An Analysis of the Effect of Transmission on MPG

Matthew Halbasch

February 28, 2019

Summary

In this report, we analyze the effect that the transmission (automatic or manual) has on the mpg a car achieves. We consider a multiple possible models to discern if there is a statistically significant relationship between the transmission and the miles per gallon after accounting for potential confounding variables. We find that while manual transmission cars do achieve higher mpg, the difference is almost entirely accounted for by the weight of the cars. Once we account for weight, there is no longer any noticeable change in mpg by type of transmission.

Data Processing and Exploratory Analysis

We use data from the 1974 *Motor Trend* magazine, which is available through the `mtcars` dataset in R. To start, we load this data in and make it slightly easier to work with.

```
data(mtcars)
mtcars %>% mutate(vs = ifelse(vs == 0, "V-Shaped", "Straight")) %>%
  mutate(am = ifelse(am == 0, "Automatic", "Manual")) ->
  cleaned_cars
```

Now we are able to conduct some exploratory analysis of this data (code and figures for this can be found in the Appendix). Looking at Figure 1, we quickly see that the mpg a car achieves can potentially depend on many continous variables, with weight and horsepower seemingly the most significant. Onthe other hand, Figure 2 shows that the relationship is not too obvious for some of the discrete variables, with the exception of the number of cylinders.

Looking specifically at the plot of mpg versus transmission in Figure 2, we see an apparent increase in mpg for manual transmissions, though the size of this effect is unclear.

Model Fitting

From the exploratory analysis above, it is clear that many variables could act as potential confounders when investigating the relationship between mpg and transmission. We must keep this in mind, but first we should try to find if there is any statistically significant relationship between mpg and transmission when we ignore these potential confounders.

To do this, we fit a simple linear model with only these two variables.

```
basic_fit <- lm(mpg ~ am, data=cleaned_cars)
tidy(basic_fit) %>%
  xtable(digits=c(0,0,2,2,2,-3)) %>% print(include.rownames=FALSE)
```

term	estimate	std.error	statistic	p.value
(Intercept)	17.15	1.12	15.25	1.134E-15
amManual	7.24	1.76	4.11	2.850E-04

The intercept printed here is the estimated value of mpg for an automatic transmission car, while the `amManual` term is the *added* mpg for a manual tranmission over an automatic one. The value of this difference in mpg is about 7.24, with a p-value of $p < .001$, so our simple model tells us the difference between these transmissions is statistically significant.

Now we return to our concern over confounding variables. In fact, Figure 3 shows that there is a strong relationship between weight and type of transmission, which makes it a potential confounder. We can investigate this relationship more clearly with an anova table:

```
tidy(aov(mpg ~ wt + am, data=cleaned_cars)) %>%
  xtable(digits=c(0,0,0,0,0,1,2,-3)) %>% print(include.rownames=FALSE)
```

term	df	sumsq	meansq	statistic	p.value
wt	1	848	847.7	88.33	2.645E-10
am	1	0	0.0	0.00	9.879E-01
Residuals	29	278	9.6		

```
# As a check, we can run it the other way around
tidy(aov(mpg ~ am + wt, data=cleaned_cars)) %>%
  xtable(digits=c(0,0,0,0,0,1,2,-3)) %>% print(include.rownames=FALSE)
```

term	df	sumsq	meansq	statistic	p.value
am	1	405	405.2	42.22	4.110E-07
wt	1	443	442.6	46.12	1.867E-07
Residuals	29	278	9.6		

The value of most interest to us is the p-value in these tables, which is an indication of whether the additional term provides a statistically significant contribution over the other terms. The two tables show illustrate the issue of a confounding variable: in the first case, once we account for weight, transmission has no noticeable effect on mpg. However, in the second case, we see that even accounting for transmission, weight is an extremely significant addition to our model. This suggests that weight is the only effect we should consider relevant between these two variables, since transmission provides only a subset of this information.

For diagnostics, we have also plotted the residuals from the fit just according to weight in Figure 4, colored by transmission. We see no noticeable patterns and no relevant difference between the two colors of dots, supporting the result that transmission has no additional effect.

Conclusion

We have found that the transmission of a car does appear to have a statistically significant relationship with the mpg a car achieves (with a manual tranmission having better mileage). However, this relationship is almost entirely explained by the weight of the car in question, which has a more significant relationship with mpg on its own.

This suggests that the transmission of the car is not on its own relevant to the mpg, but rather the weight of the car that is produced should be considered (along with potentially other variables not considered here.)

Holding these variables fixed we have found no evidence that the transmission affects the mpg of a car.

Appendix: Code and Figures

Here we provide the figures referenced in the document and some of the code used to produce them. Some code has been omitted to keep this section concise.

First we provide the code for Figure 1, an exploratory plot of the continous variables. The code for Figure 2 is similar, but omitted for space.

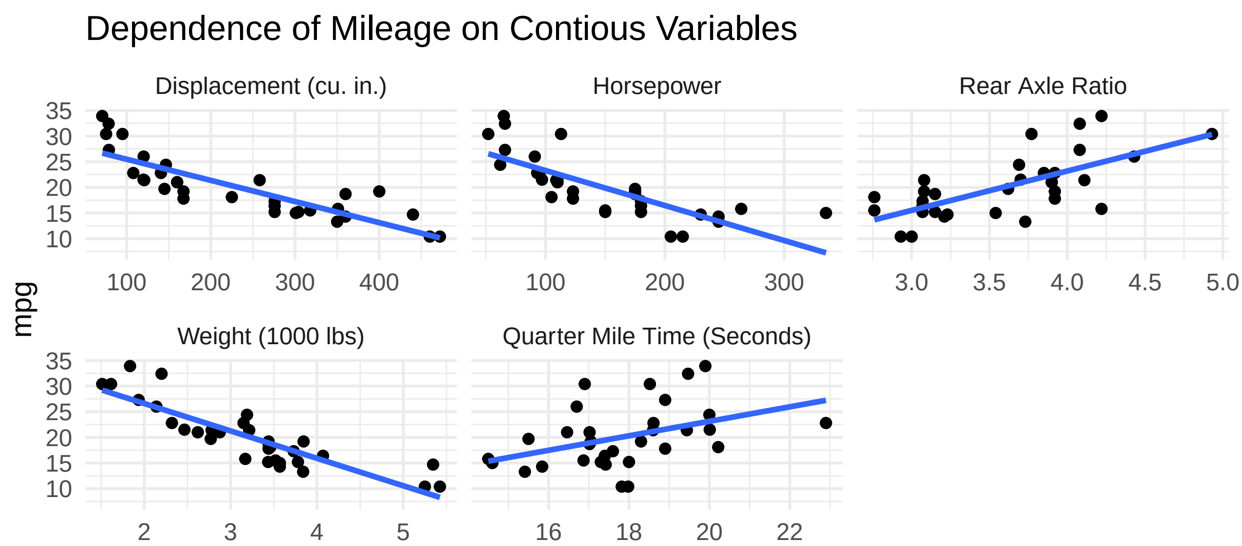

Figure 1: Plots of the continous variables in the dataset. We see that displacement, horsepower and especially weight all seem to be negatively correlated with mpg, while the rear axle ratio and to a lesser extent the 1/4 mile time are positively correlated.

```
# Split the data into continous and discrete variables for easy plotting
continuous = c("disp", "hp", "drat", "wt", "qsec")
discrete = c("cyl", "vs", "am", "gear", "carb")
melted_data <- melt(cleaned_cars, id.vars="mpg")

cont_data <- filter(melted_data, variable %in% continuous)
discrete_data <- filter(melted_data, variable %in% discrete)

cont_names <- c('disp' = "Displacement (cu. in.)",
              'hp' = "Horsepower",
              'drat' = "Rear Axle Ratio",
              'wt' = "Weight (1000 lbs)",
              'qsec' = "Quarter Mile Time (Seconds)")

ggplot(cont_data, aes(as.numeric(value), mpg)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  facet_wrap(vars(variable), scales="free_x",
             labeller=as_labeller(cont_names)) +
  theme_minimal() +
  xlab(element_blank()) +
  labs(title = "Dependence of Mileage on Contious Variables")
```

