**Chart, scatter chart

Description automatically generatedA4.1**

**A4.2**

Chart, scatter chart

Description automatically generated

● 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.

1) The R2 value for using a specific day is smaller than that of using the average yearly temperature.

The fit of 1-degree regression model is better using average yearly temperature compared to just using temperatures on a specific day.

2) A4.1 is noisy because is has a small sample size compared to its population. However, I cannot account for the noisiness of A4.2 A4.1 is more noisy since R squared value for it is smaller than A4.1.

3) Both models support the claim that global warming is causing increasingly high temperature since both models are have positive gradients which mean that temperatures are increasing over the years. However the reliabilities of these models can be up for contention since the R2 values are ridiculously low, suggesting that there might not be any relationship between the years and temperatures.

B4

Chart, scatter chart

Description automatically generated

● 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

1) R2 value for B4 is a lot closer to 1 than those in part A. Model in B4 represents a better relationship between years and temperatures than those in part A. The graph B4 supports the claim about global warming since there is a positive linear relationship between years and temperature. It is also somewhat reliable since the R2 value is relatively high.

2) More samples are used which result in less variations in the data set.

3) I would expect the R2 value very small if 3 cities are used. I would expect R2 value to be larger when 100 cities are used as compared to 21 cities. However, the difference would not be huge. SE slope ratio will have an opposite effect as compared to R2 value when number of cities used changes.

4) R2 value would be larger.

Chart, scatter chart

Description automatically generated**C4**

● 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?

1) The R2 value for this curve is much higher compared to the previous curves in A and B. Also, the SE to slop ratio is much smaller is this case compared to previous curves. This means that the curve has a higher reliability in generating a positive linear relationship (given that is has a positive gradient and it a 1st degree model) between the years and the temperature relative to the curves in A and B

2) Moving average helps to smoothen short term fluctuations, allowing the model to capture the underlying relationship between years and temperature instead of capturing noises.

Chart, scatter chart

Description automatically generatedD4-2.1A

Chart, scatter chart

Description automatically generatedD4-2.1B

Chart, scatter chart

Description automatically generated D4-2.1C

● How do these models compare to each other?

● Which one has the best R2? Why?

● Which model best fits the data? Why?

1) R2 value increases with the degree of regression model, suggesting that there is a stronger relationship between year and temperature, using models of higher degrees.

2) Model with degree of fit = 20. Increasing the complexity of model will always lead to higher R2 value. This is because at a higher degree, if the extra terms are useless, the coefficient of the term will just be 0, and if the data is noisy, the model will fit to the noise rather than just the underlying pattern in the data. Hence R2 is the largest for the model with a fit of 20 due to a high probability of over-fitting.

3) D4-2.1C since it has the largest R2 value.

D4-2.2A

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedD4-2.2B

Chart, line chart

Description automatically generatedD4-2.2C

● 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?

1) RMSE value increases as the degree of regression increases, suggesting that the fit of the model becomes worse for higher order regression models.

2) Mode D4-2.2A (degree of fit = 1) performed the best since it has the smallest RMSE value. Model D4-2.2C (degree of fit = 20) performed the worst since is has the largest RSME value. This is opposite of D4-2.1 where the model with a degree of fit = 1 performed the worst and the model with a degree of fit = 20 performed the best. D4.2.1C is an occurrence of overfitting, hence not suited to accommodate new data points in when they are introduced in D$-2.2, resulting in large RSME value. D4-2.1A captures the underlying pattern between years and temperature without overfitting, hence able to accommodate new data points when introduced in D4-2.2

3) RSME would be even larger since moving average helps to smoothen the data sets by removing short term fluctuations. Without it, there will be more variations in in the data set, resulting in a more noise and less reliability when plotting.

E4

Chart, scatter chart

Description automatically generated

● 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?

1) The result does not match the claim, instead it opposes the claim. There is a negative linear relationship between years and temperature as shown by the model having a negative gradient, suggesting that temperature variation is getting smaller over these years.

2) The R2 value in this case is small, suggesting that there may not be a strong linear relationship. We could improve the R2 value, hence the reliability of the model by using different degree of fit. Once we obtain a R2 value that is decently large (larger than 0.7), execute cross validation on testing data (sd from 2010 – 2015) to make sure that we get the best fit curve over different degrees of models without over-fitting.