Motor Trend Data Analysis Report - Coursera Regression Model Project

M.Kotsits

February 5, 2016

Executive Summary

In this report, we analyze mtcars data set and explore the relationship between a set of variables with miles per gallon (MPG). We are particularly interested in the following two questions: (1)"Is an automatic or manual transmission better for MPG? (2) Quantify the MPG difference between automatic and manual transmissions? We use regression models and exploratory data analyses to explore how automatic (am = 0) and manual (am = 1) transmissions features affect the MPG feature. We show using t-test that there is a difference in performances between cars with automatic and manual transmission. There are around 7 MPG more for cars that have manual transmission than with automatic. Later, we fit the data by few linear regression models and select the one with highest adjusted R-squared value. For given weight and time =1/4 mile as a constant, we can conclude from our analysis that manual transmitted cars are 14.079 + (-4.141)*weight more MPG (miles per gallon) on average better than automatic transmitted cars. Thus, lightre cars with a manual transmission and heavier cars with an automatic transmission have higher MPG values.

First, we load the data set mtcars. Then we change some variables from 'numeric' class to 'factor' class.

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

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

```
str(mtcars)
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)
```

Exploratory data analyses.

Please see Appendix. According to the box plot,we see higher values of MPG for the manual transmission, in general.

Inference

To check weather both samples are from the same population, we crete the null hypothesis. H0: "The MPG of the automatic and manual transmissions are from the same population." (assuming the MPG has a normal distribution)
We use the two sample T-test to check that.

```
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
```

We reject our null hypothesis, because the p-value is very small 0.00137<0.05. Thus, the automatic and manual transmissions are not from the same population. And the mean for MPG of manual transmitted cars is around 7 units higher than the mean for MPG of automatic transmitted cars.

Regression Analysis:

First, we try to fit the data with the full model, fit1:

```
fit1 <- lm(mpg ~ ., data=mtcars)
summary(fit1) # results hidden
```

In this model mpg depends on all other variables. This model has the residual standard error as 2.833 on 15 degrees of freedom. Adjusted R-squared value is not that high 0.779. This means that the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level. That is why we look for another model.

To select some statistically significant variables, we we use backward selection

```
fit2 <- step(fit1, k=log(nrow(mtcars)))
summary(fit2) # results hidden
```

This model is "mpg ~ wt + qsec + am". It has the Residual standard error as 2.459 on 28 degrees of freedom. Adjusted R-squared value is 0.8336, which is higher value than the one we got in the previous model. All coefficients are significant at 0.05 significant level.

Next, we try to fit the data with the simple model with MPG as the outcome variable and Transmission as the predictor variable.

```
fit3<-lm(mpg ~ am, data=mtcars)
summary(fit3) # results hidden
```

It shows that on average, a car has 17.147 mpg with automatic transmission, and if it is manual transmission, 7.245 mpg is increased. This model has the Residual standard error as 4.902 on 30 degrees of freedom. Adjusted R-squared value is 0.3385, which is very low. The low Adjusted R-squared value indicates that we need to add other variables to the model.

For the next model, please see the Appendix. According to the scatter plot, it indicates that there appears to be an interaction term between "wt" variable and "am" variable, since automatic cars tend to be heavier in weight than manual cars. Thus,our next model includes an interaction term:

```r
fit4<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(fit4) # results hidden
```

We can say that this is a good model because the residual standard error as 2.084 on 27 degrees of freedom and high adjusted R-squared value is 0.8804. All of the coefficients are significant at 0.05 significant level. Finally, we select the final model.

```r
anova( fit1, fit2, fit3,fit4)
confint(fit4) # results hidden
```

Finally, we end up by selecting the model with the highest Adjusted R-squared value, "mpg ~ wt + qsec + am + wt:am".

```r
summary(fit4)$coef
```

Thus, the result shows that when "wt" (weight lb/1000) and "qsec" (1/4 mile time) remain constant, cars with manual transmission add 14.079 + (-4.141)*wt more MPG (miles per gallon) on average than cars with automatic transmission. 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.

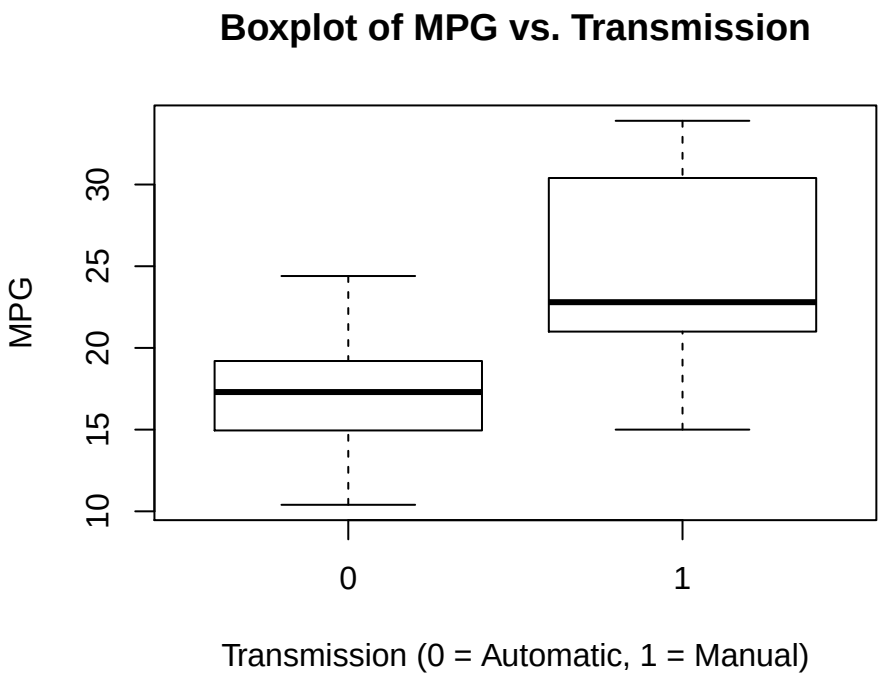
Residual Analysis and Diagnostics

Please refer to the Appendix: Figures section for the plots. According to the residual plots, we can verify the following underlying assumptions: 1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption. 2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line. 3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed. 4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands. Therefore, the above analyses meet all basic assumptions of linear regression and well answer the questions.

Appendix: Figures

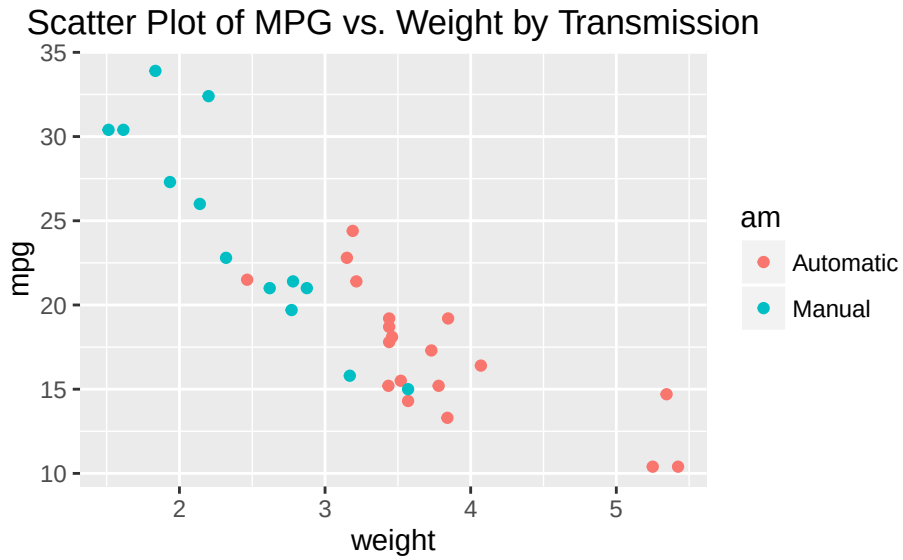
Boxplot of MPG vs. Transmission

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



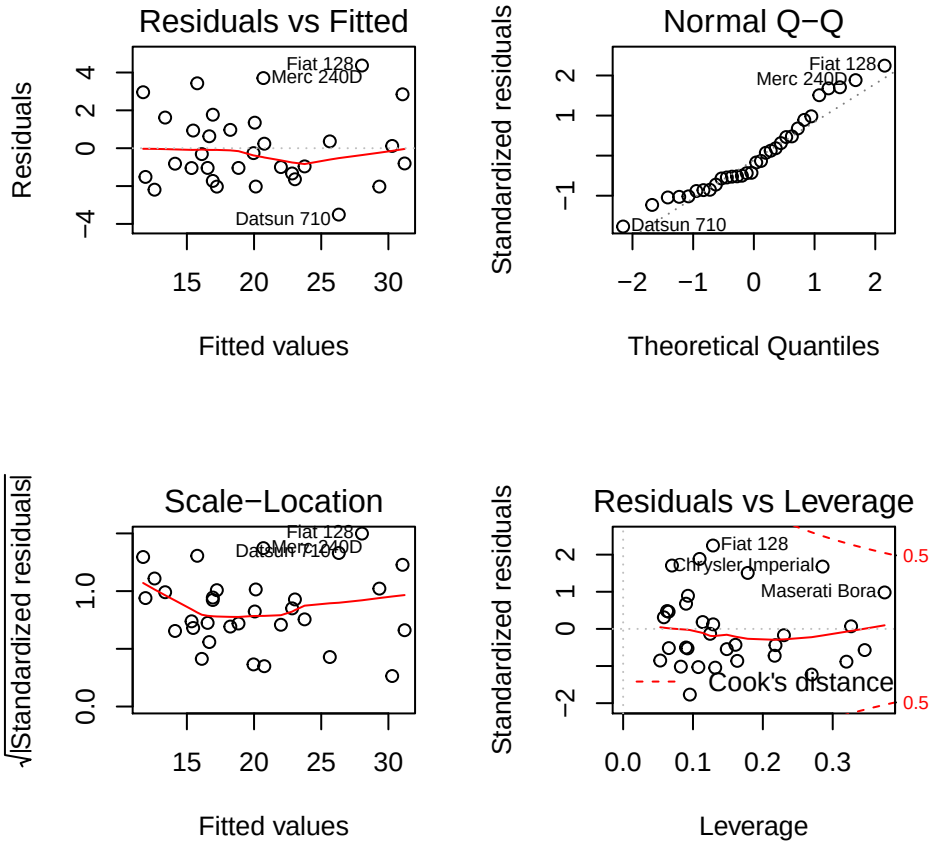
Scatter Plot of MPG vs. Weight by Transmission

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



Residual Plots

```
# diagnostic plots
par(mfrow=c(2,2))
plot(fit4)
```


Residuals vs Fitted
Residuals
Fitted values
Fiat 128
Merc 240D
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Merc 240D
Datsun 710
Scale−Location
√|Standardized residuals|
Fitted values
Fiat 128
Merc 240D
Datsun 710
Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5