PAPER DETAILS

TITLE: Regression Tool in MS Excel® Spreadsheets for Biological Data: R-BioXL

AUTHORS: Hasan Basri Öksüz, Sencer Buzrul

PAGES: 224-235

ORIGINAL PDF URL: https://dergipark.org.tr/tr/download/article-file/4451429

SİDAS MEDYA
Akademik Gıda®
ISSN Online: 2148-015X
https://dergipark.org.tr/tr/pub/akademik-gida

Akademik Gıda 22(3) (2024) 224-235, DOI: 10.24323/akademik-gida.1603881

Research Paper / Araştırma Makalesi

Regression Tool in MS Excel® Spreadsheets for Biological Data: R-BioXL

Hasan Basri Öksüz[1], Sencer Buzrul[2]

[1]Konya Technical University, Vocational School of Technical Sciences, Department of Electronics and Automation, Selçuklu, Konya, Türkiye
[2]Necmettin Erbakan University, Department of Food Engineering, Meram, Konya, Türkiye

Received (Geliş Tarihi): 31.05.2024, Accepted (Kabul Tarihi): 06.12.2024
Corresponding author (Yazışmalardan Sorumlu Yazar): sencer.buzrul@erbakan.edu.tr (S. Buzrul)
+90 332 325 2024 +90 332 325 2024

ABSTRACT

A user-friendly MS Excel® spreadsheet as a freeware (R-BioXL) was developed to fit mathematical models to experimental data. (R-BioXL is available to everyone at https://drive.google.com/drive/folders/1GyjT3Z_CJQZu6ASb4LQBlS-ajLa_nF6X?usp=sharing) Initially, users are expected to enter their *X-Y* data and define their parameters of the model. Then, a model equation should also be entered again by users. Users can visualize data (scatter plot) and model fit (line plot) with the defined initial estimates of parameters on the same graph by default. Squared differences between experimental data and model estimates are calculated automatically. Users can change the initial estimates of the parameters to make the model closer to the data instantly, and Solver Add-In of Excel® should be used to minimize the sum of squared error by changing the parameter values. After the parameters are obtained, standard errors (by using "SolverAid" macro), 95 and 99% confidence intervals of the parameters, p values to determine the statistical significance of the parameters, and goodness-of-fit indices are calculated as the last step. All results can be saved on a different Excel® working page. Whole procedure takes a couple of minutes (~3 to 10 min) depending on the Excel® experience of the user. The utility, accuracy and reliability of the spreadsheet was shown by applying two-parameter (non-linear) Michealis-Menten equation for enzyme kinetics, three-parameter (linear) van Deemter equation for chromatography, and four-parameter (non-linear) modified Gompertz equation for microbial growth. In conclusion, R-BioXL can be safely and freely used to describe the experimental data with Excel® knowledge, without any skills in programming and without additional cost for other software package.

Keywords: Curve fitting, Freeware, Mathematical modeling, Solver, SolverAid

Biyolojik Veriler İçin MS Excel® Hesap Çizelgesi Aracı: R-BioXL

ÖZ

Deneysel verilere matematik modelleri uydurmak için ücretsiz yazılım olarak kullanıcı dostu bir MS Excel® hesap çizelgesi aracı (R-BioXL) geliştirilmiştir. (R-BioXL https://drive.google.com/drive/folders/1GyjT3Z_CJQZu6ASb4LQBlS-ajLa_nF6X?usp=sharing bağlantısından herkese açıktır) Başlangıçta, kullanıcıların *X-Y* verilerini bu araca girmeleri ve model parametrelerini tanımlamaları beklenmektedir. Daha sonra model denkleminin de yine kullanıcılar tarafından girilmesi gerekmektedir. Kullanıcılar verileri (dağılım grafiği) ve model uyumunu (çizgi grafiği) girilen ilk parametre değerleri ile aynı grafik üzerinde varsayılan olarak gözlemleyebilmektedirler. Deneysel veriler ve model tahminleri arasındaki farkların karesi otomatik olarak hesaplanmaktadır. Kullanıcılar modeli verilere anında daha yakın hale getirmek için parametrelerin ilk değerlerini değiştirebilir. Excel®'in Çözücü eklentisi parametre değerlerini değiştirerek hataların karesinin toplamını en aza indirmek için kullanılmalıdır. Parametreler elde edildikten sonra son adım olarak standart hatalar ("SolverAid"

makrosu kullanılarak), parametrelerin %95 ve %99 güven aralıkları, parametrelerin istatistiksel anlamlılığını belirlemek için p değerleri ve uyum iyiliği indeksleri hesaplanır. Tüm sonuçlar farklı bir Excel® çalışma sayfasına kaydedilebilir. Tüm bu prosedür, kullanıcının Excel® deneyimine bağlı olarak birkaç dakika (~3 ila 10 dakika) sürebilir. Aracın kullanımı, doğruluğu ve güvenilirliği enzim kinetiği için iki parametreli (doğrusal olmayan) Michealis-Menten denklemi, kromatografi için üç parametreli (doğrusal) van Deemter denklemi ve mikrobiyal büyüme için dört parametreli (doğrusal olmayan) modifiye Gompertz denklemi uygulanarak gösterilmiştir. Sonuç olarak, R-BioXL, Excel® bilgisi ile deneysel verileri tanımlamak için, herhangi bir programlama becerisi gerektirmeden ve diğer yazılım paketleri için ek maliyet olmadan güvenle ve serbestçe kullanılabilir.

Anahtar Kelimeler: Eğri uydurma, Ücretsiz yazılım, Matematik modelleme, Çözücü, SolverAid

INTRODUCTION

Mathematical models are used to describe the experimental data in many fields and biological sciences (agriculture, biology, biochemistry, biotechnology, environment, food science etc.) are no exception. To estimate the parameters of a model, one should fit linear or non-linear functions to data and many software packages are available for this purpose. However, spreadsheet techniques are less effortless because in general, no programming skills are required [1]. That is why, not only undergraduate and graduate students but also scientists and researchers from various disciplines prefer highly available and user-friendly Microsoft Excel® for such data analysis [2].

If a model is linear in its parameters that is, if the derivative of the model equation with respect to a parameter does not contain that parameter, then linear regression is applied and it is possible to obtain the parameters together with the uncertainties (standard errors or confidence intervals) of the parameters by using Excel®. This can be done by using "Data Analysis" tool (under the "Data" menu of Excel®) and "Regression" application [3]. Excel® gives a detailed "Summary Output" where parameters, standard error of the parameters, 95% confidence intervals (by default) and/or any confidence intervals such as 99% (User should define it.) are listed. Moreover, coefficient of determination (R^2), adjusted coefficient of determination (R^2_{adj}) and standard error of the estimate, which is also known as root mean square error (RMSE) are calculated to judge the goodness-of-fit of the linear model. However, most of the models used in biological sciences are non-linear in parameters [1] and hence, non-linear regression is required. Excel® "Solver" routine can be used for this purpose [4–6] but, only parameters (not the uncertainties in those parameters) would be obtained which is unacceptable [7]. Parameters are uninterpretable without their uncertainties and uncertainties are as important as the parameters themselves [8] and therefore, they should be obtained as well in case of a non-linear regression.

It should be noted that depending on the non-linearity of the equation, uncertainties are not symmetric [8]. On the other hand, many software packages calculate the asymptotic standard errors or confidence intervals. Asymptotic standard errors can also be calculated by using Excel®: first, a matrix so-called "Jacobian matrix" (***J***) should be constructed and this can be done by taking the partial derivative of the model equation with respect to each parameter. Then, by using the parameter values obtained from the non-linear regression, each value of the partial derivatives for every *X* (independent variable) are calculated. Taking the transpose of the Jacobian matrix ($\boldsymbol{J}^T$) and taking the inverse of the multiplication of transpose of Jacobian matrix by Jacobian matrix itself $[(\boldsymbol{J}^T \cdot \boldsymbol{J})^{-1}]$ will give $p \times p$ variance-covariance matrix, where p is the number of parameters in the model. By using the diagonal element of this matrix, asymptotic standard errors of the parameters and confidence intervals could be calculated. This calculation may take several minutes even for the experienced Excel® users. Moreover, if the number of parameters is high in a model (≥ 4), calculations become cumbersome. Alternatively, standard errors of the parameters could also be calculated in Excel® by using macros such as the "SolverAid" provided by De Levie [9], but unfortunately it seems that it is not widely be used [7].

