HW3

Qiuying Li UNI ql2280

9/28/2017

1. **Consider the Boston dataset, in R library MASS, on Housing Values in Suburbs of Boston Fit a multiple linear regression model to predict medv (median value of owner-occupied homes in $1000s) using the following set of predictors**:

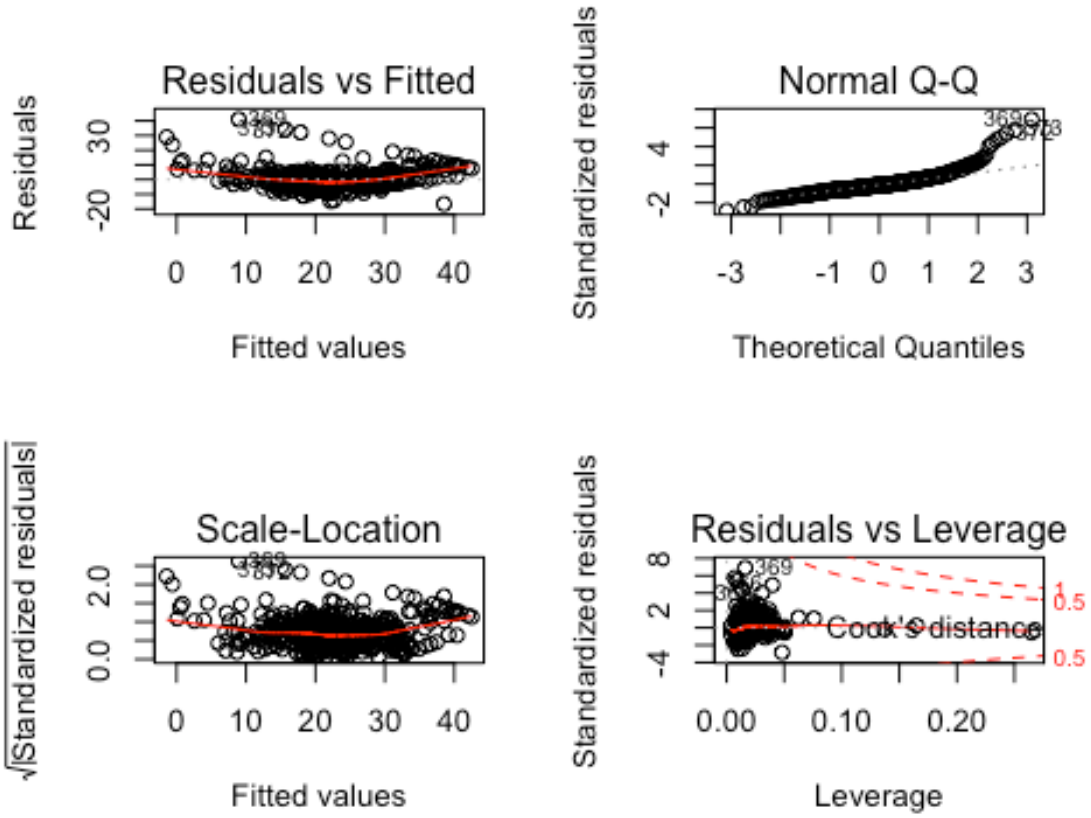
```
library("MASS", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
data = Boston
lm = lm(medv ~ crim+zn+indus+nox+rm+age+tax, data)
summary(lm)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + indus + nox + rm + age + tax,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.625  -3.161  -0.833   2.089  41.042
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.615259   3.221482  -6.089 2.27e-09 ***
## crim         -0.132538   0.038482  -3.444 0.000621 ***
## zn            0.022103   0.014823   1.491 0.136547
## indus        -0.014980   0.072282  -0.207 0.835909
## nox           0.010643   4.230468   0.003 0.997994
## rm            7.606508   0.418424  18.179  < 2e-16 ***
## age          -0.023198   0.014893  -1.558 0.119964
## tax          -0.009006   0.002662  -3.384 0.000772 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.989 on 498 degrees of freedom
## Multiple R-squared:  0.5818, Adjusted R-squared:  0.576
## F-statistic: 98.99 on 7 and 498 DF,  p-value: < 2.2e-16
```

