Regression Models-Project

Oscar Portillo

EXECUTIVE SUMMARY

It is a common belief that manual transmission results in better fuel efficiency than automatic transmission. In this work, the dataset "mtcars" of road tests performed by the *Motor Trend US* magazine is used to answer the following two questions: - Is an automatic or manual transmission better for miles per gallon (MPG)? - How different is the MPG between automatic and manual transmissions? The data was extracted from the March, April, June and July issues of the maganizine in 1974. The road test was conducted in an attempt to predict gasaoline mileage for 1973-1974 automobiles.

EXPLORATORY DATA ANALYSIS

The dataset consists of 11 variables of vehicle performance and design: 1) **mpg** Miles/(US) gallon, 2) **cyl** Number of cylinders, 3) **disp** Displacement (cu.in.), 4) **hp** Gross horsepower, 5) **drat** Rear axle ratio, 6) **wt** Weight (lb/1000), 7) **qsec** quarter mile time, 8) **vs** V/S- Engine Shape (0= V, 1= Straight), 9) **am** Transmission (0 = Automatic, 1 = Manual), 10) **gear** Number of forward gears, and 11) **carb** Number of carburetors barrels. See Figure 1 (Appendix).Technological upgrades have continue to improve vehicle's performance. For example, it is interesting to notice that the average mpg is 20.9 of the selected new vehicles in 1973-1974. Nowadays,based on the article published by the University of Michigan Transportation Research Institute via Automotive News http://www.autoblog.com/2013/09/12/average-new-car-fuel-economy-record-24-9-mpg/, the average fuel economy of new trucks, cars, SUVs and vans sold in August 2013 reached 24.9 mpg.It is observed that the average horsepower of the sample is 146.7 and the minimum was 52 hp. Today, average horsepower have climbed to the 220 hp levels (Ref. http://www.autoblog.com/2011/02/11/every-car-is-now-a-performance-car/). In the *mtcars* dataset, there are 32 observations that correspond to 32 different 1973-74 car models. The first three vehicles and their mpg and physical characteristics are listed below:

```
library(datasets); data(mtcars); head(mtcars,3)
```

```
##                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
```

An exploratory cluster plot is shown in Figure 2.In this paper, we will analize the relationship bewteen mpg and the transmission type variable, column labelled *am*. There are 19 vehicles with automatic transmission and 13 with manual transmission, see Figure 3. The average of the gasaline milage for the automatic transmission automobiles is 17.147 mpg and 24.39231 mpg for the manual transmission models. Figure 4 provides the fuel economy (in miles per galon gasoline) for the vehicles in each transmission class. It is observed that the cars with a manual transmission have 7.245 more MPGs on average then the automatic ones.

STATISTICAL INFERENCE AND LINEAR REGRESION DATA ANALYSIS

From the graph above, it appears that cars with automatic transmission have a lower mpg than manual cars. However, it is possible that this outcome happened by random chance (for instance, it could be the case that the study just happened to choose a set of automatic cars with low mpg and a group of manual cars with higher mpg). Therefore, to validate if that is the case, statistical test is used to compare two

samples and see if they have different means. One can obtain an interval estimate of the difference between two population means using the unpaired Student's T-test. The assumption is that the data populations follow the normal distribution.

```
Automatic = mtcars$am == 0; mpg.automatic = mtcars[Automatic,]$mpg; mpg.manual =
mtcars[!Automatic,]$mpg;
t.test(mpg.automatic, mpg.manual)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg.automatic and mpg.manual
## 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
```

As the p-value turns out to be 0.001374, and is less than the .05 significance level, we reject the null hypothesis. So at the .05 significance level, we conclude that there is an actual difference between the groups.The 95% confidence interval estimate of the difference between the mean gas mileage of automatic and manual transmissions is between 3.2097 and 11.2802 mpg. We begin our regression analysis by performing a simple linear regression model of gas milage (mpg) as outcome and transmission type (am) as predictor, see Figure 5. Looking at the intercepts and coefficients in the summary below, we can see that, on average, automatic cars achive 17.147 mpg and manual transmission cars do 7.245 mpg more (as mentioned earlier). Additionally, R^2 value is 0.3598, meaning that the model only explains 35.98% of the variance. The residual standard error is 4.902.

