Regression Models Course Project

Executive Summary

I work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). 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?”

To answer these questions, first, I analysed the interrelationships between the variables of our data set.Then, I studied both the simpliest and the maximal model to find out there were not satisfying. So I built the minimal adequate model (without interactions).

Finally, with the adequate model, **it appears that manual transmission is quite better for MPG**. A fact that we might have thought about at the start because cars with automatic transmissions tend to be heavier and so have a greater consumption.

Preliminary Examination of the Data

First thing, we might do is looking for interrelationships between the variables of our data set. To that extent, we should look at a correlation matrix and also a scatterplot matrix.

```
cor(mtcars)
pairs(mtcars,col=(am*1+1)) # output shown in the appendix
```

Now we have plenty to look at. As a reminder, we want to model (or predict) MPG from the other variables. And as you can see there, we have some interesting relationships. 'MPG' appears to be moderately to strongly related to 'cyl' (negatively), 'disp'(negatively), “hp” (negatively), 'drat' (positively), “wt” (neagtively), 'vs' (positevely) and 'am' (positevely). Besides, we can see a clear distinction between “am” and “non am” vehicles.

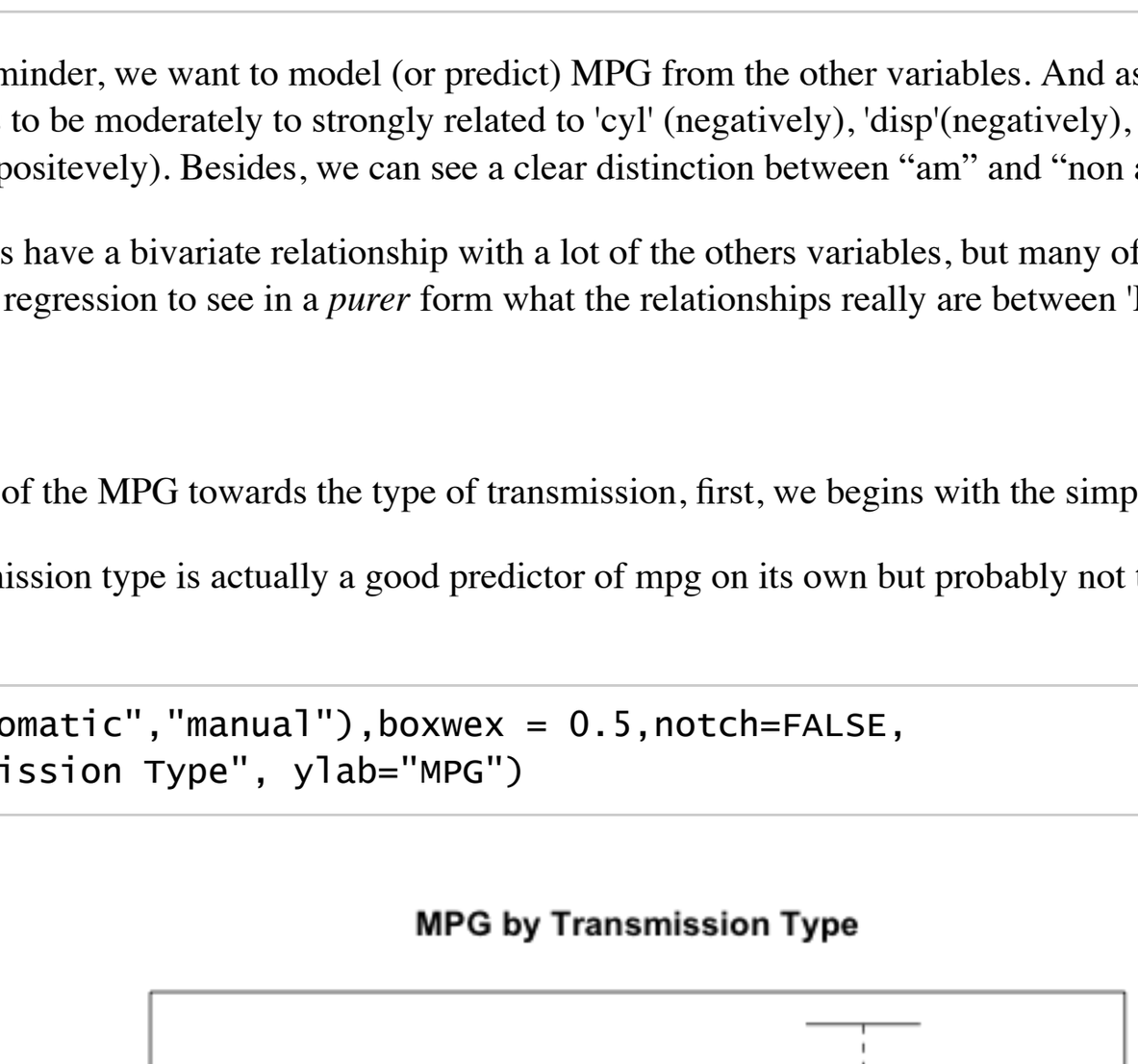
At this point of our analysis, 'MPG' does have a bivariate relationship with a lot of the others variables, but many of those variables are also related to each other. Now we are going to do multiple regression to see in a *purer* form what the relationships really are between 'MPG' and the others variables.

