Assignmet 4

Jaswanth

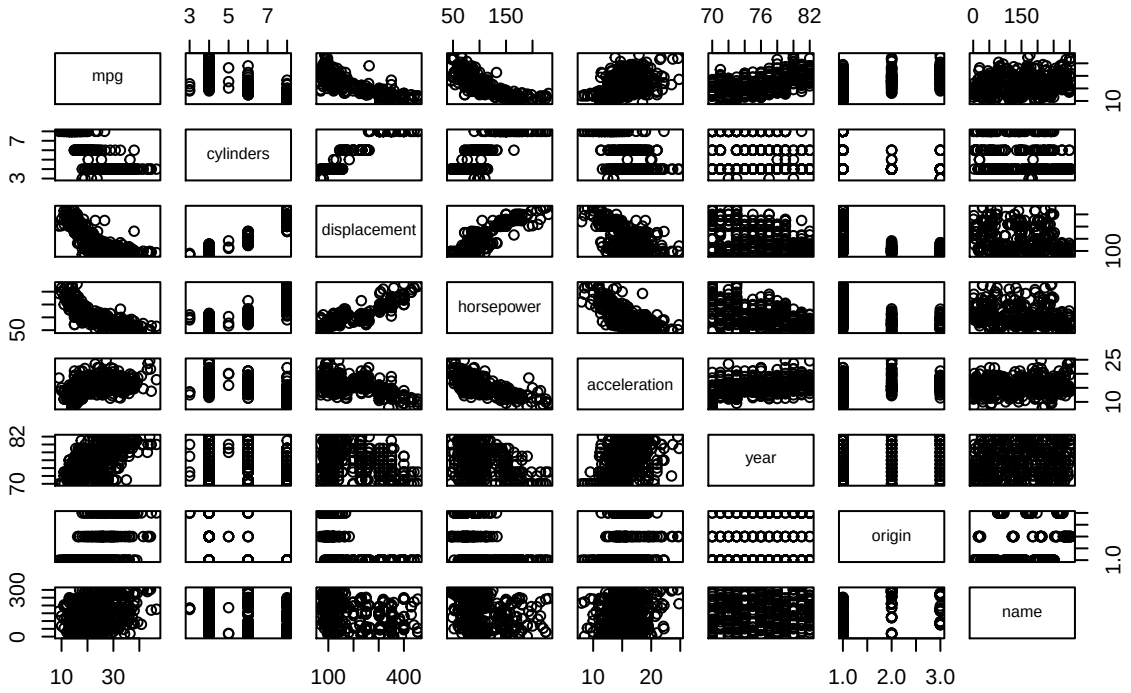
October 9, 2018

1. This question involves the use of multiple linear regression on the Auto data set from the course webpage (https://scads.eecs.wsu.edu/index.php/datasets/). Ensure that you remove missing values from the dataframe, and that values are represented in the appropriate types (num or int for quantitative variables, factor, logi or str for qualitative).

```
auto <- read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/Auto.csv", na.strings="?")
auto <- na.omit(auto)
```

   a. Produce a scatterplot matrix which includes all of the variables in the data set.

```
pairs(~mpg+cylinders+displacement+horsepower+acceleration+year+origin+name, data=auto,
    main="Three Cylinder Options")
```



   b. Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.

```
x <- auto[1:4]
y <- auto[5:8]
cor(x, y)
```

```
##                 weight acceleration       year     origin
## mpg          -0.8322442    0.4233285  0.5805410  0.5652088
```

```
## cylinders     0.8975273   -0.5046834 -0.3456474 -0.5689316
## displacement  0.9329944   -0.5438005 -0.3698552 -0.6145351
## horsepower    0.8645377   -0.6891955 -0.4163615 -0.4551715
```

c. Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output:

```
fit <- lm(mpg ~ cylinders+displacement+horsepower+ weight + year + origin, data=auto)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     year + origin, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7604 -2.1791 -0.1535  1.8524 13.1209
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.556e+01  4.175e+00  -3.728 0.000222 ***
## cylinders    -5.067e-01  3.227e-01  -1.570 0.117236
## displacement  1.927e-02  7.472e-03   2.579 0.010287 *
## horsepower   -2.389e-02  1.084e-02  -2.205 0.028031 *
## weight       -6.218e-03  5.714e-04 -10.883  < 2e-16 ***
## year          7.475e-01  5.079e-02  14.717  < 2e-16 ***
## origin        1.428e+00  2.780e-01   5.138 4.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.326 on 385 degrees of freedom
## Multiple R-squared:  0.8212, Adjusted R-squared:  0.8184
## F-statistic: 294.6 on 6 and 385 DF,  p-value: < 2.2e-16
```

