Lab6.R

rstudio-user

2021-02-27

```
#1.Load mtcars dataset
data(mtcars)
head(mtcars)
```

```
##                   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
```

```
#2.Install ridge and glmnet packages
install.packages("glmnet", dependencies=TRUE)
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-1
```

```
install.packages("ridge")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(ridge)

#3.Perform the exploratory data analysis
dim(mtcars)
```

```
## [1] 32 11
```

```
#We can see that there are 32 rows and 11 columns

class(mtcars)
```

```
## [1] "data.frame"
```

```
#We can say that the mtcars is a dataframe

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 ...
```

```
#As we can see, most the columns are all numerical

sum(is.na(mtcars))
```

```
## [1] 0
```

```
#We can conclude that there is no null value in the dataset

summary(mtcars)
```

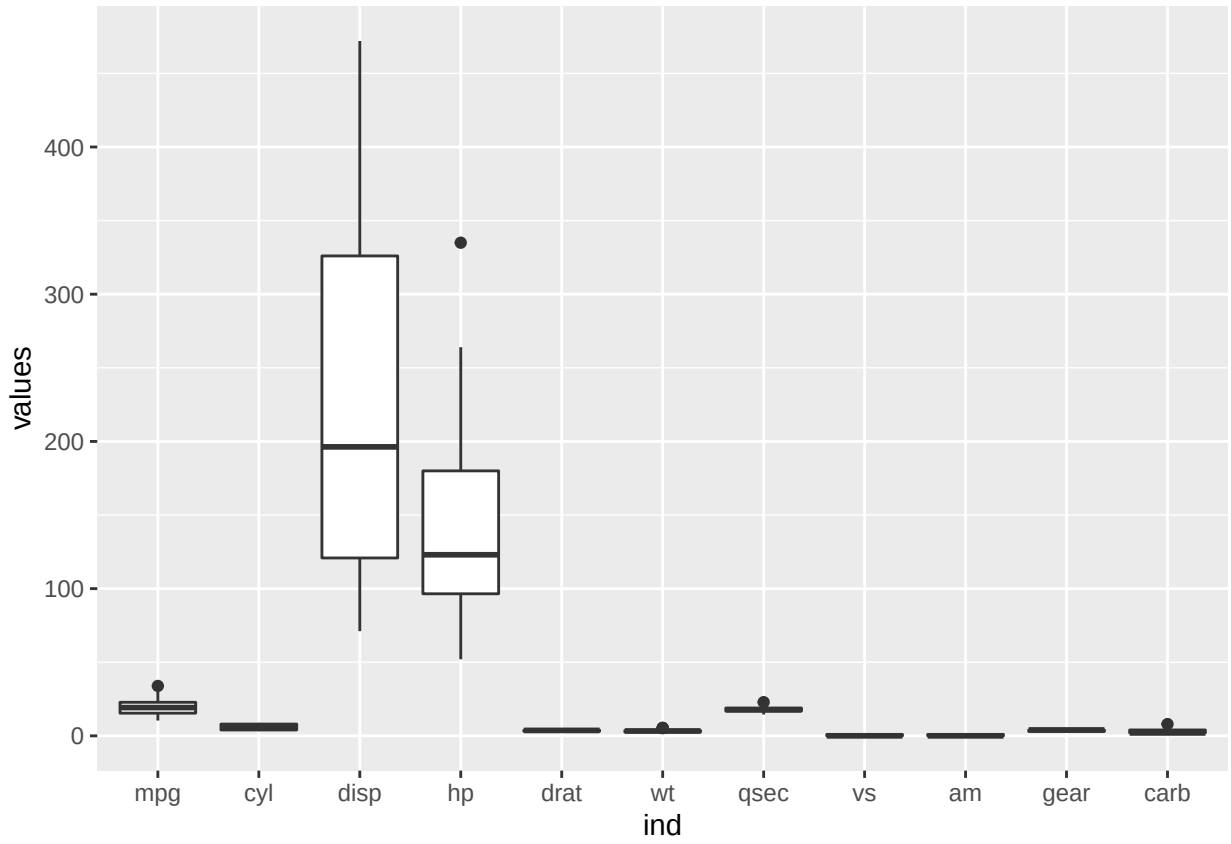
```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :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
```

```
#Central Tendency and other statistical measures of the dataset columns

install.packages("ggplot2")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(ggplot2)
ggplot(stack(mtcars), aes(x =ind, y =values)) + geom_boxplot()
```




```
#We can say that there are outliers in few columns like horsepower, mpg, etc.
```

```
install.packages("Hmisc")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

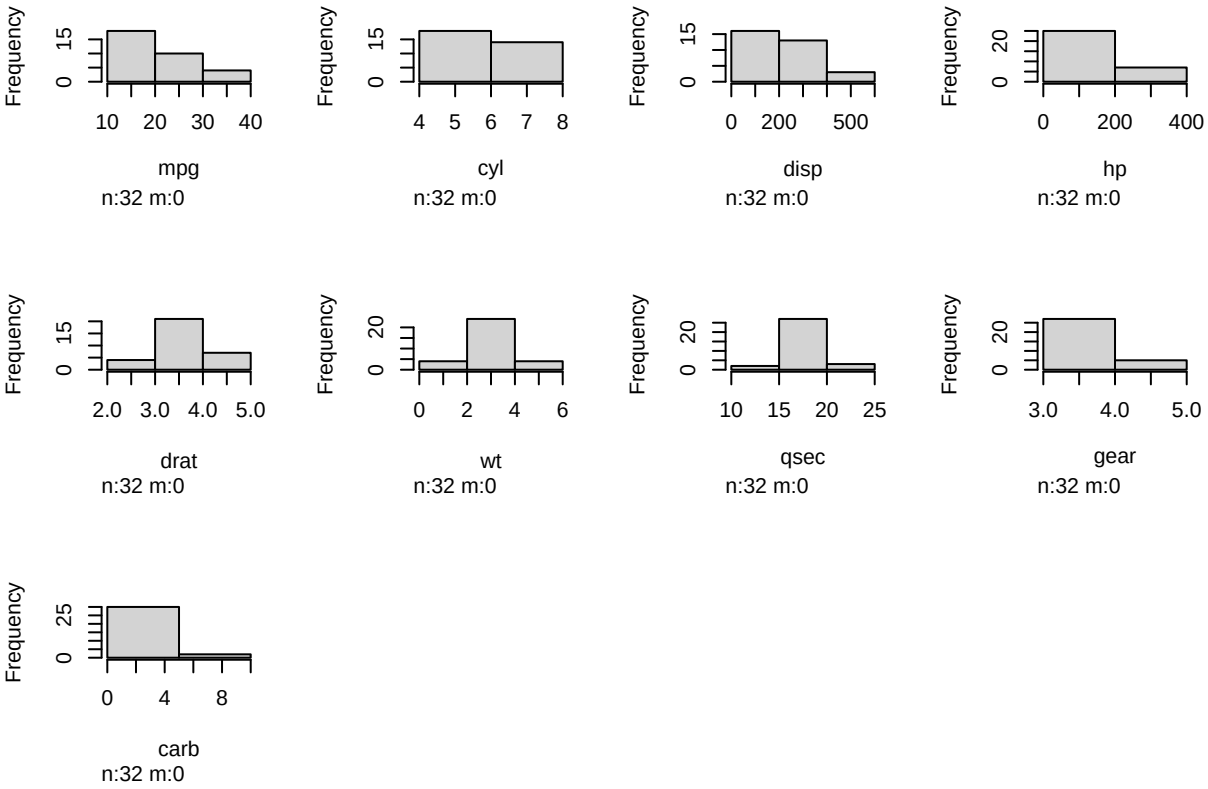
```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
hist.data.frame(mtcars)
```




