Data Analytics - Assignment-II

Anirban Biswas (Sr.No. - 14382)

September 3, 2018

Run Prediction using DLS Method

The task of this assignment is to find the best fit **run production functions** in terms of wickets-in-hand w and overs-to-go u. It is instructed to use the first innings data alone in the above data set. The model assumed is as follows:

$$Z(u,w) = Z_0(w)\left(1 - exp\frac{-Lu}{Z_0(w)}\right)$$

Data Pre-processing

The given data file contains data on ODI matches from 1999 to 2011. The data has a total of 38 columns and 126768 instances. Out of the 38 columns, 6 columns are of interest to this problem. The *Innings* data tells us whether it is first or second inning data. The count of first inning data turns out to be 67794.

The value of u is calculated by subtracting the column *Over* from *Total.Overs*. The *Wickets.in.Hand* columns directly gives us the value of w. The actual run scored in this situation i.e. when u overs and w wickets left is obtained from subtracting *Total.Runs* from *Innings.Total.Runs*. It is clear that $w \in \{i : 1 \leq i \leq 10, , i \in \mathbb{N}\}$ and $u \in \{j : 1 \leq j \leq 50, j \in \mathbb{N}\}$

Solution Approach

First Approach

To generate the run production functions, a 50x10 matrix is created. The $(i,j)^{th}$ entry of this matrix is obtained from taking the mean value of all the run data when u over and w wickets are remaining.

The loss function here squared loss. If the predicted run when u overs are remaining and w wickets in hand is P_{uw} and original mean value is A_{uw}, the loss function is as follows:

$$L_{Z,L} = \sum_{u=1}^{50}\sum_{w=1}^{10}(P_{uw} - A_{uw})^2$$

Second Approach

In this approach, instead of taking the mean of data, we are using each data points. Hence the regression process will take more time and at the same time it should be more accurate because here we are taking individual data points. Total number of data points in this case is 67794.

The loss function is same as in previous approach. Let A_i be the actual run obtained from data when u_i overs remaining with w_i wickets in hand. Consider P_i is the predicted value for i^{th} data point. The loss function can be written as:

$$L_i = \sum_{i=1}^{67794} (P_i - A_i)^2$$

I have used *scipy* package's *minimize* function for the above mentioned function optimization and *L-BGFS-B* algorithm as function parameter for minimize function. The parameters of the function is Z and L, where Z is a 10-dimensional vector and L is a scalar. Z can be written as

$$Z = [Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8, Z_9, Z_{10}]$$

Corresponding to the 10 values of Z_i, we will have 10 different run generation functions.

Results

First Approach

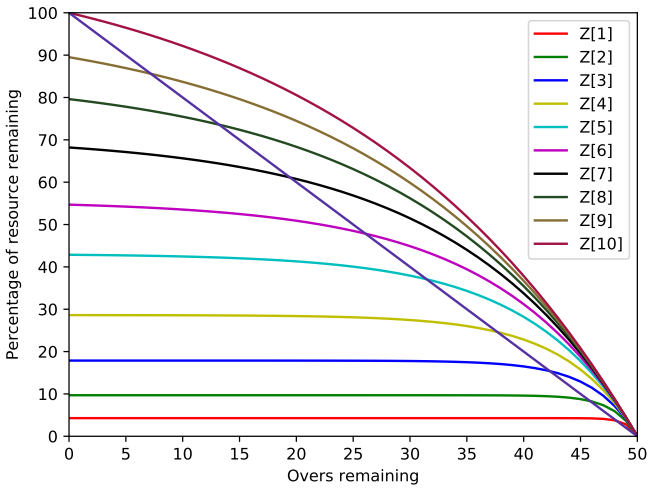
The value of parameter L is : **10.959**

The values of parameters Z_i are:

Z_i	Value
Z_1	10.25
Z_2	23.18
Z_3	42.78
Z_4	68.48
Z_5	103.07
Z_6	133.04
Z_7	169.96
Z_8	204.58
Z_9	238.03
Z_{10}	278.05

Total loss is : **29722.504**

The plot of ten functions is shown below:



Second Approach

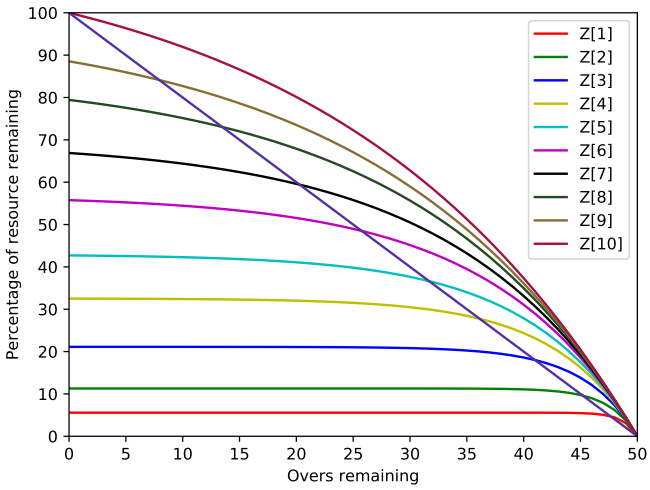
The value of parameter L is : **10.8903**

The values of parameters Z_i are:

Z_i	Value
Z_1	13.50
Z_2	27.35
Z_3	51.16
Z_4	78.83
Z_5	104.04
Z_6	137.74
Z_7	168.78
Z_8	207.48
Z_9	239.03
Z_{10}	284.08

Total loss is : **104818187.658**

The plot of ten functions is shown below:


Percentage of resource remaining
Overs remaining
Z[1]
Z[2]
Z[3]
Z[4]
Z[5]
Z[6]
Z[7]
Z[8]
Z[9]
Z[10]