How Transmission Type 'Automatic' or 'Manual' affect MPG

Oye Akinnikawe

January 17, 2017

Executive Summary

In this document, the *mtcars* dataset was used to explore the relationship between a set of variables and miles per gallon (MPG) the outcome. We were particularly interested in answering the question whether automatic or manual transmission is better for MPG, and to quantify the MPG difference between automatic and manual transmissions. From our analysis, we can show that automatic transmission is better for mpg and the MPG difference between automatic and manual transmission is **1.81 MPG**

Data Analysis

first we will load the required libraries and data, then we will take a quick peek at the data.

```
library(ggplot2)
data(mtcars)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

We need to transform some of the variables in ***mtcars*** into factors. This will help in data exploration (shown in boxplot in Appendix) and obtaining better models during regression analysis.

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

Get the mean for the factors of *am* i.e automatic and manual

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

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

Thus, we will use the hypothesies that manual cars have a **7.25 MPG** higher than automatic cars.

Split the variable *am* into "auto" and "manual"

```
auto <- mtcars[mtcars$am == "Automatic",]
manual <- mtcars[mtcars$am == "Manual",]
```

We perform a t.test to see if there is a significant difference between the mean *mpg* for automatic and manual cars

```
t.test(auto$mpg,manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  auto$mpg and manual$mpg
## 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
```

The p-value of 0.001374 states that there is a significant difference in the mpg values for automatic and manual cars. we have to quantify this using regression models.

We will perform an initial linear fit for mpg~am, then we will add more variables to the regression to see if other variables have an effect on our model.

```
## perform initial lm fit for mpg~am
fit1 <- lm(mpg ~ am, data = mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The initial model shows that mean *mpg* for automatic cars is 17.15 MPG while that of manual cars is 7.25 MPG higher. The R^2 value of 0.36 shows that only 36% of the total variability is explained by our model. We need to try using multivariate linear regression to better fit our model.

```
## perform fit2 by adding other variables
fit2 <- lm(mpg ~ am + cyl + disp, data = mtcars)
```

```
## perform fit3 by adding more variables
fit3 <- lm(mpg ~ am + cyl + disp + hp + wt, data = mtcars)
```

Next, we perform an analysis of variance to see if the second and third model fits were necessary

```
#$Analysis of Variance
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp
## Model 3: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 230.46  3    490.44 27.1724 4.919e-08 ***
## 3     25 150.41  2     80.05  6.6528  0.004825 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-values test whether or not the new variables are necessary. This model concludes that all of the added Model 3 terms are necessary over Model 2 and all of the Model 2 terms are necessary over Model 1. Therefore, we will select Model 3.

