MIT OCW 6.0002

Introduction to Computational Thinking and Data Science

Problem Set 5: Experimental Analysis

Gabriel Munoz

22 September, 2021 to 29 September, 2021

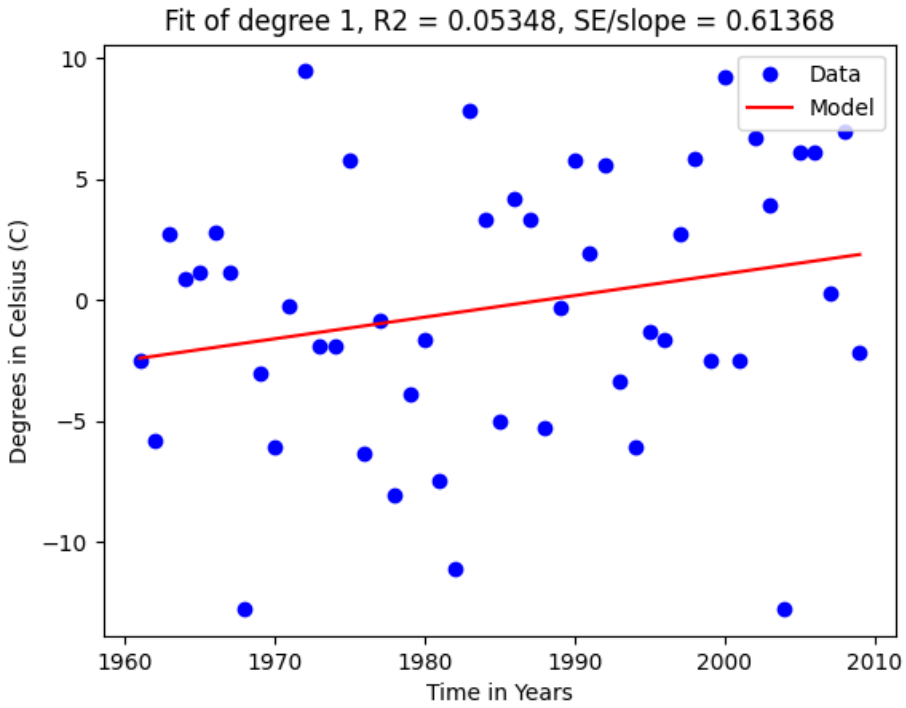
Part A: Creating Models

Problem 4: Investigating the trend

- What difference does choosing a specific day to plot the data for versus calculating the yearly average have on our graphs (i.e., in terms of the R2 values and the fit of the resulting curves)? Interpret the results.
- Why do you think these graphs are so noisy? Which one is more noisy?
- How do these graphs support or contradict the claim that global warming is leading to an increase in temperature? The slope and the standard error-to-slope ratio could be helpful in thinking about this.

