MPG in auto vs manual transmission

Omer Yavin

October 19, 2018

Executive summary: This is a closing project for the Regression Models course, which is part of the Data Science specialization. The assignments is to explore the relationship between a set of variables and miles per gallon (MPG) (outcome). Particularly, answering in the following two questions:

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

To give a real answer, the comparison should be specifically defined, including what should be accounted for (kept constant), and what is considered an inherent characteristic of a specific type of transmission. Ultimately, it seems there is some advantage to manual transmission. Here an increase of 2.94 MPG was found with the best model. The paper shows the methods used and discusses their flaws.

Data overview:

From the data description obtained from ?mtcars: *The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).*

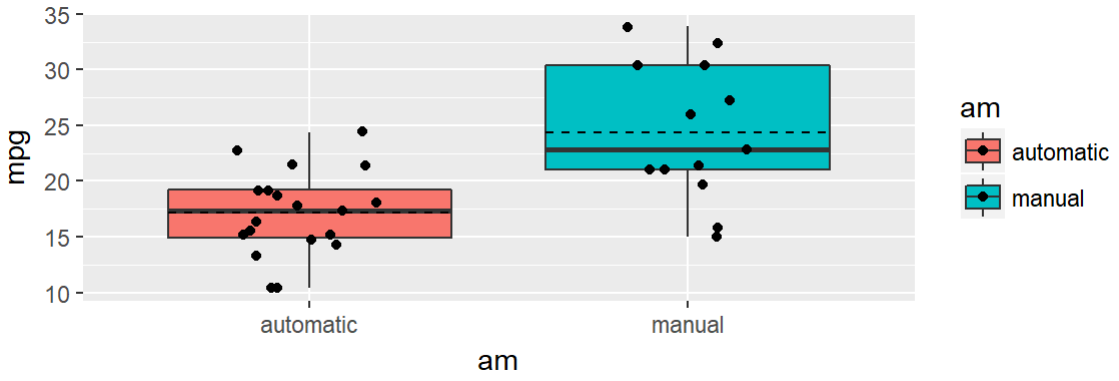
Also shown in the same data description are the various data points collected for each car: MPG, cylinder #, displacement (size?), horse power, weight, 1/4 mile time (a measure of performance) and other such factors. Also included is the type of transmission (0 for Auto and 1 for Manual), with which the relation to MPG is the focus of this work.

```
mtcars2 <- within(mtcars, {
   am <- factor(am, labels = c("automatic", "manual"))
})
```

Exploratory data analysis:

Firstly, the naive relation between transmission type and mpg will be observed.

```
p <- ggplot(mtcars2, aes(x=am, y=mpg, fill=am)) + geom_boxplot() + geom_jitter(shape=16, positio
n=position_jitter(0.2)) + stat_summary(fun.y = mean, geom = "errorbar", aes(ymax = ..y.., ymin =
 ..y..), width = .75, linetype = "dashed")
p
```



```
tapply(mtcars2$mpg,mtcars2$am,mean)
```

```
## automatic    manual
##  17.14737  24.39231
```

```
mdl_naive<-lm(mpg~am,mtcars2)
```

The affect of transmission type seems clear here. Precisely, driving a manual gear will give you an average $24.4 - 17.15 = 7.25$ extra miles per gallon, but there is some visible overlap between the mpg values. Maybe there are other factors affecting this gap. Perhaps automatic transmissions are heavier, adding drag to the car, and maybe we need to account for acceleration capabilities (qsec) to have an “apples to apples” comparison?..

These thought directions are further looked into in Appendix #2.

Stepwise regression will be used as an effective method of achieving the best selection of variables for linear regression, taking into account the likelihood function.

```
mdl_all<-lm(mpg~.,mtcars2)
mdl_step<-stepAIC(mdl_all,direction = "both",trace = FALSE)
anova(mdl_naive,mdl_step,mdl_all)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: 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     21 147.49  7     21.79  0.4432    0.8636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is expected for the most inclusive regressionto get most significance of the 3, but this probably includes some overfitting as we know some of the variables are correlated. Looking at the t-values for each seperate variable in the regression summary, it becomes clear that it is not a valid regression. Hence, as discussed further in Appendix #2, the results of the stepwise function will be used.

Questions and answers

Question #1: Is an automatic or manual transmission better for MPG?

It is shown in the boxplot graph that manual transmission is more efficient than auto transmission. This will be tested using the one-sided t.test. Our hypothesis are defined as following:

- $H_0 : MPG_{MAN} = MPG_{AUTO}$
- $H_0 : MPG_{MAN} \geq MPG_{AUTO}$

```
t.test(c(mtcars2$mpg[mtcars2$am=="manual"]),c(mtcars2$mpg[mtcars2$am=="automatic"]),alternative
 = "greater", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  c(mtcars2$mpg[mtcars2$am == "manual"]) and c(mtcars2$mpg[mtcars2$am == "automatic"])
## t = 3.7671, df = 18.332, p-value = 0.0006868
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.913256      Inf
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

Zero is *not* included in the 95% confidence interval and also $p - value << 0.05$, the null hypothesis is rejected and it is inferred that, from the data available here, cars with manual transmission is "better for MPG" than the ones with automatic transmission.

Question #2: Quantify the MPG difference between automatic and manual transmissions

The regression returned by stepwise function includes weight, 1/4 mile time and of course - transmission type (auto\manual).

```
mdl_step$coefficients
```

```
## (Intercept)          wt        qsec    ammanual
##    9.617781   -3.916504    1.225886    2.935837
```

Diagnostic plots for this regression (see Appendix #3) show normal behavior, t-values are above 2 for independent variables, and p-value (Appendix #3) for the complete regression is <<0.05 which all signifies this is statistically valid. The standard error for 'am' variable is ~1.4 and the t-value is significantly lower than the others, but still valid. Using the estimated coefficient itself it can be inferred that, holding all other variables constant, simply switching from an automatic transmission to a manual one will increase MPG by 2.94.

