Regression_model_mtcars_pdf

Suhas P K

2023-07-17

Introduction

Motor Trend is a magazine about the automobile industry. In this analysis, a data set having information of car collection is explored to understand the relationship between **miles per gallon** and **transmission type**.
The data set is available from the CRAN repository.

Libraries used.

```
library(tinytex)
library(ggplot2)
library(ggdark)
```

Reading the data set.

```
data(mtcars)
```

Basics checking of the data set.

```
summary(mtcars)
```

```
##       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             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
```

```
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Check the structure of data set.

```
str(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 ...
```

Check for the missing data.

```
colSums(is.na(mtcars))
```

```
##  mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
##    0    0    0    0    0    0    0    0    0    0    0
```

There are no missing value.

Exploratory Data Analysis

Preliminary analysis

Understand the distribution of **miles per gallon** variable.

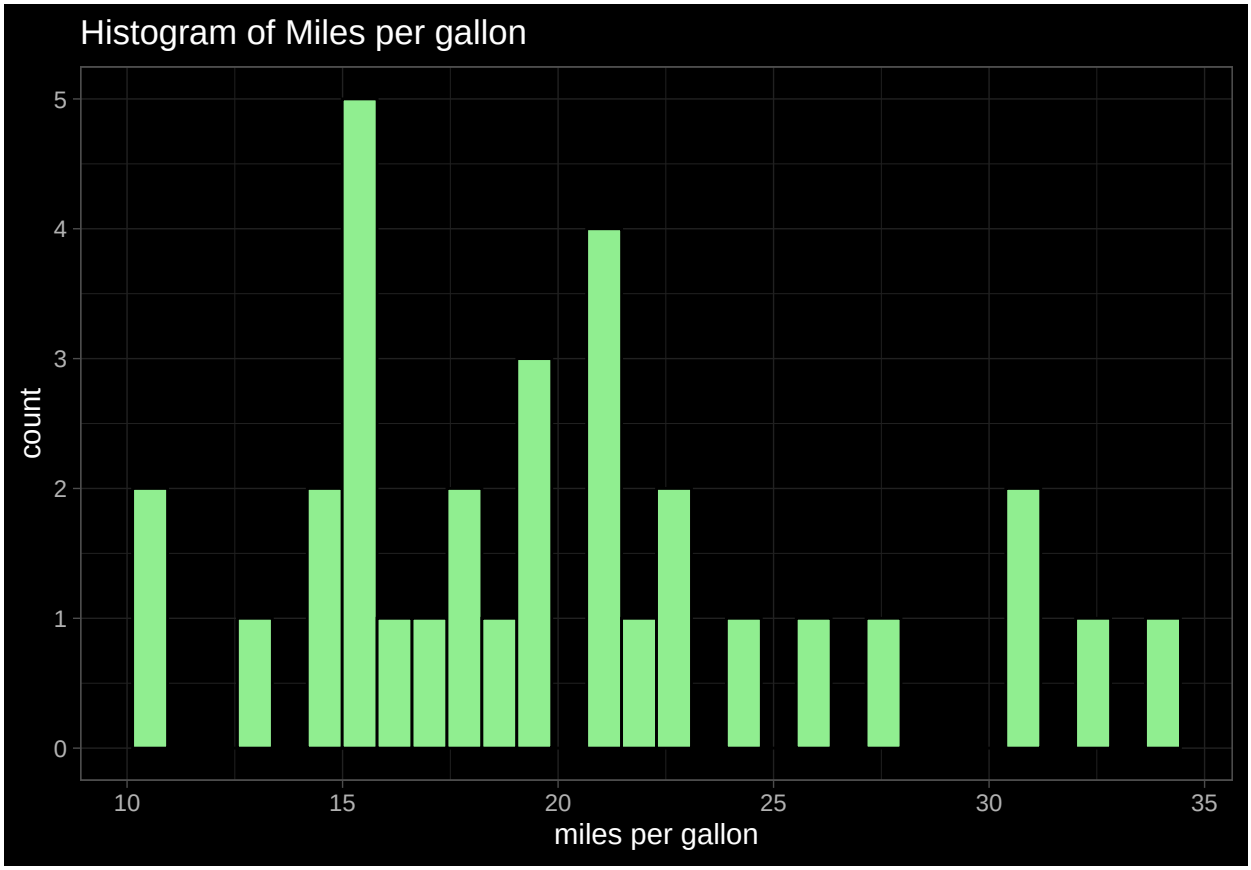
```
histplot <- ggplot(data = mtcars,
                   aes(x = mtcars$mpg)) + geom_histogram(color = "black",fill = "lightgreen") +
    xlab("miles per gallon")+
    ggtitle("Histogram of Miles per gallon")+
    dark_theme_light()
```

```
## Inverted geom defaults of fill and color/colour.
## To change them back, use invert_geom_defaults().
```

```
histplot
```

```
## Warning: Use of `mtcars$mpg` is discouraged. Use `mpg` instead.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

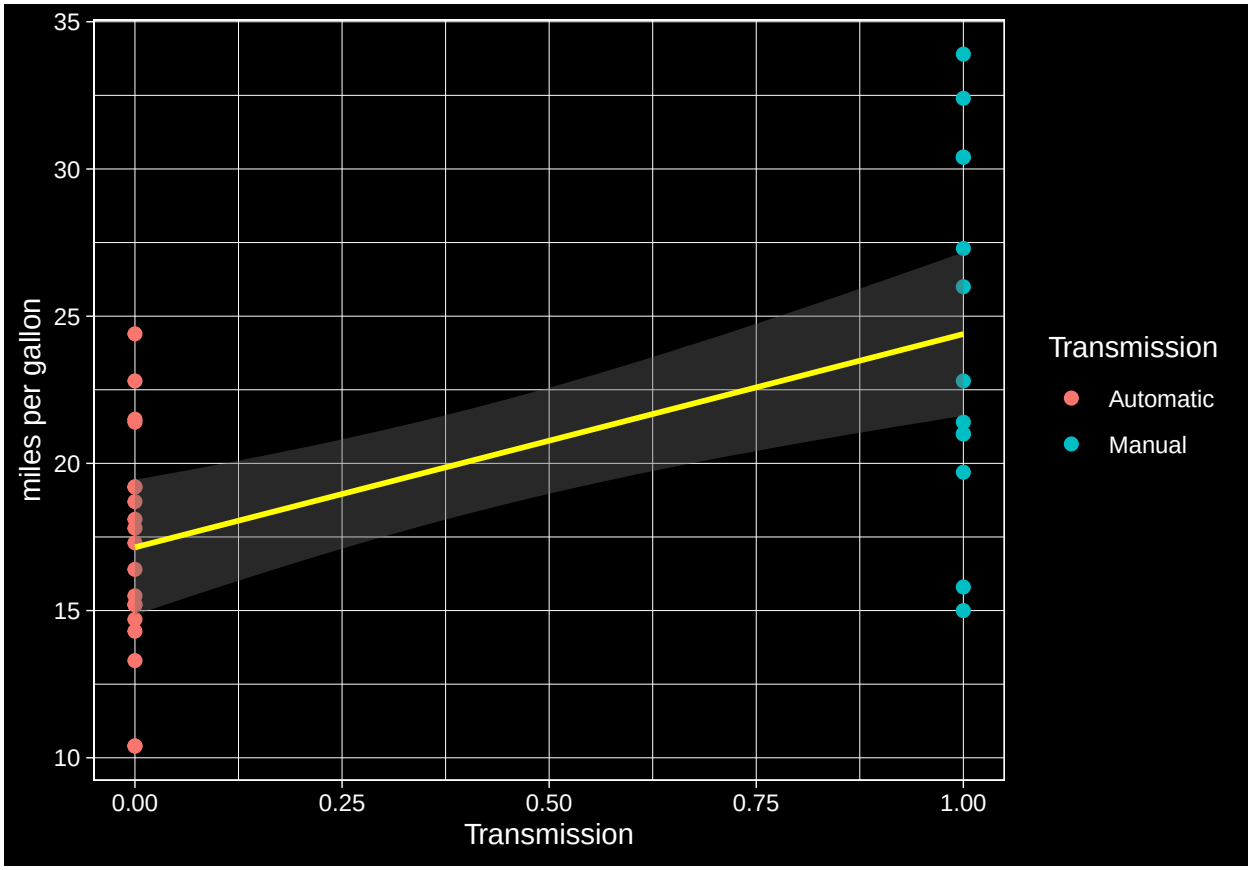


Plotting a scatter plot based on the transmission type and mpg.

```
scttrplot <- ggplot(data = mtcars,
                    aes(x = am, y = mpg, color = factor(am)))+
    geom_point(size = 2)+geom_smooth(method=lm, color = "yellow")

scttrplot +
    xlab("Transmission")+
    ylab("miles per gallon")+
    scale_colour_discrete(
        name = "Transmission",
        limits = c("0","1"),
        labels = c("Automatic",
                   "Manual")
    ) + dark_theme_linedraw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```

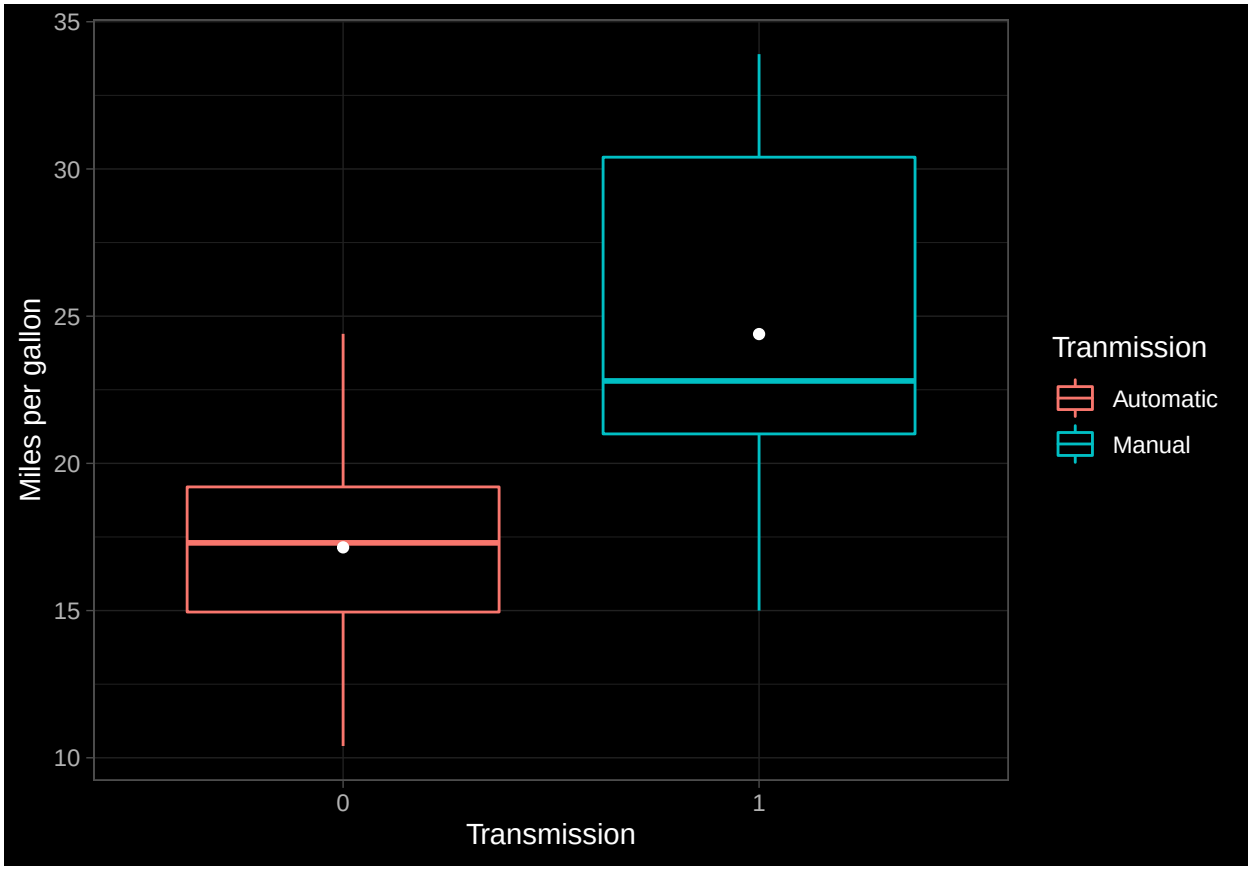


Visualizing the 'mpg' vs 'transmission' using boxplot.

