International Journal of Science, Environment and Technology, Vol. 4, No 1, 2015, 86 – 95

ISSN 2278-3687 (O)

AN APPLICATION OF NON-LINEAR REGRESSION ANALYSIS FOR FOOD SCIENCE DATA WITH MICROSOFT EXCEL SOLVER

Mohd Tarmizan Ibrahim and Ishak Ghani
School of Food Science, Faculty of Bioresources and Food Industry, Universiti Sultan ZainalAbidin, Tembila Campus, 22000 Besut, Terengganu DarulIman, MALAYSIA
E-mails: mtarmizan@unisza.edu.my; ishakghani@unisza.edu.my

Abstract: The objective of this paper was to introduce a simple, fast, robust, reliable and easily explained procedure for conducting non-linearregression analysis based on user input functions. The method described here is to use the SOLVER function in the spreadsheet programme Microsoft Excel, which employs an iterative least squares fitting protocol toproduce the optimal goodness of fit to the experimental data. The data to be used as an example is in the food science area.
Keywords: Non-linear regression; Microsoft Excel; Curve fitting; SOLVER; Sorption isotherm.

INTRODUCTION

The practice of applying curve fitting techniques to describe experimental data is widely used in all fields of research particularly in food science. The purpose of curve fitting in food science data is to describe data in the universally recognized form $y = f(x)$, where y is the dependent variable and is measured in the experiment, and x is controlled during the experiment which is called independent variable(Bowen, 1995). The relationship between x and y is described by a function f which is in form of an equation containing one or more parameters. The better the fit, the more accurately the function describes the data. Application of linear fitting to experimental data is a relatively straightforward method and can be executed with a few simple point-and-click commands. However, describing data with non-linear functions (non-linear regression) is more problematical. This can be performed using specialized computer softwares such as Sigma Plot, MATLAB, Minitab and others. However, these softwares tend to be expensive and contain an excess of unnecessary features. It is only suitable for experienced users with a mathematical background and is not applicable for a novice user to learn. Apart from that, these softwares cannot manipulate data very well and

*Received Dec 23, 2014 * Published Feb 2, 2015 * www.ijset.net*

tend to show data, graphs, results and analysis in multiple windows which may be confusing to the user.

Microsoft Excel is an alternative program to fit non-linear functions. This software is a part of the Microsoft Office package and thus no additional expense is required. It has a user-friendly interface with good data handling capabilities, built-in mathematical function and instantaneous graphing. It even contains the SOLVER function, which is ideally suited for fitting data with non-linear functions through an iterative algorithm (Bowen, 1995). The objective of this paper is to carry out non-linear regression analysis tofood sciencedata with user-input functions using the SOLVER function of Microsoft Excel. The data to be used is sorption isotherm of a food.

The sorption isotherm of a food is a curve at which the equilibrium water content (X_{eq})(kg water per kg dry solid)of a food material is plotted as a function of water activity (a_w) at a given constant temperature. Both the a_w and X_{eq} are to be determined when the system has reached equilibrium (Chen and Mujumdar, 2008). Water sorption isotherms illustrate the steady-state amount of water held (i.e, water holding capacity) by the foods as function of a_w or relative humidity at constant temperature (Barbosa-Cánovas, 2007). The knowledge and understanding of sorption isotherms of food material is very important in food science for the design and optimization of drying apparatus, design of packaging material, prediction of stability or shelf-life and for determination of moisture changes that can occur during storage of food product. The sorption isotherms of most food materials are non-linear and generally sigmoid shaped. Each food material shows different type of sorption isotherm and it is based on the chemical composition and physicochemical state of food's constituents. Brunauer*et al*. (1940) and Chen and Mujumdar (2008) described five types of isotherms according to their shape and processes, as it is shown in Figure 1.

