Regression Project

Cesar Espitia

July 9, 2015

Executive Summary - Motor Trends

The purpose of this project was to determine the relationship between a set of variables and the vehicles fuel efficiency quantified as MPG. Specifically the following:

- which transmission has better MPG, Automatic or Manual?
- expalin the differences between them

Fromt he analysis to follow, it became clear tha although there are quantifiable differences in one transmission type being better than the other, other more appropriable variabels were better estimates of fuel efficiency. In addition, the data set would benefit from having more data points (n>32; perhaps n>100) in order to have higher degrees of confidence in the results.

Data Preparation

The data is already a part of the R library of data related and doesn't require much manipulation. For the purposes of this exercise, the variables cyl, vs, gear, carb and am were coerced into factors in order to better understand the ffects on the regression model as they are binary / categorical in nature and truly continuous.

```
#load and store data
library(datasets, warn.conflicts = FALSE); library(ggplot2, warn.conflicts = FALSE); library(caret, war
```

```
## Loading required package: lattice
```

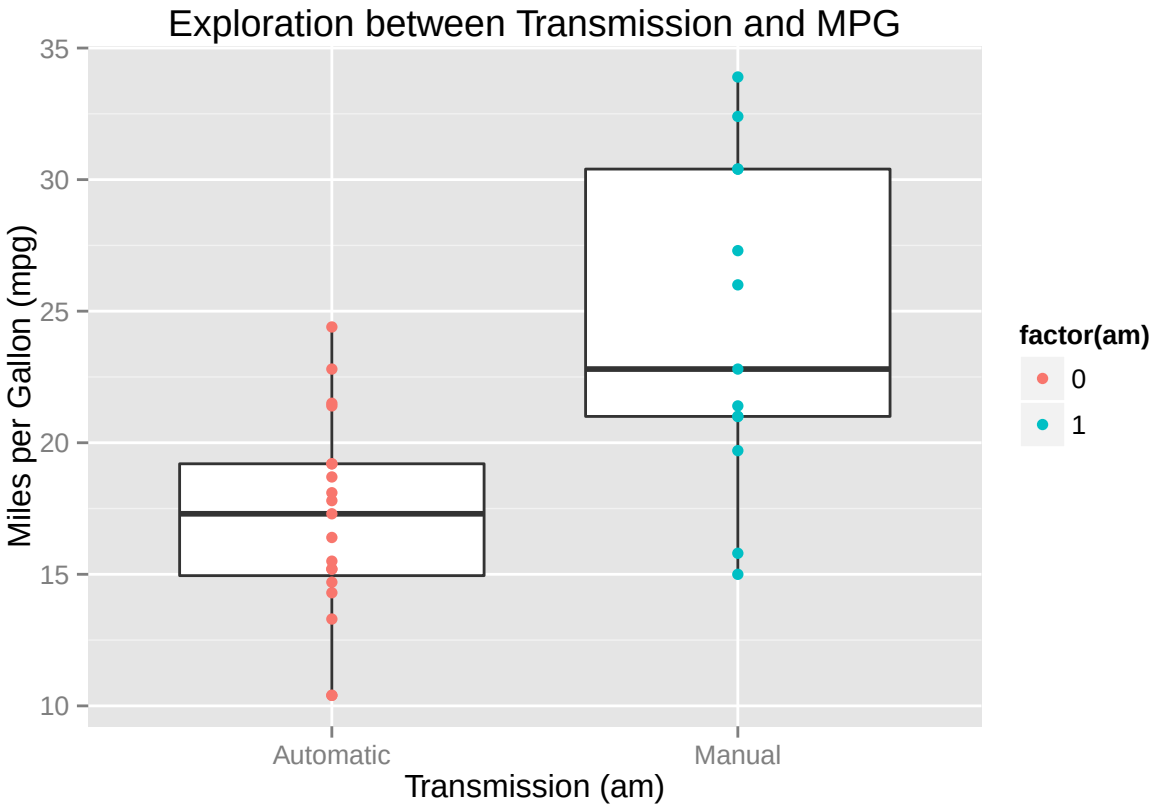
```
cars <- mtcars; cars$cyl  <- factor(cars$cyl); cars$vs   <- factor(cars$vs);cars$gear <- factor(cars$gea
```

Exploratory Analysis

Box Plot - Transmission vs. MPG

A boxplot was chosen to visualize the correlation between transmission and fuel efficiency, the data points were also overlayed to determine if any outliers exist.

```
#note Automatic = 0, Manual = 1
ggplot(mtcars, aes(factor(am), mpg))+geom_boxplot()+geom_point(aes(color=factor(am)))+scale_x_discrete(l
```



In revieweing the data, it is clear that vehicles with a Manual transmission have better MPG, but if you look closely the variability is quite large while the variability in the Automatic category is less. We can quantify this with a two-sample t-test and determine if determine if the null hypothesis (H0: no difference between Manual and Automatic) is valid or should be reject.

```
ttest <- t.test(mpg~am, data=cars)
ttest$p.value
```

```
## [1] 0.001373638
```

With a P Value of 0.0014, we can reject the null hypothesis. There is a difference between the categories.

Correlation matrix - All variables

In looking at this visual correlation matrix, ellipses that are very thin and either red or dark blue have the strongest correlations.

```
ctab <- cor(mtcars)
ctab[,1]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

In looking at the matrix, it appears as though MPG can be explained by the following variables cyl, hp and wt without really taking into consideration am. A visual of this is shown in Appendix A.

Le's explore this more with actual regression models.

Regression Models

Our initial model fit revolves around the relationship between MPG and the transmission type. This model only has a R-squared value of 0.338. Not very strong.

```
simple <- lm(mpg~am, cars)
summary(simple)$adj.r.squared
```

```
## [1] 0.3384589
```

Next we will use the step function to determine what variables do have an influence on MPG. We first build the model using all variables and then step through models removing variables until a strong correlation is found. With the use of AIC as a model quality estimator we note that the model lm(mpg ~ cyl + am + hp + wt) is the most effective. For a full detail look, please review Appendix A.

```
all <- lm(mpg~., cars)
least <- step(all, direction="backward", trace=FALSE)
```

```
summary(least)$adj.r.squared
```

```
## [1] 0.8400875
```

The point of concern on this model is that in reviewing model more closely, the variables cyl and am are used, but only for portions of factors (cyl 6/8 and Manual respectively). In my opinion, this is not a complete model although R-squared is 0.8401.

I have further whitled down my model and finalized it at mpg as a variable of weight and horsepower. This has a value of 0.815. Which is just as strong as the variable with

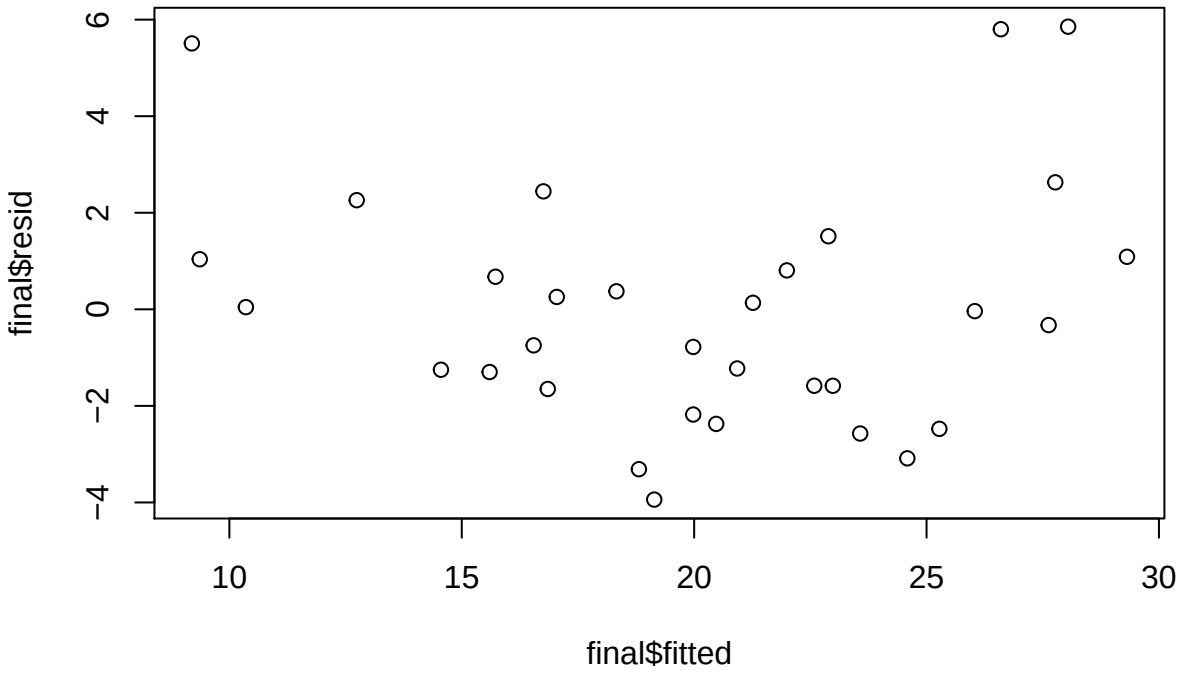
```
final <- lm(mpg ~ hp+wt, cars)
summary(final)$adj.r.squared
```

```
## [1] 0.8148396
```

Residual Plots

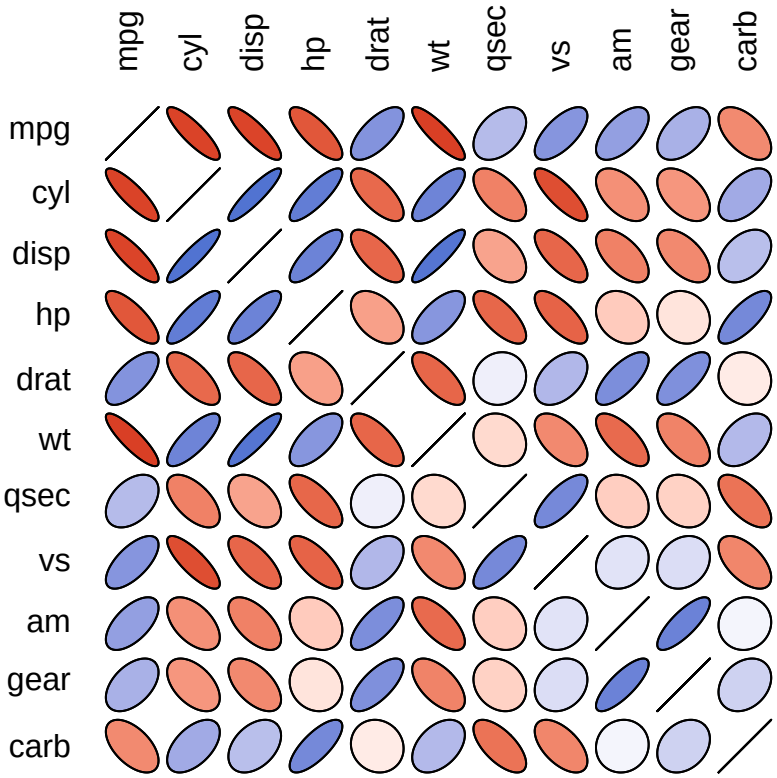
Next, we visually inspect the results of the model through its residual plots. For the other plots, including Normal Q-Q please refer to Appendix B.

```
plot(final$fitted, final$resid)
```



Appendix A

```
library(ellipse)
ctab <- cor(mtcars)
colorfun <- colorRamp(c("#CC0000","white","#3366CC"), space="Lab")
plotcorr(ctab, col=rgb(colorfun((ctab+1)/2), maxColorValue=255), mar = c(0.1, 0.1, 0.1, 0.1))
```



Appendix B

```
least <- step(all, direction="backward")
```

```
## Start:  AIC=76.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  5   13.5989 134.00 69.828
## - gear  2    3.9729 124.38 73.442
## - am    1    1.1420 121.55 74.705
## - qsec  1    1.2413 121.64 74.732
## - drat  1    1.8208 122.22 74.884
## - cyl   2   10.9314 131.33 75.184
## - vs    1    3.6299 124.03 75.354
## <none>              120.40 76.403
## - disp  1    9.9672 130.37 76.948
## - wt    1   25.5541 145.96 80.562
## - hp    1   25.6715 146.07 80.588
##
## Step:  AIC=69.83
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 67.005
## - disp  1    0.9934 135.00 68.064
```

```
## - drat  1    1.1854 135.19 68.110
## - vs    1    3.6763 137.68 68.694
## - cyl   2   12.5642 146.57 68.696
## - qsec  1    5.2634 139.26 69.061
## <none>              134.00 69.828
## - am    1   11.9255 145.93 70.556
## - wt    1   19.7963 153.80 72.237
## - hp    1   22.7935 156.79 72.855
##
## Step:  AIC=67
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1    0.9672 139.99 65.227
## - cyl   2   10.4247 149.45 65.319
## - disp  1    1.5483 140.57 65.359
## - vs    1    2.1829 141.21 65.503
## - qsec  1    3.6324 142.66 65.830
## <none>              139.02 67.005
## - am    1   16.5665 155.59 68.608
## - hp    1   18.1768 157.20 68.937
## - wt    1   31.1896 170.21 71.482
##
## Step:  AIC=65.23
## mpg ~ cyl + disp + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1    1.2474 141.24 63.511
## - vs    1    2.3403 142.33 63.757
## - cyl   2   12.3267 152.32 63.927
## - qsec  1    3.1000 143.09 63.928
## <none>              139.99 65.227
## - hp    1   17.7382 157.73 67.044
## - am    1   19.4660 159.46 67.393
## - wt    1   30.7151 170.71 69.574
##
## Step:  AIC=63.51
## mpg ~ cyl + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - qsec  1     2.442 143.68 62.059
## - vs    1     2.744 143.98 62.126
## - cyl   2    18.580 159.82 63.466
## <none>              141.24 63.511
## - hp    1    18.184 159.42 65.386
## - am    1    18.885 160.12 65.527
## - wt    1    39.645 180.88 69.428
##
## Step:  AIC=62.06
## mpg ~ cyl + hp + wt + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     7.346 151.03 61.655
## <none>              143.68 62.059
```

```
## - cyl   2    25.284 168.96 63.246
## - am    1    16.443 160.12 63.527
## - hp    1    36.344 180.02 67.275
## - wt    1    41.088 184.77 68.108
##
## Step:  AIC=61.65
## mpg ~ cyl + hp + wt + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              151.03 61.655
## - am    1     9.752 160.78 61.657
## - cyl   2    29.265 180.29 63.323
## - hp    1    31.943 182.97 65.794
## - wt    1    46.173 197.20 68.191
```

Appendix C

```
#plotting residual vs fitted; the plot shows that it is not entirely random/homoskedatic, there seems t
par(mfrow= c(2,2))
plot(final)
```