Regression Model Course Project

Nazmi Anik

Saturday, April 02, 2016

Executive Summary

This report explores the relationship between a set of variables and miles per gallon (MPG) for Motor Trend magazine. Questions that we are striving to answer are if an automatic or manual transmission better for MPG; what the quantified MPG difference between automatic and manual transmissions are.

As a resultof analysis, we have found out that fuel efficiency(mpg) is indeed related to type of transmission in cars. Cars with manual transmission are more fuel efficient than cars with automatic transmission. Mean *mpg* value for automatic transmission turns out to be 17.1, whereas it is 24.4 for manual transmission cars.

Exploratory Data Analysis

```
data(mtcars) #load mtcars data
str(mtcars) #checkout the variables in mtcars
```

See appendix for the structure display of mtcars data frame.
We can see that variables *cyl*, *vs*, *gear*, *carb* and *am* should be converted to factor variables. See appendix for conversion.

Let's check the correlation matrix for to see what kind of relationship exists between the variables and *mpg*. We will square the result for easy comparison. We are using original *mtcars* data frame with numeric values to be able to use *cor* function.

```
corCars <- cor(mtcars) #form correlation matrix
corCars[,1]^2 #display first row, squared
```

See appendix for correlations. In our preliminary analysis, it looks like *mpg* is highly correlated with *cyl*, *disp*, *hp* and *wt* variables. Since we are asked to analyze the relationship between transmission and *mpg*, we will check that relationship with a boxplot.

```
boxplot(mpg ~ am, data = mtcars2, col = (c("green","blue")), main="MPG vs Transmission",
ylab = "Miles Per Gallon", xlab = "Transmission Type")
```

See appendix for the boxplot.
Our initial analysis showed us that manual transmission may be more fuel efficient than automatic. Now we need to check the validity of our data and see if there are any dependencies between variables that might be skewing our results.

Model Fitting

Let's try to find the best model to fit. First we will fit all the variables. Then let's use *step* function to find the best fit.
Source for *step* function usage is here (http://www.stat.columbia.edu/~martin/W2024/R10.pdf)

```
fit0 <- lm(mpg ~ am, mtcars2) #our starting model
fit1 <- lm(mpg ~ ., mtcars2) #our full model
fit2 <- step(fit0, scope=list(upper=fit1),direction="both",trace=0) #create the best fit
summary(fit2)
```

See appendix for summary. Looking at the summary, we see that the best model includes *am*, *wt*, *qsec* as the regressors. R^2 value shows us that our model explains 84% of the variability. Looking at the p-values, we can see that all of these regressors are significant with an alpha = 0.05.

Diagnostics

Now let's examine our residuals to see if there are any issues with our model.

```
par(mfrow=c(2, 2))
plot(fit2)
```

See appendix for the residual plots. Properties that we find out based on the plots:
Independence: Residuals vs Fitted values plot shows no specific pattern.
Normality: Normal Q-Q plot shows us that residuals are normally distributed as they are close to the line.
Constant Covariance: Scale-Location plot affirms the constant variance, since the points are randomly distributed.
No Outliers: Residuals vs. Leverage shows all values within the 0.5 band, therefore no influential outliers are present.

Let's check some of the outliers to confirm that the outliers have no significant leverage:

```
summary(round(hatvalues(fit2),3))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.053   0.085   0.101   0.125   0.155   0.297
```

The hat values are not too far from one another, therefore, we can confirm that there are no influential outliers.

Results

Based on our analysis, we have found out that fuel efficiency(mpg) is indeed related to type of transmission in cars.
Cars with manual transmission are more fuel efficient than cars with automatic transmission.
Mean *mpg* value for automatic transmission turns out to be 17.1, whereas it is 24.4 for manual transmission cars.

