Monte Carlo simulation of parameter confidence intervals for non-linear regression analysis of biological data using Microsoft Excel

1.ABSTRACT

This report describes a method to obtain parameter confidence intervals from the fitting of nonlinear functions to experimental data, using the SOLVER and Analysis ToolPaK Add-In of the Microsoft Excel spreadsheet. Using a simple Monte-Carlo procedure within the Excel spreadsheet, SOLVER can provide parameter estimates for multiple 'virtual' data sets, from which the required confidence intervals and correlation coefficients can be obtained. The general utility of the method is exemplified by applying it to the analysis of the growth of Listeria monocytogenes, the growth inhibition of Pseudomonas aeruginosa by chlorhexidine .

2.INTRODUCTION

Our aim is to fit the Gompertz curve for finding the growth rate and understanding the effect of inhibitor.

.

The SOLVER Add-In package of Excel allows the user to conduct investigations of non-linear functions using the minimization of the sum of squares of the errors between the observed and modeled values. Hence the error analysis of the modeled and observed data was terminated at the calculation of the standard error of the fit. Confidence intervals can be calculated from knowledge of the Hessian but can also be estimated using Monte Carlo simulation. The Monte Carlo technique uses the standard error of the fit of the non-linear model to the observed data to produce sets of 'virtual' data. These data are modeled using the same non-linear model and a new group of parameters obtained for each virtual set. From the statistical distribution of these parameters, confidence intervals, as well as correlation coefficients, can be obtained.

3.COMPUTATIONAL METHODS AND THEORY

In linear regression this is solved analytically, but if using non-linear regression this is carried out numerically, based on the input of initial parameter estimates. The square
root of the mean of the square of the error (RMSE) is the standard error of the fit.

The expected value of y_i

$$(E(y_i) = y_i.$$

With linear regression, if all the prerequisite conditions are met, then the reported 95% confidence intervals will contain the true value of the regression parameters 95% of the time. With non-linear regression confidence intervals are found using linear approximations and the labelled 95% confidence intervals may not contain the true interval as often.
RMSE is an unbiased estimator of the standard deviation of the fit. A virtual data set can be calculated by adding random error to the expected value of y;

$$Y'_i = \hat{y}_i + N(0, RMSE)$$

This virtual data set can be analysed by NLR to give another set of parameters (the best fit estimates for this virtual data set).

$$SSE = n\ i=1\ (Y\ i - yfit)\ 2$$

Thus the NLR is repeated for the m sets from where the random values are generated and the below formula is applied.

$$SSEtotal = m\ j=1\ n\ i=1\ (Y\ i - yfit)\ 2$$

From the m-sets of parameters obtained, frequency analyses of the parameter values are performed and the 95% confidence intervals obtained from the normal quantiles; covariance between parameter pairs can be found by calculating the parameters' correlation coefficient.

This research paper consists of two phases:

1. Fitting the modified Gompertz equation to microbial growth data
2. Effect of growth rate in the presence of Inhibitory.

Fitting the modified Gompertz equation to microbial growth data:

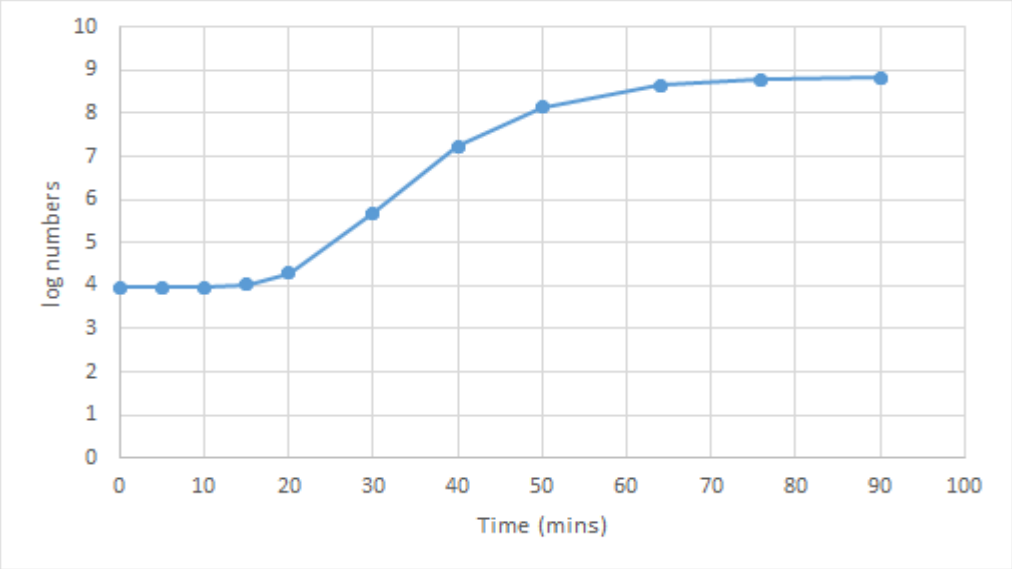
The standard empirical model for fitting of microbial growth data

$$\mathbf{\log N(t) = A + C \exp\{-\exp(B(M - t))\}}$$

N(t) : Growth Rate
A: Asymptotic Number

B: a measure of the slope
M: Time of maximum slope
A + C :maximum population density

