Regression Model Project: Manual Transmission vs Automatic Transmission

Mrugank Akarte

26 August 2016

Executive Summary

For the following analysis *mtcars* dataset was used to answer the following questions.
1.Is an automatic or manual transmission better for MPG?
2.Quantify the MPG difference between automatic and manual transmission
Initially scatterplots were plot to determine directionality and correlation between variables.Then multiple regression model was used to determine the relationship between the MPG and Transmission variable.In order to get the best model *Backward Elimination using p-values* method was used.This model has highest adjusted R square value of *0.8336* compared to other models. According to this model keeping the weight and 1/4 mile time variable constant, manual transmission gives *2.93 MPG more* than the automatic transmission on average.Also, keeping the transmission and 1/4 mile time variable constant, with increase in 1000lbs of weight the MPG *reduces by 3.91* on average.

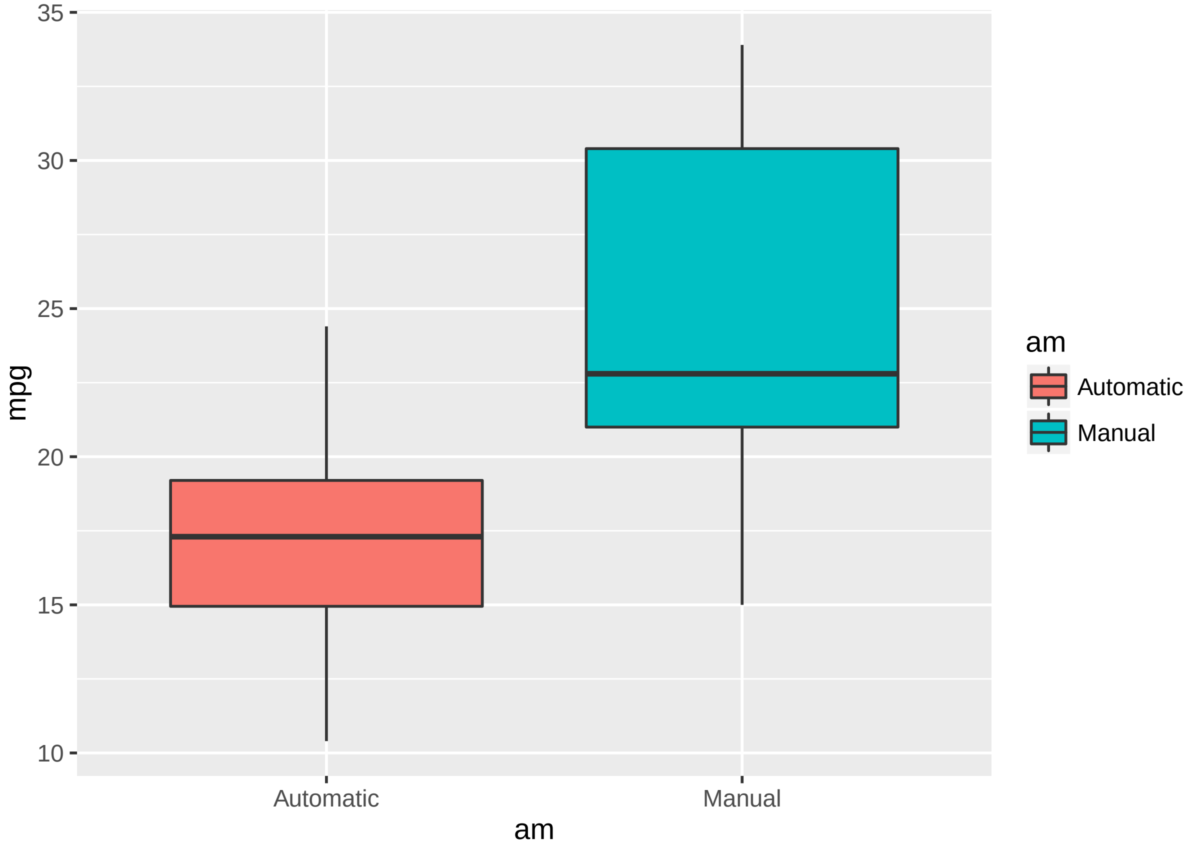
Setting up data

Transforming the data to suitable format.Converting the numerical variables into categorical variables.

Brief summary of data

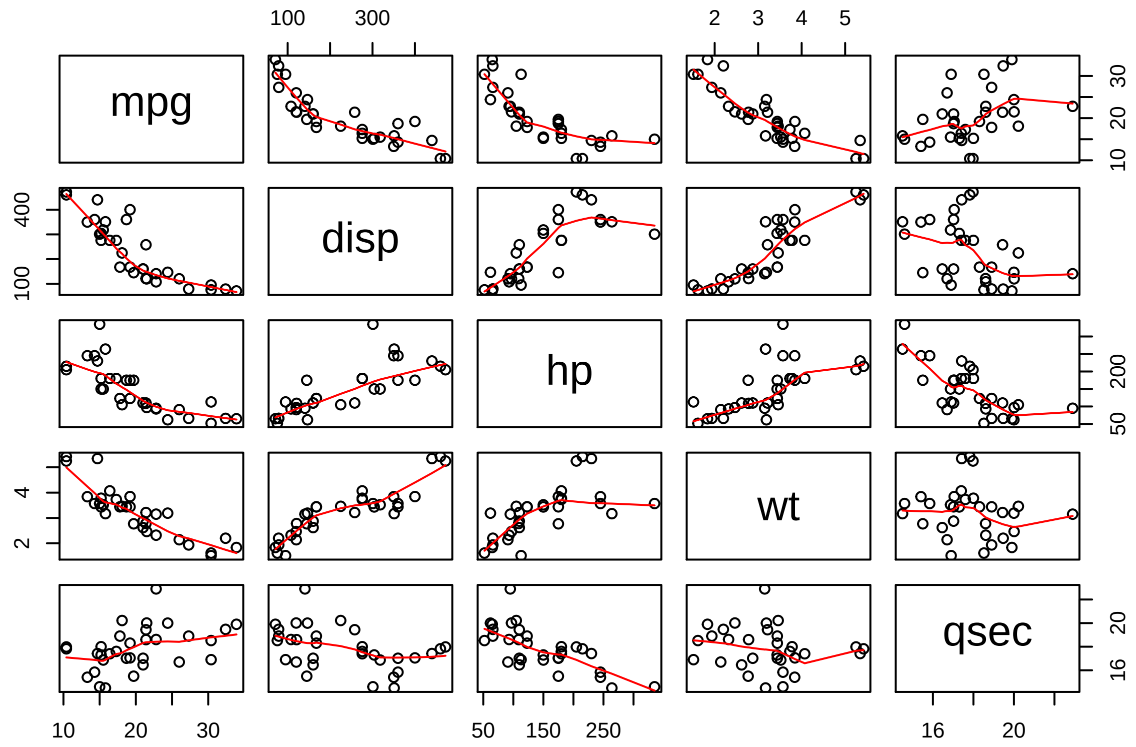
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  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 "0","1": 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: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Basic Analysis



Above box-plot shows significant difference in average MPG factor between automatic and manual transmission.

Scatterplots for directionality and correlation.



```
##              mpg       disp         hp         wt       qsec
## mpg   1.0000000 -0.8475514 -0.7761684 -0.8676594  0.4186840
## disp -0.8475514  1.0000000  0.7909486  0.8879799 -0.4336979
## hp   -0.7761684  0.7909486  1.0000000  0.6587479 -0.7082234
## wt   -0.8676594  0.8879799  0.6587479  1.0000000 -0.1747159
## qsec  0.4186840 -0.4336979 -0.7082234 -0.1747159  1.0000000
```

From above pairs graph and correlation chart we can conclude that

- Adding all the variables to regression model with result in incorrect observations since there are variables which are highly correlated to each other.Eg disp-wt, disp-hp, etc.
- Due to presence of multicollinearity, in the final model only one such variable should be included.

Simple Regression

```
sfit<-lm(mpg ~ am, data = mtcars)
summary(sfit)
```

```
##
## 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
```

Adjusted R square for above regression model is only **0.3385**, meaning only 33.85% of variability in MPG is explained by this model. Thus it is necessary to use multiple regression to find out the best model.

Multiple Regression

In order to decide best model **Backward Elimination using p-values** method was used.

Initially adding all the variables in the model.

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

In the above model there are many variables with p-value > 0.05 i.e they are statistically insignificant, thus removing the variables one by one. Eg: In first case **drat** has highest p-value of 0.98, thus removing drat variable and fitting the model again. This process is continued till only the statistically significant variables are left in the model. This leads to the following model.

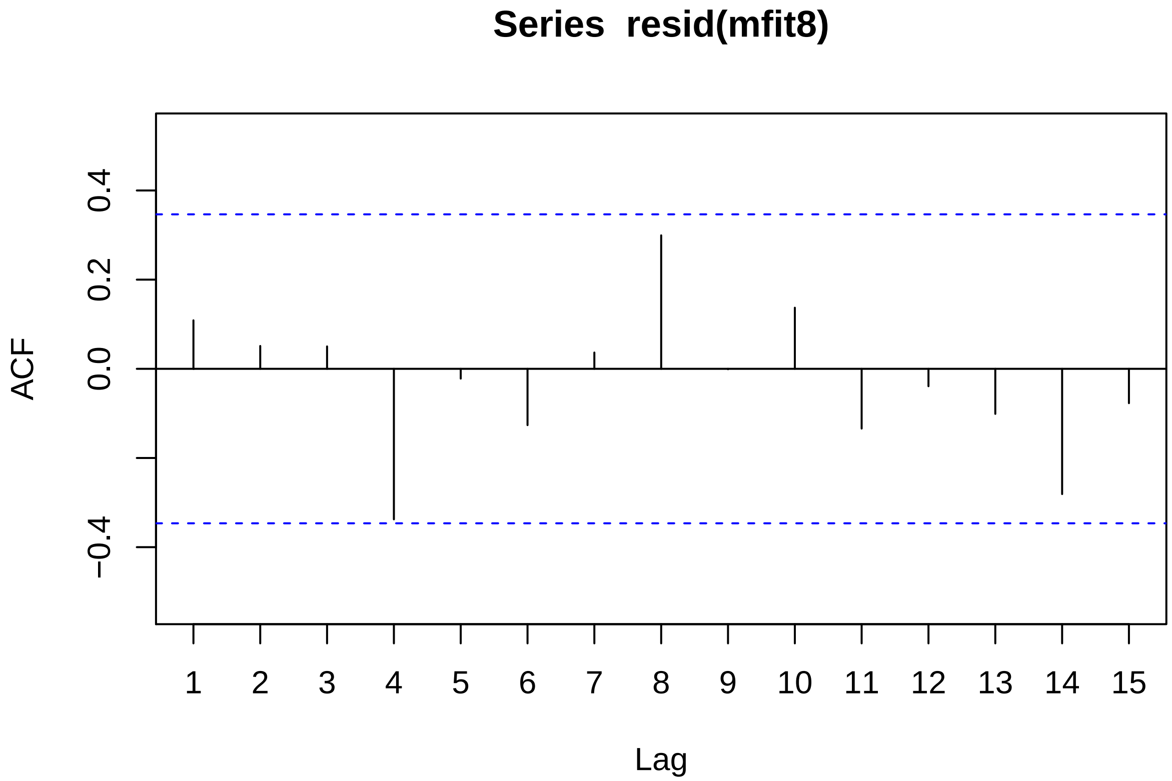
```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## amManual      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
```

The adjusted R square for the model is **0.8336** meaning 83.36% of variability in MPG is explained by this model.

