Analyzing Transmission Trends

Konstantinos Patelis

2021-02-15

Executive Summary

In this project we explore the relationship between the type of transmission and fuel efficiency in miles per gallon (mpg). To achieve this, we are using the `mtcars` dataset, seeking to answer whether automatic or manual transmission is better for fuel consumption, and to quantify this difference in consumption. We fit three models, the first one including most terms highly correlated with variable `mpg`, the second one reducing the number of variables using step-wise selection and the third one including an interaction term between `wt` and `am`. According to the third model, which seems to have the better fit, for weightless cars, manual transmissions would translate to higher efficiency, compared to cars with automatic transmissions with the same number of cylinders. In reality, cars up to a specific weight are more efficient with a manual transmission compared to cars with the same weight and number of cylinders that have automatic transmissions. Beyond that weight, cars with an automatic transmission are more efficient, compared to manual cars with the same characteristics.

Exploratory Analysis

Initially, we load the dataset and visualize whether there is a difference in fuel efficiency on average depending on the type of transmission, automatic (am = 0) or manual (am = 1). Our data contains information on miles/(US) gallon consumption (mpg), number of cylinders (cyl), weight in thousands of lbs (wt), etc.

At first glance, looking at Plot A (appendix), it seems that, on average, cars with a manual transmission have higher fuel efficiency. However, it is possible that this is an indirect effect and that in reality MPG is affected by other variables that could be correlated to the transmission type. In Plot B we can observe the degree of correlation between the different variables in the dataset.

It seems that mpg is negatively correlated with the number of cylinders, the engine displacement, the weight and horsepower, and positively correlated with the rear axle ratio, the engine shape and the type of transmission. However, the transmission type also seems to be negatively correlated with the number of cylinders, the displacement and the weight, and positively with the rear axle ratio. That could potentially mean that some of these variables affect both the type of transmission and the efficiency. Let's try fitting a linear model containing the variables that seem to be highly correlated with mpg. For modelling purposes, I will encode cyl, am and vs as factor variables. For the summary of the model, look at Model 1 in the appendix.

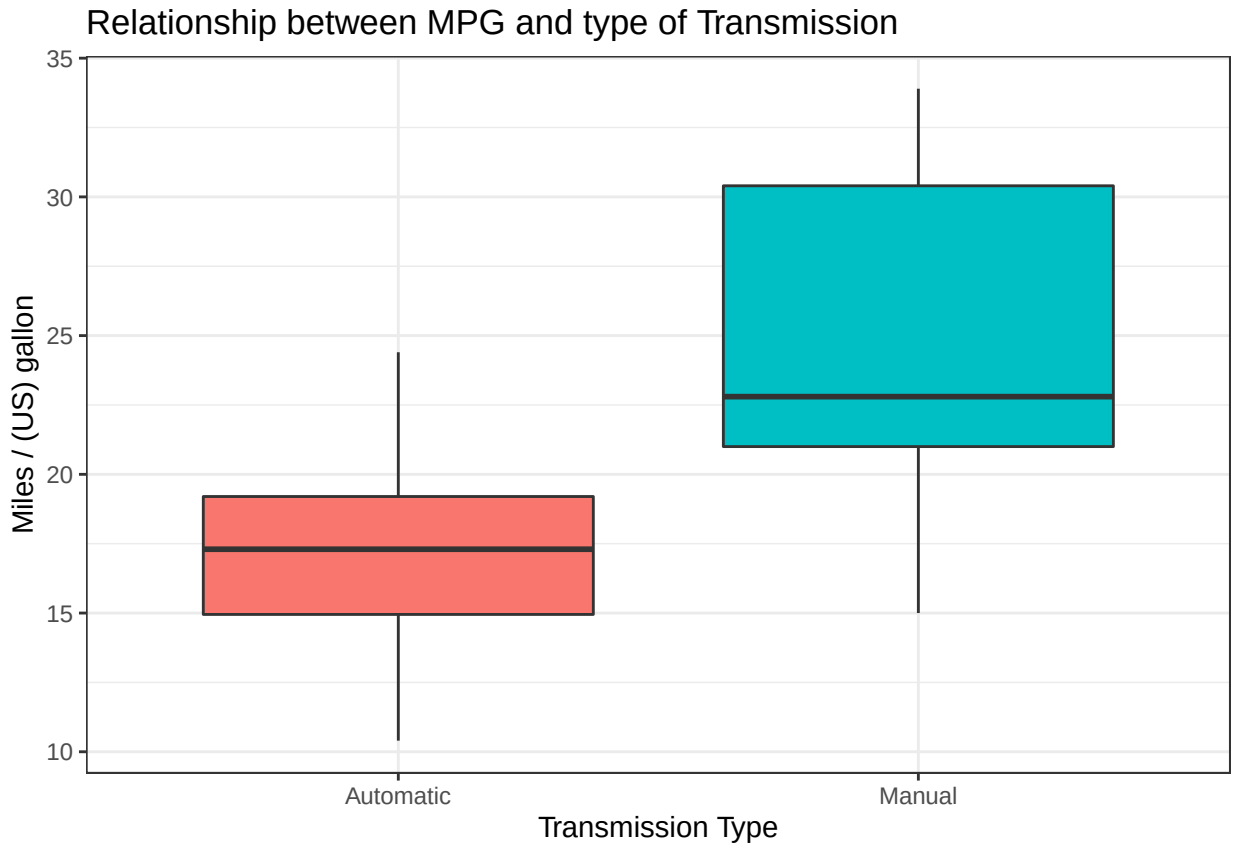
Looking at the model's p-values, it seems that displacement, the engine shape, the rear axle ratio and actually our variable of interest, the transmission type seem to not be important in explaining fuel efficiency. Potentially, displacement is explained by the number of cylinders and the overall weight, while the rear axle ratio could also be dependent on the car's weight, design-wise. Furthermore, the engine shape could also be related to the number of cylinders, which would explain its low explanatory power in the model. We can try a step-wise model search to figure out what is the best simple model that captures the variability in the data. Looking at the simplified model (Model 2, appendix), it seems that the most of the variability in mpg can be explained by the weight and the number of cylinders, while the transmission type does not seem to be important as it is dropped by the model. Plotting the residuals against the predictors can show us if there was any pattern that was missed (Model 2 - Plot A, appendix). Furthermore, we can plot the predictors

