6.0002 MIT OCW Fall 2016

Problem Set 5 Write-Up

Octavio Vega

Problem 4A

Figure A.4I: *Average temperatures on January 10 from 1961 to 2009 in New York City*

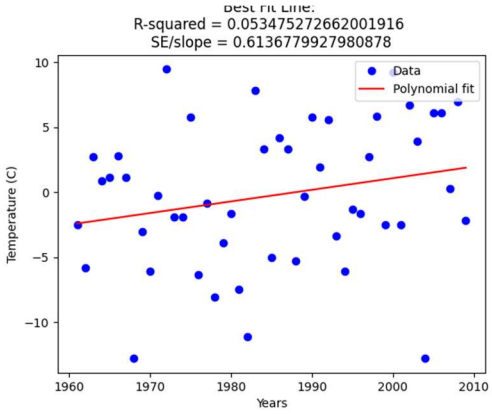
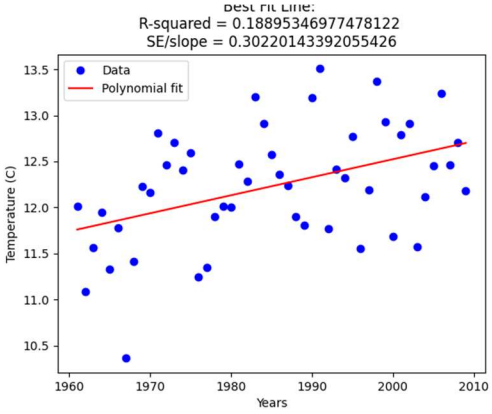


Figure A.4II: *Average annual temperatures from 1961 to 2009 in New York City*

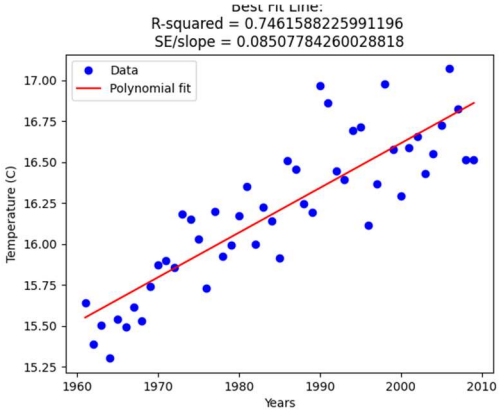


- Comparing the figures above, we see that choosing a specific day to plot the data results in a lower R^2 value than if we average over the entire year. The fit may be better in the latter case because we average over a greater amount of data, meaning that there is less random noise when we look at yearly averages versus choosing a single day, where there can be more variability in temperatures.
- Both of these graphs are likely so noisy because temperature is a noisy source of data; many random variables beyond weather contribute to average temperatures, such as pollution, machinery, density of crowds, etc.

- The fact that we see a positive slope in both plots for the best linear fit to the average temperature data over time seems to support the idea that global warming is leading to an increase in temperature. However, the standard error-to-slope ratios in both plots are large enough that there is some significant uncertainty in how well this model captures the trend in temperature changes. While not necessarily contradicting the hypothesis that global warming is leading to increase in average temperature, this does suggest that there could be other factors contributing to this data trend.

Problem 4B

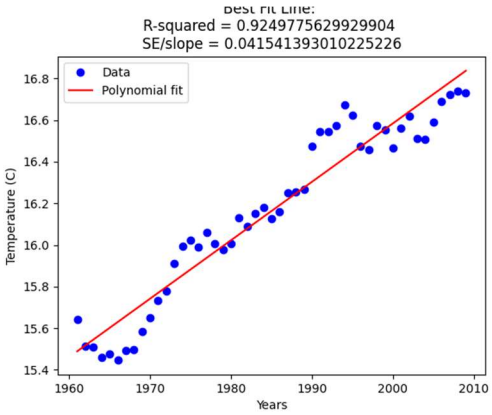
Figure B.4: *Average annual temperatures from 1961 to 2009 over all cities*



- Comparing this graph to those in part A, we see a much better fit, with a higher R^2 value and lower standard error-to-slope ratio. This graph supports much more strongly the claim that global warming is contributing to an increase in average temperature.
- Since we are no longer relying on data from a single city but rather considering all cities in our average, we have a stronger conclusion that this model fits the data well.
- If we used three different cities, we would be closer to the case in part A with fewer data points so we would expect noisier data and a worse fit. If we used 100 different cities, we could expect a better fit and less noisy data.
- If all 21 cities were in the same region of the US, then our conclusion would not be as strong. The fit might still be tighter, but the averages might be shifted down on the plot due to lower average annual temperatures in New England. This also risks not responding to the research question, since this only investigates a specific region of the world.

Problem 4C

Figure C.4: *Moving average with window size 5 of average annual temperatures over all cities*



- Compared to the graphs in parts A and B, this graph shows a better fit with a higher R^2 value and lower standard error-to-slope ratio. There is a clear upward trend in the annual temperature over time. While this does not directly contradict the hypothesis that global warming is increasing annual temperatures, once again it is difficult to claim that this model necessarily supports the claim about global warming.
- This is because here we are viewing a moving average, which means that each data point is an average of itself and (in this case) the previous five data points. Since the averages are taken in a moving window, the data points will be clustered together more tightly by virtue of the arithmetic mean bringing points closer together.

Problem 4D.I

Note: In this section, "training data" refers to the 5-year moving average of annual average temperatures over all cities from 1961 to 2009.

Figure 4D.Ia: *Best fit line to training data*

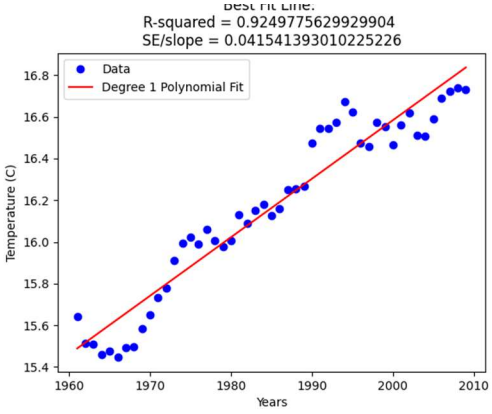


Figure 4D.Ib: *Best fit parabola to training data*

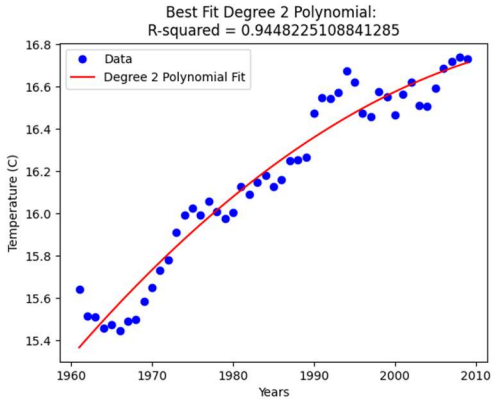
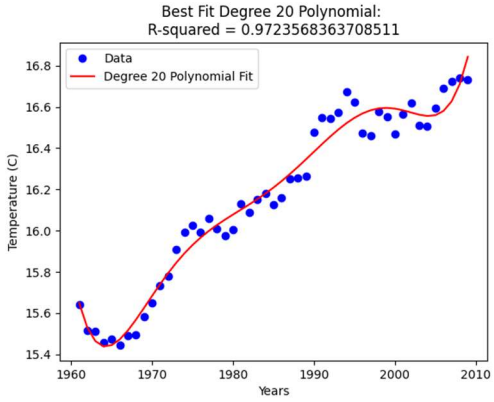


Figure 4D.Ic: *Best fit degree-20 polynomial to training data*



- These models all demonstrate good fits to the data, and the goodness of fit increases as the degree of the model increases.
- The degree-20 model has the best R2 value of approximately 0.972. This is because the parameter space of the model has dimensionality 21, meaning there are 21 free parameters to be fit to the data, giving this model the most freedom to approximate the data. The line and parabola each have two and three free parameters, respectively, to tune to the data, so they have less freedom.
- For the reasons explained above, the degree-20 model best fits the data. However, there is risk of overfitting since this model could be too complex for the data and only be fitting well since there are many more degrees of freedom.

Problem 4D.II

Note: In this section, "testing data" refers to the 5-year moving average of annual average temperatures over all cities from 2010 to 2015.

Figure 4D.IIa: *Best fit line to testing data*

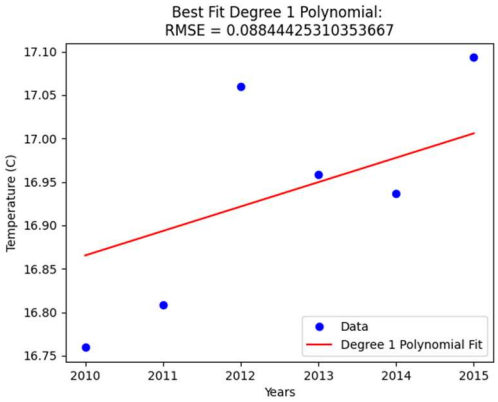


Figure 4D.IIb: *Best fit parabola polynomial to testing data*

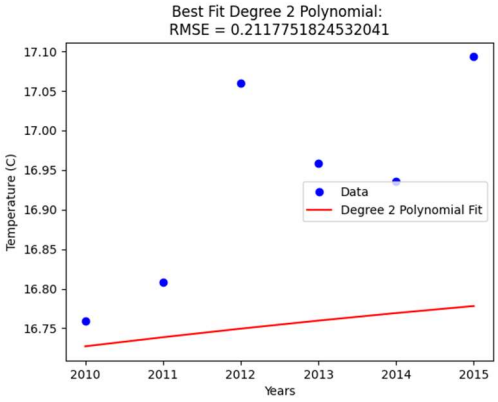
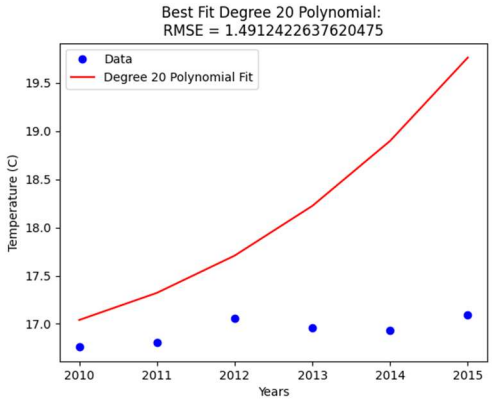


Figure 4D.IIc: *Best fit degree-20 polynomial to testing data*

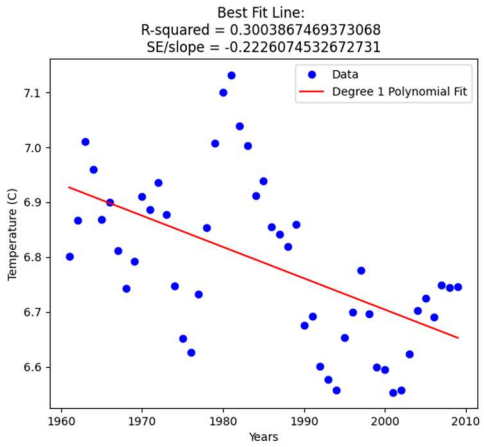


- The latter two models both performed worse than the linear model. The RMSE is quite low for the lower-degree polynomials, but increases for the degree-20 polynomial.

- The line (degree-1 polynomial) performed best, with the lowest RMSE. Similarly, the degree-20 model performed worst with the highest RMSE. This is the opposite of the pattern in part 4D.I, where model quality increased with increasing degree. This is because the degree-20 model was overfitting to the training data, and is now performing poorly on new data. The linear model was simplest, and fits best to new data.
- If we had just used the models generated from averaging over NYC data as opposed to the five-year moving average over 22 cities, then the prediction results for 2010-2015 would be much less accurate. This is because, as we previously saw, the models created from the greater volume of data stemming from the entire set of cities is more reliable than those generated from a single city are.

Problem 4E

Figure 4E: *Best fit line to moving average with window size 5 of standard deviations of average annual temperatures across all cities from 1961 to 2009*



- This result does not match the claim. Since the model displays a negative trend in the standard deviations of mean annual temperature over time, it appears as though temperature variation is decreasing over the years.
- To improve our analysis, we require more data. Specifically, we should collect temperature data not just from US cities, but also from other parts of the world. Our study would also benefit from examining temperature data not just in cities, but also in rural and suburban areas, as well as remote areas such as the arctic or deserts. This would allow for more nuance in our study, and allow us to analyze different trends and compare different combinations of subsets of the data.