회귀분석 및 실습 HW7

서울대학교 통계학과 2017-11362 박건도

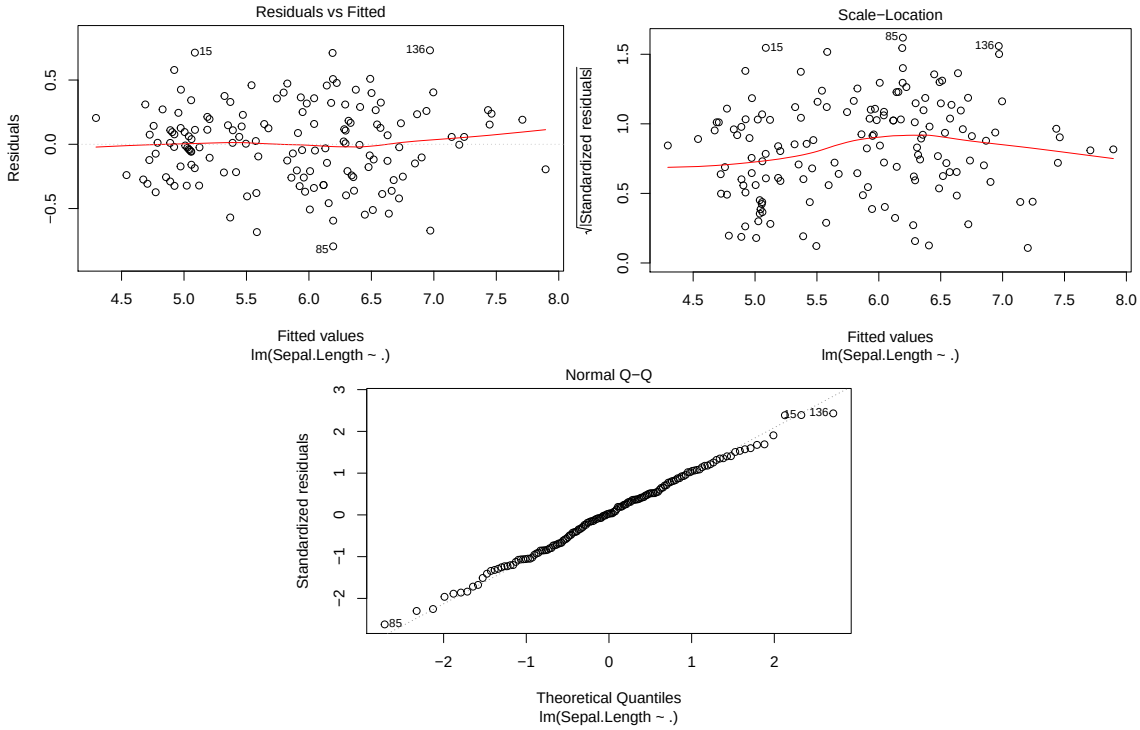
2021년 05월 10일

1. multiple linear regression

```
model <- lm(Sepal.Length ~ ., data = iris)
summary(model)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ ., data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79424 -0.21874  0.00899  0.20255  0.73103
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.17127    0.27979   7.760 1.43e-12 ***
## Sepal.Width        0.49589    0.08607   5.761 4.87e-08 ***
## Petal.Length       0.82924    0.06853  12.101  < 2e-16 ***
## Petal.Width       -0.31516    0.15120  -2.084  0.03889 *
## Speciesversicolor -0.72356    0.24017  -3.013  0.00306 **
## Speciesvirginica  -1.02350    0.33373  -3.067  0.00258 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3068 on 144 degrees of freedom
## Multiple R-squared:  0.8673, Adjusted R-squared:  0.8627
## F-statistic: 188.3 on 5 and 144 DF,  p-value: < 2.2e-16
```

```
plot(model, which = c(1, 3))
plot(model, 2)
```