Based on the results above, medv =-0.13*crim +0.02*zn -0.015 * indus +

0.01* nox + 7.6 * rm + -0.02 * age -0.009 * tax – 19.615

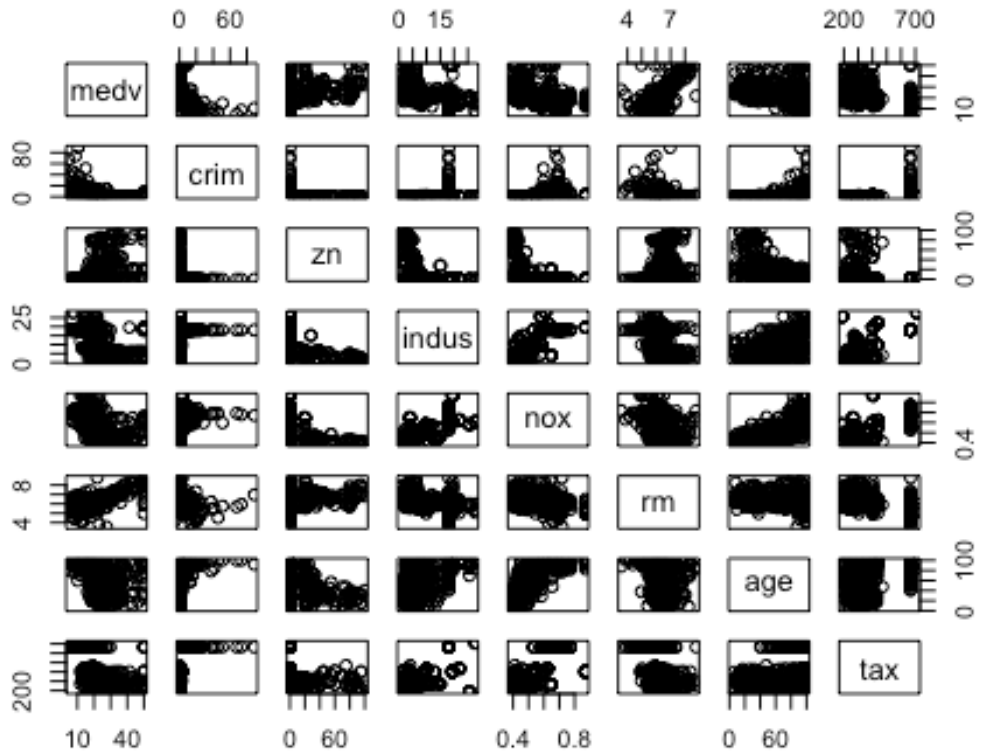
```r
par(mfrow=c(2,2))
plot(lm)
```



2. State and assess the validity of the underlying assumptions, and suggest remedial measures in case of violations of any of the underlying assumptions

