회귀분석팀

6팀

신성민
신유정
김찬영
윤주희
이혜인

INDEX

회귀 가정의 기본

기본가정 가정의 위배

정규성 가정

오차 $\epsilon_{1,}\epsilon_{2,}\cdots,\epsilon_n$ 는 **정규분포**를 따른다

등분산 가정

오차 $\epsilon_{1,}\epsilon_{2,}\cdots,\epsilon_n$ 는 평균이 0이고 **동일한 분산**을 가진다

독립성 가정

오차 $\epsilon_{1,}\epsilon_{2,}\cdots,\epsilon_n$ 는 **서로 독립**이다

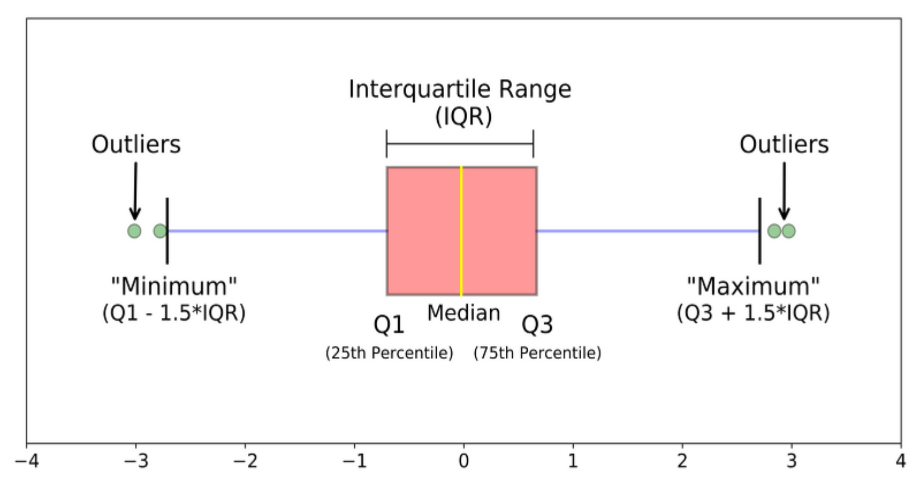
Cook's Distance

$$C_i = \frac{r_i^2}{p+1} * \frac{p_{ii}}{1-p_{ii}} (r_i\text{:표준화잔차},\ p_{ii}\text{:지레값})$$

- 표준화잔차가 클수록, 지레값이 클수록 값이 커짐
- 보통 1보다 클 경우 해당 데이터를 영향점으로 생각

특이값의 확인과 해결

데이터 삭제



IQR를 이용하여 위의 범위를 벗어난 데이터를 outlier로 평가 및 제거함

(범위 : 1분위수 -1.5*IQR < 데이터 < 3분위수+1.5*IQR)

로버스트 회귀

영향점들을 고려하는 회귀식이 따로 존재

(Least Trimmed Squares, Huber's M-estimation등)

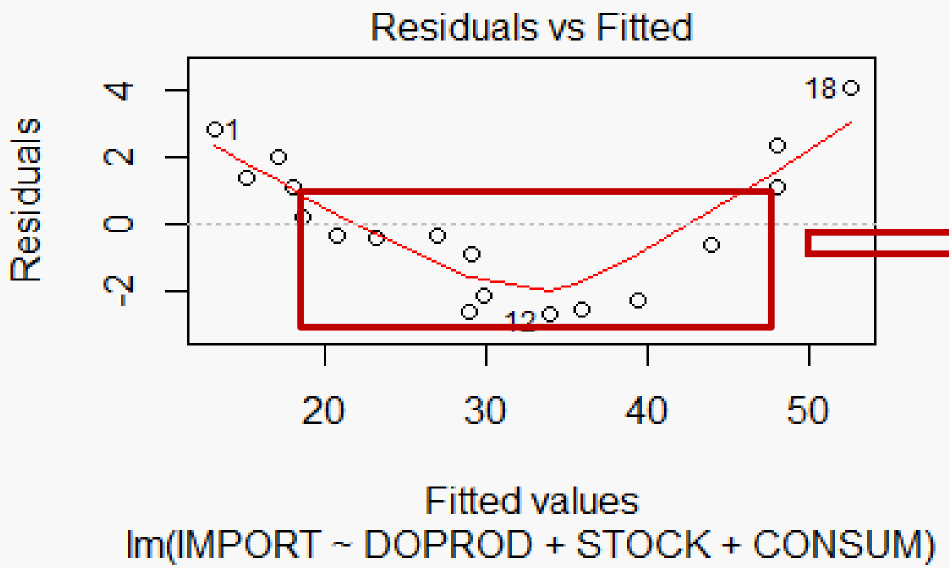
(궁금하다면, 19-2학기 회귀 교안 GO!)

