Motor Trends : Automatic or Manual transmission for better mileage ?

Girish Babu

24 October 2014

Executive summary

As part of the Coursera assignment on Regression models, I will try to address the following 2 objectives -

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

Use the `mtcars` dataset from the 1974 Motor Trend US magazine, and performed few Statistical Inference Tests and a corresponding Regression Analysis. Statistical Inference Tests show a difference in mean of about 7 miles more for the manual transmitted cars. Regression Analysis indicates, given that weight and 1/4 mile time are held constant, Manual transmission cars are (14.079 - 4.141 * weight mpg) better than Automatic Transmission cars on average and also that this result is significant. Conclusion is that lighter cars are better off with a Manual transmission, but heavier cars are better off with Automatic transmission.

Cleaning the data

First step is to load and look at the data:

```
data(mtcars)
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 ...
```

Now we coerce the "cyl", "vs", "gear", "carb" and "am" variables into factor variables:

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am)
```

Let me rename the levels of the "am" variable into "Auto" and "Manual", for better readability:

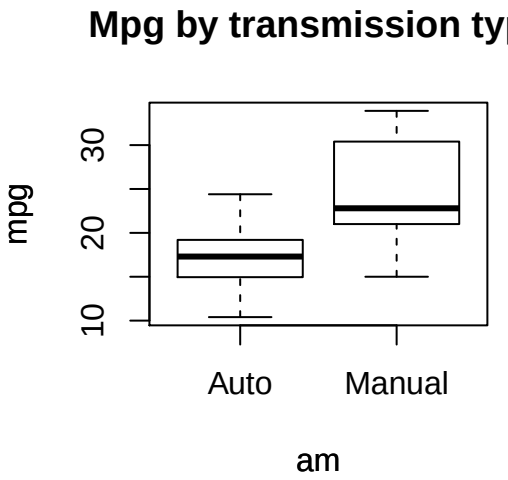
```
levels(mtcars$am) <- c("Auto", "Manual")
```

Exploratory analysis

We begin by plotting boxplots of the variable “mpg” when “am” is “Auto” or “Manual” (see below). This plot suggests an increase in mpg when cars had manual gears.

Figure : “mpg” vs. “transmission type”

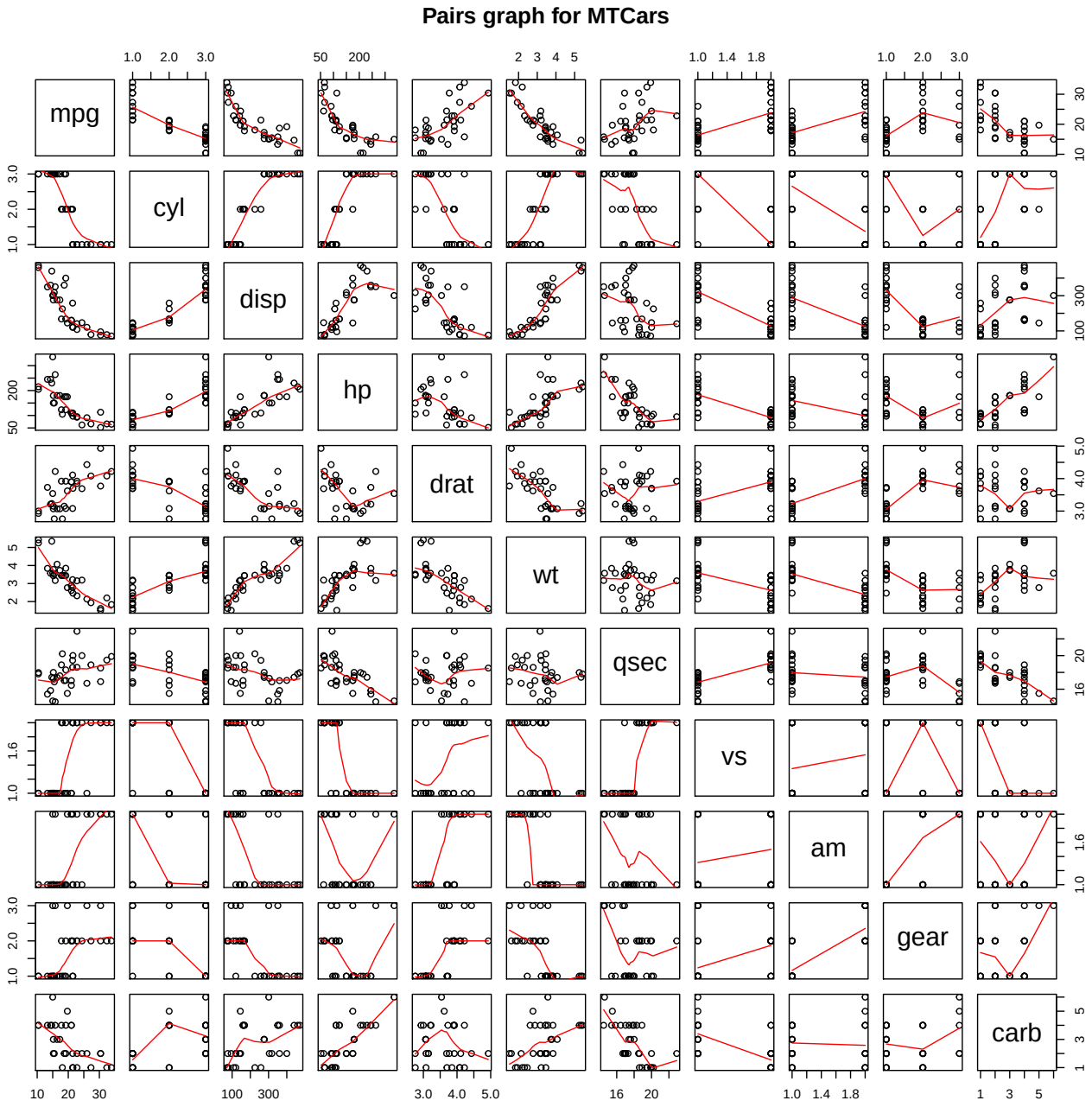
```
plot(mpg ~ am, data = mtcars)
title(main = "Mpg by transmission type", xlab = "am", ylab = "mpg")
```



We then plot the relationships between all the variables of the dataset (see below). We may note that variables like “wt”, “cyl”, “disp” and “hp” seem highly correlated together.

Figure : Manual transmission Pairs graph

