Motor Trend : Regression Models Course Project

Executive Summary

The mtcars dataset has been analyzed in this report to help Motor Trend, an automobile magazine and derive the relation between miles per gallon(MPG) and a set of variables. The analysis looks to answer two key questions.

1. Is an automatic or manual transmission better for MPG?

2. Quantifying how different is the MPG between automatic and manual transmissions?

We use exploratory analysis and regression techniques to answer these questions. t-test has been used to view the performance difference between cars having manual and automatic transmission.

During analysis, it has been captured that manual transmissions have a higher value MPG compared to automatic transmissions. The switch to manual transmission from automatic one witnessed an increase of around 1.8MPG. Also, there is a significant difference between the means of MPG for manual and automatic transmissions.

Data Processing and Transformation

The dataset, mtcars, is loaded and the required data is transformed by factoring the necessary variables.

```r
data(mtcars)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
```

Exploratory Analysis

On exploratory analysis, we found that mtcars is a dataframe with 11 variables.
mpg = Miles/(US) gallon,
cyl = Number of cylinders,
disp = Displacement (cu.in.),
hp = Gross horsepower,
drat = Rear axle ratio, wt = Weight (lb/1000),
qsec = 1/4 mile time,
vs = V/S,
am = Transmission (0 = automatic, 1 = manual), gear = Number of forward gears and carb = Number of carburetors

On examining bivariate plots of variables, we see that variables like cyl, disp, hp, drat, wt, vs and am seem to have some strong correlation with mpg. See Appendix, Figure 1. We see a relationship between transmission and fuel consumption; manual transmissions yielding higher values of MPG.

We use a boxplot to study the effects of car transmission type on mpg by plotting mpg when am is Automatic or Manual. See Appendix, Figure 2.

This plot leads us to the conclusion that there is an increase in mpg when the transmission is manual.

Regression Analysis - Model building and selection

As a first step, we fit the entire model by considering all the variables as predictors.

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

We then perform stepwise model selection to select significant predictors for the final model which turns out to be the best model. The step method which runs multiple times to build multiple regression models and selects the best variables from them.

```
bestmodel <- step(initialmodel, direction = "both")
```

The details of the best model can be found out by:

```
summary(bestmodel)

## 
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4811 -1.5555 -0.7257  1.4110  4.6610 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   9.6178     6.9596   1.382 0.177915    
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am1           2.9358     1.4109   2.081 0.046716 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336 
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

The model explains 84% of the variance, and each of the coefficients for weight and engine displacement is statistically significantly greater than 0.

Residuals and Diagnostics

The residual plots of our regression model is studied and the regression diagnostics for our model is computed to detect outliers in the dataset. See Appendix, Figure 3.

According to the residual plots, we can verify the following underlying assumptions:

1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.
2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

Statistical Inference

We perform a t-test assuming that the transmission data has a normal distribution and we clearly see that the manual and automatic transmissions are significatively different.

```
t.test(mpg ~ am, data = mtcars)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1 
##        17.14737        24.39231
```

Conclusions

From the observations of the best fit model, we can conclude the following:

1. Cars with Manual transmission get around 1.8 times more miles per gallon mpg compared to cars with Automatic transmission.
2. mpg will decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in wt.
3. mpg decreases negligibly with increase of hp.
4. If number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am).

Appendix

Figure 1 : Pairs plot for the dataset

```
pairs(mpg ~ ., data = mtcars)
```

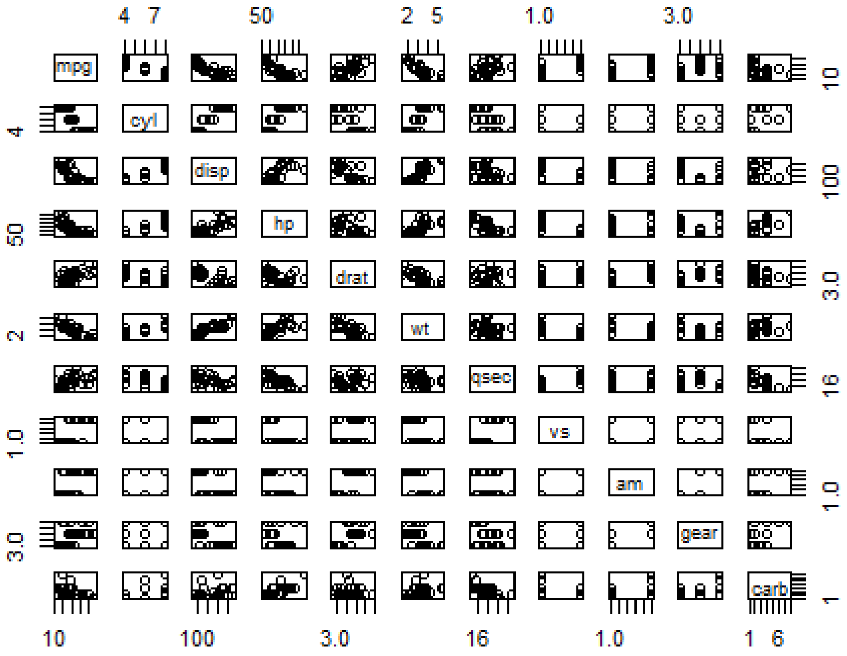


Figure 2 : Boxplot of miles per gallon by transmission type

```r
boxplot(mpg ~ am, data = mtcars, main = "Boxplot of miles per gallon by transmission type", col = (c("red","blue")), ylab = "Miles Per Gallon", xlab = "Transmission Type")
```

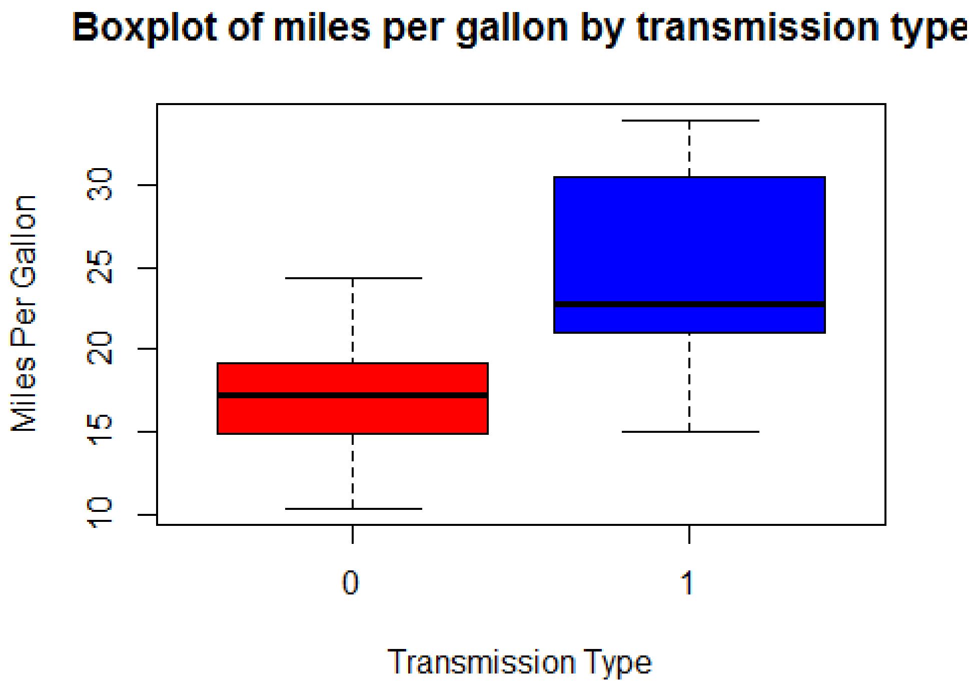
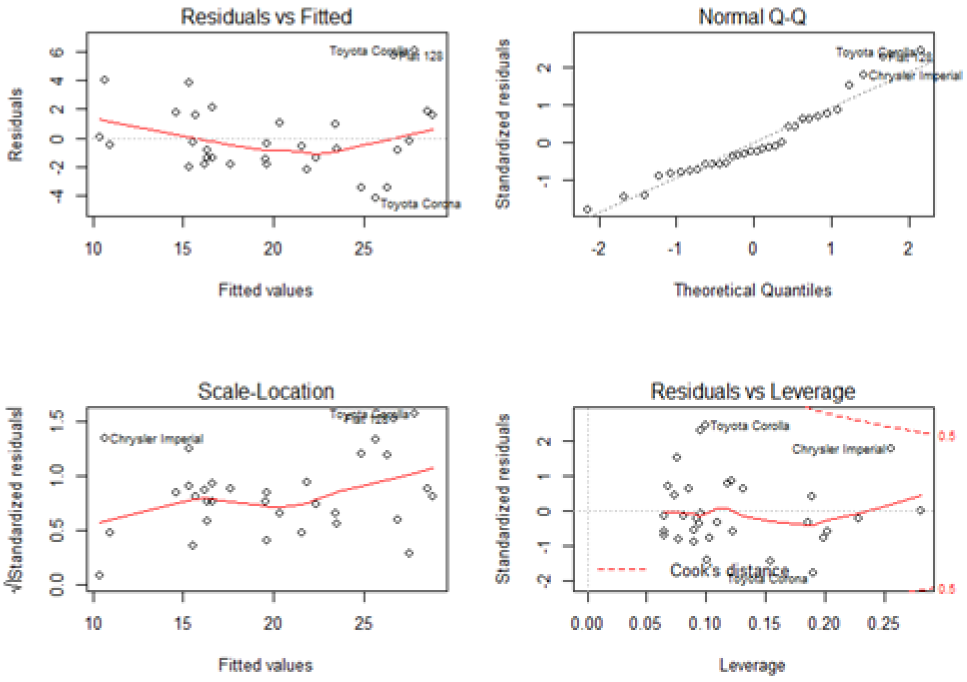


Figure 3 : Residuals and Diagnostics


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
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
0.5
0.5