

**STUDENT NAME : PAUL DINESH AUGUSTINE**

**STUDENT NO : 10515229**

**MODULE TITLE : DATA AND DATA ANALYTICS**

**MODULE CODE : B9IS106**

**LECTURER : Dr SHAHRAM AZIZI**

## Introduction:

My answers to Assignment 1 for B9IS106 are found below.

### Question 1:

#### 1.(a)

In 1.(a), To explore the general features of dataset, the commands are as follows:

#Quick inspection of raw data  
head(airquality) #To view the first 6 rows of the dataset

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 NA NA 14.3 56 5 5  
## 6 28 NA 14.9 66 5 6

tail(airquality) #To view the last 6 rows of the dataset

## Ozone Solar.R Wind Temp Month Day  
## 148 14 20 16.6 63 9 25  
## 149 30 193 6.9 70 9 26  
## 150 NA 145 13.2 77 9 27  
## 151 14 191 14.3 75 9 28  
## 152 18 131 8.0 76 9 29  
## 153 20 223 11.5 68 9 30

View(airquality) #To view the dataset  
#Type of data & its dimensions  
class(airquality) #To prints the vector of names of classes

## [1] "data.frame"

nrow(airquality) #To find the number of rows in the dataset

## [1] 153

ncol(airquality) #To find the number of columns in the dataset

## [1] 6

length(airquality[,1]) #To find the number of rows

## [1] 153

length(airquality)#To find the number of columns

## [1] 6

names(airquality) #To view names of columns

## [1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"

str(airquality) #To view the structure of the dataset

## 'data.frame': 153 obs. of 6 variables:  
## $ Ozone : int 41 36 12 18 NA 28 23 19 8 NA ...  
## $ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...  
## $ Wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...  
## $ Temp : int 67 72 74 62 56 66 65 59 61 69 ...  
## $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Day : int 1 2 3 4 5 6 7 8 9 10 ...

help("airquality")#To view the Description, Usage, Format, Details, Source, Reference, Examples in R Documentation

#### 1.(b)

In 1.(b), To perform data Cleansing the commands are as follows:

my.data<-airquality  
  
head(my.data) #To view the first 6 rows of the dataset

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 NA NA 14.3 56 5 5  
## 6 28 NA 14.9 66 5 6

tail(my.data) #To view the last 6 rows of the dataset

## Ozone Solar.R Wind Temp Month Day  
## 148 14 20 16.6 63 9 25  
## 149 30 193 6.9 70 9 26  
## 150 NA 145 13.2 77 9 27  
## 151 14 191 14.3 75 9 28  
## 152 18 131 8.0 76 9 29  
## 153 20 223 11.5 68 9 30

is.na(my.data) #To check the dataset has null values #if TRUE= null, FALSE= not null

## Ozone Solar.R Wind Temp Month Day  
## [1,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [2,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [3,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [4,] FALSE FALSE FALSE FALSE FALSE FALSE

any(is.na(my.data)) #To check the dataset has null values

## [1] TRUE

sum(is.na(my.data)) #To find total of null values in a dataset

## [1] 44

str(my.data) #To view the structure of the dataset

## 'data.frame': 153 obs. of 6 variables:  
## $ Ozone : int 41 36 12 18 NA 28 23 19 8 NA ...  
## $ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...  
## $ Wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...  
## $ Temp : int 67 72 74 62 56 66 65 59 61 69 ...  
## $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Day : int 1 2 3 4 5 6 7 8 9 10 ...

colSums(is.na(my.data)) #To view the which column has null values and its   
total per column

## Ozone Solar.R Wind Temp Month Day   
## 37 7 0 0 0 0

nrow(my.data) #No. of rows before cleansing

## [1] 153

my.data.clean<-na.omit(my.data) #To remove all rows containing NA  
  
any(is.na(my.data.clean)) #To check the dataset has any null values after   
cleansing

## [1] FALSE

nrow(my.data.clean) #To check the number of rows not affected after cleansing

## [1] 111

#### 1.(c)

In 1.(c) Computation of central and variational measures. The commands are as follows:

central measures –mean, median, mode

variational measures –range, standard deviation, variance

head(my.data.clean)

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 7 23 299 8.6 65 5 7  
## 8 19 99 13.8 59 5 8

View(my.data.clean)  
temp<-my.data.clean$Temp #To extract temp column from the cleansed dataset  
  
temp\_mean= mean(temp) #To find mean of temp  
temp\_mean

## [1] 77.79279

temp\_median=median(temp) #To find median of temp  
temp\_median

## [1] 79

temp\_occurence<-table(as.vector(temp)) #To find occurence of temp  
temp\_occurence

##   
## 57 58 59 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82   
## 1 1 2 3 2 1 2 2 2 3 4 2 1 3 3 4 2 2 6 4 4 3 3 10 7   
## 83 84 85 86 87 88 89 90 91 92 93 94 96 97   
## 3 3 3 5 3 2 2 3 1 3 2 2 1 1

names(temp\_occurence)[temp\_occurence==max(temp\_occurence)] #To find mode of   
temp

## [1] "81"

temp\_sd=sd(temp) #To find standard deviation of Temp  
temp\_sd

## [1] 9.529969

temp\_var=var(temp) #To find variance of temp  
temp\_var

## [1] 90.82031

temp\_sd\_sqrt\_var=sqrt(temp\_var) #sqrt of var = sd  
temp\_sd\_sqrt\_var

## [1] 9.529969

temp\_range=range(temp) #To find Range of temp  
temp\_range

## [1] 57 97

quantile(temp) #To find quantile of temp

## 0% 25% 50% 75% 100%   
## 57.0 71.0 79.0 84.5 97.0

summary(temp) #To find Mininum & Maximum value, 1st & 3rd quantile, mean of   
temp

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 57.00 71.00 79.00 77.79 84.50 97.00

#### 1.(d)

In 1.(d) boxplot technique to detect outlier of ‘wind’ attribute

Wind\_col=my.data.clean$Wind  
   
sort\_Wind\_col = Wind\_col[order(Wind\_col)] #To sort the column wind of   
airquality dataset  
  
fivedata\_Wind\_col =fivenum(sort\_Wind\_col) #fivenum function returns a vector of length five which contains minimum, lower-hinge, median, upper-hinge, maximum values  
  
fivedata\_Wind\_col

## [1] 2.3 7.4 9.7 11.5 20.7

IQR=fivedata\_Wind\_col[4]-fivedata\_Wind\_col[2] #InterQuartile Range of Wind  
IQR

## [1] 4.1

QLw=fivedata\_Wind\_col[2]-1.5\*IQR #Lower Quartile of Wind  
QLw

## [1] 1.25

QUw=fivedata\_Wind\_col[4]+1.5\*IQR #Upper Quartile of Wind  
QUw

## [1] 17.65

Int=c(QLw,QUw)  
vect=c()  
length(Wind) #number of rows in column Wind

## [1] 111

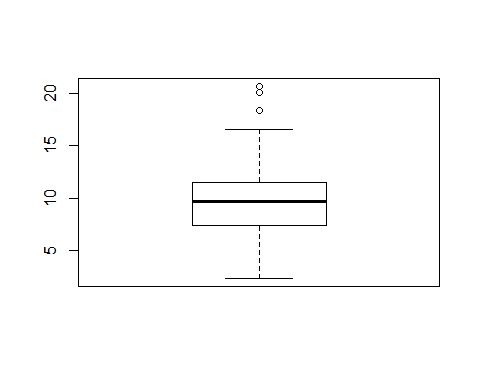
#The values outside Inter & Outer Quartile range are considered as outliers  
for(i in Wind)  
{if((i>QUw)&(i>QLw))  
{vect=c(vect,i)}  
 }  
vect

## [1] 20.1 18.4 20.7

length(vect) #To find the number of outliers in column Wind

## [1] 3

View(vect) #To view the outliers  
#Alternatively, we can find the outliers using the statistics function of   
boxplot  
boxplot(Wind)

The outliers are 20.1, 18.4, 20.7

boxplot.stats(Wind)

## $stats  
## [1] 2.3 7.4 9.7 11.5 16.6  
##   
## $n  
## [1] 111  
##   
## $conf  
## [1] 9.085135 10.314865  
##   
## $out  
## [1] 20.1 18.4 20.7

#The last line of the statistics of boxplot has the outliers 20.1, 18.4, 20.7

#### 

#### 2(a)

In 2(a) Generalised Linear Model

dataset <- read.csv("http://users.stat.ufl.edu/~winner/data/nfl2008\_fga.csv")  
library(caTools) # useful to split data to training and test datasets  
  
set.seed(1456)  
n=nrow(dataset)  
n

## [1] 1039

indexes = sample(n,n\*(80/100)) #To split the dataset  
traindataset=dataset[indexes,]# training dataset  
testdataset =dataset[-indexes,]# test dataset  
#(a)   
table(traindataset$homekick)

##   
## 0 1   
## 417 414

By identifying the output variable, the appropriate glm to be used is logistic regression as the homekick column contains binary outcomes

**2(b)**

fit=glm(homekick~togo+ydline+kicker,data=traindataset,family='binomial')  
summary(fit)

##   
## Call:  
## glm(formula = homekick ~ togo + ydline + kicker, family = "binomial",   
## data = traindataset)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3834 -1.1695 -0.9385 1.1692 1.4355   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.140714 0.209486 0.672 0.5018   
## togo -0.042133 0.017288 -2.437 0.0148 \*  
## ydline 0.011670 0.007518 1.552 0.1206   
## kicker -0.004444 0.006221 -0.714 0.4750   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1149.2 on 828 degrees of freedom  
## Residual deviance: 1141.9 on 825 degrees of freedom  
## (2 observations deleted due to missingness)  
## AIC: 1149.9  
##   
## Number of Fisher Scoring iterations: 4

To consider the significance of the variable p-value should be less than alpha. From the output, p-value or the probability is less for togo than alpha =0.05 therefore, only the input variable **togo** is considered significant.

Mathematical equation for Logistic Regression for the given problem can be represented as

#### 2(C)

Prediction of the test dataset using the trained model is done below:

pred=predict(fit,testdataset)  
predicted\_values=rep(0,length(pred))  
predicted\_values[pred>0.5]=1  
actual=testset$homekick  
df=data.frame(actual,predicted\_values)  
View(df)  
head(df)

## actual predicted\_values  
## 1 0 0  
## 2 0 0  
## 3 1 0  
## 4 1 0  
## 5 1 0  
## 6 1 0

#### 2(D)

confusion\_matrix=table(predicted\_values, actual)  
confusion\_matrix

## actual  
## predicted\_values 0 1  
## 0 108 100

View(confusion\_matrix)

Actual vs Predicted values are represented in tabular form which is obtained in confusion matrix.

Accuracy of the model can be obtained by the below formula based on the actual and the predicted values.

accuracy=mean(predicted\_values == actual) # correctness of prediction  
accuracy

## [1] 0.5192308

#### 3(A)

Assumption Validation using graphical visualization.

library(forecast)  
link='http://www.stat.ufl.edu/~winner/data/wage\_cpi.csv'  
data\_wage\_cpi=read.csv(link)  
head(data\_wage\_cpi)

## Year yr\_month ser\_month wage cpi  
## 1 2000 1 1 172.6 373.06  
## 2 2000 2 2 174.8 365.15  
## 3 2000 3 3 190.7 358.72  
## 4 2000 4 4 213.6 342.34  
## 5 2000 5 5 203.0 336.22  
## 6 2000 6 6 210.4 326.95

str(data\_wage\_cpi)

## 'data.frame': 129 obs. of 5 variables:  
## $ Year : int 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...  
## $ yr\_month : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ ser\_month: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ wage : num 173 175 191 214 203 ...  
## $ cpi : num 373 365 359 342 336 ...

x=data\_wage\_cpi$wage  
class(x)

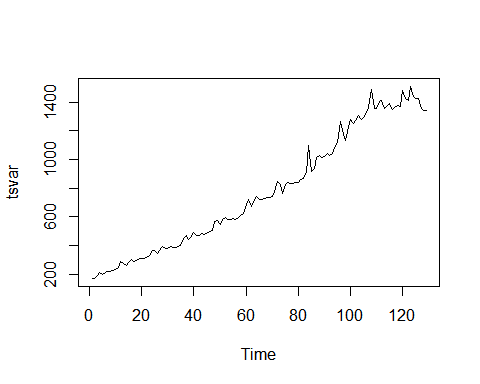
## [1] "numeric"

tsvariable <- ts(x,frequency=1)#ts function helps conversion timeseries data  
class(tsvariable)

