MT is Better than AT for Fuel Efficiency (MPG)

Exclusive Summary and Synopsis

This report tries to answer these two questions.
- *"Is an automatic or manual transmission better for MPG"*
- *"Quantify the MPG difference between automatic and manual transmissions"*
By exploring into some data, We found out that the fuel efficiency (Miles/(US) gallon, MPG) is **influnced** by the automaticity of the transmissions system.
This essay will show all steps during my analysis. All the details will be shown in the 2-page report. In order to make this report reproducible, the codes, graphs and results will be shown on the appendix.

Part1. Getting and Cleaning Data

In this step, I'll get the dataset `mtcars`. `mtcars` dataset is an embedded dataset in R `datasets` package. It's extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles.
To make the analysis more flexible, we will firstly factorize the discrete variables with `factor` function.

Part2. Exploratory Data Analysis

In this step, We may take a glimpse of the mtcars data. First use the `cor` function to get the correlations between `mpg` and each of other variables. Also we will draw a plot Figure 1 of correlations between different variables.
Table 1. The correlations between the 'mpg and other variables

```
##         cyl   disp     hp  drat     wt  qsec    vs  am gear   carb
## [1,] -0.852 -0.848 -0.776 0.681 -0.868 0.419 0.664 0.6 0.48 -0.551
```

Second we will draw the box plot Figure 2 of the `mpg` variable against the influence by factor `am`.
Figure 1 and Figure 2 will be shown in the Appendix.
As is shown in the *Table 1* and *Figure 1*, we can draw an intuitive conclusion that `am` influences the `mpg` variable. Then we will show and quantify this conclusion.

Part3. Inference with the Models

In this part we will analysis deeply into the dataset. Firstly test the `am`'s influence toward the grouped `mpg` means. Secondly find and select the optimal linear regression model.

One Way ANOVA, Test of significance of the causality between the `am` and `mpg`

ANOVA is used to analyze a factor's influence towards the grouped outputs. ANOVA is based on the assumption of homogeneity of variances. Let's test it first.

```
## The P-value is:
bartlett.test(mpg ~ am, data = mtcars.fact)$p.value
```

```
## [1] 0.07248
```

So we cannot reject the assumption of homogeneity of variances. So we will test the factor `am` with ANOVA next.

Table 2. ANOVA Table

```
##             Df Sum Sq Mean Sq F value  Pr(>F)    
## am           1    405     405    16.9 0.00029 ***
## Residuals   30    721      24                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is significantly small, thus we will draw to the conclusion that the variable `am` influences the mean of different cars' MPG.

Linear Regression Model Selection

In this sub-part, we will firstly fit the linear models for `mpg` against all other variables, and use the `step` function to delect some variable to find the optimal linear models.

```r
fit.whole <- lm(mpg ~ ., data = mtcars.fact)
fit.optimal <- step(fit.whole, direction = 'both')
```

```r
print(fit.optimal$call)
```

```
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars.fact)
```

As is printed above, the optimal linear models includes the numeric argument `hp`, `wt`, and factoral argument `cyl`, `am`.
Now we will test different models with some combinations of arguments `hp wt cyl` and `am`. We will use the R^2 criterion.

```r
fit.hpwt <- lm(mpg ~ hp + wt, data = mtcars.fact)
fit.hpwt.cyl <- lm(mpg ~ hp + wt + cyl, data = mtcars.fact)
```

Table 3. The R^2 of Each Linear Models

```
##           hp + wt      hp + wt + cyl hp + wt + cyl + am 
##            0.8148             0.8361             0.8401
```

According to the table, the linear model fitted with the variable `am` can fit better, compared to several other models. Thus `am` has the influential effects towards the `mpg`.

Part4. Diagnostics of the Optimal Linear Models

At the beginning, we will draw some graphs of the optimal regression model. These graphs contains the **Residual vs Fitted Graph**, **Q-Q Graph**, **Scale-Location Graph** and the **Residuals vs Leverage Graph**.
Figure 3 is shown on the Appendix
Take a glimpse at the Figure 2 four graphs, we can find out that some models is not quite obey the regression model. Obviously, they are **Toyota Corolla** and **Fiat 128**. The Residuals graph shows that the residuals of the two models is quite large, and the Normal Q-Q Plot shows that their residual is doesn't follow the Normal Distribution.
Regardless of the two special cases, the conclusion that the influence of the `am` towards `mpg` is significant is easy to find out.

Appendix

```
## Warning: package 'corrplot' was built under R version 3.1.1
```

Figure 1 from the Part 2

