Dengue Exercise

Dengue fever is a virus-caused disease that is spread by mosquitoes that affects millions of people in tropical environments around the globe. In this problem, you are asked to construct a simple version of the digital disease detection tool: "Google Dengue Trends" for Mexico. First, download the CSV file associated with the exercise, e.g. `MX_Dengue_Trends.csv`. The first column in the spreadsheet represents the date (in months, from 2004-2011). The second column represents the ground truth, the number of cases of Dengue in Mexico, as reported by the Mexican Ministry of Health. The remaining columns represent the number of Google searches of dengue-related terms in Mexico, in a given month. You may use any coding language (e.g. Python or R) for this problem.

(a) Plot the number of cases of Dengue as a function of time.

(b) For the training period 2004-2006 (36 months), find the best line that explains the number of cases of Dengue as a function of the number of searches of the term "dengue". You should do this by solving the least squares problem, and you should obtain the value of the y-intercept and the slope.

(c) Use the equation of the line you obtained in (b) and plot the number of cases as a function of the number of searches of the term "dengue", predicted by your method during the training period. Compare your results to the plot in (a) for such time period.

(d) For the prediction or validation period 2007-2011, use the equation of the line you obtained in (b) to predict the number of the dengue cases as a function of the number of searches of the term "dengue" from 2007-2011. Plot your predictions and compare them to the actual number of cases.

(e) Discuss your results. Could you improve this modeling approach? If so, how?