1 Code Regression and Resulting Model

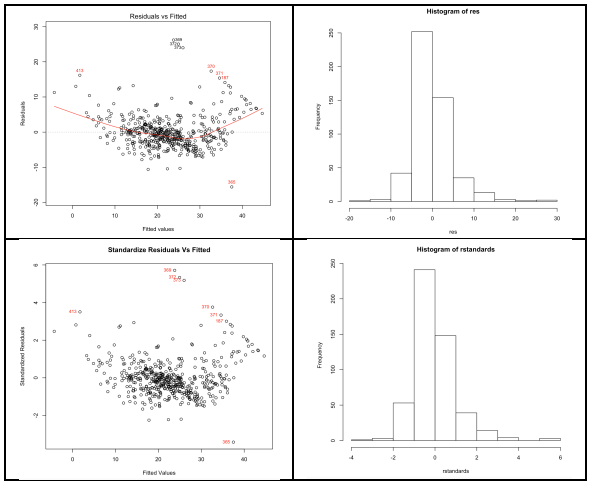
```
#Load the data set
column_name <-
c("crim","zn","indus","chas","nox","rm","age","dis","rad","tax","ptra
tio","b","lstat","medv")
housing <- read.table("HW6/data/housing.data",col.names=column_name)

#Fit the Linear Data Model
fit <- lm
(medv~crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+b+lstat,data=
housing)
plot(fit)
summary(fit)
```

```
> summary(fit)

Call:
lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age +
    dis + rad + tax + ptratio + b + lstat, data = housing)

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 ***
b            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
```

2 Diagnostic Plot & Outlier

Residuals

Thresholds:
Standard Residuals (-3 and 3)
Residuals (12 and 14)

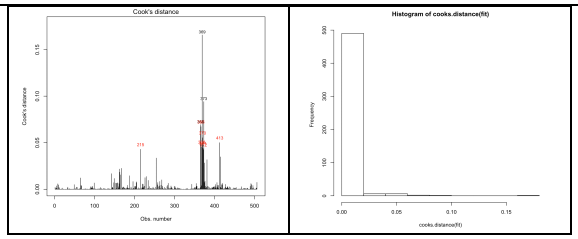


Outlier Index:[187,365,369, 370,371,372,373,413]

Reason: Clearly the residuals plot doesn't seems to be linear and there is a clear pattern of non-linearity (red line). For the initial "fitted values" (before 10) and later (after 30), the "true value" is far above the prediction line, which cause the residuals to high (**standard residuals** = (true value – prediction) > 3), except for the one point (index = 365), where the "true value" is far below the prediction line and cause the residuals to be high on negative (-ve) side(below < -3). Also the prediction between 20 & 30, there are around 3 point (index = 369, 372 and 373) which is far above the prediction line and cause high residuals (above >3). If we look into the histogram, it also observed that most of the fitted data lies within the standard residuals between -3 & 3. So, the prediction between these predicted values (<10, > 30 and between 20 & 30) violates linearity, would not follow the regression line, which would cause high residuals. Hence these points mentioned in the index (red point) needs to be reviewed as an outlier.

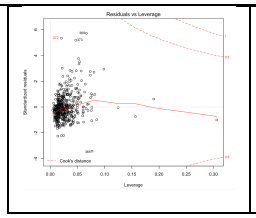
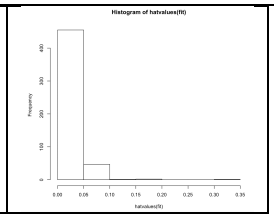
Cooks Distance

