Data Analytics Assignment - I Report by Abhishek Kumar (Sr. no. 15648)

1. Problem Statement:

Using only the first innings data in the given data set, find the best fit 'run production functions' in terms of wickets-in-hand w and overs-to-go u. Assuming that the model follows below curve.

$$Z(u, w) = Z_0(w)[1 - exp^{\{-Lu/Z_0(w)\}}] \tag{1}$$

Use the sum of squared errors loss function, summed across overs and wickets.

Note: your regression forces all slopes to be equal at u = 0. You should provide a plot of the ten functions, and report the (11) parameters associated with the (10) production functions, and the total error.

Dataset: ODI over-by-over data (Taken from the website http://www.cricketabstract.com)

2. Solution

2.1) **Data Pre-Processing**
The given data file contains data on ODI matches from 1999 to 2011. The file contains 126768 rows and 38 column but we have to estimate our curve using I^{st} innings data only. So, after removing data of II^{nd} innings, the file contains 67794 rows and 38 columns only.
To trace the above required curve we need the following data :
1) Wickets in hand (from column Wickets.in.Hand)
2) Over Remaining (from column Over)
3) Run scored (in remaining over) : after subtracting column Total.Runs from column Innings.Total.Runs

2.1) **Curve Tracing Approach**
I have used the following algorithm to generate the run production functions

Result: Parameter Z ← list of 10 numbers , l
Initialization : Intialize Z and l with some initial values;
while *While loss is not minimized* **do**
  Predicted Run Score : Calculated using the curve $Z(u, w) = Z_0(w)[1 - exp^{\{-Lu/Z_0(w)\}}]$ where Z and l are curve parameter and u is over.;
  Loss : (Actual Run score - predicted Run score)^2 (Sum over all the data points);
  Calculate new value of (Z and l) using some method such that the loss is minimized.;
  update Z and l;
end

Algorithm 1: Find curve parameter

To minimize the loss, I have used scipy.optimize.minimize function which takes the loss function, Z and l as input and produces the updated value of the parameter and total loss using these parameters.

3. Result

These are the values of the curve parameter Z and L:
The value of parameter L is : 11.54277498
The values of parameters Z are:

Z_0	10.713
Z_1	36.015
Z_2	64.032
Z_3	82.825
Z_4	103.622
Z_5	138.95
Z_6	170.237
Z_7	200.356
Z_8	241.810
Z_9	273.761

Total Loss : 147335263.72064668

Plots:

The plot of 10 functions is as below: