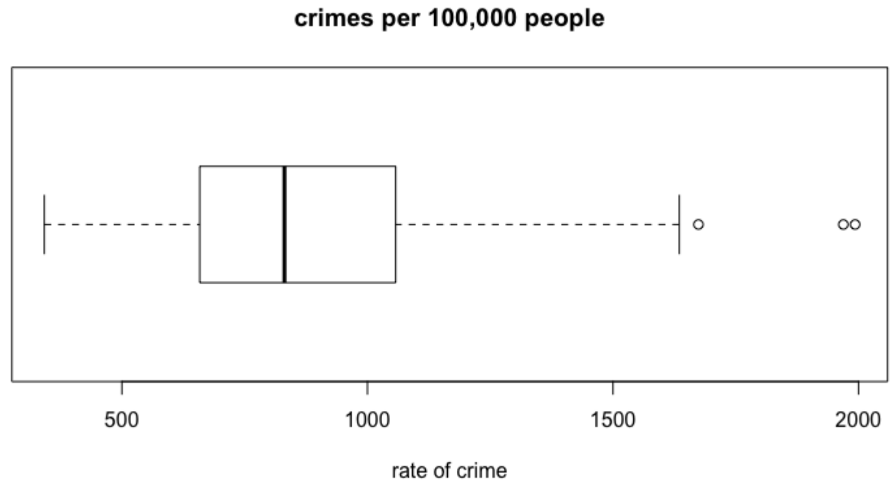
Question 5.1
Using crime data from http://www.statsci.org/data/general/uscrime.txt (description at http://www.statsci.org/data/general/uscrime.html), test to see whether there is an outlier in the last column (number of crimes per 100,000 people). Is the lowest-crime city an outlier? Is the highest-crime city an outlier? Are there others? Use the grubbs.test function in the outliers package in R.

(Please note that my R code for this question is contained in the file 'homework_3_Q5.1.R')

Response
To solve this problem, I used the Grubb's test as directed. Grubb's Test is a method of determining if there is an outlier within a normally distributed data set. The null hypothesis (H_0) of Grubb's states that there are no outliers. The alternative hypothesis (H_a) of Grubb's states that there is exactly one outlier in the data set. The R function *grubbs.test* automatically determines the most extreme point in the data set, and when implemented, returns the data point that was tested as well as the p-value associated with the likelihood that the tested data point was an outlier (the likelihood that the alternative hypothesis is true). The point on the opposite side of the dataset, that is the smallest point if *grubbs.test* determined it should test if the largest point in the set is an outlier, and the largest point if *grubbs.test* determined that it should test if the smallest point in the set is an outlier, can also be tested by setting the parameter *opposite = TRUE* in *grubbs.test*.

To begin my analysis of whether or not there are any outliers in the last column of the US crime statistics data (crimes per 100,000 people, further referred to as *crime_rate*), I first visualized the data using a box plot as may be seen below:

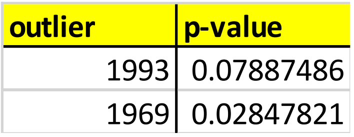


By simply observing the box plot, we can see that there appear to be several very large data points which are possible candidates to be considered outliers. We can also observe that it appears as though there are no very small points which may be possible outliers. To test these ideas, I implemented the function *grubbs.test*.

To be begin, I implemented the most standard version of *grubbs.test* and used the *crime_rate* data as the only input parameter. When implemented this way, *grubbs.test* returned a *p-value* of *0.07887* associated with the *alternative hypothesis* that the point *1993* (the most extreme point) is an outlier. The *p*-value is defined as the probability, under the null hypothesis of obtaining a result equal to or more extreme than what was actually observed, and our *null hypothesis* is that there are no outliers. Therefore, the *p-value* in this case is simply the probability that we will find a point in a normally distributed dataset that is at least *as extreme* as the most extreme point in our dataset.

The choice of whether or not to classify the point *1993* which has a *p-value* of *0.07887* as an outlier, is therefore dependent upon our choice of a value for *alpha*. If we choose to let *alpha = 0.05* (as is common practice), then we would simply determine that there are no outliers based on the grubb's test. However the plot above indicates visually that it appears as though there may be two possible outliers, so I chose to investigate what would happen if I let *alpha = 0.1* .

To explore further whether there are any outliers, I created a function called *grubbs_function* to check if there are multiple outliers with a *p-value* less than a user-defined value of *alpha* using R's *grubbs.test* function. If it is determined that there is an outlier with a *p-value* less than the user inputted value of *alpha*, then the outlier is removed, and the function is run again. The function is run repeatedly until it is determined that there are no more extreme points which could be considered outliers. When I ran *grubbs_function* and let *alpha = 0.1* , it was determined that there are two outliers in the data set. They may be observed below:



outlier	p-value
1993	0.07887486
1969	0.02847821

It is interesting to note that that after removing the data point *1993* from the set, we find a *p-value* of *0.02847821* associated with the point *1969*. This indicates that both of these points are likely outliers, and the reason we initially found the larger *p-value* of *0.07887486* associated with the point *1993* was because of its proximity to the point *1969*. This is why we found a smaller *p-value* associated with the point *1969* after removing the point *1993.*

Finally, to check if there are any very small points in the *crime_rate* data which may also be considered outliers, I ran the R function *grubbs.test* again using the *crime_rate* data, but this time I used the parameter *opposite = TRUE*. When running the test with these parameters I found a *p-value* of *1* associated with the *alternative hypothesis* that the smallest point *342* is an outlier. This indicates that there is absolutely no evidence to support that are any very small points in the crime rate data that are outliers.

Conclusion

Therefore we may conclude on a quantitative basis, that when leting *alpha = 0.1* , there are two outliers: the points *1993* and *1969.* However, it should be noted that when determining whether or not to include outliers in a model, not only quantitative aspects must be taken into consideration, but also qualitative aspects. Just because the points *1993* and *1969* are far separated from the rest of the dataset, this does not make them invalid, in fact they may be important pieces of information. Therefore, when determining how to handle these two outliers in any future models, one should first attempt to determine the reason why these points are isolated from the rest of the data, and if they are erroneous or meaningful.

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?

Response
I believe that change detection models could have many applications within finance. For example, when analyzing securities, a *cusum* model could be used to determine if the trend of a stock price is increasing or decreasing. The information gained by using a *cusum* algorithm could be used to make trading decisions. The appropriate critical and threshold values could be determined through experimentation. Using historical data to systematically back test various critical and threshold values, perhaps using cross-validation, it could be determined which critical and threshold values indicate that a stock increasing or decreasing trend is present.

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.

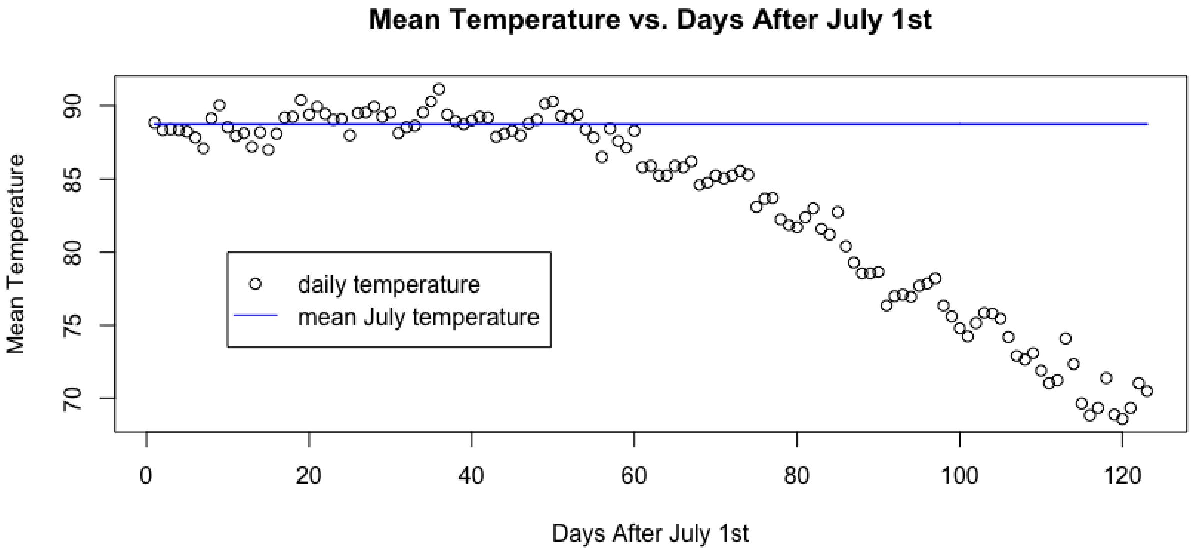
(Please note that my R code for this question is contained in the file 'homework_3_Q6.2.1.R', and if you wish to run the code, you will need to change line 8 of my R code to your local directory which contains the 'temps.txt' data)

Response
The *cumulative sum* function, or *cusum*, is a sequential analysis technique used to detect changes in a dataset. In question 6.2.1 we are asked to determine when the summer unofficially ends in Atlanta by analyzing the temperature data over a 4-month period, July, August, September, and October for each year from 1996 to 2005 (20 years of data). We determine the unofficial end of summer by determining the date in which the weather begins to cool, as defined by the *cusum* function. This means we are looking for a *decrease*. The *cusum* formula used to measure a decrease may be defined as follows:

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

Where S_t = the value returned by the *cusum* function after analyzing point *t* in the data set, S_{t-1} = the *cusum* value associated with the data point *t-1* (the previous value returned by the *cusum* function), μ = the average value of the data set being analyzed, if there were no change, x_t = data point *t* in the in the dataset being analyzed, and *c* is a user defined *critical value* parameter that determines how quickly changes accumulate.

To begin solving this problem, I started by visualizing the data. I found the mean temperature on each day over the 20-year period from 1996 to 2015. Additionally, I found the mean July temperature over the same 20-year period. I plotted these values as may be seen below:

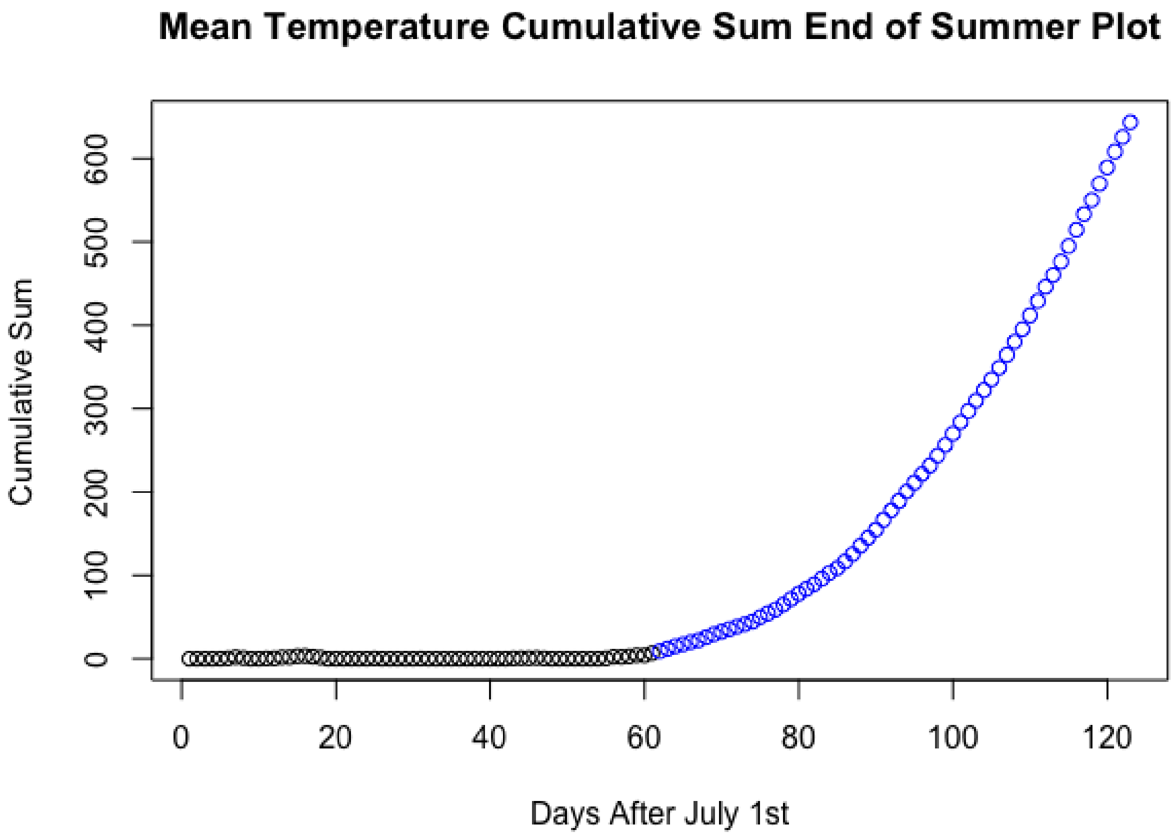


I chose to include the mean July temperature in this figure because it serves as a proxy for the mean summer temperature, since we know that the summer does not end in July. Therefore, we can say that when the daily temperature is consistently below the mean July temperature, then summer has unofficially ended, based on our definition above. Based upon a visualization analysis of the above figure, we can see that the daily temperature permanently drops below the mean July temperature at approximately day 60, which is August, 29th.

I next analyzed the above observation that summer ends on approximately August, 29th by using the *cusum* function. I created a function called *cumulative_sum* in R which works in accordance with the formula of *cusum* defined above. There are two possible methods of determining the unofficial end of summer based upon our data. The first method is to find the average temperature of each day over the 20 year period, and then apply the *cusum* function to these averages. The second method is to apply the *cusum* function to each individual year, and then determine the average day on which summer is said to end. I implemented both of these methods so that I could compare their results.

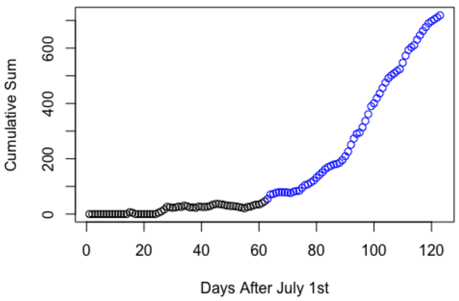
To implement the first method, I found the average temperature on each day, and then applied the *cusum* function. In my implementation of *cusum*, I let μ = the mean July temperature over the 20-year period, since as mentioned above, this serves as an appropriate proxy for the mean summer temperature without any change. I let the *c = 0.5* * the standard deviation of the mean daily temperatures in July over the 20-year period. And in order to determine, based upon the result of the *cusum* function, when summer had ended, I included a threshold parameter *T*.

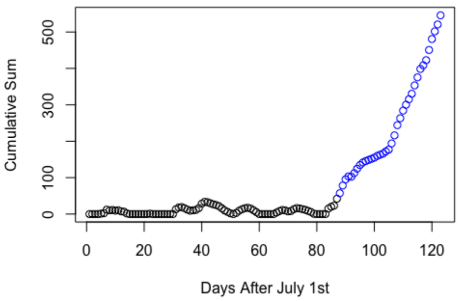
When the cumulative sum surpasses the value *T* , it may be said that a decrease has occurred and summer has ended. I let *T = 10 ** the standard deviation of the mean daily temperatures in July over the 20-year period. The results of this *cusum* analysis are shown in the figure below:

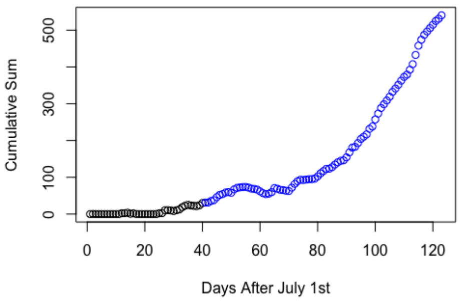


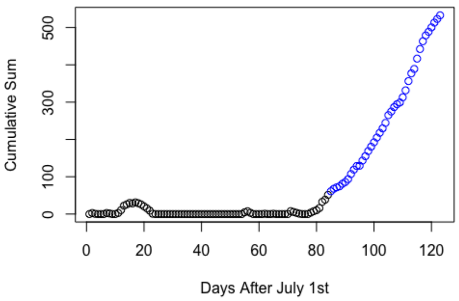
In the figure above the blue points represent the points at which the cumulative sum has surpassed the threshold *T*. As may observed in the figure (and as I verified in R), the threshold is first surpassed at day 62. This means, that based upon this approach, the unofficial end of summer is the 62nd day, or August 31st.

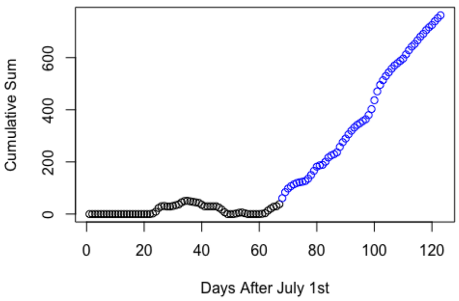
However, as mentioned above, I also attempted to determine the unofficial end of summer by applying the *cusum* function to every individual year, and then determine the average day on which summer was said to end. To implement this approach, I let μ = the mean July temperature for the given year. I let *c = 0.5 ** the standard deviation of daily temperatures in July for the given year. And in order to determine, based upon the result of the *cusum* function, when summer had ended, I again included a threshold parameter *T* as described above. I let *T = 10 ** the standard deviation of daily temperatures in July for the given year. The results of this analysis may be observed in the figures below:

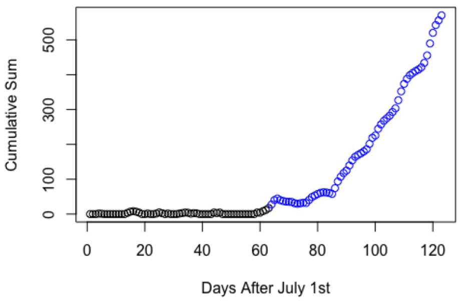

1996 Cumulative Sum End of Summer Plot

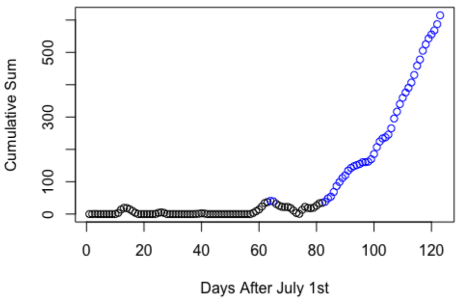

1997 Cumulative Sum End of Summer Plot

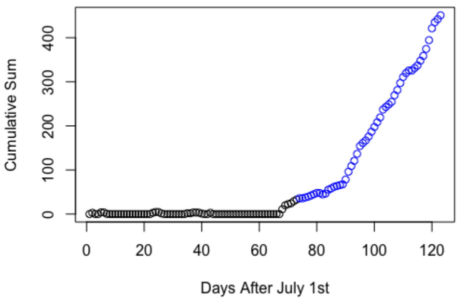

1998 Cumulative Sum End of Summer Plot

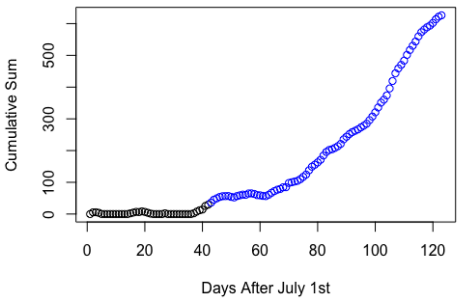

1999 Cumulative Sum End of Summer Plot

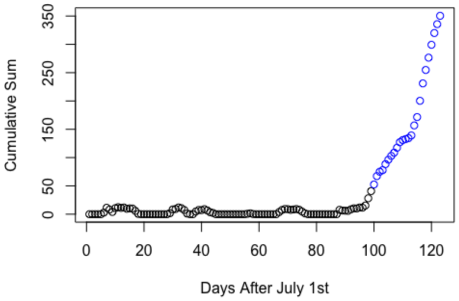

2000 Cumulative Sum End of Summer Plot

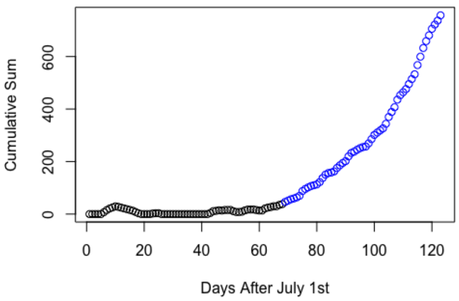

2001 Cumulative Sum End of Summer Plot

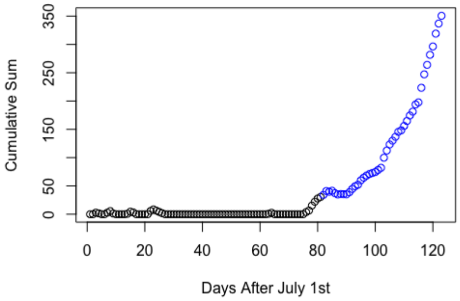

2002 Cumulative Sum End of Summer Plot

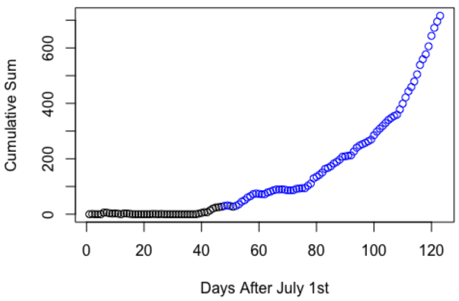

2003 Cumulative Sum End of Summer Plot

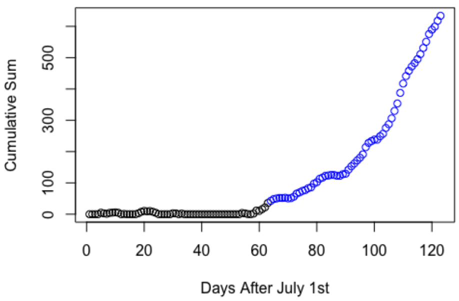

2004 Cumulative Sum End of Summer Plot

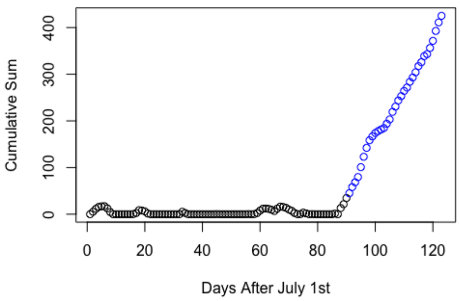

2005 Cumulative Sum End of Summer Plot

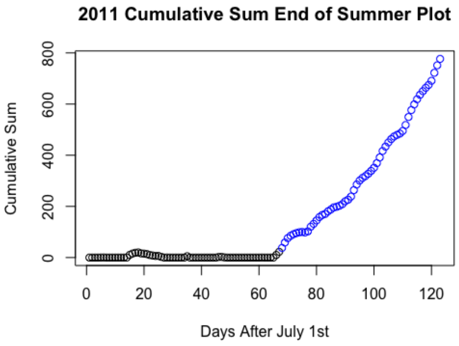

2006 Cumulative Sum End of Summer Plot

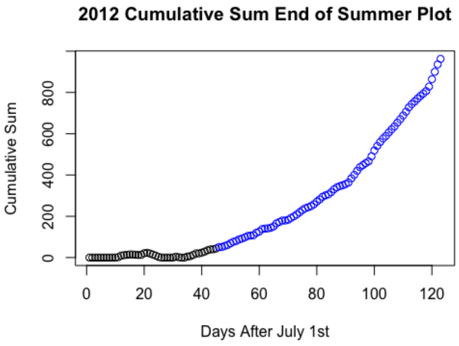

2007 Cumulative Sum End of Summer Plot

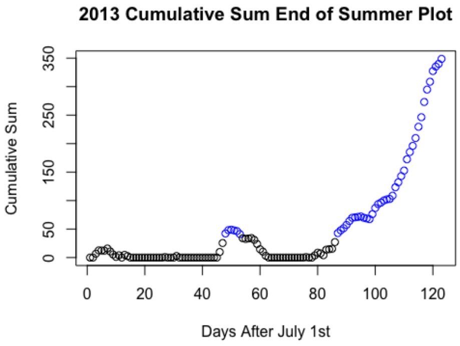

2008 Cumulative Sum End of Summer Plot

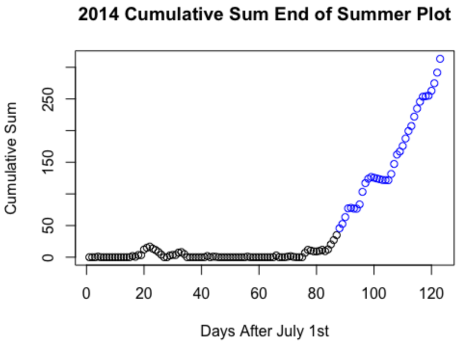

2009 Cumulative Sum End of Summer Plot

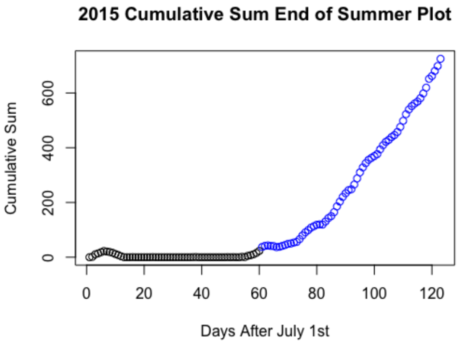

2010 Cumulative Sum End of Summer Plot


2011 Cumulative Sum End of Summer Plot


2012 Cumulative Sum End of Summer Plot


2013 Cumulative Sum End of Summer Plot


2014 Cumulative Sum End of Summer Plot


2015 Cumulative Sum End of Summer Plot

As may be observed in the above figures, there is a substantial difference in the unofficial ending of summer each year. In fact, the standard deviation of the last day of summer is equal *17.23*. As may be observed in one instance, the year 2013, the threshold *T* is crossed, and then the cumulative sum falls below the threshold again. This indicates that the temperature dropped substantially, and then subsequently increased again. This produces a false indication of when the summer has ended in this year. Overall, the mean end of summer based upon applying *cusum* to each year individually, is determined to occur on day *67.7* . Rounding *67.7* to *68* corresponds to September, 6th.

Conclusion

In conclusion, method 1 (as defined above) determined that the end of summer occurs on day 62, and method 2 (as defined above) determined that the unofficial last day of summer occurs on day 68. Since I feel that both of these methods are valid ways of solving this problem, I chose to average their results to make the determination of the last day of summer. This results in the conclusion that day 65, or September, 3rd is unofficially the last day of summer.

Question 6.2.2 Use a CUSUM approach to make a judgment of whether Atlanta’s summer climate has gotten warmer in that time (and if so, when).

(Please note that my R code for this question is contained in the file ‘homework_3_Q6.2.2.R’, and if you wish to run the code, you will need to change line 8 of my R code to your local directory which contains the ‘temps.txt’ data)

Response

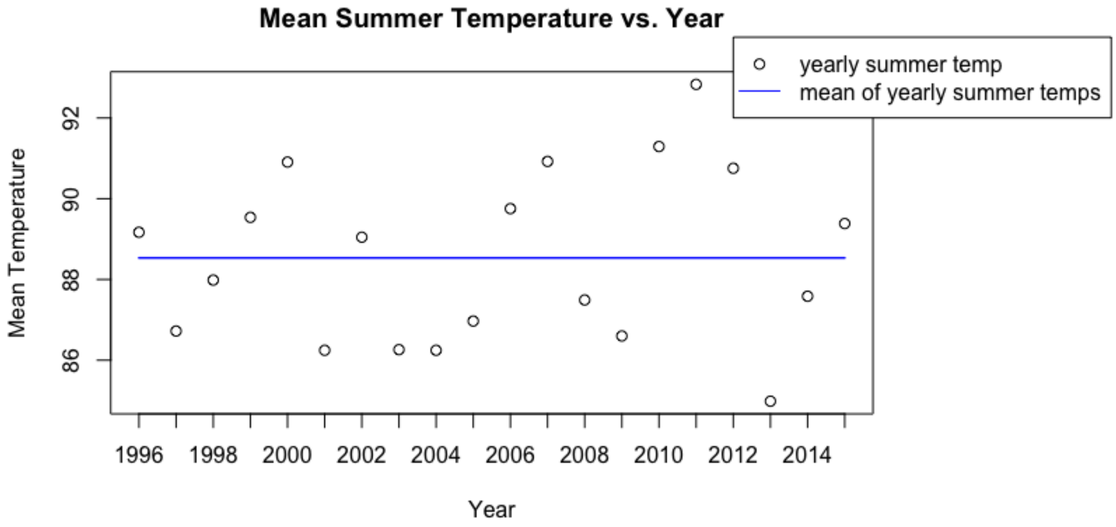
To solve this problem, I again chose to use R. My solution to question 6.2.2 relies upon my result from question 6.2.1 . In order to determine if the summer temperature in Atlanta is rising, we must first define the time period that comprises summer in Atlanta. I chose to define summer based upon the result of 6.2.1 . That is, I chose to define the all of the days from July 1st to September 3rd (or day 1 to day 65) as summertime.

I implemented the *cumulative sum*, or *cusum* function in a similar manner to the way in which it was implemented in question 6.2.1 . However, in this problem we are attempting to detect an increase in value (an increase in temperature), rather than a decrease as in problem 6.2.1 . The *cusum* formula to detect an increase is defined as follows:

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

This formula is nearly identical to the formula used to detect a decrease, except in this case we subtract μ from x_t .

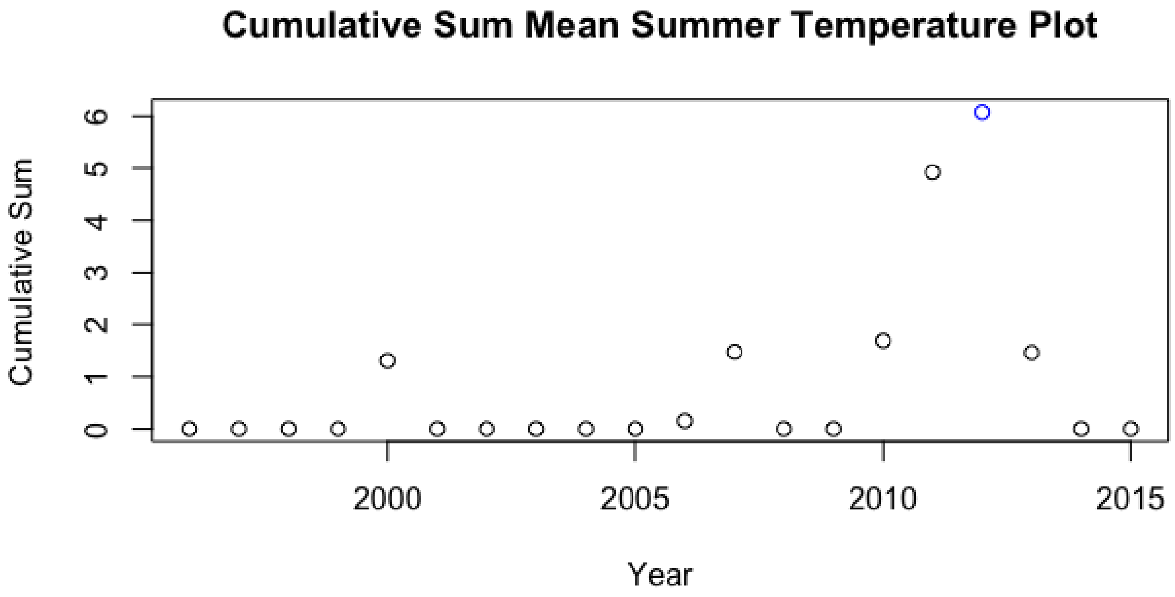
To solve question 6.2.1, I first found the average summer temperature in Atlanta for each year from 1996 to 2005. The average temperatures may be seen in the figure below:



As it may be observed in the figure above, there does not appear to be any kind of increasing (or decreasing) trend in the mean summer temperature between the years 1996 and 2015. While the warmest summer on average does appear near the end of the dataset (in 2011) it is followed two years later by the coolest year in the dataset (2013). Between the years 1996 and 2015, the mean temperature appears to follow no pattern or trend, and instead increases or decreases randomly occur from year to year. I have also found that the average of the mean summer temperaturse is approximately *88.53* and the standard deviation is of the mean summer temperatures approximately *2.13* .

While it appears clear based on looking at the data set that there is not an increasing trend, I tested this idea rigorously by a creating function in R called *cumulative_sum* which implements the *cusum* equation, as defined above, to check for an increase.

I used the following parameters when implementing the *cusum* algorithm. I let μ = the mean of the mean summer temperatures, since there was no other obvious choice of value for μ, and no indication that the mean summer temperature is increasing over the time period we are observing. I let $c = 0.5$ * the standard deviation of the mean summer temperatures. And in order to determine, based upon the result of the *cusum* function, whether the mean summer temperature was increasing, I again included a threshold parameter T. I let $T = 2.5$ * the standard deviation the mean summer temperatures. I used a coefficient of *2.5* rather *10* as done above in question 6.2.1, so that the algorithm would be more sensitive to detecting change, since it doesn't appear as though an obvious trend exists. The results of this analysis may be observed in the figures below:



In the figure above the blue points represent the points at which the cumulative sum has surpassed the threshold T. As may observed in the figure (and as I verified in R), the threshold is only surpassed once, in the year 2012. As we may recall from the plot of the mean summer temperatures versus the year, the summer temperatures in the years 2011, and 2012 were both substantially warmer than the mean summer temperature, and in fact, 2012 had the warmest mean summer temperature. This is the reason why we observe that the threshold *T* was crossed in the year 2012, and is therefore said to be in to be indicative of an increase, according to the *cusum* algorithm.

Conclusion

When making the final determination of whether or not the summers in Atlanta are actually getting warmer, we must consider a holistic analysis of the data. While *cusum* does indicate that one point (the year 2012) has crossed the (sensitive) threshold value that I set for *T*, immediately following the year 2012 we see a sharp decrease in temperature in 2013, followed in 2014 by the lowest recorded mean summer temperature in this time range. Additionally, the final two years, 2014 and 2015, both had a *cusum* value of 0, which indicates that there is no evidence of increasing temperatures at the end of the analyzed time range. Therefore, even though one year did in fact cross the threshold *T*, we cannot say that the mean summer temperatures are in general increasing. While can conclude that temperature did increase to above average levels in 2011 and 2012 there is no general trend which would indicate consistent warming over the analyzed period of time. Rather, it simply appears that within time range 1996-2015 some summers are simply randomly warmer or cooler than others.