against the transmission type, to see if there is any form of interaction between them (Model 2 - Plot B, appendix). Looking at the latter, it appears that there is some interaction effect between weight and the transmission type. I will attempt to improve the existing model (Model 2), by adding an interaction between weight and the transmission. This might also help answer which transmission type is better for fuel efficiency (Model 3, appendix). I will also perform an ANOVA test between the three models, to check whether each improves upon the previous one. Using the `glance` function from the `broom` package (Model 2 and 3), we can see that several metrics that are used to evaluate models actually improve from model 2 to model 3, adjusted `R^2` increases, Akaike's (AIC) and Bayes' (BIC) information criterion both decrease. According to the ANOVA test (see ANOVA Test, appendix)

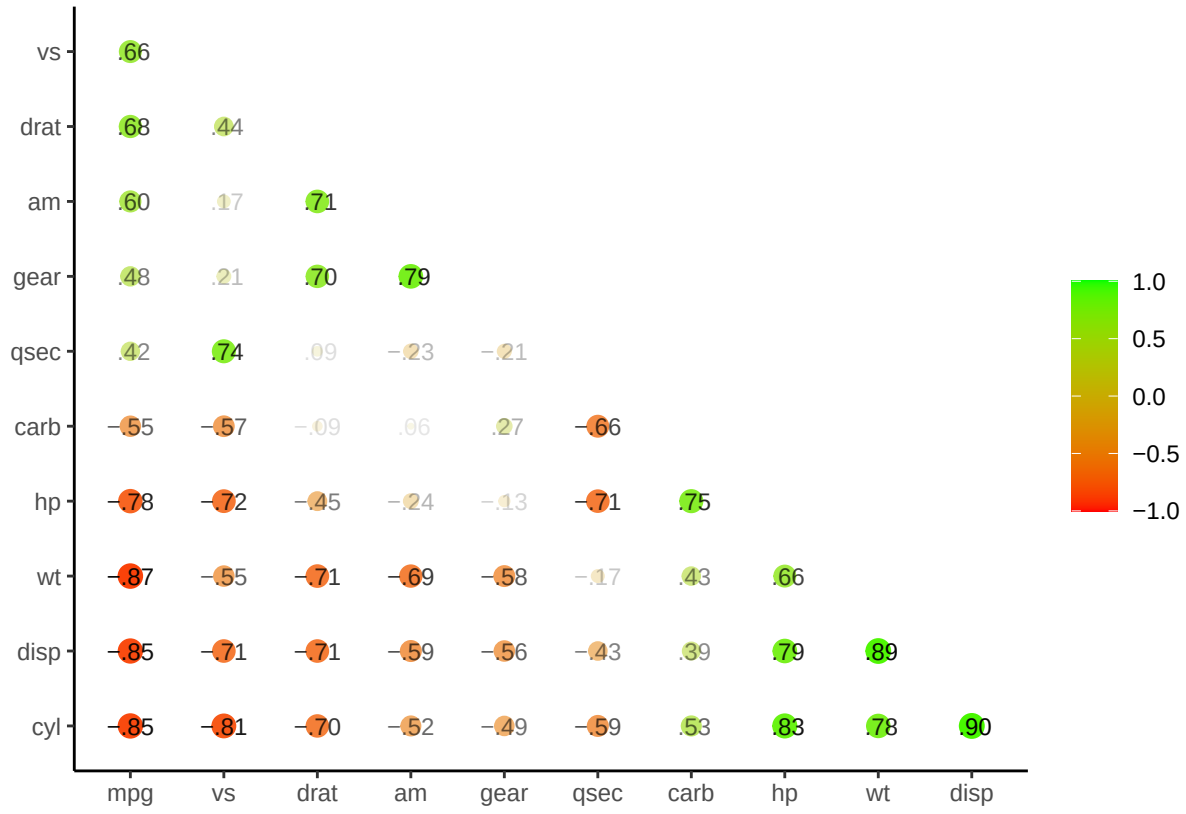
Based on Model 3, all coefficients are significant with more that 99% certainty, with the exception of the coefficient of the dummy variable indicating the number of cylinders is six. If the car was weightless, then a manual transmission would translate to higher fuel efficiency, around 11.6 miles for a gallon. Considering there is an interaction term, the effect of manual transmission on efficiency depends on the weight of the vehicle. Since the coefficient of the interaction term is negative, a car with manual transmission would have lower efficiency as weight increases, compared to cars with automatic transmissions. For cars weighing less than 11.569 / 4.068 * 1000 lbs, which is around 2840 lbs (see Model 3, appendix), the car would be more fuel efficient if it has a manual transmission, compared to cars with automatic transmissions with the same number of cylinders and same weight. For cars above the specified weight, the car would be more efficient if it had an automatic transmission, for same number of cylinders and same weight.

Appendix

Plot A



Plot B



Model 1

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + vs + drat + am, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2830 -1.2147 -0.3352  1.5081  5.5636
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.195092   7.255754   4.575 0.000122 ***
## cyl6        -3.845569   1.810914  -2.124 0.044212 *
## cyl8        -5.054698   3.532701  -1.431 0.165371
## disp         0.001251   0.014257   0.088 0.930807
## wt          -3.197287   1.292055  -2.475 0.020796 *
## vs1          1.235176   1.980999   0.624 0.538829
## drat        -0.247627   1.585552  -0.156 0.877199
## am1          0.729198   1.730045   0.421 0.677148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.737 on 24 degrees of freedom
## Multiple R-squared:  0.8404, Adjusted R-squared:  0.7938
## F-statistic: 18.05 on 7 and 24 DF,  p-value: 3.87e-08
```

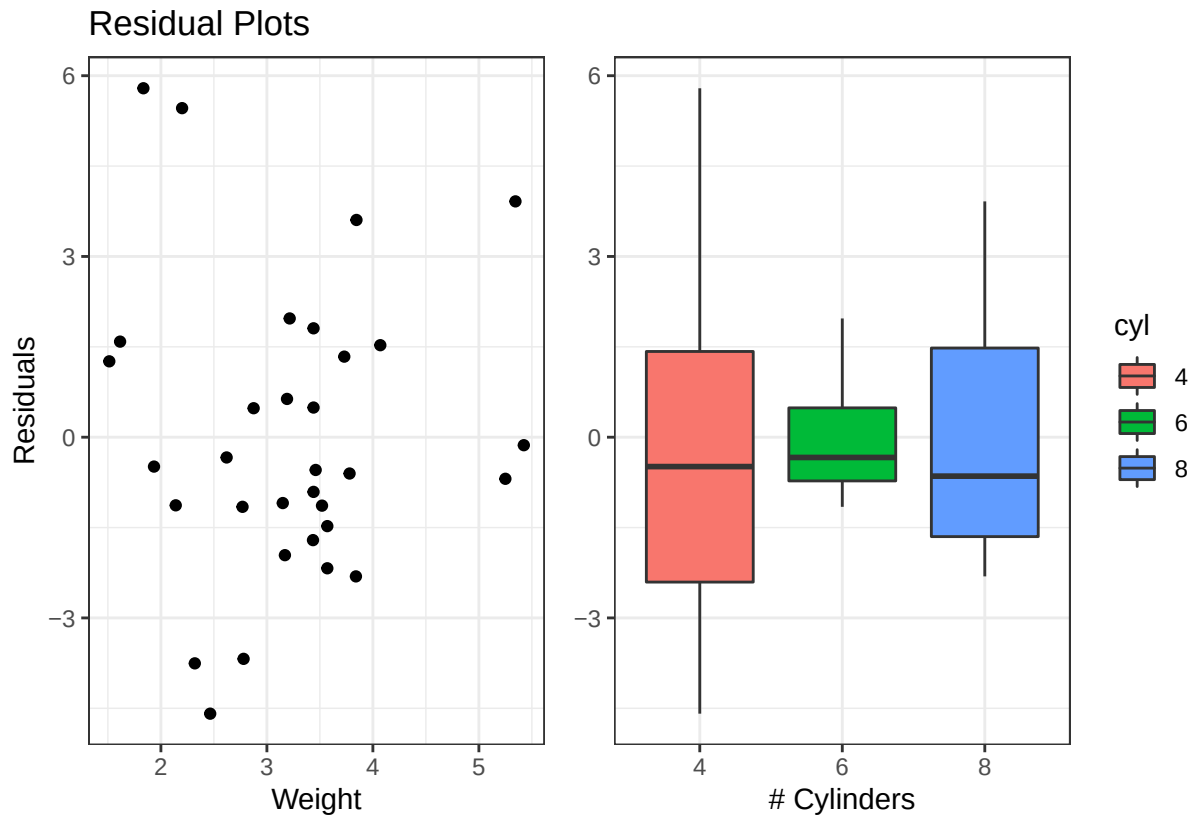
Model 2

