Regression Analysis

Kisha Taylor

Mon May 07 00:35:10 2018

```
# Machine Learning Project#1
# Prepared by: Kisha Taylor
# Due date : Nov. 20, 2017

###########################################################################
########                     BIKE DATA                           #########
###########################################################################
# Regression Analysis implementation from Scratch
# Bike Daset where output is numerical
# For the Bike Data set the following attributes were selected :
##### (1) Season
##### (2) Holiday
##### (3) Weekday
##### (4) Weather situation
##### (5) total rentals

# Dataset sourced from : https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset
setwd("C:/Users/Kisha/Downloads")
Bikedata <- read.csv("day.csv",header=TRUE)
head(Bikedata)
```

```
##   instant     dteday season yr mnth holiday weekday workingday weathersit
## 1       1 2011-01-01      1  0    1       0       6          0          2
## 2       2 2011-01-02      1  0    1       0       0          0          2
## 3       3 2011-01-03      1  0    1       0       1          1          1
## 4       4 2011-01-04      1  0    1       0       2          1          1
## 5       5 2011-01-05      1  0    1       0       3          1          1
## 6       6 2011-01-06      1  0    1       0       4          1          1
##       temp    atemp      hum windspeed casual registered  cnt
## 1 0.344167 0.363625 0.805833 0.1604460    331        654  985
## 2 0.363478 0.353739 0.696087 0.2485390    131        670  801
## 3 0.196364 0.189405 0.437273 0.2483090    120       1229 1349
## 4 0.200000 0.212122 0.590435 0.1602960    108       1454 1562
## 5 0.226957 0.229270 0.436957 0.1869000     82       1518 1600
## 6 0.204348 0.233209 0.518261 0.0895652     88       1518 1606
```

```
summary(Bikedata)
```

```
##     instant          dteday        season            yr        
##  Min.   :  1.0   2011-01-01:  1   Min.   :1.000   Min.   :0.0000  
##  1st Qu.:183.5   2011-01-02:  1   1st Qu.:2.000   1st Qu.:0.0000  
##  Median :366.0   2011-01-03:  1   Median :3.000   Median :1.0000  
##  Mean   :366.0   2011-01-04:  1   Mean   :2.497   Mean   :0.5007  
##  3rd Qu.:548.5   2011-01-05:  1   3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :731.0   2011-01-06:  1   Max.   :4.000   Max.   :1.0000  
##                  (Other)   :725                                   
##       mnth          holiday           weekday        workingday   
##  Min.   : 1.00   Min.   :0.00000   Min.   :0.000   Min.   :0.000  
##  1st Qu.: 4.00   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000  
##  Median : 7.00   Median :0.00000   Median :3.000   Median :1.000  
##  Mean   : 6.52   Mean   :0.02873   Mean   :2.997   Mean   :0.684  
##  3rd Qu.:10.00   3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:1.000  
##  Max.   :12.00   Max.   :1.00000   Max.   :6.000   Max.   :1.000  
##                                                                   
##    weathersit         temp             atemp              hum        
##  Min.   :1.000   Min.   :0.05913   Min.   :0.07907   Min.   :0.0000  
##  1st Qu.:1.000   1st Qu.:0.33708   1st Qu.:0.33784   1st Qu.:0.5200  
##  Median :1.000   Median :0.49833   Median :0.48673   Median :0.6267  
##  Mean   :1.395   Mean   :0.49538   Mean   :0.47435   Mean   :0.6279  
##  3rd Qu.:2.000   3rd Qu.:0.65542   3rd Qu.:0.60860   3rd Qu.:0.7302  
##  Max.   :3.000   Max.   :0.86167   Max.   :0.84090   Max.   :0.9725  
##                                                                      
##    windspeed          casual         registered        cnt      
##  Min.   :0.02239   Min.   :   2.0   Min.   :  20   Min.   :  22  
##  1st Qu.:0.13495   1st Qu.: 315.5   1st Qu.:2497   1st Qu.:3152  
##  Median :0.18097   Median : 713.0   Median :3662   Median :4548  
##  Mean   :0.19049   Mean   : 848.2   Mean   :3656   Mean   :4504  
##  3rd Qu.:0.23321   3rd Qu.:1096.0   3rd Qu.:4776   3rd Qu.:5956  
##  Max.   :0.50746   Max.   :3410.0   Max.   :6946   Max.   :8714  
## 
```

```
##############            EXPLORATION  ######################
## Used to select attributes wih the highest correlation

library(caret)
```

