Regression Models Project

Executive Summary

Motor Trend, a magazine about the automobile industry is interested in exploring the relationship between a set of variables and miles per gallon (MPG). They are particularly interested in the following two questions:

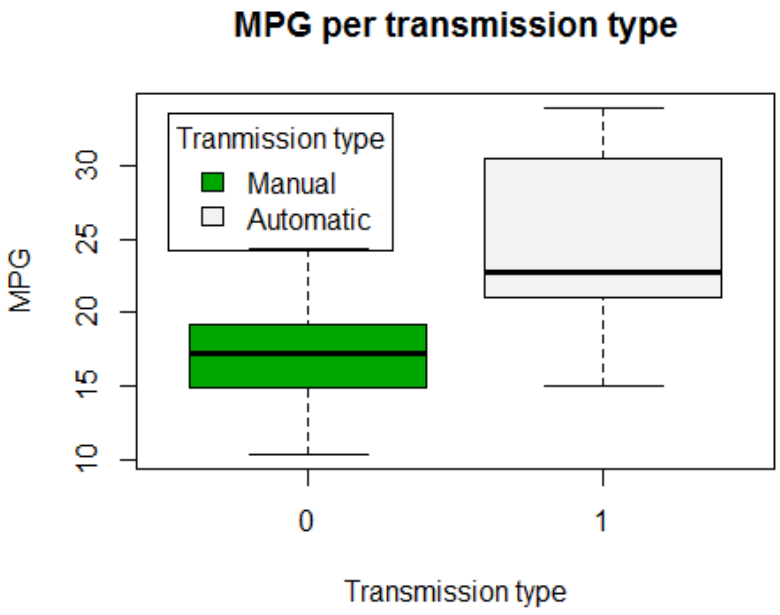
- Is an automatic or manual transmission better for MPG
- Quantifying how different is the MPG between automatic and manual transmissions?

We will do first some exploratory analysis and then come up with a model to answer this questions. An analysis of th residuals and errors will be also developed.

Is an automatic or manual transmission better for MPG

We are going to check hoy the mpg varies for the manual and automatic transmission. For that the boxplot is very useful:

```
data=mtcars
boxplot(data$mpg~data$am,main="MPG per transmission type",xlab="Transmission
type",ylab="MPG",col=terrain.colors(2))
legend("topleft",inset=0.05,title="Tranmission
type",c("Manual","Automatic"),fill=terrain.colors(2))
```



As we can see, the median and at least 50% of the points for each type lie separate, so we can say that manual transmission has less MPG than automatic in general. This is enough to answer the question for a generic car. This result makes sense because automatic transmission is designed not only to make life easier for the driver but also to improve the efficiency of the driving.

Quantifying how different is the MPG between automatic and manual transmissions?

Let's try to build a linear regression model to check how much mpg in average has each type of transmission. After some explatory analysis and backward feature selection we selected the following linear model:

```
#Results shown in appendix
#pairs(~mpg+qsec+wt+am,data)

#Backward feature selection. Hide the results because of the 2-page limit
model = step(lm(mpg~.,data),direction = "backward")
```

The features selected are and the confidence intervals for the model parameters are:

```
model$coefficient
```

```
## (Intercept)          wt        qsec          am
##       9.618      -3.917       1.226       2.936
```

```
confint(model,"am",level = 0.95)
```

```
##      2.5 % 97.5 %
## am 0.04573  5.826
```

The residuals are:

```
#Result shown in appendix due to space requirements
#plot(predict(model),resid(model))
#abline(h=mean(resid(model)),col="red")
```

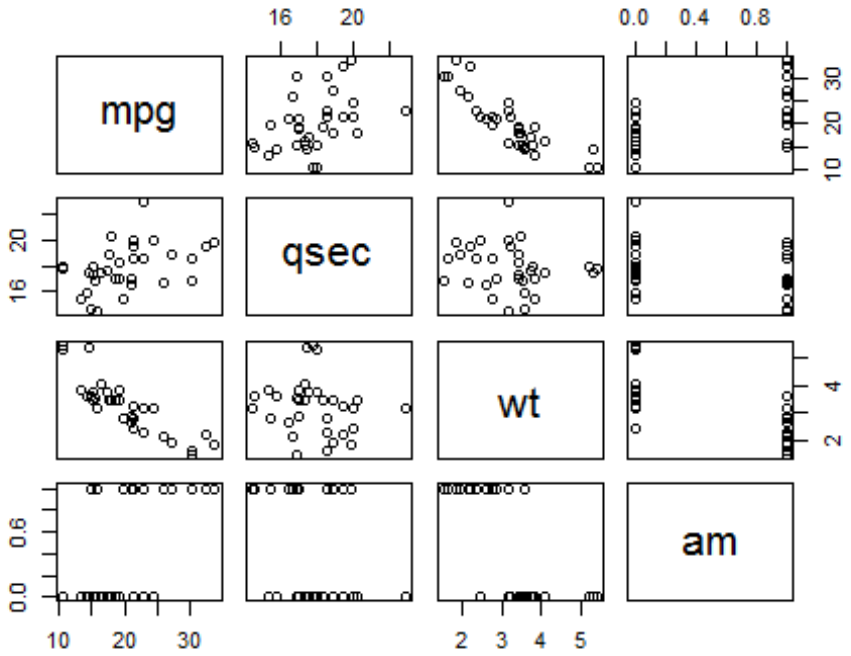
The residuals are unbiased and homoscedastic (see appendix).

In conlcusion: changing from manual to automatic increases the MPG in average 2.936 mpg, being therefore a more efficient transmission. One drawback is that because "am" is a binary variable the lm doesnt regress very good, being then necessary to include more variables. One effect is the loss in significance but a higher R^2.

Appendix

Exploratory analysis:

```
#Check linearity in variables that we want to use
pairs(~mpg+qsec+wt+am,data)
```



Summary of the model selection and model selected:

```
#Step by step computation and selection of features
step(lm(mpg~.,data),direction = "backward")
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq RSS  AIC
## - cyl   1      0.08 148 68.9
## - vs    1      0.16 148 68.9
## - carb  1      0.41 148 69.0
## - gear  1      1.35 149 69.2
## - drat  1      1.63 149 69.2
## - disp  1      3.92 151 69.7
## - hp    1      6.84 154 70.3
## - qsec  1      8.86 156 70.8
## <none>              148 70.9
## - am    1     10.55 158 71.1
## - wt    1     27.01 174 74.3
```

```
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq RSS  AIC
## - vs    1      0.27 148 67.0
## - carb  1      0.52 148 67.0
## - gear  1      1.82 149 67.3
## - drat  1      1.98 150 67.3
## - disp  1      3.90 152 67.7
## - hp    1      7.36 155 68.5
## <none>              148 68.9
## - qsec  1     10.09 158 69.0
## - am    1     11.84 159 69.4
## - wt    1     27.03 175 72.3
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq RSS  AIC
## - carb  1      0.69 148 65.1
## - gear  1      2.14 150 65.4
## - drat  1      2.21 150 65.4
## - disp  1      3.65 152 65.8
## - hp    1      7.11 155 66.5
## <none>              148 67.0
## - am    1     11.57 159 67.4
## - qsec  1     15.68 164 68.2
## - wt    1     27.38 175 70.4
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq RSS  AIC
## - gear  1       1.6 150 63.5
## - drat  1       1.9 150 63.5
## <none>              148 65.1
## - disp  1      10.1 159 65.2
## - am    1      12.3 161 65.7
## - hp    1      14.8 163 66.2
## - qsec  1      26.4 175 68.4
## - wt    1      69.1 218 75.3
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq RSS  AIC
## - drat  1       3.3 153 62.2
## - disp  1       8.5 159 63.2
## <none>              150 63.5
```

```
## - hp    1      13.3 163 64.2
## - am    1      20.0 170 65.5
## - qsec  1      25.6 176 66.5
## - wt    1      67.6 218 73.4
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq RSS  AIC
## - disp  1       6.6 160 61.5
## <none>              153 62.2
## - hp    1      12.6 166 62.7
## - qsec  1      26.5 180 65.3
## - am    1      32.2 186 66.3
## - wt    1      69.0 222 72.1
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq RSS  AIC
## - hp    1       9.2 169 61.3
## <none>              160 61.5
## - qsec  1      20.2 180 63.3
## - am    1      26.0 186 64.3
## - wt    1      78.5 239 72.3
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq RSS  AIC
## <none>              169 61.3
## - am    1      26.2 195 63.9
## - qsec  1     109.0 278 75.2
## - wt    1     183.3 353 82.8

##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = data)
##
## Coefficients:
## (Intercept)           wt         qsec           am
##        9.62        -3.92         1.23         2.94
```

```
#Summary of the model selected (higher R^2 but am not tht significant)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = data)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -3.481 -1.556 -0.726  1.411  4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## wt            -3.917      0.711   -5.51    7e-06 ***
## qsec           1.226      0.289    4.25  0.00022 ***
## am             2.936      1.411    2.08  0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
```

Analysis of the residuals:

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