MPG Analysis for Manual and Automatic cars

Niranjan Agnihotri

August 20, 2017

Executive Summary

The following document uses the `mtcars` data set in order to study the effect of transmission type of the car with the mpg of the car. This analysis also tries to find out the effect of other variables over mpg performance of the car. The data set is available in the R data sets. After analyzing we conclude that, the manual cars give better mpg than the automatic cars.

The mpg of the manual cars is more by 7.244 units than the automatic cars. Thus manual cars can be said to be more efficient.

Processing

Converting some numeric variables into factors for the sake of building models.

```
df <- data.frame(mtcars)
df$cyl <- as.factor(df$cyl)
df$vs <- as.factor(df$vs)
df$am <- as.factor(df$am)
levels(df$am) <- c("Automatic", "Manual")
df$gear <- as.factor(df$gear)
df$carb <- as.factor(df$carb)
```

Exploratory analysis

To understand the correlation between the variables, we have Fig. 1 in the appendix.This shows us the interaction between all the variables within the dataset. The Fig. 2 in the appendix is the box whisker plot that shows the observed mpg values for automatic and manual variables.

```
d <- aggregate(mpg ~ am, data = df, FUN = mean)
```

Here we quantify that the mpg for manual cars is more than the mpg for automatic cars. The difference is

```
print(abs(d[[2]][1]-d[[2]][2]))
```

```
## [1] 7.244939
```

Fitting the model

We consider that ther is no effect of am variable on the mpg. This is the null hypothesis. Let's build a model with the am variable only and examine the coefficients.

```
fit0 <- lm(mpg ~ am, data = df)
print(summary(fit0))
```

```
##
## Call:
## lm(formula = mpg ~ am, data = df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

In the above model we try to gauge the meaning of coefficients. The intercept stands for the mean of the Automatic variant and if we add the intercept to amManual's coefficient, than we get the average of Manual variants. The t value for the automatic (intercept) and amManual are sufficiently away from zero. Thus we can conclude that the variable am has effect on the mpg of the cars. The F-statistic is also 16.86 which is large enough for rejecting the null hypothesis.

The above model captures about 36% of the variance given by the R squared stastic. We try to incorporate more predictors in order to capture more variance in the model.

Building models

```
fit0 <- lm(mpg ~ am, data = df)
fit1 <- lm(mpg ~ ., data = df)
fit2 <- lm(mpg ~ am + wt, df)
fit3 <- lm(mpg ~ am + cyl + disp + hp + wt, df)

anova(fit0, fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 3: mpg ~ am + wt
## Model 4: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS  Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     15 120.40  15    600.49 4.9874 0.001759 **
## 3     29 278.32 -14   -157.92 1.4053 0.260391
## 4     25 150.41   4    127.91 3.9838 0.021341 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model number 4 that we picked has predictors who have the following co relations with mpg which are high

am - 0.9
cyl - -.852
disp - -.848
hp - -.776
wt - -.868

These values are taken from the Appendix Fig. 1 We can verify the same from the AIC score of each model below.

Uncertainity in the best model

The confidence intervals for fit3 are as follows. The results will variy by the following variations in the coefficients.

```
confint(fit3)
```

```
##                   2.5 %       97.5 %
## (Intercept) 28.31296354 39.415588581
## amManual    -1.12066818  4.732867169
## cyl6        -6.16171468 -0.110418430
## cyl8        -8.68663174  3.251069157
## disp        -0.02220684  0.030382623
## hp          -0.06127916 -0.003681192
## wt          -5.16066572 -0.316723497
```

Visualize a model

We build a sample model (fit2) and visualize it. In appendix Figure 4 we visualize it.

AIC scores

Aikaike Information Criterium A lower AIC implies a better model.

```
print(AIC(fit0))
```

```
## [1] 196.4844
```

```
print(AIC(fit1))
```

```
## [1] 169.2155
```

```
print(AIC(fit2))
```

```
## [1] 168.0292
```

```
print(AIC(fit3))
```

```
## [1] 156.3359
```

The mode mod3 has the lowest AIC and thus it's the best of models.

Appendix

Fig 1

To get insights about the interaction between different variables

```
mtcars$cyl <- as.numeric(mtcars$cyl)
ggpairs(mtcars)
```

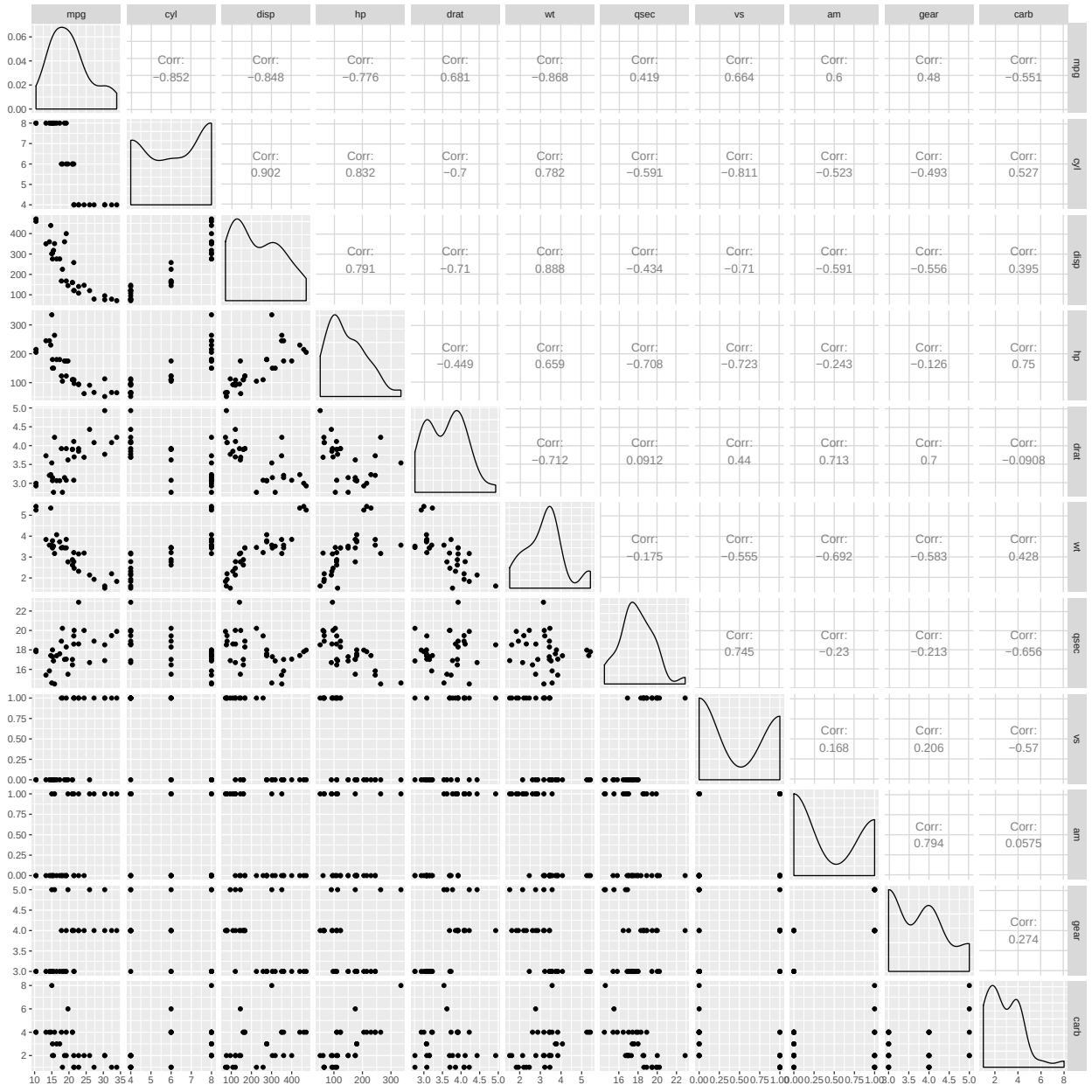

Fig 2

Box plot to find the mpg for automatic and manual cars

```
boxplot(mpg ~ am, data = mtcars, names = c("Automatic", "Manual"))
```

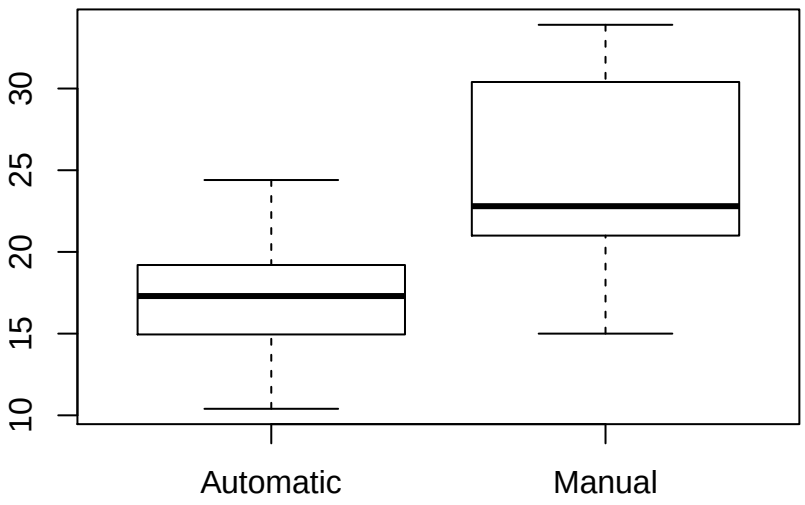

Fig 4

```
g <- ggplot(data = df, aes(wt, mpg, color=am))
g + geom_point() + stat_smooth(method = "lm", col = "blue")
```

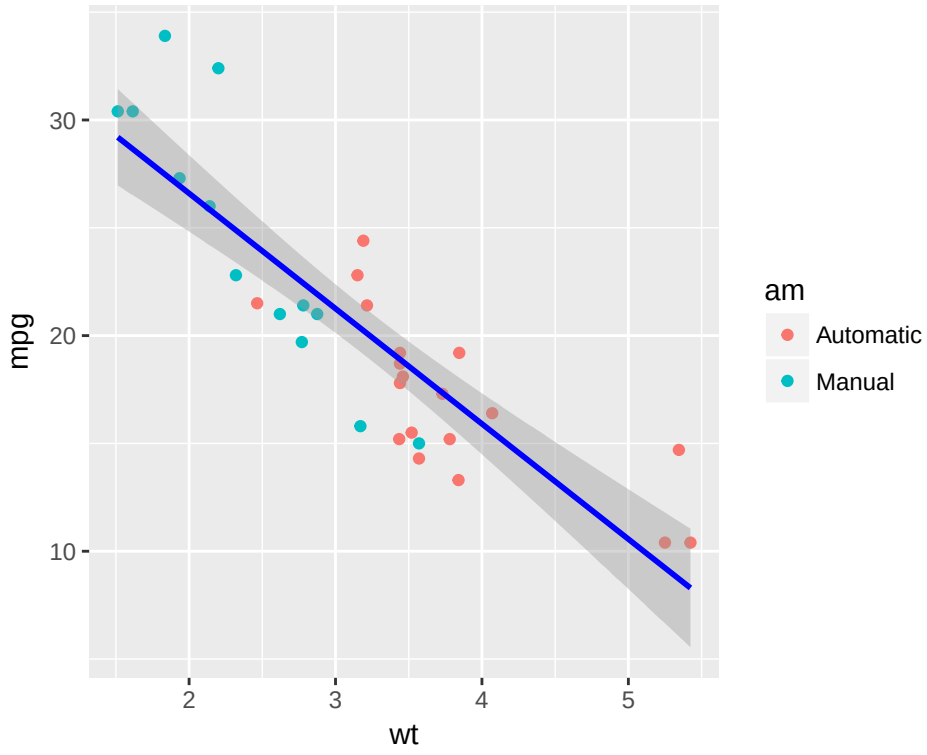

Fig 3

```
par(mfrow=c(2,2))
plot(fit3)
```