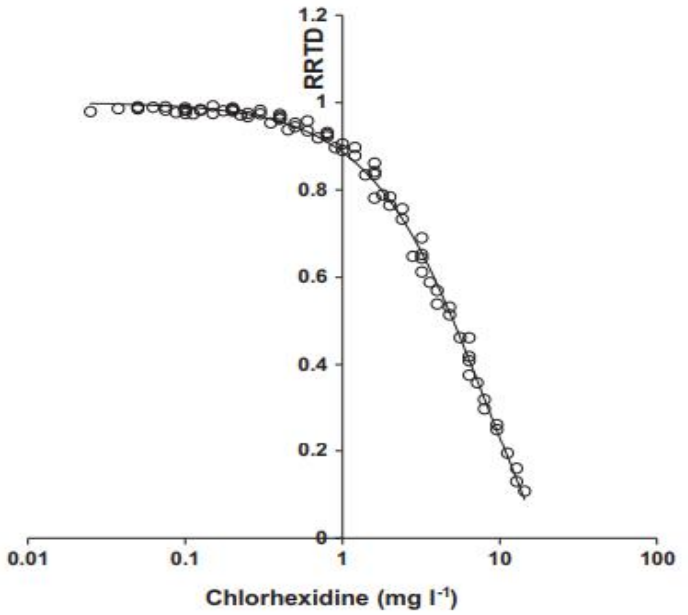
Growth Rate = CB / exp(1)
Lag Rate = M − (1/ B)



The above graph depicts that, as time tends to infinity the logarithm of the growth rate remains constant i.e. after a particular time the growth of the bacteria (Listeria monocytogenes) starts increasing constantly at a uniform rate.

Procedure:

- To find the parameters for a given set of observations.
- As for one particular set of observations, the parameters can be biased.
- Hence to make this unbiased virtual dataset is generated.
- By calculating SSE for each set and minimizing the total SSE value gives us optimized parameters.
- From the parameters obtained Standard Deviation, Upper Class Limit, Lower Class Limit, Growth Rate, and Lag Rate.



This graph above illustrates the observational data (given) in the presence of inhibitor.

RRTD: Relative Rate to Detection.

Here Relative rate means that we are considering the max as 1 and representing others with respect to the above max.

We can't be able to derive an equation for the above graph. We are able to get the value intervals based on given inhibitor concentration.

$$\text{RRTD} = \begin{cases} \text{if} & \\ & [x] = 0, 1 \\ \text{else if} & \\ & [x] < [P_1] \\ \text{then} & \\ & \exp\left(-\left(\frac{[x]}{P_1}\right)^{P_2}\right) \\ \text{else if} & \\ & \frac{1}{e}(1 - P_2(\ln[x] - \ln P_1)) < 0, 0 \\ \text{else} & \\ & \frac{1}{e}(1 - P_2(\ln[x] - \ln P_1)) \end{cases}$$

Here in the above-mentioned equation the parameters p_1 and p_2 are the parameters. p_1 is the concentration of inhibitor giving a relative inhibition of 1/e, where e is the exponential of 1, and p_2 is a slope parameter.

Assumption:

From the above graph, we have observational data. We will consider some initial values for p_1 and p_2 respectively.

Finding SSE and RMSE for these two data ranges and minimizing the total SSE in order to obtain the optimum parameters (p_1 and p_2).

The p_1 and p_2 values can be unbiased by following the same procedure which we had used in finding the growth rate. The obtained optimized values for P_1 and P_2 are 7.11 and 1.1 respectively from the given methods from the research paper.

From the research paper, the MIC and NIC values are directly given by the formula as:

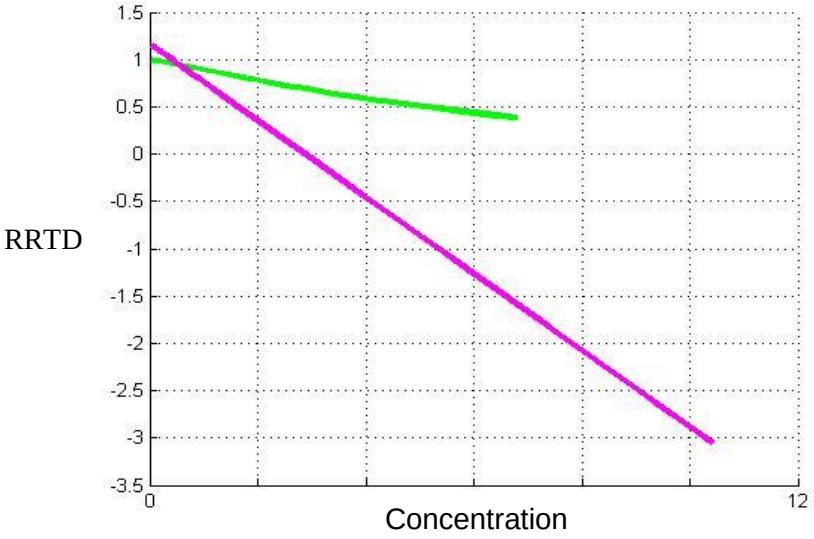
$$\text{(minimum inhibitory concentration) } MIC = P1 \exp(1 / P2) = 17$$

$$\text{(non-inhibitory concentration) } NIC = P1 \exp(1 - e / P2) = 1.4$$

Graphical Method

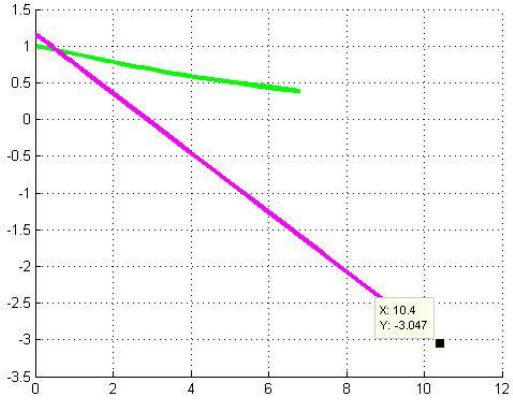
By following the method of reverse engineering, points are collected from the above graph.

The values of p1 and p2 are assigned to get the RRTD values for different values of x

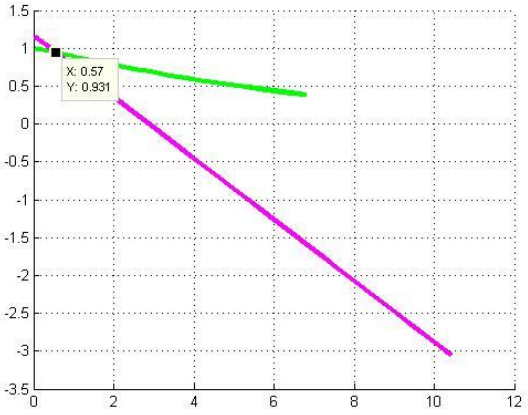


By plotting the values, two different curves are observed (green and magenta).

The green color curve is considered for the values, when concentration varies from 0 to P_1 (P_1 = 7.11). And when the concentration values are greater than P_1, magenta color line is considered. From both the curves, the Minimum Inhibitory Concentration (MIC) is 10.5 at the bottom of the magenta line (which is the minima of the curves). Non Inhibitory Concentration is at the coincidence of both the curves and the value is 0.9 .



Minimum Inhibitory Concentration



Non Inhibitory Concentration

The values that are obtained by the formula given in the paper are 17 and 1.4 for MIC and NIC respectively.

However there is a small mismatch between the values obtained from the graphical method and values obtained from the formulas.

Appendix :

Following is the MATLAB code written for plotting the graphs

```
clear();

dataset= xlsread('MonteCarloNonLinear','Sheet4','O3:Q83');

hold on

x1 = dataset(:,1);
disp(x1);

p = 7.11;
q = 1.1;

x2 = x1(x1<p);

if (x1 == 0)
     y1 = 1;
     plot(x1,y1,'-r' ,'LineWidth',3);
     hold on
end

if x2 == x1(x1<p)
   y2 = exp(-((x2/p).^(q)));
   plot(x2,y2,'-g' ,'LineWidth',3);
   hold on
end

if (1/exp(1))*(1 - q*((log(exp(x1))-log(p))/(log(exp(1))))<0
    y1=0;
    plot(x1,y1,'-o' ,'LineWidth',3);
    hold on
else
    y1 = (1/exp(1))*(1 - q*((log(exp(x1))-log(p))/(log(exp(1)))));
    plot(x1,y1,'-m' ,'LineWidth',3);
end

        grid on;
```

The results obtained from the excel sheet are pasted below :

	Mod							
P1	7.1213	8.042023	7.251323	7.723529	7.22911	7.006971	7.766054	7.11
P2	1.14796	1.2349	1.147578	1.148433	1.326278	1.128152	1.085287	1.1
	SSE	0.024641	0.024641	0.024641	0.024641	0.024641	0.024641	0.024641
	Total SSE	0.172487						
MIC	17.01679	18.07379	17.33252	18.44923	15.3652	17.00166	19.51489	17.64748
NIC	1.594048	2.000199	1.622344	1.72992	1.978916	1.527772	1.594421	1.490971

Understanding the derivation of MIC

$$\frac{d}{dx}\left(\exp\left(-\left(\frac{x}{P_1}\right)^{P_2}\right)\right) = 0$$

$$-\exp\left(-\left(\frac{x}{P_1}\right)^{P_2}\right) \times \left(1 - P_2\left(\frac{x}{P_1}\right)^{P_2-1}\right) = 0$$

$$\frac{1}{P_2} = \left(\frac{x}{P_1}\right)^{P_2-1}$$

$$\log\left(\frac{1}{P_2}\right) = \frac{(P_2-1)}{P_1}\log\left(\frac{x}{P_1}\right)$$

$$\log(x) \approx P_1 \log\left(\frac{1}{P_2}\right)$$

($\because P_2 - 1$ is negligible value).