IMT 573: Problem Set 6 - Regression

Neelam Purswani

Due: Tuesday, November 13, 2018

##### Collaborators: Dhaval Chheda

##### Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the problemset6.Rmd file from Canvas. Open problemset6.Rmd in RStudio and supply your solutions to the assignment by editing problemset6.Rmd.
2. Replace the “Insert Your Name Here” text in the author: field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. If you are using more than just a standard function that you found from another source, please credit the source in the comments. For example:
4. Collaboration on problem sets is acceptable, and even encouraged, but students must turn in an individual write-up in their own words and their own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students’ responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click Knit PDF or Knit Word, rename the R Markdown file to YourLastName\_YourFirstName\_ps6.Rmd, knit a PDF or DOC and submit both the PDF/DOC and the Rmd file on Canvas.

##### Setup:

In this problem set you will need, at minimum, the following R packages.

# Load standard libraries  
library(tidyverse)  
library(MASS) # Modern applied statistics functions  
library(mlbench)

In this problem we will use the Boston dataset that is available in the package. This dataset contains information about median house value for 506 neighborhoods in Boston, MA. Load this data and use it to answer the following questions.

### 1.

Describe the data and variables that are part of the dataset. Tidy data as necessary.

data(BostonHousing)  
housing <- BostonHousing  
str(housing)

## '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 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ 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 : num 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 ...  
## $ b : 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 ...

housing<-data.frame(housing)  
housing$chas <-as.numeric(housing$chas)

Housing data for 506 census tracts of Boston from the 1970 census. Boston housing data set is a collection of all the below mentioned properties: crim - per capita crime rate by town zn - proportion of residential land zoned for lots over 25,000 sq.ft indus - proportion of non-retail business acres per town chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) nox - nitric oxides concentration (parts per 10 million) rm - average number of rooms per dwelling age - proportion of owner-occupied units built prior to 1940 dis - weighted distances to five Boston employment centres rad - index of accessibility to radial highways tax - full-value property-tax rate per USD 10,000 ptratio- pupil-teacher ratio by town b 1000(B - 0.63)^2, where B is the proportion of blacks by town lstat - percentage of lower status of the population medv - median value of owner-occupied homes in USD 1000’s

The following section is to chek for NAs in the dataset:

var\_data <- table(is.na(housing))  
var\_data

##   
## FALSE   
## 7084

column\_wise <- sapply(housing, function(x) sum(is.na(housing)))  
column\_wise

## crim zn indus chas nox rm age dis rad   
## 0 0 0 0 0 0 0 0 0   
## tax ptratio b lstat medv   
## 0 0 0 0 0

#row\_wise<-rowSums(is.na(housing))  
#row\_wise

As there arent any NAs in the dataset, we do not need to tidy it. We might need to convert chas into numeric in order to plot correlation matrix. ### 2. Consider this data in context, what is the response variable of interest? Discuss how you think some of the possible predictor variables might be associated with this response. Response variable of interest: Median Value of owner occupied homes in USD 1000s I feel that median value of the house will be dependent on a combination of predictor variables such as weighted distances to five Boston employment centres,age, tax, average number of rooms per dwelling. case 1: The more the average rooms per dwelling the higher would be the median value of that house or the more the crime rate in a town the lower will be the median value of houses there.

### 3.

For each predictor, fit a simple linear regression model to predict the response. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

linear\_crim <- lm(medv ~ crim, data=housing)  
summary(linear\_crim)

##   
## Call:  
## lm(formula = medv ~ crim, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.957 -5.449 -2.007 2.512 29.800   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 24.03311 0.40914 58.74 <2e-16 \*\*\*  
## crim -0.41519 0.04389 -9.46 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.484 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

linear\_zn <-lm(medv ~ zn, data=housing)  
summary(linear\_zn)

##   
## Call:  
## lm(formula = medv ~ zn, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.918 -5.518 -1.006 2.757 29.082   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 20.91758 0.42474 49.248 <2e-16 \*\*\*  
## zn 0.14214 0.01638 8.675 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.587 on 504 degrees of freedom  
## Multiple R-squared: 0.1299, Adjusted R-squared: 0.1282   
## F-statistic: 75.26 on 1 and 504 DF, p-value: < 2.2e-16

linear\_indus <-lm(medv ~ indus, data=housing)  
summary(linear\_indus)