```
summary(fit3)
```

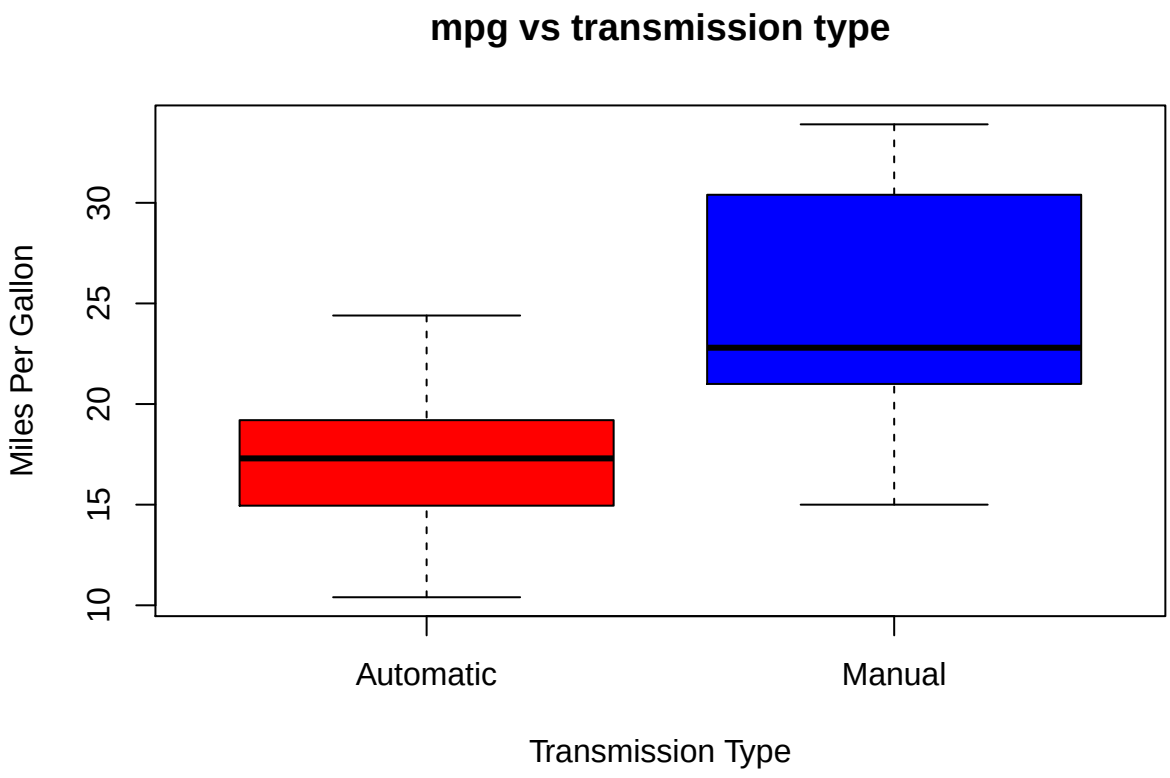
```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

The selected model *fit3* explains 87% of the total variability in our model. The addition of *cyl*, *disp*, *hp*, and *wt* affected the correlation between *mpg* and *am*. From our model, the difference between automatic and manual transmission is **1.81 MPG**

