Automatic v. Manual Transmission MPG Study

Paul Cappa

December 18, 2015

Synopsis

Looking at a data set of a collection of cars, we is interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). We are particularly interested in the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. "Quantify the MPG difference between automatic and manual transmissions"?

Preparing the Data for the Analysis

The first step in the anlysis is to prepare the data. The sub-steps performed are as follows:

1. Load the **'mtcars'** dataset.
2. Remove the **'qsec'** variable (qsec is the 1/4 mile time and is irrelevant to this analysis).
3. Convert key variables into factors. Note: Renaming the variable **'am'** factors to 'Auto' and 'Man' will make plots much more intuitive (versus '0' and '1'). Also, renaming the variable **'vs'** factors to 'V' and 'S' are done for the same reason (should we need them).
4. The remaining variables are continuous and therefore are not converted to factors.

```
data(mtcars)                                          # Load the dataset
mtcars <- mtcars[, c(1:6,8:11)]                       # Remove the 'qsec' variable.
mtcars$am   <- factor(mtcars$am,labels=c("Auto","Man"))  # Factor 'am' to 'Auto' and 'Man'
mtcars$cyl  <- factor(mtcars$cyl)                     # Convert cylinders to a factor.
mtcars$vs   <- factor(mtcars$vs, labels=c("V", "S"))  # Factor the engine type to 'V' and 'S'
mtcars$gear <- factor(mtcars$gear)                    # Convert the number of gears to a factor.
mtcars$carb <- factor(mtcars$carb)                    # Convert the number of carburators to a fa
```

Exploratory Data Analysis

Next, we take a preliminary look at the data. This will be performed with the following sub-steps.

1. First, we take a quick look at the statistics using a box plot (See: Appendix, Figure 2). At first glance, the box plot shows the manual transmission offers better performance over the automatic transmission.
2. Next, we will look for potential correlation between 'mpg'and the other factors (See: Appendix, Figure 1).
3. Compare the the means for the auto transmissions and manual transmissions, not considering any of the other variables, such as cylinders, weight, engine type, etc.)

Comparison of the Means

```
paste0("Auto Mean = ", round(mean(mtcars$mpg[mtcars$am=="Auto"]),2),
    ", Man Mean = ", round(mean(mtcars$mpg[mtcars$am=="Man"]),2))
```

```
## [1] "Auto Mean = 17.15, Man Mean = 24.39"
```

Statistical Inference

The null hypothesis is that the manual and automatic transmissions are represeted by the same distribution. The assumption is that the sample, represented within the **'mtcars'** data set is representative of the total population of vehicles.

```
t <- t.test(mpg ~ am, data = mtcars)
t$p.value
```

```
## [1] 0.001373638
```

With a p-value of .001373638, we can comfortably reject the null hypothesis and say the manual and automatic transmission data are from different distributions.

Regression Analysis

The regression analysis will happen in the following sub-steps:

1. Fit the full model.
2. Step through the full model to find the best fit.

```
fit <- lm(mpg ~ am + ., data = mtcars)                    # Fit the full model
best <- step(fit, direction = "both", trace=FALSE)      # Find the step model.
summary(best)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## amMan        1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Residuals Analysys

The final step in the analysis is to review the Residuals (See: Appendix, Figure 3).

1. The points on the **"Residuals vs Fitted"** plot are randomly distributed, indicating the model is sutiable for the dataset.
2. The points on the **"Normal Q-Q"** are relatively linear suggesting the residuals folow a normal distribution with what appear to be slight bends, possibly suggesting more variance that would be expected in a normal distribution.
3. The points on the **"The Scale-Location"** appear randomly distributed, again, suggest a good model.
4. The **"Residuals and Leverage"** plot demonstrates there is little influence of any of the control variables, thereby suggesting a fit of the model, although The Chrysler Imperial, Toyota Corona, and Toyota Corolla have an elevated influence.

Conclusions

1. There is staticical significance between

Appendix - Supporting Figures

Figure 1 - Side by Side Comparison of Manual and Automatic Transmissions

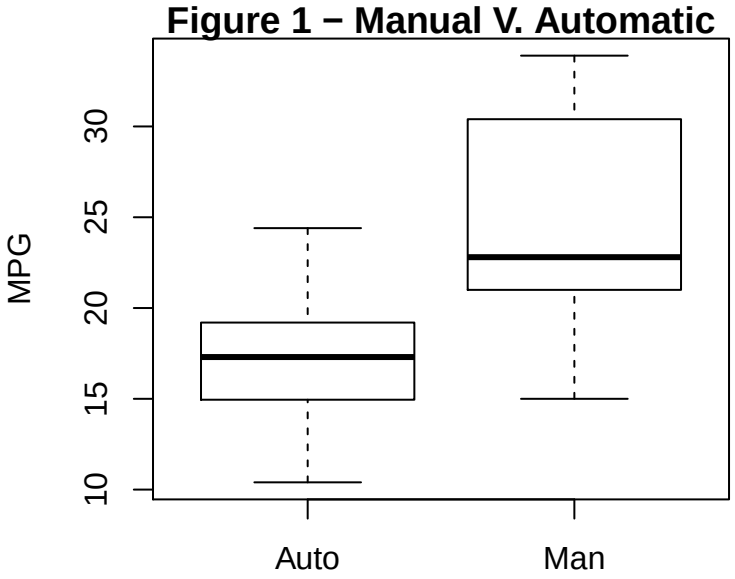

Figure 2 - Examining Potential Correlation MPG and Other Factors

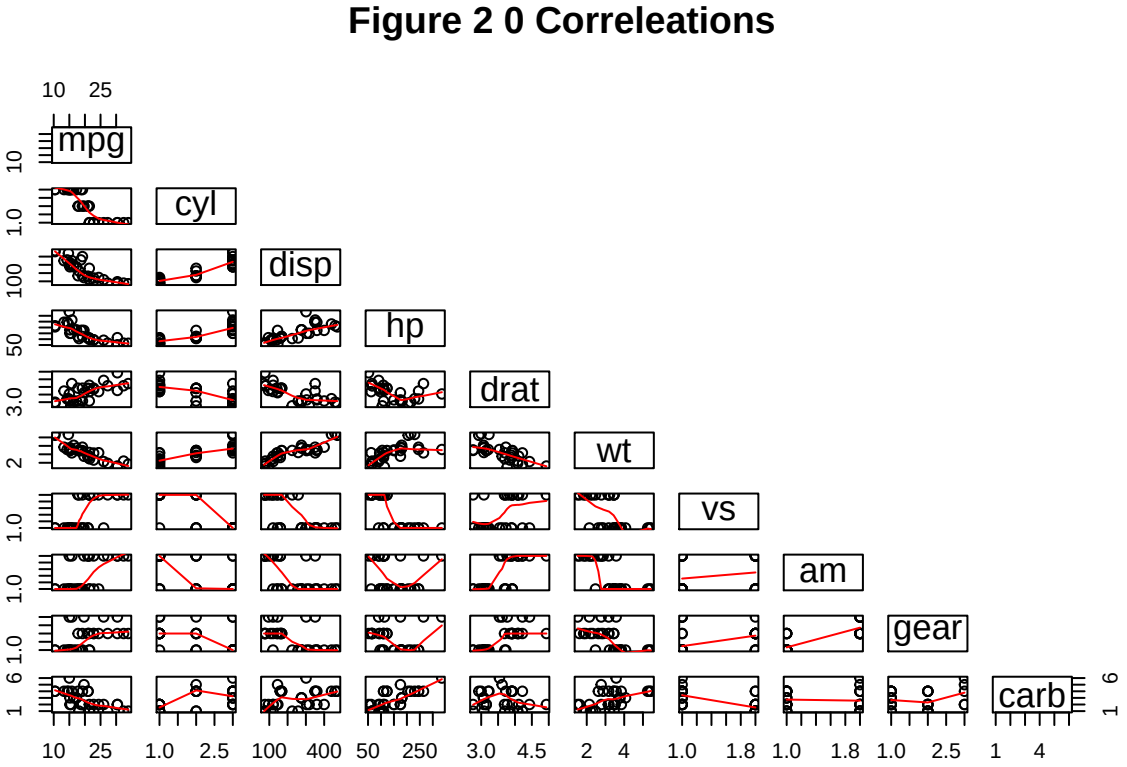

Figure 3 - Residuals and Diagnostic Plots