Regression Models - Project WriteUp

Jonathan Brant

Executive Summary

Motor Trend magazine was interested in determining the relationship between fuel economy (measured in miles per gallon, or MPG) and several other variables given in the *mtcars* dataset. In particular, they were interested in how the two transmission types (automatic and manual) affect fuel economy.

This study addressed the aforementioned question at two levels. First, we focused solely on the affect of only the transmission type on fuel economy (disregarding all other variables) and established a measure of statistical significance for that effect. It turned out that cars with manual transmissions had better fuel economy than did automatic transmissions by a statistically significant margin ($p < 0.05$).

The second portion of the study attempted to quantify the difference between transmission types in the context of other variables in the dataset. In order to do this, two different linear regression models were constructed: one representing the transmission type as a predictor of fuel economy and the other considering all variables as predictors. Stepwise regression was run on the latter (using both forward and backward selection), resulting in weight, horsepower, and the number of engine cylinders being the most significant predictors of fuel economy with transmission being a comparitvely insignifcant contributor.

As such, while transmission type does affect fuel economy, it is not the most significant driver of such.

Data Pre-Processing and Exploratory Analysis

Unfortunately, many of the attributes within the *mtcars* dataset are somewhat cryptic, so the first undertaking will simply be an effort toward making the headers more intuitive, as well as converting the binary indicator for transmission to a string value and converting discrete fields (i.e. cylinders, transmission, etc.) to factors.

In order to get a good sense of the data, an exploratory analysis needs to be conducted, examining the structure of the dataset and the range and distribution of its values for each column. The code and corresponding output for this analysis can be found in exhibit 1.1 of the appendix.

Transmission effect on Fuel Economy

In order to satisfy the initial requirements of Motor Trend, the first question that needed to be addressed was, holding all other factors constant, which transmission results in the best fueld economy (highest MPG)? The hypotheses are setup as follows:

H_0 : There is no difference in fuel economy between manual and automatic transmissions.
H_a : There is a difference in fuel economy between manual and automatic transmissions.

The box plot shown in exhibit 2.1 of the appendix depicts the median and interquartile range of the MPG for both manual and automatic transmissions. At first glance, it seems that the fuel economy for manual transmissions is much better than automatic.

In order to determine whether the result is statistically significant and not simply due to sampling error, a t-test is conducted in order to evaluate the above hypotheses, the results of which are shown exhibit 2.2 of the appendix. It turns out that the results are indeed highly significant ($p < 0.01$), so the null hypothesis that the two transmissions result in comparable fuel economy is rejected.

Quantifying Differential Transmission effect on Fuel Economy

After establishing that manual transmissions generally result in better fuel economy than automatic transmissions holding all other variables constant, we want to take a look at the effect of other variables in the dataset in order to determine the relative contribution of other factors in relation to the transmission. The hypotheses on which this effort is based are below:

H_0 : The transmission type is solely responsible for variation in fuel efficiency.
H_a : Confounders exist that contribute as much or more to variation in fuel efficiency (or lack thereof) than transmission type.

In order to examine the relationship between fuel economy and other variables, a linear model is constructed with all variables as predictors of fuel efficiency. Stepwise regression is then run on the model (in both directions) in order to select the best explanatory variables. The results are given in exhibit 3.1 of the appendix. It turns out that the best fit model includes weight, horsepower, transmission,and cylinders; however, only weight and horsepower are significant ($p < 0.05$ and $p < 0.01$ respectively). That said, the resultant coefficients with the highly significant overall p-value and the fact that adjusted R^2 indicates that 84% of variability is explained by the included parameters indicates that this is a relatively good predictive model.

The next step is to evaluate how the aforementioned model stacks up against the model using only the transmission type to predict the fuel economy by running an analysis of variance (ANOVA) against the two models (shown in exhibit 3.2 of the appendix). The results yield a residual sum of squares that is much lower for the model that includes the engine displacement, horsepower, weight, and transmission (as opposed to the one with just the transmission), indicating a better fit. Moreover, there's a highly significant p-value ($p < 0.01$ at 99% CI), suggesting (assuming accuracy of the model) that the null hypothesis that fuel economy is wholly explained by transmission can be confidently rejected.

Prior to performing inference, however, the final step is to perform model diagnostics in order to establish validity for model assumptions. Exhibit 3.3 of the appendix depicts the result of running the following four diagnostics:

- Residuals vs. Fitted plot - indicates some level of homoscedasticity in the lack of correlation between residuals and fitted values.
- Q-Q plot - depicts a tight fit of the standardized residuals to the normal line, indicating a roughly normal distribution of residuals.
- Scale-Location plot - depicts a loess curve that slopes downward, and then relatively steeply upward. This would seem to indicate a less than ideal uniformity between variances.
- Residuals vs. Leverage plot - indicates that the distribution of leverage vs. residuals s fairly well-balanced; however, the bulk of the data resides in the downard sloping portion of the line, indicating overall high leverage and low residual.

Conclusion

Despite the statistical signifance of the ANOVA, we're hesistant to reject the null hypothesis because of some of the issues found while running diagnostics, potentially indicating some shortcomings of the model itself. There are, however, some conclusions that can still be drawn:

1. Motor Trend can be reasonably assured that overall, cars with manual transmissions will have better fuel economy than cars with automatic transmissions if all other factors are held constant.
2. Transmission is not the only variable that has an effect on fuel economy, nor is its effect the most pronounced. In addition to transmission, greater mass, more horsepower, and larger engine sizes are all inversely related to fuel economy.

Appendix

Exhibit 1.1 - Exploratory Analysis

```
## Print out the dataset variables
names(mtcars)
```

```
##  [1] "MPG"             "Cylinders"       "Displacement"
##  [4] "Horsepower"      "RearAxleRatio"   "Weight"
##  [7] "QuarterMileTime" "VS"              "Transmission"
## [10] "NumForwardGears" "NumCarburetors"
```

```
## Take a look at the data set structure
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 ...
##  $ Cylinders      : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ Displacement   : num  160 160 108 258 360 ...
##  $ Horsepower     : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ RearAxleRatio  : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ Weight         : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ QuarterMileTime: num  16.5 17 18.6 19.4 17 ...
##  $ VS             : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ Transmission   : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ NumForwardGears: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ NumCarburetors : Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

```
## Print overview of data
summary(mtcars)
```

```
##       MPG        Cylinders  Displacement     Horsepower    RearAxleRatio
##  Min.   :10.40   4:11       Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7       1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14       Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09              Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80              3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90              Max.   :472.0   Max.   :335.0   Max.   :4.930
##      Weight      QuarterMileTime VS        Transmission NumForwardGears
##  Min.   :1.513   Min.   :14.50   0:18   Automatic:19    3:15
##  1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13    4:12
##  Median :3.325   Median :17.71                         5: 5
##  Mean   :3.217   Mean   :17.85
##  3rd Qu.:3.610   3rd Qu.:18.90
##  Max.   :5.424   Max.   :22.90
##  NumCarburetors
##  1: 7
##  2:10
##  3: 3
##  4:10
##  6: 1
##  8: 1
```

Exhibit 2.1 - Transmissions Box Plot

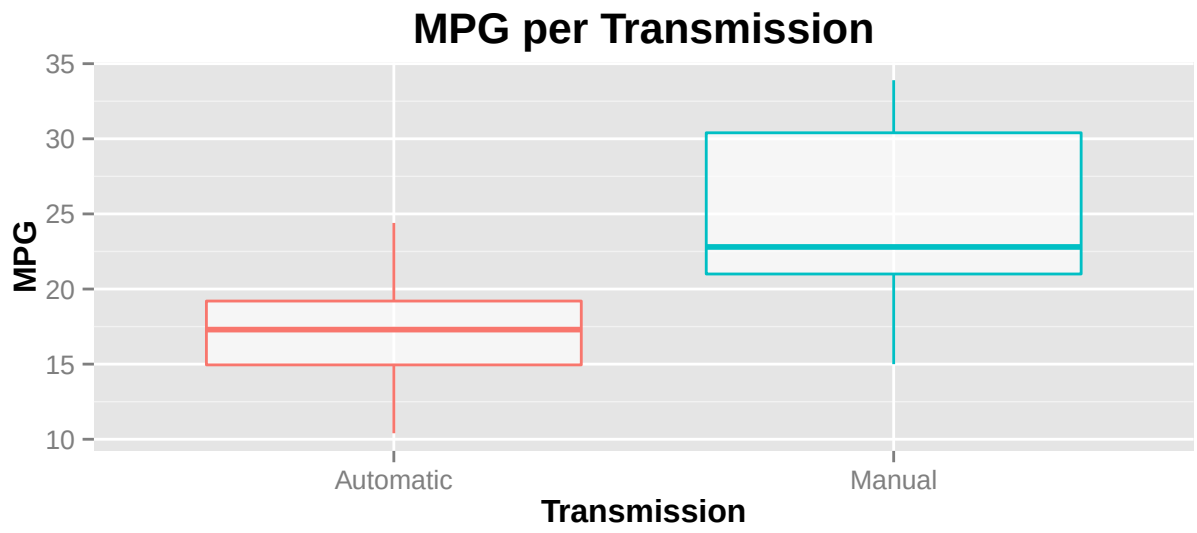

Exhibit 2.2 - Transmission t-Test

```
##
##  Welch Two Sample t-test
##
## data:  MPG by Transmission
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Exhibit 3.1 - Stepwise Regression on All Variables

```
##                      Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        33.70832390 2.60488618 12.940421 7.733392e-13
## Cylinders6         -3.03134449 1.40728351 -2.154040 4.068272e-02
## Cylinders8         -2.16367532 2.28425172 -0.947214 3.522509e-01
## Horsepower         -0.03210943 0.01369257 -2.345025 2.693461e-02
## Weight             -2.49682942 0.88558779 -2.819404 9.081408e-03
## TransmissionManual  1.80921138 1.39630450  1.295714 2.064597e-01

## $adj.r.squared
## [1] 0.8400875
```

Exhibit 3.2 - ANOVA Between Models

```
## Perform stepwise regression on model including transmission
## as the only explanatory variable
transmission.model <- lm(MPG ~ Transmission, mtcars)
stepwise.selected.transmission <- summary(step(transmission.model, direction = "both"))
```

```
## Start:  AIC=103.67
## MPG ~ Transmission
##
##                Df Sum of Sq    RSS    AIC
## <none>                       720.9 103.67
## - Transmission  1    405.15 1126.0 115.94
```

```
## Run analysis of variance on the two candidate models
anova(transmission.model, stepwise.selected.best.model)
```

```
## Analysis of Variance Table
##
## Model 1: MPG ~ Transmission
## Model 2: MPG ~ Cylinders + Horsepower + Weight + Transmission
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exhibit 3.3 - Diagnostic Plots