Common sense suggests that other fators such as weight and hp might have greatest impact on fuel consumption. Thus, a subset of predictor variables is selected using backward regression via the *step()* function with *"direction="* set to *"backward"* (other options are "forward" and "both"). The step() function basically performs a model search using the Akaike Information Criteria to find the best model. A summary of the search algorithm (intecept & coefficients in the last step) is shown below.

```
MultiLinearModel<- lm (mpg~., mtcars)
BackwardSelectionModel <- step(MultiLinearModel, direction="backward")

summary(BackwardSelectionModel)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

According to this procedure, the best model is the one that includes the variables weight, wt, quarter mile time, qsec, and transmission type, am. The coefficient of am has a p-value of 0.0467 which is less than 5%. Thus, the null hypothesis is rejected and we can claim that there is significant evidence that transmission type has an effect on gas milage mpg. The coefficient of am can be interpreted as follows, if qsec and wt remain unchanged, changing from automatic transmission to manual transmission would increase the gasoline mileage by 2.93 miles/US gallon on average. We can also notice that weight, wt, has a strong impact in fuel effeciency: keeping qsec and am constant, the expected change in mpg per 1000 lb increase in weight is -3.917 miles/US gallon, a reduction in fuel efficiency as expected. Residuals are normally distributed and homoskedastic, see Figure 6.

CONCLUSIONS

A T-test was performed to compare the difference between the mpg means of automatic and manual transmission, it was concluded that there is an actual difference between the groups mean at the .05 significance level. Based on the linear regression models, we conclude that manual transmission is more fuel efficient than automatic transmission as meassured by gas mileage mpg.The multivariable model (variables: wt, qsec, am) predicts and improvment of 2.93 MPGs for the muanual transmission cars whereas the uni-variable model predicts 7.245 more MPGs on average.

APPENDIX

Figure 1. Summary of gas milage data.

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Figure 2. Hierarchical Cluster Analysis. Dendrogram displaying a hierarchical relationship among the vehicles. It is observed that Mercedez 450SLC, 450SE and 450SL are classified as close relatives as expected.

```
d <- dist(as.matrix(mtcars)); hc <- hclust(d); plot(hc)
```

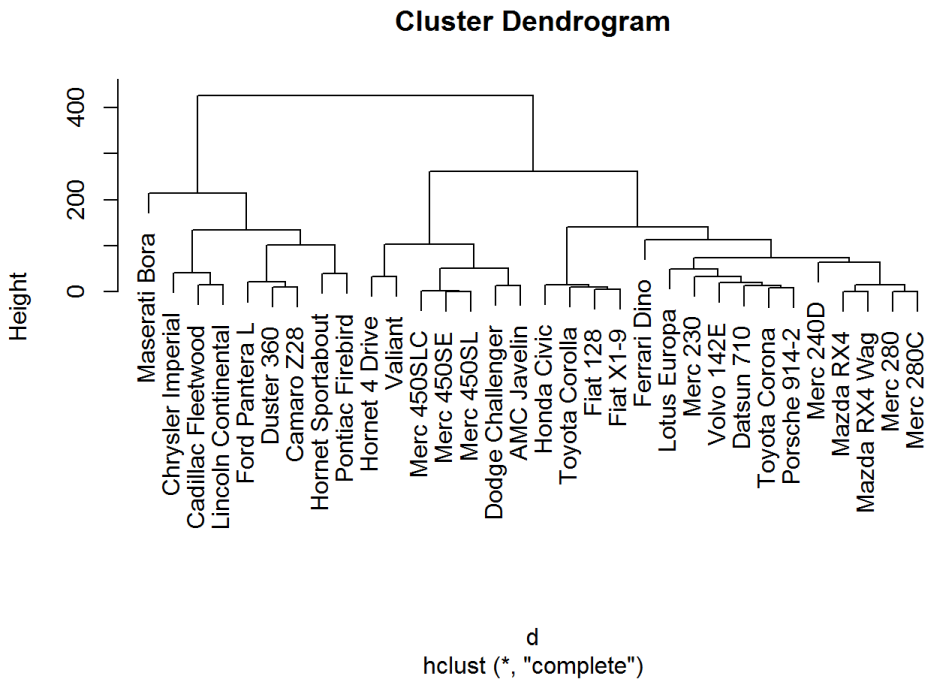


Figure 3. Frequency if automatic/manual cars in the mtcars dataset.

```
as.data.frame(table(mtcars$am)); mean(mpg.automatic); mean(mpg.manual)
```

```
##   Var1 Freq
## 1    0   19
## 2    1   13
```

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

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

Figure 4. Plot gas milage vs transmission-type.