```
    bxplot <- ggplot(mtcars, aes(x=factor(am),y = mpg, color = factor(am)))+
    geom_boxplot() +
    geom_point(stat = "summary",
               fun = "mean",
               color = "white", label = "mean")+
    xlab("Transmission")+
    ylab("Miles per gallon")
```

```
## Warning: Ignoring unknown parameters: label
```

```
bxplot +
    scale_colour_discrete(
    name = "Tranmission",
    limits = c("0","1"),
    labels = c("Automatic","Manual")
) + dark_theme_light()
```



To understand better from the scatter plot, the mean value of mpg based on the transmission type must be calculated. By far for this calculation I will use this method.

```
mean_am <- with(mtcars,
                tapply(mpg, am, mean))
mean_am
```

```
##        0        1
## 17.14737 24.39231
```

Now take the difference between the median based on the transmission type. With this we will get to know which transmission type has better mpg.

```
mean_am[2]-mean_am[1]
```

```
##        1
## 7.244939
```

In this case, the mean shows that, cars recorded with manual transmission can travel 7.24 more miles per gallon on average than the cars with automatic transmission.
Thus, manual transmission is better than the automatic.

A bit advance analysis

Performing **t-test** comparing the mean between the two transmission groups.

```
am_auto <- mtcars$mpg[mtcars$am == 0]
am_man <- mtcars$mpg[mtcars$am == 1]
t.test(
    am_auto, am_man,
```

```
    paired = FALSE,
    alternative = "two.sided",
    var.equal = FALSE
)
```

```
##
##  Welch Two Sample t-test
##
## data:  am_auto and am_man
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

The confidence interval (95%) does not contain zero (-11.28,-3.21) and p-value is greater then 0.005. Then, it can conclude that the average consumption, in miles per gallon, with automatic transmission is higher than the manual transmission. In this case, the mean analysis, it is possible to quantify the MPG difference between automatic and manual transmissions: 7.24 mpg greater, subtracting means.

Regression analysis

Single Model linear model The analysis is made to compare results from the **mean analysis**. The null hypothesis is that the difference between mean of **mpg** and **am** is zero.

```
single_model <-
    lm(mtcars$mpg ~ mtcars$am)
summary(single_model)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## mtcars$am    7.244939   1.764422  4.106127 2.850207e-04
```

The results show us that the p-value of the slope is less than 0.005. Then, it can reject the null hypothesis, and the results of the exploratory analysis were confirmed: automatic transmission results are 7.245 miles per gallon greater. If the slope is greater than zero, manual transmission is better than the automatic one.

Multivariable analysis.

```
require(MASS)
```

```
## Loading required package: MASS
```

```
multi_model <- stepAIC(
    lm(mpg~. , data = mtcars),
    direction = "both",
    trace = FALSE
)
multi_model$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
```

```
##
## Final Model:
## mpg ~ wt + qsec + am
##
##
##     Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                              21   147.4944 70.89774
## 2  - cyl  1 0.07987121        22   147.5743 68.91507
## 3   - vs  1 0.26852280        23   147.8428 66.97324
## 4 - carb  1 0.68546077        24   148.5283 65.12126
## 5 - gear  1 1.56497053        25   150.0933 63.45667
## 6 - drat  1 3.34455117        26   153.4378 62.16190
## 7 - disp  1 6.62865369        27   160.0665 61.51530
## 8   - hp  1 9.21946935        28   169.2859 61.30730
```