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

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

```
myBike_factors <- dummyVars(" ~ .",data = Bikedata)
ExploreBike <- data.frame(predict(myBike_factors,Bikedata))
class(ExploreBike)
```

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

```
explorecor <- cor(ExploreBike)

explorecor["cnt","season"]
```

```
## [1] 0.4061004
```

```
explorecor["cnt","yr"]
```

```
## [1] 0.5667097
```

```
explorecor["cnt","mnth"]
```

```
## [1] 0.2799771
```

```
explorecor["cnt","holiday"]
```

```
## [1] -0.06834772
```

```
explorecor["cnt","weekday"]
```

```
## [1] 0.06744341
```

```
explorecor["cnt","workingday"]
```

```
## [1] 0.06115606
```

```
explorecor["cnt","weathersit"]
```

```
## [1] -0.2973912
```

```
explorecor["cnt","hum"]
```

```
## [1] -0.1006586
```

```
explorecor["cnt","windspeed"]
```

```
## [1] -0.234545
```

```
explorecor["cnt","casual"]
```

```
## [1] 0.6728044
```

```
explorecor["cnt","registered"]
```

```
## [1] 0.9455169
```

```
explorecor["cnt","temp"]
```

```
## [1] 0.627494
```

```
explorecor["cnt","atemp"]
```

```
## [1] 0.6310657
```

```
## Based on exploration the following attributes were selected

#Attribute : Temperature "aTemp" Normalized temperature in Celsius. The values are derived via
 (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)

#Attribute : season


#Attribute : Weather situation

#Attribute : Windspeed


#Attribute : total rentals "cnt"


myBikedata <- data.frame(Bikedata$atemp,Bikedata$season,Bikedata$weathersit,Bikedata$windspeed,B
ikedata$cnt)

colnames(Bikedata)
```