##   
## Call:  
## lm(formula = medv ~ indus, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.017 -4.917 -1.457 3.180 32.943   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 29.75490 0.68345 43.54 <2e-16 \*\*\*  
## indus -0.64849 0.05226 -12.41 <2e-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: 0.234, Adjusted R-squared: 0.2325   
## F-statistic: 154 on 1 and 504 DF, p-value: < 2.2e-16

linear\_chas <-lm(medv ~ chas, data=housing)  
summary(linear\_chas)

##   
## Call:  
## lm(formula = medv ~ chas, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.094 -5.894 -1.417 2.856 27.906   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.748 1.745 9.025 < 2e-16 \*\*\*  
## chas 6.346 1.588 3.996 7.39e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.064 on 504 degrees of freedom  
## Multiple R-squared: 0.03072, Adjusted R-squared: 0.02879   
## F-statistic: 15.97 on 1 and 504 DF, p-value: 7.391e-05

linear\_nox <-lm(medv ~ nox, data=housing)  
summary(linear\_nox)

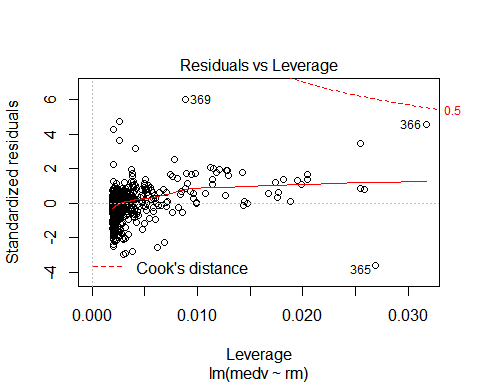
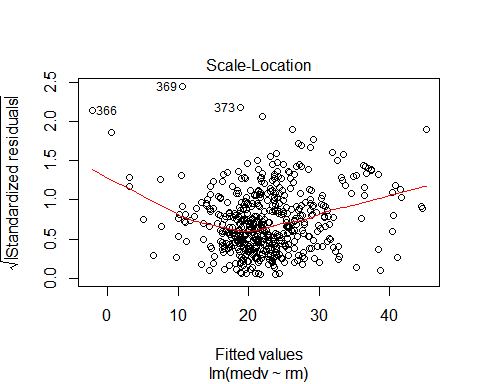
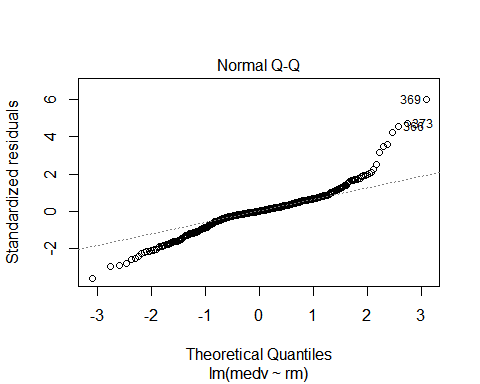
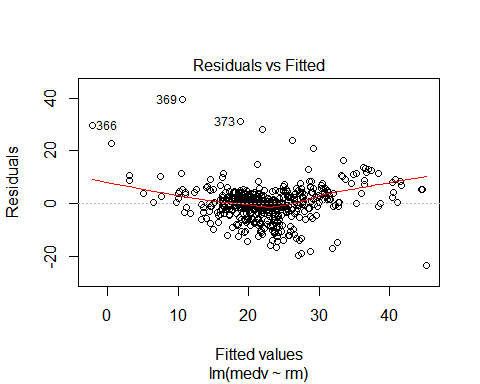
##   
## Call:  
## lm(formula = medv ~ nox, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.691 -5.121 -2.161 2.959 31.310   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 41.346 1.811 22.83 <2e-16 \*\*\*  
## nox -33.916 3.196 -10.61 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.323 on 504 degrees of freedom  
## Multiple R-squared: 0.1826, Adjusted R-squared: 0.181   
## F-statistic: 112.6 on 1 and 504 DF, p-value: < 2.2e-16

linear\_rm <-lm(medv ~ rm, data=housing)  
summary(linear\_rm)

##   
## Call:  
## lm(formula = medv ~ rm, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.346 -2.547 0.090 2.986 39.433   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -34.671 2.650 -13.08 <2e-16 \*\*\*  
## rm 9.102 0.419 21.72 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.616 on 504 degrees of freedom  
## Multiple R-squared: 0.4835, Adjusted R-squared: 0.4825   
## F-statistic: 471.8 on 1 and 504 DF, p-value: < 2.2e-16

Since the R squared value for rm is high, we can plot the model and observe the relationship:

plot(linear\_rm)



linear\_age <-lm(medv ~ age, data=housing)  
summary(linear\_age)

##   
## Call:  
## lm(formula = medv ~ age, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.097 -5.138 -1.958 2.397 31.338   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.97868 0.99911 31.006 <2e-16 \*\*\*  
## age -0.12316 0.01348 -9.137 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.527 on 504 degrees of freedom  
## Multiple R-squared: 0.1421, Adjusted R-squared: 0.1404   
## F-statistic: 83.48 on 1 and 504 DF, p-value: < 2.2e-16

linear\_dis <-lm(medv ~ dis, data=housing)  
summary(linear\_dis)

##   
## Call:  
## lm(formula = medv ~ dis, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.016 -5.556 -1.865 2.288 30.377   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 18.3901 0.8174 22.499 < 2e-16 \*\*\*  
## dis 1.0916 0.1884 5.795 1.21e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.914 on 504 degrees of freedom  
## Multiple R-squared: 0.06246, Adjusted R-squared: 0.0606   
## F-statistic: 33.58 on 1 and 504 DF, p-value: 1.207e-08

linear\_rad <-lm(medv ~ rad, data=housing)  
summary(linear\_rad)

##   
## Call:  
## lm(formula = medv ~ rad, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.770 -5.199 -1.967 3.321 33.292   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 26.38213 0.56176 46.964 <2e-16 \*\*\*  
## rad -0.40310 0.04349 -9.269 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.509 on 504 degrees of freedom  
## Multiple R-squared: 0.1456, Adjusted R-squared: 0.1439   
## F-statistic: 85.91 on 1 and 504 DF, p-value: < 2.2e-16

linear\_tax <-lm(medv ~ tax, data=housing)  
summary(linear\_tax)

##   
## Call:  
## lm(formula = medv ~ tax, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.091 -5.173 -2.085 3.158 34.058   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 32.970654 0.948296 34.77 <2e-16 \*\*\*  
## tax -0.025568 0.002147 -11.91 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.133 on 504 degrees of freedom  
## Multiple R-squared: 0.2195, Adjusted R-squared: 0.218   
## F-statistic: 141.8 on 1 and 504 DF, p-value: < 2.2e-16

linear\_ptratio <-lm(medv ~ ptratio, data=housing)  
summary(linear\_ptratio)

##   
## Call:  
## lm(formula = medv ~ ptratio, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.8342 -4.8262 -0.6426 3.1571 31.2303   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 62.345 3.029 20.58 <2e-16 \*\*\*  
## ptratio -2.157 0.163 -13.23 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.931 on 504 degrees of freedom  
## Multiple R-squared: 0.2578, Adjusted R-squared: 0.2564   
## F-statistic: 175.1 on 1 and 504 DF, p-value: < 2.2e-16

linear\_b <-lm(medv ~ b, data=housing)  
summary(linear\_b)

