회귀분석 및 실습 HW7

서울대학교 통계학과 2017-11362 박건도

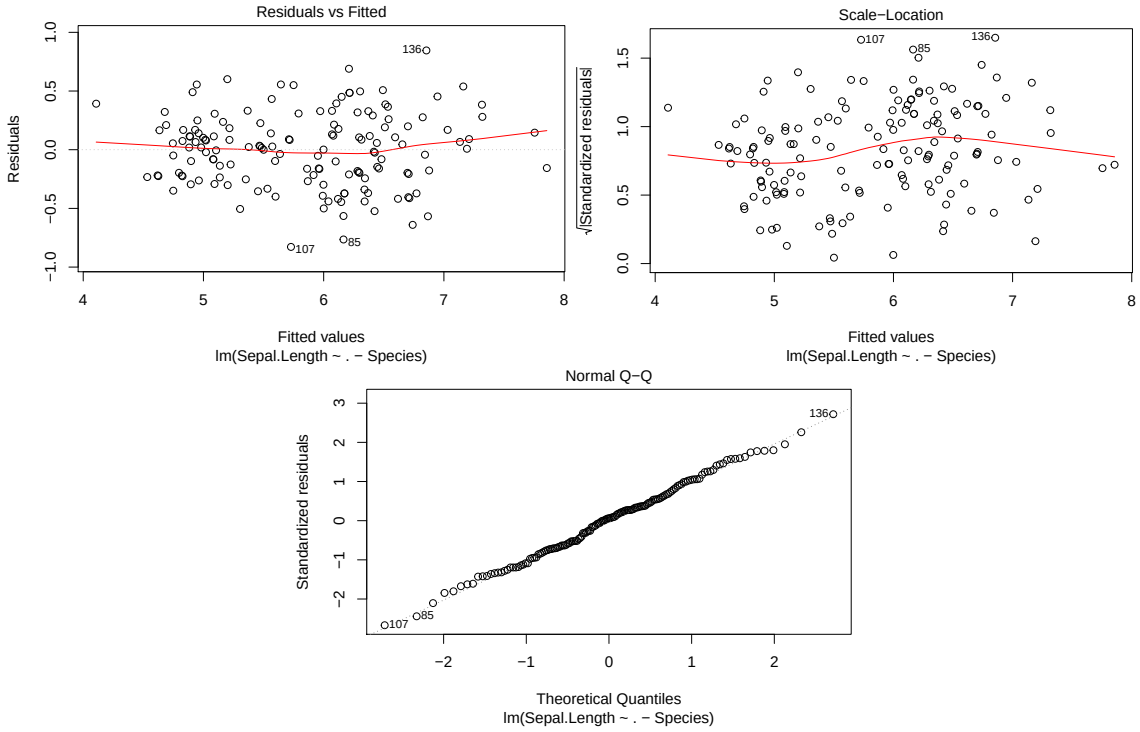
2021년 05월 11일

1. multiple linear regression

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

```
##
## Call:
## lm(formula = Sepal.Length ~ . - Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82816 -0.21989  0.01875  0.19709  0.84570
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.85600    0.25078   7.401 9.85e-12 ***
## Sepal.Width   0.65084    0.06665   9.765  < 2e-16 ***
## Petal.Length  0.70913    0.05672  12.502  < 2e-16 ***
## Petal.Width  -0.55648    0.12755  -4.363 2.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3145 on 146 degrees of freedom
## Multiple R-squared:  0.8586, Adjusted R-squared:  0.8557
## F-statistic: 295.5 on 3 and 146 DF,  p-value: < 2.2e-16
```

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



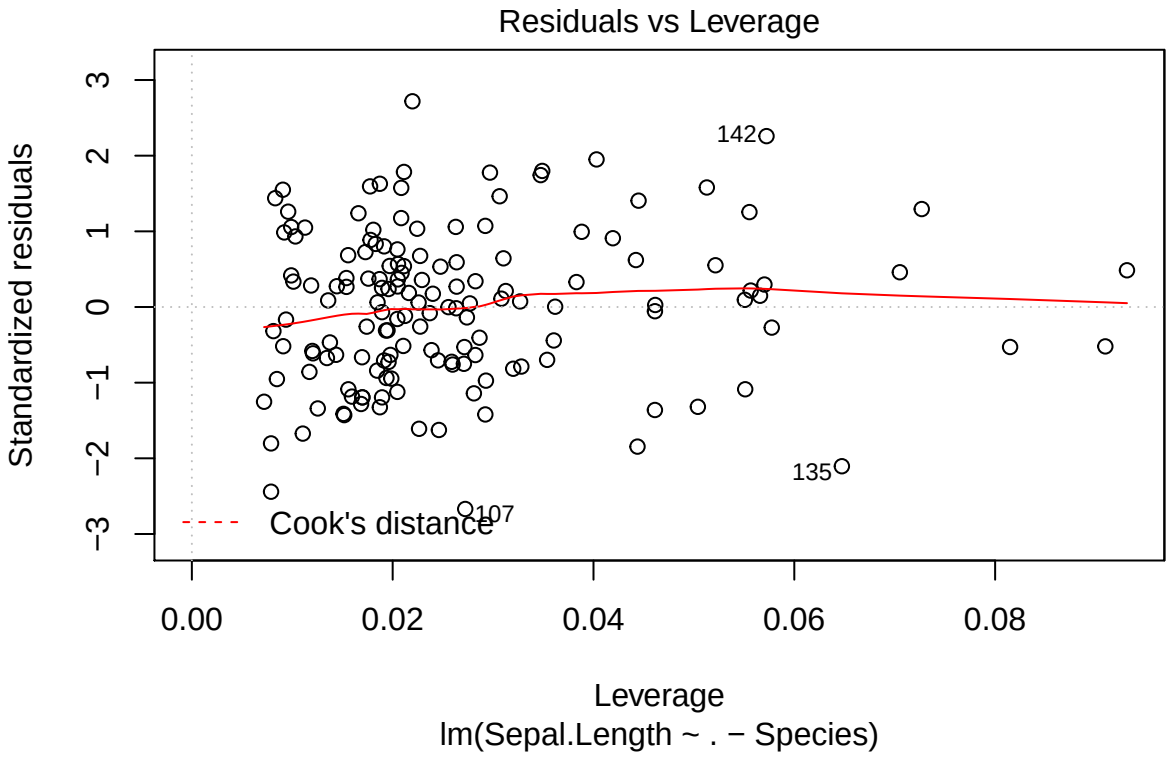
적합 결과는 위와 같다.

