Regression Model Course Project

Aayush Shah

Executive Summary

This project was created as per the requirement of the coursera peer assignment as follows;

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

“Is an automatic or manual transmission better for MPG” “Quantify the MPG difference between automatic and manual transmissions”

This project involves exploring the *mtcars* dataset in R.

Data Exploration

```
library(datasets)
data("mtcars")
library(ggplot2)
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
```

We have loaded the dataset *mtcars*. But we need to convert *vs*, *am*, *gear* and *carb* columns as factors and add a new column **transmission** type depending on *automatic* and *manual* types.

```
mtcars$vs <- factor(mtcars$vs)
mtcars$transmission <- factor(mtcars$am,labels=c("Automatic","Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
head(mtcars)
```

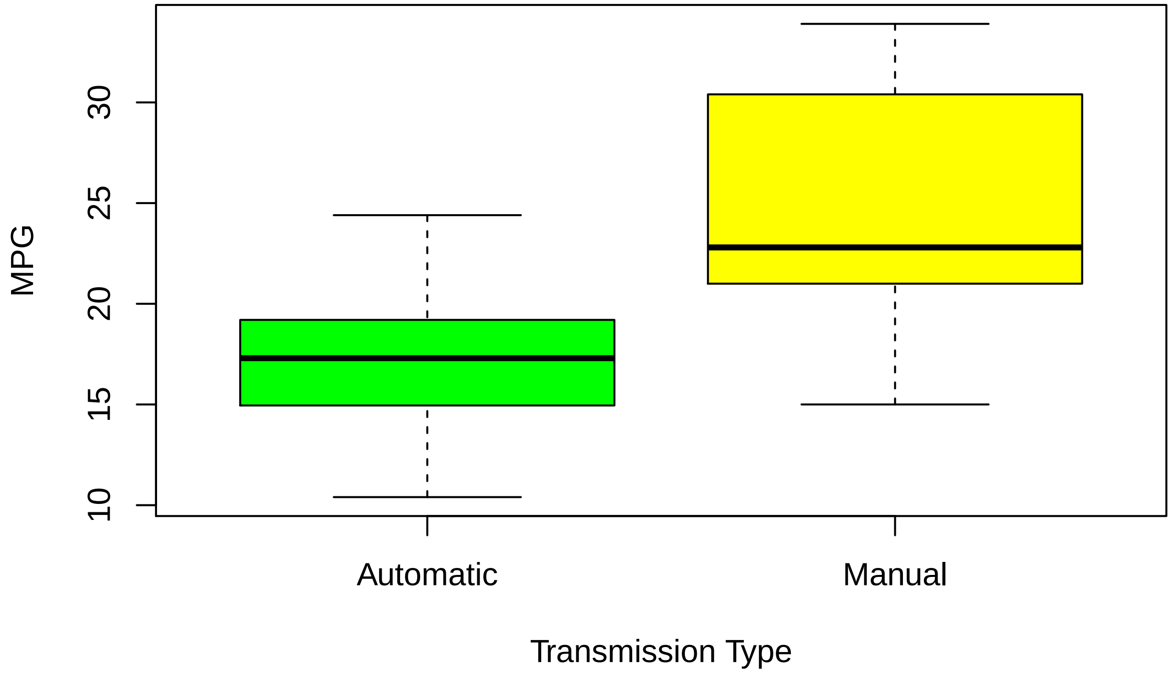
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 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
##                   transmission
## Mazda RX4               Manual
## Mazda RX4 Wag           Manual
## Datsun 710               Manual
## Hornet 4 Drive        Automatic
## Hornet Sportabout     Automatic
## Valiant               Automatic
```

```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

We now illustrate relationship between *mpg* and *transmission* variables.

```
boxplot(mpg ~ transmission, data = mtcars, col = (c("green","yellow")), ylab = "MPG", xlab = "Transmiss
```



As we see, *Manual Transmission* type gives a *better* MPG than *Automatic Transmission*.But lets explore it further.

Regression Analysis

Let *mpg* be the dependent variable and *transmission* be the independent variable. Lets fit a linear model now,

```r
fit<-lm(mpg~transmission,mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ transmission, 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 ***
## transmissionManual    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
```

As we see, The R-Squared value is 0.338 which means that only 33.8% of the regression variance can be explained by our model.Also We see that Manual transmission yields on average 7 MPG more than Automatic.

Lets explore relationship of other variables on mpg using analysis of variance.

```r
Varianceanalysis<-aov(mpg ~ ., data = mtcars)
summary(Varianceanalysis)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## cyl          1  817.7   817.7 102.591 2.3e-08 ***
## disp         1   37.6    37.6   4.717 0.04525 *
## hp           1    9.4     9.4   1.176 0.29430
## drat         1   16.5    16.5   2.066 0.16988
## wt           1   77.5    77.5   9.720 0.00663 **
## qsec         1    3.9     3.9   0.495 0.49161
## vs           1    0.1     0.1   0.016 0.90006
## am           1   14.5    14.5   1.816 0.19657
## gear         2    2.3     1.2   0.145 0.86578
## carb         5   19.0     3.8   0.477 0.78789
## Residuals   16  127.5     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the p values which are less than 0.5, we see that *cyl*, *disp*, and *wt* variables must be considered along with *transmission* type to explain relationship with mpg.

