MINI PROJECT

On

BOSTON HOUSING DATA

(Using R)

Prepared By:

Wyendrila Roy
http://in.linkedin.com/pub/wyendrila-roy/5/3a/876

Understanding the Data

Variable Name	Definition
MEDV	median value in $1000
DIS	distance to employment centers
CRIM	per capita crime rate
RAD	accessibility to radial highways
ZN	% land zoned for lots
INDUS	% nonretail business
TAX	property tax/$10,000
CHAS	1 on Charles River, 0 else
PT	pupil/teacher ratio
NOX	nitrogen oxide conc. (p.p.109)
Black	(% black - 63)2/10
RM	average number of rooms
LSTAT	% lower-status pop.
AGE	% built before 1940

Structure of the Data

> data(Boston, package="MASS") (This function reads the Boston dataset from the MASS Package in R)

> names(Boston) (This function gives the variable names in the Boston dataset, the definition has been provided at the beginning for reference)

```
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
 [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

> str(Boston) (This function gives the structure of the data, the total number of variables and observations in the data and the datatype of each variable)

```
'data.frame':	506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm     : num  6.58 6.42 7.18 7 7.15 ...
 $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

`> head(Boston)` **(To view the first few rows in the data in order to check the data as a sample)**

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

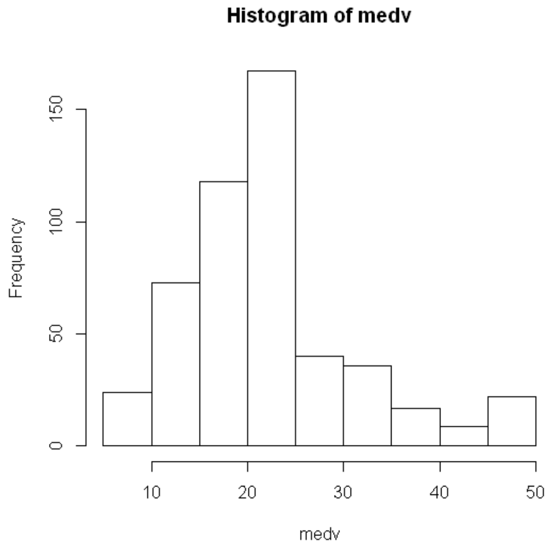
`> summary(Boston)` **(To understand the basic statistics of each individual variable)**

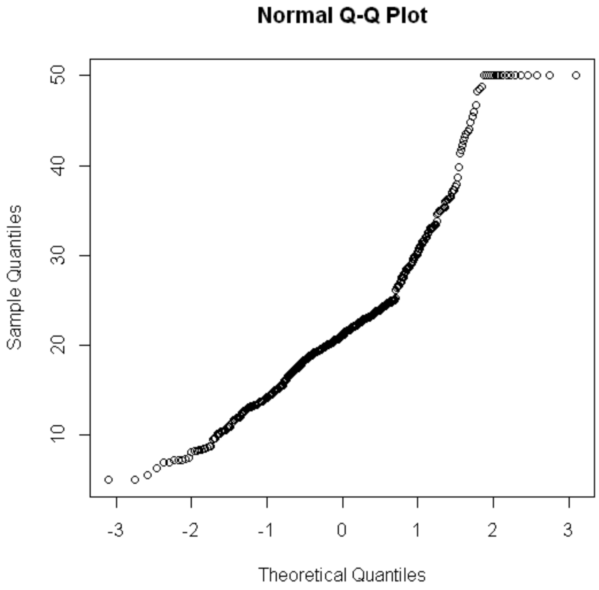
```
     crim              zn                indus              chas

 Min.   : 0.00632         Min.   :  0.00     Min.   : 0.46     Min.   :0.00000
 1st Qu.: 0.08204         1st Qu.:  0.00     1st Qu.: 5.19     1st Qu.:0.00000
 Median : 0.25651         Median :  0.00     Median : 9.69     Median :0.00000
 Mean   : 3.61352         Mean   : 11.36     Mean   :11.14     Mean   :0.06917
 3rd Qu.: 3.67708         3rd Qu.: 12.50     3rd Qu.:18.10     3rd Qu.:0.00000
 Max.   :88.97620         Max.   :100.00     Max.   :27.74     Max.   :1.00000

     nox                  rm                 age                dis

 Min.   :0.3850           Min.   :3.561      Min.   :  2.90    Min.   : 1.130
 1st Qu.:0.4490           1st Qu.:5.886      1st Qu.: 45.02    1st Qu.: 2.100
 Median :0.5380           Median :6.208      Median : 77.50  Median : 3.207
 Mean   :0.5547           Mean   :6.285      Mean   : 68.57    Mean   : 3.795
 3rd Qu.:0.6240           3rd Qu.:6.623      3rd Qu.: 94.08    3rd Qu.: 5.188
 Max.   :0.8710           Max.   :8.780      Max.   :100.00    Max.   :12.127

     rad                  tax                ptratio            black

 Min.   : 1.000           Min.   :187.0      Min.   :12.60     Min.   :  0.32
 1st Qu.: 4.000           1st Qu.:279.0      1st Qu.:17.40     1st Qu.:375.38
 Median : 5.000           Median :330.0    Median :19.05   Median :391.44
 Mean   : 9.549           Mean   :408.2      Mean   :18.46     Mean   :356.67
 3rd Qu.:24.000           3rd Qu.:666.0      3rd Qu.:20.20     3rd Qu.:396.23
 Max.   :24.000           Max.   :711.0      Max.   :22.00     Max.   :396.90
    lstat                 medv
```

