Your Mileage Might Vary

Frog

Monday, August 04, 2014

Executive Summary

This report analyzes the mtcars database of cars to determine if according to that database the transmission has a direct effect on the miles per gallon (mpg) that the car can travel. We will examine the dataset and try to see if there is a direct relationship between mpg and the type of transmission used in the card.

My first step would be to check for cupholders but I'd better check the correlation between mpg and the other variables in the dataset.

```
library(ggplot2)
data(mtcars)
cor(mtcars$mpg,mtcars)
```

```
##      mpg     cyl    disp      hp   drat      wt   qsec    vs     am   gear
## [1,]   1 -0.8522 -0.8476 -0.7762 0.6812 -0.8677 0.4187 0.664 0.5998 0.4803
##         carb
## [1,] -0.5509
```

From the correlation we observe that wt (weight) is the feature with higher correlation to mpg followed by cyl (cylinders). This makes sense as the heavier cars consume more than the light ones. I will name this observation #1

The first thing to check is if the am feature is skewed, if there are only a few cars of either class our conclusions can be very wrong.

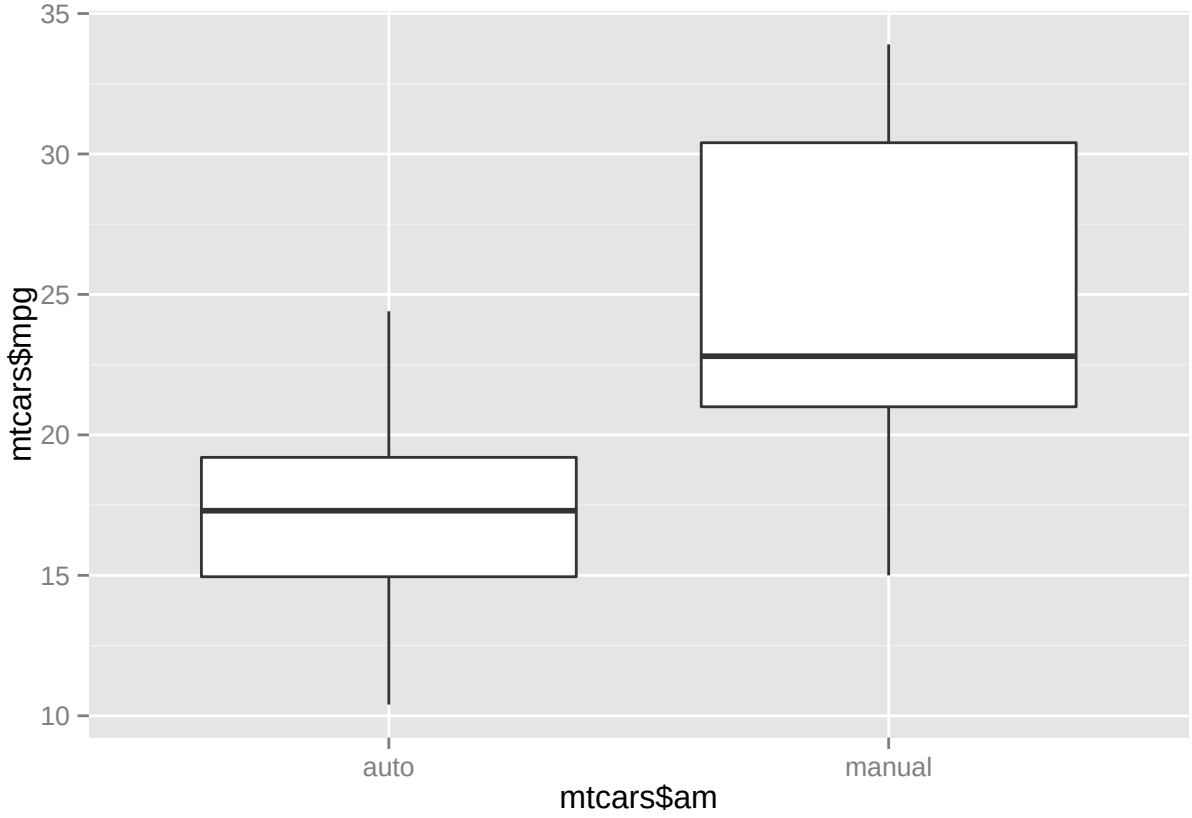
```
mtcars$am<-as.factor(mtcars$am)
levels(mtcars$am)<-c("auto","manual")
summary(mtcars$am)
```

```
##   auto manual
##     19     13
```

With 19 automatic cars and 13 manual cars in the dataset there's no skewing.

