Motor Trend Journal - February, 2015

A statistical analysis of the potential impact of transmission type on fuel efficiency

Executive Summary

This analysis attempts to answer two key questions:

1. What impact do automatic and manual transmissions have on the number of miles per gallon (MPG) obtained by a motor vehicle?
2. Can this difference, if any, be statistically quantified?

We use the 1974 Motor Trend dataset for this analysis that details fuel consumption and ten aspects of automobile design and performance for some thirty-two types of automobiles by brand and make.

Using a simple linear regression model, we conclude that, on average, a manual transmission provides a superior fuel efficiency by some **7.245** mpg.

The other most significant (and confounding) variables are cylinder count, weight, and horsepower. Once these are included in a multivariate analysis using ANOVA, it is revealed that the average improvement achieved by manual-tramsmisssion cars is around **1.809** mpg.

Interpretation of Coefficients of Interest

The analysis undertaken below yields two key coefficients: **7.245** (average improvement for manual transmission using simple linear regression) and **1.809** (reduced improvement using multivariate regression).

Exploratory Data Analyses

```
data(mtcars)
```

Preliminary data is explored upon loading, using a summary table of the mtcars dataset.

The output generated by the following code chunk can be viewed in the Appendix.

```
str(mtcars)
```

Convert `am` from numeric to factor class and provide labels for the levels - manual and automatic.

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Auto", "Manual")
```

The output generated by the following code chunk can be viewed in the Appendix.

```
hist(mtcars$mpg, breaks=10, col="blue", border = "black", xlab="Miles Per Gallon",
   main="")
```

The variable `mpg` appears to be distributed more or less normally and the data do not exhibit any significant skew. We can thus proceed with a simple linear regression analysis using `am` as the predictor variable.

Multiple Models and Selection Strategy

Hypothesis Test

Null hypothesis: The type of transmission has no influence on MPG.

```
aggregate(mpg~am, data = mtcars, mean)
```

```
##       am      mpg
## 1   Auto 17.14737
## 2 Manual 24.39231
```

Hence, on average, a manual car delivers an improvement of 7.245 mpg over an automatic.

The output generated by the following code chunk can be viewed in the Appendix.

```
boxplot(mpg ~ am, data = mtcars, col=c("steelblue","purple"),names = c("Automatic", "Manual"),las=1, font.lab=
2)
```

The box plot also reveals an approximately normal distribution with no significant outliers.

The results of the t-test below with a p-value of 0.001374 indicate that the null hypothesis is false and that the difference calculated above of **7.245 mpg** is statistically significant.

The output generated by the following code chunk can be viewed in the Appendix.

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

Selection of the most appropriate model

In order to build the most suitable model, one must determine which variables - apart from transmission type - have the most significant influence on fuel economy. The backward stepwise regression will initially include all the predictors, then progressively remove the ones that are not considered statistically significant.

The output generated by the following code chunk can be viewed in the Appendix.

```
complete.model <- lm(mpg ~ ., data = mtcars)
significant.model <- step(complete.model, direction="backward", k=2, trace=0)
summary(significant.model)
```

This model is clearly statistically significant for the following reasons:

1. The p-value is well below 0.05.
2. Adjusted R-squared is 0.8336, hence the model explains some 83% of the outcome's variance.
3. The coefficients are all non-zero, attesting to the significance of the model.
4. The summary also provides the greatest weight to the following variables: weight, horsepower, and number of cylinders

Simple Linear Regression

The output generated by the following code chunk can be viewed in the Appendix.

```
simple <- lm(mpg~am, data = mtcars)
summary(simple)
```

The Adjusted R-squared value is .3385, hence this simple linear model can explain only some 34% of the variance. We now need to ascertain whether a multiple regression model will provide more robustness.

Multiple Linear Regression

There are now two models with the same data set, hence an ANOVA will inform us if the there is a significant difference between the two models, i.e., if the one of the models is superior to the other.

The output generated by the following code chunk can be viewed in the Appendix.

```
simple <- lm(mpg~am, data = mtcars)
multivar <- lm(mpg~am + wt + hp, data = mtcars)
anova(simple, multivar)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results help us assess the appropriateness of the two models: The p-value of 3.745e-09 indicates that the simple linear regression model is clearly inferior to the multivariate one. However, we now need to review the residuals to ensure that they fit a linear model.

Residual Plot and Diagnostics

The output generated by the following code chunk can be viewed in the Appendix.

```
par(mfrow = c(2,2))
plot(multivar)
```

On examining the residuals against their fitted values, we detect no significant signs of non-linearity or heteroskedasticity, hence the multivariate regression model is the appropriate one to use.

Conclusion - Inference with Quantified Uncertainty

We can draw a number of conclusion from this more complex model:

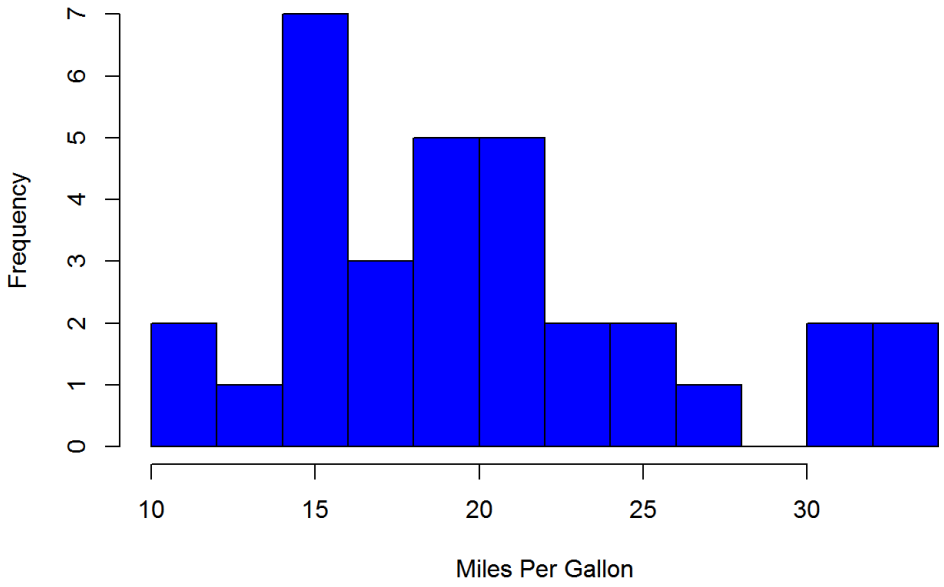
1. The model is able to account for some 84% of the variance.
2. Weight and horsepower exerted a confounding influence on the realtionship between miles per gallon and transmission type.
3. The coefficient for transmission type tells us that, on average, one can expect a superior fuel efficiency in a manual car by about 2.084 miles per gallon.
4. A level of uncertainty in this conclusion cannot be avoided: (a) the `am` variable in our model is not statistically significant; and, (b) the most important variable appears to be the weight of the vehicle and it is possible that automatic vehicles may, in general, tend to be weigh more. (The uncertainty is due to the variance in the data not being completely explainable by linear models.)
5. The 95% confidence interval of the difference in mean gas mileage is between 3.2097 and 11.2802 mpg (detailed in the t-test in the Appendix).

Appendix

Data Exploration

```
## '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  : Factor w/ 2 levels "Auto","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Histogram of mpg - near-normal distribution

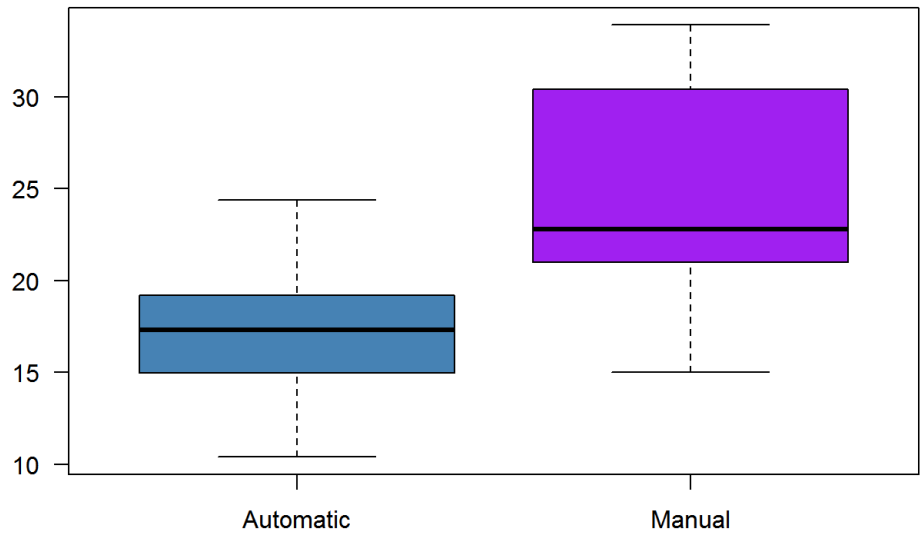


ANOVA table

```
simple <- lm(mpg~am, data = mtcars)
multivar <- lm(mpg~am + wt + hp, data = mtcars)
anova(simple, multivar)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Boxplot of mpg by transmission type



Simple linear regression

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Welch two-sample t-test

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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 Auto mean in group Manual
##             17.14737             24.39231
```

Backward Stepwise regression model

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## 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
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## 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
```

Residual plots