Relationship between Transmission Type and Fuel Efficiency

F. Alex Crofut

Saturday, April 30, 2016

Executive Summary

This analysis examines the fuel efficiency of the vehicles included in the mtcars dataset. Of special concern is the relationship between transmission type (automatic or manual) on fuel efficiency (mpg). This analysis will determine which transmission type is better for fuel efficiency and quantify the MPG difference between automatic and manual transmissions.

Assumptions

- Population is independent and identically distributed.
- Measurement error is accounted for by significant digits.
- A 95% confidence interval is sufficient to determine significance.

Exploratory Data Analysis

Required Packages

- datasets
- ggplot2
- gridExtra
- caret
- lattice

Data Structure and Processing

The dataset shows 32 observations of 11 variables concerning several key characteristics of the vehicles observed. The transmission type (am), a focus of the analysis, is denoted as either "0" (automatic) or "1" (manual). For ease of use, the numeric values of am are translated into factors "automatic" and "manual".

Data Exploration

As apparent in the box plot "Miles per Gallon by Transmission Type" (Appendix - Figure 1), there is a significant difference between the fuel efficiency in miles per gallon (mpg) for automatic and manual transmissions. Cars with manual transmissions have a mean mpg of 24.4, a full 7.2 improvement over automatics (17.1 mpg).

Models

In order to examine this relationship more closely, this analysis will apply a simple linear regression. If the linear model does not explain the regression variance as measured by the adjusted r-squared, the analysis will add additional variables to the model. The models will be compared via nested model testing to determine the best fit. That fit will be evaluated using the plot() function in R.

Simple Linear Regression

The first model utilized is a simple linear regression looking at the relationship between fuel efficiency and transmission type.

In this model, the confidence interval of [3.642, 10.848] does not include zero and the p-value of 0.000285 indicate the null hypothesis (transmission type is not related to fuel efficiency) can be rejected with a 95% confidence interval. However, based on the adjusted r-squared, only 33.8% of the variance is explained by our model.

Multivariable Regression Model

Due to the low percentage of variance explained by the simple linear model, additional variables will be added to the model. The variables were chosen utilizing a feature plot (Appendix - Figure 2) to identify weight ("wt") as being correlated to fuel efficiency. Variables correlated with weight were eliminated and quarter-mile times ("qsec") was retained because it was not correlated with weight. As the regressor of interest, the transmission type ("am") was added to evaluate significance. The summary of the multivariable regression model is shown in the Appendix as Figure 3.

The 95% confidence interval for transmission type is [0.046, 5.826],which excludes zero. With a p-value of 0.046716, the null hypothesis is again rejected in favor of the alternate hypothesis that fuel efficiency and transmission type are related. Additionally, the adjusted r-squared value shows that 33.8% of the variance is explained by this new model.

Nested Model Testing

A nested model test is utilized to determine the significance of the regressors included. Weight alone is examined in the first model, Quarter-mile time is added in the second, and transmission type is added in the final model.

the results of anova() show us the p-values of:

- Weight: 1.29e-10
- Quarter-mile Time: 0.000929
- Transmission Type: 0.046716

As seen above, each of the three regressors included in the multivariable regression model are significant with 95% confidence.

Diagnostics

The multivariable model was also tested using the plot() function in R (Appendix - Figure 4). The "Residuals vs Fitted" plot shows some, but not much, curvature, meaning this model does not require a quadratic component. The "Normal Q-Q" plot shows the residuals are normally distributed. The "Scale-Location" plot is tilted up, indicating the residuals may not be homoskedastic. The "Residuals vs Leverage" plot shows that no points were overly influential.

Conclusions

The cars with manual transmissions show a mean fuel efficiency 7.2 MPG higher than that for automatic transmissions. However, the weight of the vehicle and the quarter-mile time are confounding variables. Additionally, there is some doubt that our regressors are homoskedastic.

Appendix

Figure 1:

```
ggplot(aes(x = am, y = mpg), data = mtcars2) +
  geom_boxplot(fill = "blue") +
  labs(x = "Transmission Type", y = "Miles per Gallon") +
  ggtitle("Miles per Gallon by Transmission Type") +
  annotate(geom = "text", x = 1, y = meanMPGAuto,
           color="white", label = "-----", size=7, fontface="bold",
           angle = 0) +
  annotate(geom = "text", x = 2, y = meanMPGManual,
           color="white", label = "-----", size=7, fontface="bold",
           angle = 0)
```