Appendix #1: observe data structure

```
str(mtcars)
```

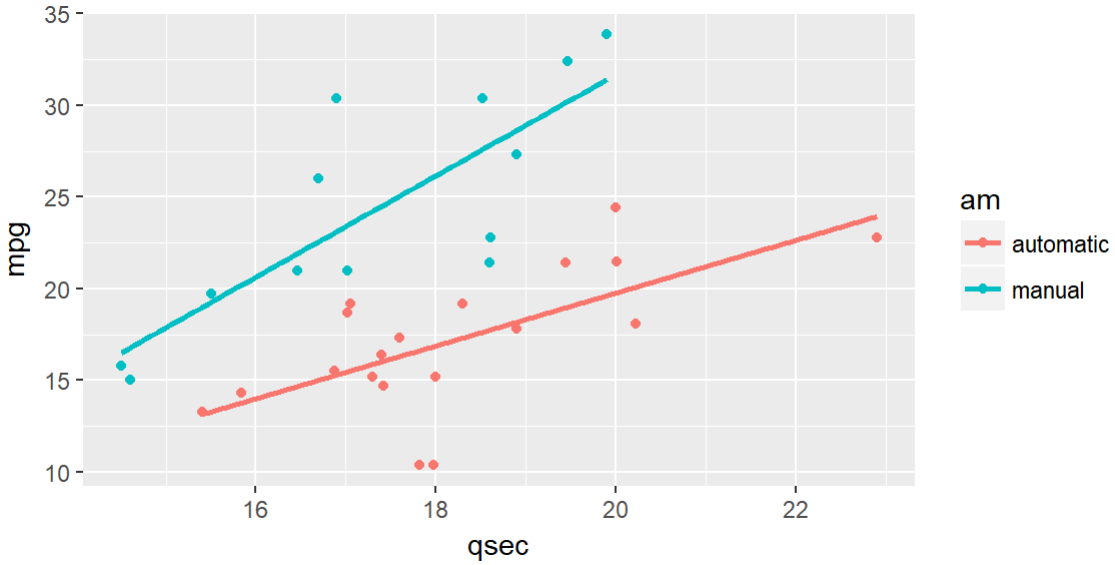
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 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 ...
```

Appendix2: more exploratory data analysis

The initial observation of auto vs manual transmission mean MPG rates showed a significant advantage to manual transmission, but raised the question - how much of that advantage can be purely attributed to the transmission type itself? Maybe weight and acceleration are non-inherent qualities of the transmission type which widen this gap, or even create it?..

Following plots will examine a couple of these thought directions.

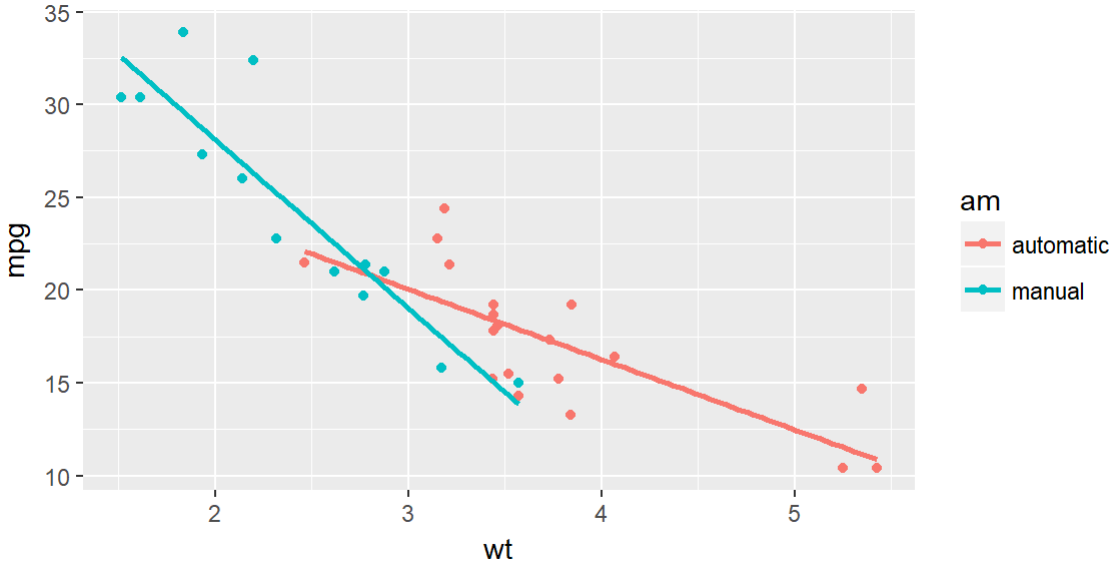
```
qsec <- ggplot(mtcars2, aes(x=qsec, y=mpg, color=am)) + geom_point() + geom_smooth(method = "lm"
, se = FALSE)
qsec
```



We see that: 1. Higher performing cars (lower qsec) generally seem to be less efficient in terms of mpg. 2. For a given qsec value, manual transmission cars are generally more efficient then auto transmission cars. 3. This also shows us - there is no clear connection between qsec and transmission type.

Can car weight explain some of this gap?

```
wtmpg <- ggplot(mtcars2, aes(x=wt, y=mpg, color=am)) + geom_point() + geom_smooth(method = "lm",
 se = FALSE)
wtmpg
```



From this plot we can infer 2 things: 1. There seems to be a clear negative connection between car weight and mpg. 2. Manual transmission cars tend to be slightly less heavy than auto transmission cars. 3. For the same weight, it actually seems auto transmission is about the same, if not more efficient than manual transmission.

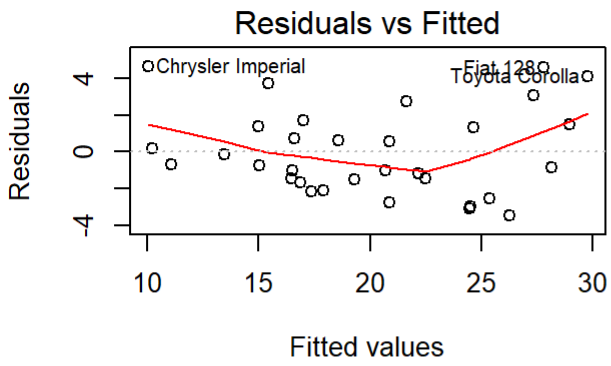
This gets slightly tricky here, since it can be asked - is the weight of the car an inherent property of the transmission type? If not, it should be accounted for. Since the assignment deals with regression models, the formal methods will be used to deduce what should be included in the regression. In an actual study of transmission inefficiencies more discussion and details would be needed here.

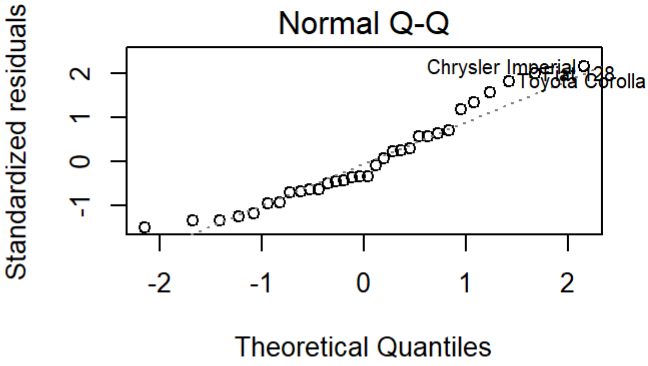
Appendix #3: Dignostic of chosen regression of MPG

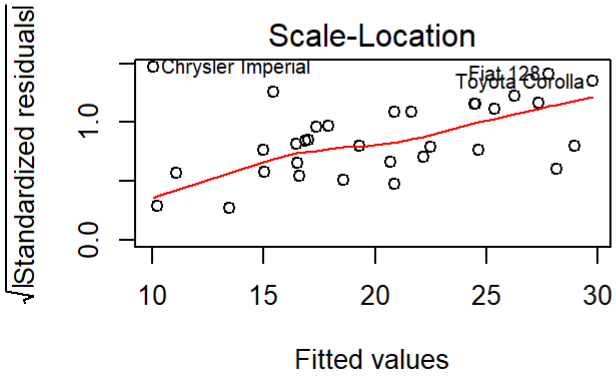
```
mdl <- lm(mpg~.,data=mtcars2)
step <- stepAIC(mdl, direction="both", trace = "FALSE")
summary(step)
```

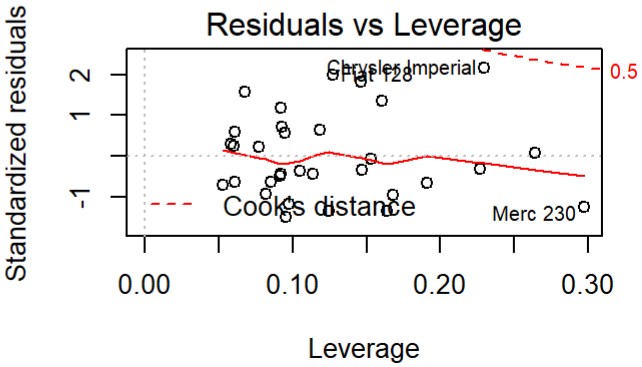
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ammanual      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
```

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


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
Fitted values


Normal Q-Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
Theoretical Quantiles


Scale-Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Chrysler Imperial
Fiat 128
0.5
Cook's distance
Merc 230
Standardized residuals
Leverage