Week3 Assignment

Mohit Mehndiratta

11/08/2021

## Question 1

### Part(a) - Upload the Advertising data set and explore it.

AdvertisingDS <- read.csv("../datasets/advertising.csv")  
attach(AdvertisingDS)  
head(AdvertisingDS)

## TV Radio Newspaper Sales  
## 1 230.1 37.8 69.2 22.1  
## 2 44.5 39.3 45.1 10.4  
## 3 17.2 45.9 69.3 9.3  
## 4 151.5 41.3 58.5 18.5  
## 5 180.8 10.8 58.4 12.9  
## 6 8.7 48.9 75.0 7.2

dim(AdvertisingDS)

## [1] 200 4

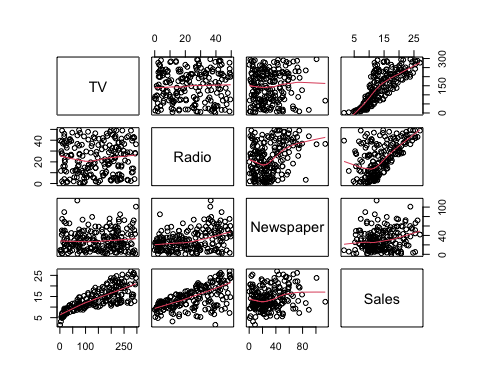
sapply(AdvertisingDS, class)

## TV Radio Newspaper Sales   
## "numeric" "numeric" "numeric" "numeric"

From the above output, we can see that Advertising data set has 200 observations of 4 variables. All the variables are numeric.

### Part(b) - Find the Covariance and Correlation Matrix of Sales, TV, Radio and Newspaper.

pairs(AdvertisingDS, panel = panel.smooth)



cor(AdvertisingDS)

## TV Radio Newspaper Sales  
## TV 1.00000000 0.05480866 0.05664787 0.7822244  
## Radio 0.05480866 1.00000000 0.35410375 0.5762226  
## Newspaper 0.05664787 0.35410375 1.00000000 0.2282990  
## Sales 0.78222442 0.57622257 0.22829903 1.0000000

cov(AdvertisingDS)

## TV Radio Newspaper Sales  
## TV 7370.94989 69.86249 105.91945 350.39019  
## Radio 69.86249 220.42774 114.49698 44.63569  
## Newspaper 105.91945 114.49698 474.30833 25.94139  
## Sales 350.39019 44.63569 25.94139 27.22185

By Looking at the plot for Sales vs TV, we can see that there is a **strong positive linear relationship** between them and the correlation matrix table we can see that the value is **0.7822244** which confirms the same.

For Sales vs Radio, there exists a **moderate positive linear relationship** as the correlation between them is **0.5762226**.

For Sales vs Newspaper, we can see that the line in the graph is not so linear thus depicting a **week linear relationship** and from the correlation matrix table we can see that the value is **0.2282990** which is very far away from 1 thus depicting a **week linear relationship**.

### Part(c) - Construct the multiple linear regression model and find the least square estimates of the model parameters.

Multiple linear regression includes building a model of multiple variable with respect to our target variable. Here Sales is our target variable and we can first start with including all the numeric variables into our model.

model1 <- lm(Sales~TV+Radio+Newspaper)  
summary(model1)

##   
## Call:  
## lm(formula = Sales ~ TV + Radio + Newspaper)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.8277 -0.8908 0.2418 1.1893 2.8292   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.938889 0.311908 9.422 <2e-16 \*\*\*  
## TV 0.045765 0.001395 32.809 <2e-16 \*\*\*  
## Radio 0.188530 0.008611 21.893 <2e-16 \*\*\*  
## Newspaper -0.001037 0.005871 -0.177 0.86   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.686 on 196 degrees of freedom  
## Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956   
## F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

By looking at the summary for the above model, we can see that the least square estimates of our model are as follows:

1. = 2.938889
2. = 0.045765
3. = 0.188530
4. = -0.001037

Now, substituting the value of these estimates in our equation:

= 2.938889 + 0.045765 TV + 0.188530 Radio - 0.001037 Newspaper

### Part(d) - Test the significance of the parameters and find the resulting model to model Sales in terms of advertising modes, TV, Radio and Newspaper

To test the significance of the parameters of our model, we will use Hypothesis testing, H: = 0 and H: 0.

From the summary of our model, we can see that p-value of TV is less than 0.05. Therefore, we get enough evidence to reject the null hypothesis H at 5% significance level. Moreover, we can also see that p-value of TV is also less than 0.01 thus depicting a **strong linear relationship**. Therefore, we can reject the null hypothesis H at 1% level of significance as well.

For Radio, we can see that p-value of Radio is less than 0.05. Therefore, we get enough evidence to reject the null hypothesis H at 5% significance level. Moreover, we can also see that p-value of Radio is also less than 0.01 thus depicting a **strong linear relationship**. Therefore, we can reject the null hypothesis H at 1% level of significance as well.

However, for Newspaper we can see that p-value of Newspaper is 0.86 which is not less than 0.05.This shows that there is a **week linear relationship**. Therefore, we can not reject the null hypothesis H here, we have to reject the alternate hypothesis H here.

Now, after removing Newspaper our resultant model becomes:

model2 <- lm(Sales~TV+Radio)  
summary(model2)

##   
## Call:  
## lm(formula = Sales ~ TV + Radio)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.7977 -0.8752 0.2422 1.1708 2.8328   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.92110 0.29449 9.919 <2e-16 \*\*\*  
## TV 0.04575 0.00139 32.909 <2e-16 \*\*\*  
## Radio 0.18799 0.00804 23.382 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.681 on 197 degrees of freedom  
## Multiple R-squared: 0.8972, Adjusted R-squared: 0.8962   
## F-statistic: 859.6 on 2 and 197 DF, p-value: < 2.2e-16

Now, by looking at the summary of our new model, we can see that through hypothesis testing, TV and Radio show **strong linear relationship**. Hence our model equation becomes:

= 2.92110 + 0.04575 TV + 0.18799 Radio

### Part(e) - Assess the overall accuracy of the model.

After looking at the summary of our new model, we can see that Residual standard error is 1.681. We can check the Sales summary and then we can compare the error.

summary(Sales)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.60 10.38 12.90 14.02 17.40 27.00

So, for the summary we can observe that the Residual standard error or standard deviation seems to be low as compared to the median of Sales, which is good.

Alternatively we can also check value from the summary which is 89.72%. That means 89.72% proportion is explained by the model.

And at the end, if we look at the p-value, it is very small, which is good.