Thresholds:
> 0.04

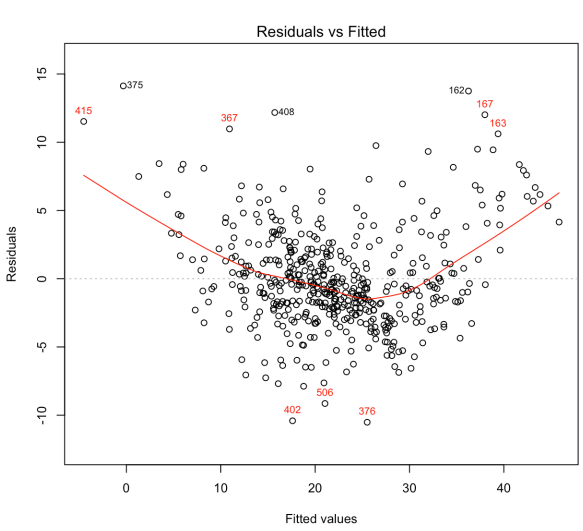


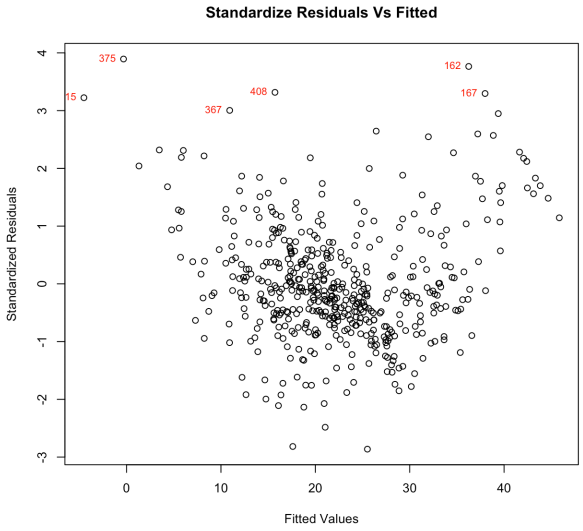
Outlier Index:[215,365,366,368,369, 370,371,372,373,413]

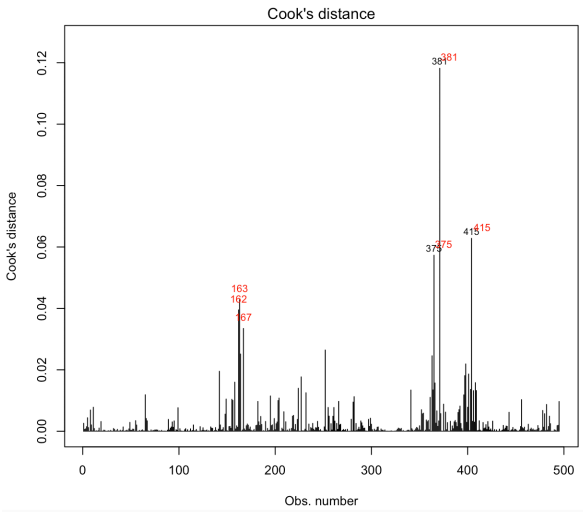
Reason: The cooks distance is the sum of all changes in the regression model when the specific observation is removed. By having a threshold(> 0.04) to verify the observation having high cook's distance, It's observed that the mentioned observation (in red) have high cooks distance, which means, have high influence/impact to the regression and observed as a potential outlier. Also from the histogram, it sees that most of the data point lies within 0.04 of cooks distance, which makes these points (red) are problem point and needs to be reviewed as an outlier.

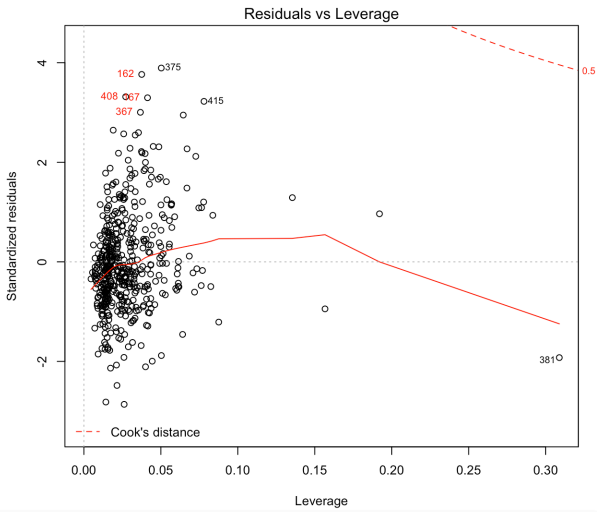
Leverage **Thresholds:** Standard Residuals (-3 and 3)	 	Outlier Index:[365,369,372,373] **Reason:** Though the leverage isn't high for any of the data points (also visible from the histogram, most of the "hat values" are under 0.10), the standardize residuals seems high (over 3 standard deviation from the mean), where the index are the same identified from the standardized residual plots from the 1st row. Also There are no observation which shows gigantic cook's distance, nor any of the observation goes beyond the 0.5 cook's distance.

