**Name:** HOANG VO

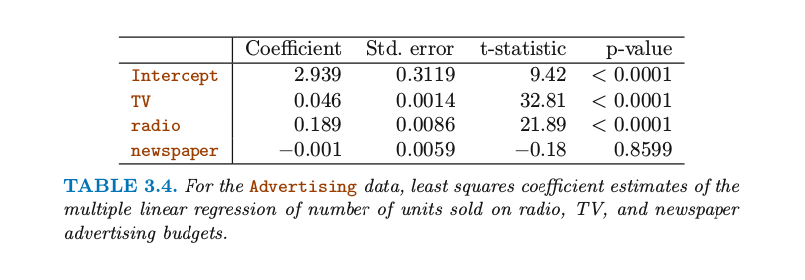
**MyUH ID:** 1671058

**Math 4322**

*Homework #2*

Conceptual

**Exercise 1:**



The Null Hypotheses is that the advertising budgets on radio, TV, and newspaper have no effect on Sales of the product.

* H0: B1=B2=B3=0

or:

H(1)0: B1= 0 , in which B1 is effect of radio advertising on sales

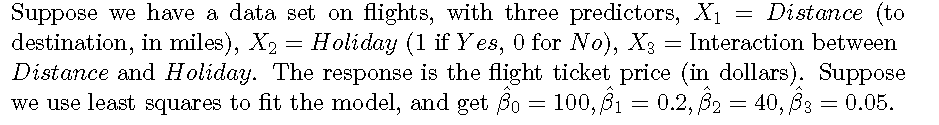
H(2)0: B2 = 0 , in which B2 is effect of TV advertising on sales

H(3)0: B3 = 0, in which B3 is effect of newspaper advertising on sales

Basing on the table 3.4, the p-values are significant for TV (B1) and radio (B2) : <0.0001 (<0.05) => reject H(1)0 and H(2)0. Therefore, we can say that advertising in radio and TV affects sales of products.

The p-value is not significant for newspaper (B3): 0.8599 => We do not reject H(3)0. The advertising in newspaper does not affect sales.

**Exercise 2:**



1. Which answer is corrected, and why?
2. For a fixed value of Distance, on average tickets are more expensive on holidays than on usual days.

Correct because:

-On Holidays, X2= 1:

the flight ticket price = 100 + 0.2 \* Distance + 40 \* Holiday +0.05 \* Distance\*Holiday

-On usual days, the X2 = 0:

the flight ticket price = 100+0.2\*Distance

* The average tickets are more expensive on Holidays than on usual Days.

1. For a fixed value of Distance, on average tickets are more expensive on usual days than on holidays.

Incorrect. It was explained in i) above.

1. For a fixed value of Distance, on average tickets are more expensive on usual days than on holidays, provided that Distance is long enough.

Incorrect. Since the equation is (100 + 0.2 \* Distance + 40 \* Holiday +0.05 \* Distance\*Holiday), the longer the distance, the more expensive the tickets are on Holidays.

1. For a fixed value of Distance, on average tickets are more expensive on holidays than on usual days, provided that Distance is long enough.

Correct. However, regardless of the value of distance, as long as the distance is fixed, the price of tickets on holidays is more expensive than the price of tickets on usual days.

1. Predict the average holiday price of a ticket for a flight that travels 1000 miles to its destination.

Distance X1 = 1000

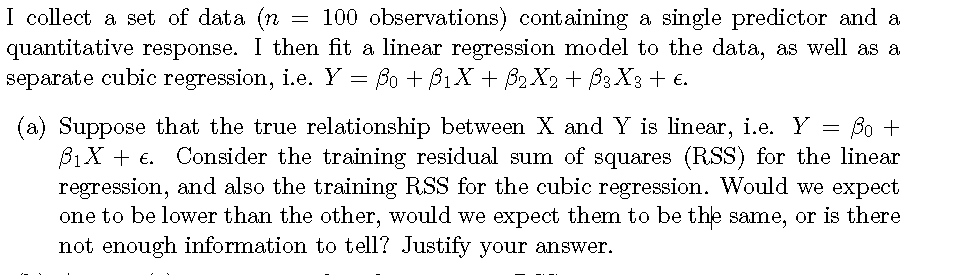
Holiday: X2 = 1

The flight ticket price = 100 + 0.2\*1000 + 40\*1 + 0.05\*1000\*1 = 490 (Dollars)

1. True or false: Since the coefficient for the Distance/Holiday interaction term is pretty small, there is no evidence of an interaction effect. Justify your answer.

FALSE: because the statistical significance of an interaction is different from the magnitude of the interaction. It's possible to have a lot of evidence for a small effect. Moreover, units of the two variables Holiday and Distance can have some roles to result to the small coefficient for the Distance/Holiday interaction.

**Exercise 3:**



The cubic regression model is more flexible than the linear regression model. Therefore, we would expect the cubic model to fit better the data, and thus to have lower training RSS.

b)

If the true relationship between X and Y is linear, a cubic regression model is excessively flexible, and we would expect the method to fit test data poorly. Therefore, we would expect the cubic model to have a higher test RSS

c)

We don't have enough information to know. It's possible that the cubic relationship could capture some of this non-linearity, but it depends on the nature of the relationship.

d)

In this case, we do not know the right amount of flexibility to fit the true underlying model. So there is not enough information to tell which model would give the lower test RSS

Applied

**Exercise 4:**

> lm.fit<-lm(mpg~horsepower,data=Auto)

> summary(lm.fit)

Call:

lm(formula = mpg ~ horsepower, data = Auto)

Residuals:

Min 1Q Median 3Q Max

-13.5710 -3.2592 -0.3435 2.7630 16.9240

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*

horsepower -0.157845 0.006446 -24.49 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.906 on 390 degrees of freedom

Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049

F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

1. p-value < 2.2e-16 -> p-value is almost ZERO -> there is a relationship between the predictor and the response.
2. How strong is the relationship between the predictor and the response?

The R-Squared value is almost 0.606 -> 60.6% of the variation in the response variable mpg is due to the predictor variable horsepower.

1. Is the relationship between the predictor and the response positive or negative?

Yes, Negative. As the coeficient of "horsepower" is negative: -0.158, so the relationship is also negative.

1. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

-predicted mpg when horsepower is 98: 24.467

-The 95% confidence and prediction intervals:

> predict(lm.fit,data.frame(horsepower=c(98)),interval="prediction")

fit lwr upr

1 24.46708 14.8094 34.12476

> predict(lm.fit,data.frame(horsepower=c(98)),interval="confidence")

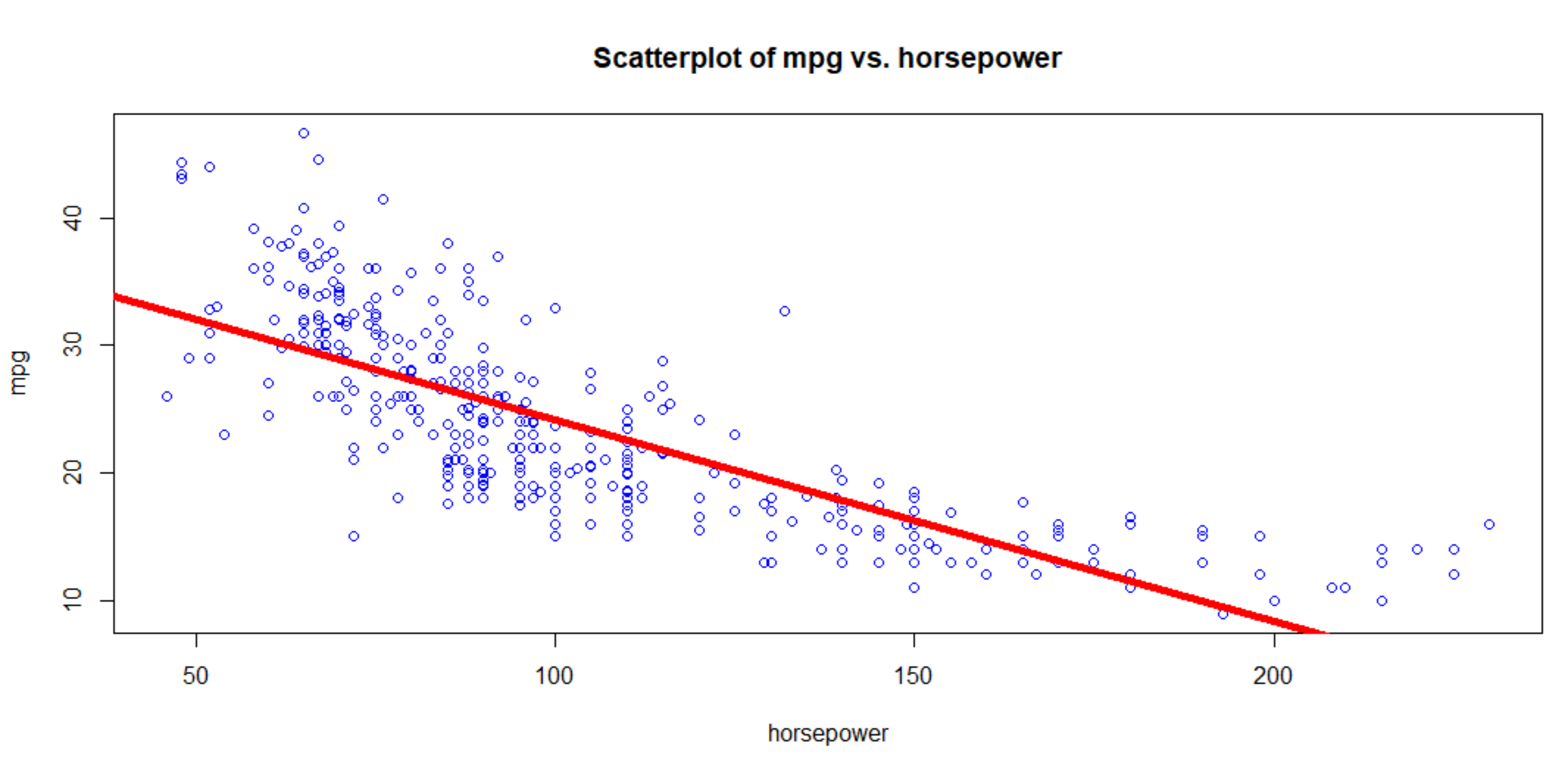
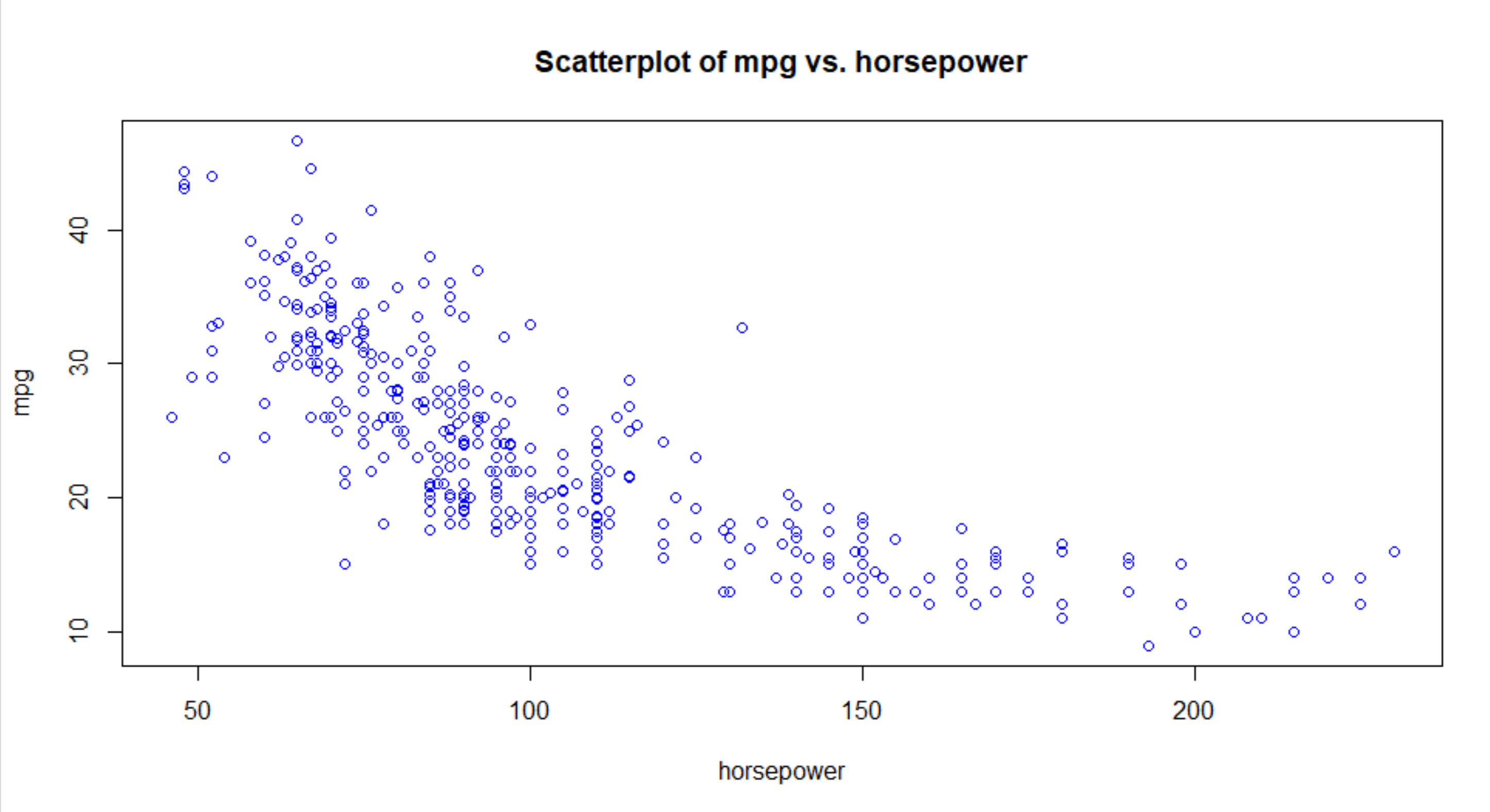
fit lwr upr