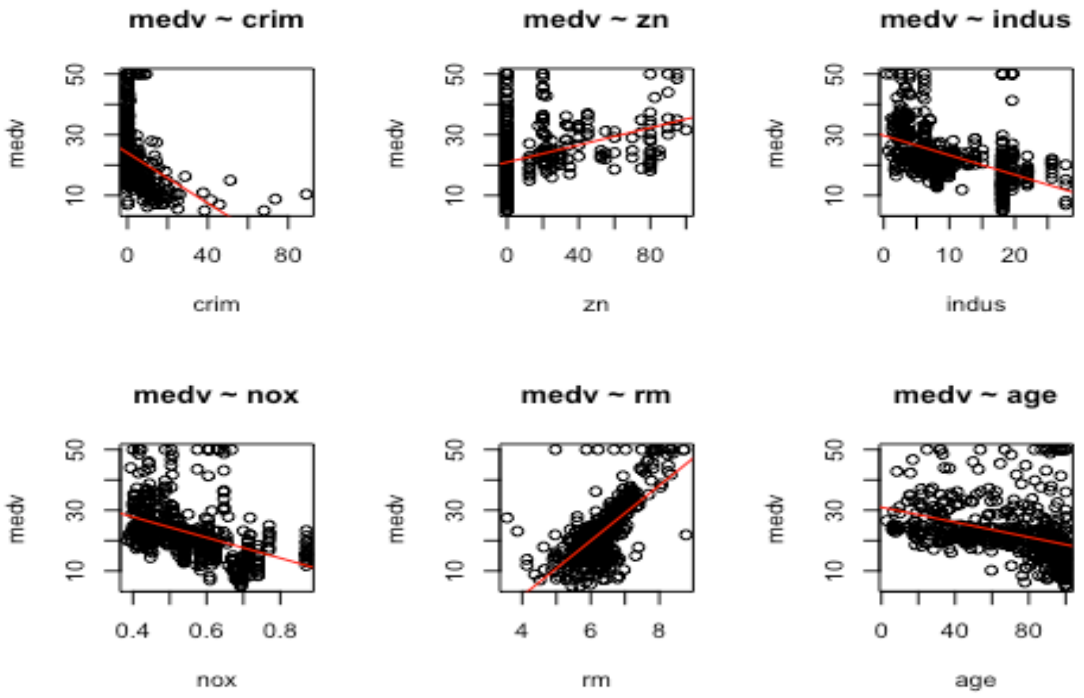
- **Linearity/functional form**

```
#(a)Check matrix scatter plot of all variables
pairs(medv ~crim+zn+indus+nox+rm+age+tax, data)
```




```
#(b)Check linearity of each predictor
lm1 = lm(medv~crim,data)
lm2 = lm(medv~zn,data)
lm3 = lm(medv~indus,data)
lm4 = lm(medv~nox,data)
lm5 = lm(medv~rm,data)
lm6 = lm(medv~age,data)
lm7 = lm(medv~tax,data)
par(mfrow = c(2,3))
plot(data$crim,data$medv, main = "medv ~ crim", xlab = "crim", ylab = "medv")
abline(lm1, col = "red")
plot(data$zn,data$medv, main = "medv ~ zn", xlab = "zn", ylab = "medv")
abline(lm2, col = "red")
```

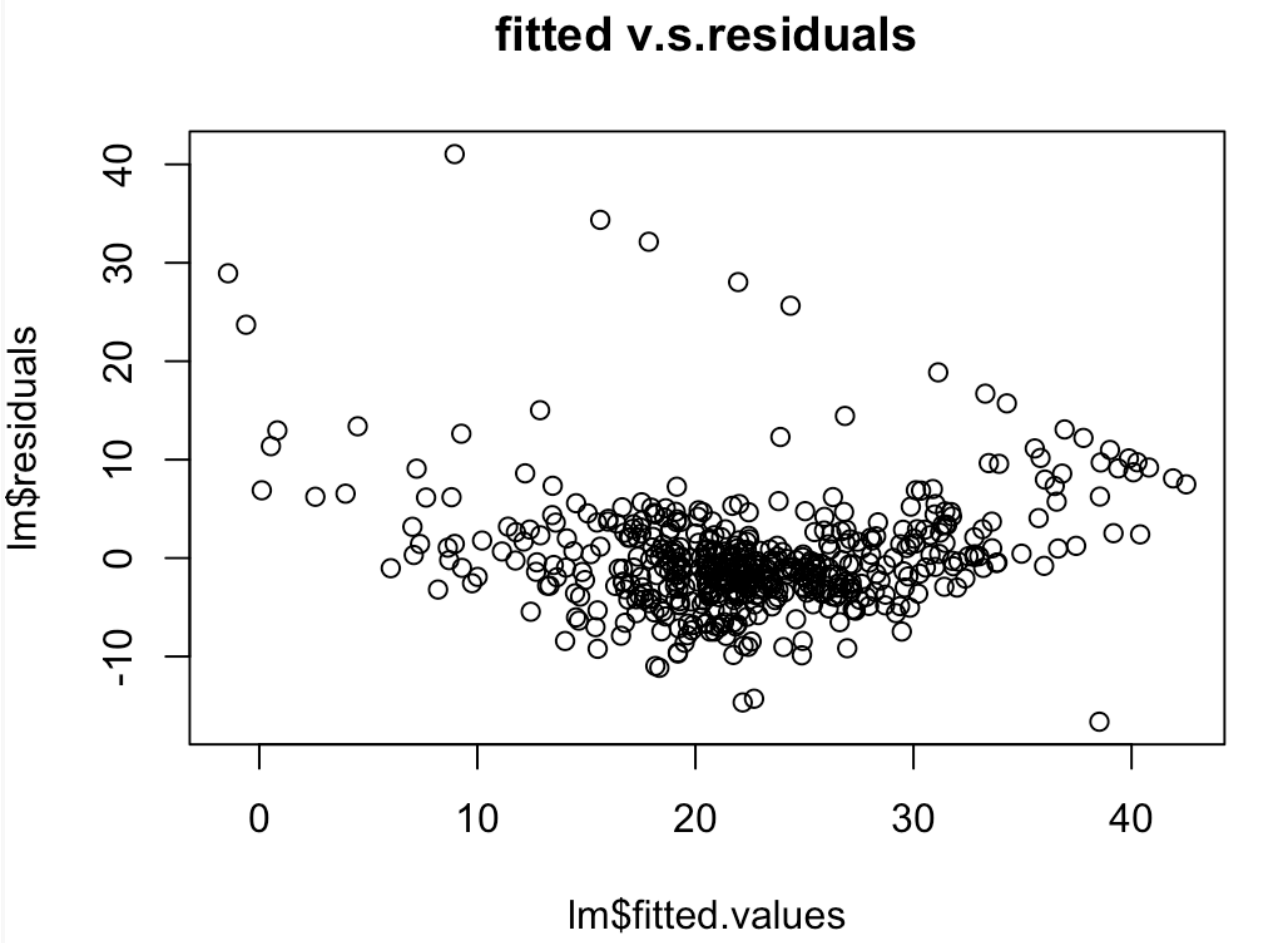
```r
plot(data$indus,data$medv, main = "medv ~ indus", xlab = "indus", ylab = "med
v")
abline(lm3, col = "red")
plot(data$nox,data$medv, main = "medv ~ nox", xlab = "nox", ylab = "medv")
abline(lm4, col = "red")
plot(data$rm,data$medv, main = "medv ~ rm", xlab = "rm", ylab = "medv")
abline(lm5, col = "red")
plot(data$age,data$medv, main = "medv ~ age", xlab = "age", ylab = "medv")
abline(lm6, col = "red")
```



