Regression Models Course Project: Analysis of Mileage and Transmission Type

juradom

April 17, 2015

Executive Summary The goal of this analysis is to answer the following two questions leveraging the mtcars dataset within R:

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

Based on my analysis I conclude that manual transmission vehicles are more efficient than automatic transmissions for MPG. Manual transmission vehicles are predicted to get 4.36 mpg better than their automatic counterparts.

Exploratory Analysis As described in the R documentation, the mtcars dataset is derived from a 1974 Motor Trend magazine.

- **mtcars Structure**:
  - This data contains 32 observation of 11 variables.
  - 19 of the observations are automatic transmission; 13 are manual

  - *Appendix A* contains other structural info about the mtcars dataset.
- Research note: Around the time of the Motor Trend study, a 1975 EPA document was published indicating that weight and engine displacement are the most important factors in fuel economy (http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=9100S2LD.txt)
- **Mean Values**: The average miles per gallon of the manual sample is: **24.39** and the automatic transmission mean is: **17.15**.
- **Variances**: The variances for manual and automatic transmission are: **38.03** and **14.7**, respectively. I suspect the variance issue is due to the small sample size. A greater sample size would likely result in identical variances.
- **Normality**: *Appendix B* shows two density plots that appear to be normal but to be sure I performed a Shapiro test for normality. The p-value for the manual plot is: **0.54**. The p-value for the automatic plot is: **0.9**. Since null hypothesis in the Shapiro tests **against** the assumption of normalty, we cannot reject the null hypothesis because both p-values are $> .05$.

Regression Analysis

Means Testing The example model can be defined as: lm(mpg~am, data=mtcars). The p-value for the t-test of this model is: **0.0014**, therefore, we can reject the null hypothesis and determine that the two population means are different.

Interpreting Coefficients And Finding The Best Fitting Model *Appendix C* summarizes the example model **lm(mpg~am)**. This model shows that manual transmissions yield about 7.2 mpg more than automatic, however it only explains 34% of the variance in mpg. This suggests there may be other factors that explain mpg. To do this we perform a Stepwise Algorithm. *Appendix D* shows the output for the best model and reveals that with this model ~83% of the variance in mpg can be explained–much better than the example model. The p-values are all < 0.05 which means that wt, qsec, and am are all significant variables.

```r
fit <- lm(mpg~., data=mtcars)
model.step <- step(fit)
best.step.model <- summary(model.step)$call
```

We conclude that the best formula via the Stepwise process is:

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

Finding Interactions Below is a test to see if any interactions are significant to be added to our fitted model so far.

Based on the different combinations (*Appendix E*), it appears that the am:wt interaction with a p-value of: 0.0018. So our best fitted model is: **lm(formula = mpg ~ wt + qsec + am + am:wt, data = mtcars)** (*See Appendix F*)

```r
anova(example.model,best.fit.model)
```

ANOVA

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am + am:wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     27 117.28  3    603.62 46.323 8.847e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

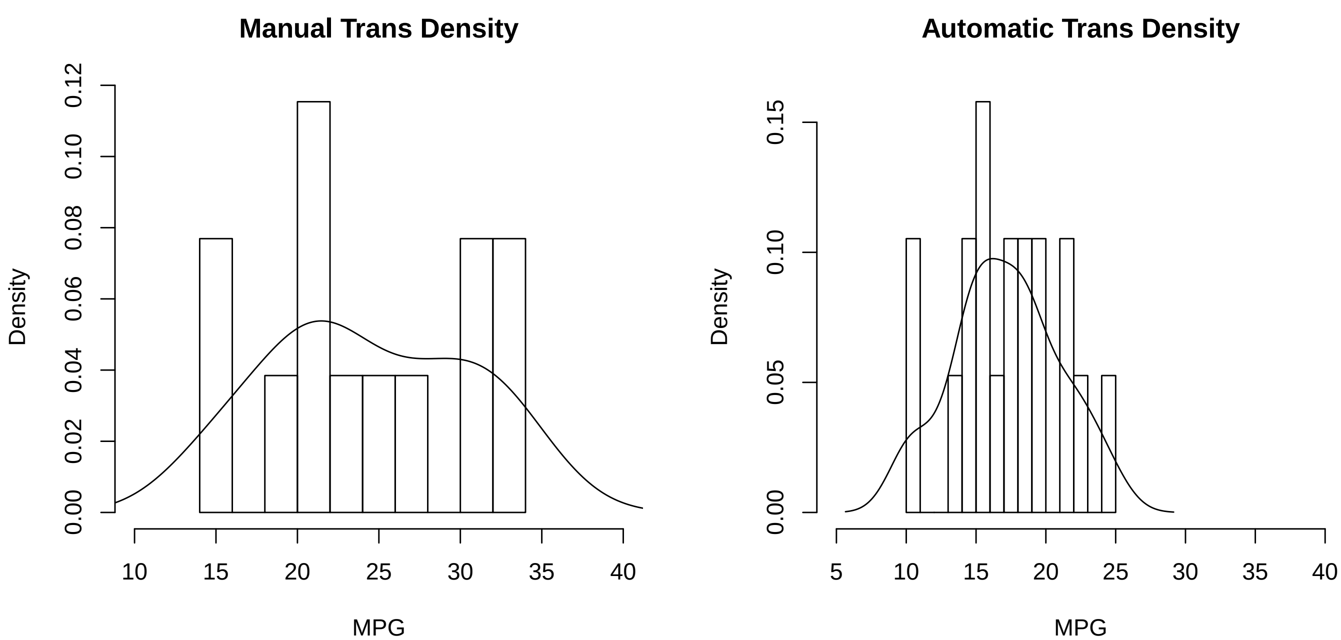
Based on our ANOVA analysis, we can can conclude that the best fitted model is significantly different from our example model, therefore we can reject the null hypothesis.

Conclusions Referring back to our best fitted model, it appears that manual transmissions are better for gas mileage than automatic transmissions. In *Appendix D*, we can see that the expected increase in mpg for manual transmissions is 4.36. By adding in other counfounding variables and interactions reduced the mpg estimate but provided greater precision over the example model **lm(mpg~am)**.

Appendix A: Structural Analysis of mtcars

```
## '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  : Factor w/ 2 levels "V","S": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "automatic","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 ...
```

Appendix B: Testing Normality



Appendix C: Summary Of The Base Model

```
##
## 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
```

Appendix D: Summary Of The Best Step 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
```

Appendix E: Interactions

```
##   interaction  inter.coef
## 1       am.wt 0.001808576
## 2     am.qsec 0.038407879
## 3     qsec.wt 0.282397339
```

Appendix F: Summary of The Best Fit Model

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + am:wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## ammanual      14.079      3.435   4.099 0.000341 ***
## wt:ammanual   -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```