```
pairs(mtcars, panel = panel.smooth, main = "Pairs graph for MTCars")
```



Statistical Inference

Let us perform few statistical tests to compare the mpg means between automatic and manual transmissions.

Wilcoxon test

Perform a non-parametric test to see if there's a difference in the population means.

```
wilcox.test(mpg ~ am, data = mtcars)
```

```
## Warning: cannot compute exact p-value with ties
```

The p-value of 0.0019 allows us to reject the null hypothesis that the mileage data of the manual and automatic transmissions are from the same population. This indicates there is a difference.

T-test

Perform a T-test assuming that mileage data has a normal distribution:

```
t.test(mpg ~ am, data = mtcars)
```

The p-value of 0.0014 clearly shows that Manual & Automatic transmissions are significatively different

Regression Analysis

Select the Bayesian Information Criteria (BIC) in a stepwise algorithm. This algorithm does not evaluate the BIC for all possible models but uses a search method that compares models sequentially. Thus it bears some comparison to the classical stepwise method; however this has the advantage that no dubious p-values are used:

```
model.all <- lm(mpg ~ ., data = mtcars)
n <- nrow(mtcars)
model.init <- step(model.all, direction = "backward", k = log(n))
```

```
summary(model.init)$coefficients
```

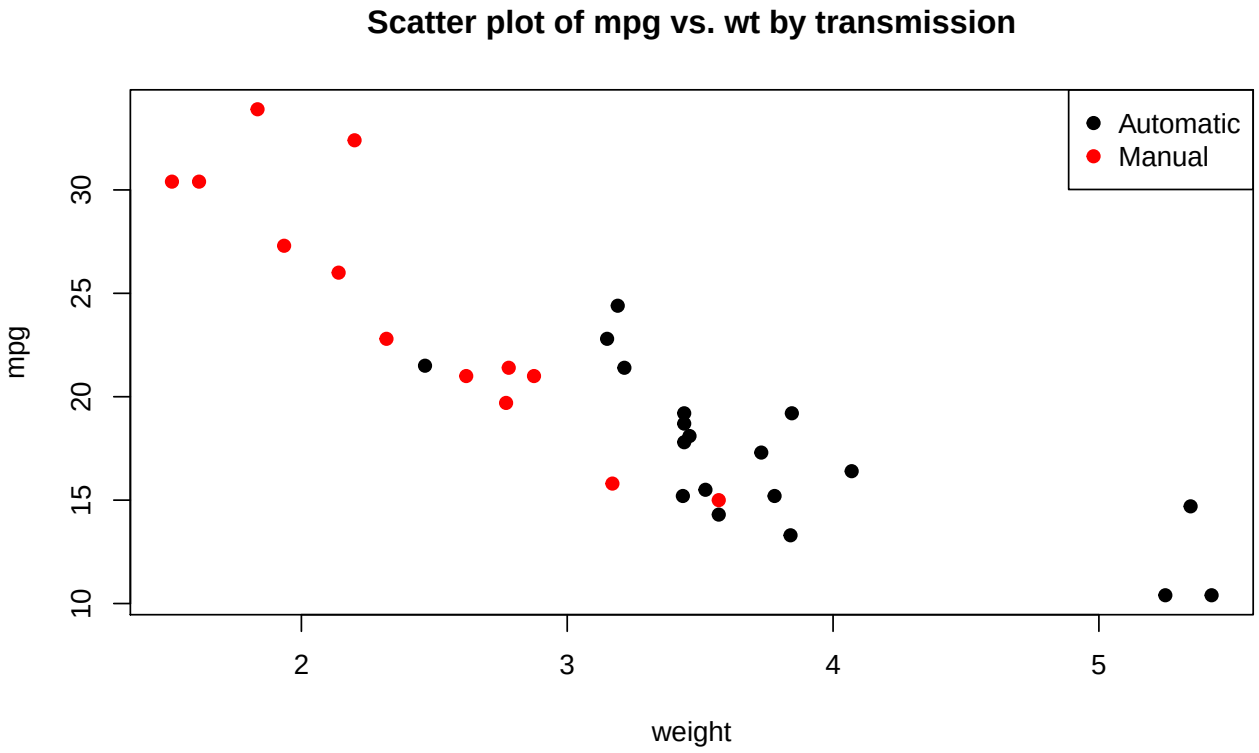
```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    9.618     6.9596   1.382 1.779e-01
## wt            -3.917     0.7112  -5.507 6.953e-06
## qsec           1.226     0.2887   4.247 2.162e-04
## amManual       2.936     1.4109   2.081 4.672e-02
```

The BIC algorithm tells us to consider “wt” and “qsec” as confounding variables. The individual p-values allows us to reject the hypothesis that the coefficients are null. The adjusted r-squared is 0.8336, so we may conclude that more than 83% of the variation is explained by the model.

However, if we take a look a the scatter plot of “mpg” vs. “wt” by transmission type (see below) we may notice that the “wt” variable depends on whether or not the car is automatic transmitted (as automatic transmitted cars tend to weigh more than manual transmitted ones). Apparently, manual transmission only confers an advantage to lighter cars. If the car is heavier than approximately 3 tons, an automatic transmission is actually more fuel-efficient than a manual one. This fact suggests that it would be appropriate to include an interaction term between “wt” and “am”.

Figure : Scatter plot of “mpg” vs. “wt” by type of Transmission

```
plot(mtcars$wt, mtcars$mpg, col = mtcars$am, pch = 19, xlab = "weight", ylab = "mpg")
title(main = "Scatter plot of mpg vs. wt by transmission")
legend("topright", c("Automatic", "Manual"), col = 1:2, pch = 19)
```




```
model <- lm(mpg ~ wt + qsec + am + wt:am, data = mtcars)
summary(model)$coefficients
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     9.723      5.899   1.648 0.1108925
## wt             -2.937      0.666  -4.409 0.0001489
## qsec            1.017      0.252   4.035 0.0004030
## amManual       14.079      3.435   4.099 0.0003409
## wt:amManual    -4.141      1.197  -3.460 0.0018086
```

The adjusted r-squared is now 0.8804, so we may conclude that more than 88% of the variation is explained by the model. We will choose this model as our final model.

```
anova <- anova(lm(mpg ~ am, data = mtcars), lm(mpg ~ am + wt, data = mtcars), model.init, model)
cbind(anova[1], anova[2], anova[3], anova[4], anova[5], anova[6])
```

```
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.9 NA        NA     NA        NA
## 2     29 278.3  1    442.58 101.89 1.161e-10
## 3     28 169.3  1    109.03  25.10 2.963e-05
## 4     27 117.3  1     52.01  11.97 1.809e-03
```

We may notice that when we compare the model with only “am” as independant variable and our chosen model, we reject the null hypothesis that the variables “wt”, “qsec” and “wt:am” don’t contribute to the accuracy of the model.

```
confint(model)[c(4, 5), ]
```

```
##                2.5 % 97.5 %
## amManual       7.031 21.128
## wt:amManual -6.597 -1.686
```

More accurately, we are 95% confident that the difference in miles per gallon between manual and automatic transmitted cars lies somewhere in the interval [7.0309 + -6.597 * wt, 21.128 + -1.6857 * wt].

Residuals and diagnostics

Residual analysis

We begin by studying the residual plots (see below). These plots allow us to verify some assumptions made before:

Figure : Residual plots

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