ISYE 6501 HOMEWORK 3

2022/09/14

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.

Script:

```
##### ISYE6501 Homework #3 ######
##### Question 5.1       ######


## Clear Work Area ##
rm(list=ls())  #clear environment
cat("\014") #clear console
#dev.off()  #clear plots


##Setup Directory
setwd("C:/Users/Mike/Documents/ISYE6501/HW03")  #set working directory to read and write data
#getwd()  #check working directory if needed


##Input data
uscd <- read.table("uscrimedata.csv", header = TRUE, sep = ",") #read csv data into a table
#uscd # check that data read in correctly
#install.packages("outliers")
library(outliers)  #load outliers package


##plot data
jpeg(file="ISYE6501HW03_51aggregatecrimedatascatterplot.jpeg")
plot(uscd[,16], main = "Aggregate 1960 US Crime Data ", xlab = "State index", ylab = "# of offenses per
100,000 people")
dev.off()
jpeg(file="ISYE6501HW03_51aggregatecrimedataboxplot.jpeg")
```

```
boxplot(uscd[,16], main = "Aggregate 1960 US Crime Data Across 47 States", ylab = "# of offenses per 100,000 people")
dev.off()

##search for outliers, null hypothesis is the data has no outliers.  P < 0.05 indicative of outliers at 95% confidence
grubanalysis <- grubbs.test(uscd[,16], type = 10) #type 10 to evaluate if max value is an outlier
Pvalue <- grubanalysis[3] #save outlier probability from grubbs.test


##Outputs
print("outlier test run with null hypothesis that there are no outliers at the 90% and 95% confidence level")
if (Pvalue < 0.05) {
  print("There are outliers in the data at the 95% confidence level")
  print(grubanalysis[2])
} else if (Pvalue < 0.1) {
  print("There are outliers in the data at the 90% confidence level")
```

Output:

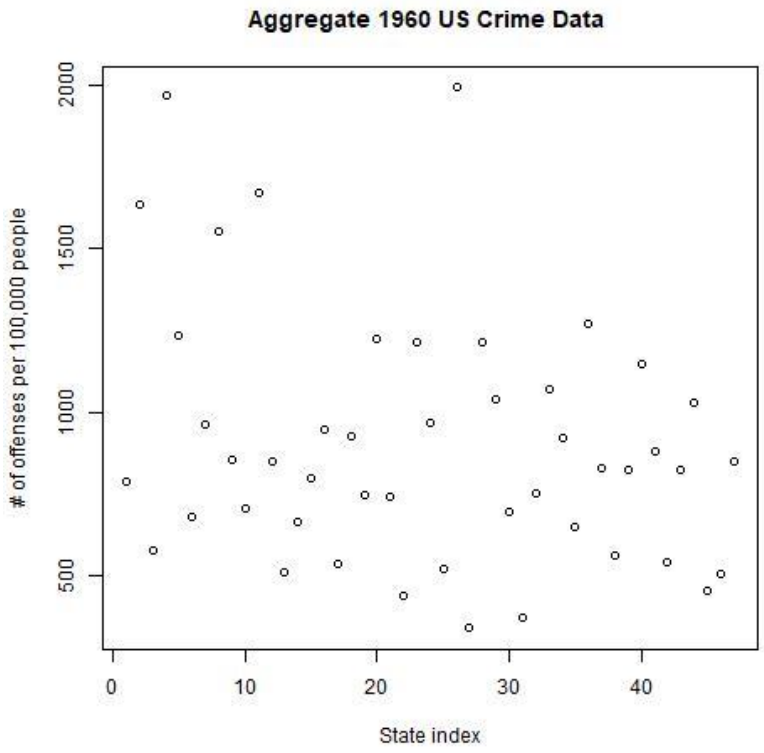
[1] "outlier test run with null hypothesis that there are no outliers at the 90% and 95% confidence level"

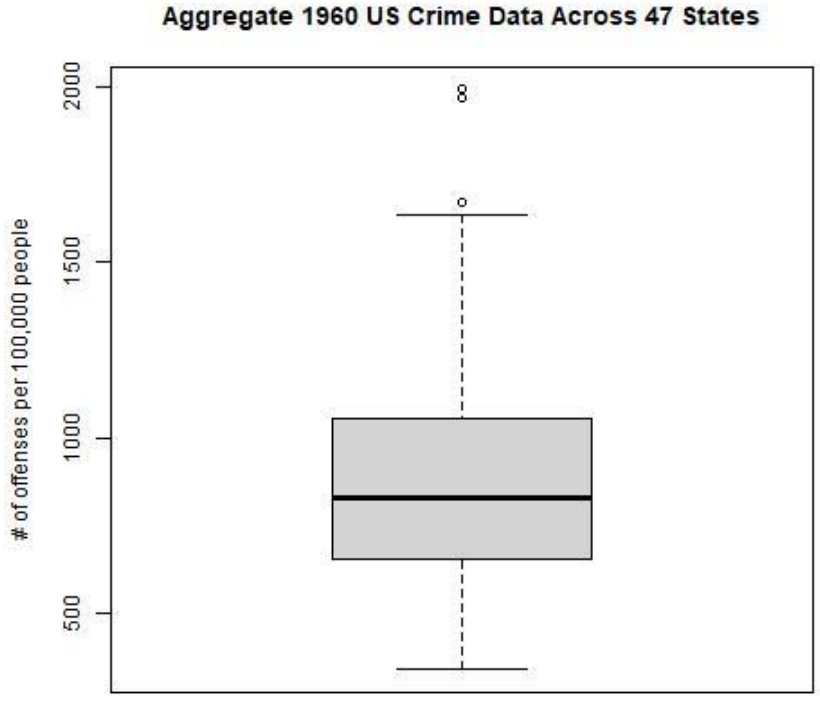
[1] "There are outliers in the data at the 90% confidence level"

$alternative

[1] "highest value 1993 is an outlier"

Data plots:


Aggregate 1960 US Crime Data
# of offenses per 100,000 people
500
1000
1500
2000
0
10
20
30
40
State index


Aggregate 1960 US Crime Data Across 47 States
# of offenses per 100,000 people
500
1000
1500
2000
8

Discussion:

This dataset aimed to investigate the effect of punishment regimes on crime rates in 47 US states during 1960. To investigate for outliers, the data was loaded into R, plotted visually, and run through the grubbs.test. The scatter plot indicates there are a few states with a high number of offenses per 100,000 people, with the max being 1993. To dive deeper, a box and whisker plot was made, showing that there may be 3 outliers. Boxplots identify outliers by identify data points that are over 1.5 times the interquartile range above quartile 3 or below quartile 1. The grubbs test takes a different approach to outlier identification – confidence testing. It tests the null hypothesis that there are no outliers in the dataset, outputting the associated probability. If that probability is below a certain threshold, say p = 0.05 (95% confidence) or p = 0.10 (90% confidence), the null hypothesis is rejected in favor of the alternate hypothesis that are outliers in the data. In this instance, the p value was 0.0789, which is borderline. There are no outliers at the 95% confidence level, but there would be outliers at the 90% confidence level. More context is needed to understand if these abnormal values are expected/predictable or true outliers.

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?

One example from my work where a change of detection model would be appropriate would be for temperature changes of furnaces during heat treatment. Furnace temperatures as typically tightly controlled with feedback controllers and sensors so that the metal alloy is subjected to specific thermal conditions that bring out its best possible microstructure and properties. If temperatures deviate, then there is potential for undesired microstructural phases to form, leading to suboptimal properties, and potential part failure during use. The part has to work, so heat treatment temperature needs to be controlled. CUSUM would be a useful technique for evaluating the temperature of the furnace over time, detecting if it crossed an allowable threshold. The thermocouple sensor would be expected to have some degree of noise, and that would become the critical value. The threshold would be the max allowed temperature deviation from the mean that results in an acceptable microstructure. The absolute value of the threshold would be material and thermal cycle dependent. The sensor (and therefore the C value) would need to be selected such that it has sufficient precision to detect important temperature changes well before they accumulate towards the threshold. This combination would mitigate occurrence of false alarms in the system, while still being sensitive to real trends and threshold exceedance.

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.

1. To answer this question, an CUSUM calculation was done in excel.

$$S_t = max\{0, S_{t-1} + (\mu - x_t - C)\}$$

With:

$$\mu = Summer\ average\ temperature\ (July)$$

$$x_t = Temperature\ Reading$$

$$S_{t-1} = Previous\ cusum$$

$$S_t = Cusum$$

$$C = critical\ value$$

Looking for:

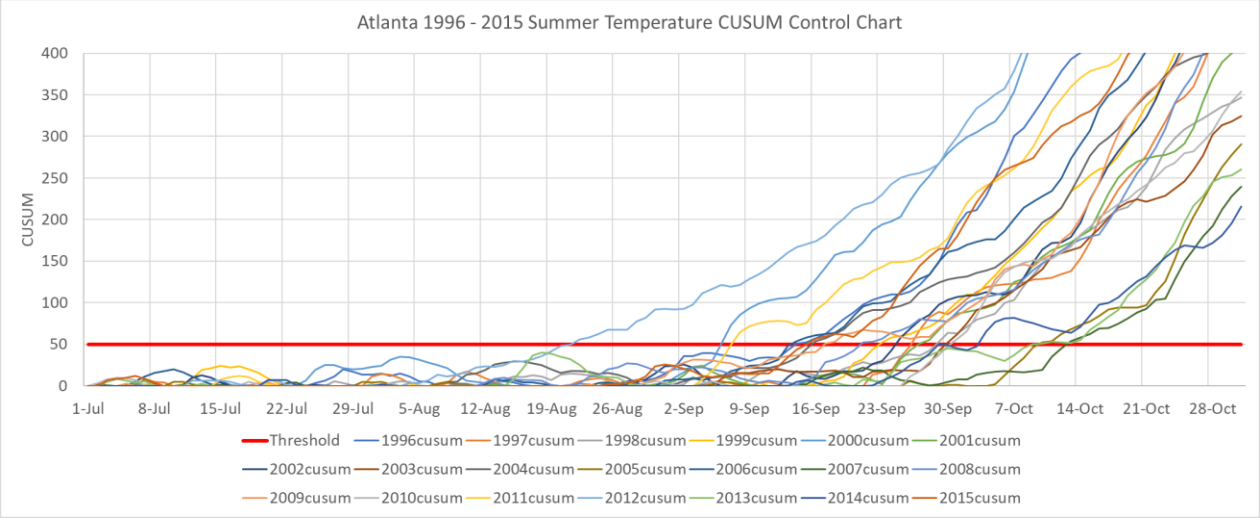
$$S_t \geq T$$

With:

$$T = threshold\ for\ end\ of\ summer$$

The data was imported and a mean value of summer temperature was calculated. The mean summer temperature (μ) was based on the July temperature data, as it was expected that July would be mid-summer, and only have random day to day fluctuations. C=4 was selected based on the average standard deviation of July temperatures being 3.9, intended to mitigate the accumulation of noise towards the cusum value and occurrence of false alarms. T = 100 was arbitrarily chosen as the initial threshold value. This cusum function was run over every year (column), returning the date when summer ended for each (the cusum exceeded the threshold value). The model parameters T and C were then tuned based on the results, targeting an average summer end date of September 22nd, which is the official summer end date. To help with this, the data was plotted to get a visual of cusum value by year. C = 4 and T=50 seemed like appropriate values that limited instances of false alarms, while still being in the vicinity of where slope of the data changed rapidly for most years. C values between 1-10 and T = 25-100 were explored. T the earliest summer end date is predicted to be 8/22 in 2012 and the latest summer end date being predicted to be 10/13 in 2007.

Cusum Control Chart:



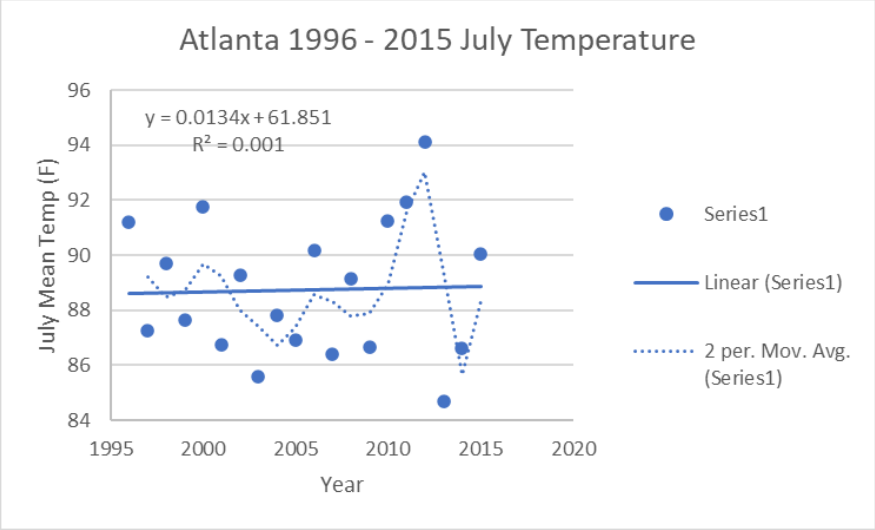
Output Summer end dates:

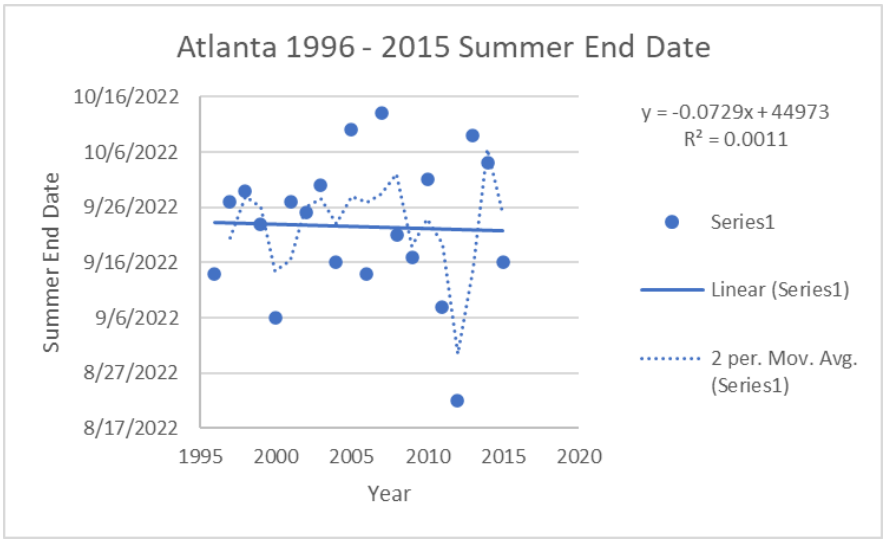
Year	Predicted Summer End Date
1996	9/14/2022
1997	9/27/2022
1998	9/29/2022
1999	9/23/2022
2000	9/6/2022
2001	9/27/2022
2002	9/25/2022
2003	9/30/2022
2004	9/16/2022
2005	10/10/2022
2006	9/14/2022
2007	10/13/2022
2008	9/21/2022
2009	9/17/2022
2010	10/1/2022
2011	9/8/2022
2012	8/22/2022
2013	10/9/2022
2014	10/4/2022

2015	9/16/2022

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)

2. Considering the data generated for part 1 of the question, it does not appear that there is a trend of Atlanta's summer climate temperatures between 1996 - 2015. The summer end date appears to fluctuate on both sides of the mean, and a similar lack of trend is noted in the July temperature averages. Attempting to do a linear regression on them yields R^2 values of ~0.001, which is too low to draw any conclusions.
    a. One datapoint of interest is 2012. During that year Atlanta's summer climate was warmer than expected. The July temperatures were hotter than any other year, and the temperature fell off quicker. This resulted in the cusum function predicting an early end to the summer. Note that 2012 did not have the highest average temperature. Looking a bit deeper, it appears there was a heat wave in Georgia during July 2012 (https://www.redandblack.com/news/july-heat-waves-cause-organization-to-rally-for-change-in-georgia/article_068ac20e-d4fa-11e1-98e6-0019bb30f31a.html)
    Ideally the model for predicting an end to summer would be robust to the effects of heat waves as they are a real phenomenon. One possible improvement would be to incorporate more data in the assessment of average summer temperature. Note that getting rid of 2012 does not appear to change the model results in a significant fashion.





Appendix:

Excel workbook for 6.2:

ISYE6501 Hw03 Workbook Q622.xlsx