Motor Trend's Investigation on Factors Affecting MPG

JBrand

Executive Summary

From *Motor Trend*, 1974: Looking at a data set of a collection of cars, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG). We are particularly interested in the following two questions:

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

On an initial review of the data using a simple hypothesis test with a linear regression model, we do see a significant difference in MPG, on average, between automatic and manual transmission cars. The manual transmission seems to provide an increase of **7.245** MPG more than the automatics. But this is a simple model and does not take into account any of the other variables in the data set that could have confounding effects on the outcome, such as weight, cylinder count, gear count, etc. The best method to test for this is a multivariate regression with ANOVA for validation. That method revealed a slight edge still for the manual transmission at **2.084**, but one that might no longer be statistically significant.

Data Processing

The data is pulled from the `mtcars` dataset in R, located in the `datasets` library.

```
data(mtcars)
## We want to use the "am" column, which means we'll need to turn it into a factor for us to use easily.
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
```

Data Analysis

As a first step to this analysis, we will do a very simple first look to just take the means of all MPG's for each of the Automatic and Manual cars.

```
simple <- aggregate(mpg ~ am, data = mtcars, FUN = mean)
round(simple[2,2]-simple[1,2],3)
```

```
## [1] 7.245
```

So on the surface it looks like there is a flat difference of **7.245** MPG for the manual transmission. But we should do a quick hypothesis test with a = .05 to see if that value is statistically significant.

```
ttest <- t.test(mtcars[mtcars$am == "Automatic",]$mpg, mtcars[mtcars$am == "Manual",]$mpg)
c(abs(ttest$statistic), ttest$p.value)
```

```
##           t
## 3.767123145 0.001373638
```

So we end up with a t-statistic of **3.77** and a p-value of **.001**. So in this case, we have to reject the null hypothesis. Now let's see if we can quantify this difference a little better.

Linear Regression

We start with a simple linear regression using our two variables, `mpg` and `am`.

```r
fit <- lm(mpg ~ am, data = mtcars)
fit$coefficients
```

```
## (Intercept)    amManual
##   17.147368    7.244939
```

```r
round(summary(fit)$r.squared,4)
```

```
## [1] 0.3598
```

So the model shows our same value as before, **7.245** MPG increase for manual transmissions, but another interesting point is the R-squared value of **0.3598**, which says that only **35.98%** of our variability is explained by the model - looks like we need to add some more variables.

Multivariate Linear Regression

The first step when adding multiple variables for adjustment, is to decide which ones to use. The easiest way to do this is to run several models, adding in additional variables one at a time, and doing an ANOVA test. So that's what we will do.

```r
fit2 <- lm(mpg ~ am + wt, data = mtcars)
fit3 <- lm(mpg ~ am + wt + hp, data = mtcars)
fit4 <- lm(mpg ~ am + wt + hp + cyl, data = mtcars)
fit5 <- lm(mpg ~ am + wt + hp + cyl + gear, data = mtcars)
anova(fit, fit2, fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + hp
## Model 4: mpg ~ am + wt + hp + cyl
## Model 5: mpg ~ am + wt + hp + cyl + gear
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 67.7975 1.026e-08 ***
## 3     28 180.29  1     98.03 15.0168 0.0006469 ***
## 4     27 170.00  1     10.29  1.5768 0.2203885
## 5     26 169.73  1      0.27  0.0416 0.8399065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the p-value results of ANOVA, and assuming our standard confidence alpha = .05, Model 3 looks to be the cutoff. In addition to transmission type, we will also include the weight and the horsepower of the vehicles. This makes sense, since the main factor in the MPG variable is number of miles traveled and the amount of fuel required. Obviously, a heavier vehicle with more horsepower is going to consume more fuel for the same distance traveled. I had thought that the number of cylinders and the gearing would make a difference as well, but since the same cylinder and gear counts are available in both manual and automatic transmissions, there is no effect when the transmission type is the first regressor.

The last step to verify this new model is to look at the residuals and check for non-normality and heteroskedasticity. That's easy to do with a plot, which you can see in the Appendix, Figure 1. As you can see from the plots, the model residuals are distributed normally and they're homoskedastic. Finally, let's get a summary of our selected model.

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## amManual     2.08371013 1.376420152  1.513862 1.412682e-01
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
```

```
## [1] 0.8399
```

As can be seen, the addition of weight and horsepower did have an affect - they were confounding the relationship between transmission type and MPG. So when adjusted, now the manual transmission only has an edge of **2.084** MPG over the automatic transmission, and the new R-squared value of **0.8399** gives us a **83.99%** coverage of the variance.

Conclusion

After analysis, it was determined that to discover the true relationship between transmission type and MPG, it was necessary to adjust for weight and horsepower. After doing so, it was determined that there was still a higher MPG for manual transmission vehicles.

Appendix

Figure 1