Exploring Gas Consumption for 32 Automobile Models in 1973-74

Jay Yanamandala

October 29, 2021

Executive Summary

This report is an analysis of the relationship between Miles/(US) gallons and various factors based on data compiled by Motor Car Trend, a magazine about the automobile industry. For this report, picked up 'mtcars' a dataset created in '1974' by the same magazine.

This dataset comprises of performances for 32 (1973-74) automobile models. We are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome).

In this analysis, we explore and analyze the mtcars data set and answer following questions:

Areas of particular interest:
1. "Is an automatic **(am=0)** or manual transmission **(am=1)** better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

We fit several linear regression models to not only answer the "Areas of particular interest", but also analyze what other factors influence MPG's "Adjusted R-squared value".
1. t-test shows 2. linear model - MPG -vs- all fatures 2. stepwise linear model - backward selection 3. manual slection - based on visual inference

Welch Two Sample t-test
1. mpg -vs- am (manual/automatic transimission)
Since the p-value is 0.00137, we reject our null hypothesis. The mileage of manual transmission is 7.25 miles more than automatic transmissions.
2. mpg -vs- engine type (v-shaped -or- standard)
Since the p-value is 0.00011, we reject our null hypothesis. The mileage of manual transmission is approx 7.94 miles more than automatic transmissions.

The t-test shows that the performance difference between cars with automatic and manual transmission. And it is about 7 MPG more for cars with manual transmission than those with automatic transmission. Then, we fit several linear regression models and select the one with highest Adjusted R-squared value. So, given that weight and 1/4 mile time are held constant, manual transmitted cars are 14.079 + (-4.141)*weight more MPG (miles per gallon) on average better than automatic transmitted cars. Thus, cars that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will have higher MPG values.

Exploratory Data Analysis

Setup 'mtcars' dataset to perform exploratory data analysis 1. load required libraries 2. Since rownames names do not add much value, we will drop them

```
library(datasets)
library(ggplot2)
library(gridExtra)
data(mtcars)
rownames(mtcars) <- NULL
```

Viewing some of the entries in dataset

```
dim(mtcars)
```

```
## [1] 32 11
```

```
head(mtcars, 5)
```

```
##    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

Convert some variables from ‘numeric’ class to ‘factor’ class, and attach ‘mtcars’ to be able to use column names as variables.

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

Inference

Welch Two Sample t-test

```
test_mpg_am <- t.test(mpg ~ am)
test_mpg_am$p.value
```

```
## [1] 0.001373638
```

```
test_mpg_am$estimate[2] - test_mpg_am$estimate[1]
```

```
## mean in group 1
##        7.244939
```

```
test_mpg_vs <- t.test(mpg ~ vs)
test_mpg_vs$p.value
```

```
## [1] 0.0001098368
```

```
test_mpg_vs$estimate[2] - test_mpg_vs$estimate[1]
```

```
## mean in group 1
##        7.940476
```

Since the p-values for both the above are less than 0.05, we reject our null hypothesis.
The difference of mileage in case of 'vs' **Engine type** -versus- 'am' **manual/automatic transimission** is 0.7 miles/(US) Gallon.

Regression Analysis

Creating a Linear Regression Model with mpg as outcome and rest of the columns as predictors

```
all_features <- lm(mpg ~ ., data=mtcars)
summary(all_features)
```

The above linear model, has an adjusted values of 0.779 on 15 degrees of freedom, which means the model can explain about 78% of the variance related to mpg variable. From the above summary, we see that none of the variables are shown as significant – as all p values are greater than the threshold limit .05 + no star notations besides the predictors. So now the question is how to fit the model?

Continuing on to Stepwise Regressioni
We can look at either 'forward selection' -or- 'backward elimination'. In case of 'forward selection', we only keep adding the features, and do not delete the already added feature. in every iteration. Only those features which increase the overall model fit, are retained.

In case of 'backward elimination', we include all predictors in the first step, and in subsequent steps, keep on removing the one which has the highest p-value (>.05 the threshold limit). After a few iterations, methos will produce the final set of features which are significant enough to predict the outcome with the desired accuracy.

There are three major criteria in the stepwise regression: Cp, AIC and BIC. We will try the 'Backward Elimination' 'BIC', method with 'k' value equal to log(nrow(mtcars))

```
stepall <- step(all_features, k=log(nrow(mtcars)))
summary(stepall)
```

This model is "mpg ~ wt + qsec + am", has the Residual standard error as 2.459 on 28 degrees of freedom. And the Adjusted R-squared value is 0.8336, which means that the model can explain about 83.4% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

Picking features based on visual inspection of plots