Next we provide the code for Figure 3, the plot of mpg versus weight, colored by transmission.

```
ggplot(cleaned_cars, aes(x=as.numeric(wt), y=mpg, color=am)) +
  geom_point() +
  xlab("Weight (1000 lbs)") +
  labs(title = "Dependence of mpg on Weight and Transmission") +
  theme_minimal() +
```

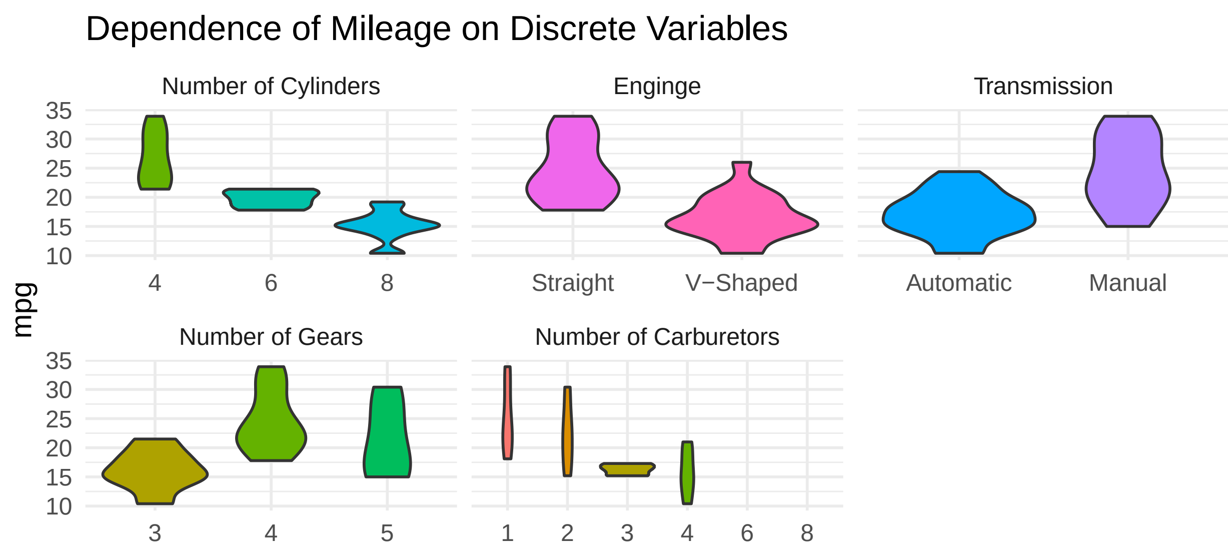

Figure 2: Plots of the discrete variables in the dataset. We see that the number of cylinders has a noticeable effect on the mpg of a car, but other variables are less clear. We are particularly interested in the Transmission plot here.

```
  labs(color = "Transmission")
```

Finally, we provide the code for Figure 4, the residual plot from the weight fit, with coloring by transmission type.

```
weight_fit <- lm(mpg ~ wt, data=cleaned_cars)

select(cleaned_cars, am) %>%
  mutate(residuals = weight_fit$residuals,
         fitted_values = weight_fit$fitted.values) %>%
  ggplot(aes(x=fitted_values, y=residuals, color=am)) +
  geom_point() +
  xlab("Fitted mpg") +
  ylab("Residuals") +
  labs(color = "Transmission") +
  labs(title = "Residuals After Accounting for Weight")
```