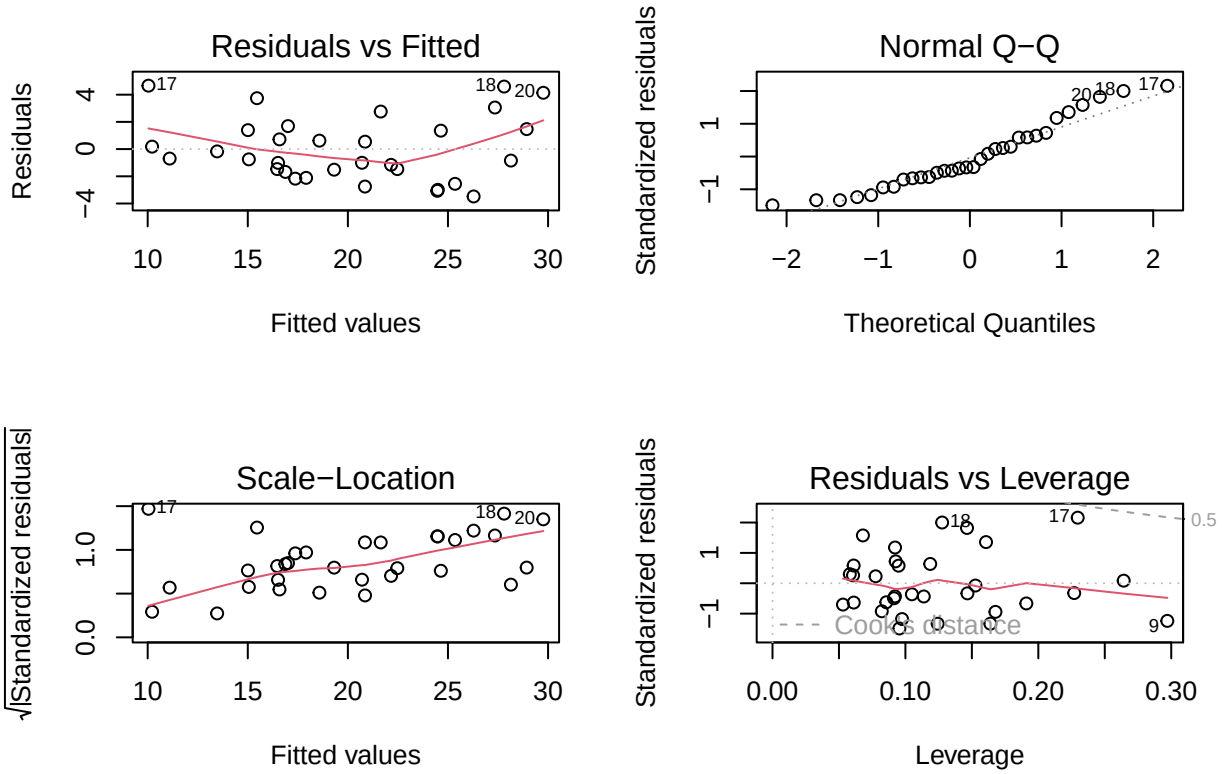
The **best model** indicated by the automated analysis consists of the variables **wt**, **qsec**, **am** and **mpg** as the outcome.

```
final_model <- lm(mtcars$mpg ~
                    mtcars$wt + mtcars$qsec + mtcars$am)
summary(final_model)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## mtcars$wt   -3.916504  0.7112016 -5.506882 6.952711e-06
## mtcars$qsec  1.225886  0.2886696  4.246676 2.161737e-04
## mtcars$am    2.935837  1.4109045  2.080819 4.671551e-02
```

Then, the regression equation is $mpg = 9.618 - 3.917wt + 1.226qsec + 1.4109am$. It is assumed that $Errors = 0$. As the two-sided p-value for the **am** coefficient is 0.04672, smaller than 0.05, it can we reject the null hypothesis.Looking at the plots,

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



Final Model Residuals , the visual analysis show us that the behavior of the best model is adequate considering normal residuals and constant variability. The leverage is within reasonable upper limit.

Conclusion

- Manual transmission is better than the automatic.
- Cars analyzed with manual transmission can travel 7.24 more miles per gallon on average than the cars with automatic transmission.
- There is a correlation between mpg and transmission, but other variables should also be considered, as qsec and wt, beyond the type of transmission.
- The obtained regression equation is **mpg = 9.618 -3.917 wt + 1.226 qsec + 1.4109 am** . Then, for the same weight (wt) and quarter mile time (qsec),manual transmission cars get 1.4109 miles per gallon more than automatic transmission cars.