The objective of this study is to present a user-friendly Excel spreadsheet where the user can enter his/her model by himself/herself and minimize the sum of squared error (SSE) by changing the parameter values of the non-linear model in Excel® with the Solver function. Then, the asymptotic standard errors, 95% and 99% confidence intervals as well as the goodness-of-fit indices (R^2, R^2_{adj} and RMSE) can be calculated automatically (with a little effort by the user). Usefulness of the Excel® spreadsheet is described by using three different models and the results are compared with other statistical software packages such as SigmaPlot 12.0, MATLAB R2017b and SPSS 22.0.

METHODOLOGY

R-BioXL

R-BioXL (Regression tool for biological data in Excel®) is a user-friendly spreadsheet application to describe the biological data in various disciplines (biology, biotechnology, biomedicine, environmental sciences, food sciences etc.). R-BioXL is a freeware and can be found at https://drive.google.com/drive/folders/1GyjT3Z_CJQZu6ASb4LQBlS-ajLa_nF6X?usp=sharing. It contains 5 working pages where users can enter 5 different models to each. Alternatively, users can use the same page by deleting the previous models or the tool can be downloaded again to use it for the next 5 models. Furthermore, users need not know any programming

skills, being familiar to some basic functions of Excel® is enough to use the tool – see below.

Figure 1 shows the blank page of R-BioXL. Buttons to be used to calculate the squared differences between the experimental data and model fit, sum of squared differences (error), and regression statistics, and to save the results are all inactive. User should enter the *X*-*Y* data to Columns A and B (starting from A2-B2), respectively where *X* is the independent variable (generally time) and *Y* is the dependent variable. Definitions of all necessary cells are also given for the users. It is possible to see the data on the graph as a scatter plot as soon as the data are entered. Then, user should enter the name(s) of the parameter(s) to Column G (starting from cell G2) and in the next cell (Column H – starting from cell H2) initial value(s) of the parameter estimate(s). Initial values for all parameters can be entered as "1" as the starting point of the iteration, but these are subjected to change before using the Solver function – see below. Before entering the model to Column C, users have two options: (i) defining the name of the parameters i.e., naming the cells, (ii) using "$" sign to fix the cells of the parameters so that using the same values of the parameters even after dragging or copying. Naming the cells are very easy and explained in different studies [4–6]. Moreover, it allows users to see the parameters on the model equation, not the name of the cells, which is easy to interpret. Nevertheless, second option is still valid: in the model equation, users should select the initial estimate of the parameters and enter "$" before and after the letter of the cell as such "$H$2". Both cases are demonstrated in the next section.

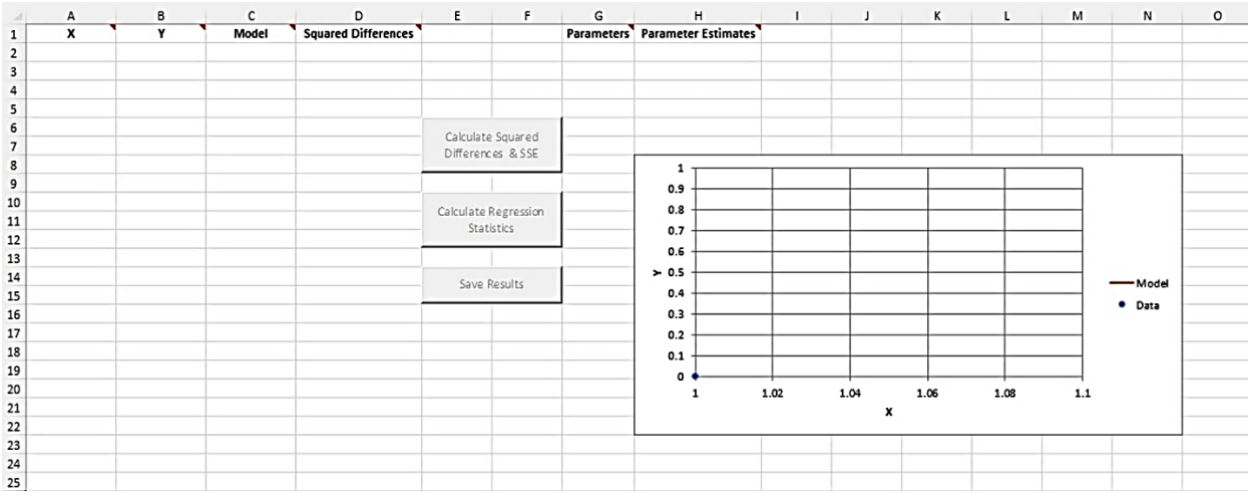


Figure 1. Blank page of R-BioXL. Buttons will be all inactive unless the necessary parts are filled in

RESULTS

Case Study I: Enzyme Kinetics

Data of initial rates of sucrose hydrolysis by the enzyme (yeast) invertase as a function substrate concentration are given in Table 1. Data were originally published by Chase et al. [10] and were also used by van Boekel [11]. Henri-Michealis-Menten equation or more widely known as Michealis-Menten equation [Equation (1)] was used to describe the data:

$$V = \frac{V_{max} \cdot S}{K_M + S} \qquad (1)$$

where *V* is the (initial) rate (dependent variable) and *S* is the sucrose concentration (independent variable). The model has two parameters: V_{max} is the maximum rate, and K_M is the substrate concentration where rate is equal to $0.5 \cdot V_{max}$.

Table 1. Initial rates of hydrolysis of sucrose by the Invertase enzyme. Original data are from Chase et al. [10]

Sucrose (M)	Initial rate (min^{-1})
0.0292	0.182
0.0580	0.258
0.0584	0.265
0.0876	0.311
0.1170	0.330
0.1170	0.342
0.1460	0.349
0.1750	0.372
0.2050	0.347
0.2340	0.371

Figure 2 shows entering the data into columns A and B. As the data are entered, graph is updated simultaneously, and data are seen as blue circles (Figure 2a). At this stage, buttons are still inactive. Then, the parameters (column G) and their corresponding (initial) values (column H) are entered. Model equation should be written to column C. When these steps are completed, model fit appears as the red line on the graph and the first button becomes active (Figure 2b). The equation can be written in Excel® as such: "=V_{max}*A2/(K_m+A2)" where V_{max} and K_m are the defined parameters and A2 is the first cell where S (substrate

concentration) was inserted. Note that, parameters were defined i.e., the cells were named (cells H2 and H3 not G2 and G3!): Formulas > Define name, so that the names of the parameters appear on the model equation (column C) [4–6].