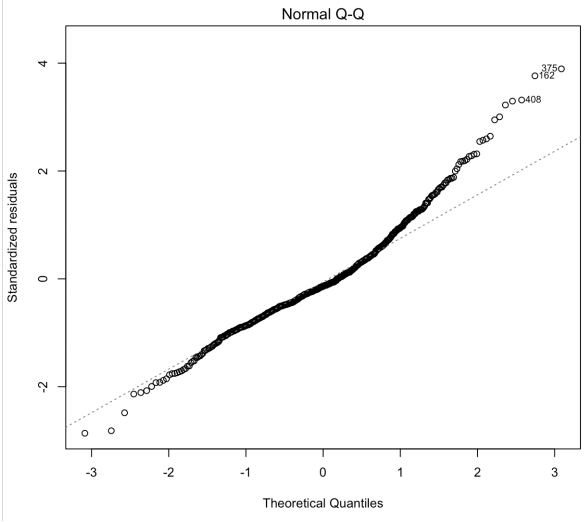
3 New Diagnostic Plot

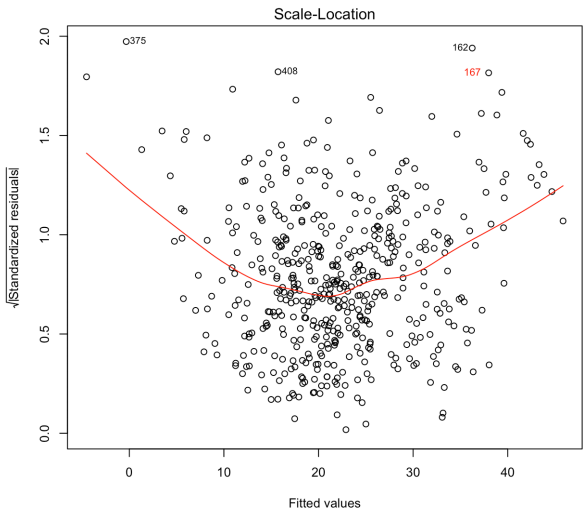










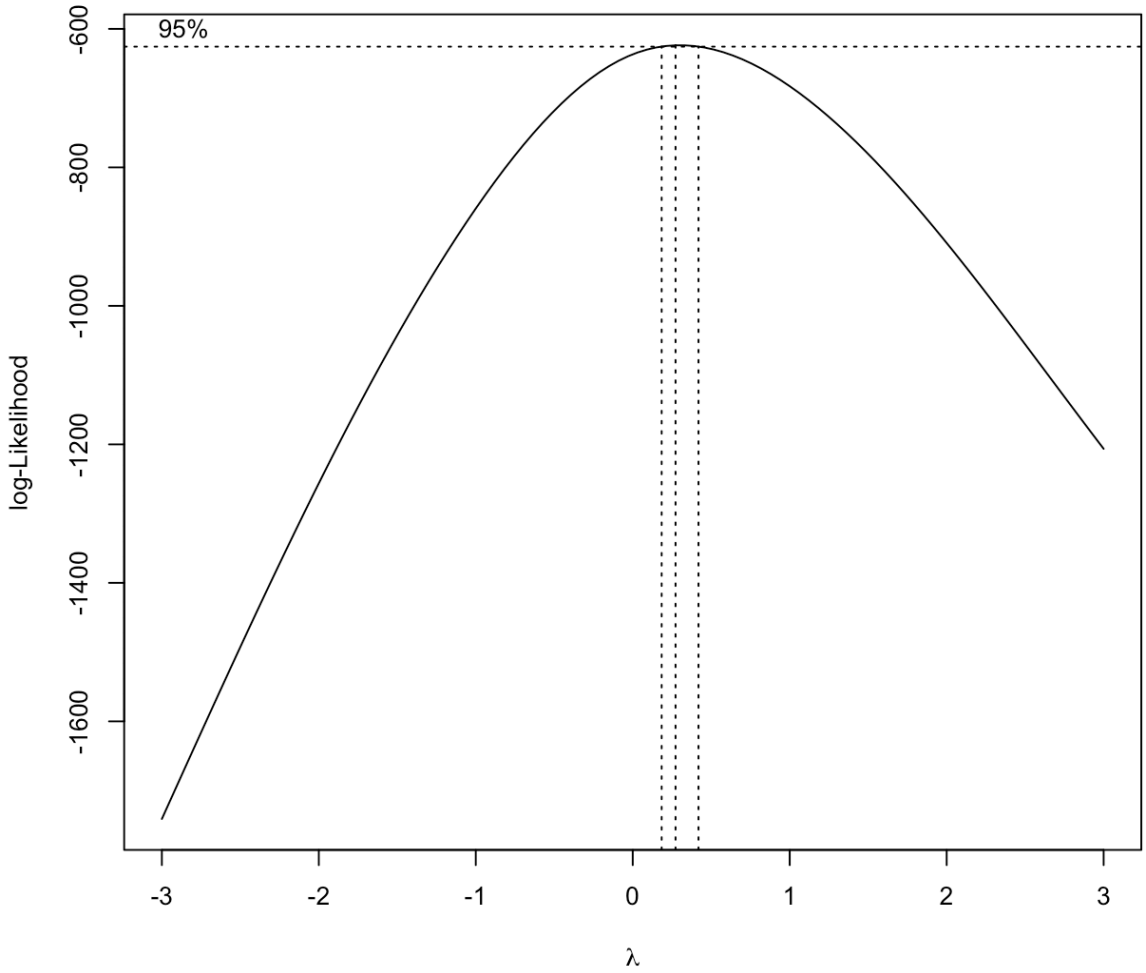


4 Code – SubProblem 2

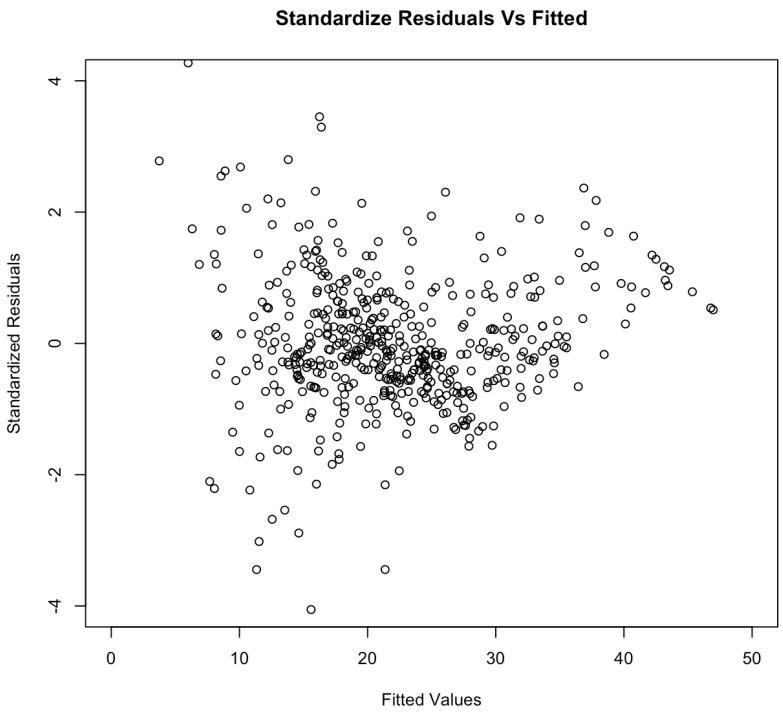
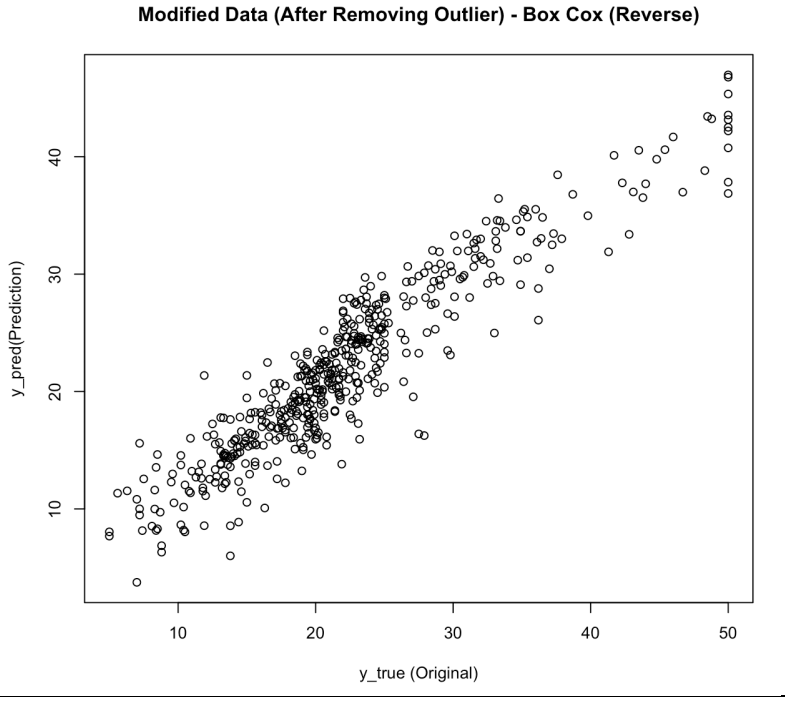
```
#############################################################################
#########################
#
# Residuals, Cooks Distance, Leverage ########  Before Outlier
#
#############################################################################
#########################
#-----------------------------------------Diagnostic Plot (Plot 1,3) ####
Residuals
par(mfrow=c(2,2))
fit <- lm
(medv~crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+b+lstat,data=housing)
res <- fit$residuals
plot(fit,which=1)
text(predict(fit),res,ifelse( ((rownames(housing)==369 | rownames(housing)==372
| rownames(housing)==373)
                                              | (!(res < -12 | res > 14))
),"",rownames(housing) ),
     cex= 0.8,pos=3,col='red')
hist(res)

rstandards = rstandard(fit)
plot(predict(fit),rstandards,xlab="Fitted Values",ylab="Standardized
Residuals",main="Standardize Residuals Vs Fitted")
text(predict(fit),rstandard(fit),ifelse((!(rstandards < -3 | rstandards >
3)),"",rownames(housing) ),
     cex= 0.8,pos=2,col='red')
hist(rstandards)

#-----------------------------------------Diagnostic Plot (Plot 4,5) ####
Cooks Distaince, Leverage
par(mfrow=c(2,2))
plot(fit,which=4)
text(rownames(housing),cooks.distance(fit),ifelse( ((rownames(housing)==369 |
rownames(housing)==373 | rownames(housing)==365)
                                                                    |
(cooks.distance(fit) < 0.04) ),"",rownames(housing) ),
     cex= 0.8,pos=3,col='red')
hist(cooks.distance(fit))
plot(fit,which=5)
text(hatvalues(fit),(rstandard(fit)),ifelse( ((rownames(housing)==369 |
rownames(housing)==373 | rownames(housing)==375)
                                                        | (!(rstandard(fit) <
-3 | rstandard(fit) > 3)) ),"",rownames(housing) ),
     cex= 0.8,pos=2,col='red')
hist(hatvalues(fit))


plot(fit,which=2)
```