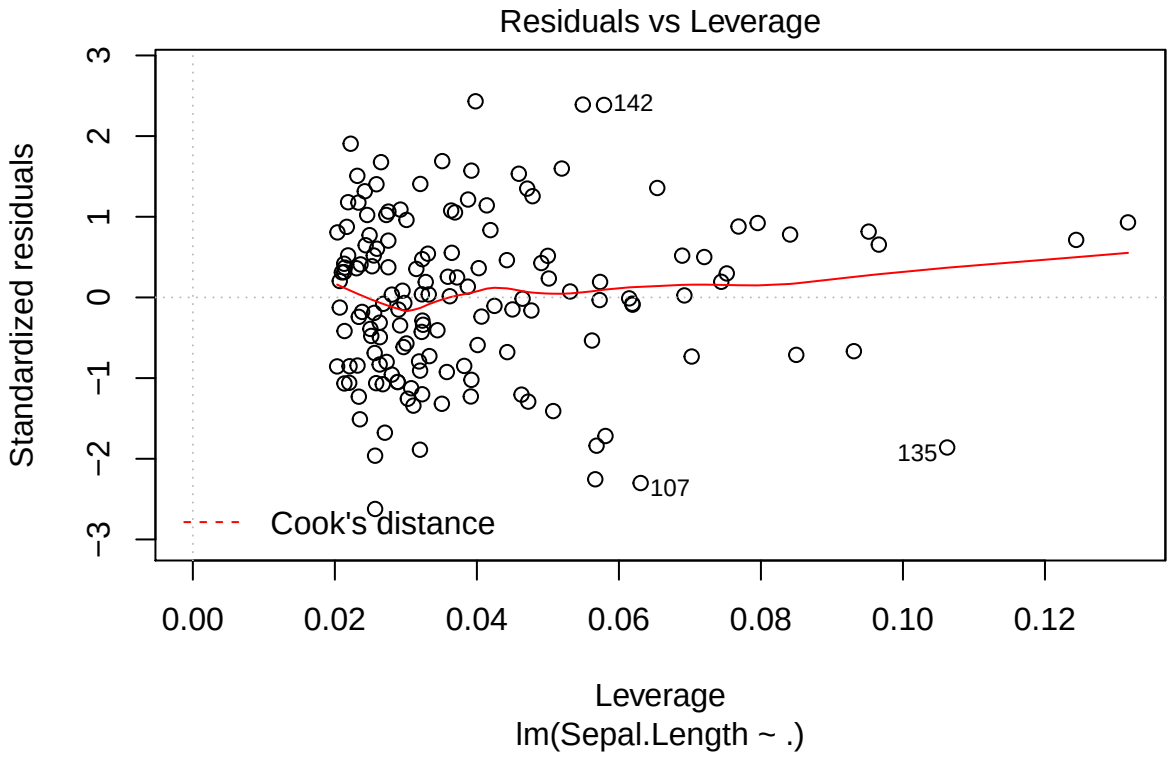
적합 결과는 위와 같다.

잔차분석을 하기 위해 1, 2번째 그래프를 보자. Fitted values와 Residuals, Standardized residuals 값을 보면 등분산성을 어느정도 만족하는 것으로 보인다. Normal Q-Q 그래프 또한 Standardized residual이 Theoretical Quantiles을 거의 만족함을 알 수 있다.

2. Outlier, high leverage point, influential point and multicollinearity

(1) Outlier

```
plot(model, 5)
```



먼저 outlier를 살펴보기 위해 Residuals vs Leverage 그래프를 보면, 107, 135, 142번 점이 outlier임을 알 수 있다. outlier들이 존재하지만, Residual이 -3을 벗어나는 점들이 없으므로 괜찮다고 할 수 있다.

