week4:선형회귀분석 실습

김현우,박주연,이주영,이지예,주진영,홍정아

2018년 4월 13일

encoding: UTF-8

```
#Sys.setlocale('LC_ALL','C')
data(cars)
tail(cars)
```

```
##    speed dist
## 45    23   54
## 46    24   70
## 47    24   92
## 48    24   93
## 49    24  120
## 50    25   85
```

```
# 목표 dist ??? β0 + β1 * speed.
model <- lm(dist ~ speed, data = cars)
# lm은 R에 내장된 함수로 linear regression을 불러들이는 함수이고, cars라는 data에서 y값으로는 dist를 x값으로는 speed를 넣으라는 명령어 입니다.
```

```
model
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
```

결론 dist = -17.579 + 3.932×speed.

이제 추가적으로 더 자세한 내용을 보기 위해서 summary라는 명령어와 plot이라는 명령어를 입력해 보겠습니다.

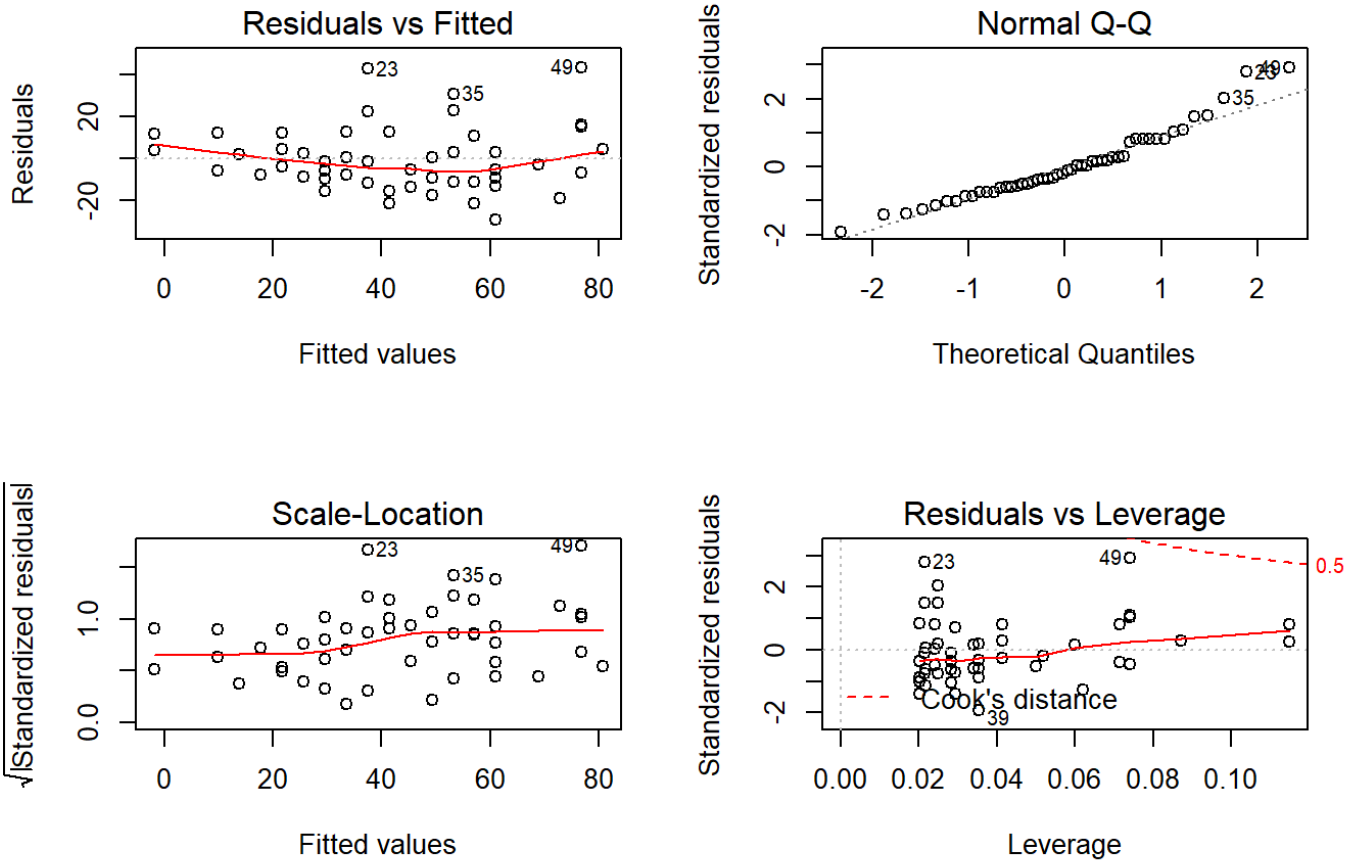
```
summary(model)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

`R-squared`는 설명의 정도를 알려주는 공식인데, 왼쪽의 R-squared는 row가 증가함에따라 값이 증가해서 Adjusted R-squared를 사용합니다. 0.64는 전체데이터의 64%를 설명한다는 의미. Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

마지막 p-value는 과연 이러한 회귀분석모델 자체가 유의미한지를 확인하는것으로 유의수준5%에서 의미가 있는것을 확인할 수 있습니다. F-statistic: 89.57 on 1 and 48 DF, `p-value` : 1.49e-12

```
par(mfrow = c(2, 2))
plot(model)
```



```
#independent errors assumption
#library(lmtest)
lmtest::dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.6762, p-value = 0.09522
## alternative hypothesis: true autocorrelation is greater than 0
```

```
#normality assumption
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.94509, p-value = 0.02152
```

```
#constant variance assumption
#library(car)
car::ncvTest(model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 4.650233    Df = 1     p = 0.03104933
```

```
# multicollinearity
# library(car)
# vif <- vif(lm(dist ~ speed, data = cars))
# 지금 내용은 단순선형회귀분석이여서 (독립변수가 한개여서) 확인을 못하지만, 다중선형회귀분석에서는 이를 통해 다중공선성이 10이하인것을 확인해야 합니다
```

위 과정을 한번에 해주는 패키지 `gvlma`

```
#library(gvlma)
assumptiontest <- gvlma::gvlma(model)
```

```
summary(assumptiontest)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
##ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
##USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
##Level of Significance =  0.05

##Call:
## gvlma::gvlma(x = model)

                    Value  p-value                   Decision
##Global Stat        15.801 0.003298 Assumptions NOT satisfied!
##Skewness            6.528 0.010621 Assumptions NOT satisfied!
##Kurtosis            1.661 0.197449    Assumptions acceptable.
##Link Function       2.329 0.126998    Assumptions acceptable.
##Heteroscedasticity  5.283 0.021530 Assumptions NOT satisfied!
```