### Part(f) - Calculate the predicted vales and residuals

Below are the predicted values for our model.

pred\_values <- predict(model2)  
pred\_values

## 1 2 3 4 5 6 7 8   
## 20.555465 12.345362 12.337018 17.617116 13.223908 12.512084 11.718212 12.105516   
## 9 10 11 12 13 14 15 16   
## 3.709379 12.551697 7.035860 17.256520 10.608662 8.810951 18.444668 20.828915   
## 17 18 19 20 21 22 23 24   
## 12.903865 23.241076 9.941215 14.153846 18.121392 14.742064 6.514172 16.544027   
## 25 26 27 28 29 30 31 32   
## 8.140352 15.608021 14.967694 17.046335 19.399541 9.159297 21.642922 11.357918   
## 33 34 35 36 37 38 39 40   
## 7.650459 18.833463 7.563028 16.992801 23.367207 15.625899 9.912578 20.440580   
## 41 42 43 44 45 46 47 48   
## 16.378721 17.298709 21.562154 13.966923 8.900997 15.162638 8.886450 21.699440   
## 49 50 51 52 53 54 55 56   
## 16.286903 8.181629 12.645694 9.319628 20.661801 19.961262 20.355124 21.308647   
## 57 58 59 60 61 62 63 64   
## 8.537748 12.762395 21.890729 18.107469 5.744971 22.904187 16.784138 13.184749   
## 65 66 67 68 69 70 71 72   
## 16.965709 7.826528 8.987035 12.020662 18.953134 21.093690 17.783507 10.633296   
## 73 74 75 76 77 78 79 80   
## 10.351138 9.913340 17.309835 11.909704 4.480148 13.792391 8.789203 9.676214   
## 81 82 83 84 85 86 87 88   
## 11.436214 14.663881 10.182720 14.416472 20.773505 15.220024 11.582034 15.618724   
## 89 90 91 92 93 94 95 96   
## 11.755103 16.931103 9.987143 4.511679 19.179730 21.262772 10.467086 16.333479   
## 97 98 99 100 101 102 103 104   
## 12.620231 15.329044 24.128426 16.946510 13.905346 23.307018 17.640341 14.751930   
## 105 106 107 108 109 110 111 112   
## 20.268099 17.953621 6.132907 7.113733 3.595686 19.663924 14.794090 21.123819   
## 113 114 115 116 117 118 119 120   
## 13.855332 16.383990 15.297256 12.937084 11.978488 6.567163 15.609467 6.816651   
## 121 122 123 124 125 126 127 128   
## 14.424501 7.860765 13.621365 15.058118 19.494043 9.129252 10.590963 6.590636   
## 129 130 131 132 133 134 135 136   
## 22.212603 7.904018 10.397700 15.600460 8.418883 19.275815 11.866030 13.966786   
## 137 138 139 140 141 142 143 144   
## 11.424198 20.877226 9.757607 19.634112 9.475405 18.438803 19.251445 8.778621   
## 145 146 147 148 149 150 151 152   
## 10.105028 9.697690 15.279189 23.260388 12.235950 9.816591 18.377596 10.036584   
## 153 154 155 156 157 158 159 160   
## 16.342517 18.222271 15.480532 5.289428 15.395226 10.019564 10.393418 12.406103   
## 161 162 163 164 165 166 167 168   
## 14.216501 13.572481 14.944003 17.320200 11.047079 14.289784 10.808694 13.360766   
## 169 170 171 172 173 174 175 176   
## 17.213351 17.921933 7.389574 14.376846 7.596578 11.960970 13.736151 24.783526   
## 177 178 179 180 181 182 183 184   
## 19.964022 12.174924 16.013844 12.378040 10.575089 13.933696 6.564088 24.163936   
## 185 186 187 188 189 190 191 192   
## 18.537949 20.779377 9.698684 17.060279 18.620097 6.051445 12.454978 8.405926   
## 193 194 195 196 197 198 199 200   
## 4.478859 18.448761 16.463190 5.364512 8.152375 12.768048 23.792923 15.157543

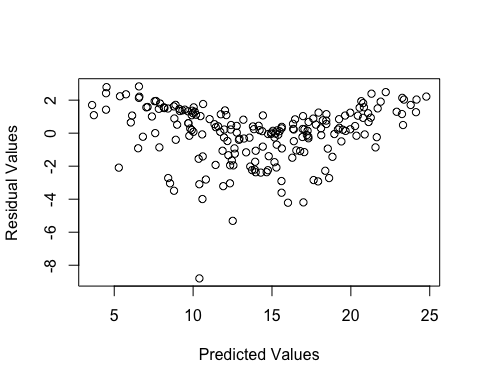
Below are the residuals values for our model.

resid\_value <- resid(model2)  
resid\_value

## 1 2 3 4 5 6   
## 1.544535367 -1.945362291 -3.037017734 0.882884040 -0.323908130 -5.312084486   
## 7 8 9 10 11 12   
## 0.081787586 1.094484471 1.090620802 -1.951696959 1.564140295 0.143479847   
## 13 14 15 16 17 18   
## -1.408661866 0.889049493 0.555332270 1.571084607 -0.403865071 1.158923744   
## 19 20 21 22 23 24   
## 1.358785237 0.446153807 -0.121391608 -2.242063573 -0.914171675 -1.044026630   
## 25 26 27 28 29 30   
## 1.559647851 -3.608020596 0.032306170 -1.146334603 -0.499541451 1.340702516   
## 31 32 33 34 35 36   
## -0.242921869 0.542081919 1.949540720 -1.433463336 1.936972365 -4.192800991   
## 37 38 39 40 41 42   
## 2.032792811 -0.925899395 0.187421706 1.059419901 0.221278776 -0.198709354   
## 43 44 45 46 47 48   
## -0.862153703 -1.066922660 -0.400997396 -0.262638136 1.713550329 1.500559540   
## 49 50 51 52 53 54   
## -1.486902684 1.518370506 -1.245694072 1.380372076 1.938198850 1.238737580   
## 55 56 57 58 59 60   
## -0.155123566 2.391352569 -3.037747831 0.437605120 1.909271424 0.292530860   
## 61 62 63 64 65 66   
## 2.355029026 1.295813425 -1.084137679 0.815251469 1.034290928 1.473471538   
## 67 68 69 70 71 72   
## 0.512965437 1.379338058 -0.053134249 1.206309625 0.516493066 1.766703949   
## 73 74 75 76 77 78   
## -1.551138436 1.086659922 -0.309835430 -3.209703991 2.419851910 0.407609409   
## 79 80 81 82 83 84   
## -3.489203290 1.323785991 0.363786363 -2.363880903 1.117279710 -0.816472350   
## 85 86 87 88 89 90   
## 0.926495319 -0.020023960 0.417966463 0.381276455 1.144897135 -0.231102643   
## 91 92 93 94 95 96   
## 1.212856709 2.788321036 0.220270248 0.937227707 1.032913773 0.566521220   
## 97 98 99 100 101 102   
## -0.920231170 0.170956016 1.271574367 0.253489836 -2.205345965 0.492982465   
## 103 104 105 106 107 108   
## -2.840340793 -0.051930368 0.431901157 1.246378970 1.067093217 1.586266534   
## 109 110 111 112 113 114   
## 1.704314319 0.136075615 -1.394089821 0.676180666 0.244667984 -0.483990226   
## 115 116 117 118 119 120   
## -0.697256259 -0.337084458 0.221512385 2.832836833 0.290532867 -0.216650951   
## 121 122 123 124 125 126   
## 1.075499440 -0.860765154 -2.021364639 0.141882108 0.205956502 1.470748337   
## 127 128 129 130 131 132   
## -3.990962886 2.209363916 2.487397216 1.795982388 -8.797699657 -2.900460134   
## 133 134 135 136 137 138   
## -2.718883323 0.324185137 -1.066029737 -2.366786133 -1.924198017 -0.077225955   
## 139 140 141 142 143 144   
## -0.157607431 1.065888227 1.424594807 0.761196780 0.848555034 1.621379336   
## 145 146 147 148 149 150   
## 1.294972321 0.602310498 -2.079188873 2.139611951 -1.335950219 0.283408806   
## 151 152 153 154 155 156   
## -2.277596261 1.563415957 0.257483143 0.777729464 0.119467630 -2.089427683   
## 157 158 159 160 161 162   
## -0.095225908 0.080436291 -3.093418211 0.493897169 0.183498981 -0.272480880   
## 163 164 165 166 167 168   
## -0.044002579 0.679800279 0.852920626 -2.389784424 -2.808694024 -1.160765654   
## 169 170 171 172 173 174   
## -0.113350833 -2.921932648 1.010426304 0.123153667 0.003421757 -0.260969784   
## 175 176 177 178 179 180   
## -2.236151161 2.216474104 0.235978373 -0.474924410 -4.213843972 0.221960441   
## 181 182 183 184 185 186   
## -0.075088946 -1.733695836 2.135912387 2.036063520 -0.937949012 1.820623371   
## 187 188 189 190 191 192   
## 0.601315505 0.239720618 -2.720096781 0.648554903 -1.654977823 1.494073900   
## 193 194 195 196 197 198   
## 1.421140939 1.151239411 0.836809799 2.235487512 1.547624795 0.031951507   
## 199 200   
## 1.707077007 -1.757542846

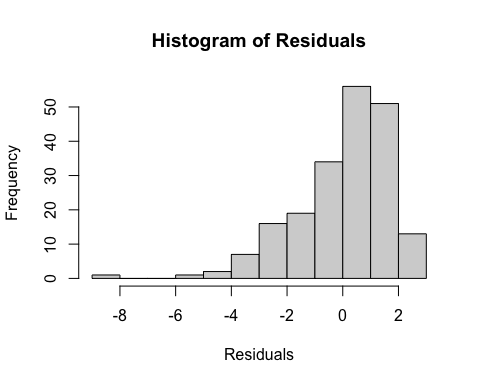
### Part(g) - Plot the residuals against the predicted values

plot(pred\_values, resid\_value, xlab="Predicted Values", ylab="Residual Values")



### Part(h) - Plot the histogram of the residuals

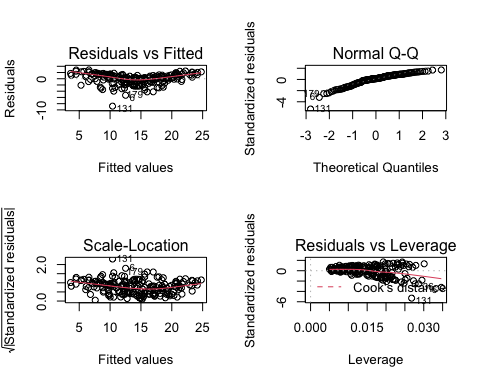
hist(resid\_value, xlab = "Residuals", main="Histogram of Residuals")



Here, we can see that the residual values are not normally distributed and are negatively skewed.

### Part(i) - Comment on the residual plots

par(mfrow=c(2,2))  
plot(model2)



From the above 4 plots, we can see that plot for Residuals vs Fitted are forming a pattern and are not scattered thus violating our **Linearity Assumption**.

In second plot, we can see that the points are not accurately lying over the straight line, this violates our **Normality Assumption**.

In Third plot, we can see that the spread of the plot is not constant throughout. This violates our **Heteroscedasticity Assumption**.

### Part(j) - Use the multivariate model for prediction

predict(model2, as.data.frame(cbind(TV = 400, Radio = 60)))

## 1   
## 32.50268

We can see that, the model predicted a sale of 32.50268.

## Question 2

### Part(a) - Add the Interaction Term TV\*Radio and test the significance of the interaction term

model3 <- lm(Sales~TV+Radio+TV\*Radio)  
summary(model3)

##   
## Call:  
## lm(formula = Sales ~ TV + Radio + TV \* Radio)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.3366 -0.4028 0.1831 0.5948 1.5246   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.750e+00 2.479e-01 27.233 <2e-16 \*\*\*  
## TV 1.910e-02 1.504e-03 12.699 <2e-16 \*\*\*  
## Radio 2.886e-02 8.905e-03 3.241 0.0014 \*\*   
## TV:Radio 1.086e-03 5.242e-05 20.727 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9435 on 196 degrees of freedom  
## Multiple R-squared: 0.9678, Adjusted R-squared: 0.9673   
## F-statistic: 1963 on 3 and 196 DF, p-value: < 2.2e-16

By adding the interactive term TV\*Radio, we can see that becomes 96.78%, that means 96.78% proportion is explained by model.

By doing Hypothesis testing for the interaction term, we can see that it is highly significant at 5% and 1% level of significance. Therefore we can reject the null hypothesis H. And this shows us that the new interaction term shows **strong positive linear relationship**.

