Regression Models - Transmission and MPG

Ryan Wissman

Friday, March 20, 2015

Executive Summary

Fuel efficienty and transmission type are both very important and greatly debated factors when selecting a new car. This report examines the relationship between tranmission type and fuel economy to determine if there is any MPG benefit to purchasing a car of either type transmission. The data used in this report is from the 1974 Motor Trend US magazine.

Exploring the Data

First the mtcars data is loaded, variables set as factors (cylinders, v/s (v or straight engine), transmission, number of gears, number of carburetors), and some brief exploratory statistics are discovered.

```
data(mtcars); attach(mtcars)
head(mtcars,1) #Examine how the data is structured in mtcars

#set relevant variables to factors
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)

mean(mpg[am=="0"]) #Automatic Transmission mean
mean(mpg[am=="1"]) #Manual Transmission mean
```

Quickly examining the data to determine the means we find that the average fuel economy among automatic cars is 17.14 MPG whereas the average among manual transmissions is 24.39 MPG. Furthermore, according to a boxplot of the data (*see* **Figure 1**) we could guess that the fuel efficiency of a manual transmission is greater than that of an automatic transmission. The average and median is MPG for manual transmission is distinctly higher than that of automatic transmissions. However, we cannot yet make a conclusion on based on this chart alone. First we will need to determine if a relationship does exist by using regression.

Regression Models

First we try a linear model using mpg as the outcome and transmission type (variable am, "0" denotes automatic whereas "1" denotes manual) as the predictor.

Model: Transmission Type Only

```
model_am <- lm(mpg ~ am)
summary(model_am)$adj.r.squared
```

Using this model we can determine that this model can only explain about 34% (Adjusted R-squared value of 0.3598) of the variance in MPG. Transmission alone does not look like it accounts for enough of the variation to be significant by itself. Therefore we should try another model to examine the other variables in the mtcars dataset.

Model: All Variables

```
model_full <- lm(mpg ~ ., mtcars)
summary(model_full)$adj.r.squared
```

```
## [1] 0.7790215
```

Fitting all the variables can explain about 81% (Adjusted R-Squared 0.8066) of the variance in MPG. Accoding to the matrix scatterplot of all the variables in mtcars (*see* **Figure 2**) the there are a number of other variables that show significant correlation with MPG. Therefore we can probably create a better model using only the most significant variables.

Model: Best

Using R step function to step through the iterations of variables to determine which model would be the best.

```
model_best <- step(model_full, trace=0)
summary(model_best)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, 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 ***
## 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 **
## am1          1.80921    1.39630   1.296  0.20646
## ---
## 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
```

This model now accounts for about 83% of the variance which is better than our model that accounts for all the variables. The final model included three variables (weight, 1/4 mile time, and transmission).

```
anova(model_am, model_full, model_best)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 3: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS  Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     15 120.40  15    600.49 4.9874 0.001759 **
## 3     26 151.03 -11    -30.62 0.3468 0.958824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model_best)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## am1          1.80921138 1.39630450  1.295714 2.064597e-01
```

Residuals

According to the Normal Q-Q plot (*see* **Figure 3**) we see that the residuals appear to be normally distributed as the data points fit quite closely to the line.

Appendix

Data and figures that accompany the report.

Figure 1: Boxplot summarizing both automatic and manual transmission types relative to MPG

```
boxplot(mpg ~ am, xlab="Transmission Type", ylab="Miles per Gallon (MPG)", xaxt="n",
        main="Transmission and MPG",
        col=c("green", "red"))
axis(1, at=1:2, labels=c("Automatic", "Manual"))
```

