Regression Models Project

Eric Erkela

11/18/2020

Executive Summary

In consumer car design, efficiency is the name of the game. As such, understanding the effect on miles per gallon (mpg) produced by a wide variety of other measures is of the utmost importance. In this analysis, we will investigate and quantify the relationship between mpg and transmission type (manual/automatic) across Motor Trend's 1974 mtcars data set, provided in the datasets package of the base R installation.

Exploratory Analysis

Before we jump to model fitting, we should take some time to explore our mtcars data set first. What can we tell about the relationship between mpg and transmission type at this stage? According to Figure 1 (see appendix), there seems to be a strong (roughly 10 mpg) difference in the median mpg of those cars that have an automatic vs a manual transmission. How much of this difference can be explained by other factors, such as weight or horsepower? To answer this, we'll need to develop a more sophisticated model to fit to our data.

Fitting a Model

In order to develop a model for our mtcars data, we use a series of nested models and ANOVA to determine which variables are worth including. Figure 2 in the appendix shows the output of doing so. Judging from it, only the inclusion of 3 variables (# of cylinders, horsepower, and weight, in addition to transmission type), have a significant positive impact on model fit, as evidenced by the Pr(>F) column. As such, our final model will include a total of 4 predictors for our mpg outcome: transmission type (our variable of interest) and the 3 confounders we identified above.

Figure 3 in the appendix shows the residuals from performing such a fit. They appear to be approximately normally distributed, which is a good sign, but how confident can we be that that is in fact the case? Quickly performing a Shapiro-Wilk normality test (see Figure 4) establishes a p-value $p = 0.07695$ in favor of residual normality. This is quite small, but still technically above the traditional 5% threshold. Still, this is the minimum we can produce with simple linear models, and any deviation from a standard normal will be small at worst, not significantly impacting the results of our analysis.

Results

Once we've fit our chosen model to the mtcars data set, all we need to do to find which transmission type produces better mpg is inspect the coefficients of our fit, which are given in Figure 5 in the appendix. The ammanual coefficient shows the estimated mpg increase by switching from an automatic to a manual

transmission. Turning these values into a 95% confidence interval for the effect of switching from an automatic to a manual transmission yields the following results:

$$-1.48 < \mu_{trans}[mpg|cyl, hp, wt] < 4.44$$

Since this interval includes 0, **we do not have enough evidence to suggest that transmission type has a significant impact on mpg** at the 5% confidence threshold. This can also be observed via the p-value associated with the ammanual coefficient in Figure 5: $p = 0.32 > 0.05$.

Appendix

Figure 1

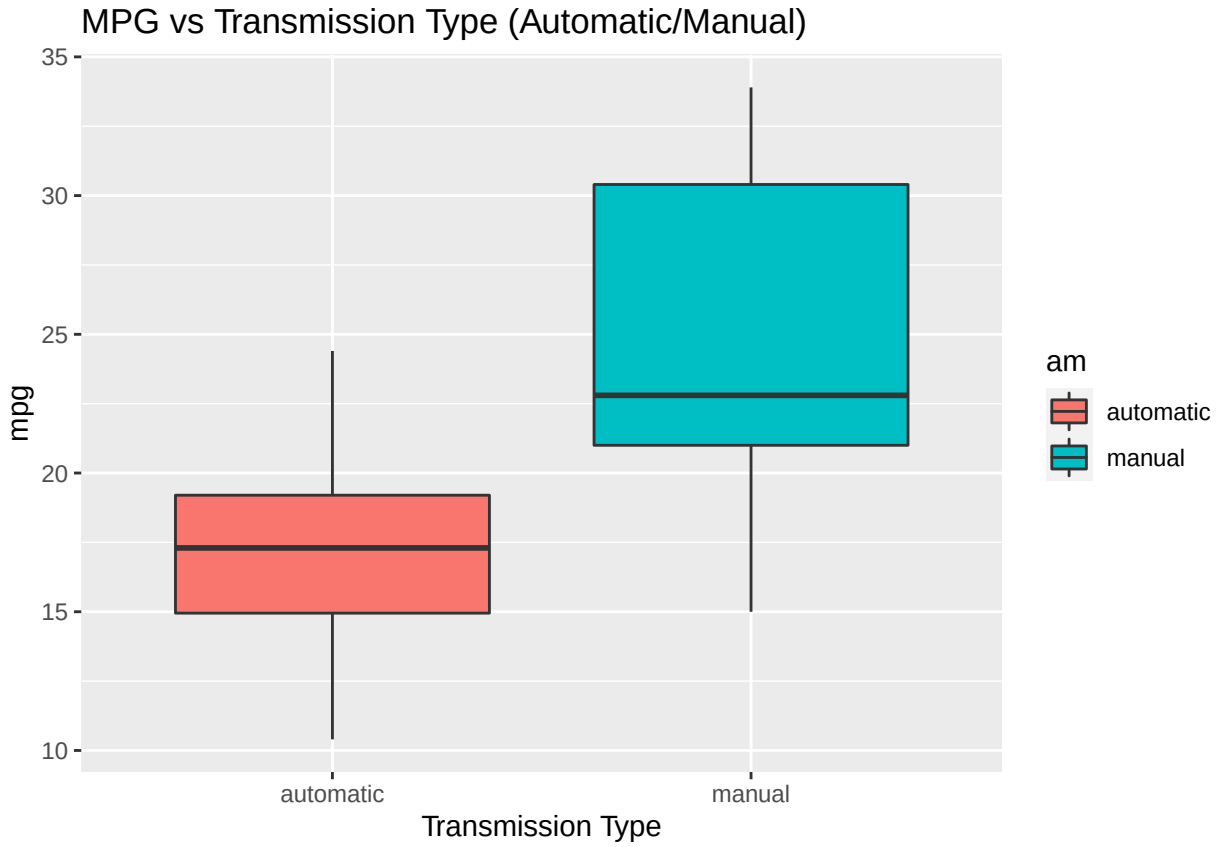

Figure 2

```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + cyl
## Model  3: mpg ~ am + cyl + disp
## Model  4: mpg ~ am + cyl + disp + hp
## Model  5: mpg ~ am + cyl + disp + hp + drat
```

```
## Model  6: mpg ~ am + cyl + disp + hp + drat + wt
## Model  7: mpg ~ am + cyl + disp + hp + drat + wt + qsec
## Model  8: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs
## Model  9: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear
## Model 10: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      30 720.90
## 2      29 271.36  1    449.53 64.0039 8.231e-08 ***
## 3      28 252.08  1     19.28  2.7452   0.11241
## 4      27 216.37  1     35.71  5.0849   0.03493 *
## 5      26 214.50  1      1.87  0.2663   0.61121
## 6      25 162.43  1     52.06  7.4127   0.01275 *
## 7      24 149.09  1     13.34  1.8999   0.18260
## 8      23 148.87  1      0.22  0.0309   0.86214
## 9      22 147.90  1      0.97  0.1384   0.71365
## 10     21 147.49  1      0.41  0.0579   0.81218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3

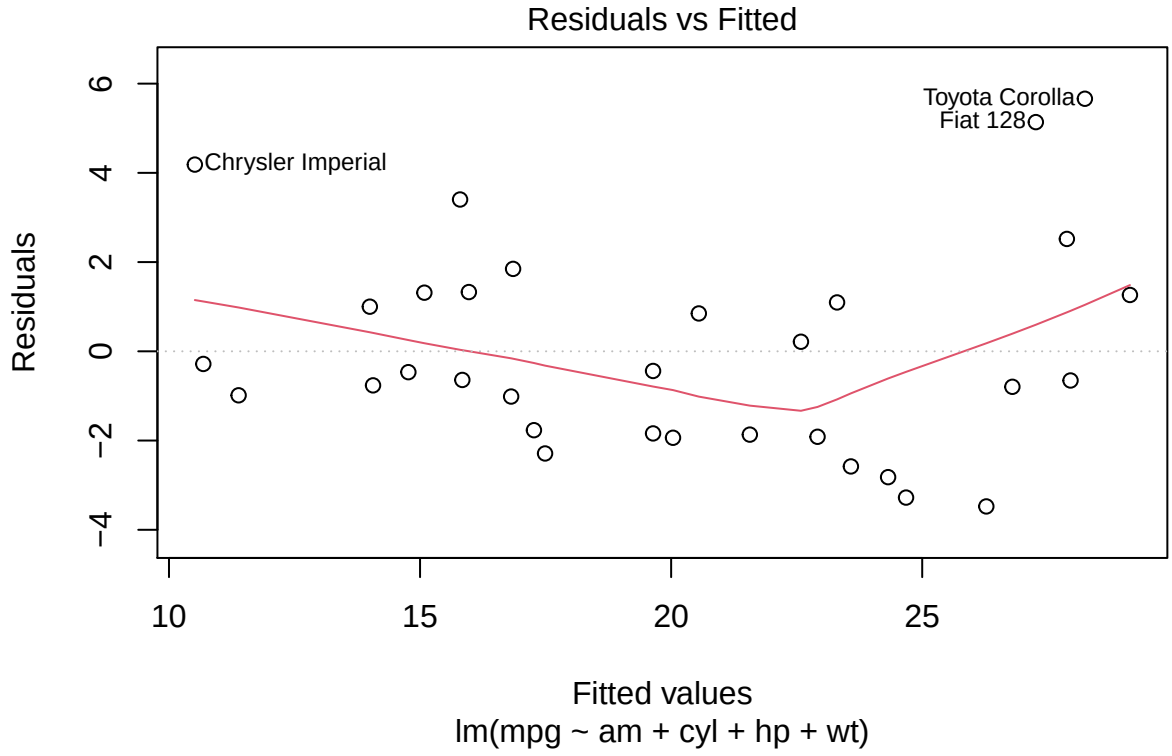

Figure 4

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  bestfit$residuals
## W = 0.94042, p-value = 0.07695
```

Figure 5

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 36.14653575 3.10478079 11.642218 4.944804e-12
## ammanual     1.47804771 1.44114927  1.025603 3.141799e-01
## cyl         -0.74515702 0.58278741 -1.278609 2.119166e-01
## hp          -0.02495106 0.01364614 -1.828433 7.855337e-02
## wt          -2.60648071 0.91983749 -2.833632 8.603218e-03
```