identifying effective variables on miles per gallon for mtcars dataset

Reza Nirumand

Wednesday, September 28, 2015

Executive Summary

In this document we have analysed a car dataset consisting of 32 observations and 11 variables. Considering the importance of transmission type we have tried to answer below two questions using exploratory data analysis and linear regression: ## *Is an automatic or manual transmission better for MPG?##*

Quantifing the MPG difference between automatic and manual transmissions

Loading requried tools & preparing data

Original *mtcars* dataset consists of 32 observations with 11 variables which all stored as number Considering the documentations i have decided to do:
-to remove qsec from the dataset as it is not relevant to fuel consumption rather it shows overall car performance.
-to convert the variables "am"={0,1},"vs"={0,1} to factor variables

```
library(datasets);library(ggplot2);library(dplyr);library(tidyr);library(GGally)
dt<-select(mtcars,-qsec) ##removing irrelevant column
dt<-mutate(dt,am=factor(am),vs=factor(vs)) ##adjusting column type
summary(dt) ##lets see the data
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt          vs     am         gear
##  Min.   :2.760   Min.   :1.513   0:18   0:19   Min.   :3.000
##  1st Qu.:3.080   1st Qu.:2.581   1:14   1:13   1st Qu.:3.000
##  Median :3.695   Median :3.325                 Median :4.000
##  Mean   :3.597   Mean   :3.217                 Mean   :3.688
##  3rd Qu.:3.920   3rd Qu.:3.610                 3rd Qu.:4.000
##  Max.   :4.930   Max.   :5.424                 Max.   :5.000
##       carb
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.812
##  3rd Qu.:4.000
##  Max.   :8.000
```

A brief description of features(variables):
mpg: Miles/(US) gallon ; ***cyl:*** Number of cylinders ; ***disp:*** Displacement (cubic inch) ; ***hp:*** Gross horsepower ;***drat:*** Rear axle ratio ; ***wt:*** Weight (lb/1000) ; ***qsec:*** 1/4 mile time ; ***vs:*** V/S ; ***am:*** Transmission (0 = automatic, 1 = manual) ; ***gear:*** Number of forward gears ; ***carb:*** Number of carburetors.

Explatory data analysis

Figure.1 shows histogram of miles per gallon for the dataset. The variable mpg has the average 20.090625 and standard deviation 6.0269481 .

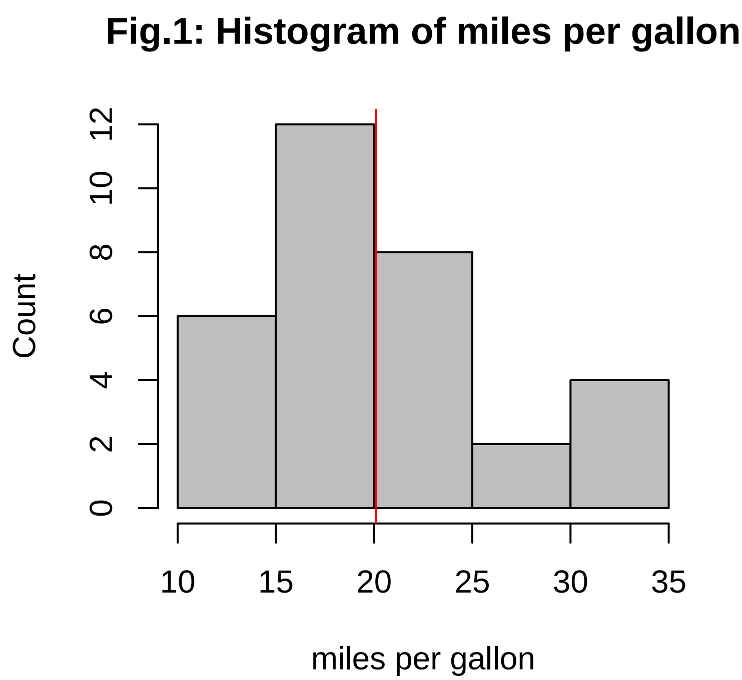
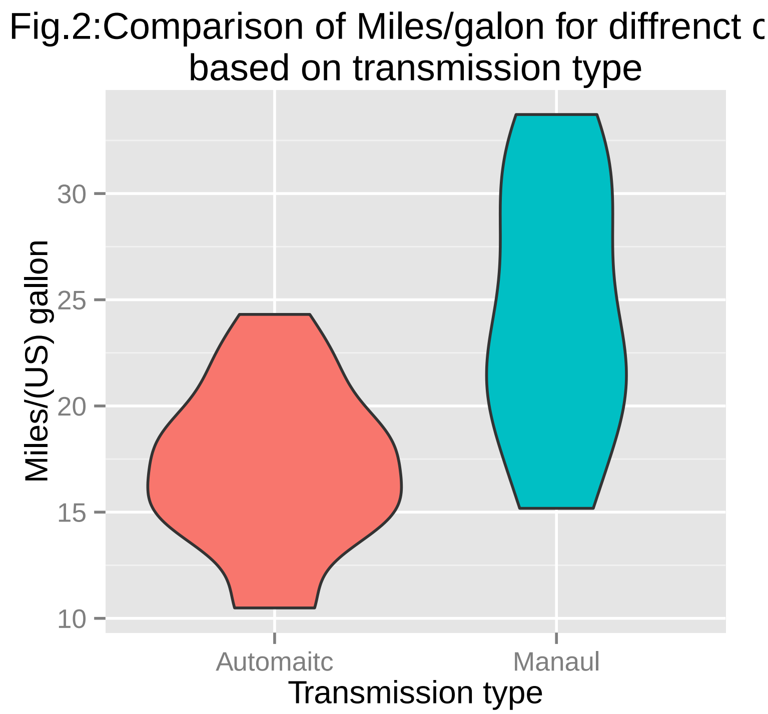


Figure.2 shows relationship between transmission type and miles per gallon.



Model Selection

In order to find a parsimonious model, we will use *nested model* technique. That means we will begin with one regressor and will add regressors one-by-one, comparing the result for each model using anova test. But Considering the correlation matrix ??? finding the best subset of regressors requires exhaustive search for the best subsets of the variables. This can be done using different r-packages such as ***"leaps"***.
But for the purpose of this project i have decided to follow the strategy of including variables which could describe other variables as well. For example hp could describe displacement,cyl,vs,carb and these described variables will not be added to the models.

```
t1<-lm(data=dt,mpg~am)
t2<-lm(data=dt,mpg~am+hp)
t3<-lm(data=dt,mpg~am+hp+wt)
t4<-lm(data=dt,mpg~am+hp+wt+gear)
t5<-lm(data=dt,mpg~am+hp+wt+drat)
anova(t1,t2,t3,t4,t5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp
## Model 3: mpg ~ am + hp + wt
## Model 4: mpg ~ am + hp + wt + gear
## Model 5: mpg ~ am + hp + wt + drat
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 71.5809 4.503e-09 ***
```

```
## 3     28 180.29  1     65.15  9.8082  0.004149 **
## 4     27 179.34  1      0.95  0.1431  0.708145
## 5     27 176.96  0      2.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Considering the anova test we will choose the model 3 since it shows significant change in RSS. Now we need to validate our model for the underling assumptions. The assumption for anova test is that the model's Residual are approximately Normal. To validate the assumption so we will use the ***Shapiro-Wilk test***. The null hypothesis on this test is that the distribution is approximately normal. Considering the following result we fail to reject the null hypothesis, hence our anova test is valid.

```
st<-shapiro.test(t3$residuals)
st
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t3$residuals
## W = 0.9453, p-value = 0.1059
```

Results

- R2=.6 means 60% of variation of outcome is explained by linear relationship with regressors.

Appendix A: Figures

Apendix B: Envirnoment Setup

Windows 10 X64 ; - R version 3.2.2 (2015-08-14) ; Rstudio Version 0.98.1103