Fuel Efficiency Analysis of ‘mtcars’ Data

Executive Summary

In this report, we analyze the data from a collection of cars in Motor Trend magazine and try to ascertain the impact of the type of transmission (automatic vs. manual) on the mileage. The single-variable model seems to suggest that cars with automatic transmission have better mileage. However, after controlling for additional factors in the multi-variable model, the effect of type of transmission dissipates while the weight of the car turns out to be the most influential factor.

Exploratory Data Analysis

Before delving into the specific question regarding fuel efficiency (mpg), we will first explore the ‘mtcars’ data set to find out more about its structure, its variables and the relationship between them.

```r
## Load data set and required packages
data("mtcars");library(knitr); library(tidyverse); library(ggplot2); library(GGally)
## View the first few observations of the data set
head(mtcars,3) %>% kable()
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1

We can see that each line of *mtcars* represents one model of car, which are labelled in the rownames. Each column is then one attribute of the specific model, such as the miles per gallon (**mpg**), the number of cylinders (**cyl**), the engine’s horsepower (**hp**), whether the car has an automatic or manual transmission (**am**), etc.

To get an better idea of the relationship between mpg and the rest of the variables, we start by exploring the correlations among them. The computed correlation coefficients are shown below. For simplicity, it only shows the correlation between mpg and the rest of the variables. See Figure 1 for the full pairwise correlation between all the variables.

```r
corr <- cor(mtcars) %>% as.data.frame() %>% rownames_to_column(var = "var") %>%
         filter(var == "mpg") %>% format(digits = 3)
kable (corr)
```

var	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1	-0.852	-0.848	-0.776	0.681	-0.868	0.419	0.664	0.6	0.48	-0.551

From the coefficients, we can see that the variable of interest **am** is positively correlated with **mpg** at 0.6, which is moderately strong. We can also detect other variables with a strong linear relationship with **mpg**.

Is an automatic or manual transmission better for MPG

Single Variable Model. Focusing the specific relationship between type of transmission and fuel efficiency, we first compare the differences in distribution between the two groups (automatic vs. manual). The boxplot in Figure 2 shows the comparison. From the visualization, we can form a hypothesis that manual cars have a higher miles per gallon, and therefore a better fuel efficiency. To confirm this hypothesis, we can fit a single linear model.

```
fit_1 <- lm(mpg ~ am, data = mtcars)
summary(fit_1)[c("coefficients", "r.squared")]
```

```
## $coefficients
##              Estimate Std. Error   t value                Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 0.000000000000001133983
## am           7.244939   1.764422  4.106127 0.000285020743935067769
##
## $r.squared
## [1] 0.3597989
```

The intercept is the average miles per gallon when am is equal to 0 (i.e. automatic transmission) and the slope of the “am”" variable show the effect of the change in transmission type (from 0 to 1) on fuel efficiency: using 1 gallon of gas, cars with manual transmission can travel 7.24 more miles than those with automatic transmission. Additionally, the low p-value suggests that the effect is not due to chance alone. However, we can see that the R-squared statistic is at 0.3598, meaning that only 35.98% of the variation in “mpg” is explained by this model.

Multiple Regresssion. Recall that in the exploratory analysis, we detected other variables with a strong correlation to the “mpg”. We can try adding “wt”, “disp”, “cyl” to the model (the variables with the highest correlation to “mpg”) and see if the model strength improves.

```
fit_m1 <- lm(mpg ~ am + wt + disp + cyl, data = mtcars)
summary(fit_m1)[c("coefficients", "r.squared")]
```

