Regression Models

Viktor M

Friday, July 24, 2015

My github page

Executive Summary

In this study we try to find relationship between cars fuel economy (mpg) and other common cars characteristics. For this study we will use widely available *mtcars*, extracted from 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). Our analysis will attempt to determine whether an automatic or manual transmission is better for MPG, and quantifies the MPG difference.

Plan of our study is:

- Utilize R language processing to build regression models and exploratory data analyses to mainly explore how automatic (am = 0) and manual (am = 1) transmissions features affect the MPG feature. The t-test shows that the performance difference between cars with automatic and manual transmission.
- Then, we fit several linear regression models and select the one with highest Adjusted R-squared value. So, given that weight and 1/4 mile time are held constant, manual transmitted cars are 14.079 + (-4.141)*weight more MPG (miles per gallon) on average better than automatic transmitted cars. Thus, cars that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will have higher MPG values.
- This conclusion holds whether we consider the relationship between MPG and transmission type alone or transmission type together with 2 other predictors: wt / weight; and qsec / 1/4 mile time.

Preparing Test Data

Here we load necessary libraries and dataset to be used in our research and we change data type to factor.

```
data(mtcars)
str(mtcars)
head(mstcars,3
library(ggplot2)
```

```
dim(mtcars)
```

And here we change data type for parameters used in research to factors.

```
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)
```

Then, we do some basic exploratory data analyses. Please refer to the **Appendix: Figures** section for the plots. According to the box plot, we see that manual transmission yields higher values of MPG in general. And as for the pair graph, we can see some higher correlations between variables like “wt”, “disp”, “cyl” and “hp”.

Inference

At this step, we make the null hypothesis as 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 show it.

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

```
result$estimate
```

Since the p-value is 0.00137, we reject our null hypothesis. So, the automatic and manual transmissions are from different populations. And the mean for MPG of manual transmitted cars is about 7 more than that of automatic transmitted cars.

Regression Analysis

First, we fit the full model as the following.

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

This model has the Residual standard error as 2.833 on 15 degrees of freedom. And 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.

Then, we use backward selection to select some statistically significant variables.

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

This model is “mpg ~ wt + qsec + am”. It has the Residual standard error as 2.459 on 28 degrees of freedom. And 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.

Please refer to the **Appendix: Figures** section for the plots again. According to the scatter plot, it indicates that there appear to be an interaction term between “wt” variable and “am” variable, since automatic cars tend to weigh heavier than manual cars. Thus, we have the following model including the interaction term:

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

This model has the Residual standard error as 2.084 on 27 degrees of freedom. And 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. This is a pretty good one.

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

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

Finally, we select the final model.