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

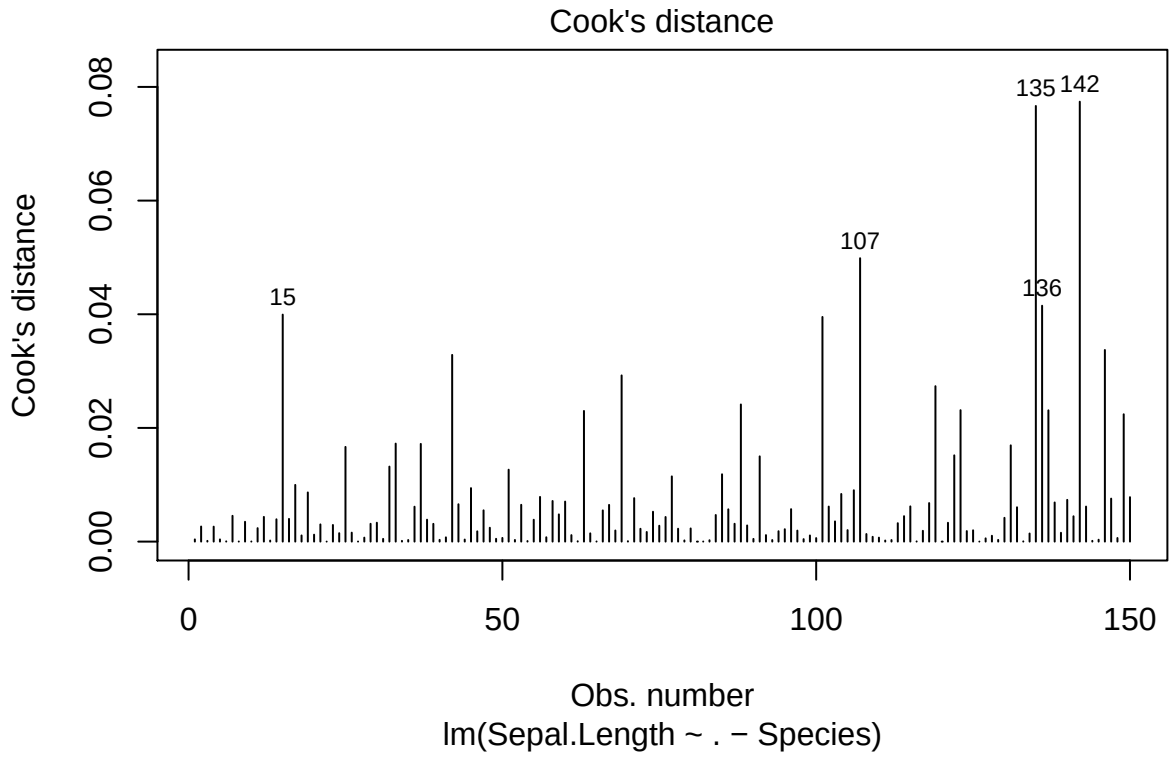
또한 Leverage 공식에 의해 $2(p+1)/n = 8/150 = 0.053$를 넘는 점들은 high leverage point라고 할 수 있는데, 다시 위의 그래프를 보면 0.053를 넘는 점들은 총 14개가 있다. 점들은 아래와 같다.

```
hats <- hatvalues(model)
names(hats[hats > 0.053])
```

```
##  [1] "16"  "33"  "34"  "42"  "61"  "108" "110" "115" "118" "123" "132" "135"
## [13] "142" "145"
```

(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)
```

```
##  Sepal.Width Petal.Length  Petal.Width
##     1.270815    15.097572    14.234335
```

위의 결과에서 Petal.Length와 Petal.Width가 10 이상의 큰 값이 나온다. 먼저 제일 큰 값인 Petal.Length를 제외하고 적합을 해보자.

```
model_1 <- lm(Sepal.Length ~ Sepal.Width + Petal.Width, data = iris)
car::vif(model_1)
```

```
## Sepal.Width Petal.Width
##    1.154799    1.154799
```

Petal.Length를 제외하고 Sepal.Length를 Sepal.Width와 Petal.Width로 적합하면 multicollinearity가 사라진 것을 볼 수 있다.

3. Ridge regression

Ridge regression 방법으로 구한 회귀계수는 아래와 같다.

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

lambda_seq <- 10^seq(0, -4, by = -.1)
ridge_cv <- cv.glmnet(x_var, y_var, alpha = 0, lambda = lambda_seq)
best_lambda <- ridge_cv$lambda.min
best_lambda
```

```
## [1] 0.001584893
```

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

```
## 4 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)   1.9201721
## Sepal.Width   0.6401970
## Petal.Length  0.6813400
## Petal.Width  -0.4957837
```