영향을 주지 않는 독립변수 제거

- 변수 선택법의 종류 :
(1) 전진선택법(forward selection) :
독립변수가 전혀 없는 상태에서 변수를 하나씩 추가하면서 회귀모델을 평가
(2) 후진제거법(backward elimination) :
모든 독립변수를 고려한 회귀모델에서 하나씩 변수를 제거하면서 모델 평가
(3) 단계적 선택법(stepwise selection) :
독립변수가 전혀 없는 상태에서 변수를 하나씩 추가하거나 제거하면서 회귀모델을 평가

- 변수선택법의 척도 : 언제까지 변수를 추가하거나 제거해야 하는가?의 기준
1) R^2_{adj} : 앞에서 설명한 전체 회귀모형의 설명력을 평가하는 결정계수, 클수록 좋음
2) MS_E : R^2_{adj}를 구할 때 한 회귀모델 분산분석에서 오차의 평균제곱합, 작을수록 좋음