Applied Regression Course Project

Bohdan Horak

Friday, August 22, 2014

## Executive Summary

The purpose of this paper is to examine the effects of a variety of car related features on the miles per gallon performance of a vehicle. In order to convincingly illustrate the effects of factors on the car's gaes mileage a variety of statistical tools will be employed, using the native "mtcars" data provided with R. This paper will examine all factors provided in the data set, however the focus will lie on the type of transmission the car has equiped. There are two types of transmissions Manual and Automatic, this paper will exame which is better or wether the differences between the two statistically insignificant given the provided data. In our analysis we will use a significance level of 5.

## Exploratory Analysis

### Data Description

The description of the variables, provided by the R help files.

In order to perform the exploratory analysis I split the variables into two camps, continous and discrete.

For continous variables I used basic scatter plots with a fitted line to get a sense of the relationship. You can see the plots in figure 1. We can tell that as the weight(wt), the displacement(disp) and Horsepower(hp) increase the mpg tends to decrease. The opposite is true for the Rear axle ratio(drat) and quarter mile time(qsec).

For discrete variable I used a series of bar plots(figure 2). The bar plots show the mean mpg for each variable seperated by factors. I also, included an error range which corresponds to a one deviation shift in the standard error of the mean. Please note, the carb variable has very few observation inclduded therefore any comparative anylysis between the factors is sure to be unreliable, so it will be ignored from now on.

## Analysis

For the regression analysis the discrete variables were coded as factor variables. The regression system in R automaticaly converts these factor variables into dummy's.For example, the regression will treat the automatic transmission as a 0,in other words the base case, and the manual transmission as a 1. So from looking at our boxplots we can expect the slope coefficient to be positive.

### Regression 1

To get things started I did a multiple regression across all variables except carb. The summary of coefficients is provided in figure 4. The R-squared ended up being 0.881, which is quite high. However, the p-values of the T-tests on wether or not the coefficients calculated by the regression are significant or not are all quite and not close to a reasonable significance level of 5%. This likely means that there are confounding variable which are not adding any explanatory power to our analysis.

In order to weed out these confounding variables I looked at the covariance matrix of all the variables.We looked for variables that have really high or low correlations, I then ran another regression to see wether this will improve our model accuracy.

### Regression 2

Looking at the caovariance matrix it can be seen that disp and wt have a high correlation correlation. Displacement was removed and a new regression was created. The summary of which can be seen figure 5. As can be seen removing the displacemendid not cause the R^2 to decresea by much, however our P values have decreased, hoever they are still quite unsatifactory.

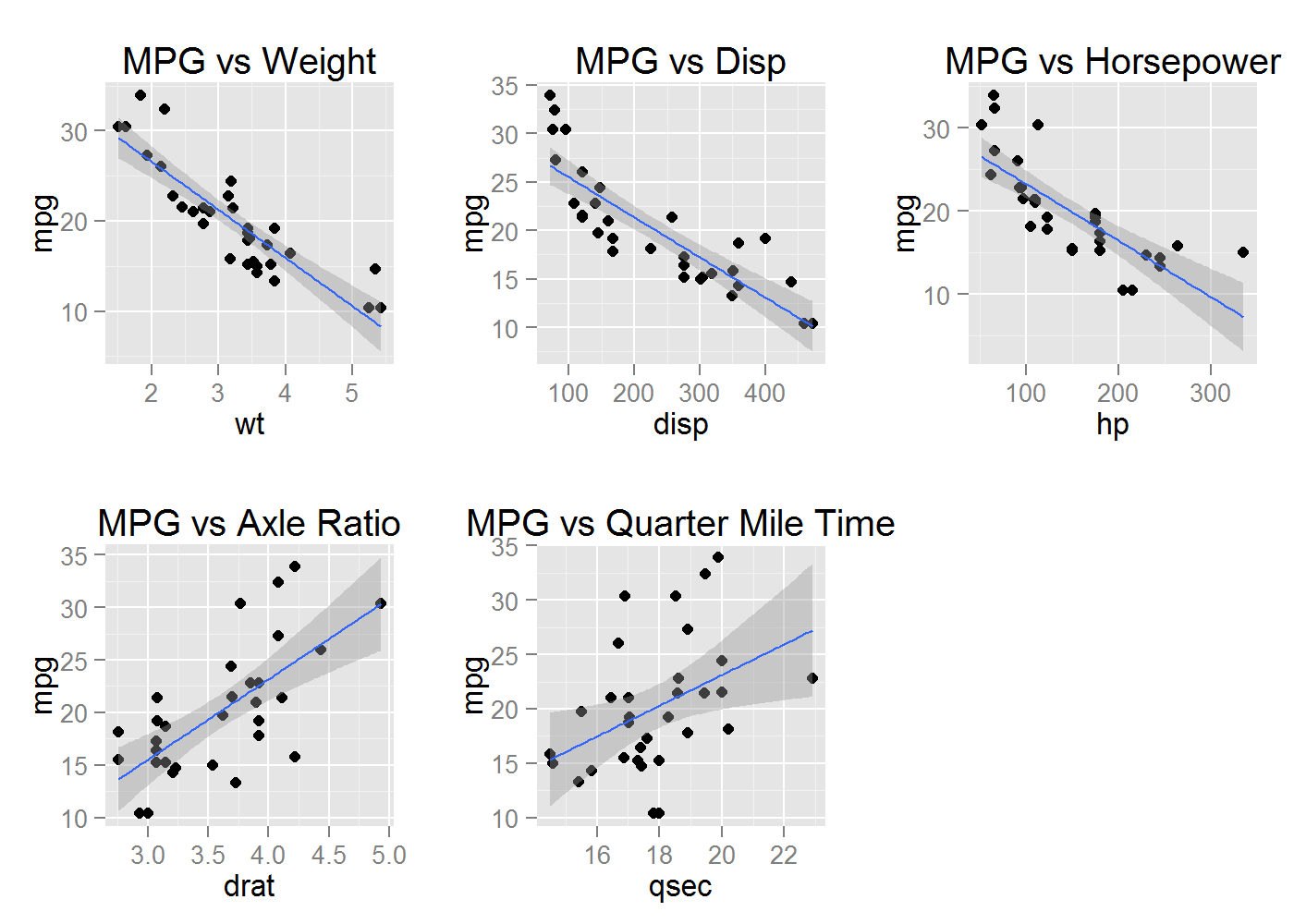
### Regression 3

Repeating the same process, the variable vs is eliminatedm, and the regression rerun. The results are similar as above. Our P values decrease and R-sq remains relatively unchanged. The summary of the regression is Figure 6.

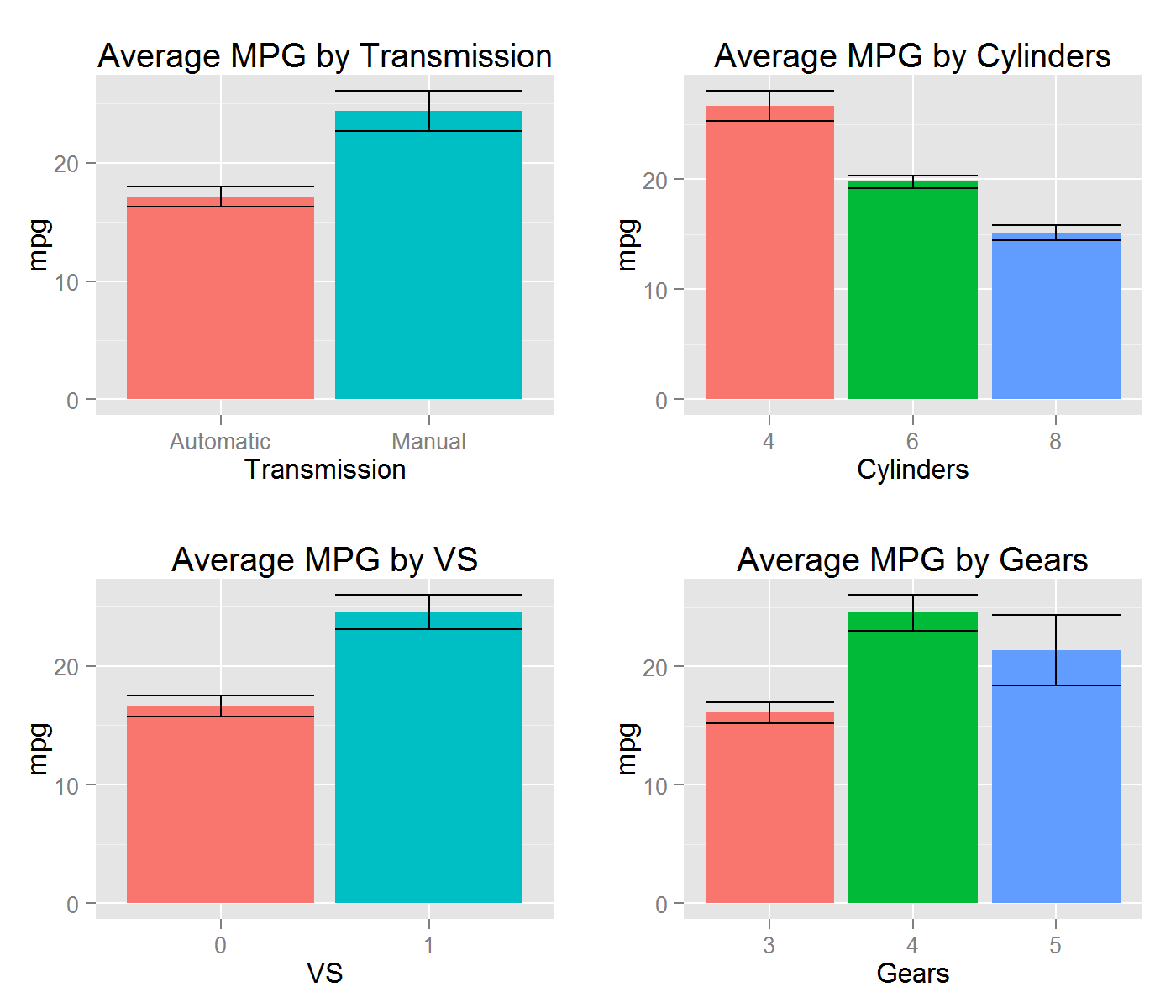
### Regressions 4-Final

In the interest of brevity I repeated the process until the p values of the model seemed reasonable. I ended up removing the all but am and wt variables. The results of which can be seen in Figure 7.We can see that both of the P values resulting from the T-test are effectively 0 meaning that they are statistically significant. At this point we can be reasonably certain that low horspower vehicles with manual transmissions tend to be more gas efficient.

### Figure 1:



### Figure 2:



### Figure 3: Anova, Summary

### Figure 4: Regression #1

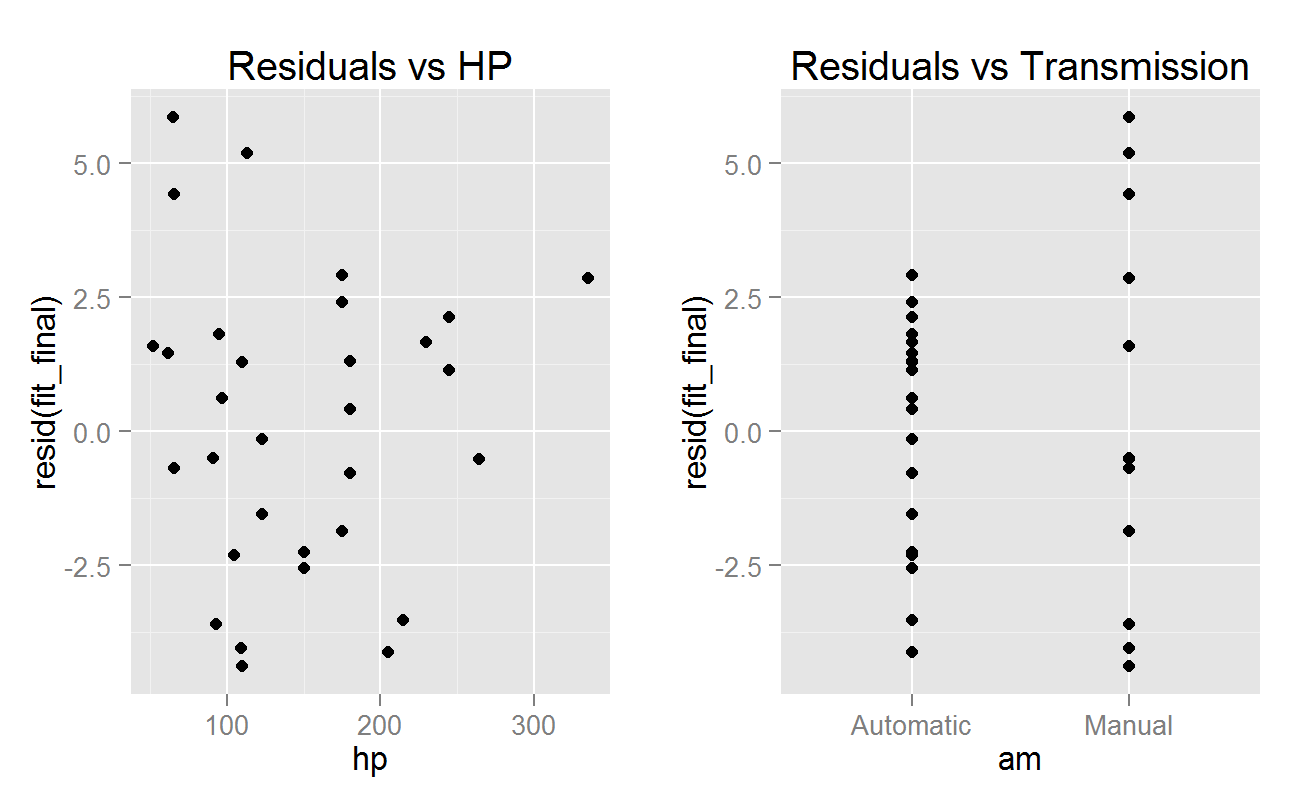
### Figure 5: Covariance Matrix

### Figure 6: Regression #2

### Figure 7: Rgression #3

### Figure 8: Regression Final

### Figure 9: Residuals



### Figure 10: Residuals vs Fitted

