Yonsei University, Department of Applied Statistics

Chapter 3 실습 (R)

YBIGTA 12기 신보현

January 25, 2019

```r
library(MASS)
library(ISLR)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

```r
# fitting all variables
lm.fit = lm(medv~. , data=Boston)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
```

```
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406,Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
names(lm.fit)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
library(car) # to compute VIF
vif(lm.fit)
```

```
##     crim       zn    indus     chas      nox       rm      age      dis
## 1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945
##      rad      tax  ptratio    black    lstat
## 7.484496 9.008554 1.799084 1.348521 2.941491
```

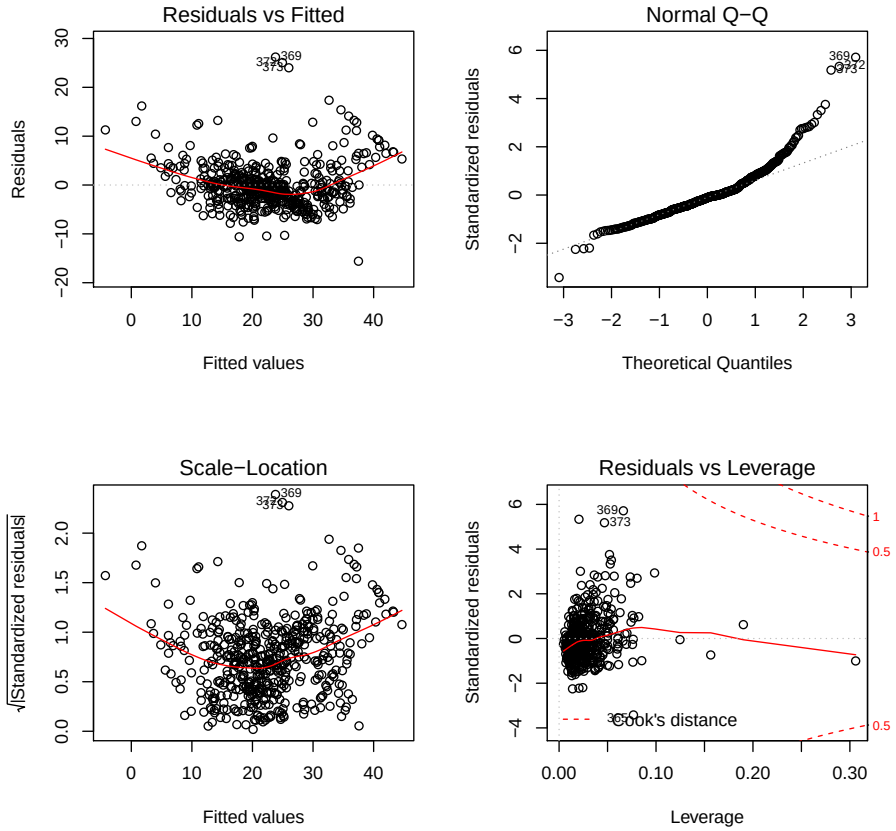
```
# regression using all variables but one
lm.fit2 = lm(medv~.-age, data=Boston)
## Interaction Terms
lm.fit.interaction = lm(medv~lstat*age, data=Boston)
summary(lm.fit.interaction) # p-value of interaction term is statistically significant
```

```
##
## Call:
## lm(formula = medv ~ lstat * age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.806  -4.045  -1.333   2.085  27.552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
## lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
## age         -0.0007209  0.0198792  -0.036   0.9711
## lstat:age    0.0041560  0.0018518   2.244   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.149 on 502 degrees of freedom
## Multiple R-squared:  0.5557,Adjusted R-squared:  0.5531
## F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
## Non-linear Transformations of the predictors
lm.fit.nonlinear = lm(medv~lstat + I(lstat^2),data=Boston)
summary(lm.fit.nonlinear)
```

```
##
## Call:
## lm(formula = medv ~ lstat + I(lstat^2), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2834  -3.8313  -0.5295   2.3095  25.4148
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.862007   0.872084   49.15   <2e-16 ***
## lstat       -2.332821   0.123803  -18.84   <2e-16 ***
```

```
## I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.524 on 503 degrees of freedom
## Multiple R-squared:  0.6407,Adjusted R-squared:  0.6393
## F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16


## important to check the assumptions
# 1) check residual plots to check linearity of variables
par(mfrow=c(2,2))
plot(lm.fit)
```



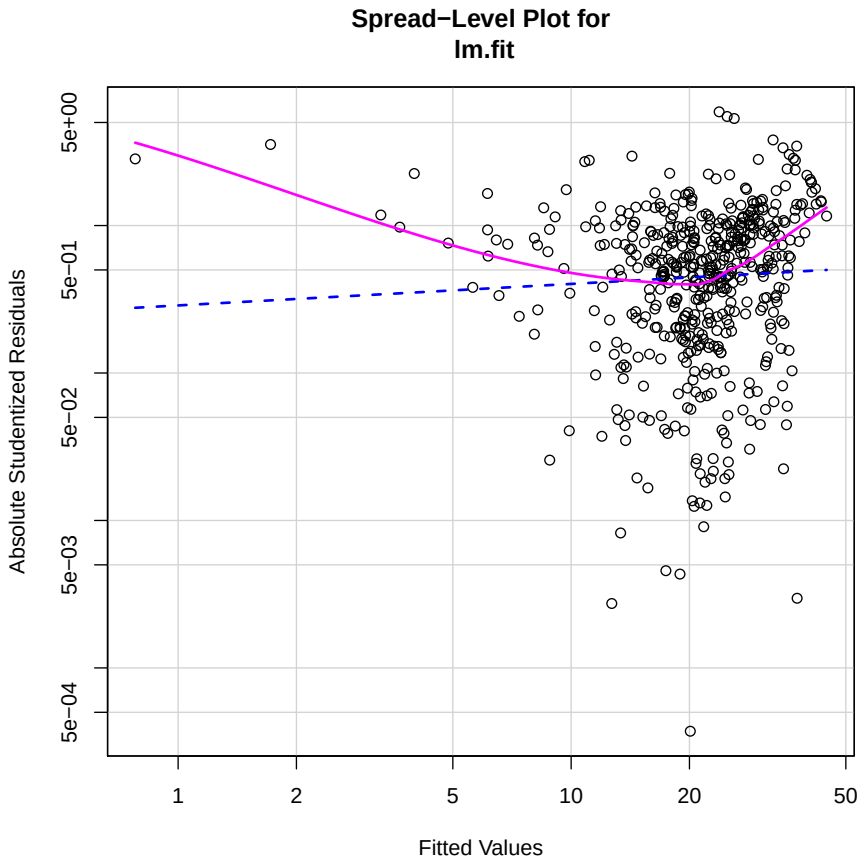

```
# 2) check correlation(independence) of error terms
durbinWatsonTest(lm.fit)


##  lag Autocorrelation D-W Statistic p-value
```

```
##    1       0.4542626      1.078375       0
##  Alternative hypothesis: rho != 0
```

```
# 3) check non-constant variance of error terms (Homoscedasticity)
library(car)
ncvTest(lm.fit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 15.24407    Df = 1     p = 9.4473e-05
```

```
par(mfrow=c(1,1))
spreadLevelPlot(lm.fit)
```




```
##
## Suggested power transformation:  0.8541632
```

```r
# Brown-Forsythe Test
library(lawstat)
length(lm.fit$residuals)
```

```
## [1] 506
```

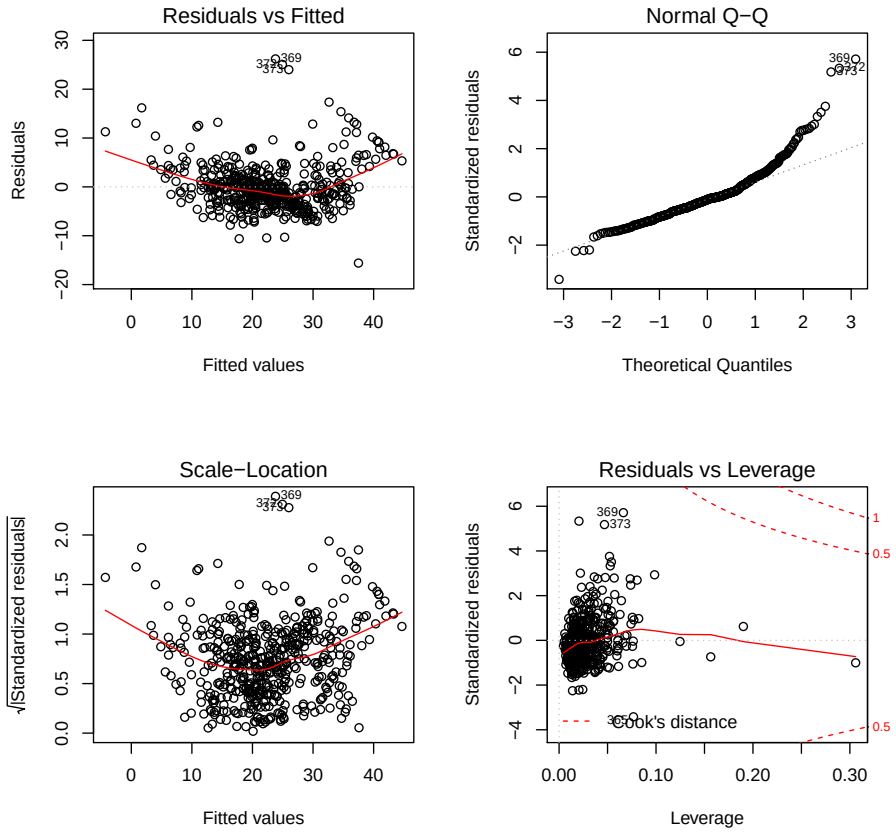
```r
sample1 = lm.fit$residuals[1:253]
sample2 = lm.fit$residuals[254:506]
group = as.factor(c(rep(1,253),rep(2,253)))
levene.test(lm.fit$residuals,group,location="median")
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  lm.fit$residuals
## Test Statistic = 4.6305, p-value = 0.03188
```

```r
# Breusch-Pagan
library(lmtest)
bptest(lm.fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm.fit
## BP = 65.122, df = 13, p-value = 6.265e-09
```

```r
# 4) Normality
par(mfrow=c(2,2))
plot(lm.fit)
```




```
# see Normal Q-Q plot
# qqPlot() from car package
qqPlot(lm.fit)
```

```
## [1] 369 372
```

```
# shapiro.test
shapiro.test(lm.fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm.fit$residuals
## W = 0.90138, p-value < 2.2e-16
```

```
# 5) outliers / influential data
# https://cran.r-project.org/web/packages/olsrr/vignettes/influence_measures.html
```

```
library(olsrr)
# cook's distance
ols_plot_cooksd_bar(lm.fit)
# deleted residual
ols_plot_resid_stud_fit(lm.fit)
# high leverage
ols_plot_resid_pot(lm.fit)
# 6) multicollinearity
# vif function from car packages
vif(lm.fit)
```

```
##     crim       zn    indus     chas      nox       rm      age      dis
## 1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945
##      rad      tax  ptratio    black    lstat
## 7.484496 9.008554 1.799084 1.348521 2.941491
```