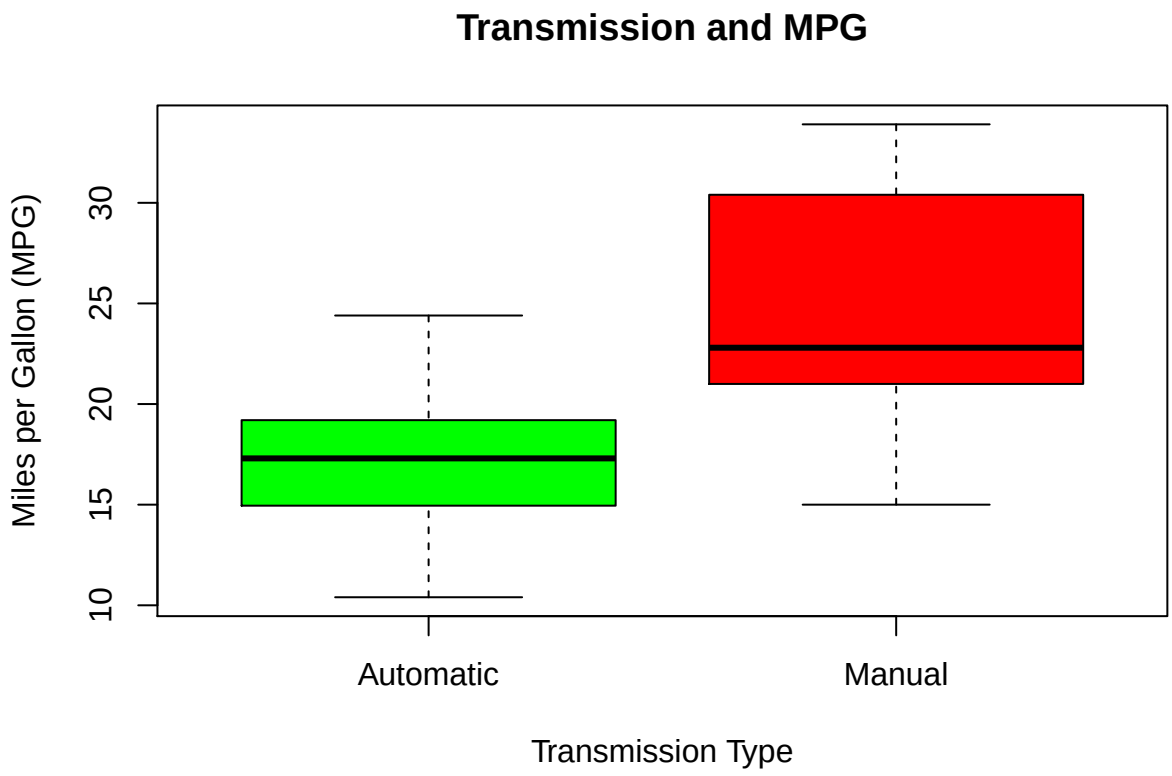


Figure 2: Matrix Scatterplot of all Variables

Matrix scatterplot covering the variances between many different variables in the mtcars dataset.

```
pairs(mtcars, main="Matrix Scatterplot of all Variables", panel=function(x,y){
  points(x,y)
  abline(lm(y~x), col="red")
})
```