```r
fit2<-lm(mpg~cyl + disp + wt + transmission, data = mtcars)
summary(fit2)
```

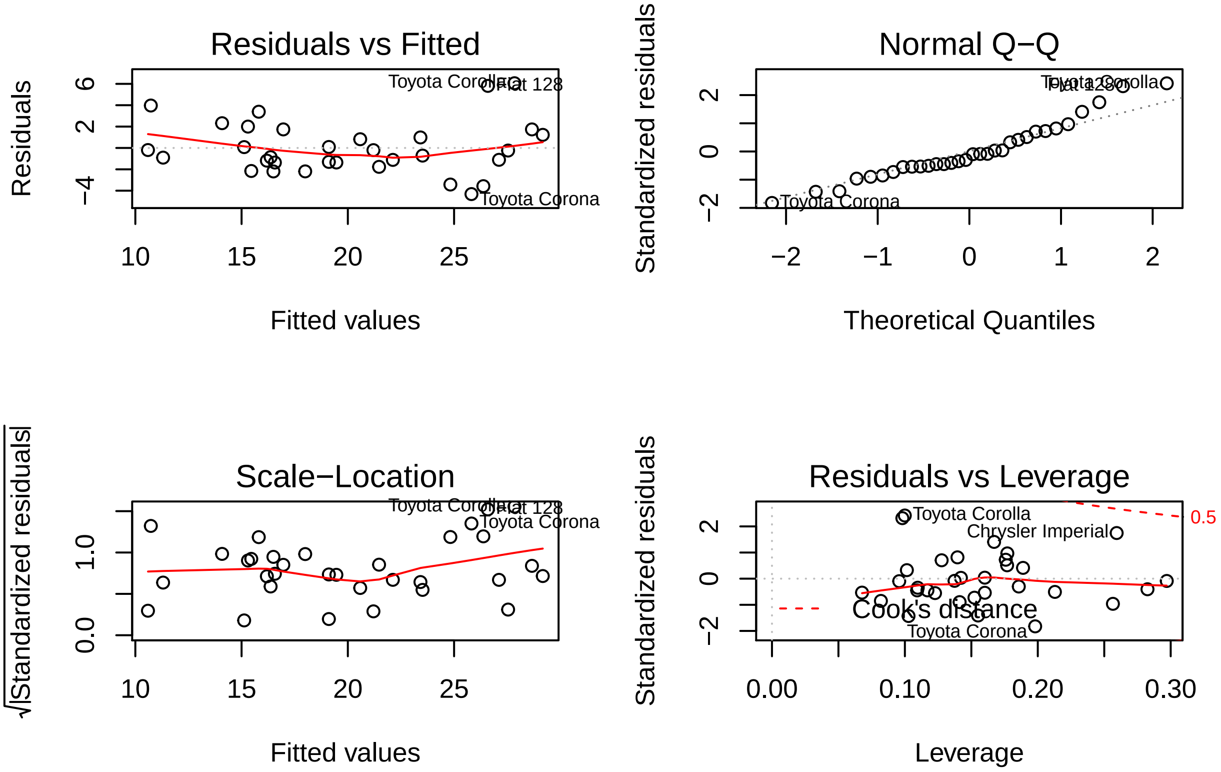
```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + transmission, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        40.898313   3.601540  11.356 8.68e-12 ***
## cyl                -1.784173   0.618192  -2.886  0.00758 **
## disp                0.007404   0.012081   0.613  0.54509
## wt                 -3.583425   1.186504  -3.020  0.00547 **
## transmissionManual  0.129066   1.321512   0.098  0.92292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

We see that now about 80% or more variance can be explained by considering variables *cyl*, *disp*, and *wt* along with *transmission*.

The P-values of *wt* and *cyl* are less than 0.5 which tells us that these are important variables in explaining relation between *transmission* type and *mpg*.

Now lets do a residual plot of this multivariable model(*fit2*)

```r
par(mfrow = c(2,2))
plot(fit2)
```



So, it seems that the fit of the multivariable model *fit2* and its residuals seem to satisfy basic requirement for a linear model to explain the variation of the variable *mpg*.

Conclusion

Is an automatic or manual transmission better for MPG? Manual transmission cars appear to be better for mpg compared to Automatic cars. But with a multivariable model with confounding variables *cyl*, *disp*, and *wt* the difference is less significant.

Quantify the MPG difference between automatic and manual transmissions? Using only *transmission* variable, manual cars yield on average 7 MPG more than automatic cars. But when variables *cyl*, *disp*, and *wt* are included the average goes down to a lesser value.