```
manual_sel <- lm(mpg ~ cyl + hp + wt + am, data=mtcars)
summary(manual_sel)
```

This model is “mpg ~ cyl + hp + wt + am”. It has the Residual standard error as 2.41 on 26 degrees of freedom. And the Adjusted R-squared value is 0.8401, which means that the model can explain about 84% of the variance of the MPG variable. In this model, am - auto/manual is not significant - dropped with inclusion of cylinder predictor

Note:
If by adding a new variable we find the impact of already added variables decrease, .i.e. p-value crosses the upper threshold of .05 for an existing feature in the mix, this feature now becomes insignificant and we remove that feature and rebuild the model

Looking at interactions between predictors
1. Horse power increases with increase in number of cylinder as well as engine type
2. Weight increases with increase in horse power as noted in (1) above
3. Auto-transimission vehicles tend to wieght more compared to manual (stick-shift) automobiles
4. There are other interaction terms but adding those will lead to over-fitting

Model generated by visual insepction and analysis + individual interactions of features

```
manual_sel_intr <- lm(mpg ~ cyl * hp + wt * am, data=mtcars)
summary(manual_sel_intr)
```

This model has the Residual standard error as 2.172 on 23 degrees of freedom. And the Adjusted R-squared value is 0.8701, which means that the model can explain about 87% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level. Except for 6 cylinder engines.

Model generated by step function + individual interactions of features

```
stepall_intr <-lm(mpg ~ qsec + wt*am, data=mtcars)
summary(stepall_intr)
```

This model has the Residual standard error as 2.084 on 27 degrees of freedom. And the Adjusted R-squared value is 0.8804, which means that the model can explain about 88% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

Summary of Results

```
summary(stepall_intr)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

```
summary(manual_sel_intr)$coef
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  36.65880801 3.98710756  9.194336 3.639216e-09
## cyl6         -7.19711004 5.60784260 -1.283401 2.121326e-01
## cyl8        -10.82118109 4.22762121 -2.559638 1.751921e-02
## hp           -0.08268006 0.03401129 -2.430959 2.326374e-02
## wt           -2.31292740 0.81181090 -2.849096 9.082257e-03
## am1           9.14282418 4.12170068  2.218216 3.669238e-02
## cyl6:hp       0.05953742 0.05035370  1.182384 2.491331e-01
## cyl8:hp       0.07633722 0.03564540  2.141573 4.304901e-02
## wt:am1       -3.04684600 1.51645879 -2.009185 5.639473e-02
```

Conclusion

Based on the analysis conducted so far, the following can be concluded:
1. In all the models created, ‘am’ p-value is more than 0.05 and so can be discarded as a feature that has impact on ‘mpg’
2. The model generated by stepwise function with ‘interactions’ predicts better RSS by a tiny fraction compared to model picked manually.
3. For the model generated by step-wise interaction, we see that manual transmission have more miles per gallon ‘approximately 7.2’ than automatic transmissions.
4. For the model generated by manual interaction, we see that manual transmission also have more miles per gallon ‘approximately 7.9’ than automatic transmissions.
5. Even though manual interaction model gives more mileage, since some of the features have to be discarded due to p-value greater than 0.5, we will choose the model generated by step-wise interaction, which is the model with highest Adjusted R-squared value, “mpg ~ wt + qsec + am + wt:am”.

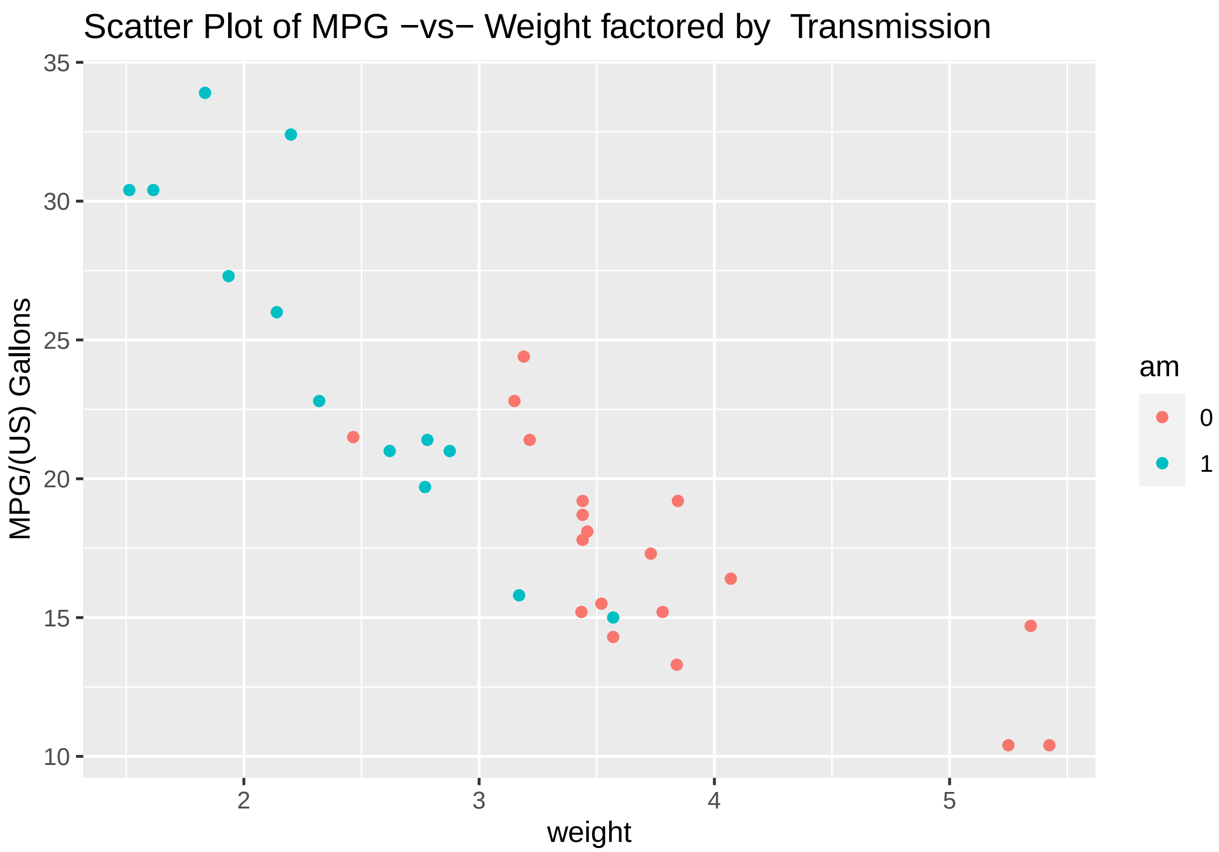
Appendix: R-Code and Plots

For Plots refer to Appendix section
From the box and scatter plots we can infer the following:
1. Miles/US gallon decreases with increase in number of cylinders
2. Miles/US gallon decreases with increase in weight of automobile
3. Miles/US gallon decreases with increase in horse power of automobile engine
4. Miles/US gallon decreases with increase in Engine Displacement 5. Miles/US gallon decreases with increase in number of carburetors, except when the number is either 6 -or- 8 6. Miles/US gallon is higher when the ‘1/4 mile time’ is higher - slow moving vehicle related to horse power of engine
7. Miles/US gallon increases if type of engine is ‘standard’
8. Miles/US gallon increases if transimission is manual compared to automatic
9. Miles/US gallon increases with increase in ‘Rear Axle Ratio’

Appendix: Figures

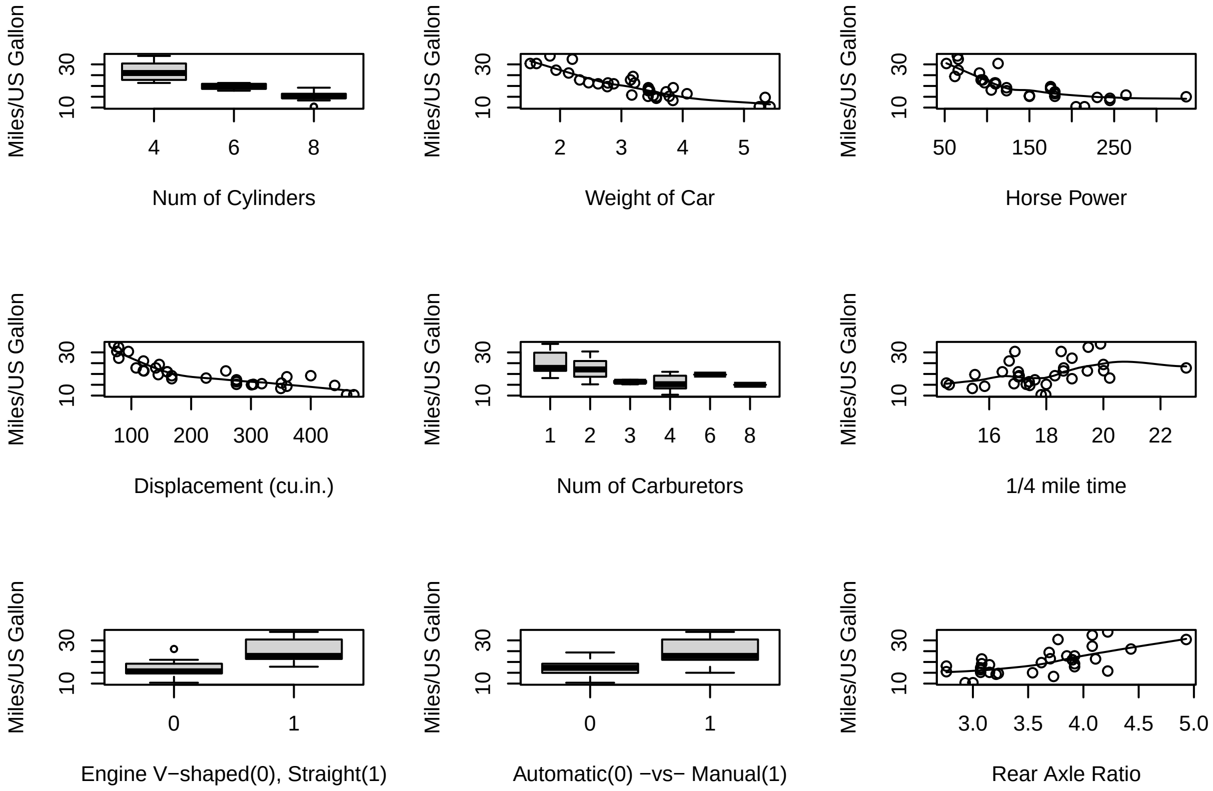
Scatter plot of mpg -vs- wt using am as factor

```r
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
  ylab("MPG/(US) Gallons") + xlab("weight") +
  ggtitle("Scatter Plot of MPG -vs- Weight factored by  Transmission")
```



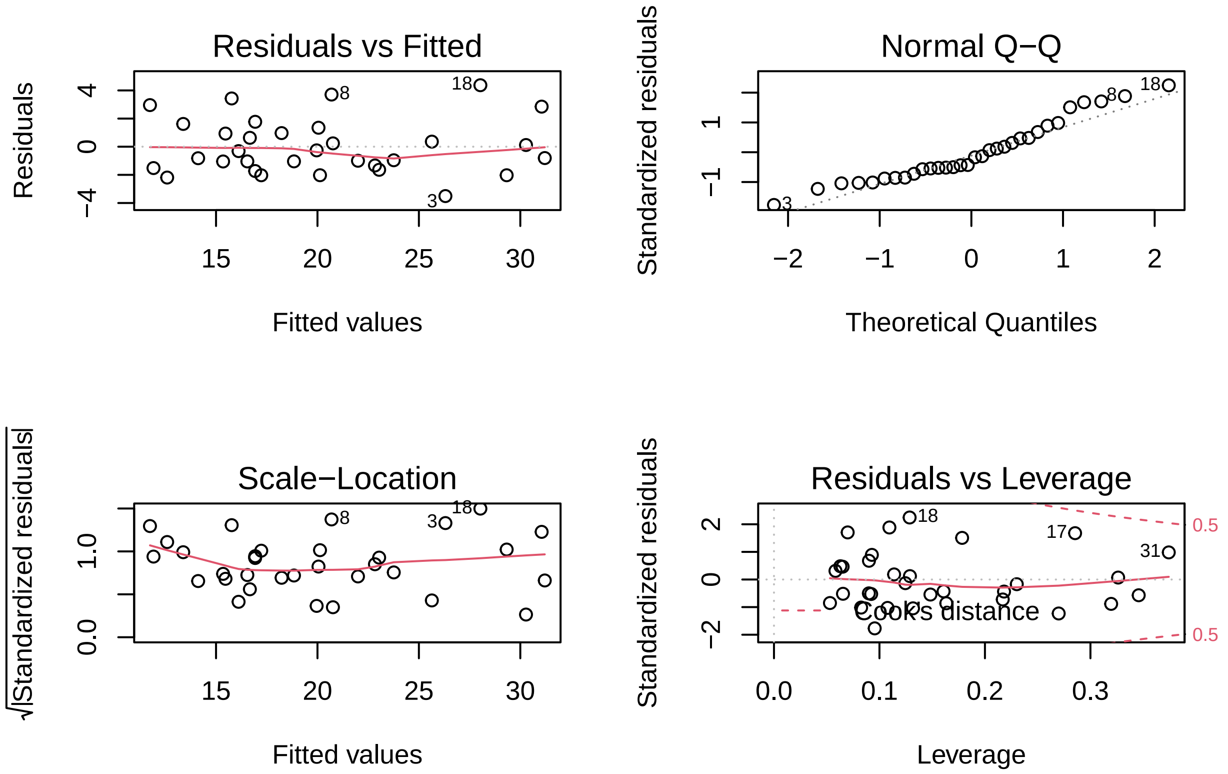
Box and Scatter plots of MPG -vs- other factors of interest in 'mtcars' dataset

```
par(mfrow=c(3,3))
plot(y=mpg, x=cyl, xlab="Num of Cylinders", ylab="Miles/US Gallon")
scatter.smooth(y=mpg, x=wt, xlab="Weight of Car", ylab="Miles/US Gallon")
scatter.smooth(y=mpg, x=hp, xlab="Horse Power", ylab="Miles/US Gallon")
scatter.smooth(y=mpg, x=disp, xlab="Displacement (cu.in.)", ylab="Miles/US Gallon")

plot(y=mpg, x=carb, xlab="Num of Carburetors", ylab="Miles/US Gallon")
scatter.smooth(y=mpg, x=qsec, xlab="1/4 mile time", ylab="Miles/US Gallon")

plot(y=mpg, x=vs, xlab="Engine V-shaped(0), Straight(1)", ylab="Miles/US Gallon")
plot(y=mpg, x=am, xlab="Automatic(0) -vs- Manual(1)", ylab="Miles/US Gallon")
scatter.smooth(y=mpg, x=drat, xlab="Rear Axle Ratio", ylab="Miles/US Gallon")
```



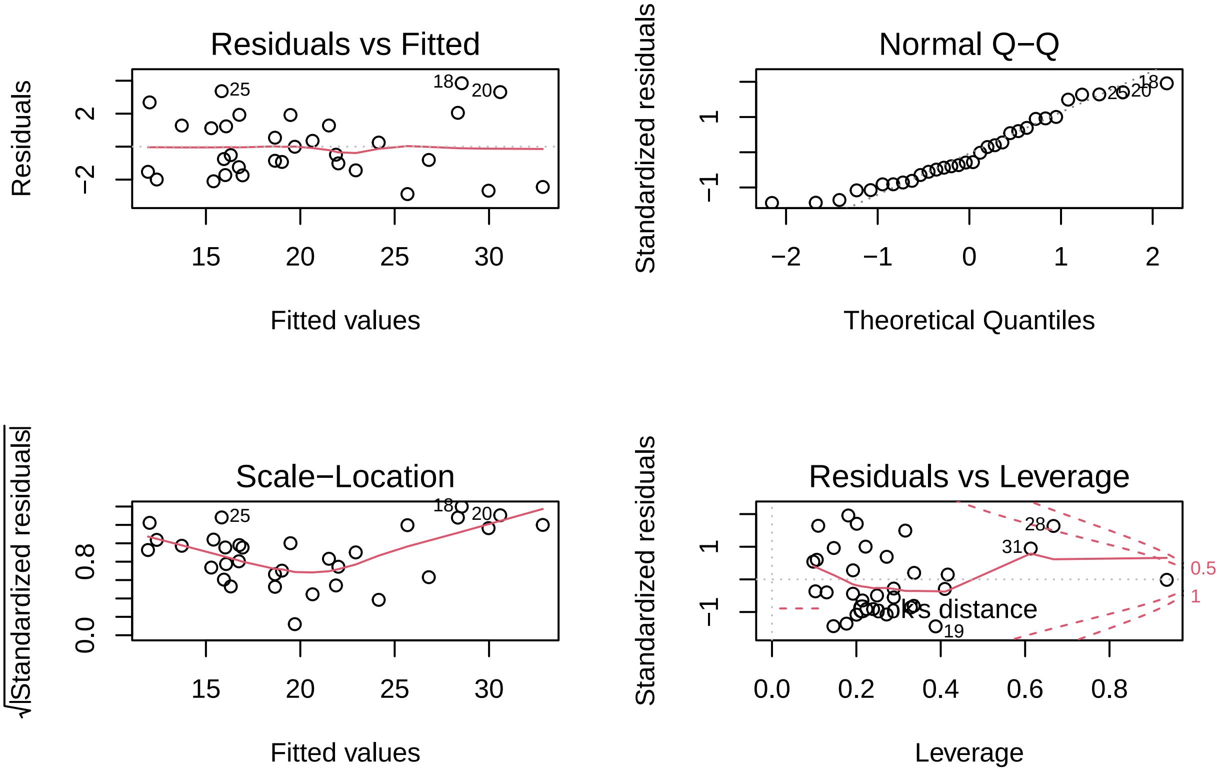
Residual Plot - Stepwise model

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



Residual Plot - Manually generated model

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



This concludes our analysis on 'mtcars' dataset