Plot description

From the liearty plot of each predictor, we can see the simple linear regression relationship is not very significant. We cannot find the perfect linear lines in above plots, which means we need more tests on linearity of these predictors.

```r
#(c)Check fitted value V.S. residuals
plot(lm$fitted.values,lm$residuals, main = " fitted v.s residuals)
```



Plot description:
From the plot of fitted value v.s. residuals, we can see that most of the residuals are around 0, and all the points are scatter around the horizontal line, except several extremely large outliers, which affect much on the regression line. Last but not the least, the residuals are horn-shaped, then we should do some transform on response variable, such as logarithm and take reciprocal In order to make sure the accuracy of the linearity, we need check R^2.

```r
#(d) Check the R^squared is 0.5818, and adjust R2 = 0.57
```

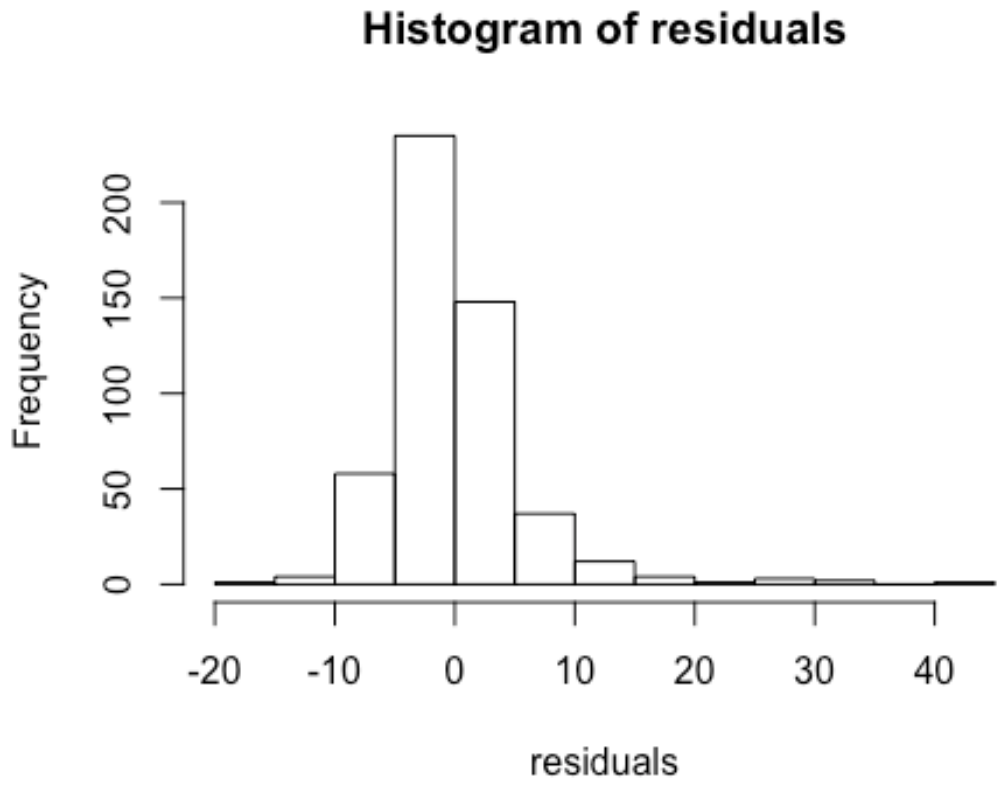
Based on the summary table of the multi-linear models, the R^2 is 0.5818, which means 58.8% change of the response variable by one unit can be explained by these predictors. This number is not a strong evidence for the linearity of the model.

