The MPG Difference between 1973 Cars with Automatic and Manual Transmissions

Submission by Connor Lenio. Email: cojamalo@gmail.com. Completion Date: Apr 28, 2017.

Executive Summary

A popular metric of a car's fuel economy is its miles-per-gallon rating (MPG). There are many features of a car that may impact its MPG. This analysis explores the following two concerns:

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

The analysis uses data for 1973 model cars and features inference for the difference between means and single linear regression to consider transmission type as sole predictor for MPG. In isolation, transmission type is a moderately weak predictor for MPG. The analysis continues by incorporating the other factors available in the data for determining MPG using multiple linear regression to determine how transmission type effects MPG when other features of a car are considered. With other more informative factors included and a stronger model fit to the data, transmission type is not expected to be a significant predictor for the MPG of 1973 cars.

Getting Started

Load packages

```
library(DAAG); library(AICcmodavg);library(pander); library(ggplot2); library(gridExtra); library(dplyr
if(!exists("tree_lm", mode="function")) source("mult_regres.R")
```

This analysis will rely on custom functions in "mult_regres.r." The code for these functions is kept separate for brevity, but can be viewed at any time at my mult_regres.R git repository.

The data

```
data(mtcars)
```

From the package datasets: "The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models)."

The explanatory variable for transmission type (0 = automatic, 1 = manual) is `am`, and the response variable for MPG in miles per gallon is `mpg`.

Exploratory Data Analysis

First, the data is summarized by `am` to compare the sample size, mean, and variance of the automatic and manual cars.

```
mtcars_table <- mtcars %>%
            tbl_df %>%
            group_by(am) %>%
            summarize(n = n(), Mean = round(mean(mpg),1),
```

```
                      Variance = round(var(mpg), 1))
pandoc.table(mtcars_table)
```

am	n	Mean	Variance
0	19	17.1	14.7
1	13	24.4	38.0

The sample sizes are unequal and have different variances. Moreover, the mean `mpg` of the automatic cars is lower than the manual cars. For a plot of the distribution and density of the mtcars data grouped by `am`, please view Figure 1 in Appendix A: Figures.

Inferential Analysis

One way to address "Is an automatic or manual transmission better for MPG" is to compare the mean of the two groups and test for statistical significance. A valid two sample t-test assumes the data represents IID samples and that the sampling distribution of the response variable under consideration in both sample populations is normal. Each observation is a single, unique car and one can assume `mpg` varies normally in the population of 1973 model cars.

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

```
##
##  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 0 mean in group 1
##        17.14737        24.39231
```

The p-value for the t-test is 0.001. Thus, 1973 manual cars have a statistically significant 7.3 mpg $\pm$ 4.0 mpg greater mpg than 1973 automatic cars.

Simple Linear Regression Model

In order to quantify this relationship, single linear regression is used with the formula, `mpg ~ am`.

```
fit1 <- lm(mpg ~ am, mtcars)
summary(fit1)
```

```
##
## 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 ***
## am             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
```

For a plot of this linear regression over the data, please view Figure 2 in Appendix A: Figures.

The linear regression suggests that MPG has a moderate linear correlation with transmission type (r = 0.5998). The model is of the form:

$$mpg = 17.15 + 7.24 * am + e$$

The p-value for the slope estimate for `am` is < 0.05, so this relationship between `am` and `mpg` is statistically significant. However, the R-squared for the model is only 0.3598, signifying that about 36.0% of the variability in MPG can be explained by transmission type using this model. Thus, if all one knew about a 1973 model car was its transmission type, then `fit1` would give them an estimate, but with a large variability in accuracy for predicting MPG. It is worth investigating other models that include `am` to utilize the additional information in `mtcars` to identifying confounding explanatory variables for `mpg` and to determine a better fit for the data.

