Homework - 3 Submission by Haritha Pulletikurti

Homework3Solutions.R

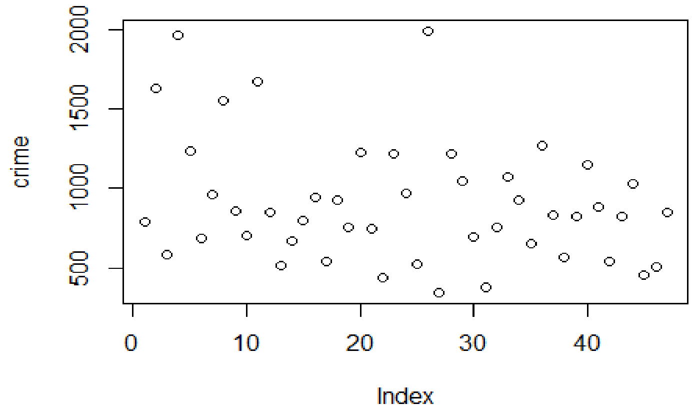
2020-09-09

Question 5.1

Using crime data from the file uscrime.txt (http://www.statsci.org/data/general/uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), test to see whether there are any outliers in the last column (number of crimes per 100,000 people).
Use the grubbs.test function in the outliers' package in R.

Implementation:

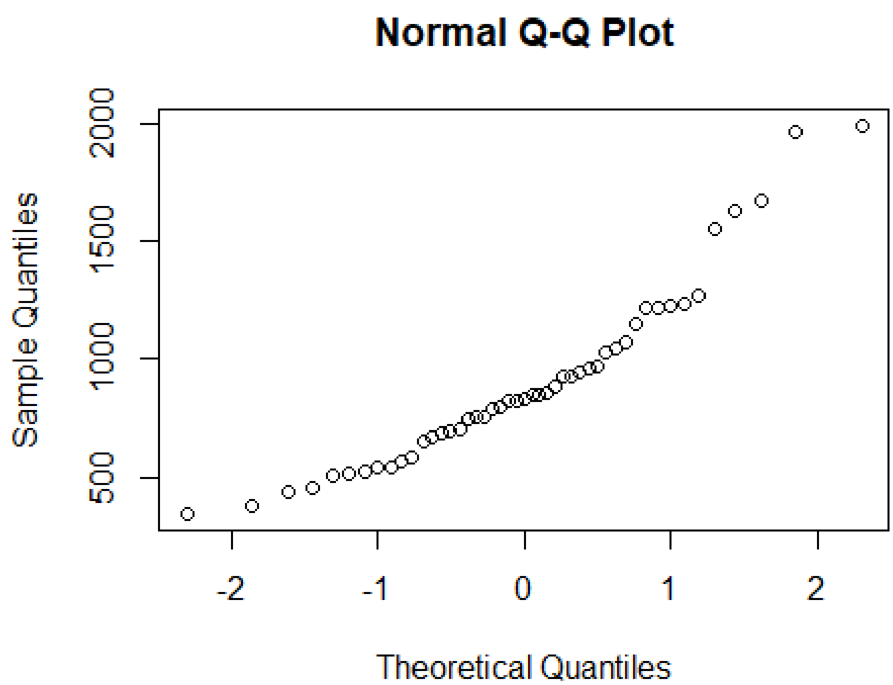
```
# Start with a clear environment
rm(list = ls())

#load the libraries
library(knitr)
library(stringr)
library(outliers)
#load the crime data with headers
crime_data <- read.table("uscrime.txt", stringsAsFactors = FALSE, header =
TRUE)

set.seed(1)
dim(crime_data)

## [1] 47 16

crime <- crime_data[,"Crime"]
plot(crime)
```



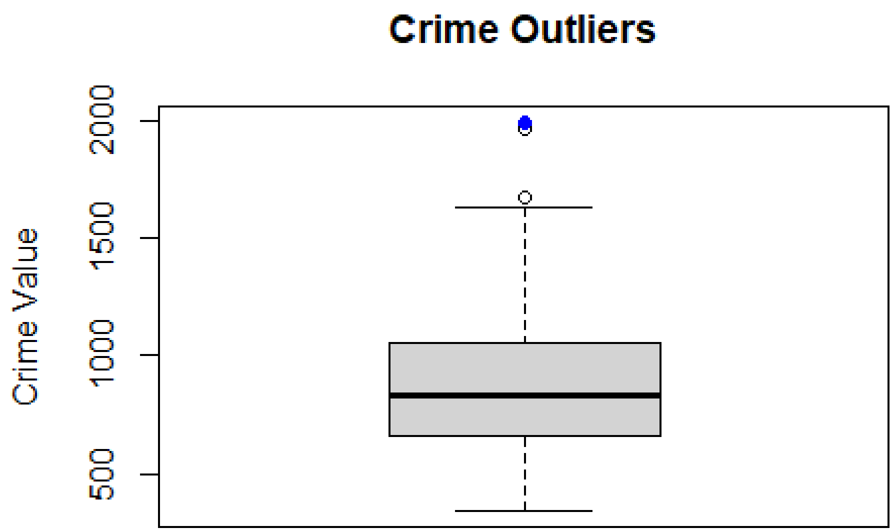

```r
shapiro.test(crime)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  crime
## W = 0.91273, p-value = 0.001882
```

```r
qqnorm(crime)
```



```r
#The qq norm show that the data is 90% of the data is almost normal

#Using Grubbs Test find the mininum and maximum outliers using type = 11
#type = 11 gives 2 outliers one- min and one-max
# I am not using type = 20 as it accepts less data around 30 data values.
# As we have more data, type = 20 gives error. So choosing type = 11


outlier_results<-grubbs.test(crime_data[,16],type=11)
outlier_results

##
## Grubbs test for two opposite outliers
##
## data:  crime_data[, 16]
## G = 4.26877, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers

#Inference: 343 and 1993 are outliers. Let us plot them in boxplot and
visualize


values<- as.numeric(str_extract_all(outlier_results$alternative, "[0-
9]+")[[1]])
Outliers_data<- subset(crime_data, Crime %in% values)
boxplot(crime_data[,16],
        ylab="Crime Value")
points(Outliers_data[,ncol(Outliers_data)], pch=19,col='blue')
title(main="Crime Outliers")
```



Inference: Based on the box plot, the blue dot is 1993. Also looks like the 1969 is also an outlier.
Since grubbs test only gets the min and max outliers (only 2) 1969 is not listed in the test result but it is shown in the boxplot next to 1993. Also 342(min value) is returned as outlier in the grubbs.test. Boxplot did not show it. So, looks like the 342 is within the reasonable range according to boxplot.

Question 6.1

Describe a situation or problem from your job, everyday life, current events, etc. for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

Answer:
The Cusum Apprach can be used in Production Assembly lines where a lot of Robots are involved in assembling the parts of the units. A small change in the movement of the Robotic Holding Parts will make a huge impact on whether the assembly takes place successfully or the process crashes.

Example:

Consider a car manufacturing plant, where in each assembly line a different car part is being assembled by the robots.

```
Suppose the robot hand must drop an item onto the conveyor belt at an angle
of 45 - 50 degrees and the item successfully land at the place


Say the angle is changed to 51 degrees instead of the allowed threshold of 50
degrees, the item fails to land successfully.


Using cusum approach on the everyday data collected of such processes, we can
determine at what point the most failures happen and we will be able to
correct the process to be more successful




Question 6.2

1. Using July through October daily-high-temperature data for Atlanta for
1996 through 2015,
use a CUSUM approach to identify when unofficial summer ends
(i.e., when the weather starts cooling off) each year.  You can get the data
that you need from the file temps.txt or online, for example at
http://www.iweathernet.com/atlanta-weather-records
or
https://www.wunderground.com/history/airport/KFTY/2015/7/1/CustomHistory.html
You can use R if you'd like, but it's straightforward enough that an Excel
spreadsheet can easily do the job too.

rm(list = ls())
data <- read.delim("temps.txt", header=T)


S= c()
DetectedDecreaseIndex = c()
DetectedDecrease = c()
S[0] = 0

# Let c= 5 and T = 20
# WE need to find when the temperrature decreases so the summer ends.

#Detecting a decrease: S(t)=MAX{0, S(t-1)+(Mean(X)-X(t)-C)}


t=20
C=5
for(j in 2:ncol(data))
{
 for(i in 1:nrow(data))
 {
    S[i] = max(0,S[i-1]+(mean(data[,j])-data[i,j] - C))
    if(S[i]>t)
```

```
    {
      DetectedDecreaseIndex[j-1] = i
      DetectedDecrease[j-1]=S[i]
      break
    }
  }
}

#Preparing Data for the cusum table


cusum_year = colnames(data[-1])
cusum_decrease_date = c()
cusum_c = c()
cusum_t =c()
cusum_st = c()
for(k in 1:length(DetectedDecreaseIndex))
{
 cusum_decrease_date[k] = data[DetectedDecreaseIndex[k],1]
 cusum_st[k]=DetectedDecrease[k]
 cusum_c[k] = C
 cusum_t[k] = t

}
#put all the values into a matrix so we can display as a table

matrix.c = cbind(cusum_year,cusum_decrease_date,cusum_st,cusum_c,cusum_t)
colnames(matrix.c) = c("Year","End of Summer", "S(t)", "C","Threshold")
matrix.c = as.table(matrix.c)
matrix.c

##   Year  End of Summer       S(t)         C Threshold
## A X1996 30-Sep        25.1463414634146 5 20
## B X1997 27-Sep        30.0243902439024 5 20
## C X1998 9-Oct         21.0406504065041 5 20
## D X1999 30-Sep        20.9349593495935 5 20
## E X2000 7-Sep         26.0650406504065 5 20
## F X2001 26-Sep        22.6585365853658 5 20
## G X2002 27-Sep        22.3414634146341 5 20
## H X2003 1-Oct         23.9186991869919 5 20
## I X2004 12-Oct        22.349593495935  5 20
## J X2005 9-Oct         26.4308943089431 5 20
## K X2006 13-Oct        32.390243902439  5 20
## L X2007 12-Oct        21.7967479674797 5 20
## M X2008 19-Oct        31.5365853658537 5 20
## N X2009 5-Oct         22.9430894308943 5 20
## O X2010 30-Sep        23.0569105691057 5 20
## P X2011 8-Sep         26.3821138211382 5 20
## Q X2012 3-Oct         25.6016260162602 5 20
## R X2013 17-Aug        24               5 20
```

```
## S X2014 29-Sep        23.6016260162601 5 20
## T X2015 26-Sep        23.8048780487805 5 20



Inference :

From the Year"1996 -2002”, Summer ended around mid to end of September
From the Year "2003-2009", Summer ended around Mid October
From Year "2010-2011 and 2014-2015" Summer ended at end of September
For Years 2013 Summer ended very soon in mid August.The Values ranged between
60’s through 70’s

#Question(2)
#   Use a CUSUM approach to make a judgment of whether Atlanta’s summer
climate has gotten warmer
#in that time (and if so, when).
# Start with a clear environment

rm(list = ls())
data <- read.delim("temps.txt", header=T)

St= c()
DetectedIncreaseIndex = c()
DetectedIncrease = c()
St[0] = 0

# Let c= 5 and T = 20
# WE need to find when the temperature rises

#Detecting a Increase: S(t)=MAX{0, S(t-1)+(X(t)-Mean(X)-C)}


t=10
C=5
for(j in 2:ncol(data))
{
  for(i in 1:nrow(data))
  {
    St[i] = max(0,St[i-1]+(data[i,j]- mean(data[,j]) - C))
    if(St[i]>t)
    {
      DetectedIncreaseIndex[j-1] = i
      DetectedIncrease[j-1]=St[i]
      break
    }
  }
}
cusum_year = colnames(data[-1])
cusum_increase_date = c()
```

```r
cusum_c = c()
cusum_t =c()
cusum_st_inc = c()
for(k in 1:length(DetectedIncreaseIndex))
{
  cusum_increase_date[k] = data[DetectedIncreaseIndex[k],1]
  cusum_st_inc[k]=DetectedIncrease[k]
  cusum_c[k] = C
  cusum_t[k] = t

}


matrix.HighSummer =
cbind(cusum_year,cusum_increase_date,cusum_st_inc,cusum_c,cusum_t)
colnames(matrix.HighSummer) = c("Year","Increased Summer", "Cusum S(t)",
"C","Threshold")
matrix.HighSummer = as.table(matrix.HighSummer)
matrix.HighSummer
```

```
##   Year  Increased Summer Cusum S(t)       C Threshold
## A X1996 3-Jul            16.5691056910569 5 10
## B X1997 4-Jul            13.9756097560976 5 10
## C X1998 8-Jul            14.4390243902439 5 10
## D X1999 23-Jul           10.5691056910569 5 10
## E X2000 4-Jul            11.9024390243902 5 10
## F X2001 11-Jul           11.6829268292683 5 10
## G X2002 8-Jul            10.0731707317074 5 10
## H X2003 21-Jul           10.1219512195122 5 10
## I X2004 8-Jul            10.1788617886179 5 10
## J X2005 23-Jul           10.4959349593496 5 10
## K X2006 4-Jul            12.8536585365854 5 10
## L X2007 4-Aug            14.609756097561  5 10
## M X2008 12-Jul           11.8780487804878 5 10
## N X2009 4-Jul            12.0243902439025 5 10
## O X2010 9-Jul            10.3658536585366 5 10
## P X2011 4-Jul            10.1707317073171 5 10
## Q X2012 3-Jul            12.6991869918699 5 10
## R X2013 10-Aug           11.3333333333333 5 10
## S X2014 7-Aug            10.2276422764228 5 10
## T X2015 10-Jul           10.7967479674797 5 10
```

Inference :
Using the CUSUM approach, I calculated the increase in Atlanta Temperatures.
The most high summers are in July(90 and above) based on the above table.
I have selected the Threshold = 10 and C=5.
Note : When The Threshold is 20 or above, few year's St value is null.
So,Threshold = 10 is choosen as ideal.