Regression Models Project - Motor Trend Data Analysis

Shayan (Sean) Taheri

June 9, 2019

Executive Summary

The main goal of this project is analysis of the “mtcars” dataset. Next, the relationship between a set of variables and miles per gallon (MPG) is explored in detail. We extracted the data from the year of 1974 (“Motor Trend U.S. Magazine”). It comprises fuel consumption and ten aspects of automobile design and performance measurement for 32 automobiles (1973â“74 models). The regresion models are used along with the exploratory data analyses. This experiment is help in exploring how **automatic** (am = 0) and **manual** (am = 1) transmissions features affect the **MPG** feature. The t-test demonstrates that the performance differenc between cars with automatic and manual transmission. It is about 7 MPG

This analysis let us know that about seven mils per gallon (MPG) more for cars with manual transmission than those with automatic transmission. Next, we fit several linear regression models and select the highest Adjusted R-Squared value. Therefore, the weight given and the portion of ¼ mile time are kept constant. Manual transmitted cars are defined as 14.079 + (-4.141)*Weight. More MPG on average is better than automatic transmitted cars. Accordingly, the cars that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will include higher MPG amounts.

Exploratory Data Analysis

Let’s load the dataset under analysis, called “mtcars”. We change some variables from the “numeric” class to the “factor” class.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
data(mtcars)
mtcars[1:3, ] # Sample Data
```

	mpg <dbl>	cyl <dbl>	disp <dbl>	hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>	vs <dbl>	am <dbl>
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1

3 rows | 1-10 of 12 columns

```
dim(mtcars)
```

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

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

Afterwards, we perform preliminary exploratory data analysis. Take a look at the **Appendix: Figures** section for the plots.According to the box plot, it is seen that the manual transmission yields higher values of MPG in general. From the pair of graph, it is seen zthat higher amount of correlations among the variables such as “wt”, “disp”, “cyl” and “hp” exist.

Inference

In the step of inference, we make the null hypothesis of having the MPG of the automatic and manual transmission stems from the same population (with the assumption of the MPG that has a normal distribution). For the purpose of graphical analysis, we use the two sample T-test.

```
result <- t.test(mpg ~ am)
result$p.value
```

```
## [1] 0.001373638
```

```
result$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Due to having the p-value equalling to 0.00137, the null hypothesis is rejected. Therefore, the automatic and manual transmissions are extracted from different populations. After completion of this step, the mean for MPG of manual transmitted cars is about seven more than that of automatic transmitted cars.

Regression Analysis

For regression analysis step, a fit to the full model is skethed according to the following:

```
fullModel <- lm(mpg ~ ., data=mtcars)
summary(fullModel) # results hidden
```

This model has the Residual Standard Error equalling to 2.833 on 15 degrees of freedom. Next, we calculate the adjusted R-squared value equal to 0.779. This means the model can explain about 78% of the variance of the MPG variable. This is not true always since none of coefficients are significant at 0.05 level of significance. Next, we use backward selection for selection of some statistically significant variables.

```
stepModel <- step(fullModel, k=log(nrow(mtcars)))
summary(stepModel) # results hidden
```

This model is written as “mpg ~ wt + qsec + am”. From this model, the residual standard error of 2.459 on 28 degrees of freedom is achieved. Besides that, the Adjusted R-squared value is 0.8336. This means that the model can explain about 83% of the variance of the MPG variable. All of the coefficients are significant at 0.05 level of significance. For more information, we can refer to the **Appendix: Figures** section for the plots again. According to the scatter plot, it is indicated that an intersection term between “wt” variable and “am” variable exists. Due to the higher level of weight in automatic cars than their manual counterparts. Accordingly, the following model is developed with the inclusion of the interaction term:

```
amIntWtModel<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(amIntWtModel) # results hidden
```

This model has the Residual Standrad Error equalling to 2.084 on 27 degrees of freedom. The adjusted R-squared value is equal to 0.8804 that provides us with an interpretation of the model that explains 88% of the variance of the MPG variable. All of the coefficients are significant at the 0.05 level of significance. Completion of this step leads a simple model with MPG as the outcome variable and Transmission as the predictor variable.

```
amModel<-lm(mpg ~ am, data=mtcars)
summary(amModel) # results hidden
```

