Fuel Economy for Manual vs. Automatic Transmissions

by: Ryan Swartz

Executive Summary

Using the 'mtcars' dataset, we conclude that vehicles with manual transmissions have greater fuel economy than those with automatic transmissions. A multiple linear regression model fit to the data indicates a 2.94 miles per gallon advantage for manual transmissions. The 95% confidence interval for this value is 0.05 to 5.83, further confirming manual transmission vehicles' higher fuel economy.

Analysis

Data Ingestion and Set Up

Loading and inspection of the 'mtcars' dataset included with R reveals a corpus of data containing eleven variable measures for thirty two different cars. As a note that will drive further analyses, each observation in the dataset is a unique car (i.e., the same car is not included twice with the only difference being the type of transmission).

```
data(mtcars)
library(ggplot2)
## add variable for a discernable label of transmission type
mtcars$tran.label[mtcars$am == 0] <- "automatic"
mtcars$tran.label[mtcars$am == 1] <- "manual"
```

Exploratory Analysis

To assess the possible difference in fuel economy between automatic transmission equipped cars and manual transmission equipped cars, we fit a simple linear regression (SLR) model to the data treating status of the vehicle transmission (variable 'am') as the predictor and miles per gallon of the vehicle (variable 'mpg') as the outcome.

```
slr.mpg <- lm(mpg ~ am, data = mtcars)
slr.mpg$coefficients

## (Intercept)          am
##      17.147       7.245
```

This model's slope indicates that having a manual transmission (am = 1) provides a 7.24 miles per gallon (mpg) advantage in fuel efficiency over an automatic transmission (am = 0). To illustrate this model, we also include (in the Appendix, as Figure 1) a scatterplot of the 'mpg' data separating the observations by transmission type and the SLR model quantifying the relationship. The line representing the model confirms that a vehicle with a manual transmission is positively correlated with fuel efficiency (mpg) as compared to vehicles with an automatic.

However, this is an oversimplification of the conditions, as the data is not comprised of pairs of the same car represented in both manual and automatic transmission variants. Many other variables of the vehicles are likely confounders to fuel efficiency including, but not limited to: weight, displacement, cylinders, and horsepower. The following section will pursue a model that balances for these other variables to provide a more accurate measure of the difference in fuel economy for a manual versus automatic transmission.

Formal Analysis

To determine this more accurate measure, we will fit a multiple linear regression (MLR) model to

the relevant variables of the data. We will then have a model that allows us to hold other factors of the vehicles constant - effectively enabling us to compare cars with equal characteristics with the exception of transmission type - and determine the true difference in fuel economy.

We employed a backwards selection strategy using p-value as the decision criteria to build the MLR model of interest (full details for this procedure presented in the Appendix).

```
## note: summary commented out for report length considerations
mlr.mpg.8 <- lm(mpg ~ am + wt + qsec, data = mtcars)
## summary(mlr.mpg.8)
## all predictors are now significant (p < 0.05)
mlr.mpg.8$coefficients

## (Intercept)          am          wt        qsec
##       9.618       2.936      -3.917       1.226
```

By fitting an MLR model, we can now say we've considered other potential factors affecting fuel economy in determining the relationship between transmission type and this measure. By definition of MLR, we can expect the fuel economy to increase by 2.94 miles per gallon for cars with manual transmissions as compared to those with automatic transmissions, with weight ('wt') and quarter-mile time ('qsec') held constant (i.e., comparing two like-cars, with only transmission type differing).

Noting that this figure differs significantly from that produced by the SLR model, we will examine the R2 and residuals to determine if the MLR model is indeed a better model for the data:

- R^2: since this is a measure of the amount of variability explained by the model, we can conclude the SLR model fit is very poor, with an R2 of 0.36. The MLR model is much better with an adjusted R2 (used for multiple variables) of 0.83.
- Residuals: a sum of the absolute value of the residuals for the SLR model is 125.9 while the MLR model registers 61.8; since residuals are a measure of the difference in actual outcomes versus those predicted by the model, the MLR model is a closer fit to the actual data.
- Residual plot: as a final check on the residuals, we present a residual plot in Figure 2 in the Appendix. Visually, the MLR model seems to have a much tighter closer fit around the x-axis as compared to the SLR residuals. The code in the Appendix below defines the approach to gather the residual data.

Results

After this analysis of the 'mtcars' data, we can conclude that vehicles with manual transmissions have greater fuel efficiency than those with automatic transmissions. Per a fit of an MLR model to the data, vehicles with a manual transmission have a predicted 2.94 mpg advantage over their automatic counterparts. Examining the 95% confidence interval for this coefficient of 0.05 on the low end, and 5.83 on the high end, we can further confirm our conclusion that manual transmission cars have higher fuel efficiency.

Appendix

Figures

```
fig.1 <- ggplot(mtcars, aes(x = tran.label, y = mpg)) +
      geom_smooth(method = "lm", color="blue", se = FALSE, aes(group = 1)) +
      geom_point(alpha = 0.75) +
      ggtitle("Transmission Type vs. Fuel Efficiency") +
      theme(plot.title = element_text(lineheight = 0.8, face = "bold")) +
      labs(x = "Transmission Type", y = "Fuel Efficiency [mpg]")
```

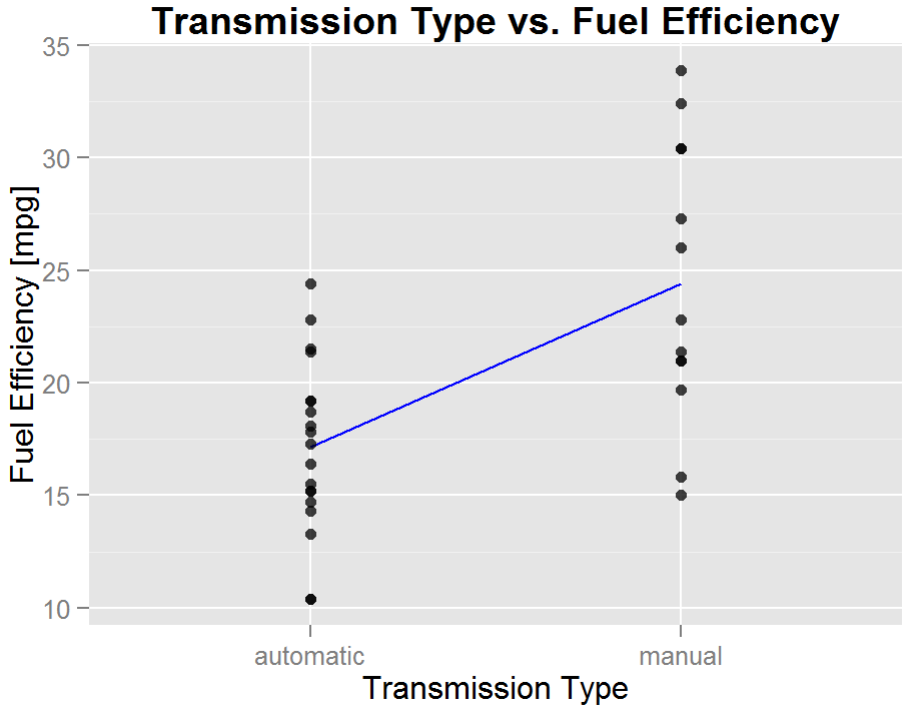

Fig. 1: Simple relationship between transmission type and fuel efficiency

```
## combine residuals in a dataframe ready for plotting
mpg.resid.1 <- data.frame(cbind(row.names(mtcars), resid(slr.mpg), "SLR"))
mpg.resid.2 <- data.frame(cbind(row.names(mtcars), resid(mlr.mpg.8), "MLR"))
mpg.resid <- rbind(mpg.resid.1, mpg.resid.2)
row.names(mpg.resid) <- seq(1:64)
## rename and clean up the class of data
names(mpg.resid) <- c("vehicle", "residual", "model")
mpg.resid$vehicle <- as.character(mpg.resid$vehicle)
mpg.resid$residual <-
as.numeric(levels(mpg.resid$residual))[mpg.resid$residual]
mpg.resid$model <- as.factor(mpg.resid$model)


fig.2 <- ggplot(mpg.resid, aes(x = vehicle, y = residual, color = model)) +
      geom_point(shape = 1, size = 2.5) +
```

```
    geom_hline(yintercept = 0) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.2)) +
    ggtitle("SLR vs. MLR Residual Comparison") +
    theme(plot.title = element_text(lineheight = 0.8, face = "bold")) +
    labs(x = "Vehicle", y = "Model Residual", color = "Model")
```