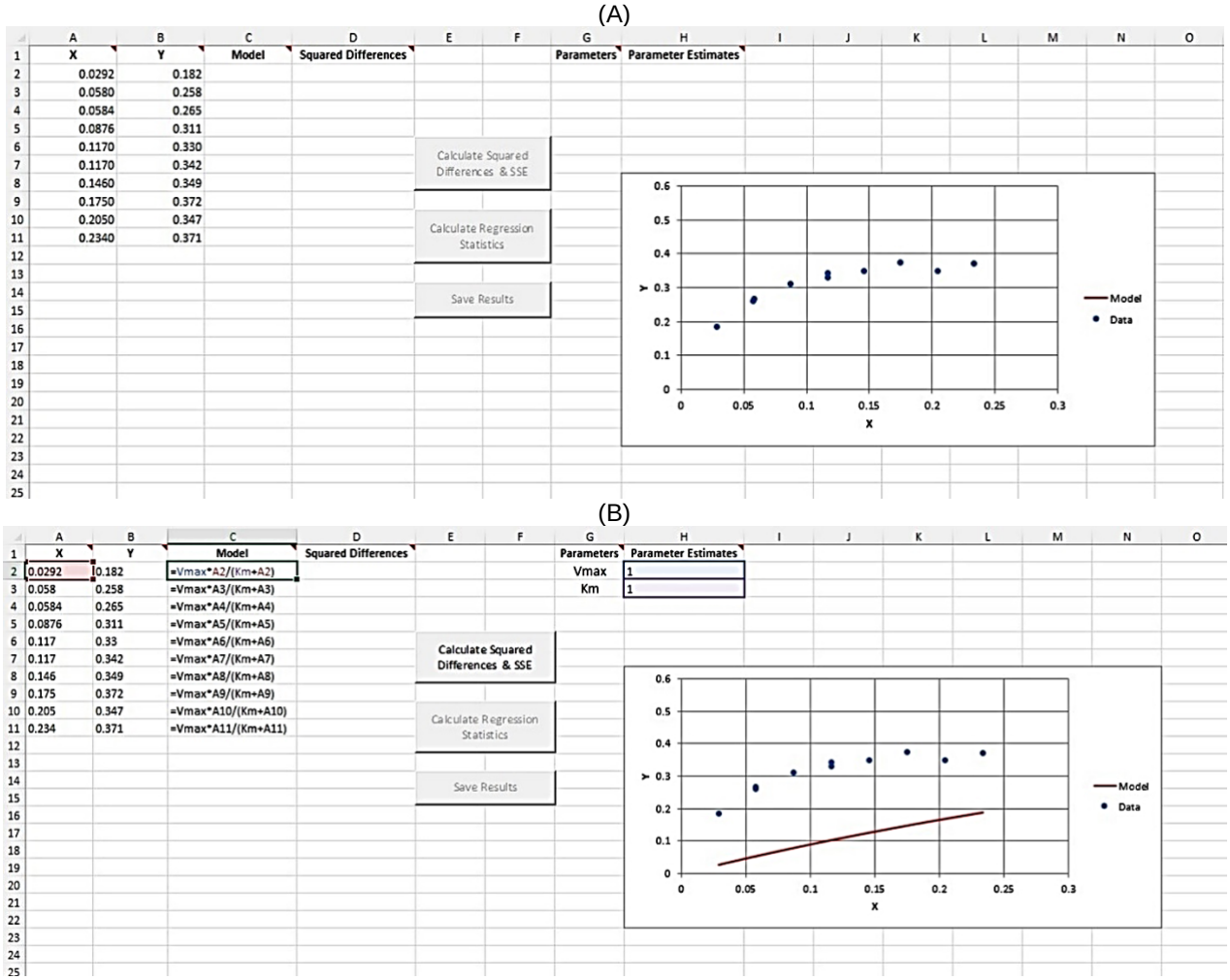


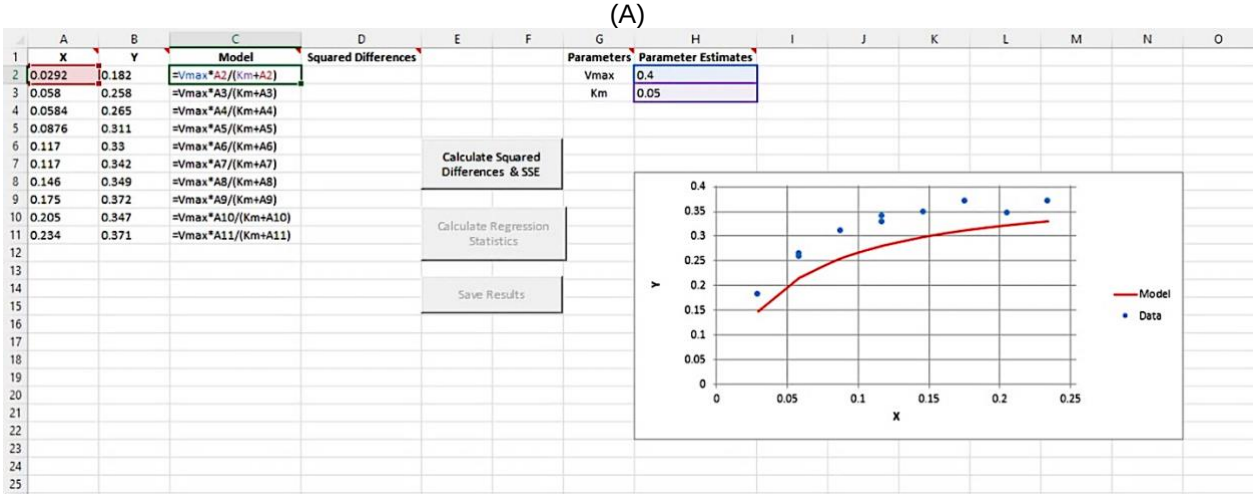
Figure 2. Entering the data given in Table 1 to R-BioXL. Data are observed as blue circles on the graph (A). Entering the model parameters (cells G2 and G3), initial estimates of the model parameters (cells H2 and H3) and the model equation [Equation (1)] (column C, starting from cell C2). Model fit is observed as the red line on the graph (B). Note that cells H2 and H3 are named.

Initial values which were entered as 1, can be changed and the updated model fit according to the new initial values is also observed instantly (Figure 3a). User can select better values and make the model fit closer to the data before using the Solver function. Then, by clicking the "Calculate Squared Differences & SSE" button, all calculations are done automatically and instantly (Figure 3b). At this stage, second button is active; however, using this button before using the Solver function for obtaining the best parameter estimates ends up with the wrong results. Therefore, a warning will appear if the user clicks this button.

Figure 4a shows the best parameter estimates by using Solver function. As soon as the "Calculate Regression Statistics" button is hit, a window asking the number of parameters in the model equation appears (Figure 4b). Since there are two parameters (V_{max} and K_M) in the equation, "2" was entered. Then, standard errors and confidence intervals of the parameters, p values, and goodness-of-fit indices were calculated and tabulated automatically (Figure 4c).

Note that the standard errors and the confidence intervals are asymptotic meaning that V_{max} = 0.4367 ± 0.0122 min^{-1} (standard error), V_{max} = 0.4367 ± 0.0282 min^{-1} (95% confidence interval) or V_{max} = 0.4367 ± 0.0411 min^{-1} (99% confidence interval). Both parameters were statistically significant since p values were ≤ 0.05 or ≤ 0.01. Moreover, high R^2 and R^2_{adj}, and low RMSE values revealed a good fit. These were compared with some other software packages such as SigmaPlot, MATLAB and SPSS and identical results were obtained (results not shown).

(A)



(B)

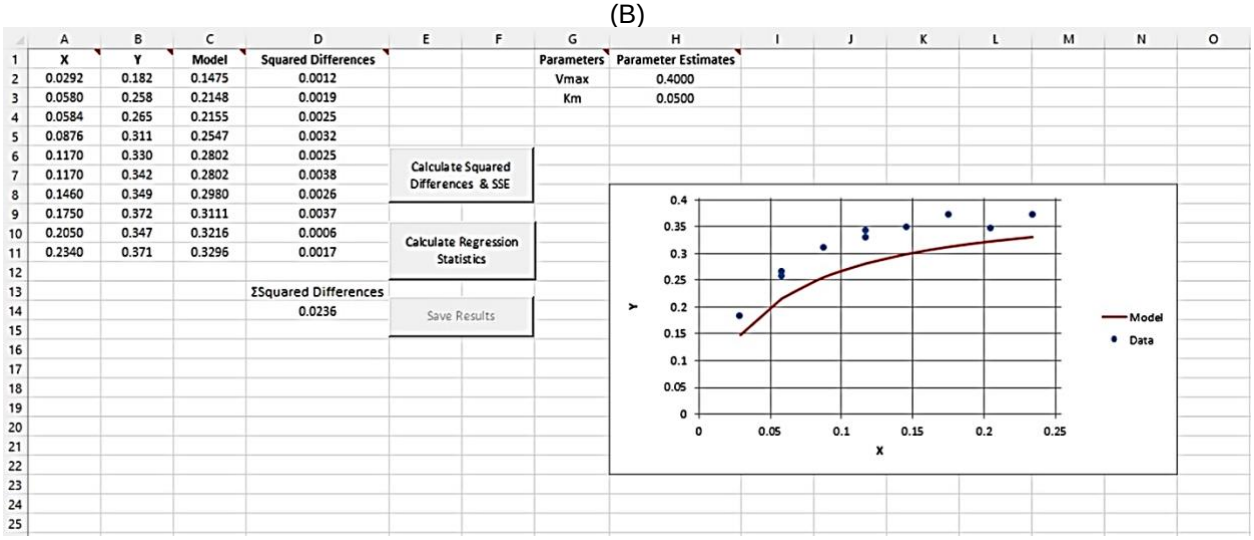


Figure 3. Changing the initial estimates of the parameters and making model fit closer to the experimental data given in Table 1 (A). Calculating the squared differences and sum of squared error (SSE) by clicking the "Calculate Squared Differences & SSE" button (B).

"Save Results" button can be used after all calculations to save and extract the output to another Excel® page. User can now arrange the graph (scaling and naming the axes, changing the color etc.).

Case Study II: Gas Chromatography

Experimental gas chromatography data published by Moody [12] are given in Table 2. Data were described by van Deemter equation [Equation (2)]:

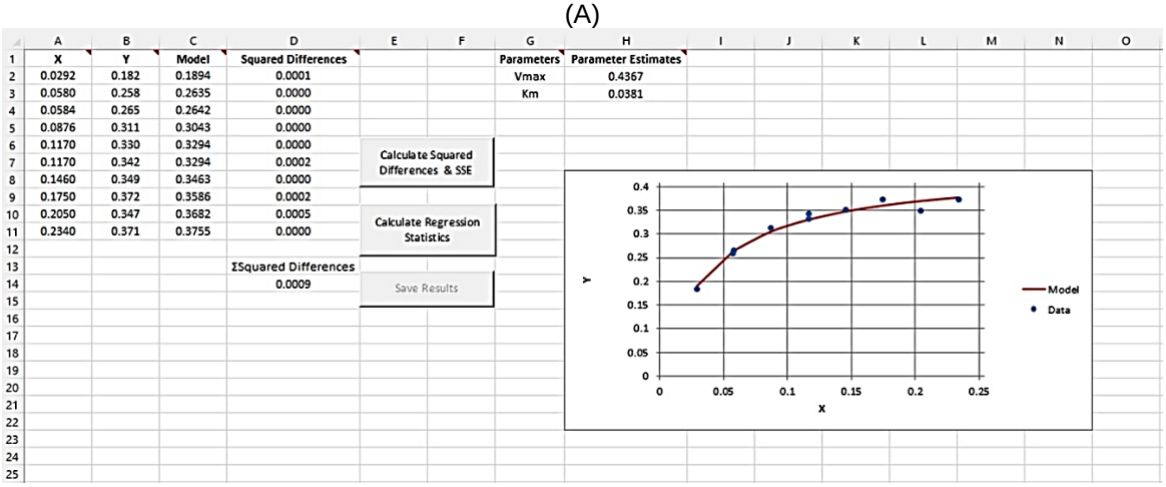
$$H = A \cdot \dot{F} + B/\dot{F} + C \qquad (2)$$

where *H* is the plate height (dependent variable) and $\dot{F}$ is the volumetric flow rate (independent variable). Note that the model has three parameters (*A*, *B* and *C*) and the parameters are linear. Therefore, linear regression can be used to obtain the parameter values; however, this example was used to show that R-BioXL could also be used linear models as well as non-linear models.

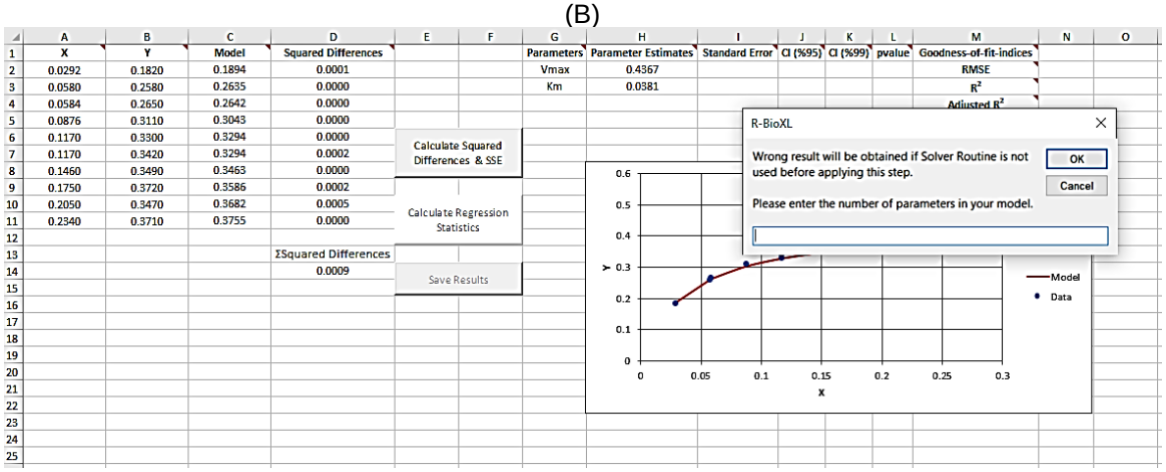
Table 2. Gas chromatography data from Moody [12]

Plate height (mm)	Flow rate (mL/min)
3.4	9.59
7.1	5.29
16.1	3.63
20.0	3.42
23.1	3.46
34.4	3.06
40.0	3.25
44.7	3.31
65.9	3.50
78.9	3.86
96.8	4.24
115.4	4.62
120.0	4.67

(A)



(B)



(C)

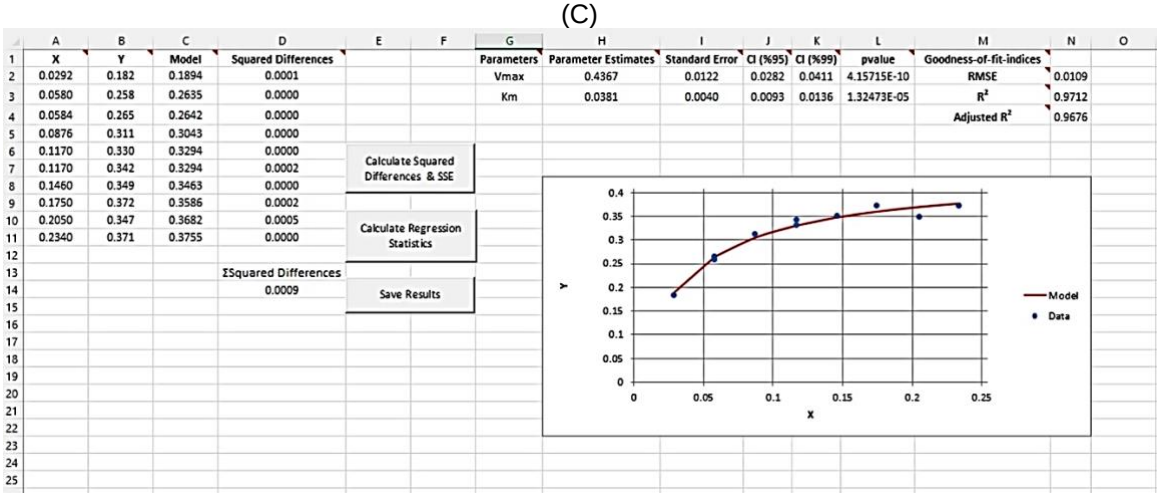


Figure 4. Obtaining the best-fitted parameter values by using Solver (A). Entering the number of parameters in the model after clicking the “Calculate Regression Statistics” button (B). Results of the model fit [Equation (1)] which can be saved and extracted to another Excel® working page by clicking the “Save Results” button (C).

Figure 5 shows entering the parameters, the initial estimates of the parameters as “1” and the model equation. The equation was written in Excel® as such: “=$H$2*A2+$H$3/A2+$H$4” where $H$2, $H$3 and $H$4 are the values of the parameters *A*, *B* and *C*, respectively in Equation (2), and again A2 is the is the first cell where $\dot{F}$ (volumetric flow rate) was inserted (Figure 5a). Note that cells were not defined but “$” ($H$2, $H$3 and $H$4) was used to fix the values of the parameters in the model equation (Figure 5a). If the Solver function is used with these initial estimates, no results will be obtained (convergence failure). Therefore, initial estimates of the parameters should be selected wisely (Figure 5b) and after obtaining the best estimates of the parameters, the necessary steps could be followed (calculating SSE, entering the number of parameters as “3”) to conclude the process (Figure 5c). Another important remark, since the van Deemter

equation [Equation (2)] is linear in parameters there is no need for initial estimates of the parameters. Data analysis > Regression tool of Excel® can be used [3]. The results of the linear regression are presented in Figure 6. The same results were also obtained for applying linear regression by using Data Analysis > Regression application of Excel®. Output gives the confidence intervals as upper and lower limit; however, in R-BioXL they are given as plus/minus the best fit value and therefore, they can be calculated by subtracting the best-fitted parameter value from the upper limit or subtracting lower limit value from the best-fitted parameter value. For example, for the parameter *C* in van Deemter equation [Equation (1)] 95% confidence interval is whether 1.7365 – 1.5681 = 0.1684 or 1.5681 – 1.3997 = 0.1684 (Figure 6) and this was the same result with the R-BioXL (Figure 5c).

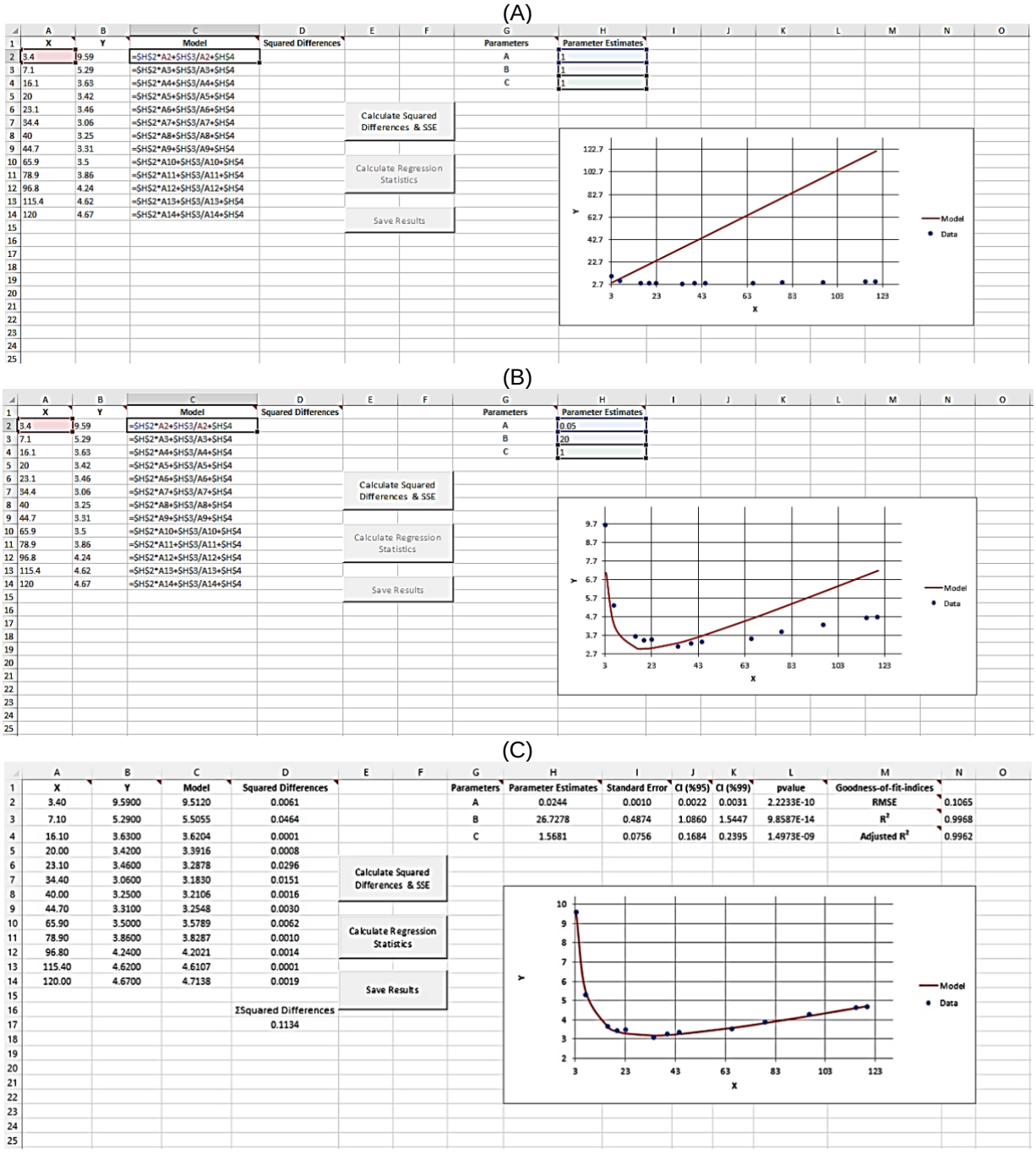


Figure 5. Entering the data given in Table 2, the model parameters (cells G2 to G4), initial estimates of the model parameters (cells H2 to H4) and the model equation [Equation (2)] (column C, starting from cell C2). (A). Changing the initial estimates of the parameters and making model fit closer to the experimental data given in Table 2 (B). Results of the model fit [Equation (2)] which can be saved and extracted to another Excel® working page by clicking the “Save Results” button (C).

	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.9984								
5	R Square	0.9968								
6	Adjusted R Square	0.9962								
7	Standard Error	0.1065								
8	Observations	13								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*Significance F*				
12	Regression	2	35.4764	17.7382	1563.7258	3.2894E-13				
13	Residual	10	0.1134	0.0113						
14	Total	12	35.5898							
15										
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
17	Intercept (C)	1.5681	0.0756	20.7494	1.4974E-09	1.3997	1.7365	1.3286	1.8076	
18	X Variable 1 (A)	0.0244	0.0010	25.1942	2.2232E-10	0.0222	0.0265	0.0213	0.0274	
19	X Variable 2 (B)	26.7279	0.4874	54.8363	9.8585E-14	25.6418	27.8139	25.1831	28.2726	
20										
21										

Figure 6. Summary output of the fit of van Deemter equation [Equation (2)] to the data given in Table 2 by linear regression tool (Data > Data Analysis > Regression) of Excel®.

Case Study III: Microbial Growth

Growth of *Listeria monocytogenes* at 30°C in high salt media was published by Lambert et al. [13] are given in Table 3. Data were described by modified Gompertz equation [Equation (3)] which was proposed by Zwietering et al. [14]:

$$\log_{10}N(t) = \log_{10}N_0 + (\log_{10}N_{max} - \log_{10}N_0) \cdot \exp\left\{-\exp\left[\frac{\mu_{max} \cdot e/2.303}{(\log_{10}N_{max} - \log_{10}N_0)} \cdot (\lambda - t) + 1\right]\right\} \quad (3)$$

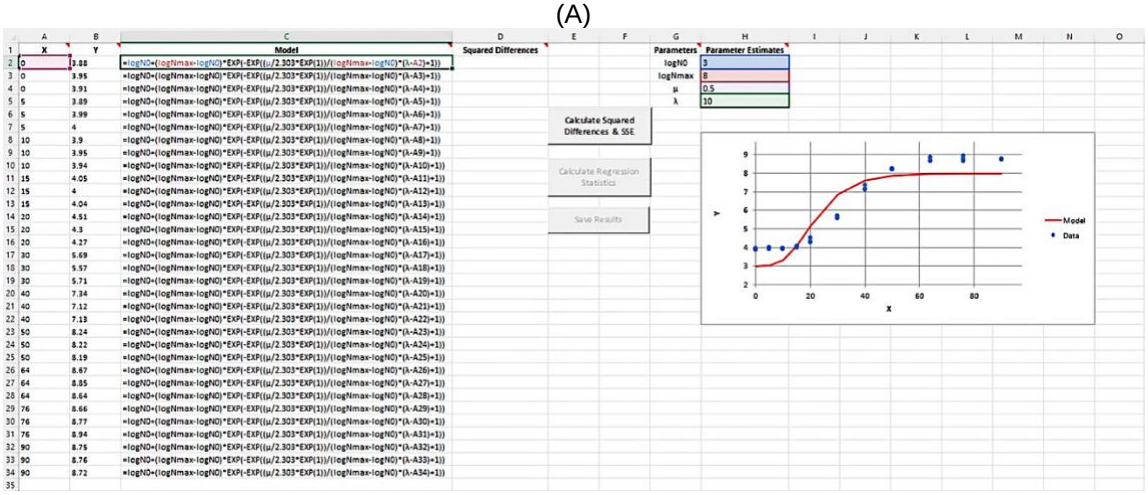
where $\log_{10}N(t)$ is the number of bacteria (dependent variable) and t is the time (independent variable). Model has four parameters: $\log_{10}N_{max}$ is the maximum bacterial load, $\log_{10}N_0$ is the initial number of bacteria, μ_{max} is the maximum growth rate and λ is the lag time.

Figure 7 shows the data, the model equation with the selected initial values of the parameters and the results. Model was written as: "=logN$_0$+(logN$_{max}$-logN$_0$)*EXP(-EXP((μ/2.303*EXP(1))/(logN$_{max}$-logN$_0$)*(λ-A2)+1))". If the user enters "1" as the initial estimates of the parameters, the model [Equation (3)] will be undefined because $\log_{10}N_{max} - \log_{10}N_0$ will be zero. Furthermore, if $\log_{10}N_{max}$ is set to "2" leaving all parameters with "1" will not enough to obtain the best parameter estimates because Solver capacity is limited and number of iterations is not sufficient to have global minimum. This was also the case with van Deemter equation [Equation (2)] as discussed above, but not for Michealis-Menten equation [Equation (1)]. Users should be careful more and more about the initial estimates as the number of parameters in a model is high (≥ 3). Once again, the parameter estimates, standard errors of the parameters, confidence intervals and goodness-of-fit indices were all same with the other software programs (results not shown).

Table 3. Growth data of *L. monocytogenes* at 30 °C in 9% salt. Original data are from Lambert et al. [13]

Time (h)	$\log_{10}N(t)$ (CFU/mL)
0	3.88
0	3.95
0	3.91
5	3.89
5	3.99
5	4.00
10	3.90
10	3.95
10	3.94
15	4.05
15	4.00
15	4.04
20	4.51
20	4.30
20	4.27
30	5.69
30	5.57
30	5.71
40	7.34
40	7.12
40	7.13
50	8.24
50	8.22
50	8.19
64	8.67
64	8.85
64	8.64
76	8.66
76	8.77
76	8.94
90	8.75
90	8.76
90	8.72

(A)



(B)

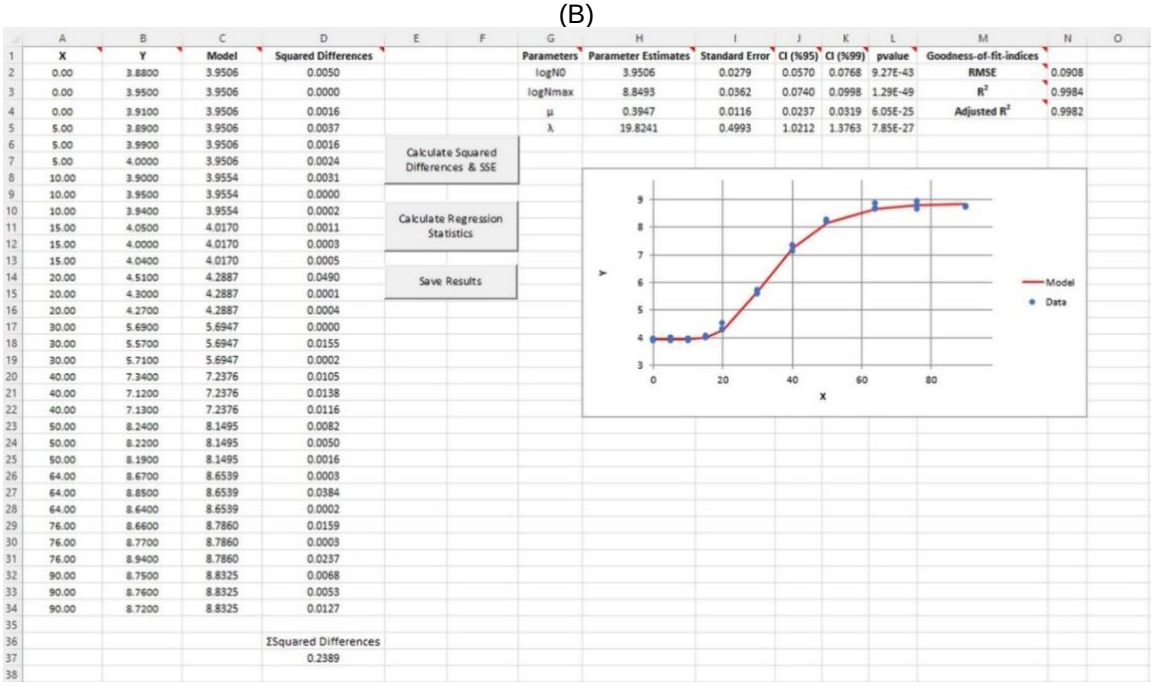


Figure 7. Entering the data given in Table 3, the model parameters (cells G2 to G5), initial estimates of the model parameters (cells H2 to H5) and the model equation [Equation (2)] (column C, starting from cell C2). (A). Results of the model fit [Equation (3)] which can be saved and extracted to another Excel® working page by clicking the "Save Results" button (B).

Workflow and Step-By-Step Guide to R-BioXL

R-BioXL is used to describe the experimental biological data in the form $y = f(x)$, where x is the independent variable and y is the dependent variable. R-BioXL focuses on models with one explanatory variable (X) and one response variable (Y) which is generally the case in biological sciences. Number of bacteria with respect to time, rate constant with respect to temperature and enzyme activity with respect to pH are the notable examples. Nevertheless, there may be cases where the response is non-linearly defined by using two explanatory variables such as lag time with respect to temperature and water activity. Users referred to other software packages which are capable to perform such fittings.

Before to use R-BioXL, experimental data in the form of X-Y and the mathematical equation to describe the data should be known by the user, so that he/she could enter the data and write down the model in Excel®. Moreover, regression assumptions are also valid for R-BioXL and these are listed below:

- Errors are normally distributed.
- All error (scatter) is in Y not in X viz., X is precisely known or there are no experimental errors in X.
- Data is homoscedastic i.e.; error or amount of scatter is the same for all Y values. If data is heteroscedastic, users need to transform the data before using R-BioXL or any other software for regression.
- The model used is the correct one and experimental errors are uncorrelated.

To use R-BioXL, the following steps should be followed:

1. Enter (or simply copy and paste) the raw data in two columns: X should be entered column A (starting from cell A2) and Y should be entered column B (starting from cell B2). After this step, data can be observed on the graph as the scatter plot – see Figs. 1 and 2a.
2. Write down the name of the parameters of the model to column G (starting from cell G2) – see Figure 2b.
3. Write down the initial values of the parameters to the adjacent cells i.e., column H (starting from cell H2). Initial values can be entered as "1" for the starting point – see Figure 2b.
4. Define the cells in column H (starting from cell H2): from Excel menu Formulas > Define name. Excel will define the cell as the name of the parameter in column G (starting from cell G2). Repeat this procedure for all parameters. Readers are referred to the works of Brown [4] and Kemmer and Keller [5] for naming the cells in Excel®.
5. Enter the model equation to column C (starting from cell C2). Since the cells in which the parameter values exist (column H) are defined in step 4, parameter names appeared in the model equation – also see Figure 2b. However, step 4 can be skipped and the model equation can be written without defining the cells. In this case, "$" should be used before and after the letter of the cell such as "$H$2" to fix the cell in the model equation – see Figure 5a.
6. The model will appear as the line plot (red line) – see Figure 2b.
7. Change the initial values of the parameters to make the red line (model equation) as close as possible to the blue data (experimental data) before using the Solver – see Figure 3a.
8. First button i.e., "Calculate the Squared Differences & SSE" is active after the step no. 5 – see Figs 2b and 3a.
9. Click on the button and R-BioXL will calculate the SSE – see Figure 3a.
10. User should use the Solver Routine of Excel® to obtain the best parameter estimates: from Excel® menu Data > Solver (Solver can be installed easily if it does not appear in Excel® menu through File > Options > Add-Ins > Excel® Add-Ins > Go > Solver), target should be selected as the cell of SSE and since this is a minimization problem i.e., our target is to minimize the SSE. Hence, minimum (Min) should be selected and this could be done by changing the parameters (By Changing Variable Cells). Several studies are present for the use of Solver in Excel® for minimizing SSE and obtaining parameter estimates [4–6]. Solver makes iteration to obtain the best parameter estimates and to be sure that the results are the global but not the local minimum, user can repeat this procedure with different initial estimates of the parameters. Nevertheless, it would be better to start the iteration with the closest possible initial values of the parameters to avoid (i) convergence failure, (ii) to obtain a local minimum – see step no. 7.
11. Click on "Calculate Regression Statistics" and R-BioXL will ask the number of parameters of the model entered to column C – see Figure 4b. Enter the number of the parameters and click "OK".
12. Asymptotic standard errors and confidence intervals (95 and 99%), p values, R^2, R^2_{adj} and RMSE values are calculated by R-BioXL – see Figure 4c.
13. Results can be saved into another worksheet by clicking on the "Save Results" button.

Features and Reliability of R-BioXL

R-BioXL is an Excel®-based application and Excel®'s Solver function need to be used before calculating parameters' precisions and goodness-of-fit statistics. Solver uses generalized reduced gradient (GRG) method as the iteration method although different algorithms such as Gauss-Newton, Marquardt-Levenberg and Neder-Mead methods are being used in non-linear regression [4] in different software programs. However, the outcome (parameter estimates and goodness-of-fit indices) would be the same for all methods.

The most popular statistic to compare the goodness-of-fit of different models is R^2; however, R^2 alone is not sufficient to judge the performance of a model's goodness-of-fit [15]. Therefore, R^2_{adj} and RMSE values are also given in R-BioXL. R^2 can be calculated as:

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \qquad (4)$$

where SSE is the sum of squared difference between experimental (observed) data (y_{exp}) and fitted (estimated) value by the model (y_{fitted}):

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_{exp} - y_{fitted})^2 \qquad (5)$$

and SST (sum square total) is the sum of squared difference between experimental (observed) data (y_{exp}) and mean value of experimental data (y_{mean}):

$$\mathrm{SST} = \sum_{i=1}^{n}(y_{exp} - y_{mean})^2 \qquad (6)$$

R^2_{adj} is calculated as:

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}/n-p}{\mathrm{SST}/n-1} = 1 - \frac{\mathrm{MSE}}{\mathrm{MST}} \qquad (7)$$

where n is the number of data points, p is the number of parameters in the model, MSE is the mean square error and MST is the mean square total.

Another way to express R^2_{adj} is:

$$R^2_{\mathrm{adj}} = 1 - (1 - R^2) \times \frac{n-p}{n-1} \qquad (8)$$

Equation (8) reveals that R^2_{adj} is almost always lower than R^2. There are two occasions that R^2_{adj} equals to R^2: (i) when the model used has one parameter (p = 1), (ii) when the model has perfect fit i.e., R^2 = 1 and hence, R^2_{adj} = 1.

Last statistic is RMSE which is also known as standard error of the estimate or the model:

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{SSE}}{n-p}} \quad (9)$$

RMSE is known as the most informative indices for the microbiological [16] and biological data.

R-BioXL also calculates asymptotic standard errors and confidence intervals (95 and 99%) together with p values. Asymptotic standard errors are calculated by using SolverAid macro [9]. Reliability of this macro were also shown in other studies [17, 18]. Confidence intervals are calculated by using the standard error values and *t*-parameter for the confidence levels 95 or 99% and degrees of freedom ($n - p$), and p values are determined using a *t* distribution $n - p$ degrees of freedom.

Availability and user-friendliness are the most important characteristics of software packages used for non-linear regression and these are satisfied by R-BioXL. However, researchers generally neglect the numerical accuracy of software programs [19]. Excel® and its statistical features including non-linear regression in Solver has been improved throughout the years [20–24]. As mentioned above, the results of the examples shown in this study were all same (same parameter estimates with the same standard errors and same goodness-of-fit indices) for SigmaPlot 12.0, MATLAB R2017b and SPSS 22.0. Moreover, this comparison has done for different datasets including drying of foods, microbial inactivation, degradation/formation kinetics of several compounds etc. (results not shown) and R-BioXL produced the same exact results with SigmaPlot, MATLAB and SPSS showing the accuracy and reliability of the tool.

DISCUSSION

In this study, we showed that R-BioXL tool can be safely used to describe the biological data with suitable mathematical models. Users should use the Solver routine to obtain the parameter values and should enter the number of parameters in the model. The rest (standard errors, confidence intervals, p values and goodness-of-fit indices) are calculated automatically and instantly. The tool could be beneficial for the ones dealing with the biological data as well as chemical or physical data; however, users are expected to have some experience in Excel® since the model equation and initial parameter estimates should be entered by the users. Results also revealed that R-BioXL can be used not only for non-linear models, but also for the linear models; however, users are again expected to enter the initial estimates of the parameters which is not required for the linear regression.

Some Excel® based tools or Add-Ins are designed for unexperienced users on modeling. GInaFiT [25] for example, can be used to describe microbial inactivation and it contains 10 different models. The user can select a model or more than one model for the same dataset and can compare the results. Parameters, asymptotic standard errors and goodness-of-fit indices are also listed. DMFiT is another Excel® Add-In in which Baranyi model [26] can be used for microbial growth. The above examples are for the ones who are not expert in regression or mathematical modeling. Although some experience in Excel® is required for the use of R-BioXL, there is no restriction for entering the model i.e., user is free to input any model. However, there are two pitfalls in the usage of R-BioXL (i) Solver can find a local instead of a global minimum and (ii) standard errors may be underestimated. In fact, first issue is not specific to R-BioXL and can be easily solved because user can change the initial values of the parameters (as we did above), observe the model together with the data on the graph and select the suitable initial estimates to start the iteration. If the same parameter values are obtained with different initial values, a global minimum has likely been found [8]. Second problem has no direct solution because standard errors and confidence intervals are only approximate in R-BioXL. Therefore, confidence intervals calculated this way may be underestimated by a factor of 2-3 [8, 27]. Of course, this underestimation depends on the model structure or non-linearity of the model equation.

