March 15, 2021

Clarke Homan
cahoman@gmail.com

MIT 6.0002 Problem Set 5 Write Up

Part A: Creating Models

Problem 4: Investigating the trend

Code Implemented or Improved:

- Class: Climate:
  - added additional commenting, formatted docstring to Google standard in __init__ constructor.
  - Added additional commenting, formatted docstring to Google standard in get_yearly_temp() and get_daily_temp() methods. Added parameter assertion checking to both methods
- Functions Implemented:
  - generate_models():
    - Implemented function with added parameter assertion checking
    - added commenting, formatted docstring to Google standard
  - r_squared():
    - Implemented function with added parameter assertion checking
    - added commenting, formatted docstring to Google standard
    - Directly calculated r**2 value, did not use python packages to calculate r**2
  - evaluate_models_on_training():
    - added commenting, formatted docstring to Google standard
    - Implemented function with added parameter assertion checking
    - Calculated r**2 and Standard Error to model slope ratio (for 1st order model fitting)

Problem 4.I: January 10th Temperatures

The three plots below show the temperatures for January 10 for the years between 1961 and 2009 for New York City. Each plot shows the actual recorded temperatures as well as a predicted temperature based upon a particular regression model. The three predictive regression models are linear, quadratic and cubic, which are 1st, 2nd and 3rd order regression models.

Each plot indicates its regression model R^2 coefficient, also known as the regression's *Coefficient of Determination or "goodness of fit"*. A regression's R^2 coefficient provides a measure of how well the regression explains the total variation of the actual samples to the expected model values. R^2 values range between 0 and 1, with 0 indicating no relationship between the model prediction and the samples. A R^2 value of 1 indicates perfect relationship between the model prediction and the samples. The greater the R^2 value, the better the model fit.

Plots involving linear models include a measure of the ratio of the standard error of the model prediction curve versus the curve's slope. The measure predicts the probability of the

prediction trend fitting the data by chance (value closer to 1.0) also validating the Ho hypothesis, versus the prediction actually fitting the data (value closer to 0.0), validating the model. If there exists a significant relationship between the years and temperature trend, either increasing or decreasing, the slope of the linear model will not be zero. A positive slope may indicate a trend in increasing temperatures over the years. For this exercise, a measure of less than 0.5 indicates the trend being significant.

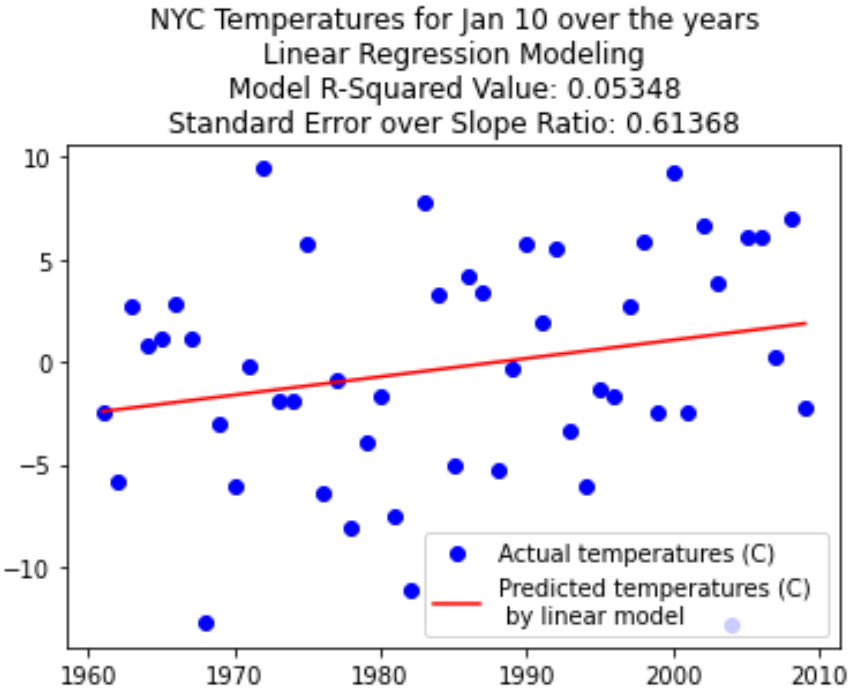
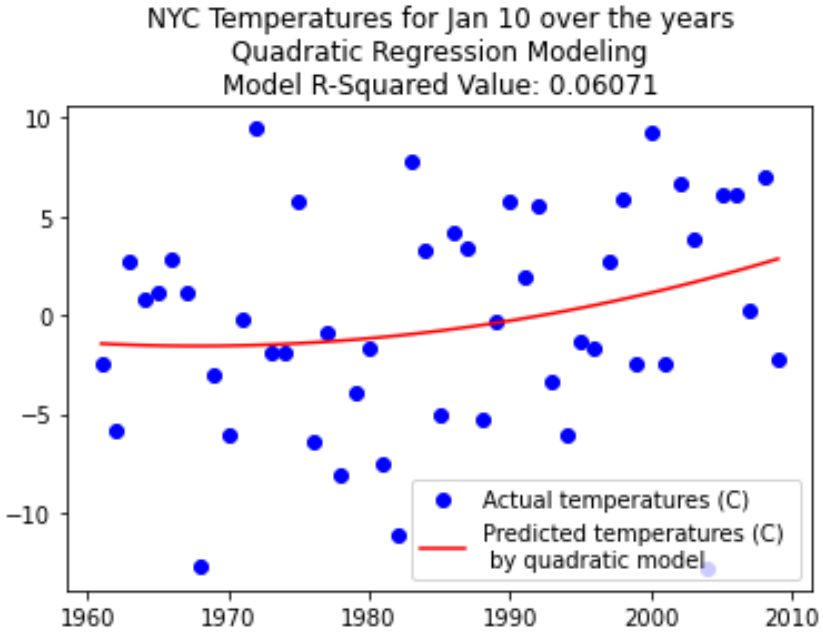


Figure 1: Jan 10 Temps Trend over the years (Linear)



Clarke Homan
cahoman@gmail.com

Figure 2: Jan 10 Temps Trend over the years (Quadratic)

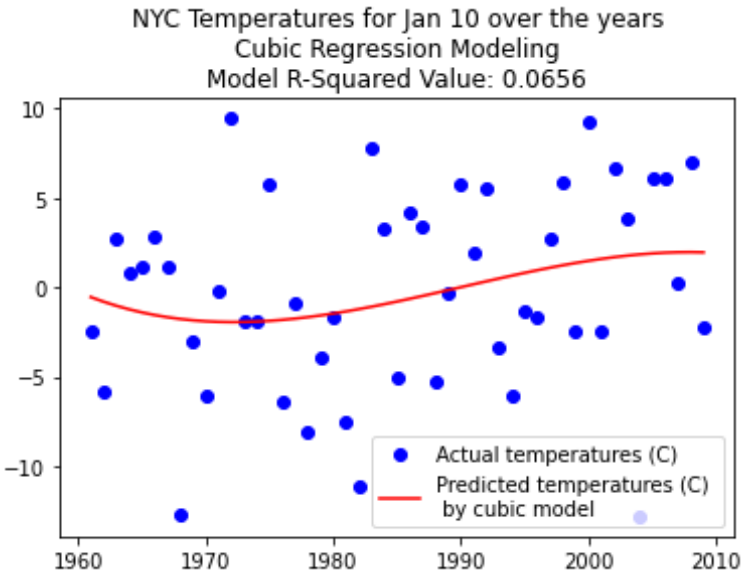


Figure 3: Jan 10 Temps Trend over the years (Cubic)

Problem 4.II: Annual Temperatures

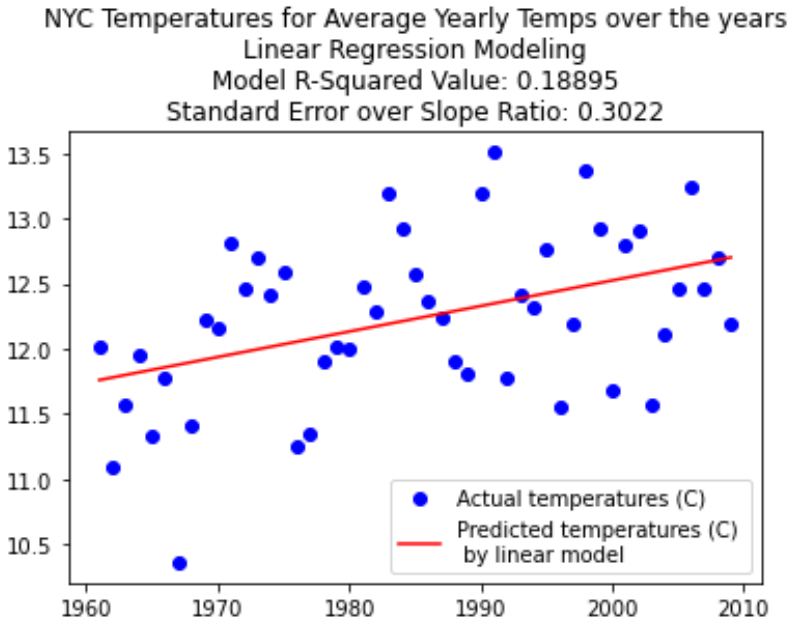


Figure 4: Average Annual Temps Trend over the years (Linear)

Clarke Homan
cahoman@gmail.com

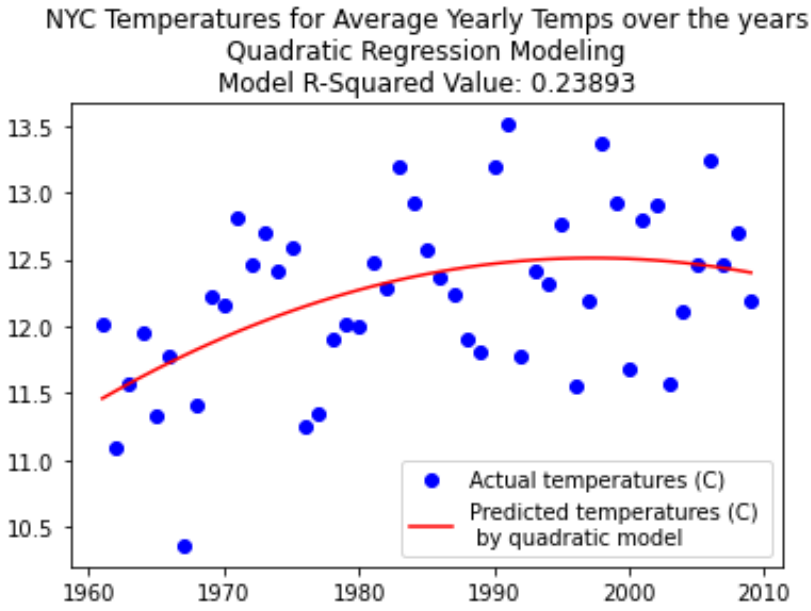


Figure 5: Average Annual Temps Trend over the years (Quadratic)

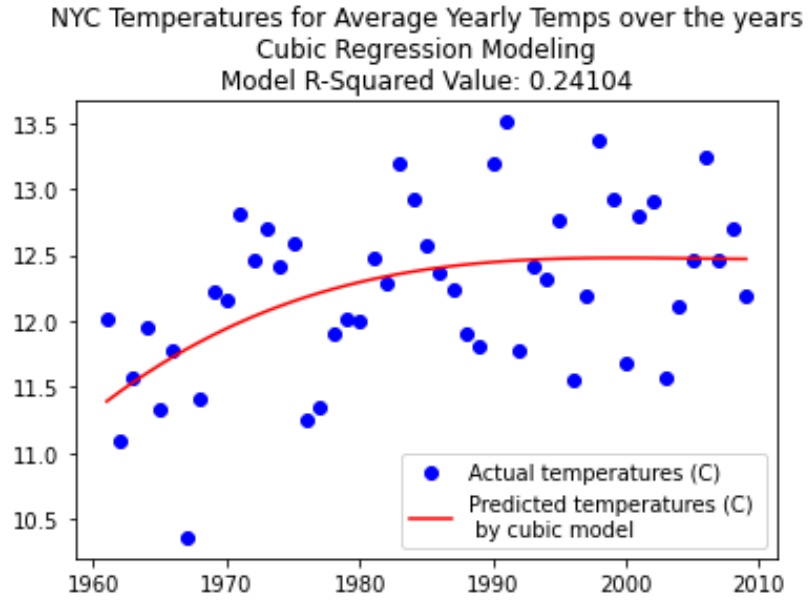


Figure 6: Average Annual Temps Trend over the years (Cubic)

Problem 4 Conclusions

The following questions were asked to be addressed in this write-up.

- What difference does selecting a specific day to plot the data versus calculating the yearly average have on the graphs (i.e. in terms of the R^2 values and the fit of the resulting curves)? Interpret the results.
    - Choosing a choosing a specific day within each year to model versus using each year's average temperature is reflected in both of each model's coefficient of determination (R^2) and the model's slope relationship with model's slope Standard Error coefficient. The R^2 coefficient provides a measure of "goodness of

fit", ranging between 0 and 1, and represents a percentage of how much of the observed temperature data can be accounted for with the model curve. Generally, a higher R^2 value can be interpreted as representing a better fit. In all three regression models, using a specific day versus the yearly average resulted in a lower R^2 value, which indicates a lower goodness of fit.

Also using the specific day versus the average temperature data when using a linear regression yielded a much higher "Standard Error over Slope" statistic, which indicates that using a specific day linear regression model was less significant in representing the temperature trend than using the annual average temperatures. Using the average annual temperatures provides a much more reliable trend analysis then using a specific day of each year.

- Why do you think these graphs as so "noisy"? Which one is noisier?
    - The graphs that modeled a specific day of each year was noisier due to the temperature range of the data. In the specific day cases, the temperatures ranged ~20° (-10° to 10°) whereas the annual average temperatures ranged ~3° (10° to 13.5°). In both cases, the actual temperature data were randomly scattered around the regression model curves, where each curve rarely contained an actual sampled datapoint.
- 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.
    - In all graphs, the regression model slope is positive in the earlier years, which may indicate that the temperatures were increasing over time, although the models that were performed on a specific day of the year provides much mode unreliable modelling (due to its lower R^2 value and higher standard error of the model slope statistics). Since the linear regression yields a straight line, the slope is constant throughout the years, whereas the quadratic and cubic models which provide a tighter fit to the data show some "leveling off" and perhaps decline in average temperatures in the latter years. In all models for both selected day of year and average annual temperatures, the goodness of fit (R^2) was poor with them struggling to claim to be more accurate than 31% for the average temperature models. Given these graphs, I would have little confidence in supporting the claim that global warming (within NYC) for these years existed.

      Another consideration is that this study only involves one city, New York City, which can hardly represent the whole of the US.

Part B: Incorporating More Data

In Part B, the average yearly temperature for 21 US cities over the years from 1961 to 2009 were modeled using linear, quadratic and cubic regression models. Note the tighter distribution of actual sample points to each regression curve for each model as well as the significantly increased R^2 coefficients for each model, suggesting better fit. Also note the more pronounced positive slope of each regression model curve for the national average versus the curves modelling only NYC.

Clarke Homan
cahoman@gmail.com

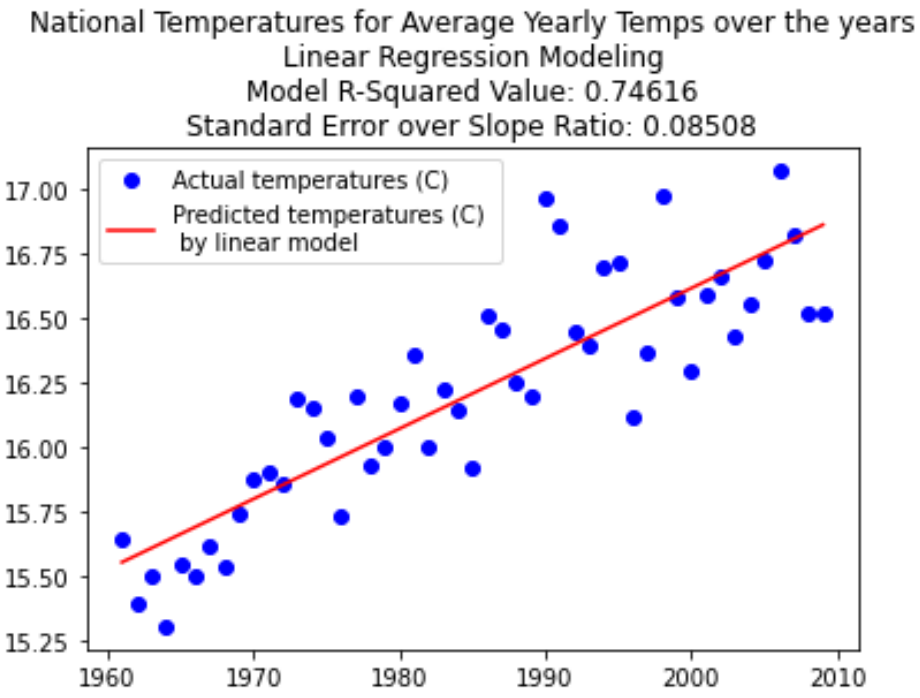


Figure 7: Average National Temperatures over the Years (Linear Regression Model)

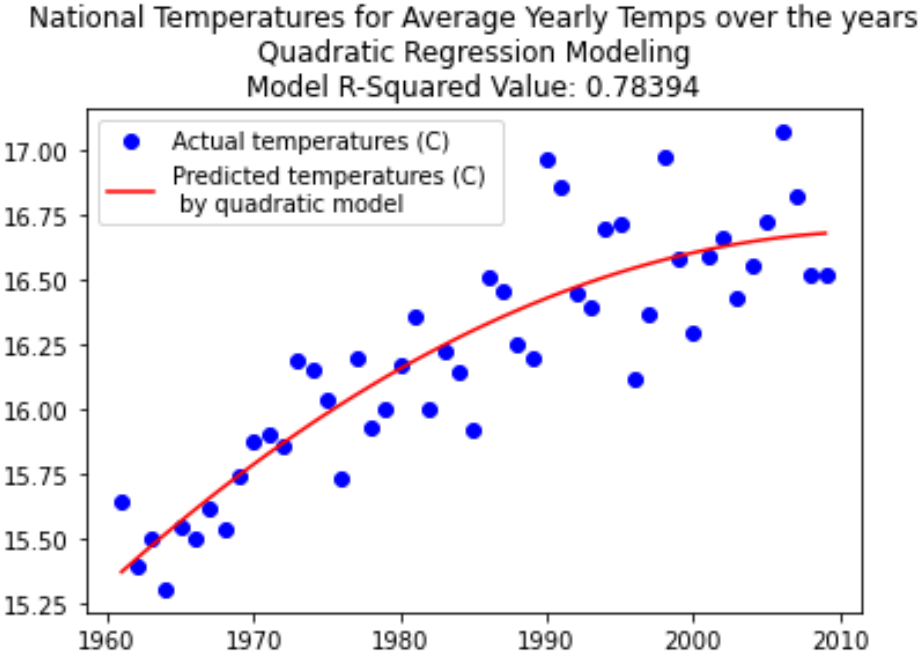


Figure 8: Average National Temperatures over the Years (Quadratic Regression Model)

Clarke Homan
cahoman@gmail.com

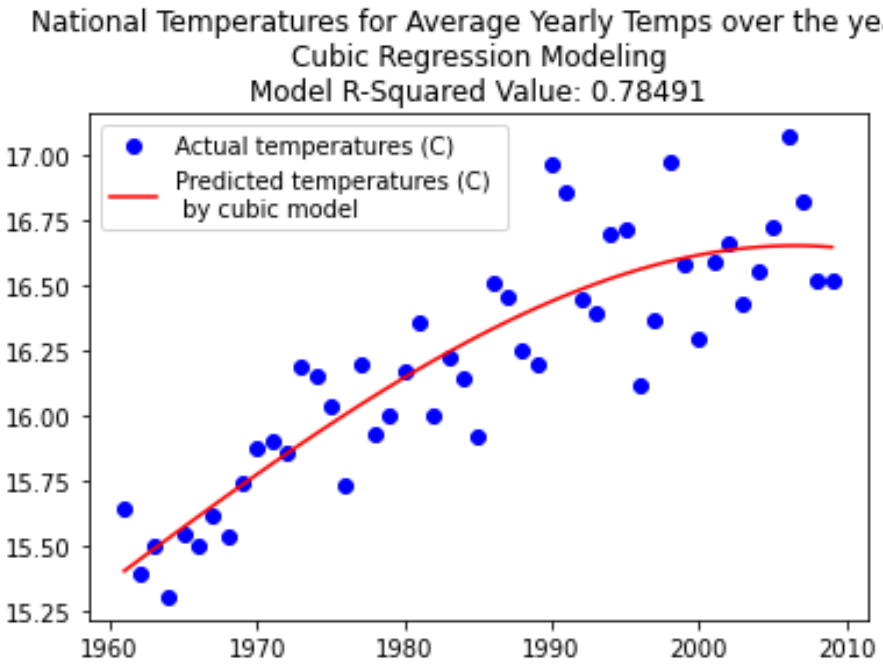


Figure 9: Average National Temperatures over the Years (Cubic Regression Model)

Part B Conclusions Write Up

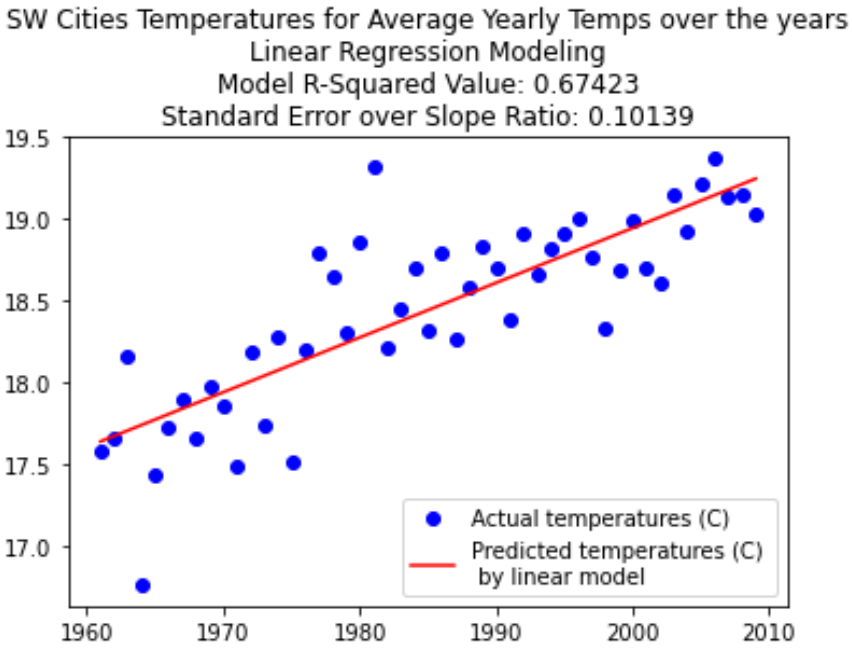
Answer the following questions:

- How do these graphs compare to the graphs in Part A (in terms of R^2 coefficients and fit of the resulting curves, and whether the graphs support or contradict our claim of global warming)? Interpret the results.
    - The first thing noticed was that the actual average annual temperatures for the larger population of cities are more "tighter" in distribution than the data for a single city's annual average temperature. My intuition suggests that having multiple cities from multiple US regions allows the data to converge to a more common average over time, perhaps following the Law of Large Numbers rule, regressing to the mean. Because of the tighter distribution of sample data, the resultant model curves have a higher probability of predicting the overall trend, as indicated with the higher R^2 values (~ 75%) which indicates that the trend curves account for 75% of the data. With the linear model, the Standard Error over the Slope statistic is near zero (8%) which indicates that the curve trend is very likely represents an accurate representation of the overall trend (that the trend predicted is not by chance), that the trend slope (average temperature over the years) is non-zero.
- Why do you think this is the case?
    - Adding more cities increases the distribution to tend (or regress) more to the mean; Law of Large Numbers. As the distribution tightens, the regression models are more effective in predicting a trend.
- How would we expect the results to differ is we used 3 different cities? What about 100 different cities?

Clarke Homan
cahoman@gmail.com

- If fewer cities were used, especially if the cities were geographically and climatically different, then the data dispersion between the cities would make for more nosier data. Adding more cities, such as 100 different cities, if they were geographically and climatically distributed would reduce the actual temperature data noise thus allowing the regression models to better predict trends.

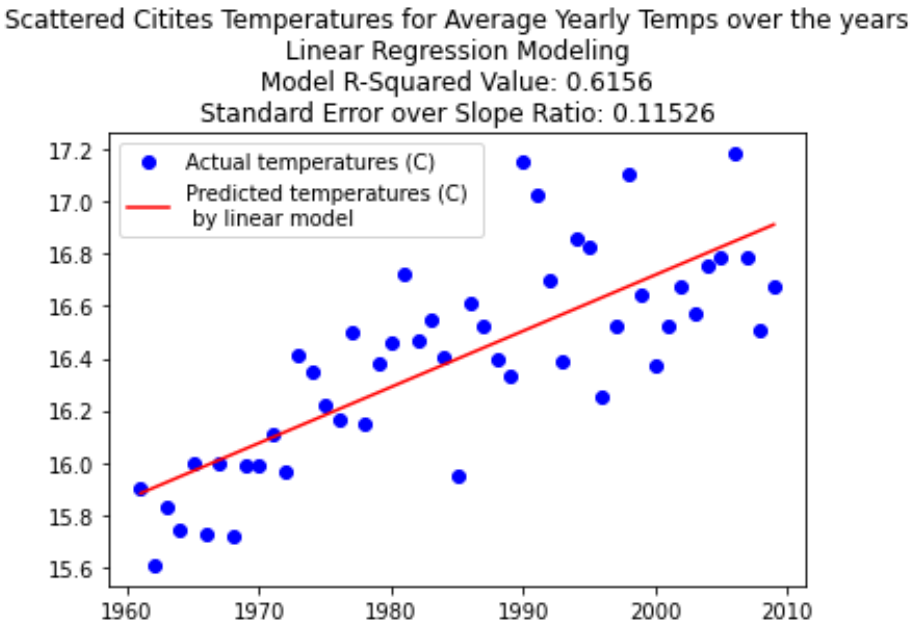
- How would the results have changed if all 21 cities were in the same region of the US (e.g., New England cities)?
  - If all of the cities modeled came from the same region, the average temperatures for each city would be similar to the other cities and thus the distribution of temperatures would tighten. This tighter distribution would possibly allow the regression models to better predict. The disadvantage to this approach would be that the samples would only be from a particular regions and not necessarily representative of the US, or North America or the world's overall trend.
  - Additional plots were made for various geographical considerations.
    - Southwest Cities Trend
      - Cities included: 'SAN DIEGO', 'PHOENIX', 'DALLAS', 'ALBUQUERQUE', 'LOS ANGELES', 'LAS VEGAS'
        - Cities located in similar region of the US
        - Notice that the R^2 value decreases for the model over the "all" cities case with a slightly lower confidence in the trend line representing any increasing (or decreasing) trend (SE over Slope)



    - Scattered Cities Trend

Clarke Homan
cahoman@gmail.com

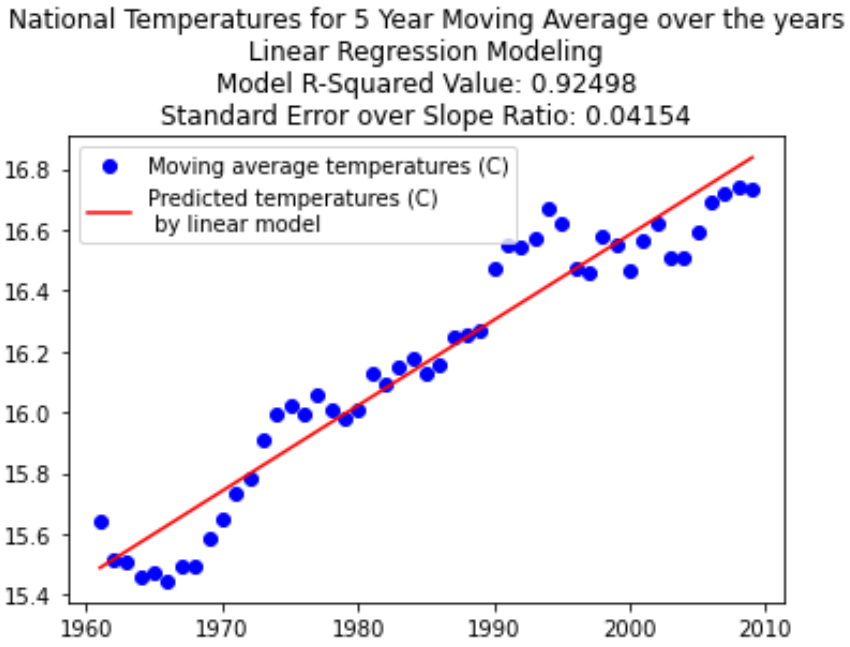
- 'BOSTON', 'SEATTLE', 'SAN DIEGO', 'MIAMI', 'NEW ORLEANS', 'ST LOUIS'
    - Cities scattered around the US with dissimilar geographies
    - Again, lower R^2 value than the "all" cities scenarios with a lower confidence in the trend line representing any increasing (or decreasing) trend (SE over Slope)
    - More data usually brings better forecasts

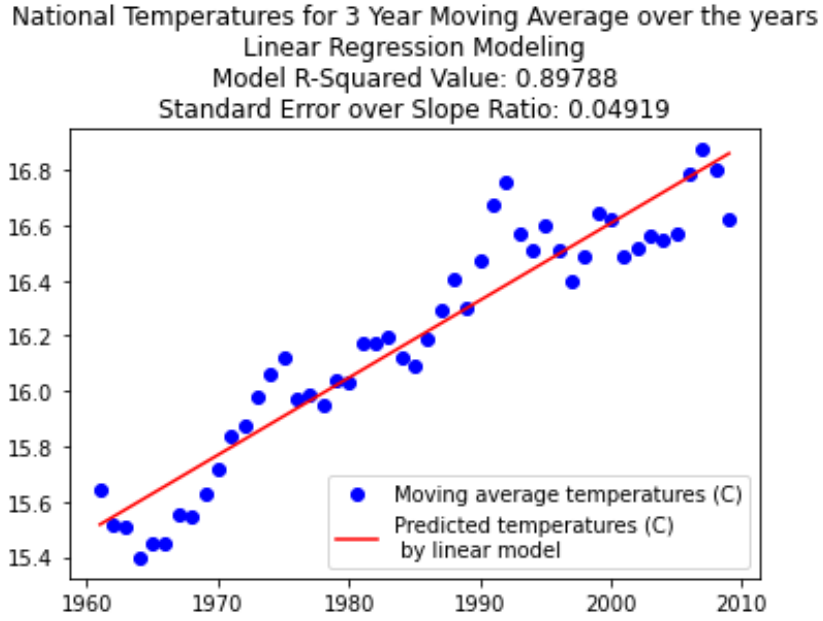


Part C: 5-year Moving Average

In Part C, the moving average temperature data was generated from the average yearly temperature for 21 US cities over the years from 1961 to 2009. The moving average temperatures were computed with a trailing sliding five-year window, starting with 1961. Note the even tighter distribution of moving average temperature data points to each regression curve for each model as well as the significantly increased R^2 coefficients for each model, suggesting better fit.

Clarke Homan
cahoman@gmail.com





Part C Conclusions Write Up

Answer the following questions:

- How does this graph compare to the graphs from part A and B (i.e., in terms of the R^2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.

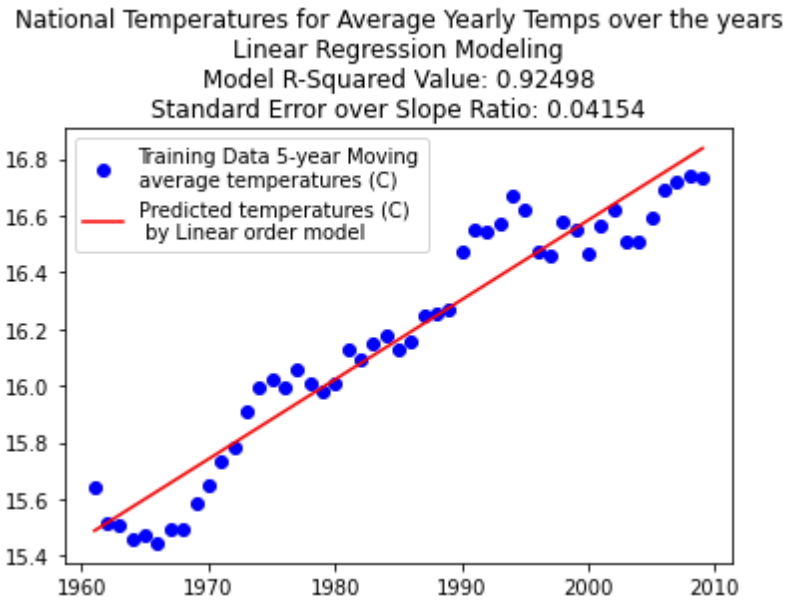
Clarke Homan
cahoman@gmail.com

- Calculating the moving average national temperatures and then plotting them clearly shows an overall upward trend in average temperatures over the trial years. There are periods of time where the temperatures dip (local minima) but the overall trend is obvious. The slope of the linear regression model is positive (indicating an increasing temperature over time) and its calculated R^2 value is very large (over 92% for the 5-year moving average model) which indicates that the linear model can account for over 92% of the moving average data. The linear model is a good fit for the moving average data.
- Also, the Standard Error over the slope statistic for the 5-year linear model is very small (around 4%) which indicates that the linear regression predicts the trend not by chance.

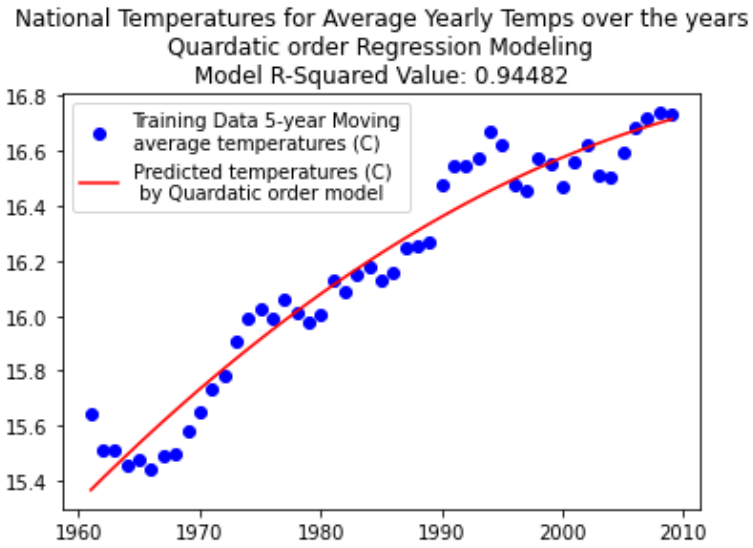
- Why do you think this is the case?
  - Using the moving average for the national temperatures smooths the yearly fluctuations in average temperatures due to its incorporation of past years average in the sliding average window averaging. Only average temperatures that have sustained (in time) deviation from the prediction model will be significantly distanced from the trend line.

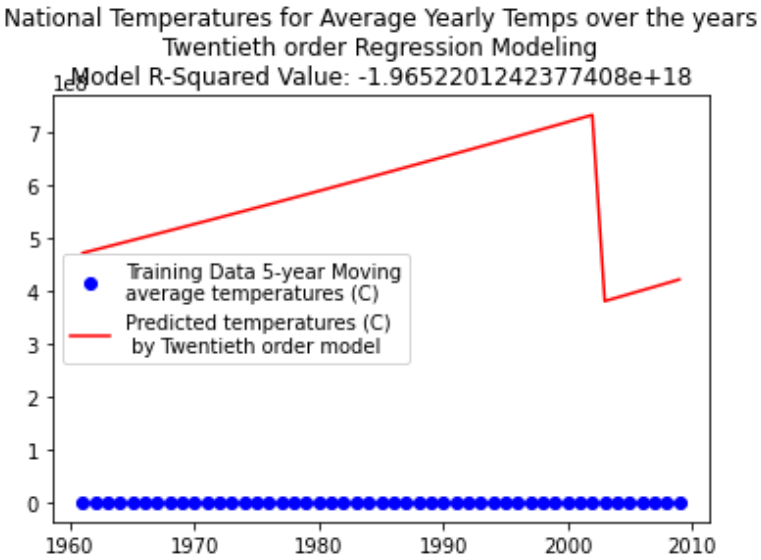
Part D: Predicting the Future

Problem 2.II Generate more models



Clarke Homan
cahoman@gmail.com





Problem 2.I Conclusions Write Up

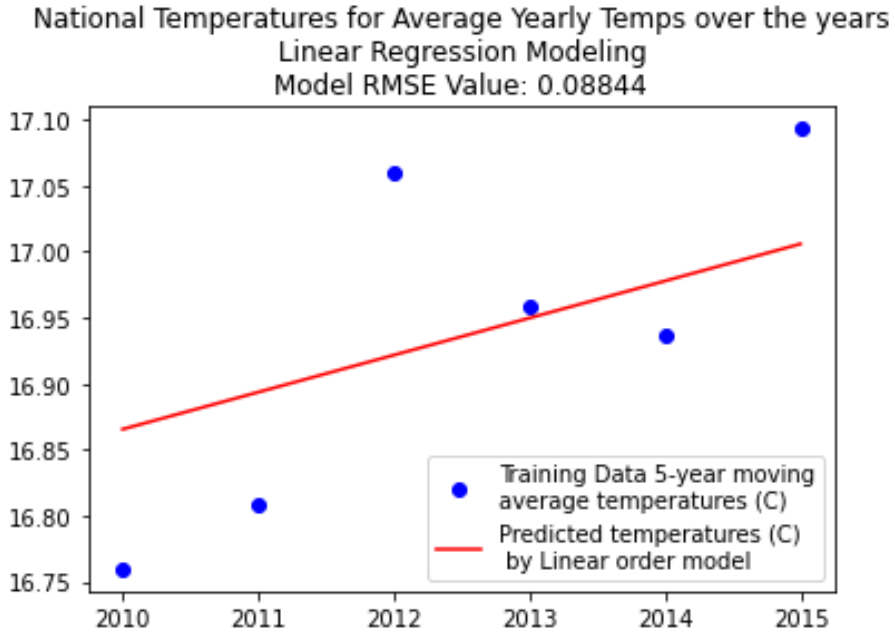
Answer the following:

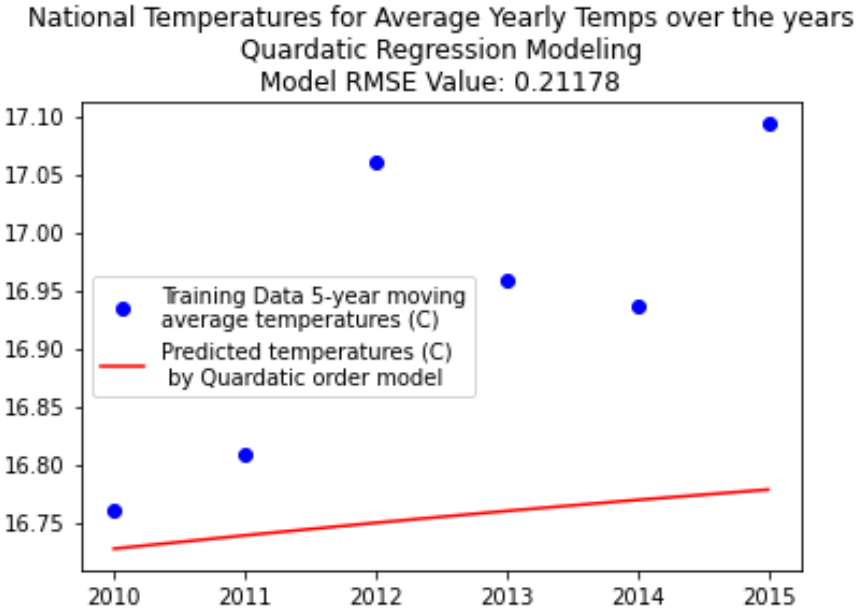
- How do these models compare to each other?
  - The first two models, the linear and quadratic, fit the moving average data relatively well. The lines on both plots were close to the overall data trend appearance. The third plot, which used a 20^{th} order polynomial fit obviously blew up, not fitting the training data well.
- Which one has the best R^2? Why?
  - The second plot, the quadratic model, produced the highest R^2 value at 94.482%, compared to the 92.498% R^2 for the linear model (although this difference is fairly minor). The R^2 value of the 20^{th} order model was again very small (near

zero with a very small negative value). R^2 values should range from 0.0 to 1.0 so this calculation, being negative was obviously was generated in error.

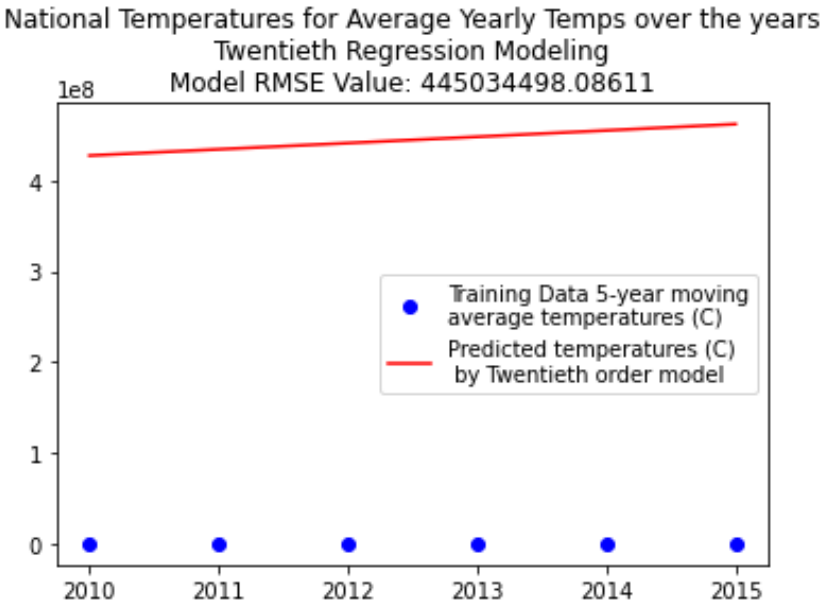
- Which model best fits the data? Why?
  - Given the R^2 value of the quadratic model, it fits the data the best because of its higher R^2 value than the linear model.

Problem 2.II Predict the results





Clarke Homan
cahoman@gmail.com



Problem 2.II Conclusions Write-up

Answer the following questions:

- How did the different models perform? How did their RMSEs compare?
    - The moving average data in the testing period (years 2010 – 2015) were closer in value (min: 16.7595 max: 17.0932) .334 °C to each other than the training period (min: 15.4458 max: 16.7386) 1.293 °C. The best model in terms of prediction as measured by the lowest RMSE value would be the linear model, with an RMSE of 0.08844. The quadratic has a significantly higher RMSE value which indicates that this model has greater prediction errors. Again the 20^{th} order regression model fails to predict the testing data with a RMSE value of 445,034,498!
- Which model performed the best? Which model performed the worst? Are they the same as those in part D.2.I? Why?
    - The linear model performs the best as evidenced with its lowest RMSE value, followed by the quadratic and lastly the 20^{th} order model. Unlike the models developed on the training data where the quadratic model performed the best as calculated by the model's R^2 value, the linear model provides the best predictor of trend in testing data with its lowest RMSE value.

      As for why the quadratic model suffers significantly for predicting the testing data over the linear model, a look at the two model's regression coefficients (in the plots below) show that the linear model near zero to minus 40 for coefficient value. For the quadratic model, the coefficients range from near zero (for the first two coefficients but then have a significantly lower value of -1300 for the third term (y-intercept of the quadratic expression). Thus, there is a lowered bias in the trend line for the quadratic model as shown with the trend line below all of the testing data. This is not the case with the linear model

Clarke Homan
cahoman@gmail.com

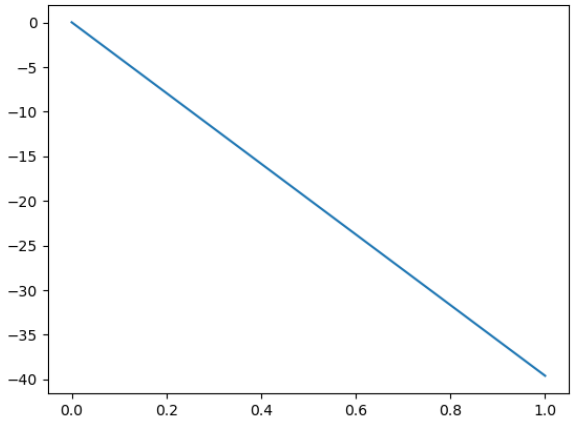


Figure 10: Prediction Coefficients on Training Data - Linear Model

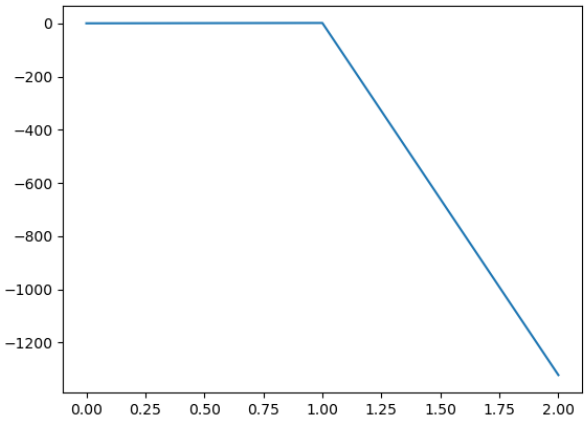
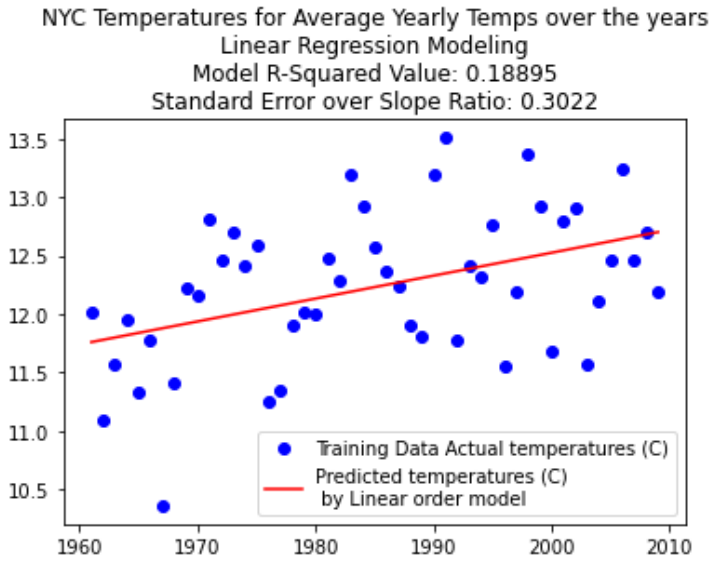


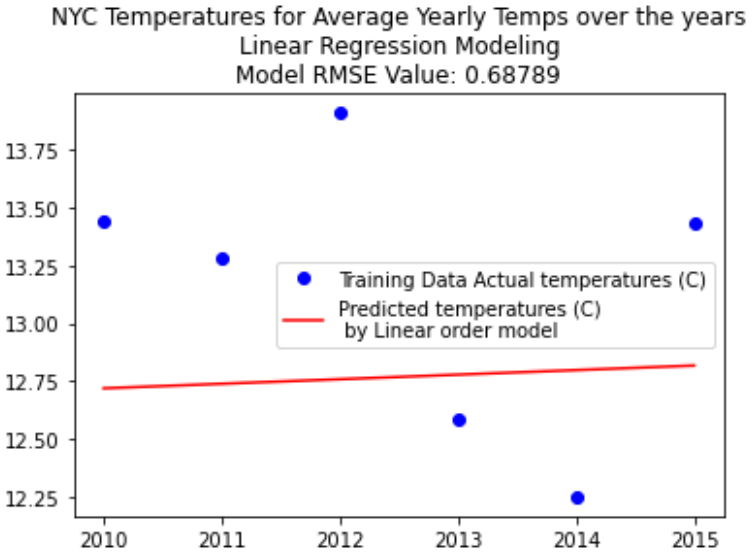
Figure 11: Prediction Coefficients on Training Data - Quadratic Model

- 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?
    - By using a single city instead of a larger sample of cities and by using yearly averages of temperatures instead of using a sliding multi-year average (like 5 years), the data is much more variable as indicated by the training model's R2 value if 18.89%, which indicates that the training model for the linear regression is a poorer fit than the national statistics. Because the training model was poorer for NYC, the prediction model suffers with a significantly higher RMSE statistic (poorer prediction fit) than the national models. For the quadratic model, the

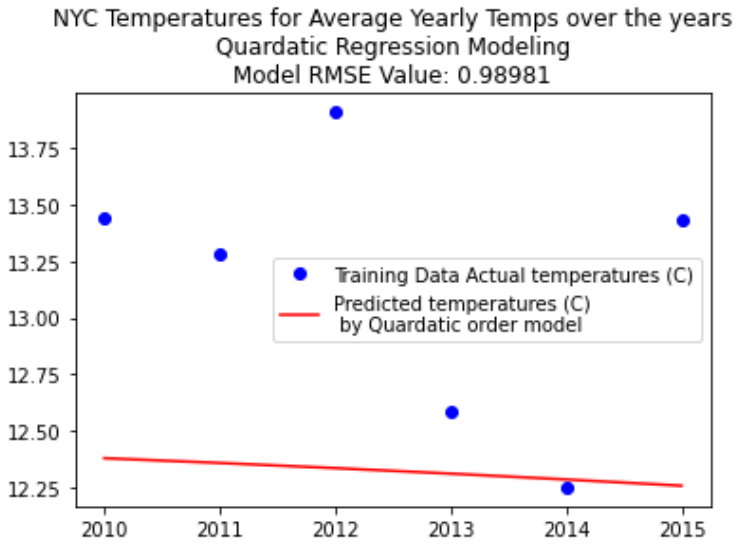
Clarke Homan
cahoman@gmail.com

prediction shows a decreasing temperature but with a highly suspect RMSE value of nearly 1.0!



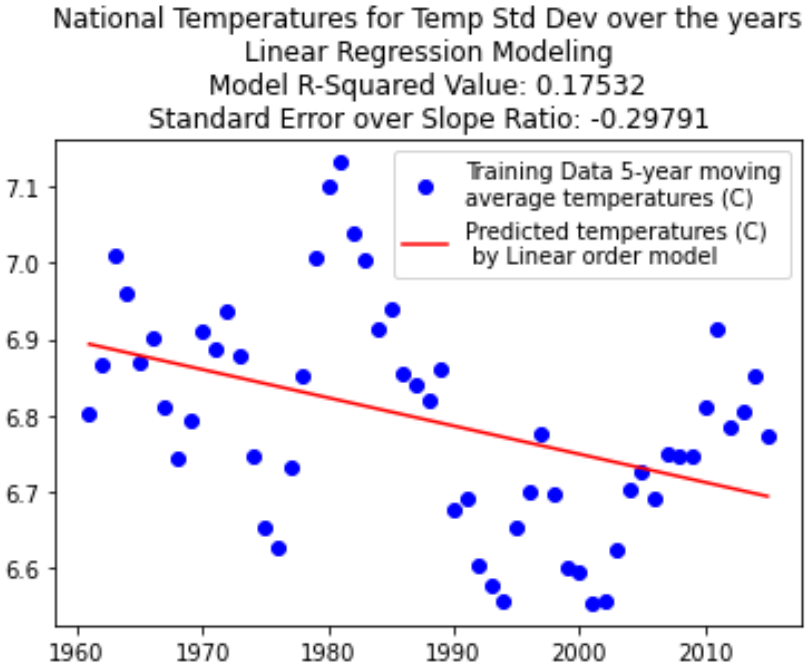


Clarke Homan
cahoman@gmail.com

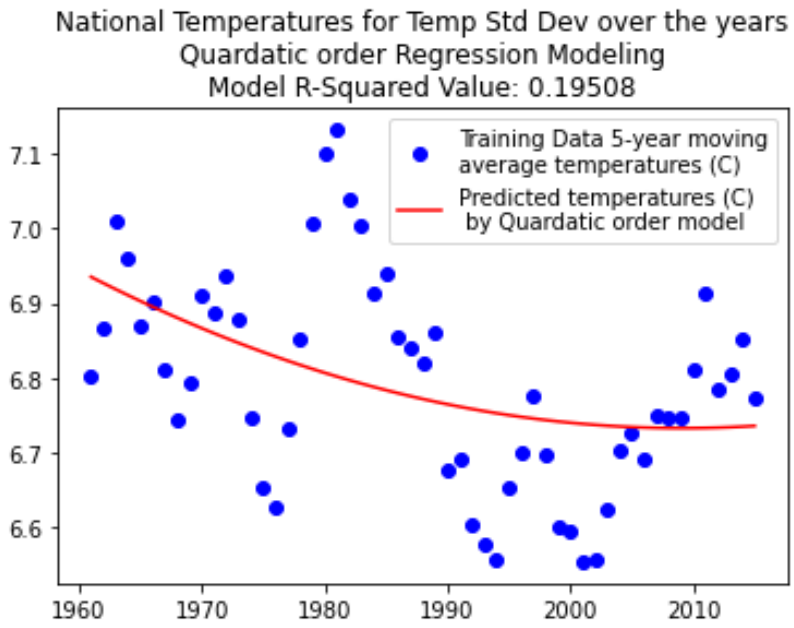


Part E: Modeling Extreme Temperatures

The plots below show the trend of average temperature variability for 21 select cities from 1961 to 2010 as measured by aggregate differences between measured annual temperatures from the period's mean temperature. The measured annual temperature was actually a sliding 5-year average temperature. Both the linear and quadratic regression prediction indicate that temperature variability has decreased over this time period, but with a low probability of confidently predicting this fact as indicated with low R^2 values.



Clarke Homan
cahoman@gmail.com



Part E: Conclusions Write-up

Answer the following questions:

- Does the result match our claim (i.e., temperature variation is getting larger over these years)?
    - Although the prediction models have a fairly low R^2 value, the model trend lines show that temperatures have become less variable (negative sloped trend lines) over the training period, instead of the thesis that temperature variability has increased over time.
- Can you think of ways to improve our analysis?
    - If the goal of improvement is gain improved confidence that the trend depicted closely models the temperature variability over time, then more annual temperature data should be added. This addition could come from additional cities within the time frame as well as additional years (before 1961 and after 2010). Ideally both additional cities and years might provide the additional data necessary to improve predictions.
    - If we want to improve the prediction model's R^2 value, then by restricting the number of cities or years will help improve the prediction models R^2 value but also lead us to a false sense of security that we have a strong national trend due to this data limitation.