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.

Dependencies

Below is the list of all dependencies required for this analysis:

```
require(datasets, quietly = TRUE)
require(dplyr, quietly = TRUE, warn.conflicts = FALSE)
require(ggplot2, quietly = TRUE)
```

Loading and Cleaning the Data

First, we need to load the mtcars data set.

```
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Consulting with the mtcars data set's help page (?mtcars) can help us decipher the somewhat cryptic variable names in the raw data set and figure out what's going on within it. As we can see, the data covers fuel

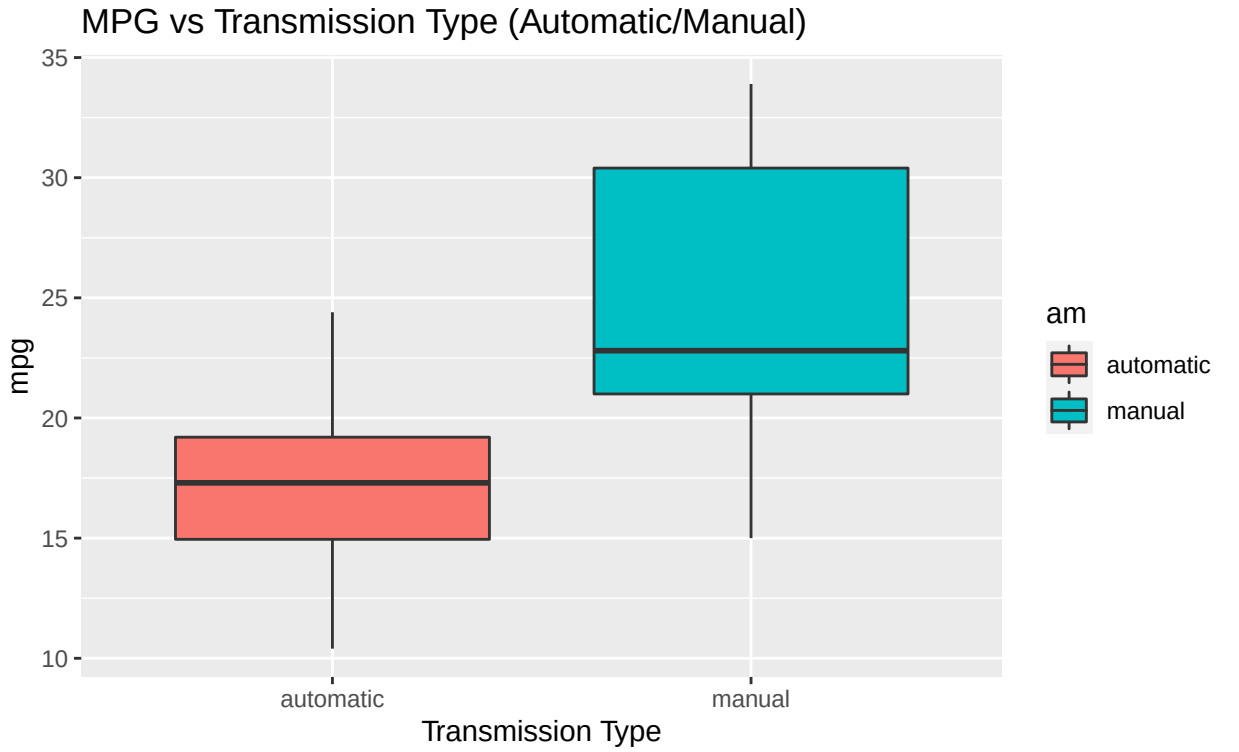
consumption and 10 other aspects of automobile design and performance for 32 models from the years 1973-74, as reported by Motor Trend. Many of the variables are self-explanatory, but the ones that aren't are as follows: drat (rear axle ratio), qsec (1/4 mile time), vs (engine shape - V-shaped/straight), and am (transmission - automatic/manual). One thing that is immediately apparent upon loading the data is that a few variables which are supposed to represent factors are encoded as numerics instead. Let's fix those now:

```
mtcars$vs <- as.factor(sapply(mtcars$vs, function(num) {
  if (num == 0) {
    "V-shaped"
  } else {
    "straight"
  }
}))
mtcars$am <- as.factor(sapply(mtcars$am, function(num) {
  if (num == 0) {
    "automatic"
  } else {
    "manual"
  }
}))
```

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?

```
g <- ggplot(data = mtcars, aes(x = am, y = mpg, fill = am))
g + geom_boxplot() +
  labs(title = "MPG vs Transmission Type (Automatic/Manual)",
       x = "Transmission Type")
```



According to the boxplot above, there seems to be a strong (roughly 10 mpg) difference in the median 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 will use a series of nested models and ANOVA to determine which variables are worth including.

```
# Nested model fits
fit1 <- lm(mpg ~ am, mtcars)
fit2 <- update(fit1, mpg ~ am + cyl)
fit3 <- update(fit1, mpg ~ am + cyl + disp)
fit4 <- update(fit1, mpg ~ am + cyl + disp + hp)
fit5 <- update(fit1, mpg ~ am + cyl + disp + hp + drat)
fit6 <- update(fit1, mpg ~ am + cyl + disp + hp + drat + wt)
fit7 <- update(fit1, mpg ~ am + cyl + disp + hp + drat + wt + qsec)
fit8 <- update(fit1, mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs)
fit9 <- update(fit1, mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear)
fit10 <- lm(mpg ~ ., mtcars)
anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fit9, fit10)
```

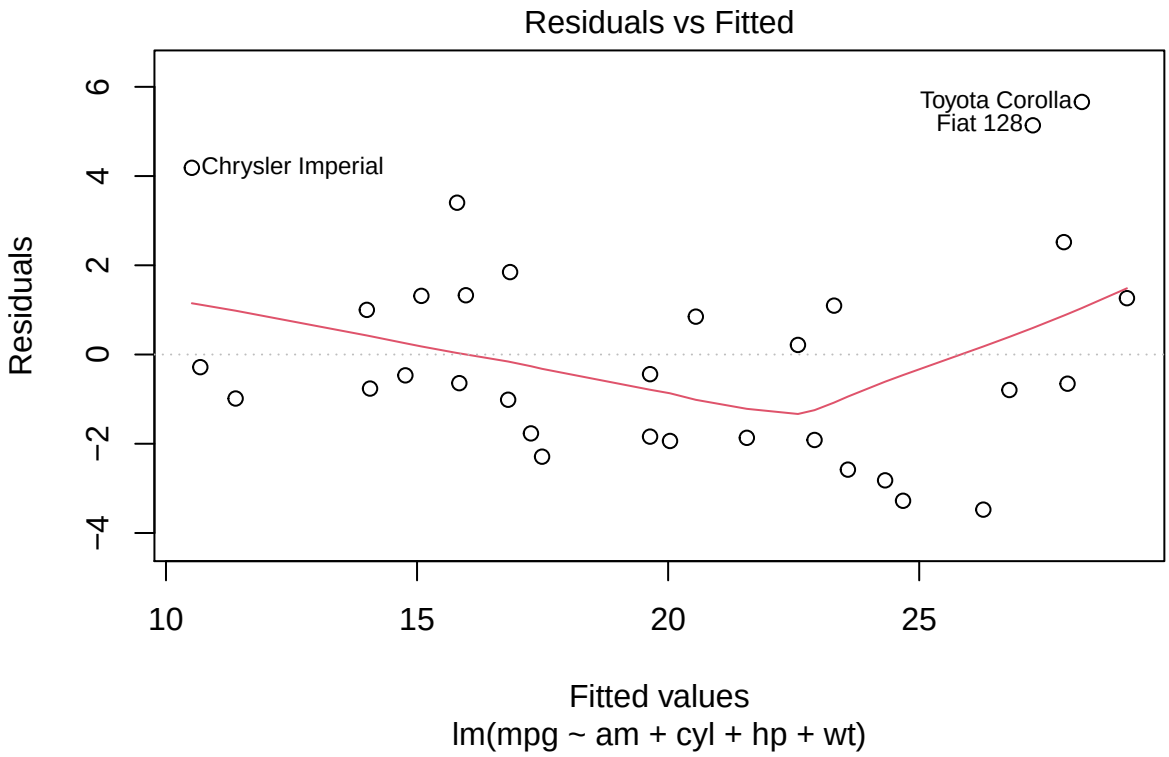
```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + cyl
```

```
## 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
```

As we can see from the output, only the inclusion of 3 variables (in addition to am - our transmission type), have a significant positive impact on model fit, as evidenced by the Pr(>F) column of the anova results. These variables are cyl, hp, and wt, in that order. As such, our final model will include 4 total predictors for our mpg outcome: am (our variable of interest) and the 3 confounders we identified above.

Let's perform a few diagnostics on our chosen model before we move on:

```
bestfit <- lm(mpg ~ am + cyl + hp + wt, mtcars)
plot(bestfit, which = 1)   # residuals vs fitted values
```



Our residuals according to this model appear to be approximately normally distributed, which is a good sign. How confident can we be that that is in fact the case?

```
shapiro.test(bestfit$residuals)
```

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

So our p-value for residual normality is small, but not quite below the traditional 5% threshold. Still, for a simple linear model, this isn't bad at all.

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.

```
summary(bestfit)$coef
```

```
##                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
```

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:

```
confint(bestfit)[2, ]
```

```
##     2.5 %    97.5 %
## -1.478946  4.435042
```

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 the bestfit summary at the beginning of this section. This is reproduced below for convenience:

```
summary(bestfit)$coef[2, 4]
```

```
## [1] 0.3141799
```