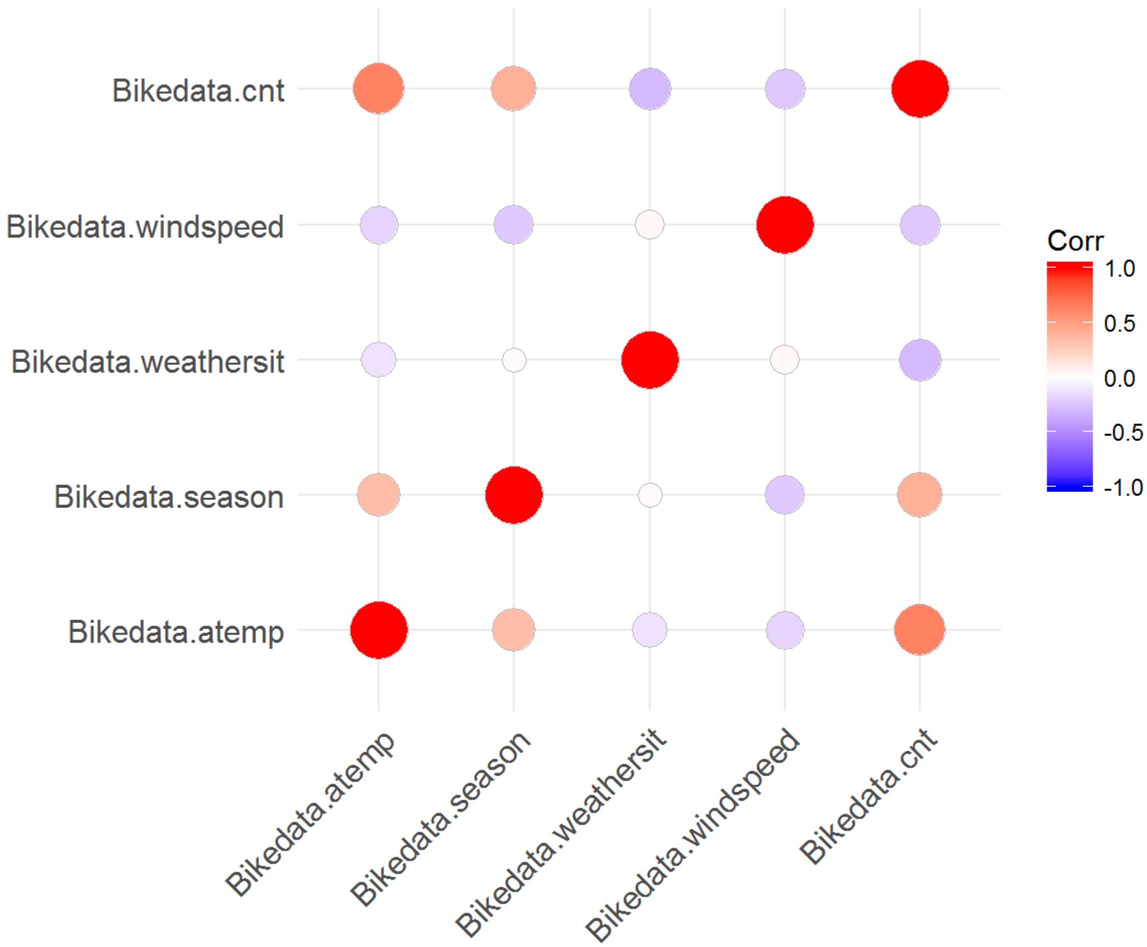
```
##  [1] "instant"    "dteday"     "season"     "yr"         "mnth"
##  [6] "holiday"    "weekday"    "workingday" "weathersit" "temp"
## [11] "atemp"      "hum"        "windspeed"  "casual"     "registered"
## [16] "cnt"
```

```
mycormat_Bikedata <- cor(myBikedata)



#install.packages("ggcorrplot")
library(ggcorrplot)

# method = "circle"
ggcorrplot(mycormat_Bikedata, method = "circle")
```



```
### Observed Dependencies
### After exploration
#### THe folllowng correlation observations were made
#### The attributes with the highest correlation were
####
#### Temp & Cnt : positive cor of 0.631
####
#### Season & Cnt : positive cor of 0.406

#### Weather situation & Cnt : negative cor of 0.297
###  Windspeed & Cnt : neg. cor of -0.235

#### Windspeed & Cnt : negative cor of 0.2345
#### Othe observations included :
### low to moderate cor b/w atemp & the following:
####     (i) season : pos cor of 0.34
###      (ii) weather situation : neg cor of -0.12
###      (iii) windspeed : neg cor of -0.184

###  low to moderate cor b/w season & the foll. :
###       (i) windspeed : -0.229



### Applying multivariate regresson analysis
# soving for parameters w in w= (XT.X)^-1 . XT.r
#where XT rep. X transpose and ^-1 represents inverse and r rep. the output value
# we will apply this to the entire training dataset

# Spliting data set into training and test set
dim(myBikedata)
```

```
## [1] 731   5
```

```
bike_rn <- nrow(myBikedata)
bike_cn <- ncol(myBikedata)
Tr_nrows <- round(0.75*bike_rn,digit=0)
BikeInput_Tr <- myBikedata[1:Tr_nrows,-bike_cn]
dim(BikeInput_Tr)
```

```
## [1] 548   4
```

```
myBikedata_Tr <- myBikedata[1:Tr_nrows,]
dim(myBikedata_Tr)
```

```
## [1] 548   5
```

```
myBikedata_Test <- myBikedata[((Tr_nrows + 1):bike_rn),]
dim(myBikedata_Test)
```

```
## [1] 183   5
```

```
colnames(myBikedata_Test)
```

```
## [1] "Bikedata.atemp"      "Bikedata.season"     "Bikedata.weathersit"
## [4] "Bikedata.windspeed"  "Bikedata.cnt"
```

```
##add column to input data set to solve for w using training input set

#BikeInput_Tr_mod <- BikeInput_Tr_mod$col1
BikeInput_Tr_mod <-  data.frame(wo_constant=rep(1,nrow(BikeInput_Tr)),BikeInput_Tr)


head(BikeInput_Tr_mod)
```

```
##   wo_constant Bikedata.atemp Bikedata.season Bikedata.weathersit
## 1           1       0.363625               1                   2
## 2           1       0.353739               1                   2
## 3           1       0.189405               1                   1
## 4           1       0.212122               1                   1
## 5           1       0.229270               1                   1
## 6           1       0.233209               1                   1
##   Bikedata.windspeed
## 1          0.1604460
## 2          0.2485390
## 3          0.2483090
## 4          0.1602960
## 5          0.1869000
## 6          0.0895652
```

```
dim(BikeInput_Tr_mod)
```

```
## [1] 548   5
```

```
colnames(BikeInput_Tr_mod)
```

```
## [1] "wo_constant"         "Bikedata.atemp"      "Bikedata.season"
## [4] "Bikedata.weathersit" "Bikedata.windspeed"
```

```
#applying formula to derive w
r <- myBikedata_Tr[,ncol(myBikedata_Tr)]
head(r)
```

```
## [1]  985  801 1349 1562 1600 1606
```

```
BikeSales <- solve(t(as.matrix(BikeInput_Tr_mod))%*%as.matrix(BikeInput_Tr_mod))%*%t(as.matrix(B
ikeInput_Tr_mod))%*%(as.matrix(r))
head(BikeInput_Tr_mod)
```

```
##   wo_constant Bikedata.atemp Bikedata.season Bikedata.weathersit
## 1           1       0.363625               1                   2
## 2           1       0.353739               1                   2
## 3           1       0.189405               1                   1
## 4           1       0.212122               1                   1
## 5           1       0.229270               1                   1
## 6           1       0.233209               1                   1
##   Bikedata.windspeed
## 1          0.1604460
## 2          0.2485390
## 3          0.2483090
## 4          0.1602960
## 5          0.1869000
## 6          0.0895652
```

```
dim(BikeSales)
```

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

```
head(BikeSales)
```

```
##                            [,1]
## wo_constant          2429.81225
## Bikedata.atemp       6661.03345
## Bikedata.season       -55.69886
## Bikedata.weathersit  -769.00010
## Bikedata.windspeed  -1641.08953
```

```
# So, our model is as follows:

# cnt_bikeSales <- 2429.81225+ 6661.03345*atemp-55.69886*season-769.00010*weathersituation
#                   -1641.08953*windspeed


# For testing, dset rep. by myBikedata_Test
# model rep. by BikeSales

sales_predict <- c()
sales_predict<- (as.matrix(myBikedata_Test[,-5]))%*%as.matrix(BikeSales[-1]) + BikeSales[1]

r_test <- myBikedata_Test[,ncol(myBikedata_Test)]


error_cal <- function(msales,actualsales){
  m_rnum <- length(msales)
  error <- rep(0,m_rnum)
  sq_error <- rep(0,m_rnum)

  for (i in 1:m_rnum){
    error[i] <- msales[i] - actualsales[i]
    error[i]
    sq_error[i] <- (error[i])^2
    sq_error[i]
  }#end for loop
  sqrtmeansq_error <- sqrt(((sum(sq_error))/m_rnum))
  return(sqrtmeansq_error);
}#end function

error_model <- error_cal(sales_predict,r_test)
error_ck <- data.frame(sales_predict,r_test)


error_model # root of the mean squared error of all test data based on model predictions
```

```
## [1] 2168.17
```

```
# result is an error of 2168.17 -not good. This means that on average the predictions will be of
f by about this about.
# This could be exxplained by the fcat that the variables used, though the highest from what was
 available,
#the correlations were not very strong. Highet was temperature at a pos cor of 0.631 to the resp
onse variable.
```