ST462/ST662 Advanced Regression Analysis
Assignment 1

Due: 11:59pm on Tuesday, January 24, 2024

ST462: Hand-in Questions 1-2

Hand-in Questions

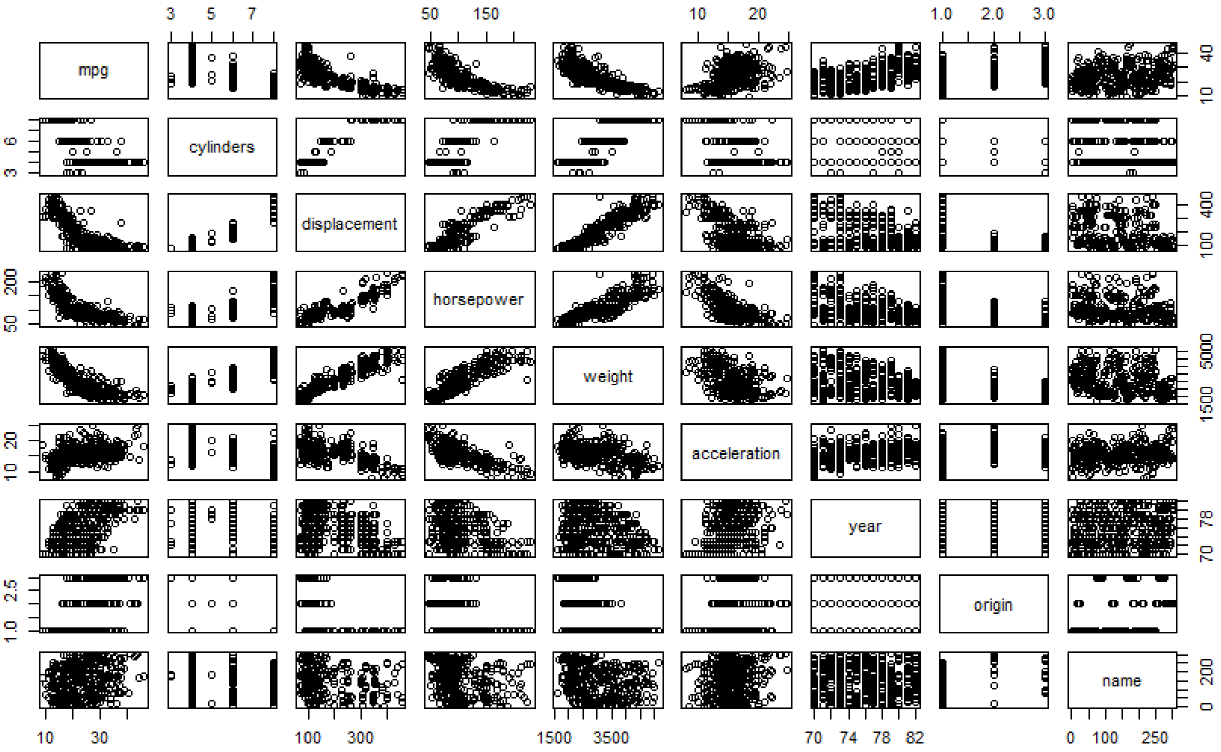
1. This question involves the use of multiple linear regression on the *Auto* data set (install *library(ISLR2)*, and use *data(Auto)* to extract the data set).

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

Command:

```
pairs(auto)
```

Result:



