대안탐색

변수 변환

- 가정이 만족되지 않는 경우의 첫 번째 대안
  - 동일 분산 혹은 정규성 가정의 문제: 반응변수의 변환
  - 선형관계의 문제: 설명변수의 변환 혹은 다항회귀모형

- 반응변수의 변환: Box-Cox 변환

- 설명변수의 변환: Box-Tidwell 변환

- 반응변수의 변환: Box-Cox 변환
  - 양의 값을 갖는 반응변수에만 적용 가능한 변환 방법
  - 변환 방식: Y를 $g_\lambda(Y)$로 변환

$$g_\lambda(y) = \begin{cases} y^\lambda & , \lambda \neq 0 \\ \log y & , \lambda = 0 \end{cases}$$

λ는 우도함수를 최대화시키는 조건으로 계산

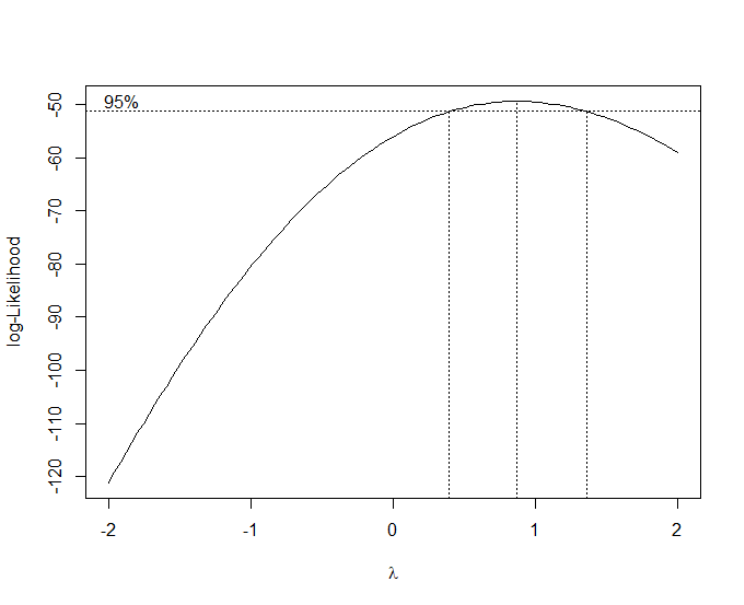
- R 함수
  - MASS :: boxcox( )
  - car :: powerTransform( )

- 예: states 자료에 대한 회귀모형

```
> library(dplyr)
> states <- as.data.frame(state.x77) %>%
    select(Murder, Population, Illiteracy, Income, Frost)
> fit_s <- lm(Murder ~ ., states)
```

- 함수 MASS::boxcox( )에 의한 변환 탐색

```
> library(MASS)
> bc <- boxcox(fit_s)
> names(bc)
[1] "x" "y"
> bc$x[which.max(bc$y)]
[1] 0.8686869
```



- $\hat{\lambda} = 0.8686$으로 추정
- 변환된 반응변수의 해석 문제
- λ의 95% 신뢰구간에 1 포함
- 변환이 필요 없는 경우

- 함수 car::powerTransform( )에 의한 변환 탐색

```
> library(car)

> summary(powerTransform(fit_s))
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.8653           1       0.3853       1.3453

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                             LRT df      pval
LR test, lambda = (0) 13.14361  1 0.0002885

Likelihood ratio test that no transformation is needed
                              LRT df    pval
LR test, lambda = (1) 0.2990438  1 0.58448
```

- 설명변수의 변환: Box-Tidwell 변환

  - Box와 Tidwell이 제안한 모형

$$Y_i = \beta_0 + \beta_1 X_{1i}^{\lambda_1} + \cdots + \beta_k X_{ki}^{\lambda_k} + \varepsilon_i$$

  모수 $\lambda_1, \ldots, \lambda_k$를 MLE로 추정

- R 함수

  - car :: boxTidwell( )

  - 사용 예: boxTidwell(y ~ x1 + x2 , other.x = ~ x3 + x4 , data=df)

    - 반응변수: y , 설명변수: x1, x2, x3, x4
    - 변환 필요 여부를 확인하려는 변수: x1, x2
    - 변환을 고려하지 않는 변수: x3, x4

- 예: states 자료의 모형 fit_s에서 Population, Illiteracy의 변환 필요 여부 확인

```
> library(car)

> boxTidwell(Murder ~ Population + Illiteracy,
             other.x= ~ Frost + Income, data=states)
            MLE of lambda Score Statistic (z) Pr(>|z|)
Population         0.8451              -0.3709   0.7107
Illiteracy         1.4024               0.6893   0.4907

iterations =  21
```

- $\hat{\lambda}_1 = 0.84, \quad \hat{\lambda}_2 = 1.4$로 추정
- Score 검정 결과 두 모수는 모두 비유의적
- 변환이 필요 없는 상황

- 예: women 자료의 fit_w에서 height의 변환 필요 여부 확인

```
> boxTidwell(weight~height, data=women)
 MLE of lambda Score Statistic (z)  Pr(>|z|)
        4.2008                13.067 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  2
```

- $\hat{\lambda} = 4.2$로 추정
- height 대신 height^4을 설명변수로 사용한 회귀모형 설정
- 최고차항인 height^4만을 설명변수로 사용한 회귀모형은 큰 의미가 없음

  height $\rightarrow$ height+a 로 위치 이동하면

$$Y = \beta_0 + \beta_1(height + a)^4 + \varepsilon$$
$$= \beta_0 + \beta_1 a^4 + 4\beta_1 a^3 height + 6\beta_1 a^2 height^2 + 4\beta_1 a\, height^3 + \beta_1 heigth^4 + \varepsilon$$

- women 자료에 4차 다항회귀모형 적합

```
> mod1.fit_w <- lm(weight~poly(height,degree=4,raw=TRUE),women)
> summary(mod1.fit_w)
```

```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          8.824e+03  4.550e+03   1.939   0.0812
poly(height, degree = 4, raw = TRUE)1 -5.552e+02  2.814e+02  -1.973   0.0768
poly(height, degree = 4, raw = TRUE)2  1.319e+01  6.515e+00   2.024   0.0705
poly(height, degree = 4, raw = TRUE)3 -1.389e-01  6.693e-02  -2.076   0.0646
poly(height, degree = 4, raw = TRUE)4  5.507e-04  2.574e-04   2.140   0.0581

(Intercept)                          .
poly(height, degree = 4, raw = TRUE)1 .
poly(height, degree = 4, raw = TRUE)2 .
poly(height, degree = 4, raw = TRUE)3 .
poly(height, degree = 4, raw = TRUE)4 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2244 on 10 degrees of freedom
Multiple R-squared:  0.9999,      Adjusted R-squared:  0.9998
F-statistic: 1.669e+04 on 4 and 10 DF,  p-value: < 2.2e-16
```

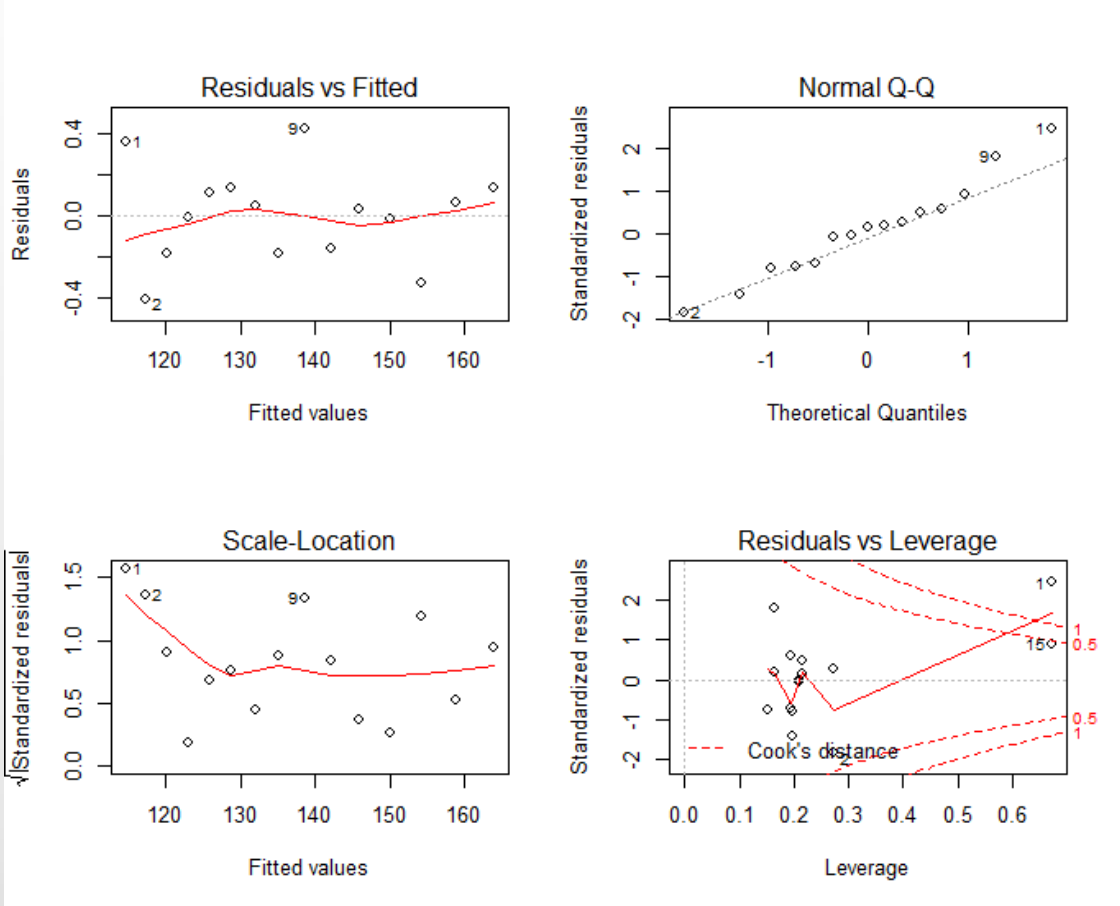
- 어떤 문제가 있는가?

- women 자료에 3차 다항회귀모형 적합

```
> mod2.fit_w <- lm(weight~poly(height,degree=3,raw=TRUE),women)
> summary(mod2.fit_w)

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -8.967e+02  2.946e+02  -3.044  0.01116
poly(height, degree = 3, raw = TRUE)1  4.641e+01  1.366e+01   3.399  0.00594
poly(height, degree = 3, raw = TRUE)2 -7.462e-01  2.105e-01  -3.544  0.00460
poly(height, degree = 3, raw = TRUE)3  4.253e-03  1.079e-03   3.940  0.00231

(Intercept)                         *
poly(height, degree = 3, raw = TRUE)1 **
poly(height, degree = 3, raw = TRUE)2 **
poly(height, degree = 3, raw = TRUE)3 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2583 on 11 degrees of freedom
Multiple R-squared:  0.9998,       Adjusted R-squared:  0.9997
F-statistic: 1.679e+04 on 3 and 11 DF,  p-value: < 2.2e-16
```

- 모든 변수 유의적

- women 자료의 3차 다항회귀모형의 진단



- 큰 문제가 없는 것으로 보임

- 예제: Ericksen

```
> Ericksen_1 <- select(Ericksen, -undercount, -conventional)
> set.seed(1234)
> x.id <- sample(1:nrow(Ericksen_1), size=6)
> df_1 <- Ericksen_1[x.id,]
> df_2 <- Ericksen_1[-x.id,]
```

```
> fit_1 <- lm(crime ~ . , data=df_2)
```

- 모형 fit_1을 대상으로 반응변수의 변환 필요 여부 확인

- 필요한 것으로 결정되면 반응변수 변환 후 적합, 가정 만족 여부 및 이상값 존재 여부 다시 확인

- 반응변수 변환 후 적합된 모형을 사용하여 df_1에 대한 예측 실시 및 예측 오차 산출