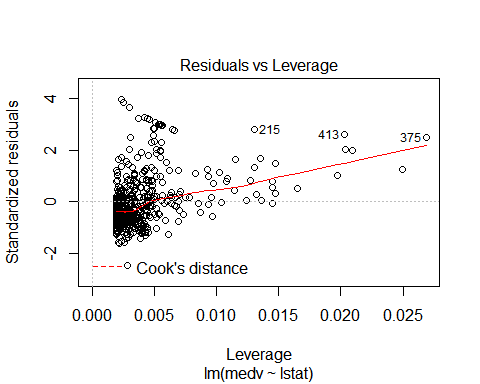
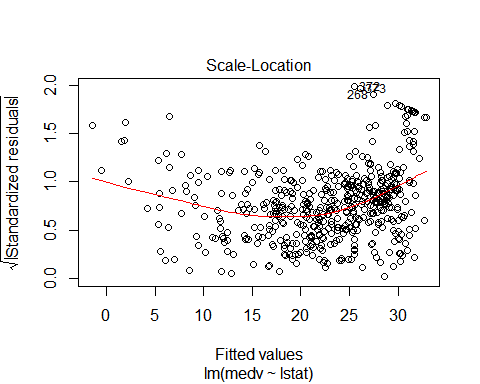
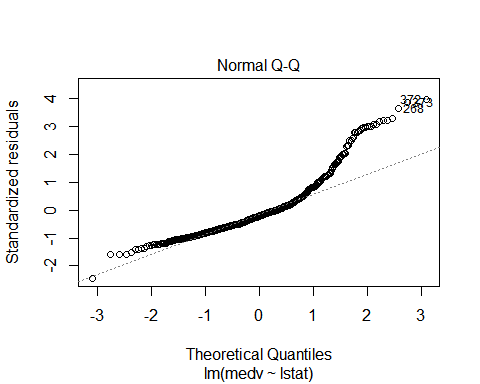
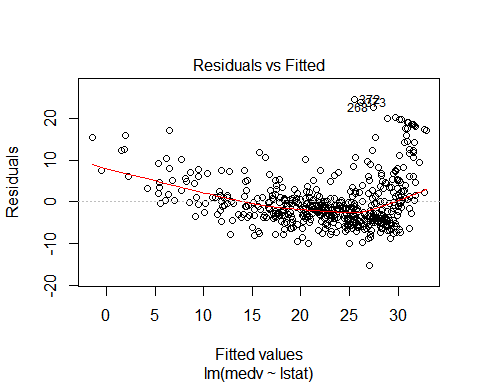
##   
## Call:  
## lm(formula = medv ~ b, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.884 -4.862 -1.684 2.932 27.763   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.551034 1.557463 6.775 3.49e-11 \*\*\*  
## b 0.033593 0.004231 7.941 1.32e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.679 on 504 degrees of freedom  
## Multiple R-squared: 0.1112, Adjusted R-squared: 0.1094   
## F-statistic: 63.05 on 1 and 504 DF, p-value: 1.318e-14

linear\_lstat <-lm(medv ~ lstat, data=housing)  
summary(linear\_lstat)

##   
## Call:  
## lm(formula = medv ~ lstat, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.168 -3.990 -1.318 2.034 24.500   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.55384 0.56263 61.41 <2e-16 \*\*\*  
## lstat -0.95005 0.03873 -24.53 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.216 on 504 degrees of freedom  
## Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432   
## F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16

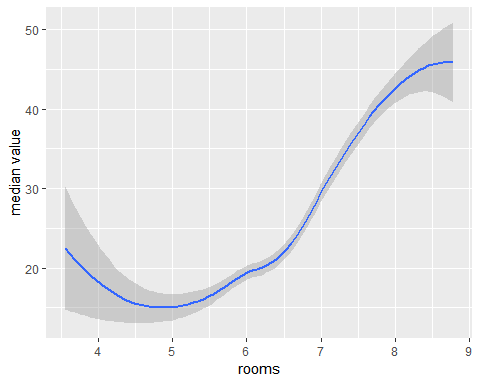
For lstat,the R squared value is high, we can prepare a plot for this one:

plot(linear\_lstat)

 #linear\_rm, linear\_lstat

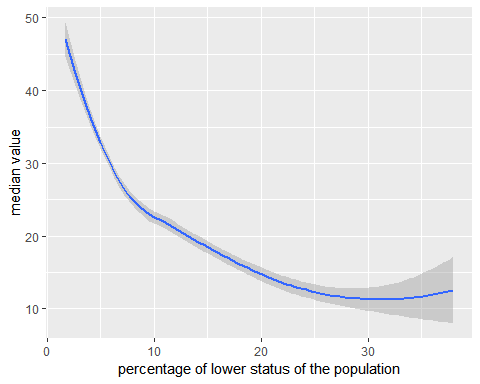
p <-ggplot(data = housing) + geom\_smooth(mapping = aes(x =rm , y = medv))  
p+labs(x="rooms", y="median value")

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



p <-ggplot(data = housing) + geom\_smooth(mapping = aes(x =lstat , y = medv))  
p+labs(x="percentage of lower status of the population", y="median value")

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

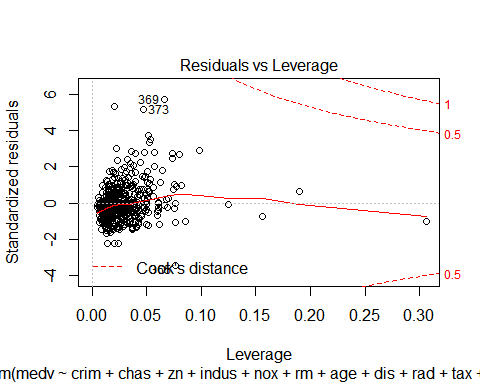
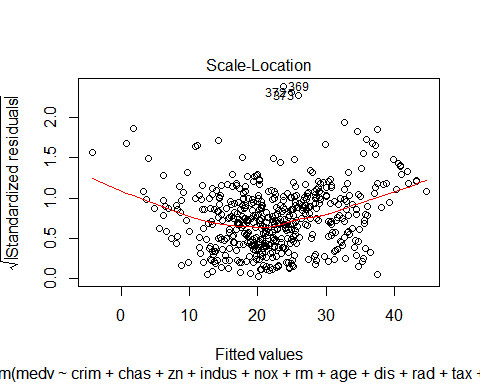
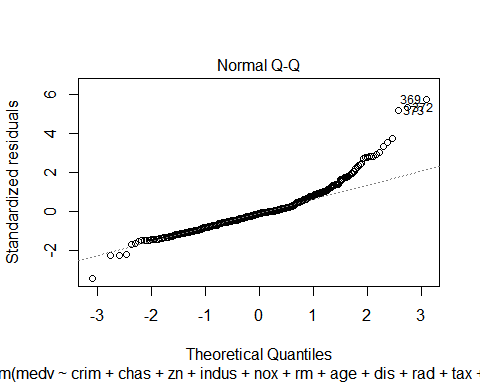
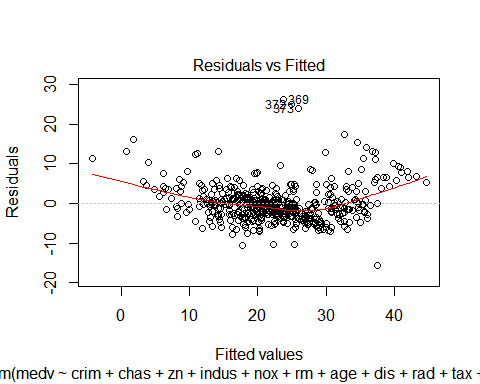
 For the above two graphs, since the r squared value was high, we plotted the residuals and also the graph. The graphs shows somewhat positive linear relationship for rm and lstat appears to be negatively linearly related to medv. ### 4. Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis ?

multiple\_lin\_reg <- lm(medv ~ crim+chas+zn+indus+nox+rm+age+dis+rad+tax+ptratio+b+lstat, data=housing)  
summary(multiple\_lin\_reg)

##   
## Call:  
## lm(formula = medv ~ crim + chas + zn + indus + 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) 3.377e+01 5.199e+00 6.496 2.02e-10 \*\*\*  
## crim -1.080e-01 3.286e-02 -3.287 0.001087 \*\*   
## chas 2.687e+00 8.616e-01 3.118 0.001925 \*\*   
## zn 4.642e-02 1.373e-02 3.382 0.000778 \*\*\*  
## indus 2.056e-02 6.150e-02 0.334 0.738288   
## 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

all predictors are significant except indus, and age.And we can reject the null hypothesis for the rest.

plot(multiple\_lin\_reg)



### 5.

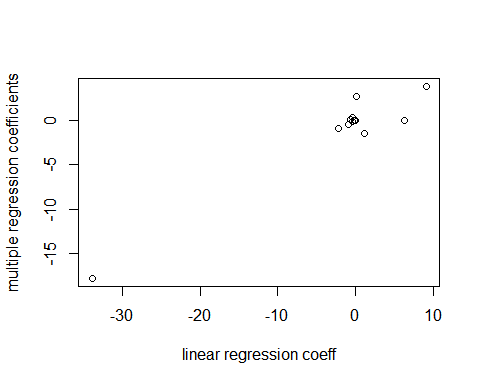
How do your results from (3) compare to your results from (4)? Create a plot displaying the univariate regression coefficients from (3) on the x-axis and the multiple regression coefficients from part (4) on the y-axis. Use this visualization to support your response.

coeff\_crim<- linear\_crim$coefficients  
linear<-as.data.frame(coeff\_crim)  
linear$zn <- linear\_zn$coefficients  
linear$indus <- linear\_indus$coefficients  
linear$chas <- linear\_chas$coefficients  
linear$nox <- linear\_nox$coefficients  
linear$rm <- linear\_rm$coefficients  
linear$age <- linear\_age$coefficients  
linear$dis <- linear\_dis$coefficients  
linear$rad <- linear\_rad$coefficients  
linear$tax <- linear\_tax$coefficients  
linear$ptratio <- linear\_ptratio$coefficients  
linear$b <-linear\_b$coefficients  
linear$lstat <-linear\_lstat$coefficients  
#linear

# multiple regression coeff

multiple <- multiple\_lin\_reg$coefficients[c(-1)]  
#multiple

plot(as.numeric(linear[2,]), as.numeric(multiple), xlab="linear regression coeff", ylab="multiple regression coefficients")



Coefficients from linear regression are similar to coefficients of multiple regression .Thus we can conclude that [4] supports [3].

### 6.

Is there evidence of a non-linear association between any of the predictors and the response? To answer this question, for each predictor fit a model of the form:

predictors <- names(housing[,-ncol(housing)])  
r\_sq<-NULL  
for(i in predictors){  
 placeholder<- lm(housing$medv ~ housing[,i] + housing[,i]^2 + housing[,i]^3)  
 r\_sq[i] <-summary(placeholder)$r.squared  
}  
r\_sq

## crim zn indus chas nox rm   
## 0.15078047 0.12992084 0.23399003 0.03071613 0.18260304 0.48352546   
## age dis rad tax ptratio b   
## 0.14209474 0.06246437 0.14563858 0.21952592 0.25784732 0.11119612   
## lstat   
## 0.54414630

Since the values of rm and lstat are higher than other and also we saw their adjusted R square value was also higher, they have some associatiom. But others do not seem to have any association. ### 7. Consider performing a stepwise model selection procedure to determine the best fit model. Discuss your results. How is this model different from the model in (4)?

steplinear <- step(lm(medv ~ ., data=housing))

## Start: AIC=1589.64  
## medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +   
## tax + ptratio + b + lstat  
##   
## Df Sum of Sq RSS AIC  
## - age 1 0.06 11079 1587.7  
## - indus 1 2.52 11081 1587.8  
## <none> 11079 1589.6  
## - chas 1 218.97 11298 1597.5  
## - tax 1 242.26 11321 1598.6  
## - crim 1 243.22 11322 1598.6  
## - zn 1 257.49 11336 1599.3  
## - b 1 270.63 11349 1599.8  
## - rad 1 479.15 11558 1609.1  
## - nox 1 487.16 11566 1609.4  
## - ptratio 1 1194.23 12273 1639.4  
## - dis 1 1232.41 12311 1641.0  
## - rm 1 1871.32 12950 1666.6  
## - lstat 1 2410.84 13490 1687.3  
##   
## Step: AIC=1587.65  
## medv ~ crim + zn + indus + chas + nox + rm + dis + rad + tax +   
## ptratio + b + lstat  
##   
## Df Sum of Sq RSS AIC  
## - indus 1 2.52 11081 1585.8  
## <none> 11079 1587.7  
## - chas 1 219.91 11299 1595.6  
## - tax 1 242.24 11321 1596.6  
## - crim 1 243.20 11322 1596.6  
## - zn 1 260.32 11339 1597.4  
## - b 1 272.26 11351 1597.9  
## - rad 1 481.09 11560 1607.2  
## - nox 1 520.87 11600 1608.9  
## - ptratio 1 1200.23 12279 1637.7  
## - dis 1 1352.26 12431 1643.9  
## - rm 1 1959.55 13038 1668.0  
## - lstat 1 2718.88 13798 1696.7  
##   
## Step: AIC=1585.76  
## medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +   
## b + lstat  
##   
## Df Sum of Sq RSS AIC  
## <none> 11081 1585.8  
## - chas 1 227.21 11309 1594.0  
## - crim 1 245.37 11327 1594.8  
## - zn 1 257.82 11339 1595.4  
## - b 1 270.82 11352 1596.0  
## - tax 1 273.62 11355 1596.1  
## - rad 1 500.92 11582 1606.1  
## - nox 1 541.91 11623 1607.9  
## - ptratio 1 1206.45 12288 1636.0  
## - dis 1 1448.94 12530 1645.9  
## - rm 1 1963.66 13045 1666.3  
## - lstat 1 2723.48 13805 1695.0

summary(steplinear)

##   
## Call:  
## lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +   
## tax + ptratio + b + lstat, data = housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.5984 -2.7386 -0.5046 1.7273 26.2373   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.622429 5.153549 6.524 1.70e-10 \*\*\*  
## crim -0.108413 0.032779 -3.307 0.001010 \*\*   
## zn 0.045845 0.013523 3.390 0.000754 \*\*\*  
## chas 2.718716 0.854240 3.183 0.001551 \*\*   
## nox -17.376023 3.535243 -4.915 1.21e-06 \*\*\*  
## rm 3.801579 0.406316 9.356 < 2e-16 \*\*\*  
## dis -1.492711 0.185731 -8.037 6.84e-15 \*\*\*  
## rad 0.299608 0.063402 4.726 3.00e-06 \*\*\*  
## tax -0.011778 0.003372 -3.493 0.000521 \*\*\*  
## ptratio -0.946525 0.129066 -7.334 9.24e-13 \*\*\*  
## b 0.009291 0.002674 3.475 0.000557 \*\*\*  
## lstat -0.522553 0.047424 -11.019 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.736 on 494 degrees of freedom  
## Multiple R-squared: 0.7406, Adjusted R-squared: 0.7348   
## F-statistic: 128.2 on 11 and 494 DF, p-value: < 2.2e-16

#plot(steplinear)

steplinear$anova

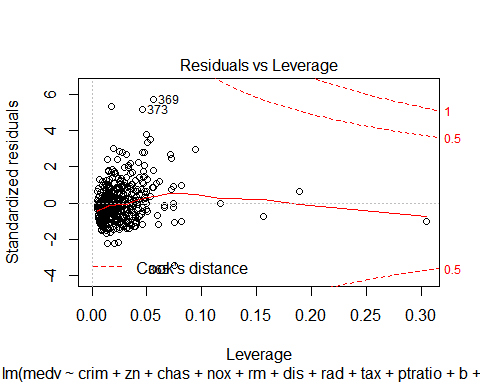
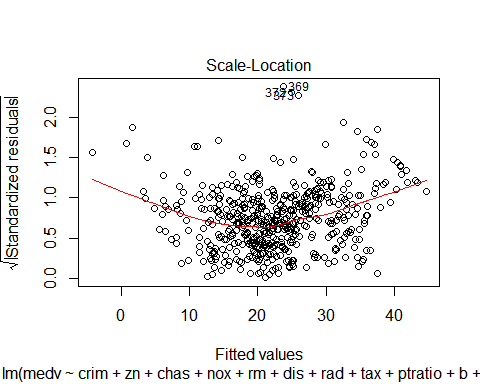
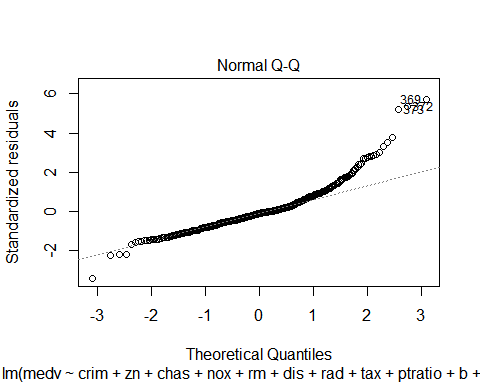
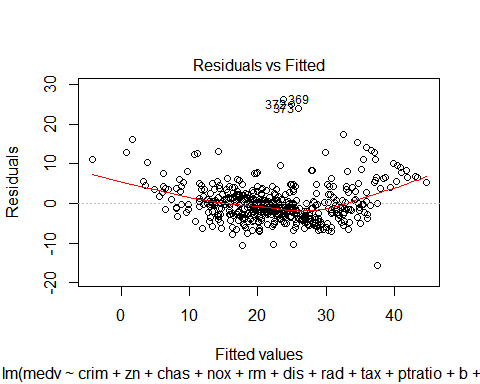
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 NA NA 492 11078.78 1589.643  
## 2 - age 1 0.06183435 493 11078.85 1587.646  
## 3 - indus 1 2.51754013 494 11081.36 1585.761

We see that the predictors suggested by stepwise model selection are similar to the ones we saw in multiple linear regression. Stepwise model and multiple regression model suggests crim, zn, chas, nox, rm, dis, rad, tax, ptratio, black, lstat.

### 8.

Evaluate the statistical assumptions in your regression analysis from (7) by performing a basic analysis of model residuals and any unusual observations. Discuss any concerns you have about your model.

plot(steplinear)

 We observe that the residual line is not in line with zero, that indicates some outliers, error variance. There could be some assumptions around linear relationship and multi-collinearity for this data model.