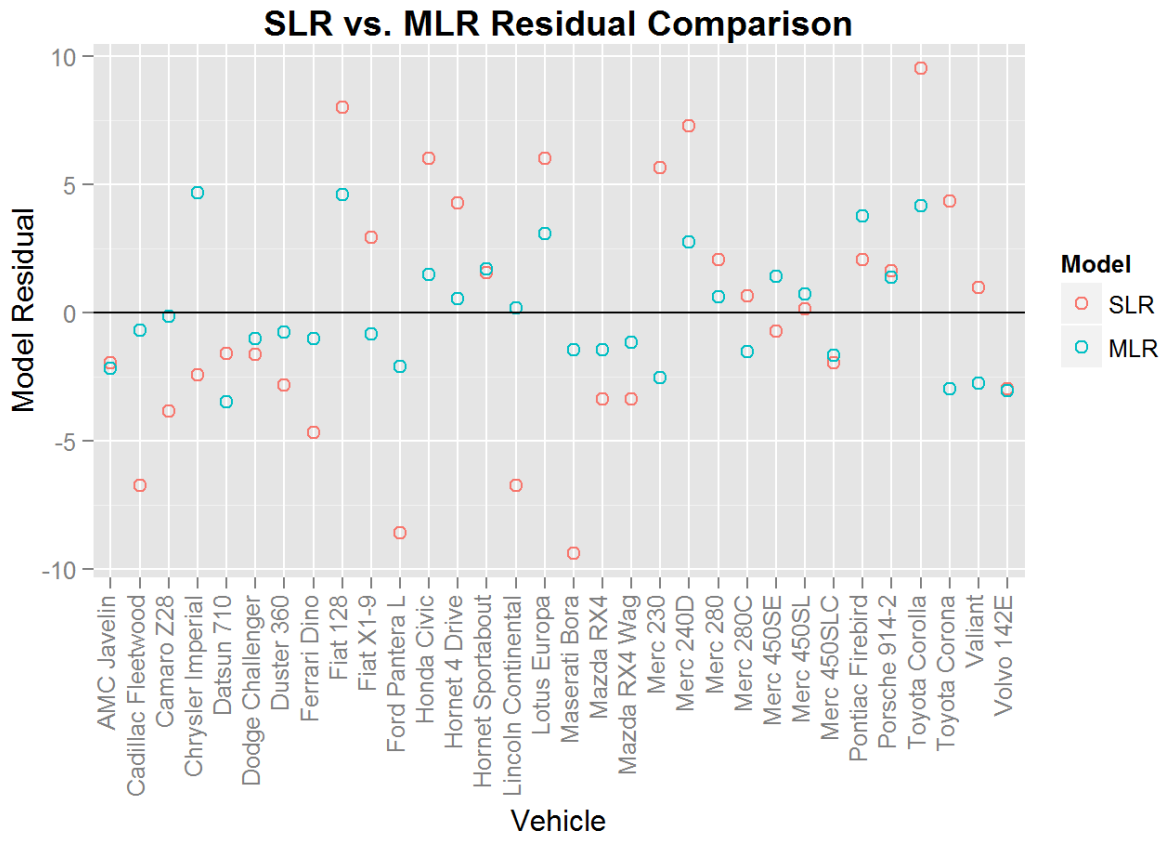

Fig. 2: Comparing residuals for SLR and MLR models

Model Iterations

We'll start with a model that has 'mpg' as the outcome and all other variables as predictors, and systematically remove variables from the model until only those predictors with a significant p-value are left.

```
## create initial model with all variables
## note: summary steps commented out for report length considerations
mlr.mpg.1 <- lm(mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear +
carb, data = mtcars)
## summary(mlr.mpg.1)
## drop 'cyl' predictor as it has largest p-value
mlr.mpg.2 <- lm(mpg ~ am + disp + hp + drat + wt + qsec + vs + gear + carb,
data = mtcars)
## summary(mlr.mpg.2)
## drop 'vs' predictor as it has largest p-value
mlr.mpg.3 <- lm(mpg ~ am + disp + hp + drat + wt + qsec + gear + carb, data =
mtcars)
```

```
## summary(mlr.mpg.3)
## drop 'carb' predictor as it has largest p-value
mlr.mpg.4 <- lm(mpg ~ am + disp + hp + drat + wt + qsec + gear, data =
mtcars)
## summary(mlr.mpg.4)
## drop 'gear' predictor as it has largest p-value
mlr.mpg.5 <- lm(mpg ~ am + disp + hp + drat + wt + qsec, data = mtcars)
## summary(mlr.mpg.5)
## drop 'drat' predictor as it has largest p-value
mlr.mpg.6 <- lm(mpg ~ am + disp + hp + wt + qsec, data = mtcars)
## summary(mlr.mpg.6)
## drop 'disp' predictor as it has largest p-value
mlr.mpg.7 <- lm(mpg ~ am + hp + wt + qsec, data = mtcars)
## summary(mlr.mpg.7)
## drop 'hp' predictor as it has largest p-value
mlr.mpg.8 <- lm(mpg ~ am + wt + qsec, data = mtcars)
## summary(mlr.mpg.8)
## all predictors are now significant (p < 0.05)
```