1 24.46708 23.97308 24.96108

1. Plot the response and the predictor. Use the abline() function to display the least squares regression line.

plot(Auto$horsepower, Auto$mpg, main = "Scatterplot of mpg vs. horsepower", xlab = "horsepower", ylab = "mpg", col = "blue")

abline(lm.fit,lwd=5,col="red")



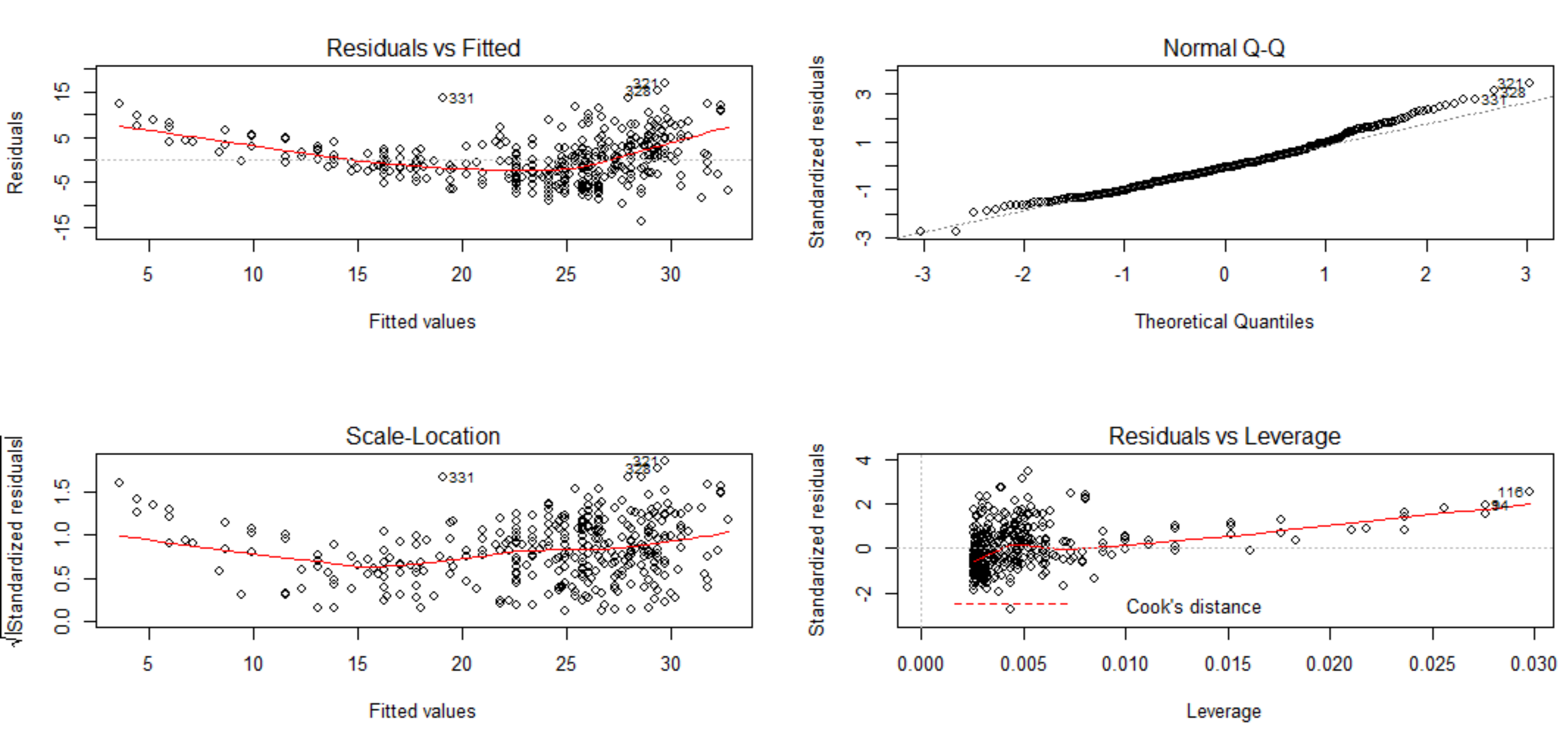
1. plot

> which.max(hatvalues(lm.fit))

116

par(mfrow = c(2,2))

plot(lm.fit)



The plot of Residuals vs Fitted indicates the presence of non-linear values in the data.

The plot of Residuals vs Leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and a few high leverage points. Most standardized residuals gather within 0.075 Leverage.

**Exercise 5:**

1. Produce a scatterplot matrix which includes all of the numerical variables in the data set. Exclude all the qualitative variables (Hint: use function str(data) to determine which variables of data frame data are numerical and which are factors).

> str(Carseats)

'data.frame': 400 obs. of 11 variables:

$ Sales : num 9.5 11.22 10.06 7.4 4.15 ...

$ CompPrice : num 138 111 113 117 141 124 115 136 132 132 ...

$ Income : num 73 48 35 100 64 113 105 81 110 113 ...

$ Advertising: num 11 16 10 4 3 13 0 15 0 0 ...

$ Population : num 276 260 269 466 340 501 45 425 108 131 ...

$ Price : num 120 83 80 97 128 72 108 120 124 124 ...

$ ShelveLoc : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...

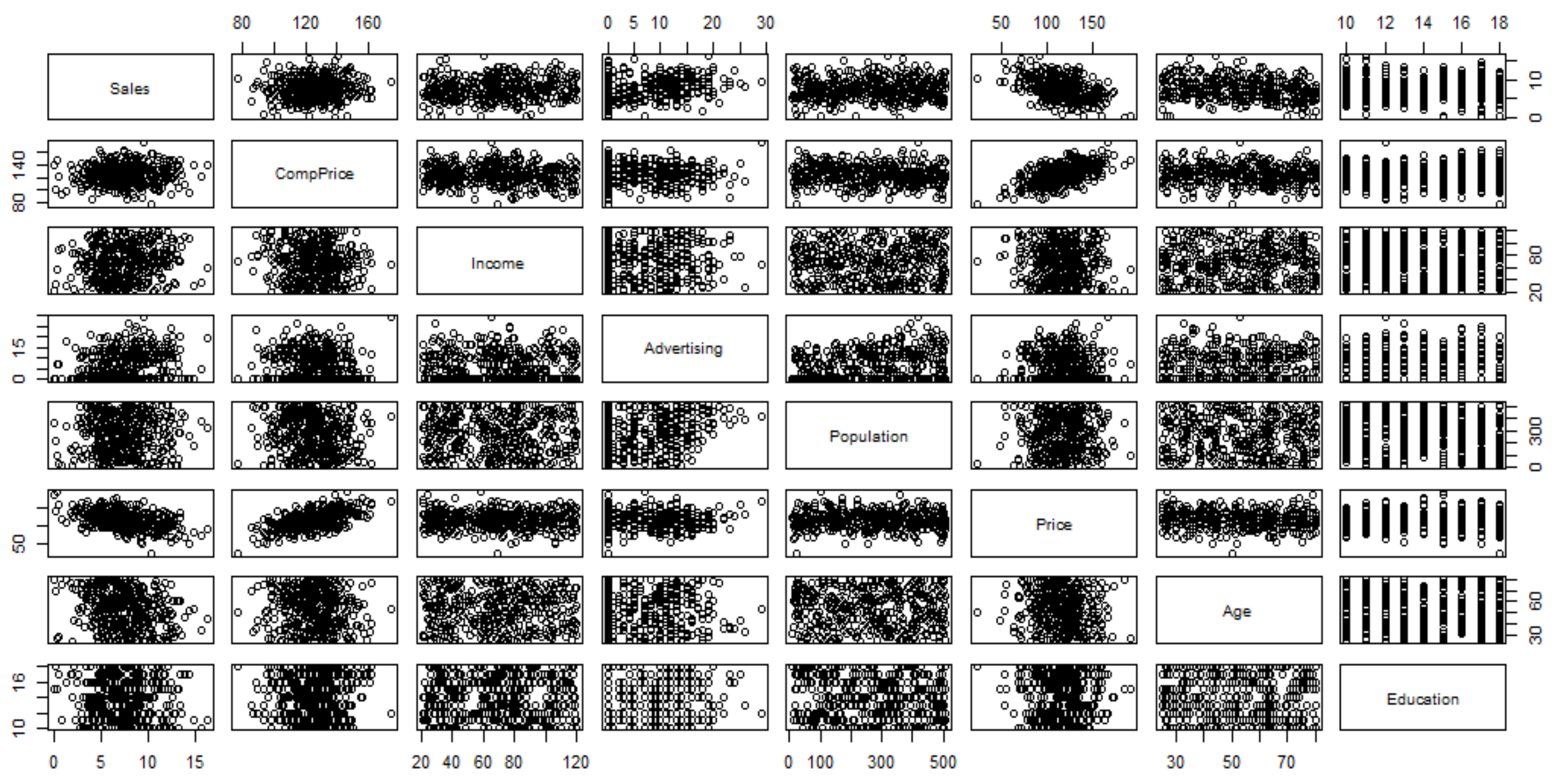
$ Age : num 42 65 59 55 38 78 71 67 76 76 ...

$ Education : num 17 10 12 14 13 16 15 10 10 17 ...

$ Urban : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...

$ US : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...

pairs(Carseats[,c(1,2,3,4,5,6,8,9)])



> cor(Carseats[,c(1,2,3,4,5,6,8,9)])

Sales CompPrice Income Advertising Population

Sales 1.00000000 0.06407873 0.151950979 0.269506781 0.050470984

CompPrice 0.06407873 1.00000000 -0.080653423 -0.024198788 -0.094706516

Income 0.15195098 -0.08065342 1.000000000 0.058994706 -0.007876994

Advertising 0.26950678 -0.02419879 0.058994706 1.000000000 0.265652145

Population 0.05047098 -0.09470652 -0.007876994 0.265652145 1.000000000

Price -0.44495073 0.58484777 -0.056698202 0.044536874 -0.012143620

Age -0.23181544 -0.10023882 -0.004670094 -0.004557497 -0.042663355

Education -0.05195524 0.02519705 -0.056855422 -0.033594307 -0.106378231

Price Age Education

Sales -0.44495073 -0.231815440 -0.051955242

CompPrice 0.58484777 -0.100238817 0.025197050

Income -0.05669820 -0.004670094 -0.056855422

Advertising 0.04453687 -0.004557497 -0.033594307

Population -0.01214362 -0.042663355 -0.106378231

Price 1.00000000 -0.102176839 0.011746599

Age -0.10217684 1.000000000 0.006488032

Education 0.01174660 0.006488032 1.000000000

> lm.fit2<-lm(Sales~CompPrice + Income + Advertising + Population + Price + Age + Education,data=Carseats)

> summary(lm.fit2)

Call:

lm(formula = Sales ~ CompPrice + Income + Advertising + Population +

Price + Age + Education, data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-5.0598 -1.3515 -0.1739 1.1331 4.8304

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.7076934 1.1176260 6.896 2.15e-11 \*\*\*

CompPrice 0.0939149 0.0078395 11.980 < 2e-16 \*\*\*

Income 0.0128717 0.0034757 3.703 0.000243 \*\*\*

Advertising 0.1308637 0.0151219 8.654 < 2e-16 \*\*\*

Population -0.0001239 0.0006877 -0.180 0.857092

Price -0.0925226 0.0050521 -18.314 < 2e-16 \*\*\*

Age -0.0449743 0.0060083 -7.485 4.75e-13 \*\*\*

Education -0.0399844 0.0371257 -1.077 0.282142

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.929 on 392 degrees of freedom

Multiple R-squared: 0.5417, Adjusted R-squared: 0.5335

F-statistic: 66.18 on 7 and 392 DF, p-value: < 2.2e-16

-Is there a relationship between the predictors and the response? Yes, there are. R-quared = 0.54 explains that 54% change in the response can be explained by the predictors in this regression model. Both Population and Age may have negative relationship with Sales: the co-efficient values < 0. The predictors Age and Education have relationship with the response Sales because their P-values are not significant. Meanwhile P-values of other numerical variables are too significant (<0.05)

-Predictors with statistically significant relationships with response Sales are : CompPrice, Income, Advertising, Price, and Age.

-Price and CompPrice:

CompPrice 0.0939149 0.0078395 11.980 < 2e-16 \*\*\*

Price -0.0925226 0.0050521 -18.314 < 2e-16 \*\*\*

correlation: sales~ price is -0.45; sales~Comprice is 0.064

Both Price and CompPrice are significant predictors toward the response Sales due to the small P-values.

Price and CompPrice are continuous variables. When all other predictors remain as constants, each difference in Price or CompPrice predict the value of Sales.

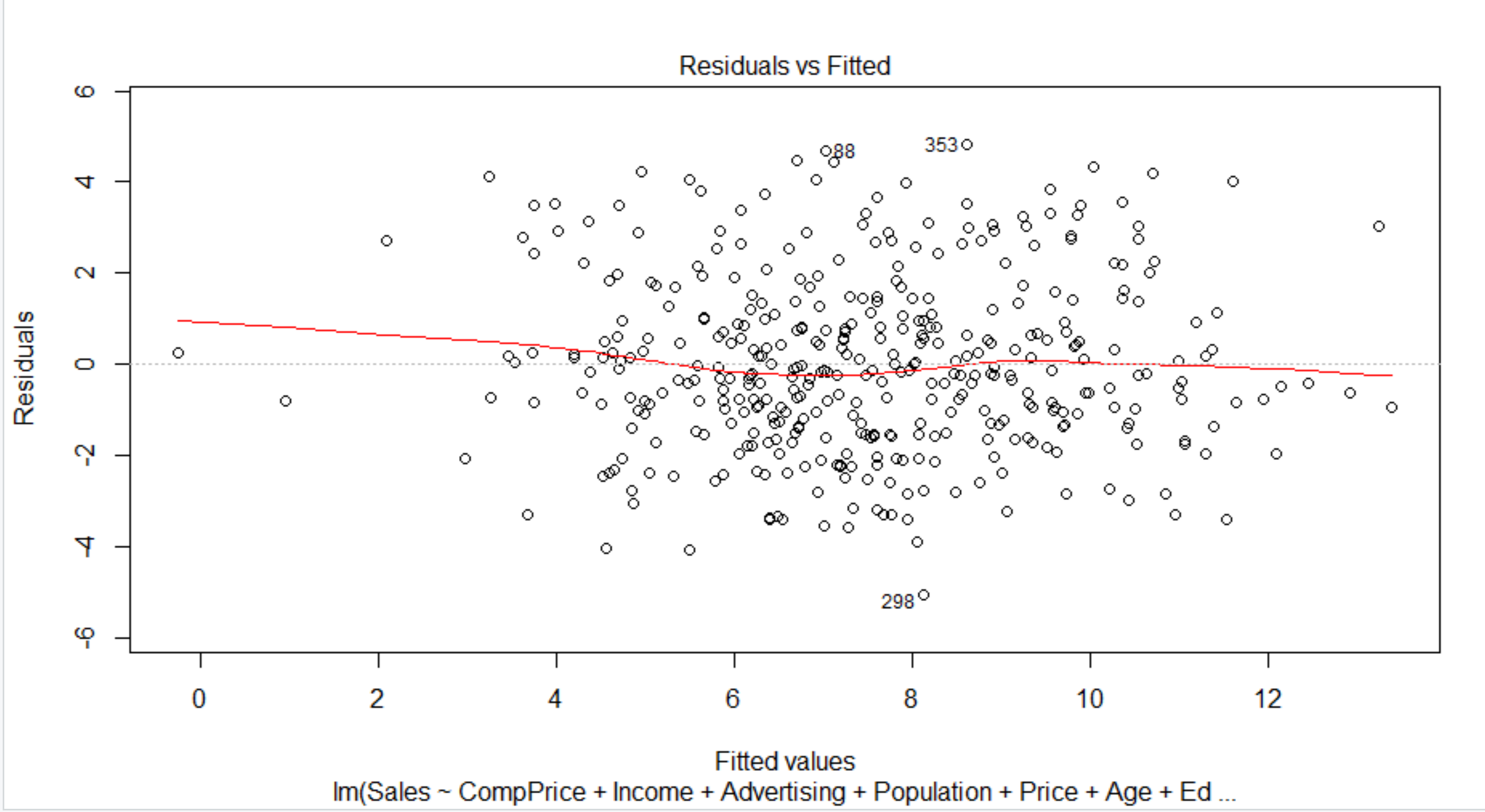
Price has a negative relationship with Sales.

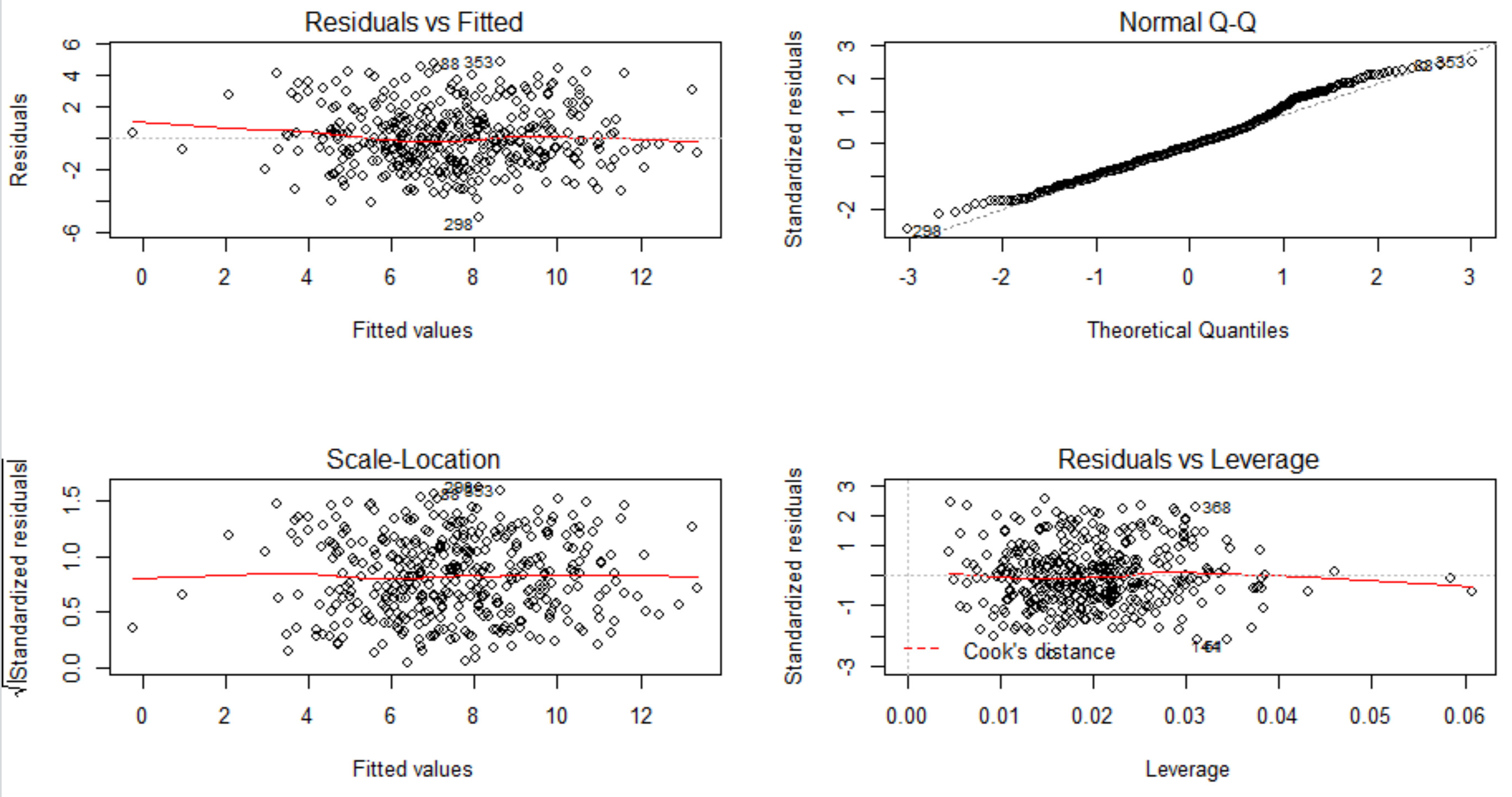
CompPrice has a positive relationship with Sales.

> plot(lm.fit2)

par(mfrow = c(2,2))

plot(lm.fit2)





The first graph shows that there is a non-linear relationship between the responce and the predictors.

The second graph shows that the residuals are normally distributed and right skewed.

