

Corso di studio in Amministrazione Aziendale

Corso “BIG DATA & ANALYTICS” 2020/2021

Prof. Roberto Marazzato

Relazione operativa:

INVESTING

“Analysis stock prices and banking crises”

Redatta da:

Giovanni Vignola – Matricola 892615



**Abstract**

Analisi azionaria del portafoglio S&P500, che comprende le 500 aziende americane più importanti, effettuando degli esercizi con modelli statistici e seguendo le richieste degli esercizi speculari assegnati. Si vuole inoltre esaminare se ci sono delle differenze nelle combinazioni di esportazioni e riserve, con eventuali dipendenze dallo stato di crisi o no del settore bancario. Verranno imposte delle variabili che andranno a specificare il tipo di situazione in cui si trova la banca, insieme alle due combinazioni che potrebbero inficiare o no nella stessa.

Keywords: Missing values. Stocks. Modello lineare. Gaussiana. Correlogramma. Arima. Residuals. Outliers. Correlazione. Manova.

**Indice**

1. Analysis stock prices
   * Missing values
   * Prezzi di chiusura del portafoglio azionario S&P500
   * Rendimento medio giornaliero
   * Modello di regressione lineare semplice
   * Test di normalità
   * Media mobile
   * Correlogramma
   * Modello Arima
2. Banking crises

* Verifica preventiva
* Modello univariato e multivariato
* Modello di regressione lineare semplice
* Test di Pearson
* Manova

1. Bibliografia
2. Materiali utilizzati
3. **Analysis stock prices**

In prima istanza si vogliono analizzare i prezzi delle azioni, il primo passo è caricare il dataset (dopo aver cercato quello più adeguato) in locale, tramite la funzione *read.csv()*, inoltre si effettua una verifica del caricamento stampando le prime 9 righe, e tutte le colonne *head (...,9)*.

SP500Stock <- read.csv("C:/Users/vgiov/Downloads/SP 500 Stock Prices 2014-2017.csv", sep = "," ,header = T)  
head(SP500Stock,9)

## symbol date open high low close volume  
## 1 AAL 2014-01-02 25.0700 25.8200 25.0600 25.3600 8998943  
## 2 AAPL 2014-01-02 79.3828 79.5756 78.8601 79.0185 58791957  
## 3 AAP 2014-01-02 110.3600 111.8800 109.2900 109.7400 542711  
## 4 ABBV 2014-01-02 52.1200 52.3300 51.5200 51.9800 4569061  
## 5 ABC 2014-01-02 70.1100 70.2300 69.4800 69.8900 1148391  
## 6 ABT 2014-01-02 38.0900 38.4000 38.0000 38.2300 4967472  
## 7 ACN 2014-01-02 81.5000 81.9200 81.0900 81.1300 2405384  
## 8 ADBE 2014-01-02 59.0600 59.5300 58.9400 59.2900 2746370  
## 9 ADI 2014-01-02 49.5200 49.7500 49.0400 49.2800 2799092

Un passaggio fondamentale da eseguire quando si scaricano dei dataset, è il controllo dei missing values *is.na()*.

sum(is.na(SP500Stock))

## [1] 27

Successivamente, si può passare all’eventuale eliminazione degli stessi *na.omit()*, verificando che il processo sia avvenuto con successo.

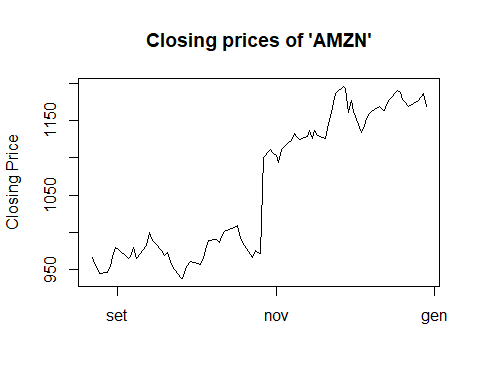
SP500Stock <- na.omit(SP500Stock)  
#check all  
summary(SP500Stock)

## symbol date open high   
## Length:497461 Length:497461 Min. : 1.62 Min. : 1.69   
## Class :character Class :character 1st Qu.: 41.69 1st Qu.: 42.09   
## Mode :character Mode :character Median : 64.97 Median : 65.56   
## Mean : 86.35 Mean : 87.13   
## 3rd Qu.: 98.41 3rd Qu.: 99.23   
## Max. :2044.00 Max. :2067.99   
## low close volume   
## Min. : 1.50 Min. : 1.59 Min. : 101   
## 1st Qu.: 41.28 1st Qu.: 41.70 1st Qu.: 1080183   
## Median : 64.36 Median : 64.98 Median : 2085013   
## Mean : 85.55 Mean : 86.37 Mean : 4253695   
## 3rd Qu.: 97.58 3rd Qu.: 98.42 3rd Qu.: 4271999   
## Max. :2035.11 Max. :2049.00 Max. :618237630

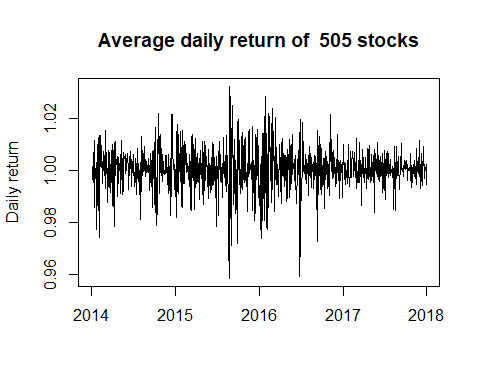
Ora, per iniziare l’analisi, decidiamo di concentrare l’attenzione sui prezzi di chiusa del portafoglio S&P500 usando *reshape ()* per raggruppare i dati del dataset.

stock\_closeSP500 <- reshape(SP500Stock[c("symbol", "date", "close")], timevar = "symbol", idvar = "date", direction = "wide")  
colnames(stock\_closeSP500) <- c("date" , as.character(unique(SP500Stock$symbol)))  
stock\_closeSP500$date <- as.Date(stock\_closeSP500$date)  
stock\_closeSP500 <- stock\_closeSP500[with(stock\_closeSP500, order(date)),]

Dopo aver raggruppato i dati, l’idea è quella di estrapolare e visualizzare in un grafico a linea *type= 'l'*, dell’andamento del prezzo di chiusura di Amazon degli ultimi 90 giorni.

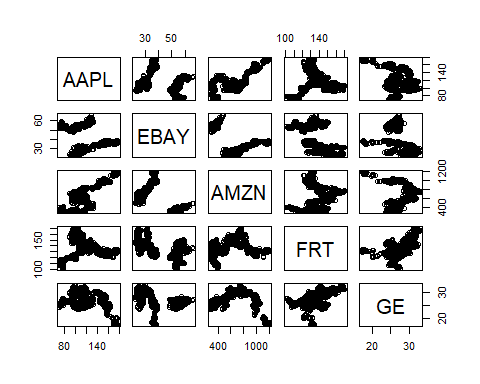
Last90D <- stock\_closeSP500[(NROW(stock\_closeSP500)-90):NROW(stock\_closeSP500),]  
plot(type = 'l', x = Last90D$date, y = Last90D$AMZN, ylab= "Closing Price", xlab="", main= "Closing prices of 'AMZN'" )

Adesso vogliamo calcolare e tracciare il rendimento medio giornaliero del portafoglio prescelto, supponendo di avere quantità eguale di azioni in un dataframe [-1].

stock\_returnSP500 <- data.frame(date = stock\_closeSP500$date[-1], sapply(stock\_closeSP500[-1], function(x){  
 diff(x)/x[-length(x)]+1  
}))  
  
stock\_returnSP500$date <- as.Date(stock\_returnSP500$date)  
  
plot(x= stock\_returnSP500$date, y= rowMeans(stock\_returnSP500[-1], na.rm = T), type = c("l"),  
 xlab= "", ylab="Daily return", main= paste("Average daily return of ",NCOL(stock\_returnSP500)-1, "stocks"))

Generiamo un *pairwise scatter* *plots* per comparazione di cinque azioni a nostra scelta, all’interno del portafoglio, per una prima generica analisi.

pairs(stock\_closeSP500[, c("AAPL", "EBAY", "AMZN", "FRT", "GE")])

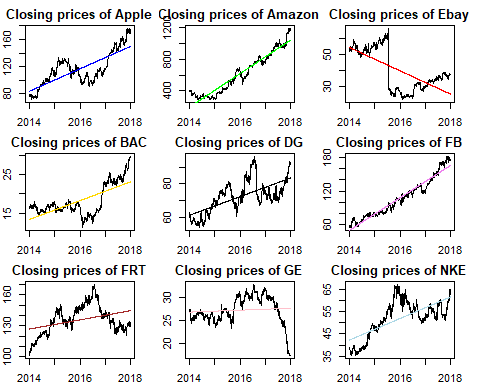


A questo punto possiamo analizzare delle azioni del portafoglio attraverso il modello di regressione lineare (semplice), questo poiché la statistica inferenziale propone di valutare se esiste una relazione fra le variabili e se questa relazione è significativa. In questo caso abbiamo selezionato le seguenti azioni:

* Apple
* Amazon
* Ebay
* Back of American Corporation
* Dollar General Corporation
* Facebook
* Federal Investiment Trust
* General Electric Company
* Nike

tracceremo il tipo di relazione con la retta di regressione *lines()* tra il prezzo di chiusura con le medesime date, tutte nello stesso grafico tramite la modifica del *par()*.

par(mfrow = c(3,3))  
par(mar= c(2,2,2,2))  
  
#Apple  
plot(type = "l", x = stock\_closeSP500$date, y = stock\_closeSP500$AAPL, ylab= "Closing price", xlab = "", main= "Closing prices of Apple")  
lines(x= stock\_closeSP500$date, y = lm(stock\_closeSP500$AAPL ~ stock\_closeSP500$date)$fit, col= "blue")  
#Amazon  
plot(type= "l", x = stock\_closeSP500$date, y=stock\_closeSP500$AMZN, ylab= "Closing price", xlab ="", main= "Closing prices of Amazon")  
lines(x=stock\_closeSP500$date, y = lm(stock\_closeSP500$AMZN ~ stock\_closeSP500$date)$fit, col = "green")  
#Ebay  
plot(type= "l", x=stock\_closeSP500$date, y=stock\_closeSP500$EBAY, ylab="Closing price", xlab="", main= "Closing prices of Ebay")  
lines(x= stock\_closeSP500$date, y= lm(stock\_closeSP500$EBAY ~ stock\_closeSP500$date)$fitted.values, col = "red")  
#Back of american Corporation  
plot(type= "l", x= stock\_closeSP500$date, y = stock\_closeSP500$BAC ,ylab= "Closing price", xlab="", main = "Closing prices of BAC")  
lines(x= stock\_closeSP500$date, y = lm(stock\_closeSP500$BAC ~ stock\_closeSP500$date)$fit, col= "gold")  
#Dollar General Corporation  
plot(type = "l", x = stock\_closeSP500$date, y = stock\_closeSP500$DG, ylab= "Closing prince", xlab="", main= "Closing prices of DG")  
lines(x= stock\_closeSP500$date, y = lm(stock\_closeSP500$DG~ stock\_closeSP500$date)$fit, col= "black")  
#Facebook  
plot(type = "l", x = stock\_closeSP500$date, y = stock\_closeSP500$FB, ylab= "Closing prince", xlab="", main= "Closing prices of FB")  
lines(x= stock\_closeSP500$date, y = lm(stock\_closeSP500$FB~ stock\_closeSP500$date)$fit, col= "violet")  
#Federal Investiment Trust  
plot(type = "l", x = stock\_closeSP500$date, y = stock\_closeSP500$FRT, ylab= "Closing prince", xlab="", main= "Closing prices of FRT")  
lines(x= stock\_closeSP500$date, y = lm(stock\_closeSP500$FRT~ stock\_closeSP500$date)$fit, col= "brown")  
#General Electric Company  
plot(type = "l", x = stock\_closeSP500$date, y = stock\_closeSP500$GE, ylab= "Closing prince", xlab="", main= "Closing prices of GE")  
lines(x= stock\_closeSP500$date, y = lm(stock\_closeSP500$GE~ stock\_closeSP500$date)$fit, col= "pink")  
#Nike  
plot(type = "l", x = stock\_closeSP500$date, y = stock\_closeSP500$NKE, ylab= "Closing prince", xlab="", main= "Closing prices of NKE")  
lines(x= stock\_closeSP500$date, y = lm(stock\_closeSP500$NKE~ stock\_closeSP500$date)$fit, col= "light blue")



par(mfrow= c(1,1))

Per verificare a quanto ammonta la variazione del prezzo di Ebay spiegata tramite il modello lineare semplice *lm()*, all’adattamento con il rendimento giornaliero .

summary(lm(stock\_closeSP500$EBAY~ stock\_closeSP500$date ))$r.squared

## [1] 0.4253385

Approssimativamente è spiegata dal 42,5%.

Impostiamo un modello di regressione lineare semplice del prezzo di chiusura di Apple nel 2017.

Ricordando che questo modello di basa su due variabili di cui abbiamo campioni accoppiati, variabili e , la è indipendente, la è quella dipendente e ipotizziamo che ci sia una relazione lineare tra le due. Ora verificheremo se è statisticamente significante attraverso il metodo *p-value* di α=0.05.

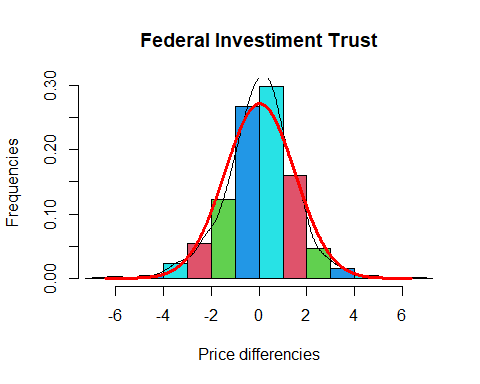
y <- subset(stock\_closeSP500, stock\_closeSP500$date >= '2017-01-01')$AAPL  
x <- subset(stock\_closeSP500, stock\_closeSP500$date >= '2017-01-01')$date  
summary(lm(y ~ x))

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.329 -4.016 1.133 4.054 12.164   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.098e+03 5.834e+01 -35.96 <2e-16 \*\*\*  
## x 1.296e-01 3.362e-03 38.54 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.555 on 249 degrees of freedom  
## Multiple R-squared: 0.8564, Adjusted R-squared: 0.8559   
## F-statistic: 1485 on 1 and 249 DF, p-value: < 2.2e-16

Dall’analisi risulta la presenza di una significanza statistica, trovata con il metodo del p-value.

Un altro fattore da tener presente riguarda la distribuzione delle azioni, quindi bisogna stabilire se si comportano come delle normali (Gaussiana μ=0; σ=1), o diversamente in modo tale da poter approfondire l’analisi. Per essere più precisi tracciamo anche la linea dell’andamento stesso. Questa volta utilizziamo l’azione FRT, perché da una preliminare osservazione è risultata quella con le caratteristiche idonee al nostro lavoro.

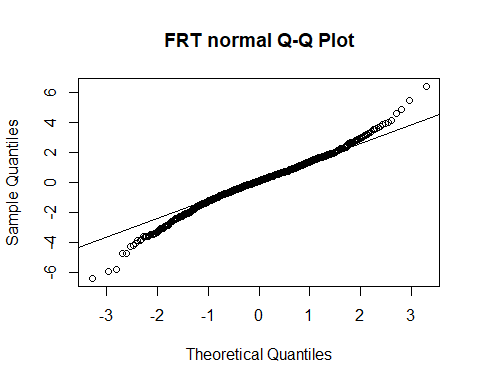
#Distribution  
FederalI.norm <- diff(stock\_closeSP500$FRT)  
hist(FederalI.norm, probability = "T", col = 2:5, ylab= "Frequencies", xlab = "Price differencies", main = "Federal Investiment Trust")  
lines(density(FederalI.norm))  
  
#Normality  
mu <- mean(FederalI.norm)  
sigma <- sd(FederalI.norm)  
  
x <- seq(min(FederalI.norm), max(FederalI.norm), length = length(FederalI.norm))  
y <- dnorm(x, mu, sigma)  
lines(x, y, lwd = 3, col= "red")



Vogliamo vedere l’interpretazione grafica dei quantili di una variabile a confronto con quelli della distribuzione normale di una variabile teorica, con la stessa media e la stessa deviazione standard, prenderemo il prezzo di chiusura dell’azione FRT, usando la *normal Q-Q Plot*, infatti precedentemente sono stati ricavati la μ e la σ della .

Nell'asse delle X saranno rappresentati i quartili della distribuzione normale (i valori attesi in caso di normalità), su quello delle Y i valori campionari (quindi quelli di FRT).

qqnorm(FederalI.norm, main = "FRT normal Q-Q Plot")  
qqline(FederalI.norm)



Il test di Shapiro-Wilk è considerato uno dei test più potenti per la verifica della normalità, soprattutto per piccoli campioni. La verifica della normalità avviene confrontando due stimatori alternativi della varianza σ2: uno stimatore non parametrico e lo stimatore parametrico, quindi quello campionario.

Il test Kolmogorov-Smirnov, invece, permette di stabilire il grado di somiglianza di due distribuzioni e stabilisce a quale livello di significatività le due distribuzioni possono considerarsi diverse, basandosi sulla frequenza cumulativa relativa dei dati.

Si può proseguire con l’analisi eseguendo i due tipi di test.

shapiro.test(FederalI.norm)

##   
## Shapiro-Wilk normality test  
##   
## data: FederalI.norm  
## W = 0.98504, p-value = 1.256e-08

ks.test(FederalI.norm, "pnorm", mean(FederalI.norm), sd(FederalI.norm))

## Warning in ks.test(FederalI.norm, "pnorm", mean(FederalI.norm),  
## sd(FederalI.norm)): ties should not be present for the Kolmogorov-Smirnov test

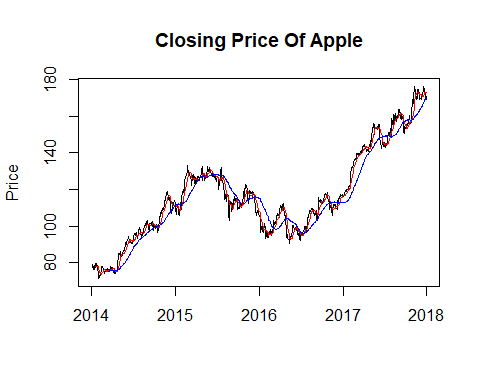
##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: FederalI.norm  
## D = 0.043996, p-value = 0.04071  
## alternative hypothesis: two-sided

Il p-value è decisamente basso rispetto ai livelli di significatività a cui di solito si fa riferimento: ciò ci fa propendere per l’ipotesi alternativa, ovvero la non normalità della distribuzione.

Oltre al metodo prettamente analitico per la stima del trend, ci sono altre metodologie più semplici da utilizzare per detrendizzare una serie temporale. Noi ci focalizzeremo sulle medie mobili, con le quali si pone il problema dell’esatta determinazione del numero dei termini da usare. In linea generale, il trend può essere stimato con un’opportuna ponderazione dei valori della serie:

Dovendo tracciare il grafico della media mobile della stima di 12 e 50 giorni dei prezzi di chiusura di Apple, è necessario fare un *plot()* dei prezzi di chiusura, e di seguito impostare il *filter()* fissando il parametro corrispondente ad una serie mensile (esempio “1/12” per “12” volte per un valore di “a” come dalla formula di sopra ), e il metodo della *convolution* che consiste in un'operazione tra due funzioni di una variabile, e quindi nell'integrare il prodotto tra la prima e la seconda traslata di un certo valore.

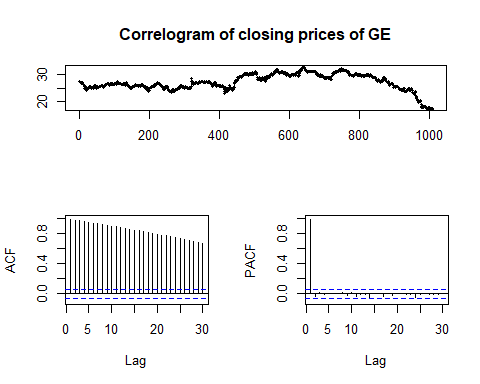
#1.Closing prices of APPL  
plot(x = stock\_closeSP500$date, y = stock\_closeSP500$AAPL, type= "l",  
 main= "Closing Price Of Apple", xlab= "", ylab="Price")  
#2. 12 days moving average of the closing prices of APPL  
lines(x= stock\_closeSP500$date , y= filter(stock\_closeSP500$AAPL,   
 filter = rep(1/12,12), method = "convolution", sides = 1), col= "red")  
#3.50 days moving average of the closing prices of APPL  
lines(x= stock\_closeSP500$date, y= filter(stock\_closeSP500$AAPL,   
 filter = rep(1/50, 50),method = "convolution", sides = 1), col= "blue")

In un’analisi di questo tipo si può avere un fenomeno di autocorrelazione temporale, causato dall'inerzia o dalla stabilità dei valori osservati, quindi ogni valore è influenzato da quello precedente e determina in parte rilevante quello successivo.

Per stabilire se una serie presenta autocorrelazione può essere quella di tracciare il correlogramma con la funzione *acf()*. Se si verificasse l’assenza di autocorrelazione, la distribuzione asintotica della stima del coefficiente di autocorrelazione sarebbe di tipo normale ed avremmo una banda di confidenza del tipo:

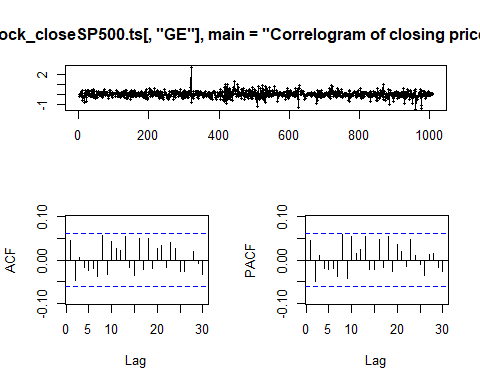
i valori esterni a tale intervallo indicano che vi è la presenza di autocorrelazione significativa.

Ora, per creare una serie storica, utilizzeremo la funzione *ts()*, dove la difficoltà è solo nell’ impostare quando farla iniziare e finire, così poi da riuscire a visualizzare il prezzo di chiusura di GE in un correlogramma.

stock\_closeSP500.ts <- ts(stock\_closeSP500[,-1])  
tsdisplay(stock\_closeSP500.ts[, "GE"],   
 main="Correlogram of closing prices of GE")

Le linee tratteggiate di colore azzurro indicano la banda di confidenza pari al 95%. Con il variare del *lag* temporale i coefficienti di autocorrelazione dei residui risultano essere tutti esterni alla banda di confidenza, indicando quindi autocorrelazione serie e non stazionaria.

Invece se effettuiamo la differenziazione *diff()* della serie temporale.

tsdisplay(diff(stock\_closeSP500.ts[,"GE"],   
 main = "Correlogram of closing prices of GE"))

In questo caso, come vediamo, la serie temporale è stazionaria senza autocorrelazione di serie.

I modelli ARIMA (autoregressivi integrati a media mobile) di Box e Jenkins presuppongono che tra due osservazioni di una serie, quello che altera il livello della stessa è il cosiddetto disturbo.

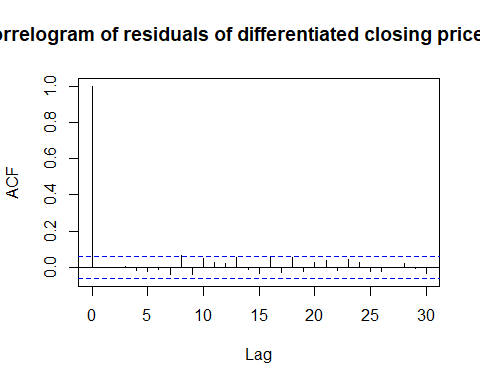
Se la serie non è stazionaria (la media e la varianza non sono costanti nel tempo) viene integrata a livello 1 o 2, dopo aver eseguito un'eventuale trasformazione dei dati (solitamente quella logaritmica). Quindi questo tipo procedura è di tipo iterativo e serve per identificare la stima di una serie storica, che sia adatta e che rappresenti il processo generatore della serie osservata.

Per trovare il modello Arima più adatto ai prezzi di chiusura di GE utilizziamo *auto.arima()*.

fit.GE <- auto.arima(diff(stock\_closeSP500.ts[, "GE"]), stepwise = F)  
fit.GE

## Series: diff(stock\_closeSP500.ts[, "GE"])   
## ARIMA(0,0,2) with zero mean   
##   
## Coefficients:  
## ma1 ma2  
## 0.0486 -0.0483  
## s.e. 0.0315 0.0319  
##   
## sigma^2 estimated as 0.09723: log likelihood=-254.14  
## AIC=514.29 AICc=514.31 BIC=529.03

Proviamo a scoprirne il periodo utilizzando la funzione di autocorrelazione. Il secondo massimo avviene in corrispondenza del valore del periodo, in modo da non agire per tentativi. Di conseguenza visualizzeremo i residui del modello.

acf(residuals(fit.GE), main= "Correlogram of residuals of differentiated closing price of GE")

Nell’analisi delle azioni, spesso conviene fare previsioni sull’andamento dei prezzi futuri. In questo caso utilizzeremo l’80% della serie temporale dei prezzi di GE, per effettuare una previsione dei restanti 20%.

Avvalendoci dei dati ottenuti tramite il modello Arima (0,0,2), si riesce a inizializzare la previsione.

effective <- stock\_closeSP500.ts[1:(0.8 \*length(stock\_closeSP500.ts[, "GE"]))]  
prediction <- predict(arima(effective, order = c(0, 0 , 2 )),  
  
 n.ahead = (0.2 \* length(stock\_closeSP500.ts[,"GE"])))$pred

Fondamentale è effettuare un controllo sull’accuratezza della previsione stimata, con la funzione *accuracy()* si può eseguire un confronto fra i test.

try <- stock\_closeSP500.ts[(0.8\* length(stock\_closeSP500.ts[,"GE"])+1):length(stock\_closeSP500.ts[,"GE"])]  
  
Ac.pack <- accuracy(prediction,try)[2]  
tdev<- sd(try)  
  
ifelse(Ac.pack <= tdev, "Yes", "No")

## [1] "No"

1. **Banking crises**

La seconda parte verte sull’analisi delle banche in crisi, delle differenze nelle combinazioni di esportazioni e riserve, con eventuali dipendenze dallo stato di crisi o no del settore bancario. Anche in questo caso si procede con il caricamento del dataset in locale, ma non si effettua il controllo dei missing values a causa dell’utilizzo di un'altra funzione.

bank <- read.csv("C:/Users/vgiov/Downloads/banking-crises-data.csv", sep = ",", header = T)  
head(bank)

## crisis export reserves  
## 1 No 24 130  
## 2 No 16 220  
## 3 No 11 190  
## 4 Yes 18 100  
## 5 No 6 130  
## 6 No 20 210

Per poter verificare i dati delle banche e stabilire i criteri, è necessario adottare l’analisi di MANOVA (scopo di identificare le variabili dipendenti specifiche che hanno contribuito all’effetto globale), per condurla sono necessari almeno due campioni, quindi dovremo verificare la loro esistenza.

aggregate(.~crisis , data= bank , FUN = function(x){sum(!is.na(x))}, na.action = na.pass)

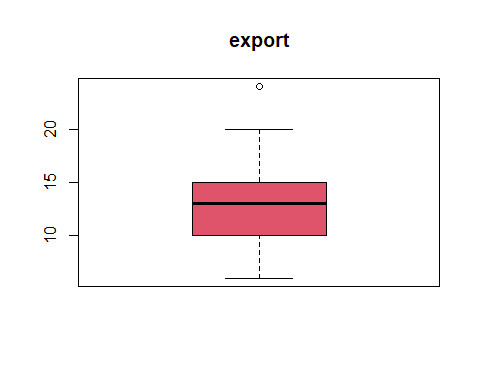
## crisis export reserves  
## 1 No 67 67  
## 2 Yes 12 12

Successivamente si vuole verificare la presenza degli outliers tramite un’analisi esplorativa dei dati, utilizzando un modello univariato (concentrazione su un singolo attributo alla volta, matrice n x 1) e quello multivariato (considerazione di tutti gli attributi numerici d X1, X2, ..., Xd, matrice n x d).

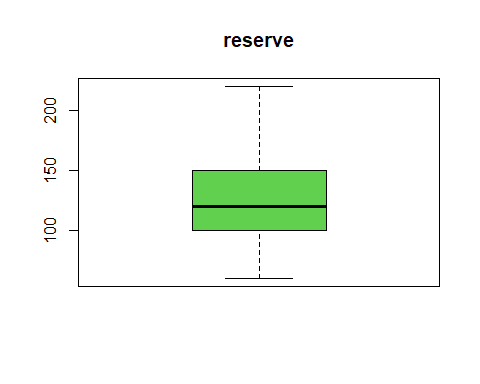
Si procede prima eseguendo delle funzioni grafiche per avere una percezione del comportamento dei dati, poi tracciando un *boxplot()* e ancora un *mvn()*.

* Univariati outliers:

boxplot(bank$export, main = "export", col = 2:5)$out



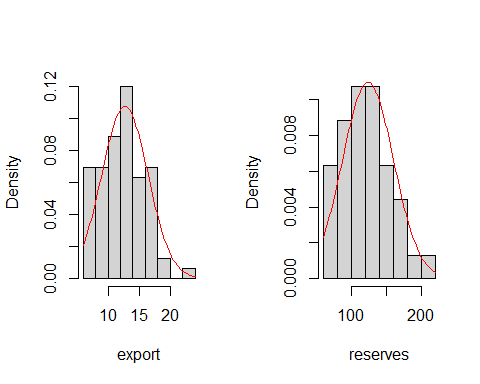
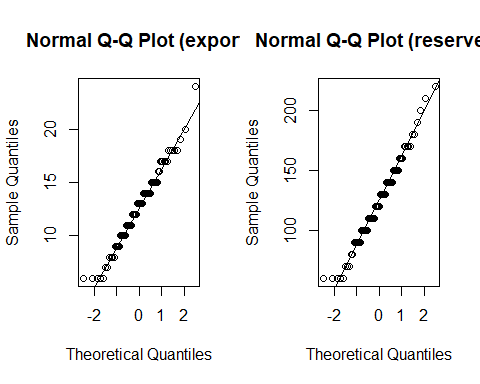
## [1] 24

boxplot(bank$reserves, main= "reserve", col= 3:7)$out

## numeric(0)

result <- mvn(data = bank[,2:3], mvnTest = "royston", univariatePlot = "qqplot", showOutliers = T)

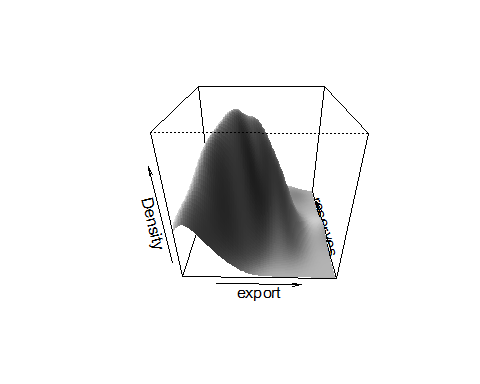
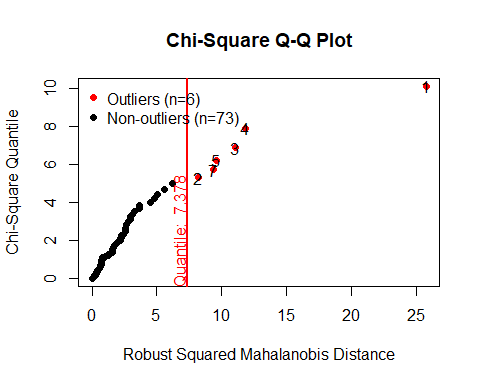
result <- mvn(data = bank[,2:3], mvnTest = "royston", univariatePlot = "histogram",showOutliers = T)



* Multivariati outliers:

In questo caso preferiamo fare anche un “*persp*” per vedere una trama prospettica.

result <- mvn(bank[,2:3], mvnTest = "royston", multivariatePlot = "persp", multivariateOutlierMethod = "quan")



Come si può vedere ci sono sia univariati che multivariati outliers.

A questo punto si stima la normalità univariata e multivariata delle variabili dipendenti. Verificheremo tutti i dati perché tramite i test *mvn*, si possono ottenere risultati differenti. A causa di ciò può essere utile esaminare i grafici *MVN* insieme ai test di ipotesi per avere una prospettiva un po’ più chiara.

Ricordando che “*mardia*” viene utilizzata per calcolare la multivariata di coefficienti asimmetria e curtosi e la loro corrispondente significatività statistica (può essere utilizzata anche per calcolare la versione corretta del coefficiente di asimmetria per campioni n <20), “*hz*” e “*royston*” la normalità delle distribuzioni, tutte e tre hanno un livello di significanza del 5%, questo per poi sottoporre a verifica l’ipotesi nulla della MANOVA.

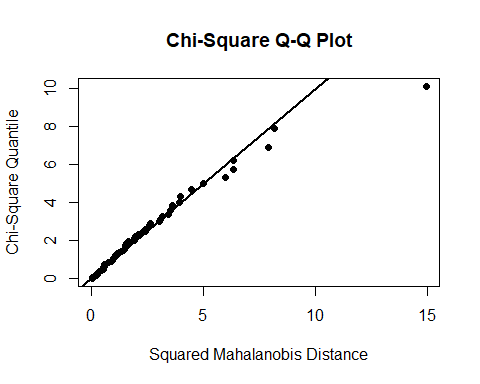
mvn(data = bank[2:3], mvnTest ="mardia", multivariatePlot = "qq")

## $multivariateNormality  
## Test Statistic p value Result  
## 1 Mardia Skewness 18.0648459678829 0.00119860027864038 NO  
## 2 Mardia Kurtosis 1.52528130643459 0.127188948329479 YES  
## 3 MVN <NA> <NA> NO  
##   
## $univariateNormality  
## Test Variable Statistic p value Normality  
## 1 Shapiro-Wilk export 0.9785 0.2023 YES   
## 2 Shapiro-Wilk reserves 0.9793 0.2278 YES   
##   
## $Descriptives  
## n Mean Std.Dev Median Min Max 25th 75th Skew Kurtosis  
## export 79 12.72152 3.703556 13 6 24 10 15 0.1802891 -0.1335514  
## reserves 79 124.17722 36.289703 120 60 220 100 150 0.2669754 -0.2826640

mvn(data = bank[2:3], mvnTest = "hz", multivariatePlot ="qq")

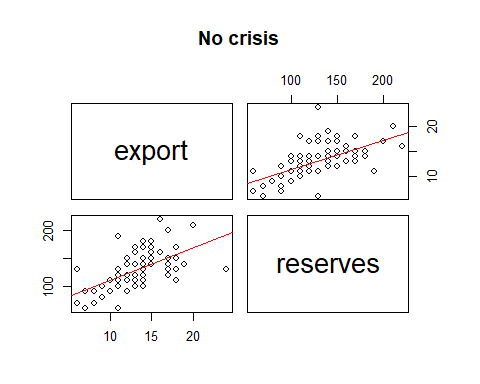
## $multivariateNormality  
## Test HZ p value MVN  
## 1 Henze-Zirkler 0.9748434 0.03951718 NO  
##   
## $univariateNormality  
## Test Variable Statistic p value Normality  
## 1 Shapiro-Wilk export 0.9785 0.2023 YES   
## 2 Shapiro-Wilk reserves 0.9793 0.2278 YES   
##   
## $Descriptives  
## n Mean Std.Dev Median Min Max 25th 75th Skew Kurtosis  
## export 79 12.72152 3.703556 13 6 24 10 15 0.1802891 -0.1335514  
## reserves 79 124.17722 36.289703 120 60 220 100 150 0.2669754 -0.2826640

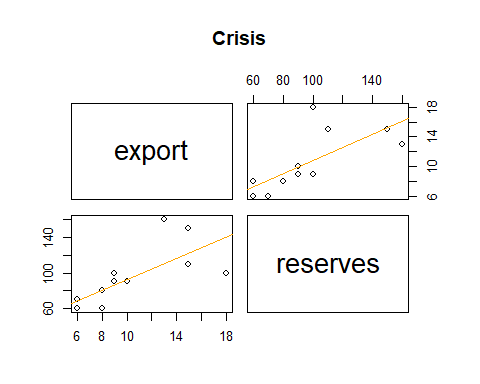
mvn(data = bank[2:3], mvnTest = "royston", multivariatePlot ="qq")

## $multivariateNormality  
## Test H p value MVN  
## 1 Royston 3.071635 0.2144106 YES  
##   
## $univariateNormality  
## Test Variable Statistic p value Normality  
## 1 Shapiro-Wilk export 0.9785 0.2023 YES   
## 2 Shapiro-Wilk reserves 0.9793 0.2278 YES   
##   
## $Descriptives  
## n Mean Std.Dev Median Min Max 25th 75th Skew Kurtosis  
## export 79 12.72152 3.703556 13 6 24 10 15 0.1802891 -0.1335514  
## reserves 79 124.17722 36.289703 120 60 220 100 150 0.2669754 -0.2826640

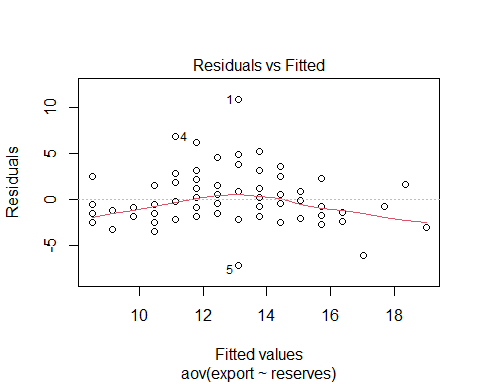
Entrambe sono univariate normalmente distribuite, ma non multivariate.

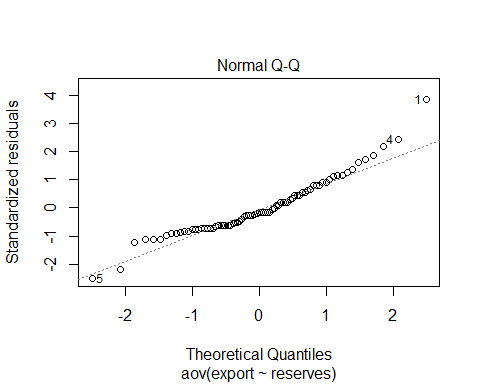
Usando *pairs()* si vuole controllare la linearità tra le variabili dipendenti esportazioni e riserve per ogni categoria di variabili indipendenti e anche la dispersione delle variabili accoppiate.

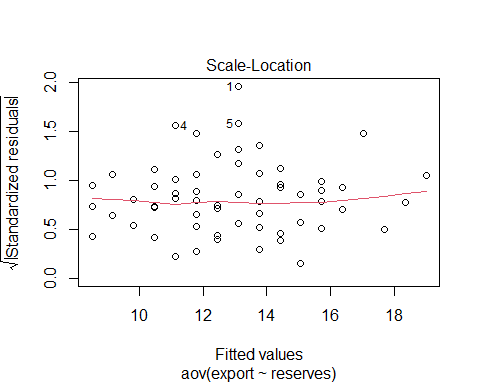
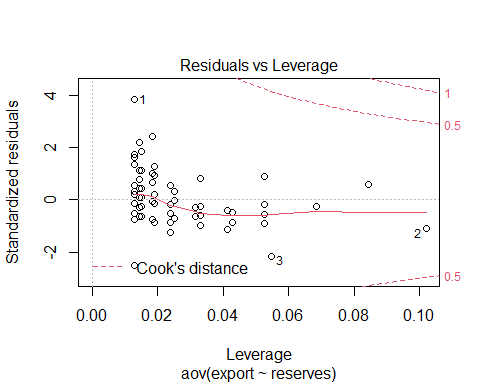
pairs(bank[bank$crisis=="No", 2:3],main= "No crisis",panel= function(x,y,...){   
 points(x,y);   
 abline(lm(y~x),col='red')})

pairs(bank[bank$crisis== "Yes", 2:3], main = "Crisis", panel = function(x,y,...){  
   
 points(x,y);  
 abline(lm(y~x), col= 'orange')})

*Aov()* serve per verificare la presenza di eventuali errori, infatti serve per adattare i dati all’analisi del modello della varianza.

aov.mod <- aov(export ~ reserves, data=bank)  
plot(aov.mod)





Per continuare l’analisi, bisogna calcolare la correlazione tra le due variabili, in modo da stabilire se è in grado di giustificare l’utilizzo della MANOVA.

cor.test(bank$export, bank$reserves)

##   
## Pearson's product-moment correlation  
##   
## data: bank$export and bank$reserves  
## t = 7.3508, df = 77, p-value = 1.79e-10  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4907221 0.7559335  
## sample estimates:  
## cor   
## 0.6421562

Correlazioni non elevate tra le variabili dipendenti (MANOVA), in modo da evitare il problema della collinearità. Questa dovrebbe quindi essere minore dell’80%, infatti il risultato giustifica il proseguimento della stessa.

Adesso si può verificare se ci sia omogeneità nelle covarianze delle variabili dipendenti esportazione e riserve attraverso i gruppi.

boxM(bank[,2:3], group = bank$crisis)

##   
## Box's M-test for Homogeneity of Covariance Matrices  
##   
## data: bank[, 2:3]  
## Chi-Sq (approx.) = 1.0465, df = 3, p-value = 0.79

C’è omogeneità. Quindi si verificano anche le varianze.

leveneTests(bank[,2:3], group = bank$crisis, center = mean)

## Warning in leveneTest.default(x, group = group, center = center, ...): group  
## coerced to factor.  
  
## Warning in leveneTest.default(x, group = group, center = center, ...): group  
## coerced to factor.

## Levene's Tests for Homogeneity of Variance (center = mean)  
##   
## df1 df2 F value Pr(>F)  
## export 1 77 0.8434 0.3613  
## reserves 1 77 0.0807 0.7770

Si in entrambe (metodo del *p-value*).

Bisogna ora verificare l’effetto della crisi bancaria sull’esportazione e sulla combinazione delle riserve bancarie, ci affidiamo all’argomento test = ”*Wilks*”, che serve per valutare la significatività mediante la statistica di *Wilks*. Questo a causa del test MANOVA con risultato significativo, e occorre esaminare le singole variabili dipendenti per capire come esse contribuiscano alla significatività del test globale, inoltre implementare confronti multivariati tra vettori di medie tra i vari gruppi o dei test di *Tukey* sui risultati dei singoli test ANOVA.

man <- manova(cbind(bank$export, bank$reserves) ~ bank$crisis)  
summary(man, test="Wilks")

## Df Wilks approx F num Df den Df Pr(>F)   
## bank$crisis 1 0.87021 5.6677 2 76 0.005078 \*\*  
## Residuals 77   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Si può notare che contribuiscono con un valore *p-value* basso. Quindi verifichiamo quando contribuiscono alla significatività totale.

summary(man)$SS

## $`bank$crisis`  
## [,1] [,2]  
## [1,] 86.42939 1049.473  
## [2,] 1049.47336 12743.285  
##   
## $Residuals  
## [,1] [,2]  
## [1,] 983.444 5682.425  
## [2,] 5682.425 89978.234

In ultima istanza bisogna controllare quanta della varianza nelle variabili dipendenti è spiegata dalle banche in crisi.

etasq(man, test= "Wilks")

## eta^2  
## bank$crisis 0.1297908

É spiegata al 13%.

Valutando anche le esportazioni quando le banche sono in crisi e quando no.

summary.aov(man)

## Response 1 :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## bank$crisis 1 86.43 86.429 6.7671 0.01113 \*  
## Residuals 77 983.44 12.772   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response 2 :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## bank$crisis 1 12743 12743.3 10.905 0.001456 \*\*  
## Residuals 77 89978 1168.5   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

In conclusione, l’analisi ha dimostrato che esportazioni e le riserve differiscono rispetto a quando il settore bancario non è in crisi.

1. **Bibliografia**

Dell’Omodarme M., *Esercitazioni di statistica biomedica. Alcune note su R,* 2012.

Massarotto G., *Analisi delle Serie Temporali. Lucidi delle lezioni,* 2003.

Mineo M. A., *Una guida all’utilizzo dell’ambiente statistico R*, 2003.

Micciolo R. e Espa G., *Analisi esplorativa dei dati con R*, 2012.

Peng R.D., *R programming for data science*, 2015.

R.H. Shumway and D.S. Stoffer, *Time Series Analysis and its Applications*. *Springer*-*Verlag*, ,2000.Ricci V., *Analisi delle serie storiche con R,* 2005.

**IV. Materiali utilizzati**

Documentazione RStudio.

Testi e risorse del corso di Big Data and Analytics.

<https://cran.r-project.org/web/packages/MVN/vignettes/MVN.pdf>

[https://bookdown.org/yihui/rmarkdown/notebook.html#using-notebooks](https://bookdown.org/yihui/rmarkdown/notebook.html%23using-notebooks)

<https://data.library.virginia.edu/understanding-q-q-plots/>

<https://www.youtube.com/watch?v=10cuDKGytMw>

<https://www.youtube.com/watch?v=16Jh_t4QkRA>

<https://www.rdocumentation.org/packages/heplots/versions/1.3-5/topics/boxM>

<http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance>