STA302H1F: Final Project

Member name	*Contributed part(s)*
Aleksa Zatezalo	Introduction/Data analysis/Conclusion
Arhum Ahmad	Model Development/Discussion
Mital Topiwala	Model Development/Discussion
Wei Qiao	Data analysis/Discussion

Introduction

In recent years, global climate change has been recognized as an issue of increasing importance. Not only is it driving detrimental changes in rainfall, extreme weather, seasonal shifts, and the biosphere as a whole, it can have detrimental health effects on human life. As recently as 2019, newly occurring heat waves in Paris have been responsible for the deaths of 1,500 people and been responsible for the hospitalization of many more. What is more shocking is that the scientific community has concluded that shifts in the climate are largely due to large scale human behaviour, most specifically carbon commissions.

However, these emissions can be controlled. If the carbon footprint of each human is reduced enough, the consensus in the scientific community states that the Climate Change can not only be stopped but reversed. As it stands, transportation is responsible for 28% of man made greenhouse emissions making it the number one producer of greenhouse gas in the United States. Thus, learning how to reduce greenhouse gas emission of personal transportation is critical in helping contain emissions.

Prior to the writing of this paper many States in the US have passed anti emissions laws, most notably in California. All vehicles must be smog tested , and large fees must be paid for those who emit lots of CO2. Moreover vehicles that emit too much are not approved. Since the introduction of these laws, they have seen a significant improvement in air quality.

However, fuel mileage has not yet been worked into legislation. Since 90% of the fuel used in personal transport is petroleum based, reducing emissions is intrinsically tied to obtaining good fuel mileage. Certain characteristics like car weight or number of cylinders are closely tied to the miles achieved per gallon. Thus they can be critical in understanding both in the creation of new legislation and the engineering of new fuel efficient cars. It is for this reason, that we will look at the *mtcars* data set in R, in order to identify how each characteristic of a car works to predict the miles per galon spent by a vehicle.

Data Analysis

As can be seen in the basic data summary section of the R appendix. Ten predictor variables described below.

- Cylinders: The Number of Cylinders in engine.
- Displacement: The size of the engine in cubic inches.
- Gross Horsepower: The Horsepower.
- Rear Axis Ratio: The ratio between the gears on the front axil and rear axil.
- Weight: The weight of the car in pounds.
- ¼ Mile Time: The time it can travel a quarter mile in seconds.
- Engine Shape: How the cylinders are arranged in the engine.
- Transmission: Automatic or manual transmission.
- Number of Forward Gears: The number of gears used to push the car forward.
- Number of Carburetors: The number of carburetors used to inject oxygen into the fuel.

The effect of these variable on miles per gallon(mpg) can be seen in the table below. This calculation can be seen in the R Appendix (page 1).