Multiple Regression Modelling

Before, evaluating more models, each explanatory variable in `mtcars` is compared to `mpg` to determine if any linear transformations may be performed to better fit the data such as logistic or reciprocal transformations. The custom functions `find_best_trans` and `add_best_trans` are used from my mult_regres.R git repository. In short, the functions determine if transformations of the variables lead to any better fits, then adds these transformed variables to the data set, `mtcars`, yielding `new_mtcars`.

```
new_mtcars <- add_best_trans(find_best_trans(mpg,mtcars),mtcars)
glimpse(new_mtcars[,(ncol(mtcars)+1):ncol(new_mtcars)])
```

```
## Observations: 32
## Variables: 8
## $ recip_disp <dbl> 0.006250000, 0.006250000, 0.009259259, 0.003875969,...
## $ log_wt     <dbl> 0.9631743, 1.0560527, 0.8415672, 1.1678274, 1.23547...
## $ recip_hp   <dbl> 0.009090909, 0.009090909, 0.010752688, 0.009090909,...
## $ log_cyl    <dbl> 1.791759, 1.791759, 1.386294, 1.791759, 2.079442, 1...
## $ drat^2     <dbl> 15.2100, 15.2100, 14.8225, 9.4864, 9.9225, 7.6176, ...
## $ log_carb   <dbl> 1.3862944, 1.3862944, 0.0000000, 0.0000000, 0.69314...
## $ recip_gear <dbl> 0.2500000, 0.2500000, 0.2500000, 0.3333333, 0.33333...
## $ recip_qsec <dbl> 0.06075334, 0.05875441, 0.05373455, 0.05144033, 0.0...
```

A glimpse of the newly added transformed variables reveals eight variables that may produce better linear fits if transformed.

Next, the custom function, `tree_lm`, from my mult_regres.R git repository is used to search using recursion for the best formulas for predicting `mpg` using `new_mtcars`. The function prioritizes parsimony by constructing candidate formulas step by step and returning the best models by the quality and the number of predictors included. In this particular case, the Akaike information criterion (AICc) method is used to find the "best" models.

```
pandoc.table(tree_lm("mpg ~ am", new_mtcars, "AICc", kfold=TRUE), round = 3)
```

```
## [1] "Processing, please wait...."
##
## ------------------------------------------------------------------------------------------
##             model               terms   adj_R_2    BIC      AICc    LOOCV    KFOLD
## ------------------------------ ------- --------- -------- -------- -------- --------
## mpg ~ am + recip_disp + log_wt    4    0.8893981 150.6710 145.2365 5.133758 5.247844
##           + recip_hp
##
##  mpg ~ am + recip_hp + log_wt     3    0.8817608 150.5057 145.4847 5.305080 5.379610
##
##     mpg ~ am + recip_disp         2    0.8502837 155.7159 151.3345 5.851149 5.867531
## ------------------------------------------------------------------------------------------
```

Three models are suggested by `tree_lm`:

```
fit2 <- lm(mpg ~ am + recip_disp,new_mtcars)
fit3 <- lm(mpg ~ am + recip_hp + log_wt,new_mtcars)
fit4 <- lm(mpg ~ am + recip_disp + log_wt + recip_hp,new_mtcars)
```

An ANOVA is run, starting with `fit1`, in order of a number of predictors to determine the point at which adding another predictor no longer yields a significantly improved model.

```
anova(fit1, fit2, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + recip_disp
## Model 3: mpg ~ am + recip_hp + log_wt
## Model 4: mpg ~ am + recip_disp + log_wt + recip_hp
##   Res.Df    RSS Df Sum of Sq        F    Pr(>F)
## 1     30 720.90
## 2     29 157.71  1    563.19 140.1825 3.373e-12 ***
## 3     28 120.26  1     37.45   9.3224   0.00504 **
## 4     27 108.47  1     11.79   2.9334   0.09823 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The fourth model is no longer significantly improved over model three.

Thus, `fit3` is selected as the best model for this analysis.

```
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am + recip_hp + log_wt, data = new_mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5736 -1.3726 -0.3296  1.3279  4.3481
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.7465     3.9488   6.773 2.35e-07 ***
```

```
## am            0.2045     1.1249   0.182 0.857049
## recip_hp    634.0900   140.3124   4.519 0.000103 ***
## log_wt      -10.7888     2.4054  -4.485 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.072 on 28 degrees of freedom
## Multiple R-squared:  0.8932, Adjusted R-squared:  0.8818
## F-statistic: 78.06 on 3 and 28 DF,  p-value: 1.032e-13
```

The multiple regression suggests that MPG has a strong linear correlation with transmission type, the reciprocal of horsepower, and the log of weight. (r = 0.945). The model is of the form:

$$mpg = 26.747 + 0.205 * am + \frac{634.09}{disp} - 10.789 * \log(wt) + e$$

The adjusted R-squared for the model is 0.8818. Thus, about 88.2% of the variability in MPG can be explained by the explanatory variables in this model.

However, the role of transmission in this model is not significant as its p-value is greater than 0.05. Therefore, it is not unlikely that there is no relationship between `am` and `mpg` when considering the information provided by the other two predictors. The interesting conclusion from the multiple regression is that there is not likely to be a significant difference in `mpg` for 1973 automatic and manual cars when accounting for other attributes of 1973 cars that effect `mpg`.

Residual Analysis

For the residual plots for `fit3`, please view Figure 3 in Appendix A: Figures.

A valid linear analysis involves:

(1) A linear relationship between each (numerical) explanatory variable and the response

Looking at the bottom row of Figure 4 in in Appendix A: Figures, there is a linear relationship between `mpg` and each of the explanatory variables.

(2) Nearly normal residuals with a mean of zero

Yes, the quantile-quantile plot of the residuals in Figure 3 confirms a normal distribution and residuals are centered around zero.

(3) Constant variability of residuals

Yes, but heteroscedasticity may still be an issue. The Scale-Location plot in Figure 3 does not show a straight line from left to right, so there may be differences in variability of residuals. One must be wary that the model may result in biased prediction accuracies for different cars.

(4) Independence of residuals (and hence observations)

Yes, the residuals versus fitted plot in Figure 3, the residuals all appear independent

Conclusion

If all that was known about a 1973 model car was its transmission type, then a manual car is expected to have 7.245 greater MPG than an automatic car. However, transmission type is not the most informative feature of a 1973 car for determining its MPG. Thus, when more informative variables like the car's weight and engine power (horsepower) are considered, transmission type is not likely to effect a 1973 car's MPG. In that case, there is no MPG difference between automatic and manual transmissions.

Appendix A: Figures

Figure 1: The following is a plot of the distribution and density of the mtcars data grouped by `am`:

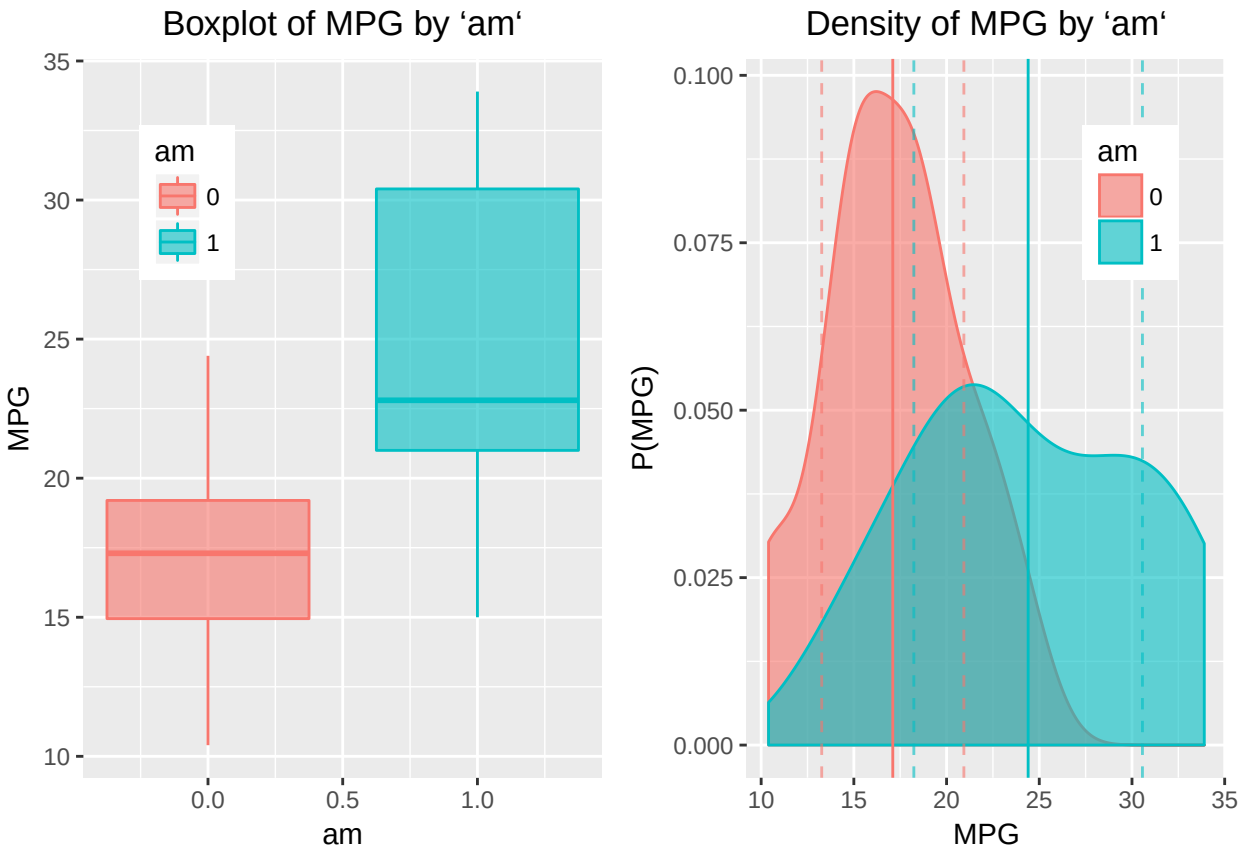

Figure 2: Plot of this linear regression using formula `mpg ~ am`:

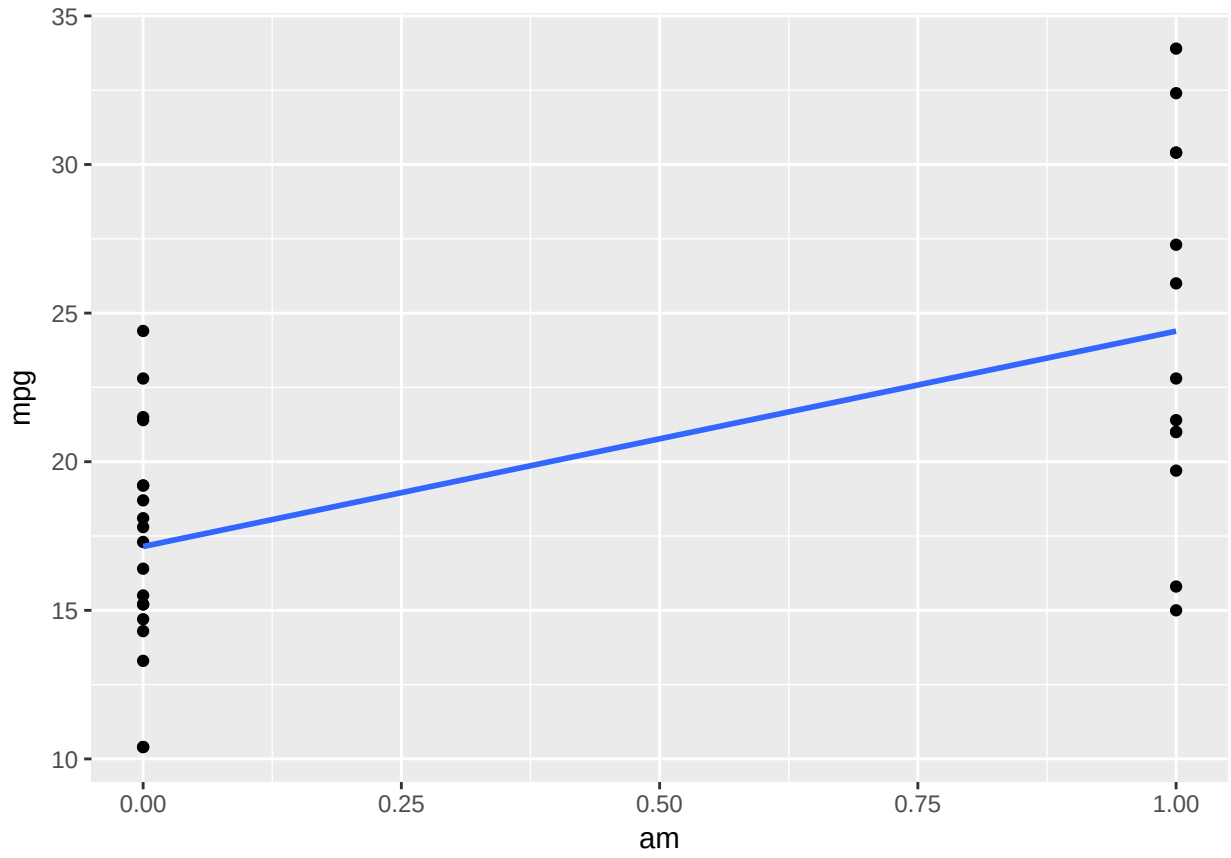

Figure 3: Four-panel Residual Plot of `fit3`, for formula `mpg ~ am + recip_hp + log_wt`:

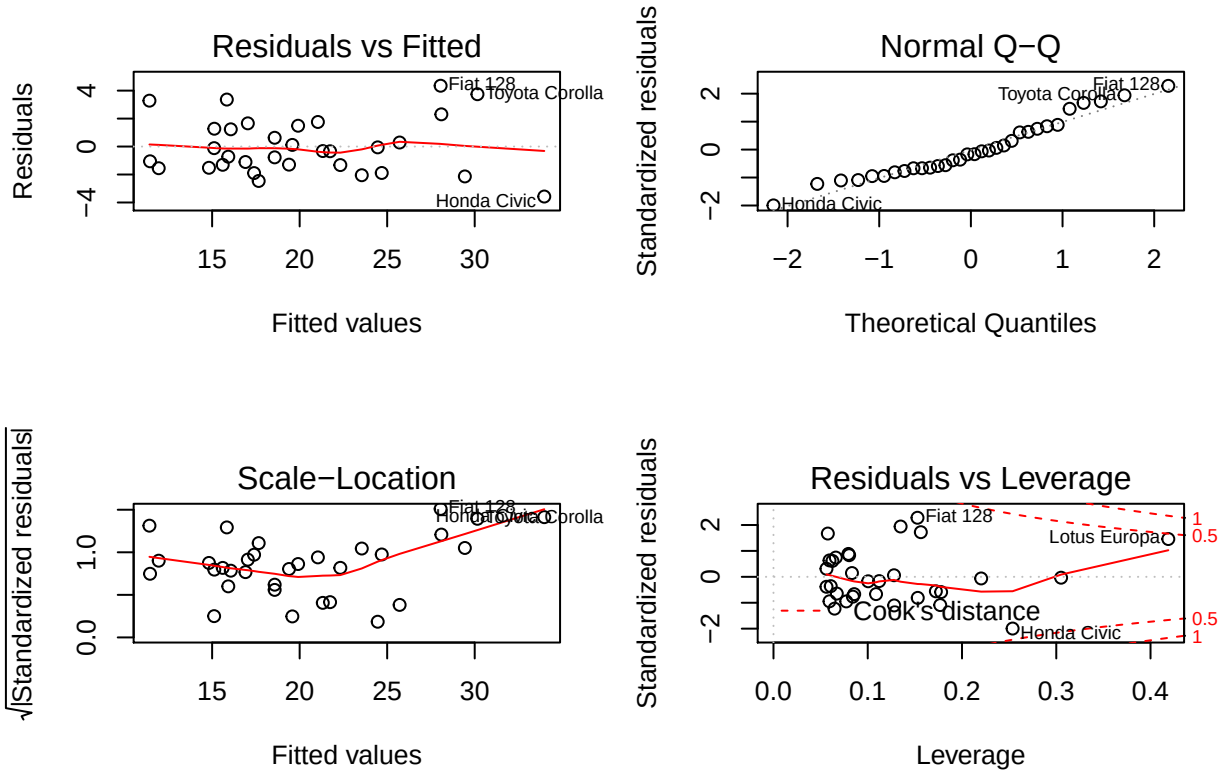

Figure 4: GGpairs plot of `fit3`, for formula `mpg ~ am + recip_hp + log_wt`: