Regression Course Project

Adam Ficke

5/11/2020

Abstract

This paper is interested in addressing the relationship between car transmission type and fuel economy. Specifically, it answers:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

The conclusions of this paper at that these data do not show a significant effect of transmission type on the fuel economy of a vehicle, despite the superficial relationship initially observed.

Data Exploration

First examine the shape of the response data (fig. 1) We can see that the data are all positive, but aren't counts (they can be fractions). For that reason we'll use a Gamma distribution in our link function.

Next we'll look at the superficial relationship between transmission type and fuel economy. (fig. 2) We can see on average manual transmission cars have higher fuel economy.

Modeling

Here we'll go through a series of models to see if the superficial relationship we found in the first step holds when we control for the other variables. To start, we'll use all of the variables, and trim them down later. (Model 1)

Clearly this model is overfit - none of the variables show as significant, and we can see from the correlation exhibit (fig. 3) that there's some collinearity going on.

We'll next use resampling to reduce our predictors. The data are relatively thin and we aren't resource limited, so we'll use cross validation to prevent overfitting. (Table 1)

We see that disp, wt, hp, and cyl were chosen by the recursive feature elimination. This doesn't bode well for the variable of interest, transmission type, but we'll keep that in the model to see if it shows significance.

We'll produce nested models so we can test the incremental significance of removing several of the categorical variables. (Models 2,3,& 4)

From the final Anova test, we can see that the significant model predictors end up being weight, displacement, and horsepower, but not transmission type.

Though transmission type doesn't seem to drive fuel economy, we'll still go back to the model with transmission type included, to obtain a confidence interval for its potential impact. First we'll check the residuals first for that final model to check our assumptions (fig. 4)

The residuals are well dispersed - we'll keep the Gamma distribution.

Now we'll look at the confidence intervals for the final model which includes transmission type (am) (Model 4)

Results

You can see from the confidence interval that the multiplicative effect of a car having an automatic transmission relative to manual (am1) is 0.0133% - which is small. When you look at the confidence intervals that coefficient falls on either side of 1, which shows that within the CI we aren't sure which direction the effect is. For that reason, these data do not show a significant effect of transmission type on the fuel economy of a vehicle, despite the superficial relationship observed.

Appendix

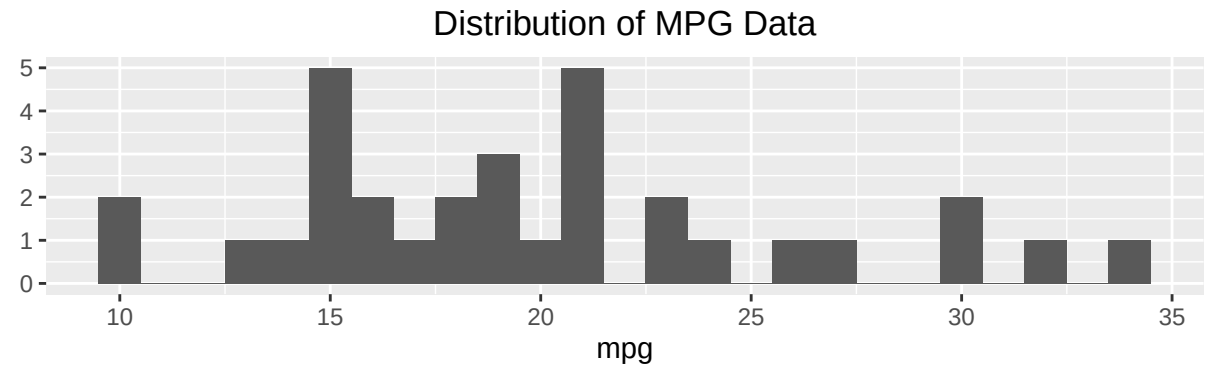

Figure 1: Response Distribution

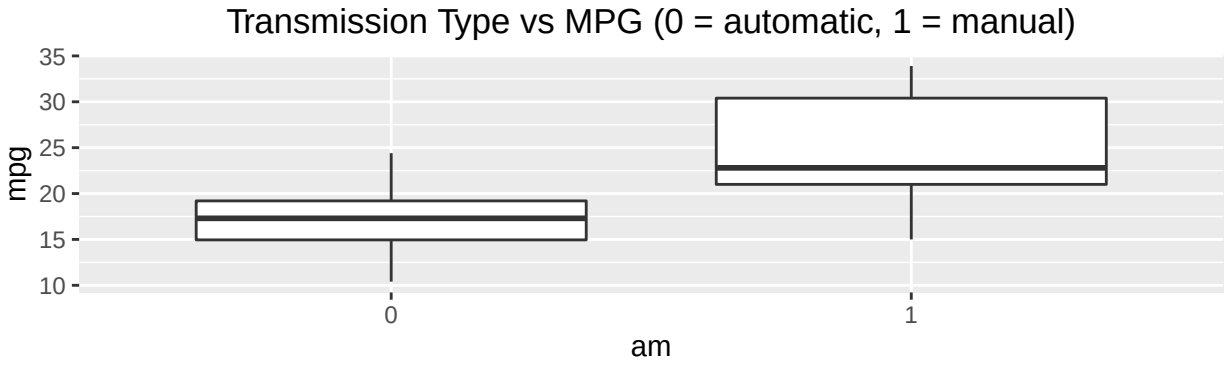

Figure 2: Transmission Boxplot

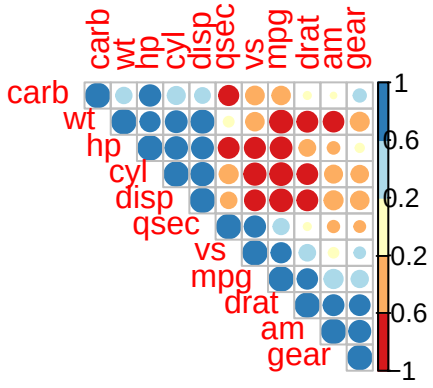

Figure 3: Correlation Matrix

```
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 1.0282205   1.041562 1.9806183 1.656602
## cyl6        0.9982546   1.007395 0.7888977 2.260922
## cyl8        1.0070804   1.016600 1.5350405 1.962730
## disp        0.9999429   1.000077 0.4779194 1.602746
## hp          1.0000402   1.000092 1.5489053 1.950232
## drat        0.9989719   1.004898 0.8101416 2.307293
## wt          1.0122395   1.006169 7.2276953 1.068877
## qsec        0.9996476   1.001874 0.8284233 2.347189
## vs1         1.0024835   1.005582 1.5614747 1.939109
## am1         1.0051573   1.007270 2.0343603 1.629848
## gear4       0.9870249   1.008093 0.1978543 1.134296
## gear5       0.9886755   1.008055 0.2417897 1.192614
## carb2       1.0031782   1.004223 2.1234192 1.588975
## carb3       0.9974591   1.010000 0.7743933 2.229270
## carb4       1.0118379   1.009853 3.3210954 1.282279
## carb6       1.0076850   1.014058 1.7304411 1.806701
## carb8       1.0071254   1.019809 1.4361461 2.059428
```

Model 1: All Variables

```
cl <- makePSOCKcluster(16)
registerDoParallel(cl)
set.seed(100)

subsets <- c(1:11)

ctrl <- rfeControl(
  functions = rfFuncs,
  method = "repeatedcv",
  repeats = 5,
  verbose = FALSE
)

lmProfile <- rfe(
  x = mtcars.fac[, 2:11],
  y = mtcars.fac$mpg,
  sizes = subsets,
  rfeControl = ctrl,
)
lmProfile
```

```
##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold, repeated 5 times)
##
## Resampling performance over subset size:
##
##  Variables  RMSE Rsquared   MAE RMSESD RsquaredSD  MAESD Selected
##          1 2.984   0.8507 2.563 1.3721     0.2053 1.1440
##          2 2.551   0.8872 2.245 1.0234     0.1707 1.0121
```

```
##          3 2.212   0.9189 1.941 0.9634     0.1385 0.9458
##          4 2.109   0.9314 1.838 0.9897     0.1086 0.9425        *
##          5 2.299   0.9168 2.043 0.9705     0.1294 0.8901
##          6 2.221   0.9208 1.979 0.9624     0.1232 0.9090
##          7 2.256   0.9160 2.007 0.9801     0.1288 0.9064
##          8 2.270   0.9157 2.049 0.9947     0.1323 0.9139
##          9 2.184   0.9263 1.953 0.9985     0.1216 0.9263
##         10 2.213   0.9258 1.968 0.9874     0.1184 0.9156
##
## The top 4 variables (out of 4):
##    disp, wt, hp, cyl
```

Table 1: Resampled Variable Reduction

```
##             Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 1.011900   1.007111  5.310236 1.113455
## am1         1.000099   1.003306  1.030342 2.654858
## disp        1.000011   1.000035  1.378903 2.118403
## wt          1.008488   1.002945 17.705053 1.008192
## hp          1.000070   1.000036  7.220173 1.060971
## cyl6        1.001938   1.003497  1.741079 1.793485
## cyl8        1.002719   1.007433  1.442889 2.048212
```

Model 2: Reduced Variables

```
## Analysis of Deviance Table
##
## Model 1: .outcome ~ disp + wt + hp
## Model 2: .outcome ~ am1 + disp + wt + hp
## Model 3: .outcome ~ am1 + disp + wt + hp + cyl6 + cyl8
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1        28    0.32860
## 2        27    0.32858  1 0.0000213   0.9678
## 3        25    0.32454  2 0.0040403   0.8573
```

Models 3 & 4: Testing for Significance of Categorical Variables

---

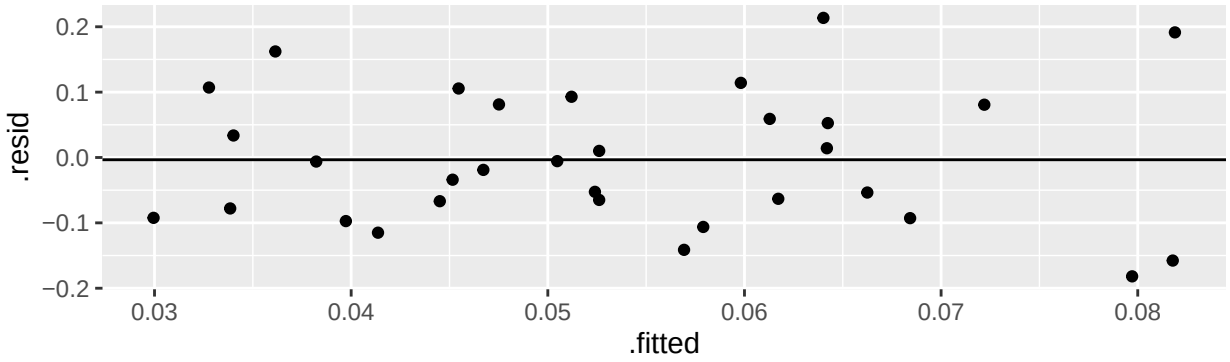

Figure 4: Residual Plot

```
## (Intercept)         am1        disp          wt          hp
```

```
##    1.010619    1.000133    1.000017    1.008726    1.000076
```

```
##                 2.5 %   97.5 %
## (Intercept) 0.9978349 1.023674
## am1         0.9938345 1.006415
## disp        0.9999682 1.000066
## wt          1.0034224 1.014026
## hp          1.0000140 1.000139
```

Model 4: Coefficients and Confidence Intervals