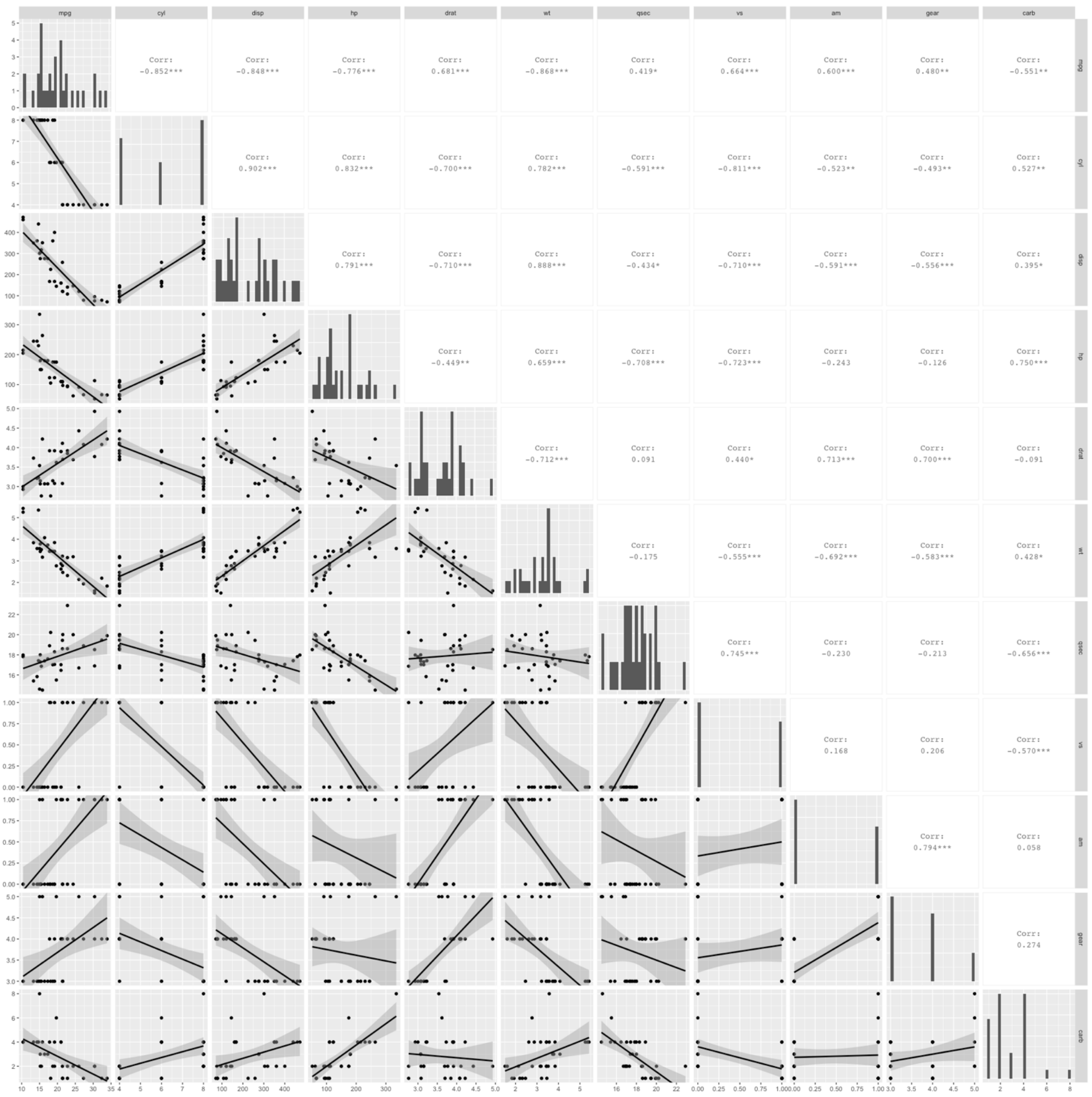
	Estimate	*Std. Error*	*T Value*	*Pr(>|t|)*
Intercept	12.303	18.718	0.657	0.5181
Cylinders	-0.111	1.045	-0.107	0.9161
Displacement	0.0133	0.0178	0.747	0.4635
Gross Horsepower	-0.02148	0.0217	-0.987	0.3350
Rear Axis Ratio	0.78711	1.6353	0.481	0.6353

Weight	-3.515	1.89441	-1.961	0.0633
¼ Mile Time	0.82104	0.73084	1.123	0.2739
Engine Shape	0.31776	2.10451	0.151	0.8814
Transmission	2.52023	2.05665	1.225	0.2340
Number of Forward Gears	0.65541	1.49326	0.439	0.6652
Number of Carburetors	-0.19942	0.82875	-0.241	0.8122

At the first glance, almost all predictor variables seem not useful, they all come with relatively high p-value, which means they are likely not able to reject the null hypothesis of which the predictor variable has not impact on the response variable.

This could be caused by dependency issue between the predictor variables. For example, engine shape is largely influenced by the number of cylinders present in the motor. An engine where the cylinders are arranged with an offset will have what is called a “v shape” and is labeled as 0 in the data, and is general created when we have an odd number of cylinder pairs (For example 3 or five pairs of cylinders). An engine with an even number of cylinder pairs takes on a “straight” shape and is ascribed the value 1. Because engine performance is based on number of cylinders & not the shape of the engine, it will be discarded as a potential predictor variable. Moreover, cars with manual transmission will be disregarded as the car’s fuel economy is largely subject to the drivers practice regarding shifting gears. Although manual cars are often considered to have better mileage, this presumes optimal gear shifting practices.

To exam the data further, we drew scatter plots for each predictor variables vs mpg and calculate the correlation between each predictor variables.



It clearly shows a linear trend between mpg and most predictor variables. For example, a larger number of cylinders, displacement, horsepower, and weight definitely decrease mpg, and a larger rear axis ratio. Some of the variance of predictor variables are visually too large, such number of gears and transmission type. This confirms our thought on the multicollinearity issue addressed before. To exclude some of these predictors later, we would take a look at the correlation matrix.

	mpg	*cyl*	*disp*	*hp*	*drat*	*wt*	*Qsec*	*Vs*	*am*	*Gear*	*Carb*
mpg	1.000	-0.852	-0.848	-0.776	0.681	-0.868	0.419	0.664	0.600	0.480	-0.551
cyl	-0.852	1.000	0.902	0.832	-0.700	0.782	-0.591	-0.811	-0.523	-0.493	0.527
disp	-0.848	0.902	1.000	0.791	-0.710	0.888	-0.434	-0.710	-0.591	-0.556	0.395
hp	-0.776	0.832	0.791	1.000	-0.449	0.659	-0.708	-0.723	-0.243	-0.126	0.750
drat	0.681	-0.700	-0.710	-0.449	1.000	-0.712	0.091	0.440	0.713	0.700	-0.091
wt	-0.868	0.782	0.888	0.659	-0.712	1.000	-0.175	-0.555	-0.692	-0.583	0.428
qsec	0.419	-0.591	-0.434	-0.708	0.091	-0.175	1.000	0.745	-0.230	-0.213	-0.656
vs	0.664	-0.811	-0.710	-0.723	0.440	-0.555	0.745	1.000	0.168	0.206	-0.570
am	0.600	-0.523	-0.591	-0.243	0.713	-0.692	-0.230	0.168	1.000	0.794	0.058
gear	0.480	-0.493	-0.556	-0.126	0.700	-0.583	-0.213	0.206	0.794	1.000	0.274
carb	-0.551	0.527	0.395	0.750	-0.091	0.428	-0.656	-0.570	0.058	0.274	1.000

Generally, a correlation larger than 0.9 is considered too high. Some of the values in the matrix are around 0.9, but we can't conclude they are the issue for sure. Thus, we calculated VIFs for predictor variables.

cyl	*disp*	*hp*	*drat*	*wt*	*Qsec*	*Vs*	*am*	*Gear*	*Carb*
15.373833	21.620241	9.832037	3.374620	15.164887	7.527958	4.965873	4.648487	5.357452	7.908747

A correlation with higher value than 0.9 would generate a VIF larger than 10. Cyl, disp and wt might be the main reason of multicollinearity issue with VIFs of 15, 21 and 15 respectively. However, this doesn't mean we have to exclude them prior to model building, it is possible that two of them are highly dependent on the third one. Also, we would also pay attention on those predictors with a VIF between 5 and 10, allowing too many of such predictors in our model wouldn't affect the fitting, but will generate poor prediction results. Although variables gear, carb, and am have a low VIF rate, they do not seem to display characteristics of a normal distribution when graphed on a scatter plot. This is especially true for cyl, and carb.

Model Development & Discussion

Through discussion above we will develop a model using variables a combination of variables discussed above. Because of the discussion involving the relevance of certain predictors engine shape(*vs*) and transmission(*am*) will be not be included in the development of the model. Because of the fact that they do not appear to be normally distributed in relation to miles per gallon(*mpg*), the number of carbonators (*carb*) will also not be included in the model development.