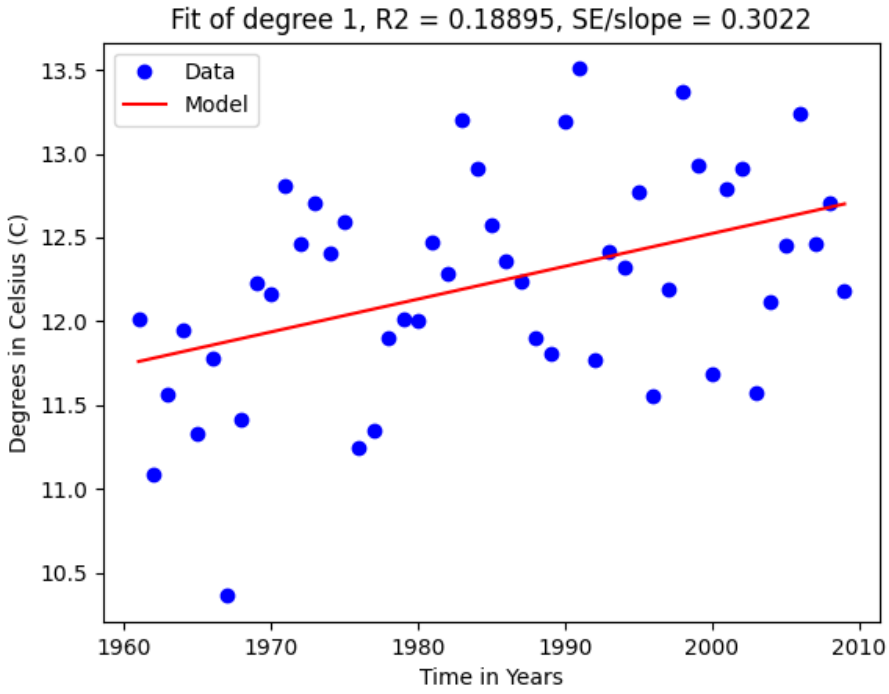
I. January 10th

Temperature on January 10th in New York City from 1961-2009



II. Annual Temperature

Average Yearly Temperature in New York City from 1961-2009



Choosing a specific day to plot the data for leads to a very noisy graph and a poor linear fit to the data, with an R-squared of 0.053. That is, the model only accounts for about 5.3% of the variability in the data. Using the average yearly temperature for that city instead of a single day--particularly January 10th, a day straight in winter--is more representative of the climate across years. It would capture the warming effects seen most starkly in the hotter seasons and therefore a better fit to the data is expected. Indeed, using the average yearly temperature leads to a linear model with an R-squared of 0.189 that captures a more defined upward trend in the temperature across time and 18.9% of the variability in the data.

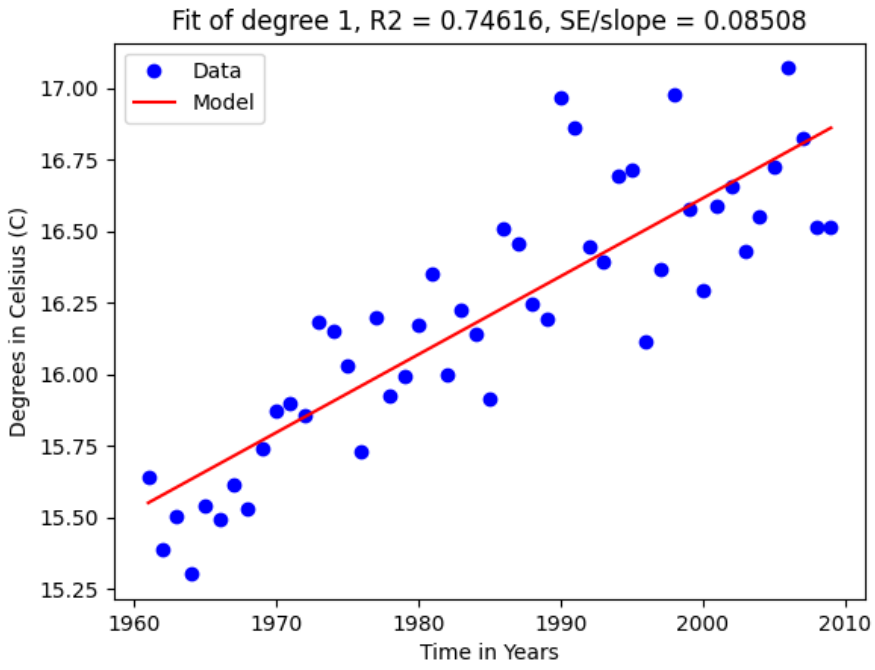
The graphs are quite noisy mainly because we are only using the temperature of one city (New York) which has a high variability in its climate every year, with many outliers. The first graph is more noisy than the one using average yearly temperatures because changes in temperature from year to year probably don't correspond day to day, but rather to similar time periods--in other words, if January 10th of 1975 is a given temperature, we can expect to see a higher temperature 5 years later, but not necessarily on the 10th, but perhaps on January 11th or 9th or so of 1980.

The standard error/slope ratio for the first graph indicates the slight trend captured there is likely just by chance, but the trend captured by the linear model on the average yearly temperature is very likely not by chance, with a value well below 0.5 at 0.302 (compared to 0.614).

Part B: Incorporating More Data

- How does this graph compare to the graphs from part A (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
- Why do you think this is the case?
- How would we expect the results to differ if we used 3 different cities? What about 100 different cities?
- How would the results have changed if all 21 cities were in the same region of the United States (for ex., New England)?