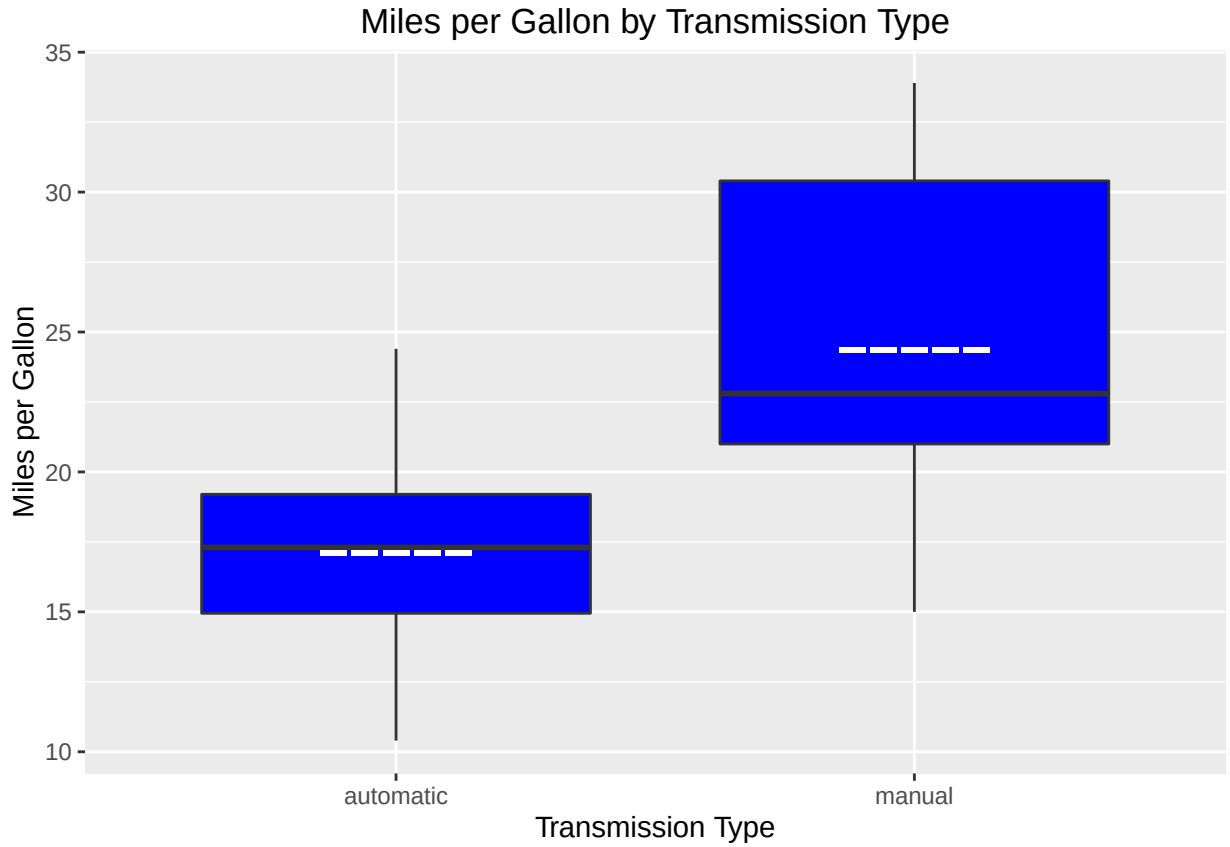

Figure 2:

```
featurePlot(x = mtcars2[,c(1, 3, 4, 6, 7)], y = mtcars2$am, plot = "pairs", col = c("red", "blue"),
            title = "Feature Plot for Variables in Analysis")
```

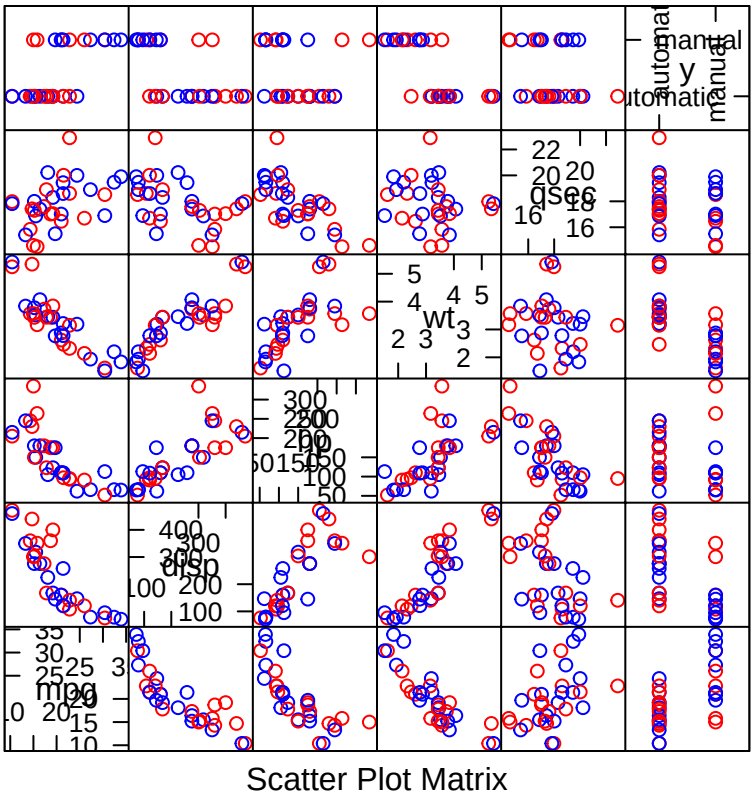

Figure 3:

Summary of Multivariable Regression Model

```
summary(fitM)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## ammanual      2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
```

```
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Figure 4:

Diagnostic Plots for Multivariable Regression Model

```
par(mfrow = c(2,2))
par(oma = c(0, 0, 2, 0))
plot(fitM)
```