Week 3: Statistics and Exploratory Data Analysis

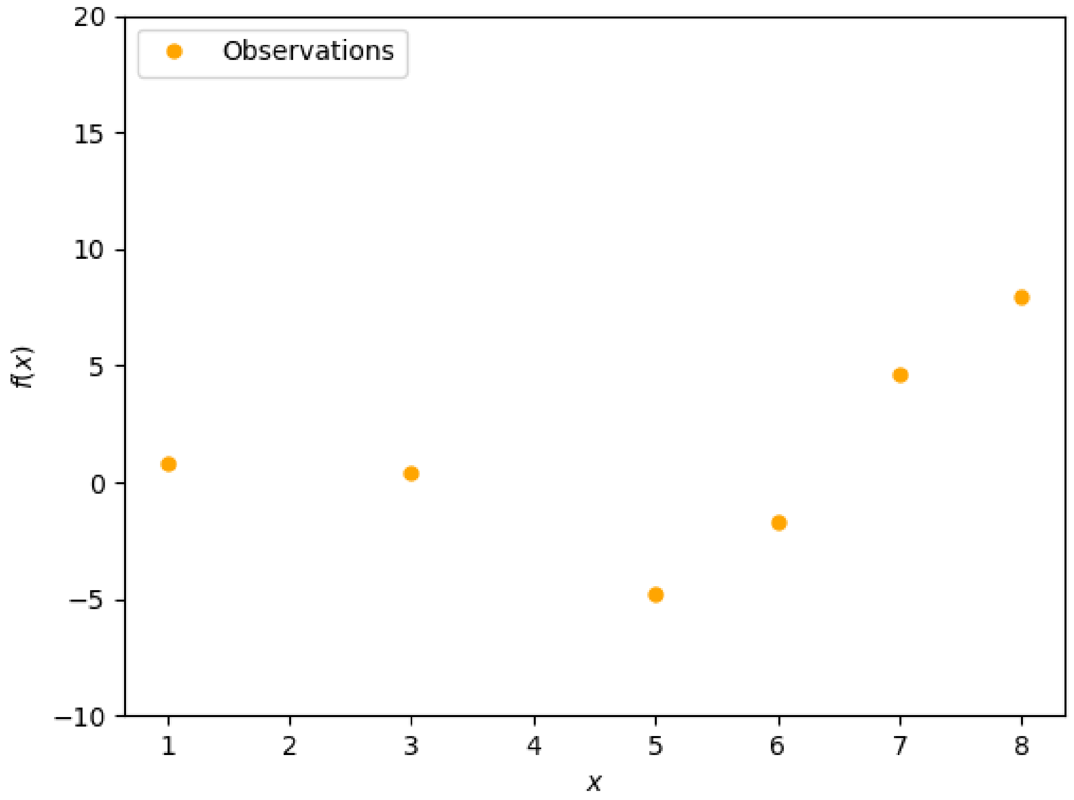
SESSION 2: MAKING STATISTICS WORK FOR YOU

STARFISH SCHOOL 2021

Smoothing and optimization

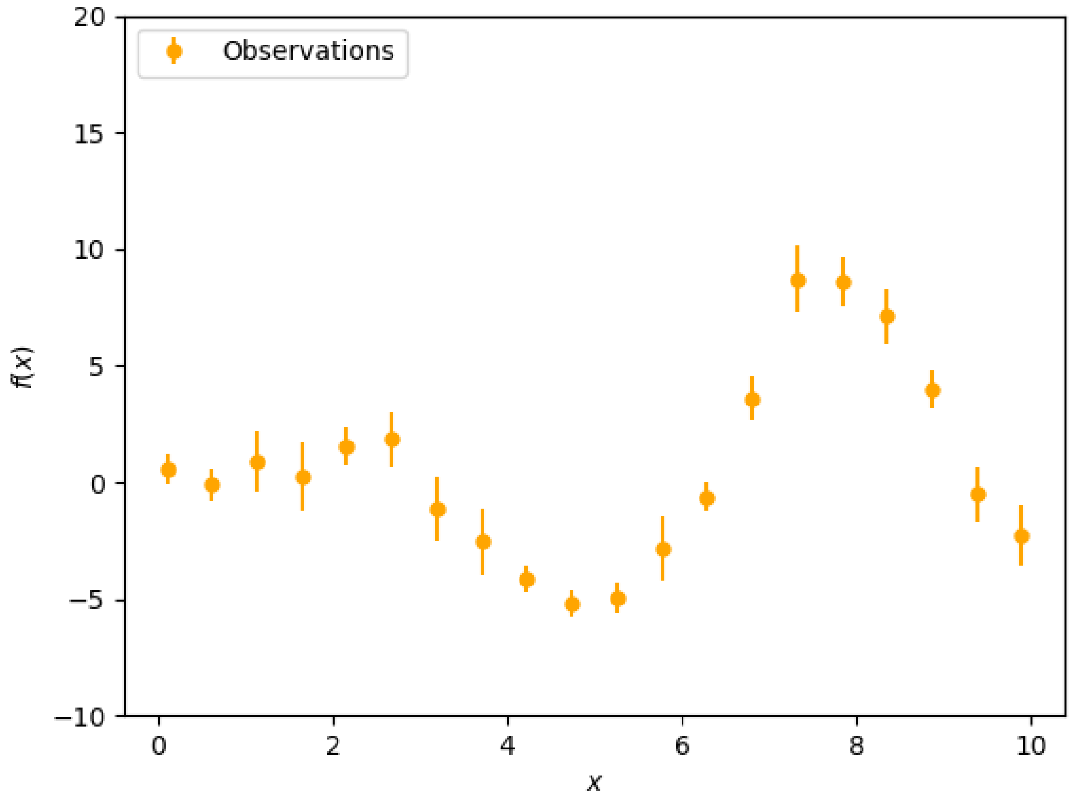
How to make data more pliable

- Sometimes you'll get data that looks like this:



How to make data more pliable

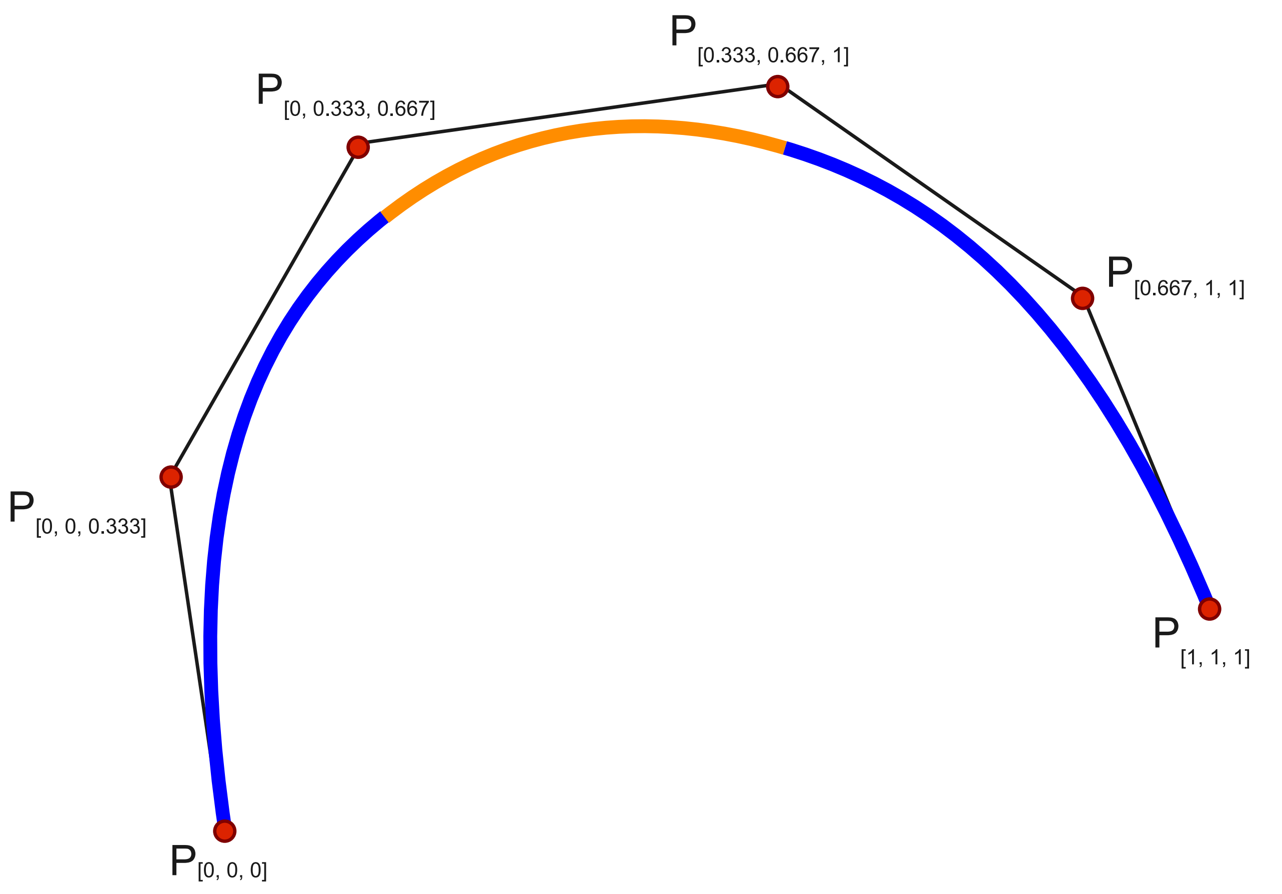
- or this (with error bars)



