Relationship Between Miles Per Gallon and Transmission (Automatic vs Manual)

Bhaskar S

08th Jun 2016

Executive Summary

In this project, we look at the **mtcars** dataset to explore the relationship between a set of variables and miles per gallon (MPG) (outcome). In particularly, we are interested in answering the following two questions:

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

Using Exploratory Data Analysis, Hypothesis Testing, and Linear Regression Modelling, we can conclude that the cars with Manual Transmission provide a better MPG compared to the cars with Automatic Transmission.

Setup

For this analysis, we will be using the R packages `ggplot2` and `GGally`. We load the desired libraries and the **mtcars** dataset:

```
library(ggplot2); library(GGally); data('mtcars')
```

Exploratory Data Analysis

We will display the dimensions and the structure **mtcars**:

```
str(mtcars)
```

From **Output.1** in **APPENDIX**, we see there are **32** rows and **11** columns and all the columns are `numeric`. Now, we compute and display the *mean* mpg for automatic (`am` = 0) and manual (`am` = 1):

```
aggregate(mpg ~ am, data=mtcars, mean)
```

From **Output.2** in **APPENDIX**, we observe that the *mean* mpg for automatic is about `17` and that for manual is about `24`. The **Plot 1** in **APPENDIX** also illustrates and confirms this fact.

The initial assessment indicates that the **mpg** from manual is better compared to automatic.

Hypothesis Testing

We collect the `mpg` values for automatic (`am` = 0) and manual (`am` = 1) into vectors *auto* and *manual* respectively and display their lengths:

```r
auto <- mtcars[mtcars$am==0, 'mpg']; length(auto)
manual <- mtcars[mtcars$am==1, 'mpg']; length(manual)
```

From the above, we see the lengths are less than **30**. Also, we do not have any knowledge of their population variances. As a result, we will be conducting a **t** hypothesis test to find the **p-value**. We will test the *null* hypothesis that the mean `mpg` for automatic and manual are equal with a 95% **Confidence Interval**. Since the *null* hypothesis is testing for equality, this is a **two-tail** test. If the **p-value** is < 0.05, we reject the *null* hypothesis.

The following code performs a two-tailed **t** hypothesis test:

```r
t.test(auto, manual, var.equal = FALSE, paired = FALSE, conf.level = 0.95)
```

From **Output.3** in **APPENDIX**, we observe a **p-value** of **0.0014** that is less than **0.05** and hence we reject the *null* hypothesis concluding that there is a major difference between the mean `mpg` for manual vs automatic.

Linear Regression Modelling

We start with the `pairs` plot. From **Plot.2** in **APPENDIX**, we observe that the variables `cyl` (-0.852), `disp` (-0.848), and `wt` (0.868) are highly correlated to `mpg`. We will create few models with `mpg` as the outcome and `am`, `cyl`, `disp`, and `wt` as explanatory variables:

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

To compare the various models, we perform an *ANOVA* analysis on the models:

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

From **Output.4** in **APPENDIX**, we observe that models *fit2* and *fit3* have lower **p-values** and hence are better choices.

To pick the best model, we compare the `R-Squared` values for the models *fit2* and *fit3* and pick the one with a larger value:

```r
summary(fit2)$r.squared; summary(fit3)$r.squared
```

From **Output.5** in **APPENDIX**, we choose model *fit3* as it has a larger `R-Squared` value.

Finally, we look at the residual plots:

From **Plot.3** in **APPENDIX**, we infer that the residuals are randomly scattered and follow a normal distribution.

Now that we have the best model, we look at the coefficients for the model *fit3*:

```r
summary(fit3)$coef
```

From **Output.6** in **APPENDIX**, we conclude that the *mean* `mpg` for manual is `2.567` more than for automatic.