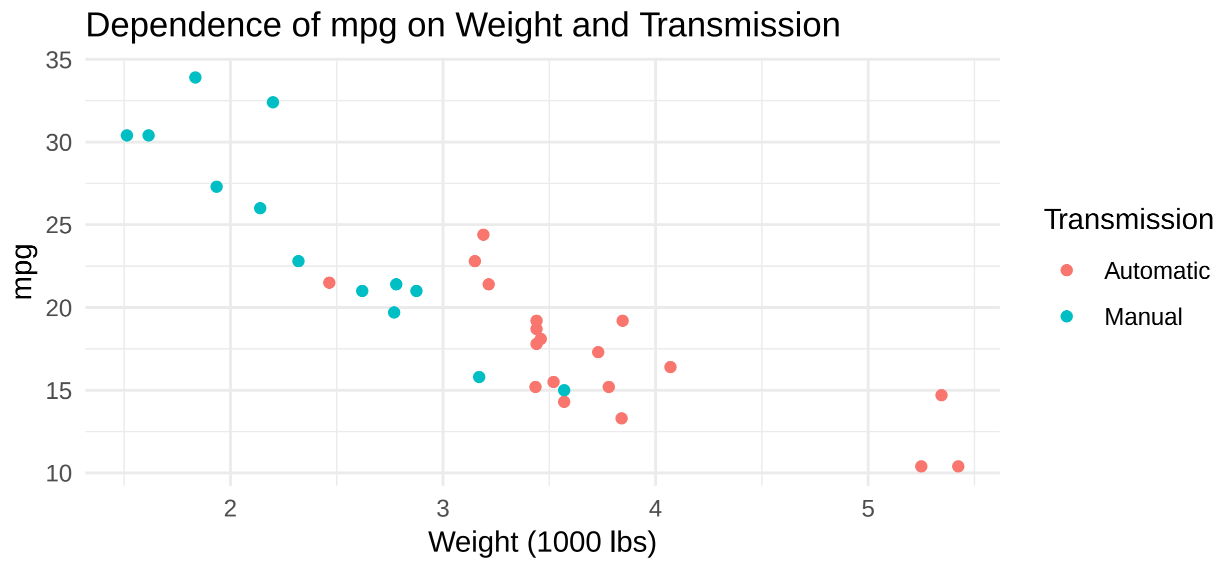

Figure 3: A plot of mpg versus weight, colored by the type of transmission. The plot shows that manual transmission cars tend to be significantly lighter than automatic transmission cars. Since there is a strong dependence of mpg on weight, this indicates that weight could be a confounding variable for the effect of transmission on mpg.

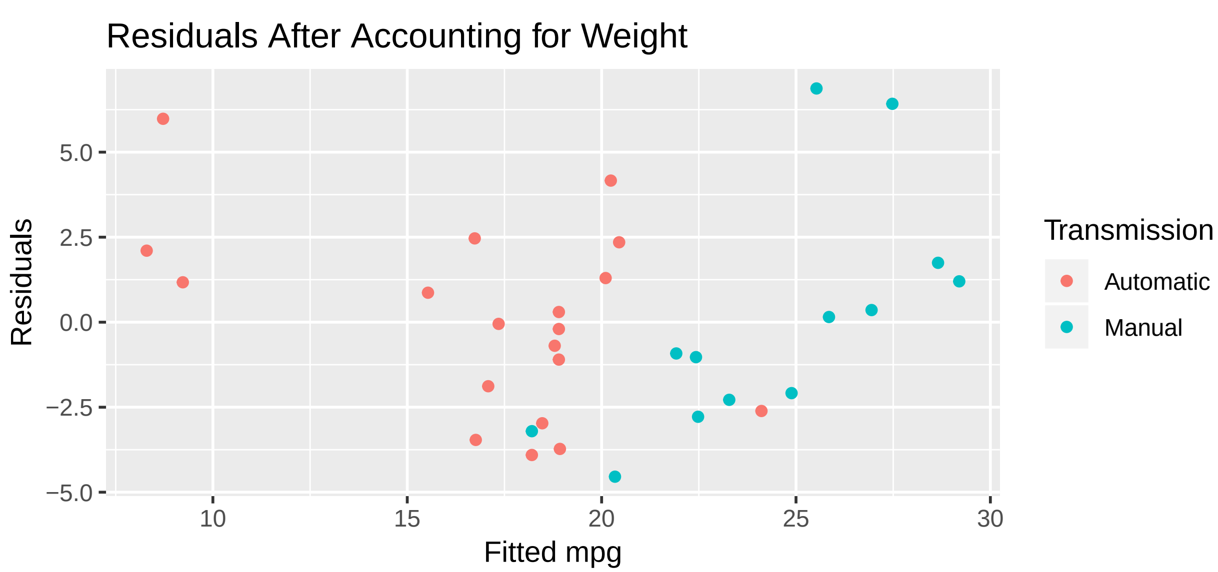

Figure 4: A plot of the residuals from the linear fit of just the weight of the car, colored according to the transmission. We see that higher mpg values correspond to Manual transmission, but the residuals appear no different than those of automatic transmissions.