ST – 512 Final Project Report

Name – Saurabhkumar Makwana

Email – makwanas@oregonstate.edu

Introduction:

The dataset I have used for this final project is auto-mpg dataset, which stands for mileage per gallon performance of various cars. The dataset is downloaded from UCI Machine Learning Repository. The dataset contains 7 columns (all of them being numeric) and 398 instances/rows. It contains miles per gallons(mpg) as the response variable whereas the other 6 explanatory variables named as weight of the vehicle, number of cylinders, horsepower, displacement, acceleration and the model year of the car.

After analysis of this dataset, I tried to answer these 2 research questions:
1. Finding out the statistically significant factors that are responsible for the mean miles per gallon variable after accounting for the weight of the vehicle?

2. Does the effect of weight on mean miles per gallon depend on the horsepower of the vehicle?

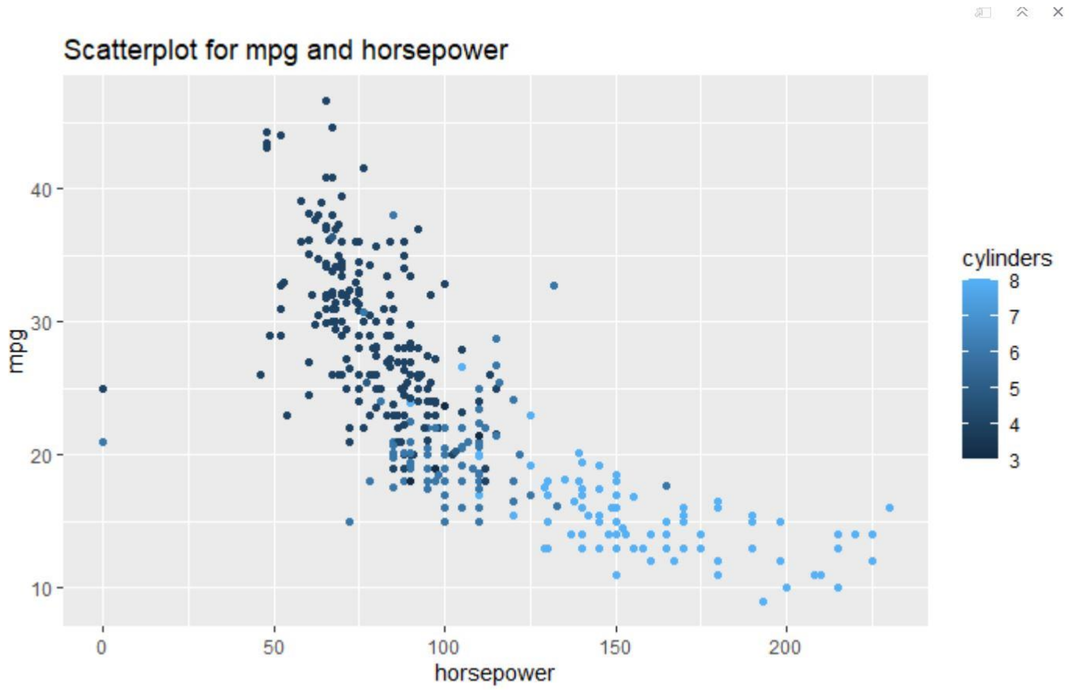
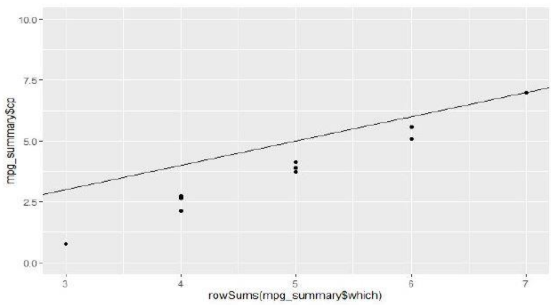
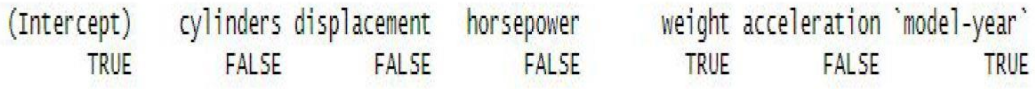


Figure 1. shows the scatter plot trying to answer the second research question

Methods:

For addressing the first research question, I have considered several models. I have used all the 4 methods (forward selection, backward selection, best subsets and sequential replacement) for choosing the best model to include and excluding the significant and insignificant explanatory variables respectively. The idea behind using these methods was to use automatic variable selection methods named earlier which would allow us to gauge which of the variables would be suitable to preserve in the model after keeping weight as one of the variables. After performing p vs cp analysis, we find that the model with p=3 would be the best fit





```
(Intercept)    cylinders displacement   horsepower       weight acceleration `model-year`
       TRUE        FALSE        FALSE        FALSE         TRUE        FALSE         TRUE
```

Figure 2. shows the p vs cp analysis and the significance of the variables

For addressing the second research question, the following null and alternate hypothesis were taken into consideration.
H0 -> There is no effect of weight on mean mpg for different values of horsepower.

Ha -> There is an effect of weight on mean mpg for different values of horsepower.

Results:

As we can see from figure 2, the Akaike information criterion (AIC) is the lowest. The values of BIC and Cp are also minimum, which states that we have a likelihood of choosing the variables with the “TRUE” tag into our final model. After the analysis of the adjusted R2 square obtained from the summary, we find that the following values gives us the maximum value. This states that choosing the model-year model along with weight as explanatory variables would give the model statistical significance w.r.t the response variable. The following values correspond to the

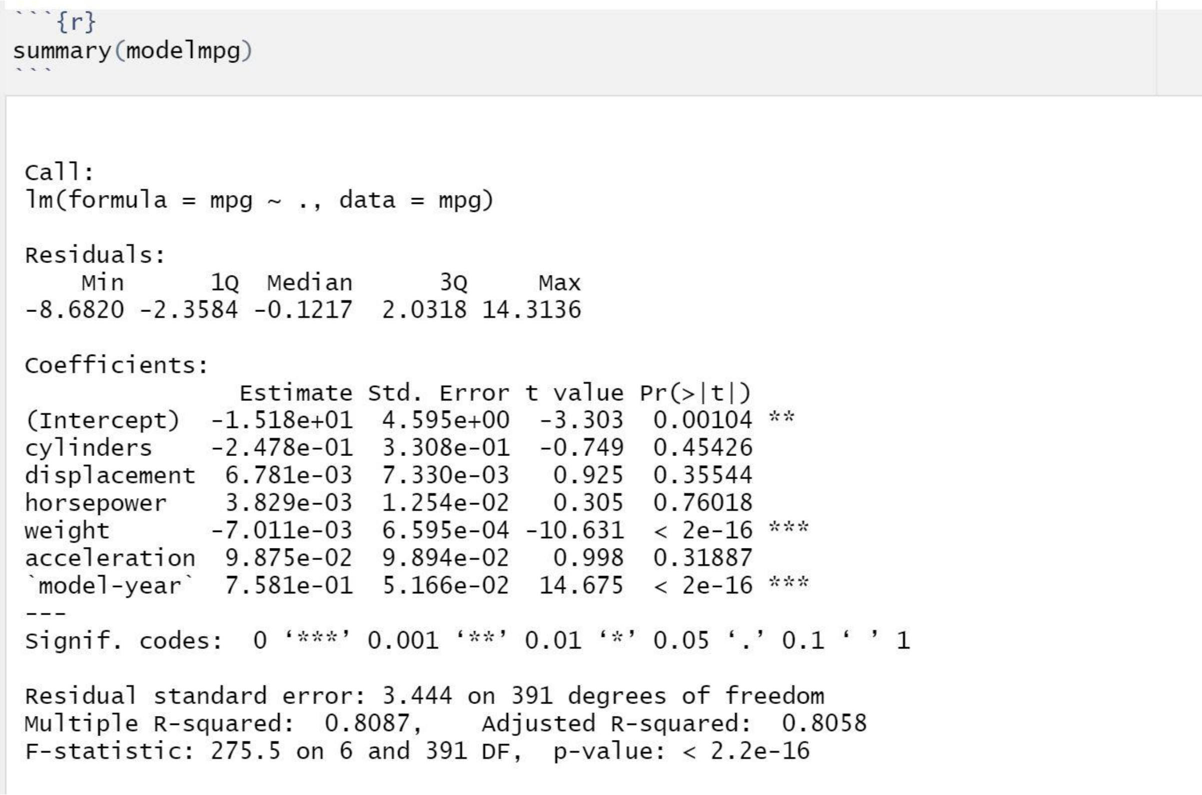
above scenario:
Minimum BIC = -638.55

Minimum AIC = 549.81

Minimum Cp = 0.78

Maximum AdjR2 = 0.80

On further analysis, we see that only the p-values for those 2 explanatory values are low (less than 0.05) and hence are statistically significant. The other variables such as acceleration, displacement, cylinder and so on do not seem to be significant.



~~~
```{r}
summary(modelmpg)
```
~~~

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

Residuals:
    Min      1Q  Median      3Q     Max 
-8.6820 -2.3584 -0.1217  2.0318 14.3136 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.518e+01  4.595e+00  -3.303  0.00104 ** 
cylinders    -2.478e-01  3.308e-01  -0.749  0.45426    
displacement  6.781e-03  7.330e-03   0.925  0.35544    
horsepower    3.829e-03  1.254e-02   0.305  0.76018    
weight       -7.011e-03  6.595e-04 -10.631  < 2e-16 ***
acceleration  9.875e-02  9.894e-02   0.998  0.31887    
`model-year`  7.581e-01  5.166e-02  14.675  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.444 on 391 degrees of freedom
Multiple R-squared:  0.8087,	Adjusted R-squared:  0.8058 
F-statistic: 275.5 on 6 and 391 DF,  p-value: < 2.2e-16
```

Figure 3. shows the summary of the full model

The normal Q-Q plot indicates that the residual values are not normally distributed. Additionally, the high variance inflation factors (VIF > 10) suggesting multi-collinearity amongst the variables. Also, the same could be suggested for the residual vs fitted values which indicates some relationship amongst the variables.

The final model equation looks like:

$$\mu(mpg|weight, model\text{-}year) = \beta0 + \beta1*Weight+ \beta2*model\text{-}year$$

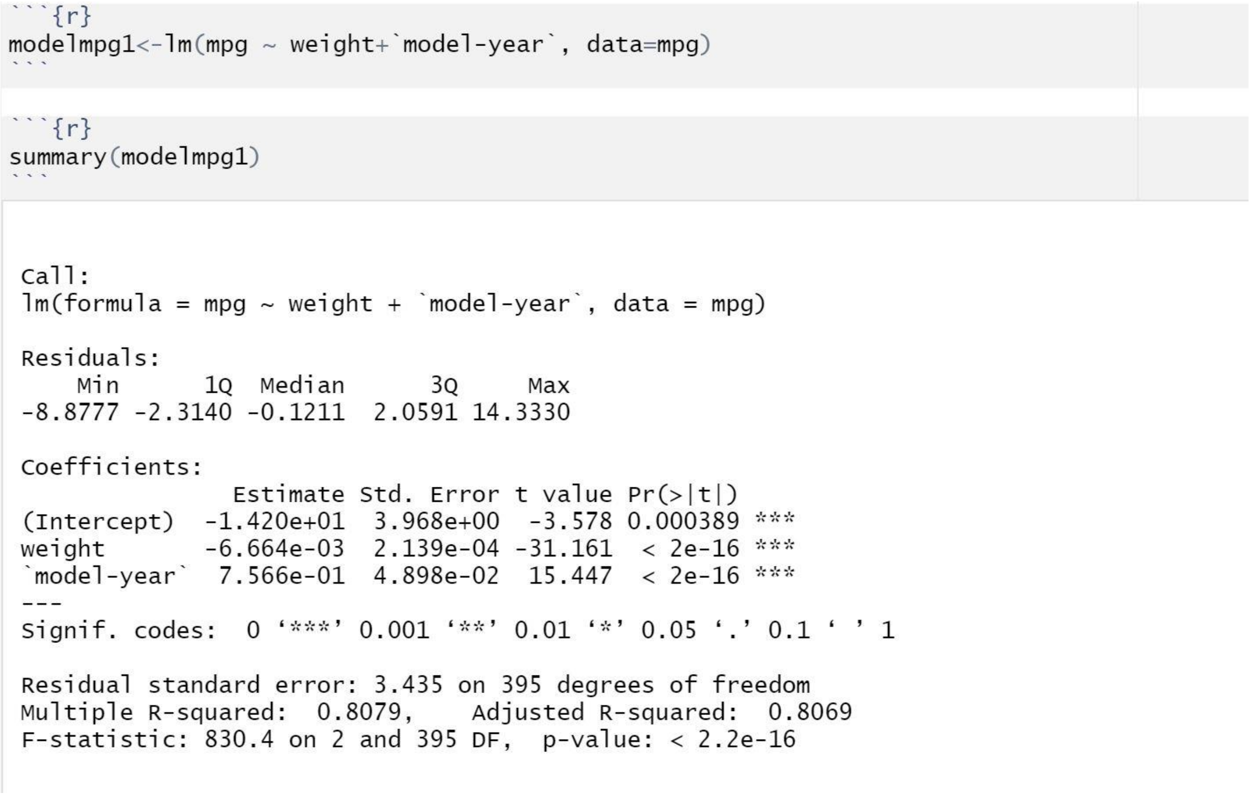
The full-model equation for the null hypothesis looks like:

$$\mu(mpg|weight, horsepower) = \beta0 + \beta1*Weight+ \beta2*horsepower + \beta3*weight*horsepower$$

Whereas the reduced model equation for the alternate hypothesis looks like:

$$\mu(mpg|weight, horsepower) = \beta0 + \beta1*Weight+ \beta2*horsepower$$

where β0 denotes the intercept parameter, β1 denotes slope parameter for weight considering other terms are also there. β2 denotes slope parameter for horsepower. β3 denotes coefficient of interaction parameter between weight and horsepower.



~~~
```{r}
modelmpg1<-lm(mpg ~ weight+`model-year`, data=mpg)
```
~~~

~~~
```{r}
summary(modelmpg1)
```
~~~

```
Call:
lm(formula = mpg ~ weight + `model-year`, data = mpg)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8777 -2.3140 -0.1211  2.0591 14.3330 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.420e+01  3.968e+00  -3.578 0.000389 ***
weight       -6.664e-03  2.139e-04 -31.161  < 2e-16 ***
`model-year`  7.566e-01  4.898e-02  15.447  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.435 on 395 degrees of freedom
Multiple R-squared:  0.8079,	Adjusted R-squared:  0.8069 
F-statistic: 830.4 on 2 and 395 DF,  p-value: < 2.2e-16
```

Figure 4. shows the summary of the reduced model

As we can see from the above figure that the p-value of the interaction term seems to be very low(less than 0..01) and hence significant. Therefore, we can reject the null hypothesis and this suggests that the β3 parameter is not zero.

Discussion and Conclusion:

From our results, we can conclude that there is an effect of weight on mean mpg for different values of horsepower. Hence, it seems more likely that the effect of weight on mean mpg is dependent on horsepower. It is evident from the p-value, which is less than 0.01 and is shown in the figure and obtained to assess the interaction between weight and horsepower. We also see that there might be some sort of level relationships observed amongst some of the independent variables. It can be seen that the miles per gallon(mpg) of an auto-car is largely dependent on the weight of the vehicle and even the time at which the vehicle was manufactured and hence the model-year. We can also see that the older vehicles seem to have a lower mean mpg whereas the newer ones seem to have a higher value for mean mpg.

The collection of data seems to find a conclusive evidence for establishing that heavier vehicles who have been manufactured recently have a positive effect on the mean miles per gallon(mpg), and hence a higher value. It also seems that this was an observational study since all of the data for the explanatory variables was collected based on randomness and observations. It does not seem to be causal study because any causal effects were not considered from the way the experiment was conducted and hence no treatment effect was observed.