APPENDIX

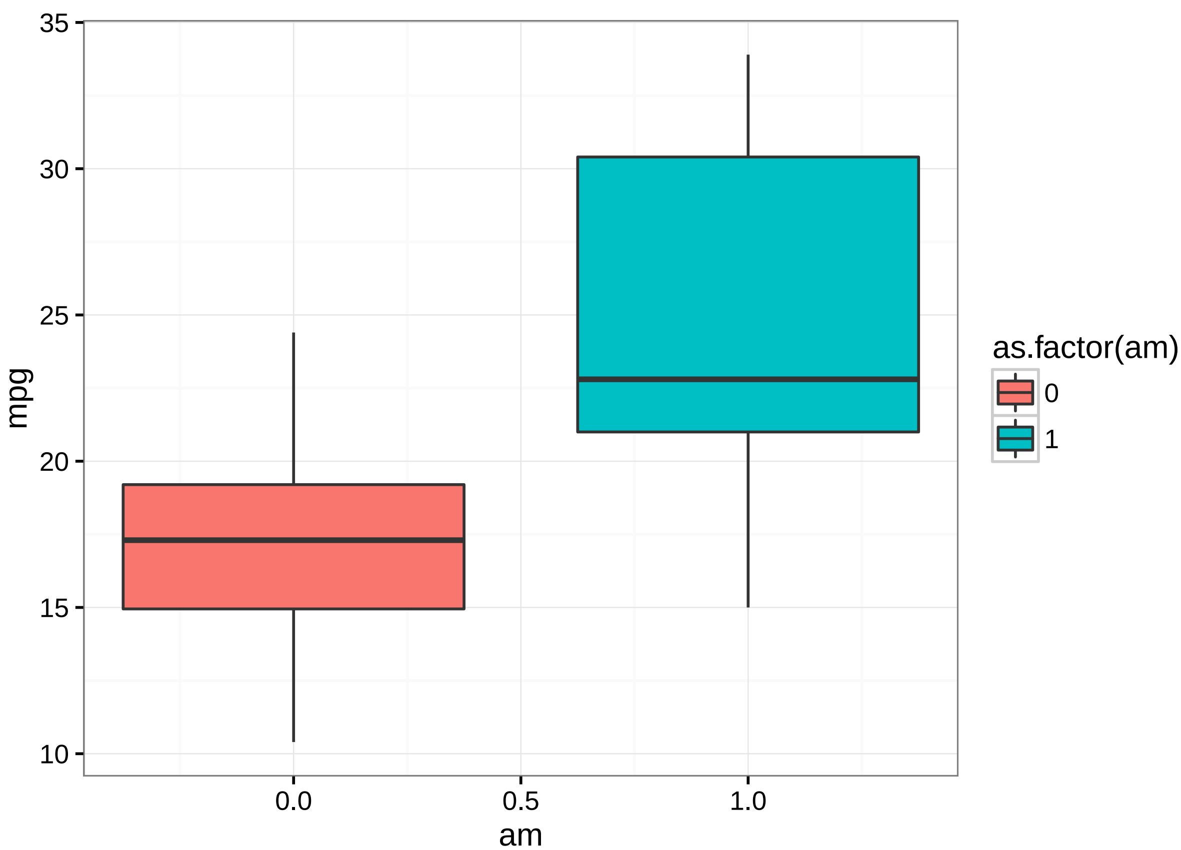
Output.1

```
## '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 ...
```

Output.2

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

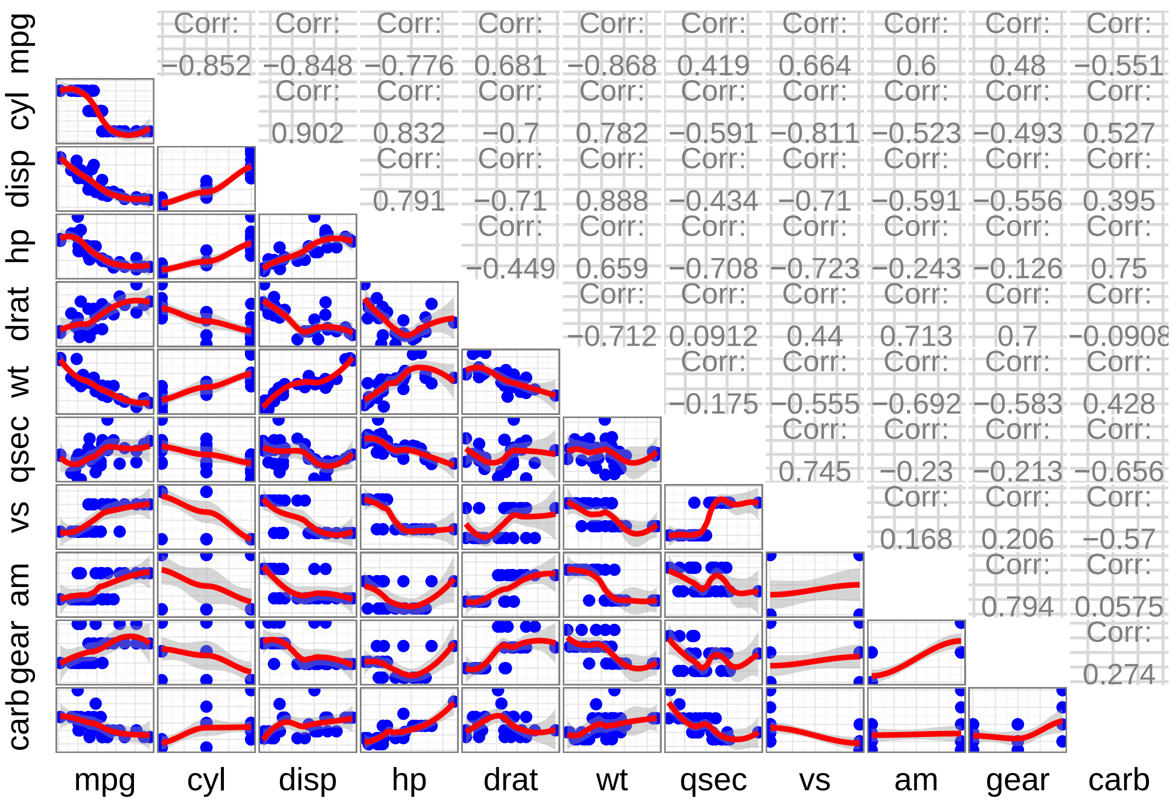
Plot.1



Output.3

```
##
##  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
```

Plot.2



Output.4

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am
## Model 3: mpg ~ am + cyl
## Model 4: mpg ~ am + cyl + disp
## Model 5: mpg ~ am + cyl + disp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     21 147.49
```

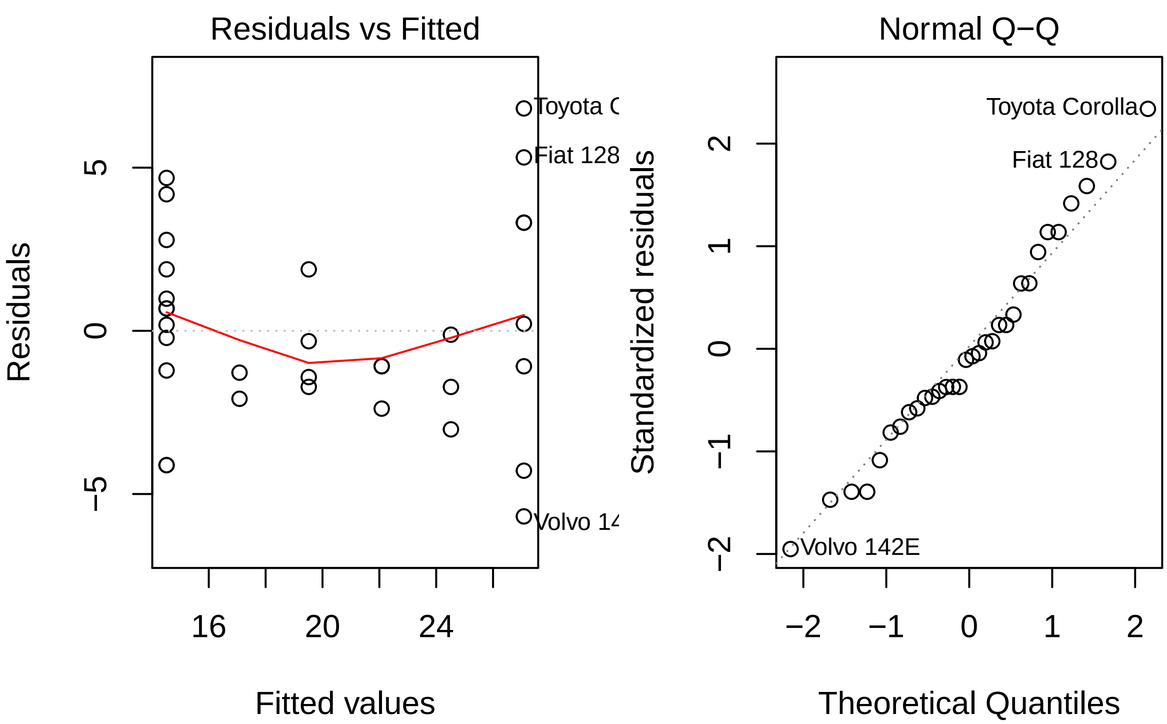
```
## 2     30 720.90 -9   -573.40  9.0711 1.779e-05 ***
## 3     29 271.36  1    449.53 64.0039 8.231e-08 ***
## 4     28 252.08  1     19.28  2.7452   0.11241
## 5     27 188.43  1     63.66  9.0631   0.00666 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output.5

```
## [1] 0.3597989
```

```
## [1] 0.7590135
```

Plot.3



Output.6

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 34.522443  2.6031842 13.261621 7.694408e-14
## am           2.567035  1.2914280  1.987749 5.635445e-02
## cyl         -2.500958  0.3608282 -6.931159 1.284560e-07
```