```
## $coefficients
##                Estimate Std. Error    t value             Pr(>|t|)
## (Intercept) 40.898313414 3.60154037 11.3557837 0.000000000008677574
## am           0.129065571 1.32151163  0.0976651 0.922919644373322301
## wt          -3.583425472 1.18650433 -3.0201537 0.005468412488207290
## disp         0.007403833 0.01208067  0.6128661 0.545092996564771282
## cyl         -1.784173258 0.61819218 -2.8861142 0.007581533437721448
##
## $r.squared
## [1] 0.8326661
```

It turns out that, after accounting for the effects of other key variables on fuel efficiency, the type of transmission ceases to be a significant variable. The most significant variable seems to be weight and the coefficients suggest that as the weight of the car increase 1000 lbs, the car travels 3.58 miles less using 1 gallon of gas. Additionally, the R-squared statistic improved to 0.833.

Residual Diagnostics

In Figures 3 and 4, we can see the diagnostics plots of the two models.

- **Residuals vs Fitted**: the residuals spread more or less equally around a horizontal line without distinct patterns, which is a good indication that there are not non-linear relationship left unexplained by our models.
- **Normal Q-Q**: the residuals more or less follow a straight line in both cases, which indicates normal distribution of the residuals.
- **Scale-Location**: the residuals spread more or less equally along the fitted values. This suggests that the assumption of equal variance are met.
- **Residuals vs Leverage**: we can barely see the Cook’s distance lines, which suggests that there are no influential outliers.

Appendix

Figure 1. Pairwise Correlation Plot

```
corr_plot <- mtcars %>%
        ggpairs(upper = list(continuous = wrap("cor", size = 4)),
                lower = list(continuous = wrap("smooth", method = "lm"))
                ) %>% print(progress = FALSE)
```

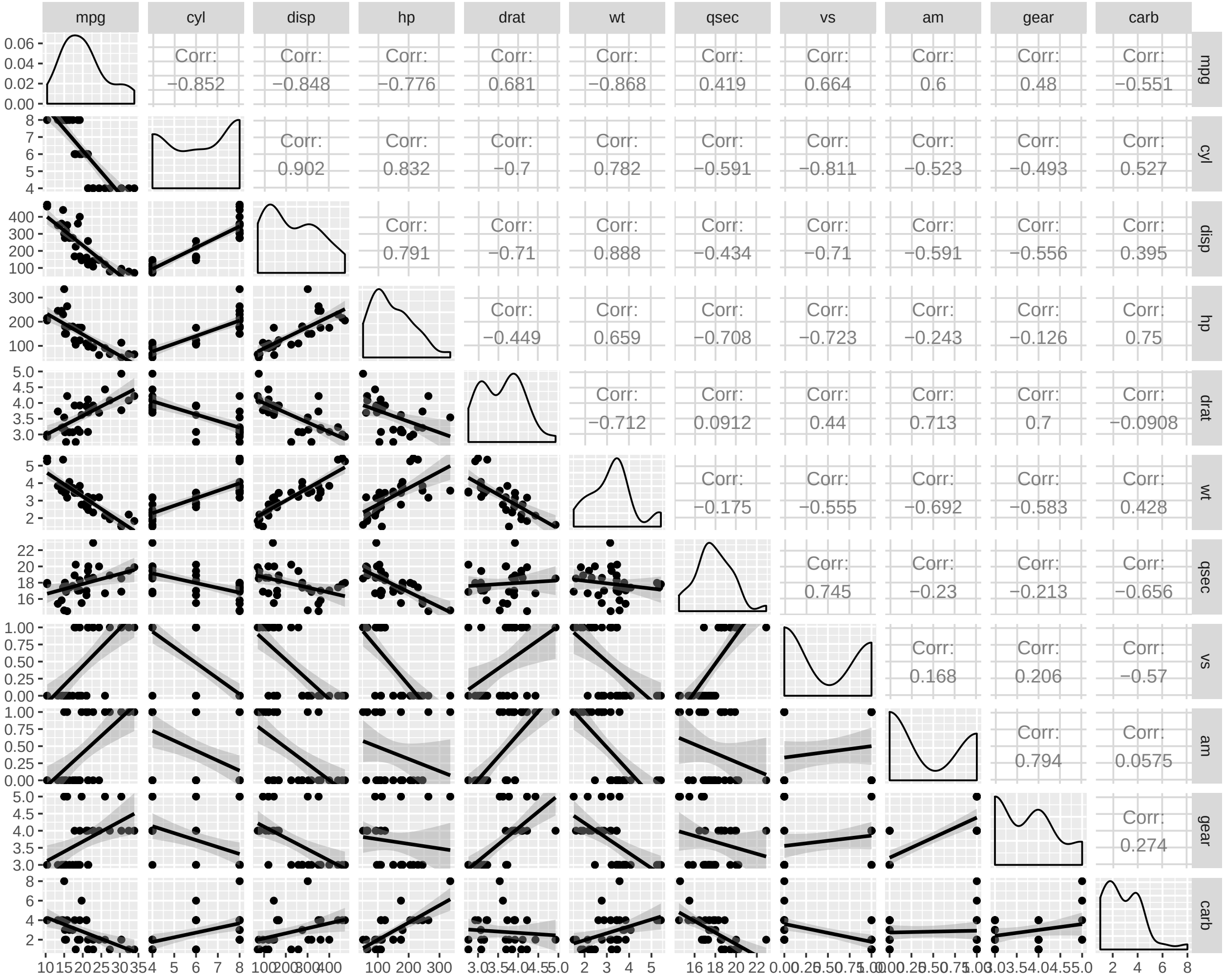


Figure 2. Comparison of MPG - Automatic and Manual Transmissions

```
boxplot_am <- mtcars %>% ggplot(aes(x = factor(am), y = mpg)) + geom_boxplot() +
        scale_x_discrete(labels = c("Automatic", "Manual")) +
        labs(x = "Transmission Type", y = "Miles per Gallon (MPG)")
boxplot_am
```

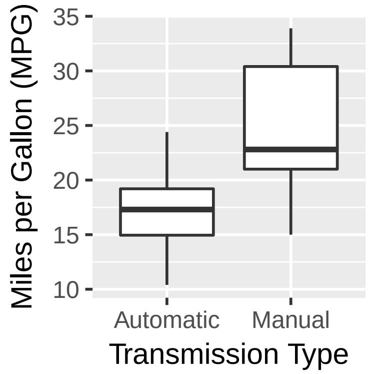


Figure 3. Residual Plots - Single Variable Model

```r
par(mfrow=c(2,2));plot(fit_1)
```

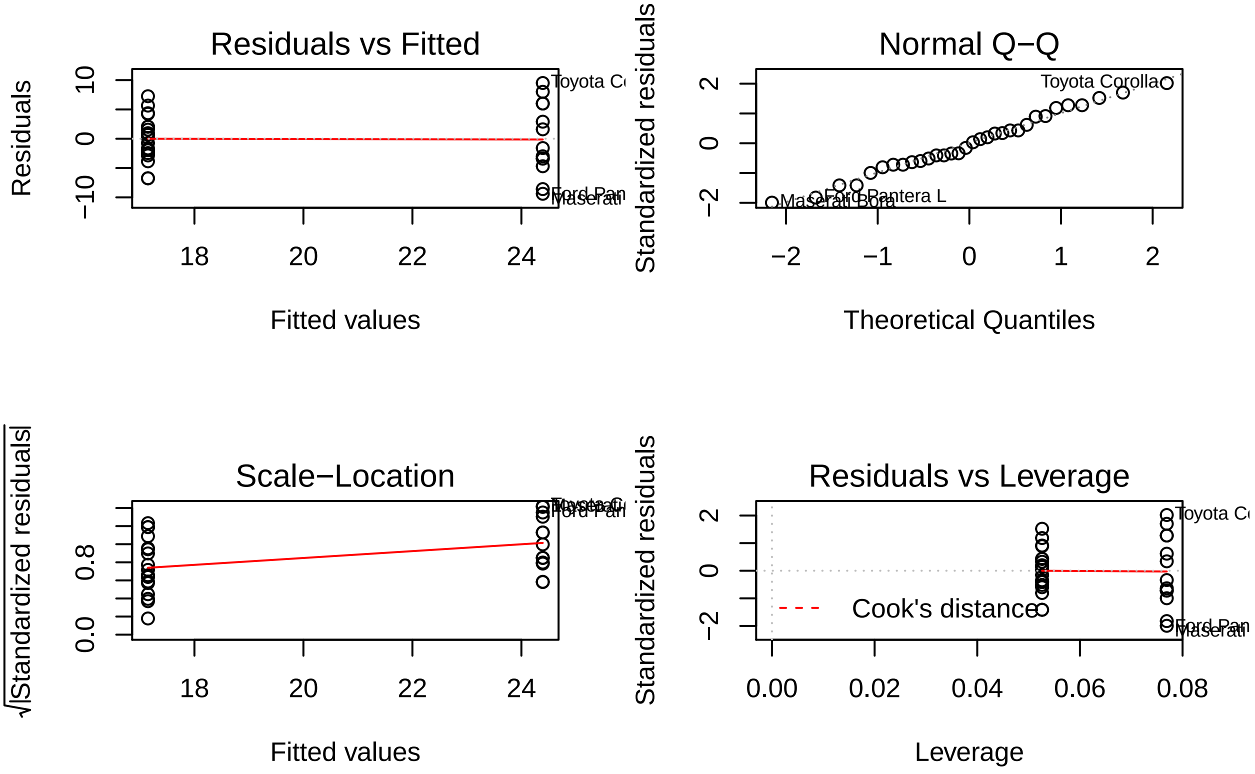
