Is automatic or manual transmission better for mpg?

Rodrigo

May 31, 2019

Overview

In this report, we investigate which type of transmission, manual or automatic, is better fot MPG. Better is defined as the transmission having a higher mpg. We use the mtcars dataset included in the R base package for analysis. The analysis consisted of fitting a regression line on mpg, the response variable, with weight (wt), number of cylinders (cyl) and transmission (am) as the independent variables. In the end, we found, **am**, is not statistically significant i.e we can't reject the null that both transmission types have the same effect on mpg. An ANOVA test confirmed this and a shapiro-wilk test was conducted to validate the ANOVA results.

Glance at the Dataset

```
data("mtcars")
```

```
summary(mtcars[,1:7])
```

```
##       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      
##  Min.   :2.760   Min.   :1.513   Min.   :14.50  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89  
##  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  
```

The first step is to look at the data for anything that might warrant concern during the analysis phase. The summary function reveals there are no missing or erroneous data points.

```
str(mtcars[,1:9])
```

```
## 'data.frame':    32 obs. of  9 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 ...
```

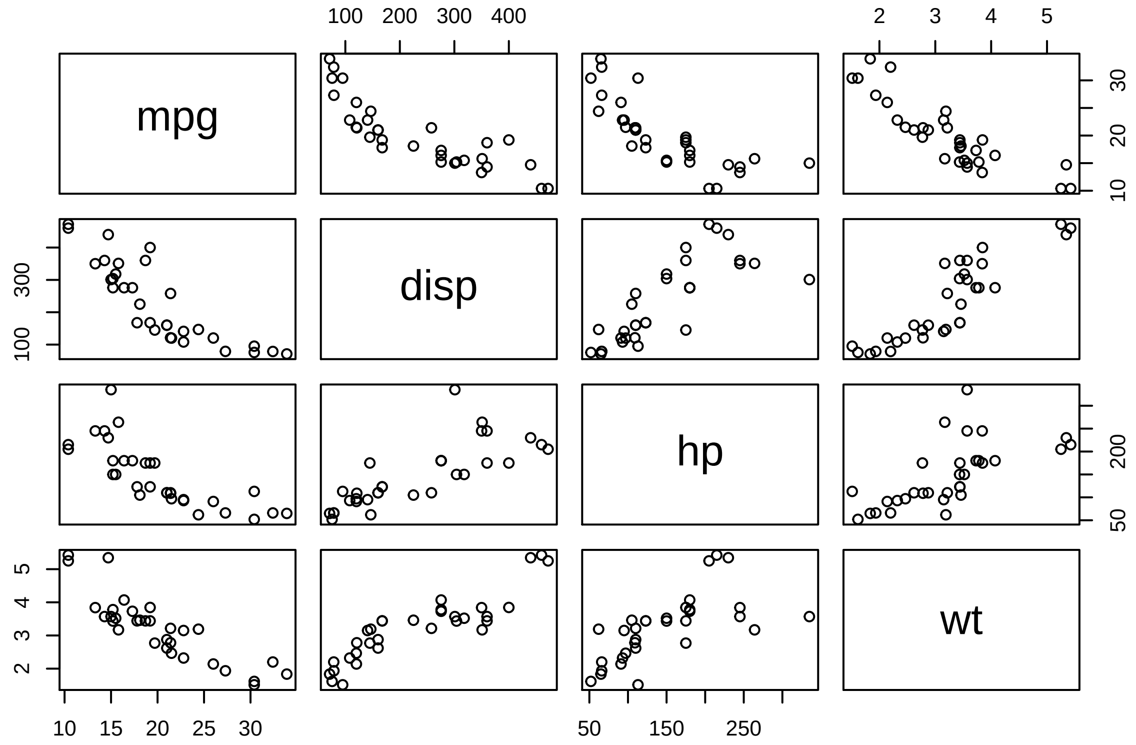
```
mtcars$am <- factor(mtcars$am,
                    ordered = F,
                    labels = c("Automatic", "Manual"))
```

```
mtcars$cyl <- factor(mtcars$cyl,
                     ordered = F)
```

However, a look at the structure of the dataset reveals **am** and **cyl**, two variables of interest, are of type numeric. We know these are categorical variables, so we convert them to factors. The data is now tidy and analysis-ready.

Simple Linear Regression

```
pairs(mtcars[, c(1,3,4,6)])
```



From the pairwise plots, it appears **mpg** and **wt** have nice linear relationship. We begin the analysis with a simple linear regression on mpg versus weight.

