# Regression Models Course Project

*A. Swarup*

*May 20, 2018*

## Executive Summary

In this report we analyze the relatioship between transmission type (automatic or manual) and miles per gallon (MPG). The report sets out to determine which tranamission type produces higher MPG. We use the mtcars (Motor Trend Car Road Tests) dataset for the analysis. The data was extracted from the 1974 Motor Trend US Magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). We perform the analysis using exploratory data analysis and regression modelling techniques. Simple linear regression yields the result that manual transmission cars give, on average, 7.245 more miles per gallon than automatic transmission cars. However, after further analysis using multivariate regression techniques we find that other variables like "weight" and "1/4 mile time" (mostly weight) contribute to MPG, and manual transmission cars, on average, give only 2.94 miles more per gallon than automatic transmission cars.

## Initial Data Processing

Load the mtcars dataset and covert some categorical variables to factors.

```
rm(list=ls()) # remove all data store in the Data Environment
library(knitr)
library(ggplot2)
data("mtcars")
# Change some numeric predictor variables to factors and assign labels to levels
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
```

A Basic Summary of 'mtcars' data can be seen in the Appendix.

## Exploratory Data Analysis

Since we need to analyze MPG difference between automatic and manual transmissions, we perform exploratory analysis with am and mpg variables.

First, we compare the means of MPG for automatic and manual transmissions. Please refer to Appendix for the code. The mean MPG for Manual (24.39) is greater than Automatic (17.15)

Next, we show in Figure 1 of Appendix a boxplot of MPG versus Transmission. Looking at this also it is observed that manual transmission is better than automatic. Next, we will be running a linear regression test on this data. For Linear Regression, we need to ensure that the following basic assumptions are met:

- The distribution of mpg is approximately normal
- Outliers are not skewing the data

We plot dependent variable mpg to check its distribution - please see Figures 2a and 2b of the Appendix. By these plots we confirm that distribution of mpg is approximately normal and there are no apparent outliers skewing our data.

## Detailed Data Analyses

### t-test

Null hypothesis is that the mean MPG is the same for both Manual and Automatic cars. We set our alpha-value at 0.5 and run a t-test to analyse further.

```
autoData <- mtcars[mtcars$am == "Automatic",]
manualData <- mtcars[mtcars$am == "Manual",]
ttest <- t.test(autoData$mpg, manualData$mpg); ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  autoData$mpg and manualData$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

The p-value is 0.0014, so we can reject the null hypothesis and conclude automatic has low mpg compared with manual cars. This ratifies our observations as seen in the boxplot of Figure 1. However, this conclusion would be incomplete without considering other characteristics of auto and manual cars. Therefore, we explore further using multiple linear regression analyses techniques.

### Simple Linear Regression

Let us perform a linear regression on the data and see what the model says.

```
fit1 <- lm(mpg ~ am, data = mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

From the summary of model fit1, the intercept (am = 0 for Automatic) is 17.147 and the coefficient of amManual is 7.245. This means the mean for manual is 7.245 more than that of automatic. However,

R-squared for this model is 0.3598 which means this model is explaining only 36% of the variance. Other variables should be added in to get a higher Adjusted R-Squared value.

### Multivariate Regression Analysis

We use a stepwise algorithm to choose the best linear model by using step().

```
fittotal <- lm(mpg ~ ., data = mtcars)
fitstep <- step(fittotal, direction="both", trace=FALSE, steps=10000)
summary(fitstep)
```

```
##
## 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 ***
## amManual      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
```

Resulting model: formula = mpg ~ wt + qsec + am, shows that in adddition to transmission (am), wt (weight) & qsec (1/4 mile time) are most significant in explaining the variations in mpg.

### Best Model - am + wt + qsec

To quantify the mpg difference between automatic and manual transmission, we include 3 variables am, wt, and qsec.

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

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, 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
## amManual      2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
```

```
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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 adjusted $R^2$ is 0.8497 which means that the model handles 84.97% of the variation in mpg. We can safely conculde that this is a robust and highly predictive model. The p-value is 1.21 x $10^{-11}$

### Residuals Diagnostics of Final Model

Please refer to the residual plots in the Appendix. Here we see that 'Normal Q-Q' plot looks ok, but the 'Residuals vs Fitted' and 'Scale-Location' both show worrysome trends. That is for this model which is based on only 32 observations to train on, we cannot say with confidence that the model will fit all future observations.

### Conclusion

Cars with manual transmission get better miles per gallon compared to those with automatic transmission.

- The t-test shows that manual transmission cars get an average of 7.25 MPG more than cars with automatic transmission.
- Several linear regression models were fitted to evaluate different aspects that could impact MPG. The best fitted model lm(formula = mpg ~ am + wt + qsec, data = mtcars) showed that when "wt" (weight (lb/1000)) and "qsec" (1/4 mile time) remain constant, manual transmission cars get an average of 2.94 more MPG than those with automatic transmission.

# APPENDIX

### Basic Summary of 'mtcars' Data

```
kable(summary(mtcars[1:5])); kable(summary(mtcars[6:10]))
```

| mpg | cyl | disp | hp | drat |
|---|---|---|---|---|
| Min. :10.40 | Min. :4.000 | Min. : 71.1 | Min. : 52.0 | Min. :2.760 |
| 1st Qu.:15.43 | 1st Qu.:4.000 | 1st Qu.:120.8 | 1st Qu.: 96.5 | 1st Qu.:3.080 |
| Median :19.20 | Median :6.000 | Median :196.3 | Median :123.0 | Median :3.695 |
| Mean :20.09 | Mean :6.188 | Mean :230.7 | Mean :146.7 | Mean :3.597 |
| 3rd Qu.:22.80 | 3rd Qu.:8.000 | 3rd Qu.:326.0 | 3rd Qu.:180.0 | 3rd Qu.:3.920 |
| Max. :33.90 | Max. :8.000 | Max. :472.0 | Max. :335.0 | Max. :4.930 |

| wt | qsec | vs | am | gear |
|---|---|---|---|---|
| Min. :1.513 | Min. :14.50 | Min. :0.0000 | Automatic:19 | Min. :3.000 |
| 1st Qu.:2.581 | 1st Qu.:16.89 | 1st Qu.:0.0000 | Manual :13 | 1st Qu.:3.000 |
| Median :3.325 | Median :17.71 | Median :0.0000 | NA | Median :4.000 |
| Mean :3.217 | Mean :17.85 | Mean :0.4375 | NA | Mean :3.688 |
| 3rd Qu.:3.610 | 3rd Qu.:18.90 | 3rd Qu.:1.0000 | NA | 3rd Qu.:4.000 |
| Max. :5.424 | Max. :22.90 | Max. :1.0000 | NA | Max. :5.000 |

### Means and Boxplot of MPG versus Transmission

```
mean(mtcars[mtcars$am=="Manual", "mpg"])
```

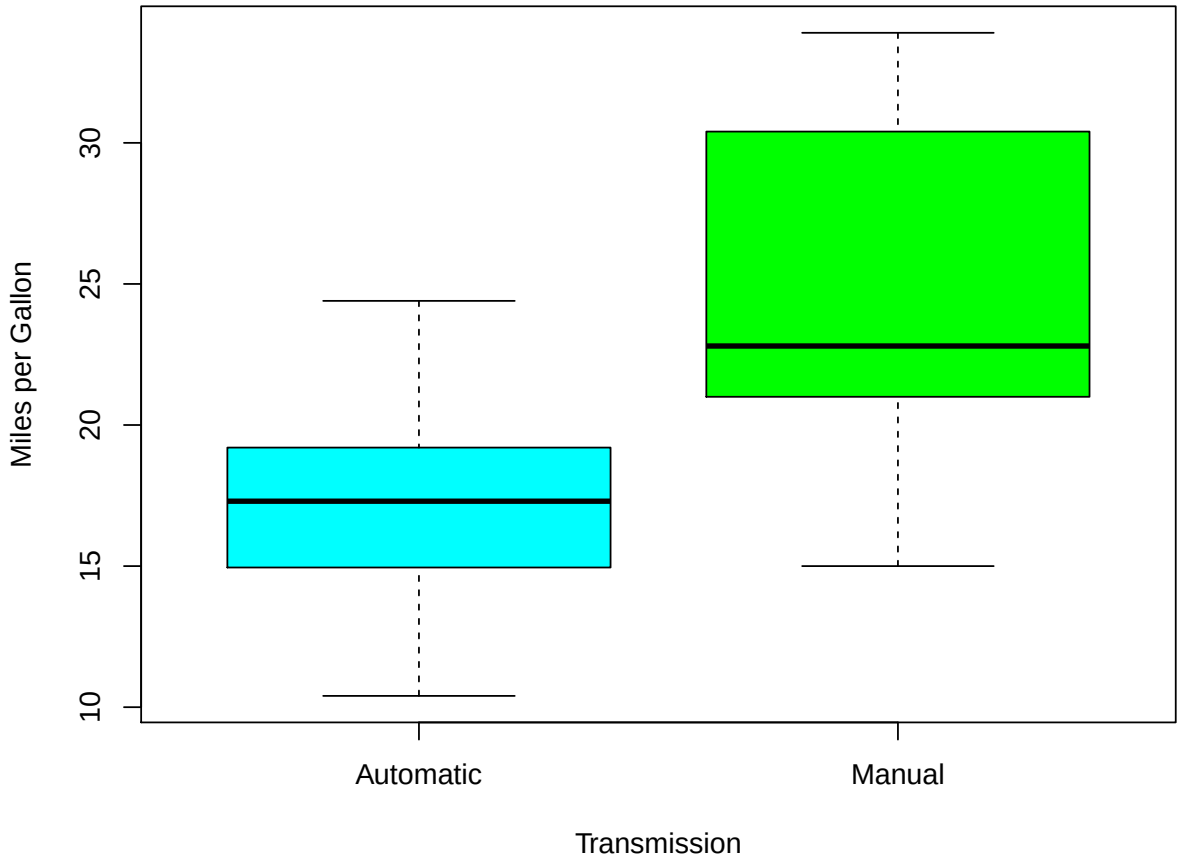
```
## [1] 24.39231
```

```
mean(mtcars[mtcars$am=="Automatic", "mpg"])
```

```
## [1] 17.14737
```

```
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission",
        ylab = "Miles per Gallon",
        main = "Figure 1: Boxplot", col = c("cyan", "green"))
```

**Figure 1: Boxplot**




### Historgram and Density Plots of MPG

```
par(mfrow = c(1, 2))
g <- mtcars$mpg
# Overlay normal curve to histogram of MPG
h <- hist(g, breaks=10, density = 10,
          col="lightgray", xlab="Miles Per Gallon",
```

```
         main="Figure 2a. Histogram of MPG")
xfit <- seq(min(g), max(g), length=40)
yfit <- dnorm(xfit, mean=mean(g), sd=sd(g))
yfit <- yfit * diff(h$mids[1:2]) * length(g)
lines(xfit, yfit, col="black", lwd=2)
# Kernel Density Plot (smoothed histogram) of MPG
d <- density(mtcars$mpg)
plot(d, xlab = "Miles per Gallon (MPG)", main ="Figure 2b. Density Plot of MPG")
```

**Figure 2a. Histogram of MPG**

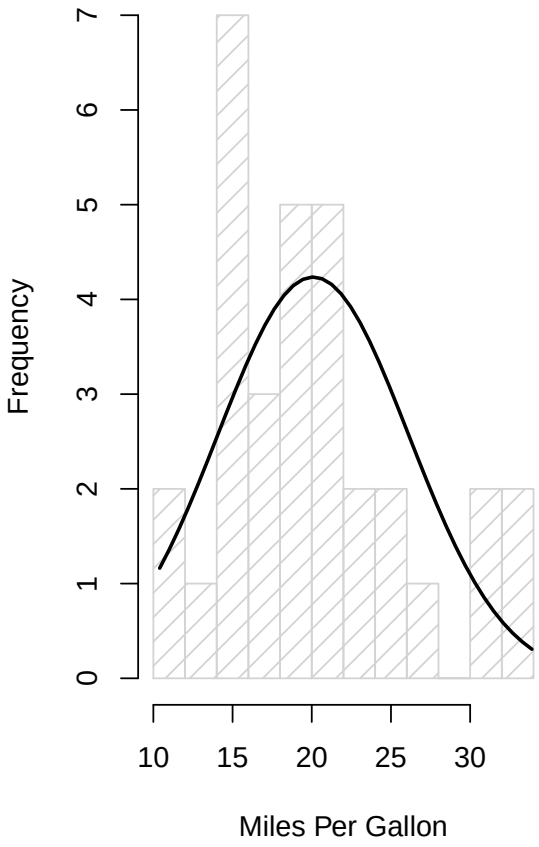
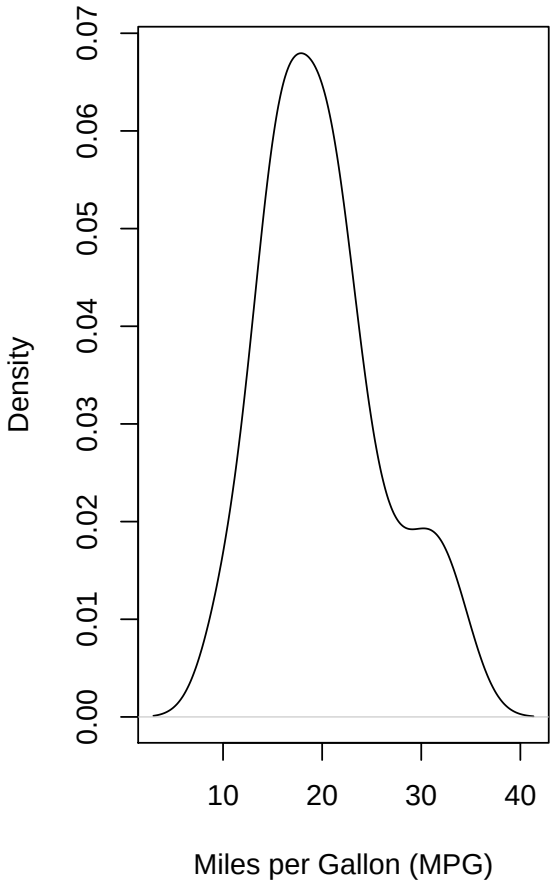
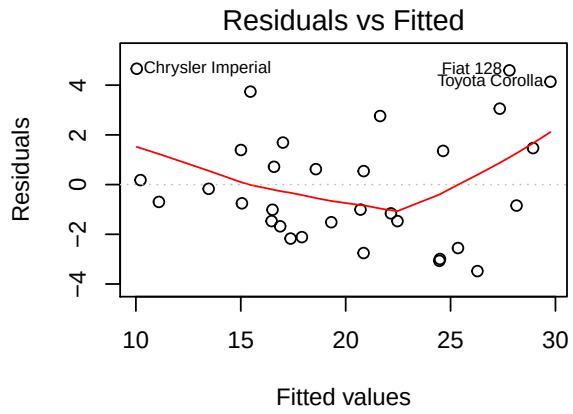


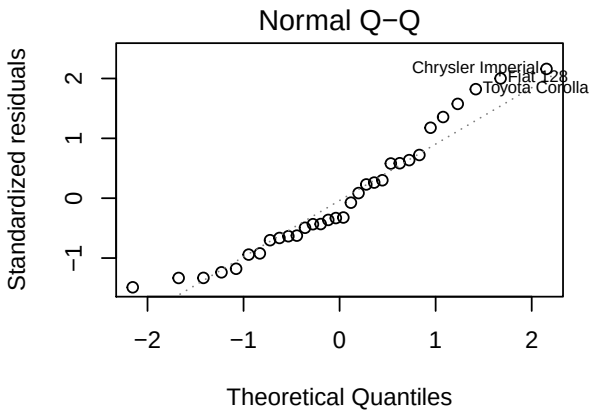

**Figure 2b. Density Plot of MPG**

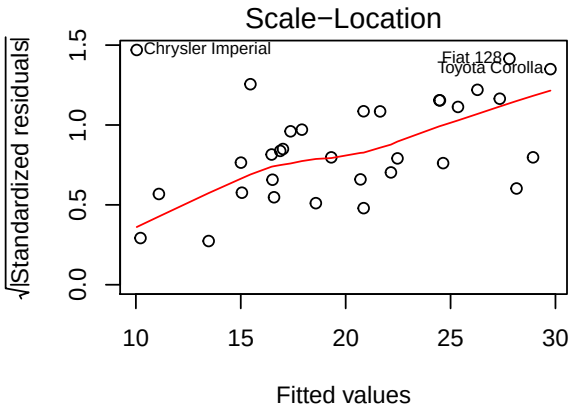


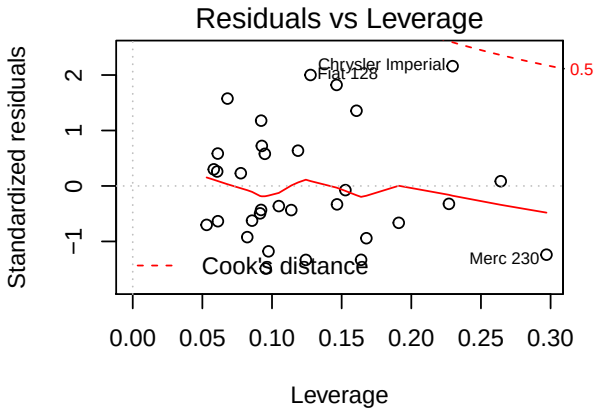

**Residual Plots of Best Fit Model**

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


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
Fitted values


Normal Q−Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
Theoretical Quantiles


Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Chrysler Imperial
Fiat 128
Merc 230
0.5
Cook's distance
Standardized residuals
Leverage