Motor Trend Data Analysis Project

Carlos Martinez Reyes

30/10/2020

Executive Summary

In this report, we analyze the 1974 US Motor Trend magazine mtcars data set to evaluate the effect of transmission type on **MPG** (*miles per gallon*) performance. The database includes the fuel consumption and 10 design and performance aspects of 32 cars (1973–74 models). We use **MPG** as the response variable and fit a regression model considering a set of variables as predictors.

Exploratory Data Analysis

```
library(ggplot2)
data(mtcars)
dim(mtcars)
```

```
## [1] 32 11
```

The data consists of 32 samples (different automobiles) and 11 variables (10 control variables, 1 target variable (mpg). The data are numeric which is not correct for our finishers, so the next step is the transformation of variables.

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

Lets check the result

```
aggregate(mpg ~ am, data = mtcars, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

The automatic type car travels less miles per gallon compared to the manual transmission and this is also confirmed visually by referring to the **Figure 1** in the appendix. Confirm this analytically by proving that the difference between the **MPG** averages is statistically significant, we use the two-sample T-test to prove it. ON AVERAGE the automatic type travels less miles per gallon compared to the manual transmission and we also confirm this analytically by proving that the difference between the MPG averages is statistically significant (Null hypothesis: the difference is not significant). We use the two-sample T-test to prove it.

```
Auto <- mtcars[mtcars$am == "Automatic",]$mpg
NonAuto <- mtcars[mtcars$am == "Manual",]$mpg
t.test(Auto, NonAuto)
```

```
##
##  Welch Two Sample t-test
##
## data:  Auto and NonAuto
## 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
```

Quantifying the difference: since the p-value is 0.001374, we reject the null hypothesis at 5% and 1% significance level and the mean MPG of cars with manual transmission is different (7.245 times more) from the average performance of cars with automatic transmission.

Regression Analysis

We have already tested that there is a significant difference at 5% and 1% in the average **MPG** performance and we roughly quantify this difference. Now let's see what kind of relationship it has with the rest of the factors. We will fit two linear regression models to the data, one simple and the other multiple to see if there is any change in **MPG** based on the transmission and how it is affected considering other variables.

Model 1: MPG and Transmission

```
fitam <- lm(mpg ~ am, mtcars)
summary(fitam)
```

```
##
## 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
```

The R^2 value for this model is only 0.3598, which means that adjusting mpg only with am explains about 36% of the variation in mpg because of its linear relationship. From the model we get an adjusted R-squared of 33.85% this quite a low variance explained by the model. Due to little variance explained by the model let examine other variable that are might be relevent to explain more variance to build a multivariate linear regression.

Model 2: BESTFIT

From **Figure 2** of the appendix there are predictor variables correlated with the **am** factor, which is why the exclusion of the variables that are correlated with the type of transmission. Including unnecessary regressors will inflate the variance of the model.

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

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9511 -1.4244 -0.1767  1.3666  4.2187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.57617   11.27271   1.914   0.0671 .
## cyl6        -1.90950    1.72992  -1.104   0.2802
## cyl8        -0.22716    2.87047  -0.079   0.9376
## hp          -0.02481    0.01515  -1.637   0.1141
## wt          -2.96274    0.97728  -3.032   0.0056 **
## qsec         0.61917    0.55987   1.106   0.2793
## amManual     2.83270    1.67020   1.696   0.1023
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.4 on 25 degrees of freedom
## Multiple R-squared:  0.8721, Adjusted R-squared:  0.8414
## F-statistic: 28.42 on 6 and 25 DF,  p-value: 5.196e-10
```

This works as expected the model has improved significantly to attain a R^2 of 84.14% and reduced our residual standard error to 2.4 from 2.8.

Residuals Normality Test

```
Sbest=shapiro.test(bestfit$resid)
print(Sbest)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bestfit$resid
## W = 0.96809, p-value = 0.4484
```

The p-value > 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality. Hence, our evaluation of the residuals is valid and the model is a good fit for the data. This is supported by the graphs in **Figure 3**.

Appendix

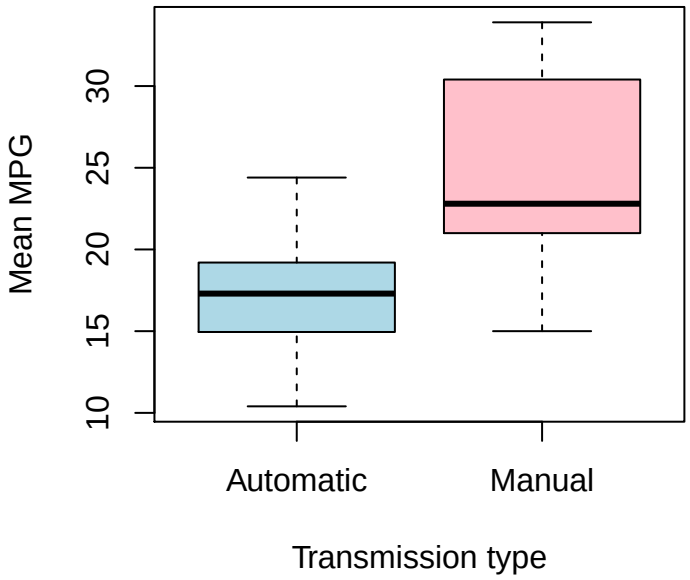

Figure 1: The graph corroborates the t-test performed earlie, the manual transmission provides a better MPG overall.

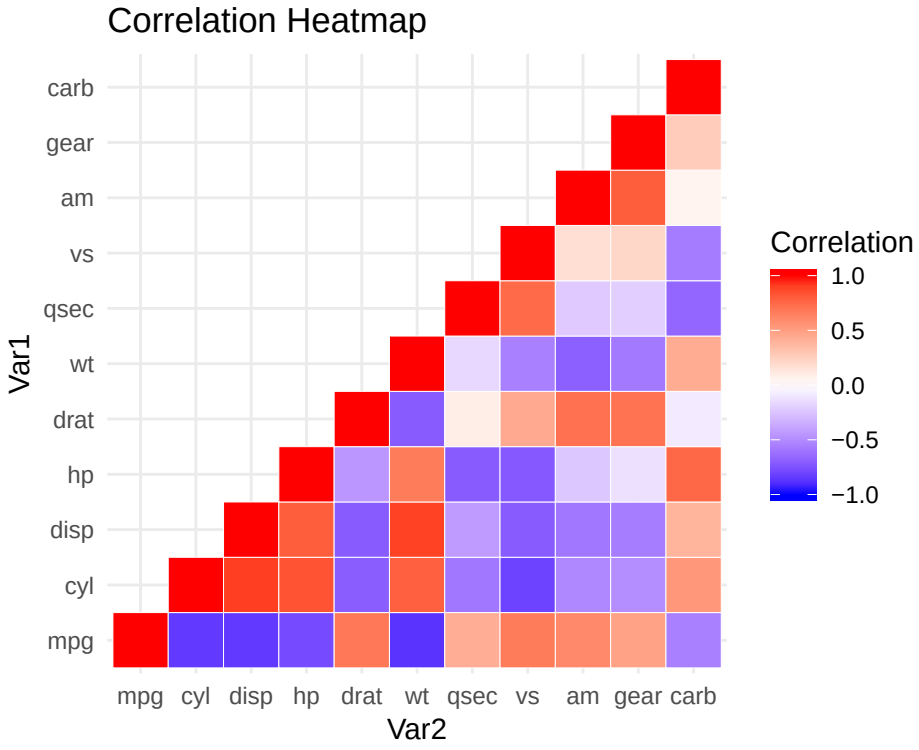

Figure 2: Correlation Matrix

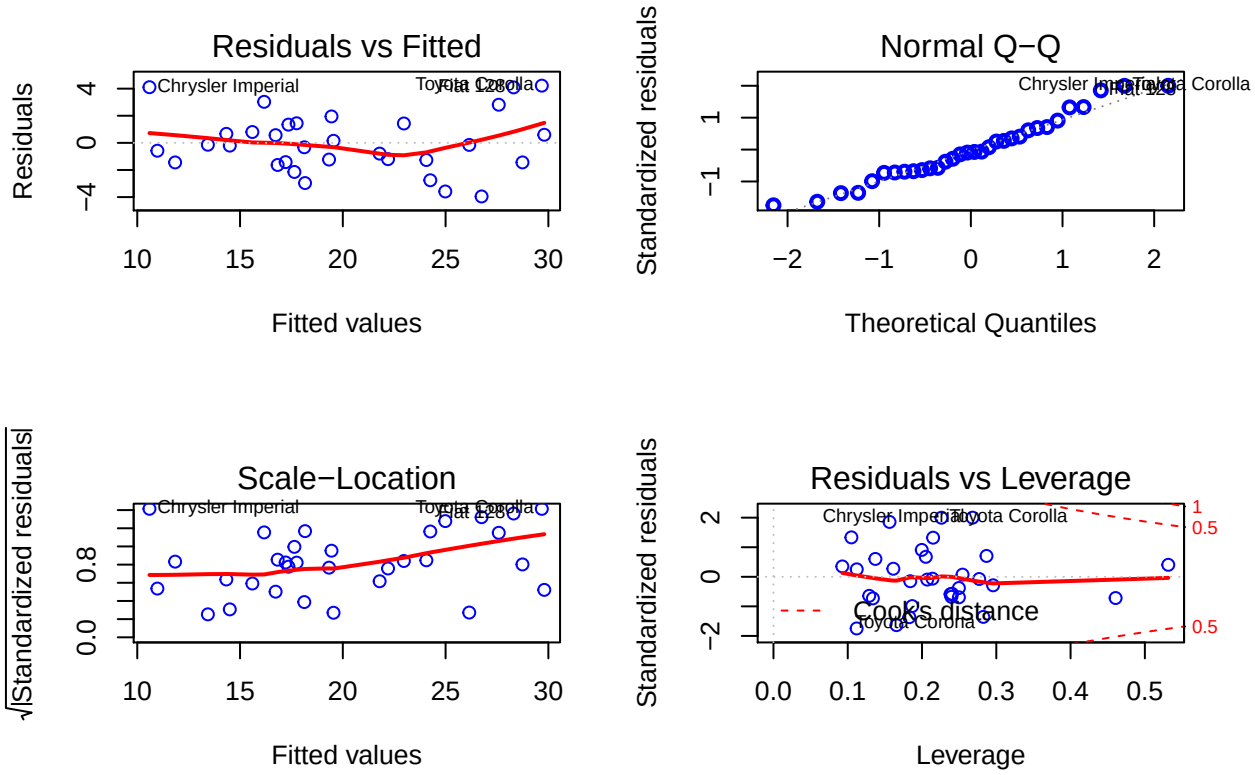

Figure 3: The Residual Fit Plot looks how we would expect it to look if residuals were independently and almost identically distributed with zero mean, and were uncorrelated with the fit. The highest residuals were for the outliers. The QQ Plot shows how the outliers, the Chrysler Impala, Lotus Europa and Fiat 128 affect the curve.

Conclusions

Collectively for all control variables considered together, there is significant effect, while for each control variable including transmission design effect is insignificant. Individually, transmission design shows significant difference on MPG.