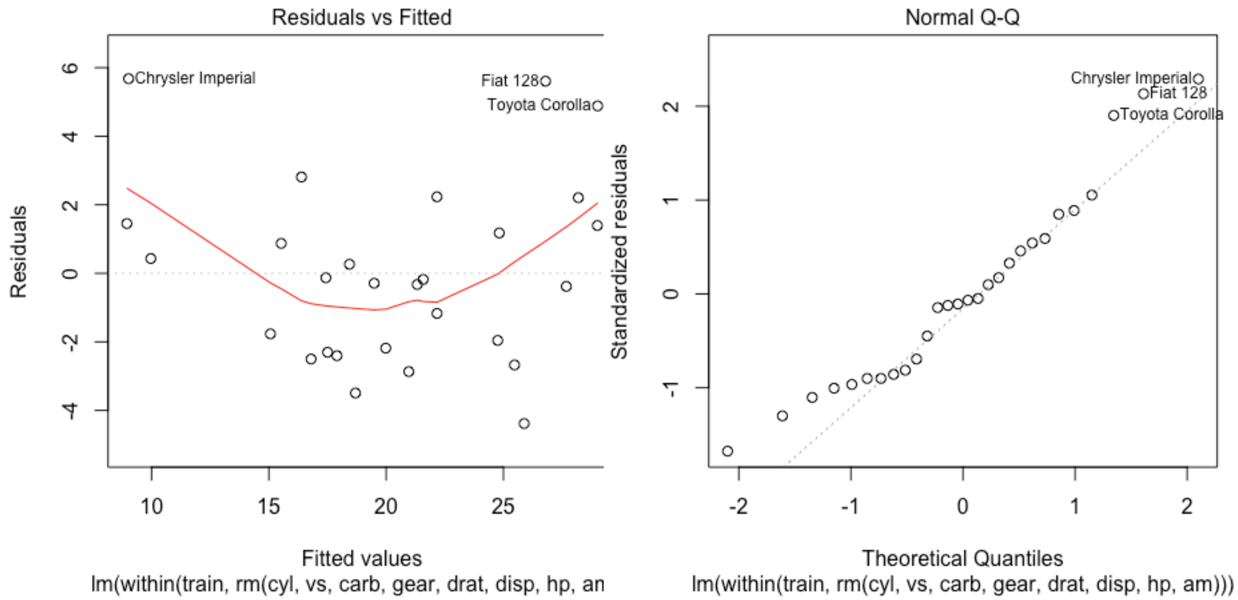
This leaves the following variables to be used in model development:

- Displacement (disp)
- Gross Horsepower (hp)
- Rear Axis Ratio (drat)
- Weight (wt)
- Quarter Mile Time (qsec)
- Number of Forward Gears (gear)

However when this model is created it has a very high Mean Squared Prediction Error(of 16.1). It is significantly higher than the Mean Squared Residual of approximately 0.3. As seen in the R appendix below, the Mean Squared Prediction Error of 16 is lowest when we use all 6 specified variables indicating that we might need to re-introduce the number of carbonators, engine shape, and transmission as possible predictors.

As the two remaining variables were added to the model, the difference between the MSPE and the MSRes drastically decreased. With all eight variables in question the difference between the two indicators was only 1.736. This difference is reasonably small. Because the difference had been minimised, we can consider the model in question as being a good fit. Moreover the regression has an R^2 value of 0.8, which is considered an exceptional fit.

To ensure the assumptions of multiple linear regression are not violated, we checked plot of residual vs fitted value and residual qq-plot.



It shows that residuals are normally distributed, independent from variables, have a fairly consistent finite variance and an expected value of 0.

Conclusion

Through this analysis we can induce that the most useful model for predicting miles per gallon expenditure involves all eight predictor variables present in the mtcars data set. Moreover, the relationship seems to be linear in nature.

The model is useful as it allows us to observe all aspects of a cars gas expenditure and not only predict its expenditure and relative carbon emissions but also the impact of each predictor on expenditures. In future cases, consumers may use this model to make a more informed purchase, when it comes to motorized vehicles. It may also be used to help engineers build more efficient vehicles as the impact of each component is better understood.

However, this model comes with limitations. Consumers may not always have access to all predictors necessary to make an informed decision, and fully utilizes this model. Moreover, engineers would need more information on how all variables co-vary in order to better engineer efficient vehicles. All in all this model provides valuable insight into how fuel economy is impacted.

Appendix - R code

```r
#To plot easily
library(GGally)
library(ggplot2)

#To use VIF function
library(car)

#Plot scatter for each independent variable vs dependent variable, show trend line with confidence interval, and histogram for each variable.
ggpairs(mtcars, lower = list(continuous = "smooth"), diag=list(continuous="bar"))

#A printed version of correlations between each variable
correlation_matrix <- round(cor(mtcars),3)
correlation_matrix

model <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb, data = mtcars)

#Measure multicollinearity, a VIF > 10 means severe multicollinearity
vif(model)

#Build model
train <- mtcars[1:28,]
test <- mtcars[29:32,]

##### First 6 Combinations

#iteration 1
summary(lm(within(train, rm(disp))))

#iteration 2
summary(lm(within(train, rm(disp, hp))))

#iteration 3
summary(lm(within(train, rm(disp, hp, drat))))
```

```
#iteration 4
summary(lm(within(train, rm(disp, hp, drat, wt))))

#iteration 5
summary(lm(within(mtcars, rm(disp, hp, drat, wt, qsec))))

#iteration 6
summary(lm(within(mtcars, rm(disp, hp, drat, wt, qsec, gear))))

model <- lm(within(mtcars, rm(disp, hp, drat, wt, qsec, gear)))
MSPE <- sum(((test$mpg - predict(model, test))^2))/4

#the MSPE is significantly higher than the MSRes so we check for collinearity

df <- within(train, rm(lm(mtcars, rm(disp, hp, drat, wt, qsec, gear))))
cor(df)
#high correlation between am and wt
#wt has higher t value so remove am

model <- lm(within(train, rm(disp, hp, drat, wt, qsec, gear)))
MSPE <- sum(((test$mpg - predict(model, test))^2))/4 #2.0389
MSRes <- (2.752^2)/25

MSPE - MSRes


#### All Eight

#iteration 1
summary(lm(mtcars))

#iteration 2
summary(lm(within(train, rm(cyl))))

#iteration 3
```

```
summary(lm(within(train, rm(cyl, vs))))

#iteration 4
summary(lm(within(train, rm(cyl, vs, carb))))

#iteration 5
summary(lm(within(train, rm(cyl, vs, carb, gear))))

#iteration 6
summary(lm(within(mtcars, rm(cyl, vs, carb, gear, drat))))

#iteration 7
summary(lm(within(mtcars, rm(cyl, vs, carb, gear, drat, disp))))

#iteration 8
summary(lm(within(mtcars, rm(cyl, vs, carb, gear, drat, disp, hp))))

model <- lm(within(train, rm(cyl, vs, carb, gear, drat, disp, hp)))
MSPE <- sum(((test$mpg - predict(model, test))^2))/4

#the MSPE is significantly higher than the MSRes so we check for collinearity

df <- within(train, rm(cyl, vs, carb, gear, drat, disp, hp))
cor(df)
#high correlation between am and wt
#wt has higher t value so remove am

summary(lm(within(train, rm(cyl, vs, carb, gear, drat, disp, hp, am))))
model <- lm(within(train, rm(cyl, vs, carb, gear, drat, disp, hp, am)))
MSPE <- sum(((test$mpg - predict(model, test))^2))/4 #2.0389
MSRes <- (2.752^2)/25 #0.3029

MSPE - MSRes #1.736

#MSPE and MSRes are reasonably close, so we accept the model as valid
```