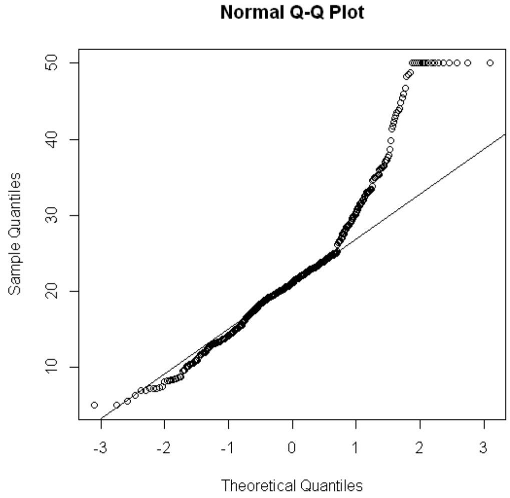
```
Min.   : 1.73          Min.   : 5.00
1st Qu.: 6.95          1st Qu.:17.02
Median :11.36          Median :21.20
Mean   :12.65          Mean   :22.53
3rd Qu.:16.95          3rd Qu.:25.00
Max.   :37.97          Max.   :50.00
```

Viewing the median value of owner occupied homes in $1000 by using histogram and normal Q-Q Plot

```
attach(Boston)
hist(medv)
qqnorm(medv)
qqline(medv)
```







> cor(Boston) (To understand the correlation between each individual variable)

	crim	zn	indus	chas	nox
crim	1.00000000	-0.20046922	0.40658341	-0.055891582	0.42097171
zn	-0.20046922	1.00000000	-0.53382819	-0.042696719	-0.51660371
indus	0.40658341	-0.53382819	1.00000000	0.062938027	0.76365145
chas	-0.05589158	-0.04269672	0.06293803	1.000000000	0.09120281
nox	0.42097171	-0.51660371	0.76365145	0.091202807	1.00000000
rm	-0.21924670	0.31199059	-0.39167585	0.091251225	-0.30218819
age	0.35273425	-0.56953734	0.64477851	0.086517774	0.73147010
dis	-0.37967009	0.66440822	-0.70802699	-0.099175780	-0.76923011
rad	0.62550515	-0.31194783	0.59512927	-0.007368241	0.61144056
tax	0.58276431	-0.31456332	0.72076018	-0.035586518	0.66802320
ptratio	0.28994558	-0.39167855	0.38324756	-0.121515174	0.18893268
black	-0.38506394	0.17552032	-0.35697654	0.048788485	-0.38005064
lstat	0.45562148	-0.41299457	0.60379972	-0.053929298	0.59087892
medv	-0.38830461	0.36044534	-0.48372516	0.175260177	-0.42732077