5 Box Cox Transformation

Best Value of Lambda = 0.2727273



6 Plot Standardize Residuals Vs Fitted Price

Plot - Standardize Residuals Vs Fitted Values	
Plot – True Price against Predicted House Price (After Box Cox Transformation)	

7 Code – Sub Problem3 & 4

```
#----------------------------------------Load the data set
column_name <-
c("crim","zn","indus","chas","nox","rm","age","dis","rad","tax","ptra
tio","b","lstat","medv")
housing <- read.table("HW6/data/housing.data",col.names=column_name)
#----------------------------------------Fit the Linear Data Model
fit <- lm
(medv~crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+b+lstat,data=
housing)
plot(fit)
summary(fit)
#####################################################################
###################################
#
# Residuals, Cooks Distance, Leverage ########  Before Outlier
#
#####################################################################
###################################
#----------------------------------------Diagnostic Plot (Plot
1,3) #### Residuals
par(mfrow=c(2,2))
fit <- lm
(medv~crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+b+lstat,data=
housing)
res <- fit$residuals
plot(fit,which=1)
text(predict(fit),res,ifelse( ((rownames(housing)==369 |
rownames(housing)==372 | rownames(housing)==373)
                              | (!(res < -12 | res > 14))
),"",rownames(housing) ),
     cex= 0.8,pos=3,col='red')
hist(res)

rstandards = rstandard(fit)
plot(predict(fit),rstandards,xlab="Fitted Values",ylab="Standardized
Residuals",main="Standardize Residuals Vs Fitted")
text(predict(fit),rstandard(fit),ifelse((!(rstandards < -3 |
rstandards > 3)),"",rownames(housing) ),
     cex= 0.8,pos=2,col='red')
hist(rstandards)

#----------------------------------------Diagnostic Plot (Plot
4,5) #### Cooks Distaince, Leverage
par(mfrow=c(2,2))
plot(fit,which=4)
text(rownames(housing),cooks.distance(fit),ifelse(
((rownames(housing)==369 | rownames(housing)==373 |
rownames(housing)==365)
                                                    |
(cooks.distance(fit) < 0.04) ),"",rownames(housing) ),
     cex= 0.8,pos=3,col='red')
hist(cooks.distance(fit))
plot(fit,which=5)
text(hatvalues(fit),(rstandard(fit)),ifelse( ((rownames(housing)==369
| rownames(housing)==373 | rownames(housing)==365)
                                                | (!(rstandard(fit) < -
3 | rstandard(fit) > 3)) ),"",rownames(housing) ),
     cex= 0.8,pos=2,col='red')
```

```
hist(hatvalues(fit))


plot(fit,which=2)


#########################################################################
###################################
#
# Residuals, Cooks Distance, Leverage ########  After Outlier being
removed
#
#########################################################################
###################################
#----------------------------------------Diagnostic Plot (Plot
1,3) #### Residuals
point_exclude <- c(187,215,365,366,368,369,370,371,372,373,413)
housing_remove <- housing[-(point_exclude),]
par(mfrow=c(2,2))
#rownames(housing_remove) = 1:nrow(housing_remove) #reset the
rownames
fit_after <- lm
(medv~crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+b+lstat,data=
housing_remove)
res_after <- fit_after$residuals
plot(fit_after,which=1)
text(predict(fit_after),res_after,ifelse(
((rownames(housing_remove)==375 | rownames(housing_remove)==162 |
rownames(housing_remove)==408)
                                          | (!(res_after < -9 |
res_after > 10)) ),"",rownames(housing_remove) ),
     cex= 0.8,pos=3,col='red')
#hist(res_after)

rstandards_after = rstandard(fit_after)
plot(predict(fit_after),rstandards_after,xlab="Fitted
Values",ylab="Standardized Residuals",main="Standardize Residuals Vs
Fitted")
text(predict(fit_after),rstandard(fit_after),ifelse((!(rstandards_aft
er < -3 | rstandards_after > 3)),"",rownames(housing_remove) ),
     cex= 0.8,pos=2,col='red')
#hist(rstandards_after)

#----------------------------------------Diagnostic Plot (Plot
4,5) #### Cooks Distaince, Leverage
#par(mfrow=c(2,2))
plot(fit_after,which=4)
text(rownames(housing_remove),cooks.distance(fit_after),ifelse(
((rownames(housing_remove)==366 | rownames(housing_remove)==372 |
rownames(housing_remove)==405)
                                                                  |
(cooks.distance(fit_after) < 0.03) ),"",rownames(housing_remove) ),
     cex= 0.8,pos=3,col='red')
#hist(cooks.distance(fit_after))
plot(fit_after,which=5)
text(hatvalues(fit_after),(rstandard(fit_after)),ifelse(
((rownames(housing_remove)==375 | rownames(housing_remove)==415)
                                                        |
(!(rstandard(fit_after) < -3 | rstandard(fit_after) > 3))
),"",rownames(housing_remove) ),
     cex= 0.8,pos=2,col='red')
```

```
#hist(hatvalues(fit_after))


par(mfrow=c(2,2))
plot(fit_after,which=2)
plot(fit_after,which=3)
text(predict(fit_after),sqrt(rstandard(fit_after)),ifelse(
((rownames(housing_remove)==375 | rownames(housing_remove)==408 |
rownames(housing_remove)==162)
                                                                  |
(!(sqrt(rstandard(fit_after)) < -1.8 | sqrt(rstandard(fit_after)) >
1.8)) ),"",rownames(housing_remove) ),
     cex= 0.8,pos=2,col='red')

#####################################################################
###################################
#
# Box Cox Transformation (After Removing Outlier)
#
#####################################################################
###################################
#library(MASS)
bc = boxcox(fit_after,lambda = seq(-3,3))
best_lam=bc$x[which((bc$y == max(bc$y)))]
fit_modified_after_boxcox <- lm((((medv^best_lam)-
1)/best_lam)~crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+b+lsta
t,data=housing_remove)
par(mfrow=c(2,2))
plot(fit_modified_after_boxcox)

#####################################################################
###################################
#
# Plotting the Data (Original Vs Precited)
#
#####################################################################
###################################
par(mfrow=c(2,2))
xlim=c(0,50)
ylim=c(-4,4)
plot
((1+(predict(fit_modified_after_boxcox))*best_lam)^(1/best_lam),rstan
dard(fit_modified_after_boxcox),xlab="Fitted
Values",ylab="Standardized Residuals",main="Standardize Residuals Vs
Fitted",xlim=xlim,ylim=ylim)
y_pred <-
(1+(predict(fit_modified_after_boxcox)*best_lam))^(1/best_lam)
plot(housing_remove$medv,y_pred,xlab = "y_true
(Original)",ylab="y_pred(Prediction)",main = "Modified Data (After
Removing Outlier) - Box Cox (Reverse)")
```