Solution: There are some significant relationships between the response and predictors. Weight, year and origin are significant at 0.001 level. So the null hypothesis is rejected fot these predictors. For the cylinder variable null hypothesis cannot be rejected. And also the r squared value is very high which also implies a good fit.

i. Which predictors appear to have a statistically significant relationship to the response, and how do you determine this?

Solution: Asterisks in a regression table indicate the level of the statistical significance of a regression coefficient. weight= (***), year= ()***, origin= (**), *horsepower= ()*, displacement= (*),

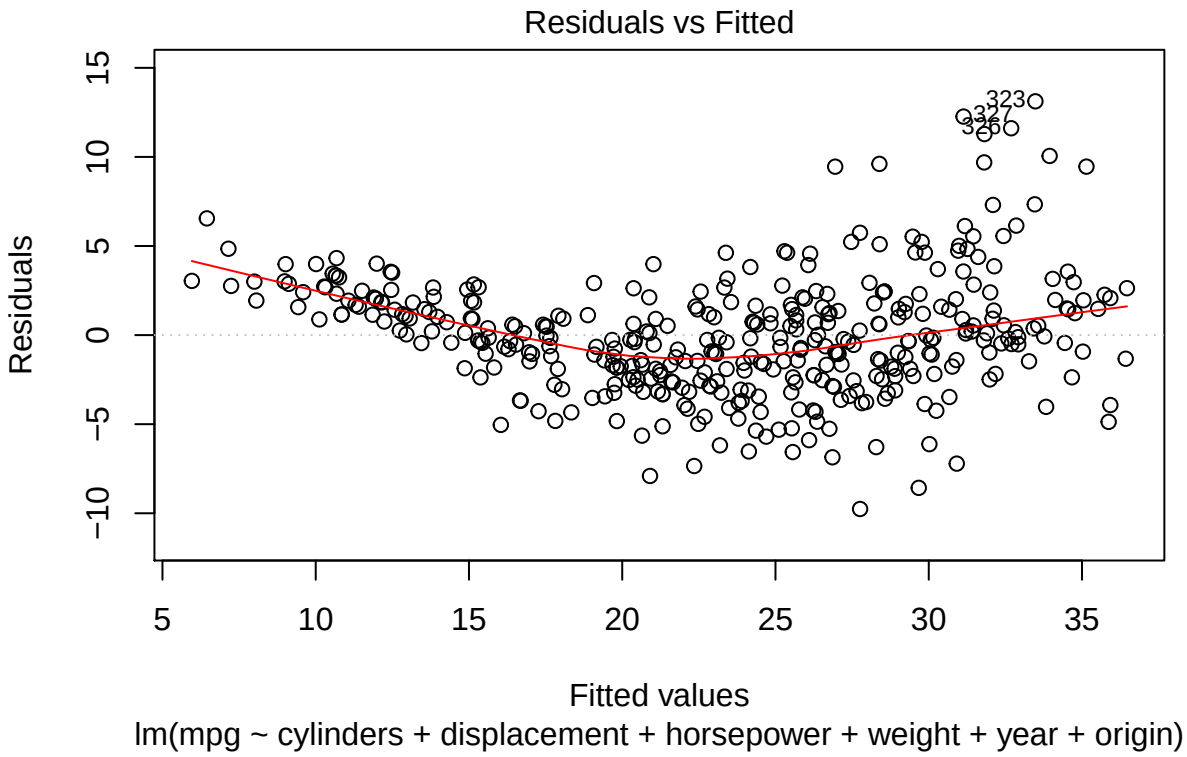
The asterisks in a regression table correspond with a legend at the bottom of the table. Weight,year and origin are at 0.001 level. Horsepower and displacement are at 0.5 level significanc and for rest of the variables we cannot reject the null hypothesis.

