Analysis of the effect of transmission on the fuel economy of motor vehicles

Ryan Barley

Tuesday, January 20, 2015

Executive Summary

This report provides an analysis of the role of transmission type on fuel economy in motor vehicles. The data used was taken from the 'mtcars' dataset in R, which was extracted from the 1974 *Motor Trends* US magazine. The data includes fuel consumption and 10 aspects of automobile design for 32 vehicle models. In this report we want to conclude whether or not an automatic or manual transmission yields a better fuel economy in a motor vehicle. If there is a difference, we want to quantify it. This report includes an exploratory analysis of the 'mtcars' dataset, and a linear additive regression analysis to build a minimally adequate regression model.

This report finds that, on average, vehicles with manual transmissions have a better fuel economy. However, there are other factors that mitigate the difference, including gross horsepower of the vehicle, the weight of the vehicle, and the number of cylinders in the vehicles engine. The analysis conducted has limitations, first among those being a very small dataset. Also, the data used is fairly outdated, and modern advances in vehicle manufacturing may change the differences found significantly.

Exploratory Analysis

We start by loading the `mtcars` dataset, and taking a cursory glance at the data. We will do this using the `summary()` function. We will also make a pair-wise scatter plot looking at each variable in relation to each other (Figure 1)

```r
data(mtcars); attach(mtcars); #summary(mtcars)
```

By looking at this breakdown, it is clear that some of these variables are dummy variables and some are just coded as numeric when they should be factors. This is easily fixed.

```r
mtcars$cyl <- factor(mtcars$cyl) # Number of cylinders
mtcars$vs <- factor(mtcars$vs) # V-line/Straight-line engine type
mtcars$gear <- factor(mtcars$gear) # Number of forward gears
mtcars$am <- factor(mtcars$am) # Transmission
mtcars$carb <- factor(mtcars$carb) # Number of carburetors
```

Our first question is whether or not there is a clear difference between the fuel economy of a vehicle with a manual transmission and of one with an automatic transmission. Looking at a box plot of the variables *mpg* and *am* (Figure 2), we can see there is a substantial difference. We can quantify that difference by looking at the mean miles fuel economy of each type of transmission.

```r
round(aggregate(mpg ~ am, FUN = mean), 2)
```

```
##   am   mpg
## 1  0 17.15
## 2  1 24.39
```

Where 0 and 1 represent automatic and manual transmission, respectively.

We can also use a two-sided t-test to look at the difference. We define the null hypothesis as there being no difference between automatic and manual transmission in terms of fuel economy (the difference in means is zero):

```
t <- t.test(mpg ~ am, data = mtcars)
```

Based on a p-value of 0.0014, we can reject the null hypothesis. On the surface, this seems like a cut-and-dry solution, but there might be other factors influencing the fuel economy of a vehicle besides the transmission type.

Regression Analysis

Before deciding what regression method would be the best, we will look at a linear additive model using *mpg* as the outcome and every other variable as a predictor. We will also look at the model that only considers *am* as a predictor.

```
options(show.signif.stars=F)
model0 <- lm(mpg ~ am)
model1 <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb, data = mtcars)
summary(model1)$coefficients
```

```
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 23.87913244 20.06582026  1.19004018 0.25252548
## cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
## disp         0.03554632  0.03189920  1.11433290 0.28267339
## hp          -0.07050683  0.03942556 -1.78835344 0.09393155
## drat         1.18283018  2.48348458  0.47627845 0.64073922
## wt          -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec         0.36784482  0.93539569  0.39325050 0.69966720
## vs1          1.93085054  2.87125777  0.67247551 0.51150791
## am1          1.21211570  3.21354514  0.37718957 0.71131573
## gear4        1.11435494  3.79951726  0.29328856 0.77332027
## gear5        2.52839599  3.73635801  0.67670068 0.50889747
## carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3        2.99963875  4.29354611  0.69863900 0.49546781
## carb4        1.09142288  4.44961992  0.24528452 0.80956031
## carb6        4.47756921  6.38406242  0.70136677 0.49381268
## carb8        7.25041126  8.36056638  0.86721532 0.39948495
```

From our linear additive model, we do not see anything truly significant, but we are fairly close with the *hp* and *wt* variables. These are the gross horsepower of the vehicle and the weight of the vehicle (lb/1000), respectively. Next, we will build a minimal adequate model by eliminating insignificant variables one by one. For the sake of space, I have hidden the output of this code.

```
model2 <- update(model1, .~. -carb)
model3 <- update(model2, .~. -drat)
model4 <- update(model3, .~. -gear)
model5 <- update(model4, .~. -disp)
model6 <- update(model5, .~. -qsec)
model7 <- update(model6, .~. -vs)
summary(model7)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## am1          1.80921138 1.39630450  1.295714 2.064597e-01
```

```
a <- anova(model0, model1, model7)
```

Removing those extra confounding variables contributes to a small difference in our R-squared value, which means we are accounting for less variance, but this happens when predictors are removed. Our adjusted R-squared value, on the other hand, has moved from 0.779 to 0.84. Nearly each variable is now significant. We also see from the `anova` test, that removing the extra confounding variable from our models had an extremely insignificant effect on our final model.

We can check the confidence interval of our model by

```
confint(model7, level = .95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 28.35390366 39.062744138
## cyl6        -5.92405718 -0.138631806
## cyl8        -6.85902199  2.531671342
## hp          -0.06025492 -0.003963941
## wt          -4.31718120 -0.676477640
## am1         -1.06093363  4.679356394
```

Thus, we are 95% certain that the difference between fuel economy for different transmissions is between -1.061 and 4.679 miles per gallon, accounting for 86.6% of variance. We also see that the number of cylinders an engine has, the gross horsepower of the vehicle, and the weight of the vehicle all contribute to fuel economy.

Diagnostics

We see in our Residual vs. Fitted plot, the points bounce around the 0 line, thus they are reasonably linear. The variance in the error terms are not equal, due to 3 outliers: The Toyota Corolla, Fiat 128, and Datsun 710, these points warrant further investigation. Based on the Normal Q-Q Plot, our resulting model is roughly normal. (See appendix for these plots)

Appendix

Figure 1

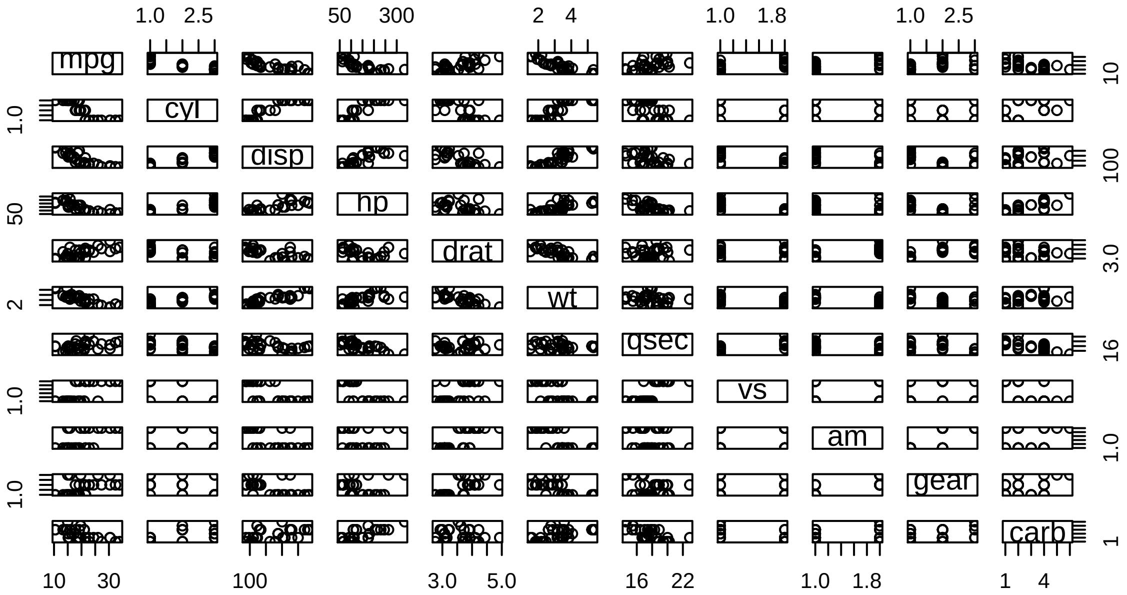
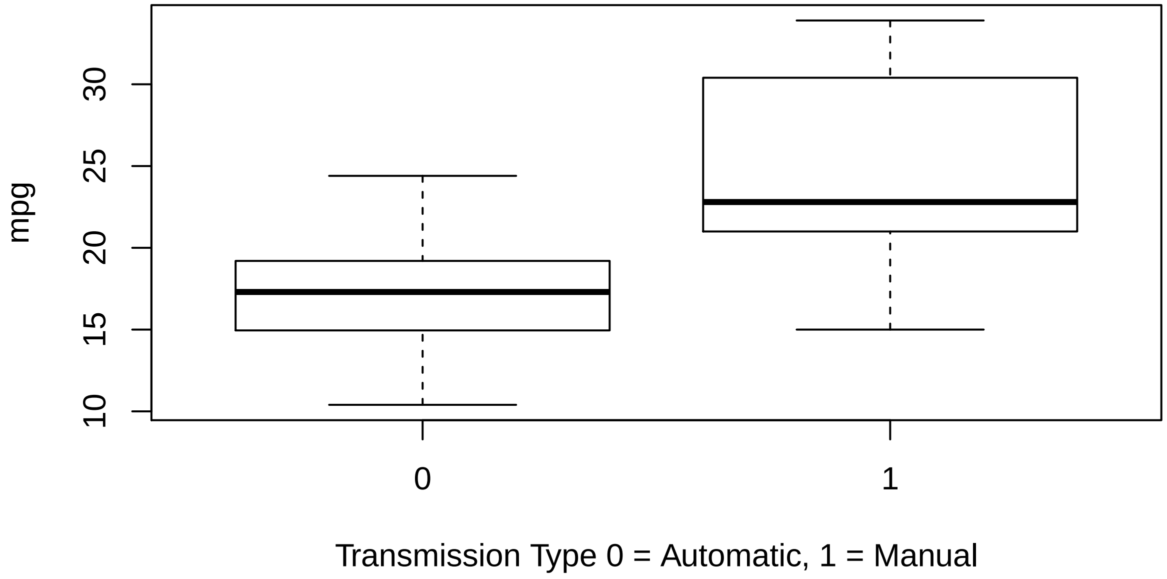
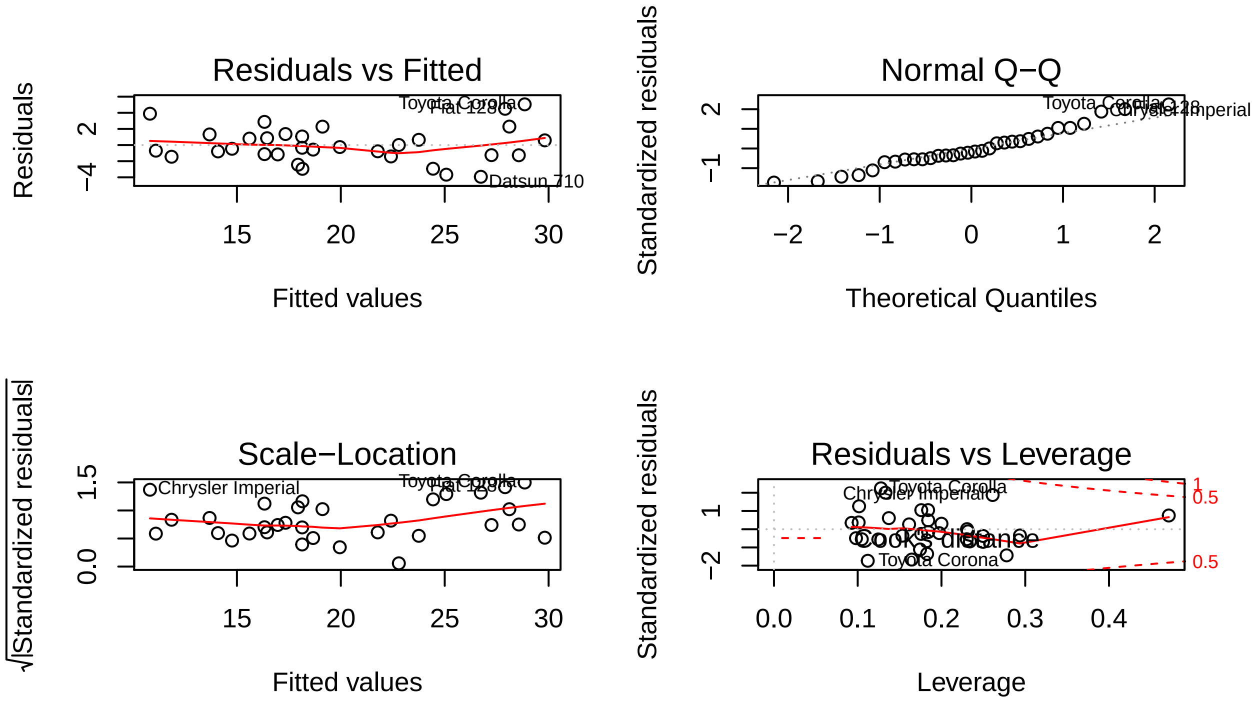


Figure 2




Residuals vs Fitted
Residuals
2
−4
15
20
25
30
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
2
−1
−2
−1
0
1
2
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
1.5
0.0
15
20
25
30
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
1
−2
0.0
0.1
0.2
0.3
0.4
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5