Exploring the relationship between a transmission type and miles per gallon

Oleg Tsarev

EXECUTIVE SUMMARY

Motor Trend is a magazine about the automobile industry. Looking at a data set of a collection of cars, in this article we explore the relationship between a set of variables and miles per gallon (MPG). Particularly we will answer on the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions?

Based on our analyses, corresponding answers are:

1. Manual transmission are better for MPG.
2. The cars with manual transmission have 2.9 more miles per one gallon than cars with automatic transmission.

DATA PROCESSING

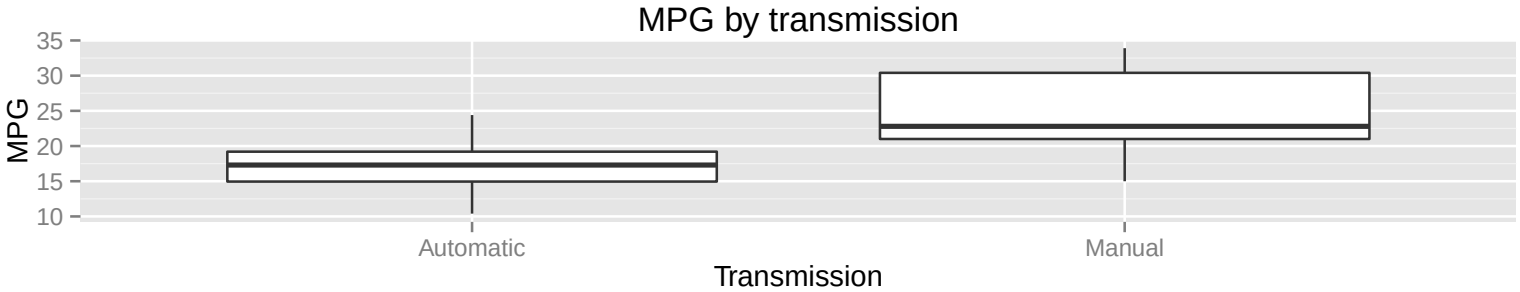
We will analyze dataset *mtcars* which comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). We transform variable with type of transmission (automatic or manual) from numeric type to factors.

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

EXPLORATORY DATA ANALYSIS

At first, let's look on the MPG distribution by automatic and manual transmission.

```
library(ggplot2)
q <- ggplot(data = mtcars_f, aes(x = am, y = mpg)) + geom_boxplot()
q + ggtitle('MPG by transmission') + labs(x = "Transmission", y = "MPG")
```



Above you can see that the mean of MPG with automatic transmission is lower than with manual transmission. Let's use hypothesis test to investigate it.

```
t.t <- t.test(mtcars_f[mtcars_f$am == "Automatic","mpg"],mtcars_f[mtcars_f$am == "Manual","mpg"]); t.t
```

Code listing you can see in the Appendix 1. The null hypothesis here is that the difference between the mean of MPG with automatic transmission and the mean of MPG with manual transmission is zero. As you can see p-value = 0.0013736 $< 5\%$, therefore we reject the null hypothesis. So we can say with 95% probability that mean of MPG with automatic transmission (17.15) is lower than mean of MPG with manual transmission (24.39). Therefore we can answer on the first question: **manual transmission is better for MPG**.

REGRESSION MODELS

Now let's quantify the MPG difference between automatic and manual transmissions. For this purpose we will create two regression models: a. simple linear regression, where MPG depends only on the transmission type; b. multiple linear regression. Our **strategy for model selection** is to choose model with the higher rate of explained variance and check it by using *anova* fucntion.

Simple linear regression

```
slr <- lm(mpg ~ am, data = mtcars_f); summary(slr)
```

Code listing you can see in the Appendix 2. Intepreting the outcome you can see that p-value $< 5\%$, so we can say that on average cars with manual transmission have 7.2 miles more per one gallon of fuel. But adjusted $R^2 = 0.3385$, so it means that we can explain only 34% of variance of mpg. It is not enough.

Multiple linear regression Let's use *step* function - automatic variable selection function:

```
mlr <- step(lm(mpg ~ ., data = mtcars),trace = 0, steps = 10000,direction = "both");
summary(mlr)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Intepreting the outcome you can see that in this model adjusted $R^2 = 0.8336$, so it means that we can explain 83% of variance of mpg.

Now let's compare these 2 models:

```
anova(slr,mlr)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Intepreting the outcome you can see p-value < 0.05. So second model better explains the variance of MPG.
In order to make final decision let's check residuals of the second model - Appendix 3. The residual diagnostics show normal randomly scatereness.

CONCLUSION

As we proved above, manual transmission is better for MPG. But quantifying the MPG difference between automatic and manual transmissions depends on the considering other variables in the model. We showed that multiple linear regression better explained variance of MPG than simple linear regression. So, based on the multiple regression we can say that **the cars with manual transmission have 2.9 more miles per one gallon than cars with automatic transmission**. And weight and acceleration have more impact than transmission type on MPG. So every 1000lb will cause a decrease of -3.9 miles per one gallon and every increase of 1/4 mile time will cause an increase of 1.2 miles per one gallon.

Appendix 1

```
t.t <- t.test(mtcars_f[mtcars_f$am == "Automatic","mpg"],mtcars_f[mtcars_f$am == "Manual","mpg"]); t.t
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars_f[mtcars_f$am == "Automatic", "mpg"] and mtcars_f[mtcars_f$am == "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
```

Appendix 2

```
slr <- lm(mpg ~ am, data = mtcars_f); summary(slr)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars_f)
##
## 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
```

Appendix 3

```
par(mfrow=c(2,2),mar=c(2,2,2,2))
plot(mlr)
```