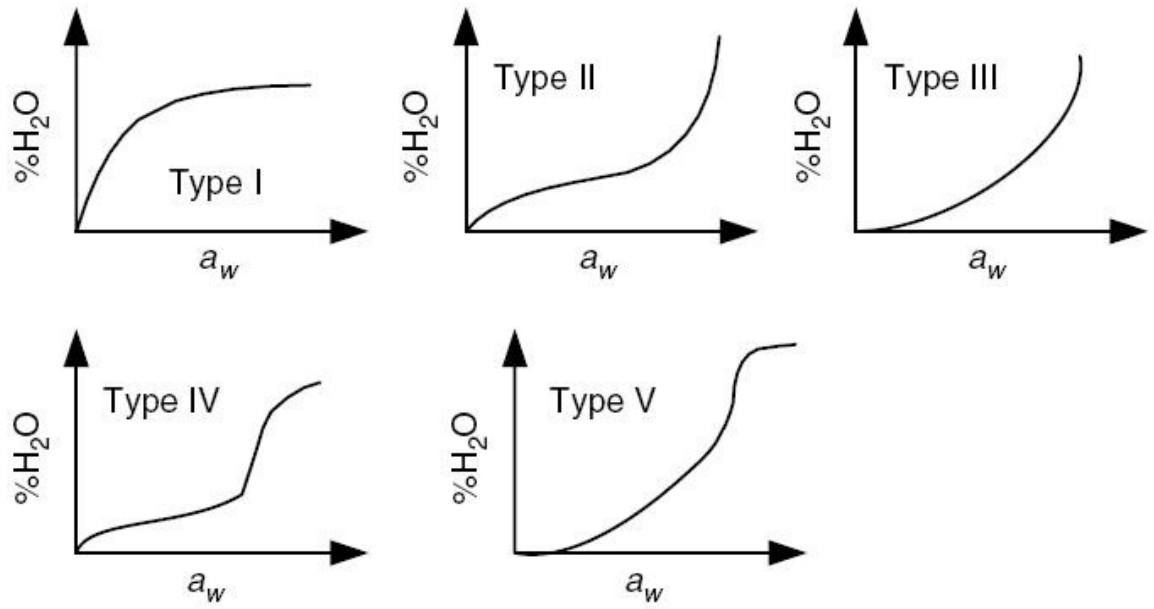

Figure 1: General shape of isotherms observed in food materials
(Chen and Mujumdar, 2008, pg. 76)

Several mathematical models have been developed to describe the sorption isotherms. For prediction and data fitting of sorption isotherms of foods some of semi-empirical equations with two or three fitting parameters such as BET, GAB and Oswinare most commonly used. In this paper the sorption isotherm data of banana chip at 25 °C from Kim (2014) as shown in Table 1 will be used for non-linear regression analysis.

a_w	X_{eq}
0.229	0.0247
0.385	0.0434
0.462	0.0456
0.599	0.0615
0.621	0.0794
0.735	0.1085
0.788	0.1412
0.892	0.2249

Table 1: Sorption isotherm data of banana chip (Kim, 2014, pg. 16)

METHOD

To perform a curve fitting protocol in an Excel spreadsheet, the method described in this paper was carried out on a Microsoft Pentium i5 laptop and Excel 2013. The method involves manual data entry and graphing of data, followed by curve fitting and displaying the resulting curve fitted to the data. The goodness of fit can be assessed by calculating R^2value.The R^2 value is called the coefficient of determination and itsvalue represents the fraction of the overall variance of the dependent variable that is explained by the independent variable (Bowen, 1995).

Traditionally, the non-linear data could be change into a linear form and afterward analysed by linear regression. This transformation tends to give inaccurate analysis since the linear regression was performed on transformed data, which can also change the experimental error. This method is outdated and should not be used anymore. One of the suitable methods for analysing non-linear data is called iterative non-linear least squares fitting. This method uses the same objective as explained for linear regression, which is to minimize the squared sum (SS) value of the difference between the experimental data and fit. The detail about the SS value is explained by Billo (2011). It is an iterative or cyclical method which is different than the linear regression method. In this method the user needs to provide initial estimation of the parameter. The first iteration step calculates an initial SS value based on the initial value given by the user. The second iteration step changes the parameter value by a small amount and recalculates SS value. This processing step is repeated many times to make surethat changes in the parameter values give the smallest possible value of SS. SOLVER function in Excel uses the generalized reduced gradient (GRG) iteration programming method. A detailed description of the evolution and implementation of this code are described by Smith and Lasdon (1992).

