Regression Models Project

Christian J. Lagares Nieves

Executive summary:

In this report we will look to evaluate the following question, does automatic transmission (among other variables) explain MPG in different car models?. In order to provide some light about this we will work employing the R package datasets. Within this package, we will summon the `mtcars` dataset that contains different car models characteristics. From this report, as we will son see, we can conclude that weight and cylinders are relevant and important variables and those characteristics, not transmission, explain MPG in different car models. For us to answer this question, we will use ANOVA which allows to determine if a variable can be dropped in a multivariated model. For aditional information on the Analysis of Variance Methodology please visit the following url https://www.calvin.edu/~scofield/courses/m145/materials/handouts/anova1And2.pdf. The information on ANOVA is made available to you through the Calvin College, Grand Rapids, Michigan.

1st Step: Load data and determine principal variables:

- `datasets` library provides royalte-free databases
- `mtcars` dataset provides information on a wide variety of car models and key characteristics.

```
library(datasets) #This library provides free databases
data(mtcars) #The database I will use
str(mtcars) #str displays variables names and displays basic information
```

There are several variables in this dataset. ANOVA (Analysis of Variance) will be used to determine whose variables are relevant apart from transmission type. I will proceed from general to particular, so I won't lose generality after dropping variables.

ANOVA

ANOVA explains the sources of variance so it can help to determine which variables have significant effects. According to Wikipedia: "In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal".

```
analysis <- aov(mpg ~ ., data = mtcars) #I run ANOVA
summary(analysis) #this returns a summary containing relevant statistics
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## cyl          1  817.7   817.7 116.425 5.03e-10 ***
## disp         1   37.6    37.6   5.353  0.03091 *  
## hp           1    9.4     9.4   1.334  0.26103    
## drat         1   16.5    16.5   2.345  0.14064    
## wt           1   77.5    77.5  11.031  0.00324 ** 
## qsec         1    3.9     3.9   0.562  0.46166    
## vs           1    0.1     0.1   0.018  0.89317    
## am           1   14.5    14.5   2.061  0.16586    
## gear         1    1.0     1.0   0.138  0.71365    
## carb         1    0.4     0.4   0.058  0.81218    
## Residuals   21  147.5     7.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This analysis suggest I have to include CYL, DISP and WT within a linear model as those are significant variables.

2nd Step: Model's specification

I considered the following model (Details and specific numbers are provided at the end of the document) as I need to determine transmission and significant varibles effects over MPG:

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

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## wt           -3.1251     0.9109  -3.431  0.00189 **
## am            0.1765     1.3045   0.135  0.89334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

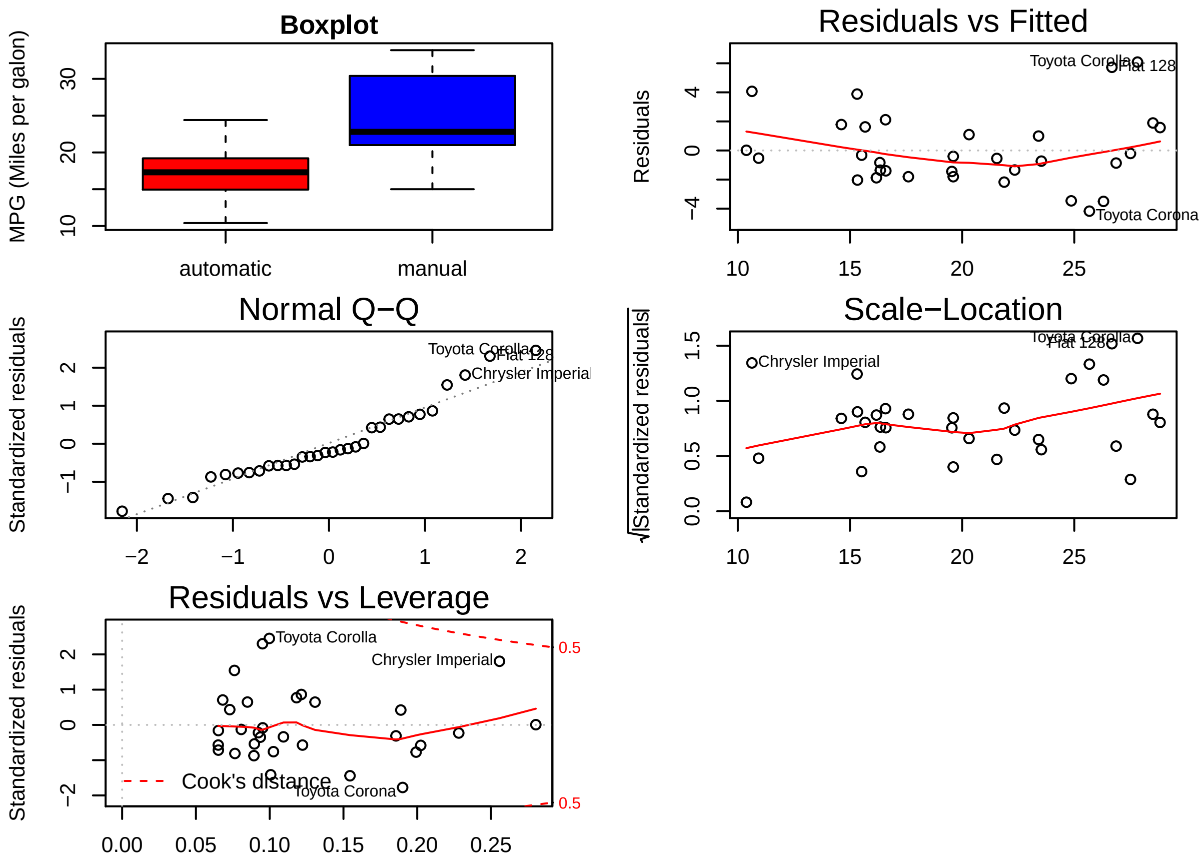
AM is *not significant* and has *a large p-value*, but it is not possible to reject the hypothesis that the coefficient of AM is 0.

3rd step: Box plot and residual plots

Automatic transmission versus manual transmission related to MPG (the question is to determine is transmission is a relevant variable to explain MPG). The residual is the difference between the observed data of the dependent variable MPG and the fitted values $\widehat{MPG}$.

```
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
par(mfrow=c(3,2))
par(mar=c(2.5, 5.5, 1.5, 1.5))
boxplot(mpg ~ am, data = mtcars, xlab = "AM (Transmission type)",
        ylab = "MPG (Miles per galon)", main="Boxplot", xaxt="n", col=c("red","blue"))
```

```r
axis(1, at=c(1,2), labels=c("automatic", "manual"))
par(mar=c(2.5, 5.5, 1.5, 1.5))
plot(lm)
```



As the boxes in the plot do not superpose it means automatic and manual cars are different.

4th: Conclusion

AM, WT and CYL are relevant variables that explain MPG. The R^2 is 0.83 so the model has a *desirable goodness* of fit and we can explain MPG related to the model's variables. AM by its own cannot explain MPG but is an important variable that explains (partially) car's performance.

APPENDIX: Descriptive statistics

In order to obtain an idea of how to proceed I considered the statistical momentums (mean, median, quartiles and variance-covariance matrix)

```
summary(mtcars)  #mean, median and quatiles
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec         vs     am    
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   0:18   0:19  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1:14   1:13  
##  Median :3.695   Median :3.325   Median :17.71                 
##  Mean   :3.597   Mean   :3.217   Mean   :17.85                 
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90                 
##  Max.   :4.930   Max.   :5.424   Max.   :22.90                 
##       gear            carb      
##  Min.   :3.000   Min.   :1.000  
##  1st Qu.:3.000   1st Qu.:2.000  
##  Median :4.000   Median :2.000  
##  Mean   :3.688   Mean   :2.812  
##  3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :5.000   Max.   :8.000  
```

```
var(mtcars)  #variance-covariance matrix
```

```
##              mpg         cyl        disp          hp         drat
## mpg    36.324103  -9.1723790  -633.09721 -320.732056   2.19506351
## cyl    -9.172379   3.1895161   199.66028  101.931452  -0.66836694
## disp -633.097208 199.6602823 15360.79983 6721.158669 -47.06401915
## hp   -320.732056 101.9314516  6721.15867 4700.866935 -16.45110887
## drat    2.195064  -0.6683669   -47.06402  -16.451109   0.28588135
## wt     -5.116685   1.3673710   107.68420   44.192661  -0.37272073
## qsec    4.509149  -1.8868548   -96.05168  -86.770081   0.08714073
## vs      2.017137  -0.7298387   -44.37762  -24.987903   0.11864919
## am      1.803931  -0.4657258   -36.56401   -8.320565   0.19015121
## gear    2.135685  -0.6491935   -50.80262   -6.358871   0.27598790
## carb   -5.363105   1.5201613    79.06875   83.036290  -0.07840726
##               wt         qsec           vs           am        gear
## mpg   -5.1166847   4.50914919   2.01713710   1.80393145   2.1356855
## cyl    1.3673710  -1.88685484  -0.72983871  -0.46572581  -0.6491935
## disp 107.6842040 -96.05168145 -44.37762097 -36.56401210 -50.8026210
## hp    44.1926613 -86.77008065 -24.98790323  -8.32056452  -6.3588710
## drat  -0.3727207   0.08714073   0.11864919   0.19015121   0.2759879
## wt     0.9573790  -0.30548161  -0.27366129  -0.33810484  -0.4210806
## qsec  -0.3054816   3.19316613   0.67056452  -0.20495968  -0.2804032
## vs    -0.2736613   0.67056452   0.25403226   0.04233871   0.0766129
## am    -0.3381048  -0.20495968   0.04233871   0.24899194   0.2923387
## gear  -0.4210806  -0.28040323   0.07661290   0.29233871   0.5443548
## carb   0.6757903  -1.89411290  -0.46370968   0.04637097   0.3266129
##             carb
## mpg  -5.36310484
## cyl   1.52016129
## disp 79.06875000
## hp   83.03629032
## drat -0.07840726
## wt    0.67579032
## qsec -1.89411290
## vs   -0.46370968
## am    0.04637097
## gear  0.32661290
## carb  2.60887097
```

APPENDIX: Model specification

If I take this model direct from ANOVA data:

```
lm <- lm(mpg ~ cyl + disp + wt + am, data = mtcars)
summary(lm)
```

```
## 
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am, data = mtcars)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -4.318 -1.362 -0.479  1.354  6.059 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## cyl         -1.784173   0.618192  -2.886  0.00758 ** 
## disp         0.007404   0.012081   0.613  0.54509    
## wt          -3.583425   1.186504  -3.020  0.00547 ** 
## am1          0.129066   1.321512   0.098  0.92292    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079 
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

Then DISP is not significant, so there is evidence that the model can be improved by dropping variables as it is done with the final specification considered in this report.