## [1] "ts"

Validate the assumption 1)Stationarity 2)Normality

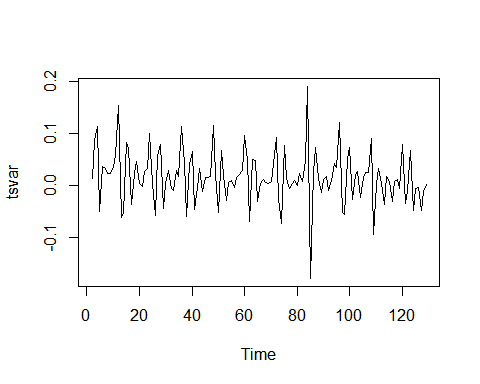
plot(tsvariable)



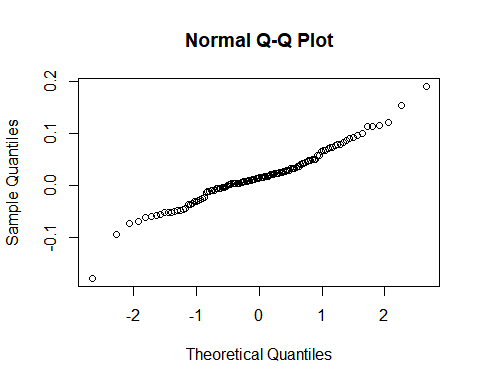
By plotting the timeseries data we can see that it is non-stationary in mean

To make the timeseries data to stationary in mean diff function is applied

tsvariable=diff(log(tsvariable)) # to make ts stationary in mean and variance use diff & log  
plot(tsvariable)



qqnorm(tsvariable) # if the graph is straight slope line the time series data is   
normal



From the above plots the assumption of timeseries data to be 1) Stationarity 2) Normality is true

#### 3(b)

The optimized model for ‘wage’ and the coefficient estimates for the fitted model is found using auto arima function, the one with the lowest values of B IC and AIC should be our choice of the optimized model

#Seasonal =F & T test is performed to check AIC & BIC   
auto.fit<-auto.arima(tsvariable,seasonal=T) #automatic way to find   
auto.fit

## Series: tsvariable   
## ARIMA(2,1,1)   
##   
## Coefficients:  
## ar1 ar2 ma1  
## -0.3022 -0.4719 -0.9350  
## s.e. 0.0794 0.0787 0.0323  
##   
## sigma^2 estimated as 0.001986: log likelihood=214.55  
## AIC=-421.1 AICc=-420.77 BIC=-409.72

auto.fit<-auto.arima(tsvariable,seasonal=F) #automatic way to find   
auto.fit

## Series: tsvariable   
## ARIMA(2,1,1)   
##   
## Coefficients:  
## ar1 ar2 ma1  
## -0.3022 -0.4719 -0.9350  
## s.e. 0.0794 0.0787 0.0323  
##   
## sigma^2 estimated as 0.001986: log likelihood=214.55  
## AIC=-421.1 AICc=-420.77 BIC=-409.72

Akaike information criterion (AIC) & Bayesian information criterion(BIC) is same for both the tests. Any one can be considered as optimized model. The coefficients can be seen in the output.

#### 3 (c)

Based on the output of auto arima

The order of AR is 2

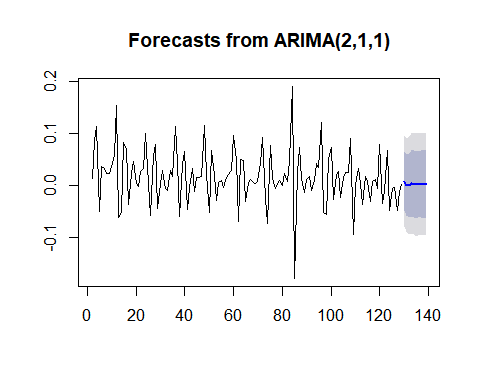
The order of MA is 1

#### 3 (d)

Forecast h=10 step ahead prediction of wage on the plot of the original time series

Visualization the prediction for 10 years

auto.fcast<-forecast(auto.fit,h=10)  
plot(auto.fcast)



#### 4(a)

library(MASS)  
dataset <- read.csv("http://users.stat.ufl.edu/~winner/data/nfl2008\_fga.csv")  
View(dataset)  
dataset.lda<-lda(qtr~togo+ydline+kicker,data=dataset)  
dataset.lda