On average, a car has 17.147 MPG using automatic transmission and 7.245 MPG with manual transmission. The model has the Residual Standard Error as 4.902 on 30 degrees of freedom. Besides that, the Adjusted R-squared value is equal to 0.3385. This means that the model is capable of explaining 34% of the variance of the MPG variable. The low Adjusted R-squared value shows the need to add other variables to the model. After this step, the model is complete and ready to be chosen.

```
anova(amModel, stepModel, fullModel, amIntWtModel)
confint(amIntWtModel) # results hidden
```

The model with the highest Adjusted R-squared value is selected that is presented as: “mpg ~ wt + qsec + am + wt:am”.

```
summary(amIntWtModel)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

That is, a manual transmitted car that weighs 2000 lbs have 5.797 more MPG than an automatic transmitted car that has both the same weight and 1/4 mile time. Based on this analysis, the result shows the importance of “wt” (weight lb/1000) and “qsec” (1/4 mile time) when they stay constant. Cars with the manual transmission has an additional level of 14.079 + (-4.141)*wt more MPG (miles per gallon) on average than cars with automatic transmission. A manual transmitted car with the weight of 2000 lbs has 5.797 more MPG than an automatic transmitted car. This implies that both have the same weight and ¼ mile time.

Residual Analysis and Diagnostics

The plots are shown in the **Appendix: Figures** section. The residual plots help us to verify the following underlying assumptions: A. The residual versus Fitted plot shows no consistent pattern, supporting the accuracy of the independent assumption. B. The Normal Q-Q plot shows us the normal distribution of the residuals due to the positioning of the points close to the line. C. The Scale-Location plot confirms that the consistent variance assumption as the point sare randomly distributed. D. The Residuals versus Leverage argues that no outliers are present. All the values fall well within the 0.5 bands. E. As for the Dfbetas, the measure of how much an observation has been affected by the estimate of a regression coefficient.

```
sum((abs(dfbetas(amIntWtModel)))>1)
```

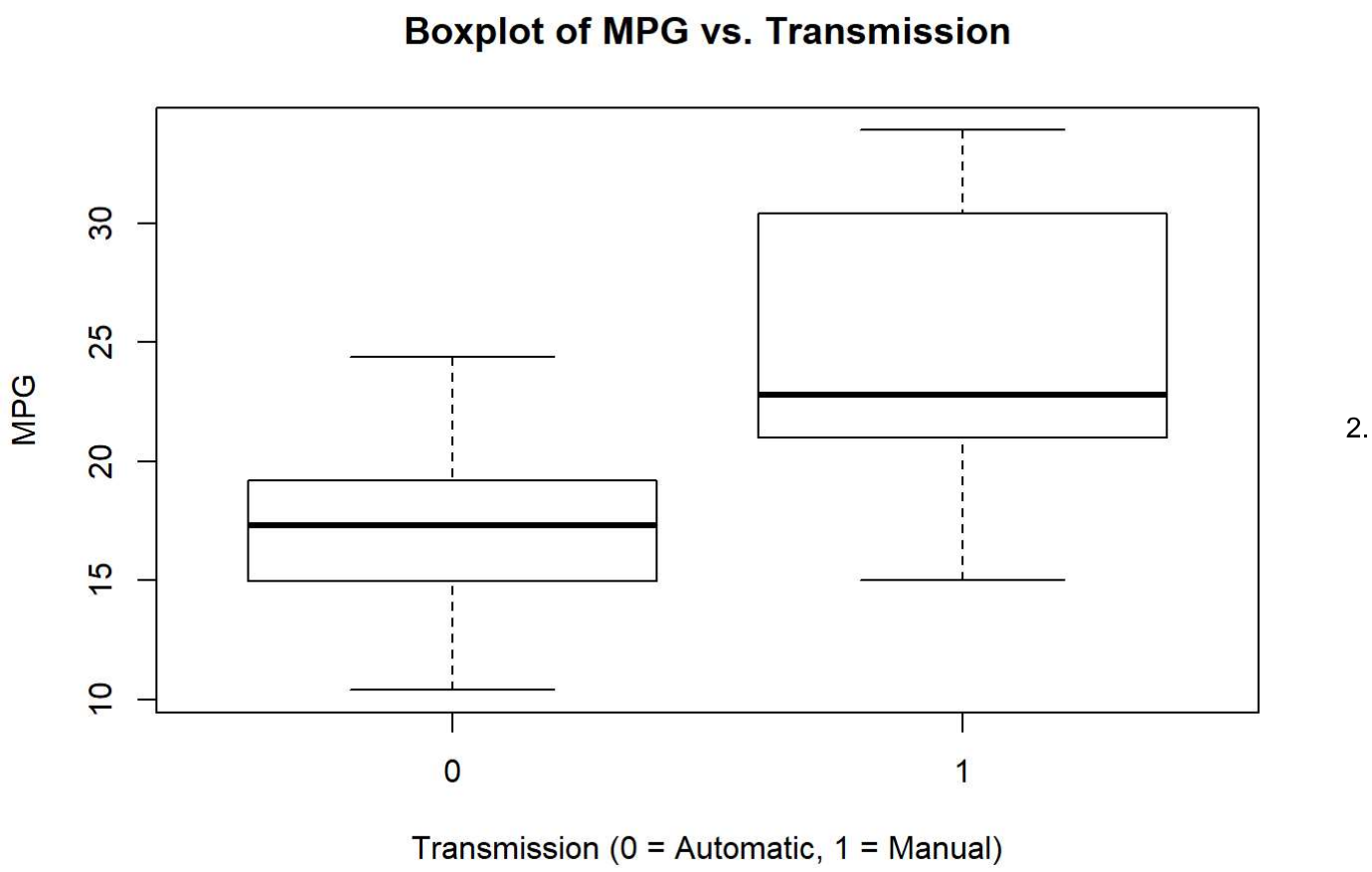
```
## [1] 0
```

With these assumptions, the analyses given above will meet all the basic assumptions of linear regression.

Appendix: Figures

1. Boxplot of MPG Vs. Transmission:

```
boxplot(mpg ~ am, xlab="Transmission (0 = Automatic, 1 = Manual)", ylab="MPG",
        main="Boxplot of MPG vs. Transmission")
```

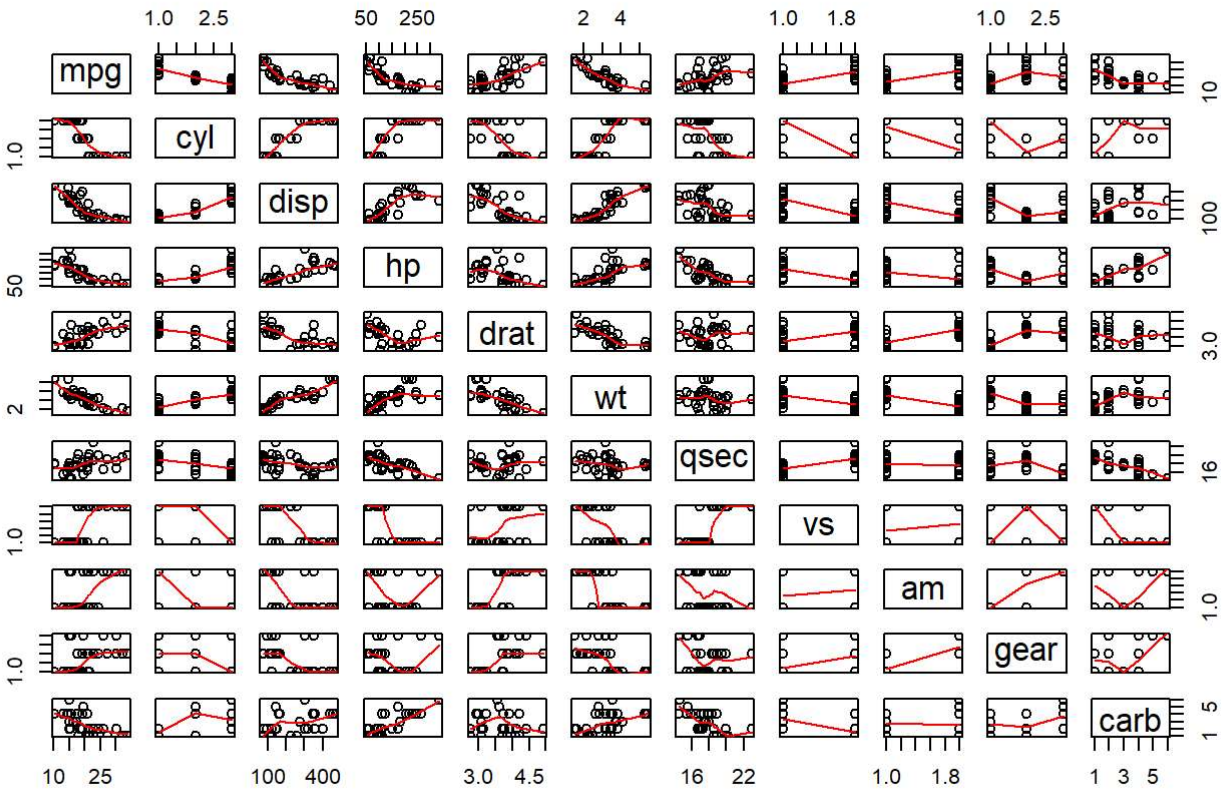


2.

Pair Graph of Motor Trend Car Road Tests:

```
pairs(mtcars, panel=panel.smooth, main="Pair Graph of Motor Trend Car Road Tests")
```

Pair Graph of Motor Trend Car Road Tests



3.

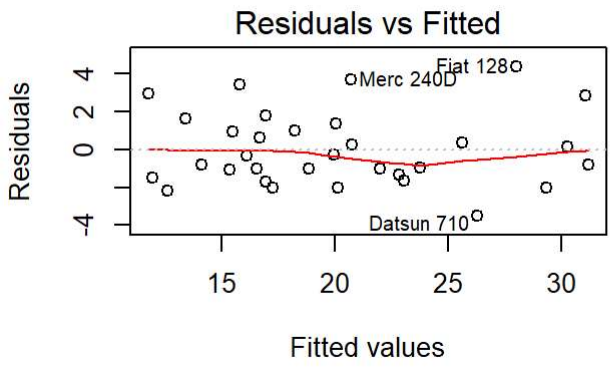
Scatter Plot of MPG Vs. Weight by Transmission:

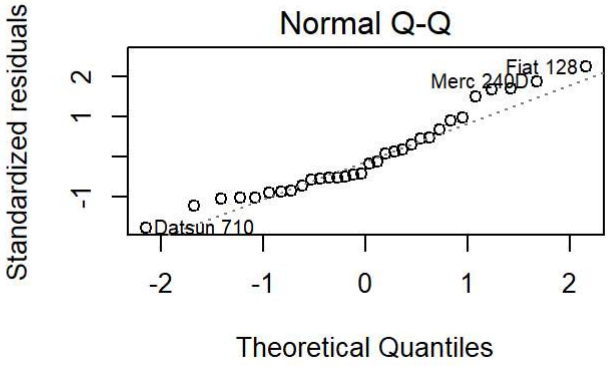
```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("Scatter Plot of MPG vs. Weight by Transmission")
```

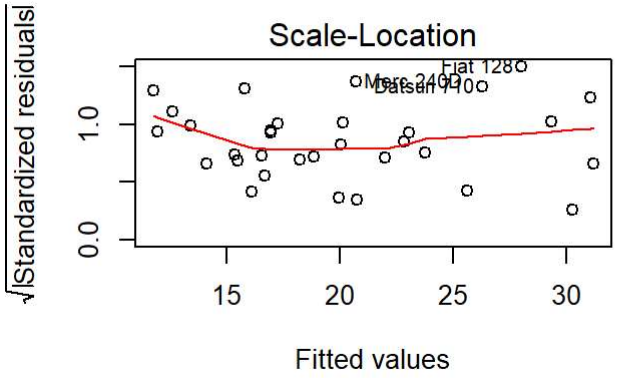
4.

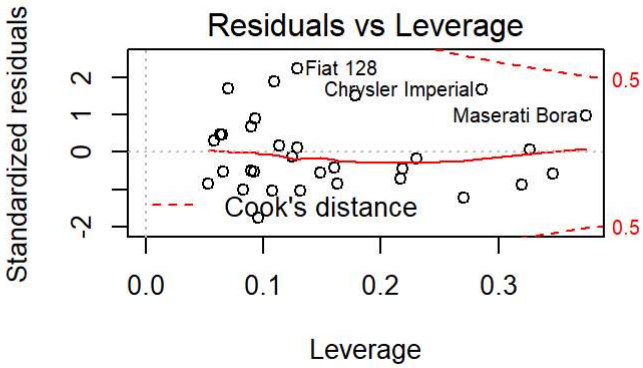
Residual Plots:

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


Residuals vs Fitted
Residuals
Fitted values
Merc 240D
Fiat 128
Datsun 710
-4
0
2
4
15
20
25
30


Normal Q-Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Merc 240D
Datsun 710
-1
1
2
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
Fiat 128
Merc 240D
Datsun 710
0.0
1.0
15
20
25
30


Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
-2
0
1
2
0.0
0.1
0.2
0.3