Problem 4 (Modeling Changes in Republicans and Sunspots, 15pts)

The objective of this problem is to learn about linear regression with basis functions by modeling the number of Republicans in the Senate. The file `data/year-sunspots-republicans.csv` contains the data you will use for this problem. It has three columns. The first one is an integer that indicates the year. The second is the number of Sunspots observed in that year. The third is the number of Republicans in the Senate for that year. The data file looks like this:

```
Year,Sunspot_Count,Republican_Count
1960,112.3,36
1962,37.6,34
1964,10.2,32
1966,47.0,36
```

You can see scatterplots of the data in the figures below. The horizontal axis is the Year, and the vertical axis is the Number of Republicans and the Number of Sunspots, respectively.

(Data Source: http://www.realclimate.org/data/senators_sunspots.txt)

Make sure to include all required plots in your PDF.

1. In this problem you will implement ordinary least squares regression using 4 different basis functions for **Year (x-axis)** v. **Number of Republicans in the Senate (y-axis)**. Some starter Python code that implements simple linear regression is provided in `T1_P4.py`.

   First, plot the data and regression lines for each of the following sets of basis functions, and include the generated plot as an image in your submission PDF. You will therefore make 4 total plots:

   (a) $\phi_j(x) = x^j$ for $j = 1, \ldots, 5$
   ie, use basis $y = a_1 x^1 + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5$ for some constants $\{a_1, ..., a_5\}$.

   (b) $\phi_j(x) = \exp \frac{-(x-\mu_j)^2}{25}$ for $\mu_j = 1960, 1965, 1970, 1975, \ldots 2010$

   (c) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 5$

   (d) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 25$

   * Note: Be sure to add a bias term for each of the basis functions above.

   Second, for each plot include the residual sum of squares error. Submit the generated plot and residual sum-of-squares error for each basis in your LaTeX write-up.

Problem 4 (cont.)

2. Repeat the same exact process as above but for **Number of Sunspots (x-axis)** v. **Number of Republicans in the Senate (y-axis)**. Now, however, only use data from before 1985, and only use basis functions (a), (c), and (d) – ignore basis (b). You will therefore make 3 total plots. For each plot make sure to also include the residual sum of squares error.

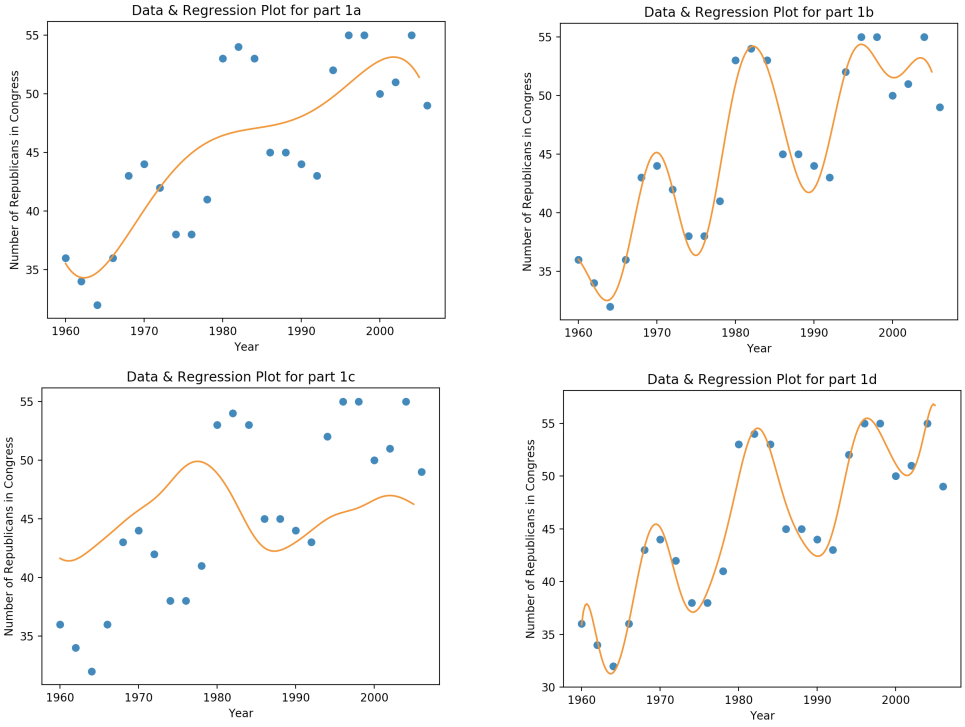
   Which of the three bases (a, b, d) provided the "best" fit? **Choose one**, and keep in mind the generalizability of the model.

   Given the quality of this fit, do you believe that the number of sunspots controls the number of Republicans in the senate (Yes or No)?

Solution

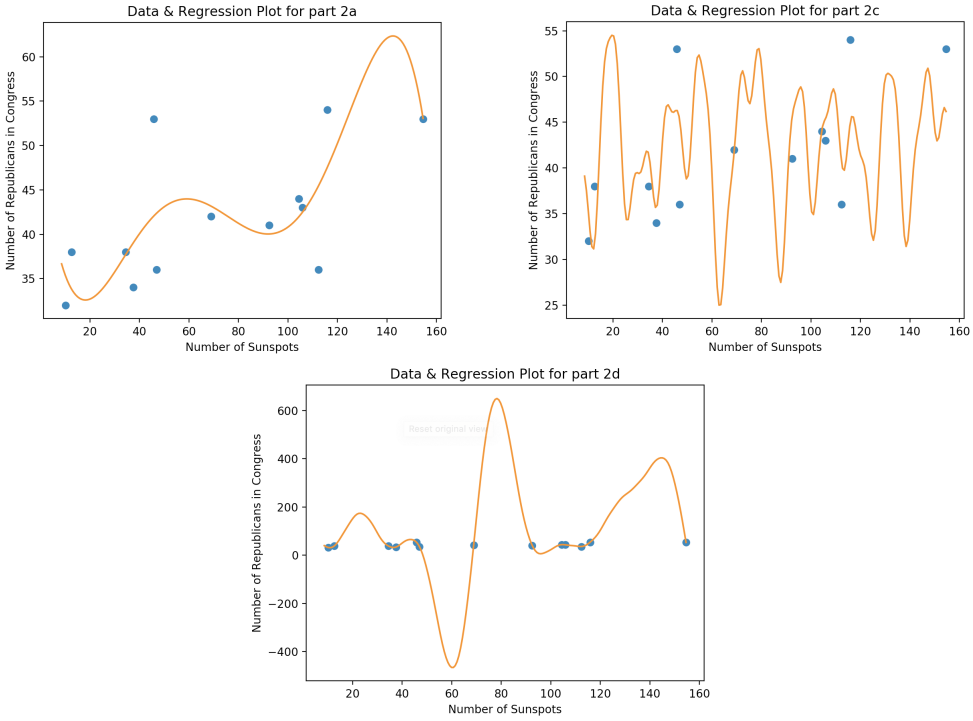
1. The residual sum of squares error for each basis are as follows in the table (organized based on the question part a-d of the basis):

Question Part (a-d)	L2 Loss
a	394.9803839890865
b	54.2730966167196
c	1082.8088559867185
d	39.001226916562295



2. The residual sum of squares error for each basis are as follows in the table (organized based on the question part a-d of the basis):

Question Part (a-d)	L2 Loss
a	351.22793577417474
c	375.10675778167393
d	8.622599569654343e-22



I would choose basis (a) as providing the best fit. Observing the plot and the loss for (d) shows that the basis for (d) is severely over-fit, as the regression line almost perfectly predicts every point in our training data set. By the bias-variance decomposition, basis (d) is an undesirable basis because although it has small bias on our training data, it will not be very generalizable to new unseen data. Moreover, we can see in the plot of basis (d) that we have meaningless predictions, as the range of our regression runs from -400 to 600, which is far too large to predict for number of Republicans in the Senate. Basis (c) has too much variance to be a good choice of basis, as evidenced in the plot. In particular, basis (c) is not very generalizable, since for example, a prediction for a value of 60 sunspots is drastically different than a prediction for a value of 65 sunspots as evidenced in the plot. This is due to the excessive variance of basis (c). Therefore, (a) is the best fit to use as it is the most generalizable model.

Regardless of the quality of the fit, I don't believe the number of sunspots controls the number of Republicans in the Senate because my real-world intuition tells me that the two should be uncorrelated. A more likely explanation is that both the Number of Republicans in the Senate and the number of sunspots follow a cyclical pattern over time similar to a cosine wave, thus explaining their correlation to each, when in fact they both have similar correlations to Years which is the more likely explanation.