선형성 진단&처방

진단 처방

Residuals vs fitted

```
data <- read.delim('P241.txt')
lm_fit <- lm(IMPORT~DOPROD+STOCK+CONSUM, data=data)
plot(lm_fit, which = 1)
```



분포가 고르지 않으니
비선형적이다!

이건 그래프가
우릴 보고 웃고있는 수준..
행복하니..?

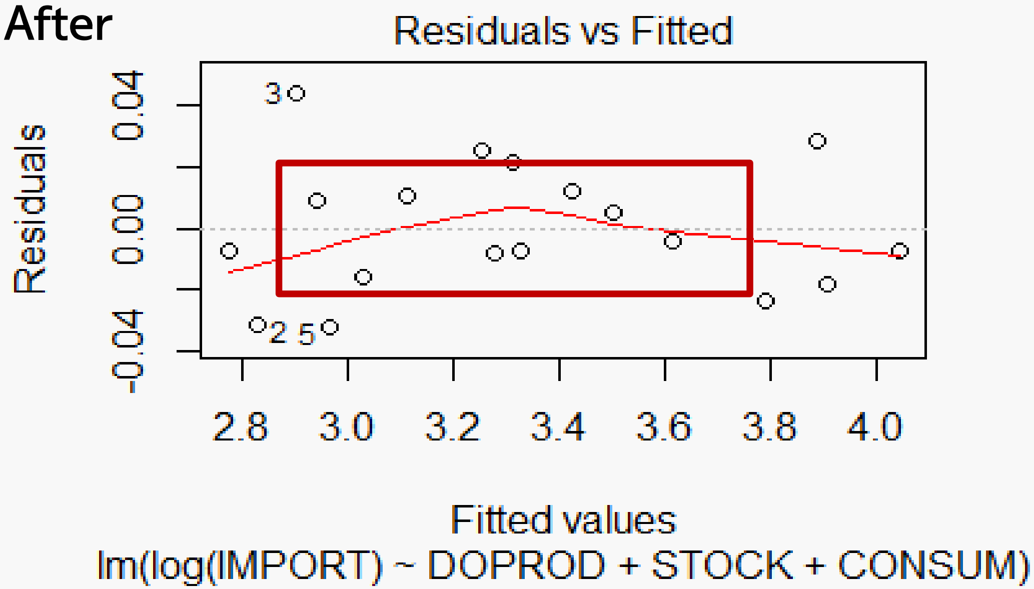
선형성 진단&처방

잔차플롯 이용 **처방**

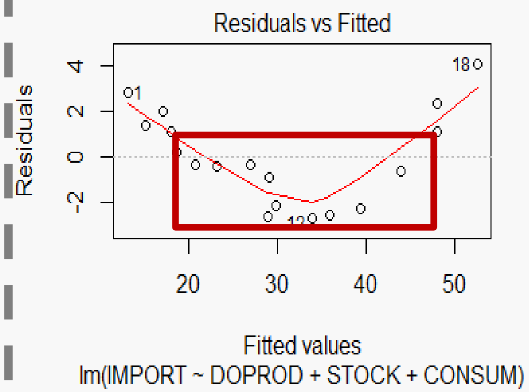
변수 변환 – log 변환

```
# 처방 1 - log 변환
log_fit <- lm(log(IMPORT)~DOPROD+STOCK+CONSUM,data=data)
plot(log_fit, which = 1)
```