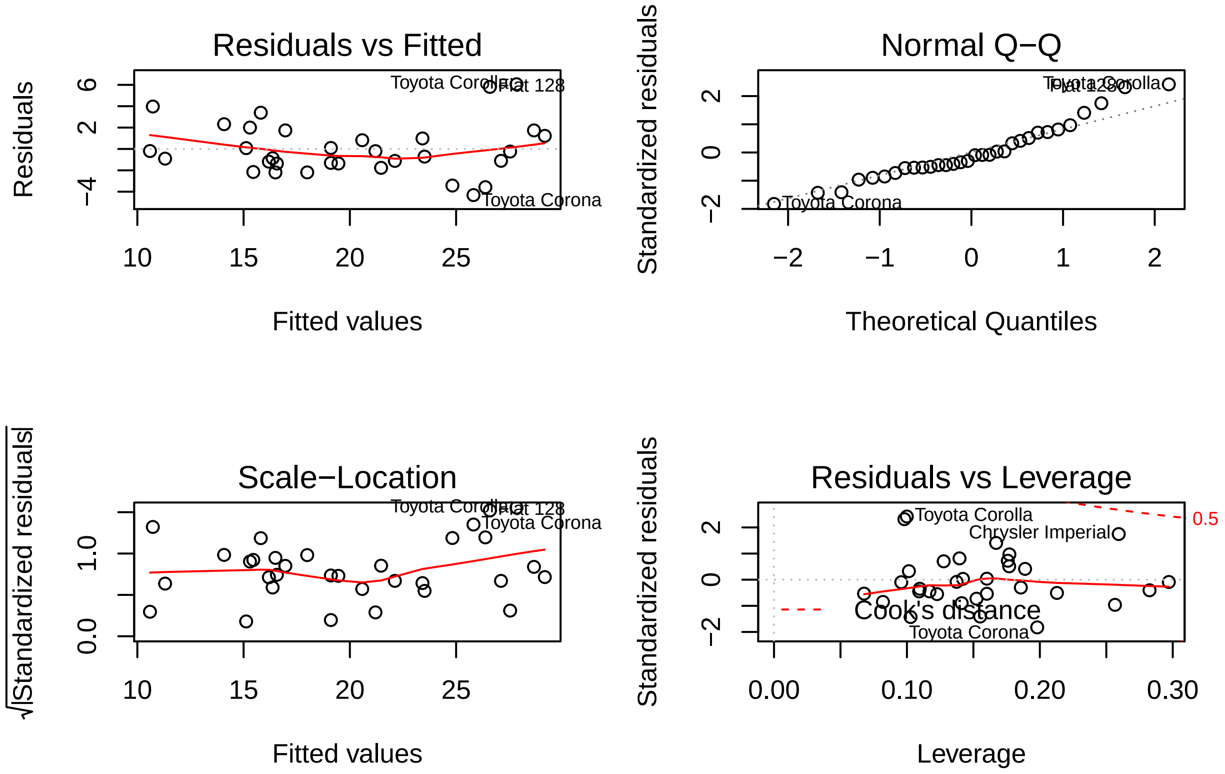


Figure 4. Residual Plots - Multi-variable Model

```r
par(mfrow=c(2,2));plot(fit_m1)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
6
2
−4
10
15
20
25
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Toyota Corona
2
0
−2
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
0.0
1.0
10
15
20
25
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5
2
0
−2
0.00
0.10
0.20
0.30