The following example illustrates how to use the SOLVER function in Excel to fit the sorption isotherm data as in Table 1 with user-input non-linear functions. In this example Oswin sorption isotherm equation as in Eq. 1 is used to fit the data. The equation describes equilibrium moisture content of a material (food) at a certain relative humidity or water activity. The Oswin model is shown below (Chen and Mujumdar, 2008):

$$X_{eq} = k\left(\frac{a_w}{1-a_w}\right)^n \quad (1)$$

whereX_{eq} is the equilibrium moisture content (dependent variable), a_w is water activity (independent variable) and k and n are the fitting parameters, which will be determined.

To perform non-linear regression analysis using the Oswin equation, the following proceduresare carried out:

1. Insert onto a spreadsheet the experimental data in two columns, column C containing the independent variable (a_w) and column D containing the dependent variable (X_{eq}). This is illustrated in Fig. 2.

2. Graph the inserted data contained in cells C25 to D33 in a Scatter plot. The data points are displayed as filled squares.

3. Enter labels in cells I25 to I32 to describe the contents of the adjacent cells. In cell I25 enter k, which will describe the parameter in cell J25. By selecting the cells I25 and J25 select FORMULAS in the top menu and in the Name Manager click “Create Names from Selection”. A small window dialogue will appear and tick the option ‘Left column’. This will assign the name in cell I25 to the cell J25. Similarly, for cells I26 to I32 enter n, Mean of y, df, SE of y, R^2, Critical t and CI respectively. To assign the names in the cells J26 to J32, repeat the same procedure as explained for cell J25.

4. Input initial estimation value of the parameters k and n into cells J25 and J26, respectively. In this example initial estimates are 0.1 and 0.2, respectively.

5. In column E ($X_{eq,fit}$) key in the equation describing the Oswin function. This has been rearranged from Eq. 1 into a form that Excel recognizes:

=k*(C26/(1-C26))^n, where k and n refer to the parameter values in cells J25 and J26.

6. Copy the equation from cell E27 down to E33. Please take note that C26 is a relative reference, which specifies the location of a cell relative to the cell in which the calculation will be carried out, in this case cell E26. Thus copying from Rows 25 to 33, changes the value of C26 to reflect the appropriate row.

7. Tocalculate mean value of the y, enter the following formula in J27.

=AVERAGE(D26:D33)

8. The degrees of freedom is calculated by entering the following formula in J28. It is defined as the number of data points minus the number of parameters in the functions.

= COUNT(D26:D33)-COUNT(J25:J26)

9. The standard error of the y values is defined as (Billo, 2011)

$$SE = \sqrt{\frac{\sum(y-y_{fit})^2}{df}} \quad (2)$$

and is calculated by entering the following formula in J29.

=SQRT(SUM((D26:D33-E26:E33)^2)/df)

However as this formula must be expressed as an array formula, press Ctrl + Shift + Enter. This encloses the whole formula within a pair of curly brackets ({}), denoting it as an array formula.

10. The R^2 value is calculated by entering the following formula in J30 and expressing it as an array formula as described above.

=1-SUM((D26:D33-E26:E33)^2)/SUM((D26:D33-Mean_of_y)^2)

11. To determine the confidence interval of the fit, the critical t value at a significance level 95% is calculated by entering the following formula in J31.

=TINV(0.05,df)

The confidence interval (CI) is defined in J32.

=Critical_t*SE_of_y

Enter the following formula in F26

=E26+Cl

and copy it down to F33. Similarly enter

=E26-Cl

In G26 and copy down to G33. This will give the upper and lower confidence limit values of the fit.

12. The SE of the y values, R^2 and CI are automatically calculated: 0.052, 0.458 and 0.128, respectively.

13. Figure 2 shows the spreadsheet template with the formulas used in this procedure.Plot the graphof columns E, F and G versus column C such that they are displayed as lines on the graph as shown in Figure 3. It is clearly seen that the initial estimation values(thick line) are not a good fit of the data with large confidence limits (thin lines).

14. Open the SOLVER function, which can be found under the Data menu. The dialogue box illustrated in Figure 5 appears. If SOLVER is not in this menu it should be installed. Refer Excel documentation for installation procedure.

15. In “Set Target Cell” box enter J30

16. Set the “Equal To” option to “Max”. This means SOLVER tries to calculate the maximum values of R^2.

17. In “By Changing Cells” box enter J25:J26. This means SOLVER tries to change the values of k and n until it get the maximum value of R^2 (R^2 expresses the proportion of variance in the dependent variable explained by the independent variable. The correlation

index of 0 means that x does not help to predict y. As the R^2 value increases toward 1 the more accurately the function fits the data).