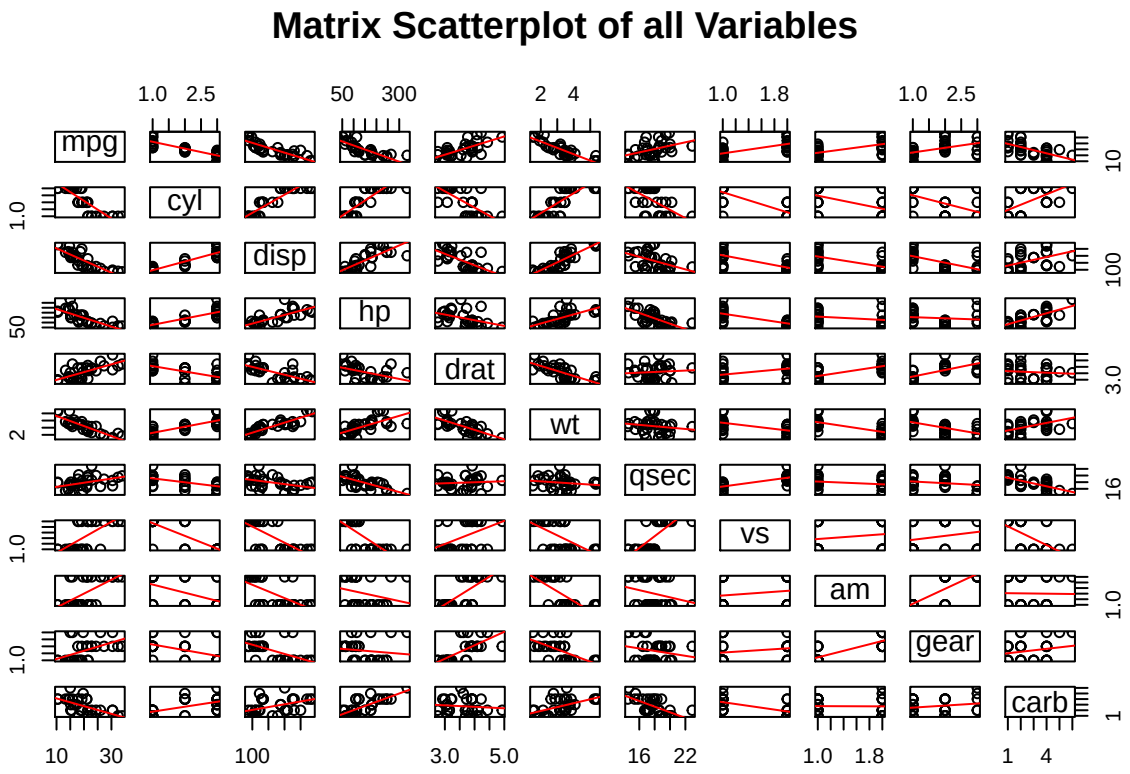
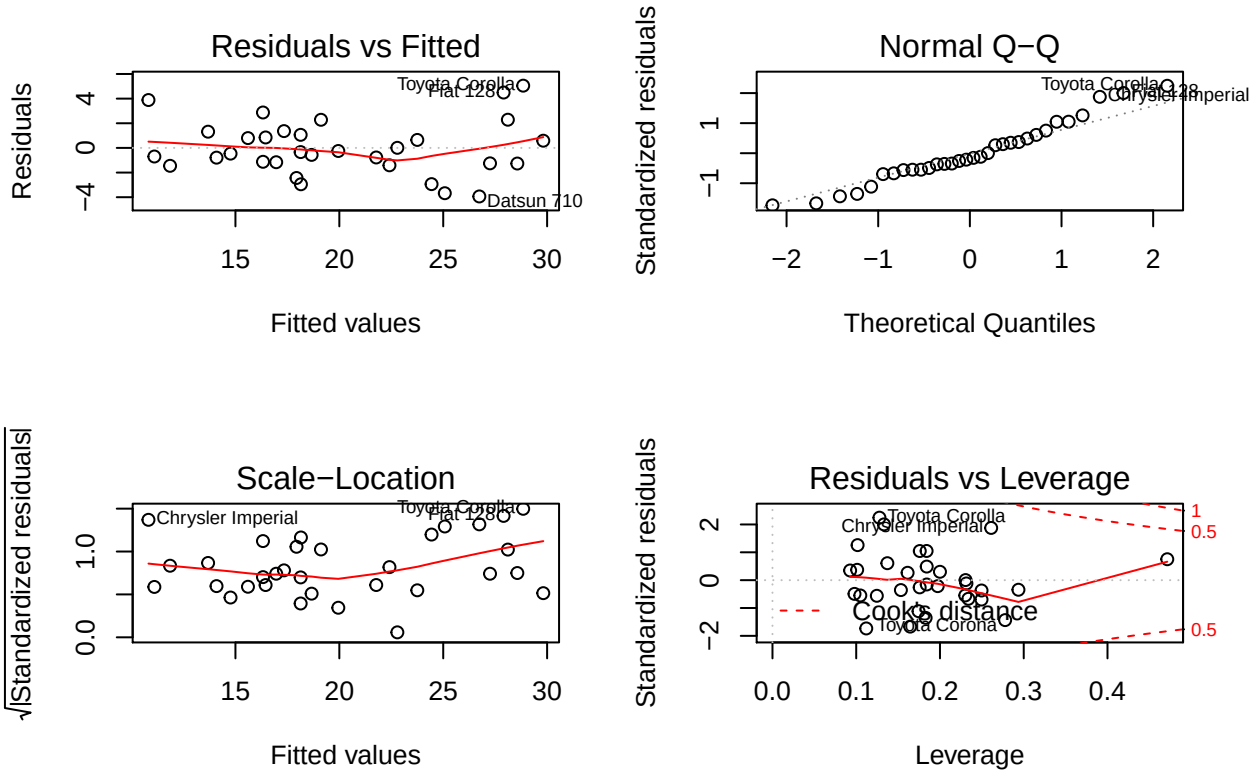


Figure 3: Residual Plots

Residuals plots for the best fit model.

```
par(mfrow=c(2,2))
plot(model_best)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Corolla
Chrysler Imperial
Toyota Corona