Average Yearly Temperature Across U.S. (21 Cities) from 1961-2009



Compared to the graphs from part A, the R-squared is dramatically higher and the standard error-to-slope ratio is much lower. Those measures, combined with the visually distinct trend of the data and the fit generated, suggest global warming is observed and not random. This is due partly to having more data to train the model on, and further, to having a wider breadth of

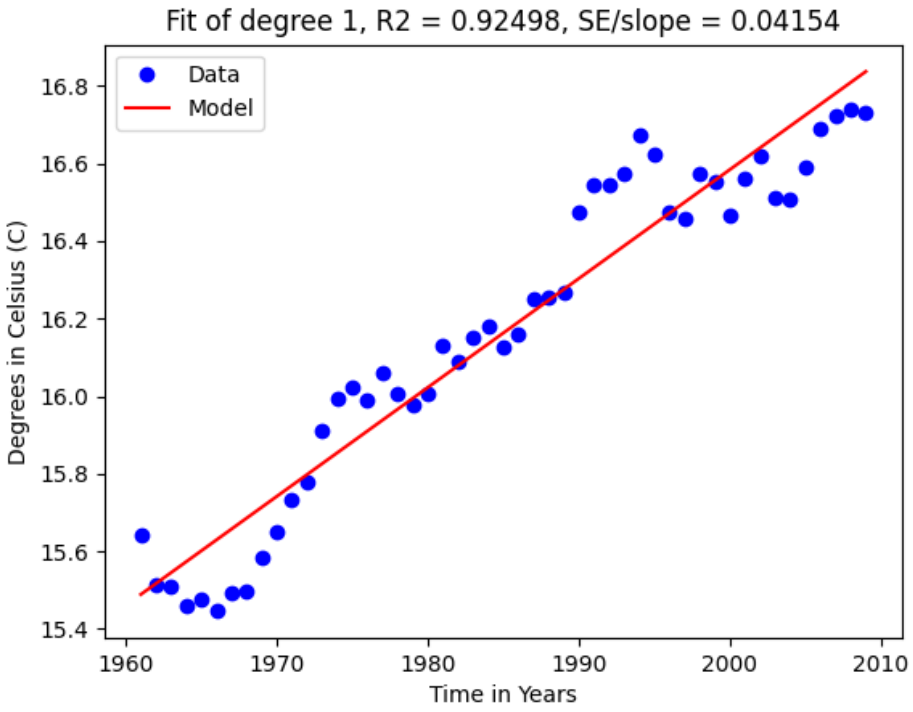
climates represented by the different cities in which the effects of climate change are more adverse and apparent.

Had we used 3 cities, we would have a better model than with just NYC, but not as good a model as the one with 21. If 100 cities were used, I suspect there would be diminishing returns, potential overfitting, and we would have a model with higher R-squared and lower SE/slope, but not necessarily a more useful one. On this note, had we used 21 cities all in the same region of the United States, we would be getting something very similar to the NYC models, but better due to having more data. It wouldn't be as representative of the whole country and not as predictive for unseen data since the training data would not have much geographic variation.

Part C: 5-year Moving Average

- How does this graph compare to the graphs from part A and B (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
- Why do you think this is the case?

5-Year Moving Average for U.S. Average Yearly Temperature from 1961-2009



The graph for the 5-year moving average has the highest R-squared and lowest SE/slope yet. It captures about 92.5% of the variation over time. The 5 year average fluctuates up and down from year to year, but consistently trends upward. This window used in a moving average smoothes out the data (less noise, fewer outliers, etc.). It allows us to focus on the trend and not the year to year fluctuations. This linear fit strongly supports the global warming claim, considering only the U.S. data that we have.

