Problem Set 5

John L. Jones IV

MIT 6.0002 as Taught in Fall 2016 - July 10, 2020

Part A: Creating Models

Problem 4. Investigating the trend

Problem 4.I January 10th

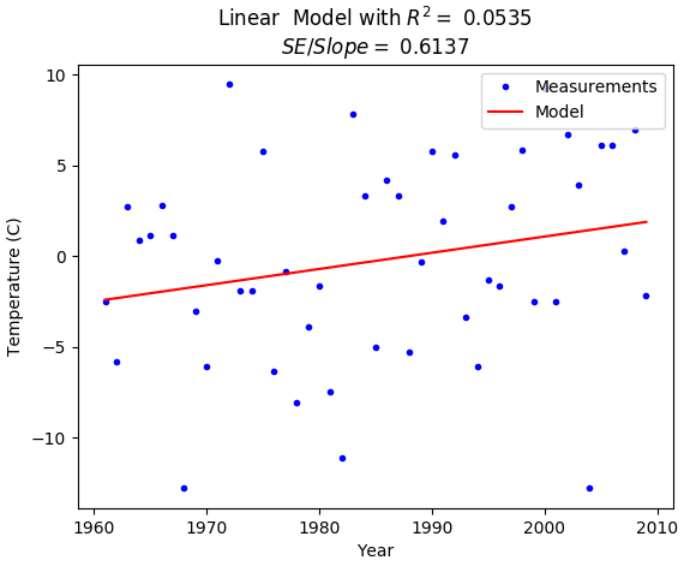

Figure 1: Temperature on January 10th in New York through training years.

The best fitting linear model for a random day (January 10) resulted with $R^2 = 0.0535$ as shown in Figure 1. The $SE/slope$ for Figure 1 is 0.6137 for the purposes of this study $R^2 > 0.5$ indicates the trend is likely due to chance. When observing climate change, a yearly average should remove some of the "day-to-day" noise. Weather is more difficult to predict than climate. Daily weather changes introduce high frequency noise. Average with respect to time acts as a low-pass filter.

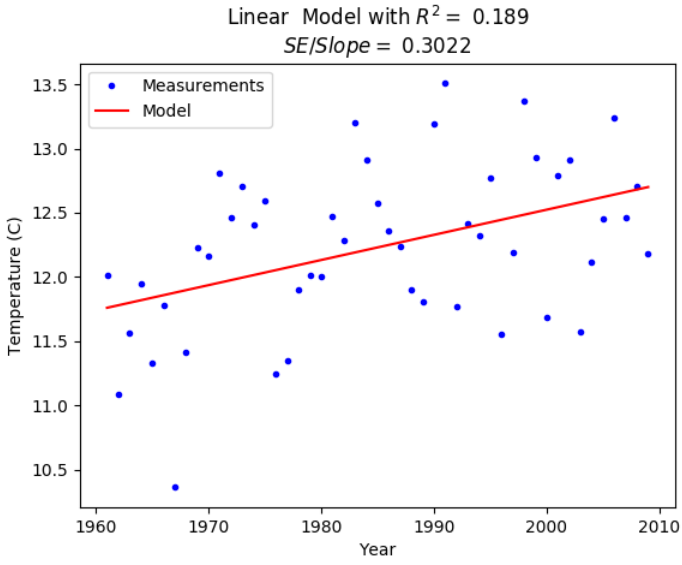

Figure 2: Annual average temperature in New York through training years.

Problem 4.II Annual Temperature

The same model-generating algorithm was used on the yearly average data and produced a best fitting line with an $R^2 = 0.189$ as shown in Figure 2. The $SE/slope$ for Figure 2 is 0.3022 indicating the trend may not be due to chance. There is likely less variance due to noise in the yearly average data allowing a trend over time to be more apparent. Both Figure 1 and 2 indicate temperate increasing on the order of $\frac{1}{2}$ degrees Celsius per 50 years. The annual temperature study more strongly supports this trend than the random day study.

Part B: Incorporating More Data

When including more cites a better picture of the global climate can be observed, especially cities that are geographically diverse. Much of the noise due to local weather variations can be averaged across multiple cities located in different regions. One would expect as more geographically independent cities are used in the model, the less power local weather noise has. There appears to be much less variance in this data compared to the single-city data previously analyzed. The $R^2 = 0.7462$ and $SE/Slope = 0.0851$ shown in Figure 3 make a strong argument that this model for these cities indicating increasing yearly temperate is not likely to be due to chance or randomness.

Part C: 5-year Moving Average

More high frequency noise can be removed with a moving average filter in addition to geographic averaging. We are only interested in the long term climate change so any loss of high frequency signals such as weather variations or an outliers is acceptable. The $R^2 = 0.925$ and $SE/Slope = 0.0415$

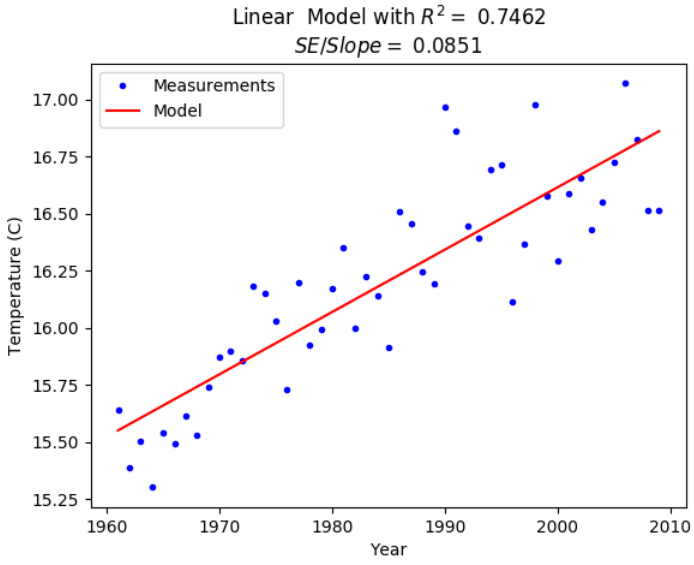

Figure 3: Yearly average across multiple cities.

shown in Figure 4 make a strong argument that this model for these cities indicating increasing yearly temperate is not likely to be due to chance or randomness. This stronger R^2 and $SE/Slope$ is likely the result of reducing variance in the training data with the moving average filter.

Part D: Predicting the Future

Problem 2.I Generate More Models

The 20-degree model shown in Figure 6 has the lowest $R^2 = 0.9724$ value. A higher R^2 value indicates the model variance is small in comparison to the data variance i.e a better fit. However, the 20-degree model likely has been over-fit to the training data. It is very unlikely the underlying system of climate has any x^{20} terms. Although the 20-degree model closely fits the training data, new climate data may not be accurately predicted with this high-order model. The linear and quadratic models shown in Figures 4 and 5 seem to be the simplest and best solutions for predicting future values, despite the slightly lower values of $R^2 = 0.925$ and $R^2 = 0.9448$ respectively.

Problem 2.II Predict the Results

When tested against future data, the linear model results in the lowest Root-Mean-Squared Error $RMSE = 0.3687$. The 2-degree model also performs well with $RMSE = 0.4617$. The 20-degree model quickly diverges away from the future data as seen in Figure 9. When evaluated using future data, the linear model appears to be the best predictor of climate change over time. As demonstrated here, a model that is the best fit for a set of training data, does not always perform the best at predicting future data. If the models were trained using only the unfiltered temperatures of New

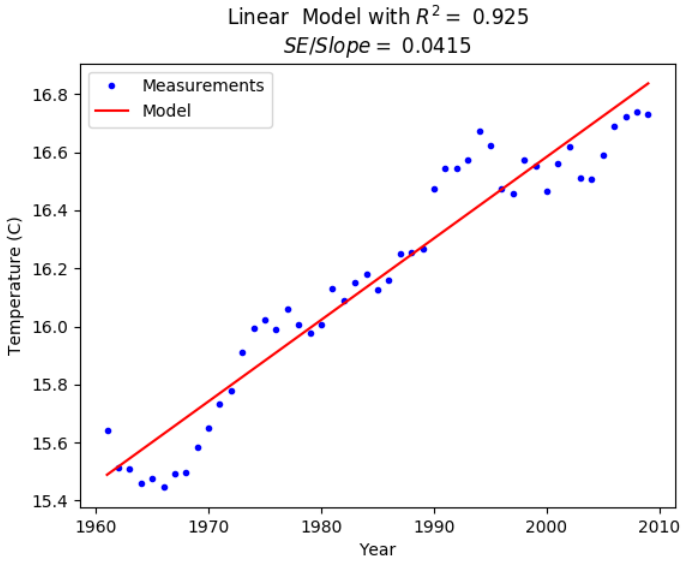

Figure 4: Yearly average across multiple cities using 5-year moving average filter.

York, there would likely be a "DC" bias since NYC seems to have a lower average temperature than world average temperature. Out of curiosity, I plotting the result in Figure 10. This same linear model from average New York data does a better job predicting temperatures in New York, see Figure 11. There is still a much higher $RMSE$ than the models produced using many cities.

Part E: Modeling Extreme Temperatures

This result does not support the claim that temperatures are getting more extreme. The negative slope in Figure 12 indicates there is less variance in temperature as time progresses. It may be better to average the standard deviations of locations instead of taking the standard deviation of the averages. The average temperature of many cities to indicate may not indicate extreme changes within the individual cities.

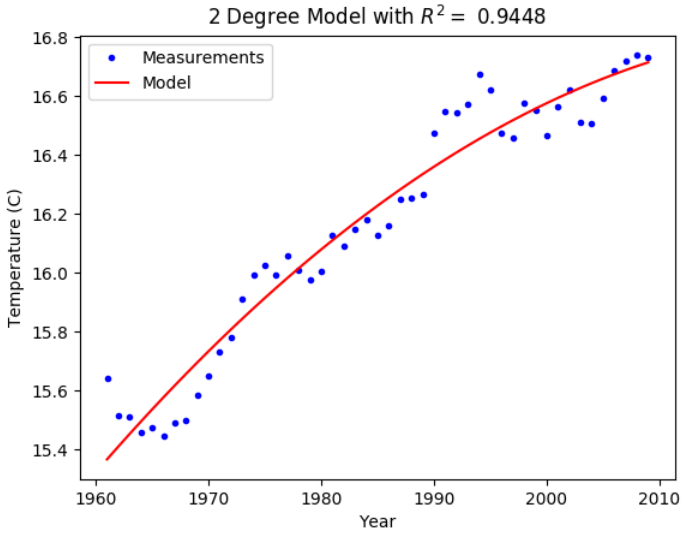

Figure 5: Training evaluation with 2-degree model.

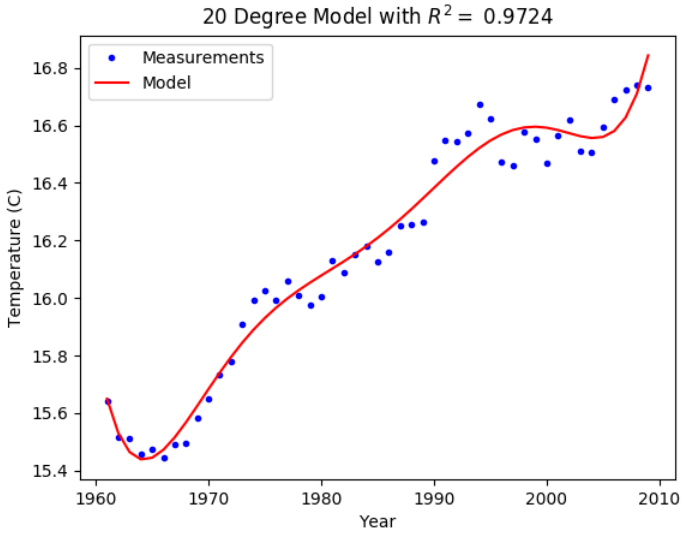

Figure 6: Training evaluation with 20-degree model.

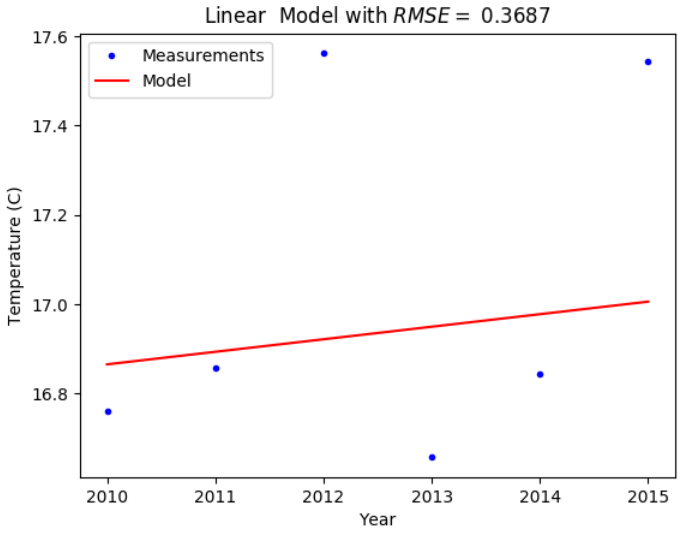

Figure 7: Testing evaluation with linear model.

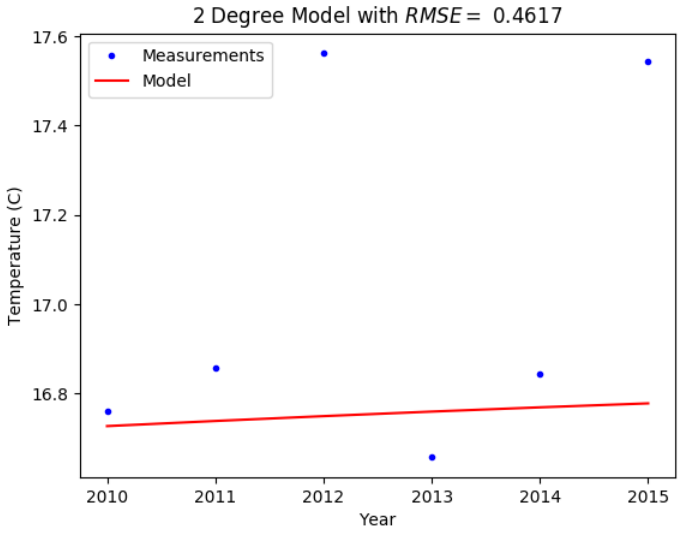

Figure 8: Testing evaluation with 2-degree model.

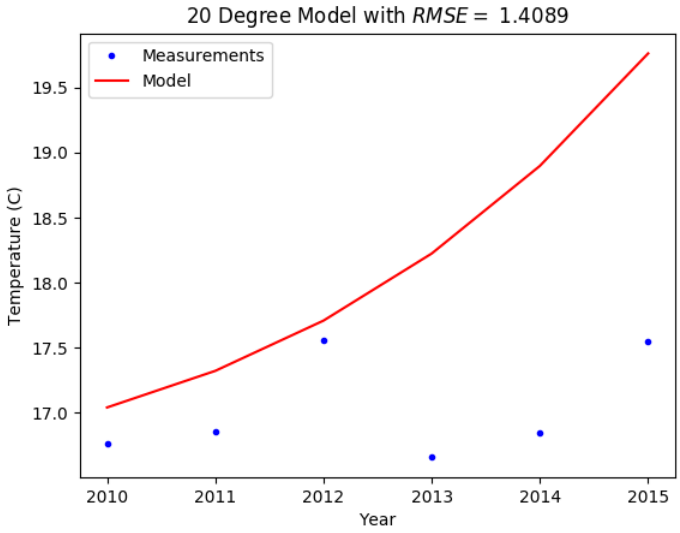

Figure 9: Testing evaluation with 20-degree model.

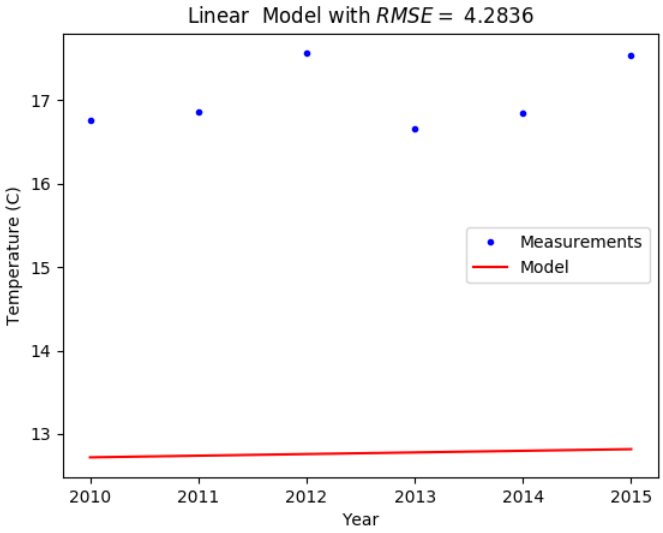

Figure 10: NYC annual average model tested against world average temperatures.

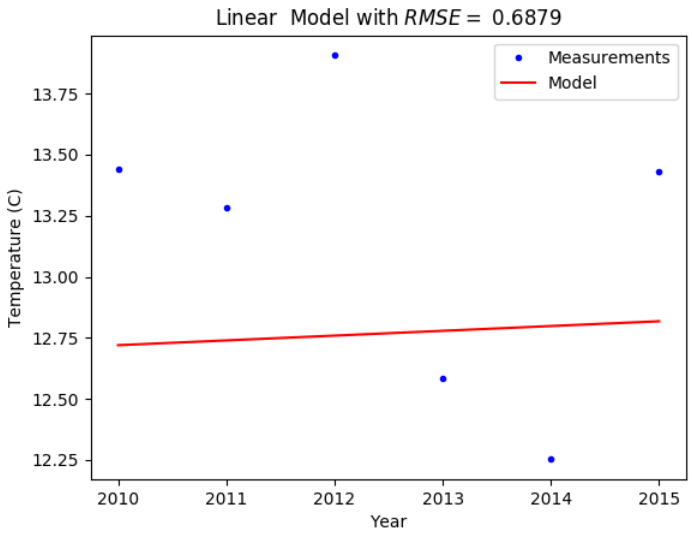

Figure 11: NYC annual average model tested against NYC average temperatures.

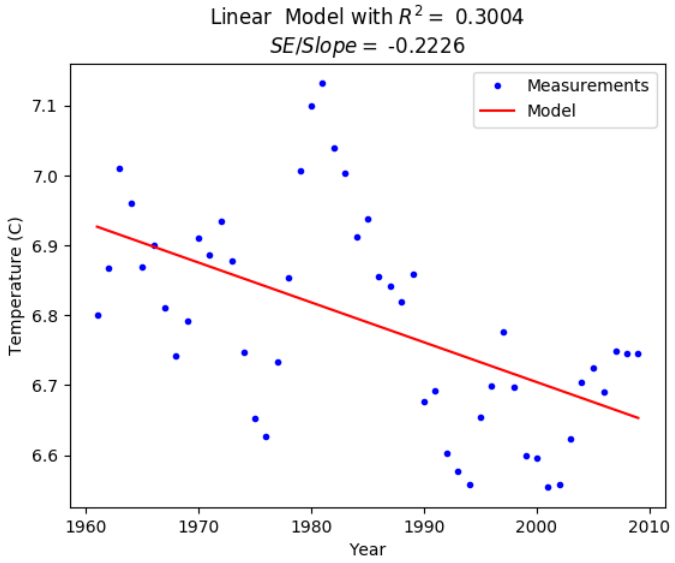

Figure 12: Model trained on temperature standard deviation over multiple cities.