Splining

- The first thing you might think to do is to spline the data.
- Spline is a polynomial fit between points that ensure that the curve you fit goes through each point you have.

- The smoothness depends on the order of the polynomial (e.g. linear, quadratic, cubic)
- Splines are very bad at *extrapolation* and can suffer if you have large gaps between points



Python's curve_fit

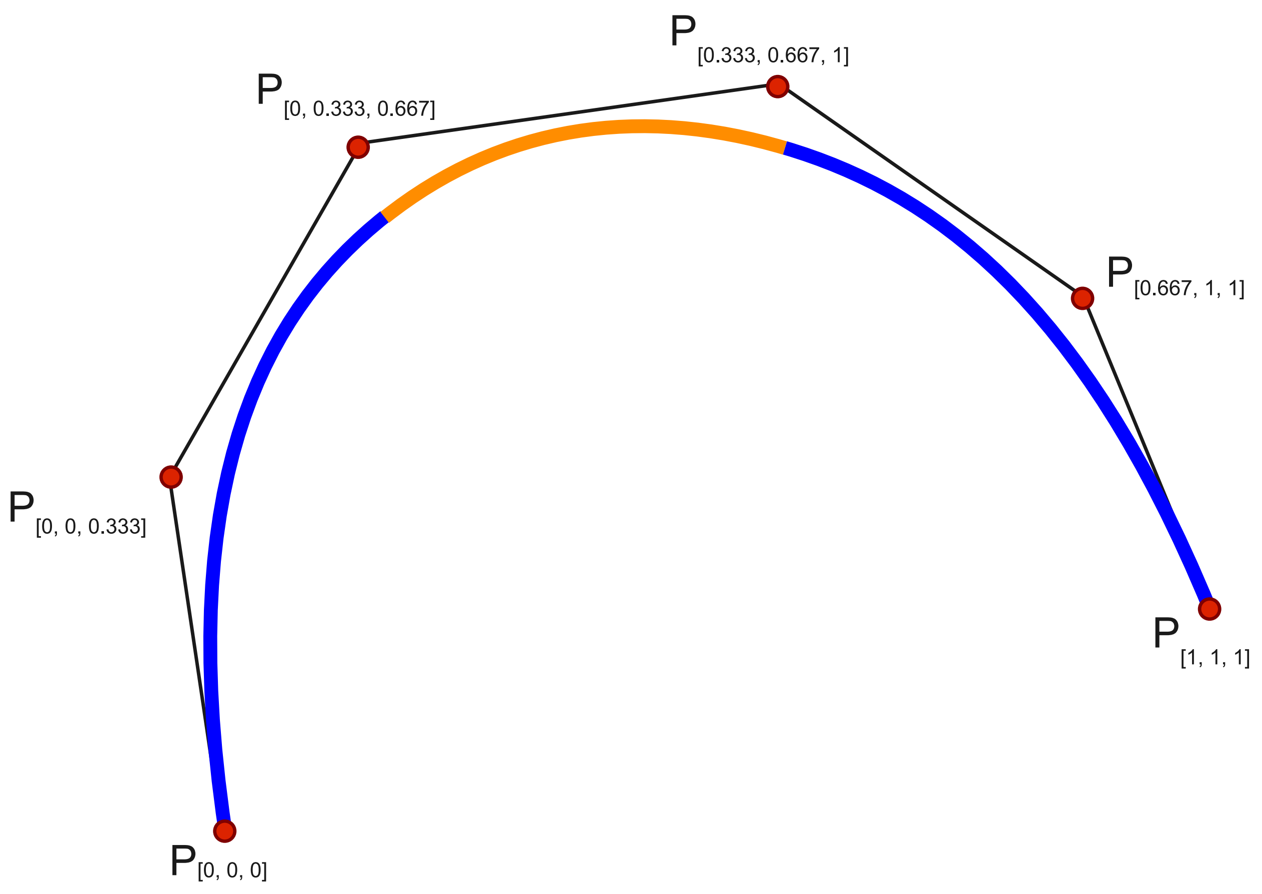
- If you think you can guess the functional form of the curve you can always fit for the parameters of that curve.
- In python that is through functions like scipy.curve_fit

"""

e_fit(ff,x,y)

HOT TIP

Curve Fitting (also known as "optimization") is an open ended problem, that leads all the way to cutting-edge deep learning of today!



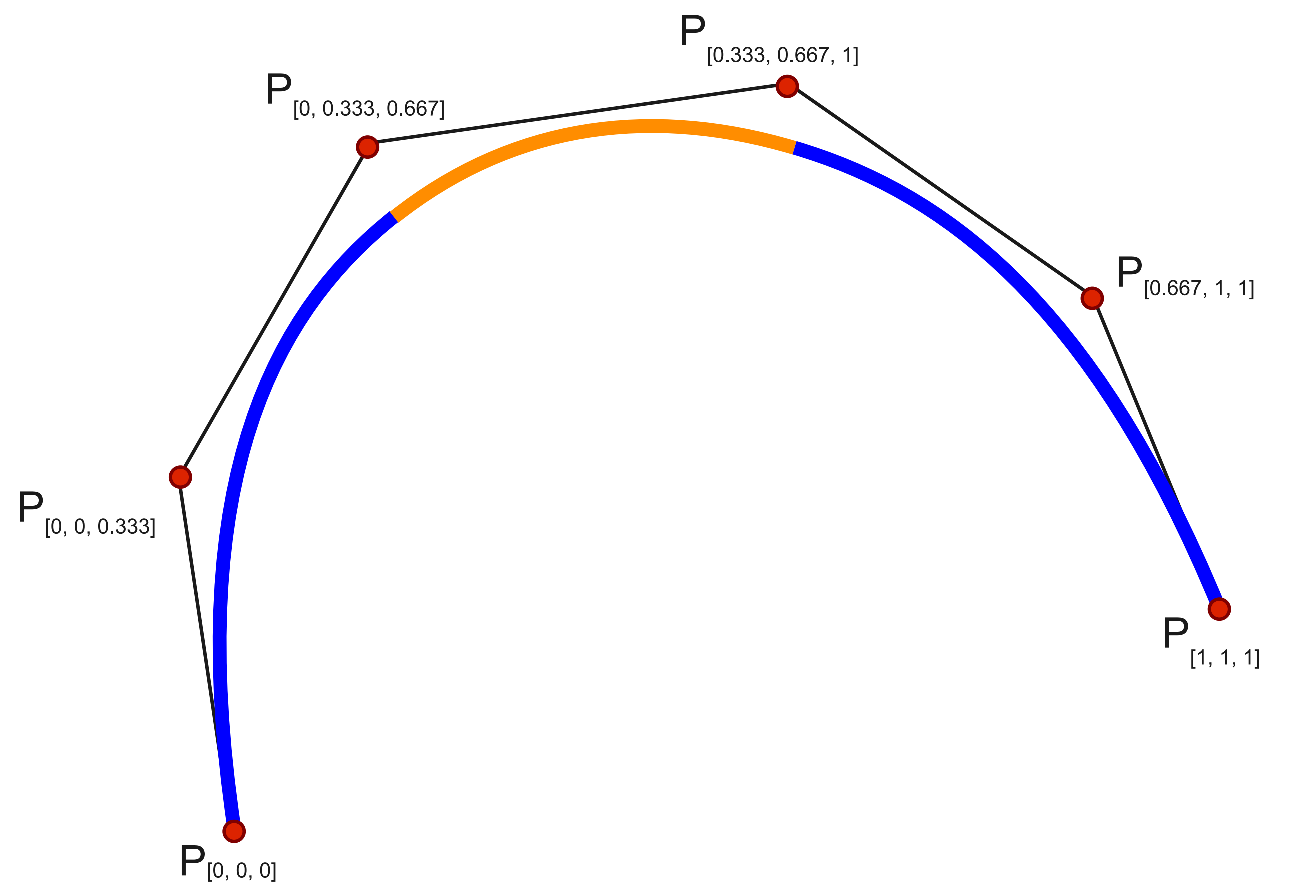
Python's curve_fit

- If you think you can guess the functional form of the curve you can always fit for the parameters of that curve.
- In python that is through functions like scipy.curve_fit

- def ff(x,a, b):

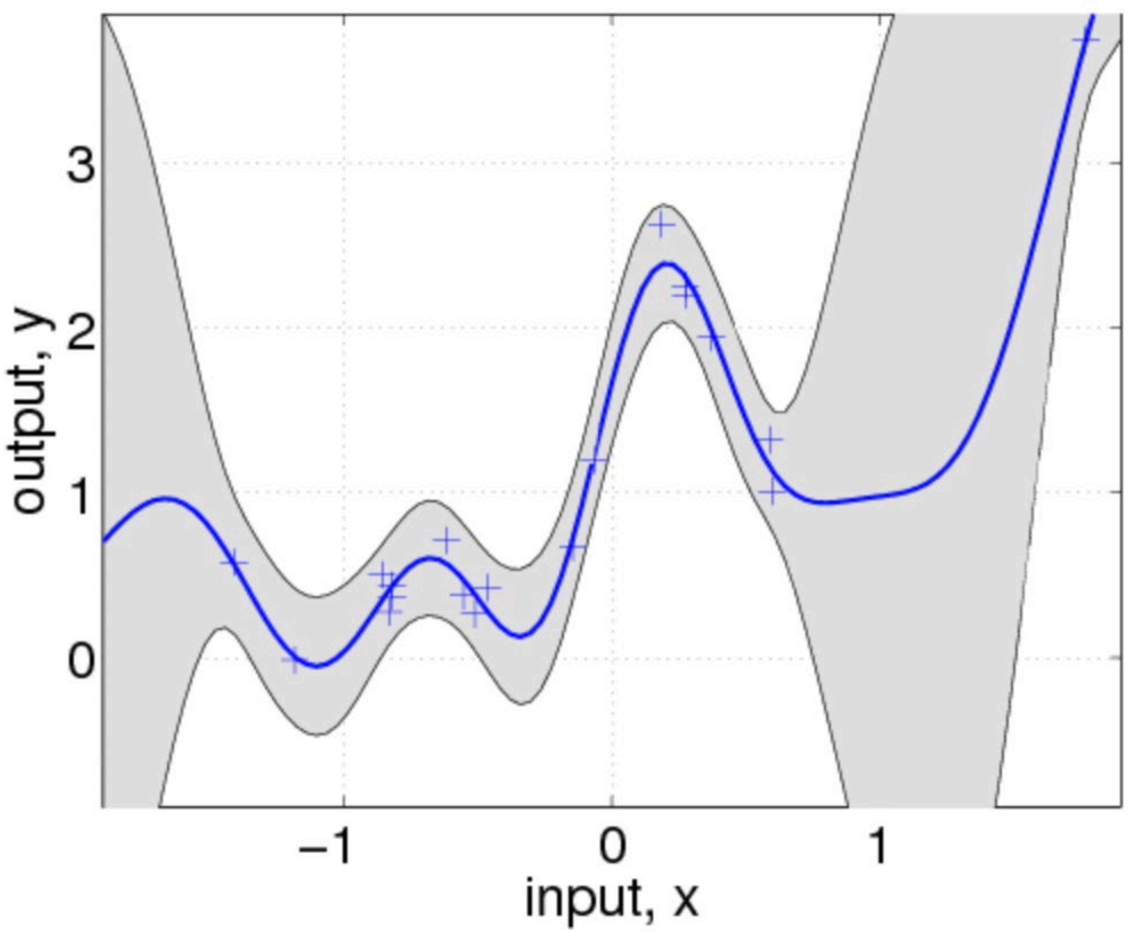
  """The function to predict."""

  return a*x * np.sin(b*x)
- fitparams, fiterror = curve_fit(ff,x,y)



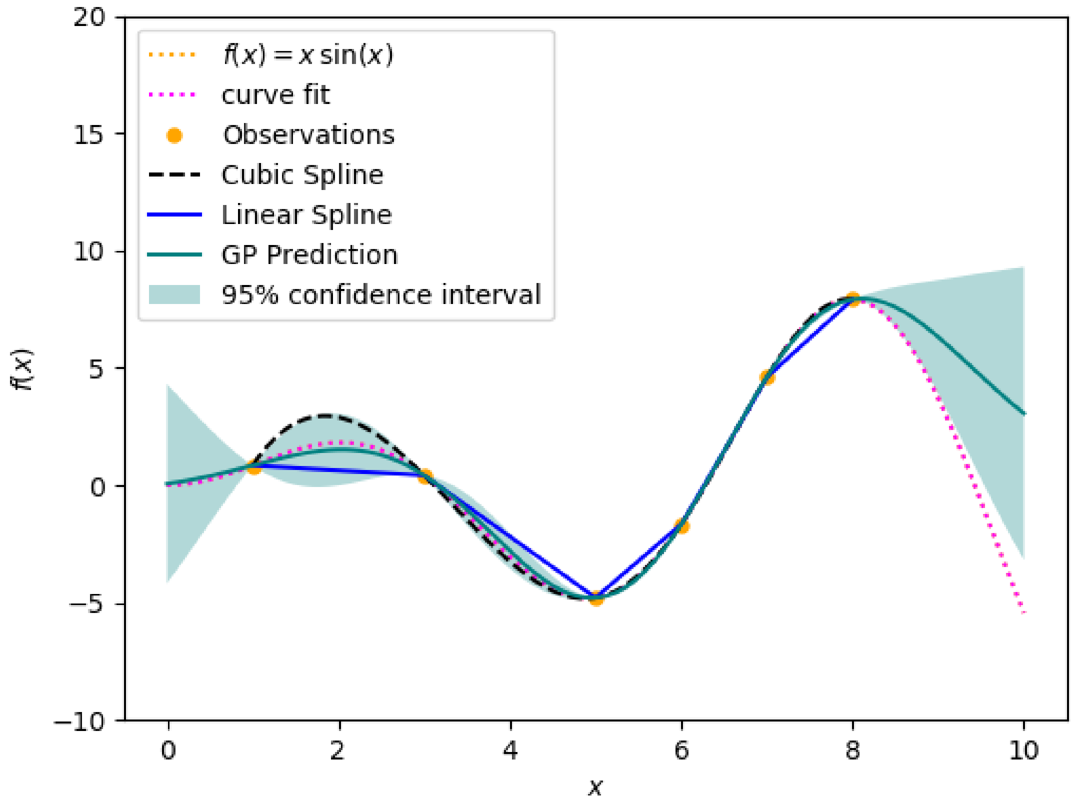
Gaussian process

- Modelling data as a Gaussian Process (GP) assumes that every point is modelled as a series of multi-variate Gaussians in a linear combination. It gives you the error on your fit -- [I like to call it the 'sausage of uncertainty'
- The key thing with GP modelling is specifying the "sigma" of the GP – or in multi-dimensional space, the *kernel*