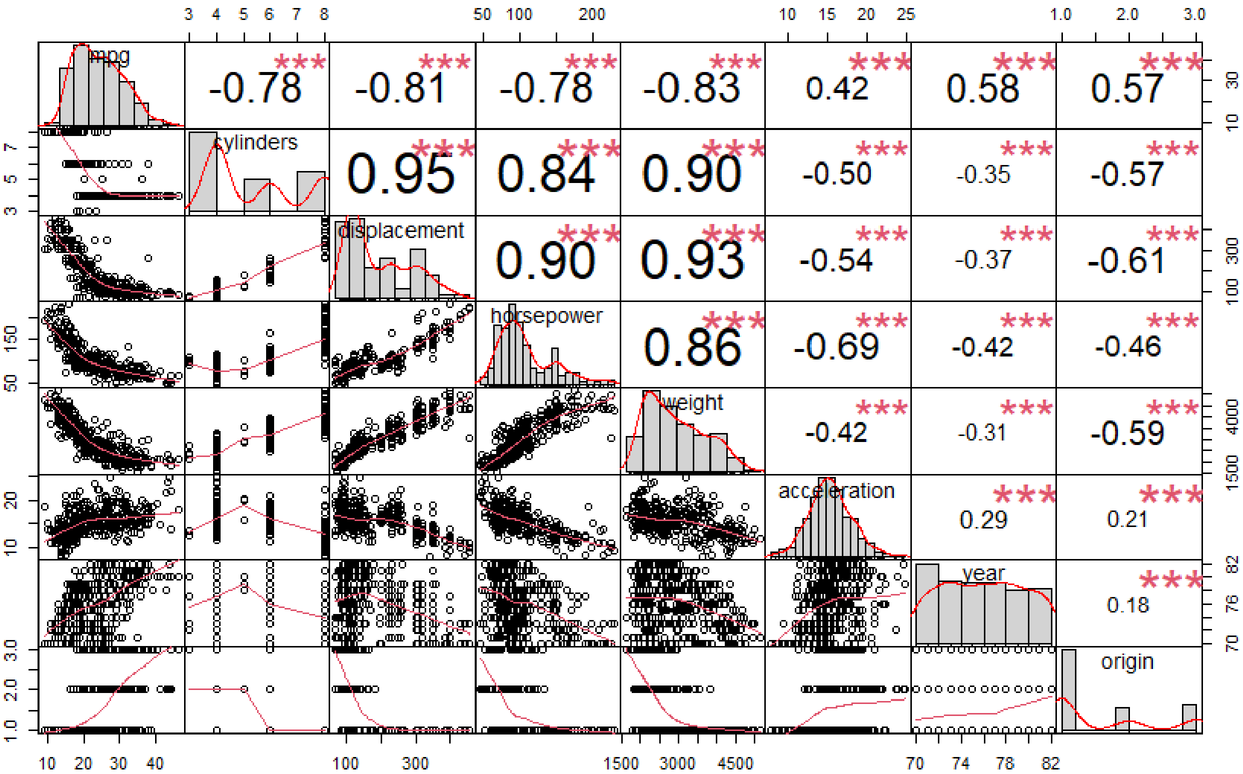
(b). Compute the matrix of correlations between the variables using the function *cor()*. (Hint: exclude variable *name*, which is qualitative.)

Command:

```
chart.Correlation(Auto[, -9])
```

(Dr.Wang said it was okay as long as it produced a correlation matrix)

Results:

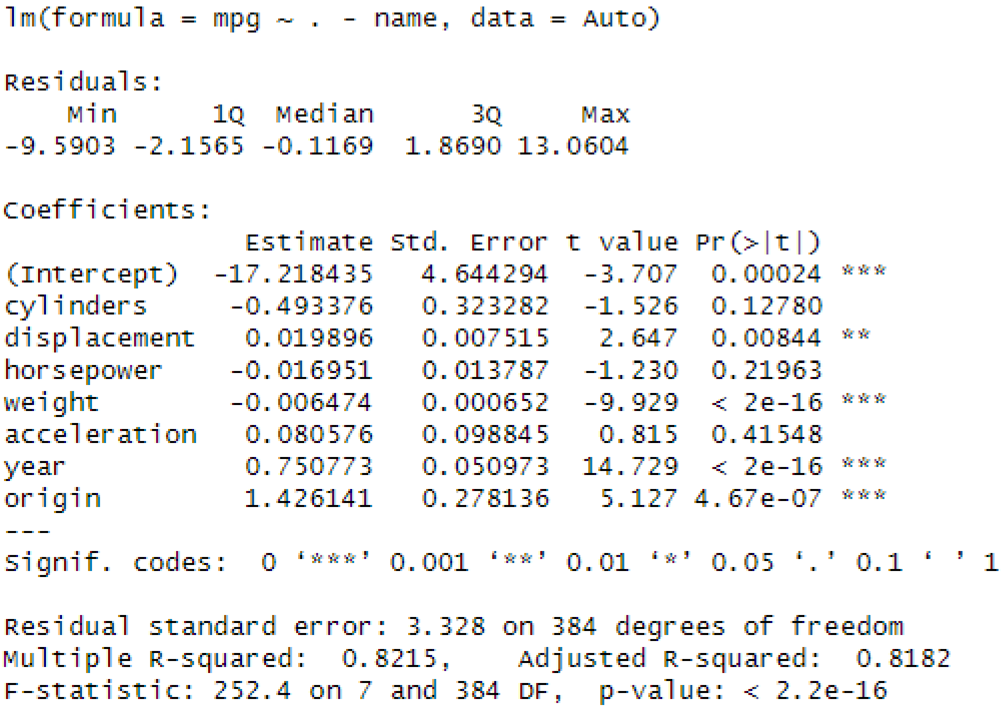


(c). Use the *lm()* function to perform a multiple linear regression with *mpg* as the response and all other variables except *name* as the predictors. Use the *summary()* function to print out results. Comment on the outputs, for instance:

Command:

```
model <- lm(mpg ~ . - name, data = Auto)

# Print the summary of the regression model

summary(model)
```

Results:



```
lm(formula = mpg ~ . - name, data = Auto)

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
```

- Is there a relationship between the predictors and the response?

There is a relationship between the predictors and the response as we can see that the displacement, weight, year and origin are all statically significant (at least at the 0.05 level). We also see that these predictors have a R2 of 82.15% and adjusted R2 of 81.82 indicating there is a great fit of the model on the data.

- Which predictors appear to have a statistically significant relationship to the response?

The displacement, weight, year and origin appear to be significant although they don’t seem to be practically significant as they don’t have much of an effect on the prediction of the model. The intercept appears to be very high, with low effects from the predicting variables indicating that although this regression is a good start for conceptualizing strategies for future models, this model is not very good.

- What does the coefficient for the *year* variable suggest?

The year model suggests that every unit increase of the “year” variable, there will be a “year” val * the coefficient of year βi to produce the overall effect on the miles per gallon variable. As a semi-dummy variable because year starts at 70 and goes up until 82 it shows the effect from 1970-1982. The coefficient βi being positive shows that you get more miles per gallon as the year goes up which is true to real data. Overall, the $\Delta year$ is equal to 1.43 per year.

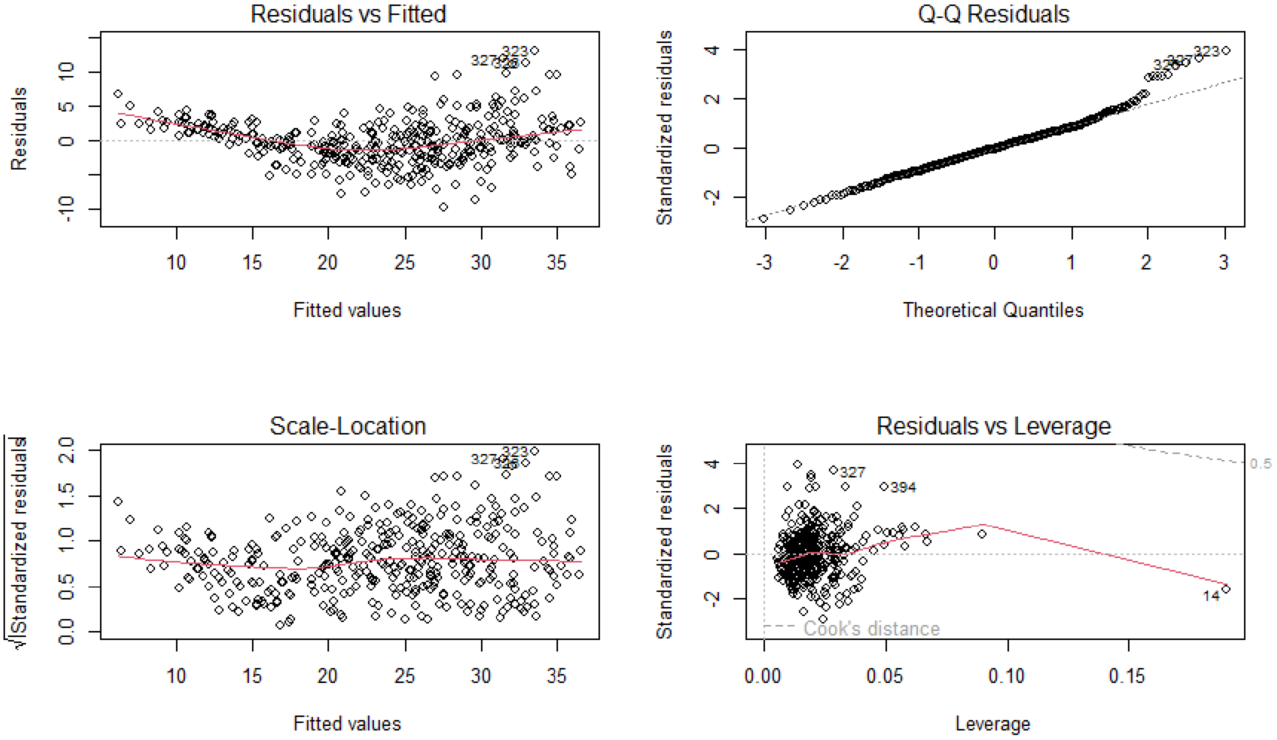
(d). Use the *plot()* function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