## Call:  
## lda(qtr ~ togo + ydline + kicker, data = dataset)  
##   
## Prior probabilities of groups:  
## 1 2 3 4 5   
## 0.20636451 0.35969142 0.17550627 0.24590164 0.01253616   
##   
## Group means:  
## togo ydline kicker  
## 1 6.481308 17.22897 19.64486  
## 2 6.973190 19.30027 18.77212  
## 3 6.543956 19.03297 19.96703  
## 4 6.792157 18.53725 20.20000  
## 5 5.923077 19.53846 22.61538  
##   
## Coefficients of linear discriminants:  
## LD1 LD2 LD3  
## togo 0.06665269 0.12498308 0.20996464  
## ydline 0.07726467 -0.07173243 -0.02257770  
## kicker -0.04134867 -0.06009657 0.05013225  
##   
## Proportion of trace:  
## LD1 LD2 LD3   
## 0.615 0.322 0.063

Based on the Proportion of trace, two LDA’s can be chosen as most important as the sum of LD1+LD2 is >90 % of variance so select LD1 and LD2.

No. of LDA = min(# of categories in output-1,# of input variable) = min(5-1,3) = min(4,3)

No. of LDA =3

Linear Discriminant 1 & 2 can be expressed as follows

LD1 = 0.066 \* togo + 0.077 \* ydline - 0.041 \* kicker

LD2 = 0.124 \* togo - 0.071 \* ydline - 0.060 \* kicker

#### 4(b)

PCA

head(dataset)

## GameDate AwayTeam HomeTeam qtr min sec kickteam def down togo kicker  
## 1 20081130 IND CLE 1 47 2 IND CLE 4 11 15  
## 2 20081005 IND HOU 1 54 47 IND HOU 4 3 15  
## 3 20081228 TEN IND 1 45 20 IND TEN 4 3 15  
## 4 20081012 BAL IND 1 45 42 IND BAL 4 1 15  
## 5 20080907 CHI IND 1 50 56 IND CHI 4 21 15  
## 6 20081116 HOU IND 1 50 43 IND HOU 4 7 15  
## ydline name distance homekick kickdiff timerem offscore defscore  
## 1 12 A.Vinatieri 30 0 -3 2822 0 3  
## 2 28 A.Vinatieri 46 0 0 3287 0 0  
## 3 10 A.Vinatieri 28 1 7 2720 7 0  
## 4 19 A.Vinatieri 37 1 14 2742 14 0  
## 5 21 A.Vinatieri 39 1 0 3056 0 0  
## 6 22 A.Vinatieri 40 1 -3 3043 0 3  
## season GOOD Missed Blocked  
## 1 2008 1 0 0  
## 2 2008 1 0 0  
## 3 2008 1 0 0  
## 4 2008 1 0 0  
## 5 2008 1 0 0  
## 6 2008 1 0 0

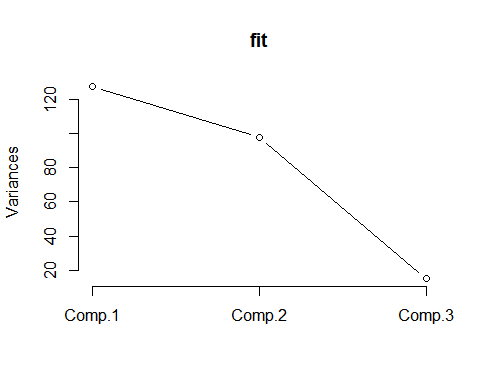
data=cbind(dataset$togo,dataset$ydline,dataset$kicker)  
data.cleansed=na.omit(data)  
fit<-princomp(data.cleansed)  
summary(fit) #print variance accounted for,

## Importance of components:  
## Comp.1 Comp.2 Comp.3  
## Standard deviation 11.2926696 9.8769392 3.90137351  
## Proportion of Variance 0.5306904 0.4059689 0.06334073  
## Cumulative Proportion 0.5306904 0.9366593 1.00000000

loadings(fit) #pc loadings

##   
## Loadings:  
## Comp.1 Comp.2 Comp.3  
## [1,] 0.157 0.988  
## [2,] 0.987 -0.157  
## [3,] -1.000   
##   
## Comp.1 Comp.2 Comp.3  
## SS loadings 1.000 1.000 1.000  
## Proportion Var 0.333 0.333 0.333  
## Cumulative Var 0.333 0.667 1.000

plot(fit,type="lines") #scree plot



This shows that first principal component explains 33.33% variance. Second component explains 66.7% variance. Third component explains 100% variance. So, we select first two components as important principle components involving at least 90% of dataset variation. It also can be answered based on scree plot. Based on the scree plot the second component exceeds 90%. Therefore first 2 components are chosen.