```
## Start:  AIC=82.95
## mpg ~ cyl + disp + wt + vs + drat + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     0.058 179.82 79.499
## - drat  1     0.183 179.94 79.522
## - am    1     1.331 181.09 79.725
## - vs    1     2.912 182.68 80.003
## - cyl   2    33.872 213.63 81.547
## <none>              179.76 82.955
## - wt    1    45.866 225.63 86.761
##
## Step:  AIC=79.5
## mpg ~ cyl + wt + vs + drat + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     0.203 180.02 76.070
## - am    1     1.394 181.21 76.281
## - vs    1     2.929 182.75 76.551
## - cyl   2    34.972 214.79 78.255
## <none>              179.82 79.499
## - wt    1    78.897 258.72 87.674
##
## Step:  AIC=76.07
## mpg ~ cyl + wt + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - am    1     1.193 181.22 72.815
## - vs    1     2.945 182.97 73.123
## - cyl   2    36.292 216.32 75.015
## <none>              180.02 76.070
## - wt    1    78.871 258.89 84.231
##
## Step:  AIC=72.82
## mpg ~ cyl + wt + vs
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     1.842 183.06 69.673
## - cyl   2    42.877 224.09 72.680
## <none>              181.22 72.815
## - wt    1   118.800 300.02 85.482
##
## Step:  AIC=69.67
## mpg ~ cyl + wt
##
##        Df Sum of Sq    RSS    AIC
## <none>              183.06 69.673
## - cyl   2    95.263 278.32 76.149
## - wt    1   118.204 301.26 82.149
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = data)
```

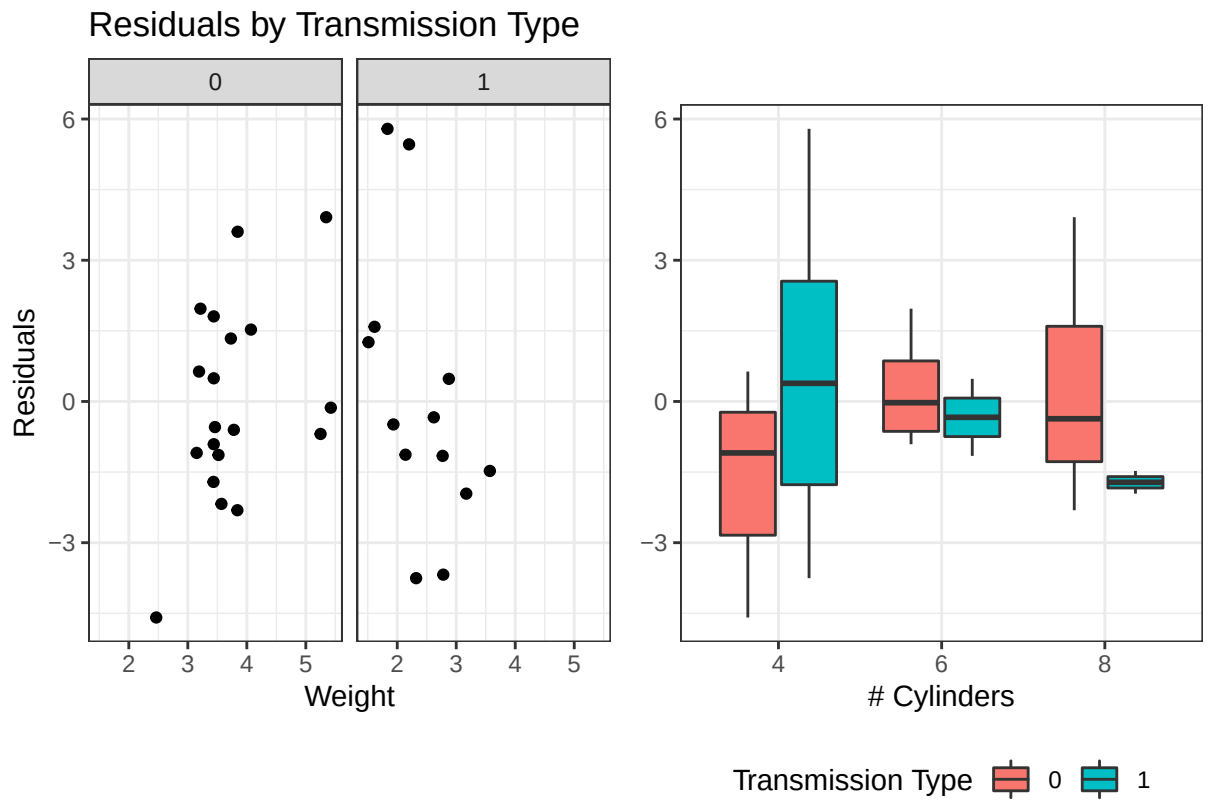
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.9908     1.8878  18.006  < 2e-16 ***
