Regression Models Course Project

Vince Glick

Saturday, May 23, 2015

Synopsis

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”

Data Processing

Load/Import Required Libraries & Data Files

```
echo = TRUE
options(scipen = 1)
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
data(mtcars)
```

As per the mtcars help file:

The variables ‘am=0’ for automatic transmissions and ‘am=1’ for manual transmissions are used to render distributions for each respectively based upon their mpg.

Exploratory Analysis

In order to deduce whether manual or auto is better for MPG, we must first infer as to whether the two distinctive transmission have normalized distributions across their mpg values.

```
mtauto<-subset(mtcars, mtcars$am==0)
mtman<-subset(mtcars, mtcars$am==1)

nmtauto<-rnorm(mtauto$mpg)
tmtauto<-rt(mtauto$mpg, df=Inf)
shapiro.test(nmtauto);shapiro.test(tmtauto)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  nmtauto
## W = 0.94822, p-value = 0.3684
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  tmtauto
## W = 0.94928, p-value = 0.3843
```

```
nmtman<-rnorm(mtman$mpg)
tmtman<-rt(mtman$mpg, df=Inf)
shapiro.test(nmtman);shapiro.test(tmtman)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  nmtman
## W = 0.97449, p-value = 0.9414
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  tmtman
## W = 0.95225, p-value = 0.6328
```

Utilizing the Shapiro-Wilk Normality Test:

Because the p-value of ‘mtauto’ and ‘mtman’ mpg distributions are greater than 0.05, you cannot reject the hypothesis that the sample comes from a population which has a normal distribution. As a result, we are at liberty to utilize the t-test to compare the means of mpg for both automatic and manual transmission data sets.

```
t.test(mtauto$mpg, mtman$mpg, alternative = "greater", paired=FALSE)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  mtauto$mpg and mtman$mpg
## t = -3.7671, df = 18.332, p-value = 0.9993
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -10.57662       Inf
## sample estimates:
## mean of x mean of y 
##  17.14737  24.39231
```

With a p-value of 0.9993, the confidence in this comparison cannot reject the hypothesis that this sample delivers a comparison which entitles manual transmission automobiles have a stronger MPG than those with automatic transmissions.

```
model1 <- lm(mpg ~ ., data = mtcars)
model2 <- step(model1, direction = "both")
summary(model2)
model3 <- lm(mpg ~ am, data=mtcars)
```

Greatest R-Squared Model

```
lrm1<-lm(data=mtcars, mpg ~ wt + qsec + am)
```

Attain the variables with the highest correlations

```
cor <- round(cor(mtcars)[1,], 2)
corsort <- names(sort(abs(cor),decreasing=T))
lrm2<-lm(data=mtcars, mpg~wt+cyl+disp+hp)
```

Linear regression model with the most correlated variables and greatest R squared values

```
lrm3<-lm(data=mtcars, mpg~wt+cyl+disp+hp+qsec+am)
```

We compare the original ‘model3’ with the ‘lrm1’, ‘lrm2’, and ‘lrm3’ regression models to determine any significant difference amongst the models for a best fit.

```
anova(model3, lrm1, lrm2, lrm3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ wt + cyl + disp + hp
## Model 4: mpg ~ wt + cyl + disp + hp + qsec + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.6658 4.498e-09 ***
## 3     27 170.44  1     -1.16
## 4     25 150.99  2     19.45  1.6105    0.2198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 4.498e-09, we have determined that wt, qsec, and am are the most influential variables in determining the quantification of mpg difference between automatic and manual transmissions. The makes ‘lrm1’ our best fitting model.

