Institute of Actuaries of India

Subject CS2B – Risk Modelling and Survival Analysis (Paper B)

May 2023 Examination

INDICATIVE SOLUTION

Introduction

The indicative solution has been written by the Examiners with the aim of helping candidates. The solutions given are only indicative. It is realized that there could be other points as valid answers and examiner have given credit for any alternative approach or interpretation which they consider to be reasonable.

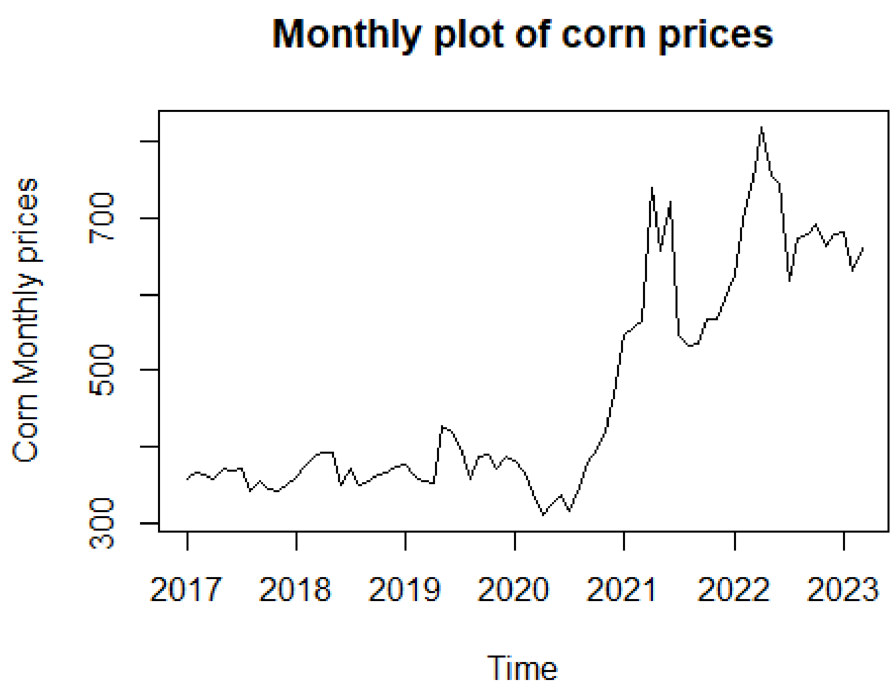
Solution 1:

```
Corn_Prices<-read.csv("D:\\Monthly_corn.csv")
```

(i) `Close<-ts(Corn_Prices$Close,start = c(2017,1),frequency=12)`

[1]

(ii) `plot(Close, ylab = "Corn Monthly prices", main = "Monthly plot of corn prices")`