Now I plot the mpg by the type of transmission

```
qplot(mtcars$am,mtcars$mpg,geom="boxplot")
```



This plot shows that manual cars seem to have better mpg than automatic cars. I can verify this by a simple linear model.

```
reg1<-lm(mtcars$mpg~mtcars$am)
summary(reg1)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$am)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        17.15       1.12   15.25  1.1e-15 ***
## mtcars$ammanual     7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

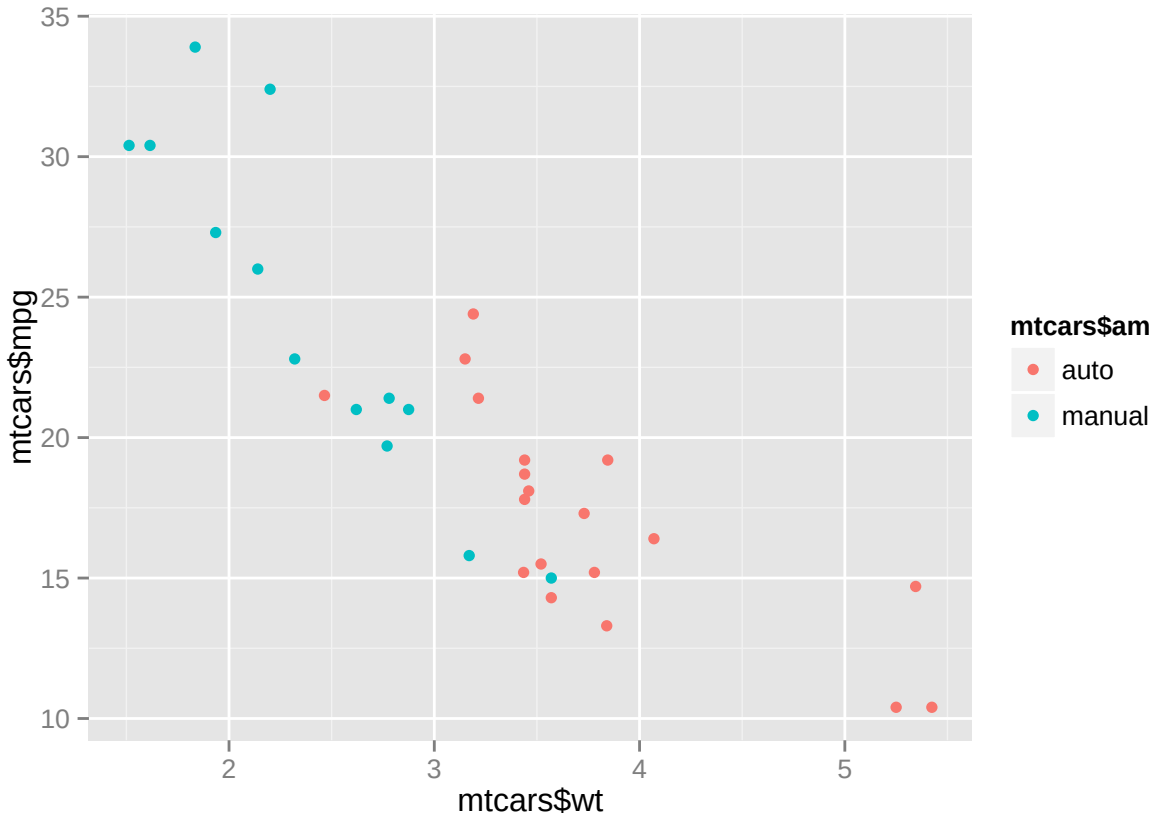
This model suggests that manual cars do 7.245mpg better than automatic cars.

From observation #1 now I would like to check if there is a correlation between weight and the car being automatic or not.

```
cor(as.numeric(mtcars$am),mtcars$wt)
```

```
## [1] -0.6925
```

```
qplot(mtcars$wt,mtcars$mpg,color=mtcars$am)
```



We observe there's a strong correlation between weight and transmission. Manual cars are lighter than automatic cars. This leads to my second model of regression trying this time both weight and transmission as predictors for mpg.

```
reg2<-lm(mtcars$mpg~as.numeric(mtcars$am) + mtcars$wt)
summary(reg2)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ as.numeric(mtcars$am) + mtcars$wt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.530 -2.362 -0.132  1.403  6.878
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            37.3452     4.3692    8.55  2.0e-09 ***
## as.numeric(mtcars$am)  -0.0236     1.5456   -0.02     0.99
## mtcars$wt              -5.3528     0.7882   -6.79  1.9e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.1 on 29 degrees of freedom
## Multiple R-squared:  0.753,  Adjusted R-squared:  0.736
## F-statistic: 44.2 on 2 and 29 DF,  p-value: 1.58e-09
```

Aha! It seems that now weight has a -5.35 influence over mpg but transmission is 0.02, in other words neutral.

I create a third model adding the cylinders (from observation #1) to check if this holds.

```
reg3<-lm(mtcars$mpg~as.numeric(mtcars$am)+mtcars$wt+mtcars$cyl)
summary(reg3)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ as.numeric(mtcars$am) + mtcars$wt +
##     mtcars$cyl)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.173 -1.534 -0.539  1.586  6.081
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             39.241      3.722   10.54    3e-11 ***
## as.numeric(mtcars$am)    0.176      1.304    0.14   0.8933
## mtcars$wt               -3.125      0.911   -3.43   0.0019 **
## mtcars$cyl              -1.510      0.422   -3.58   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 28 degrees of freedom
## Multiple R-squared:  0.83,   Adjusted R-squared:  0.812
## F-statistic: 45.7 on 3 and 28 DF,  p-value: 6.51e-11
```

It seems the weight impacts the mpg in 3.12mpg per ton, the number of cylinders also has something to say but the transmission has a very low impact.

Now I compare the three models using anova which curiously has nothing to do with astronomy.

```
anova(reg1,reg2,reg3)
```

```
## Analysis of Variance Table
##
## Model 1: mtcars$mpg ~ mtcars$am
## Model 2: mtcars$mpg ~ as.numeric(mtcars$am) + mtcars$wt
## Model 3: mtcars$mpg ~ as.numeric(mtcars$am) + mtcars$wt + mtcars$cyl
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
```

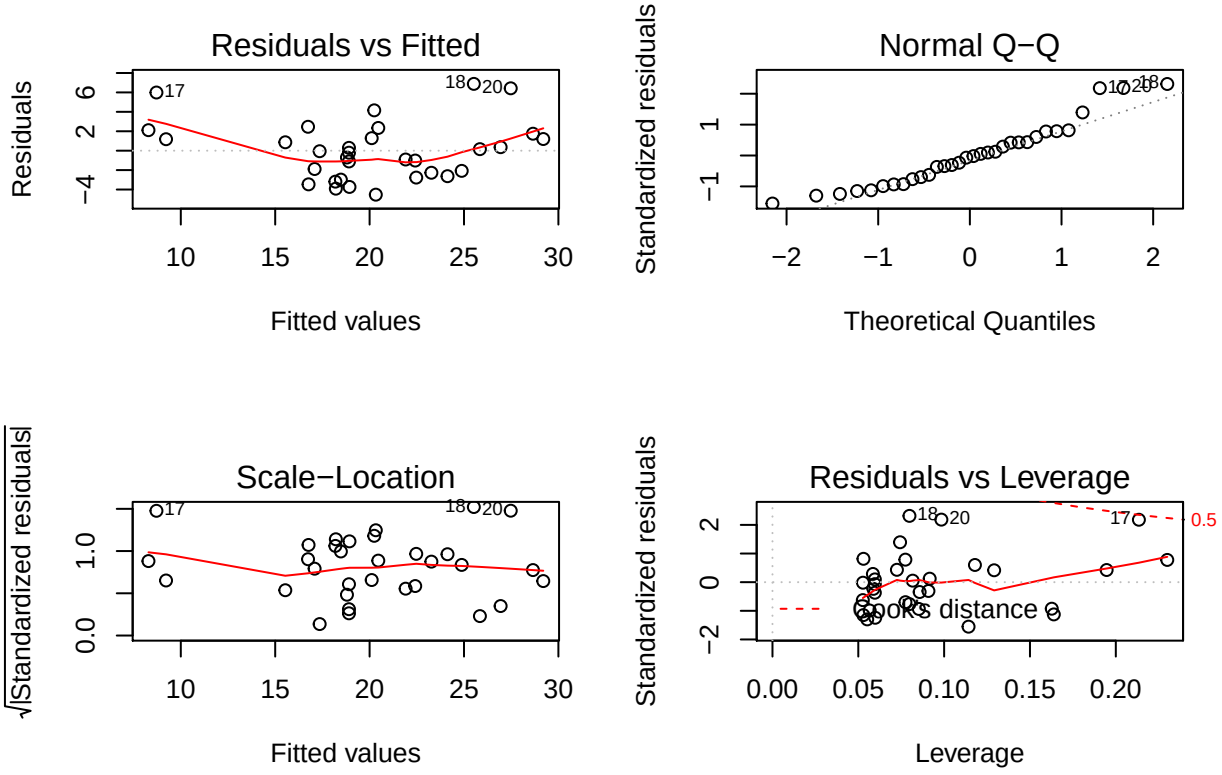
```
## 1     30 721
## 2     29 278  1       443 64.9 9.1e-09 ***
## 3     28 191  1        87 12.8  0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We observe model2 is better than 1 and 3, this doesn't mean that it is the best model for this dataset, we could have found that using fit.step but we are trying to buy a car and not a regression mode aren't we?

Since the second model seems to be the better one, let's diagnose its residuals.

Model 2: mpg ~ wt

```
par(mfrow=c(2,2))
plot(reg2)
```



Outliers! the world would be much better without them. We can see that even considering outliers there's a good fit for the residuals meaning our model is accurate.

Conclusion

With this dataset we can't conclude if they type of transmission affects the fuel consumption of a car, lighter cars have better mpg than heavier cars. Manual transmission cars tend to be lighter than automatic cars and that's why they show better fuel consumption.

Buy a light car it doesn't matter if it has a stick or not.

Finally: I know this went further than the 2 requiered pages and I could solve it moving the figures to appendixes but that would impact the clarity of the report so I prefer to lose a point than to lose my clarity. Thank you for understanding.