Linearity Remedies:

Since the3 tests(linearity regression plot, fitted v.s. residual plot and R^2) conclude that the response and explanatory variables are not in perfectly linearly form, thus we should try to transform the data, or deal with the outliers. Maybe we could also exclude some variables or add more variables. Moreover, we can also try some simple-linear form on the data Check

- **Normality**

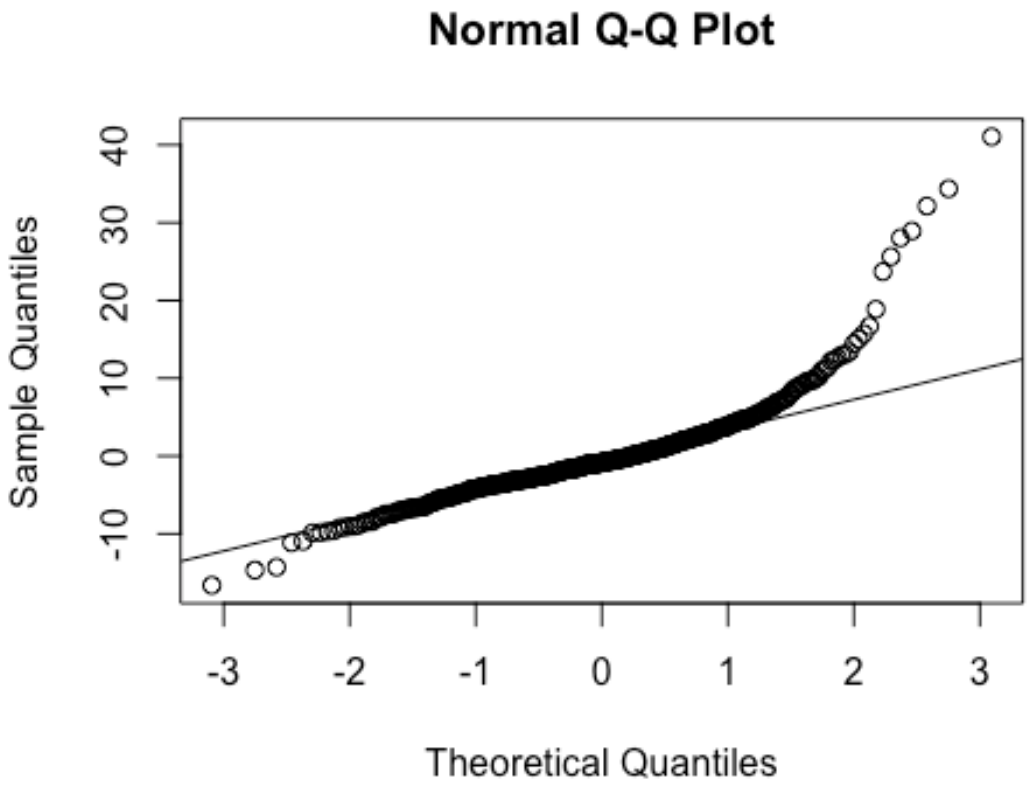
```
# (a) check histogram
hist(lm$residuals, xlab = "residuals", main = "Histogram of residuals")
abline(density(lm$residuals),col = "red")
```