	C	D	E	F	G	H	I	J
25	a_w	X_{eq}	$X_{eq,fit}$	Upper CI	Lower CI		k	0.100
26	0.229	0.0247	=k*(C26/(1-C26))^n	=E26+CI	=E26-CI		n	0.200
27	0.385	0.0434	=k*(C27/(1-C27))^n	=E27+CI	=E27-CI		Mean of y	=AVERAGE(D26:D33)
28	0.462	0.0456	=k*(C28/(1-C28))^n	=E28+CI	=E28-CI		df	=COUNT(D26:D33)-COUNT(J25:J26)
29	0.599	0.0615	=k*(C29/(1-C29))^n	=E29+CI	=E29-CI		SE of y	=SQRT(SUM((D26:D33-E26:E33)^2)/df)
30	0.621	0.0794	=k*(C30/(1-C30))^n	=E30+CI	=E30-CI		R^2	=1-SUM((D26:D33-E26:E33)^2)/SUM((D26:D33-Mean_of_y)^2)
31	0.735	0.1085	=k*(C31/(1-C31))^n	=E31+CI	=E31-CI		Critical t	=TINV(0.05,df)
32	0.788	0.1412	=k*(C32/(1-C32))^n	=E32+CI	=E32-CI		CI	=Critical_t*SE_of_y
33	0.892	0.2249	=k*(C33/(1-C33))^n	=E33+CI	=E33-CI			

Figure2. Spreadsheet template for non-linear regression: The dates are entered into Column C and D used to generate the fit based on the parameters in cells J25 and J26. Column F and G calculate the 95% confidence interval around the fit.

18. The "Subject to the Constraints" option can be left empty. Constraints are used to impose limits over the range values used to define parameters. If for example the value of parameter k is below 10, then it can be defined in this option by clicking the Add button and inserting "k < 10" in the box.

19. Click "Solve" to perform the fit. The SOLVER will iteratively cycle through the fitting routine, adjusting the parameter values of k and n to maximize the value of R^2. The optimal values of k and n are 0.055 and 0.675, respectively, and the maximum value of R^2 is 0.993.The thick line in Fig. 4shows the best fit and it is clear that it is an improvement over the fit provided by the initial parameter values. Additionally the confidence intervals (thin lines) around the fit have been reduced. So the final equation is

$$X_{eq} = 0.055\left(\frac{a_w}{1-a_w}\right)^{0.675} \quad (3)$$

20. This procedure can be repeated with different initial values of k and n to see if SOLVER finds the same solution.

The default SOLVER setting can be changed by clicking the "Options" and the "Solver Options" dialogue box will appear. Each option has a default setting which is appropriate for most situations but it can be changed. This option is designedonly for experienced user and therefore will be not discussed in this paper.

Note that, thestandard error of the data around the regressionline is calculated, which is also known as the standard error of theresiduals. In the procedure above, thestandard error of the residuals is used tocalculate the confidence interval. The confidence interval is an indicator of the probability that thetrue value lies within the range specified by theprobability formula

(Bowen, 1995). The common value of confidence interval is 95 %, which means that there is a95% probability that the true value lies within theinterval. To calculate the confidence interval the critical t-value must be determined and it depends on thedegrees of freedom. Microsoft Excel has a built-in function (tinv) to determine the critical t-value and the input formula in cell J31 calculates this value for the desired confidence intervaland degrees of freedom.

