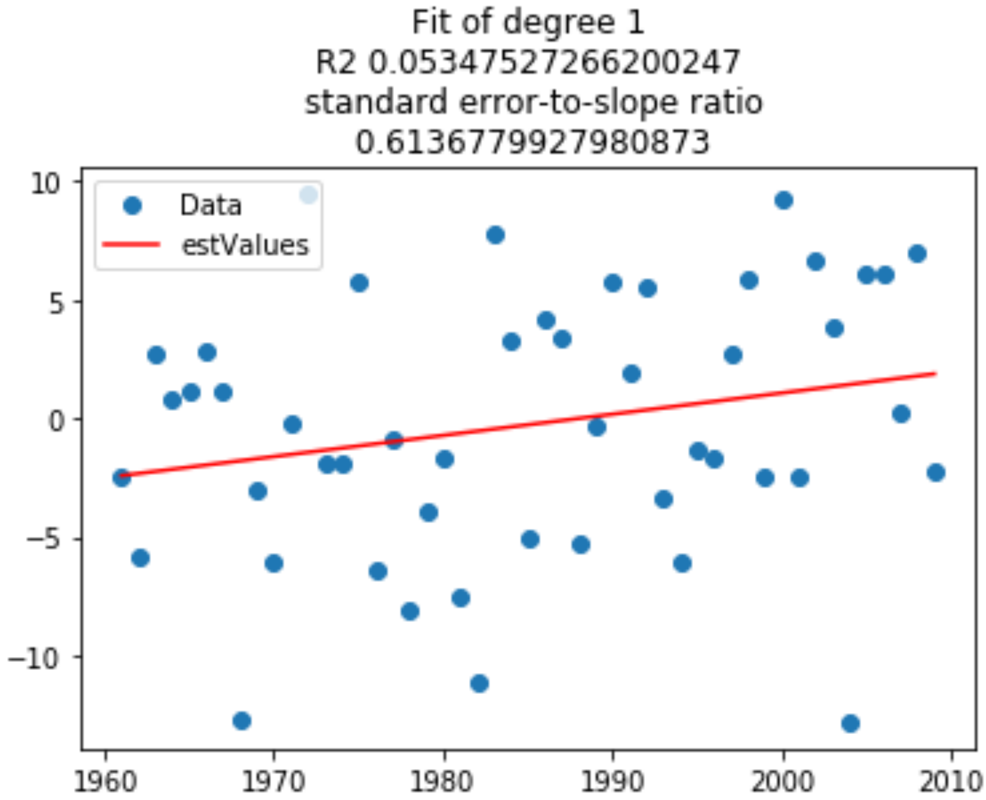
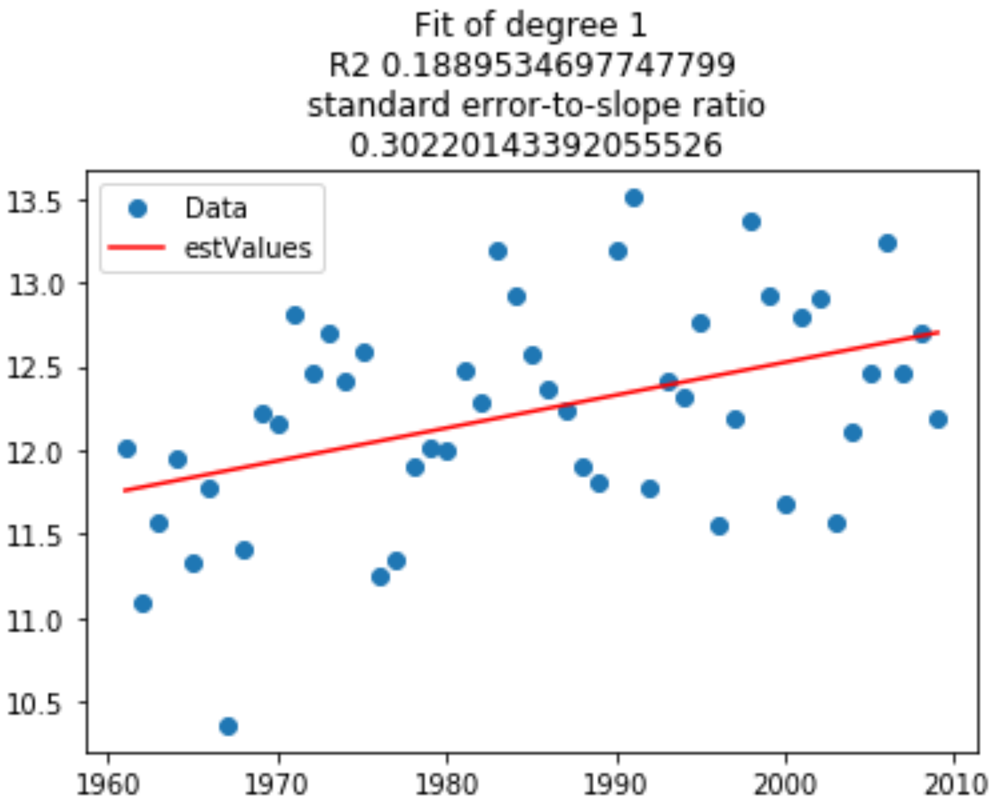
PART A