### Part(b) - Give the resulting model after considering this interaction term.

Now, after considering the interaction term, the resultant model now become -

= 6.750 + 0.001910 TV + 0.00288 Radio - 0.001086 TV \* Radio

### Part(c) - Construct the Polynomial Regression Model of order 3 and test the model significance.

So, to consider a term for polynomial regression, we can check the correlation matrix of the data set. By looking at the correlation matrix we can see that TV has the highest correlation and thus most **significant** with respect to Sales. Therefore, after considering this our new model becomes:

model4 <- lm(Sales~TV+I(TV\*TV)+I(TV\*TV\*TV))  
summary(model4)

##   
## Call:  
## lm(formula = Sales ~ TV + I(TV \* TV) + I(TV \* TV \* TV))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.9734 -1.8900 -0.0897 2.0189 7.3765   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.420e+00 8.641e-01 6.272 2.23e-09 \*\*\*  
## TV 9.643e-02 2.580e-02 3.738 0.000243 \*\*\*  
## I(TV \* TV) -3.152e-04 2.022e-04 -1.559 0.120559   
## I(TV \* TV \* TV) 5.572e-07 4.494e-07 1.240 0.216519   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.232 on 196 degrees of freedom  
## Multiple R-squared: 0.622, Adjusted R-squared: 0.6162   
## F-statistic: 107.5 on 3 and 196 DF, p-value: < 2.2e-16

Now, for significance we can do hypothesis testing here. And by seeing at summary we can clearly state that p-value for and are 0.120559 and 0.216519 respectively, which is greater than 0.05. So from this we have to reject our alternate hypothesis H at 5% level of significance. Therefore, we can say that this polynomial model of order 3 is **not significant**.

### Part(d) - Give the resulting selected model

So, if we check the value of for all of the 4 models, these are:

1. - 89.72%
2. - 89.72%
3. - 96.78%
4. - 62.2%

Therefore, from above data we can conclude that model 3 is the best model out of these. Thus, the resulting selected model is:

= 6.750 + 0.001910 TV + 0.00288 Radio - 0.001086 TV \* Radio

## Question 3

### Part(a) - Upload the Advertising data set and explore it.

AutoDS <- read.csv("../datasets/auto.csv")  
attach(AutoDS)  
head(AutoDS)

## mpg cylinders displacement horsepower weight acceleration year origin  
## 1 18 8 307 130 3504 12.0 70 1  
## 2 15 8 350 165 3693 11.5 70 1  
## 3 18 8 318 150 3436 11.0 70 1  
## 4 16 8 304 150 3433 12.0 70 1  
## 5 17 8 302 140 3449 10.5 70 1  
## 6 15 8 429 198 4341 10.0 70 1  
## name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

dim(AutoDS)

## [1] 397 9

sapply(AutoDS, class)

## mpg cylinders displacement horsepower weight acceleration   
## "numeric" "integer" "numeric" "character" "integer" "numeric"   
## year origin name   
## "integer" "integer" "character"

From the above output, we can see that Advertising data set has 397 observations of 9 variables, out of which some are numeric, integers and characters.

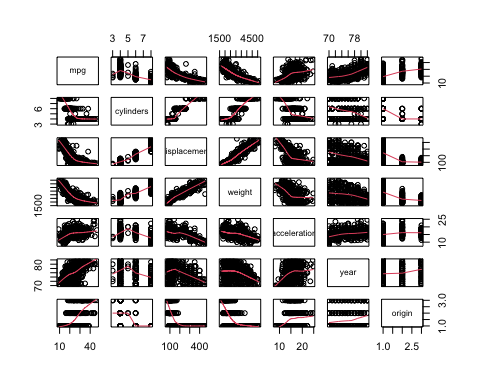
As we only need numeric variables at this point, therefore taking out the subset of only numeric variabes.

AutoDS <- subset(AutoDS, select = c(mpg, cylinders, displacement, weight, acceleration, year, origin))  
head(AutoDS)

## mpg cylinders displacement weight acceleration year origin  
## 1 18 8 307 3504 12.0 70 1  
## 2 15 8 350 3693 11.5 70 1  
## 3 18 8 318 3436 11.0 70 1  
## 4 16 8 304 3433 12.0 70 1  
## 5 17 8 302 3449 10.5 70 1  
## 6 15 8 429 4341 10.0 70 1

### Part(b) - Find the Covariance and Correlation Matrix of the variables.

pairs(AutoDS, panel = panel.smooth)



cor(AutoDS)

## mpg cylinders displacement weight acceleration  
## mpg 1.0000000 -0.7762599 -0.8044430 -0.8317389 0.4222974  
## cylinders -0.7762599 1.0000000 0.9509199 0.8970169 -0.5040606  
## displacement -0.8044430 0.9509199 1.0000000 0.9331044 -0.5441618  
## weight -0.8317389 0.8970169 0.9331044 1.0000000 -0.4195023  
## acceleration 0.4222974 -0.5040606 -0.5441618 -0.4195023 1.0000000  
## year 0.5814695 -0.3467172 -0.3698041 -0.3079004 0.2829009  
## origin 0.5636979 -0.5649716 -0.6106643 -0.5812652 0.2100836  
## year origin  
## mpg 0.5814695 0.5636979  
## cylinders -0.3467172 -0.5649716  
## displacement -0.3698041 -0.6106643  
## weight -0.3079004 -0.5812652  
## acceleration 0.2829009 0.2100836  
## year 1.0000000 0.1843141  
## origin 0.1843141 1.0000000

cov(AutoDS)

## mpg cylinders displacement weight acceleration  
## mpg 61.243207 -10.3368388 -657.11264 -5519.0297 9.0882305  
## cylinders -10.336839 2.8953642 168.89278 1294.1927 -2.3586654  
## displacement -657.112642 168.8927785 10895.09741 82583.3601 -156.1980396  
## weight -5519.029673 1294.1927334 82583.36009 718941.3958 -978.1671450  
## acceleration 9.088231 -2.3586654 -156.19804 -978.1671 7.5624741  
## year 16.791242 -2.1769776 -142.43418 -963.3497 2.8707357  
## origin 3.540358 -0.7715251 -51.15522 -395.5422 0.4636561  
## year origin  
## mpg 16.7912418 3.5403582  
## cylinders -2.1769776 -0.7715251  
## displacement -142.4341781 -51.1552172  
## weight -963.3496870 -395.5422423  
## acceleration 2.8707357 0.4636561  
## year 13.6161362 0.5458298  
## origin 0.5458298 0.6440857

By Looking at the plot for mpg vs cylinders, we can see that there is a **strong negative linear relationship** between them and the correlation matrix table we can see that the value is **-0.7762599** which confirms the same.

By Looking at the plot for mpg vs displacement, we can see that there is a **strong negative linear relationship** between them and the correlation matrix table we can see that the value is **-0.8044430** which confirms the same.