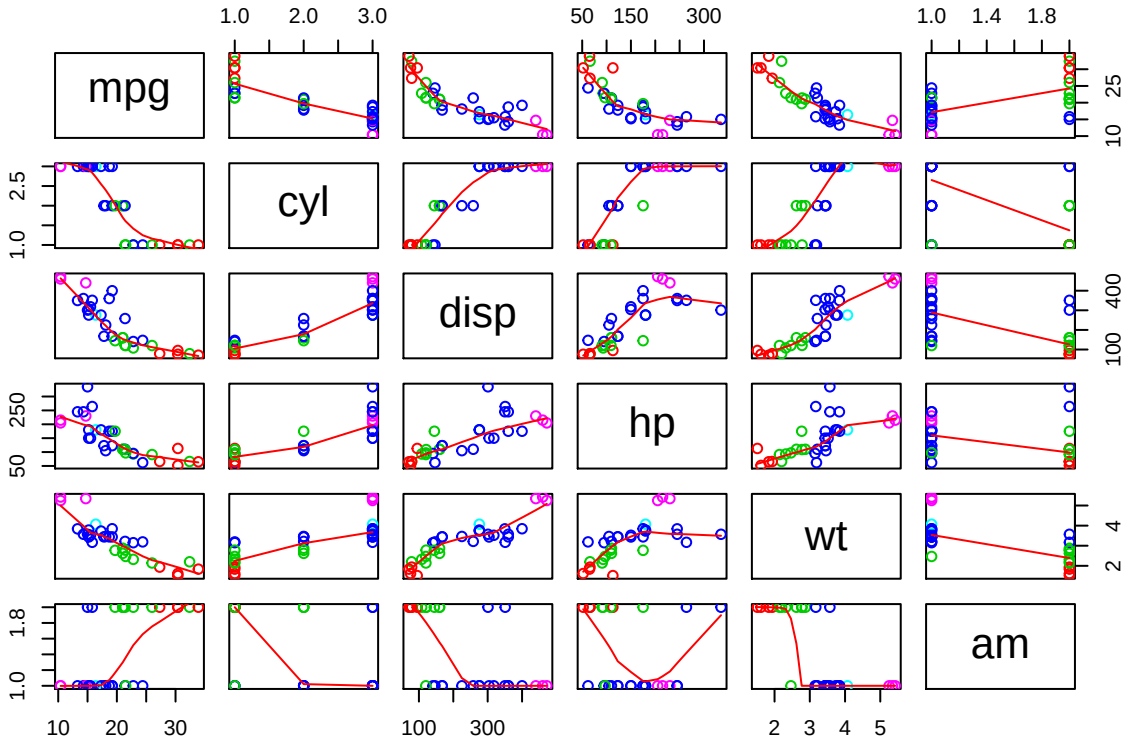
Appendix

The appendix shows some of the exploratory data analysis that was performed. Plot 1 shows the box plot of mpg vs transmission type, Plot 2 shows a pairs plot for selected variables, and plot 3 shows the residual plot used for investigating model fit.

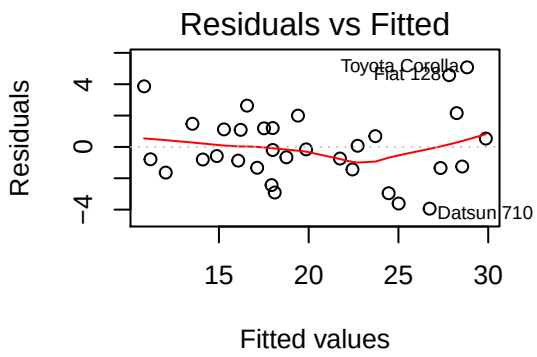
```
boxplot(mpg ~ am, data = mtcars, ylab = "Miles Per Gallon",
        xlab = "Transmission Type", main = "mpg vs transmission type",
        col = (c("red","blue")))
```

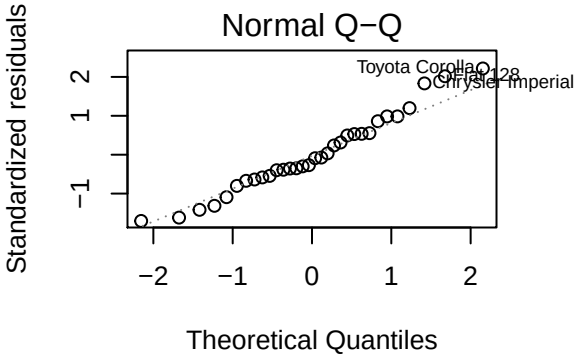


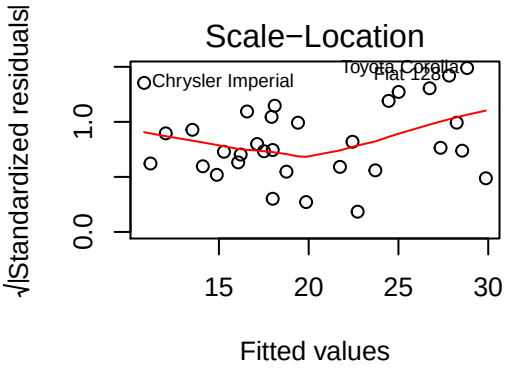

```
#Pairs plot
mtcars_vars <- mtcars[, c(1,2,3,4,6,9)]
par(mar = c(1,1,1,1))
pairs(mtcars_vars, panel = panel.smooth, col = mtcars$wt + 1)
```

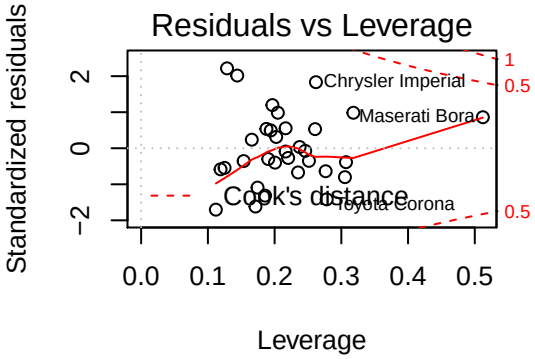



```
##residual plot
par(mar = c(4,4,4,4))
par(mfrow=c(2,2))
plot(fit3)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
−4
0
4
15
20
25
30


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
−1
1
2
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
0.0
1.0
15
20
25
30


Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5
−2
0
2
0.0
0.1
0.2
0.3
0.4
0.5