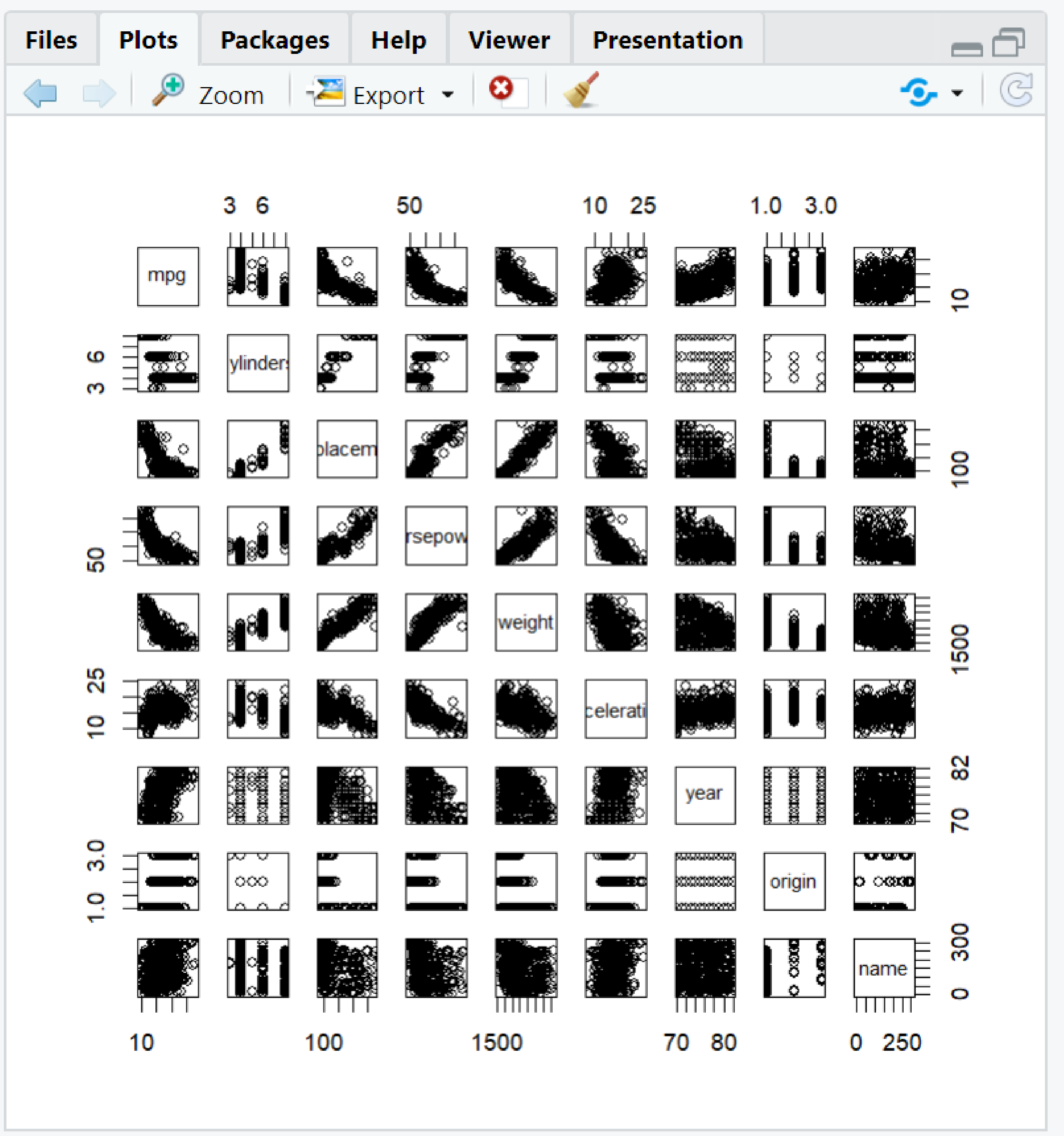
Meher Shrishti Nigam

20BRS1193

EDA LAB – 2 (Q1)