```r
M <- cor(mtcars)
corrplot.mixed(M, lower = "number", upper = "circle", title = "Correlations between Different Variances
```

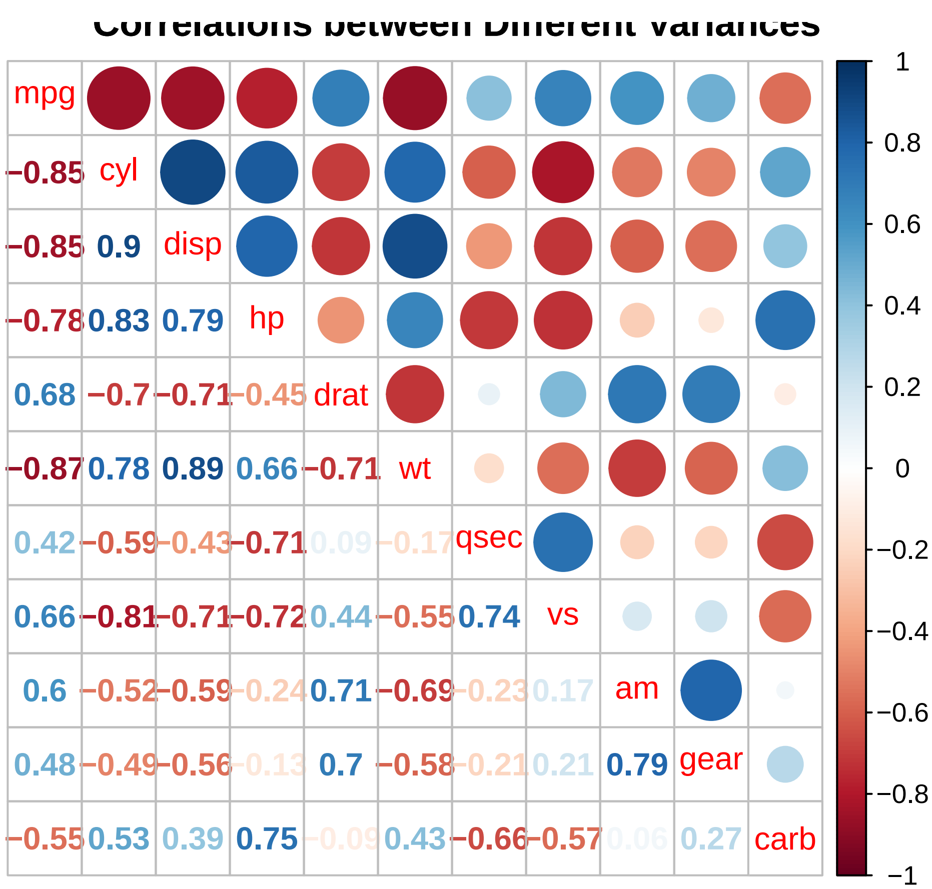


Figure 2 from the Part 2

Figure 2. The box plots of the `mpg` variable against the influence by factor `am`

```r
boxplot(mpg ~ am, data = mtcars, names = c("Automatic", "Manual"))
```

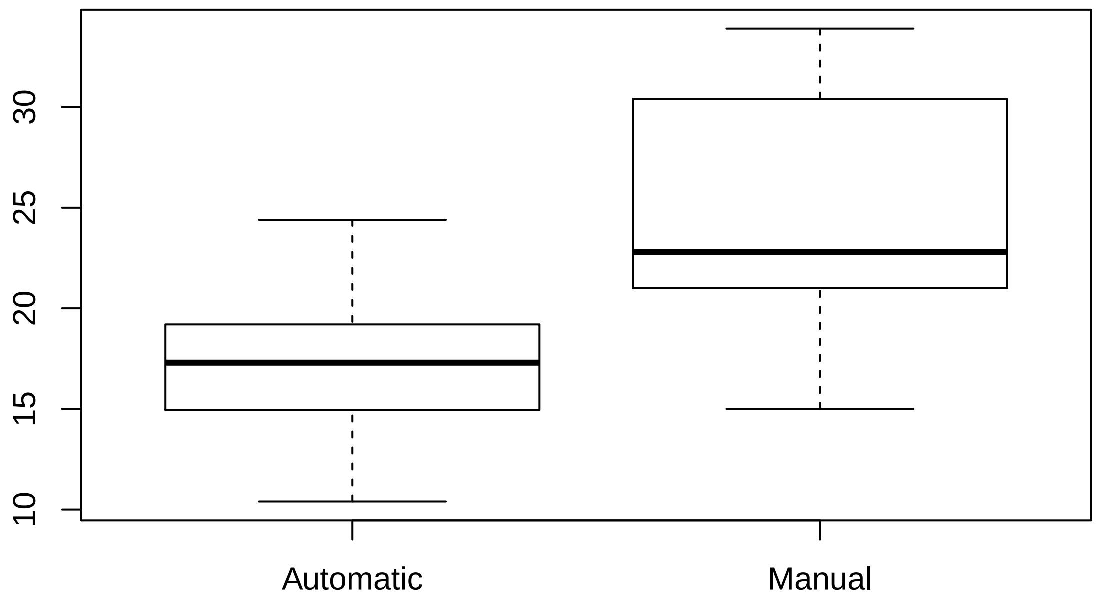
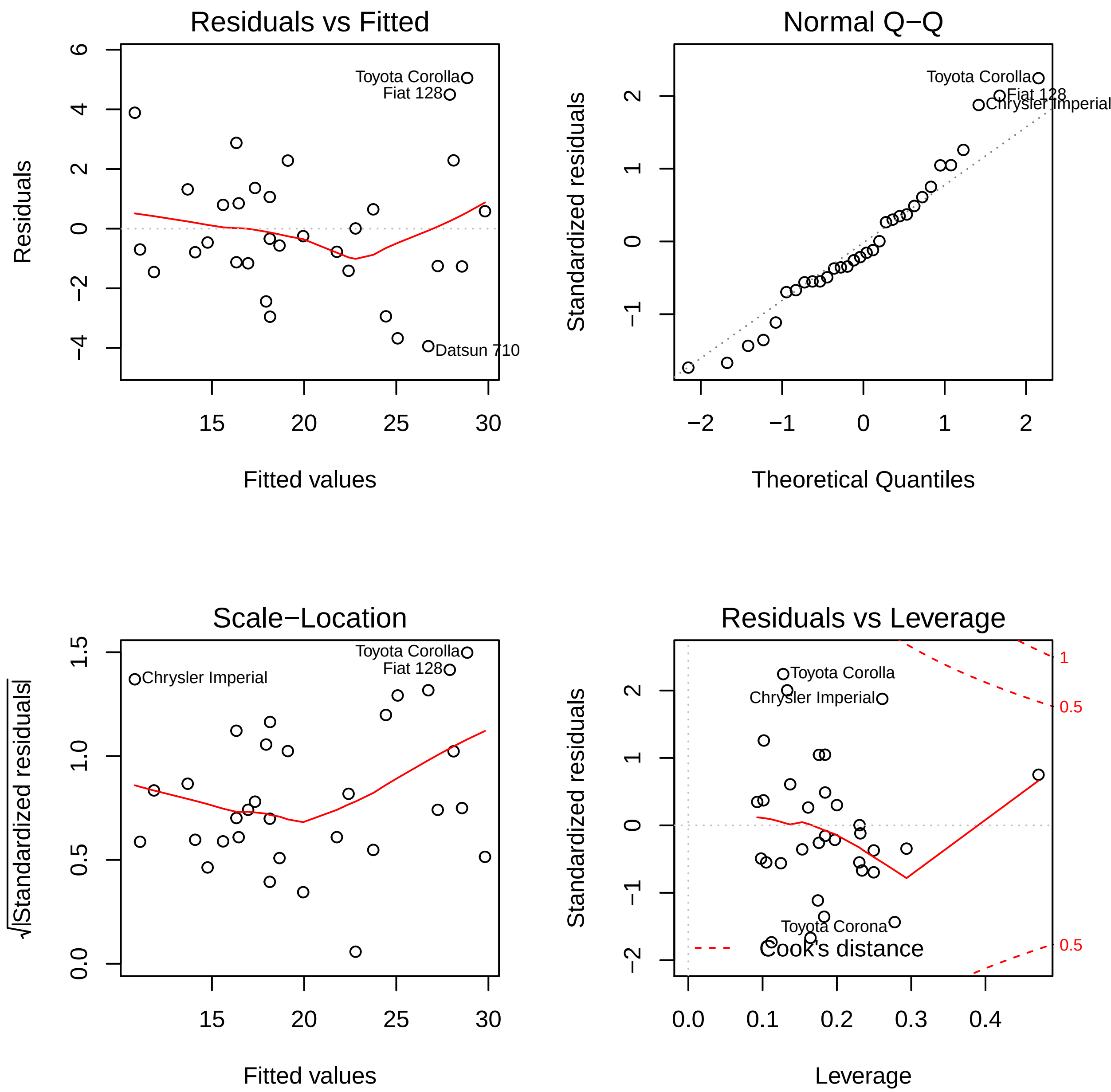


Figure 3 from the Part 4

Figure 3. Diagnostical Plots of the Optimal Linear Model

```
par(mfrow = c(2,2))
plot(fit.optimal)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance