Comparison of automatic and manual transmissions

Simone

2/23/2020

Executive summary

The purpose of this project is to look at the **mtcars** dataset to answer the following two questions:
- Is an automatic or manual transmission better for miles per gallon?
- What is the miles per gallon difference between automatic and manual transmissions?

Using the best model that we examined, our conclusion is that **manual transmissions are better than automatic transmissions in terms of miles per gallon**, all the rest being equal (significative result), and in particular **they can achieve *4.300* miles per gallon more**.

Loading and exploring the dataset

```
library(datasets)
data(mtcars)
```

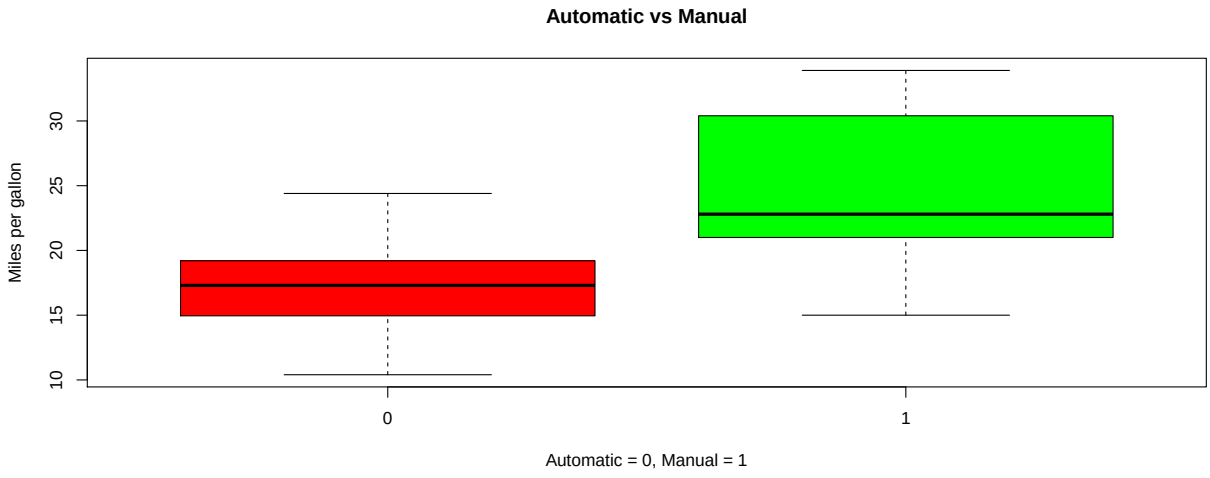
The data cointains 32 observations on 11 (numeric) variables.

```
colnames(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

The columns are: (1) **mpg** (miles per US gallon), (2) **cyl** (number of cylinders), (3) **disp** (displacement in cu.in.), (4) **hp** (gross horsepower), (5) **drat** (rear axle ratio), (6) **wt** (weight with 1000 lbs unit), (7) **qsec** (1/4 mile time), (8) **vs** (engine: 0 = V-shaped, 1 = straight), (9) **am** (transmission: 0 = automatic, 1 = manual), (10) **gear** (number of forward gears), (11) **carb** (number of carburetors)

```
with(mtcars, boxplot(mpg ~ factor(am), col = c("red", "green"), main = "Automatic vs Manual",
                     xlab = c("Automatic = 0, Manual = 1"), ylab = "Miles per gallon"))
```



```r
delta_mean_manual_automatic <- with(mtcars, mean(mpg[am == 1]) - mean(mpg[am == 0]))
```

Here we can see that if we base our analysis on the variable **am** alone, on average manual cars have an **mpg** value ***7.245*** higher than automatic cars. However, like this we are ignoring the effect of the other variables, which could be significative.

```r
round(cor(mtcars)[1,sort(abs(cor(mtcars)[1,]), decreasing = TRUE, index.return = TRUE)$ix], 3)
```

```
##    mpg     wt    cyl   disp     hp   drat     vs     am   carb   gear   qsec
##  1.000 -0.868 -0.852 -0.848 -0.776  0.681  0.664  0.600 -0.551  0.480  0.419
```

In fact, **mpg** is mostly correlated, in decreasing order, with **wt** (negative), **cyl** (negative), **disp** (negative), **hp** (negative), **drat** (positive), **vs** (positive) and only then **am** (positive, as expected from the previous boxplot and delta mean calculation).

Model selection (strategy 1)

In the first strategy, we define models by progressively adding new variables based on the correlation order with **mpg** and then we select the best model running the anova test (in ***Appendix***).

```r
fit1 <- lm(mpg ~ am, data = mtcars)
fit2 <- lm(mpg ~ am + wt, data = mtcars)
fit3 <- lm(mpg ~ am + wt + hp, data = mtcars)
fit4 <- lm(mpg ~ am + wt + hp + cyl + disp, data = mtcars)
fit5 <- lm(mpg ~ am + wt + hp + cyl + disp + drat + vs, data = mtcars)
fit6 <- lm(mpg ~ ., data = mtcars)
```

Moving from fit1 to fit2 and from fit2 to fit3 the p-value is less than 0.05, so with a 95% confidence level we can say that the 2 new variables introduced (**wt** and **hp**) are both significative, however after that the next new variables are not significative anymore (p-value > 0.05), therefore we select model **fit3**, which has a residual standard error of ***2.538*** and an adjusted R-squared of ***0.823***.

```r
round(t(summary(fit3)$coeff[,c(1,4)]), 3)
```

```
##             (Intercept)    am     wt     hp
## Estimate         34.003 2.084 -2.879 -0.037
## Pr(>|t|)          0.000 0.141  0.004  0.001
```

Looking at the coefficients of the final model, **am** appears to have a positive effect on **mpg** (in particular the coefficient is ***2.084***, indicating that manual cars tend to have higher miles per gallon than automatic cars, all the rest being equal.

The problem with this conclusion, though, is that **am** doesn't appear to be significative (because the p-value for it's coefficient is ***0.141***) and therefore we should accept the null hypothesis (**am** doesn't impact **mpg**).

Model selection (strategy 2)

In this second strategy we start with all the variables (the full model fit6) and then we progressively remove them one by one, selecting everytime the one with the highest p-value, until only significant variables are left (the summary of all the models is in ***Appendix***).

```
fit7 <- lm(mpg ~ . - cyl, data = mtcars)
fit8 <- lm(mpg ~ . - cyl - vs, data = mtcars)
fit9 <- lm(mpg ~ . - cyl - vs - carb, data = mtcars)
fit10 <- lm(mpg ~ . - cyl - vs - carb - gear, data = mtcars)
fit11 <- lm(mpg ~ . - cyl - vs - carb - gear - drat, data = mtcars)
fit12 <- lm(mpg ~ . - cyl - vs - carb - gear - drat - disp, data = mtcars)
fit13 <- lm(mpg ~ . - cyl - vs - carb - gear - drat - disp - hp, data = mtcars)
fit14 <- lm(mpg ~ . - cyl - vs - carb - gear - drat - disp - hp - 1, data = mtcars)
```

The final model **fit14** has a residual standard error of ***2.497*** and an adjusted R-squared of ***0.986***, which are both better than the previous best model **fit3**.

```
round(t(summary(fit14)$coeff[,c(1,4)]), 3)
```

```
##              wt qsec  am
## Estimate -3.185  1.6 4.3
## Pr(>|t|)  0.000  0.0 0.0
```

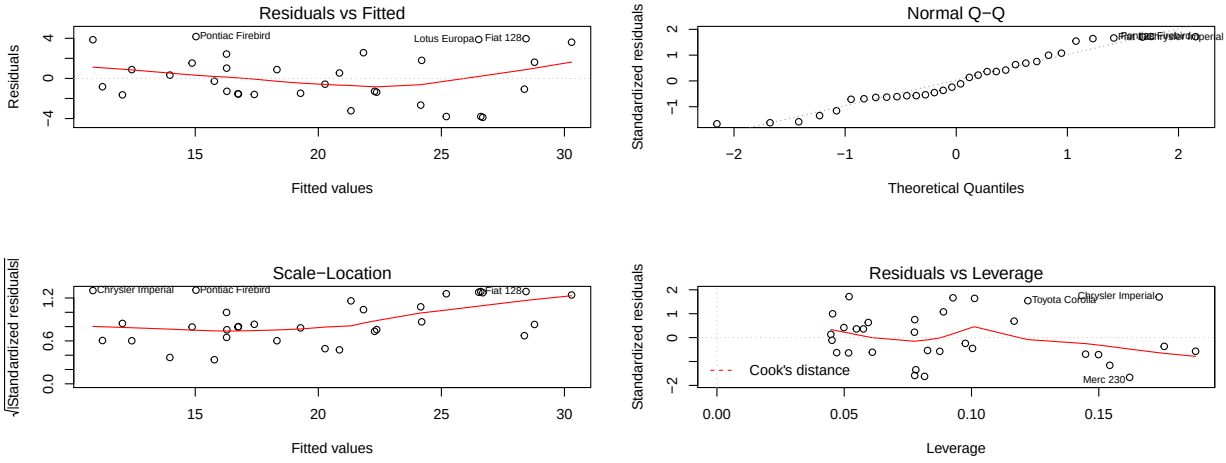
Furthermore here **am** is significative (p-value < 0.05) and still has a positive effect on **mpg** (the coefficient is ***4.300***), indicating that manual cars have higher miles per gallon than automatic cars all the rest being equal, albeit less than what initially forecasted by looking simply at the effect of **am** alone. The other significative variables are **wt** (negative effect) and **qsec** (positive effect), which appears to make sense.

Conclusion

Both from a model performance perspective, and **am** significance perspective, we finally select model **fit14**.

However, as a last step, we run some diagnostics on the model, but we do not observe any big problem with any of the plots.

```
par(mfrow = c(2,2))
plot(fit14)
```



Appendix

```
anova(fit1, fit2, fit3, fit4, fit5, fit6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + hp
## Model 4: mpg ~ am + wt + hp + cyl + disp
## Model 5: mpg ~ am + wt + hp + cyl + disp + drat + vs
## Model 6: 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     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     28 180.29  1     98.03 13.9571  0.001219 **
## 4     26 163.12  2     17.17  1.2224  0.314639
## 5     24 158.65  2      4.47  0.3179  0.731119
## 6     21 147.49  3     11.16  0.5296  0.666844
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
round(t(summary(fit6)$coeff[,c(1,4)]), 3)
```

```
##          (Intercept)    cyl  disp     hp  drat     wt  qsec    vs    am  gear
## Estimate      12.303 -0.111 0.013 -0.021 0.787 -3.715 0.821 0.318 2.520 0.655
## Pr(>|t|)       0.518  0.916 0.463  0.335 0.635  0.063 0.274 0.881 0.234 0.665
##            carb
## Estimate -0.199
## Pr(>|t|)  0.812
```

```
round(t(summary(fit7)$coeff[,c(1,4)]), 3)
```

```
##          (Intercept)  disp     hp  drat     wt  qsec    vs    am  gear   carb
## Estimate      10.960 0.013 -0.022 0.835 -3.693 0.842 0.390 2.577 0.712 -0.220
## Pr(>|t|)       0.427 0.454  0.306 0.592  0.057 0.233 0.843 0.198 0.608  0.783
```

```
round(t(summary(fit8)$coeff[,c(1,4)]), 3)
```

```
##          (Intercept)  disp     hp  drat     wt  qsec    am  gear   carb
## Estimate       9.768 0.012 -0.021 0.875 -3.712 0.911 2.524 0.760 -0.248
## Pr(>|t|)       0.420 0.459  0.304 0.563  0.050 0.132 0.193 0.569  0.747
```

```
round(t(summary(fit9)$coeff[,c(1,4)]), 3)
```

```
##          (Intercept)  disp     hp  drat     wt qsec    am  gear
## Estimate       9.198 0.016 -0.025 0.810 -4.131 1.01 2.590 0.606
## Pr(>|t|)       0.433 0.213  0.135 0.582  0.003 0.05 0.171 0.620
```

```r
round(t(summary(fit10)$coeff[,c(1,4)]), 3)
```

```
##          (Intercept)  disp     hp  drat     wt  qsec    am
## Estimate      10.711 0.013 -0.022 1.021 -4.045 0.991 2.985
## Pr(>|t|)       0.338 0.244  0.149 0.462  0.003 0.050 0.080
```

```r
round(t(summary(fit11)$coeff[,c(1,4)]), 3)
```

```
##          (Intercept)  disp     hp     wt  qsec    am
## Estimate      14.362 0.011 -0.021 -4.084 1.007 3.470
## Pr(>|t|)       0.152 0.299  0.156  0.002 0.044 0.027
```

```r
round(t(summary(fit12)$coeff[,c(1,4)]), 3)
```

```
##          (Intercept)     hp     wt  qsec    am
## Estimate      17.440 -0.018 -3.238 0.811 2.926
## Pr(>|t|)       0.072  0.223  0.001 0.076 0.046
```

```r
round(t(summary(fit13)$coeff[,c(1,4)]), 3)
```

```
##          (Intercept)     wt  qsec    am
## Estimate       9.618 -3.917 1.226 2.936
## Pr(>|t|)       0.178  0.000 0.000 0.047
```