Motor Trend Car Road Tests

Robert Blaser

Friday, December 19, 2014

Executive Summary:

Looking at a data set of a collection of cars, we are 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" and 2) "Quantify the MPG difference between automatic and manual transmissions".

The mtcars data set consists of 32 observations on 11 variables: -mpg: Miles/(US) gallon, -cyl: Number of cylinders, -disp: Displacement (cu.in.), -hp: Gross horsepower, -drat: Rear axle ratio, -wt: Weight (lb/1000), -qsec: 1/4 mile time, -vs: V/S, -am: Transmission (0 = automatic, 1 = manual), -gear: Number of forward gears, -carb: Number of carburetors.

In this report, I applied simple linear regression techniques to show relationships of between the regression variables, with specific focus on "mpg" versus "am" variable. This report shows that there is a significant relationship between "mpg" and the "am" variable where "manual" has a significant increase in "mpg" versus "automatic".

Throughout the report, please see the "Appendix" section for script outputs and supporting details.

Data Processing:

In the following, I load the data transform the some of the appropriate discrete variables into factors:

```
data(mtcars)
mtcars$am <- as.factor(mtcars$am)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
levels(mtcars$am) <- c("Automatic", "Manual")
```

Exploratory Analysis:

In this section, I perform exploratory analysis to gain insights into the data set and the various relationships. As shown in the **Appendix**, I calculate the means and standard deviations of "mpg" for "automatic" and "manual" transmissions. The output shows that the mean for "mpg" with different transmission types are significantly different as well as the standard deviations for "mpg" with different transmission types being significantly different. In addition, I generate a boxplot of "mpg" as a function of "am" which clearly shows the difference in "mpg" versus transmission type.

Model Investigations:

As follows, I build and demonstrate regression models to gain further insight into the relationships. I perform a linear regression of the data set for "mpg" as a function of the remaining variables as follows:

```
lm(mpg ~ ., data = mtcars)
```

Taking a coefficient summary of the above as shown in the **Appendix**, I find that "am", "wt" and "hp" have strong relationships with "mpg". Thus, I will compare doing a simple linear regression for "mpg" as function of "am" with no adjusters and compare this to a linear regression with "wt" and "hp" added to see the effects:

```
lm(mpg ~ am, data = mtcars)
lm(mpg ~ am + wt + hp, data = mtcars)
```

Taking the summary of the above as shown in the **Appendix**, including "wt" and "hp" in the model, increases the Adjusted R-squared from 34% to 81%. Thus, including "wt" and "hp" in the model captures an acceptable number of contributors and therefore will be used as my linear model going forward. as follows, I perform an ANOVA of the two models for further insight into my decision:

```
anova(mpg_am_model,mpg_am_wt_hp_model)
```

As seen in the **Appendix**, the p-value is highly significant, and thus I reject the null hypothesis that the confounder variables "hp" & "wt" do not contribute to the accuracy of the model.

Model Residuals & Diagnostics:

In this section, I investigate the Residuals and the Influencers by investigating the hatvalues and dfbeta of the model as follows:

```
hatvalues(mpg_am_wt_hp_model)
dfbetas(mpg_am_wt_hp_model)
hist( mpg_am_wt_hp_model$residuals)
```

As seen in the results in the **Appendix**, the highest leverage vehicles are shown.

Statistical Inference:

To ascertain the statistical inference of "mpg" versus "am" as shown in the **Appendix**, I first due an aggregate of "mpg" versus "am" to show the delta in average "mpg" versus "am". Then I conduct a t-test as follows:

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

Based on the t-test results as shown in the **Appendix**, I reject the null hypothesis that the mpg distributions for manual and automatic transmissions are the same.

Conclusion:

In this report, I has shown that there is a significant difference in "Miles per Gallon" (mpg) between "Automatic" versus "Manual" transmissions. I have performed an exploratory analysis, built & analyzed a linear model that captures 81% of the data set contributions to "mpg", explored the residuals and influencers within the data set model and performed a statistical inference to determine that there is a significant statistical difference in "mpg" between "Automatic" versus "Manual" transmissions. **The findings show that: 1) the "mpg" difference between manual versus automatic transmissions is 2.08 mpg (manual has higher mpg tham automatic) and 2) the "mpg" change versus "wt" is -2.82 mpg per 1000 lb increase in weight.**

Please see the **Appendix** for supporting details and calculation results.

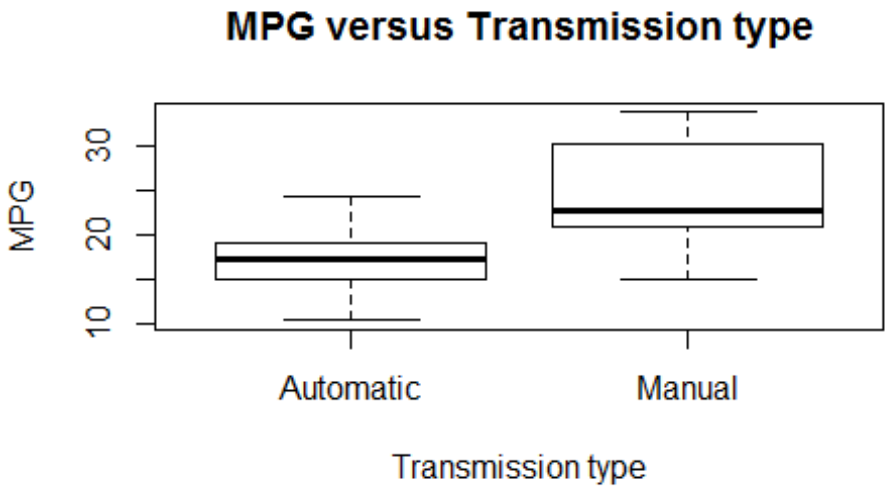
Appendix:

Exploratory Analysis support material:

```
summary(mtcars$mpg[mtcars$am=="Automatic"])
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.4    15.0    17.3    17.1    19.2    24.4
summary(mtcars$mpg[mtcars$am=="Manual"])
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    21.0    22.8    24.4    30.4    33.9

sd(mtcars$mpg[mtcars$am=="Automatic"])
## [1] 3.834
sd(mtcars$mpg[mtcars$am=="Manual"])
## [1] 6.167

boxplot(mpg ~ am, data = mtcars, main = "MPG versus Transmission type",
        xlab = "Transmission type", ylab = "MPG")
```



Model Investigations support material:

```
overview_mod <- lm(mpg ~ ., data = mtcars)
summary(overview_mod)$coeff
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 23.87913   20.06582  1.19004  0.25253
## cyl6        -2.64870    3.04089 -0.87103  0.39747
## cyl8        -0.33616    7.15954 -0.04695  0.96317
## disp         0.03555    0.03190  1.11433  0.28267
## hp          -0.07051    0.03943 -1.78835  0.09393
## drat         1.18283    2.48348  0.47628  0.64074
## wt          -4.52978    2.53875 -1.78426  0.09462
## qsec         0.36784    0.93540  0.39325  0.69967
## vs           1.93085    2.87126  0.67248  0.51151
## amManual     1.21212    3.21355  0.37719  0.71132
## gear4        1.11435    3.79952  0.29329  0.77332
## gear5        2.52840    3.73636  0.67670  0.50890
## carb2       -0.97935    2.31797 -0.42250  0.67865
## carb3        2.99964    4.29355  0.69864  0.49547
## carb4        1.09142    4.44962  0.24528  0.80956
```

```
## carb6        4.47757    6.38406  0.70137  0.49381
## carb8        7.25041    8.36057  0.86722  0.39948
```

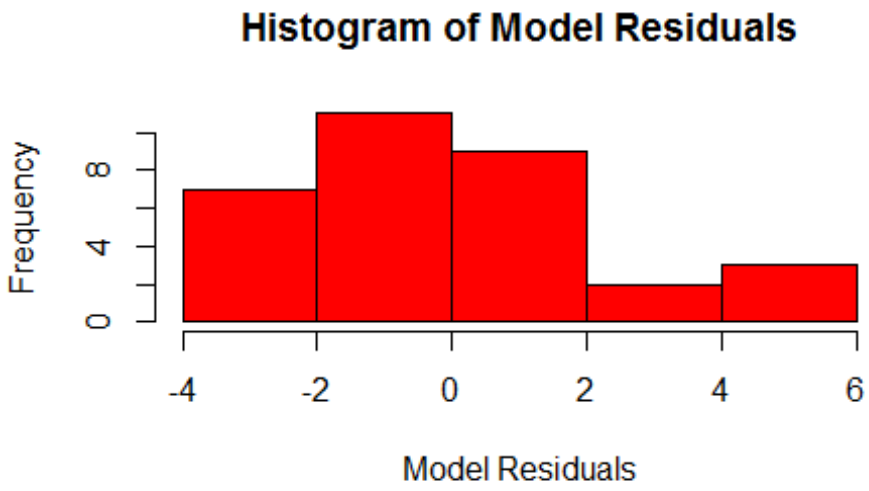
```
mpg_am_model <- lm(mpg ~ am, data = mtcars)
summary(mpg_am_model)$coeff
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.134e-15
## amManual       7.245      1.764   4.106 2.850e-04
summary(mpg_am_model)$adj.r.squared
## [1] 0.3385
```

```
mpg_am_wt_hp_model <- lm(mpg ~ am + wt + hp, data = mtcars)
summary(mpg_am_wt_hp_model)$coeff
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 34.00288   2.642659  12.867 2.824e-13
## amManual     2.08371   1.376420   1.514 1.413e-01
## wt          -2.87858   0.904971  -3.181 3.574e-03
## hp          -0.03748   0.009605  -3.902 5.464e-04
summary(mpg_am_wt_hp_model)$adj.r.squared
## [1] 0.8227
```

```
anova(mpg_am_model,mpg_am_wt_hp_model)
## 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     30 721
## 2     28 180  2       541 42 3.7e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Residuals & Diagnostics support material:

```
lev_meas <- hatvalues(mpg_am_wt_hp_model)
tail(sort(lev_meas),4)
##   Chrysler Imperial  Cadillac Fleetwood Lincoln Continental
##              0.2303              0.2350              0.2726
##       Maserati Bora
##              0.4122
inf_meas <- dfbetas(mpg_am_wt_hp_model)
tail(sort(inf_meas[,4]),4)
##    Volvo 142E Mazda RX4 Wag  Lotus Europa Maserati Bora
##        0.1847        0.2183        0.2364        0.5810
```



hist(data_subset$Global_active_power, col = "red", main = "Global Active Power", xlab = "Global Active Power(kilowatts)")

Statistical Inference support material:

```
aggregate(mpg~am, data = mtcars, mean)
##          am   mpg
## 1 Automatic 17.15
## 2    Manual 24.39
t.test(mpg ~ am, data = mtcars)
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean in group Automatic    mean in group Manual
##                   17.15                   24.39
```