Regression Models Project

Executive Summary

In this particular report, I looked at a data set of collection of cars and was interested in exploring the relationships between a set of variables and miles per gallon (the outcome). The goal was to answer two particular questions: 1) "Is an automatic or manual transmission better for MPG?" 2) "Quantify the MPG difference between automatic and manual transmissions." These two questions were answered in which we saw that automatic transmission is worse for MPG and that the manual is better for mpg in the given boxplot. From the regression analysis, we were able to quantify these values in which automatic had a slope of 17.14737 and that manual had a slope of 24.39231. The difference between these two values is 7.24494. The interpretation of each slope is that for there is that for every 17.14737 miles gone, 1 gallon is used. Also, for every 24.39231 miles gone, one gallon is used. However, the R^2 value represents that only 35.98% of the variation in y (mpg) can be explained by x (transmission). Therefore, I created more a multivariable model and included this predictor since it was indeed significant at a .05 alpha significance level. From the multivariable model, the model was also significant and the R^2 value much larger where adjusted R^2 was 0.8344. 83.44% of the variation in mpg could be explained by the tranmsission, cylinder, horsepower and weight of the car which were significant at the .05 level. Also, the AIC and BIC values were lower than that of the single variable model which is a better model. The residuals were also approximately normal through plots and the Shapiro test. So overall, both questions were answered and the multivariable model was accurate.

Exploratory Data Analysis

```
#Get some information regarding the data set mtcars
?mtcars
#Load the mtcars data set
data(mtcars)
#See the structure of the data along with each variable
#and their particular class
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
#See the first 6 values
head(mtcars)
```

```
##                   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
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
#See the names of the variables for the data set
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
#Convert transmission, engine shape, carbeurator, cylinder
#and gear to factor variables
mtcars[["am"]] <- factor(mtcars[["am"]], levels = c(0,1), labels = c("Auto", "Manual"))
mtcars[["vs"]] <- factor(mtcars[["vs"]])
mtcars[["carb"]] <- factor(mtcars[["carb"]])
mtcars[["cyl"]] <- factor(mtcars[["cyl"]])
mtcars[["gear"]] <- factor(mtcars[["gear"]])
#Exploratory plots are in the appendix
```

Regression Analysis

```
#Subset the data of mtcars outcome mpg as a linear model
#or function of transmission type and calculate the mean
#for the subset
aggregate(mtcars[["mpg"]] ~ mtcars[["am"]], mtcars, mean)
```

```
##   mtcars[["am"]] mtcars[["mpg"]]
## 1           Auto        17.14737
## 2         Manual        24.39231
```

```
#Create two variables to set boolean equal to character string of Auto and
#manual
a <- mtcars[mtcars[["am"]] == "Auto",]
m <- mtcars[mtcars[["am"]] == "Manual",]
#Conduct t-test both subsetted data
t.test(a[["mpg"]], m[["mpg"]])
```

```
##
##  Welch Two Sample t-test
##
## data:  a[["mpg"]] and m[["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
```

```r
auto_manual_t_test <- t.test(a[["mpg"]], m[["mpg"]])
#Create confidence interval
auto_manual_t_test[["conf.int"]]
```

```
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

```r
#See the validity of a linear model with just mpg and transmission type
linear_model_1 <- lm(mtcars$mpg ~ mtcars[["am"]], mtcars)
#Calculate the summary and AIC and BIC values
summary(linear_model_1)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars[["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 ***
## mtcars[["am"]]Manual     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
```

```r
AIC(linear_model_1)
```

```
## [1] 196.4844
```

```r
BIC(linear_model_1)
```

```
## [1] 200.8816
```

```r
#Values indicate a positive outcome, add more predictors to the model from the
#pairs plot in the appendix
multivariable_model <- lm(mtcars$mpg ~ mtcars[["am"]] + mtcars[["cyl"]] +
mtcars[["disp"]] + mtcars[["hp"]] + mtcars[["wt"]], mtcars)
#See analysis of variance summary on both models
anova(linear_model_1, multivariable_model)
```

```
## Analysis of Variance Table
##
## Model 1: mtcars$mpg ~ mtcars[["am"]]
```

```
## Model 2: mtcars$mpg ~ mtcars[["am"]] + mtcars[["cyl"]] + mtcars[["disp"]] +
##     mtcars[["hp"]] + mtcars[["wt"]]
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#See just the summary for the multi-variable model and see the AIC and BIC values
summary(multivariable_model)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars[["am"]] + mtcars[["cyl"]] +
##     mtcars[["disp"]] + mtcars[["hp"]] + mtcars[["wt"]], data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          33.864276   2.695416  12.564 2.67e-12 ***
## mtcars[["am"]]Manual  1.806099   1.421079   1.271   0.2155
## mtcars[["cyl"]]6     -3.136067   1.469090  -2.135   0.0428 *
## mtcars[["cyl"]]8     -2.717781   2.898149  -0.938   0.3573
## mtcars[["disp"]]      0.004088   0.012767   0.320   0.7515
## mtcars[["hp"]]       -0.032480   0.013983  -2.323   0.0286 *
## mtcars[["wt"]]       -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

```
AIC(multivariable_model)
```

```
## [1] 156.3359
```

```
BIC(multivariable_model)
```

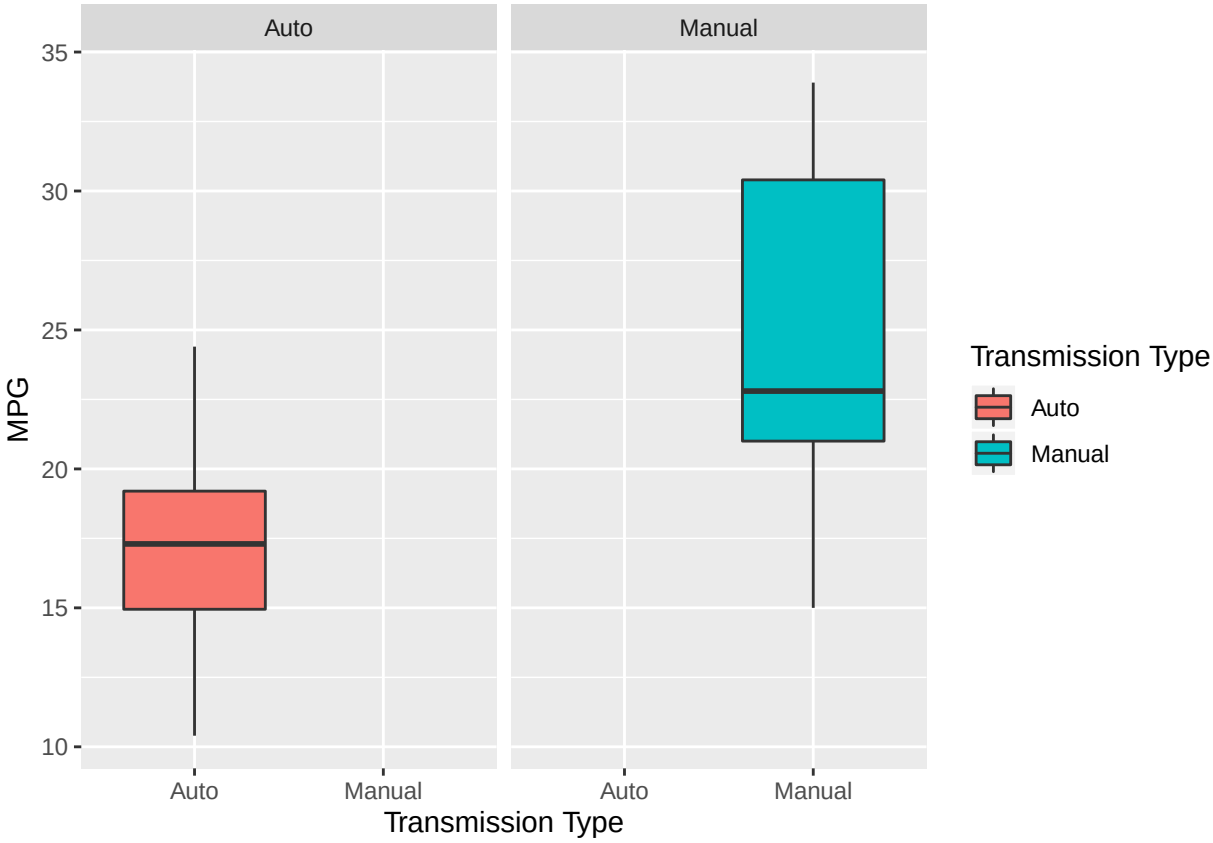
```
## [1] 168.0618
```

```
#Calculate the residuals of the multi-variable model
r <- resid(multivariable_model)
#Regression plots are in the appendix
```

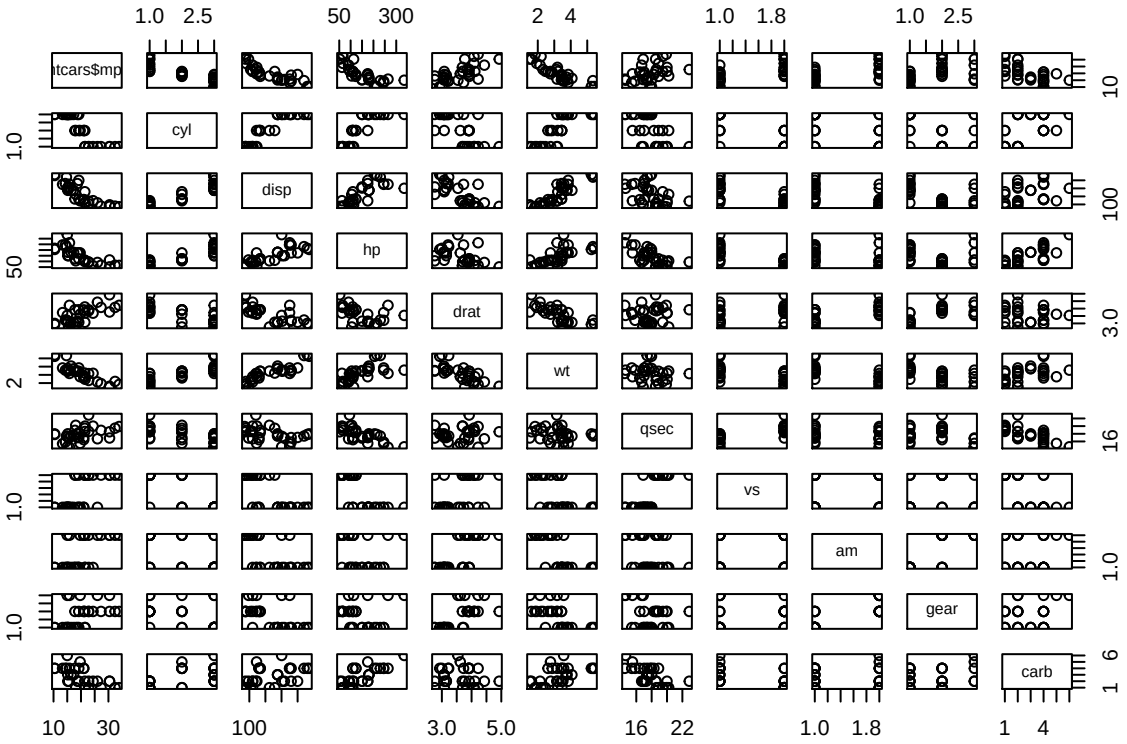
Appendix

```r
#Exploratory Analysis Plots

#Create boxplot to see quartiles and if there are any outliers
library(ggplot2)
ggplot(mtcars, aes(mtcars[["am"]], mtcars[["mpg"]], fill = mtcars[["am"]])) +
geom_boxplot() + facet_wrap(. ~ mtcars[["am"]]) + xlab("Transmission Type") +
ylab("MPG") + guides(fill = guide_legend(title = "Transmission Type"))
```




```r
#Create plot to see which variables seem to have a correlation
pairs(mtcars$mpg ~ ., mtcars)
```



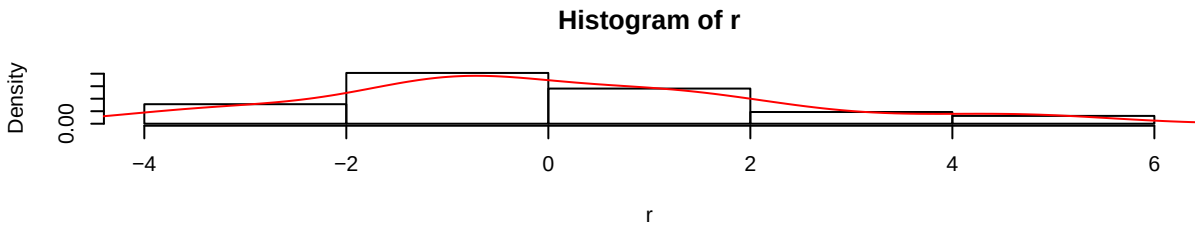

```
#Regression Analysis Plots
#Analyze the residuals and conduct tests for normality
par(mfrow = c(3,1))
hist(r, probability = TRUE)
lines(density(r), col = "red")
qqnorm(r)
qqline(r, col = "red")
var(r)
```

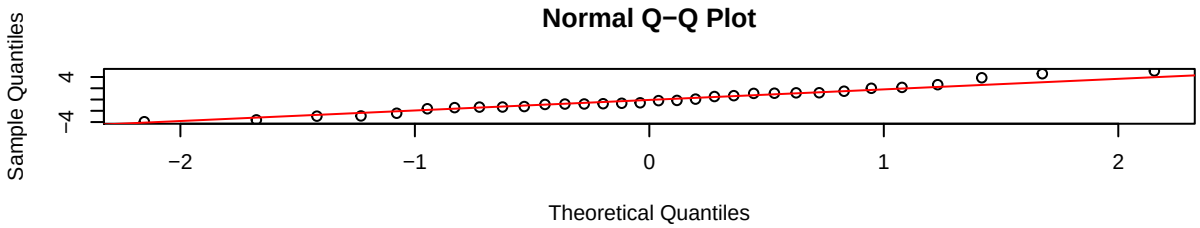
```
## [1] 4.851897
```

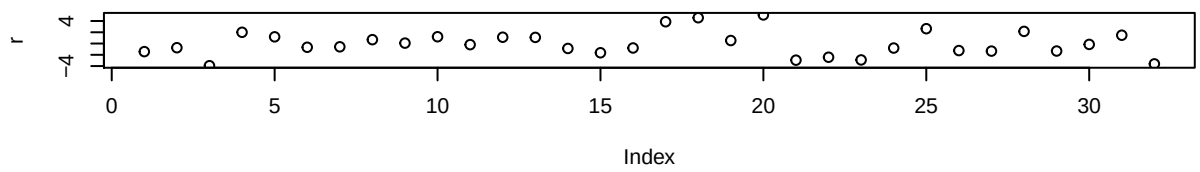
```
shapiro.test(r)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  r
## W = 0.971, p-value = 0.5274
```

```
plot(r)
```


Histogram of r
Density
0.00
−4
−2
0
2
4
6
r


Normal Q−Q Plot
Sample Quantiles
−4
4
−2
−1
0
1
2
Theoretical Quantiles


r
−4
4
0
5
10
15
20
25
30
Index