Regression Models Project

Kyle Maurice

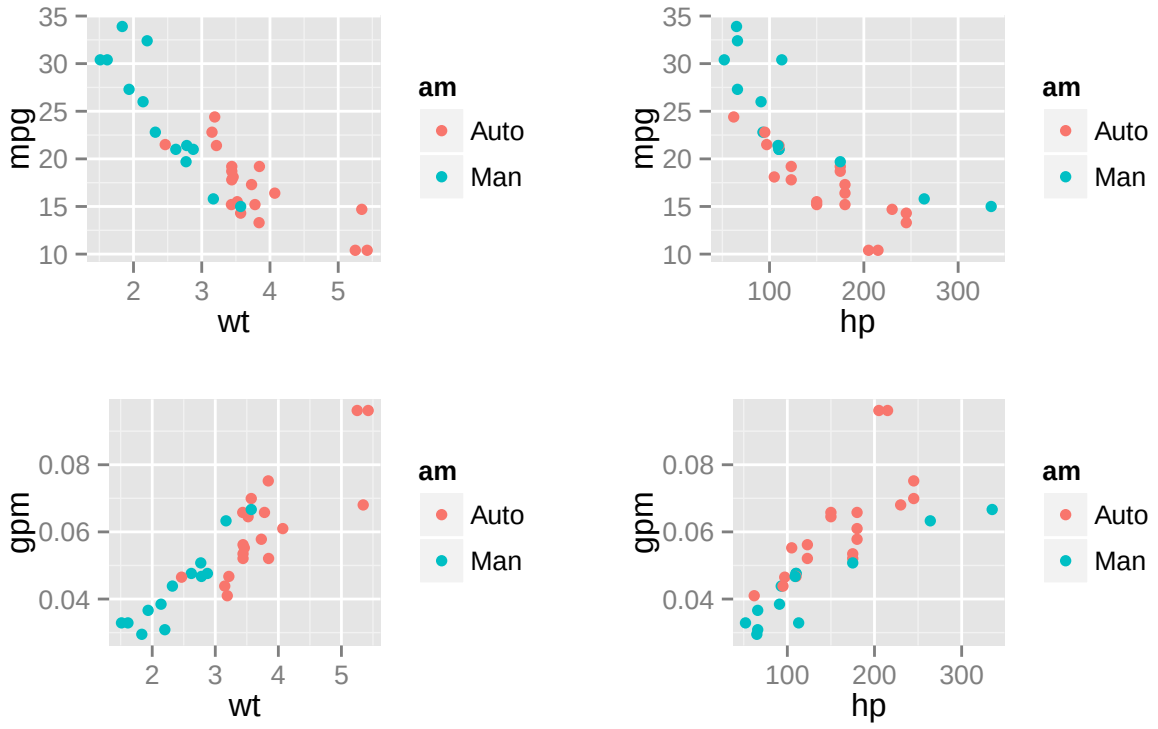
December 23, 2015

Executive Summary

In an analysis of varied population of cars from the mtcars dataset in R, it was found that there was no significant statistical difference between the manual and automatic transmission with respect to MPG. In the analysis GPM, a transformation of MPG, was used as a proxy for MPG to avoid non-linearity in the data.

Introduction

Starting from the mtcars dataset in R, an anlysis will be done to quamtify the difference between manual and automatic transimission on an population fo cars. The types of cars represented in the data set range from american muscle and luxury cars to exotic sports cars to small economy cars. As a starting point in the anaylsis, a few exploratory plots will be created to examine the relationships between the parameters of the data set.



From the plots we can see a strong negative influence of displacement, cylinders, weight and horsepower. There is a weak to moderate positive influence of gears, am, qsec and vs on mpg. It is important to note that the plot of MPG with respect to weight, displacement and horsepower all look asymptotic at that approach the origin. The relationship looks like 1/x and a transformation might be helpful in removing the non linearity. As shown in the figure below, the relationships in the data are more linear when MPG is transformed to GPM (i.e. GPM = 1/MPG). The inputs representing the number of cylinders, horsepower and misplacement are all correlated with each other. This follows from knowledge of cars. only one of those parameters should be necessary for the model. Clearly weight should be part of the model but the selection of other parameters for the model is more difficult.

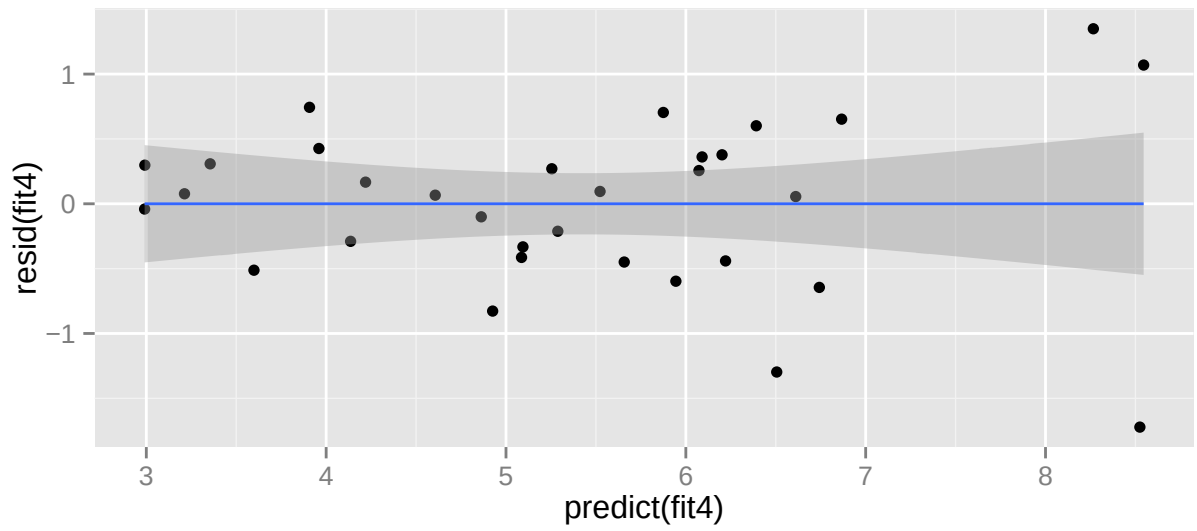
Analysis

The model should start from the variable in question and add parameters until the model in accurate enough to provide some inference about the question of the impact of transmission type can be answered. The model of the transformed and scaled variable will be built to remove the nolineatiry in the MPG measurements.

R squared for the model is insufficient for an accurate model. More parameters must be added. The first candidate for additional parameters is weight, since the exploratory plots show a strong correlation with GPM.

Adding weight has removed significant error from the model. Searching in the residual plot for the next parameter to add. The residual plots below show that there is a slight upward trend in residuals for quarter mile time, displacement, horsepower, and number of cylinders. Since qsec has the smallest correlation to wt, that parameter will be added next.

```
fit4 <- lm(I(gpm*100)~am+wt+qsec, data = mtcars )
summary(fit4)$coefficients
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  4.94158611 1.91472775  2.5808296 1.538878e-02
## amMan       -0.05860867 0.38816897 -0.1509875 8.810681e-01
## wt           1.40149203 0.19566626  7.1626659 8.542054e-08
## qsec        -0.22432958 0.07941896 -2.8246352 8.626708e-03
summary(fit4)$r.squared
## [1] 0.8467749
qplot(predict(fit4),resid(fit4))+geom_smooth(method="lm")
```



The residuals look random and the models appears to be healthy. From the coefficients we can perform hypothesis test on the difference between the Automatic and Manual transmission types. The measure of the difference between the value of the coefficient in the regression model are not statistically different from one another. A hypothesis test for for the difference between the transmission types will return the null hypothesis.

Appendix

```
fit1 <- lm(I(gpm*100)~am, data = mtcars )
summary(fit1)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  6.144642  0.3224107 19.058429 2.588825e-18
## amMan       -1.777029  0.5058396 -3.513028 1.426570e-03
```

```
summary(fit1)$r.squared
```

```
## [1] 0.2914731
```

R squared for the model is insufficient for an accurate model. More parameters must be added. The first candidate for additional parameters is weight, since the exploratory plots show a strong correlation with GPM.

```
fit3 <- lm(I(gpm*100)~am+wt, data = mtcars )
summary(fit3)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -0.1280333  0.7429582 -0.1723291 8.643759e-01
## amMan        0.4829529  0.3759364  1.2846664 2.090762e-01
## wt           1.6643275  0.1917190  8.6810787 1.473699e-09
```

```
summary(fit3)$r.squared
```

```
## [1] 0.8031136
```