It may still be possible to calculate the asymmetrical confidence intervals in Excel® by using Monte Carlo (MC) simulation [13, 28] which can be considered as the best method to do that [8, 29, 30]. An Excel® based tool ÖK-BUZ GRoFiT [31] for microbial growth modeling can be given as an example. ÖK-BUZ GRoFiT has three growth models in it (Baranyi, modified Gompertz and three-phase linear models) and there are three versions of this tool. The third version of ÖK-BUZ GRoFiT is available both in Turkish and in English (https://drive.google.com/drive/folders/1X_sNdpdQ2dT3KKl6KIYGNW0dq_KzJV_Q) and uses linear approximation to calculate the confidence intervals just like R-BioXL. First version (available only in Turkish) on the other hand, calculates the confidence intervals by 100 MC simulations and depending on the computer speed, it may take about 10 to 30 seconds to finish the simulations. Normally, between 1000 and 10000 simulations are performed in MC analysis [7]; however, using higher number of simulations would not affect the results in microbial growth modeling [31, 32]. Moreover, increasing the number of simulations requires more time. Therefore, performing MC simulations in Excel® to calculate the asymmetrical confidence intervals for the non-linear models is an option, but this would affect the speed of the analysis. The protocol described by Kemmer and Keller [5] can also be used to determine confidence intervals of the parameters in Excel®; however, the protocol takes about an hour which makes it difficult to implement the analyses of the data.

CONCLUSION

We introduced R-BioXL (Regression tool for biological data in Excel®) which is an Excel®-based user-friendly tool for regression models to define *X*-*Y* data. Researchers who deal with mostly biological data can safely use it to describe their data with suitable (non-linear) mathematical models and obtain parameter estimates by using Excel® Solver. Moreover, parameters' precisions and goodness-of-fit of the models can be determined accurately in R-BioXL and

this was demonstrated with three different examples in this study.

REFERENCES

[1] Hu, W., Xie, J., Chau, H.W., Si, B.C. (2015). Evaluation of parameter uncertainties in nonlinear regression using Microsoft Excel Spreadsheet. *Environmental Systems Research*, 4, 4.

[2] Serment-Moreno, V. (2021). Microbial Modeling Needs for the Nonthermal Processing of Foods. *Food Engineering Reviews*, 13, 465–489.

[3] Leylak, C., Yurdakul, M., Buzrul, S. (2020). Use of Excel in food science 1: Linear regression. *Food and Health*, 6, 186–198.

[4] Brown, A.M. (2001). A step-by-step guide to non-linear regression analysis of experimental data using a Microsoft Excel spreadsheet. *Computer Methods and Programs Biomedicine*, 65, 191–200.

[5] Kemmer, G., Keller, S. (2010). Nonlinear least-squares data fitting in Excel spreadsheets. *Nature Protocols*, 5, 267–281.

[6] Yurdakul, M., Leylak, C., Buzrul, S. (2020). Use of Excel in food science 2: Non-linear regression. *Food and Health*, 6, 199–212.

[7] van Boekel, M.A.J.S. (2022). Kinetics of heat-induced changes in dairy products: Developments in data analysis and modelling techniques. *International Dairy Journal*, 126, 105187.

[8] van Boekel, M.A.J.S. (1996). Statistical aspects of kinetic modeling for food science problems. *Journal of Food Science*, 61, 477–485.

[9] de Levie, R. (2004). Advanced Excel for scientific data analysis. New York, USA, Oxford University Press.

[10] Chase, A.M., von Meier, H.C., Menna, V.J. (1962). The non-competitive inhibition and irreversible inactivation of yeast. *Journal of Cellular and Comparative Physiology*, 59, 1–13.

[11] van Boekel, M.A.J.S. (2008). Kinetic Modeling of Reactions in Foods. Boca Raton, CRC Press.

[12] Moody, H.W. (1982). The evaluation of the parameters in the van Deemter equation. *Journal of Chemical Education*, 59, 290–291.

[13] Lambert, R.J.W., Mytilinaios, I., Maitland, L., Brown, A.M. (2012). Monte Carlo simulation of parameter confidence intervals for non-linear regression analysis of biological data using Microsoft Excel. *Computer Methods and Programs Biomedicine*, 107, 155–163.

[14] Zwietering, M.H., Jongenburger, I., Rombouts, F.M., Van't Riet, K. (1990). Modeling of the bacterial growth curve. *Applied Environmental Microbiology*, 56, 1875–1881.

[15] Alcantara, I.M., Naranjo, J., Lang, Y. (2022). Model selection using PRESS statistic. *Computational Statistics*, 38, 285–298.

[16] Öksüz, H.B., Buzrul, S. (2020). Monte Carlo analysis for microbial growth curves. *Journal of Microbiology, Biotechnology and Food Sciences*, 10, 418–423.

[17] de Levie R (2012). Collinearity in least-squares analysis. *Journal of Chemical Education*, 89, 68–78.

[18] de Levie R (2012). Nonisothermal analysis of solution kinetics by spreadsheet simulation. *Journal of Chemical Education*, 89, 79–86.

[19] Bergtold, J,S,, Pokharel, K.P., Featherstone, A.M., Mo, L. (2018). On the examination of the reliability of statistical software for estimating regression models with discrete dependent variables. *Computational Statistics*, 33, 757–786.

[20] McCullough, B.D., Wilson, B. (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis*, 31, 27–37.

[21] McCullough, B.D., Wilson, B. (2000). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. *Computational Statistics and Data Analysis*, 40, 713–721.

[22] McCullough, B.D., Wilson, B. (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics and Data Analysis*, 49, 1244–1252.

[23] McCullough, B.D., Heiser, D.A. (2008). On the accuracy of statistical procedures in Microsoft Excel 2007. *Computational Statistics and Data Analysis*, 52, 4570–4578.

[24] Mélard, G. (2014). On the accuracy of statistical procedures in Microsoft Excel 2010. *Computational Statistics*, 29, 1095–1128.

[25] Geeraerd, A.H., Valdramidis, V.P., van Impe, J.F. (2005). GInaFiT, a freeware tool to assess non-log-linear microbial survivor curves. *International Journal of Food Microbiology*, 102, 95–105.

[26] Baranyi, J., Roberts, T.A. (1994). A dynamic approach to predicting bacterial growth in food. *International Journal of Food Microbiology*, 23, 277–294.

[27] Johnson, M.L. (1992). Why, when, and how biochemists should use least squares. *Analytical Biochemistry*, 206, 215–225.

[28] Buzrul, S. (2021). Monte Carlo simulation in Microsoft Excel: Confidence intervals of model parameters for non-linear regression used in food sciences. *Akademik Gıda*, 19, 291–299.

[29] Press, W.H.,. Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (1989). Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, New York.

[30] Straume, M., Johnson, M.L. (1992). Monte Carlo Method for determining complete confidence probability distributions of estimated model parameters. *Methods in Enzymology*, 210, 117–129.

[31] Öksüz, H.B., Buzrul, S. (2021). An Excel-based, user-friendly freeware tool to describe microbial growth curves: ÖK-BUZ GRoFiT. *Journal of Tekirdag Agricultural Faculty*, 18, 521–532

[32] Buzrul, S. (2024). Fen Bilimleri ve Mühendislik Uygulamalarında Deneysel Verilerin Matematik Modellerle Tanımlanması. *Excel Uygulamalı Anlatım*. Ankara, Türkiye, *Akademisyen Kitabevi*.