Regression

Tarun

4/21/2021

Overview

Our goal here to explore the relationship between the variables of the “mtcars” dataset. Our research will focus on 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.” I am going to use the following R libraries to assist in my analysis:

```
require(datasets)
require(ggplot2)
```

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

```
require(GGally)
```

```
## Loading required package: GGally
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
require(ggthemes)
```

```
## Loading required package: ggthemes
```

```
require(plotly)
```

```
## Loading required package: plotly
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
require(ggcorrplot)
```

```
## Loading required package: ggcorrplot
```

```
## Warning: package 'ggcorrplot' was built under R version 4.0.5
```

```r
require(Amelia)
```

```
## Loading required package: Amelia
```

```
## Warning: package 'Amelia' was built under R version 4.0.5
```

```
## Loading required package: Rcpp
```

```
## ## 
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.6, built: 2019-11-24)
## ## Copyright (C) 2005-2021 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

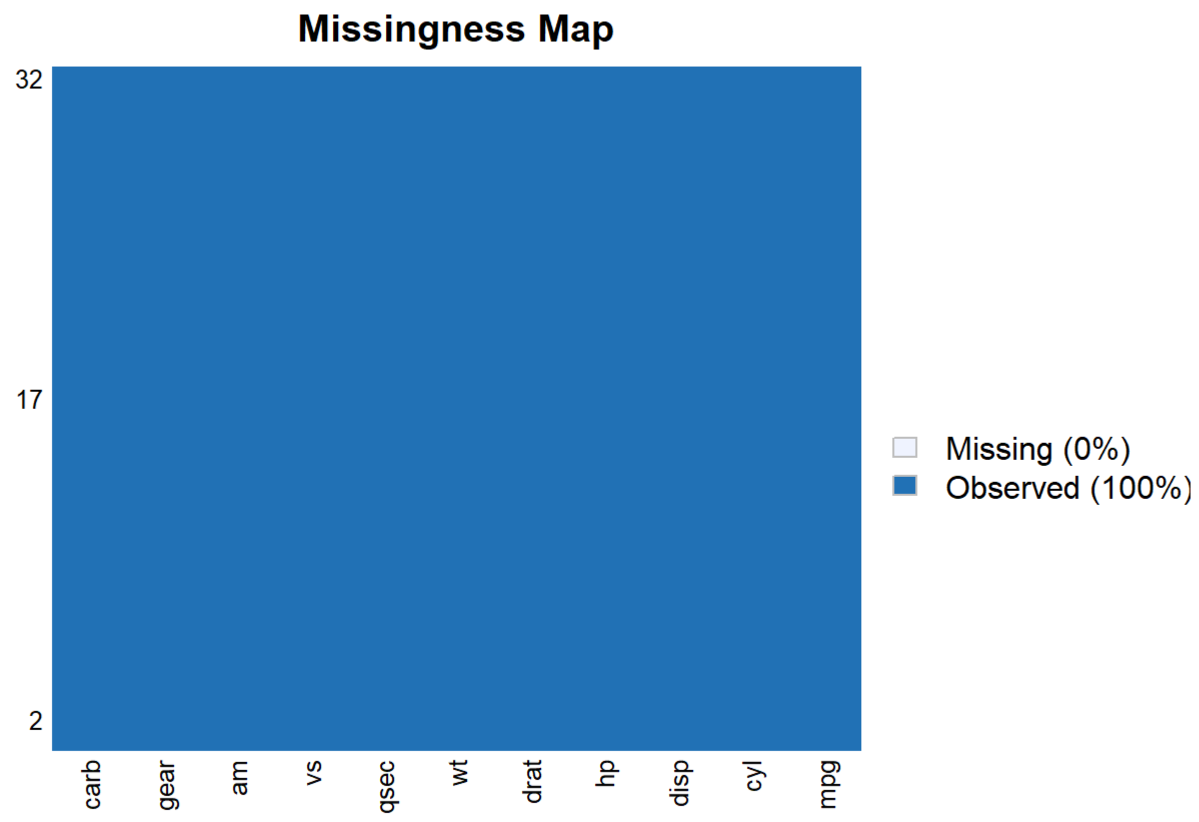
```r
require(leaps)
```