After



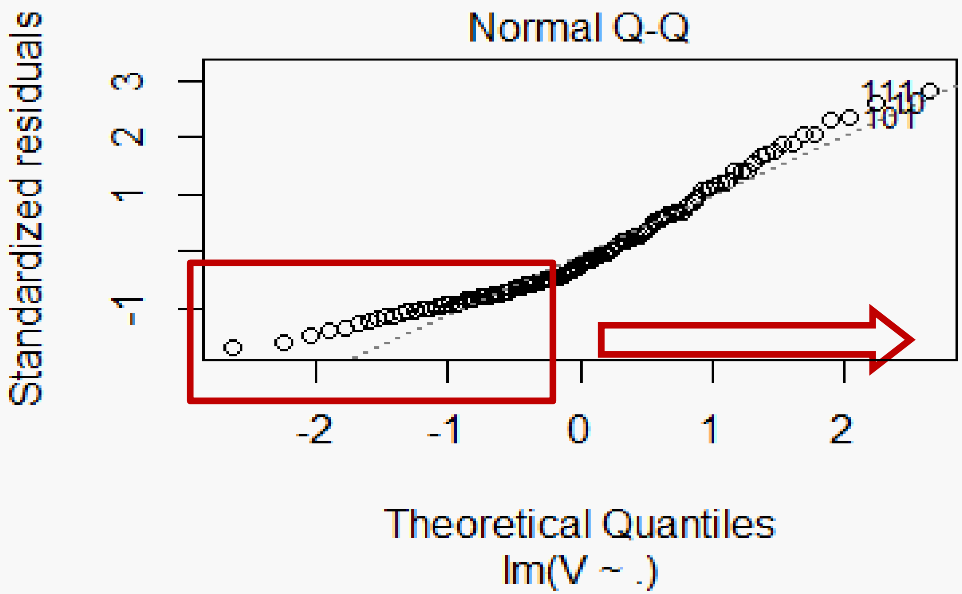
Before



정규성 진단&처방 잔차플롯 이용 통계적 기법 이용 처방

Normal Q-Q

```
data <- read.delim('P188.txt')
lm_fit2 <- lm(V~., data=data)
plot(lm_fit2, which=2)
```



정규분포 사분위수의
Y=X 그래프와 가깝지 않으니
정규성을 만족하지 않는다!

6 오차의 정규성

정규성 진단&처방

잔차플롯 이용 **통계적 기법 이용** 처방

Shapiro-wilk Test

```
> shapiro.test(lm_fit2$residuals)

        Shapiro-Wilk normality test

data:  lm_fit2$residuals
W = 0.94169, p-value = 5.577e-05
```

Jarque-Bera Test

```
> jarque.bera.test(lm_fit2$residuals)

        Jarque Bera Test

data:  lm_fit2$residuals
X-squared = 11.304, df = 2, p-value = 0.003511
```

Anderson-Darling Test

```
> ad.test(lm_fit2$residuals)

        Anderson-Darling normality test

data:  lm_fit2$residuals
A = 2.327, p-value = 6.303e-06
```

정규성 진단&처방 잔차플롯 이용 통계적 기법 이용 처방

Yeo-johnson Transformation

Box-Cox Transformation의 한계를 극복하는 변환 방식

Y가 음수일 때에도 전부 변환이 가능

λ(Lamda) 구하는 방식은 Box-Cox와 동일

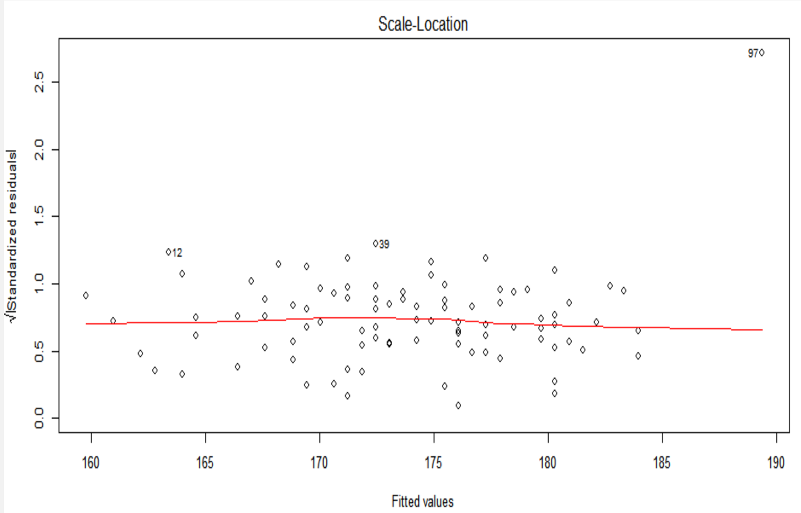
수식

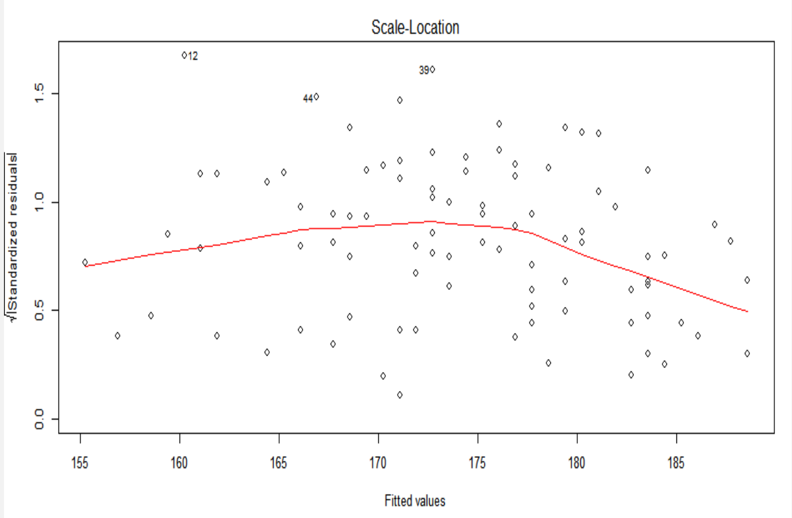
$$\psi(\lambda, y) = \begin{cases} ((y+1)^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0 \\ -[(-y+1)^{2-\lambda} - 1)]/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

오차의 등분산성

등분산성 진단   잔차플롯   통계적기법 1)BP-Test   2)MLT

<Scale-Location plot>





➡ plot의 잔차들이 패턴 없이 random하게 분포되어 있는지를 확인!

- 좌측: 등분산성 만족, 우측: X값에 따라 분산이 달라지는 경향이 보임
- 우측의 경우, 다른 진단 필요없이 이분산성 의심가능함

등분산성 진단 잔차플롯 통계적기법 1)BP-Test 2)MLT

Breusch-Pagan Test

- 기본가정: 잔차가 설명변수 X에 의해 영향 받는 확률변수
- X와 잔차제곱 간의 선형결합식 만들고 이를 검정해 둘의 연관성 파악

$$e_i^2 = b_0 + b_1 x_1 i + \ldots + b_K x_K i + \nu_i$$

이 선형회귀식을 추정하여 R^2를 구함

$$F = \frac{\frac{R^2_{\hat{\varepsilon}^2}}{1}}{\frac{(1 - R^2_{\hat{\varepsilon}^2})}{n-2}} \text{ or } \chi^2 = nR^2_{\hat{\varepsilon}^2}$$

R^2 ↑ F통계량 ↑ p-value ↓
귀무가설 기각!
=> 이분산성

등분산성 위배 처방 **WLS** 변수변환

Weighted Linear Square

- 데이터마다 분산이 다른 것을 고려, 기존 LSE에 분산에 대한 가중치 부여

$$\sum_{i=1}^{n} w_i(y_I - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

- 위의 식을 세운 뒤, 회귀식을 찾는 과정
- w_i 는 경험적으로 찾아야함

8 오차의 독립성

독립성 진단 잔차플롯 통계적기법 1)DW Test 2)run_test

Durbin Watson Test

- 바로 앞,뒤 오차들 간의 **자기상관 존재 여부**를 확인하는 검정

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad \hat{p} = Cor(e_t, e_{t-1}) = \frac{\sum_{t=2}^{n}(e_t - \bar{e})(e_{t-1} - \bar{e})}{\sum_{t=1}^{n}(e_t - \bar{e})^2} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

오차 e_t, e_{t-1} 간의 상관관계

수식을 D와 $\hat{p}$ 이용해 새로 풀어내면

$$\boxed{D} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{\sum_{t=2}^{n}(e_t^2 + e_{t-1}^2 - 2e_t e_{t-1})}{\sum_{t=1}^{n} e_t^2} = 1 - 2\hat{p} + \frac{\sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=1}^{n} e_t^2} \approx \boxed{2(1-p)}$$

독립성 위배 처방 시계열 가변수

독립성 위배 처방 방법

- 시계열 분석
- 가변수 만들기

  Ex) 뚜렷한 계절성을 보이는 데이터가 있을 경우, 가변수 만들어 새롭게 설명

- Cochrane-Orcutt

  - 더빈왓슨 통계량의 가정인 '앞, 뒤의 오차항들이 독립이 아니다'를 기반으로 변환