Appendix

Structure of mtcars Data Frame

```
str(mtcars) #checkout the variables in mtcars
```

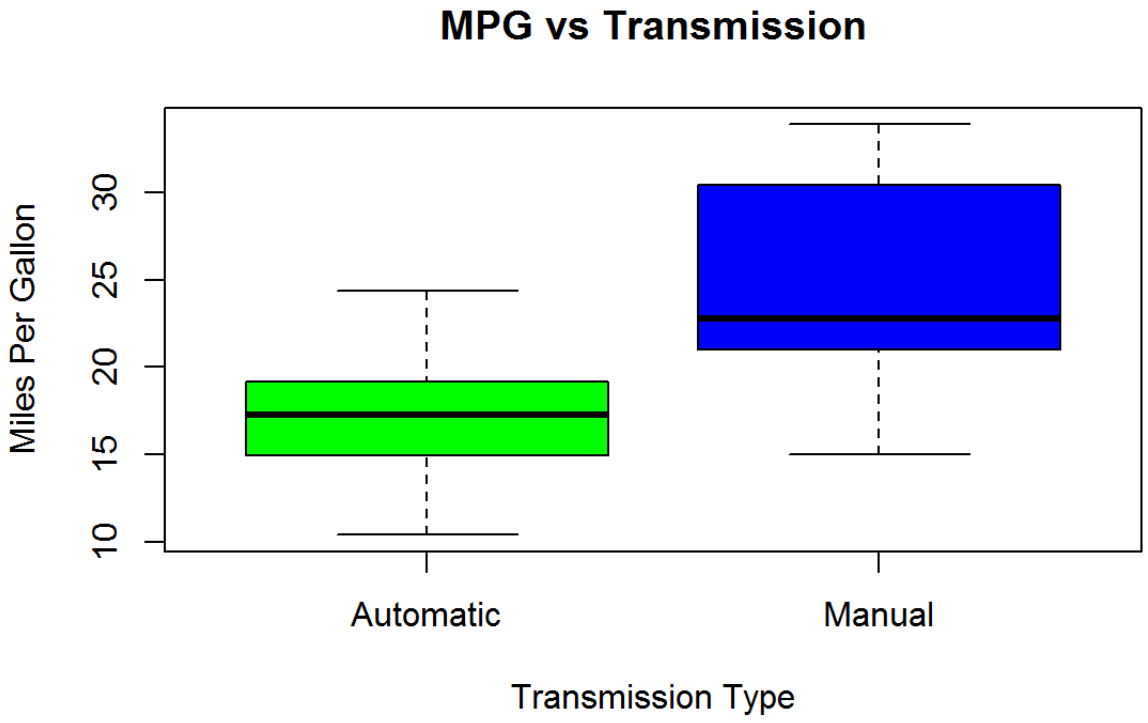
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 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  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Factor Conversion

```
mtcars2 <- mtcars
mtcars2$cyl <- factor(mtcars$cyl)
mtcars2$vs <- factor(mtcars$vs)
mtcars2$gear <- factor(mtcars$gear)
mtcars2$carb <- factor(mtcars$carb)
mtcars2$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

Boxplot Showing Miles per Gallon vs Transmission Type

```
boxplot(mpg ~ am, data = mtcars2, col = (c("green","blue")), main="MPG vs Transmission",
ylab = "Miles Per Gallon", xlab = "Transmission Type")
```



Correlation

```
corCars <- cor(mtcars) #form correlation matrix
corCars[,1]^2 #display first row, squared
```

