Effect of Vehicle Transmission on Fuel Efficiency

Jerry Lakin

11/18/2020

Executive Summary

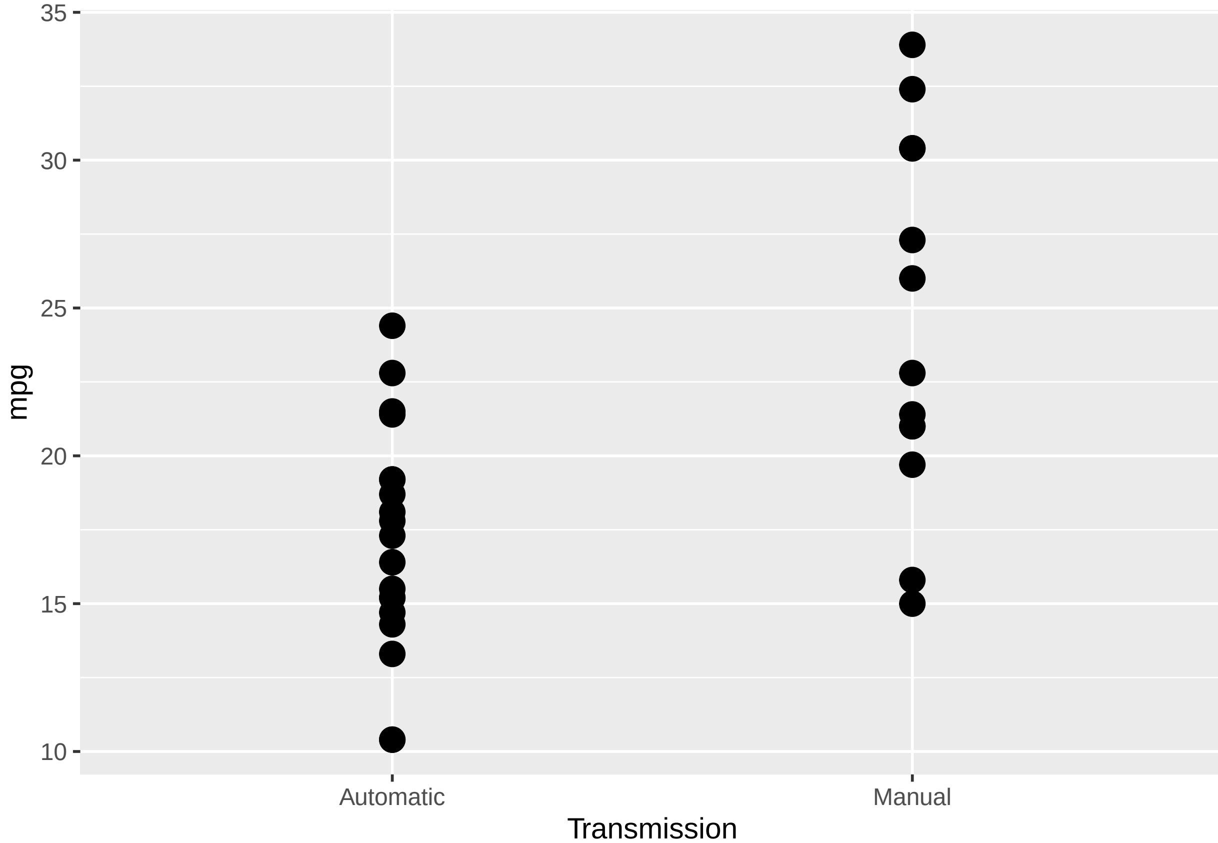
In this project we will be investigating the relationship between a vehicle's transmission and fuel efficiency, measured in miles per gallon. We will be using the mtcars dataset which comes preloaded in R. This data comes from the 1974 Motor Trend magazine in the United States. In this dataset, miles per gallon is contained in the column mpg, while transmission is contained in the column am, with a value of 0 representing automatic transmission and a value of 1 representing manual transmission.

At first glance it would appear that transmission does indeed impact mpg, with manual cars tending to be more fuel efficient by 7.2 mpg. However, after accounting for other variables in the dataset that correlate with both mpg and transmission, this impact shrinks significantly to only 1.2 mpg. Furthermore, in the model that incorporates additional variables, the relationship between transmission and mpg is not deemed to be statistically significant. Therefore the conclusion of our study is that we cannot confirm the degree to which transmission impacts mpg with the data provided. The only variables that we confirm to have a statistically significant impact on mpg are the weight of the car and the number of cylinders in the engine.

Analysis

We will begin by looking at a simple scatter plot comparing transmission and mpg in order to get an idea of how they might be related.

```r
data(mtcars)
library(ggplot2)
ggplot(mtcars, aes(x = factor(am, levels = c(0,1), labels = c("Automatic", "Manual")), y = mpg)) +
  geom_point(size = 4, color = "black") +
  xlab("Transmission")
```



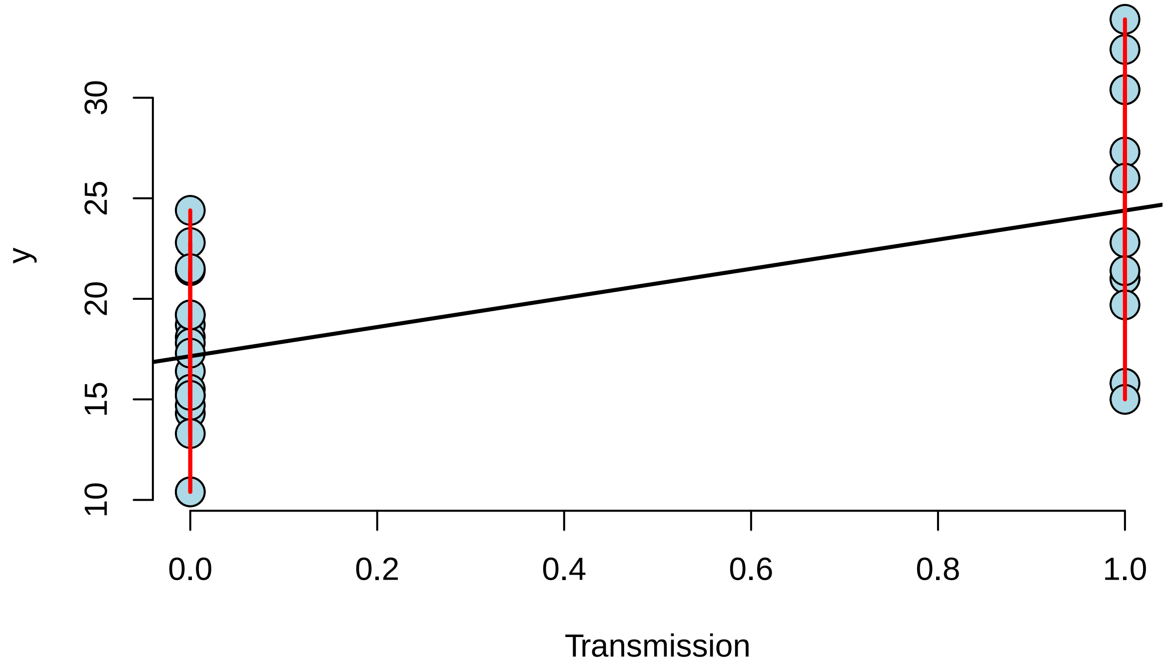
Based on this first glance, it appears that manual transmission cars may have a higher mpg than automatic transmission cars. As a simplistic first pass, we will fit a linear model using only transmission.

```
fit1 <- lm(mpg ~ am, data = mtcars)
fit1$coefficients
```

```
## (Intercept)          am
##   17.147368    7.244939
```

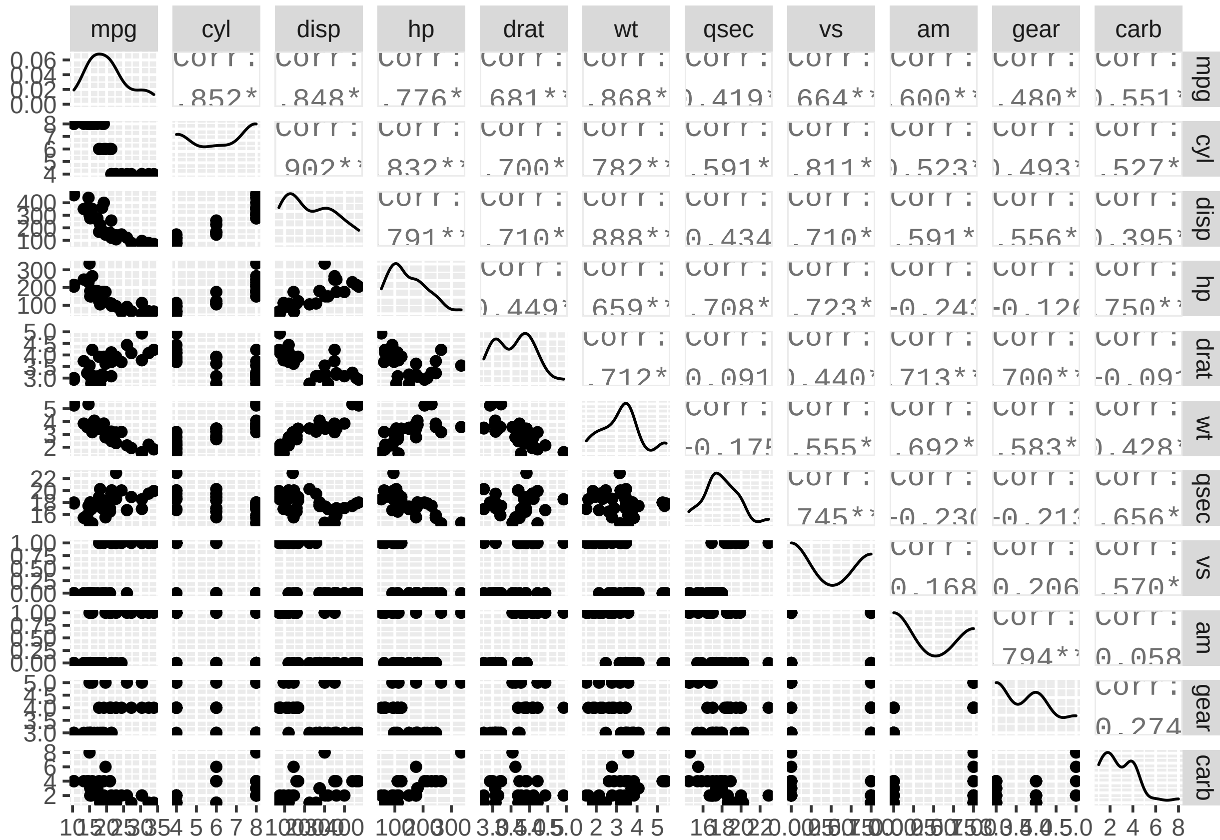
Now let's look at a residual graph to see how close our predictions came to the actual values.

```
e <- resid(fit1)
yhat <- predict(fit1)
x <- mtcars$am; y <- mtcars$mpg
n <- length(x)
plot(x, y,
     xlab = "Transmission",
     bg = "lightblue",
     col = "black", cex = 2, pch = 21,frame = FALSE)
abline(fit1, lwd = 2)
for (i in 1 : n)
  lines(c(x[i], x[i]), c(y[i], yhat[i]), col = "red" , lwd = 2)
```



The fit doesn't appear to be very good. Furthermore, in order to be thorough and avoid Simpson's paradox we must investigate the other variables in the dataset and how they might correlate with both mpg and transmission. In order to start, we'll look at a pair plot showing the relationship between each of our variables.

```
library(GGally)
ggpairs(mtcars, lower = list(continous = "smooth"))
```



Looking at these pair plots we can see that each of the variables in the dataset appears to be correlated with mpg, though some are more strongly correlated than others. Furthermore, some of these variables are also correlated with transmission. For example, the weight of the car is strongly correlated with mpg, with heavier cars exhibiting lower fuel efficiency. The weight of the cars are also correlated with transmission, with automatic cars tending to be heavier than manuals. When considering the impact of transmission, we will need to account for variables like weight.

Rather than fitting a model with all of the variables in the dataset, we will only look at those variables which are strongly correlated with both mpg and transmission. As a cutoff, let's only look at variables that have correlation coefficients with mpg and transmission that are at least 0.25 in magnitude. This leaves us with cylinder count (cyl), displacement (disp), rear axle ratio (drat), weight (wt), and number of forward gears (gear).

```r
fit2 <- lm(mpg ~ am + cyl + disp + drat + wt + gear, data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + drat + wt + gear, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.636 -1.446 -0.466  1.674  5.604
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.849311   7.740361   5.536 9.38e-06 ***
## am           1.181576   1.878725   0.629  0.53510
## cyl         -1.740815   0.654740  -2.659  0.01348 *
## disp         0.005705   0.012481   0.457  0.65154
## drat         0.331103   1.618600   0.205  0.83957
## wt          -3.425362   1.226050  -2.794  0.00985 **
## gear        -1.072365   1.153165  -0.930  0.36130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.699 on 25 degrees of freedom
## Multiple R-squared:  0.8383, Adjusted R-squared:  0.7995
## F-statistic:  21.6 on 6 and 25 DF,  p-value: 9.105e-09
```

After accounting for these additional variables, the model is still predicting that manual transmission cars will have a higher mpg than automatic cars, but the predicted difference fell from 7.2 mpg to 1.2 mpg. More importantly, the p-value for the relationship between transmission and mpg in this model is just 0.54. This means that we cannot state with an acceptable degree of confidence that transmission has an impact on mpg, after accounting for other variables. The only two variables that we can confirm to have a statistically significant impact on mpg are the weight of the car and the number of cylinders in the engine.