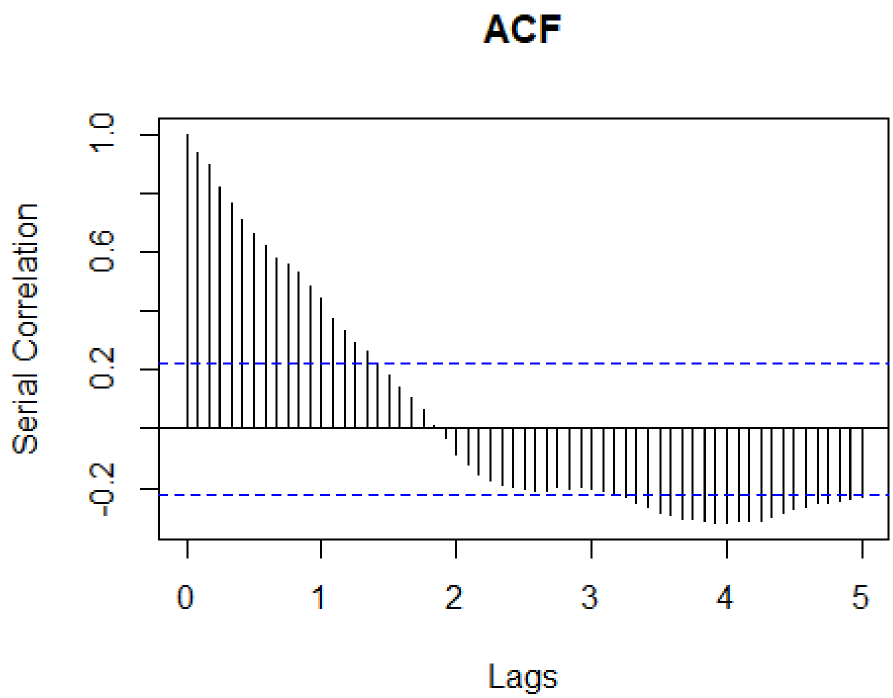


[1.5 marks for the plot, 0.5 marks for proper labeling of title and axes,**Max 2]**

(iii) The series is not stationary as it appears to be trending upwards with time. The mean at different periods appears not to be constant For periods be tween 2017 to 2020, it appeared to be stationary.

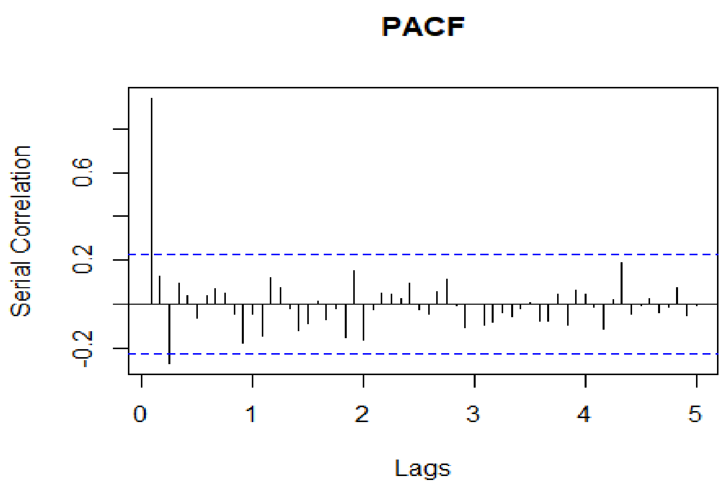
[1]

(iv) `acf(Close, main = "ACF", xlab = "Lags", ylab = "Serial Correlation", lag.max=60)`



[1.5 marks for the plot, 0.5 marks for proper labeling of title and axes,2 marks]

```
pacf(Close, main = "PACF", xlab = "Lags", ylab = "Serial Correlation", lag.ma
x=60)
```



[1.5 for the plot, 0.5 for proper labeling of title and axes,2 marks]

[Max 4]

```
(v)   If a series is stationary, the ACF should decay to zero quickly and sho
uld not display any oscillation.The series is not stationary because auto cor
relation function is not decaying to zero quickly. Also, if the number of lag
s is increased, oscillation is also observed, hence cannot be stationary
```

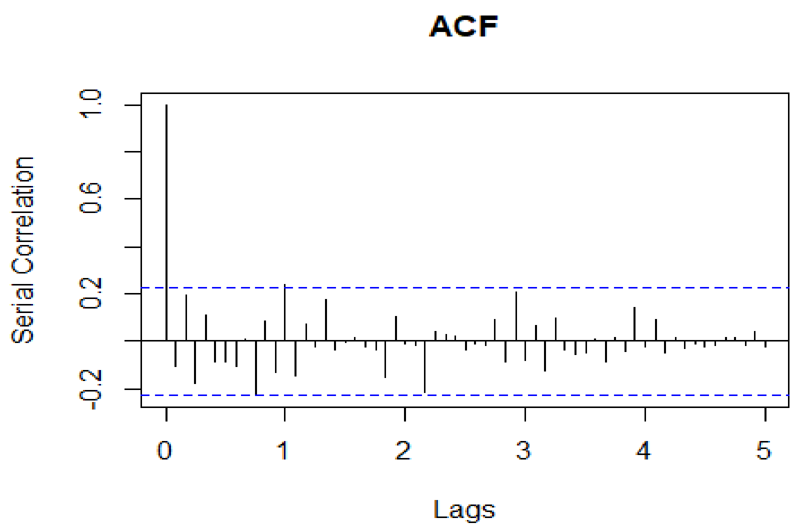
[1 Mark for each valid reason,**Max 2]**

(vi) Create another time series "monthly_returns" by using the formula log(Pt/Pt-1), where Pt and Pt-1 correspond to the closing prices of month t and t-1 respectively.

```
returns<-diff(log(Close))
```

[2]

(vii) `acf(returns, main = "ACF", xlab = "Lags", ylab = "Serial Correlation", lag.max=60)`



[1 for the plot, 0.5 for proper labeling of title and axes]

```
pacf(returns, main = "PACF", xlab = "Lags", ylab = "Serial Correlation", lag.max=60)
```

[1 for the plot, 0.5 for proper labeling of title and axes]

[3]

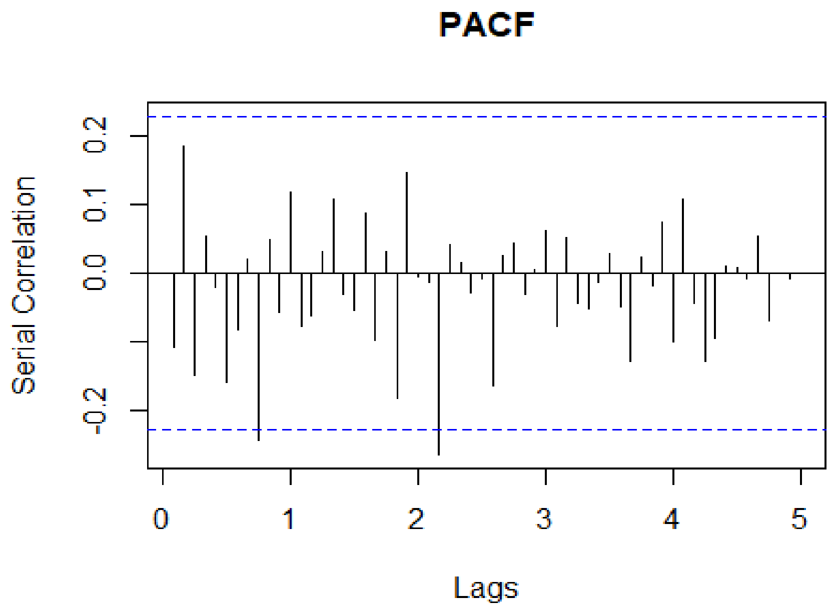
(viii) The monthly returns appear stationary. If a series is stationary, the ACF should decay to zero quickly and should not display any oscillation.

[2]

(ix) The most appropriate ARMA will be ARMA (0,0) as the ACF and PACF do not show significance at any of the lags [1]

(x) `library(forecast)`

```
## Warning: package 'forecast' was built under R version 4.0.5

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```




```
m1<-arima(returns,order = c(0,0,0))
m2<-arima(returns,order = c(1,0,0))
m3<-arima(returns,order = c(0,0,1))
m4<-arima(returns,order = c(1,0,1))
m1$aic

## [1] -164.6838

m2$aic

## [1] -163.5281

m3$aic

## [1] -163.3079

m4$aic

## [1] -165.2896

auto.arima(returns)

## Series: returns
## ARIMA(0,0,0)(1,0,0)[12] with zero mean
##
## Coefficients:
##          sar1
##        0.2548
## s.e.   0.1120
##
## sigma^2 estimated as 0.005687:  log likelihood=86.37
## AIC=-168.74   AICc=-168.58   BIC=-164.14

# ARMA (1,1) has the lowest AIC.
```

[0.5 for fitting each of the four models, 2 for identifying the best model]

[4]

(xi) The deviation is observed because seasonality is not considered in the model. ACF is actually demonstrating a small significant serial correlation at lag 12, which might give rise to seasonal ARMA.
If we are forcefully fitting non seasonal models to a seasonal data, the discrepancy is possible.

[1 mark for each valid observation, **Max 2]**

[24 Marks]

Solution 2:

```
CSK<-read.csv("D:\\CSK.csv")
rownames(CSK)<-CSK$Player
CSK$Player<-NULL
CSK[,1:6]<-scale(CSK[,1:6])
```

(i)
```
cent<-stats::aggregate(.~Initial_Class,data = CSK, FUN = "mean")
cent
```

```
##   Initial_Class    Bat_Avg     Bat_SR Bound_Sixes   Bowl_Avg  Bowl_Econ
## 1       Batsman  0.7358645  0.4665582   0.4410948  0.4163466  0.4205526
## 2        Bowler -1.0118137 -0.6415175  -0.6065053 -0.5724766 -0.5782599
##      Bowl_SR
## 1  0.4267018
## 2 -0.5867150
```

[0.5 marks for correct values of each column]

[Max 3]

(ii) The batsman has higher batting average, higher batting strike rate, higher percentage of runs scored in boundaries and sixers, higher bowling average, higher bowling economy and higher bowling strike rate. All the values are comparatively lower for bowlers

[1 for each valid observation,**Max 2]**

(iii)

(a) & (b):

```
newD1<-c()
newD2<-c()
for (i in 1:38) {

    d1<-sqrt(sum((CSK[i,1:6]-cent[1,2:7])^2))
    d2<-sqrt(sum((CSK[i,1:6]-cent[2,2:7])^2))
    newD1<-c(newD1,d1)
    newD2<-c(newD2,d2)

}
newD1
```

```
##  [1] 2.434718 1.729955 2.997824 1.568778 3.348392 2.958461 2.311089 2.4109
48
##  [9] 1.772017 1.799161 1.791762 1.729785 2.272338 3.004869 1.710662 5.6257
47
## [17] 1.844042 3.434897 1.950693 1.868280 2.269412 2.204044 1.642882 2.0192
07
## [25] 4.154325 3.522683 5.453950 1.895283 5.360322 1.765092 1.405719 1.5410
04
## [33] 2.974695 2.349785 2.502810 1.633570 1.549088 1.203394

newD2

##  [1] 2.5460797 3.6709312 0.7248067 3.9317258 0.7657979 1.5270051 2.2706863
##  [8] 1.2766249 2.3503867 4.4885763 4.3502845 4.4537767 2.4004672 1.1346542
## [15] 4.4894753 3.1527952 4.4756309 1.6133179 2.1024193 3.5411385 1.6613139
## [22] 1.1997969 4.3415921 2.3017138 2.0161058 0.8954538 2.9284766 1.1502851
## [29] 3.1197124 3.6388673 2.8807912 4.1211401 1.1314002 2.8696064 1.1010748
## [36] 3.9474648 3.8648224 3.0541624

CSK$D1<-newD1
CSK$D2<-newD2
```

[2 marks for the fomula and 2 marks for the output]

[4+4=**8**]

(iv) `CSK$Iteration1<-ifelse(CSK$D1<CSK$D2,"Batsman","Bowler")` [3]

(v) `table(CSK$Initial_Class,CSK$Iteration1)`

```
##
##          Batsman Bowler
##   Batsman      19      3
##   Bowler        2     14

sum(CSK$Initial_Class!=CSK$Iteration1)/nrow(CSK)

## [1] 0.1315789
```

[2]

(vi)
```
CSK<-CSK[,c(1:7,10)]
cent<-stats::aggregate(.~Iteration1,data = CSK[,-7], FUN = "mean")
cent

##   Iteration1    Bat_Avg     Bat_SR Bound_Sixes   Bowl_Avg  Bowl_Econ    Bo
wl_SR
## 1    Batsman  0.6892993  0.5433875   0.5505359  0.6280737  0.5811715  0.61
36204
## 2     Bowler -0.8514873 -0.6712434  -0.6800737 -0.7758557 -0.7179177 -0.75
80017

newD1<-c()
newD2<-c()
for (i in 1:38) {

    d1<-sqrt(sum((CSK[i,1:6]-cent[1,2:7])^2)) # 3 Marks
```

```
    d2<-sqrt(sum((CSK[i,1:6]-cent[2,2:7])^2)) # 3 Marks
    newD1<-c(newD1,d1)
    newD2<-c(newD2,d2)

}

CSK$D1<-newD1
CSK$D2<-newD2

CSK$Iteration2<-ifelse(CSK$D1<CSK$D2,"Batsman","Bowler")                  [

table(CSK$Iteration1,CSK$Iteration2)

##
##            Batsman Bowler
##   Batsman       18      3
##   Bowler         0     17
```

[2 Mark for computing new means, 3 marks for computing the D1, 3 Marks for computing D2 and 2 Marks for assigning the player to “Batsman or “Bowler” category, **Max 10]**

(vii) `rownames(CSK)[CSK$Iteration1!=CSK$Iteration2]`

```
## [1] "MM Ali"    "RA Jadeja" "JA Morkel"
```

[2]

(viii)

```
set.seed(100)
m1<-kmeans(CSK[,1:6],2)
m1$centers

##      Bat_Avg     Bat_SR Bound_Sixes   Bowl_Avg  Bowl_Econ    Bowl_SR
## 1  0.8514137  0.4421142   0.4803985  1.2071033  1.2126534  1.2242183
## 2 -0.4966580 -0.2578999  -0.2802325 -0.7041436 -0.7073812 -0.7141274
```

[1 Mark for setting the seed and 2 marks for correctly printing the cluster means, **Max 3]**

(ix)
```
CSK$cluster<-m1$cluster
CSK$cluster<-ifelse(CSK$cluster ==1, "Batsman", "Bowler")
```

[2]

(x)

```
table(CSK$Iteration2,CSK$cluster)

##
##            Batsman Bowler
##   Batsman       14      4
##   Bowler         0     20
```

[2]

```
rownames(CSK)[CSK$Iteration2!=CSK$cluster]
```

```
## [1] "SM Curran" "MS Gony"   "SK Raina"  "DR Smith"
```

(xi) After few more iterations, the convergence of iteration 2 with the kmea ns cluster solution will occur.

[1 mark for each valid point, **Max 2**]

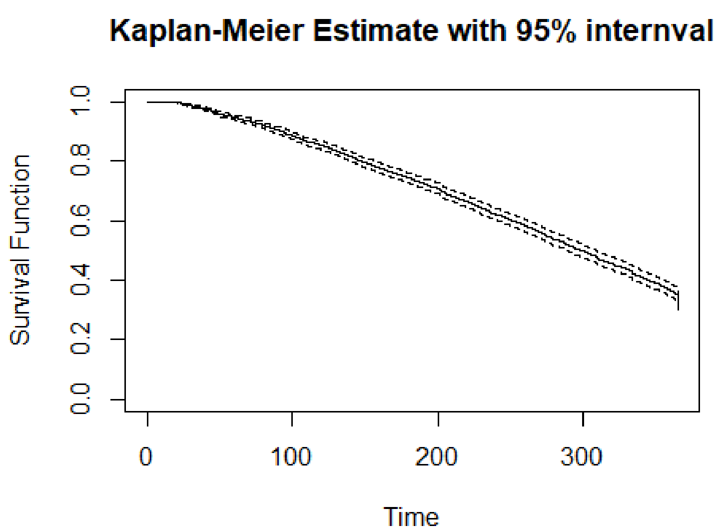
[39 Marks]

Solution 3:

```
library(survival)

## Warning: package 'survival' was built under R version 4.0.5

crackers<-read.csv("D:\\Crackers.csv")
```

(i)

```
KMfit = survfit(Surv(crackers$Time, crackers$Status) ~ 1, conf.int = 0.95)
plot(KMfit,xlab = "Time",ylab = "Survival Function",main = "Kaplan-Meier Esti
mate with 95% internval")
```



[1 Mark for using Surv function, 2 marks for survfit function, 3 marks for the correct plot with appropriate labels.]

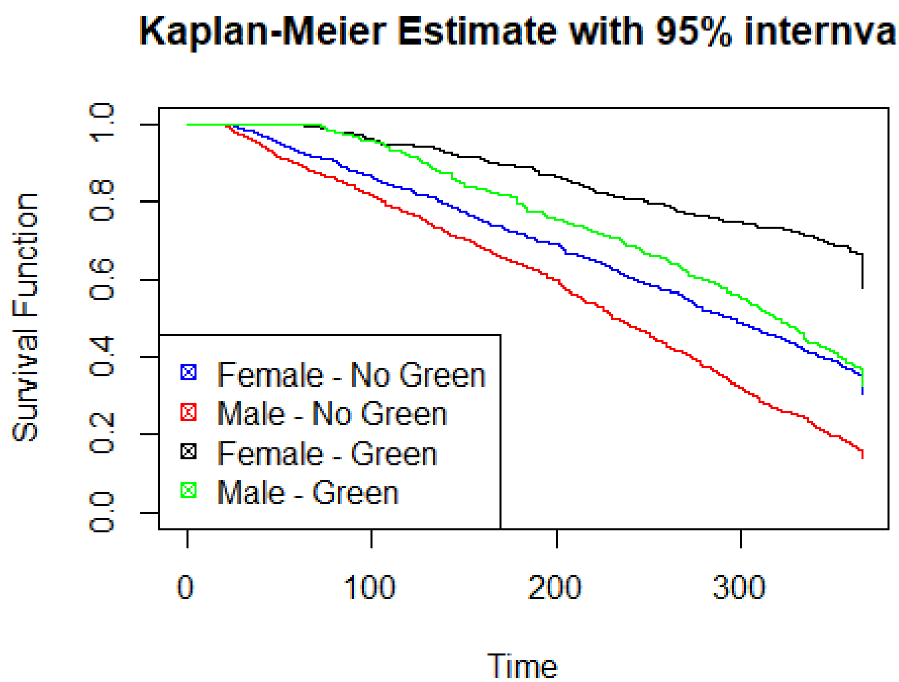
[6]

(ii)

```
summary(KMfit, time = 365)$surv

## [1] 0.3095438

# OR

min(KMfit$surv)

## [1] 0.3095438
```

[2]

```
(iii)
KMfit =survfit(Surv(crackers$Time,crackers$Status)~Green+Male,data = crackers
)
plot(KMfit,col = c("blue", "red", "black", "green"), xlab = "Time",ylab = "Su
rvival Function",main = "Kaplan-Meier Estimate with 95% internval")
legend("bottomleft",legend = c("Female - No Green", "Male - No Green", "Femal
e - Green", "Male - Green"),col = c("blue", "red", "black", "green"),pch = 7)
```



[8]

```
(iv)
# Males are more susceptible to respiratory disorders in general compared to
Females. Females have a better survival function
# The green crackers has a positive effect and increased the survival for bot
h males and females
                                          [1 Mark for each valid reason, Max 2]


(v)
coxph(Surv(Time, Status) ~ Green + Male, data = crackers,ties="breslow")

## Call:
## coxph(formula = Surv(Time, Status) ~ Green + Male, data = crackers,
##     ties = "breslow")
##
##           coef exp(coef) se(coef)       z      p
## Green -0.71233   0.49050  0.05716 -12.463 <2e-16
## Male   0.53867   1.71372  0.05429   9.922 <2e-16
##
## Likelihood ratio test=284.6  on 2 df, p=< 2.2e-16
## n= 2500, number of events= 1452
```

[4]

(vi)

```
# The results suggest that Green crackers reduces the respiratory disorder ra
te of the workers by around 51% irrespective of males and females
# The results also suggests the hazard rate of males is 71% more than that of
the females
# The interaction effect is not clearly visible as the decrease in hazard rat
e is similar among males and females.
# The p-values of both the coefficients are less than 0.05 indicating that bo
th the effects are statistically significant at 5%.
```

[1 Mark for each valid point, **Max 4]**

(vii)

```
mod<-coxph(Surv(Time, Status) ~ Green * Male, data = crackers,ties="breslow")
```

[4]

(viii)

```
female_hazard_red<-1-exp(mod$coefficients[1])
female_hazard_red

##    Green
## 0.571342

male_hazard_red<-1-exp(mod$coefficients[1]+mod$coefficients[2])
male_hazard_red

##     Green
## 0.3158954
```

 # The green crackers reduce the respiratory disorder hazard rate of females by 1-0.42866 = 57%
*# The green crackers reduce the respiratory disorder rate of males by 1-1.59592*0.42866 = 31.5%*

[3]

(ix) The green crackers reduced in a reduction in respiratory disorder for both males and females though the decrease was slightly lesser in case of males

The p-value of the interaction effect is >0.05 indicating that the interaction effect is not statistically significant.

[2]

(x) Both the results are communicating that the green crackers reduce the respiratory disorders

[2]

[37 Marks]

****