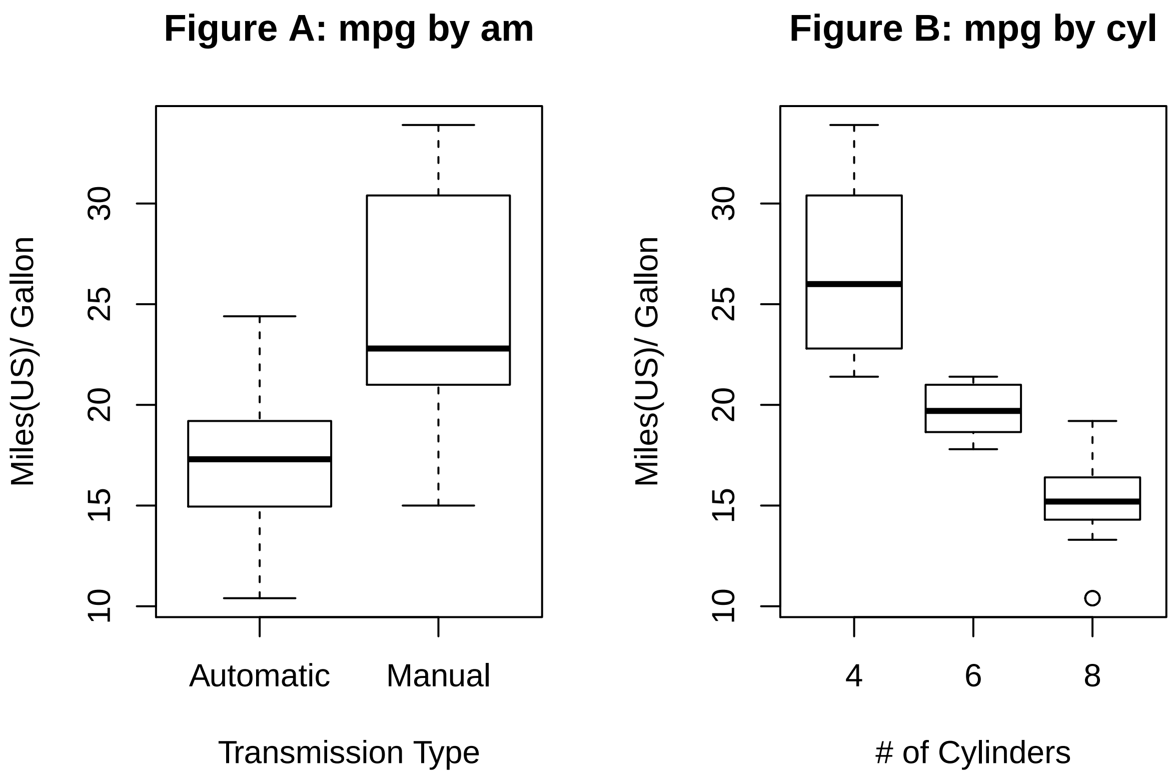
```
simple <- lm(mpg ~ wt, data = mtcars)
summary(simple)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 37.285126   1.877627 19.857575 8.241799e-19
## wt          -5.344472   0.559101 -9.559044 1.293959e-10
```

Interpretation: The coefficient estimate for weight (wt) is -5.34. Put differently, a 1000 lbs increase in weight is associated with a 5.34 decrease in mpg. A p-value samller than 0.001 suggests the coefficient is statistically significant. This makes sense, heavier cars consume more gasoline and thus have a lower mpg.

Multivariate Linear Regression

```
old.par <- par(mfrow=c(1, 2))
plot(mtcars$am, mtcars$mpg, ylab = "Miles(US)/ Gallon", xlab = "Transmission Type", main = "Figure A: mp
plot(mtcars$cyl, mtcars$mpg, ylab = "Miles(US)/ Gallon", xlab = "# of Cylinders", main = "Figure B: mpg
```



```
par(old.par)
```

Transmission type (am) and number of cylinders (cyl) are two categorical variables we wish to add next. We generate boxplots for an indication that miles per gallon (mpg) varies by **am** and **cyl**. Figure A suggests manual transmission vehicles have a higher mpg than automatic transmission vehicles.

```
auto = subset(mtcars, am == "Automatic")$mpg
manual = subset(mtcars, am == "Manual")$mpg
t.test(auto, manual, paired = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  auto and manual
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

After conducting a t-test, at a .05 significance level, we can reject the null that the true difference in mpg means between the groups (automatic and manual) is equal to zero. In fact, a 95% confidence interval suggests the true difference is between 17.14 and 24.39. We add them to the model, one at a time, to quantify their effect on **mpg**.

```
multi <- lm(mpg ~ wt + am, data = mtcars)
multi2 <- lm(mpg ~ wt + am + cyl, data = mtcars)
summary(multi2)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 33.7535920  2.8134831 11.9970836 2.495549e-12
## wt          -3.1495978  0.9080495 -3.4685309 1.770987e-03
## amManual     0.1501031  1.3002231  0.1154441 9.089474e-01
## cyl6        -4.2573185  1.4112394 -3.0167231 5.514697e-03
## cyl8        -6.0791189  1.6837131 -3.6105432 1.227964e-03
```

Interpretation: The results are surprising. Manual transmission vehicles, on average, have 0.15 more mpg than manual transmission cars, holding weight and number of cylinders constant; a trivial difference, but even more interesting is the insignificance of the transmission coeffient. At a .05 significance level, we fail to reject the null that automatic and manual transmission cars have the same effect on mpg.

```
anova(simple, multi, multi2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + am
## Model 3: mpg ~ wt + am + cyl
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 278.32
## 2     29 278.32  1     0.002 0.0003 0.985627
## 3     27 182.97  2    95.351 7.0353 0.003473 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

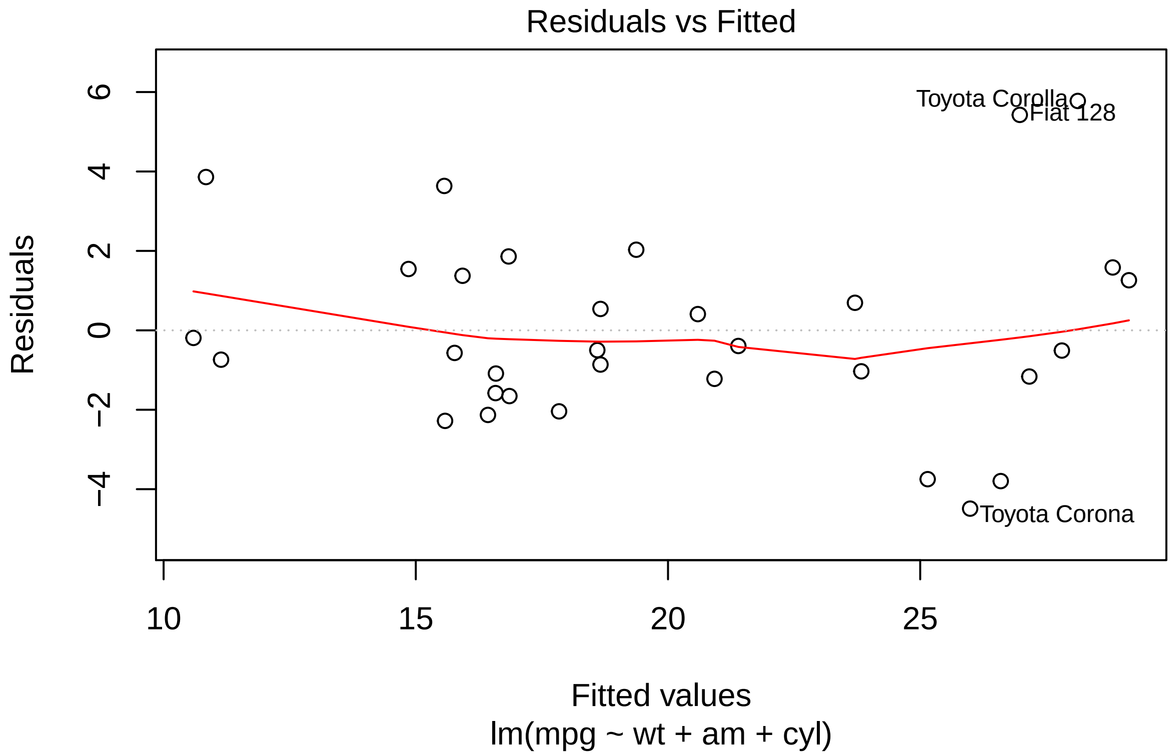
This inference is strengthened when we perform an anova test and observe that we fail to reject the null that the added regressor, namely **am**, is not significant.

Validating ANOVA Results

```
shapiro.test(multi2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  multi2$residuals
## W = 0.95915, p-value = 0.2602
```

```
plot(multi2, which = 1)
```



Since analysis of variance assumes that the residuals are approximately normal, we conduct an shapiro-wilk test as well as plot the residuals to confirm normality. The plot shows there is no pattern in the residual line. Also, at a significance level of .05, we fail to reject the null hypothesis that the residuals are normally distributed in the Shapiro-test.

Conclusion

Initially, a t-test suggested transmission type (am) was significat by rejecting the null that the difference between the two groups was zero. However, once in the multiregression model, **am** became insignificant, most likely because most of the variation in the response variable is explained by **wt** and **cyl**. In other words, in the end, we fail to reject that the true difference between manual and automatic transmission vehicles is zero.