# Predicting the network usage, using the linear regression algorithm

## Reading & Cleaning the data

Data: A month’s web traffic data with 2 columns

* Time (Starting 1st hour in Day-1, week-1 till nth hour in Day-7, week-4)
* Hits/Hour



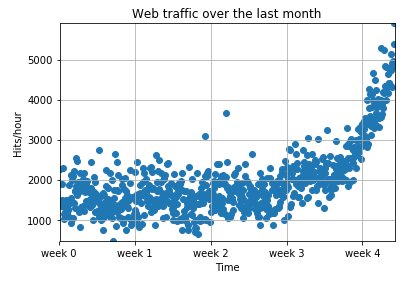
As part of data cleaning, we have eliminated NULL values in the 2nd column

## Finding the best-fit model using Ordinary Least Squares method

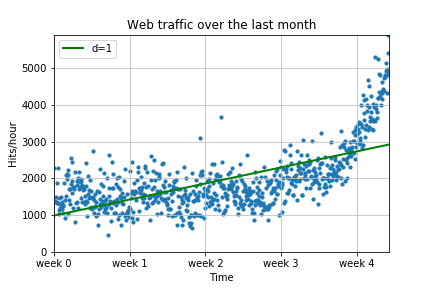
Y = aX + b + error

Predicting the value of Y based on the variable X, ‘a’ is the slope of the line and ‘b’ is the Y intercept at X=0. Error is the regression error for each observation. This error will guide us in choosing the right model. This error will be calculated as the squared distance of the model's prediction to the real data.

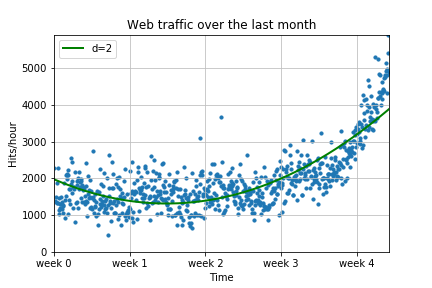
Plotting a graph with X as Time and Y as Hits/hr



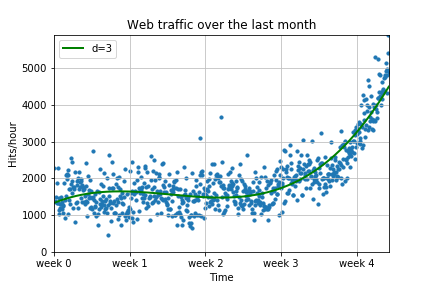
Starting with polynomial degree of 1



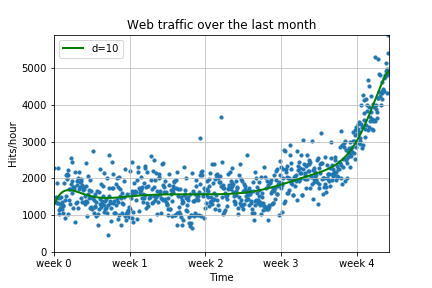
With polynomial degree of 2



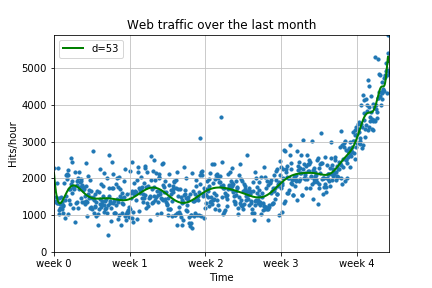
With polynomial degree of 3



With polynomial degree of 10



With polynomial degree of 100



Error for above each polynomial:

F1 - 317389767.34

F2 - 179983507.878

F3 - 139350144.032

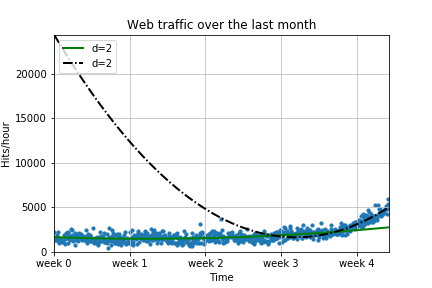
F10 - 121942326.364

F100 - 109452402.923

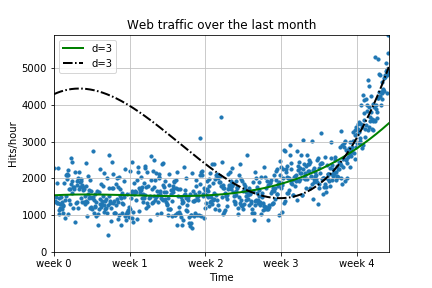
Looking at the graphs, model - F10 and F100 are over-fitting as the linear-regression line has more fitted even with the noise. F1 is general fit, as we distance value is too much.

Out of F2 and F3, if we observe more on the variation of summed squared distance across the weeks, we see 2 different sets i.e. week 1 to week 3.5 and week 3.5 & further. So to understand it further we will inflect the data at week 3.5, i.e. separating data into 2 sub sets and calculating the Summed squared distance for each set separately and summing them up.

With polynomial degree of 2



With polynomial degree of 3



error value for above each polynomial on 2 sub-sets of Data:

F2 - 122584668.146112

F3 - 121352906.776585

## Conclusion

Model F3 is the best fit model