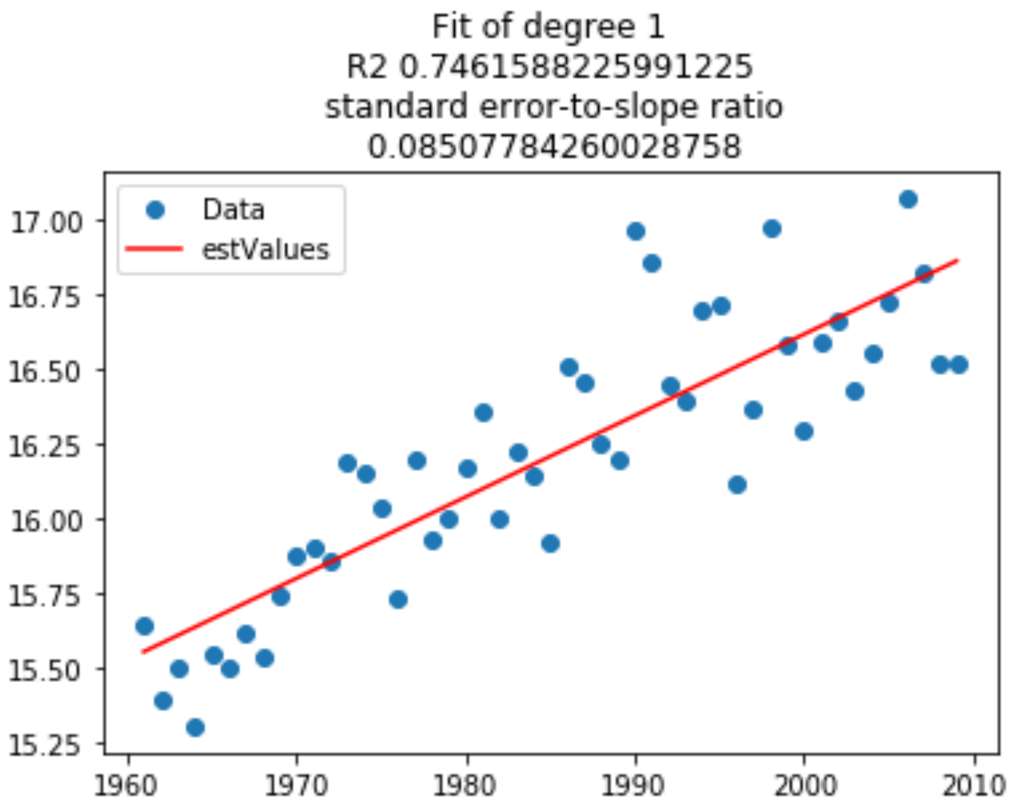

Fit of degree 1
R2 0.05347527266200247
standard error-to-slope ratio
0.6136779927980873
Data
estValues
10
5
0
−5
−10
1960
1970
1980
1990
2000
2010