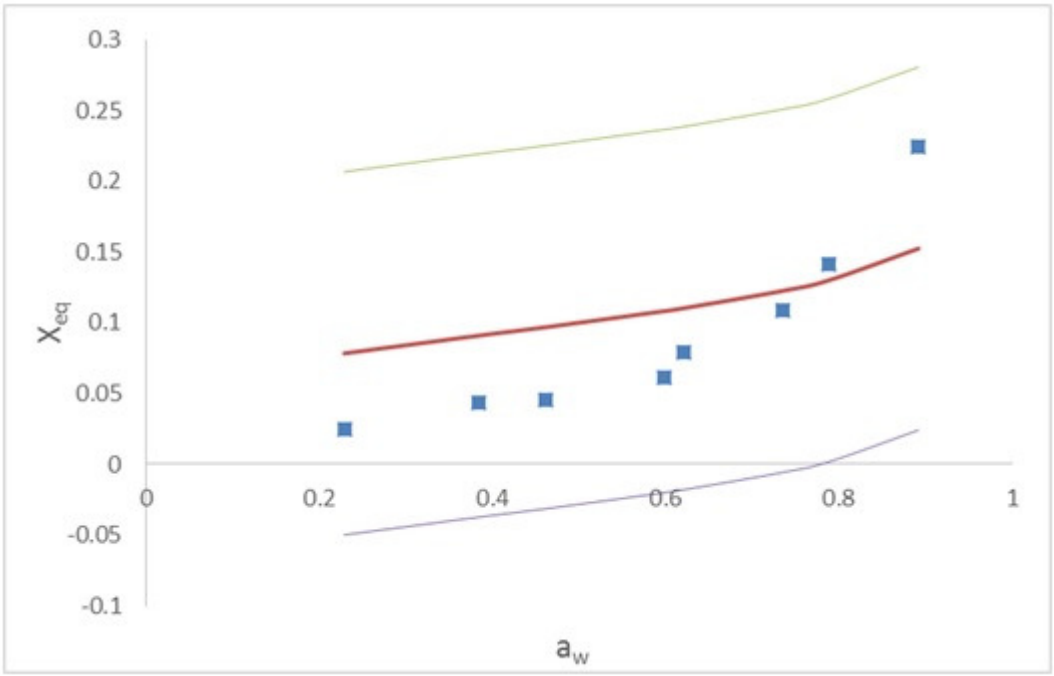

Figure 3: Oswin fit of sorption isotherm data based on the initial parameter estimates (thick line), and the 95% confidence interval (thin line) around the fit.

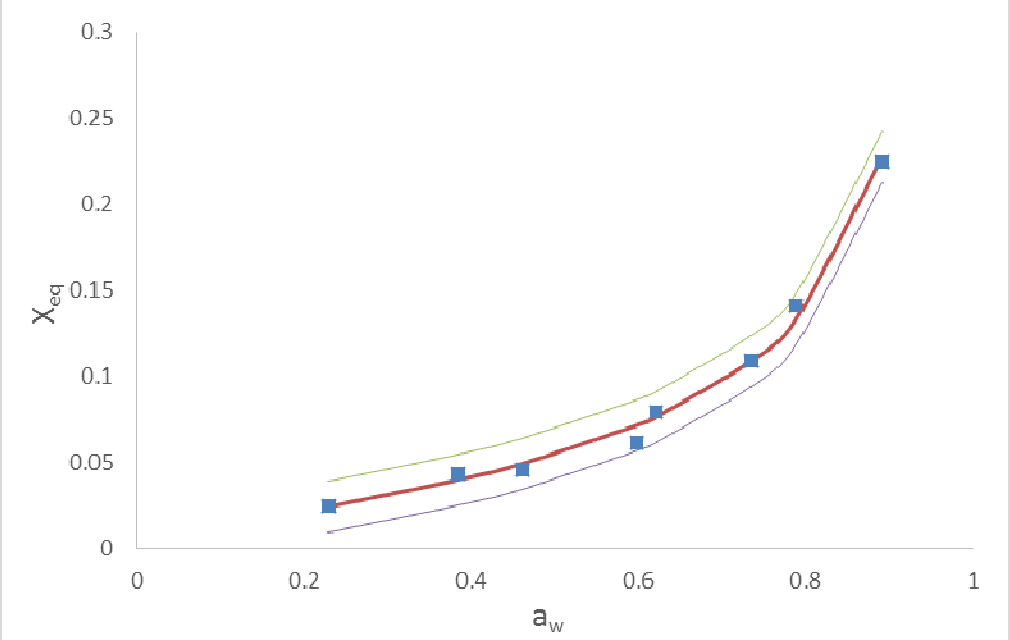

Figure 4: Oswin fit of sorption isotherm data calculated by the SOLVER.

