Maria Ruwe, Leah Gillespie, Austin Hardesty

Mini-Project #2

Dr. Long

16 Feb. 2018

1) We certainly agreed with the first group about the minimum temperature increasing. However, we disagreed about maximum temperature decreasing. We do not believe we can say that maximum temperature is increasing or decreasing. Also, we believe that the first group could have provided some statistics to support their conclusions.

2) Dapaong is located in northern Togo. Its population is 58, 071. Its elevation is 1082 feet. It is in the Savanas region.

3) Looking at the monthly minimum and maximum temperatures of Dapaong over the last 50 years, it is difficult to find any sort of trend. After taking a closer look at the minimum monthly temperatures, it appears that November, December, and January had significantly lower temperatures before the year 1980. I think it explains why the first group's model displays a significant upward trend for minimum yearly data.

   When comparing the yearly means to the monthly means, I could only find 2 mismatches. The years 1978 and 1979 contain different values in the maximum temperature datasets. Both temperatures are only off by about 1 degree Celsius (21.1 vs 22.1 and 21.2 vs 22.1). This could have been a typo.

Dapaong Temperature Differences		
Max Data	**dataset 1**	**dataset 2**
1978	21.1	22.1
1979	21.2	22.1

4) We determined which outliers to eliminate by finding the "outer fences" for the data set. This told us to eliminate all values over a certain number and all values under another certain number. From the maximum data, we eliminated the data points from April 1985, June 1966, and June 2008. From the maximum data, we eliminated the data point for March 1995. We considered eliminating the data points from October 1995, October 1996, October 1999, and October 2000. But since those data points were not above or below the "outer fences" values, we decided to not eliminate those values.

5) For minimum temperature, we believe that there is a significant increase in temperature over time. The coefficient on the x term is positive, which would indicate that the temperature is increasing over time. Also, the t-statistic for the x term was 11.6531, which means that we are over 11 standard deviations away from zero. The p-value was very small: $1.21288*10^{-28}$. This supports the hypothesis that the x term does not equal

zero. Using this evidence, we are able to state that there is a significant increase in temperature over time.

For maximum temperature, we do not believe that there is a significant increase in temperature over time. The coefficient on the x^2 term is negative. The coefficient on the x^3 term is positive, but is extremely small. Also, by eyeballing the data graph with the model, it does not appear to be increasing.

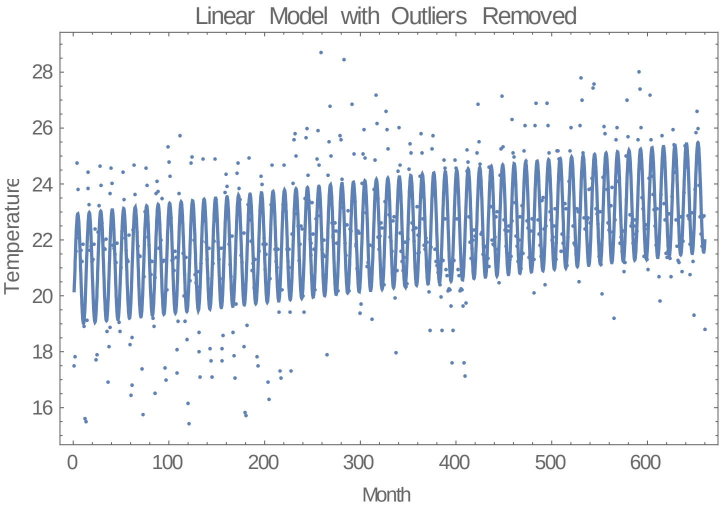
6) For minimum temperature, we believe that the best modest is the linear model. The linear model shows a significant increase over time and has a positive slope of .0038. Also, the confidence intervals did not include zero for any of the terms, which indicates that none of the terms equal zero. This was further supported by the small p-values and the large t-statistics. Also, the residuals did not show any pattern or "smiley faces." The R^2 value was a bit disappointing--it was only around .46 for the original dataset and around .48 for the dataset with the outlier removed. We would like to have seen a higher R^2 value so that we could know that the variation in our data was properly represented.

   We also considered using higher order models. We looked at the quadratic and the cubic models. We rejected the quadratic model because the confidence intervals included zero for the x^2 term (which would mean that the x^2 term probably equals zero). This was further supported by the relatively large p-value and relatively small t-statistic for the x^2 term. We rejected the cubic model because the confidence intervals included zero for the x^3, x^2, and x terms. The hypothesis that those three terms would equal zero was further supported by the relatively high p-values and relatively small t-statistics for those three terms.

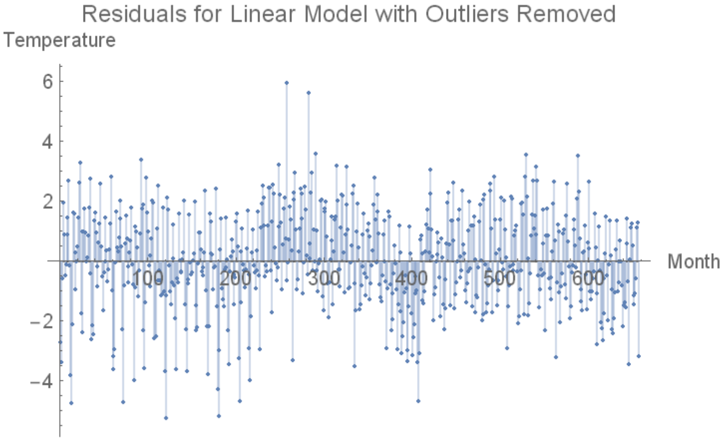
   For maximum temperature, we believe that the cubic model is the best model. We chose this model because the confidence intervals did not include zero for any of the terms, which implies that none of the terms equal zero. This was further supported by fairly small p-values and fairly large t-statistics. The R^2 value was the highest for the cubic model. For the original dataset, the R^2 value was .5076 and for the dataset with the outlier removed, the R^2 value was .6359. Of course, we would use the dataset with the outlier removed. Also, there was no pattern to the residuals for the cubic model.

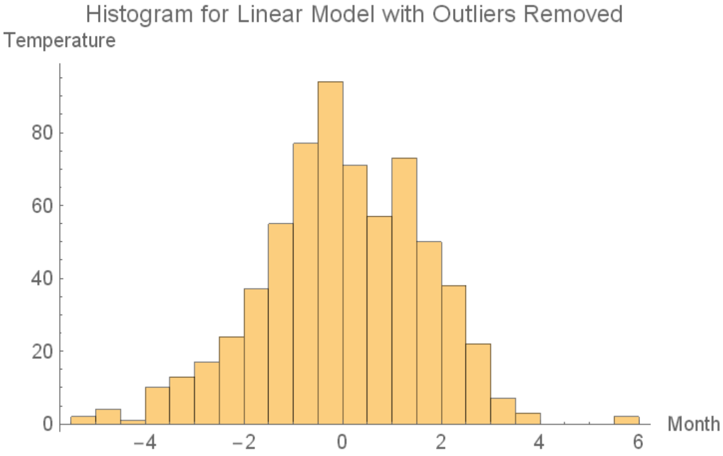
   We considered the linear and the quadratic model as well. We rejected the quadratic model because the confidence intervals included zero for the x^2 term (which would mean that that the x^2 term probably equals zero). We chose the cubic model instead of the linear model because the cubic model had a higher R^2 value. A higher R^2 values indicates that the model represents the variation in the data better than a lower R^2 value.

7) The graph of the linear model for the **minimum** dataset with the outlier removed is:

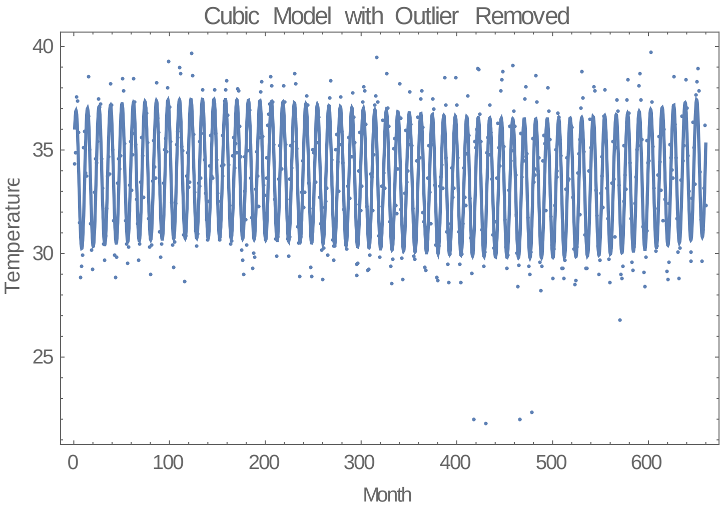


8) The graph of the residuals for the **minimum** dataset is below. We also made a histogram.

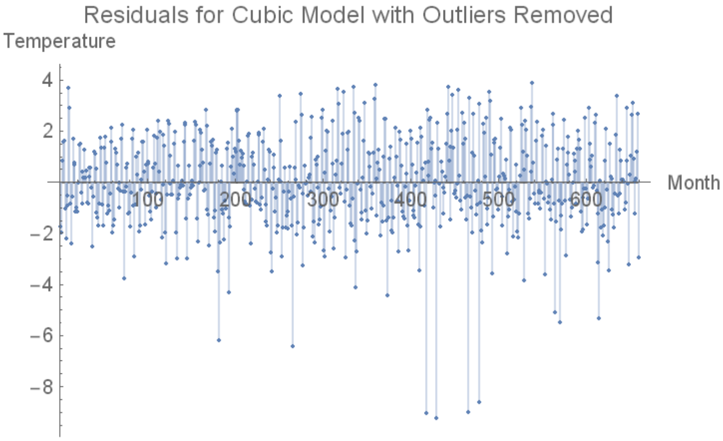


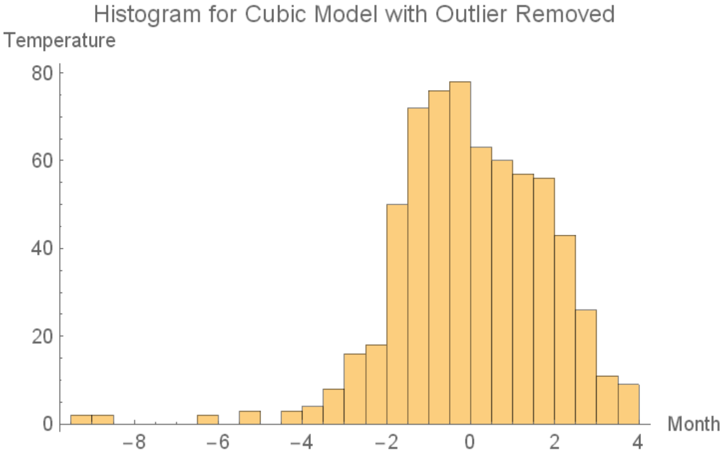


7) The graph of the cubic model for the **maximum** dataset with the outlier removed is:



8) The graph and histogram of the residuals for the **maximum** dataset are:





9) We could always use better data to strengthen our observations. Daily temperature data would give us a closer look at the temperature trends. Hourly temperature data would be even better for the same reasons. Having data that goes further back in time would help

us have a better perspective of the data. However, it may not be possible to obtain this, and the measurements become more unreliable. Other kinds of data to consider would be overcast data, humidity levels, CO2 levels, and precipitation.

The accuracy of the data could be improved as well. There are two occurrences in the maximum data that concern us. There are two consecutive years that have identical data for each month (1972-1973 & 1993-1994). The exact same thing occurred in the minimum data as well. The same years (1972-1973 & 1993-1994) again have the exact same data for each month. It would be nearly possible that there would be the exact same data for each month for two years in a row. The fact that this issue has occurred twice in both the minimum and maximum data for the same years makes us question how reliable the data is. If we are using data that is questionable from the start, our models will probably not be accurate or representative of what actually happened. There was also a 0 degrees Celsius temperature recorded in March 1995 in the monthly maximum dataset, which is clearly not accurate. If there are at least 4 rows that are obviously inaccurate, there could also be a mess of other errors and inaccuracies in the data.

In both minimum and maximum datasets respectively, some data values have one or two decimal places while others have up to 13 decimal places, which causes the data overall to be inconsistent. This lack of consistent precision can cause models to become less accurate, and it also makes us curious about how the temperatures were measured and recorded. We also wonder if the temperatures were taken by the same person each time. If the temperatures were recorded by different people, there is a possibility that the temperatures were measured and recorded differently each time. All of these concerns may have caused some problems in the data.