	rm	age	dis	rad	tax	ptratio
crim	-0.21924670	0.35273425	-0.37967009	0.625505145	0.58276431	0.2899456
zn	0.31199059	-0.56953734	0.66440822	-0.311947826	-0.31456332	-0.3916785
indus	-0.39167585	0.64477851	-0.70802699	0.595129275	0.72076018	0.3832476
chas	0.09125123	0.08651777	-0.09917578	-0.007368241	-0.03558652	-0.1215152
nox	-0.30218819	0.73147010	-0.76923011	0.611440563	0.66802320	0.1889327
rm	1.00000000	-0.24026493	0.20524621	-0.209846668	-0.29204783	-0.3555015
age	-0.24026493	1.00000000	-0.74788054	0.456022452	0.50645559	0.2615150
dis	0.20524621	-0.74788054	1.00000000	-0.494587930	-0.53443158	-0.2324705
rad	-0.20984667	0.45602245	-0.49458793	1.000000000	0.91022819	0.4647412
tax	-0.29204783	0.50645559	-0.53443158	0.910228189	1.00000000	0.4608530
ptratio	-0.35550149	0.26151501	-0.23247054	0.464741179	0.46085304	1.0000000
black	0.12806864	-0.27353398	0.29151167	-0.444412816	-0.44180801	-0.1773833
lstat	-0.61380827	0.60233853	-0.49699583	0.488676335	0.54399341	0.3740443
medv	0.69535995	-0.37695457	0.24992873	-0.381626231	-0.46853593	-0.5077867

	black	lstat	medv
crim	-0.38506394	0.4556215	-0.3883046
zn	0.17552032	-0.4129946	0.3604453
indus	-0.35697654	0.6037997	-0.4837252
chas	0.04878848	-0.0539293	0.1752602
nox	-0.38005064	0.5908789	-0.4273208
rm	0.12806864	-0.6138083	0.6953599
age	-0.27353398	0.6023385	-0.3769546
dis	0.29151167	-0.4969958	0.2499287
rad	-0.44441282	0.4886763	-0.3816262
tax	-0.44180801	0.5439934	-0.4685359
ptratio	-0.17738330	0.3740443	-0.5077867
black	1.00000000	-0.3660869	0.3334608
lstat	-0.36608690	1.0000000	-0.7376627
medv	0.33346082	-0.7376627	1.0000000

If we look at the median value of owner occupied homes in $1000 (medv) and try to understand its correlation between other variables then we can see that the average number of rooms (rm) has the highest positive correlation with medv and pupil/teacher ratio (ptratio) and % lower-status pop. (lstat) has high negative correlation with medv.

Regression Model

```
> RegModel.2 <- lm(medv~black+chas+crim+dis+lstat+nox+ptratio+rad+rm+zn,
+   data=Boston)

> summary(RegModel.2)

Call:
lm(formula = medv ~ black + chas + crim + dis + lstat + nox +
    ptratio + rad + rm + zn, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-16.2609  -2.9888  -0.5083   1.8041  26.2482  (The Residuals are quite high ranges from -16.26 to 26.24, we will
have to bring this down)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.712342   5.102742   6.803 2.97e-11 ***
black         0.009700   0.002701   3.591 0.000363 ***
chas          2.967868   0.860830   3.448 0.000614 ***
crim         -0.104843   0.033132  -3.164 0.001650 **
dis          -1.429370   0.186922  -7.647 1.08e-13 ***
lstat        -0.528147   0.047930 -11.019  < 2e-16 ***
nox         -20.314416   3.472292  -5.850 8.92e-09 ***
ptratio      -1.014914   0.129006  -7.867 2.30e-14 ***
rad           0.128761   0.040788   3.157 0.001692 **
rm            3.977104   0.407731   9.754  < 2e-16 ***
zn            0.036634   0.013412   2.731 0.006532 **
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.79 on 495 degrees of freedom
Multiple R-squared:  0.7342,            Adjusted R-squared:  0.7288
F-statistic: 136.7 on 10 and 495 DF,  p-value: < 2.2e-16

> vif(RegModel.2)
   black     chas     crim      dis    lstat      nox  ptratio      rad
1.338982 1.052428 1.787963 3.410587 2.579040 3.564036 1.717222 2.776775
      rm       zn
1.806735 2.154054
```

The Variance Inflation Factors are fine because all the variables are below 4.

```
> outlierTest(RegModel.2)
    rstudent unadjusted p-value Bonferonni p
369 5.825701         1.0263e-08   5.1932e-06
372 5.394982         1.0650e-07   5.3886e-05
```

```
373 5.205160        2.8478e-07   1.4410e-04
```

There are 3 outliers we will now remove these outliers from the data and re-run the model

```
> Boston <- Boston[-c(369,372,373),]

> RegModel.2 <- lm(medv~black+chas+crim+dis+lstat+nox+ptratio+rad+rm+zn,
+   data=Boston)

> summary(RegModel.2)

Call:
lm(formula = medv ~ black + chas + crim + dis + lstat + nox + 
    ptratio + rad + rm + zn, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.6641  -2.8410  -0.3751   1.8066  19.2247 (The range of the Residuals have come down considerably)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  28.640030   4.680205   6.119 1.92e-09 ***
black         0.009106   0.002454   3.711  0.00023 ***
chas          2.405876   0.792192   3.037  0.00252 ** 
crim         -0.095376   0.030100  -3.169  0.00163 ** 
dis          -1.229979   0.170831  -7.200 2.27e-12 ***
lstat        -0.433398   0.044591  -9.719  < 2e-16 ***
nox         -19.218002   3.154500  -6.092 2.25e-09 ***
ptratio      -1.025144   0.117171  -8.749  < 2e-16 ***
rad           0.073739   0.037418   1.971  0.04932 *  
rm            4.668916   0.377795  12.358  < 2e-16 ***
zn            0.030425   0.012193   2.495  0.01291 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.349 on 492 degrees of freedom
Multiple R-squared:  0.7699,            Adjusted R-squared:  0.7653 
F-statistic: 164.6 on 10 and 492 DF,  p-value: < 2.2e-16
```

We will create another model eliminating the insignificant variables such as zn and rad

```
> RegModel.3 <- lm(medv~black+chas+crim+dis+lstat+nox+ptratio+rm, data=Boston)

> summary(RegModel.3)

Call:
lm(formula = medv ~ black + chas + crim + dis + lstat + nox + 
    ptratio + rm, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.4796  -2.6529  -0.6067   1.6867  20.2662 (The range of the Residuals have slightly increased but still acceptable)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.706692   4.450328   5.776 1.35e-08 ***
black         0.008151   0.002442   3.338 0.000907 ***
chas          2.382992   0.799459   2.981 0.003017 **
crim         -0.059040   0.027282  -2.164 0.030936 *
dis          -0.991221   0.149398  -6.635 8.56e-11 ***
lstat        -0.425606   0.044940  -9.470  < 2e-16 ***
nox         -16.937158   2.937265  -5.766 1.43e-08 ***
ptratio      -1.014710   0.101895  -9.958  < 2e-16 ***
rm            4.943351   0.372196  13.282  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.39 on 494 degrees of freedom
Multiple R-squared:  0.7646,	Adjusted R-squared:  0.7608 
F-statistic: 200.5 on 8 and 494 DF,  p-value: < 2.2e-16
```

```
> vif(RegModel.3)
   black     chas     crim      dis    lstat      nox  ptratio       rm 
1.302030 1.051365 1.440841 2.569929 2.686973 3.024642 1.270217 1.778336 
```

The Variance Inflation Factors are fine because all the variables are below 4.

Residual Analysis

Residual analysis is usually done graphically using:

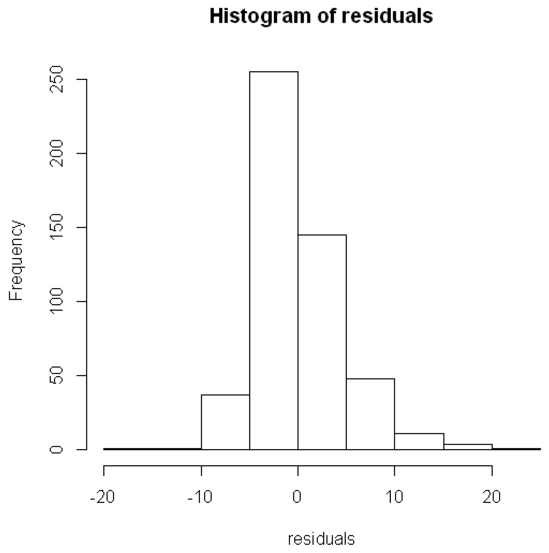
- **Quantile plots: to assess normality**
- **Histograms and boxplots**
- **Scatterplots: to assess model assumptions, such as constant variance and linearity, and to identify potential outliers**

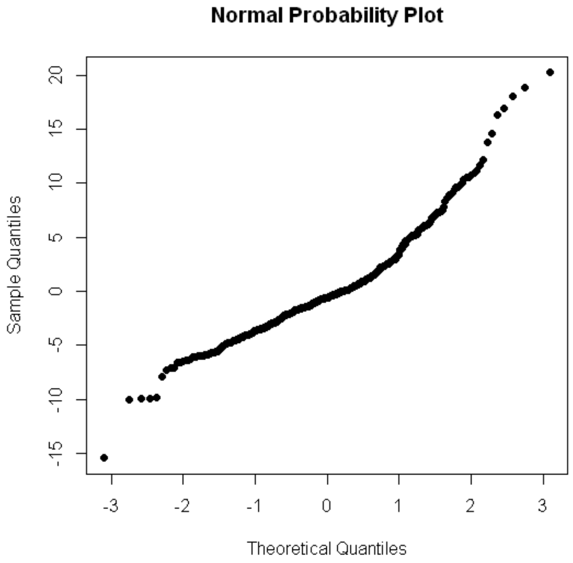
```
> mean(RegModel.3$residuals)
[1] 2.127881e-17
```

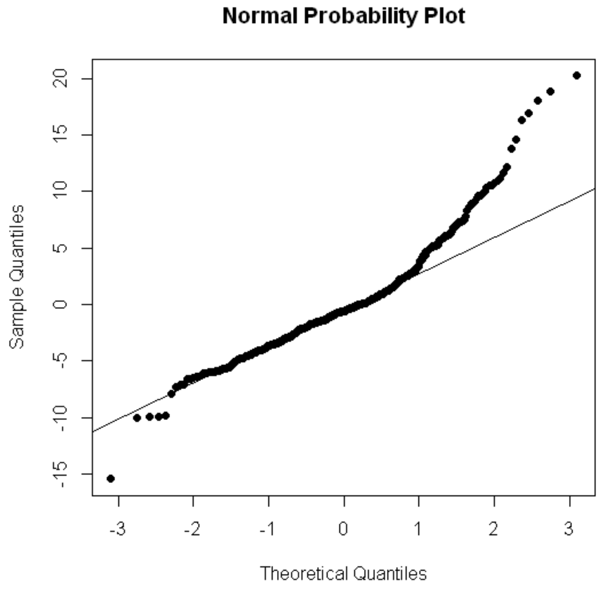
```
hist(RegModel.3$residuals, xlab="residuals", main="Histogram of residuals")
```

```
qqnorm(RegModel.3$residuals, main="Normal Probability Plot", pch=19)
```

```
qqline(RegModel.3$residuals)
```







Checking for linear relationship amongst the important variables, error should have a constant variance and error terms should not be independent

Plotting residuals against key predictor variables, lstat, ptratio and rm

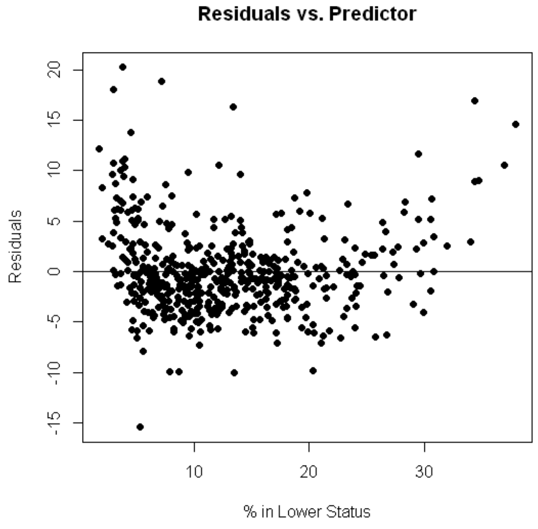
```
plot(lstat, RegModel.3$residuals, main="Residuals vs. Predictor", xlab="% in Lower Status", ylab="Residuals",
pch=19)
abline(h=0)

plot(rm, RegModel.3$residuals, main="Residuals vs. Predictor", xlab="Average No of Rooms", ylab="Residuals",
pch=19)
```