6 / 1 / 23

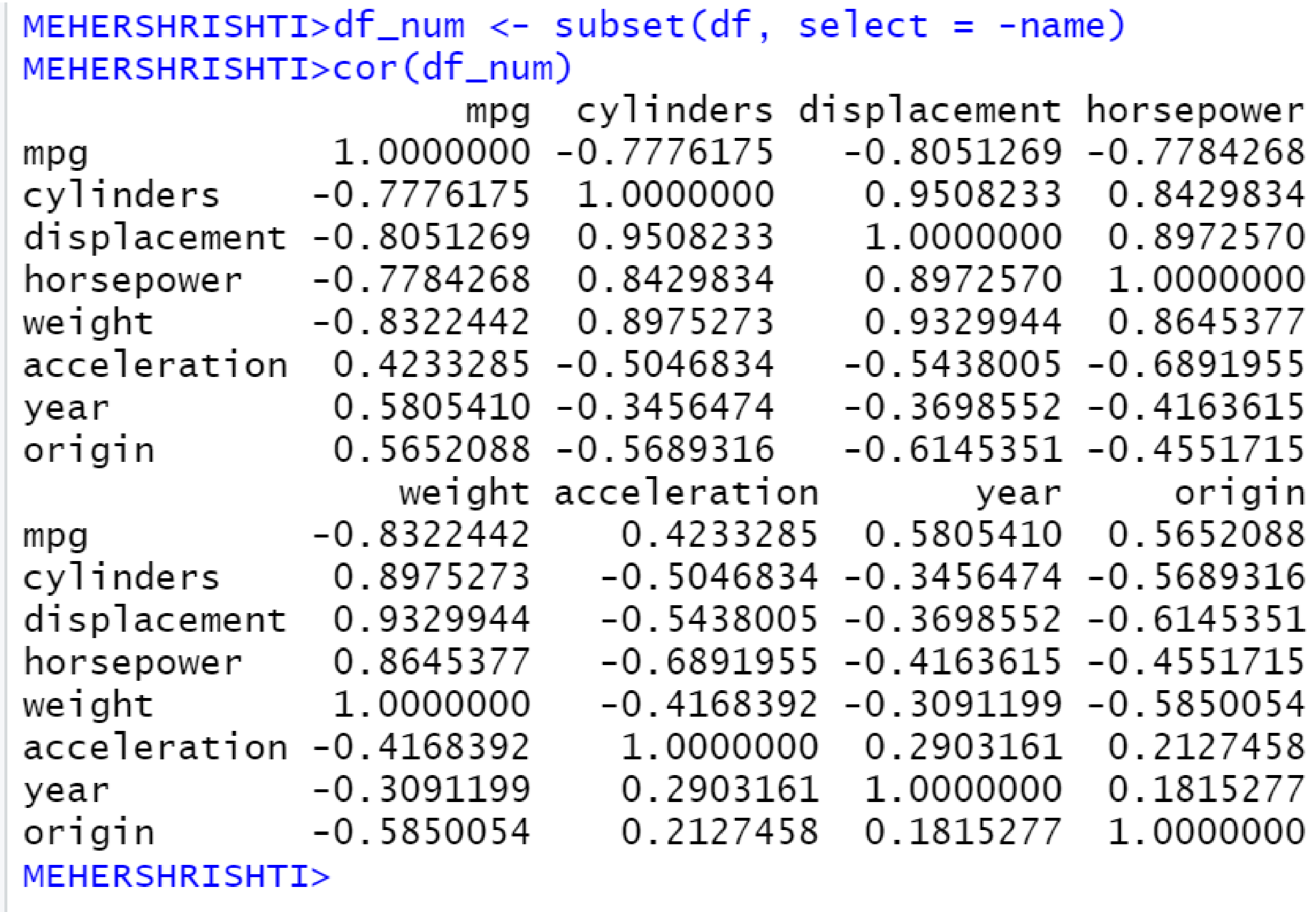
```r
# Meher Shrishti Nigam

# 20BRS1193

# EDA Lab 2

options(prompt="MEHERSHRISHTI>", continue =" ")

# options(prompt=">", continue =" ")

# EDA-LAB-EXPERIMENT-2 (Date-6/1/2023)

library(ISLR)

# Q1. This question involves the use of multiple linear regression on the Auto data set.

df <- Auto

df <- na.omit(df)

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

pairs(df)
```


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
mpg
ylinders
placem
rsepow
weight
celerati
year
origin
name
3 6
50
10 25
1.0 3.0
10
100
1500
70 80
0 250
300
0
82
70

