Modeling Global Warming

November 2017

1 Introduction

Donald Trump claimed that climate change was a hoax created by China, we would like to find evidence to prove him wrong. In this project, we use regression analysis to model the climate difference in US. The main road map is the following:

- First clean and visualize data from small set to large, and try to make data less noisy and obtain long term trends.
- Second, using historical data to predict future temperatures.
- Third, we want to investigate extremity of temperatures rather than just warming.

The data is obtained from the National Centers for Environmental Information (NCEI). It is stored in data.csv, contains daily maximum and minimum temperatures in 21 US cities from 1961 to 2015.

2 Model one city's temperatures

We aimed to see if there is any trend over the years, our schema is to fit a ploynomial for $(x^{(i)}, y^{(i)})$,where $x^{(i)}$is year like 1997, $y^{(i)}$is temperature in Celsius. Specifically we estimate y through:

$$y = \beta \cdot f(x) + \epsilon$$

Notice $f(x)$ is determined by the degree of polynomials, it captures the complexity of models. To avoid overfitting, we evaluate models by computering R^2 as follows:

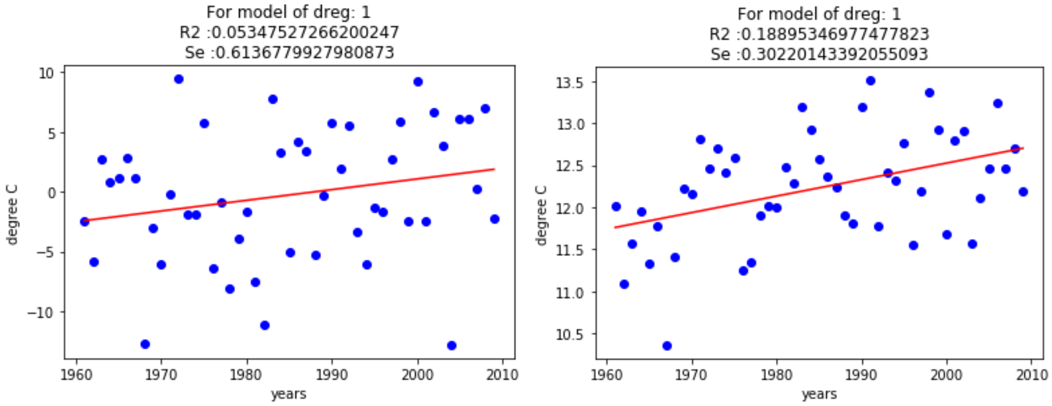
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)} - \overline{y})^2}$$

The training set is from 1961 to 2009, we now grasp training data from two methods:

- One is randomly select one day (10, January) in one particular city . Figure for polynomial fitting.
- Another way is to computer average temperature of each year.

The outcome in **??** leads us to the following observations:

Figure 1



1. Choosing specified day doesn't fit as much well as average temperature, from figures above, the latter fits more well, and R2 is large $0.18 > 0.05$.

2. Because there are a lot of other affects except year, so year does not explain quite much. Also, the specific day plot is more noisy.

3. In fact, the latter one can support the claim (global warming), the slope is obversely larger than 0, and the *se* is $0.3 < 0.5$, so the trend is not by chance.

3 Incorporating more cities

One way to decrease noise is adding data. Now we includes 20 more cities and do the same curve fitting.

1. This plot 2a fits data quite well, it has much lager R^2 (0.74), and *se* is smaller (0.08), so evidently we can see the trend, so it strongly supports the claim.

2. Because if we average on the nation level, the noise can effectively be lowed.

3. So if only 3 cities, it could not fit that well, if 100, it could fit much more due to the fact of Law of Large Number.

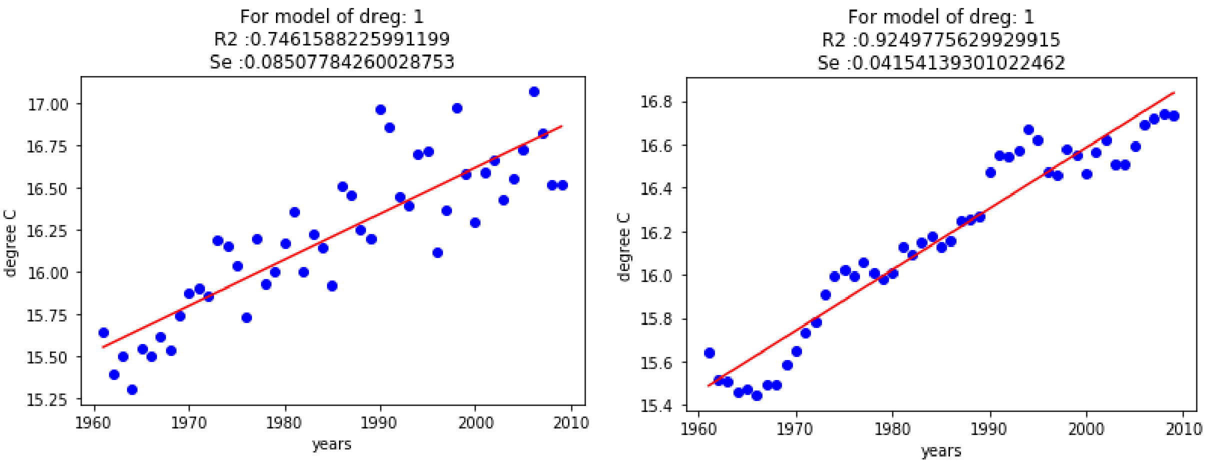
4. If all the cities adjacent, that is a biased sample which doesn't give us more help than only New England, therefore the result will fit not much well.

Notice that moving average is another method to decrease noise and emphasize general trends over local fluctuation.We computer moving average as follows: suppose $y = [1, 2, 3, 4, 5]$, then output is $[\frac{1}{1}, \frac{1+2}{2}, \frac{1+2+3}{3}, \frac{2+3+4}{3}, \frac{3+4+5}{3}]$. After this operation, the outcome of fitting shows in figure 2b.

1. This plot fits data more well, R^2 is high than previous ones, it interprets 90% of y, and *se* is smaller.

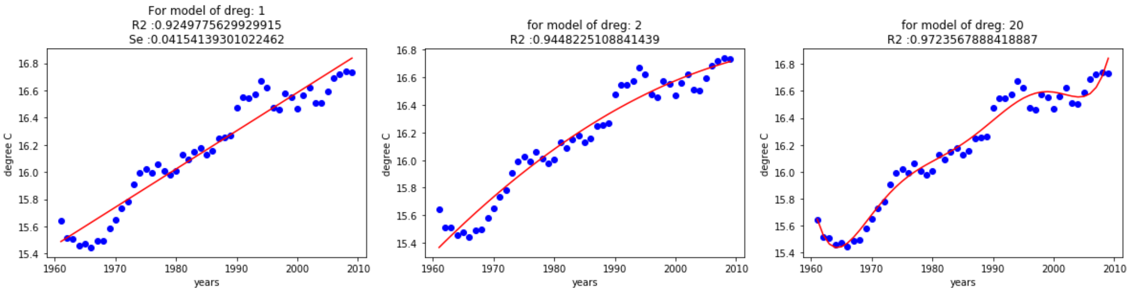
2. By moving the average, we can decrease noise not only between areas, but between adjacent years.

Figure 2


(a) More data fitting

(b) Fitting after moving average

Figure 3: 1,2,20 degree of polynomials



4 Prediction and test

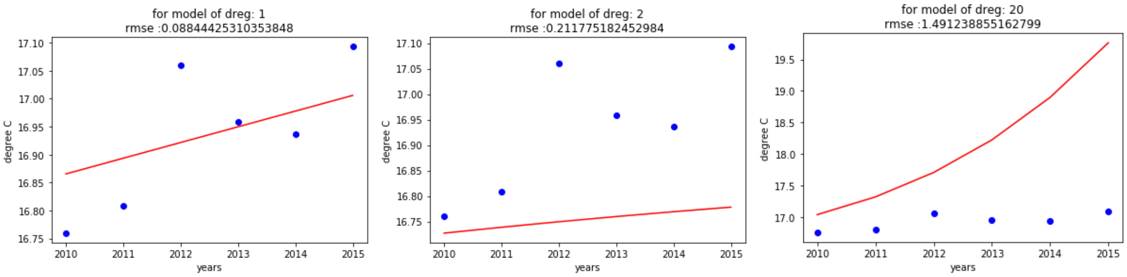
We generate $1, 2, 20$ degree of ploynomials for the training set, the result is **??**. Then we using temperatures from 2011 to 2015 to evaluate models and related metric is *root mean square error (RMSE):*

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - e_i)^2}{n}}$$

The prediction result is in **??**.
Some conclusions from the test:

1. Model_1 is better than model_2, and model_2 is better than model_3. And *rmse* is model_1 < model_2 < model_3.

Figure 4: Predictions in test set



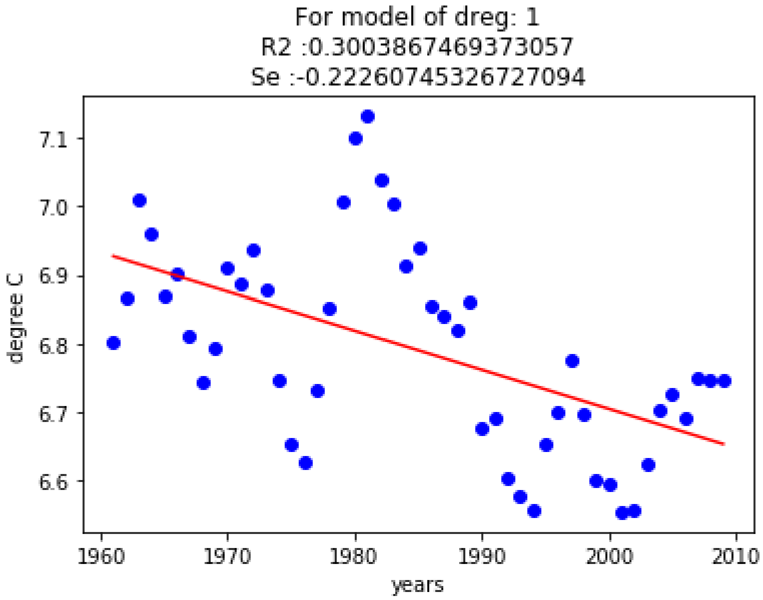
2. Model_1 is best, model_3 is the worst, because model_3 has higher degree, and cause overfitting problem.

3. It's basically worse when increase model complexity, because that model using much noisy data.

5 Extreme temperatures

The consequence of global warming includes extreme temperatures, like being very hot or very cold. We measure the standard deviation to model this affect. We expect that over time the standard deviation increases.

We computer the moving average of *std* from 1960 to 2010, surprisingly, the following figure does not support our claim, the *std* has minor decreasing trend.


For model of dreg: 1
R2 :0.3003867469373057
Se :-0.22260745326727094
degree C
years