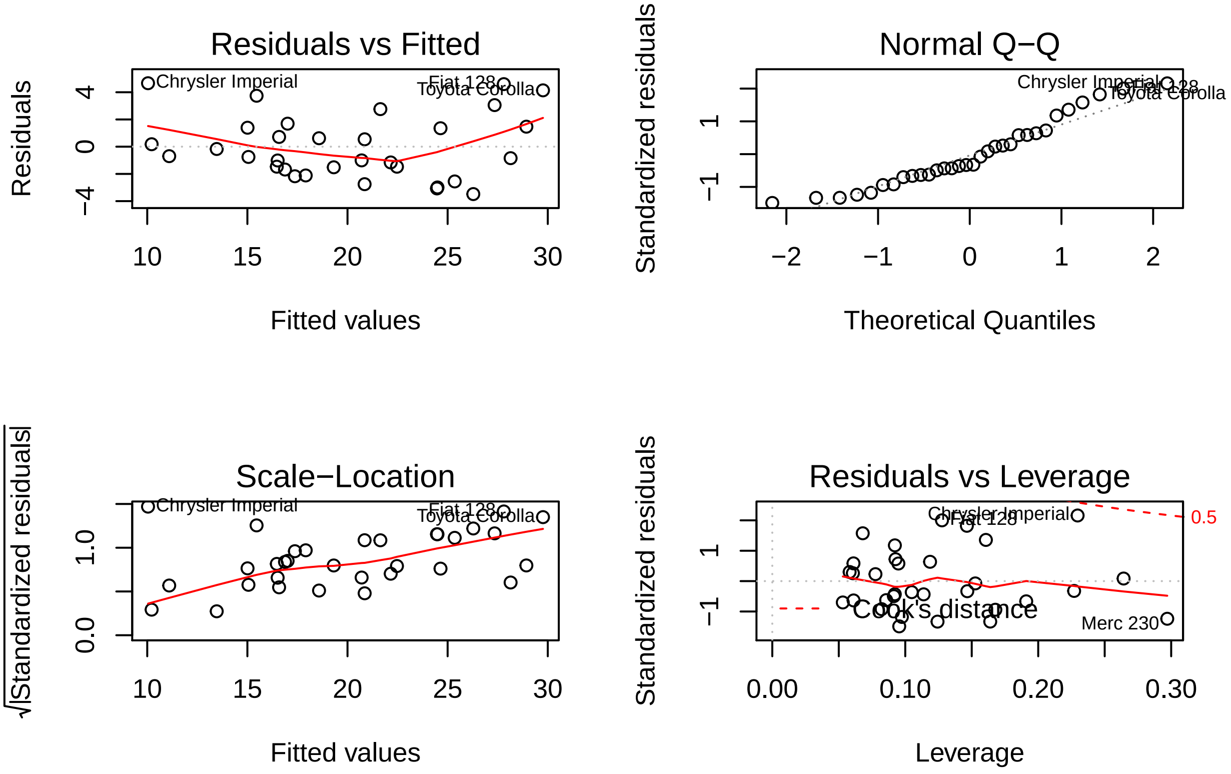
The estimates of the above model gives following observations.

- Keeping weight and 1/4 mile time constant, manual transmission gives **2.93 MPG more** compared to automatic transmission, on average.
- Keeping the transmission and 1/4 mile time constant, with increase in 1000lbs of weight the MPG is **reduced by 3.91**, on average.

Residual Diagnostics



- The Acf graphs shows that the residuals are not correlated to each other.
- Multicollinearity was eliminated by selecting suitable variables.



- Above Residuals vs Fitted plots shows that residuals are randomly distributed, hence are not correlated with each other.
- The residuals lie roughly on line in Q-Q plot indicating normal distribution.
- The Scale-Location plot has random distribution of residuals indicating constant variance.
- Residuals vs Leverage plot show no presence of outliers.

Conclusion

Based on above analysis we can conclude that manual transmission is better than automatic transmission for MPG.