# (b) Compute the matrix of correlations between the variables using the function cor().

df_num <- subset(df, select = -name)

cor(df_num)



```
MEHERSHRISHTI>df_num <- subset(df, select = -name)
MEHERSHRISHTI>cor(df_num)
                    mpg  cylinders displacement horsepower
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834
displacement -0.8051269  0.9508233    1.0000000  0.8972570
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000
weight       -0.8322442  0.8975273    0.9329944  0.8645377
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955
year          0.5805410 -0.3456474   -0.3698552 -0.4163615
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715
                 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
weight        1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration -0.4168392    1.0000000  0.2903161  0.2127458
year         -0.3091199    0.2903161  1.0000000  0.1815277
origin       -0.5850054    0.2127458  0.1815277  1.0000000
MEHERSHRISHTI>
```

# (c) Use the lm() function to perform a multiple linear regression with mpg as

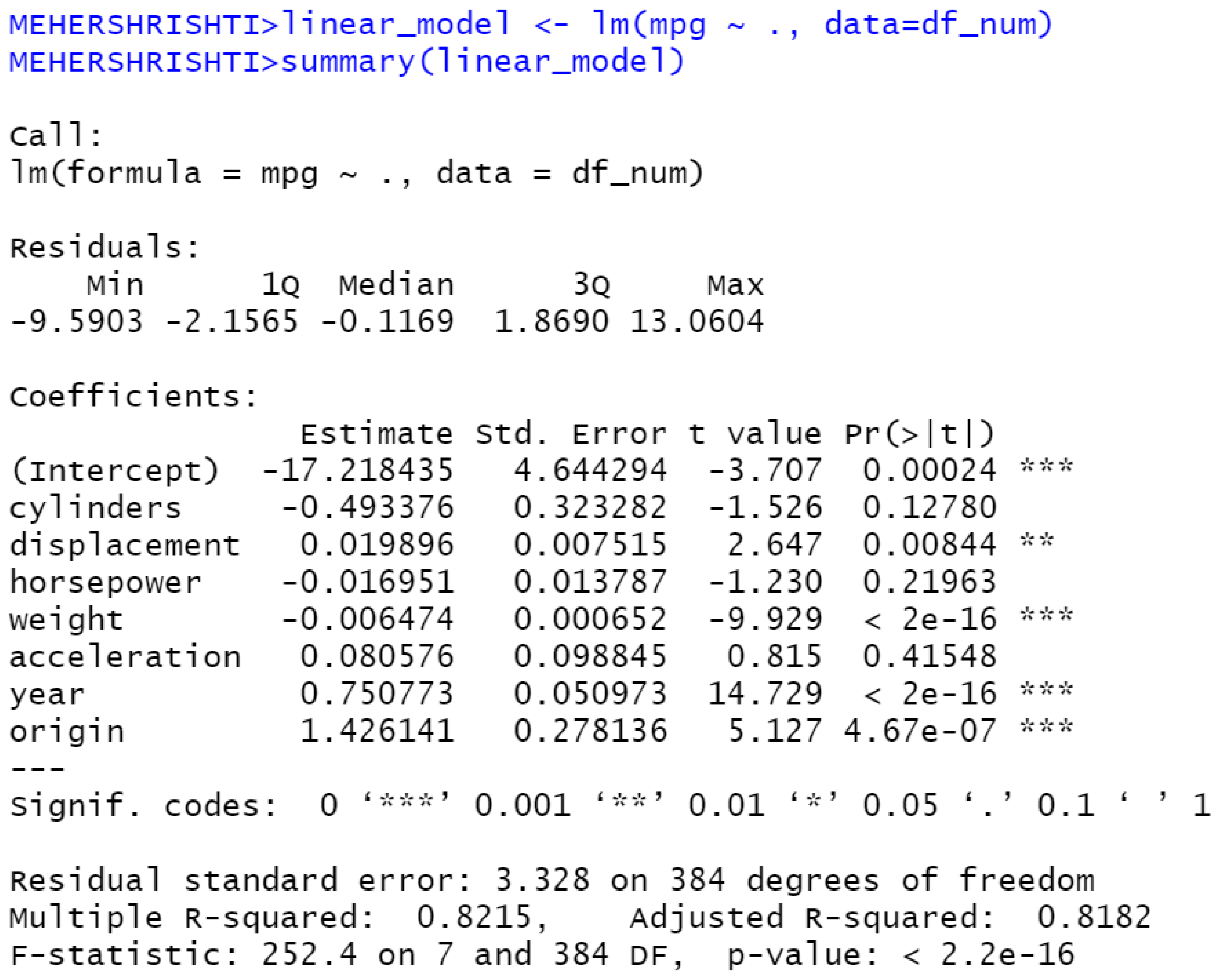
# the response and all other variables except name as the predictors.