## cyl6         -4.2556     1.3861  -3.070 0.004718 **
## cyl8         -6.0709     1.6523  -3.674 0.000999 ***
## wt           -3.2056     0.7539  -4.252 0.000213 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.837         0.820  2.56      48.1 3.59e-11     3  -73.3  157.  164.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Model 2 - Plot A



Model 2 - Plot B



Model 3

```
##
## Call:
## lm(formula = mpg ~ cyl + wt * am, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5409 -1.5377 -0.6783  1.3160  5.2831
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.775      2.840  10.483 7.87e-11 ***
## cyl6          -2.710      1.357  -1.996  0.05647 .
## cyl8          -4.776      1.556  -3.070  0.00496 **
## wt            -2.399      0.844  -2.842  0.00860 **
## am1           11.569      4.088   2.830  0.00885 **
## wt:am1        -4.068      1.397  -2.911  0.00730 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.304 on 26 degrees of freedom
## Multiple R-squared:  0.8775, Adjusted R-squared:  0.8539
## F-statistic: 37.23 on 5 and 26 DF,  p-value: 4.743e-11
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.877         0.854  2.30      37.2 4.74e-11     5  -68.8  152.  162.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
anova(mod1, mod2, mod3)
```

ANOVA Test

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + wt + vs + drat + am
## Model 2: mpg ~ cyl + wt
## Model 3: mpg ~ cyl + wt * am
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     24 179.76
## 2     28 183.06 -4    -3.296 0.1100 0.97784
## 3     26 137.99  2    45.067 3.0084 0.06826 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```