Plot description:

This histogram of the residuals is not in a symmetric bell- curved shape, and we can see that it is skewed. The assumption of normality was not supported by this histogram.

```r
#(b) check QQ-PLOT
qqnorm(lm$residuals, main = "Normal Q-Q Plot",
       xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
qqline(lm$residuals)
```



Plot description:

The QQ line does not follow the diagonal line of the QQ plot, which means we cannot support the assumption of the normality.

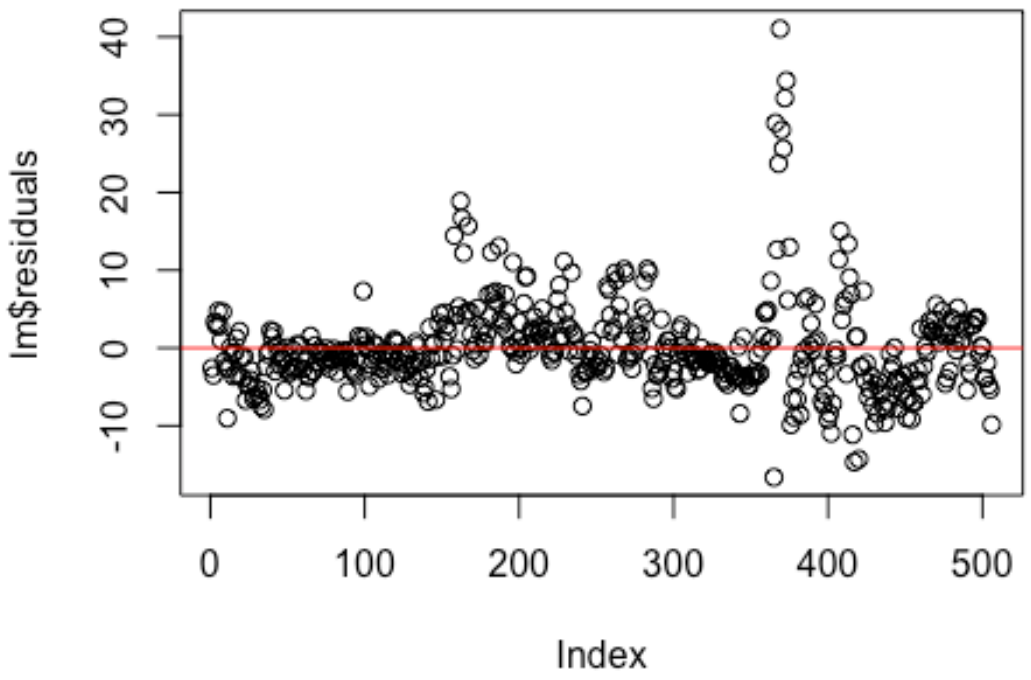
```r
# DO shapiro test: H0 : residuals ~ Normal Ha : residuals not ~ Norma

shapiro.test(lm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm$residuals
## W = 0.83945, p-value < 2.2e-16
```

Results: From the results above, p value is closed to 0, we have strong evidence to reject the null, which means the residuals are not normal.