The Third graphs shows that there are no leverage points. However, there on observation that stands out as a potential leverage point

**Exercise 6:**

1. Fit a multiple regression model to predict Sales using Price, Income, and US.

> x6 <- lm(Sales~Price + Income + US, data=Carseats)

> summary(x6)

Call:

lm(formula = Sales ~ Price + Income + US, data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-6.4956 -1.6166 0.0417 1.6041 7.1873

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.219893 0.706169 17.304 < 2e-16 \*\*\*

Price -0.053670 0.005206 -10.309 < 2e-16 \*\*\*

Income 0.011011 0.004415 2.494 0.013 \*

USYes 1.139703 0.257901 4.419 1.28e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.453 on 396 degrees of freedom

Multiple R-squared: 0.251, Adjusted R-squared: 0.2454

F-statistic: 44.24 on 3 and 396 DF, p-value: < 2.2e-16

b) Provide an interpretation of each coefficient in the model. Be careful - there is a quantitative variable in the model!

Price is continuous variable and its coefficient can be interpreted as average increase in sales for unit change in prices, keeping other variables constant. Price and Sales have a negative relationship.

Income is continuous variable and its coefficient can be interpreted as average increase in sales for unit change in incomes, keeping other variables constant. Income and Sales have a positive relationship.

US is a categorical variable with YES coded as 1 and NO as base 0. The coefficient can be interpreted as average sales more for 1 compared to 0 when other variables are kept constant.

c)

Sales = 12.2199 -0.054\*Price + 0.011\*Income + 1.1397 \*(1 if US is YES, 0 if US is NO)

=>Sales = 12.22 – 0.054\*Price + 0.011\* Income if US is NO

Sales = 12.22 – 0.054\*Price + 0.011\*Income + 1.14 if US is YES

d)

For which of the predictors can you reject the null hypothesis H0: \_j = 0:

i. at significance level B0 = 0:05?

ii. at significance level B0= 0:01?

At significance level B0=0.05, the null Hypothesis can be rejected for all three predictor Price , Income, and US.

At significance level B0 =0.01, the null Hypothesis is rejected on Price and US, not Income.

e)

x7 <- lm(Sales~Price + Income + US + Price:US + Income:US, data=Carseats)

> summary(x7)

Call:

lm(formula = Sales ~ Price + Income + US + Price:US + Income:US,

data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-6.5013 -1.6259 0.0394 1.6135 7.1929

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.2471022 1.0594708 11.560 < 2e-16 \*\*\*

Price -0.0541741 0.0081214 -6.671 8.65e-11 \*\*\*

Income 0.0114750 0.0074572 1.539 0.125

USYes 1.0908207 1.4150528 0.771 0.441

Price:USYes 0.0008223 0.0106128 0.077 0.938

Income:USYes -0.0006829 0.0092760 -0.074 0.941

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.46 on 394 degrees of freedom

Multiple R-squared: 0.2511, Adjusted R-squared: 0.2415

F-statistic: 26.41 on 5 and 394 DF, p-value: < 2.2e-16

After adding extension, we can see that there is no statistically significant interaction.

1. We use the Residual Standard Error to compare between two models in a) and e): the model with lower Residual Standard Error is fitter.

-Residual Standard Error in a : 2.453

-Residual Standard Error in e: 2.46

->Model a has lower Residual Standard Error -> **model in A is fitter.**

**Exercise 7:**

a) This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

> boston.zn<-lm(crim~zn,data=Boston)

> summary(boston.zn)

Call:

lm(formula = crim ~ zn, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-4.429 -4.222 -2.620 1.250 84.523

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.45369 0.41722 10.675 < 2e-16 \*\*\*

zn -0.07393 0.01609 -4.594 5.51e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.435 on 504 degrees of freedom

Multiple R-squared: 0.04019, Adjusted R-squared: 0.03828

F-statistic: 21.1 on 1 and 504 DF, p-value: 5.506e-06

> boston.indus<-lm(crim~indus,data=Boston)

> summary(boston.indus)

Call:

lm(formula = crim ~ indus, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-11.972 -2.698 -0.736 0.712 81.813

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.06374 0.66723 -3.093 0.00209 \*\*

indus 0.50978 0.05102 9.991 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.866 on 504 degrees of freedom

Multiple R-squared: 0.1653, Adjusted R-squared: 0.1637

F-statistic: 99.82 on 1 and 504 DF, p-value: < 2.2e-16

> boston.chas<-lm(crim~chas,data=Boston)

> summary(boston.chas)

Call:

lm(formula = crim ~ chas, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-3.738 -3.661 -3.435 0.018 85.232

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.7444 0.3961 9.453 <2e-16 \*\*\*

chas -1.8928 1.5061 -1.257 0.209

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.597 on 504 degrees of freedom

Multiple R-squared: 0.003124, Adjusted R-squared: 0.001146

F-statistic: 1.579 on 1 and 504 DF, p-value: 0.2094

> boston.nox<-lm(crim~nox,data=Boston)

> summary(boston.nox)

Call:

lm(formula = crim ~ nox, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-12.371 -2.738 -0.974 0.559 81.728

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -13.720 1.699 -8.073 5.08e-15 \*\*\*

nox 31.249 2.999 10.419 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.81 on 504 degrees of freedom

Multiple R-squared: 0.1772, Adjusted R-squared: 0.1756

F-statistic: 108.6 on 1 and 504 DF, p-value: < 2.2e-16

> boston.rm<-lm(crim~rm,data=Boston)

> summary(boston.rm)

Call:

lm(formula = crim ~ rm, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-6.604 -3.952 -2.654 0.989 87.197

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 20.482 3.365 6.088 2.27e-09 \*\*\*

rm -2.684 0.532 -5.045 6.35e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.401 on 504 degrees of freedom

Multiple R-squared: 0.04807, Adjusted R-squared: 0.04618

F-statistic: 25.45 on 1 and 504 DF, p-value: 6.347e-07

> boston.age<-lm(crim~age,data=Boston)

> summary(boston.age)

Call:

lm(formula = crim ~ age, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-6.789 -4.257 -1.230 1.527 82.849

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.77791 0.94398 -4.002 7.22e-05 \*\*\*

age 0.10779 0.01274 8.463 2.85e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.057 on 504 degrees of freedom

Multiple R-squared: 0.1244, Adjusted R-squared: 0.1227

F-statistic: 71.62 on 1 and 504 DF, p-value: 2.855e-16

> boston.dis<-lm(crim~dis,data=Boston)

> summary(boston.dis)

Call:

lm(formula = crim ~ dis, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-6.708 -4.134 -1.527 1.516 81.674

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.4993 0.7304 13.006 <2e-16 \*\*\*

dis -1.5509 0.1683 -9.213 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.965 on 504 degrees of freedom

Multiple R-squared: 0.1441, Adjusted R-squared: 0.1425

F-statistic: 84.89 on 1 and 504 DF, p-value: < 2.2e-16

> boston.rad<-lm(crim~rad,data=Boston)

> summary(boston.rad)

Call:

lm(formula = crim ~ rad, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-10.164 -1.381 -0.141 0.660 76.433

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.28716 0.44348 -5.157 3.61e-07 \*\*\*

rad 0.61791 0.03433 17.998 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.718 on 504 degrees of freedom

Multiple R-squared: 0.3913, Adjusted R-squared: 0.39

F-statistic: 323.9 on 1 and 504 DF, p-value: < 2.2e-16

> boston.tax<-lm(crim~tax,data=Boston)

> summary(boston.tax)

Call:

lm(formula = crim ~ tax, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-12.513 -2.738 -0.194 1.065 77.696

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -8.528369 0.815809 -10.45 <2e-16 \*\*\*

tax 0.029742 0.001847 16.10 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.997 on 504 degrees of freedom

Multiple R-squared: 0.3396, Adjusted R-squared: 0.3383

F-statistic: 259.2 on 1 and 504 DF, p-value: < 2.2e-16

> boston.ptratio<-lm(crim~ptratio,data=Boston)

> summary(boston.ptratio)

Call:

lm(formula = crim ~ ptratio, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-7.654 -3.985 -1.912 1.825 83.353

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -17.6469 3.1473 -5.607 3.40e-08 \*\*\*

ptratio 1.1520 0.1694 6.801 2.94e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.24 on 504 degrees of freedom

Multiple R-squared: 0.08407, Adjusted R-squared: 0.08225

F-statistic: 46.26 on 1 and 504 DF, p-value: 2.943e-11

> boston.black<-lm(crim~black,data=Boston)

> summary(boston.black)

Call:

lm(formula = crim ~ black, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-13.756 -2.299 -2.095 -1.296 86.822

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.553529 1.425903 11.609 <2e-16 \*\*\*

black -0.036280 0.003873 -9.367 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.946 on 504 degrees of freedom

Multiple R-squared: 0.1483, Adjusted R-squared: 0.1466

F-statistic: 87.74 on 1 and 504 DF, p-value: < 2.2e-16

> boston.lstat<-lm(crim~lstat,data=Boston)

> summary(boston.lstat)

Call:

lm(formula = crim ~ lstat, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-13.925 -2.822 -0.664 1.079 82.862

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.33054 0.69376 -4.801 2.09e-06 \*\*\*

lstat 0.54880 0.04776 11.491 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.664 on 504 degrees of freedom

Multiple R-squared: 0.2076, Adjusted R-squared: 0.206

F-statistic: 132 on 1 and 504 DF, p-value: < 2.2e-16

> boston.medv<-lm(crim~medv,data=Boston)

> summary(boston.medv)

Call:

lm(formula = crim ~ medv, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-9.071 -4.022 -2.343 1.298 80.957

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 11.79654 0.93419 12.63 <2e-16 \*\*\*

medv -0.36316 0.03839 -9.46 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.934 on 504 degrees of freedom

Multiple R-squared: 0.1508, Adjusted R-squared: 0.1491

F-statistic: 89.49 on 1 and 504 DF, p-value: < 2.2e-16

With p-value=0.2094, only the predictor **chas** variable is not significant in predicting the response (per capita crime rate). Based on the p-value of its t statistic we cannot reject the null hypothesis.

b)

> boston.all<-lm(crim~.,Boston)

> summary(boston.all)

Call:

lm(formula = crim ~ ., data = Boston)

Residuals:

Min 1Q Median 3Q Max

-9.924 -2.120 -0.353 1.019 75.051

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 17.033228 7.234903 2.354 0.018949 \*

zn 0.044855 0.018734 2.394 0.017025 \*

indus -0.063855 0.083407 -0.766 0.444294

chas -0.749134 1.180147 -0.635 0.525867

nox -10.313535 5.275536 -1.955 0.051152 .

rm 0.430131 0.612830 0.702 0.483089

age 0.001452 0.017925 0.081 0.935488

dis -0.987176 0.281817 -3.503 0.000502 \*\*\*

rad 0.588209 0.088049 6.680 6.46e-11 \*\*\*

tax -0.003780 0.005156 -0.733 0.463793

ptratio -0.271081 0.186450 -1.454 0.146611

black -0.007538 0.003673 -2.052 0.040702 \*

lstat 0.126211 0.075725 1.667 0.096208 .

medv -0.198887 0.060516 -3.287 0.001087 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom

Multiple R-squared: 0.454, Adjusted R-squared: 0.4396

F-statistic: 31.47 on 13 and 492 DF, p-value: < 2.2e-16

From the summary we can say that the null hypothesis H0: Bj = 0 can be rejected for variables zn, dis, rad, black, medv because their p-value is less than 0.05.

c)

In a), we found on chas as not significant, but in b) we have indus, chas, nox, rm, age, tax, ptratio, and lstat. Even the coefficient values are different.

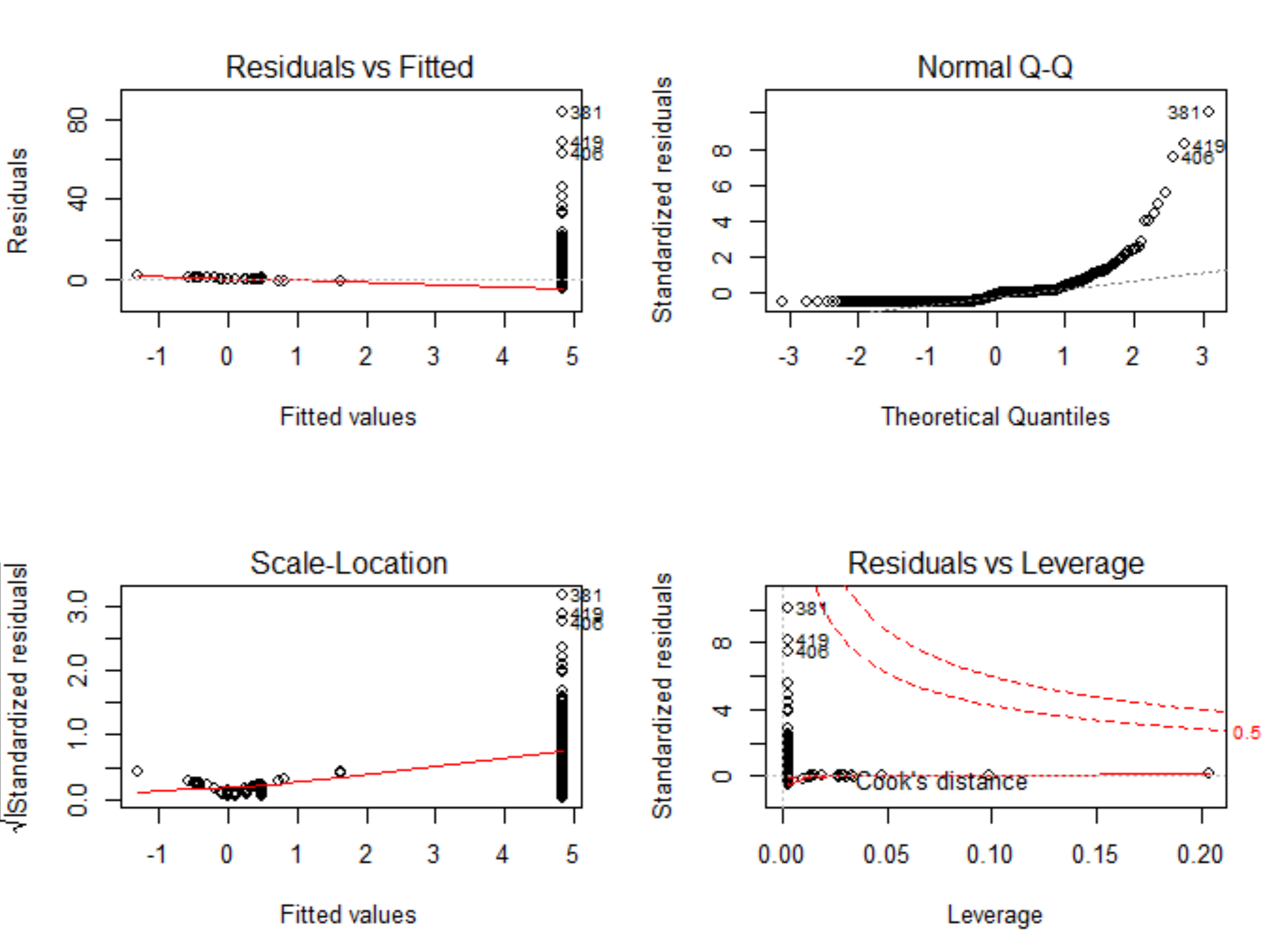
d)

boston.zn1<-lm(crim~poly(zn,3),data=Boston)

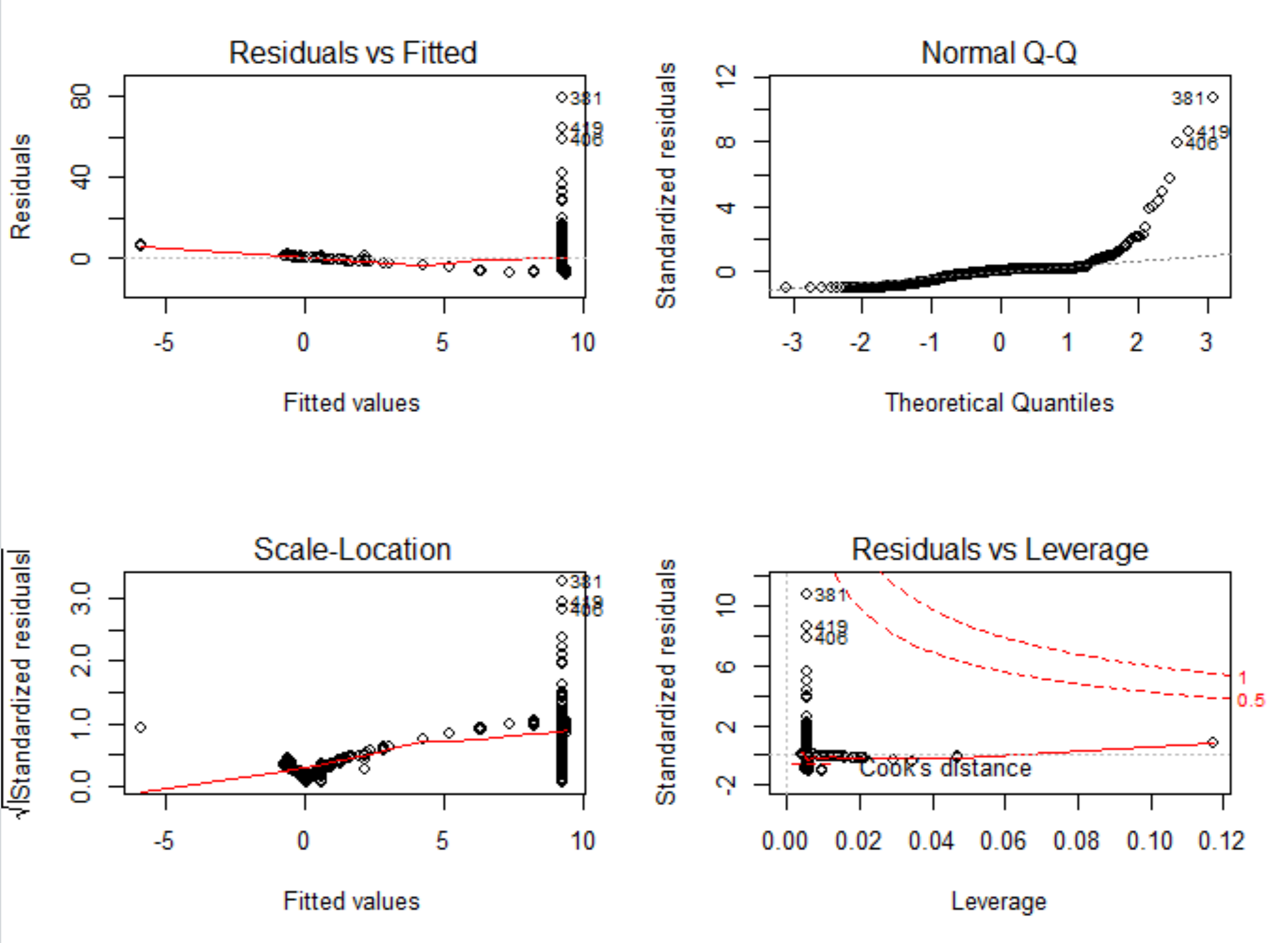
summary(boston.zn1)

par(mfrow=c(2,2))

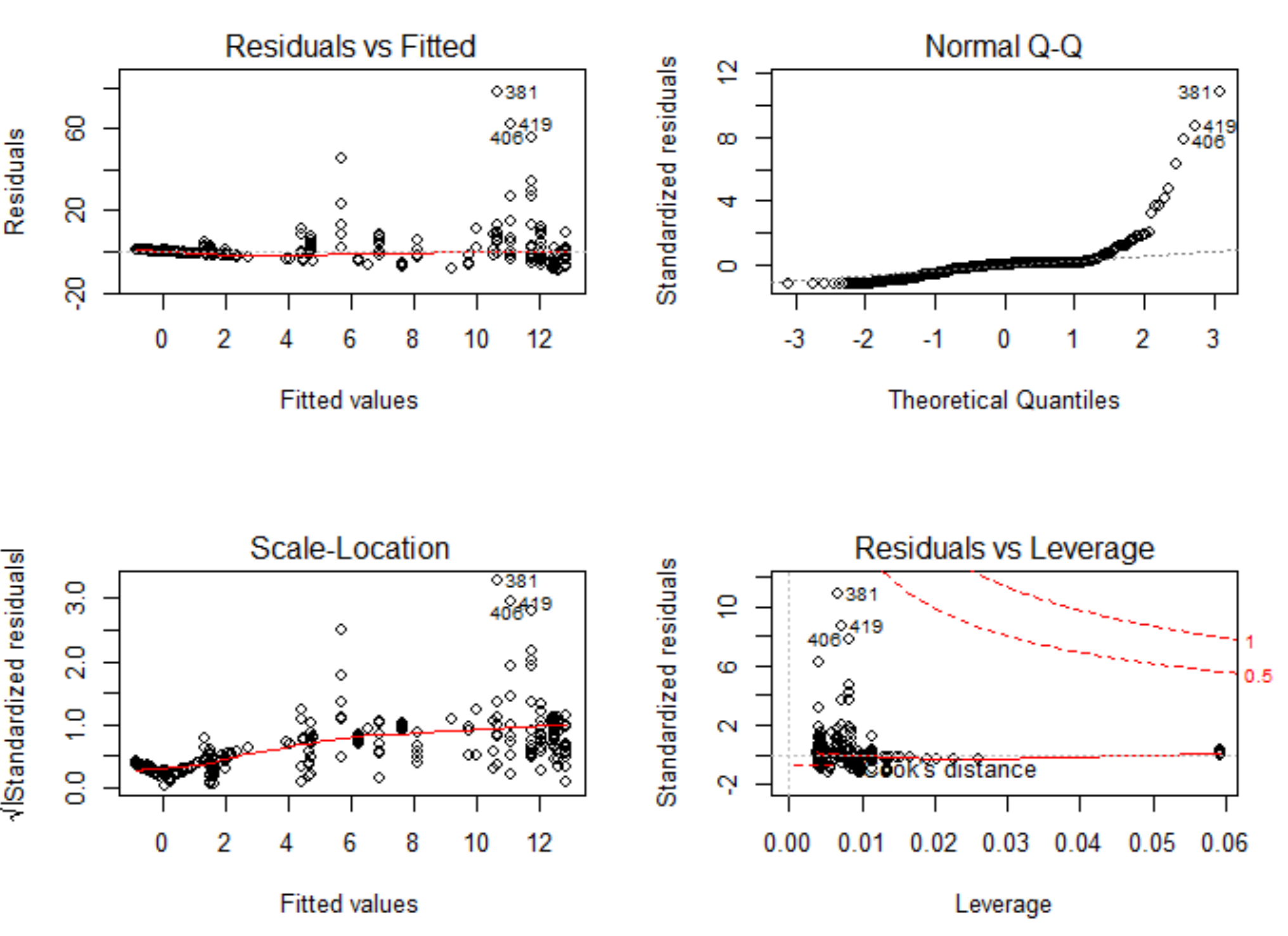
plot(boston.zn1)



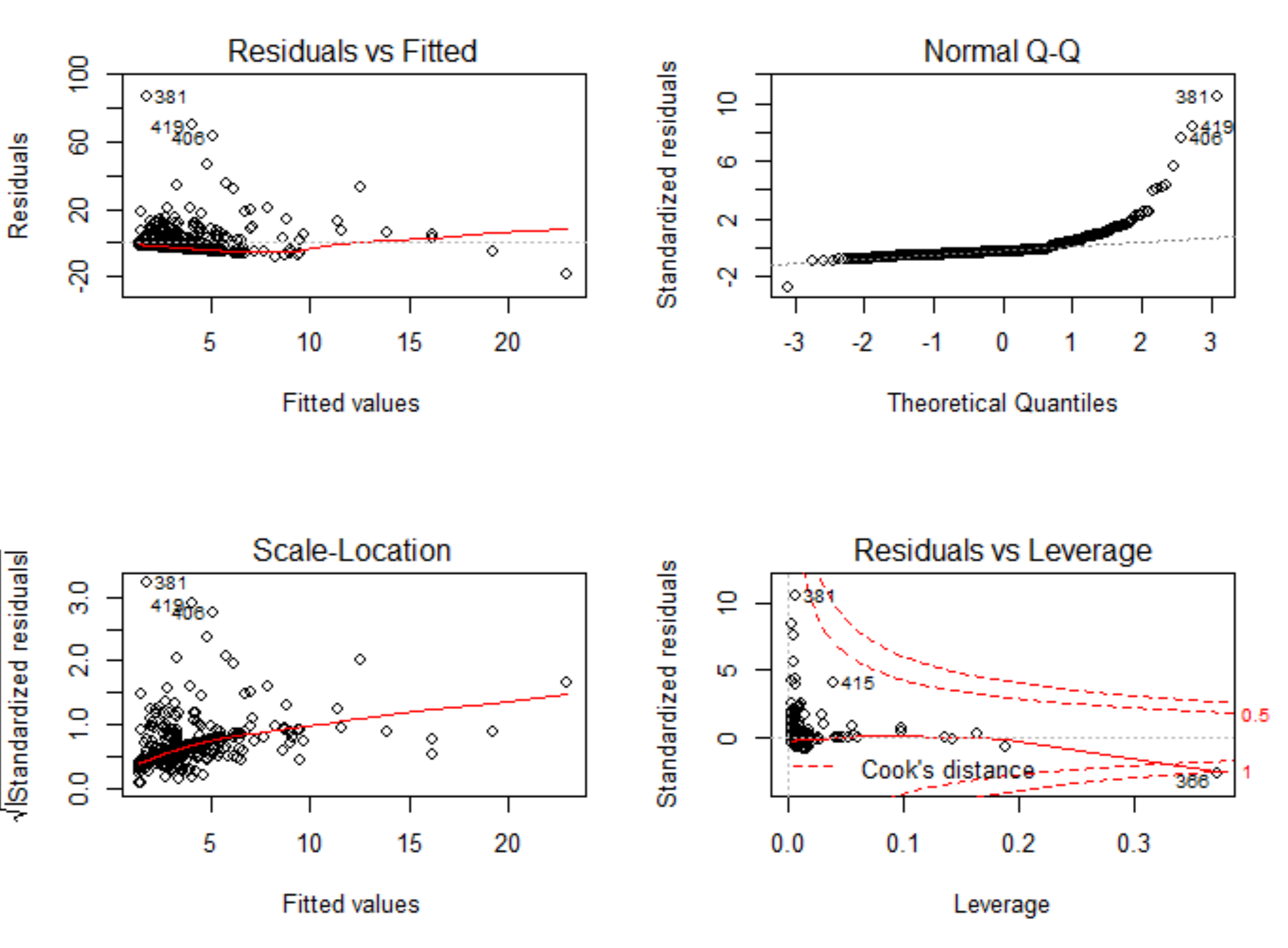
plot(boston.indus1)



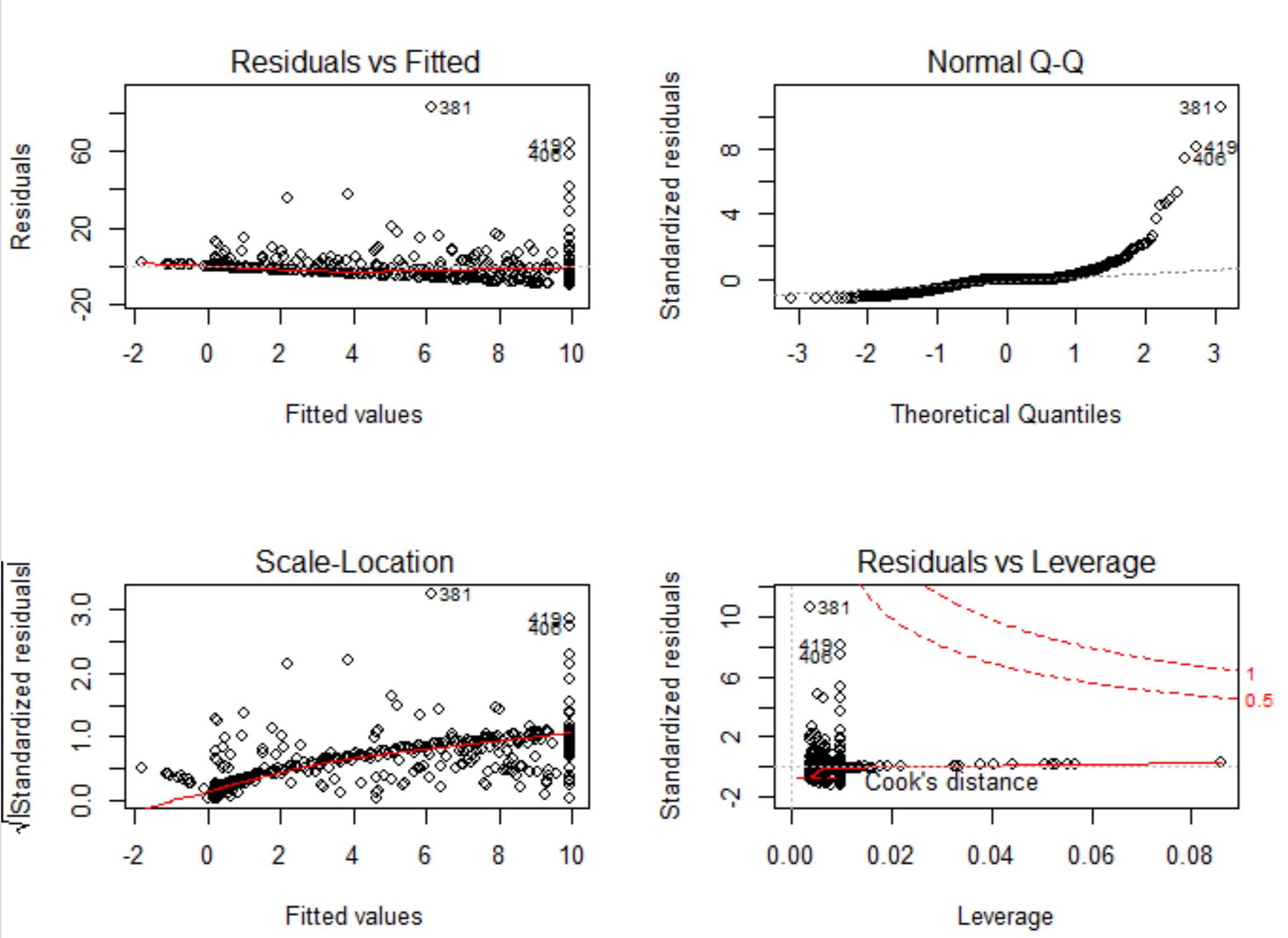
plot(boston.nox1)



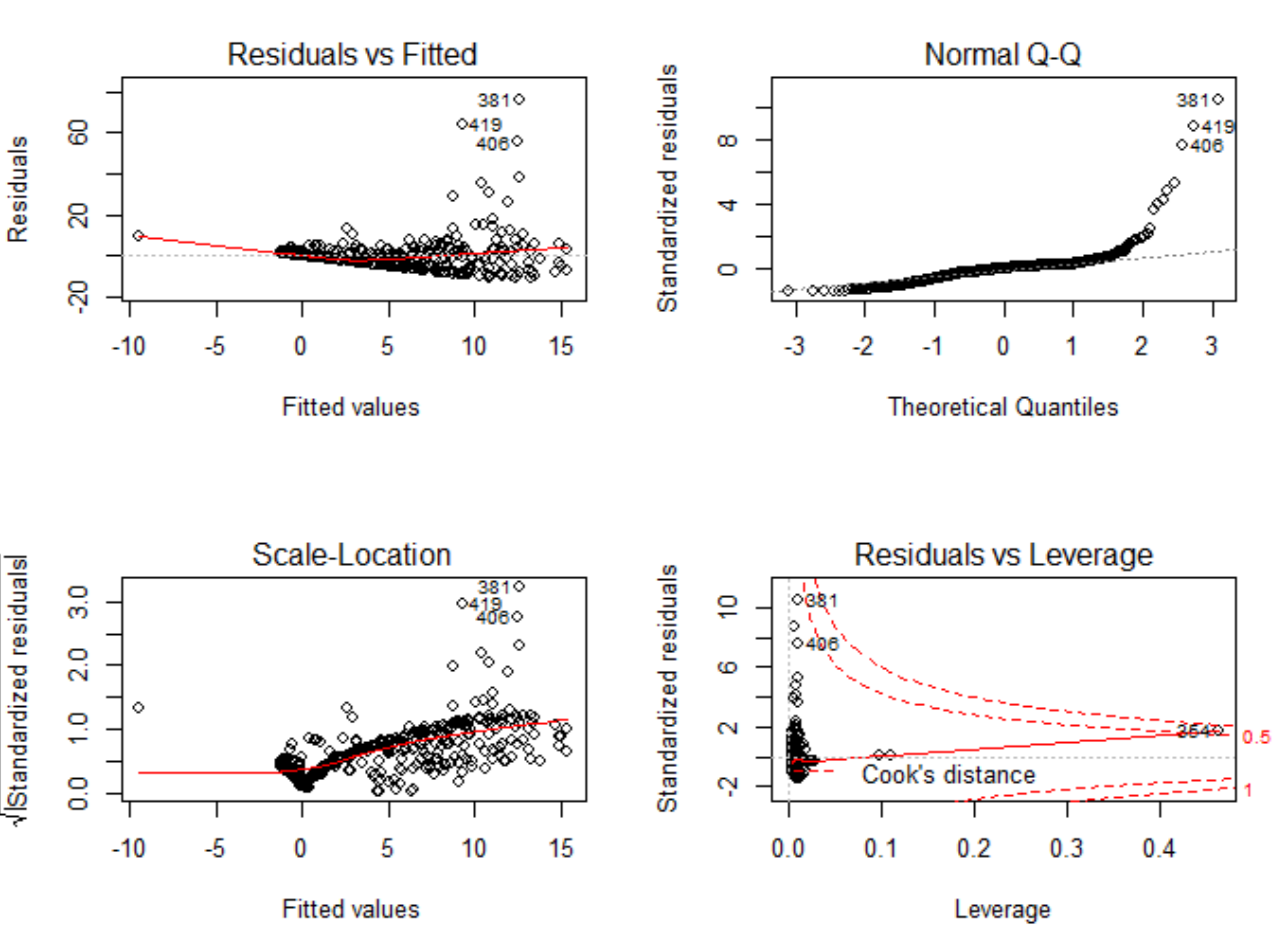
plot(boston.rm1)