The Minimal Model

As we are interested in the dependency of the MPG towards the type of transmission, first, we begins with the simpliest model **mpg ~ am**

From this model, we see that the transmission type is actually a good predictor of mpg on its own but probably not the best predictor (the adjusted R-squared is not very large, 0.34).

```
boxplot(mpg~am,names=c("automatic","manual"),boxwex = 0.5,notch=FALSE,
        main="MPG by Transmission Type", ylab="MPG")
```



Moreover, when we do a boxplot, there is a clear distinction between 'am' and 'non am' groups.

The Maximal Model (without intercations)

We begin by throwing all the (possible) predictors into our model.

```
maximalModel <- lm(mpg ~factor(cyl)+disp+hp+drat+wt+qsec+vs+am+factor(gear)+factor(carb)-1); summary(maximalModel)
```

It appears that all 'MPG' is somewhat related to 'hp' and 'wt'. But in this **maximal model**, no predictors seems really significant. Now we need to start winnowing down our model to a minimal adequate one.

The Minimal Adequate Model (without intercations)

We want to reduce our model to a point where all the remaining predictors are significant, and we want to do this by throwing out one predictor at a time. However we want to keep 'am' in our model and keep it significant if possible to answer to our issue. To that extent, we use the **step()** function to automate the process.

```
minimalModel <- step(maximalModel,direction="backward")
summary(minimalModel)
summary(lm(formula = mpg ~ wt + qsec + am - 1))
finalModel <- lm(formula = mpg ~ wt + qsec + am - 1)
```

So, the **step()** function ends up with the following model: **mpg ~ disp + wt + qsec + am - 1** by minimizing the *Akaike information criterion*. But in this model, 'disp' appears non significant (p.value = 0.19) so I prefer the following model **mpg ~ wt + qsec + am - 1** where the AIC is quite the same (61.419 instead of 61.395) and all the predictors are significant. Moreover this change doesn't affect too much the adjusted R-squared.

When we plot the **final model**, we see no pattern in the residuals vs fitted values plot and the residuals are nearly normal and homoscedastic. So we decide to keep this model and don't add any interactions terms.

Conclusion

```
summary(finalModel)$coef
```

```
##      Estimate Std. Error t value  Pr(>|t|)
## wt     -3.185     0.4828  -6.598 3.129e-07
## qsec    1.600     0.1021  15.665 1.092e-15
## am      4.300     1.0241   4.198 2.329e-04
```

From our modelisation, we can infer that **manual transmission is better for MPG** (as we have seen in our minimal model and the boxplot). Indeed, **you can travel 4.3 miles per gallon more with a manual transmission**.

