Motor Trend: Which is better for mpg: Automatic or Manual Transmission?

Pankaj Goyal

Tuesday, July 22, 2014

Summary

We have performed a set of regression test to study the relation between the type of transmission and the MPG on cars. We found that the number of cylinders, weight have the most significant effects on the MPG. Meanwhile, for smaller cylinders and weight, we found that the manual transmission has better MPG than automatic transmission when taking the most important effects into account. However, the overall difference is negligible. Therefore, the type of transmission is not the key variable that can significantly affact the mpg.

Data Analysis

Read in data and perform exploaratory analysis:

```
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 ...
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.4   Min.   :4.00   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.4   1st Qu.:4.00   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.2   Median :6.00   Median :196.3   Median :123.0  
##  Mean   :20.1   Mean   :6.19   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.8   3rd Qu.:8.00   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.9   Max.   :8.00   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.76   Min.   :1.51   Min.   :14.5   Min.   :0.000  
##  1st Qu.:3.08   1st Qu.:2.58   1st Qu.:16.9   1st Qu.:0.000  
##  Median :3.69   Median :3.33   Median :17.7   Median :0.000  
##  Mean   :3.60   Mean   :3.22   Mean   :17.8   Mean   :0.438  
##  3rd Qu.:3.92   3rd Qu.:3.61   3rd Qu.:18.9   3rd Qu.:1.000  
##  Max.   :4.93   Max.   :5.42   Max.   :22.9   Max.   :1.000
```

```
##        am              gear            carb
##  Min.   :0.000   Min.   :3.00   Min.   :1.00
##  1st Qu.:0.000   1st Qu.:3.00   1st Qu.:2.00
##  Median :0.000   Median :4.00   Median :2.00
##  Mean   :0.406   Mean   :3.69   Mean   :2.81
##  3rd Qu.:1.000   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :1.000   Max.   :5.00   Max.   :8.00
```

.1.Coercing "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)
```

.2.Rename the levels of the "am" variable into "Auto" and "Manual".

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

Graphics

We begin by plotting boxplots of the variable "mpg" when "am" is "Auto" or "Manual" (see Figure 1 in the appendix).

This plot hints at an increase in mpg when gearing was manual but this data may have other variables which may play a bigger role in determination of mpg.

We then plot the relationships between all the variables of the dataset (see Figure 2 in the appendix).

We may note that variables like "cyl", "disp", "hp", "drat", "wt", "vs" and "am" seem highly correlated to "mpg".

Inference

We may also run a some tests to compare the mpg means between automatic and manual transmissions.

T-test We begin by using a t-test assuming that the mileage data has a normal distribution.

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

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
##   mean in group Auto mean in group Manual
##                17.15                24.39
```

The test results clearly shows that the manual and automatic transmissions are significatively different.

Wilcoxon test Determine if there's a difference in the population means.

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

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  mpg by am
## W = 42, p-value = 0.001871
## alternative hypothesis: true location shift is not equal to 0
```

The Wilcoxon test also rejects the null hypothesis that the mileage data of the manual and automatic transmissions are from the same population (indicating a difference).

Regression analysis

First we need to select a model, we proceed by using AIC in a stepwise algorithm.

```
model.all <- lm(mpg ~ ., data = mtcars)
model <- step(model.all, direction = "both")
summary(model)
```

The AIC algorithm tells us to consider "cyl", "wt" and "hp" as confounding variables. The individual p-values allows us to reject the hypothesis that the coefiicients are null. The adjusted r-squared is 0.8401, so we may conclude that more than 84% of the variation is explained by the model.

```
model0 <- lm(mpg ~ am, data = mtcars)
anova(model0, model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     26 151  4       570 24.5 1.7e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 "cyl", "hp" and "wt" don't contribute to the accuracy of the model.

The regression suggests that, other variables remaining constant, manual transmitted cars can drive 1.8092 more miles per gallon than automatic transmitted cars, and the results are not statistically significant.

Residuals and diagnostics

Residual analysis We begin by studying the residual plots (see Figure 3 in the appendix). These plots allow us to verify some assumptions made before.

.1.The Residuals vs Fitted plot seem to verify the independance assumption as the points are randomly scattered on the plot. .2.The Normal Q-Q plot seem to indicate that the residuals are normally distributed as the points hug the line closely. .3.The Scale-Location plot seem to verify the constant variance assumption as the points fall in a constant band.

Dfbetas Next we look at the Dfbetas of the observations.

```
influential <- dfbetas(model)
```

Are any of the observations in the dataset influential ? We find the influential observations by selecting the ones with a dfbeta > 1 in magnitude.

```
influential[which(abs(influential) > 1)]
```

```
## numeric(0)
```

Conclusion

While a subtle relationship exists between am and mpg, this is insignificant and we cannot concluded that shifting from automatic to manual gearing will result in a car which gives better mileage. The reader is encourage to consider the car's weight and number of cylinders instead to determine the mileage of the vehicle.

Appendix

```
plot(mpg ~ am, data = mtcars, main = "Mpg by transmission type",
     xlab = "Transmission type", ylab = "Miles per gallon")
```

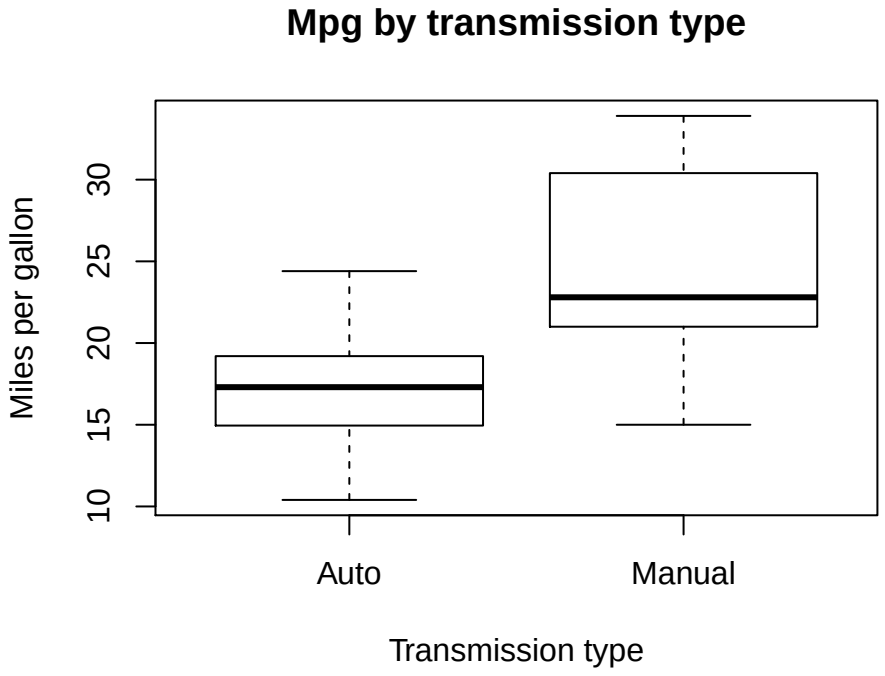

Figure 1 : Boxplots of "mpg" vs. "am"

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

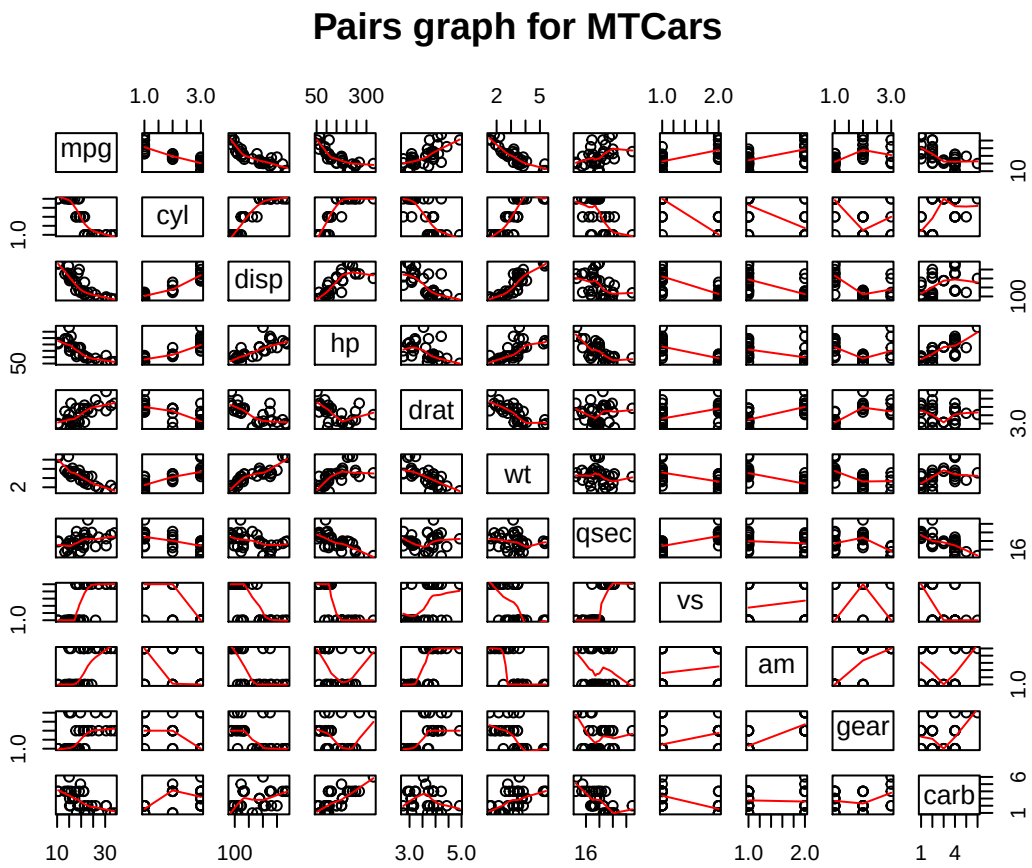

Figure 2 : Pairs graph

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

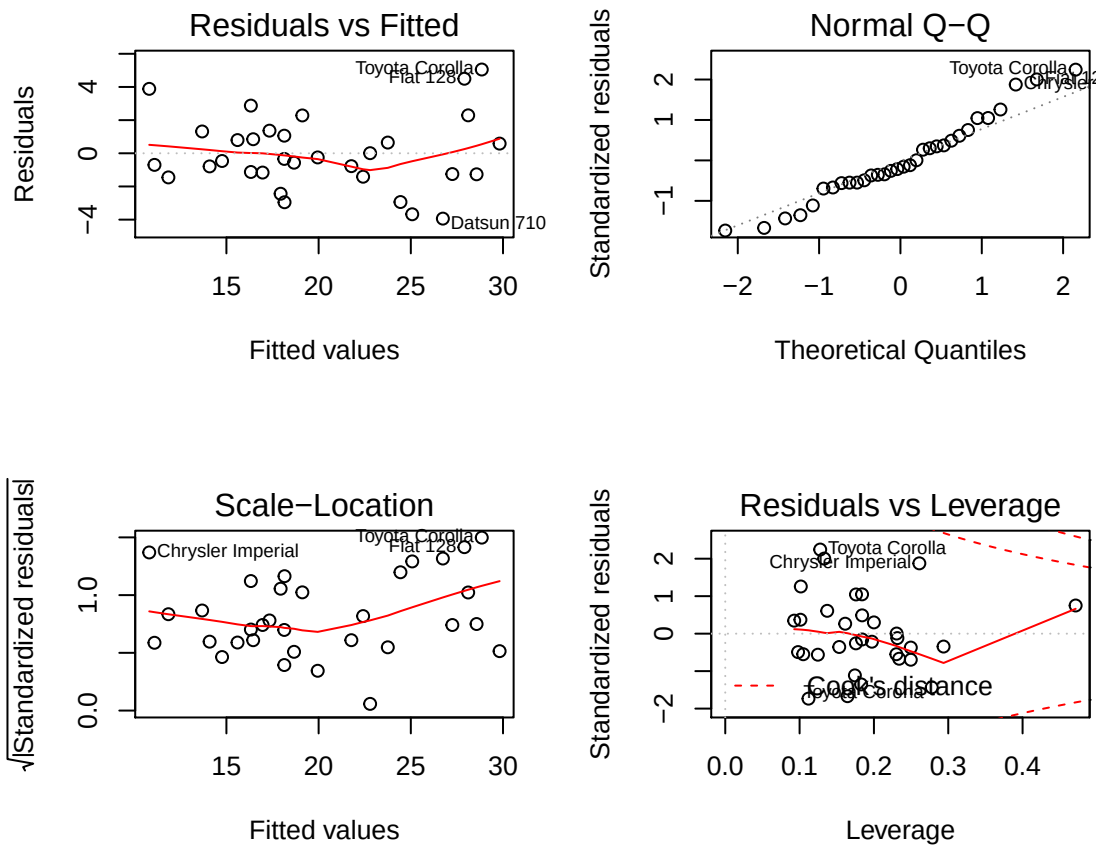

Figure 3 : Residual plots