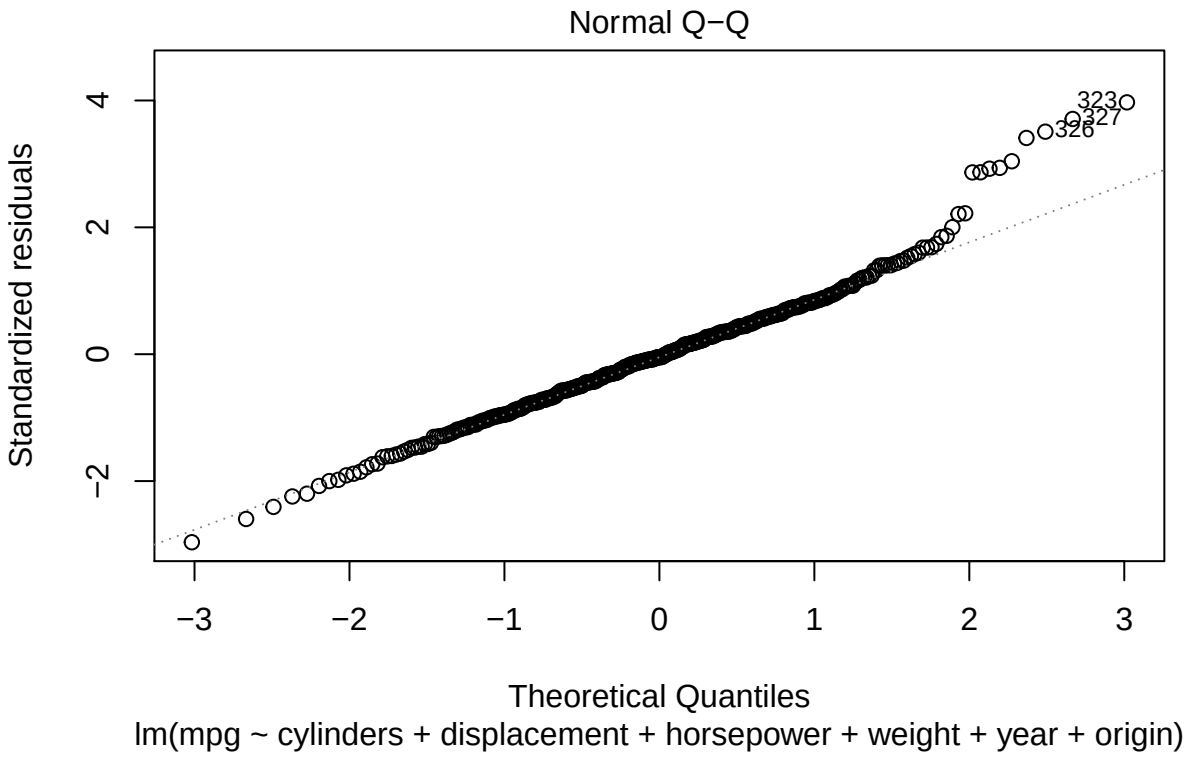
ii. What does the coefficient for the cylinders variable suggest, in simple terms?

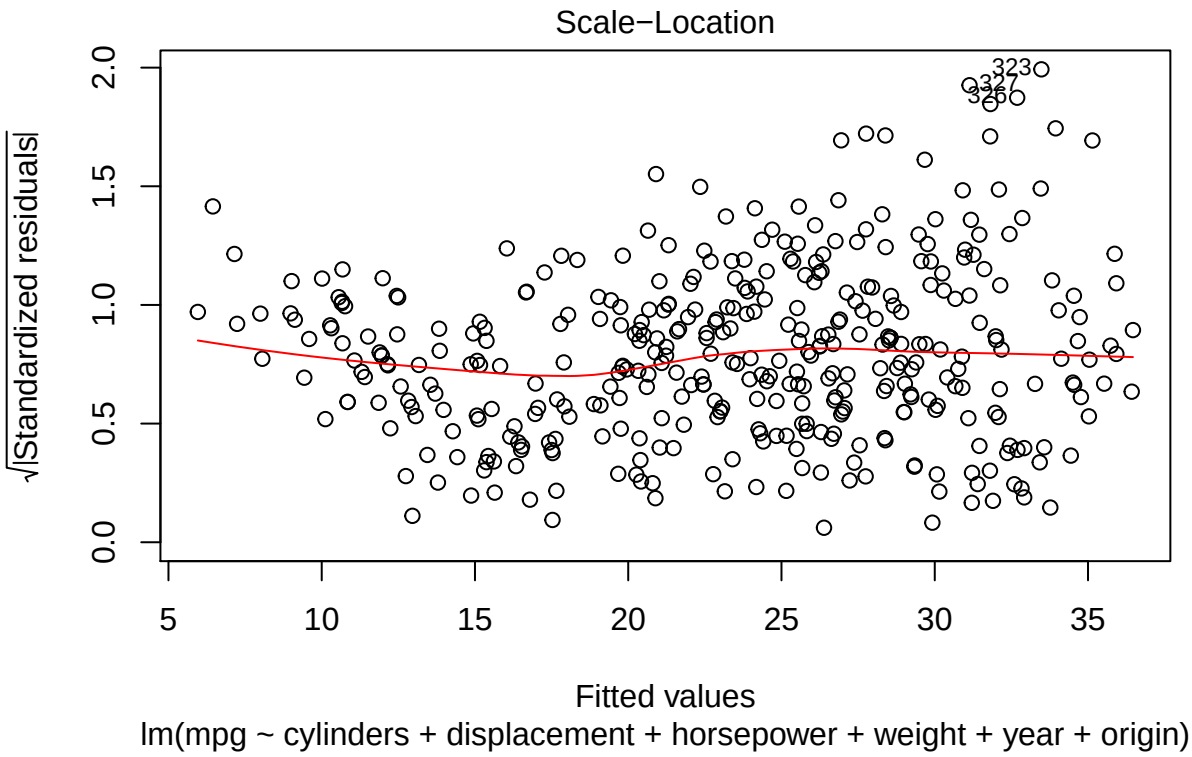
Solution: The predictor variable accepts the null hypothesis and there is no significant relationship between the cylinder and response.

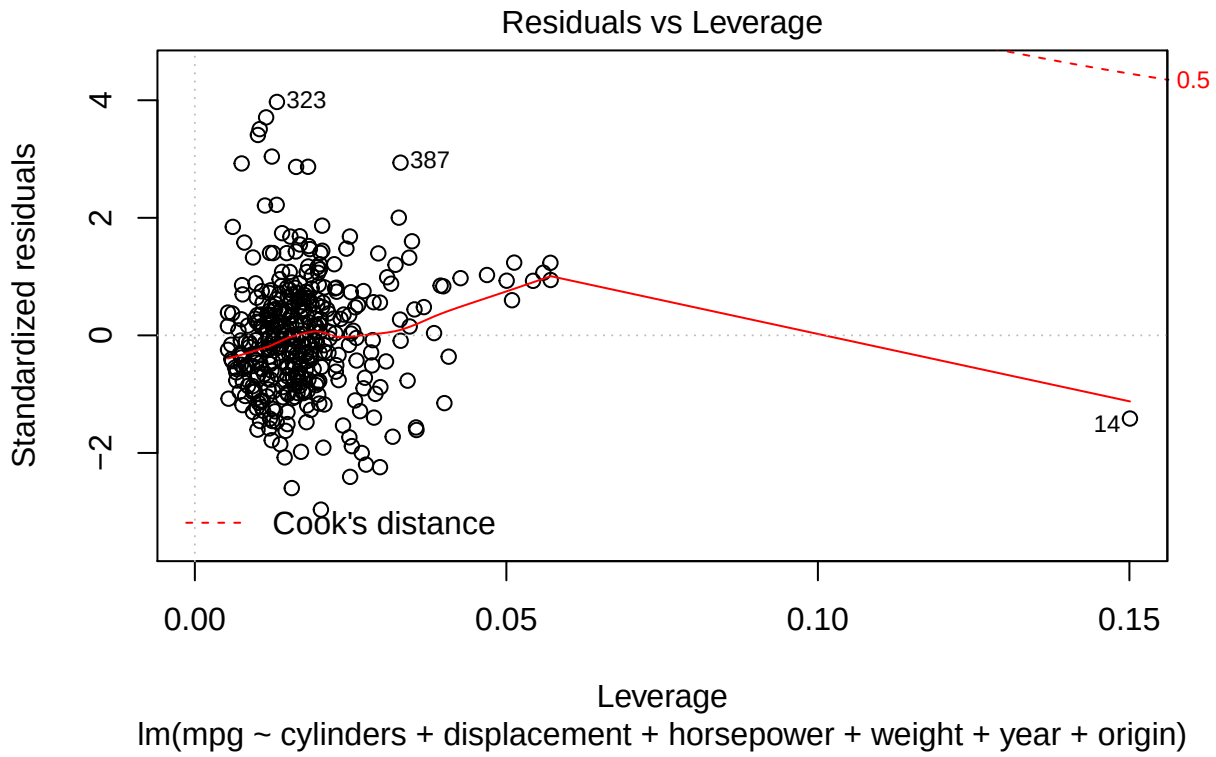
d. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```r
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ cylinders + displacement + horsepower + weight + year + origin)
323
327
326
15
10
5
0
−5
−10
5
10
15
20
25
30
35