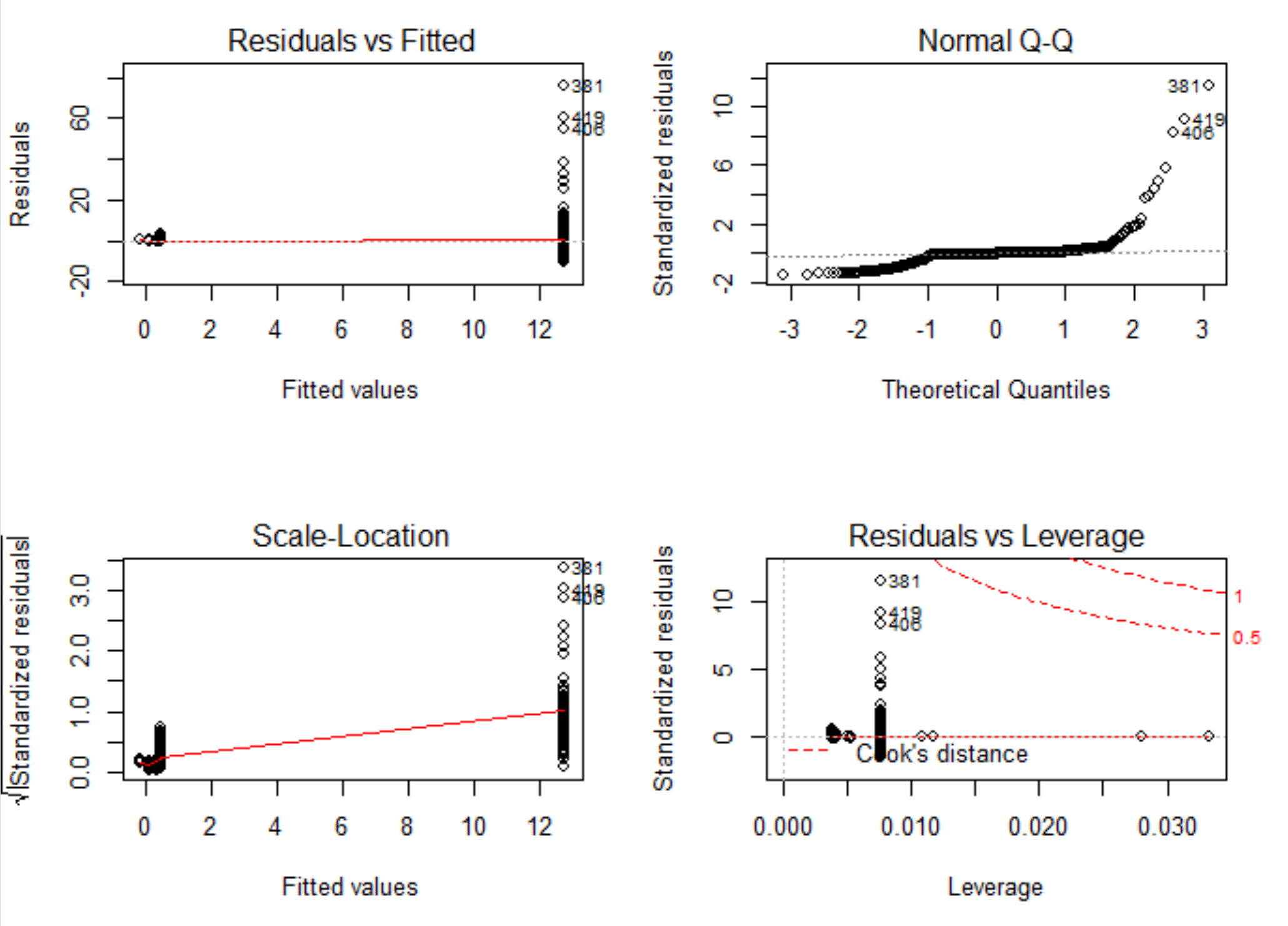
plot(boston.age1)



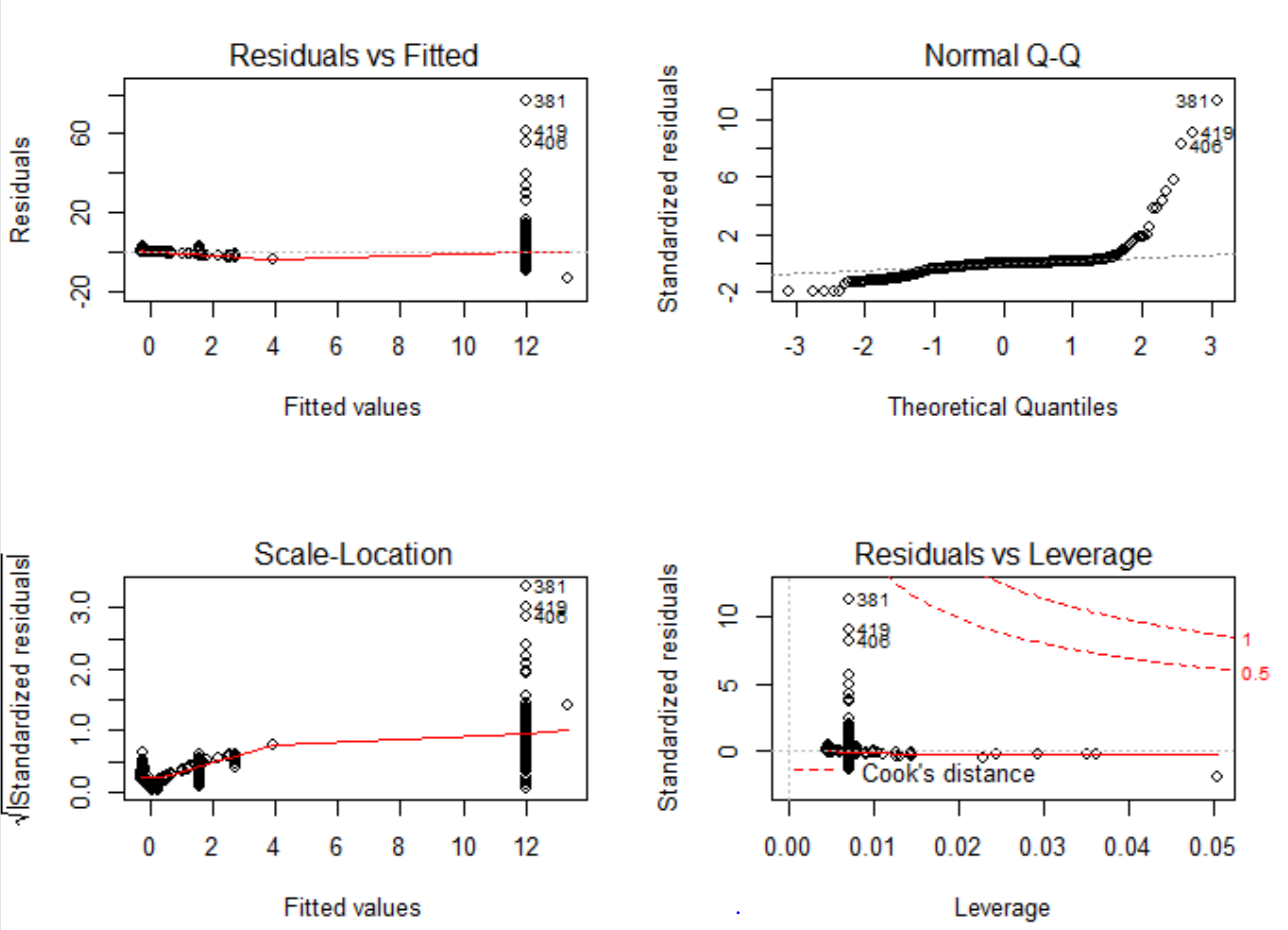
plot(boston.dis1)



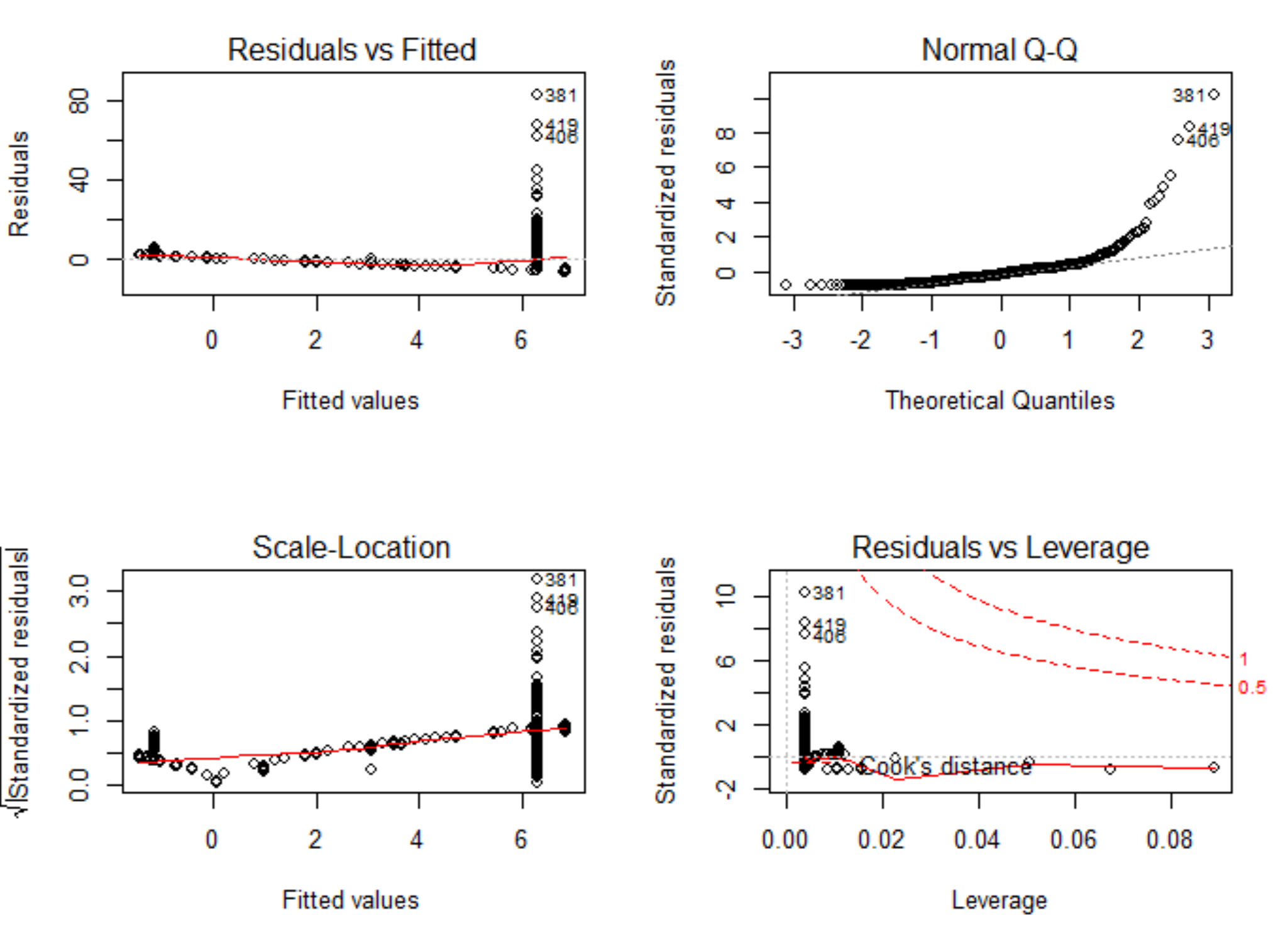
plot(boston.rad1)



