CS 498: Homework 06

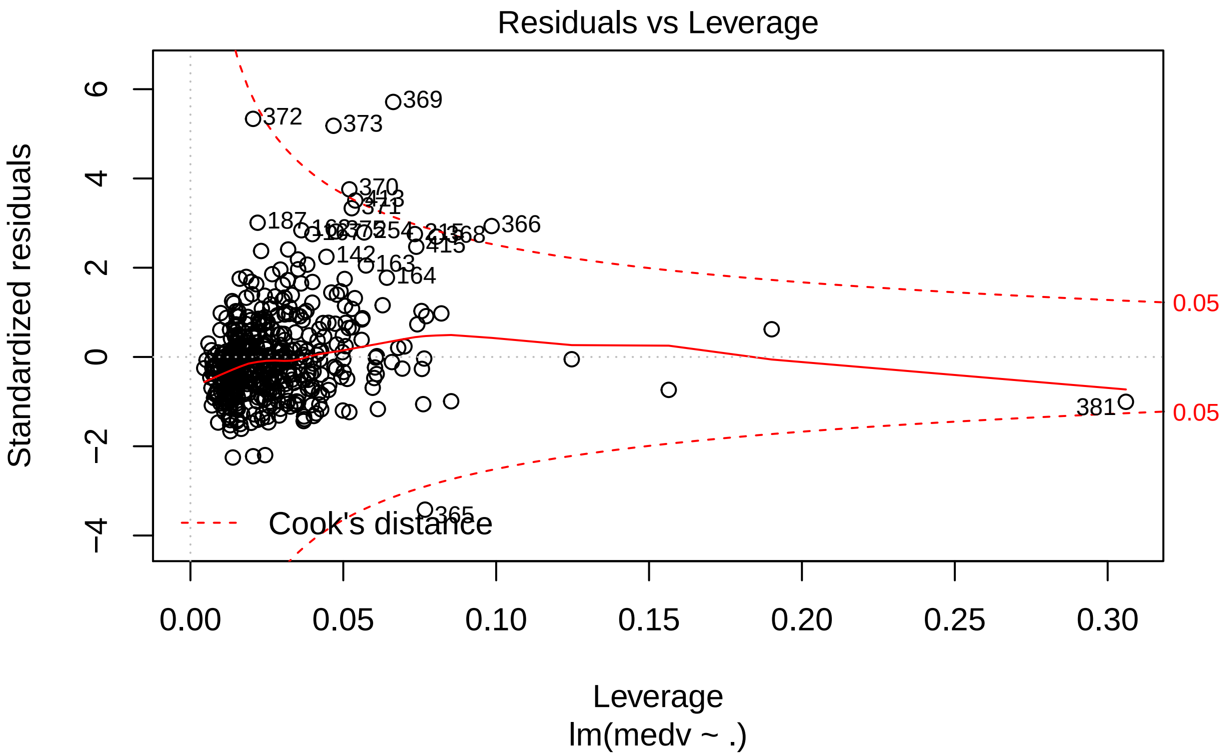
Spring 2019, Guangya Wan, Sizhi Tan

Contents

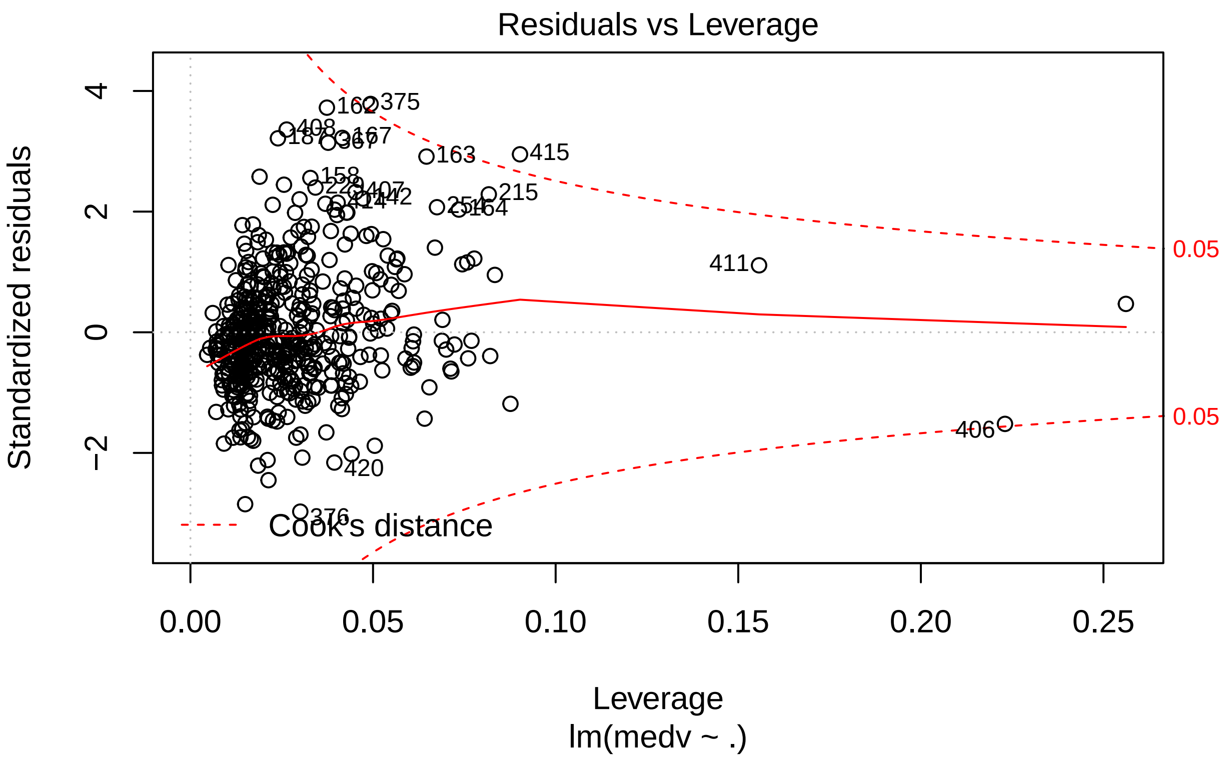
```r
library('MASS')
data_set = Boston
model = lm(medv ~ .,data = data_set) # regression here
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ ., data = data_set)
##
## 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
```

code for building regression model and the summary of model



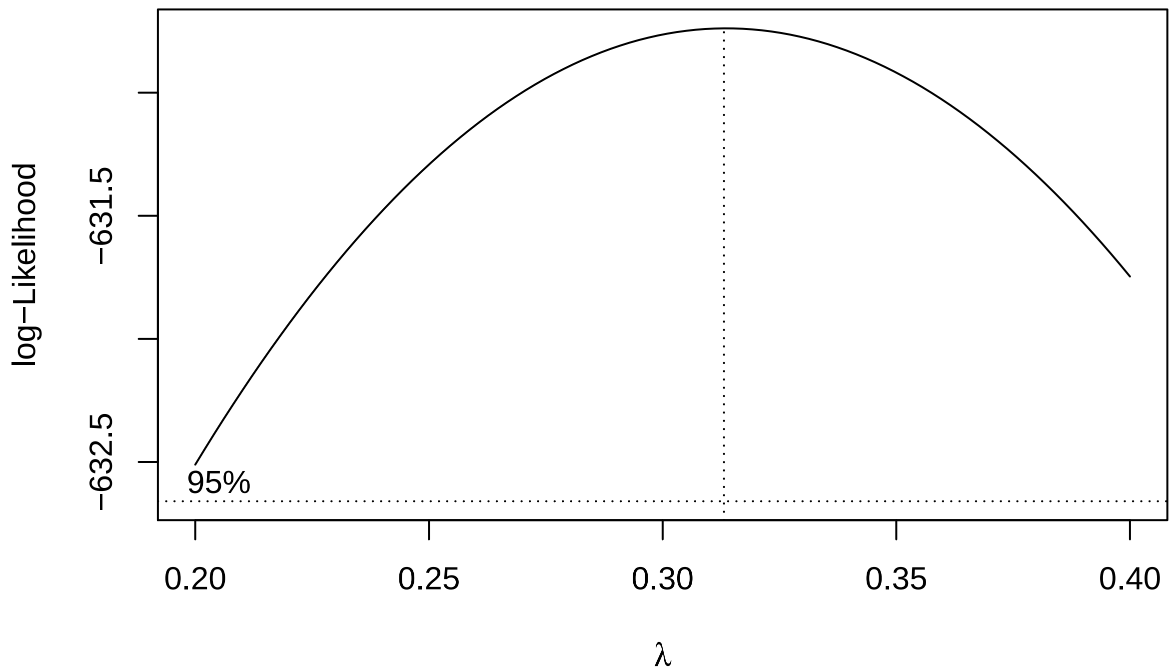
Based on this plot, I identified that there are 10 points(365,381,369,373,372,370,366,371,413,368) here that are outliers.Here are my reasonings: Point 381 have an very large leverage compared to the rest of points. Points 365, 369, 373,370, 366 are outside of my curve cut-off cook's distance value curve which is 0.05. Points 370,413,371,368 are outside but very close to my cook's distance cut-off, and they have a large standardlized residuals(great than 3), and their leverage are also larger than the leverage of other majority of data, which is about 0.025.



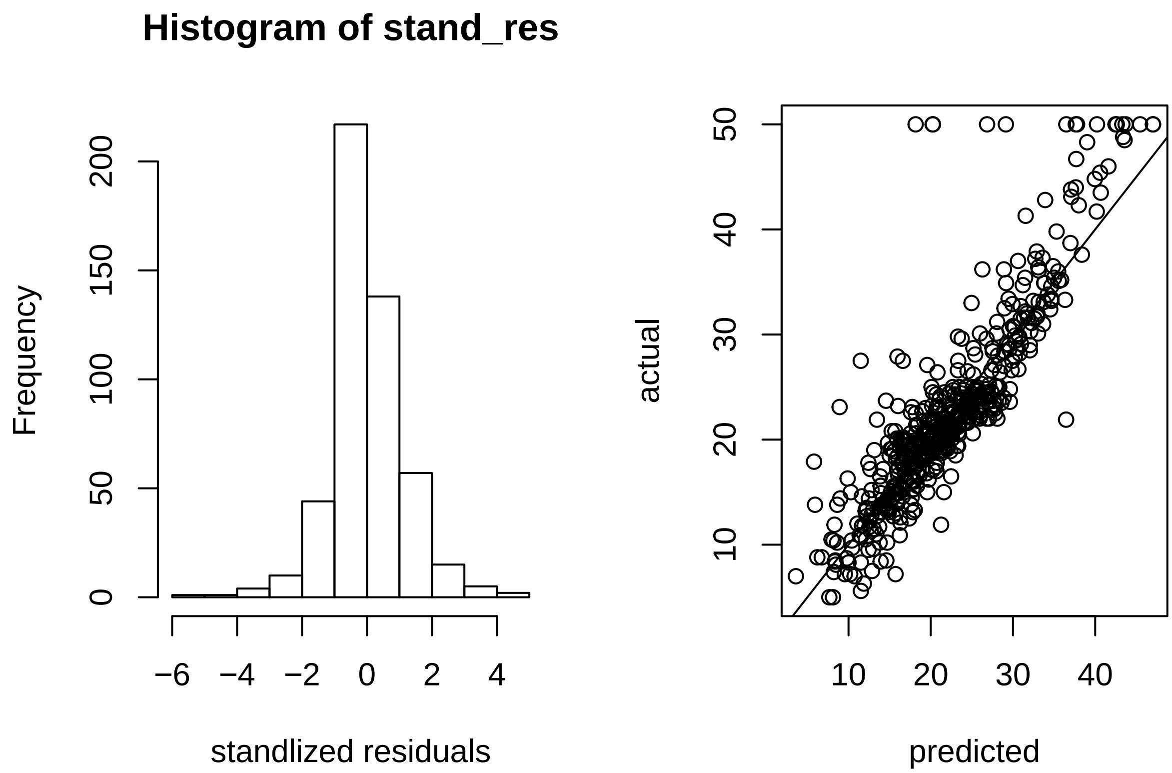
diagnostic plot after dropping selected outliers

```r
dropped_index = c(365,369,372,373,381,413,371,370,366,368)
dropped_data = data_set[-dropped_index,]
dropped_model = lm(medv ~ .,data = dropped_data)
plot(dropped_model, which=5, cook.levels=cutoff,id.n = 20)
```

codes for generating the plot



According to the plot, best value of lambda is 0.315 as it maximizes the log likelihood



Left is histogram of residuals after transformation and right is actual vs predicetd plot plus an x=y line

```r
boxcox(medv ~ ., data = dropped_data,lambda = seq(0.2, 0.4, length = 10)) # problem 3
par(mfrow=c(1,2))# the rest are for problem 4
dropped_data[,'medv'] = (dropped_data[,'medv'] ** 0.315 - 1) / 0.315
dropped_model = lm(medv ~ .,data = dropped_data)
x = as.matrix(data_set[,1:13])
y = predict(dropped_model,data_set[,1:13])
M = x %*% (solve(t(x) %*% x)) %*% t(x)
stand_res = rep(0,nrow(M))
cons =  (t(resid(dropped_model)) %*%  resid(dropped_model)) / nrow(M)
for (i in 1:nrow(M)){
    stand_res[i] = resid(dropped_model)[i] / (cons * 1-M[i,i])**0.5
}
hist(stand_res,xlab = "standlized residuals")
y = (y * 0.315 + 1)**(1/0.315)
true_price = data_set[,'medv']
plot(y,true_price,xlab="predicted",ylab="actual")
abline(a=0,b=1)
```

Code for subproblem 3 and 4