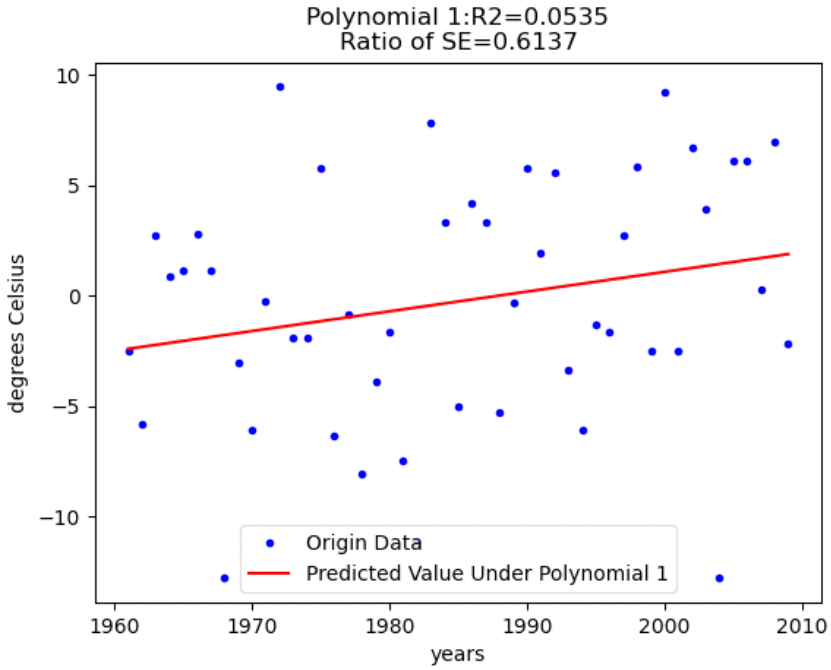
ps5_writeup

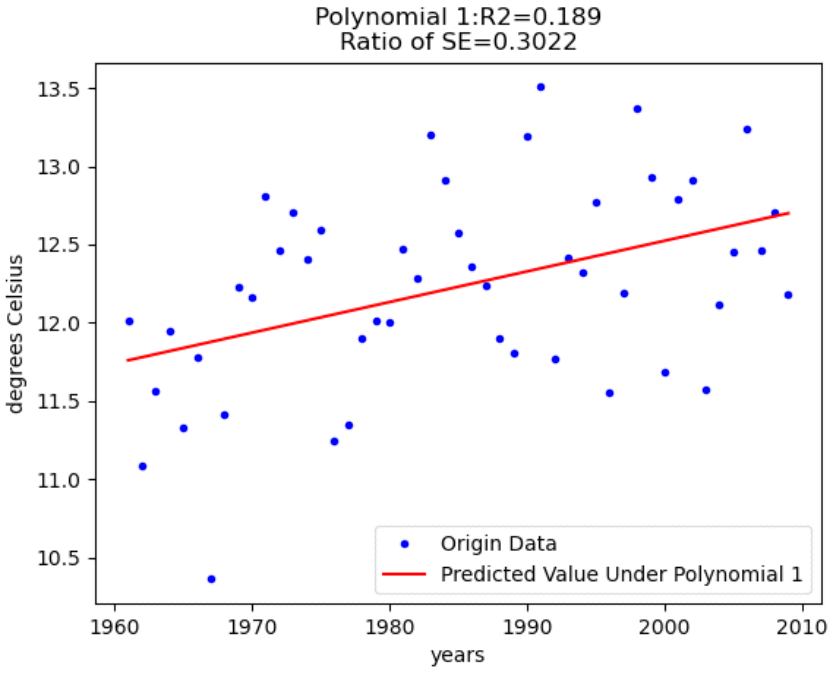
2021年4月5日 20:57

PA4

P4_I Jan 10th



P4_II Annual Temperature



●What difference does choosing a specific day to plot the data for versus calculating the yearly

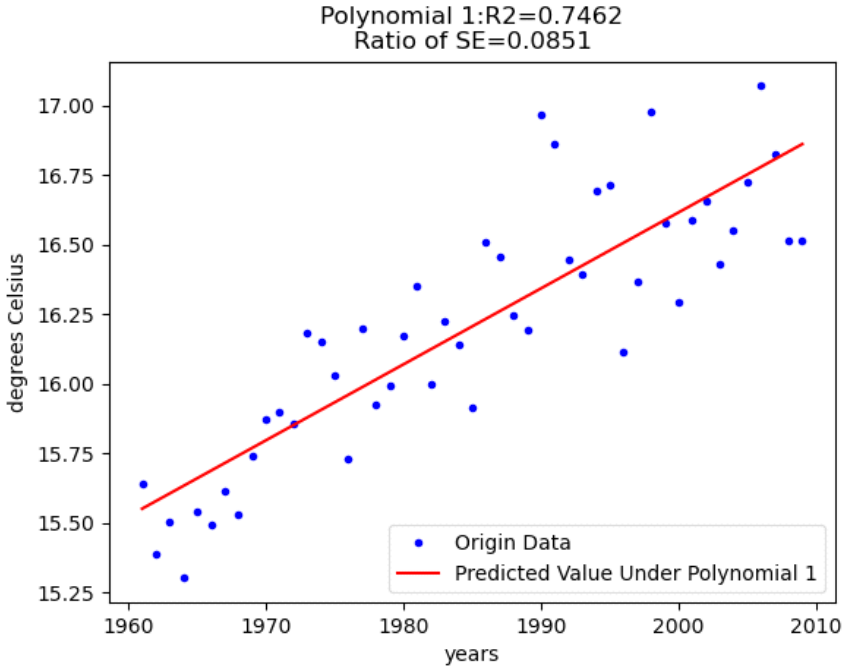
average have on our graphs (i.e., in terms of the R2 values and the fit of the resulting curves)? Interpret the results.

每年平均气温当做测试数据的R2更加高，也就是说基于平均温度的模型在训练集上表现得更好一些。

● Why do you think these graphs are so noisy? Which one is more noisy?
图的尺度比较细，这样来看，感觉点都很散，但事实上，如果不见刻度弄这么细的话，是看不出来noisy。当然很明显，Jan10th的数据更noisy，它的点的变动的温度范围更广，可能是因为它是某一天，偶然性比较大。

● 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
从Jan10th 来看，它的值是0.6，也就是看不出来趋势。从每年平均气温来看，它的值是0.3，也就是趋势是值得确信的，也就是说从图的红线来看，温度的确是上升的。

PB



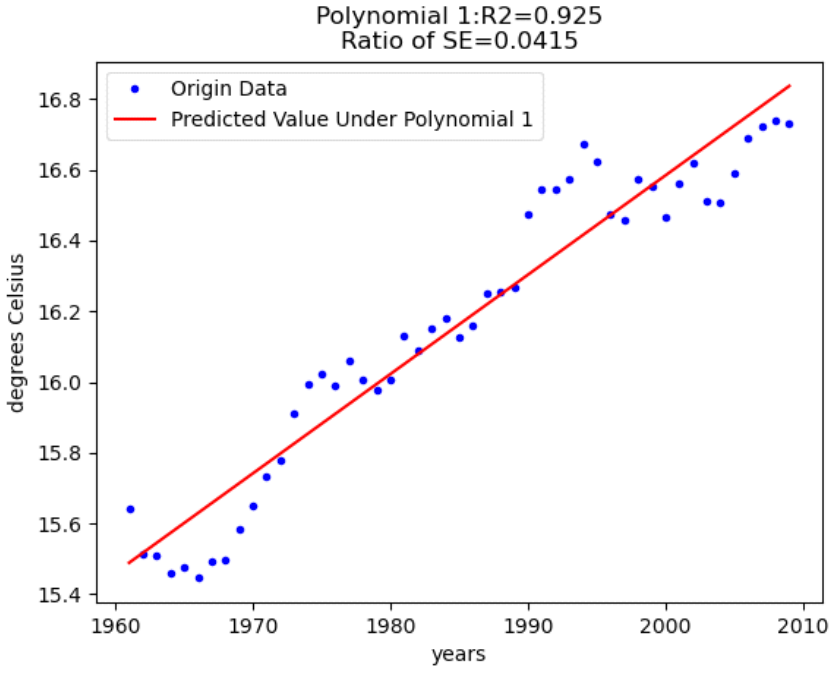
Answer the following questions with a short paragraph in ps5_writeup.pdf.
● How does this graph compare to the graphs from part A (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
R2更大，该模型更加好，而且Ratio of SE 更加地小，对我们的假设：全球变暖提供了有力支持。

● Why do you think this is the case?
采样数，从 Jane10th，到纽约每年的平均温度，到美国21城市的平均温度，可以发现数据越来越多。

● How would we expect the results to differ if we used 3 different cities? What about 100 different cities?
城市越多应该更加能支持我们的推断，毕竟数据来自不同城市，样本更加random。

● How would the results have changed if all 21 cities were in the same region of the United States (for ex., New England)?
结果应该没有这个好，因为不是随机了，但是应该比PA4的单独纽约要好。

PC

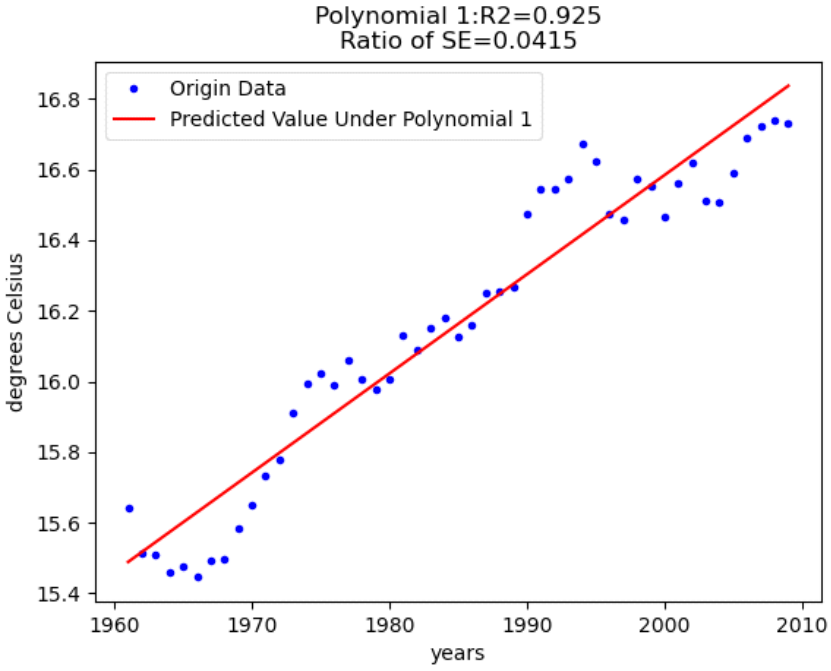


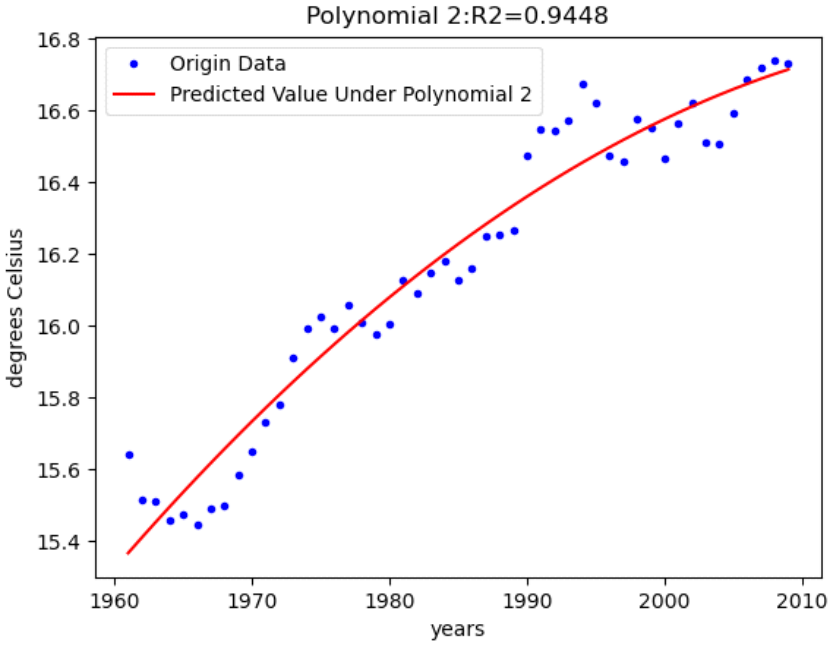
Answer the following questions with a short paragraph in ps5_writeup.pdf.
● How does this graph compare to the graphs from part A and B (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
R2更趋近1，也就是说光是在测试集的模型相比A和B更加好，当然趋势的可信度也更好，毕竟SE更加小。我觉得还是兼顾样本更多的原因。

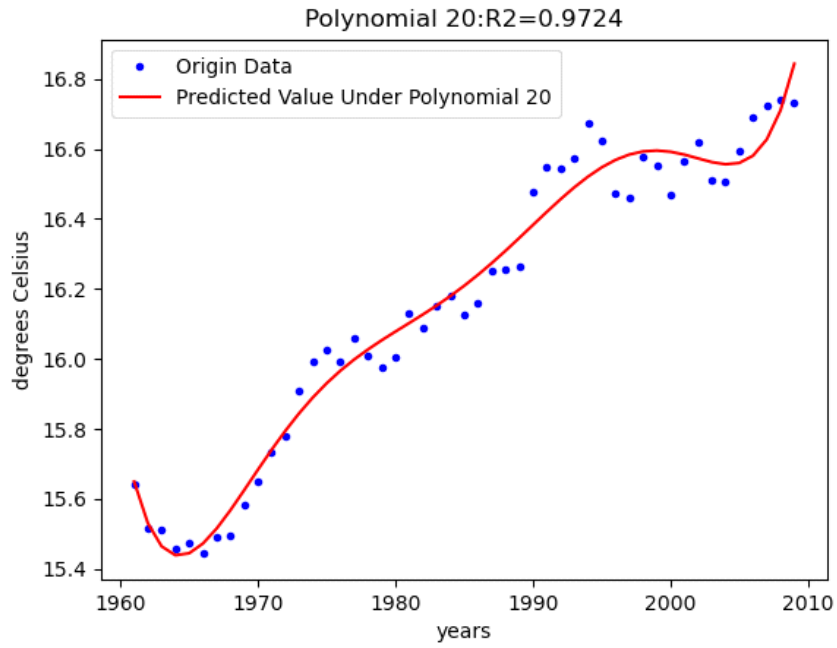
● Why do you think this is the case?
兼顾样本更多

PD

Problem 2.I Generate more models







Answer the following questions with a short paragraph in ps5_writeup.pdf.

- How do these models compare to each other?

维度更高，在训练集的性能更好，因为R2更接近于1。

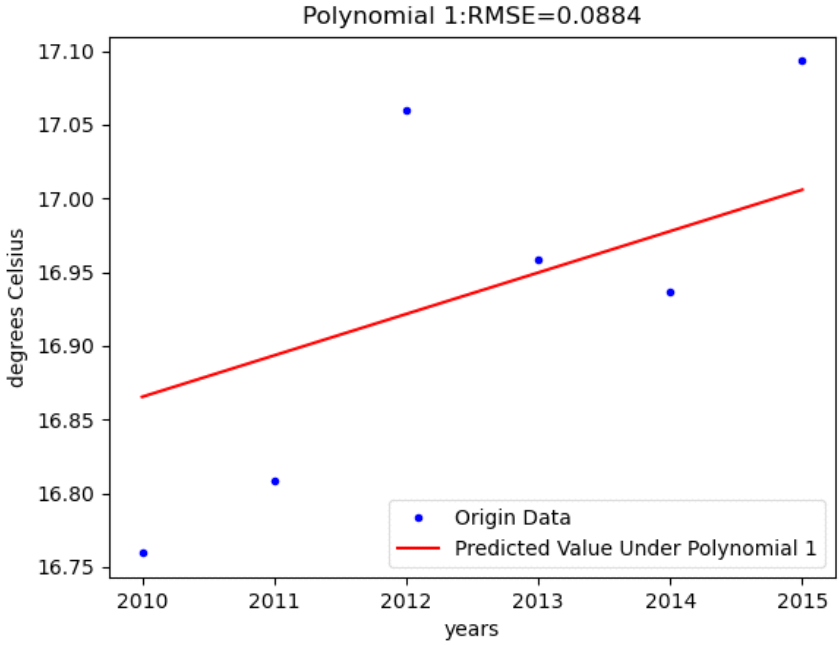
- Which one has the best R2 ? Why?

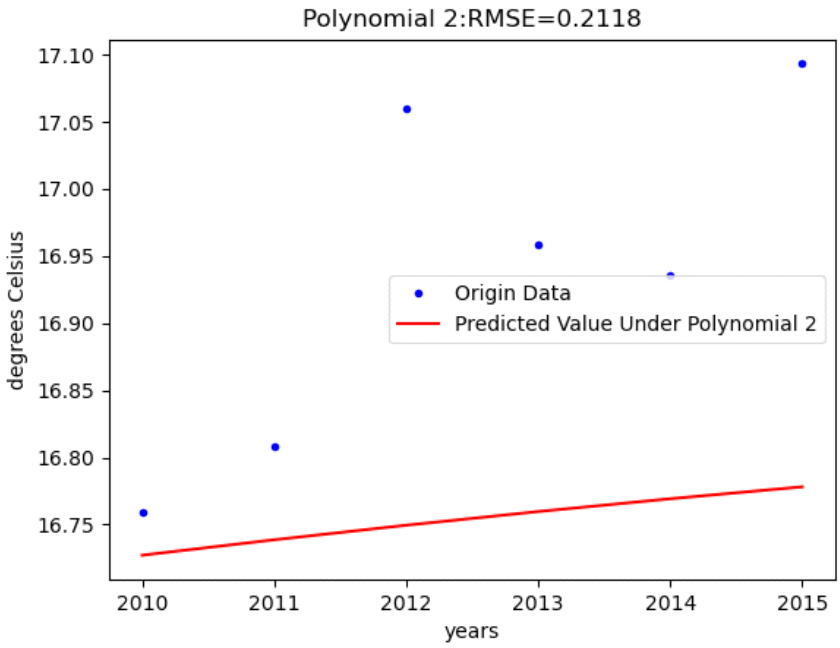
维度为20的，因为模型更复杂，参数更多。

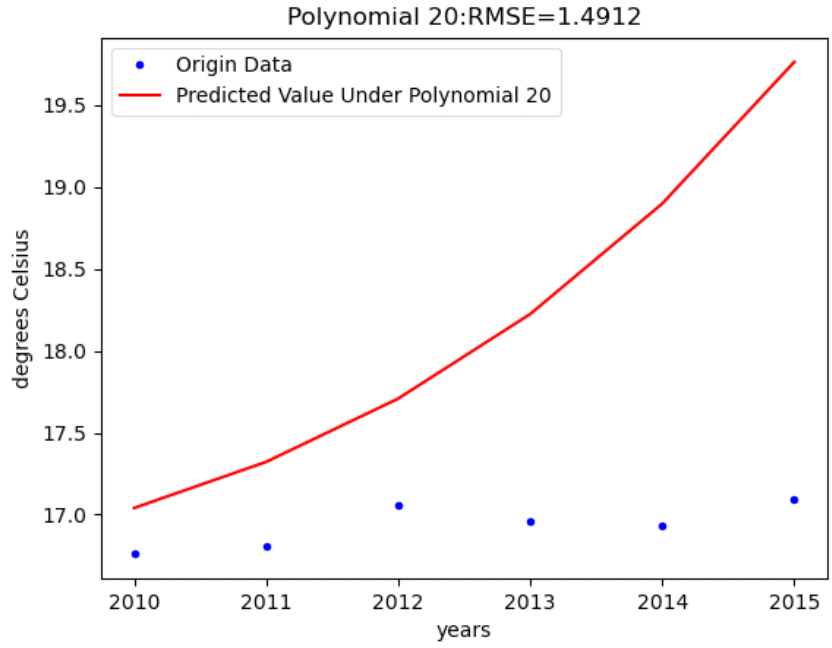
- Which model best fits the data? Why?

维度20，因为他们都是用训练集去测试模型，而模型更复杂的，显然对这些训练集的泛化能力更好。

Problem 2.II Predict the results







Answer the following questions with a short paragraph in ps5_writeup.pdf.

- How did the different models perform? How did their RMSEs compare?

维度1更好，RMSE更低

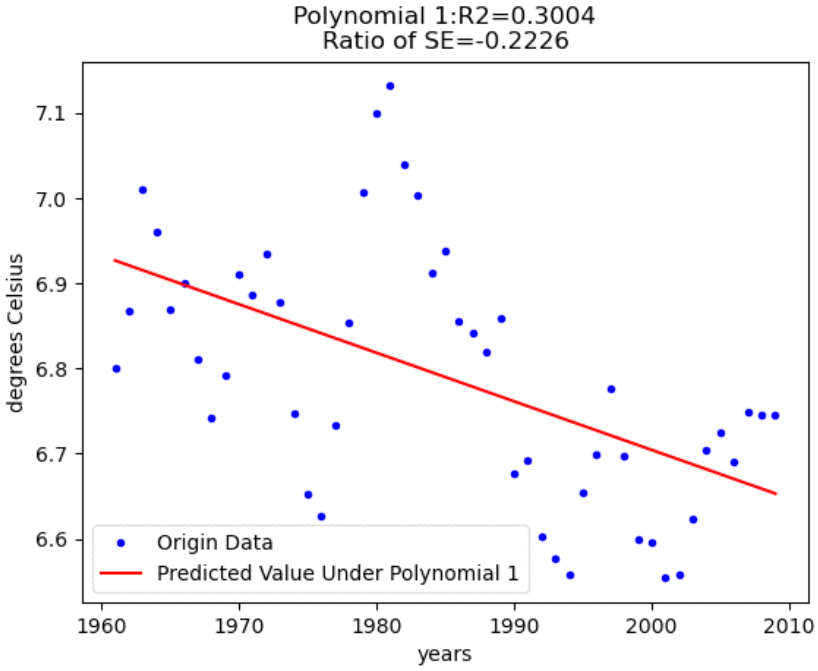
- Which model performed the best? Which model performed the worst? Are they the same as those in part D.2.I? Why?

维度1更好，维度20最差。

因为是在测试集上测试，这是真正测试泛化能力的时候。维度20存在过拟合问题，当然维度2都存在过拟合问题。

● 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?
整体 维度1 2 20的测试效果相对都要差一些。

PE



注意这个图的纵坐标应该是 **每年的温度的标准差。**

Plot the the resulting graph and include it in ps5_writeup.pdf. Answer the following questions with a short paragraph in ps5_writeup.pdf.
● Does the result match our claim (i.e., temperature variation is getting larger over these years)?
与我们的假设不符

● Can you think of ways to improve our analysis?
提高window length 可以提高R2，但是这样做，我觉得没有意义。
标准差不应该孤立来看。 如果mean 本身就处于极端的边缘（比如极寒 ）,那么标准差很小，也代表是极端天气。