```
summary(lrm1)
```

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

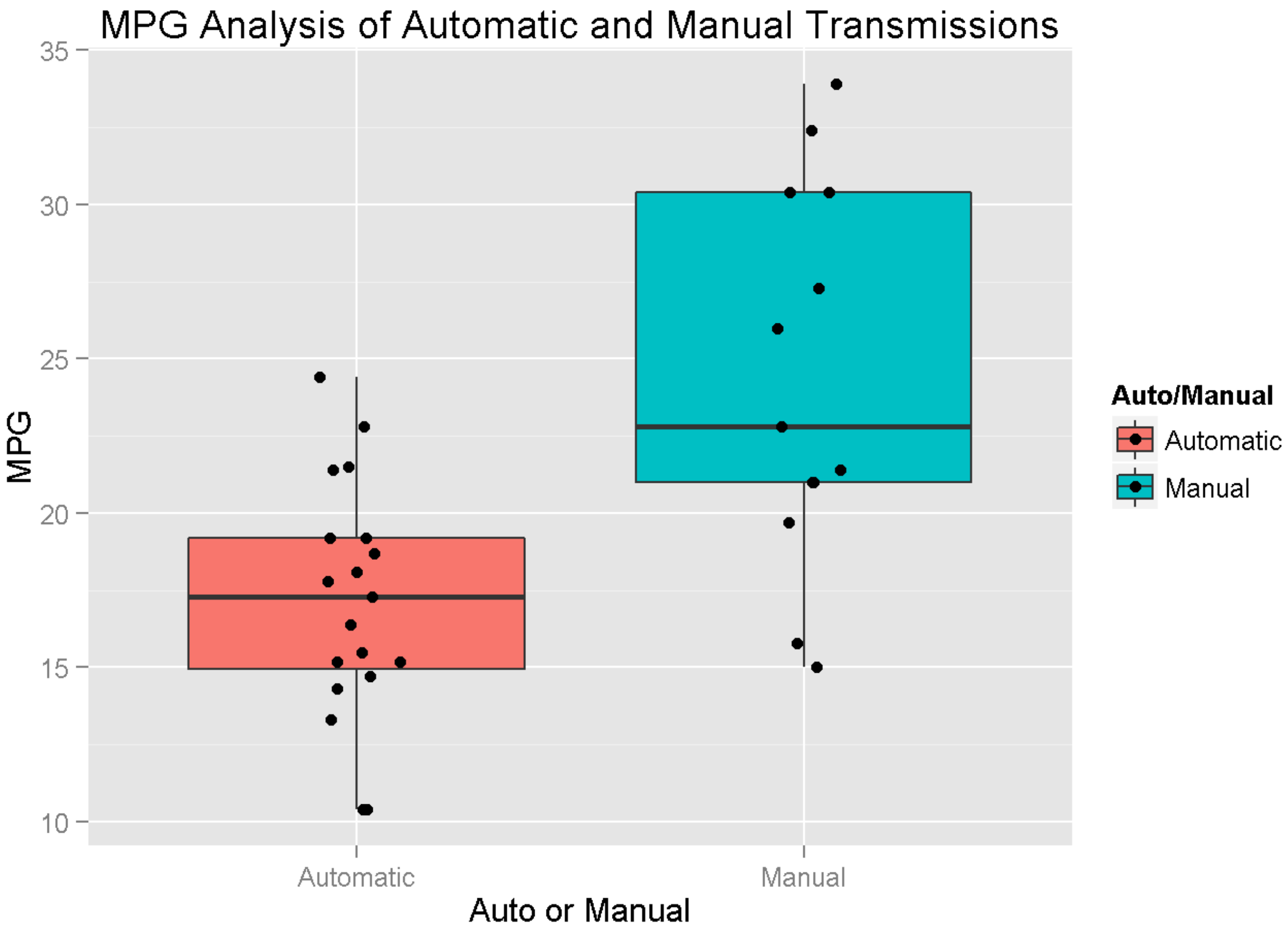
Conclusion

Based on the summary above:

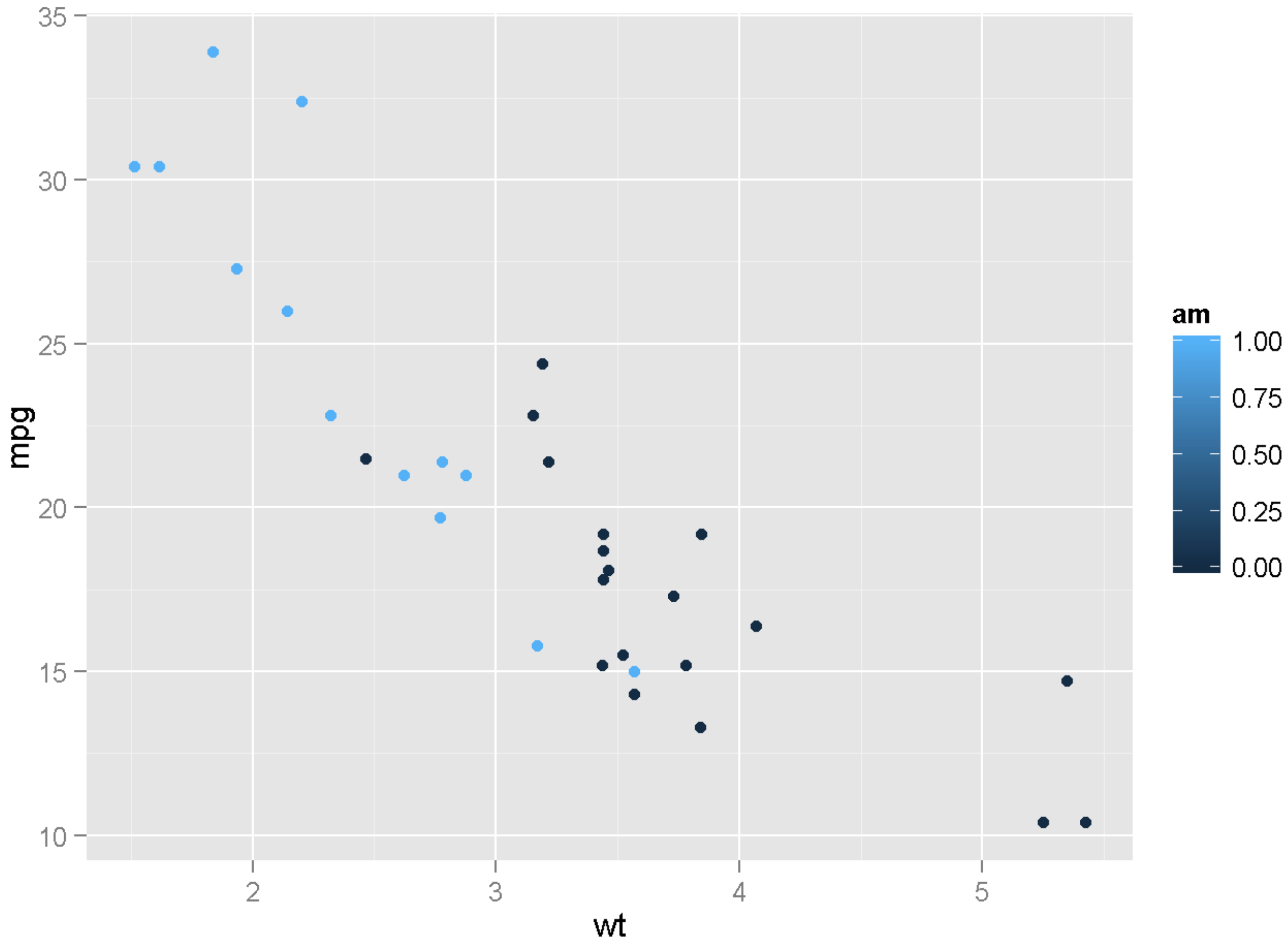
83.36% of the variance is explained by this model as manual transmissions gaining 2.94 MPG over automatic transmission vehicles for every 1,000 lbs

Appendix - Statistical Inference

The box plot below provides the direct correlation between automatic and manual transmissions for the original ‘model3’



The plot below provides the heavy correlation between weight and MPG for automatic and manual transmissions.


mpg
wt
am