CS 4731: Project 1

Chase Perry

September 18, 2019

1 Problem Definition

Global warming has become a hotly debated topic these last couple of years. Governments across the globe have conducted research, passed legislature, and changed their efforts (and where their money is going) to help in the research and to combat the effects the human race is having. However, not every country is making the same effort, either due to distrust in the reports/studies or a simple lack of conviction that the studies are correct.

In this project I hope to produce a clear model of humans effect on the global average world temperature using linear regression, allowing the temperature to be estimated some years out into the future given some error. Since countries are starting to take efforts to reduce and even reverse our effect, and since the factors effecting global warming are constantly changing predictions made by the model are likely to vary in success but they'll be important nonetheless. Such factors would include research and an increase in the use of "green" energy, an increase in the industrialization of the world, more efficient cars, etc.

As global warming is something that effects are entire planet it is something that every person on the planet should be interested in. While each of us may not feel it's effects currently it has the potential to effect every aspect of life on the planet, as this is not a class issue, or race issue, or nationality issue. It is a global issue.

2 How Linear Regression Can Help Model/Predict

One issue with convincing or even educating people about global warming is visualization about the issue. Linear regression serves this effort perfectly, as it enables us to produce clear and accurate graphs of the change in temperatures over the years. Obviously, it's possible to simply look at a scatter-plot of the temperatures over the years and see some sort of general increase, but in order to truly visualize this change a graph serves best. Visualization of the change isn't the only important factor, but also having an accurate visualization is very important. It is entirely possible to simply look at a scatter-plot of temperatures over the years and draw one straight line through as many points as possible, but this doesn't accurately represent the data being shown, nor does it properly model the data. This brings me to my next point, being able to properly model and hopefully even predict global temperatures.

One of the largest fears when speaking in reference to global warming is: "What is going to happen in X years?" If we can already see a dramatic change in the average world temperature now, then what will we see in 5 years? 10 years? 20? Through this project I hope to produce a model that can answer such questions. Obviously it won't be spot on, as

to do that it would be necessary to input all kind of information that has an effect on this topic, but since that isn't possible an estimate will have to do. As stated previously, the constant change in factors such as livestock numbers, amount of renewable energy, number of fossil fuel automobiles active, and more make this a difficult process to model.

3 Data Set Source

The data used to create the model was retrieved and compiled by Earth Policy Institute, using data obtained from the National Aeronautics and Space Administration, Goddard Institute for Space Studies spanning from 1880 to 2013. The temperatures recorded are a global land-ocean average measured in Fahrenheit for a span of 133 years, allowing for a solid and trustworthy data set. Since this model will not have been created using data from 2014 and onward these years could be used a test of the model and it's reliability to see how well it's able to accurately predict the temperature at future years. The data for years 2014-2018 was obtained from the National Oceanic and Atmospheric Administration. Every month and year a report is released detailing the temperature changes on a regional and global level, as well as averages over the ocean and land both. Most often, the reports are released with temperature changes being compared to the average global temperature of the 20^{th} century.

4 Proposed Solution

If possible, I aim to use linear regression and data collected by NASA to estimate the world's temperature for future years. To start off a basic model will be used, then analyzed for how close it matched the data. After this, a series of more complex models using more complex base equations will be used to estimate the temperatures. For example, a linear system similar to that in Figure (1) will be used as a starting point, and will be added to in order to come closer to something like Figure (2) assuming this increases it's accuracy. However, "standard" polynomials like those found in these figures will not be the only base equations used to produce the linear systems. More complex polynomials will also be explored to investigate their potential use for this particular problem.

