E0-259-O Assignment 1 - Data Analytics

Authored by Deepank Dixit : SR - 20269, M.Tech AI (Online)

May 22, 2022

1 Assignment Problem Statement

Here's the data file you will use: ODI over-by-over data (.csv file): This file contains data on ODI matches from 1999 to 2011. It is taken from this site. There is an R code for finding the 'run production functions' in this site, but you will do something marginally different in the following assignment. Discussion is encouraged. But write your own code. Please comply with the ethics policy. You must sign the submission statement and click the submit button to submit your work.

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) = Z0(w)[1 - exp-Lu/Z0(w)]. Use the sum of squared errors loss function, summed across overs, wickets, and data points.

Note that 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 normalised squared error (total squared error summed across overs, wickets, and data points, and normalised by the number of data points) in your pdf file.

Feel free to use tools for nonlinear regression available in Python. Some date fields are in different format with an extra comma. Write a short script to clean this up. This clean-up code should a part of the main program. You may create a temporary data file, but remove the temporary data file after the output data has been generated.

2 Cleaning the Data

The csv file 04_cricket_1999to2011.csv has 126769 rows and 38 filled columns. I created a data_cleanup(df) function, and began by removing all rows with Innings value != 1. Then, all rows that have value for Error.In.Data = 1 are being dropped. It was ensured that the Wickets.in.Hand filed no values less than 0 and greater than 11, and the index was reset after each of these operations. It was noticed that the important columns for this task are 'Match', 'Innings', 'Innings.Total.Runs', 'Total.Runs', 'Runs.Remaining', 'Over', 'Total.Overs', 'Wickets.in.Hand', and 'Runs', therefore the df was updated to have just these columns. Thereafter, few more columns were added and cleaned up -

Overs_left = Total.Overs - Over
Runs_to_score = Innings.Total.Runs - Total.Runs

Finally, we keep only these features in our dataframe -
'Match', 'Innings', 'Wickets.in.Hand', 'Runs_to_score', and 'Overs_left' majorly need 4 features to compute the run predictor function

3 Squared Error Loss Function

We defined a loss function and used it as input to the minimize() function along with the initialized parameters and features innings, Remaining_overs, Runs_Remaining_cleaned, Wickets_in_Hand. Function name in the Python code is loss_sqr_error(param, args).

4 Optimizer used to minimize the Loss func

minimize() function from scipy.optimize package is being used to minimize the loss function. The parameters are initialized as follows -

params_initial = [20, 45.0, 75.0, 100.0, 125.0,150.0, 175.0, 200.0, 225.0, 250.0, 10]

Function name in the Python code is optimizer(method_name, innings, Remaining_overs, Runs_Remaining_cleaned, Wickets_in_Hand):

5 Final Output of the code

Time taken to run = Under 1 min 30 seconds

Total Squared Err Loss Normalized = 1608.5293926629638
L = 11.110891473151492
Z1 = 12.988030825556528
Z2 = 27.642284472586674
Z3 = 51.94362858807882
Z4 = 79.11881473553713
Z5 = 104.88577626639699
Z6 = 137.7534747552696
Z7 = 168.34328825803766
Z8 = 205.9051323761605
Z9 = 237.22367683301383
Z10 = 280.6031715275687

It is evident from the values above that - Z10 > Z9 > Z8 > Z7 > Z6 > Z5 > Z4 > Z3 > Z2 > Z1

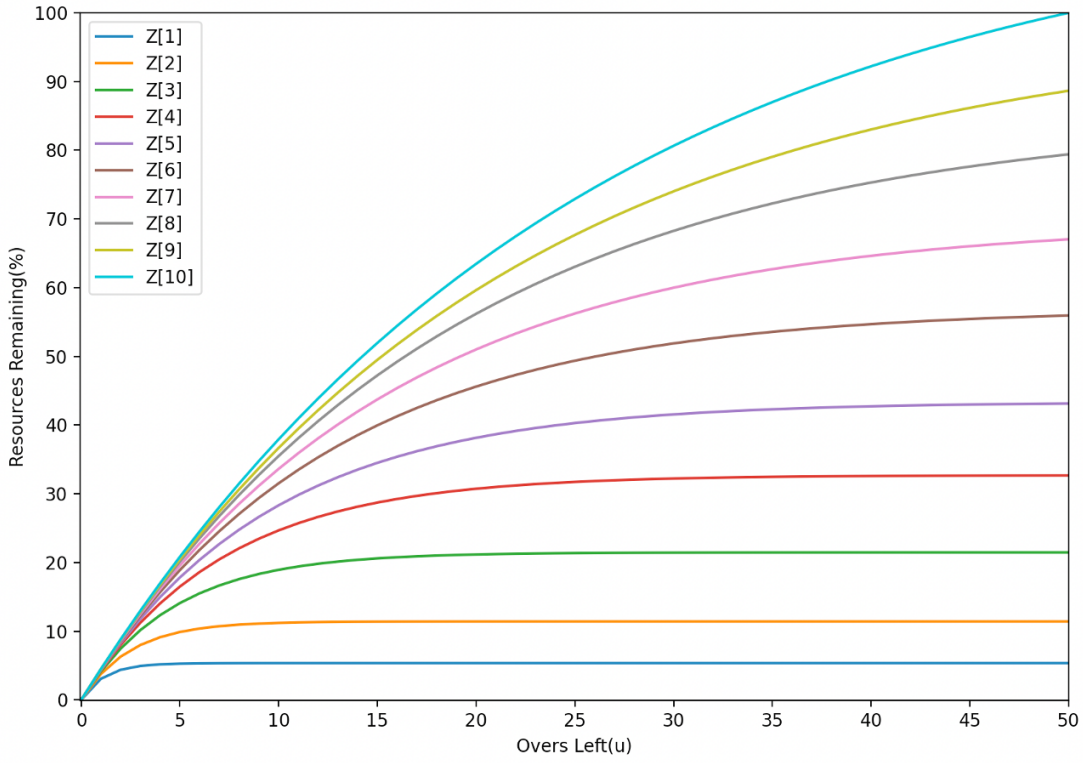

Figure 1: Plot Generated