Automobile Transmissions and Influence on MPG

Ryan Gavin

5/10/2017

Executive Summary

Fuel efficiency is important to automobile makers and consumers alike. A common measure of fuel efficiency is miles per gallon (**mpg**). In this report, we will aim to answer the following two items:

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

We will use the data included in the *mtcars* dataset* to complete our study. Conducting a Students' *t-test*, we find that cars with **manual** transmissions are more fuel efficient, statistically significantly so (p-*value* $= 0.001374$), than **automatic** transmissions by **7.245 MPG**.

However, cars are complicated machinces with many moving parts. To consider the influence of these confounding variables, we use simple multivariate linear regession to re-evaluate the effect of transmission type on fuel efficiency (mpg). From the analysis found below, by including the features `wt`, `hp`, `cyl` as well as `am`, we find an increase of only **1.478 MPG**. Although this is a smaller increase (yet still statistically significant), this multivariate model better describes the variation in `mpg` by **two and a half times** than the model consisting of only one regressor, `am`.

Data Processing

Import `mtcars` dataset, convert it to a tibble (personal preference), and convert `am` to a *factor* variable.

```
data("mtcars")
mtcars <- tbl_df(mtcars)
mtcars2 <- mtcars
mtcars2$am <- as.factor(mtcars$am)
levels(mtcars2$am) <- c("Automatic", "Manual")
```

Exploratory Analysis

First, we create a boxplot (Fig. A) showing `mpg` and transmission type (`am`) since we are looking at the relationship between `mpg` and `am`. Is there a *statistically* significant difference in `mpg` per `am`? Looking at the average `mpg` for each transmission type and the distribution of `mpg` for each `am` (Fig. B), we can begin to gain some insight and think about how to move forward in our analysis.

```
summarize(group_by(mtcars2,am),avg_mpg = mean(mpg))
```

```
## # A tibble: 2 × 2
##          am  avg_mpg
##      <fctr>    <dbl>
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

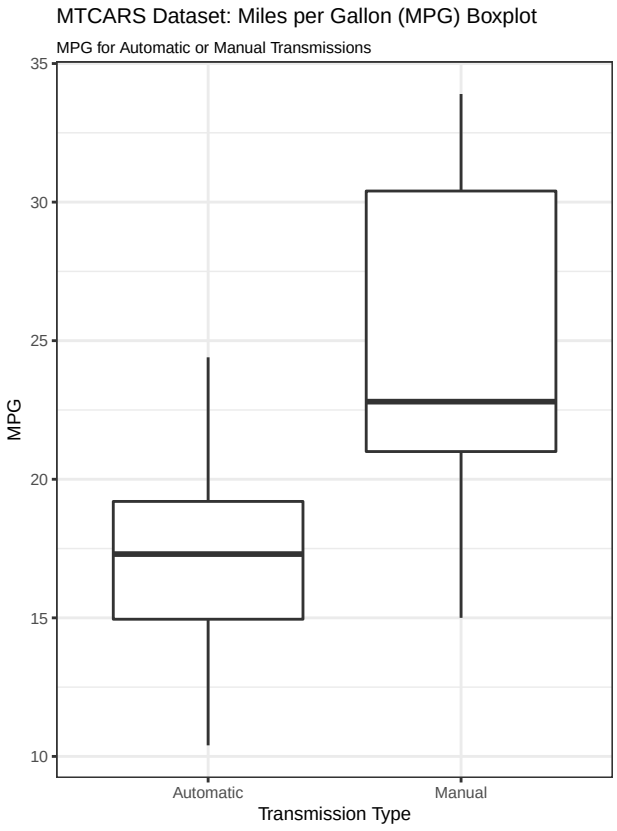


Fig. A

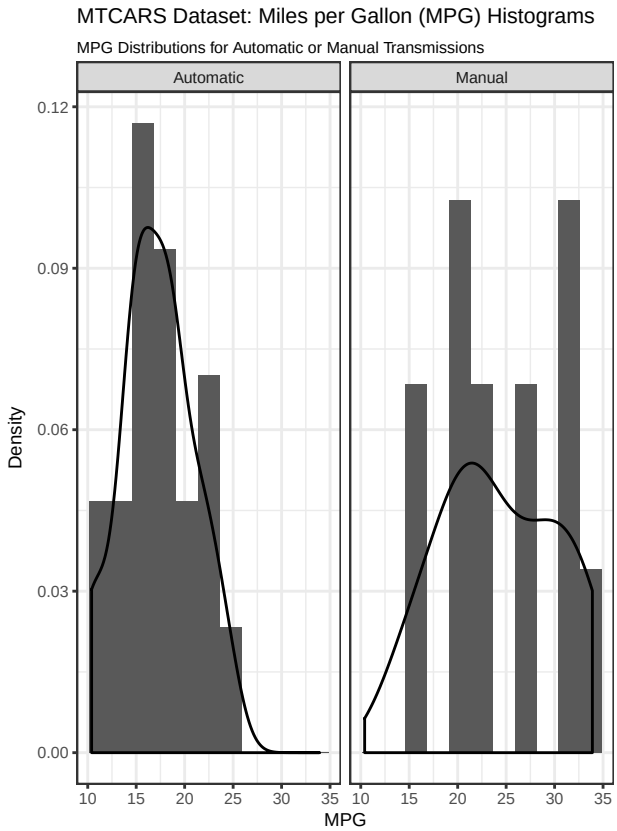


Fig. B

The distributions look relatively normal around their respective means, so we'll use Students' t-test to find the statistical significance.

Hypothesis Testing

With an $\alpha = 0.05$, we run the t-test:

```
t.test(filter(mtcars2,am=="Automatic")$mpg,filter(mtcars2,am=="Manual")$mpg,
             paired = FALSE,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(mtcars2, am == "Automatic")$mpg and filter(mtcars2, am == "Manual")$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
```

We see that the null hypothesis is rejected, and there is a statistically significant difference between MPG for automatic and manual transmission. It appears that manual transmissions, on average, get 7.245 mpg more than automatic transmissions. The resultant *p-value* is 0.001374.

Linear Regression Analysis

We conduct a similar analysis using linear regression.

```
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 ***
## am             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
```

Our simple linear regression seems to show what we already know:
1. y-intercept = average mpg for automatic transmissions
2. *slope* = average mpg increase when going from automatic transmission to manual

However, $R^2 = 0.3598$. This tells us that the variable `am` only explains ~36% of the variation in `mpg`. In some respect, this should make sense. The "type of transmission" is only *one* variable that can influence miles per gallon. Therefore, we should expand our analysis to include any confounding variables to help explain the variation in `mpg`.

Multivariate Regression Analysis

First, let's examine the correlation between various factors and `mpg` (Fig. C), as well as factors and `am` (Fig. D). Correlations have been arranged from greatest to least.

```
cor_list_mpg <- cor(mtcars)[order(desc(abs(cor(mtcars)[,1]))),1][-1]
cor_list_am <- cor(mtcars)[order(desc(abs(cor(mtcars)[,9]))),9][-1]
```

In building our model, we want to include variables that are highly correlated with `mpg`, but also those that are uncorrelated with `am`. This is because variables that we include that are highly correlated with the transmission type introduces unnecessary linear dependence between regressors.

So, we start with the 5 most correlated variables to `mpg` (`wt`, `cyl`, `disp`, `hp`, `drat`) and the 5 least correlated variables to `am` (`carb`, `vs`, `qsec`, `hp`, `cyl`) and pick those that overlap: `cyl` and `hp`. Given the high correlation between `wt` and `mpg` (-0.8676594), we choose to include `wt` as well.

```
fit2 <- lm(mpg ~ am + cyl + hp + wt, data = mtcars)
```

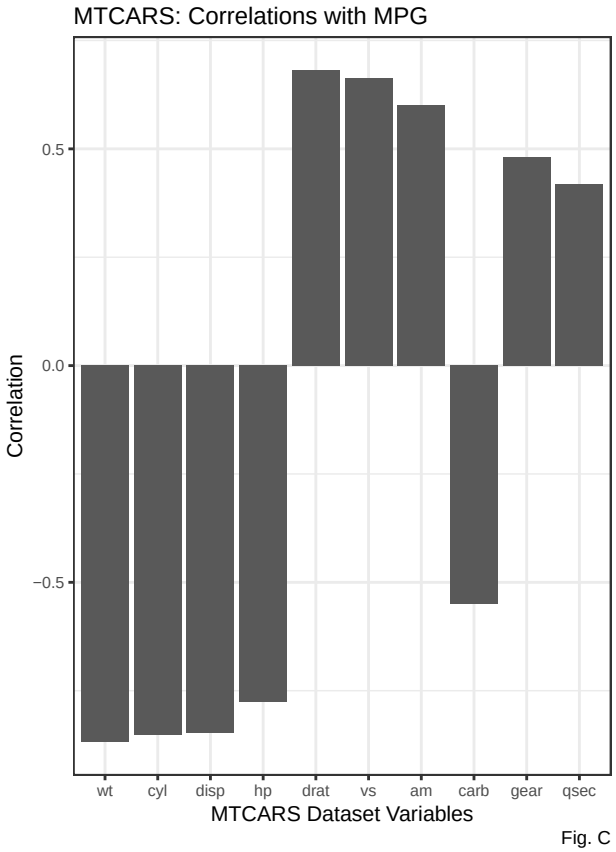


Fig. C

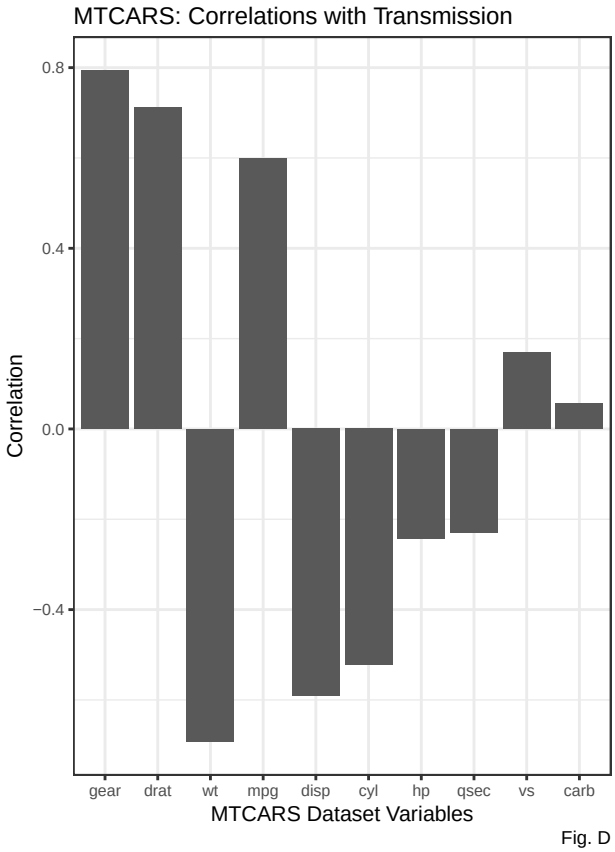


Fig. D

We now have two models that were constructed with the same data. Let's perform an **ANOVA** to compare the models.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + hp + wt
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.9
## 2     27 170.0  3     550.9 29.166 1.274e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly the multivariate model does a statistically significant better job of fitting the data (p-value = 1.274e-08). The summary of our multivariate model tells us that the average increase in miles per gallon when going from automatic to manual transmission is **1.478 MPG**. Our model also does a nice job of describing the variation in `mpg`, with an R^2 value of 0.849.

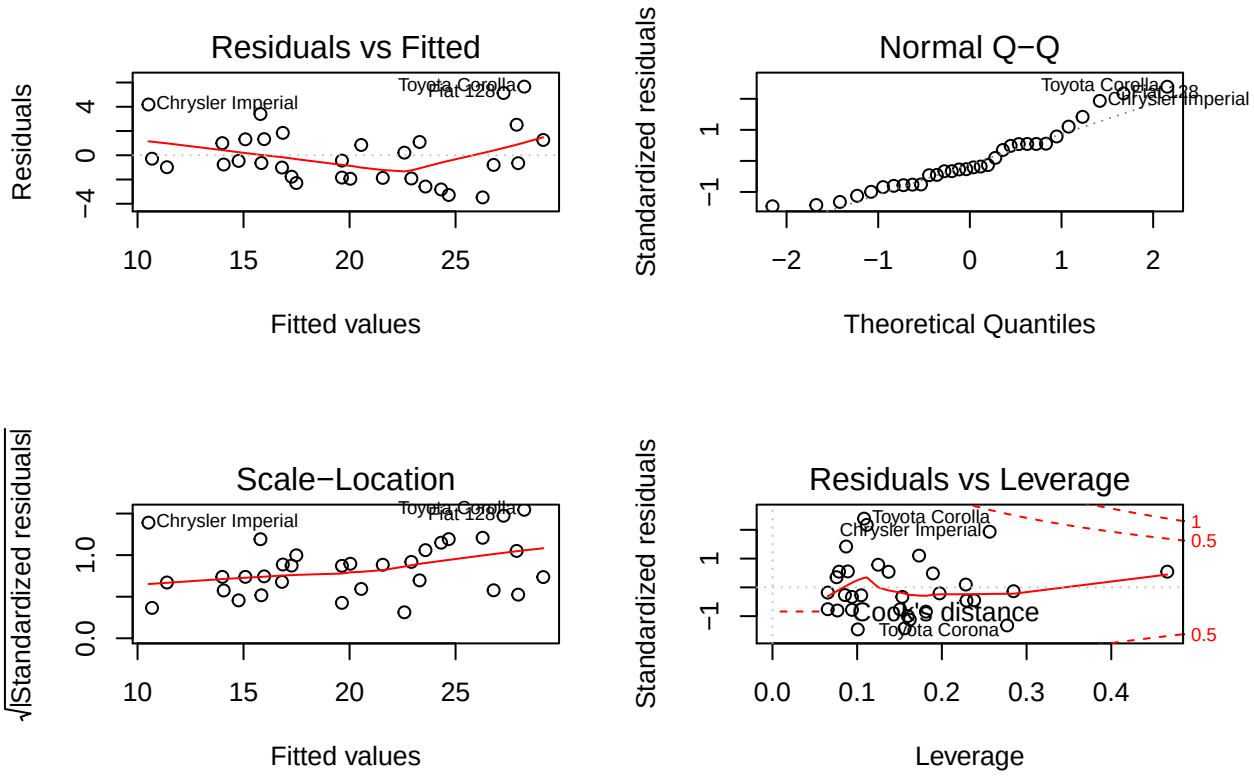
```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## am           1.47805    1.44115   1.026   0.3142
## cyl         -0.74516    0.58279  -1.279   0.2119
## hp          -0.02495    0.01365  -1.828   0.0786 .
## wt          -2.60648    0.91984  -2.834   0.0086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

Model Diagnostics

```
par(mfrow = c(2,2))
plot(fit2)
```



* The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, **37**, 391–411.