Multivariate Analysis on fuel efficiency of car

Issues:

Our data-set includes numerical data on several features of cars, including their weight, engine performance, displacement, acceleration and miles per gallon(mpg).

The mpg metric can be used to examine how the various car components are related to each other, as well as to develop a predictive model for a car's fuel efficiency based on various factors.

We Address the following questions:

1. Is at least one of the predictors useful in predicting the response?
2. Do all the predictors help to explain the response, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Findings:

According to the analysis results, displacement, horsepower, and weight have a negative correlation with mpg, while acceleration has a positive correlation with mpg.

Although the overall p-value indicates that the model fits the data well at the multivariate level, when considering individual factors, only horsepower and weight are found to be more significant, while displacement and acceleration are not as significant based on their respective p-values.

Discussions:

Given a dataset of 391 car samples, it may not be possible to draw a definitive conclusion as there could be several other variables that could impact the car's fuel efficiency (mpg). However, we can still analyze the given dataset and establish relationships between the variables.

Through analyzing the p-value, we can observe that the model is a good fit for the dataset. We can also generate a correlation matrix to determine the significance of each independent variable on the dependent variable, which is mpg. Each coefficient value not only affects the dependent variable, but it also impacts the coefficient of other variables as well.

Appendix A: Method

To conduct the analysis of dependence of fuel efficiency on various other factors, we used multiple linear regression, as there were multiple factors on which dependent variable was depending. We imported the .csv file into R studio and carried out the analysis. mpg refers miles per gallon which defines efficiency, horsepower defines the power of engine, acceleration is rate at which car can change its speed, weight is total weight of the car and displacement defines measure of cylinder volume swept.

We applied various statistical tools to analyze the relationship between multiple factors which include, correlation scatter plots, linear model plot and used Pearson's method to analyze the correlation between variables, to find out significance of each variable on the dependent variable.

Appendix B: Results

Initially we import the .csv file having 391 samples, into R studio. Since we have multiple variables, we use multiple linear regression model.

In order to fit multiple linear regression model using least squares, we use linear model function(lm()). The result is shown in below Figure 1.

```
> lm.fit <- lm(mpg~displacement+horsepower+weight+acceleration , data=mydata1)
> summary(lm.fit)

Call:
lm(formula = mpg ~ displacement + horsepower + weight + acceleration,
    data = mydata1)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9928 -3.0044 -0.4073  2.0833 15.9453

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.0191300  2.4269599   20.20  < 2e-16 ***
displacement -0.0102288  0.0066426   -1.54  0.12441
horsepower   -0.0519771  0.0163466   -3.18  0.00159 **
weight       -0.0049139  0.0008096   -6.07 3.07e-09 ***
acceleration -0.2088236  0.1273537   -1.64  0.10188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.368 on 386 degrees of freedom
Multiple R-squared:  0.728,     Adjusted R-squared:  0.7252
F-statistic: 258.3 on 4 and 386 DF,  p-value: < 2.2e-16
```

Figure 1. Result of linear model function

From the above figure 1, we have p-value of “displacement” is 0.12441 which is greater than significant value 0.05, p-value of “horsepower” is 0.00159 which is less than significant value 0.05, p-value of “weight” is very much less than 0.001, and p-value of “acceleration” is 0.10188 which is greater than 0.05.

From the values we can infer “weight” and “horsepower” are significant factors because as they are having p-value less than significant value 0.05, which is strong evidence against null hypothesis, by which we can conclude these two are useful in explaining mpg. But for displacement and acceleration the p-value is above significant level 0.05, so you can’t conclude anything about these two factors.

From figure 1 we can say the mpg is related to factors in below way:

$$mpg = 49.019 - (0.01023 * displacement) - (0.0519 * horsepower) - (0.00491 * weight) - (0.2088 * acceleration)$$

Then we have correlation plots in Figure 2 and corresponding matrix in Figure 3.

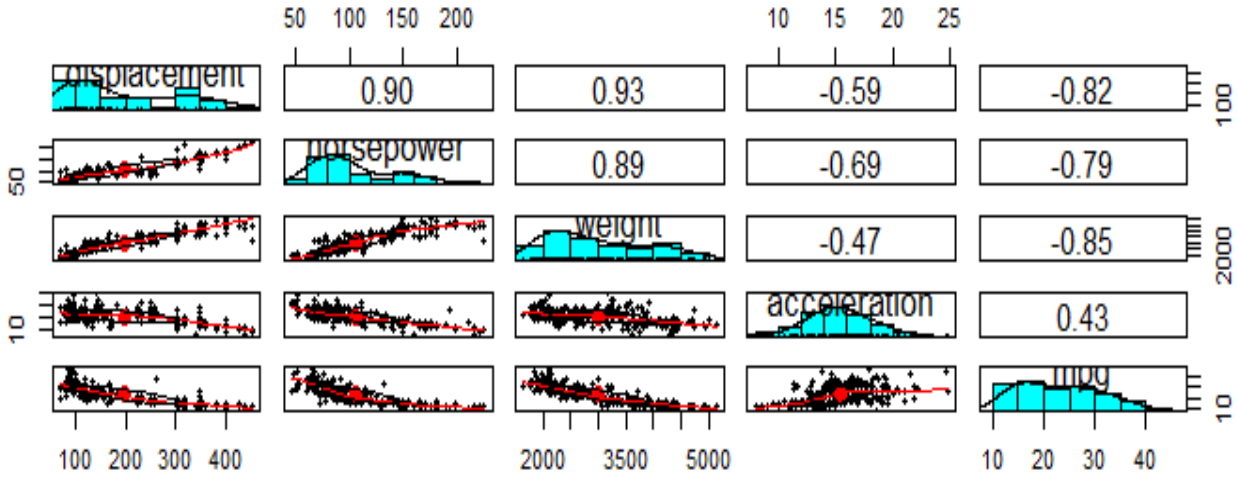

Figure 2. Correlation plots for corresponding variables

```
> cor(mydata1 , method = "pearson")
             displacement horsepower     weight acceleration        mpg
displacement    1.0000000  0.9028999  0.9336773   -0.5923890 -0.8154218
horsepower      0.9028999  1.0000000  0.8853996   -0.6887626 -0.7942153
weight          0.9336773  0.8853996  1.0000000   -0.4691266 -0.8460638
acceleration   -0.5923890 -0.6887626 -0.4691266    1.0000000  0.4318968
mpg            -0.8154218 -0.7942153 -0.8460638    0.4318968  1.0000000
>
```

Figure 3. Correlation values using Pearson's method

From Figure 2 and Figure 3, we conclude that displacement, horsepower and weight are negatively related to mpg and acceleration is positively related to mpg, but considering the value acceleration doesn't fit to model well.

The R-squared value of the model is 0.728, indicating that approximately 72.8% of the variation in mpg can be attributed to the independent variables in the model. The adjusted R-squared value, which considers the number of independent variables and the sample size, is slightly lower at 0.7252. This value provides a more conservative estimate of the model's fit.

From the model, a very small p-value (2.2e-16)(less than significant value 0,05), which is an evidence against null hypothesis ,indicating that at least one of the independent variables in the model is statistically significant in predicting mpg.

Appendix C: Code

In this appendix we will document the code written in R studio to plot correlations and code use for linear model and Pearson's coefficient.

CODE:

```
install.packages("readxl")
install.packages("psych")
library(psych)
library(readxl)
# Reading excels
file <- "C:/Users/DELL/Downloads/auto_data.xls"
mydata1 <- read_excel(file)
View(mydata1)
#Correlations matrix and plot
cor(mydata1, method = "pearson")
pairs.panels(mydata1, cex.cor = 0.5)
#Linear model
lm.fit<-lm(mpg~displacement+horsepower+weight+acceleration, data=mydata1)
summary(lm.fit)
```