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

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

```
summary(amIntWtModel)$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.

As for the Dfbetas, the measure of how much an observation has effected the estimate of a regression coefficient, we get the following result:

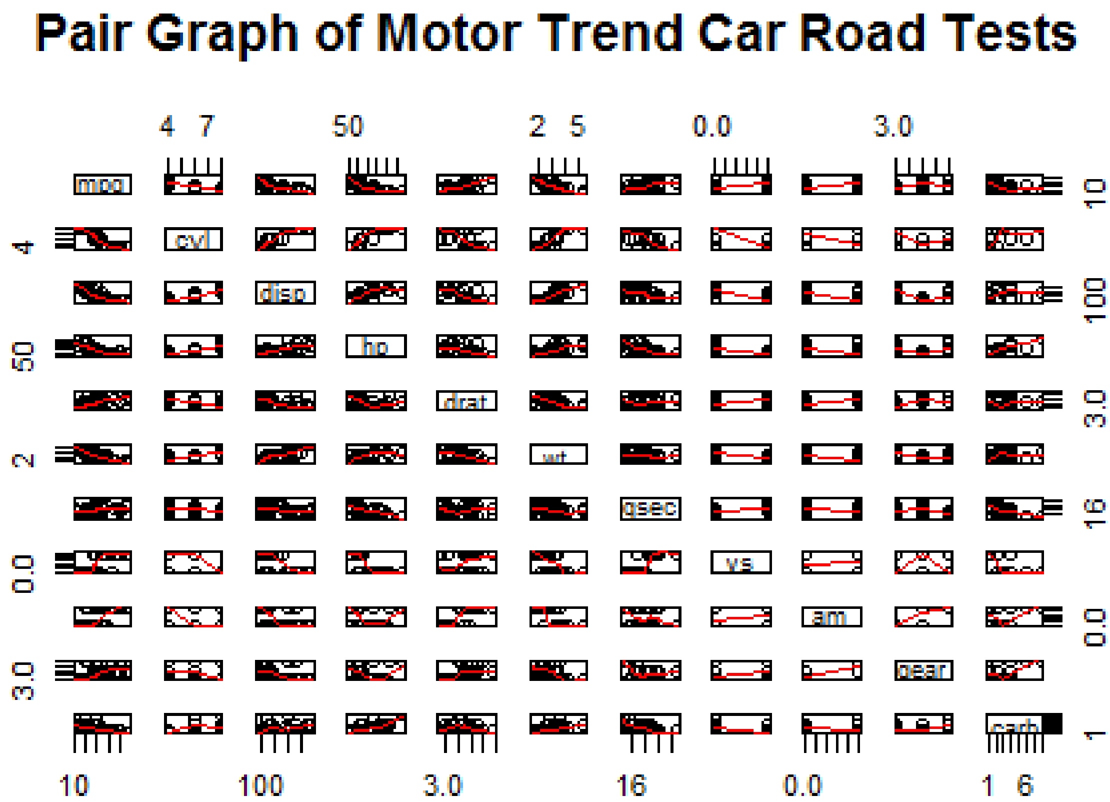
```
sum((abs(dfbetas(amIntWtModel)))>1)
```

Therefore, the above analyses meet all basic assumptions of linear regression and well answer the questions.

Appendix: Figures and results

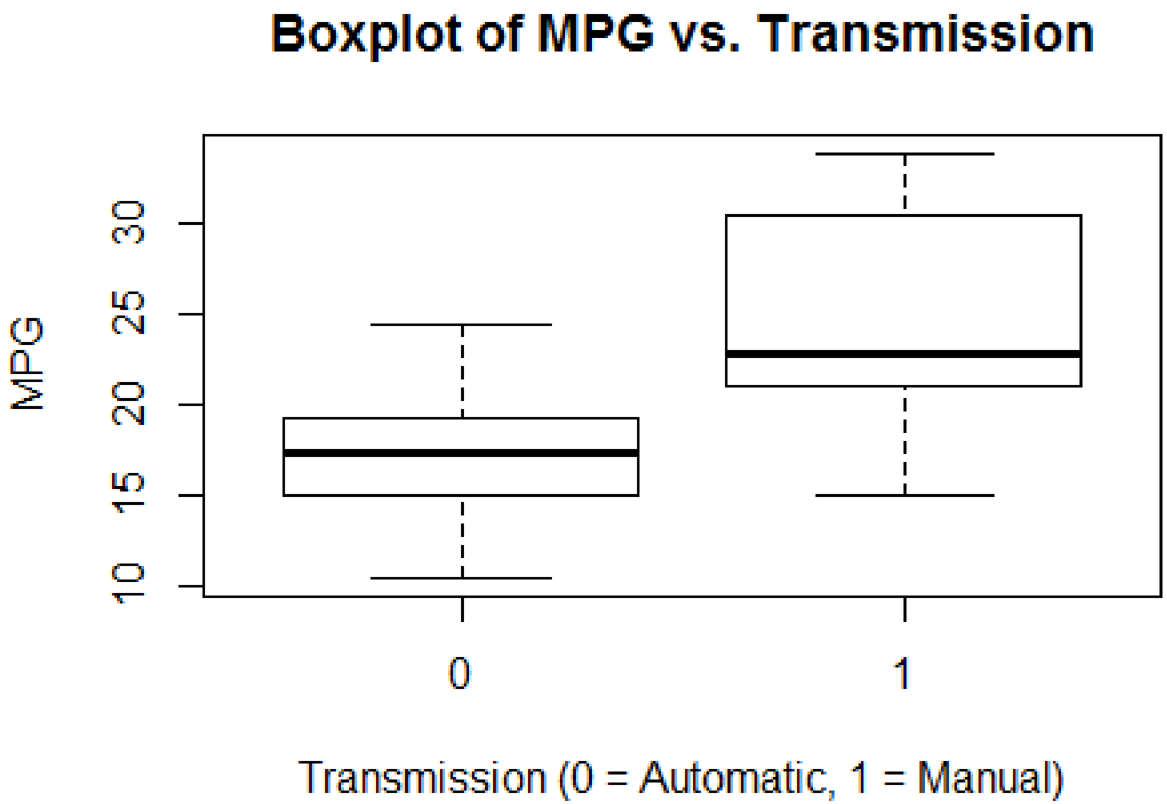
1. Pair Graph of Motor Trend Car Road Tests

    ```
    pairs(mtcars, panel=panel.smooth, main="Results of Pairs f Motor Trend Car Road Tests")
    ```



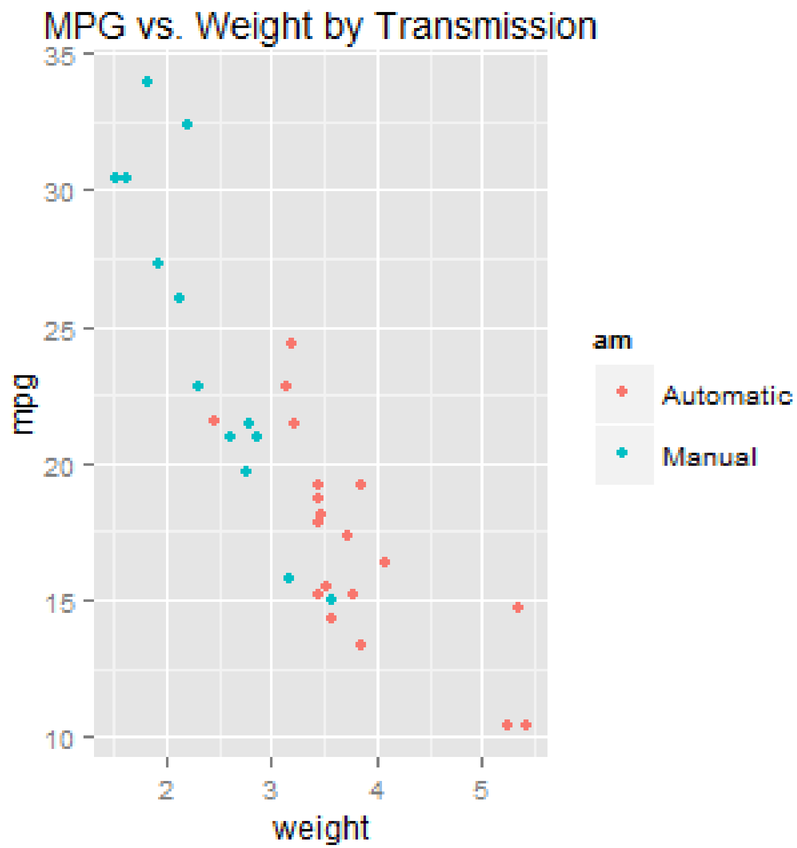
2. Boxplot of MPG vs. Transmission

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



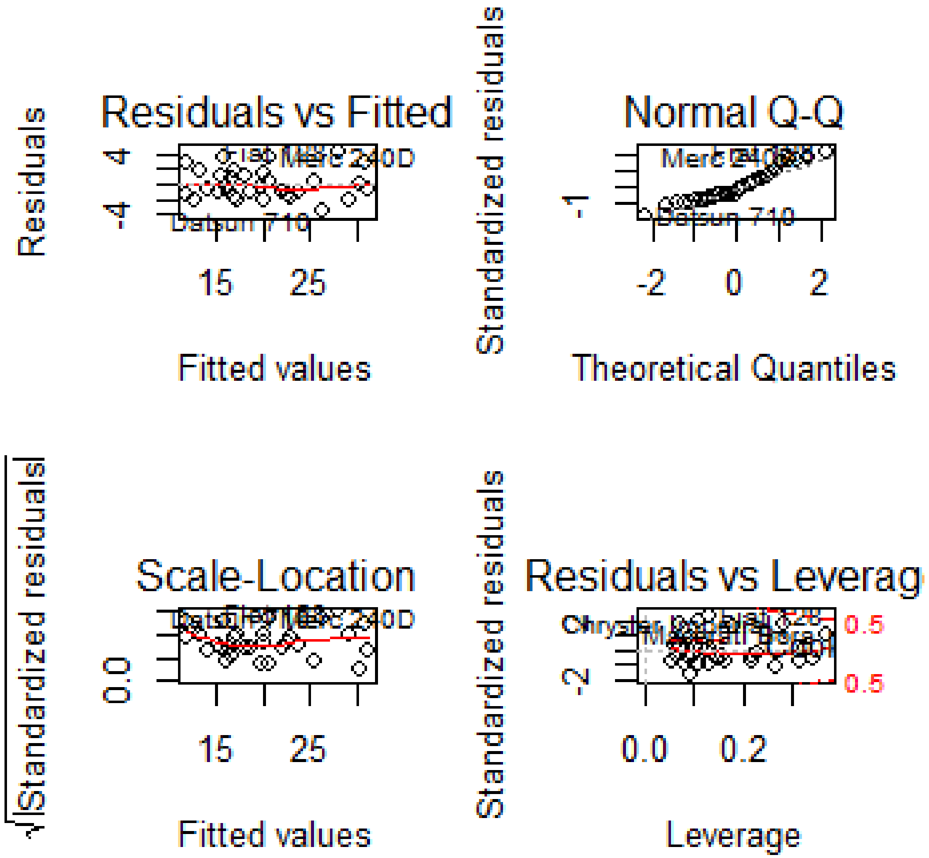
3. Pr Plot of MPG vs. Weight by Transmission

```
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("MPG vs. Weight by Transmission")
```



4. Residual Plots

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



Conclusion

Based on the observations from our best fit model, we can conclude the following, - If number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am). - Cars with Manual transmission get more miles per gallon mpg compared to cars with Automatic transmission. (1.8 adjusted by hp, cyl, and wt). - mpg will decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in wt. - mpg decreases negligibly with increase of hp.