- **Homoscedasticity**

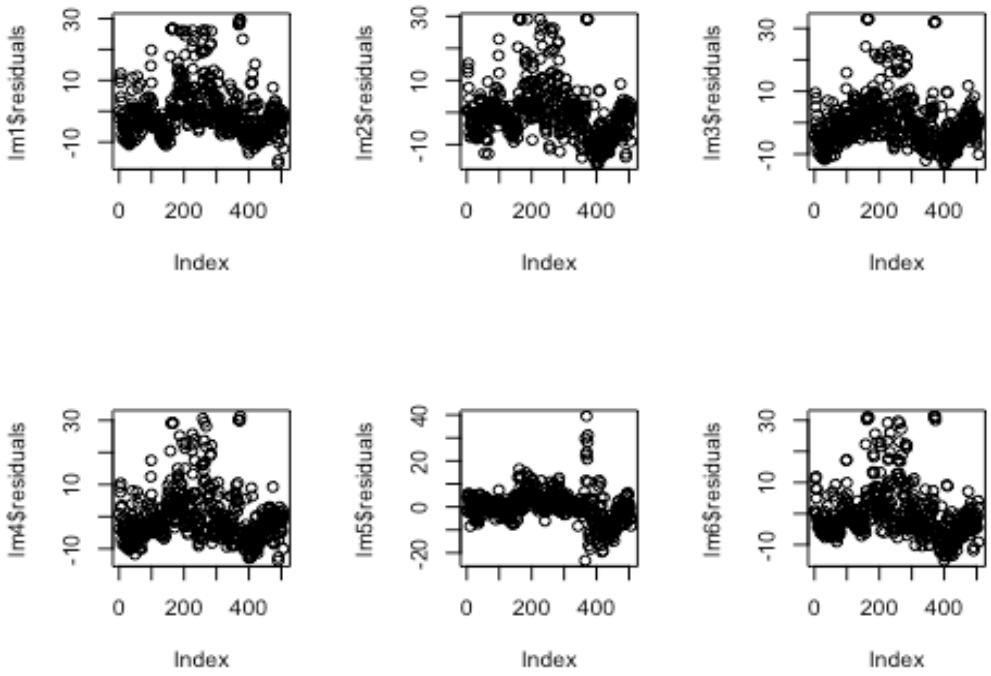
```
#(a) Check the residuals scatterplot of the multi-linear models
plot(lm$residuals)
abline(0,0, col = "red")
```



Plot description:

Most of the residuals of the multi-linear model are scatter around 0, except from several extremely large values. In order to make sure the assumption of Homoscedasticity, we need more tests.

```
#(b) Check the residuals of each predictor
par(mfrow = c(2,3))
plot(lm1$residuals)
plot(lm2$residuals)
plot(lm3$residuals)
plot(lm4$residuals)
plot(lm5$residuals)
plot(lm6$residuals); plot(lm7$residuals)
```



Plot description:

From above 7 plots, we cannot find a constant variance for any of the predictor. Although all predictors have residuals around the 0, all of predictors have extremely large values affect the assumption of the Homoscedasticity.

Homoscedasticity Remedies: From the previous plot, we can conclude that the data are not satisfying homoscedasticity. First, we could transform the data. In addition to data transformation, we can try weighted least sum of squares since the data residuals are non-constancy variance.

- **Uncorrelated error**

```
library("car", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
durbinWatsonTest(lm)
```

$H_0 : \rho = 0$ $H_a : \rho \neq 0$

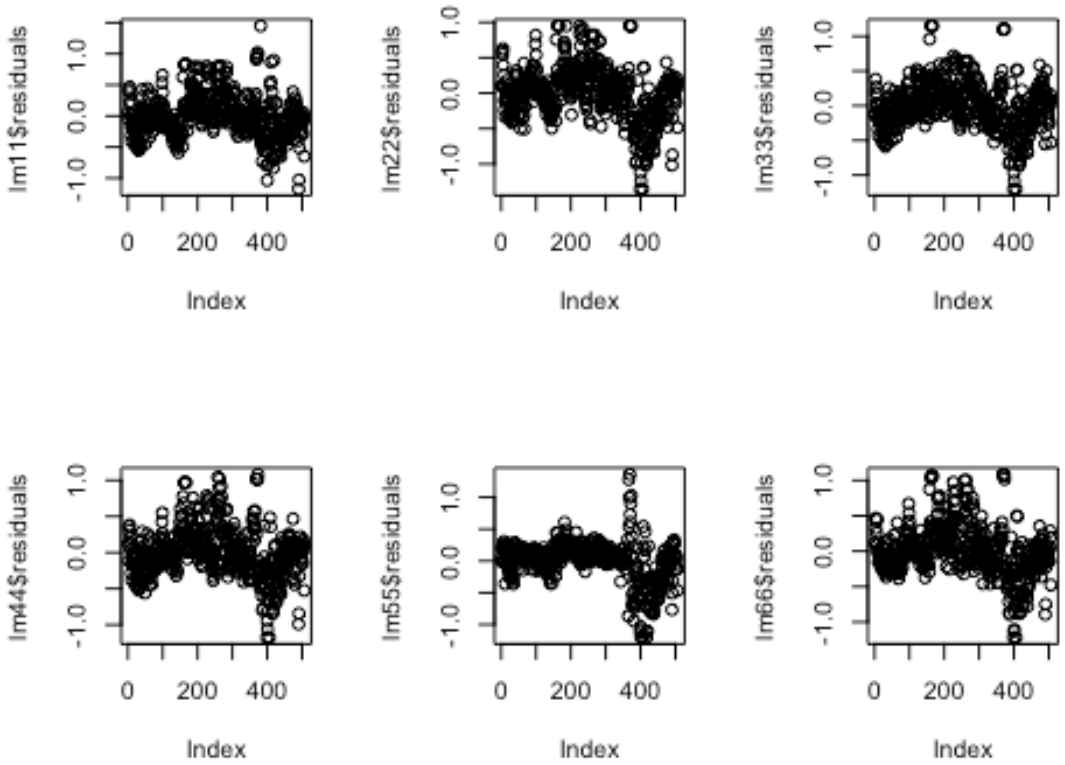
```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.6326847     0.7288349       0
##  Alternative hypothesis: rho != 0
```

Results: p value is very closed to 0, which means that we can reject null, and conclude that there are correlated errors in this regression.

Uncorrelated Error Remedies: from the test results, there are correlated errors, we should process Cochrane-Orcutt data transformation, and also use Generalized Estimating Equations to avoid the correlated errors.

- **Remedial measures: log data transformation**

```
lm11 = lm(log(medv)~crim,data)
lm22 = lm(log(medv)~zn,data)
lm33 = lm(log(medv)~indus,data)
lm44 = lm(log(medv)~nox,data)
lm55 = lm(log(medv)~rm,data)
lm66 = lm(log(medv)~age,data)
lm77 = lm(log(medv)~tax,data)
par(mfrow = c(2,3))
plot(lm11$residuals)
plot(lm22$residuals)
plot(lm33$residuals)
plot(lm44$residuals)
plot(lm55$residuals)
plot(lm66$residuals)
```



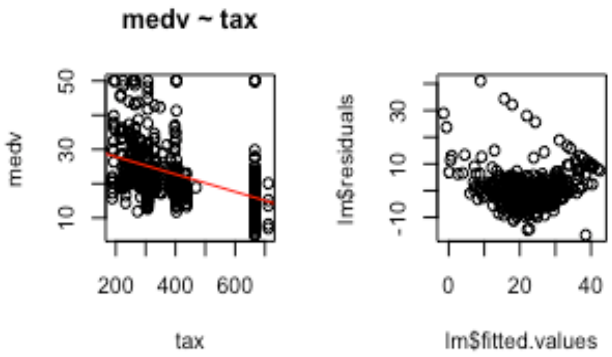
3. use least median square method do linear regression

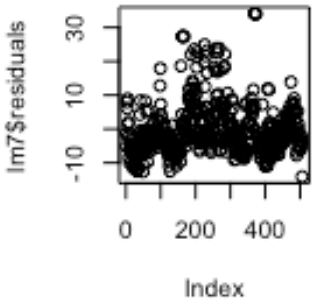
```
lm_median<-lmsreg(medv ~ crim+zn+indus+nox+rm+age+tax, data=Boston)
lm_median
```

```
## Call:
## lqs.formula(formula = medv ~ crim + zn + indus + nox + rm + age +
##     tax, data = Boston, method = "lms")
##
## Coefficients:
## (Intercept)         crim           zn        indus          nox
##  -31.504274    -1.235229    -0.013121     0.078284     9.341964
##          rm          age          tax
##    8.268715    -0.070780     0.005653
##
## Scale estimates 3.695 3.488
```

The least median of squares regression is:
$medv = -27.99 - 0.71crim + 0.03zn + 0.02indus + 3.70nox + 8.02rm - 0.04age - 0.001tax$

Compares to the multiple linear regression:
$medv = -19.62 - 0.13crim + 0.02zn - 0.01indus + 0.1nox + 7.61rm - 0.02age - 0.01tax$


medv ~ tax
medv
50
30
10
200
400
600
tax
lm$residuals
30
10
-10
0
20
40
lm$fitted.values


lm7$residuals
30
10
-10
0
200
400
Index