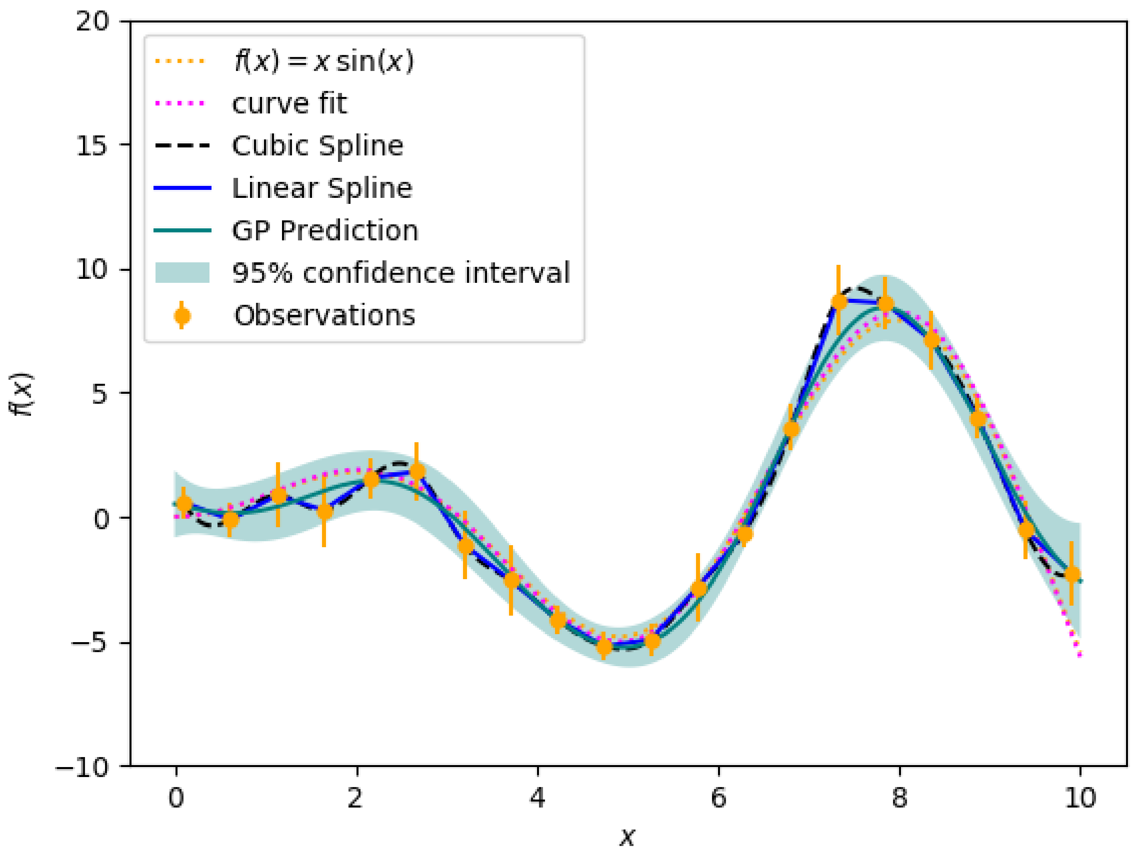
Examples

- Simple example from Vanderplaas ++
- This code is provided in your exercise set – and shows a combination of methods



Examples

- Simple example from Vanderplaas ++
- This code is provided in your exercise set – and shows a combination of methods



LOWESS

- Lowess and Loess, Clearly Explained!!! - YouTube

Exercise

1. Download the data set `xvalues.csv` from the website
2. Generate a histogram for these values using bin widths of 2, from -8 to 4. *Before going to part b),* what do you notice about this distribution? Would you hypothesize what distribution the data came from?
3. Generate a new histogram for these values using bin widths of 2, starting instead from -7.
4. Make a boxplot of these data and find the summary statistics
5. Make a kernel density estimate plot of the distribution. How does this compare to the other options?
6. Based on your figures, comment on the pros and cons of each estimate of the distribution (histogram, boxplot, KDE)
7. Standardize the data from question 1, and make a new histogram and boxplot. Compare these to your histogram and boxplot in question 1.
8. What are the mean and standard deviation of the standardized data?
9. Check the 68-95-99 rule using the standardized data. Is the empirical rule applicable here? Why or why not?

Stretch Goal:

Make an empirical CDFs of the data and compare to the CDF of a normal. Or make a Q-Q plot (look up what this is)!