Effect of Transmission Type on MPG

Connor Gooding

9/7/2017

Executive Summary

When isolated from car weight (in tons) and number of cylinder pairs in the engine, **transmission type has no significant impact on fuel efficiency.** While the average manual-transmission car gets 7 miles per gallon more than the average automatic-transmission car, this variance is largely explained by the fact that automatic cars are heavier and have more cylinders in their engines than manual cars. Because the transmission type on fuel efficiency is not significant, it is not possible to quantify the MPG difference.

Analysis

Dataset

The `mtcars` dataset provides the data used for the study. It is found by default in R's datasets library and contains 11 different variables on 32 different car models observed in 1974 by *Motor Trend* magazine.

```
library(datasets)
data(mtcars)
mtcars$vs <- as.logical(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
```

Simplest Regression

At first glance, it would seem that manual transmission would provide a significant advantage, as the mean mpg for manual cars in the datasets is a full 7 miles per gallon higher than the mean mpg for automatic cars.

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

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

However, this model is overly simplistic and fails to account for any confounding variables.

Selecting the Right Model

After some exploratory analysis, it becomes evident that weight of the vehicle and pairs of cylinders in the engine confound transmission type as a predictor for MPG.

```
with_wt <- lm(mpg ~ am + I(wt / 2), mtcars)
best_fit <- lm(mpg ~ am + I(wt / 2) * I(cyl / 2), mtcars)
anova(over_simple, with_wt, best_fit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + I(wt/2)
```

```
## Model 3: mpg ~ am + I(wt/2) * I(cyl/2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 77.438 2.046e-09 ***
## 3     27 154.31  2    124.01 10.849 0.0003484 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of variance between the simple model and one including the weight of the car shows determines that weight is extremely necessary to include as a predictor. Pairs of cylinders is also an important regressor, to a lesser extent than weight but still well within the $\alpha = 0.05$ threshold of significance. The interaction between weight and number of cylinders is also significant enough to include as its own term.

In the Appendix, additional exploratory data analysis is shown to prove that horsepower, quarter mile time, and number of carburetors come close but do not qualify as significant enough for inclusion in the final model ("Honorable Mention Predictors").

Residual Analysis

The residual vs. fitted plot for this three-predictor model is provided in the Appendix ("Residual Plot"). Overall, the residuals do not appear to have a pattern, which suggests that this model is probably a good fit of fuel efficiency. In quantifiable terms, the three-predictor model describes 86% of the variance found in the data.

From this plot, it is evident that there are three cars in the dataset that could be considered outliers to the model. Of these three cars, the Toyota Corolla and Fiat 128 outperform the predictions by more than 2 standard deviations, while the Toyota Corona underperforms by more than 2 standard deviations.

```r
rstudent(best_fit)[c("Fiat 128", "Toyota Corolla", "Toyota Corona")]
```

```
##       Fiat 128 Toyota Corolla  Toyota Corona
##       2.621692       2.306827      -2.436256
```

Conclusions

After analyzing all of the available predictors for fuel efficiency, it is evident that transmission type has no significant relationship with fuel efficiency when isolated from car weight and number of cylinders. The aforementioned additional variables are not only key confounding variables, but they also serve as the two best predictors of fuel efficiency in the dataset.

As far as transmission type alone, it is not possible to reject the null hypothesis that transmission choice has zero impact on fuel efficiency. In fact, 0 is well within the confidence interval for its slope.

```r
confint(best_fit)[2,]
```

```
##     2.5 %    97.5 %
## -3.451589  1.727878
```

The full summary for the final model is provided in the Appendix ("Full Summary of Model").

Appendix

Honorable Mention Predictors

Of all the variables not included in the final model, horsepower (hp), quarter mile time (in seconds), and number of carburetors come the closest to being significant predictors to include with weight and pairs of cylinders, falling between $\alpha = .10$ and $\alpha = .05$ when included as additional regressors.

```
with_qsec <- update(best_fit, mpg ~ am + I(wt * .5) + I(cyl/2) + qsec, mtcars)
with_hp <- update(best_fit, mpg ~ am + I(wt * .5) + I(cyl/2) + hp, mtcars)
with_carb <- update(best_fit, mpg ~ am + I(wt * .5) + I(cyl/2) + carb, mtcars)
anova(best_fit, with_qsec)[2,]
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + I(wt/2) * I(cyl/2)
## Model 2: mpg ~ am + I(wt * 0.5) + I(cyl/2) + qsec
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 2     27 167.78  0   -13.473
```

```
anova(best_fit, with_hp)[2,]
```

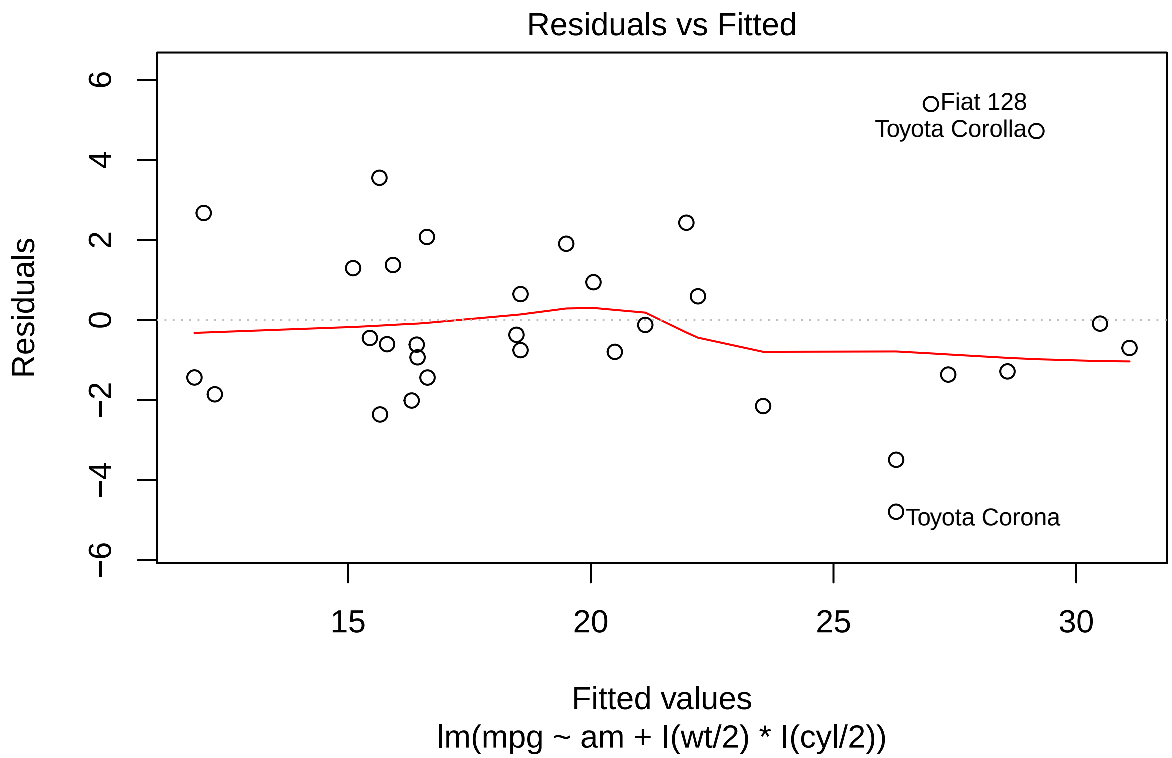
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + I(wt/2) * I(cyl/2)
## Model 2: mpg ~ am + I(wt * 0.5) + I(cyl/2) + hp
##   Res.Df RSS Df Sum of Sq F Pr(>F)
## 2     27 170  0   -15.686
```

```
anova(best_fit, with_carb)[2,]
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + I(wt/2) * I(cyl/2)
## Model 2: mpg ~ am + I(wt * 0.5) + I(cyl/2) + carb
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 2     27 168.71  0   -14.394
```

Residual Plot

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



Full Summary of Model

```
summary(best_fit)
```

```
##
## Call:
## lm(formula = mpg ~ am + I(wt/2) * I(cyl/2), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7887 -1.3812 -0.5268  1.3155  5.3945
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        57.0173     7.3507   7.757 2.43e-08 ***
## amManual           -0.8619     1.2622  -0.683 0.500524
## I(wt/2)           -19.0007     5.2984  -3.586 0.001308 **
## I(cyl/2)           -8.0222     2.1188  -3.786 0.000777 ***
## I(wt/2):I(cyl/2)    3.5432     1.3976   2.535 0.017337 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.391 on 27 degrees of freedom
## Multiple R-squared:  0.863,  Adjusted R-squared:  0.8427
## F-statistic: 42.51 on 4 and 27 DF,  p-value: 2.815e-11
```