Data Analytics:Assignment-1

Lubhawan Parihar (SR Number-19239)

Duckworth-Lewis-method

Task :

Using the first innings data alone in the above data set, find the best fit 'run production functions' in terms of wickets-in-hand *w* and overs-to-go *u*. Assume the model $Z(u,w) = Z_0(w)[1 - \exp\{-Lu/Z_0(w)\}]$. Use the sum of squared errors loss function, summed across overs, wickets, and data points for those overs and wickets

$$Z(u,w) = Z_0(w)[1 - exp(-Lu/Z_0(w)]$$

Report following:

- Plot of the ten functions
- Report the (11) parameters associated with the (10) production functions.
- The normalised squared error (total squared error summed across overs, wickets, and data points for those overs and wickets, and normalised by the total number of data points across all overs and wickets).

Methodology :

1. Extracted the data columns from the csv file which is required that is **innings number, remaining runs, remaining overs, wicket remaining.**

2. In this step I cleaned the data using 'Error in data' and removed the rows with value 1, also removed the rows with wickets remaining 0.

3. In the next step I took formed the leastsq function from scipy.optimize library and created a function for finding the residue and defined the loss function as sum of squared loss function. The function is defined as :

$$\underset{Z_0(1),Z_0(2),\ldots,Z_0(10),L}{\text{minimize}} \sum_{n=1}^{N} (y_n - Z(u_n, w_n, Z, L))^2$$

   Where N=number of all the first inning data-points. y_n is the actual output/run and $Z(u_n,w_n,Z,L)$ is the predicted run.

4. Then I found out MSE (mean squared error) using **error_func** that I created for **leastsq**.

Results :

From above observation I found:

TOTAL LOSS: 1559.30(MSE)

Z[1]	Z[2]	Z[3]	Z[4]	Z[5]	Z[6]	Z[7]	Z[8]	Z[9]	Z[10]	L
11.69	26.8	50.7	78.7	104.03	137.87	169.15	207.62	239.4	284.7	10.84

Table1: Optimized values of All Z and L Parameters.

Plots :

Plots for 10 functions as given below:-