```
library(ggplot2);ggplot(mtcars, aes(x=factor(mtcars$am), y=mpg)) + geom_boxplot()  +
labs(x = "Transmission Type: 0-Automatic, 1-Manual", y = "Gasoline Milage:
mpg[miles/(US) gallon]")+ ggtitle("Gas Milage vs Automatic/Manual Transmission")
```

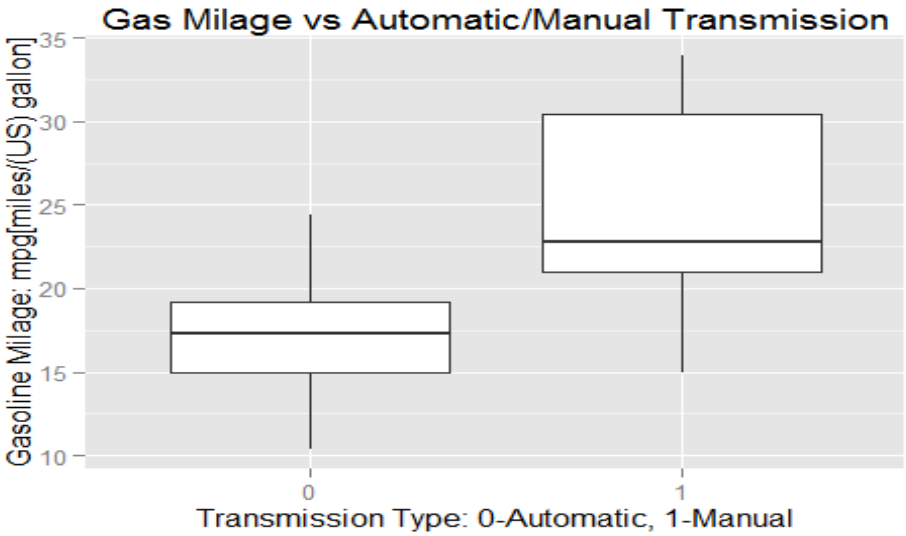
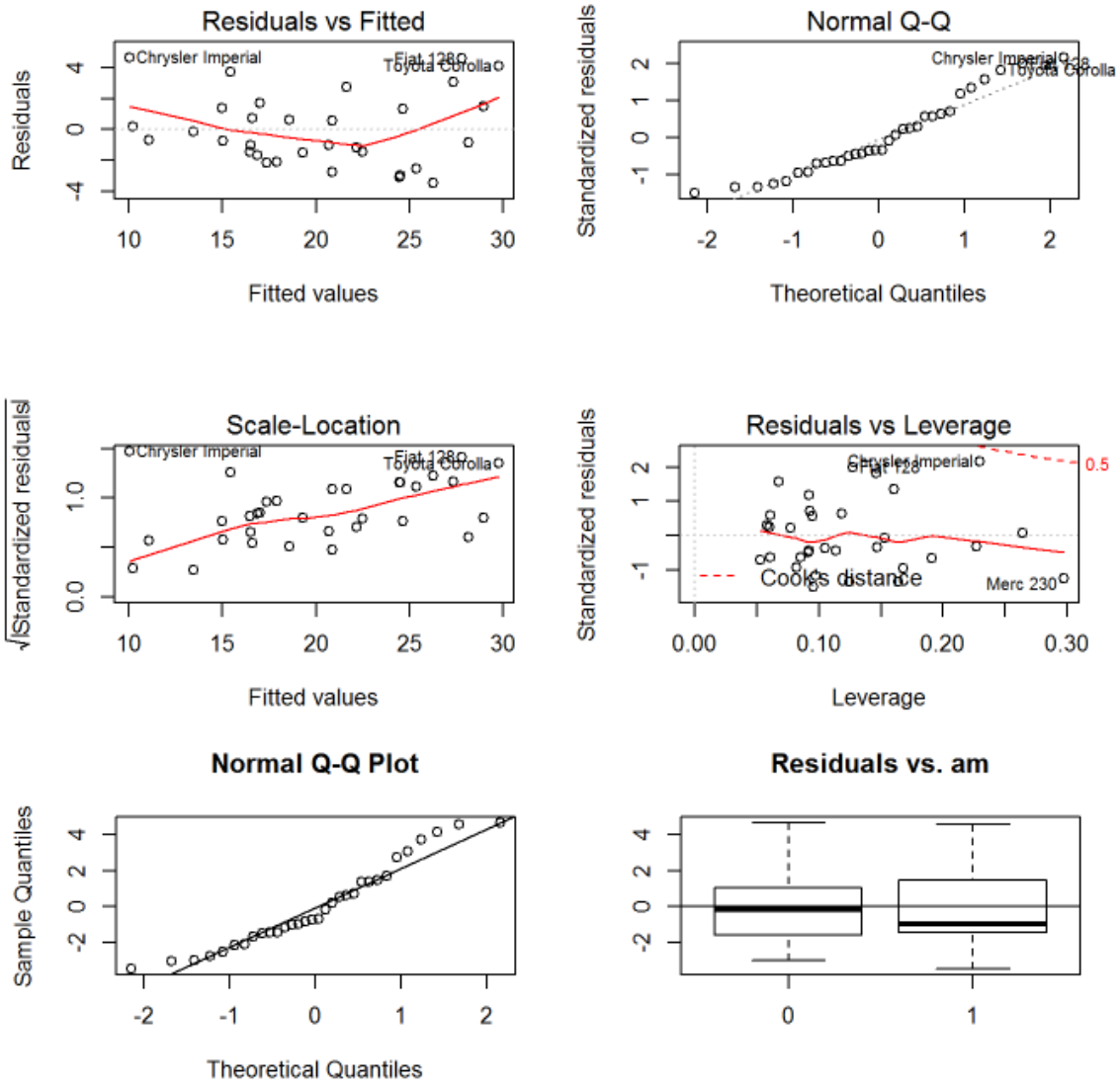


Figure 5. Linear Regression model.

```
mpg_am<-lm(mpg ~ factor(am), mtcars); summary(mpg_am);

##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

Figure 6. Normality and Residuals plots for the multivariable model.



Note: This report was created in Rmarkdown and knitWord, and converted to pdf from MS Word.