By Looking at the plot for mpg vs weight, we can see that there is a **strong negative linear relationship** between them and the correlation matrix table we can see that the value is **-0.8317389** which confirms the same.

For mpg vs acceleration, there exists a **moderate positive linear relationship** as the correlation between them is **0.4222974**.

For mpg vs year, there exists a **moderate positive linear relationship** as the correlation between them is **0.5814695**.

For mpg vs origin, there exists a **moderate positive linear relationship** as the correlation between them is **0.5636979**.

### Part(c) - Construct the multiple linear regression model and find the least square estimates of the model parameters.

Multiple linear regression includes building a model of multiple variable with respect to our target variable. Here mpg is our target variable and we can first start with including all the numeric variables into our model.

model1 <- lm(mpg~cylinders+displacement+weight+acceleration+year+origin)  
summary(model1)

##   
## Call:  
## lm(formula = mpg ~ cylinders + displacement + weight + acceleration +   
## year + origin)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.5573 -2.1745 -0.0456 1.8454 12.9946   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.014e+01 4.145e+00 -4.858 1.72e-06 \*\*\*  
## cylinders -4.198e-01 3.203e-01 -1.311 0.1908   
## displacement 1.742e-02 7.189e-03 2.423 0.0158 \*   
## weight -6.928e-03 5.781e-04 -11.983 < 2e-16 \*\*\*  
## acceleration 1.591e-01 7.741e-02 2.055 0.0405 \*   
## year 7.703e-01 4.934e-02 15.613 < 2e-16 \*\*\*  
## origin 1.356e+00 2.691e-01 5.040 7.16e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.333 on 390 degrees of freedom  
## Multiple R-squared: 0.8214, Adjusted R-squared: 0.8186   
## F-statistic: 298.9 on 6 and 390 DF, p-value: < 2.2e-16

By looking at the summary for the above model, we can see that the least square estimates of our model are as follows:

1. = -0.2014
2. = -0.4198
3. = 0.001742
4. = -0.0006928
5. = 0.01591
6. = 0.07703
7. = 1.356

Now, substituting the value of these estimates in our equation:

= - 0.2014 - 0.4198 cylinders + 0.001742 displacement - 0.0006928 weight + 0.062462 acceleration + 0.07703 year + 1.356 origin

### Part(d) - Test the significance of the parameters and find the resulting model to model mpg in terms of advertising modes, displacement and acceleration.

To test the significance of the parameters of our model, we will use Hypothesis testing, H: = 0 and H: 0.

For cylinders we can see that p-value of acceleration is 0.1908 which is not less than 0.05.This shows that there is a **week linear relationship**. Therefore, we can not reject the null hypothesis H here, we have to reject the alternate hypothesis H here.

From the summary of our model, we can see that p-value of displacement is less than 0.05. Therefore, we get enough evidence to reject the null hypothesis H at 5% significance level. Moreover, we can also see that p-value of displacement is not less than 0.01 thus depicting a **moderate strong linear relationship**.

From the summary of our model, we can see that p-value of weight is less than 0.05. Therefore, we get enough evidence to reject the null hypothesis H at 5% significance level. Moreover, we can also see that p-value of weight is also less than 0.01 thus depicting a **strong linear relationship**. Therefore, we can reject the null hypothesis H at 1% level of significance as well.

From the summary of our model, we can see that p-value of acceleration is less than 0.05. Therefore, we get enough evidence to reject the null hypothesis H at 5% significance level. Moreover, we can also see that p-value of acceleration is not less than 0.01 thus depicting a **moderate strong linear relationship**.

From the summary of our model, we can see that p-value of year is less than 0.05. Therefore, we get enough evidence to reject the null hypothesis H at 5% significance level. Moreover, we can also see that p-value of year is also less than 0.01 thus depicting a **strong linear relationship**. Therefore, we can reject the null hypothesis H at 1% level of significance as well.

From the summary of our model, we can see that p-value of origin is less than 0.05. Therefore, we get enough evidence to reject the null hypothesis H at 5% significance level. Moreover, we can also see that p-value of origin is also less than 0.01 thus depicting a **strong linear relationship**. Therefore, we can reject the null hypothesis H at 1% level of significance as well.

Now, after removing week relationship our resultant model becomes:

model2 <- lm(mpg~displacement+weight+acceleration+year+origin)  
summary(model2)

##   
## Call:  
## lm(formula = mpg ~ displacement + weight + acceleration + year +   
## origin)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3199 -2.1634 -0.1032 1.8140 12.9334   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.090e+01 4.108e+00 -5.089 5.61e-07 \*\*\*  
## displacement 1.115e-02 5.370e-03 2.077 0.0385 \*   
## weight -6.984e-03 5.771e-04 -12.102 < 2e-16 \*\*\*  
## acceleration 1.551e-01 7.742e-02 2.003 0.0459 \*   
## year 7.697e-01 4.938e-02 15.589 < 2e-16 \*\*\*  
## origin 1.330e+00 2.686e-01 4.951 1.10e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.336 on 391 degrees of freedom  
## Multiple R-squared: 0.8206, Adjusted R-squared: 0.8183   
## F-statistic: 357.7 on 5 and 391 DF, p-value: < 2.2e-16

Now, by looking at the summary of our new model, we can see that through hypothesis testing, there are some mix of moderate and strong relationship.

= -0.2099 + 0.001115 displacement - 0.0006984 weight + 0.01551 acceleration + 0.07697 year + 1.330 origin.

### Part(e) - Assess the overall accuracy of the model.

After looking at the summary of our new model, we can see that Residual standard error is 3.351 We can check the mpg summary and then we can compare the error.

summary(mpg)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 9.00 17.50 23.00 23.52 29.00 46.60

So, for the summary we can observe that the Residual standard error or standard deviation seems to be low as compared to the median of mpg, which is good.

Alternatively we can also check value from the summary which is 81.8%. That means 81.8% proportion is explained by the model.

And at the end, if we look at the p-value, it is very small, which is good.

### Part(f) - Calculate the predicted vales and residuals

Below are the predicted values for our model.

pred\_values <- predict(model2)  
pred\_values

## 1 2 3 4 5 6 7 8   
## 15.124126 14.206207 15.566602 15.586502 15.219866 10.329230 10.362156 10.421816   
## 9 10 11 12 13 14 15 16   
## 10.032543 13.090644 15.249451 14.138583 13.978752 19.383462 23.990843 19.137322   
## 17 18 19 20 21 22 23 24   
## 19.560500 20.955100 25.424902 27.085679 20.920251 22.111612 22.927446 23.326365   
## 25 26 27 28 29 30 31 32   
## 20.362889 8.266584 9.499701 9.347864 7.522845 26.194651 23.233925 25.611152   
## 33 34 35 36 37 38 39 40   
## 24.830713 21.288263 15.976143 17.023106 17.211660 17.108710 11.449997 10.149223   
## 41 42 43 44 45 46 47 48   
## 12.077839 12.037364 6.530752 8.257406 5.505903 19.365132 22.771027 17.273798   
## 49 50 51 52 53 54 55 56   
## 18.194907 23.107883 25.048125 25.830585 25.218471 29.095629 30.069076 27.629814   
## 57 58 59 60 61 62 63 64   
## 25.620752 26.264323 24.726075 26.164281 23.618309 24.223393 11.765818 11.548202   
## 65 66 67 68 69 70 71 72   
## 12.612289 12.944643 15.379409 9.984608 10.406175 10.727417 11.367345 25.111555   
## 73 74 75 76 77 78 79 80   
## 13.998102 12.825549 11.711156 13.094865 20.294519 23.784288 20.737452 25.754201   
## 81 82 83 84 85 86 87 88   
## 22.965648 26.248668 24.595075 24.156148 27.484036 13.905762 16.149158 14.687916   
## 89 90 91 92 93 94 95 96   
## 14.008152 15.727069 8.604150 11.766254 12.080251 12.824785 10.164701 8.823529   
## 97 98 99 100 101 102 103 104   
## 15.655542 19.891462 19.306425 21.121084 20.868589 21.028272 28.669355 8.354173   
## 105 106 107 108 109 110 111 112   
## 8.757075 10.148409 11.041810 22.055447 27.391399 24.436942 26.428047 27.319906   
## 113 114 115 116 117 118 119 120   
## 24.716657 23.255522 25.627843 14.031466 12.677524 28.693008 26.575798 23.359890   
## 121 122 123 124 125 126 127 128   
## 21.673261 18.134238 22.893234 23.509162 16.640443 20.492818 22.177910 22.198108   
## 129 130 131 132 133 134 135 136   
## 19.516065 30.258003 24.191602 31.275045 23.834354 16.408404 17.693220 17.225324   
## 137 138 139 140 141 142 143 144   
## 14.009001 10.724987 11.903097 10.848327 13.453813 26.873899 28.294742 25.986953   
## 145 146 147 148 149 150 151 152   
## 31.938924 29.932481 25.801261 27.404795 26.498401 26.330826 26.958929 28.113884   
## 153 154 155 156 157 158 159 160   
## 20.354785 19.271778 20.235657 21.916542 11.803583 13.225932 12.541587 11.644129   
## 161 162 163 164 165 166 167 168   
## 16.706620 16.600663 17.933650 17.181572 21.837916 20.680003 21.256591 29.219921   
## 169 170 171 172 173 174 175 176   
## 23.926703 22.877071 24.487525 25.536605 27.526505 27.008477 21.475413 29.136126   
## 177 178 179 180 181 182 183 184   
## 20.958034 24.283451 22.812673 22.519964 24.278381 32.011412 26.647731 28.669158   
## 185 186 187 188 189 190 191 192   
## 24.838710 27.018320 28.379491 14.910108 15.229659 16.805343 15.392044 21.247984   
## 193 194 195 196 197 198 199 200   
## 20.549181 22.853706 22.700744 29.107520 28.336105 29.936888 32.765654 18.685526   
## 201 202 203 204 205 206 207 208   
## 20.013748 18.773607 22.267466 30.486969 31.274937 30.163432 24.686010 22.144531   
## 209 210 211 212 213 214 215 216   
## 17.006550 22.156402 25.269567 18.044386 14.120068 16.374202 17.596095 18.422551   
## 217 218 219 220 221 222 223 224   
## 32.038155 28.181695 32.048414 27.400293 32.327931 17.941800 17.190892 16.457129   
## 225 226 227 228 229 230 231 232   
## 15.382346 20.447308 20.976403 19.601929 20.815559 16.409668 16.282726 15.846970   
## 233 234 235 236 237 238 239 240   
## 15.587382 30.810264 24.728463 30.444112 24.470035 29.104159 28.765983 32.120378   
## 241 242 243 244 245 246 247 248   
## 29.002360 26.575844 26.205163 26.348241 32.273152 31.223600 33.143451 32.503858   
## 249 250 251 252 253 254 255 256   
## 34.115169 22.271453 19.977667 20.889503 21.334615 23.487601 24.442312 25.422181   
## 257 258 259 260 261 262 263 264   
## 21.690848 23.305277 21.889839 23.848602 20.596578 21.872867 21.997591 21.063754   
## 265 266 267 268 269 270 271 272   
## 23.190380 17.645889 29.070091 28.946055 30.672034 28.314255 29.353352 25.627595   
## 273 274 275 276 277 278 279 280   
## 24.943212 29.969779 25.960916 23.796213 26.062816 22.251809 31.203655 31.884764   
## 281 282 283 284 285 286 287 288   
## 23.540335 25.409629 25.299357 23.845998 22.856403 20.210327 20.669847 19.579029   
## 289 290 291 292 293 294 295 296   
## 20.394113 17.003167 19.058233 21.365685 19.753129 32.287773 33.420971 31.190245   
## 297 298 299 300 301 302 303 304   
## 26.267209 23.073337 20.603244 25.708191 23.696044 29.091925 29.704711 33.715822   
## 305 306 307 308 309 310 311 312   
## 30.987001 26.756806 26.797088 26.311923 27.118745 31.737039 34.831292 30.698946   
## 313 314 315 316 317 318 319 320   
## 34.069523 27.548220 26.332828 25.836813 23.805135 31.589185 29.632628 30.966856   
## 321 322 323 324 325 326 327 328   
## 31.322261 32.410826 33.666622 26.426351 33.857055 33.145312 31.709564 27.171142   
## 329 330 331 332 333 334 335 336   
## 25.649490 34.902323 34.153148 33.560367 33.817787 27.986280 30.485966 29.580561   
## 337 338 339 340 341 342 343 344   
## 25.499158 32.504204 29.328400 28.602753 28.453131 27.630314 29.627486 36.682394   
## 345 346 347 348 349 350 351 352   
## 33.185401 36.545727 34.857784 35.600591 34.795795 35.070441 30.790281 32.101004   
## 353 354 355 356 357 358 359 360   
## 30.459321 32.186071 31.470564 33.429465 32.835080 30.796982 31.211190 26.286063   
## 361 362 363 364 365 366 367 368   
## 26.695461 29.011942 28.743191 23.954663 23.612960 26.289710 23.662633 29.643324   
## 369 370 371 372 373 374 375 376   
## 29.243836 30.861757 29.325608 29.900245 28.922255 27.643633 34.592929 35.901993   
## 377 378 379 380 381 382 383 384   
## 36.193047 32.157532 32.482644 34.708860 34.249638 34.353604 35.824795 35.933341   
## 385 386 387 388 389 390 391 392   
## 35.801368 27.542146 28.049577 29.482810 28.615427 31.356774 30.517494 27.312255   
## 393 394 395 396 397   
## 28.043344 34.898298 30.824167 29.437797 28.887264

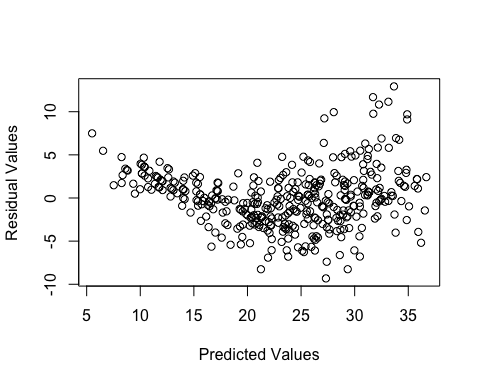
Below are the residuals values for our model.

resid\_value <- resid(model2)  
resid\_value

## 1 2 3 4 5 6   
## 2.875873813 0.793793460 2.433398309 0.413497845 1.780133597 4.670770372   
## 7 8 9 10 11 12   
## 3.637844123 3.578184230 3.967457271 1.909356434 -0.249450782 -0.138583236   
## 13 14 15 16 17 18   
## 1.021248368 -5.383462212 0.009157249 2.862678455 -1.560499691 0.044900439   
## 19 20 21 22 23 24   
## 1.575097564 -1.085679103 4.079749454 1.888388227 2.072553516 2.673635228   
## 25 26 27 28 29 30   
## 0.637110571 1.733416248 0.500299369 1.652136066 1.477154660 0.805348582   
## 31 32 33 34 35 36   
## 4.766074729 -0.611152164 0.169287328 -2.288263046 0.023857360 -0.023105549   
## 37 38 39 40 41 42   
## 1.788339689 0.891289880 2.550002998 3.850776770 1.922161442 1.962635794   
## 43 44 45 46 47 48   
## 5.469248283 4.742594021 7.494096842 -1.365132265 -0.771027257 1.726201978   
## 49 50 51 52 53 54   
## -0.194907425 -0.107882922 2.951874588 4.169415025 4.781528606 1.904371188   
## 55 56 57 58 59 60   
## 4.930923529 -0.629814145 0.379248012 -2.264323110 0.273924590 -3.164281403   
## 61 62 63 64 65 66   
## -3.618308722 -3.223393262 1.234182146 2.451798038 2.387711206 1.055357200   
## 67 68 69 70 71 72   
## 1.620591181 1.015392273 2.593824907 1.272583461 1.632655414 -6.111554506   
## 73 74 75 76 77 78   
## 1.001898348 0.174450675 1.288844011 0.905135161 -2.294519391 -1.784287865   
## 79 80 81 82 83 84   
## 0.262548297 0.245798854 -0.965647640 1.751331716 -1.595075449 3.843852354   
## 85 86 87 88 89 90   
## -0.484035624 -0.905762490 -2.149157801 -1.687915576 -0.008151965 -0.727069281   
## 91 92 93 94 95 96   
## 3.395850400 1.233746322 0.919749487 1.175215247 2.835299469 3.176471458   
## 97 98 99 100 101 102   
## -2.655542499 -1.891461655 -3.306424924 -3.121084394 -2.868589467 1.971728213   
## 103 104 105 106 107 108   
## -2.669354918 2.645827058 3.242925125 2.851591145 0.958190362 -4.055446941   
## 109 110 111 112 113 114   
## -7.391398786 -3.436942273 -4.428047211 -9.319906485 -5.716657362 -2.255521875   
## 115 116 117 118 119 120   
## 0.372156877 0.968534336 3.322475934 0.306992363 -2.575797986 -3.359890278   
## 121 122 123 124 125 126   
## -2.673261470 -3.134238497 1.106765962 -3.509161699 -5.640442785 -0.492818332   
## 127 128 129 130 131 132   
## -1.177909664 -3.198107803 -4.516065378 0.741996867 1.808398455 0.724955419   
## 133 134 135 136 137 138   
## 1.165646028 -0.408403563 -1.693220063 0.774676133 1.990998996 2.275013387   
## 139 140 141 142 143 144   
## 2.096903364 3.151673382 0.546186524 2.126101483 -2.294741634 0.013046774   
## 145 146 147 148 149 150   
## -0.938923914 2.067518587 2.198738584 -3.404794538 -0.498400666 -2.330825562   
## 151 152 153 154 155 156   
## -0.958929111 2.886115742 -1.354785250 -1.271777701 -5.235657294 -6.916541502   
## 157 158 159 160 161 162   
## 4.196416818 1.774067876 3.458413313 2.355870889 0.293380212 -0.600662870   
## 163 164 165 166 167 168   
## -2.933650062 0.818428403 -0.837916407 -0.680003477 -8.256591479 -0.219920899   
## 169 170 171 172 173 174   
## -0.926702514 -2.877071159 -1.487524919 -1.536604561 -2.526504872 -3.008476860   
## 175 176 177 178 179 180   
## -3.475412537 -0.136126229 -1.958033746 -1.283451197 0.187326552 -0.519964222   
## 181 182 183 184 185 186   
## 0.721619087 0.988588447 1.352268697 -3.669158285 0.161289783 -1.018319602   
## 187 188 189 190 191 192   
## -1.379491409 2.589892494 0.770340793 -1.305342856 -0.892043902 0.752015947   
## 193 194 195 196 197 198   
## 1.450818738 1.146294220 -0.200743930 -0.107520125 -3.836104836 -0.936888205   
## 199 200 201 202 203 204   
## 0.234345962 1.314473578 -2.013747900 -0.273606873 -4.767465791 -0.986968874   
## 205 206 207 208 209 210   
## 0.725062781 -2.163432023 1.813989671 -2.144531339 -4.006549624 -3.156402230   
## 211 212 213 214 215 216   
## -6.269566886 -1.544385609 2.379931748 -3.374202407 -4.596095063 -5.422550891   
## 217 218 219 220 221 222   
## -0.538154575 1.818304891 3.951585686 -1.900293490 1.172068856 -0.441799754   
## 223 224 225 226 227 228   
## -0.190891974 -0.957129150 -0.382345928 -2.947307561 -0.476403065 -0.601929108   
## 229 230 231 232 233 234   
## -2.315558925 -0.409667504 -0.782726343 -0.346970449 0.412617836 -1.810263664   
## 235 236 237 238 239 240   
## -0.228462844 -4.444111878 1.029964606 1.395840914 4.734016610 -2.120377654   
## 241 242 243 244 245 246   
## 1.497639747 -4.575843780 -4.705162590 -4.848241373 10.826848042 4.876399910   
## 247 248 249 250 251 252   
## -0.343451136 6.896141646 1.984830512 -2.371452813 -0.577667077 -0.689503011   
## 253 254 255 256 257 258   
## -2.134615403 -2.987600838 -4.242311761 -0.322181230 -1.190847547 -3.905277196   
## 259 260 261 262 263 264   
## -1.289838521 -3.048601713 -1.996578154 -3.772866743 -2.797591125 -3.363754472   
## 265 266 267 268 269 270   
## -5.090380104 -0.145888714 0.929909423 -1.446055361 -3.472034260 2.585745067   
## 271 272 273 274 275 276   
## -8.253352278 -2.427595122 -1.143212157 -6.069778736 -5.660916474 -6.796213403   
## 277 278 279 280 281 282   
## -4.462815583 -6.051808711 0.296345442 -2.384764249 -2.040335326 -5.609628920   
## 283 284 285 286 287 288   
## -2.999357000 -3.645998111 -2.256403228 -3.210326509 -3.069846970 -3.079029357   
## 289 290 291 292 293 294   
## -2.194112570 -0.103167299 -3.558233360 -2.165684501 -1.253128672 -0.387773201   
## 295 296 297 298 299 300   
## 0.679029054 4.509754543 1.132790690 2.326662642 2.396755825 1.491809098   
## 301 302 303 304 305 306   
## 0.203955964 5.108075250 4.795289323 -1.915821801 6.312998663 1.643193513   
## 307 308 309 310 311 312   
## 2.002912006 0.488076578 6.381255313 9.762961300 3.268708404 1.401053586   
## 313 314 315 316 317 318   
## 3.130476539 0.451780125 0.067171851 -1.536813102 -4.705134936 2.710815338   
## 319 320 321 322 323 324   
## 0.167372182 0.333144142 5.677738584 -0.210825737 12.933378472 1.473649369   
## 325 326 327 328 329 330   
## 6.942945089 11.154688054 11.690435545 9.228857597 4.350510228 9.697676944   
## 331 332 333 334 335 336   
## 6.746851850 0.239633004 -4.017787416 4.713720468 -6.785965526 5.419438997   
## 337 338 339 340 341 342   
## -1.899158438 -0.104203860 -2.128399528 -2.002753278 -2.653131359 -4.130314154   
## 343 344 345 346 347 348   
## 0.372513859 2.417606029 5.814599142 -1.445726597 -2.557783761 1.399409305   
## 349 350 351 352 353 354   
## 2.904204666 -0.970441155 3.909719488 2.298995900 -0.559320700 0.813928731   
## 355 356 357 358 359 360   
## 3.029436414 0.270535294 -0.435080341 2.103017850 0.388809588 1.813937372   
## 361 362 363 364 365 366   
## 4.004539372 -3.611941571 -4.543190619 -1.554662629 2.987039717 -6.089709743   
## 367 368 369 370 371 372   
## -6.062632675 -1.643323570 -2.243835764 3.138243340 1.674392025 -0.900245161   
## 373 374 375 376 377 378   
## -1.922255236 -3.643633219 1.407071250 1.098007325 -5.193046729 5.842467627   
## 379 380 381 382 383 384   
## 3.517356230 1.291140372 1.750362266 -0.353604339 2.175204635 -3.933340842   
## 385 386 387 388 389 390   
## 2.198631966 -2.542146464 9.950423267 -3.482809677 -6.615426846 0.643225980   
## 391 392 393 394 395 396   
## 5.482505734 -0.312255175 -1.043344474 9.101702108 1.175833460 -1.437796999   
## 397   
## 2.112735523

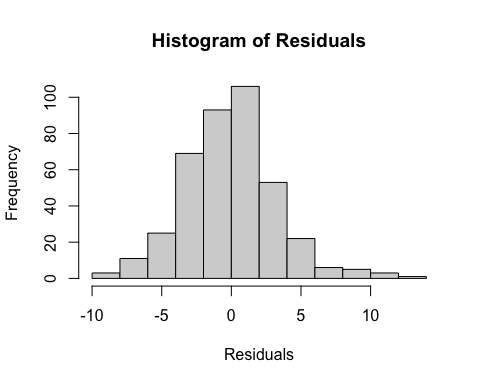
### Part(g) - Plot the residuals against the predicted values

plot(pred\_values, resid\_value, xlab="Predicted Values", ylab="Residual Values")



### Part(h) - Plot the histogram of the residuals

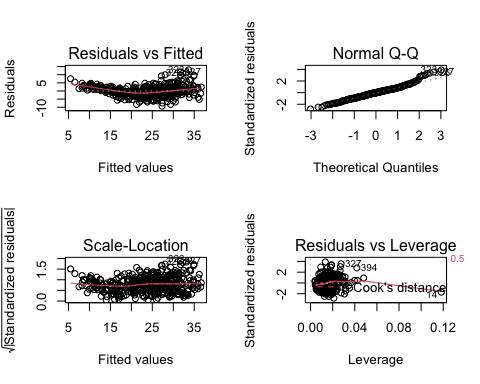
hist(resid\_value, xlab = "Residuals", main="Histogram of Residuals")



Here, we can see that the residual values looks normally distributed and have mean around 0.

### Part(i) - Comment on the residual plots

par(mfrow=c(2,2))  
plot(model2)



From the above 4 plots, we can see that plot for Residuals vs Fitted are forming a pattern and are not scattered thus violating our **Linearity Assumption**.

In second plot, we can see that the points are not accurately lying over the straight line, this violates our **Normality Assumption**.

In Third plot, we can see that the spread of the plot is somewhat constant throughout. Therefore we can say that this doesn’t violates our **Heteroscedasticity Assumption**.

### Part(j) - Use the multivariate model for prediction

predict(model2, as.data.frame(cbind(displacement = 500, weight = 1500, acceleration = 26, year = 85, origin = 3)))

## 1   
## 47.64788

We can see that, the model predicted mpg of 47.64788.