(2) High leverage point

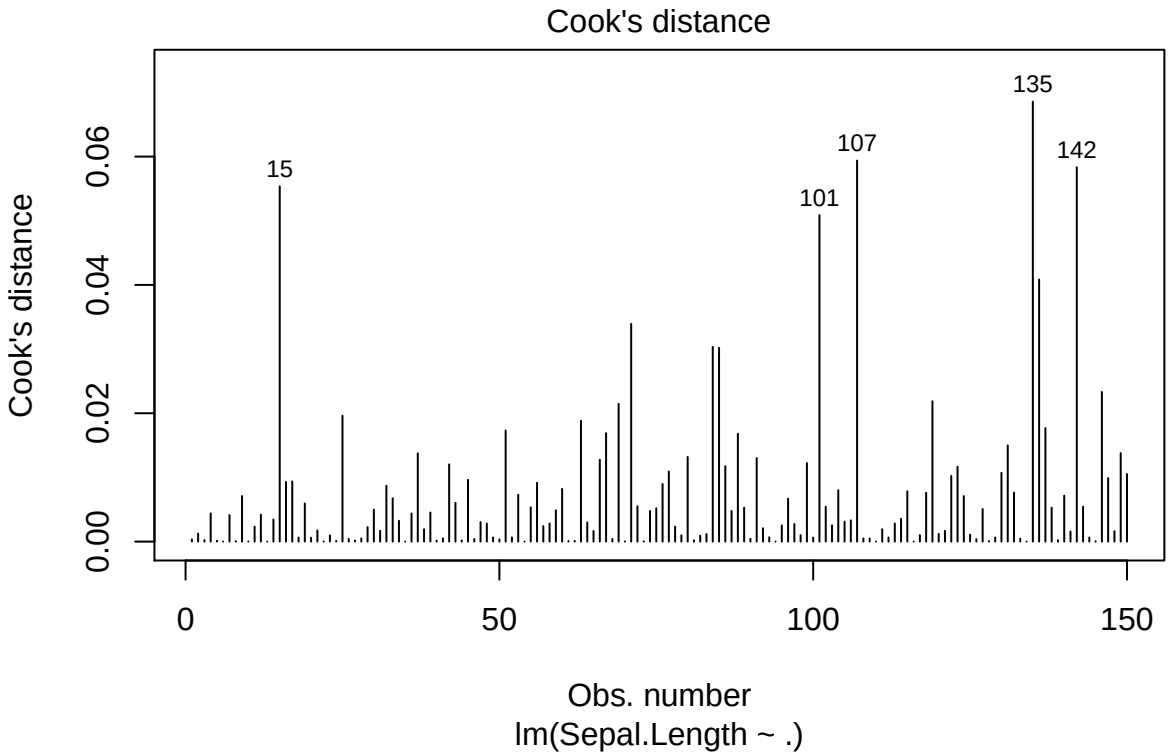
또한 Leverage 공식에 의해 $2(p+1)/n = 6/150 = 0.04$를 넘는 점들은 high leverage point라고 할 수 있는데, 다시 위의 그래프를 보면 0.04를 넘는 점들은 총 51개가 있다. 점들은 아래와 같다.

```
hats <- hatvalues(model)
names(hats[hats > 0.04])
```

```
##  [1] "15"  "16"  "17"  "33"  "34"  "42"  "44"  "57"  "58"  "61"  "63"  "65"
## [13] "68"  "69"  "71"  "73"  "78"  "80"  "82"  "84"  "86"  "94"  "99"  "101"
## [25] "106" "107" "108" "109" "110" "114" "115" "116" "118" "119" "120" "123"
## [37] "126" "127" "128" "130" "131" "132" "134" "135" "137" "139" "141" "142"
## [49] "145" "146" "149"
```

(3) Influential point

```
plot(model, 4, id.n = 5)
```



Cook's distance가 1을 넘어가면 influential point라고 할 수 있는데, 위 그래프에서는 1을 넘어가는 값이 없기 때문에 influential point가 없다고 할 수 있다.

(4) multicollinearity

```
car::vif(model)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## Sepal.Width    2.227466  1        1.492470
## Petal.Length 23.161648  1        4.812655
## Petal.Width  21.021401  1        4.584910
## Species      40.039177  2        2.515482
```

위의 결과에서 VIF는 GVIF^(1/(2*Df)) 인데, 모두 5 이하의 작은 값이므로 multicollinearity가 존재하지 않는다.

3. Ridge regression

Ridge regression 방법으로 구한 회귀계수는 아래와 같다.

```
x_var <- data.matrix(iris[, c("Sepal.Width", "Petal.Length", "Petal.Width", "Species")])
y_var <- iris[, "Sepal.Length"]

lambda_seq <- 10^seq(0, -4, by = -.1)
```

```r
ridge_cv <- cv.glmnet(x_var, y_var, alpha = 0, lambda = lambda_seq)
best_lambda <- ridge_cv$lambda.min
best_lambda
```

```
## [1] 0.001995262
```

```r
best_ridge <- glmnet(x_var, y_var, alpha = 0, lambda = best_lambda)
coef(best_ridge)
```

```
## 5 x 1 sparse Matrix of class "dgCMatrix"
##                      s0
## (Intercept)   2.1303170
## Sepal.Width   0.6155471
## Petal.Length  0.7106515
## Petal.Width  -0.3522732
## Species      -0.2085259
```