Command:

par(mfrow = c(2,2))

```
plot(model)
```

Results:



Analysis:

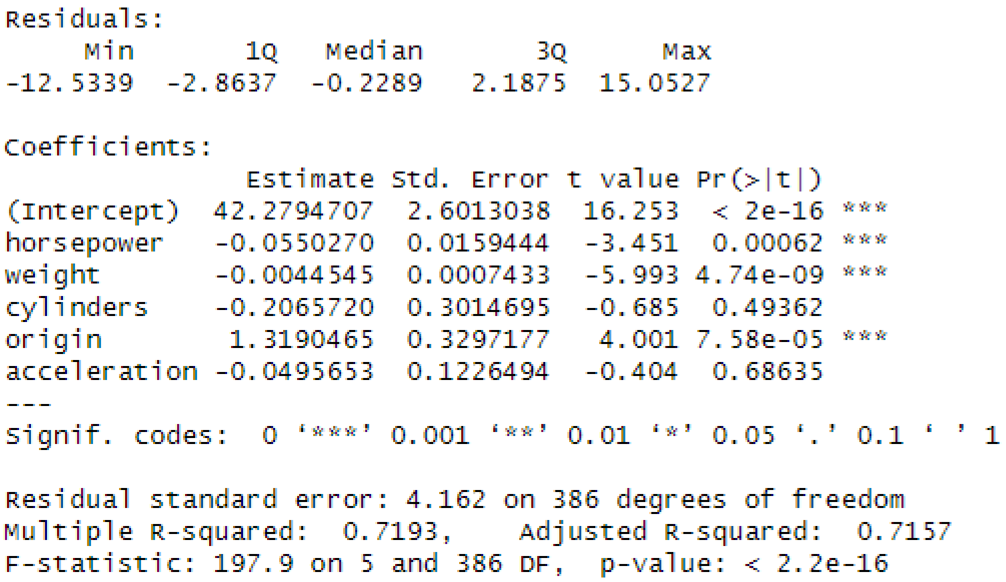
While looking at the 1st graph it shows a non-linear relationship between the response variables and the predictors. The 2nd graph shows that the residuals are normally distributed and rightward skewed. The 3rd graph shows that the constant variance of error assumption is not true for this model. Finally, the 4th graph: Shows there are no leverage points although there is a potential for point 14 to be a leverage point of the graph as it is far away from the rest of the data points that are clustered near the origin.

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

Command:

```
model_square_accel = lm(mpg ~ horsepower + weight + cylinders + origin +
acceleration*acceleration, data = auto)

summary(model_square_accel)
```

Result:



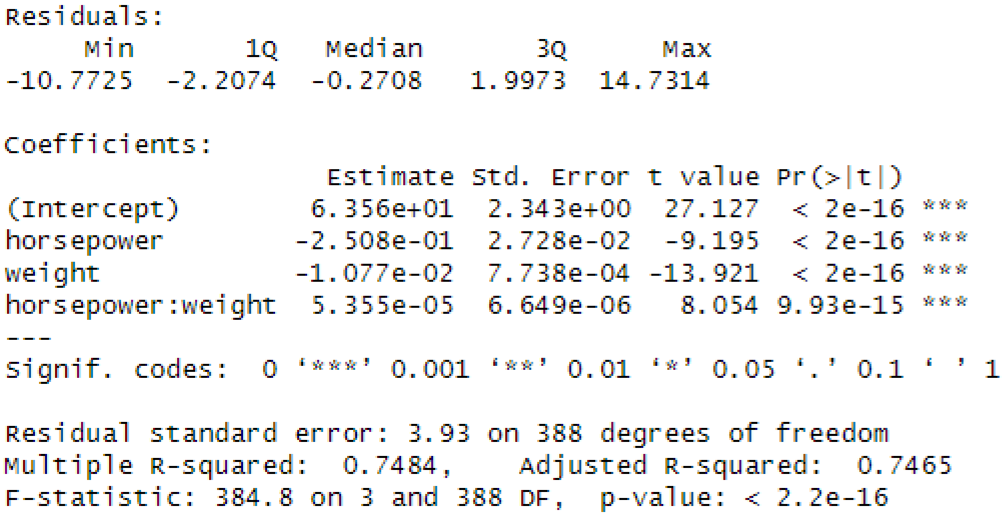
```
Residuals:
     Min       1Q   Median       3Q      Max 
-12.5339  -2.8637  -0.2289   2.1875  15.0527 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  42.2794707  2.6013038  16.253  < 2e-16 ***
horsepower   -0.0550270  0.0159444  -3.451  0.00062 ***
weight       -0.0044545  0.0007433  -5.993 4.74e-09 ***
cylinders    -0.2065720  0.3014695  -0.685  0.49362    
origin        1.3190465  0.3297177   4.001 7.58e-05 ***
acceleration -0.0495653  0.1226494  -0.404  0.68635    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.162 on 386 degrees of freedom
Multiple R-squared:  0.7193,	Adjusted R-squared:  0.7157 
F-statistic: 197.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

We see that most terms are significant at the 0.05 level expect for cylinders and acceleration. By effectively squaring the acceleration term, I thought that it would help with the curve of accelerating but it in fact is not significant. Therefore, I reject using this model and will investigate interaction terms as they are most likely to produce results when it comes to cars as most physics formulas consider interaction instead of multiplication.

Command:

model_horseweight <- lm(mpg ~ horsepower + weight + horsepower:weight, data = auto)

Result:



```
Residuals:
     Min       1Q   Median       3Q      Max 
-10.7725  -2.2074  -0.2708   1.9973  14.7314 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        6.356e+01  2.343e+00  27.127  < 2e-16 ***
horsepower        -2.508e-01  2.728e-02  -9.195  < 2e-16 ***
weight            -1.077e-02  7.738e-04 -13.921  < 2e-16 ***
horsepower:weight  5.355e-05  6.649e-06   8.054 9.93e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.93 on 388 degrees of freedom
Multiple R-squared:  0.7484,	Adjusted R-squared:  0.7465 
F-statistic: 384.8 on 3 and 388 DF,  p-value: < 2.2e-16
```

There is a high level of significance in the predictors in the model “model_horseweight” which captures the interaction term between horsepower and weight. While this result is good, as they are all significant and practically significant, the adjusted R2 is only 74.65% which indicates there can be improved fit of this model. Doing more research into cars will allow for my result to be improved.

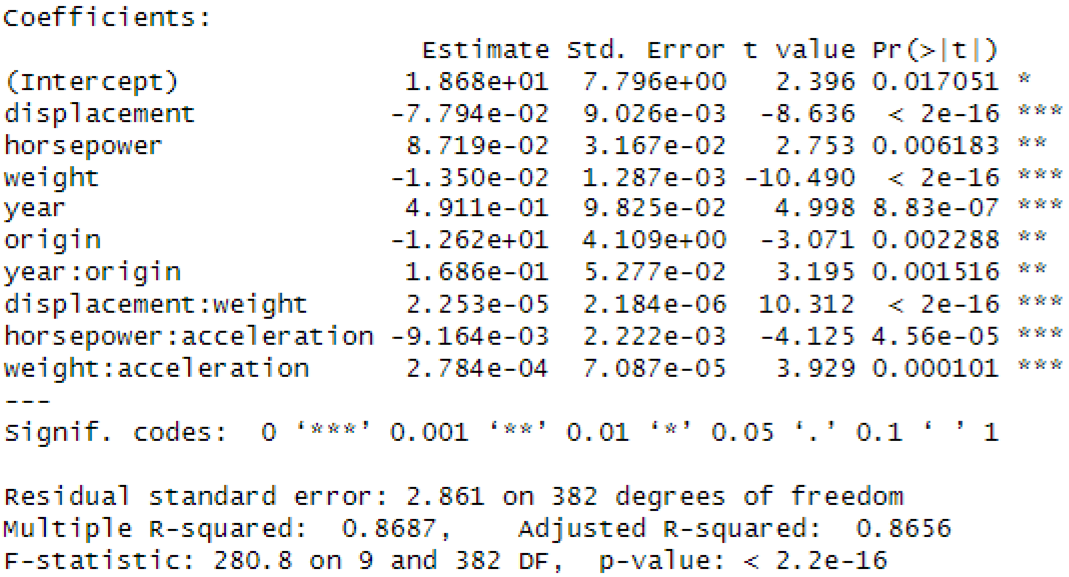
See below for the improved model.

Command:

model_triple = lm(mpg ~.-name-cylinders-acceleration+year:origin+displacement:weight+

displacement:weight+acceleration:horsepower+acceleration:weight, data=auto)

Result:



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.868e+01  7.796e+00   2.396 0.017051 *  
displacement            -7.794e-02  9.026e-03  -8.636  < 2e-16 ***
horsepower               8.719e-02  3.167e-02   2.753 0.006183 ** 
weight                  -1.350e-02  1.287e-03 -10.490  < 2e-16 ***
year                     4.911e-01  9.825e-02   4.998 8.83e-07 ***
origin                  -1.262e+01  4.109e+00  -3.071 0.002288 ** 
year:origin              1.686e-01  5.277e-02   3.195 0.001516 ** 
displacement:weight      2.253e-05  2.184e-06  10.312  < 2e-16 ***
horsepower:acceleration -9.164e-03  2.222e-03  -4.125 4.56e-05 ***
weight:acceleration      2.784e-04  7.087e-05   3.929 0.000101 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.861 on 382 degrees of freedom
Multiple R-squared:  0.8687,	Adjusted R-squared:  0.8656 
F-statistic: 280.8 on 9 and 382 DF,  p-value: < 2.2e-16
```

This seems to be the best regression model as all the coefficients seem to be significant as well as the coefficients are practically significant. Looking at the Adjusted R2 that investigates the R2 while punishing for more predictors, it shows 86.87% of the changes in the response variable can be explained by the predictors within the regression model “model_triple”. The regression has all terms (predictors) within the model achieve a 0.05 level of significance which allows for everything to be included. This is the model I select for the best interaction term and multiplier term throughout all the iterations that I have included in the tests.

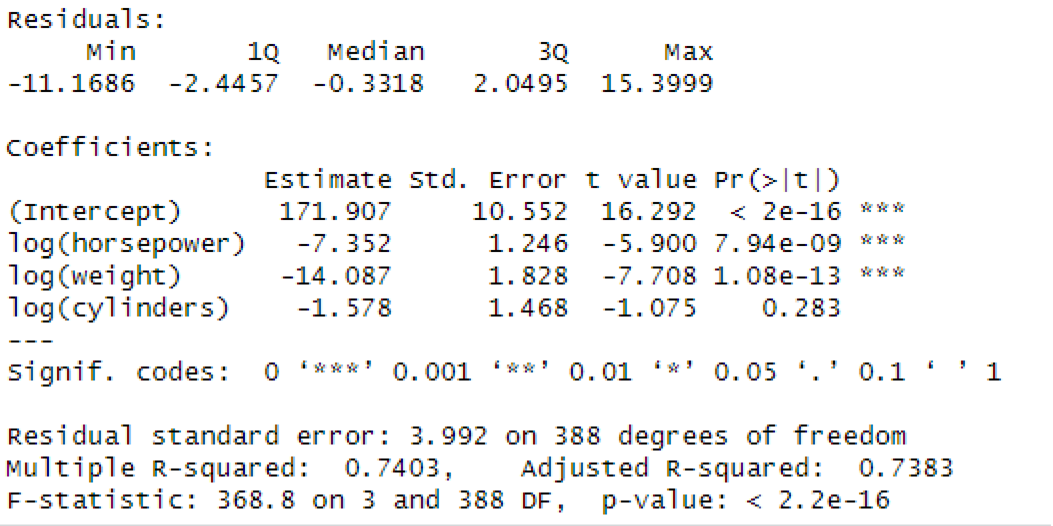
(f). Try a few different transformations of the variables, such as log(X), $\sqrt{X}$, X^2. Comment on your findings.

Command:

Log-log regression

model_log <- lm(mpg ~ log(horsepower) + log(weight) + log(cylinders), data = Auto)
summary(model_log)

Result:



```
Residuals:
     Min       1Q   Median       3Q      Max 
-11.1686  -2.4457  -0.3318   2.0495  15.3999 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       171.907     10.552  16.292  < 2e-16 ***
log(horsepower)    -7.352      1.246  -5.900 7.94e-09 ***
log(weight)       -14.087      1.828  -7.708 1.08e-13 ***
log(cylinders)     -1.578      1.468  -1.075    0.283    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.992 on 388 degrees of freedom
Multiple R-squared:  0.7403,	Adjusted R-squared:  0.7383 
F-statistic: 368.8 on 3 and 388 DF,  p-value: < 2.2e-16
```

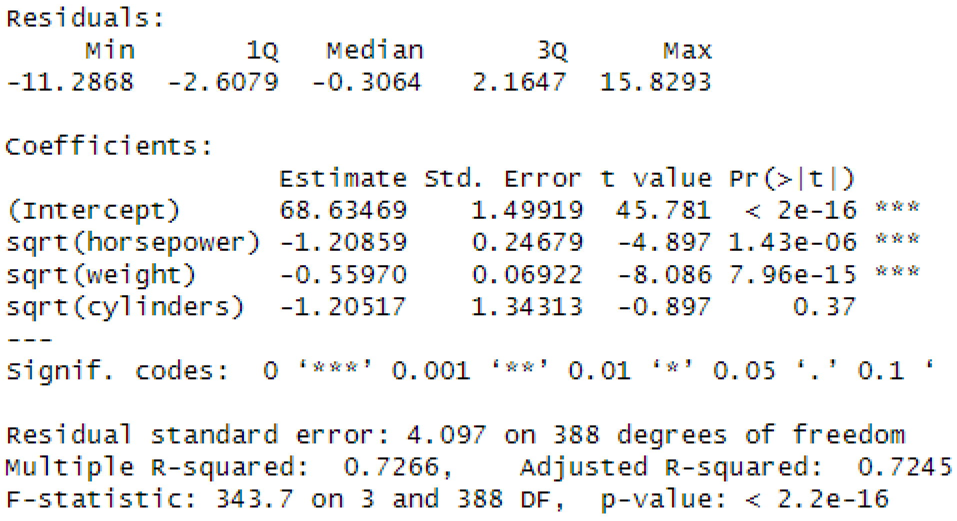
The log model is the best of the models you will see from question f but still is not as good as the model_triple above. This is because although weight and horsepower are significant, the cylinder varible is not. We can see the best R2 and adjusted R2 values out of the question f with it being 74% and 73% respectively.

Command:

Square-root regression

model_sqrt <- lm(mpg ~ sqrt(horsepower) + sqrt(weight) + sqrt(cylinders), data = Auto)
summary(model_sqrt)

Result:



```
Residuals:
     Min       1Q   Median       3Q      Max 
-11.2868  -2.6079  -0.3064   2.1647  15.8293 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      68.63469    1.49919  45.781  < 2e-16 ***
sqrt(horsepower) -1.20859    0.24679  -4.897 1.43e-06 ***
sqrt(weight)     -0.55970    0.06922  -8.086 7.96e-15 ***
sqrt(cylinders)  -1.20517    1.34313  -0.897     0.37    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.097 on 388 degrees of freedom
Multiple R-squared:  0.7266,	Adjusted R-squared:  0.7245 
F-statistic: 343.7 on 3 and 388 DF,  p-value: < 2.2e-16
```

