**

**Data and Data Analytics CA**

**Naghma Khan**

**Student ID:10514575**

**Lecturer Name: Shahram Azizi Sazi**

**Question 1**

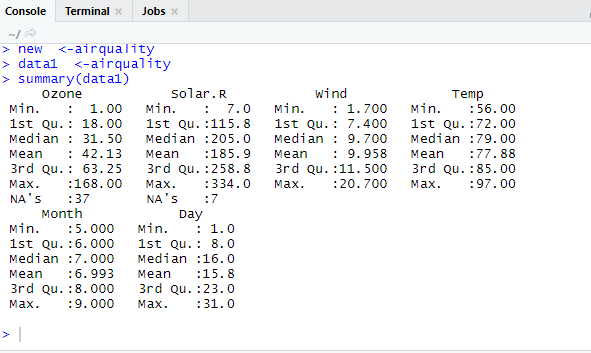
Use in-built dataset ‘airquality’,

a)Explore the general feature of dataset using appropriate R functions.

**Code:**

data1 <-airquality

summary(data1)



General features of a dataset as shown in the above screenshot are Min, Median, Mean and Max values of a dataset provided.

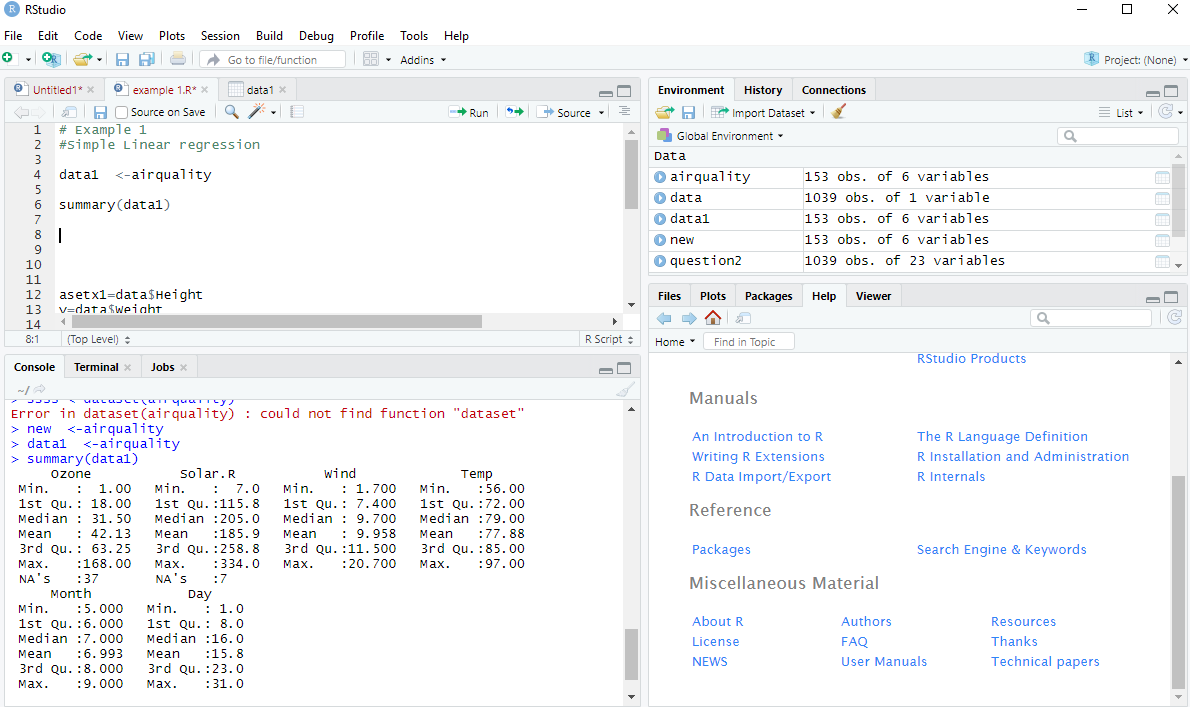
b)Perform data cleansing if required.

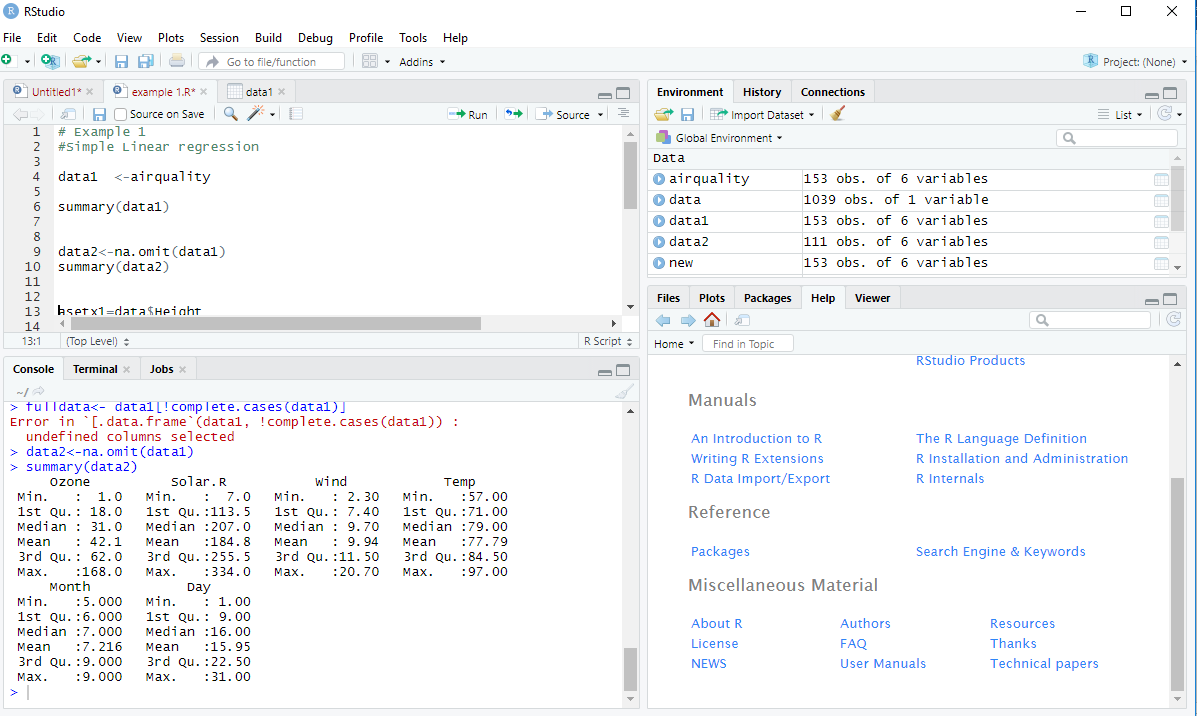
**Code:**

data2<-na.omit(data1)

summary(data2)

attach(data2)





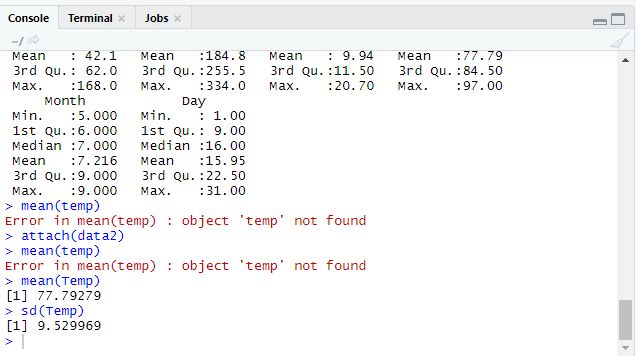
The Data Cleansing has been done by omitting the data and all the NA’s present in the first screenshot have been removed and is shown in the second screenshot. The reason Data needs to be cleaned is when we perform any operation or try to take the mean or median or any other value of the data then the NA’s will be counted and the value might be incorrect.

**c) Consider ‘Temp’ attributes and compute the central and variational measures.**

**Code:**

mean(Temp)

sd(Temp)

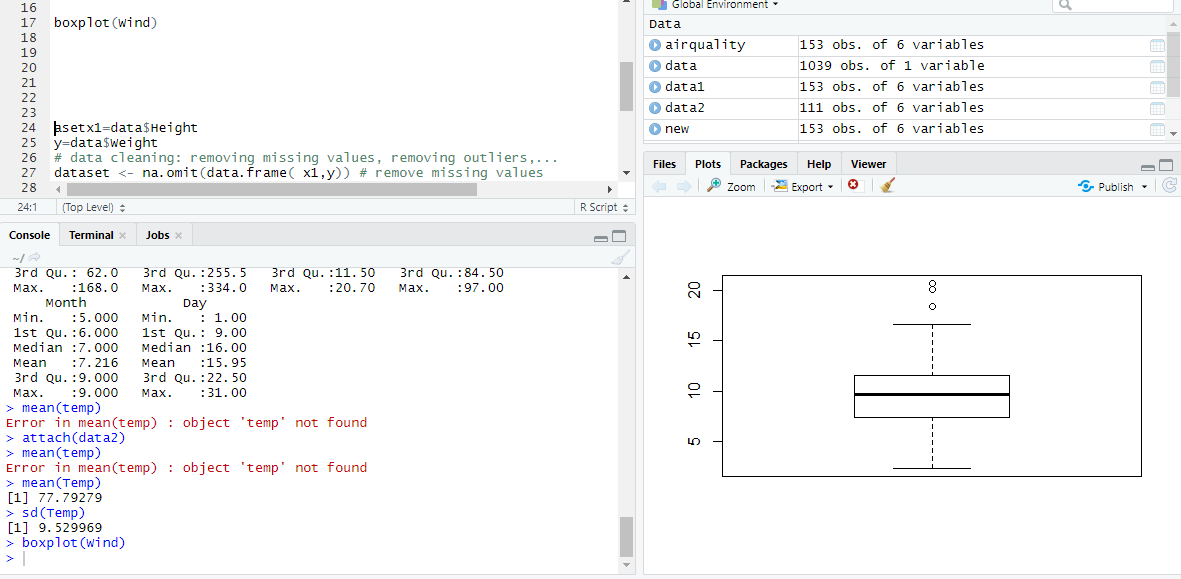


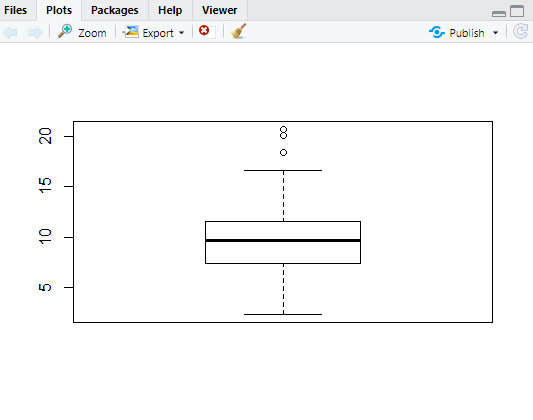
‘Temp’ is a column in a dataset for which we are calculating central and variational measure, which is also called as Mean and sd(Standard Deviation).The values are in the baove screenshot. Mean is a numerical measure of ll the values that are present in the central location. Standard Deviation is to show the amount of variation present in data values.

**d) Apply boxplot technique to detect outlier of ‘wind’ attribute if any.**

**Code:**

boxplot(Wind)





As shown in the above screenshot we have plotted a Boxplot technique-we have plotted it just for wind. There are three outlayers which means they are outside the boundary.

**Question 2**

Use dataset available on <http://users.stat.ufl.edu/~winner/data/nfl2008_fga.csv> , then:

**a)Train the model using 80% of this dataset and suggest an appropriate GLM to model homekick to togo, ydline and kicker variables.**

**Code:**

d<-read.csv('C:/Users/Nagma Khan/Downloads/nfl2008\_fga.csv',header=TRUE)

e<-d[c(10:12,15)]

set.seed(1456)

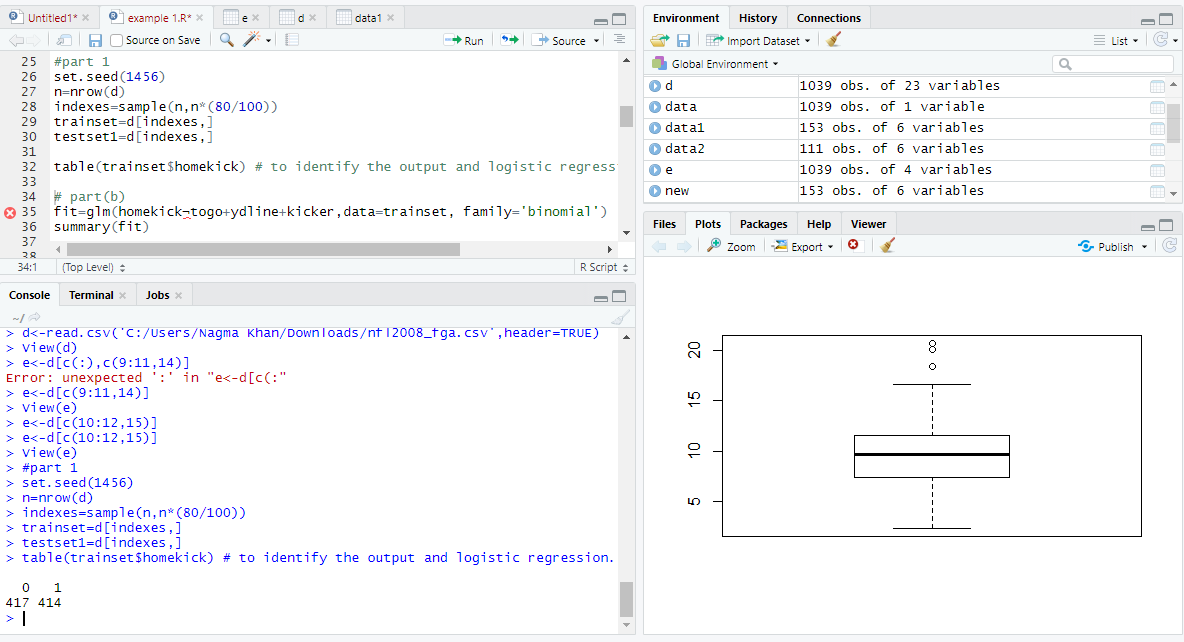
n=nrow(dataset)

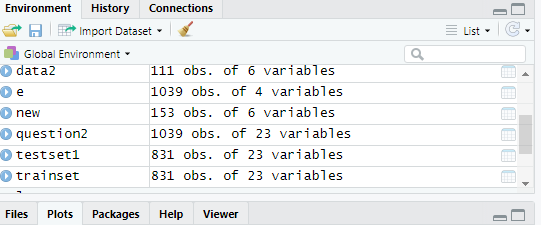
indexes=sample(n,n\*(80/100))

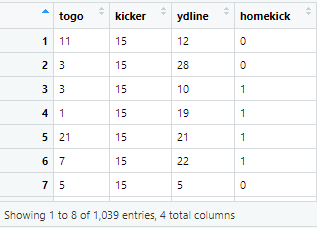
trainset=dataset[index,]

testset1=dataset[index,]

table(trainset$homekick)







We first execute this command, e<-d[c(10:12,15)], and remove all the columns except for **homekick** to **togo, ydline** and **kicker** variables. And we are training the model just to remove 80% of the data, by using this command indexes=sample(n,n\*(80/100)) initially as seen in the screenshot the data was 1039 of 4 variables and in the new testset1 and trainset it is 831 variables. We must take 3 variables and predict it against **homekick.** The appropriate GLM is Logistic Regression as the values of homekick are binary.

0 1

417 414

Values of homekick.

**b) Specify the significant variables on homekick at the level of 𝛼=0.05 and estimate the parameters of your model.**

**Code:**

fit=glm(homekick-togo+ydline+kicker,data=trainset, family='binomial')

summary(fit)

Call:

glm(formula = homekick ~ togo + ydline + kicker, family = "binomial",

data = trainset)

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

By using this code, we are retrieving the Coefficients of all the variables in the dataset. The iterations say it is 4 because we have 4 columns in the dataset. The Pr value for togo is less than 0.05(alpha value). Therefore, only togo is significant, and other variables are not as they are greater than alpha value.

**c) Predict the test dataset using the trained model.**

**Code:**

pred=predict(fit,testset)

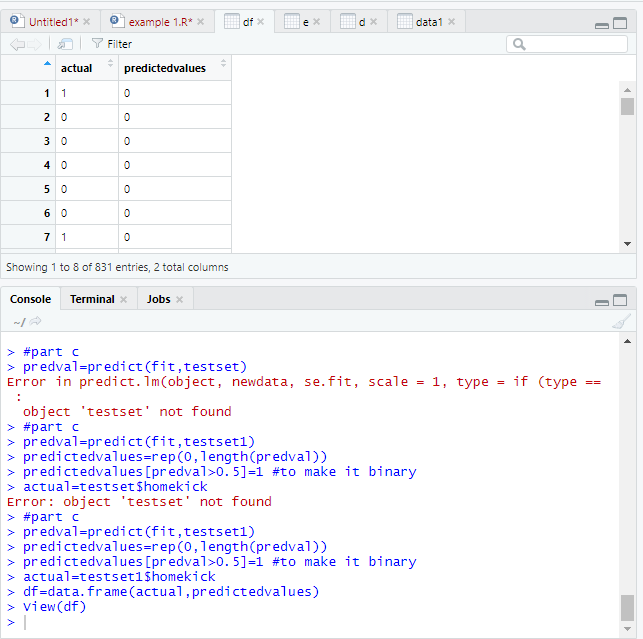
predictedvalues=rep(0,length(pred))

predictedvalues[pred>0.5]=1 #to make it binary

actual=testset$homekick

df=data.frame(actual,predictedvalues)

View(df)



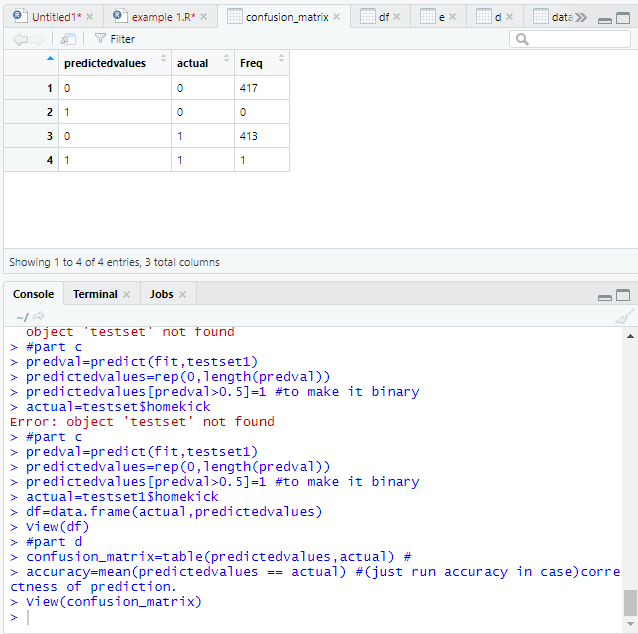
Here we are trying to predict the dataset, actual values and predicted values are calculated.

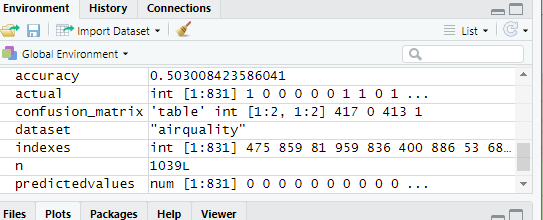
**d) Provide the confusion matrix and obtain the probability of correctness of predictions.**

**Code:**

confusion\_matrix=table(predictedvalues, actual)

accuracy=mean(predictedvalues == actual





Confusion matrix is been calculated here, so this matrix is to evaluate true and false of any data set or data values provided. Confusion matrix values are provided in the above screenshot, and we are calculating accuracy just in case of correctness of prediction.

**Question 3**

Use dataset available on <http://www.stat.ufl.edu/~winner/data/wage_cpi.csv> , apply time series analysis, consider ‘***wage’*** as your time series variable:

**a)Validate the assumptions using graphical visualization.**

**Code:**

Data3<-read.csv('C:/Users/Nagma Khan/Downloads/Copy of wage\_cpi.csv',header=TRUE)

View(Data3)

library(forecast)

head(Data3)

View(Data3)

str(Data3)

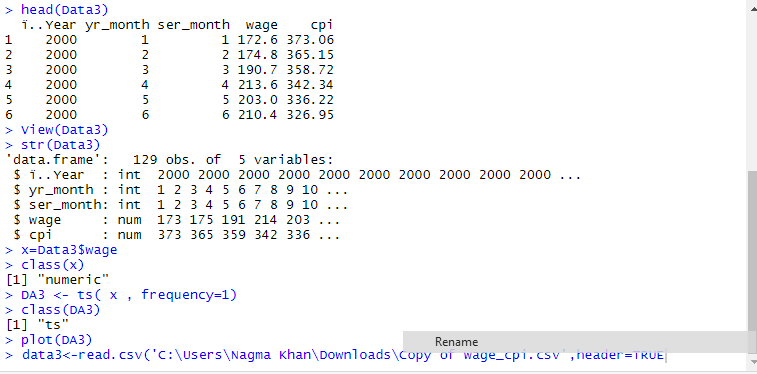
x=Data3$wage

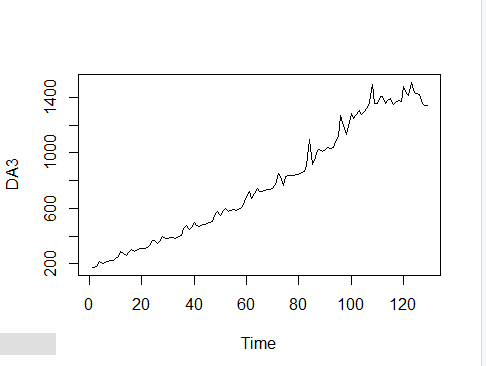
class(x)

DA3 <- ts( x , frequency=1)

class(DA3)

plot(DA3)

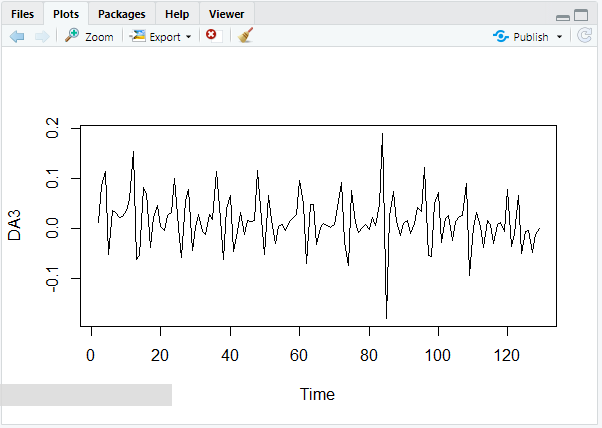




#non-stationary in mean

DA3=diff(log(DA3)) # to make ts stationary in mean and variance

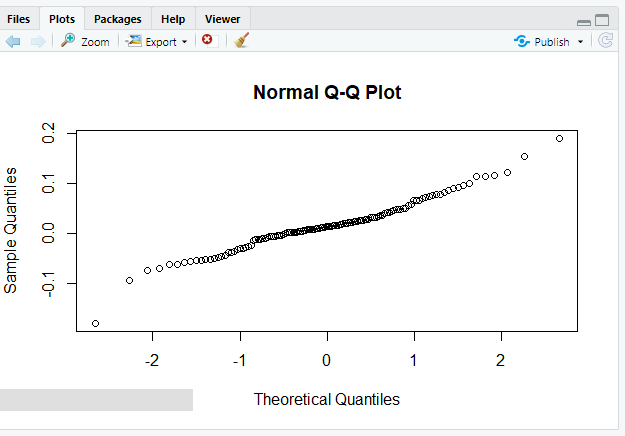
plot(DA3)



#log() for variance

ts(DA3)

qqnorm(DA3)



We need to check which type of graph is most suited for visualization.

Stationary -plot

Normality -qqnorm

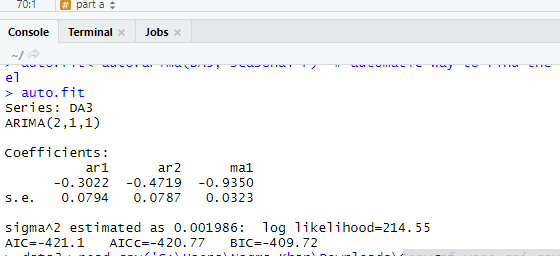
The Time series is normal as the graph is straight.

**b) Fit the optimized model for ‘*wage’* and provide the coefficient estimates for the fitted model.**

**Code:**

auto.fit<-auto.arima(DA3, seasonal=F)

auto.fit



This is the automatic way to find out the optimized model, so by using auto Arima we choose the lowest BIC and AIC should be the best choice.

**c)What is the estimated order for AR and MA?**

ARIMA(2,1,1)

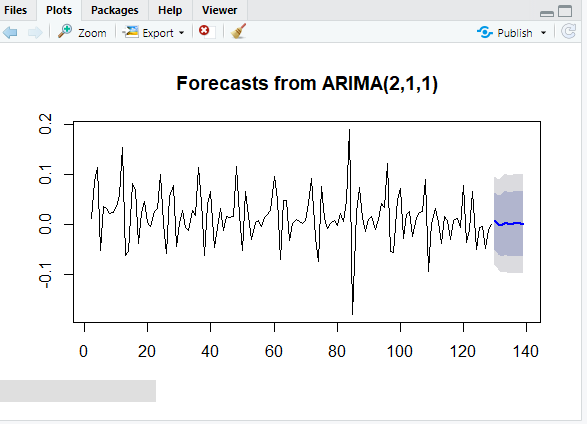
P=2,d=1,q=1

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

**Code:**

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

plot(auto.fcast)



Here we are trying to predict the next 10 future values for wages.

This is also an automatic way to find out fit.

### Question 4

Use dataset available on <http://users.stat.ufl.edu/~winner/data/nfl2008_fga.csv>

**a)Use LDA to classify the dataset into few classes so that at least 90% of information of dataset is explained through new classification. (Hint: model the variable “qtr” to variables “togo”, “kicker”, and “ydline”). How many LDs do you choose? Explain the reason.**

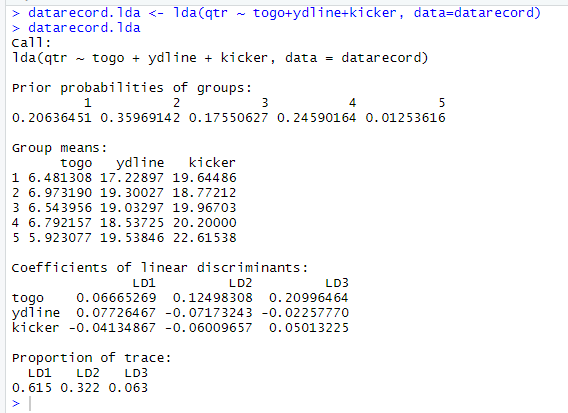
**Code:**

library(MASS)

d<-read.csv('C:/Users/Nagma Khan/Downloads/nfl2008\_fga.csv',header=TRUE)

View(d)

d.lda<-lda(qtr~togo+ydline+kicker,data=dataset)



As shown in the screenshots above, the coefficients of linear discriminants for togo, ydline , kicker are 0.615,0.322 and 0.063. The sum of 0.615+0.322 the values are greater than 0.9 therefore, LD1 and LD2 is selected.

**b)Apply PCA and identify the important principle components involving at least 90% of dataset variation. Explain your decision strategy? Plot principle components versus their variance (Hint: to sketch the plot use the Scree plot).**

**Code:**

head(d)

data=cbind(dataset$togo,dataset$ydline,dataset$kicker)

data1=na.omit(data)

fit<-princomp(data1)

summary(fit) #print variance accounted for

loadings(fit) #pc loadings

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