Part D: Predicting the Future

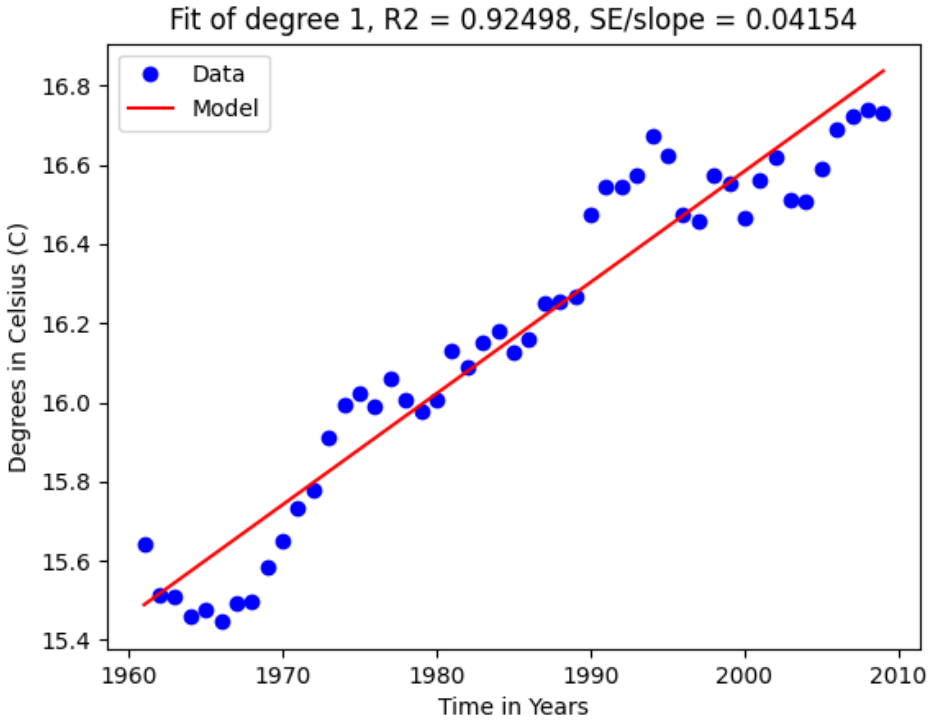
Problem 2: Predicting

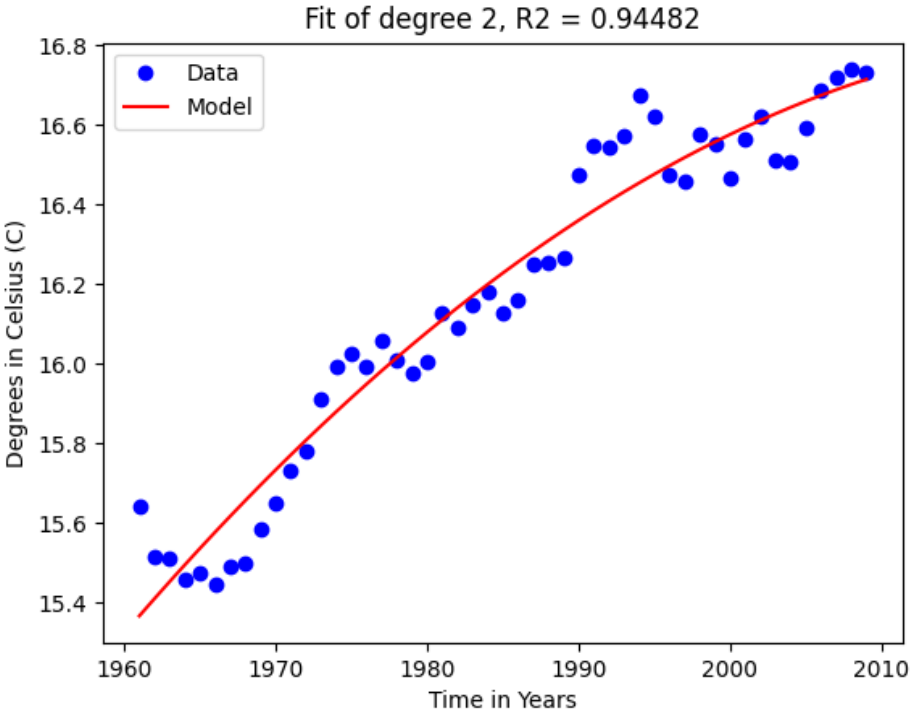
I. Generate more models

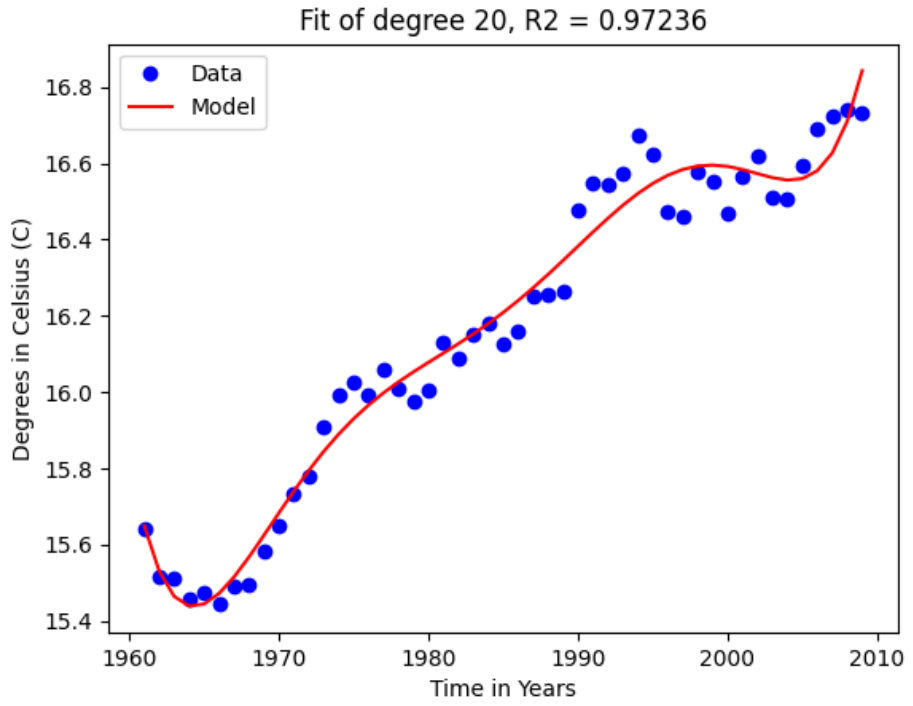
- How do these models compare to each other?
- Which one has the best R2? Why?
- Which model best fits the data? Why?

5-Year Moving Average for U.S. Average Yearly Temperature from 1961-2009

With Polynomial Fits of Varying Degrees (1, 2, 20)







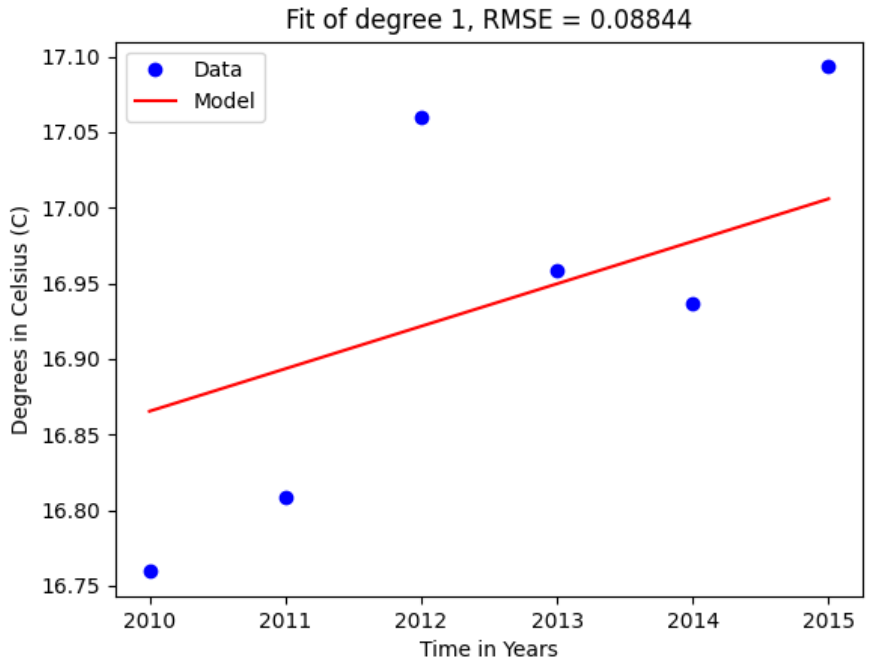
The first graph is the exact same one as the graph from C. The following two graphs are for the same data, but with a polynomial fit of degree 2 and 20. As expected, a higher degree polynomial

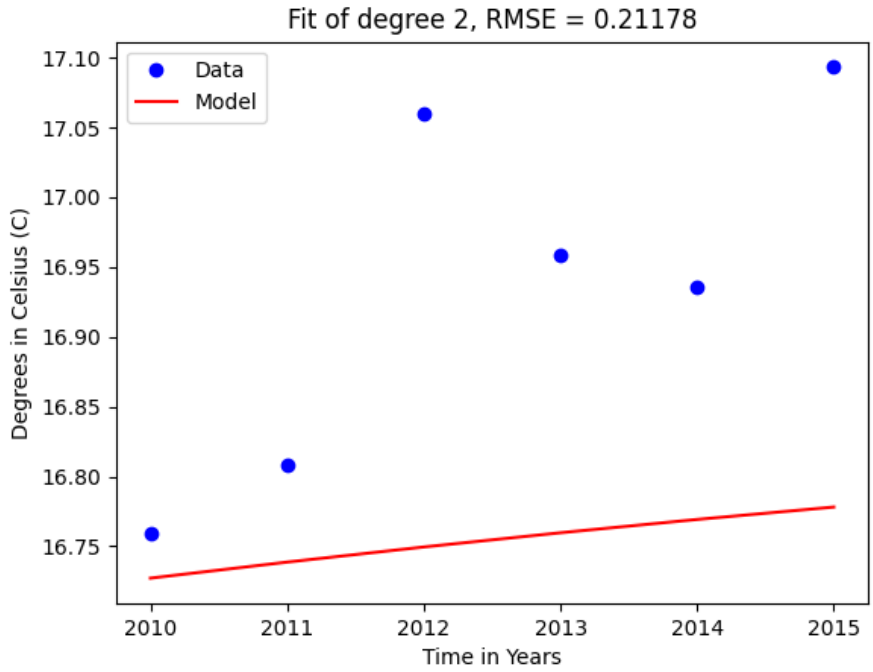
fit yields a higher R-squared and we indeed see that in the graphs, especially the one for degree 20, as the curve almost mirrors the data, reducing the error for every point. However, a greater R-squared is not always desirable nor does it make the model more useful or accurate. After all, the polynomial fit is to the training data and what we are really interested in is the model's performance on the unseen test data. A fit of degree 20 certainly will have overfitting and not be very accurate in testing. That leaves degree 1 and 2 and while the degree-2 fit indeed has a higher R-squared, it doesn't appear to be a significant difference, and it is unclear whether it has greater predictive power than the linear fit. The linear fit is probably the most adequate.

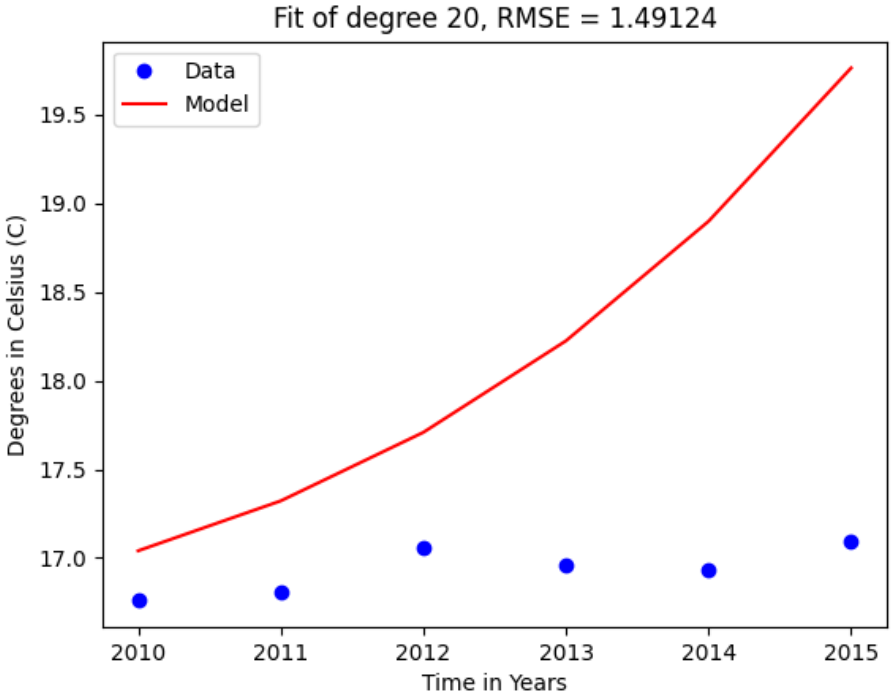
II. Predict the results

- How did the different models perform? How did their RMSEs compare?
- Which model performed the best? Which model performed the worst? Are they the same as those in part D.2.I? Why?
- If we had generated the models using the A.4.II data (i.e. average annual temperature of New York City) instead of the 5-year moving average over 22 cities, how would the prediction results 2010-2015 have changed?

Predicting U.S. Average Yearly Temperature from 2010-2015 using Models Trained on 5-Year Moving Averages from 1961-2009 (Polynomial Fits of Degrees 1, 2, 20)







The model of degree 1 performed the best, with the lowest RMSE (Root Mean Square Error). This was expected since the other two models of higher degrees would be more tightly fit to the

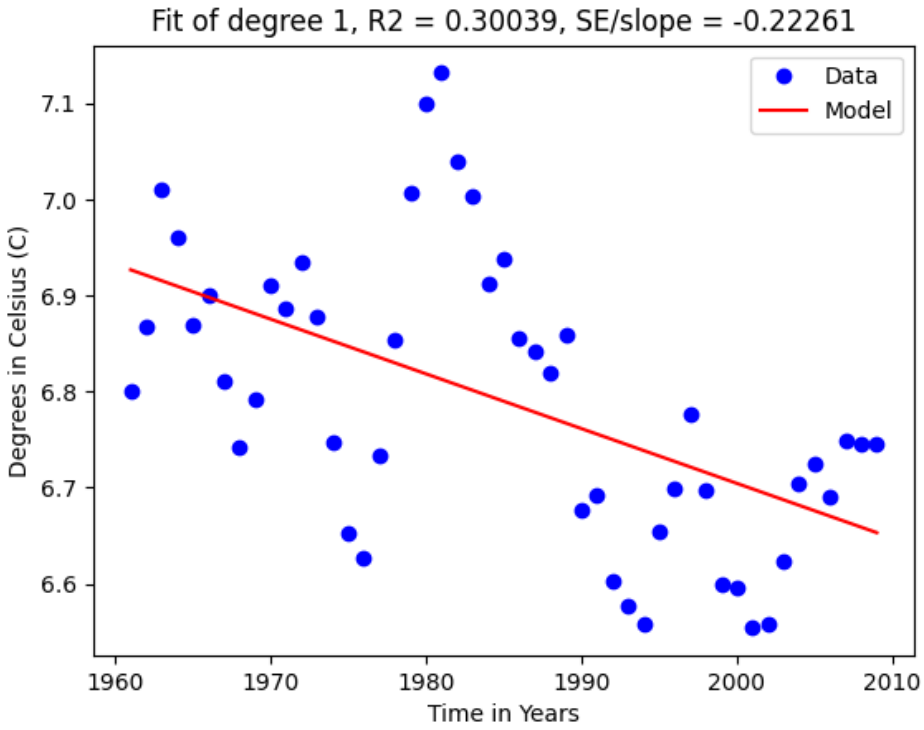
training data, resulting in models with some overfitting--in the case of the degree 20 model, a high degree of overfitting--and hence not useful for making predictions.

Had we used average yearly temperatures in New York City (A.4.II data) to generate our models and tried to predict the temperatures for 2010-2015, our predictions would be even worse. This is because we'd be trying to predict average temperatures across the whole country, with all of its climate variability across regions, using models that only learned from a single location in the country, New York City. That would not be representative of the whole country and therefore highly inaccurate predictions are expected.

Part E: Modeling Extreme Temperatures

- Does the result match our claim (i.e., temperature variation is getting larger over these years)?
- Can you think of ways to improve our analysis?

5-Year Moving Average of the Standard Deviations of U.S. Average Yearly Temperatures from 1961-2009



Based on this analysis, it appears the claim of increasing temperature variation is not true. While the R-squared is not exactly high, it is not insignificant, and the SE/slope is low enough that the trend observed appears to be non-random. This analysis, based on the average yearly temperatures of all 21 cities we have data for, seems to suggest that temperature variability is actually decreasing. What we do know about climate change is that there are more frequent extreme hot weather events such as droughts, heat waves, storms, and floods, which are associated with hot temperatures. This makes sense given that the rightward shift of the distribution of temperatures we are observing--global warming--would lead to more frequent extreme hot weather and less frequent extreme cold weather (relative to what we used to consider extreme hot and extreme cold).

In order to improve and further explore this claim of increasing temperature variation, we could look at particular regions of the U.S. where the effects of climate change might currently be more pronounced. Furthermore, we can visualize the temperature distribution and how it has been shifting over time (and as our claim would imply or our current analysis suggests, flattening or squeezing, respectively) as well as searching more granularly the frequency of extreme weather events (for example, days in which the temperature exceeded the mean by 2 or 3 standard deviations in either direction).