Fit of degree 1
R2 0.1889534697747799
standard error-to-slope ratio
0.30220143392055526
Data
estValues
13.5
13.0
12.5
12.0
11.5
11.0
10.5
1960
1970
1980
1990
2000
2010

1) Yearly average does almost 3 times better than choosing a specific day in terms of R2 value. First plot is more widely ranged than the second plot(-10 to +10 vs 10.5 to 13.5).
2) For both plots, seems like we don't have enough data to make predictions. There are only 48 January 10th daily temperatures total for the first plot and 48 yearly averages total for the second plot in the data we use. Also extreme days seems like they cause more effect than they should.

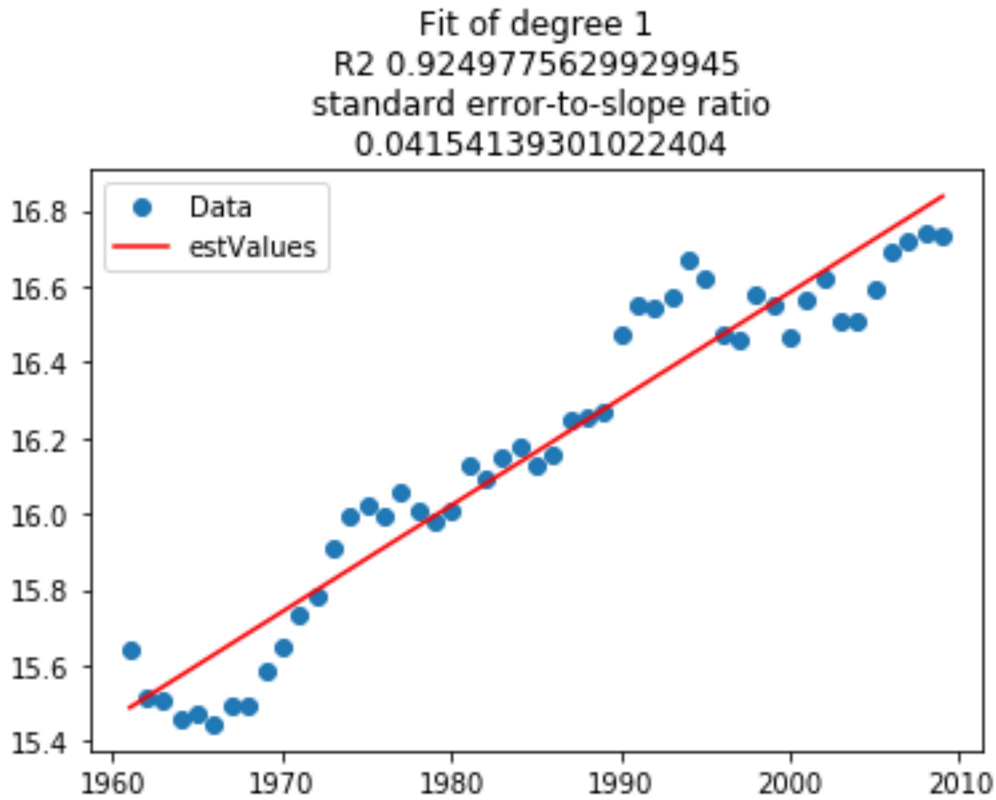
3) If we look at the scope of the model(even if it's not a very good one) we can see that it is positive so I think it supports the idea of an increasing trend in temperature.

PART B



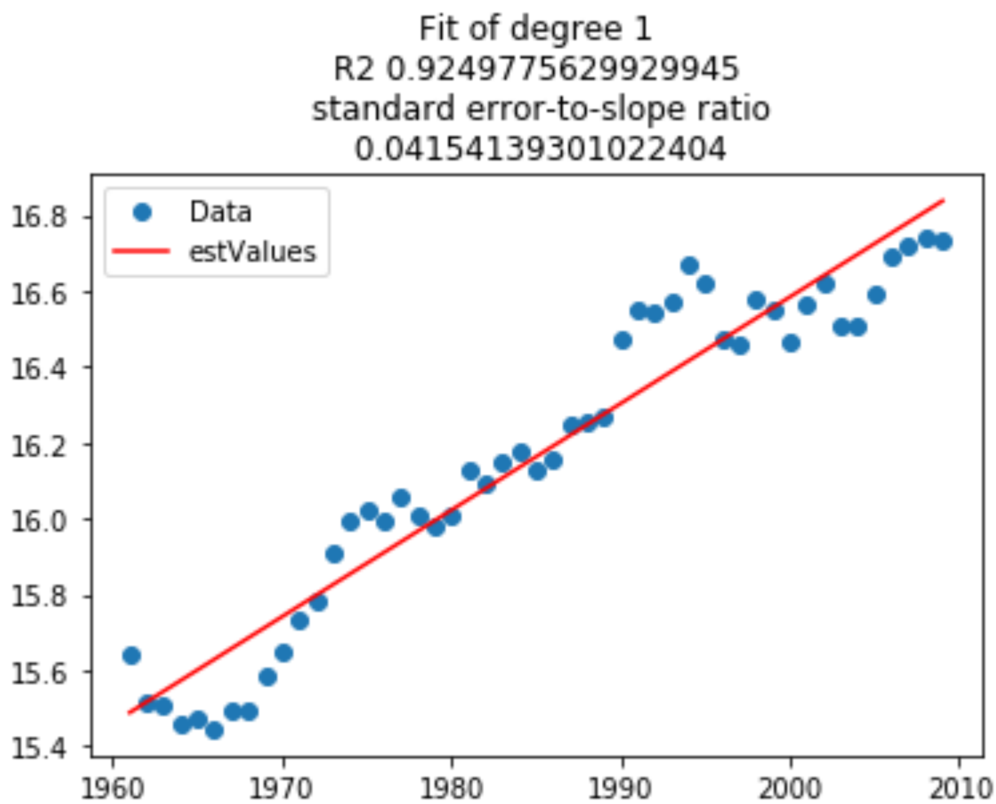
1) This graph does a lot better job than the graphs from part A in terms of R2 value. It is clear that this graph has more steady data points so model does better job about fitting the data. Scope is more easy to become aware of in this graph so it supports the increasing trend in temperature idea.
2) Because we have more data points than the first examples. Also we use average values so this means that we hava data that less widely ranged.
3) As the data size increase more likely we have a better result.
4) If all 21 cities were in same region that means less variation so that is going to cause better fit in our models.

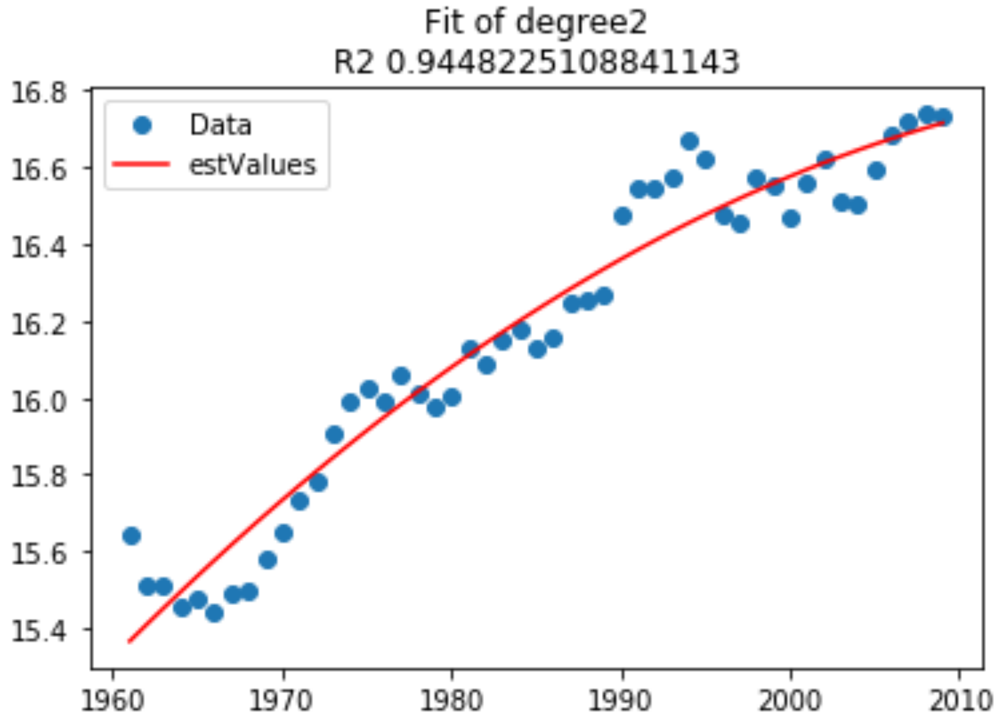
PART C

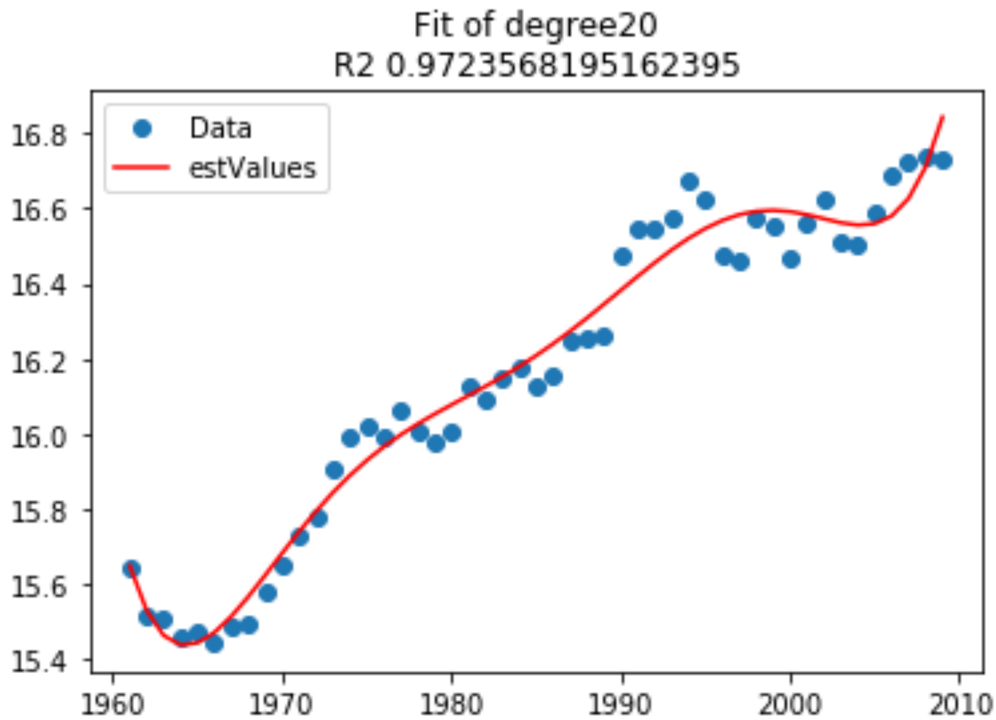


1) It is even better than the graph from part B and there is a huge difference between this graph and the graph from part A in terms of the R2 value. And it is even more clear that temperature has increasing trend.
2) With using moving averages we decreased the effect of the extreme temperatures.

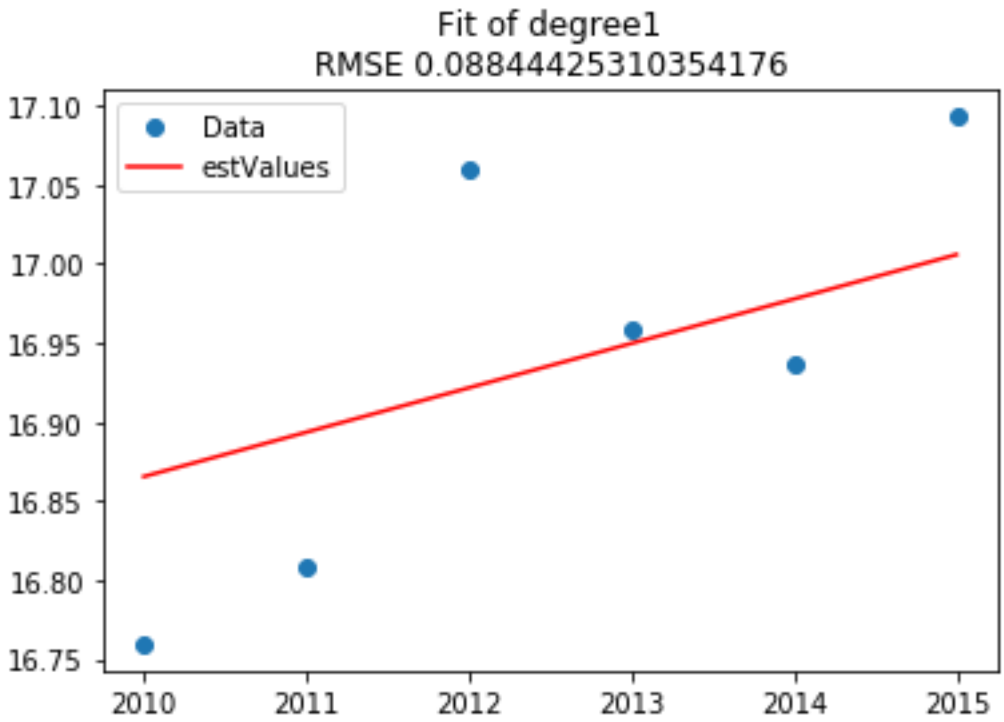
PART D

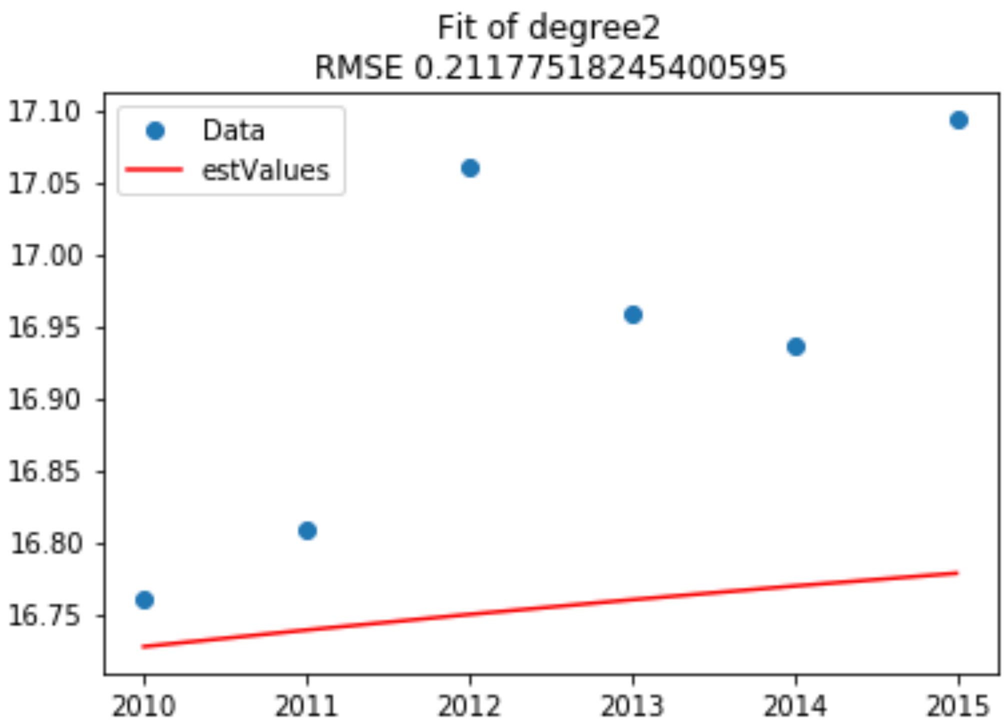


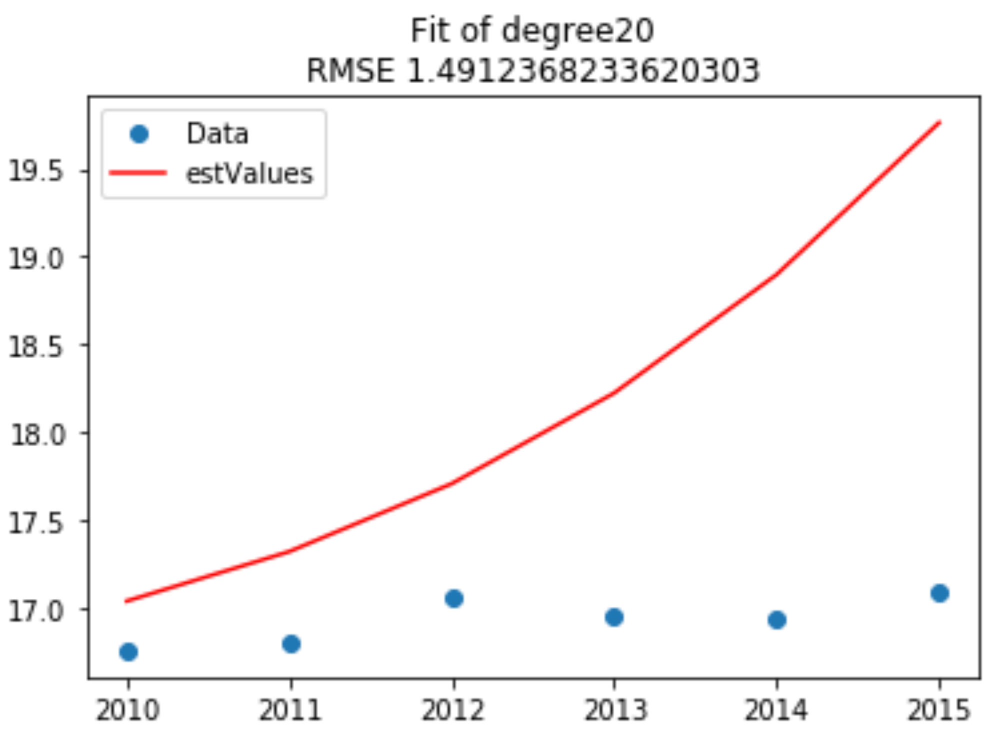




1) All models do well on training data but the best one is degree 20 polynomial.
2) Degree 20 polynomial model has the best R2 value. Since we try the guess the data we trained our model, more degree it has the better R2 value it is going to get.
3) Degree 20 polynomial fits the data best. Because we are overfitting the data.

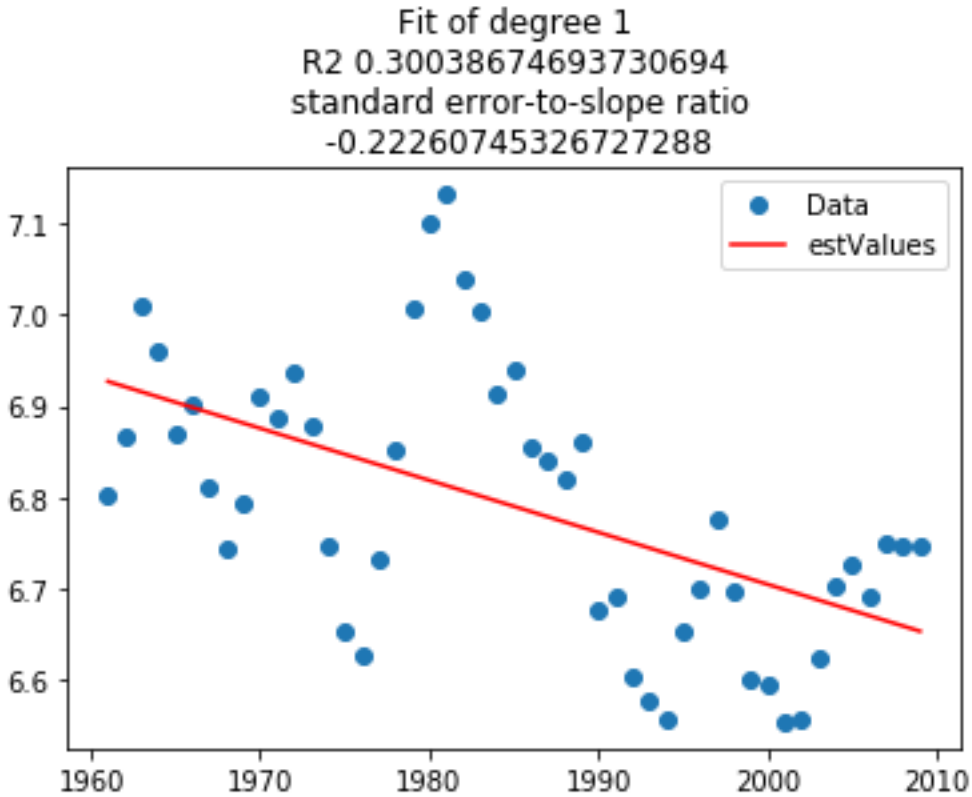

Fit of degree1
RMSE 0.08844425310354176
Data
estValues
17.10
17.05
17.00
16.95
16.90
16.85
16.80
16.75
2010
2011
2012
2013
2014
2015


Fit of degree2
RMSE 0.21177518245400595
Data
estValues
17.10
17.05
17.00
16.95
16.90
16.85
16.80
16.75
2010
2011
2012
2013
2014
2015


Fit of degree20
RMSE 1.4912368233620303
Data
estValues
19.5
19.0
18.5
18.0
17.5
17.0
2010
2011
2012
2013
2014
2015

1) None of the models seem like they fit the data well but degree 1 model has a lot smaller RMSE value than the degree 20 polynomial.
2) Degree 1 model seems like it has the best fit. Degree 20 model has the worst fit. The result is opposite of the result above. Because we overfit the data more and more as the degree of a model increases.
3) If we had generated the models using the data from Part A, since we have less data to train our model there would be more variation and as a result we would have worse results than we have right now.

PART E



1) Result doesn’t match out claim that temperature variation is getting larger over these years.
2) With calculating standart deviation we let small variations in the daily temperature to affect the result. Even though this variations small effect on the result, number of small variations is a lot more than the extreme ones. So I think calculating standart deviation is not a perfect choice for this kind of claim. Instead we should work with a model that we can tolerate small variations and more focus on the real extreme data points.