Regression Models Course Project

Alicia Rodriguez

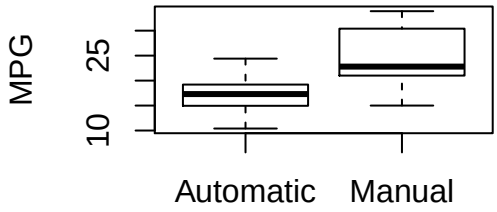
Executive Summary

This report explores the relationship between a set of variables related to cars and miles per gallon (MPG). More precisely, the focus is on the following questions:

- Is an automatic or manual transmission (represented by the variable *am*) better for MPG?
- Quantify the MPG difference between automatic and manual transmissions.

Exploratory Analysis

After transforming variables *cyl*, *vs*, *am*, *gear*, and *carb* to factors (see the data structure in the appendix), and taking a quick look at the relationship of variable MPG with respect to *am*, we could expect that manual transmission ($am = 1$) leads to a higher MPG. Let's test this hypothesis and check in which cases holds true.



Modelling MPG

Let's fit a first model of MPG with all available predictors (see appendix for the specific results of $mpg \sim .$). No significant coefficient is obtained (all p-values>0.05), thus we fail to reject the null hypothesis that any of the transmissions is better than the another one for MPG (and the same applies to the rest of predictors). So let's try different models to see the simplest one (with least predictors involved so as to decrease the standard error) that fits reasonably well *mpg* (i.e., with low bias), where the coefficients are significant to support our hypothesis.

We start by simply fitting by the predictor of interest, *mpg am*, an also by each of the rest of the available predictors, in order to compare the resulting R^2. Although R^2 is not a perfect measurement of how good the fit is, it can serve as a first approximation on which predictors provide more information so as to predict *mpg*.

```
##       am am+cyl am+disp am+hp am+drat am+wt am+qsec am+vs am+gear am+carb
## R^2 0.36  0.765   0.733 0.782   0.491 0.753   0.687 0.686   0.494   0.722
```

From the table we see that the variables that provide a higher R^2 are *hp*, *cyl*, *wt*, *disp* and *carb*, in that order. So, let's fit models adding each of these predictors at a time, and comparing the resulting models.

Taking a look at the results (see appendix for the results of the anova comparison), we are going to consider the predictors *hp*, *cyl*, and *wt*, besides *am*, since the new added variable at each model provides significant results (p-values<0.05). Let's analyze what happens more in detail with the inclusion of each one (centered at their mean for the cases of *hp* and *wt* so that the interpretation is more intuitive):

```
##                    am        am+hp     am+hp+cyl am+hp+cyl+wt
## Res. Std. Err.     "4.902"   "2.9092"  "2.7025"  "2.4101"
## Adjusted R2        "0.3385"  "0.767"   "0.7989"  "0.8401"
## intercept          "17.1474" "17.9468" "20.8059" "20.9653"
```

```
## coeff. am          "7.2449"  "5.2771"  "4.1579"  "1.8092"
## coeff. hp          "-"       "-0.0589" "-0.0442" "-0.0321"
## coeff. cyl6        "-"       "-"       "-3.9246" "-3.0313"
## coeff. cyl8        "-"       "-"       "-3.5334" "-2.1637"
## coeff. wt          "-"       "-"       "-"       "-2.4968"
## Pr(>|t|) intercept "0"       "0"       "0"       "0"
## Pr(>|t|) am        "3e-04"   "0"       "0.0027"  "0.2065"
## Pr(>|t|) wt        "-"       "-"       "-"       "0.0091"
```

In the previous table we can observe several things:

- As the number of predictors increases, the residual standard error decreases and the adjusted R^2 increases (as theoretically expected).
- Regarding the **intercept, its value changes when a factor predictor** (*cyl* in our case) is added. With *hp* and *wt* the value does not change much because these predictors are centered in their respective mean values.
- Regarding the *am* **coefficient (i.e., increase in** *mpg* **when manual transmission is used,** $am = 1$), we can see that its value decreases as we take into account more predictors. If only the use of manual or automatic transmission is considered ($mpg \sim am$), **using a manual transmission increases** *mpg* **an average of 7.2449 miles per galon, being a significant value (p-value<0.05)**. This value decreases when considering also horse power, *hp*, and number of cylinders, *cyl*, being still signficant values. However, **when a more influential predictor is added to the model, such as the vehicle weight,** *wt*, **the** *am* **coefficient is only of 1.8092 miles per galon and it is not a significant value** anymore.

We can perform a quick test, fitting the model $mpg \sim am + I(wt - mean(wt)))$ (see appendix for the results). We can see that an increase in one lbs in *wt* has an impact on *mpg* two orders of magnitude higher than the fact of using manual transmission. Therefore, *am* is not a significant predictor anymore (p-value=0.988>>0.05), and its coefficient even reverse the sign, which we know makes no sense.

Analyzing the residuals of the model $mpg \sim am + hp + cyl + wt$ (see last plots of the appendix), we can see that there is **no relationship between the residuals and fitted values**, thus suggesting that the model does not miss any important relationship. Besides, **there are some outliers for which the normality of the resiluals do not apply** (the ones in the upper and lower part of the Normal Q-Q plot), which coincide with the ones with higher standardized residuals in the leverage plot. **These points have low leverage in general**, but their errors do not follow normal distributions, thus leading to the slight slopes in the Residuals vs. Fitted and Scale-Location plots towards those points.

Conclusions

- *Is an automatic or manual transmission better for MPG?* Only considering automatic or manual **transmission as an isolated predictor for MPG**, we reject the null hypothesis that both of them lead to the same MPG with a p-value<0.05, observing that **using a manual transmission leads to a higher (thus, better) MPG**. However, **if other variables with high impact in MPG are considered**, such as the vehicle weight, we fail to reject the null hypothesis that both manual and automatic transmission lead to the same MPG, and we **could not surely determine which transmission is better for MPG**.
- *Quantify the MPG difference between automatic and manual transmissions.* Considering **manual transmission (ignoring the rest of predictors), it provides in average an increase of 7.245 MPG, with an standard error of 1.764** and a p-value=0.000285. However, as said before, **in case of considering other variables with high impact in MPG** such as the vehicle weight, **we do not observe an increase in MPG anymore, but a slight decrease of 0.02 MPG**. But, again, in presence of weight as predictor, we fail to reject the null hypothesis that automatic and manual transmissions really lead to different MPGs, as the rest of variables has a deeper impact on MPG.

Appendix

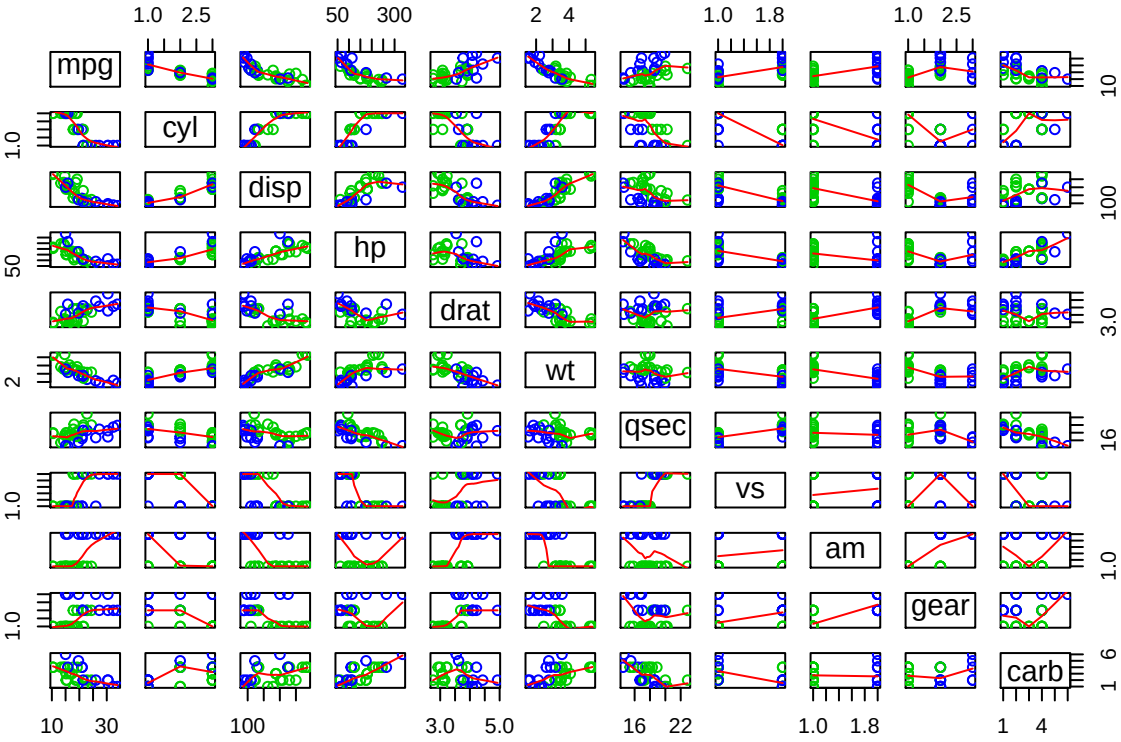
Exploratory Analysis

Data structure:

```
## '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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Let's quickly explore the correlations between each pair of variables. For the plot we can see that almost every variable when plotted against *mpg* shows a negative slope, except for *vs* and *am*.

```
pairs(mtcars,panel=panel.smooth,col=3+(mtcars$am==1))
```



Modelling MPG

We can expect to see these negative correlations in the overall regression:

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## am1          1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

Results of comparing different models adding one variable each time:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + I(hp - mean(hp))
## Model 3: mpg ~ am + I(hp - mean(hp)) + cyl
## Model 4: mpg ~ am + I(hp - mean(hp)) + cyl + I(wt - mean(wt))
## Model 5: mpg ~ am + I(hp - mean(hp)) + cyl + I(wt - mean(wt)) + I(disp -
##     mean(disp))
## Model 6: mpg ~ am + I(hp - mean(hp)) + cyl + I(wt - mean(wt)) + I(disp -
##     mean(disp)) + carb
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 69.0515 6.45e-08 ***
## 3     27 197.20  2     48.24  3.5030  0.04962 *
## 4     26 151.03  1     46.17  6.7058  0.01752 *
## 5     25 150.41  1      0.62  0.0896  0.76780
## 6     20 137.71  5     12.70  0.3688  0.86392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelling *mpg* as the a combination of *am* and a relevant predictor as *wt*. Notice the sign of *am* coefficient is reversed.

```
##
## Call:
## lm(formula = mpg ~ am + I(wt - mean(wt)), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      20.10022    0.83318  24.125  < 2e-16 ***
## am1              -0.02362    1.54565  -0.015    0.988
## I(wt - mean(wt)) -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

Residual plots for the model $mpg \sim am + hp + cyl + wt$.