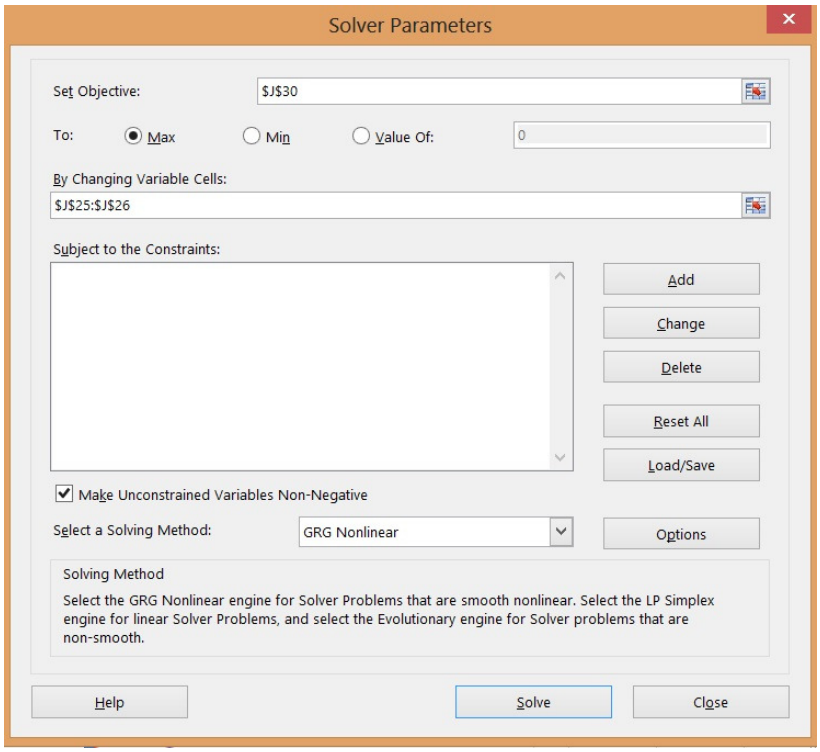

Figure 5: The built-in SOLVER function

CONCLUSION

Non-linear regression technique is a useful and reliable method for standardizing experimental data analysis. The procedure in this paper involves manual data entering and graphing it. This method is suitable for the user with a basic knowledge of Excel and do not

require the user to understand the mathematics behind the processes involved in curve fitting technique. However, it is important that the user understands enough about the data to be fitted, uses the correct type of analysis and can judge goodness of fit from the results. In this paper, the Oswin model is applied to the sorption isotherm data and the result is shown in Eq. 3.

The R^2 value calculated in this paper is assessed to givethe goodness offit of the function to the data. Assume that a certain function is used to describe the data, the accuracy of thefunction describes orfits the data can be determined based on the R^2 value. In this paper the R^2 value was 0.993 which means that 99.3% of the variation of the independent variable can be explained by the variation of the dependent variable.

Although this technique is considered as robust and reliable, a few points however should be taken into consideration. First, when the number of parameters in a function is larger the SOLVER will take longer the time to find the optimum values. Second, initial parameter values should be sensible and if they are inappropriate, the iteration process may lead in the wrong direction and the solution may never be found.

REFERENCES

[1] Barbosa-Cánovas, G.V. (2007). *Water activity in foods: fundamentals and applications* (1st ed.). Ames, Iowa: Blackwell Pub.

[2] Billo, E.J. (2011). *Excel for chemists: a comprehensive guide*. New York: Wiley-VCH.

[3] Bowen, A.M. (1995). Nonlinear regression analysis of data using a spreadsheet. *Trends in Pharmacological Sciences, 16*, 413-417.

[4] Brunauer, S., Deming, L.S., Deming, W.E., & Teller, E. (1940). On theory of the van der Waals adsorption of gases. *Journal of the American Chemical Society, 62*, 1723-1732.

[5] Chen, X.D., & Mujumdar, A.S. (2008). *Drying technologies in food processing*. Oxford: Blackwell Pub.

[6] Kim, S.K. (2014). *Determination of sorption isotherm of local dehydrated products (Kerepokkeping, KerepekUbi and Kerepek Pisang).* (Bachelor), Universiti Sultan Zainal Abidin, Terengganu.

[7] Smith, S., & Lasdon, L. (1992). Solving large sparse nonlinear programs using GRG. *ORSA Journal of Computing, 4*, 2-15.