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ cylinders + displacement + horsepower + weight + year + origin)
323
327
326
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
323
327
326
5
10
15
20
25
30
35
Fitted values
lm(mpg ~ cylinders + displacement + horsepower + weight + year + origin)



Solution: The Residuals vs Fitted graph does not seem no pattern, so it points no strong evidence of non-linearity. The Residuals vs Fitted graph assumes a bit of funnel shape, so it presents a bit of heteroscedasticity. Specifically the observation 14 is a highly leverage point as shown in Residuals vc Leverage graph.

e. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
symbol<-lm(mpg ~ cylinders+displacement+horsepower+ weight * year + origin, data=auto)
symbol1<-lm(mpg ~ cylinders+displacement+horsepower+ weight + year * origin, data=auto)
summary(symbol)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight *
##     year + origin, data = auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.330 -1.927 -0.172  1.544 11.835
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.139e+02  1.307e+01  -8.716  < 2e-16 ***
## cylinders    -1.519e-01  3.032e-01  -0.501   0.6166
## displacement  1.193e-02  7.004e-03   1.703   0.0893 .
## horsepower   -4.094e-02  1.030e-02  -3.976 8.37e-05 ***
## weight        3.026e-02  4.660e-03   6.494 2.59e-10 ***
```

```
## year          2.055e+00  1.726e-01  11.911  < 2e-16 ***
## origin        1.182e+00  2.602e-01   4.542 7.47e-06 ***
## weight:year  -4.795e-04  6.085e-05  -7.880 3.42e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.09 on 384 degrees of freedom
## Multiple R-squared:  0.8461, Adjusted R-squared:  0.8433
## F-statistic: 301.5 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
summary(symbol1)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     year * origin, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7687 -2.0167 -0.1185  1.6846 12.4129
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.012e+01  8.875e+00   1.140 0.254936
## cylinders    -5.192e-01  3.188e-01  -1.629 0.104182
## displacement  1.500e-02  7.495e-03   2.002 0.046030 *
## horsepower   -2.193e-02  1.072e-02  -2.045 0.041493 *
## weight       -6.063e-03  5.663e-04 -10.705  < 2e-16 ***
## year          4.185e-01  1.125e-01   3.720 0.000229 ***
## origin       -1.389e+01  4.695e+00  -2.958 0.003287 **
## year:origin   1.969e-01  6.026e-02   3.268 0.001181 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.285 on 384 degrees of freedom
## Multiple R-squared:  0.826,  Adjusted R-squared:  0.8228
## F-statistic: 260.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

Solution: There are 2 computation which have significance with the response i.e weight* year and year*origin and have 0.001 and 0.01 level of significance respectively. The r squared valu is also high that means the distribution is close to line.

f. Try transformations of the variables with X3 and log(X). Comment on your findings.

```
auto$mpgpower3=(auto$mpg)^3
auto$mpglog= log(auto$mpg)
```

Solution: If statistically significant variables are transformed then the new transformed variable is also significant. It looks that the log transformation gives the most linear looking plot

2. This problem involves the Boston data set, which we saw in the lab. 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.

```
library(MASS)
data("Boston")
```

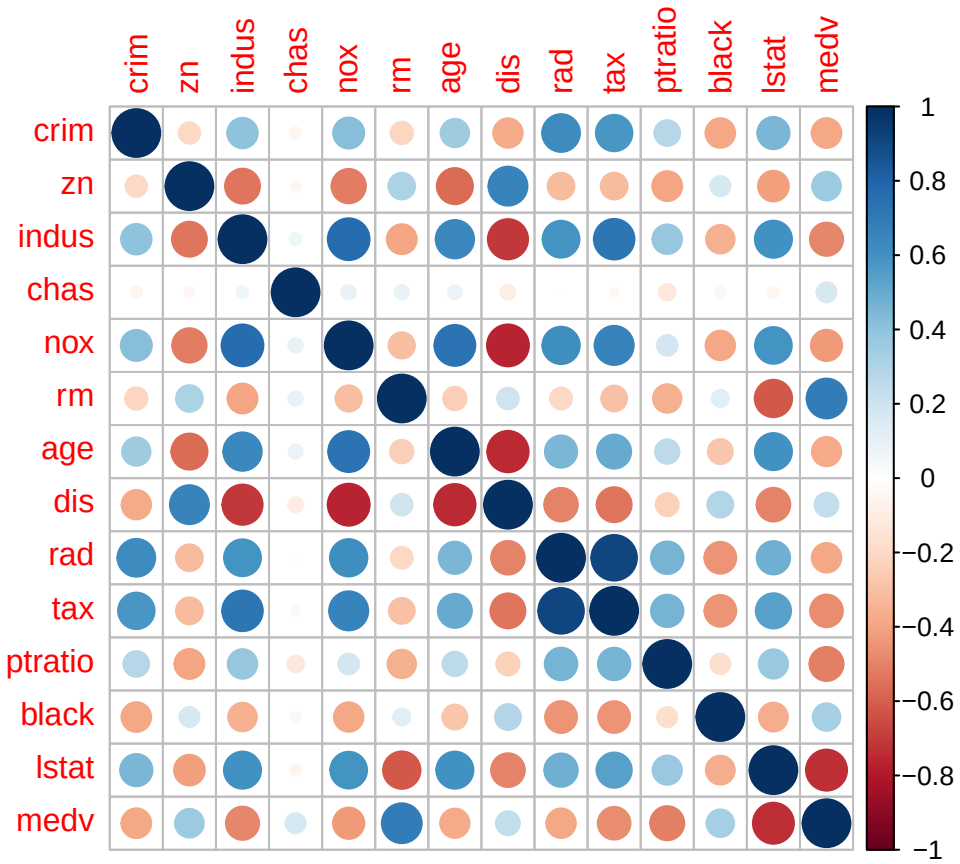
a. For each predictor, fit a simple linear regression model to predict the response. Include the code, but

not the output for all models in your solution. In which of the models is there a statistically significant association between the predictor and the response? Considering the meaning of each variable, discuss the relationship between crim and nox, chas, medv and dis in particular. How do these relationships differ?

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
B <- cor(Boston)
corrplot(B, method='circle')
```




```
percap13 <- lm(crim~ zn, data=Boston)
percap1 <- lm(crim~ indus, data=Boston)
percap2 <- lm(crim~ chas, data=Boston)
percap3 <- lm(crim~ nox, data=Boston)
percap4 <- lm(crim~ rm, data=Boston)
percap5 <- lm(crim~ age, data=Boston)
percap6 <- lm(crim~ dis, data=Boston)
percap7 <- lm(crim~ rad, data=Boston)
percap8 <- lm(crim~ tax, data=Boston)
percap9 <- lm(crim~ ptratio, data=Boston)
percap10 <- lm(crim~black , data=Boston)
percap11 <- lm(crim~ lstat, data=Boston)
percap12 <- lm(crim~ medv, data=Boston)
```

Solution: All the predictors have a significant relationship except the chas variable. nox,medv,dis all the predictors have r squared value which is close to 0.2 that means the points are not very close to the line. But

these predictors are significant with crim.

b. 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 H0 : ??j = 0?

```
percap <- lm(crim~.-crim, data=Boston)
summary(percap)
```

```
##
## Call:
## lm(formula = crim ~ . - 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
```

Solution: dis,rad,medv,zn,black are staitistically significant in this regression model. For all the rest of the predictors we now fail to reject the null hypothesis. we can reject null hypothesis for the following predictors: dis rad medv zn black

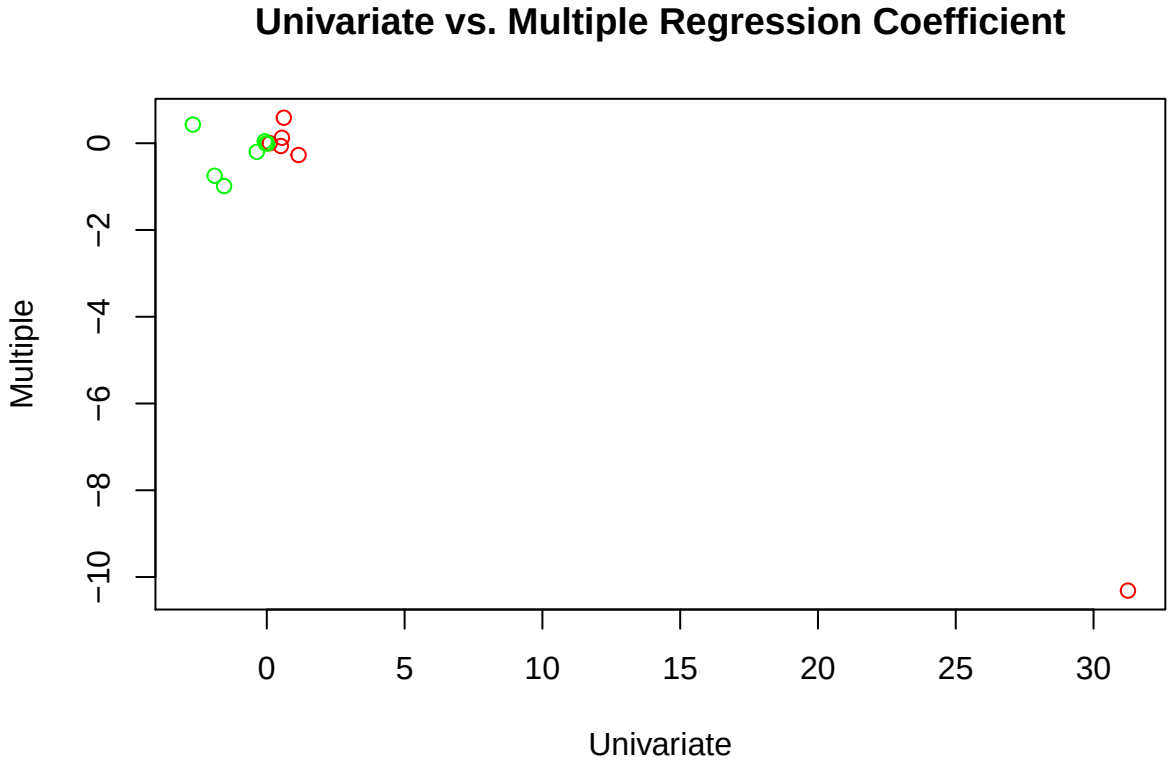
c. How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis. What does this plot tell you about the various predictors?

```
uni <- lm(crim ~ zn, data = Boston)$coefficients[2]
uni <- append(uni, lm(crim ~ indus, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ chas, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ nox, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ rm, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ age, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ dis, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ rad, data = Boston)$coefficients[2])
```

```
uni <- append(uni, lm(crim ~ tax, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ ptratio, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ black, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ lstat, data = Boston)$coefficients[2])
uni <- append(uni, lm(crim ~ medv, data = Boston)$coefficients[2])
percap$coefficients[2:14]
```

```
##            zn         indus          chas           nox            rm
##   0.044855215  -0.063854824  -0.749133611 -10.313534912   0.430130506
##           age           dis           rad           tax       ptratio
##   0.001451643  -0.987175726   0.588208591  -0.003780016  -0.271080558
##         black         lstat          medv
##  -0.007537505   0.126211376  -0.198886821
```

```
plot(uni, percap$coefficients[2:14], main = 'Univariate vs. Multiple Regression Coefficient', xlab = 'U
```



Solution: There is a difference between the simple and multiple regression coefficients. This is due to that in the simple regression case, the slope term represents the average effect of an increase in the predictor, ignoring the other predictors in the table.If it is the multiple regression case, the slope term represents the average effect of an increase in the predictor, while keeping the predictors as constant.So in some cases there is a significant relationship between the predictor and the response in simple linear regression but that is not the case in multiple linear regression

d. Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form Y = ??0 + ??1X + ??2X2 + ??3X3

- ?? Hint: use the poly() function. Again, include the code, but not the output for each model in your solution, and instead describe any non-linear trends you uncover.

```
x1 <- lm(crim ~ medv + I(medv^2) + I(medv^3), data = Boston)
x2 <- lm(crim ~ zn + I(zn^2) + I(zn^3), data = Boston)
x3 <- lm(crim ~ rad + I(rad^2) + I(rad^3), data = Boston)
x4 <- lm(crim ~ dis + I(dis^2) + I(dis^3), data = Boston)
x5 <- lm(crim ~ indus + I(indus^2) + I(indus^3), data = Boston)
x6 <- lm(crim ~ chas + I(chas^2) + I(chas^3), data = Boston)
x7 <- lm(crim ~ nox + I(nox^2) + I(nox^3), data = Boston)
x8 <- lm(crim ~ rm + I(rm^2) + I(rm^3), data = Boston)
x9 <- lm(crim ~ tax + I(tax^2) + I(tax^3), data = Boston)
x10 <- lm(crim ~ ptratio + I(ptratio^2) + I(ptratio^3), data = Boston)
x11 <- lm(crim ~ black + I(black^2) + I(black^3), data = Boston)
x12 <- lm(crim ~ lstat + I(lstat^2) + I(lstat^3), data = Boston)
```

Solution: For chas variable, we obtain NA values for the squared and cubed values. This makes it clear that as chas is a dummy variable, composed of only 0s and 1s, and these values will not change if they are squared or cubed.

The other variables indus, nox, dis, ptracio, and medv, there is possibilty of a non-linear relationship, as each of these variables squared and cubed terms are statistically signficant that means here we have to reject the null hypothesis.

Age also has a non-linear relationship as once squared-age and cubed-age are computed, linear age becomes statistically insignficant.

3. An important assumption of the linear regression model is that the error terms are uncorrelated (independent). But error terms can sometimes be correlated, especially in time-series data.

a. What are the issues that could arise in using linear regression (via least squares estimates) when error terms are correlated? Comment in particular with respect to

i) regression coefficients
ii) the standard error of regression coefficients
iii) confidence intervals

Solution: Standard errors which are calculated for the estimated regression coefficients are on the basis of assuming the uncorrelated error terms. But if there exist a correlation among the error terms, then the estimated standard errors will underestimate the true standard errors. If this happens then the confidence intervals will become narrow than they should have been. For example, a higher confidence interval may in reality have a much lower probability than what it tends to have of containing the true value of the parameter. If the error terms are correlated, we may have of confidence in our model which is more than it has to be.

b. What methods can be applied to deal with correlated errors? Mention at least one method?

Solution: Generalized least squares(Gls) and linear mixded effect models(Lme) methods are used manipulate data and handle the correlation betweene the errors.