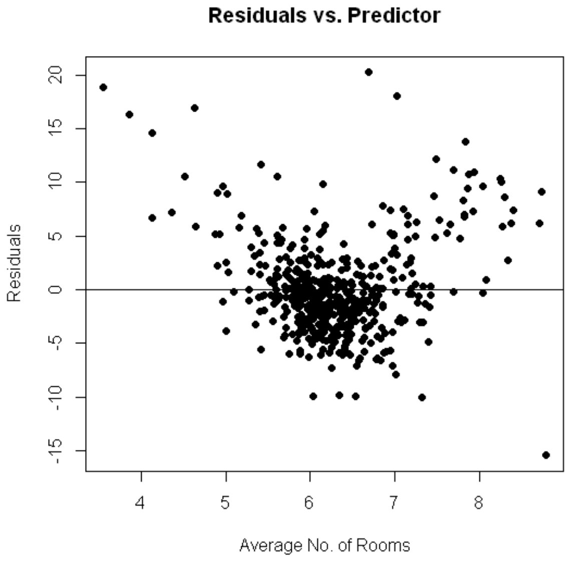
```
abline(h=0)
```

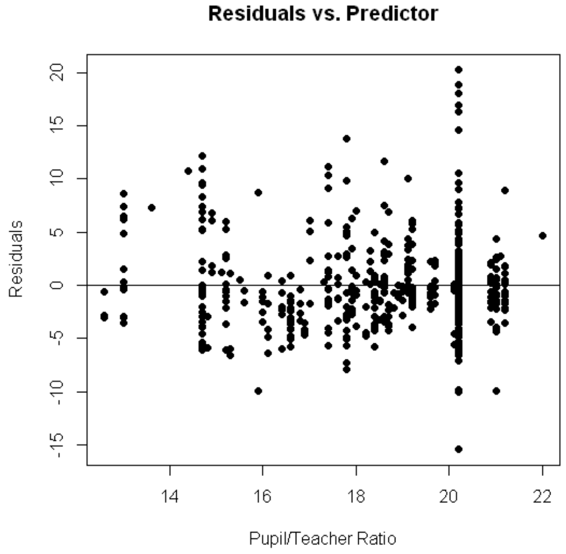
```
plot(ptratio, RegModel.3$residuals, main="Residuals vs. Predictor", xlab="Pupil/Teacher Ratio",
ylab="Residuals", pch=19)
abline(h=0)
```

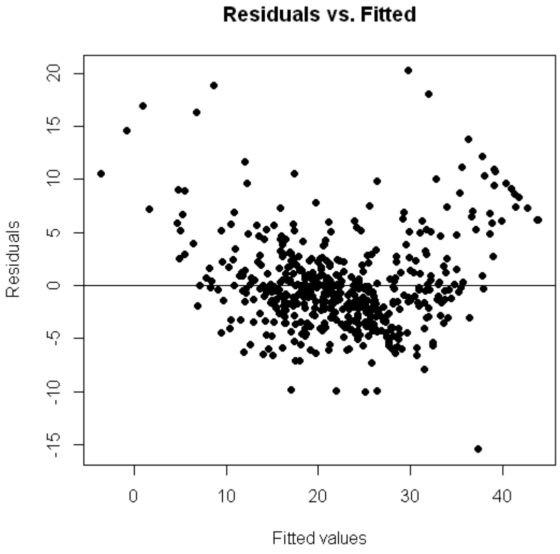
Plotting residuals against fitted values

```
plot(RegModel.3$fitted.values, RegModel.3$residuals, main="Residuals vs. Fitted", xlab="Fitted values",
ylab="Residuals", pch=19)
abline(h=0)
```









```
> anova(RegModel.2,RegModel.3)

Analysis of Variance Table

Model 1: medv ~ black + chas + crim + dis + lstat + nox + ptratio + rad +
    rm + zn
Model 2: medv ~ black + chas + crim + dis + lstat + nox + ptratio + rm
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    492 9303.9
2    494 9520.3 -2    -216.4 5.7216 0.003496 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```