Regression Models Course Project

Executive Summary

The goal of this analysis is to look at a data set of a collection of cars, explore the relationship between a set of variables and miles per gallon (MPG). Particularly following two questions:

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

In this analysis we will prove using linear models that there is a strong relation between MPG and Transmission and that the statistical difference is a 2.94 increase in MPG when we switch from automatic transmission to manual transmission.

Exploratory Analysis

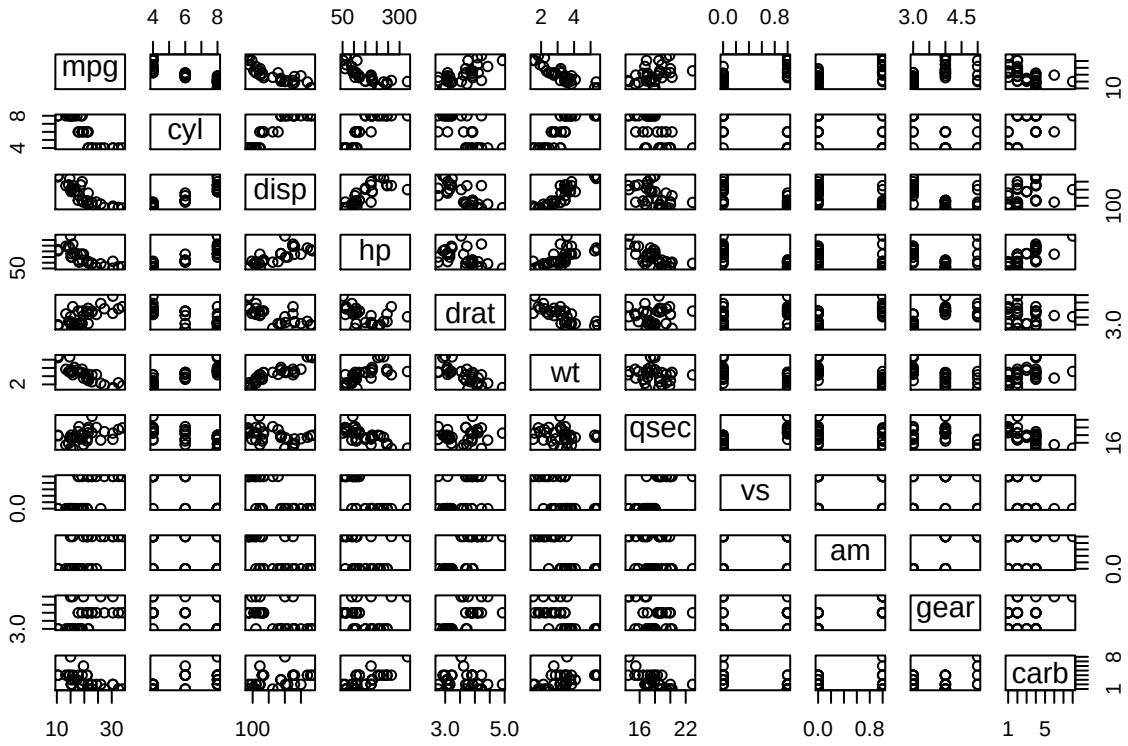
First of all, we load the data from the datasets package and perform some exploratory analysis to have an overall understanding of the data before we do some regressions. We can see here a summary of the datasets, composed of 32 observations and 11 variables.

```
library(datasets); data(mtcars); str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

We can also perform a "Pairs" plot to have an overall view of the relationship between the variables. The images are very small because there are many variables but we can see a rough shape of the relations:

```
pairs(mtcars)
```



Regression analysis

First we are going to fit a linear model between MPG and Transmission:

```
firstmdl <- lm(mpg ~ am, data = mtcars); summary(firstmdl)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

The results suggest that the relationship is highly significant with a P-value of 0.000285 The model indicates that for automatic transmission the mean value is 17.147 and for the manual transmission we have a 7.245 increase in Miles Per Gallon.

The next step will be to fit a linear model between MPG and al the variables:

```
allmdl <- lm(mpg ~ ., data = mtcars); summary(allmdl)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

In this case we see that we have a 2.52 difference between automatic and manual transmission. However, looking at the P-Value, we see that none of the variables have a value lower than our alpha: 0.05 We will simplify this model taking the variables with the lowest P-Values: wt, qsec, am and hp.

```
testmdl <- lm(mpg ~ wt + hp + qsec + am, data = mtcars); summary(testmdl)$coef
```

```
##                Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 17.44019110  9.3188688  1.871492 0.072149342
## wt          -3.23809682  0.8898986 -3.638726 0.001141407
## hp          -0.01764654  0.0141506 -1.247052 0.223087932
## qsec         0.81060254  0.4388703  1.847021 0.075731202
## am           2.92550394  1.3971471  2.093913 0.045790788
```

However, we can see in this model that the significance of hp is again very low, with a P-Value much greater that 0.05 So our final model will be with the variables: wt + qsec + am

```
finalmdl <- lm(mpg ~ wt + qsec + am, data = mtcars); summary(finalmdl)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Conclusion

We see in our analysis that MPG and Transmission are strongly related, and also that we have to take into account these other variables: wt and qsec. Our final model suggest a 2.94 increase in MPG when switching from automatic to manual transmission. We can perform an Analysis of Variance (ANOVA) to reassure if our final model is the best one.

```
anova(firstmdl, finalmdl,testmdl,allmdl)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
```

```
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ wt + hp + qsec + am
## Model 4: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 39.2687 8.025e-08 ***
## 3     27 160.07  1      9.22  1.3127    0.2648
## 4     21 147.49  6     12.57  0.2983    0.9308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and some plots of the residuals and normality tests and so on.

```
par(mfrow=c(2,2)); plot(finalmdl)
```