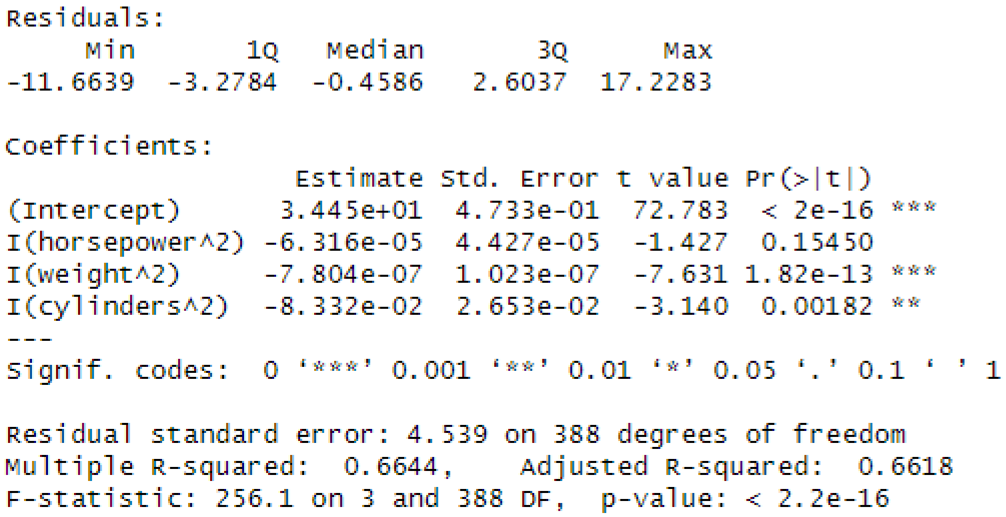
The square root model has two coefficients at the 0.05 signfiance level including weight and horsepower. The model has a good R2 value of 72% but since all of the terms are not significant, we reject the model and are going to still use the interaction term model above (model_triple).

Command:

Sqaured regression model

model_squared <- lm(mpg ~ I(horsepower^2) + I(weight^2) + I(cylinders^2), data = Auto)
summary(model_squared)

Result:



```
Residuals:
     Min       1Q   Median       3Q      Max 
-11.6639  -3.2784  -0.4586   2.6037  17.2283 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.445e+01  4.733e-01  72.783  < 2e-16 ***
I(horsepower^2) -6.316e-05  4.427e-05  -1.427  0.15450    
I(weight^2)     -7.804e-07  1.023e-07  -7.631 1.82e-13 ***
I(cylinders^2)  -8.332e-02  2.653e-02  -3.140  0.00182 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.539 on 388 degrees of freedom
Multiple R-squared:  0.6644,	Adjusted R-squared:  0.6618 
F-statistic: 256.1 on 3 and 388 DF,  p-value: < 2.2e-16
```

Overall, the model squared has evidence to suggest that weight and cylinder squared has significance at least on the 0.05 level. Moreover, the regression has a very low R2 which indicates a lower line of best fit accuracy on the real data. Therefore this model is not suggested to proceed with.

2. Let $X_1, X_2, \ldots, X_n$ be iid *Bernoulli*(p) with probability mass function

$$f_X(x|p) = p^x(1-p)^{1-x},$$

where $x = 0$ or 1 and $0 < p < 1$. Note that $E(X) = p$ and $Var(X) = p(1-p)$.

(a). Write a likelihood function for this random sample, then the log likelihood function.

Answer below

(b). Calculate the MLE of p.

Answer below

(c). What is the MLE for $Var(X)$ (you must justify your answer).

Answer below

(d). Find the (Fisher) expected information of the sample.

Answer below

(e). Show that $\bar{X}$ attains the Cramer-Rao Lower Bound and thus is the Uniform Minimum Variance Unbiased Estimator (UMVUE, or "best" unbiased estimator) of p.

Answer below

(f). Find the UMVUE for $Var(X)$.

Answer below

(g). Let $h(p) = \frac{p}{1-p}$ be the odds of observing $X = 1$. The MLE of $h(p)$ is $h(p\hat{})$. Using

the result from part (c), find the asymptotic variance of the MLE, $h(p\hat{})$ (i.e. find $Var_p(h(\hat{p})))$.

This is now a bonus question.

Answer below

SEE ANSWERS WRITTEN BELOW