```
#We can infer that a few columns are normally distributed while some are not

#4.Choose optimum lamba value
#Predictor variables
x <- data.matrix(mtcars[, c("hp", "wt", "drat")])
#Target variable
y <- mtcars[, "mpg"]

lamb <- 10^seq(2, -2, by = -.1)
# Using glmnet function to build the ridge regression in r
fit <- glmnet(x, y, alpha = 0, lambda  = lamb)
summary(fit)
```

```
##           Length Class     Mode
## a0          41   -none-    numeric
## beta       123   dgCMatrix S4
## df          41   -none-    numeric
## dim          2   -none-    numeric
## lambda      41   -none-    numeric
## dev.ratio   41   -none-    numeric
## nulldev      1   -none-    numeric
## npasses      1   -none-    numeric
## jerr         1   -none-    numeric
## offset       1   -none-    logical
## call         5   -none-    call
## nobs         1   -none-    numeric
```

```
#Using cross validation glmnet
ridge_cv <- cv.glmnet(x, y, alpha = 0, lambda = lamb)
```

```r
# Best lambda value
opt_lamb <- ridge_cv$lambda.min
opt_lamb
```

```
## [1] 0.7943282
```

```r
#The optimal value of lambda is 0.7943282.

#5.Extract the model using k-cross validation
extracted_model <- ridge_cv$glmnet.fit
extracted_model
```

```
##
## Call:  glmnet(x = x, y = y, lambda = lamb, alpha = 0)
##
##    Df  %Dev  Lambda
## 1   3 17.98 100.000
## 2   3 21.67  79.430
## 3   3 25.89  63.100
## 4   3 30.60  50.120
## 5   3 35.74  39.810
## 6   3 41.20  31.620
## 7   3 46.81  25.120
## 8   3 52.39  19.950
## 9   3 57.75  15.850
## 10  3 62.69  12.590
## 11  3 67.09  10.000
## 12  3 70.86   7.943
## 13  3 73.98   6.310
## 14  3 76.48   5.012
## 15  3 78.42   3.981
## 16  3 79.89   3.162
## 17  3 80.98   2.512
## 18  3 81.78   1.995
## 19  3 82.35   1.585
## 20  3 82.76   1.259
## 21  3 83.05   1.000
## 22  3 83.25   0.794
## 23  3 83.39   0.631
## 24  3 83.49   0.501
## 25  3 83.56   0.398
## 26  3 83.60   0.316
## 27  3 83.63   0.251
## 28  3 83.65   0.200
## 29  3 83.66   0.158
## 30  3 83.67   0.126
## 31  3 83.68   0.100
## 32  3 83.68   0.079
## 33  3 83.68   0.063
## 34  3 83.69   0.050
## 35  3 83.69   0.040
## 36  3 83.69   0.032
## 37  3 83.69   0.025
## 38  3 83.69   0.020
## 39  3 83.69   0.016
```

```
## 40  3 83.69   0.013
## 41  3 83.69   0.010
```

```
#6.Build the final model and interpret
final_mod <- glmnet(x, y, alpha = 0, lambda = 0.5011872)
coef(final_mod)
```

```
## 4 x 1 sparse Matrix of class "dgCMatrix"
##                      s0
## (Intercept) 27.09643213
## hp          -0.03164043
## wt          -2.88955100
## drat         1.92735603
```

```
#These are the intercept and coefficients of the model.
```