Figure (1) :
$$\begin{cases} w_0 + w_1 * x \\ w_0 + w_1 * x \\ . \\ . \\ . \\ w_0 + w_1 * x \end{cases}$$

Figure (2) :
$$\begin{cases} w_0 + w_1 * x + w_2 * x^2 + w_3 * x^3 + w_4 * x^4 + w_5 * x^5 \\ w_0 + w_1 * x + w_2 * x^2 + w_3 * x^3 + w_4 * x^4 + w_5 * x^5 \\ . \\ . \\ . \\ w_0 + w_1 * x + w_2 * x^2 + w_3 * x^3 + w_4 * x^4 + w_5 * x^5 \end{cases}$$

For this problem, the standard equation of Ax = b will be used alongside ordinary least squares to determines the appropriate weights to accurately model, and hopefully predict, the temperatures. The 'A' matrix will be produced as the collection of equations in the following format: $1 + x + x^2 + x^3 + x^4 + ...$ where x is a particular year. The 'x' matrix will be obtained used ordinary least squares to produce the best weight for each particular matrix. Finally, the 'b' matrix will be the temperatures for each year and used as the ground truth to test the accuracy of the model.

5 Solution

There are a number of keys pieces to my attempt at a solution for this problem which I will explain here, however nothing can beat looking at the full source code which will be attached to this report. To begin, each attempt to model the problem uses lambda functions similar to those below. The lambda function f2 defined on line 1 is used to produce the 'A' matrix in the linear equation Ax=b, where x is expected to be a matrix of size any number

of rows but only 1 entry per row. The lambda function f2_full is used to produce an estimate of b (known as Y hat), the expected input is the same as that for f2 but also a matrix of weights produce using ordinary least squares.

```
1 f2 = lambda x: np.power(x, [0, 1, 2])
2 f2_full = lambda x, w: w[0,0] + w[1,0]*x + w[2,0]*(x**2)
```

The following function is used in combination with the above two lambda functions, along with the matrix of years and the matrix of temperatures. As each attempt at modeling the problem requires similar steps (using ordinary least squares to estimate the weights, using these weights to estimate the temperatures) this function was heavily used throughout this project as by passing in two different lambda functions the rest of the work can be done without much extra work.

```
1 def solve(f, s, x, y):
2   x = x[np.newaxis].T
3   y = y[np.newaxis].T
4
5   weights = np.linalg.lstsq(f(x), y, rcond=None)[0]
6
7   test_output = s(x, weights)
8
9   return test_output, weights
```

Finally, the last part of the source code that I feel deserves focus here is the following section. This function expects two matrices of the same size and uses mean squared error to quantify the loss between the estimate (which is expected to be the first matrix) and the ground truth (the second matrix).

```
1 def loss(yhat, y):
2   return np.sum((yhat - y)**2)/y.size
```

6 Evaluation

While attempting to model the data used there were 3 major styles used, with some variations, each of course with its own level of success and failure. These styles were to use "regular" polynomials to model the problem similar to that in Figure (3), "inverse" polynomials similar to that in Figure (4), and "mixed" polynomials similar to that in Figure (5).

Figure (3)= $w_0 * x^0 + w_1 * x^1 + w_2 * x^2 + w_3 * x^3$

Figure (4)= $w_0 * x^0 + w_1 * x^{-1} + w_2 * x^{-2} + w_3 * x^{-3}$

Figure (5)= $w_0 * x^0 + w_1 * x^{-3} + w_2 * x^{-2} + w_3 * x^{-1} + w_4 * x^1 + w_5 * x^2 + w_6 * x^3$

For the each "style" of equation, attempts ranged from 1^{st} degree polynomials to 6^{th} degree polynomials, for the "inverse" polynomials, this refers to the absolute value of the minimum exponent (for instance, Figure (4) is considered to be of 3^{rd} degree).

Generally speaking, while researching this topic graphs similar to that found in Figure (6) was predominantly found, which is the plot of the "normal" 1^{st} degree attempt. This is somewhat ironic as out of the "normal" styles a linear equation had the greatest mean squared error at around 0.067. Meanwhile, a 5^{th} degree "regular" polynomial was found to best model the data set with a mean squared error at around .0377, matching the data set nearly twice as well which can be seen in Figure (7).

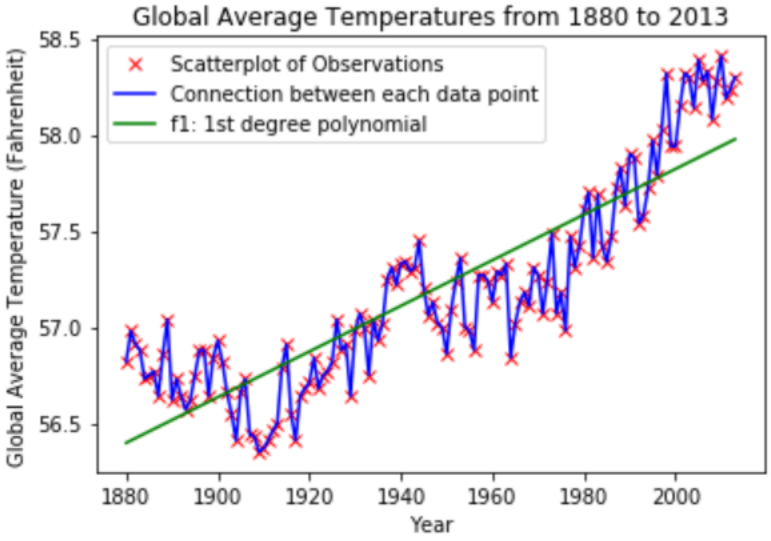

Figure (6): 1st Degree "Regular" Polynomial

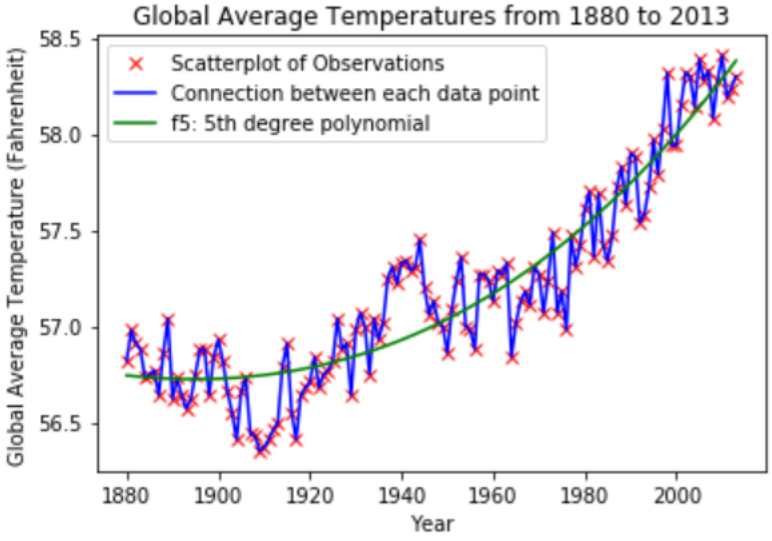

Figure (7): 5th Degree "Regular" Polynomial

Out of all "styles" and attempts for each style, the 5^{th} degree "mixed" polynomial best fit the data set with a MSE (mean squared error) just barely above that of the "regular" 5^{th} degree polynomial. In fact, the difference between their errors is around $2*10^{-15}$. The "mixed" styles were consistently more accurate than their "regular" degree counterparts, except for the 1^{st} and 2^{nd} degree attempts. The "inverse" styles were consistently less accurate than either of the other styles, being on average most inaccurate when compared to the original data set.

However, while "inverse" style was more inaccurate than the other two when compared to the 1880-2013 data they actually predicted the temperatures for 2014-2018 better than any of the other styles or attempts, with the exception of the 1^{st} degree "inverse" attempt which was last in accuracy. The 3^{rd} degree "inverse" attempt predicted the temperatures best with an error of around .0607, very closely followed by the 4^{th}, 5^{th}, 6^{th}, and 2^{nd} degree attempts. Similar to the original data set the "mixed" and "regular" styles had very close errors for their predictions with the "mixed" style normally being very slightly more accurate. Figures (8) and (9) plot each function with respect to its error to the original data set and new predictions, with the error in respect to the modeling data being on the x-axis and the new prediction errors being on the y-axis. Figure (8) represents the "regular" style and (9) represents the "mixed" style, a figure for the "inverse" was not made due to the errors being nearly the same for all attempts except for the 1^{st} degree. In general the 1^{st} degree attempts were not plotted as they were drastically different in error to the other attempts in the same style that had a greater degree.

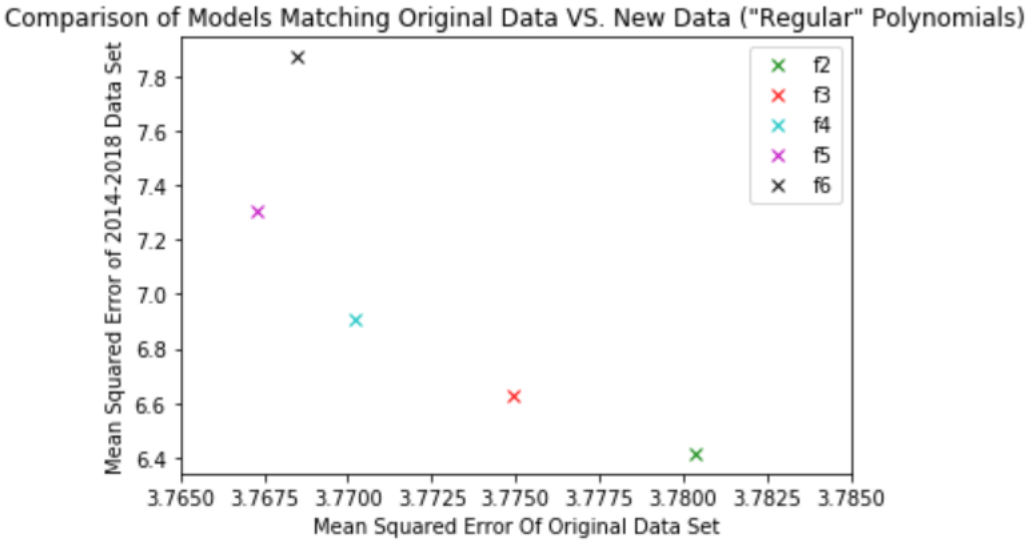

Figure (8): Regular Attempts and Their Errors

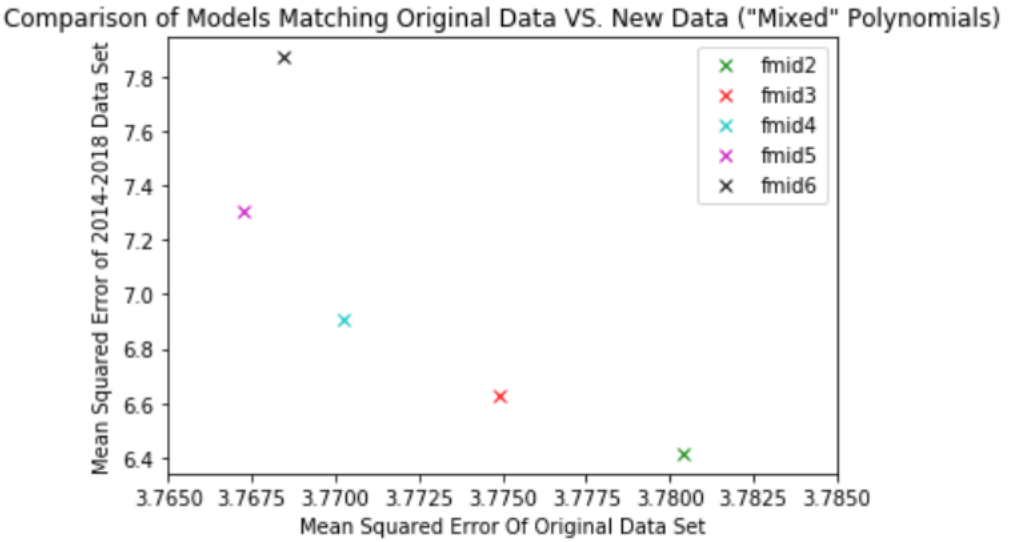

Figure (9): Mixed Attempts and Their Errors

7 Discussion

For the sake of simplicity and since this was an initial effort to model the global change in temperature the degree of the polynomials used was kept low in order to be more manageable. Similarly, for the "mixed" style of attempts the degrees match up (-1 to 1, 2 to -2, -3 to 3 and so on) as applicable so that the number of possible equations didn't grow too large. This forced system on the data seemed to model the data well, but as was shown by the "inverse" style it didn't hold up when it came to making predictions on the 2014-2018 data.

As stated above, while some error is to be expected with both the original data set and the predictions for future years there are some explanations to find why this error might error and where it might come from. As the problem of global warming is a problem caused by man, and thus is constantly changing, it is impossible to model completely perfectly using linear regression. Factors such as amount of livestock, amount of renewable energy, weather systems, number of fossil fuel vehicles, amount of travel at both regional levels and globally all have an effect on this problem. Anything that could contribute to greenhouse gas creation would generate some error, and accurately quantifying these emissions would be nearly impossible.

8 Future Work

Future work in this area could branch off into two main styles in my opinion, one being more complex equations to model the changes in temperature as being dependent on the year while the other would be to attempt to model the temperatures as a dependent result of some combination of a number of variables. Using more than just exponents would be a good place to start, for example using logarithmic or trigonometric functions, to produce the 'A' matrix in Ax = b. Other than using these other functions, trying different variations of the "mixed" style would also be a fair place to start as it would allow for more variations and possibilities in the produced weight matrix.

Similar to some of what was referenced in Section 7, attempting to model the change in temperature as more than just being dependent on the year would also be an interesting area. For example, one such model could attempt to model the temperature as being dependent on both the year and estimated population during that year. Incorporating more variables could also potentially introduce new areas for error, such as not properly estimating the population, but overall I believe the gain would likely outweigh the risk.

Of course, as time progresses rebuilding each model with the new data from each passing year would be a simple starting point that could potentially offer enormous benefits. While the models built used 133 data points and made estimates on an additional 4 years, I believe that even more accurate predictions could be made if the original data set to build the models was larger. It obviously isn't possible to go back and record this data, so only time will tell as we move forward.

References

[1] Larsen, J. (2014, February 4). 2013 Marked the Thirty-seventh Consecutive Year of Above-Average Temperature. Retrieved from http://www.earth-policy.org/indicators/C51

[2] NOAA National Centers for Environmental Information, State of the Climate: Global Climate Report for Annual 2014, published online January 2015, retrieved on September 21, 2019 from https://www.ncdc.noaa.gov/sotc/global/201413.

[3] NOAA National Centers for Environmental Information, State of the Climate: Global Climate Report for Annual 2015, published online January 2016, retrieved on September 21, 2019 from https://www.ncdc.noaa.gov/sotc/global/201513.

[4] NOAA National Centers for Environmental Information, State of the Climate: Global Climate Report for Annual 2016, published online January 2017, retrieved on September 21, 2019 from https://www.ncdc.noaa.gov/sotc/global/201613.

[5] NOAA National Centers for Environmental Information, State of the Climate: Global Climate Report for Annual 2017, published online January 2018, retrieved on September 22, 2019 from https://www.ncdc.noaa.gov/sotc/global/201713.