HW6 : Outlier Analysis

Apoorva Srinivasa, Julia Tcholakova

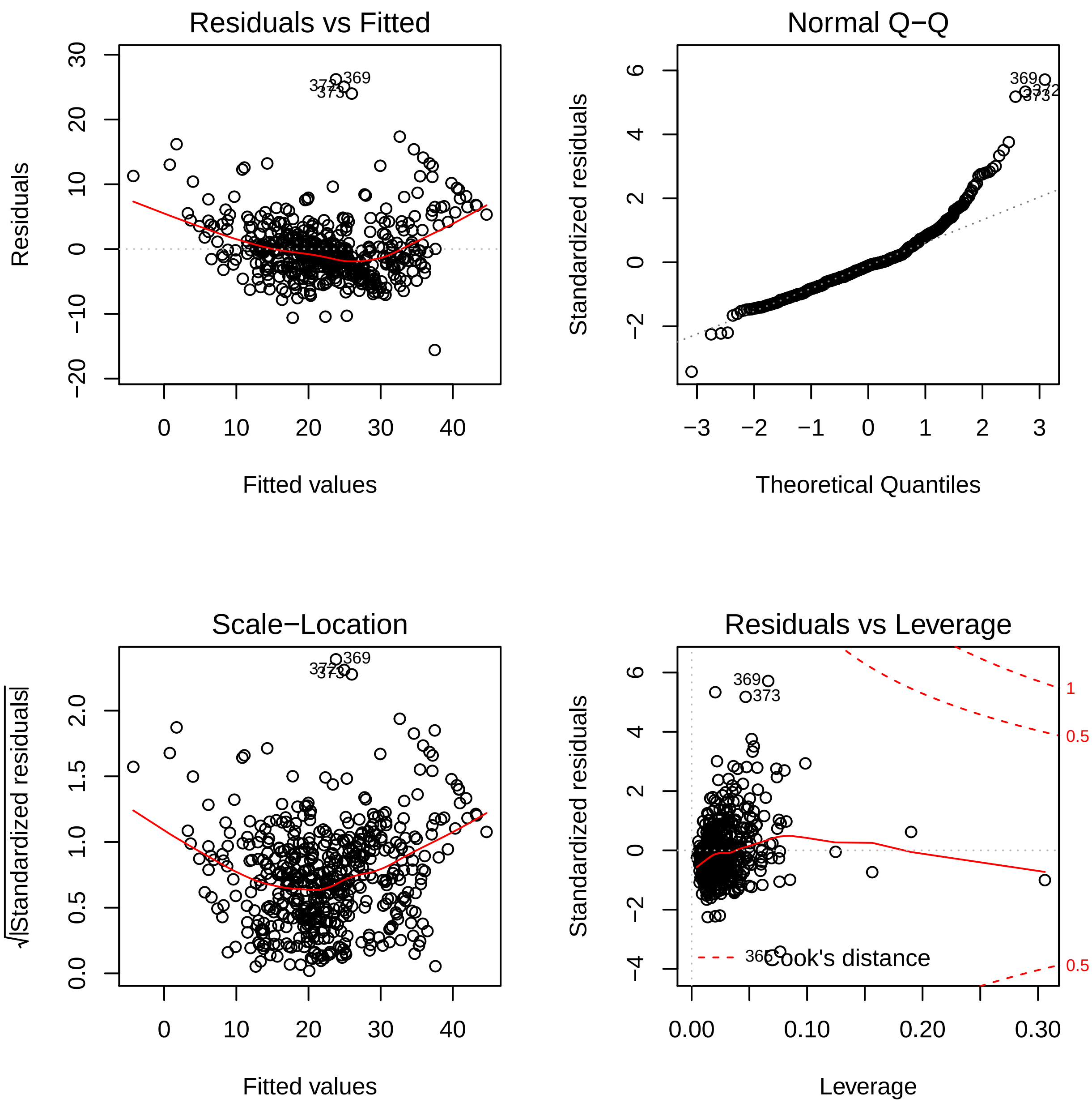
3/4/2019

Code for regression and resulting model.

```
housing_data <- read.csv("housing.data.csv", header = FALSE)
colnames(housing_data) <- c("CRIM","ZN","INDUS","CHAS" ,"NOX" , "RM" , "AGE" , "DIS" ,"RAD" ,
                            "TAX" ,"PTRATIO", "B" , "LSTAT" , "MEDV" )
fit <- lm(MEDV~ ., data =housing_data)
summary(fit)
```

```
##
## Call:
## lm(formula = MEDV ~ ., data = housing_data)
##
## 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
```

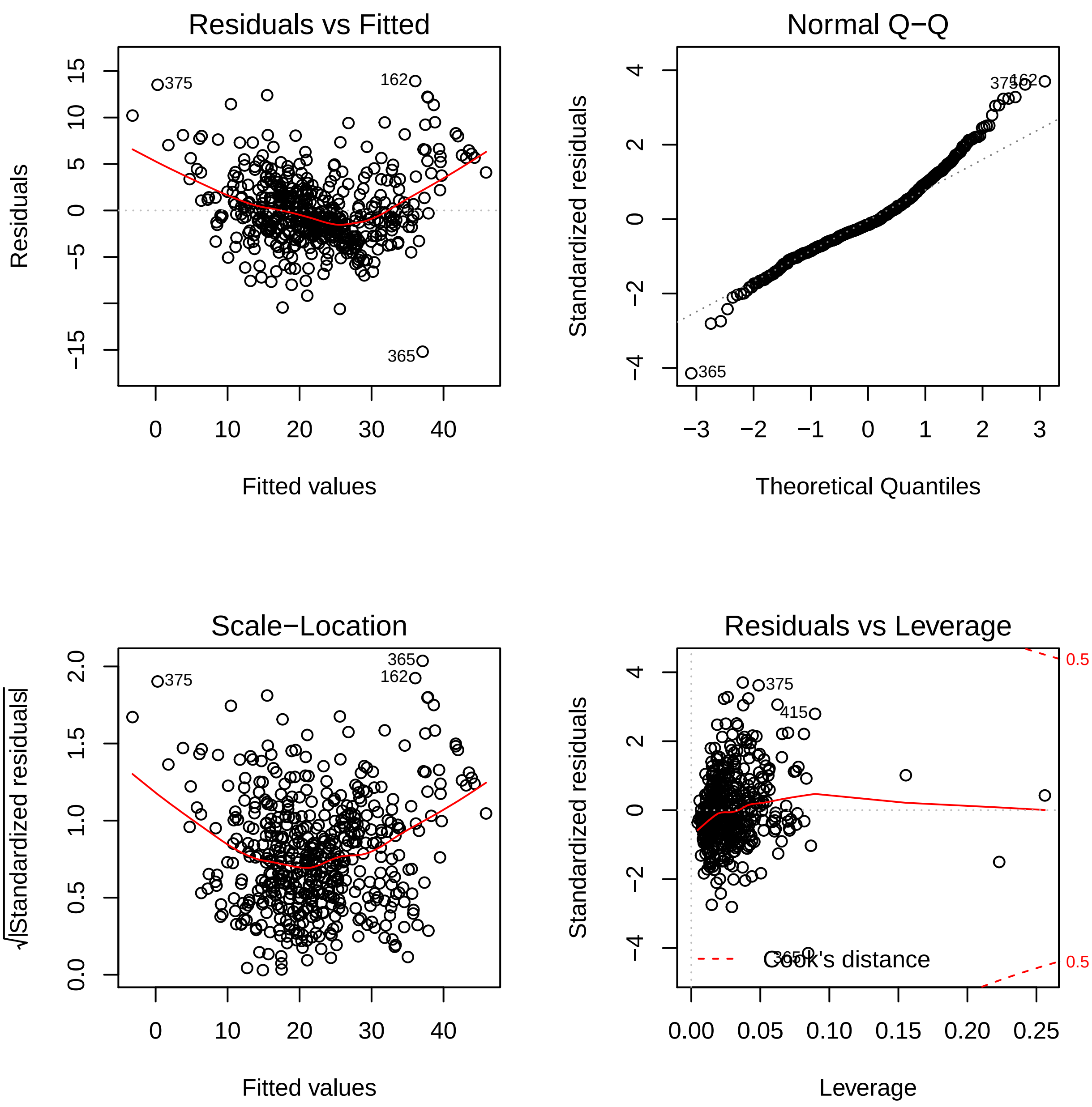
Diagnostic plot



On closely observing the Std Residuals Vs Fitted plot, we see that the point indexes - 369, 373, 372 have standardized residuals that are more than 4 std.deviations away from the mean and also high cook's distance. Additionally, the rows 366, 381 have both high Cook's distance (Those points that are greater than 4/length(all cooksdistances)) as well as high leverage(Those points that have more than 3 times the mean of the leverages), which is not favourable . Hence we will consider these 5 points as outliers and remove them from our dataset. After removing the five points that have questionably high standardized residuals, levergae and Cook's distance, we build a new model and observe the resulting plots.

On examining the plots from the new model, we find more point indexes, 368, 370, 371 and 413, have very high Standarized residuals and Cook's distance. Hence the final model is built after removing all these observations.

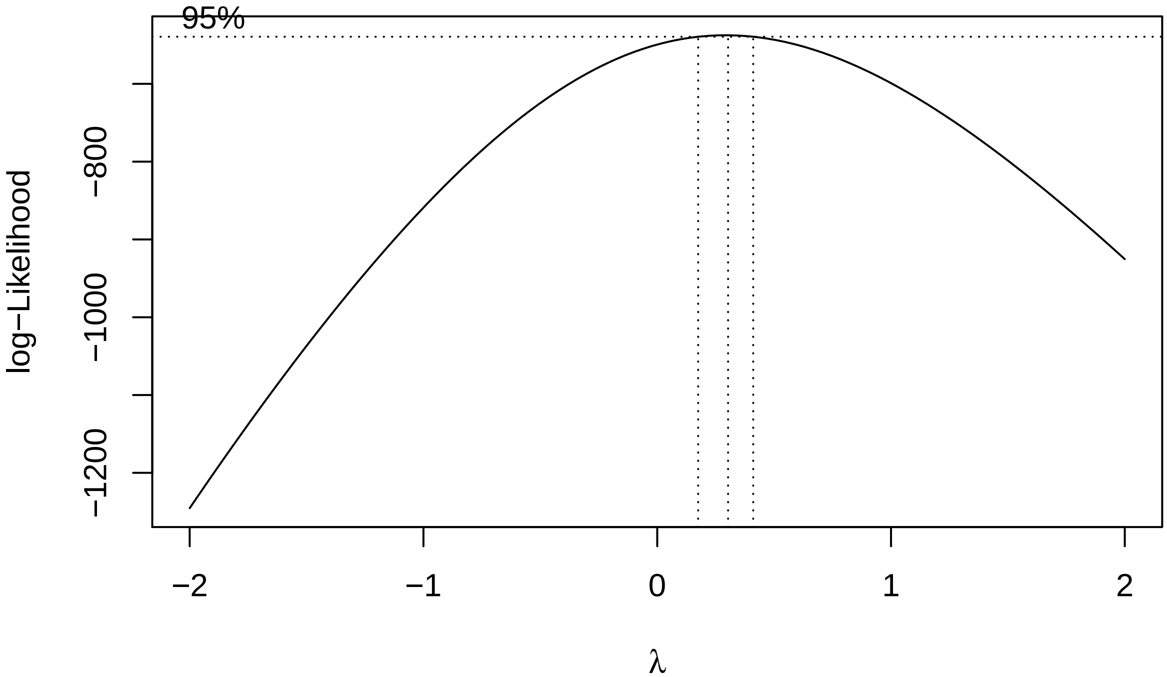
New Diagnostic plot



Code for regression model after removing outliers

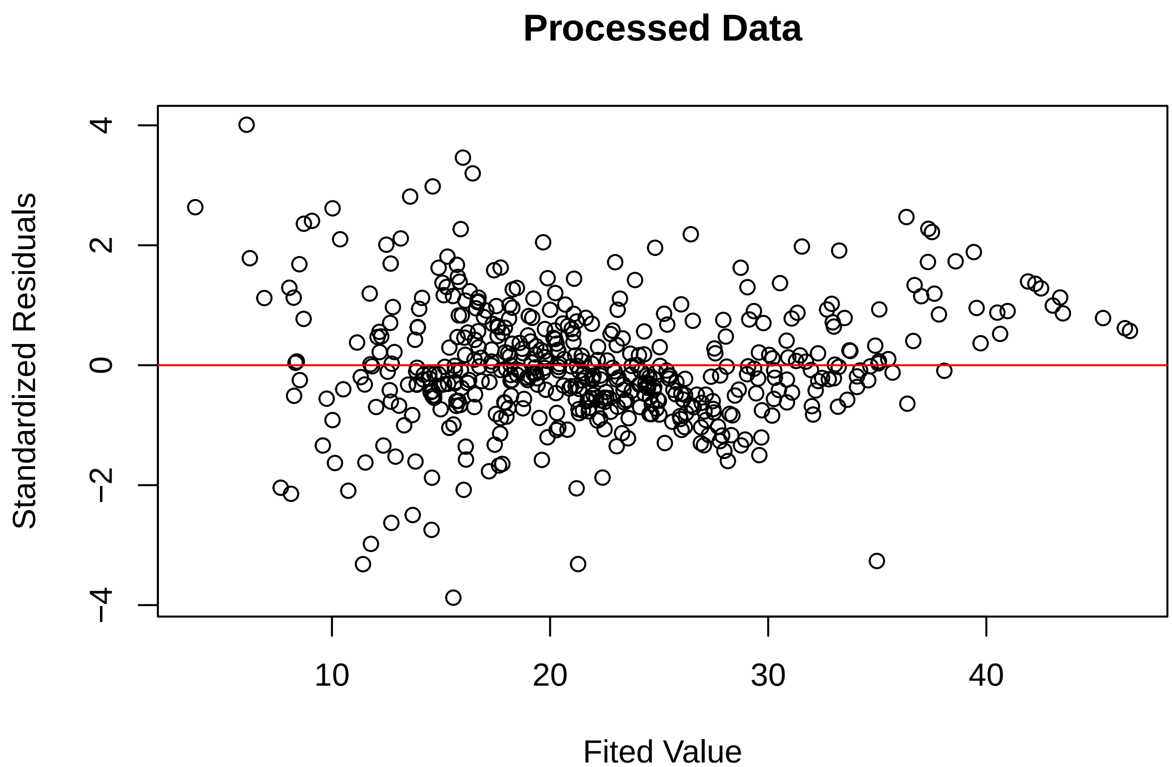
```
# Checking for outliers using leverage, Cook's distance and Standardized residuals
# Leverage cut off 3 times mean value
high_lev <- as.numeric(names(hatvalues(fit)[hatvalues(fit) > 3 * mean(hatvalues(fit))]))
# Std residuals
std.res <- rstandard(fit)[abs(rstandard(fit)) > 4]
possible_outliers <- as.numeric(names(std.res))
# Cooks distance cut off greater than 4
high_cooks <- as.numeric(names(cooks.distance(fit)[cooks.distance(fit) > 4 /
                                                     length(cooks.distance(fit))]))
# Points which have both high leverage and high Cook's distance
high_lev_cooks <- high_lev[high_lev %in% high_cooks]
possible_outliers <- c(possible_outliers,high_lev_cooks)


outlier_treated_housing_data <- housing_data[-c(possible_outliers),]
outlier_fit <- lm(MEDV~ ., data =outlier_treated_housing_data)

par(mfrow=c(2,2))
plot(outlier_fit)

# Repeating outlier removal step
std.res1 <- rstandard(outlier_fit)[abs(rstandard(outlier_fit)) > 4]
possible_outliers1 <- as.numeric(names(std.res1))
outlier_treated_housing_data <- housing_data[-c(possible_outliers,possible_outliers1),]
outlier_fit <- lm(MEDV~ ., data =outlier_treated_housing_data)
possible_outliers1

# Diagnistic Plot
par(mfrow=c(2,2))
plot(outlier_fit)
```

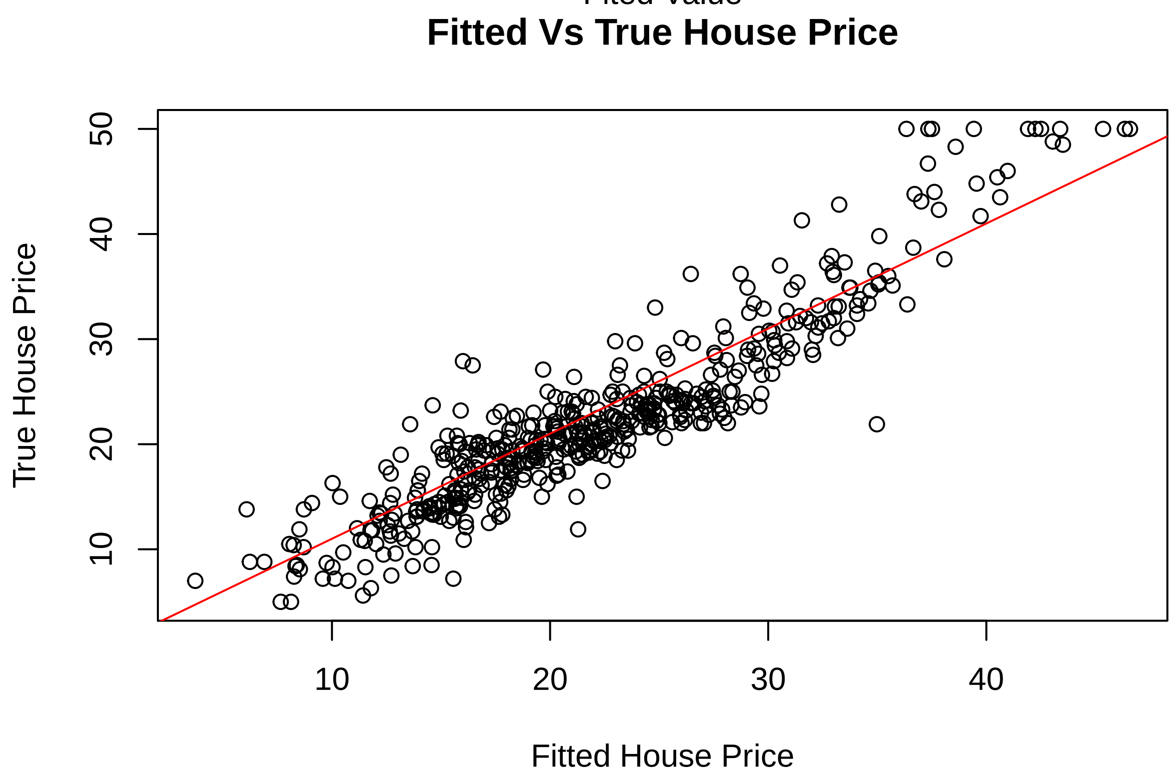
Box__Cox transformation Plot and choosing best value for Lamda



The best value of Lamda is 0.3030303

Standardized Residuals Plot and True Vs Fitted plot after Box-Cox





Code for Box-Cox transformation and regression

```r
bc <- boxcox(outlier_fit, plotit = TRUE)
#Code to obtain best lamda
best_lamda <- bc$x[which(bc$y ==max(bc$y))]

boxcox_fit <- lm ((( MEDV ^ best_lamda) - 1 )/ best_lamda ~ .,
                  data = outlier_treated_housing_data)

# Standardized Residuals Vs Fitted
reverse_transformed_y <- ((boxcox_fit$fitted.values *best_lamda)+1)^(1/best_lamda)
plot(reverse_transformed_y, rstandard(boxcox_fit), ylab="Standardized Residuals",
     xlab="Fited Value", main="Processed Data")
abline(0, 0, col='red')

#True Vs Fitted Plot
plot(reverse_transformed_y, outlier_treated_housing_data$MEDV,
     main ="Fitted Vs True House Price", xlab="Fitted House Price", ylab="True House Price")
abline(1,1, col='red')
```