```
## Loading required package: leaps
```

```r
cars <- mtcars
head(cars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
missmap(cars) # to check the missing values
```



There is no missing values in the dataset but, we need to change some of the varible types.

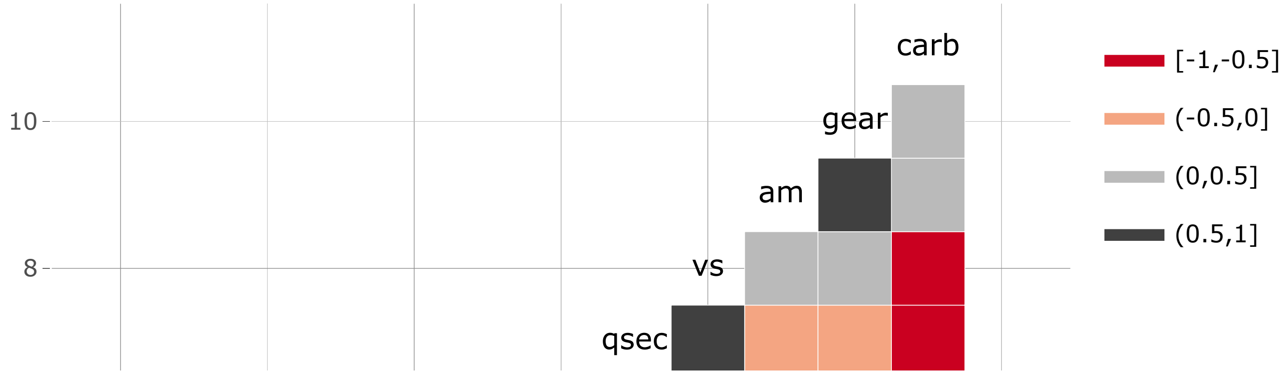
Data Processing

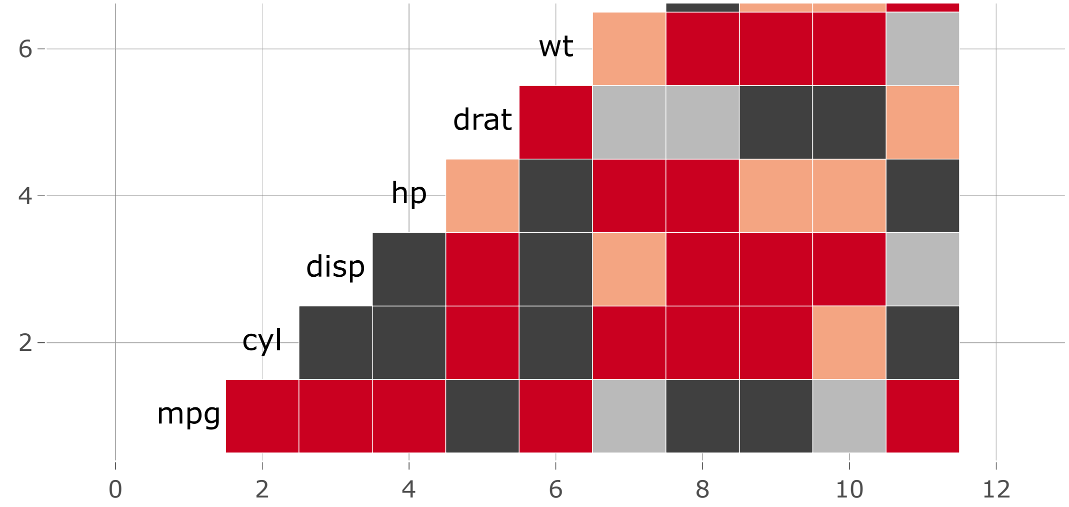
```
cars$cyl <- as.factor(cars$cyl)
cars$vs <- as.factor(cars$vs)
cars$am <- as.factor(cars$am)
str(cars)
```

```
## '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 "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

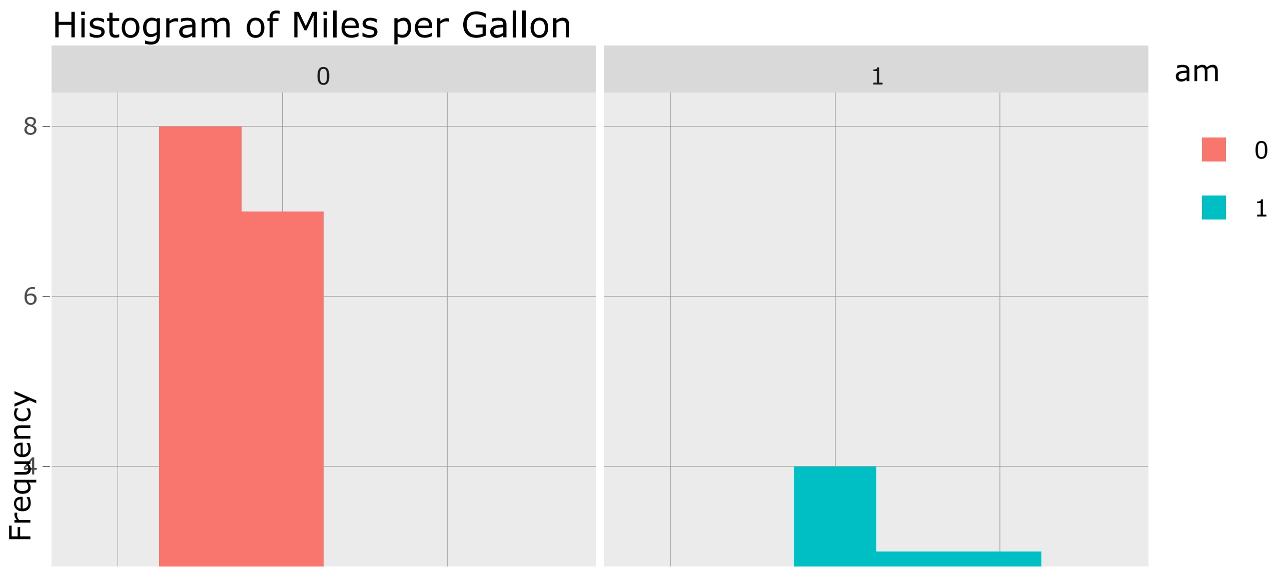
Exploratory Analysis

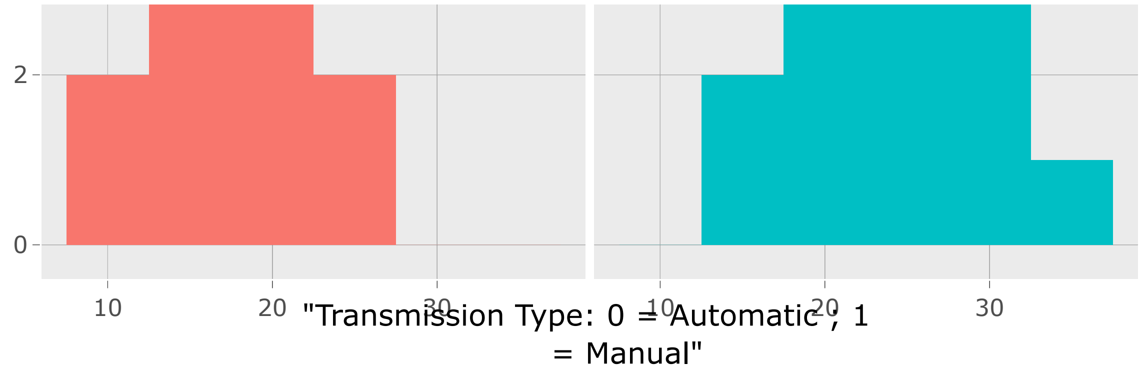
```
 # To check the correlation b/w the variables
plot1 <- ggplotly(ggcorr(mtcars, nbreaks = 4, palette = "RdGy"))
plot1
```



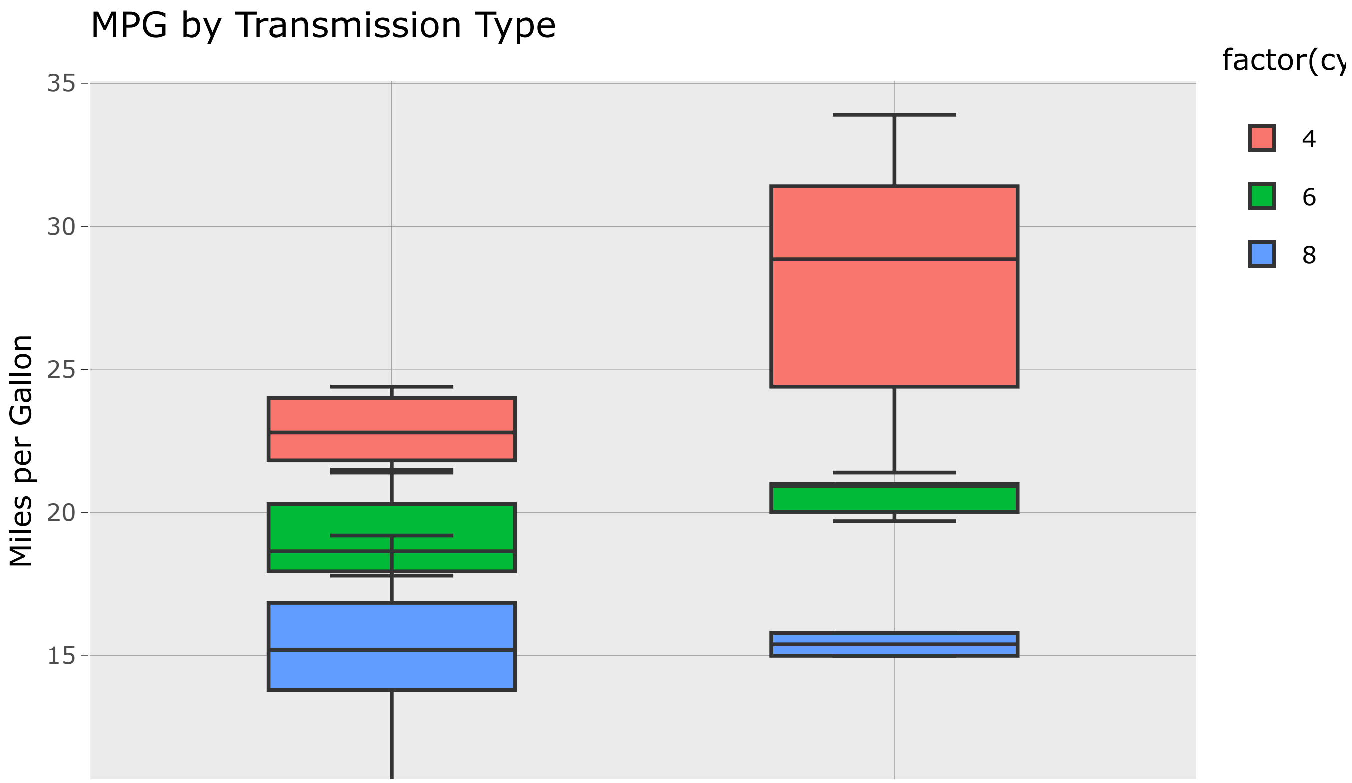


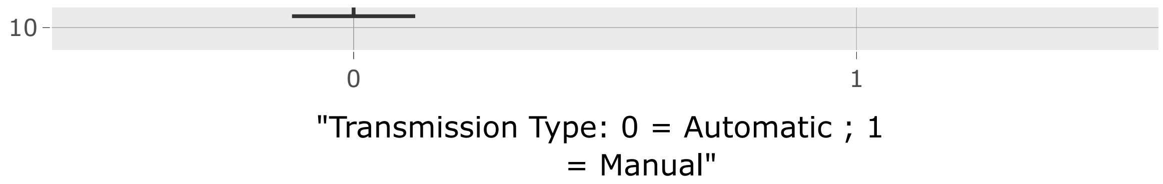
```
plot2 <- ggplot(cars, aes(mpg)) +
  geom_histogram(aes(fill=am),binwidth = 5) +facet_grid(.~am)+
  labs(title = "Histogram of Miles per Gallon",
       x='"Transmission Type: 0 = Automatic ; 1
       = Manual"', y = "Frequency")

ggplotly(plot2)
```





```
plot3 <- ggplot(data = cars, aes(x = am, y = mpg)) +
  geom_boxplot(aes(fill=factor(cyl))) +
  labs(x='"Transmission Type: 0 = Automatic ; 1
       = Manual"',y="Miles per Gallon",
       title='MPG by Transmission Type')
ggplotly(plot3)
```





```
aggregate(mpg ~ am, data = mtcars, mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

Correlation Analysis - As we can see in the correlation graph, that “Miles per Gallon” have - correlation with cyl, hp, and wt. That means car mileage will decrease as car weight and horsepower increase.

Histogram Analysis - Manual transmission cars have higher MPG than automatic transmission.

Boxplot Analysis - Manual transmission is 7.25 MPG higher than automatic transmission.

Regression modelling

Here I will use the Best Subset Regression method to check which variable affects the mpg more. To run the regression with the best subset selection, we will use the regsubsets function from the package leaps

```
fit <- regsubsets(mpg~., cars, nvmax = 11)
# nvmax is the maximum number of predictors to be included in the model by default it is set to 8
sum <- summary(fit)
#to assess the goodness-of-fit we will use the adjusted R square
# because it is automatically computed for each model

sum$adjr2
```

```
##  [1] 0.7445939 0.8148396 0.8335561 0.8413067 0.8478136 0.8440579 0.8412051
##  [8] 0.8378688 0.8319505 0.8252176 0.8164807
```

```
which.max(sum$adjr2)# find the model with maximum adjusted R squared
```

```