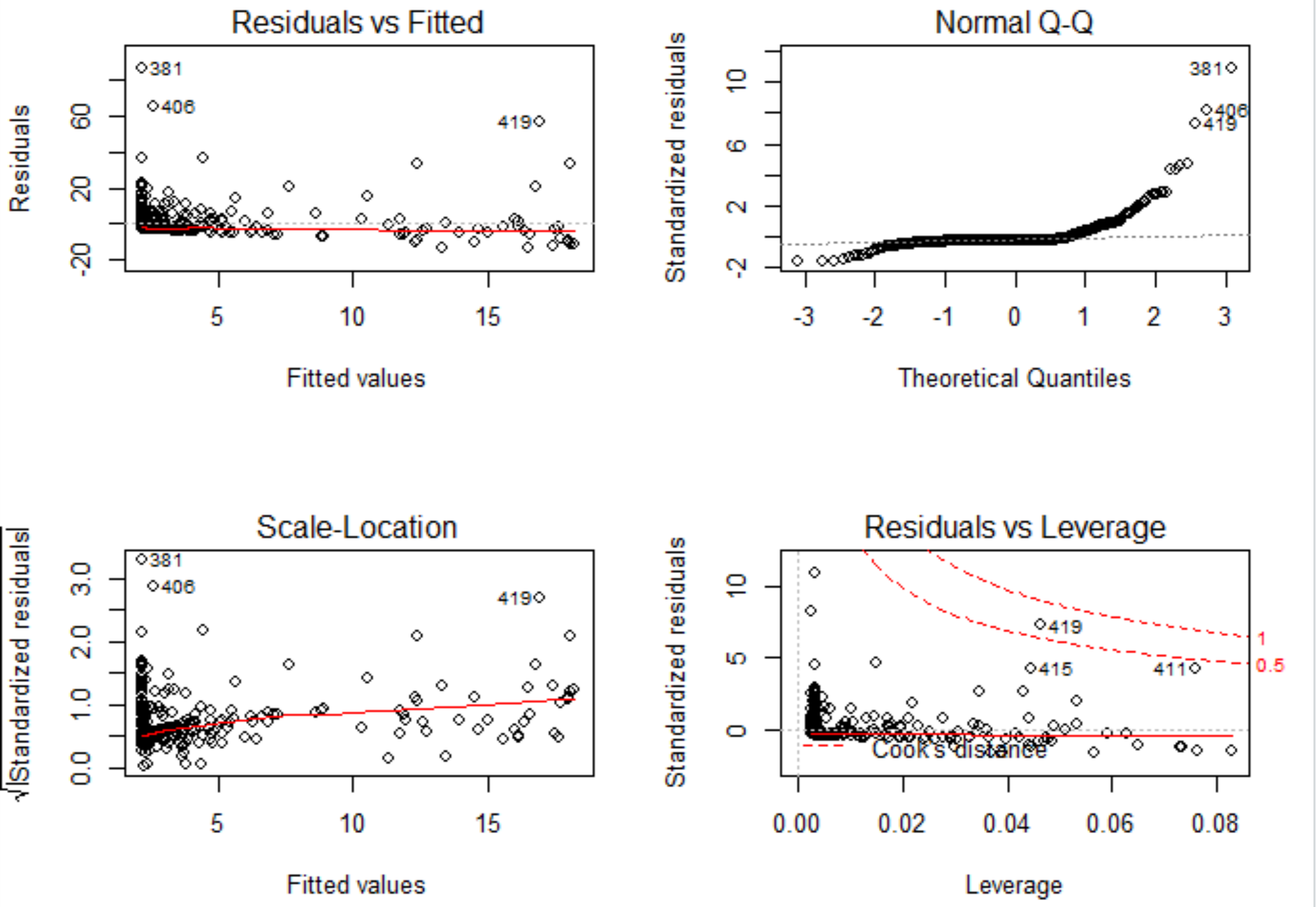
plot(boston.tax1)



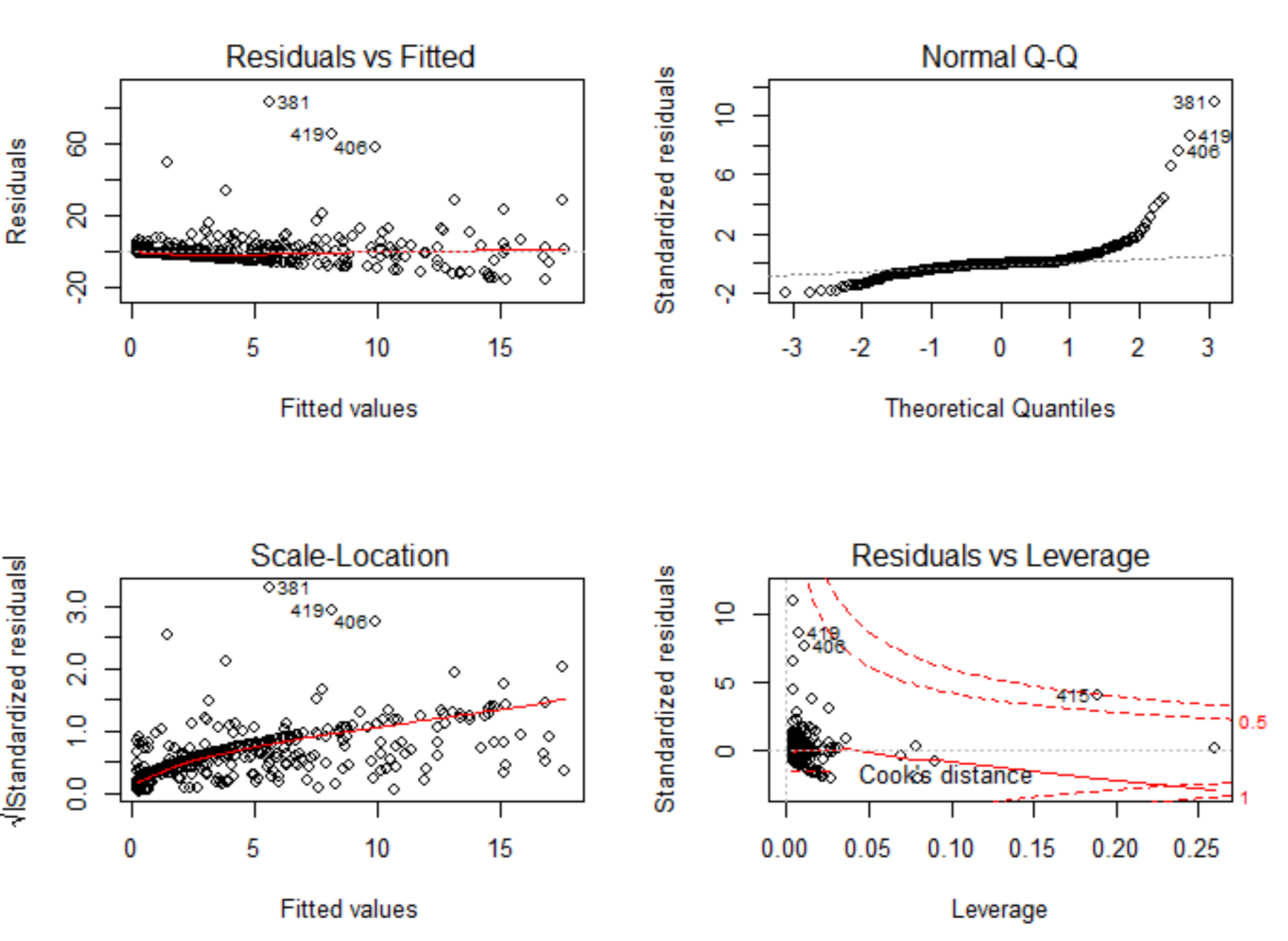
plot(boston.ptratio1)



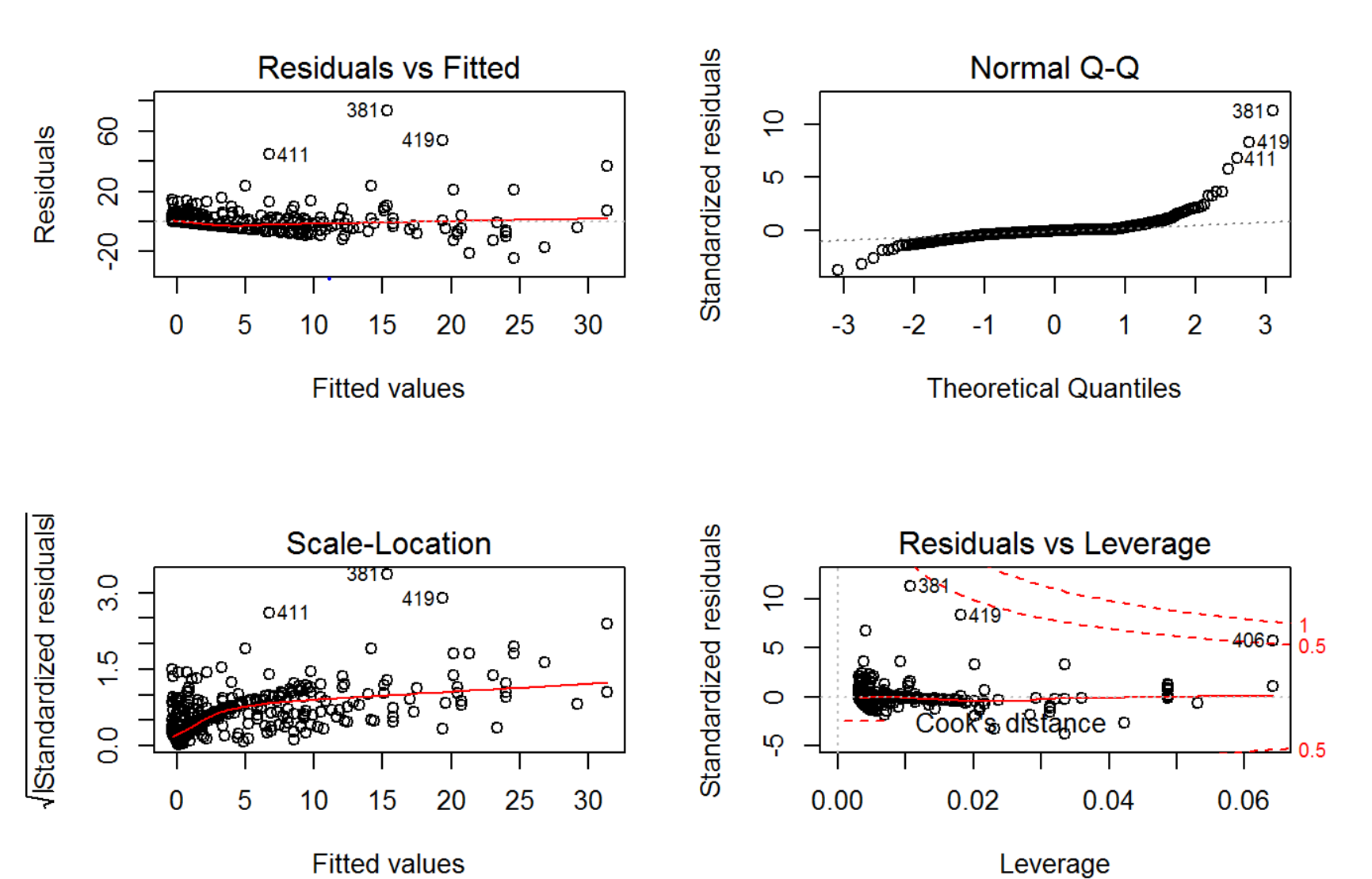
plot(boston.black1)



plot(boston.lstat1)



plot(boston.medv1)



Cubic relationship between predictor and response is significant for indus, nox, age, dis, ptratio, and medv variables, indicating non linear relationship.

For black variable neither cubic nor quadratic coefficient is significant ->suggesting **NO** non linear relationship visible.