motor_trend_project

Monica

19 de marzo de 2017

Executive summary

In this report, we will analyze mtcars data set and explore the relationship between a set of variables and miles per gallon (MPG). We try to know: *"Is an automatic or manual transmission better for MPG" and "Quantify the MPG difference between automatic and manual transmissions"*

We use regression models and exploratory data analyses to explore how automatic and manual transmissions features affect the MPG feature. We determine that there is a signficant difference between the mean MPG for automatic and manual transmission cars, manual transmission better for MPG (miles per gallon) than automatic transmission.

Processing the data

We load the data set mtcars and change some variables from numeric class to factor class.

```
data(mtcars)
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)

str(mtcars)
```

Exploratory data analysis

We do some basic exploratory data analyses and we can see the figures in "Appendix of Figures"

With the boxplot we can see that manual transmission yields higher values of MPG in general.

In the pairs,we can see some higher correlations between variables like "wt", "disp", "cyl" and "hp".

The Scatter Plot indicates that there appear to be an interaction term between "wt" variable and "am" variable

Regression Analysis

First, we fit the simple model with MPG as the outcome variable and Transmission as the predictor variable.

```
first_Model<-lm(mpg ~ am, data=mtcars)

summary(first_Model)
```

We can see the summary's result this model has the Residual standard error as 4.902 on 30 degrees of freedom. The Adjusted R-squared value is 0.3385, which means that the model can explain about 34% of the variance of the MPG variable. The low Adjusted R-squared value also indicates that we need to add other variables to the model.

We fit the full model:

```
full_Model <- lm(mpg ~ ., data=mtcars)
summary(full_Model)
```

We can see the summary´s result, this model has the Residual standard error as 2.833 on 15 degrees of freedom. The Adjusted R-squared value is 0.779, which 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.

For to select some statistically significant variables, we use backward selection

```
step_Model <- step(full_Model, k=log(nrow(mtcars)))
```

We can consult the summary´s result, this model is "mpg ~ wt + qsec + am". The Residual standard error as 2.459 on 28 degrees of freedom. The Adjusted R-squared value is 0.8336, which means that the model can explain about 83% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

In the Scatter Plot we can see an interaction between "wt" and "am", we include the interaction in our model:

```
new_Model<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
```

We consult the summary's result: Residual standard error as 2.084 on 27 degrees of freedom. The Adjusted R-squared value is 0.8804, which means that the model can explain about 88% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

We run an ANOVA to compare the models and see if they are significantly different.

```
anova(first_Model,full_Model, step_Model , new_Model)
```

For all the above, we select the model with the highest Adjusted R-squared value, "mpg ~ wt + qsec + am + wt:am" is the *best model*.

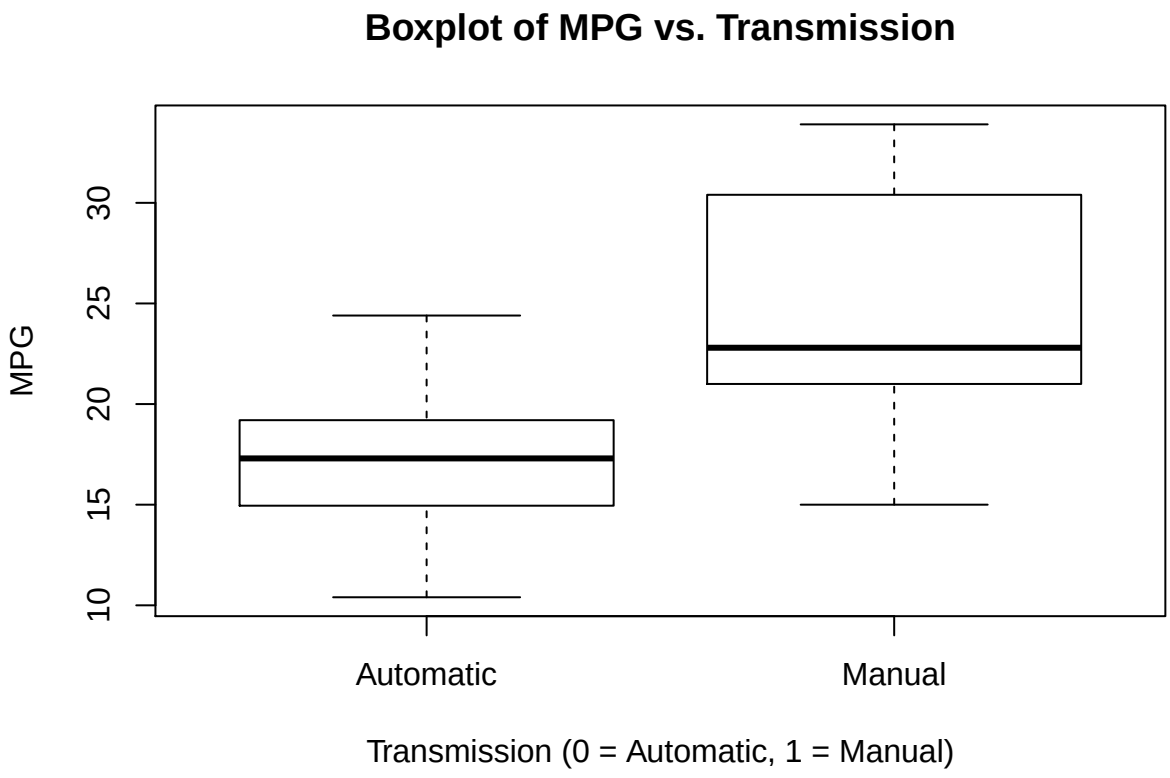
Now, check residuals to see whether they are normally distributed and examine residual vs fitted values plot to spot for heteroskedasticity.

We can see in the appendix figures the graphic and we can verify:The Residuals vs. Fitted plot support the accuracy of the independence assumption. The Normal Q-Q plot indicates that the residuals are normally distributed. The Scale-Location plot as the points are randomly distributed, confirms the constant variance assumption. For last, the Residuals vs. Leverage argues that no outliers are present.

Appendix Figures

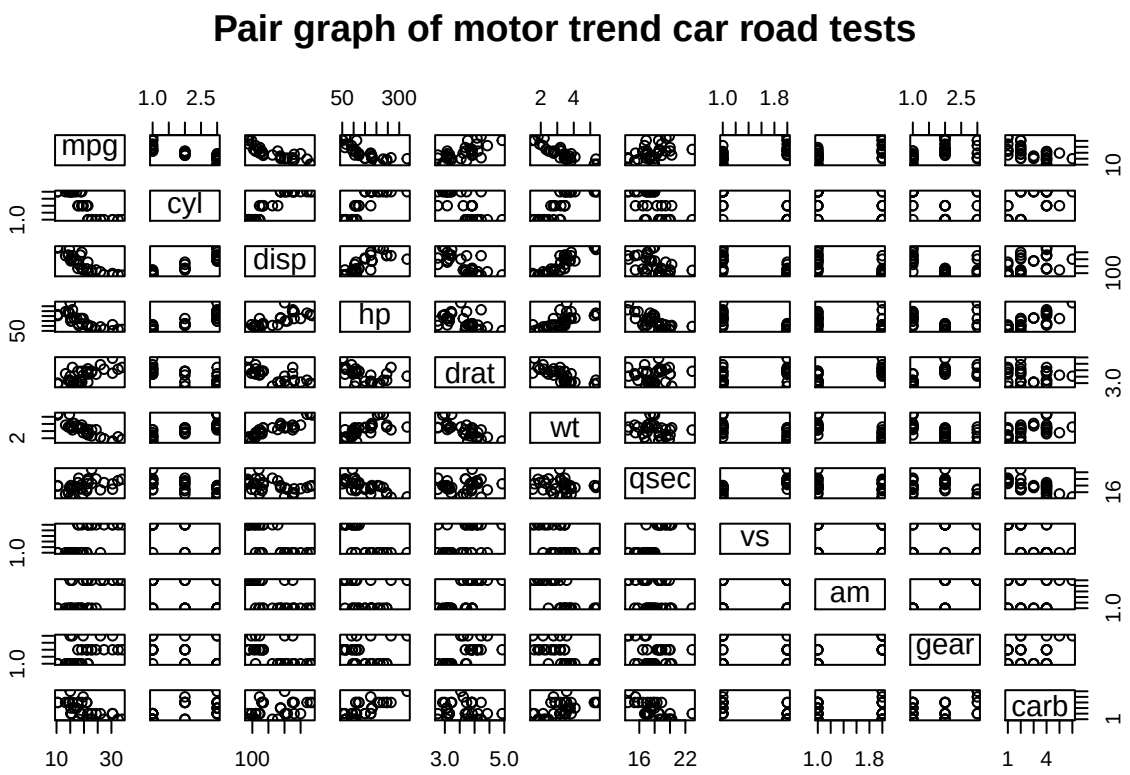
Boxplot:

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



We do a pair graph of Motor Trend Car Road Tests:

```
pairs(mpg ~ ., data=mtcars, main="Pair graph of motor trend car road tests")
```



We do a Scatter Plot of MPG vs. Weight by Transmission

```
library(ggplot2)
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")
```

We do a Residual Plots

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