## [1] 5
```

```
coef(fit, 5) # To print the coefficients of this model
```

```
## (Intercept)        cyl6          hp          wt         vs1         am1 
## 31.28240981 -2.20519611 -0.03393442 -2.36781111  1.87741318  2.62111773
```

```
model <- lm(mpg~ cyl+hp+wt+vs+am, data=cars)# Now applying the regression model on best fit coefficients
summary(model)
```

```
## 
## Call:
## lm(formula = mpg ~ cyl + hp + wt + vs + am, data = cars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.3405 -1.2158  0.0046  0.9389  4.6354 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 31.18461    3.42002   9.118    2e-09 ***
## cyl6        -2.09011    1.62868  -1.283   0.2112    
## cyl8         0.29098    3.14270   0.093   0.9270    
## hp          -0.03475    0.01382  -2.515   0.0187 *  
## wt          -2.37337    0.88763  -2.674   0.0130 *  
## vs1          1.99000    1.76018   1.131   0.2690    
## am1          2.70384    1.59850   1.691   0.1032    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.397 on 25 degrees of freedom
## Multiple R-squared:  0.8724, Adjusted R-squared:  0.8418 
## F-statistic: 28.49 on 6 and 25 DF,  p-value: 5.064e-10
```

The results suggest that the best model includes cyl6, cyl8, hp, wt, and manual variables. This model explains about 86.59% of the variance. Cylinders change negatively with mpg (-3.03miles and -2.16miles for cyl6 and cyl8, respectively), so do with horsepower (-0.03miles) and weight (-2.5miles for every 1,000lb). On the other hand, the manual transmission is 1.81mpg better than the automatic transmission.

Conclusion

On average, the manual transmission is better than the automatic transmission by 1.81mpg. However, transmission type is not the only factor accounting for MPG; cylinders, horsepower, and weight are the important factors affecting the MPG.