```
##       mpg       cyl      disp        hp      drat        wt      qsec
## 1.0000000 0.7261800 0.7183433 0.6024373 0.4639952 0.7528328 0.1752963
##        vs        am      gear      carb
## 0.4409477 0.3597989 0.2306734 0.3035184
```

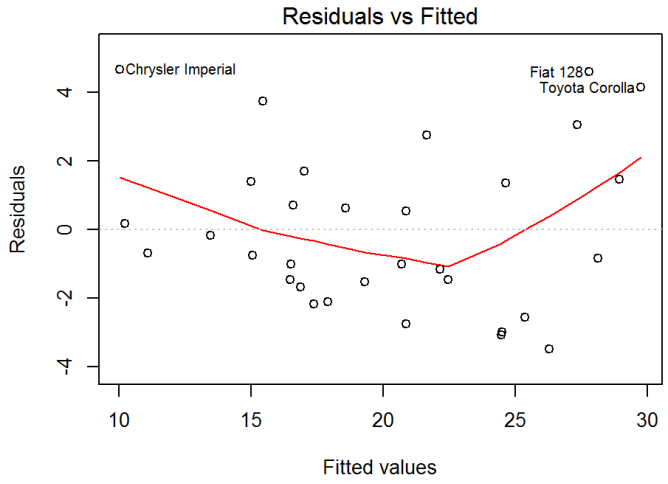
Summary for Model Fit

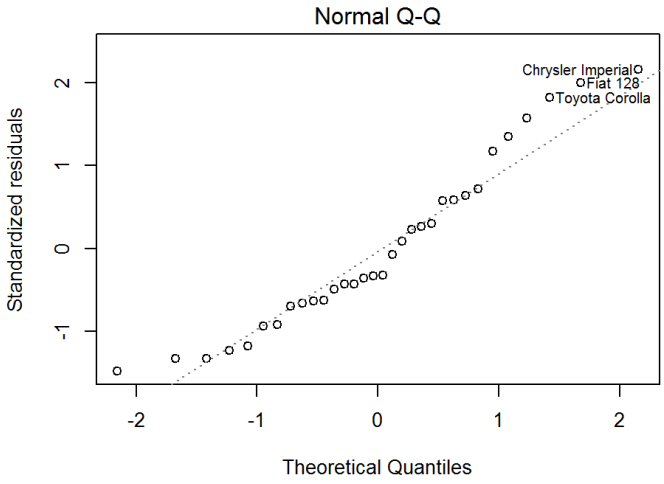
```
summary(fit2)
```

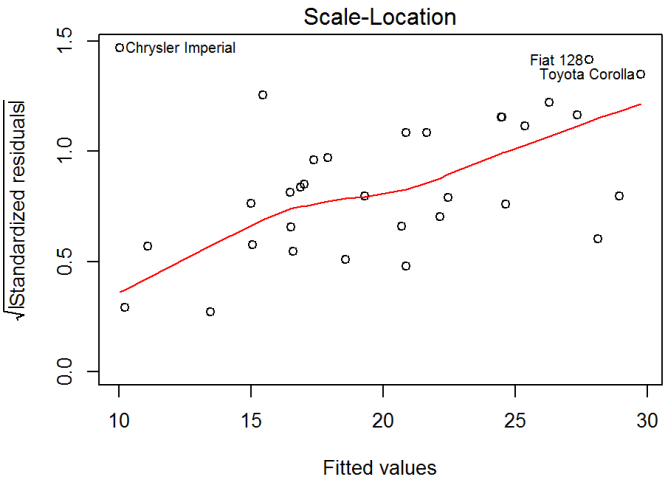
```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## amManual      2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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
```

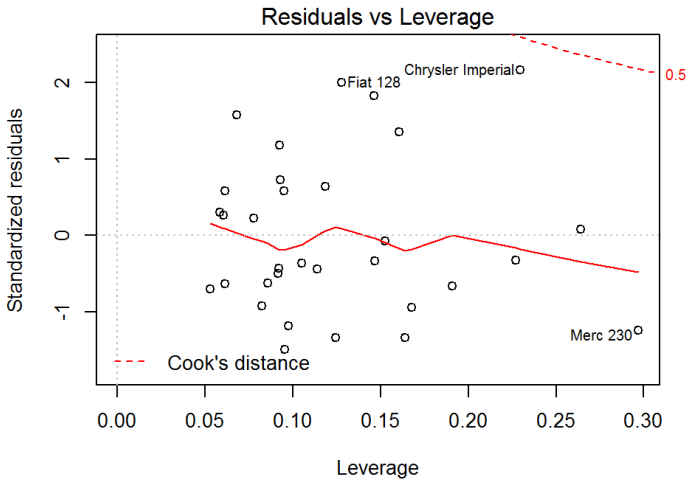
Residual Plots for Model Fit

```
par(mfrow=c(2, 2))
plot(fit2)
```


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
Fitted values


Normal Q-Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
Theoretical Quantiles


Scale-Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Chrysler Imperial
Fiat 128
Merc 230
Cook's distance
0.5
Standardized residuals
Leverage