# Use the summary() function to print the results.

# Comment on the output. For instance:

linear_model <- lm(mpg ~ ., data=df_num)

summary(linear_model)



```
MEHERSHRISHTI>linear_model <- lm(mpg ~ ., data=df_num)
MEHERSHRISHTI>summary(linear_model)

Call:
lm(formula = mpg ~ ., data = df_num)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5903 -2.1565 -0.1169  1.8690 13.0604 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780    
displacement   0.019896   0.007515   2.647  0.00844 ** 
horsepower    -0.016951   0.013787  -1.230  0.21963    
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548    
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182 
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

# i. Is there a relationship between the predictors and the response?

# We test whether the null hypothesis of all regression coefficients are zero.

# This helps us test whether there is a relationship between predictors and response.

# P-value is low and F-statistic is not close to 1, thus we can refute the null hypothesis.

# ii. Which predictors appear to have a statistically significant relationship to the response?

# Displacement, Weight, Year, Origin have statistically significant relationships with the response.

# Whereas Cylinders, Horsepower, Acceleration do not have a statistically significant relationship.

# This can be determined using their p-values of a predictor's t-statistic.

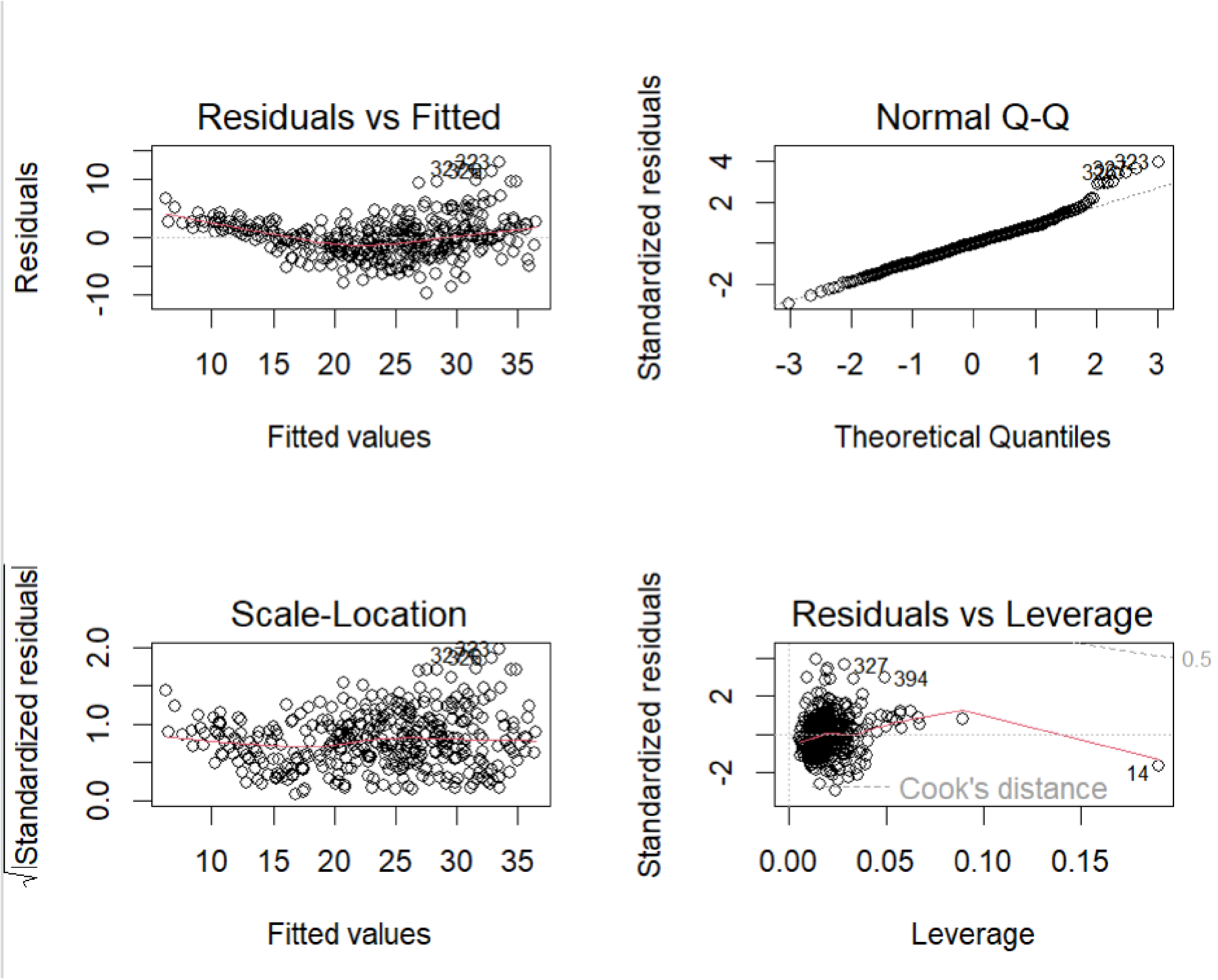
# iii. What does the coefficient for the year variable suggest?

# The coefficient for the year variable is 0.750773.

```
# This tells us that every passing year, mpg (miles per gallon) increases by the coeffcient 0.75 approximately.
```

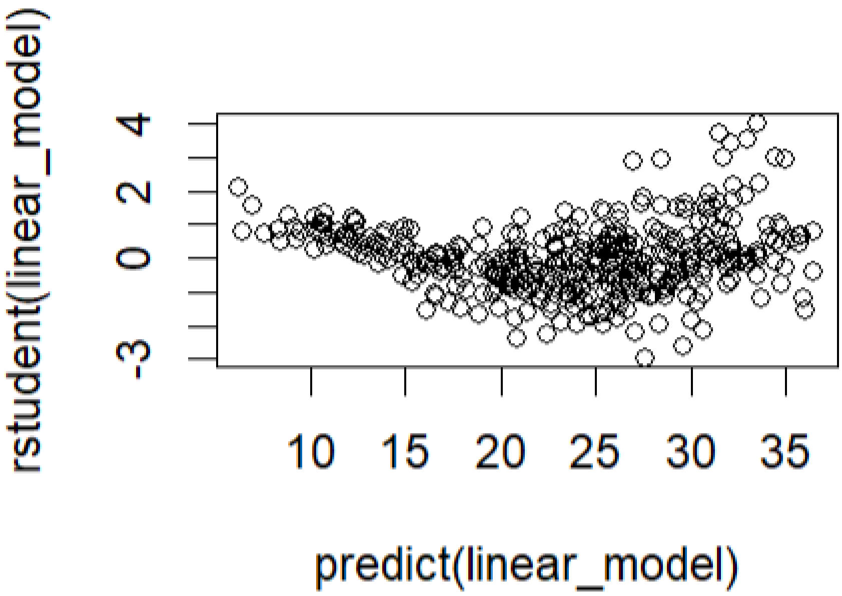
```
# (d) Use the plot() function to produce diagnostic plots of the linear regression

# fit. Comment on any problems you see with the fit.

# Do the residual plots suggest any unusually large outliers?

# Does the leverage plot identifies any observations with unusually high leverage?
```

```
par(mfrow = c(2, 2))

plot(linear_model)
```



```
# The Residuals vs Fitted Plot suggests that a linear model is not the best fit for the given dataset.

# The Residuals vs Fitted Plot does not suggest any unusually large outliers.

# The Residuals vs Leverage plot shows data point 14 has a unusually high leverage. It's residual value is low however.
```

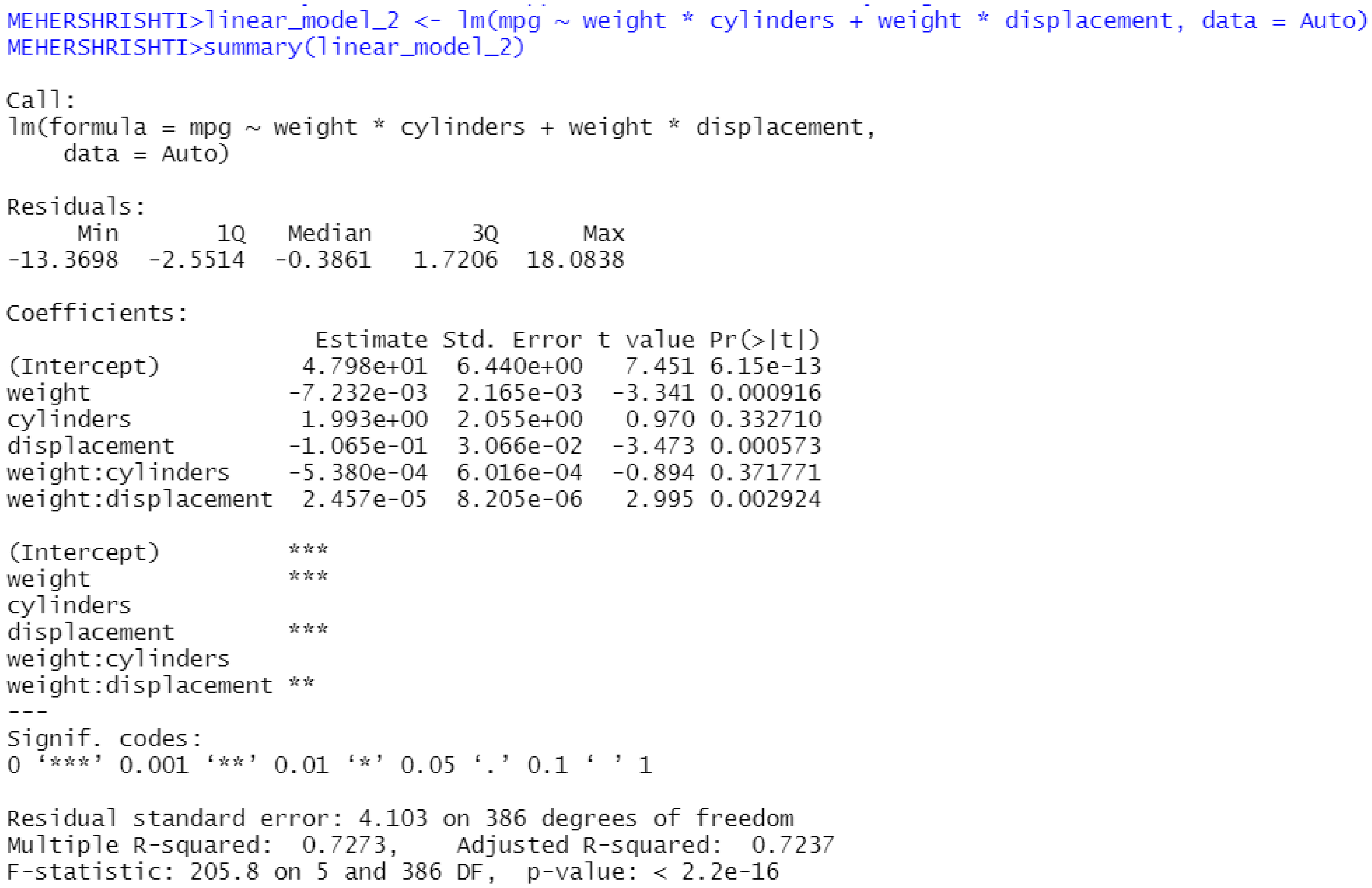
```
plot(predict(linear_model), rstudent(linear_model))
```



# (e) Use the * and : symbols to fit linear regression models with interaction effects.

# Do any interactions appear to be statistically significant?

```
linear_model_2 <- lm(mpg ~ weight * cylinders + weight * displacement, data = Auto)

summary(linear_model_2)
```

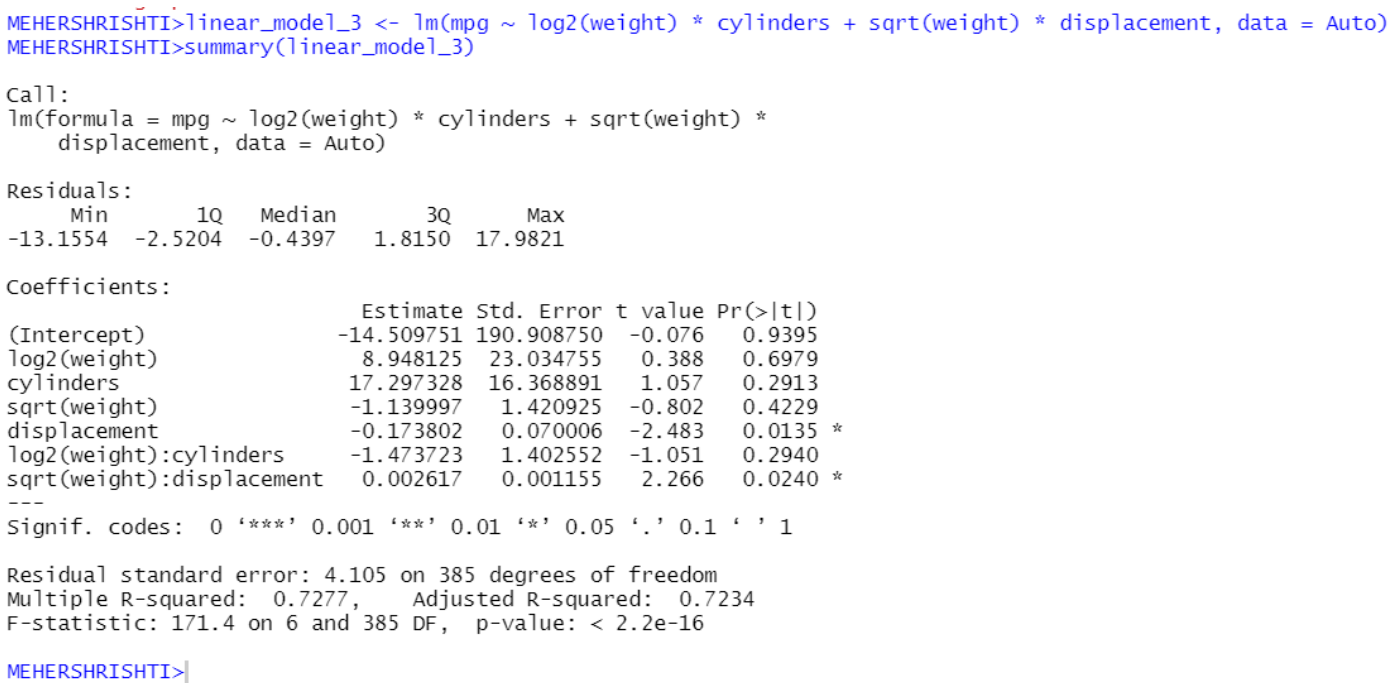


```
MEHERSHRISHTI>linear_model_2 <- lm(mpg ~ weight * cylinders + weight * displacement, data = Auto)
MEHERSHRISHTI>summary(linear_model_2)

Call:
lm(formula = mpg ~ weight * cylinders + weight * displacement,
    data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3698  -2.5514  -0.3861   1.7206  18.0838

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.798e+01  6.440e+00   7.451 6.15e-13
weight              -7.232e-03  2.165e-03  -3.341 0.000916
cylinders            1.993e+00  2.055e+00   0.970 0.332710
displacement        -1.065e-01  3.066e-02  -3.473 0.000573
weight:cylinders    -5.380e-04  6.016e-04  -0.894 0.371771
weight:displacement  2.457e-05  8.205e-06   2.995 0.002924

(Intercept)         ***
weight              ***
cylinders
displacement        ***
weight:cylinders
weight:displacement **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.103 on 386 degrees of freedom
Multiple R-squared:  0.7273,	Adjusted R-squared:  0.7237
F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

# Interaction between weight and displacement is statistically significant, while the interaction between cylinders and weight is not.

# (f) Try a few different transformations of the variables, such as log(X), √X, X2. Comment on your findings.

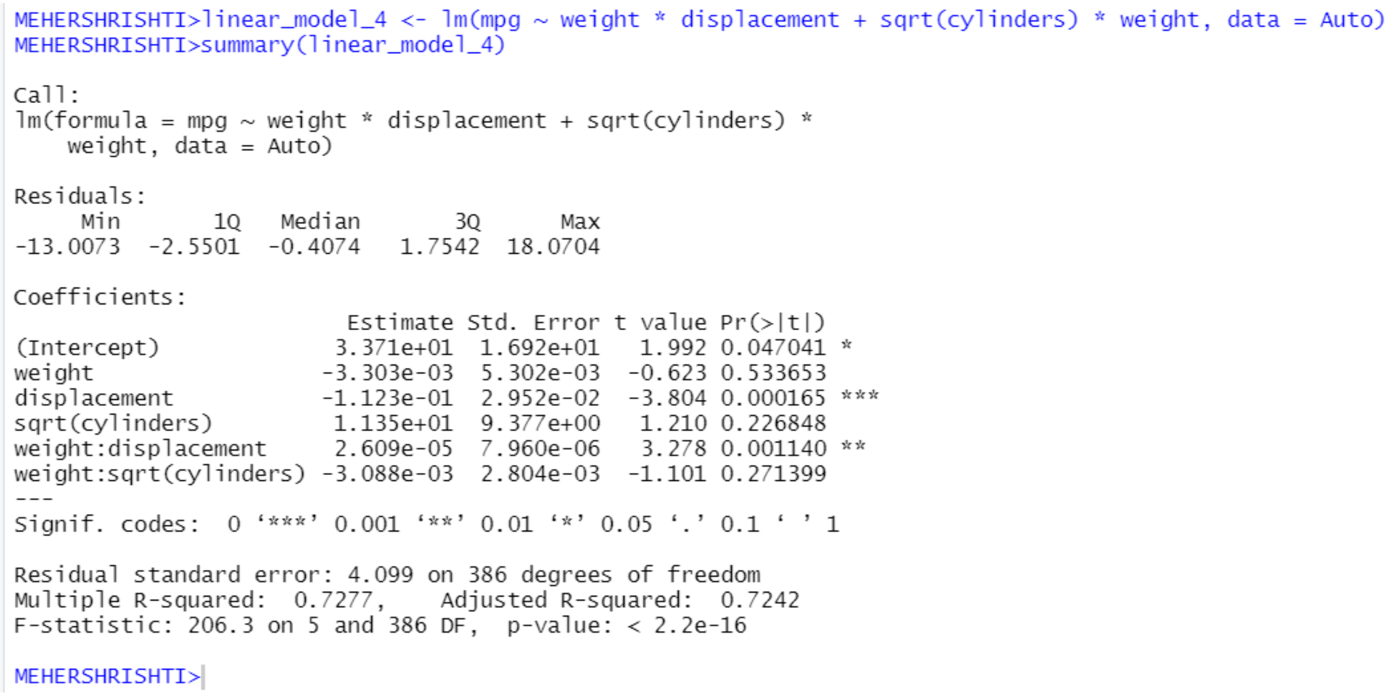
```
linear_model_3 <- lm(mpg ~ log2(weight) * cylinders + sqrt(weight) * displacement, data = Auto)

summary(linear_model_3)
```



```
MEHERSHRISHTI>linear_model_3 <- lm(mpg ~ log2(weight) * cylinders + sqrt(weight) * displacement, data = Auto)
MEHERSHRISHTI>summary(linear_model_3)

Call:
lm(formula = mpg ~ log2(weight) * cylinders + sqrt(weight) * 
    displacement, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.1554  -2.5204  -0.4397   1.8150  17.9821 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)  
(Intercept)              -14.509751 190.908750  -0.076   0.9395  
log2(weight)               8.948125  23.034755   0.388   0.6979  
cylinders                 17.297328  16.368891   1.057   0.2913  
sqrt(weight)              -1.139997   1.420925  -0.802   0.4229  
displacement              -0.173802   0.070006  -2.483   0.0135 *
log2(weight):cylinders    -1.473723   1.402552  -1.051   0.2940  
sqrt(weight):displacement  0.002617   0.001155   2.266   0.0240 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.105 on 385 degrees of freedom
Multiple R-squared:  0.7277,	Adjusted R-squared:  0.7234 
F-statistic: 171.4 on 6 and 385 DF,  p-value: < 2.2e-16

MEHERSHRISHTI>
```

# Interaction between sqrt(weight) and displacement is statistically significant, while the interaction between cylinders and log2(weight) is not.

```
linear_model_4 <- lm(mpg ~ weight * displacement + sqrt(cylinders) * weight, data = Auto)

summary(linear_model_4)
```



```
MEHERSHRISHTI>linear_model_4 <- lm(mpg ~ weight * displacement + sqrt(cylinders) * weight, data = Auto)
MEHERSHRISHTI>summary(linear_model_4)

Call:
lm(formula = mpg ~ weight * displacement + sqrt(cylinders) * 
    weight, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.0073  -2.5501  -0.4074   1.7542  18.0704 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             3.371e+01  1.692e+01   1.992 0.047041 *  
weight                 -3.303e-03  5.302e-03  -0.623 0.533653    
displacement           -1.123e-01  2.952e-02  -3.804 0.000165 ***
sqrt(cylinders)         1.135e+01  9.377e+00   1.210 0.226848    
weight:displacement     2.609e-05  7.960e-06   3.278 0.001140 ** 
weight:sqrt(cylinders) -3.088e-03  2.804e-03  -1.101 0.271399    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.099 on 386 degrees of freedom
Multiple R-squared:  0.7277,	Adjusted R-squared:  0.7242 
F-statistic: 206.3 on 5 and 386 DF,  p-value: < 2.2e-16

MEHERSHRISHTI>
```

# Interaction between weight and displacement is statistically significant, while the interaction between sqrt(cylinders) and weight is not.