Automatic versus Manual Cars Miles per Galon Performance Comparison

Regression Model Course Project by Liliana, June 5, 2017

Project Objective

This report explores the relationship between automatic and manual transmission cars in relationship with miles per gallon (MPG) (outcome) to answer the question (1) is automatic or manual transmission better for MPG? The report also (2) quantifies the MPG difference between automatic and manual transmissions and their importance in relationship with other variables such as weight (wt), number of cylinders (cyl) and horse power (hp) among others.

The mtcars data can be obtained from the R datasets package,as described at https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html)

The first two pages of this report present the main summary of the analysis and conclusions. For more detail information and results see the Appendix section.

Data Exploration

The mtcars data includes the following 11 variables: Miles per gallon (mpg), Number of cylinders (cyl), displacement in cu.in (disp), horsepower (hp), rear axle ration (drat), weight (wt), 1/4 mile time (qsec), transmission (am=0, for automatic, am=1 for manual), number of forward gears (gear), number of carburetors (carb) for 32 automobiles (1973-74 models).

Selected variables

In the ggpairs plots (see Fig A1 in the appendix), we can observe that the variables with highest correlation with mpg are (in order from highest to lowest): wt, cyl, disp, hp. In addition, the variables with the lowest correlations are (in order of lowest to highest): qsec, gear, am, carb. Figure 1 shows the analysis obtained from pairs of highest correlation variables with data for Automatic in red, and the data for Manual in blue.

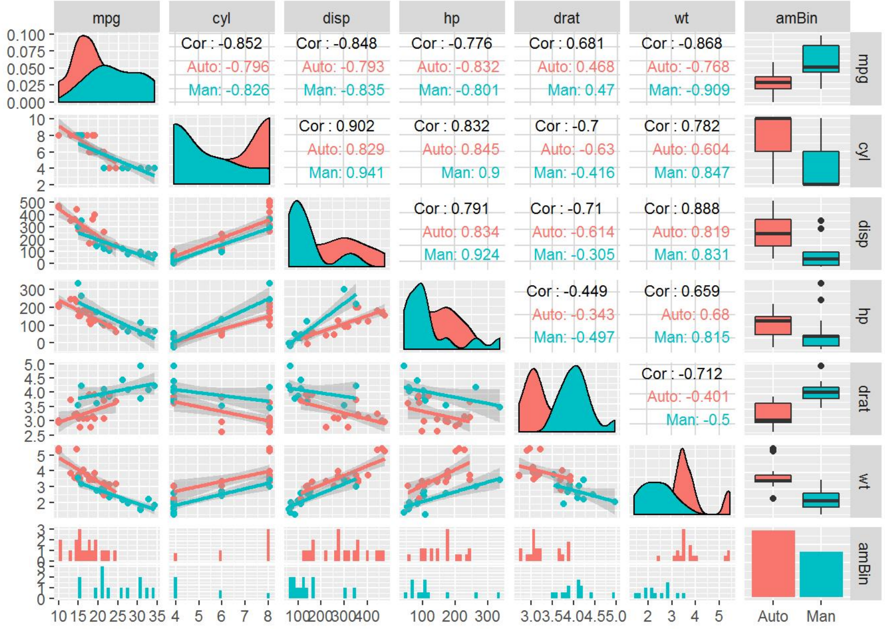


Fig. 1 Matrix describing the relation between pairs of variables, diagonal figures show density results. Upper triangle shows the correlations (black=all data, red=Automatic, blue=Manual)

Evaluating MPG Highest Correlation Variable Weight (wt) for Automatic and Manual

Figure 2 below shows mpg against weight (wt) plot with linear fits for three cases: 1. All cars (black line), 2. Automatic cars only (red line), 3. Manual cars only (blue line). A first look at the line fits (intersect and slope) suggest an “apparent” advantage of Manual compared to Automatic. However, further evaluations show that this difference is not significant and that one can NOT make a conclusion that Manual is better than Automatic or vice versa.

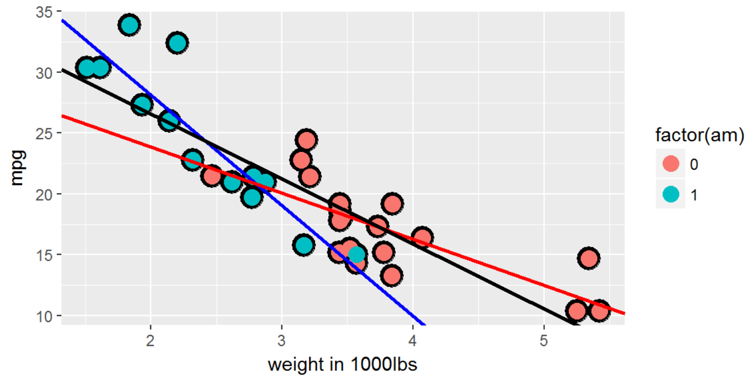


Fig. 2 Evaluating fit lm(mpg~wt) for All cars (black line), Automatic(red line), or Manual (blue line)

Models Evaluation

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + cyl
## Model 3: mpg ~ wt + cyl + am
## Model 4: mpg ~ wt + cyl + am + drat
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)   
## 1     30 278.32                                 
## 2     29 191.17  1    87.150 12.3199 0.001592 **
## 3     28 191.05  1     0.125  0.0177 0.895273   
## 4     27 191.00  1     0.051  0.0072 0.932807   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

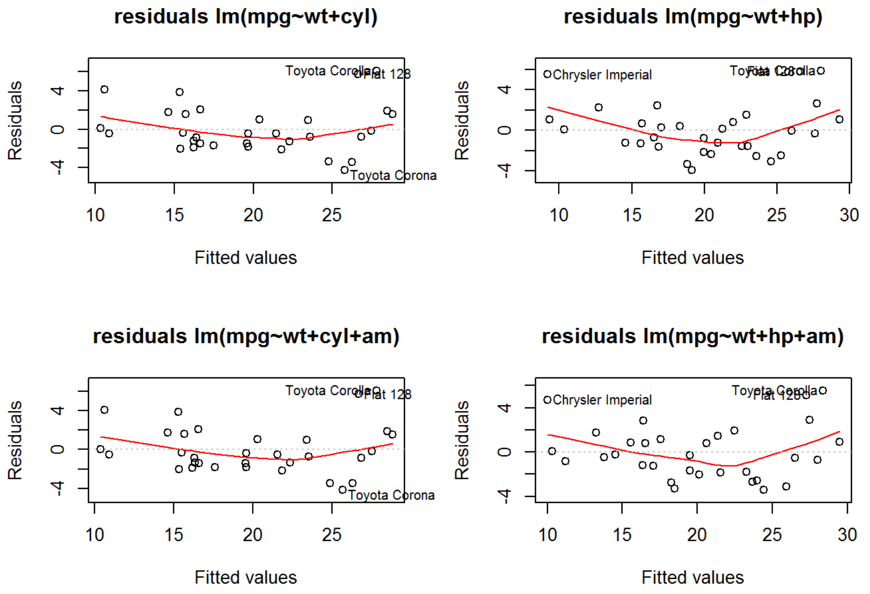


Figure 3. Models Residual Comparison (wt+cyl), (wt+hp),(wt+cyl+am), (wt+hp+am)

From the residuals results we can say that the residuals fit for (wt+cyl) has a smoother fit line compared to that of (wt+hp). In addition when the (am) variable is added (lower plots), there is not significant improvement to the residuals fit.

Executive Summary

At first when evaluating MPG with the variable (wt) for all cars, and the cars separated in automatic and manual, the line fit comparisons seem to suggest an “apparent” advantage of Manual compared to Automatic. However, further evaluations show that this difference is not significant when (am) was included as a second variable (p=0.895 >0.05). After performing multiple model comparisons, it was shown that when horse power or number of cylinders are added separately to weight, significantly contribute (p=0.00159 <0.05) to the linear model. On the other hand, adding (am) had no model improvements but inflated the model’s variance, as shown in the ivf analysis in the appendix section. In summary, for the 1974 Motor Trend data, one can NOT make a conclusion that Manual is better than Automatic or vice versa for MPG. This can also be observed in Figure 2, in the regions where Automatic and Manual overlap (in weight for example), there is no significant difference between the red and blue points in the MPG values.

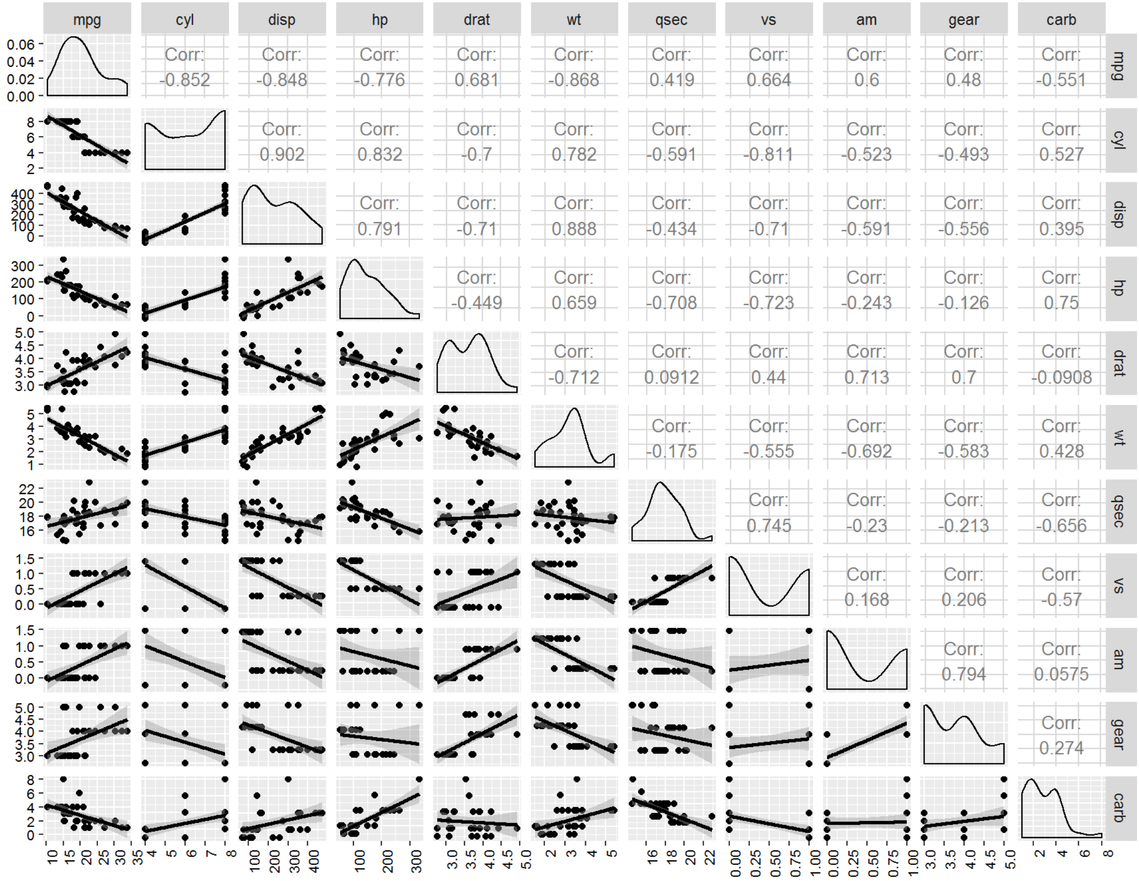
Appendix

Pages below include the code and all the analysis results that led to the summary and conclusions. ##Exploratory

```
require(datasets); data(mtcars);require(GGally);require(ggplot2);library(rgl);library(car); library(knitr)
```

Figure A1 below shows the first data exploration done to select the highest correlation variables (in order from highest to lowest): wt, cyl, disp, hp.

```
#boxplot(mtcars);#names(mtcars);#summary(mtcars);#hist(mtcars$am)
g <- ggpairs(mtcars, lower = list(continuous = "smooth",alpha = 0.2, size=0.1),params = NULL, upper = list(contin
uous = wrap("cor", size = 3.5, alignPercent = 1)))
g<-g+ theme(axis.text.x = element_text(size=8, angle = 90, vjust = 1, color = "black"), axis.text.y = element_tex
t(size=8, color = "black"))
g
```



Creating a categorical variable based on Automatic Manual

```
mtcars$amBin = ifelse(mtcars$am >0.2, "Man", "Auto")
mtcars1<-mtcars[,c(-7,-8,-9,-10,-11)]
g <- ggpairs(mtcars1, lower = list(continuous = "smooth"), params = NULL, mapping = ggplot2::aes(colour=amBin), u
pper = list(continuous = wrap("cor", size = 4, alignPercent = 1)))
#g #Same plot as in Figure 1
```

Comparison between All data, Automatic, and Manual Line Fits

```
###Plot the data coloring Automatic=0 (red) and Manual=1 (blue)
g1 = ggplot(mtcars, aes(x = wt, y = mpg, colour = factor(am)))
g1 = g1 + geom_point(size = 6, colour = "black") + geom_point(size = 4)
g1 = g1 + xlab("weight in 1000lbs") + ylab("mpg")
fitwt = lm(mpg ~ wt, data = mtcars)
fitMwt <-lm(mpg[mtcars$am==1]~wt[mtcars$am==1], mtcars)
fitAwt<-lm(mpg[mtcars$am==0]~wt[mtcars$am==0],mtcars)
####Plotting based on wt
g2 = g1
g2 = g2 + geom_abline(intercept = coef(fitMwt)[1], slope = coef(fitMwt)[2], size = 1, colour = "blue")
g2 = g2 + geom_abline(intercept = coef(fitAwt)[1], slope = coef(fitAwt)[2], size = 1, colour = "red")
g2 = g2 + geom_abline(intercept = coef(fitwt)[1], slope = coef(fitwt)[2], size = 1)
#g2 # Same plot as in Figure 2
```

Model Comparison

```
###Combining wt with disp
fitwt <-lm(mpg ~ wt, mtcars)
fitwtdisp <-lm(mpg~wt+disp, mtcars)
fitwtdispam <-lm(mpg~wt+disp+am, mtcars)
###combining wt with hp
fitwthp <-lm(mpg~wt+hp, mtcars)
fitwthpam <-lm(mpg~wt+hp+am, mtcars)
fitwthpamdrat <-lm(mpg~wt+hp+am+drat, mtcars)
###combining wt with cylinders
fitwtcyl <-lm(mpg~wt+cyl, mtcars)
fitwtcylam <-lm(mpg~wt+cyl+am, mtcars)
fitwtcylamdrat <-lm(mpg~wt+cyl+am+drat, mtcars)
fitwtcyldrat <-lm(mpg~wt+cyl+drat, mtcars)
fitwtcylhp<-lm(mpg~wt+cyl+hp, mtcars)
###Model Comparison
anova(fitwt,fitwtcyl,fitwtcylam,fitwtcylamdrat)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + cyl
## Model 3: mpg ~ wt + cyl + am
## Model 4: mpg ~ wt + cyl + am + drat
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 278.32
## 2     29 191.17  1    87.150 12.3199 0.001592 **
## 3     28 191.05  1     0.125  0.0177 0.895273
## 4     27 191.00  1     0.051  0.0072 0.932807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fitwt,fitwthp,fitwthpam,fitwthpamdrat)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + hp
## Model 3: mpg ~ wt + hp + am
## Model 4: mpg ~ wt + hp + am + drat
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 278.32
## 2     29 195.05  1    83.274 12.7054 0.001383 **
## 3     28 180.29  1    14.757  2.2515 0.145092
## 4     27 176.97  1     3.326  0.5075 0.482341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fitwt,fitwtdisp,fitwtdispam)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + disp
## Model 3: mpg ~ wt + disp + am
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     30 278.32
## 2     29 246.68  1    31.639 3.5931 0.06839 .
## 3     28 246.56  1     0.126 0.0143 0.90555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fitwt,fitwtcyl,fitwtcylhp)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + cyl
## Model 3: mpg ~ wt + cyl + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 191.17  1    87.150 13.8161 0.0008926 ***
## 3     28 176.62  1    14.551  2.3069 0.1400152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing the residuals between models with: lm(mpg~wt+cyl+am) (same as Figure 3)

```
#par(mfrow=c(2,2))
#plot(fitwtcylam, which=1, title(main="residuals lm(mpg~wt+cyl+am)"))
#plot(fitwthpam, which=1, title(main="residuals lm(mpg~wt+hp+am)"))
#plot(fitwtcyl, which=1, title(main="residuals lm(mpg~wt+cyl)"))
#plot(fitwthp, which=1, title(main="residuals lm(mpg~wt+hp)"))
#dev.off()
```

Comparing the variance inflaction factor (vif), hatvalues, and dfbeta for each model

```
###Testing the variance inflaction factor when lm(mpg~wt+cyl+am)
kable(vif(fitwtcylam), digits=3)
```

wt	3.609
cyl	2.584
am	1.925

```
###Testing the variance inflaction factor when lm(mpg~wt+cyl), without (am)
kable(vif(fitwtcyl), digits=3)
```

wt	2.579
cyl	2.579

```
###Testing the variance inflaction factor when lm(mpg~wt+hp+am)
kable(vif(fitwthpam), digits=3)
```

wt	3.775
hp	2.088
am	2.271

```
###Testing the variance inflaction factor when lm(mpg~wt+hp), without (am)
kable(vif(fitwthp), digits=3)
```

wt	1.767
hp	1.767

```
###Testing hatvalues
c(max(hatvalues(fitwtcylam)),max(hatvalues(fitwthpam)),max(hatvalues(fitwtcyl)),max(hatvalues(fitwthp)))
```

```
## [1] 0.2803808 0.4121968 0.2421225 0.3942082
```

```
###Testing the dfbeta
c(max(dfbeta(fitwtcylam)),max(dfbeta(fitwthpam)),max(dfbeta(fitwtcyl)),max(dfbeta(fitwthp)))
```

```
## [1] 1.0050783 0.9552774 1.2829757 1.1746276
```