Report on MPG and Transmission

Raymond Dineen

10/23/2020

Executive Summary

This exploration into the relationship between transmission type (automatic and manual) and miles per gallon (MPG) shows on the surface there may be important connection between the two variables. However, once we digger a little deeper into the `mtcars` dataset, we can easily realize that transmission is a relatively insignificant variable for predicting MPG. Our models, after accounting for other variables such are weight and number of cylinders, fail to reject the null hypothesis that the transmission's effect on MPG is 0. Instead, it is better to use the other variables and disregard transmission. There is also the issue of the data not having high-weight manual transmission cars or low-weight automatic transmission cars. We can conclude that **technically**, based on this data, manual cars are more fuel efficient than automatic cars but also that this correlation is unlikely to be causation.

Exploratory Analysis

Our first step after loading in the data is to make some basic plots to get a better feel for it. Since we are looking specifically for the difference in MPG between transmission types, we can make a boxplot (Figure 1) to get an initial impression. Based on this boxplot, we might assume that manual transmission generally have better MPG. Before we jump to conclusions, it may help to see which other variable in the dataset are correlated with mpg since that is ultimately what we are interested in.

```r
data(mtcars)
cor(mtcars$mpg, mtcars[,-1])
```

```
##             cyl       disp         hp      drat         wt     qsec        vs
## [1,] -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684 0.6640389
##             am      gear       carb
## [1,] 0.5998324 0.4802848 -0.5509251
```

We can see here that weight (wt) is the most correlated with MPG while transmission (am) is around middle of the pack. We can make another plot (Figure 2) that shows MPG vs. weight with different colored points representing the two transmission types. There is also a regression line fitted over it which predicts MPG using log-transformed weight as the regressor. The graph on the right shows separate regression lines for each transmission type. This plot makes apparent an important shortcoming of the data. We don't have data on high-weight manual transmission cars or low-weight automatic transmission cars. This data could help quite a bit in coming to a sure conclusion. Regardless, our individually grouped regression models have significant overlap. It is clear from looking at this that weight likely explains more of the variance than transmission.

Modeling MPG

Our initial model for MPG can be basic using only the variable of interest, transmission.

```
fit1 <- lm(mpg ~ am, data = mtcars); coef(summary(fit1)); summary(fit1)$adj.r.squared
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

```
## [1] 0.3384589
```

We get an estimated 7.24 MPG increase in MPG from using a manual transmission with a passing t-test to back it up. However, our adjusted r-squared is quite low. We already saw that weight was an important regressor so we will add that to the model and see the effect.

```
fit2 <- lm(mpg ~ am + log(wt), data = mtcars); coef(summary(fit2)); summary(fit2)$adj.r.squared
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  41.360850   2.928252 14.124757 1.567231e-14
## am           -1.253659   1.392451 -0.900326 3.753629e-01
## log(wt)     -18.507804   2.188421 -8.457150 2.554396e-09
```

```
## [1] 0.802571
```

This model shows that for every 1000 lb increase in weight, MPG is expected to decrease by about 18.5%. Our r-squared has jumped up a very significant amount. We also see that the sign has flipped on our manual transmission coefficient and it's effect has decreased. The t-test for this variable also now strongly suggests that transmission is insignificant in the model when we adjust for weight. It would likely be best to remove it from the model to reduce variance. Before we do that, we can check a residual plot of this model (Figure 3) to see if there are any trends. The residuals look decent but could maybe be better if we added a good regressor. Saw that cylinders was the second-highest correlated variable with MPG so we can try adding it, this time without transmission.

```
fit3 <- lm(mpg ~ log(wt) + cyl, data = mtcars); coef(summary(fit3)); summary(fit3)$adj.r.squared
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  40.649777  1.6103983 25.242065 2.747211e-21
## log(wt)     -11.523812  2.2226703 -5.184670 1.518181e-05
## cyl          -1.233526  0.3951429 -3.121721 4.049499e-03
```

```
## [1] 0.8480977
```

Our 2 regressors now both pass the t-test and the adjusted r-squared further rose to about 0.85. Upon looking at the residual plot (Figure 4), we see that the fit see better. It's not perfect, but at this point we've answered our questions about transmission's effect on MPG.

Conclusion

We've shown confidently that transmission doesn't has a statistically insignificant impact on MPG. Although there is a correlation, our models show that it is unlikely to be causation. There is some uncertainty in this claim since our data is missing high-weight manual and low-weight automatic cars. Based on what we have though, we can say that weight and number of cylinders are much better predictors of MPG.

Appendix

Figure 1:

```
data(mtcars)
mtcars$am[mtcars$am == 0] <- "automatic"
mtcars$am[mtcars$am == 1] <- "manual"

library(ggplot2)
ggplot(data = mtcars, mapping = aes(factor(am), mpg)) +
    geom_boxplot() +
    xlab("transmission")
```

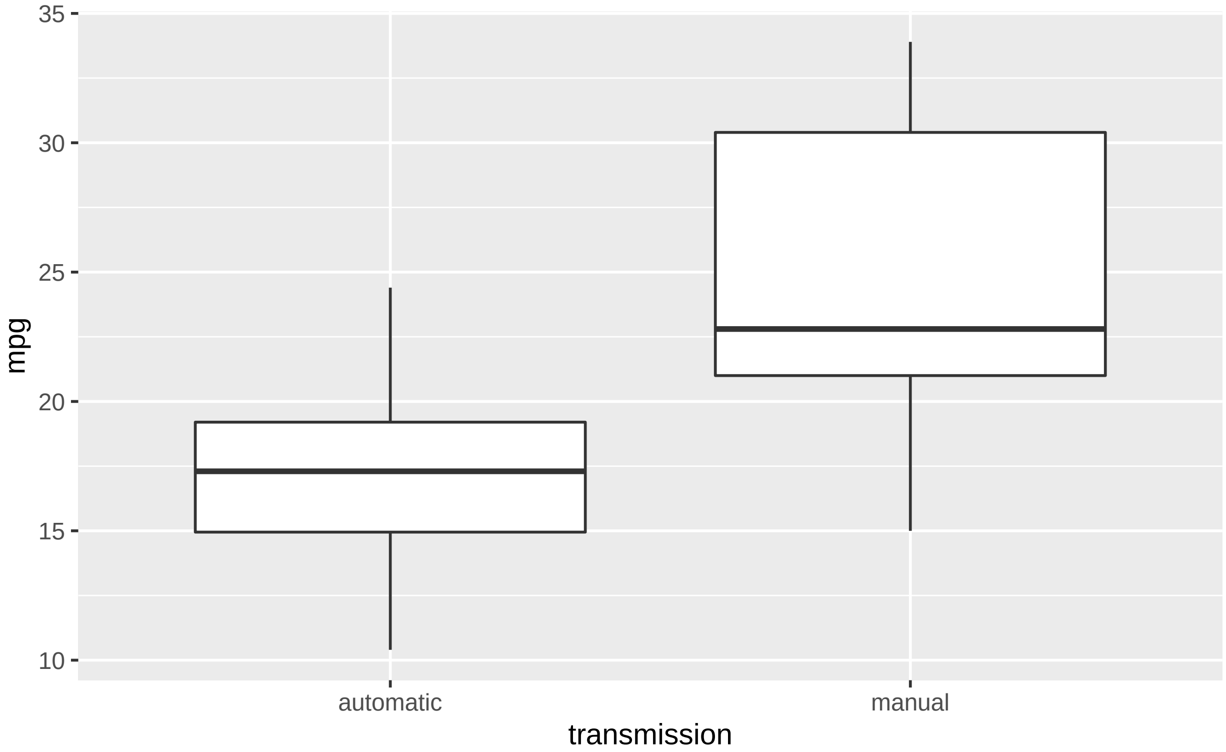


Figure 2:

```
library(ggpubr)
plot1 <- ggplot(data = mtcars, mapping = aes(wt, mpg)) +
    geom_point(size = 3, color = "black") +
    geom_point(size = 2, aes(color = factor(am))) +
    geom_smooth(method = "lm", formula = y ~ log(x)) +
    labs(x = "Weight (in 1000s of lbs)", y = "MPG", color = "Transmission")
plot2 <- ggplot(data = mtcars, mapping = aes(wt, mpg, color = factor(am))) +
    geom_point(size = 3, color = "black") +
    geom_point(size = 2) +
    geom_smooth(method = "lm", formula = y ~ log(x)) +
    labs(x = "Weight (in 1000s of lbs)", y = "MPG", color = "Transmission")
ggarrange(plot1, plot2, common.legend = TRUE)
```

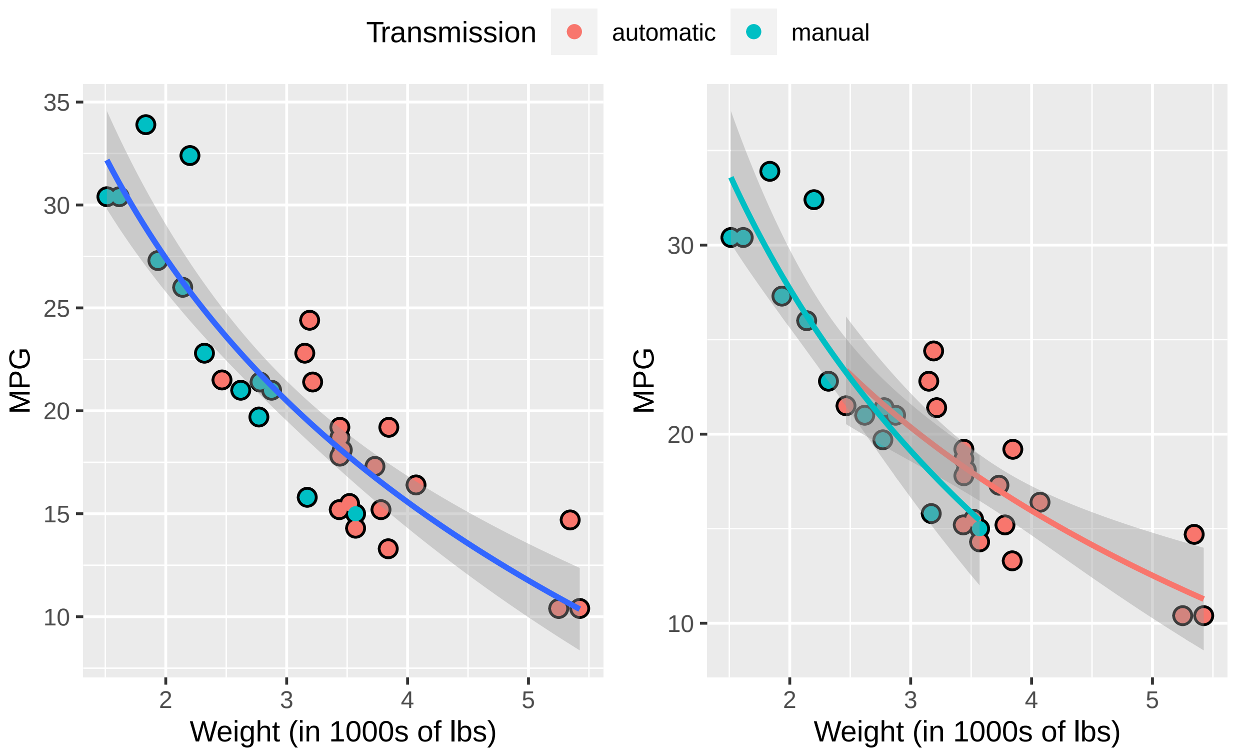


Figure 3:

```r
plot(fit2, which = 1)
```

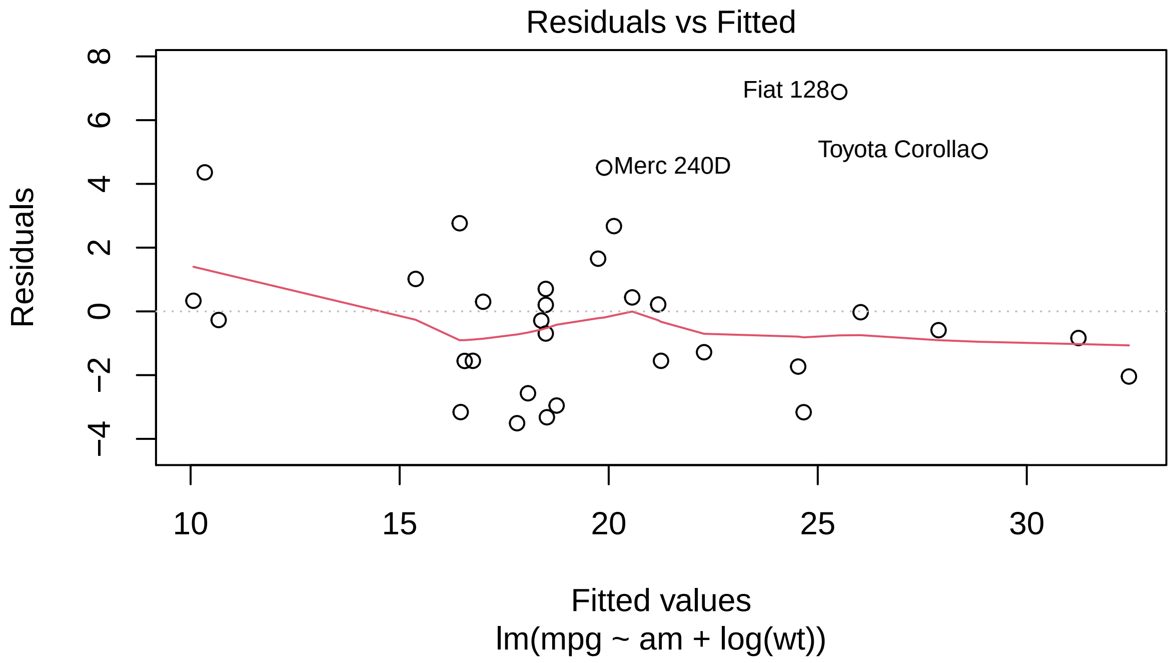
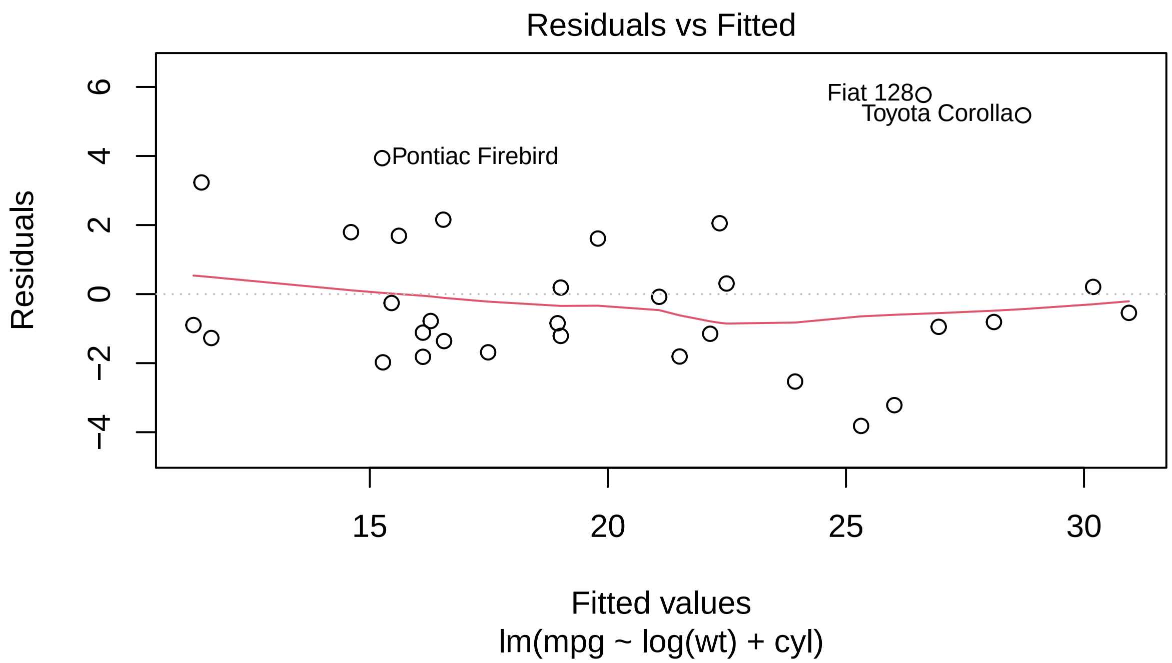


Figure 4:

```
plot(fit3, which = 1)
```


Residuals vs Fitted
Residuals
Fiat 128
Toyota Corolla
Pontiac Firebird
Fitted values
lm(mpg ~ log(wt) + cyl)