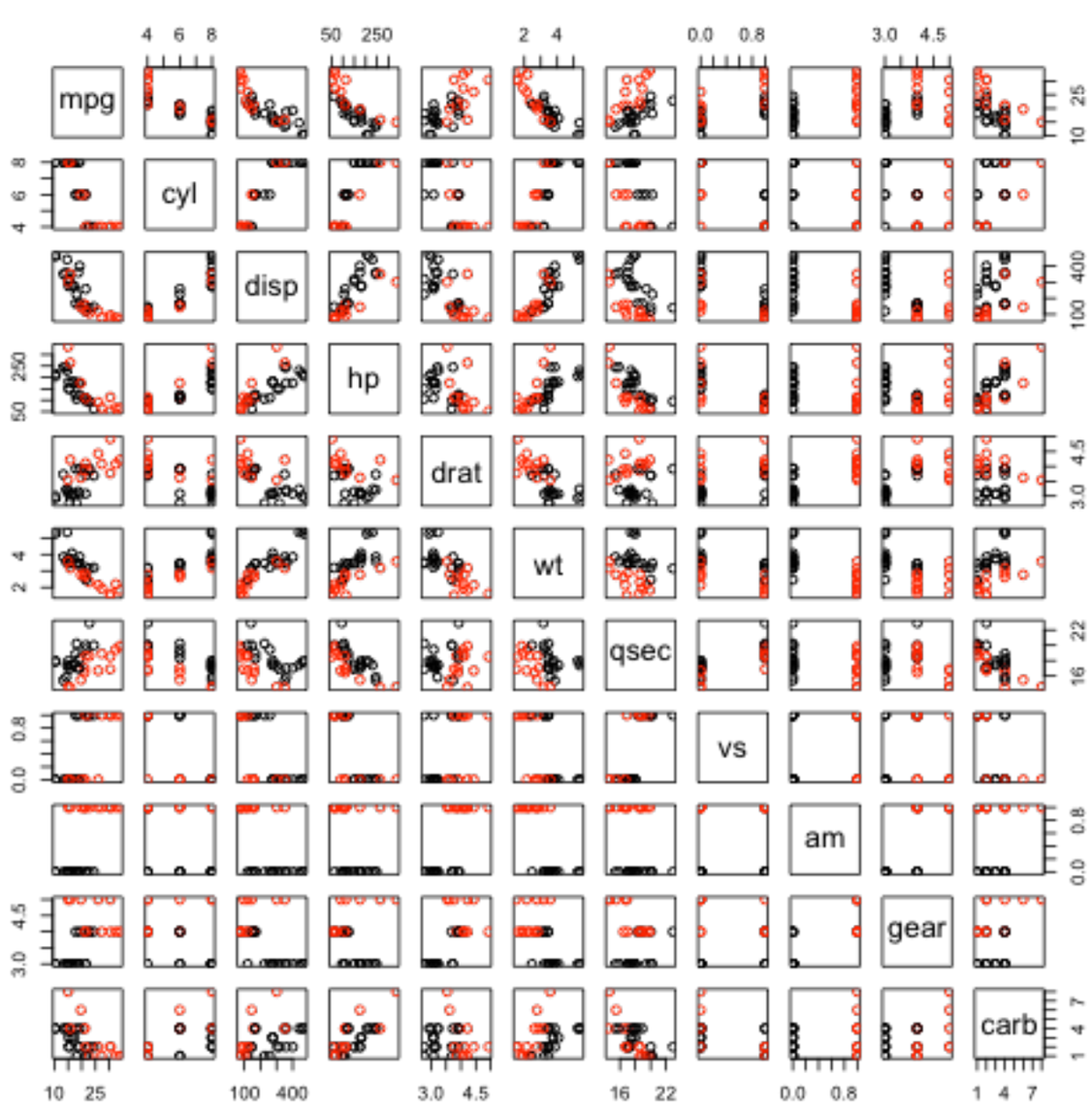
Appendix

Preliminary Examination of the Data

```
cor(mtcars)
```

```
##             mpg     cyl    disp      hp     drat      wt    qsec      vs
## mpg    1.0000 -0.8522 -0.8476 -0.7762  0.68117 -0.8677  0.4187  0.6640
## cyl   -0.8522  1.0000  0.9020  0.8324 -0.69994  0.7825 -0.5912 -0.8108
## disp  -0.8476  0.9020  1.0000  0.7909 -0.71021  0.8880 -0.4337 -0.7104
## hp    -0.7762  0.8324  0.7909  1.0000 -0.44876  0.6587 -0.7082 -0.7231
## drat   0.6812 -0.6999 -0.7102 -0.4488  1.00000 -0.7124  0.0912  0.4403
## wt    -0.8677  0.7825  0.8880  0.6587 -0.71244  1.0000 -0.1747 -0.5549
## qsec   0.4187 -0.5912 -0.4337 -0.7082  0.09120 -0.1747  1.0000  0.7445
## vs     0.6640 -0.8108 -0.7104 -0.7231  0.44028 -0.5549  0.7445  1.0000
## am     0.5998 -0.5226 -0.5912 -0.2432  0.71271 -0.6925 -0.2299  0.1683
## gear   0.4803 -0.4927 -0.5556 -0.1257  0.69961 -0.5833 -0.2127  0.2060
## carb  -0.5509  0.5270  0.3950  0.7498 -0.09079  0.4276 -0.6562 -0.5696
##              am    gear     carb
## mpg    0.59983  0.4803 -0.55093
## cyl   -0.52261 -0.4927  0.52699
## disp  -0.59123 -0.5556  0.39498
## hp    -0.24320 -0.1257  0.74981
## drat   0.71271  0.6996 -0.09079
## wt    -0.69250 -0.5833  0.42761
## qsec  -0.22986 -0.2127 -0.65625
## vs     0.16835  0.2060 -0.56961
## am     1.00000  0.7941  0.05753
## gear   0.79406  1.0000  0.27407
## carb   0.05753  0.2741  1.00000
```

```
pairs(mtcars,col=(am*1+1)) # output shown in the appendix
```



The Minimal Model

```
simply <- lm(mpg~am); summary(simply)
```

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## am              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
```

The Maximal Model (without intercations)

```
summary(maximalModel)$coef
```

```
##               Estimate Std. Error t value Pr(>|t|)
## factor(cyl)4  23.87913   20.06582  1.1900  0.25253
## factor(cyl)6  21.23044   18.33416  1.1580  0.26498
## factor(cyl)8  23.54297   18.22250  1.2920  0.21592
## disp           0.03555    0.03190  1.1143  0.28267
## hp            -0.07051    0.03943 -1.7884  0.09393
## drat           1.18283    2.48348  0.4763  0.64074
## wt            -4.52978    2.53875 -1.7843  0.09462
## qsec           0.36784    0.93540  0.3933  0.69967
## vs             1.93085    2.87126  0.6725  0.51151
## am             1.21212    3.21355  0.3772  0.71132
## factor(gear)4  1.11435    3.79952  0.2933  0.77332
## factor(gear)5  2.52840    3.73636  0.6767  0.50890
## factor(carb)2 -0.97935    2.31797 -0.4225  0.67865
## factor(carb)3  2.99964    4.29355  0.6986  0.49547
## factor(carb)4  1.09142    4.44962  0.2453  0.80956
## factor(carb)6  4.47757    6.38406  0.7014  0.49381
## factor(carb)8  7.25041    8.36057  0.8672  0.39948
```

The Minimal Adequate Model (without intercations)

```
summary(minimalModel)
```

```
##
## Call:
## lm(formula = mpg ~ disp + wt + qsec + am - 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.717 -1.464 -0.538  1.783  4.357
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## disp  0.01202    0.00889    1.35  0.18724
## wt   -4.61279    1.15817   -3.98  0.00044 ***
## qsec  1.70551    0.12749   13.38  1.1e-13 ***
## am    4.18085    1.01362    4.12  0.00030 ***
## ---
## 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:  0.988,  Adjusted R-squared:  0.986
## F-statistic:  572 on 4 and 28 DF,  p-value: <2e-16
```

```
summary(finalModel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am - 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.882 -1.540 -0.425  1.662  4.171
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## wt     -3.185      0.483    -6.6  3.1e-07 ***
## qsec    1.600      0.102    15.7  1.1e-15 ***
## am      4.300      1.024     4.2  0.00023 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.5 on 29 degrees of freedom
## Multiple R-squared:  0.987,  Adjusted R-squared:  0.986
## F-statistic:  741 on 3 and 29 DF,  p-value: <2e-16
```

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