American-Eurasian J. Agric. & Environ. Sci., 5 (2): 236-243, 2009
ISSN 1818-6769
© IDOSI Publications, 2009

Fitting a Nonlinear Regression Model to Gauge Heavy Metal Uptake in Spinach (*Amaranthus hybridus* L.)

R. Ahmad-Mahir, W.D. Wan-Rozita, J. Khairiah and B.S. Ismail

Faculty of Science and Technology, Universiti Kebangsaan Malaysia,
43600 UKM, Bangi, Selangor, Malaysia

Abstract: The main focus of the study was to model and analyze the data on the uptake of heavy metals specifically iron in spinach (*Amaranthus hybridus* L). The most common empirical model used is the simple regression but it was found that this model was not adequate based on the relationship and the value of R^2. By examining the plot of the accumulation of the heavy metal uptake it was found that it followed a nonlinear trend. Hence, a number of empirical models such as exponential, logistic, Gompertz, Weibull and piecewise regression were considered. SAS was used to estimate the model parameters, test significance and access goodness of fit of the statistical data. The simple change-point piecewise regression model was considered more appropriate modeling technique to be employed.

Key words: Nonlinear regression model . heavy metals . spinach . *Amaranthus hybridus*

INTRODUCTION

In general, heavy metals become toxic by forming complexes or 'ligands' with organic compounds. These modified biological molecules lose their ability to function properly and result in malfunction or death of the affected cells [1]. The study of heavy metal uptake in spinach showed the presence of considerably high amounts of iron (Fe) compared to other hazardous metals like Cadmium (Cd), Cromium (Cr), Copper (Cu), Manganese (Mn), Plumbum (Pb) and zinc (Zn). The concentration of iron is the highest and varies as much as four times more compared to Mn and five hundred times more compared to Cd. Iron is one of the heavy metals that is found in large amounts in the earth. Plant uptake is one of the major pathways by which iron in soils enters the human food chain. Iron in the form of supplements is very beneficial to our body. High concentration of iron exceeding the normal level may result in malfunction of human internal organs especially the liver and kidneys [2]. According to the Malaysian Food Act [3], concentration of heavy metals permitted in vegetables and fruits were 1 mg/kg and 300-500 mg/kg intake is considered critical. The "Top 20 Hazardous Substances" list according to The Agency for Toxic Substances and Diseases Registry (ATSDR) in Atlanta, USA does not include iron [4]. Nevertheless, iron is a heavy metal of concern, particularly because ingesting dietary iron supplements may acutely poison young children, such as a five to nine-mg iron tablet for a 30-lb child. In Dietary Reference Intake (DRI) the amount of safe intake for an adult is listed to be as high as 18 mg/day and for children below 8 years old, 10 mg/day [5].

MATERIALS AND METHODS

Non linear regression model: The main objective of the study was to model the relationship between dependent variables and the independent or explanatory variables. In this context, Fe uptake in mg/kg by spinach will be the dependant variable and number of days, the independent variable.

The nonlinear regression extends the linear regression model for use with much larger and more general classes of function. According to Myers *et al*. [6], any model that is not linear in the unknown parameters is a nonlinear regression model. It is used when a curvilinear relationship exists between the mean response and a predictor variable. For example, with a sample of n observations, the regression relationship is of the form

$$y_i = g(x_i;\beta) + \varepsilon_i \text{ for } i = 1,2,\ldots,n \qquad (1)$$

where $g(x_i;\ \beta)$ is some specific function which is nonlinear in one or more of x and $\beta = [\beta_0\ \beta_1 \ldots \beta_p]^T$ are parameters and the errors are normally and independently distributed with constant variance, that is $\varepsilon_i \sim$ i.i.d. $N(0,\sigma^2)$.

Corresponding Author: Dr. B.S. Ismail, School of Environmental and Natural Resource Sciences. Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia

Piecewise regression models: This model is known as the segmented regression model. It is made up of two sets of lines separated at a change point somewhere along the range of the predictor variable. When the change point is unknown the model is highly nonlinear. Segmented regression models can be continuous if the segments converge at the change point. This model becomes discontinuous if the change points suddenly drop or jump in the mean responses. For continuous models, the segmentation can be smooth or abrupt, depending on whether the derivatives with respect to the variable, exist at the change point.

In the simplest case, one of the straight lines is a horizontal plateau. This is called a plateau model, also referred to as a threshold model when the horizontal response corresponds to some threshold phenomenon. For a left-plateau model

$$g(x;\beta) = \begin{cases} \beta_0 & \text{if } x_i \le \tau \\ \beta_1 + \beta_2(x_i - \tau) & \text{if } x_i > \tau \end{cases} \qquad (2)$$

for use in equation (1). In the special case of $\beta_0 = 0$, equation (2) is called a truncated linear regression model. The mean response at the change point is assumed to be discontinuos in (2). Thus $g(x_i;\, \beta)$ jumps (or drops) a distance of β_1 - β_0 at $\tau = 0$. For unknown τ, any plateau model is inherently nonlinear. To fit the (left) threshold model the PROC NLIN was used with the Levenberg-Marquardt fitting algorithm. To find the point estimates for the β parameters and τ, nonlinear least squares were used. Inferences on these parameters followed the large-sample method [7].

Another type of piecewise regression models is the simple change-point model where the response is essentially flat for most of the x values, except it changes abruptly for some values of x. Thus this data was considered as two segments of the simple plateaus of the form:

$$g(x;\beta) = \begin{cases} \beta_1 & \text{if } x_i \le \tau \\ \beta_2 & \text{if } x_i > \tau \end{cases} \qquad (3)$$

Parameter estimation in a nonlinear system: Parameter estimation can be done via several methods namely nonlinear least squares, the geometry of linear and nonlinear least squares, maximum likelihood estimation, linearization and the Gauss-Newton method, Marquardt compromise and others. Some of the most commonly used methods are:

Non-linear Least Squares (LS): The least squares function of (1) is of the form

$$s(\beta) = \sum_{i=1}^{n} [y_i - f(x_i, \beta)]^2 \qquad (4)$$

To find the LS estimates, the equation (4) must be differentiated with respect to β. This will provide a set of p normal equations. The expectation of the function is a nonlinear function. The normal equations are

$$\sum_{i=1}^{n} [y_i - f(x_i, \beta)] \left[\frac{\delta f(x_i, \beta)}{\delta \beta_j} \right]_{\beta = b} = 0 \quad \text{for } j = 1, 2, \ldots, p \qquad (5)$$

As an example, consider the nonlinear regression model of the form:

$$y = \beta_1 e^{\beta_2 x} + \varepsilon$$

The LS normal equations for this model will be as follows:

$$\sum_{i=1}^{n} \left[y_i - b_1 e^{b_2 x_i} \right] e^{b_2 x_i} = 0$$

$$\sum_{i=1}^{n} \left[y_i - b_1 e^{b_2 x_i} \right] b_1 e^{b_2 x_i} = 0$$

These equations are not linear in b_1 and b_2 and no closed-form solution exists. In general, iterative methods must be used to find the values of b_1 and b_2.

Maximum Likelihood Estimation (MLE): If the distribution of the error is known then the method of likelihood estimation can be used. If the errors are normally and independently distributed with constant variance, application of the method of maximum likelihood to the estimation problem will lead to least squares.

In considering model (1), if the errors are normally and independently distributed with mean zero and variance σ^2, then the likelihood function is

$$L(\beta, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left[\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - f(x_i, \beta)]^2 \right] \qquad (6)$$

Maximizing the likelihood function is equivalent to maximizing the log-likelihood:

$$L(\beta, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - f(x_i, \beta)]^2 \qquad (7)$$

Choosing the vector of parameters b that maximizes the log-likelihood is equivalent to minimizing the residual sum of squares. Therefore, least squares in nonlinear regression are the same as MLE.

Marquardt's compromise: The method that will be used in this study is the SAS procedure PROC NLIN that employs the Marquardt compromise techniques in parameter estimation. Marquardt proposed computing the vector of increments at the kth iteration from

$$(D_k' D_k + \lambda I_p)\hat{\theta}_k = D_k'(y - f_k) \text{ where } \lambda > 0$$

Marquardt [8] used a search procedure to find a value of λ that would reduce the residual sum of squares at each stage. The PROC NLIN in SAS begins with $\lambda = 10^{-8}$. A series of trial-and -error computation are done at each iteration with λ repeatedly multiplied by 10 until

$$S(b_{k+1}) < S(b_k)$$

This general procedure is often called Marquardt's compromise because the resulting vector of increments produced usually lies between the Gauss-Newton vector in the linearization procedure and the direction of the steepest descent [6].

Statistical inferences from nonlinear regression: Inferences on the unknown parameters in a nonlinear regression should be base on F statistics such as

$$F_{calc} = \frac{SSE(RM) - SSE(FM)/\Delta e}{SSE(FM)/df_e(FM)} \qquad (8)$$

Where SSE is the sum of the squared errors calculated under the full model (FM) or under the reduced model (RM). The degree of freedom associated with each error terms are df_e (FM) and df_e (RM), with ?e+df_e(RM) = df_e (FM) [7]. For model adequacy assessment, residual plots can provide a good indicator for the test.

Polynomial regression model: This model is a linear regression model with an extension of simple linear relationships with the addition of higher order polynomial terms. A *p*-th order polynomial regression model is:

$$\begin{aligned} Y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \beta_2(x_i - \overline{x})^2 + ... \\ + \beta_p(x_i - \overline{x})^p + \varepsilon_i, i = 1,. . . ,n \end{aligned} \qquad (9)$$

with the usual homogeneous-variance, normal-error assumption: ε ~ i.i.d. $N(0,\sigma^2)$, Also, it is assumed that $p<n-1$ and that there should be at least p+1 distinguishable values among the x_i's. This model is used to represent simple curvilinear relationships between $E[Y_i]$ and x_i. Parameter estimates and statistical inferences follow the normal linear regression procedure.

RESULTS

Example: A set of data on heavy metal uptake by spinach leaves, stems and roots was used as an example. The dependent variable was the concentration of Fe (in mg/kg) absorbed by the spinach leaves, stems and roots and the independent variable was the number of days after sowing [9] Scatterplot from the raw data is as shown in Fig. 1.

This provided a guideline to initially recognize the shape of the model whether it is linear, nonlinear or polynomial, for later steps in modeling. From the plot, is assumed to be the shape more toward the nonlinear model, specifically the piecewise characteristic, as the first two points look like one straight line with another straight line below. Three types of regression models will be tested to model the data which are polynomial, exponential, simple change-point piecewise and right-plateau piecewise regressions models. The conclusion will be based on the results of the model fitted, residual plots and comparison of R^2 and P-values. Finally, the best fit models shall be known. The SAS PROC NLIN will be used for nonlinear regressions and the PROC REG for polynomial regression. The first step will be to find the equation for the first part and the second part of the piecewise regression model, their initial estimated values and their derivatives for β_0, β_1, β_2. The values of β and τ to be defined as part of an input in the SAS NLIN procedure.

The piecewise regression model for the right-plateau model would be of the form

$$g(x;\beta) = \begin{cases} \beta_0 + \beta_1(x_i - \tau) & \text{if } x_i \le \tau \\ \beta_2 & \text{if } x_i > \tau \end{cases} \qquad (10)$$

According to Piegorsch and Bailer [7], to fit the left threshold model the PROC NLIN should be employed with the Levenberg-Marquardt fitting algorithm. This requires the partial derivatives of (8) with respect to the unknown parameters. The first derivatives are simply:

$$\frac{\delta g}{\delta \beta_0} = 1, \ \frac{\delta g}{\delta \beta_1} = x - \tau, \ \frac{\delta g}{\delta \beta_2} = 1, \ \frac{\delta g}{\delta \tau} = -\beta_1$$

The initial estimates from the data of Fe absorption in leaves, suggest that segmentation occurs between the second and the third data points. Hence, initial

$$\tau_0 = \frac{1}{2}(8 + 11) = 9.5$$

For the slope of the right linear segment, the two data pairs are regressed to get the slope and

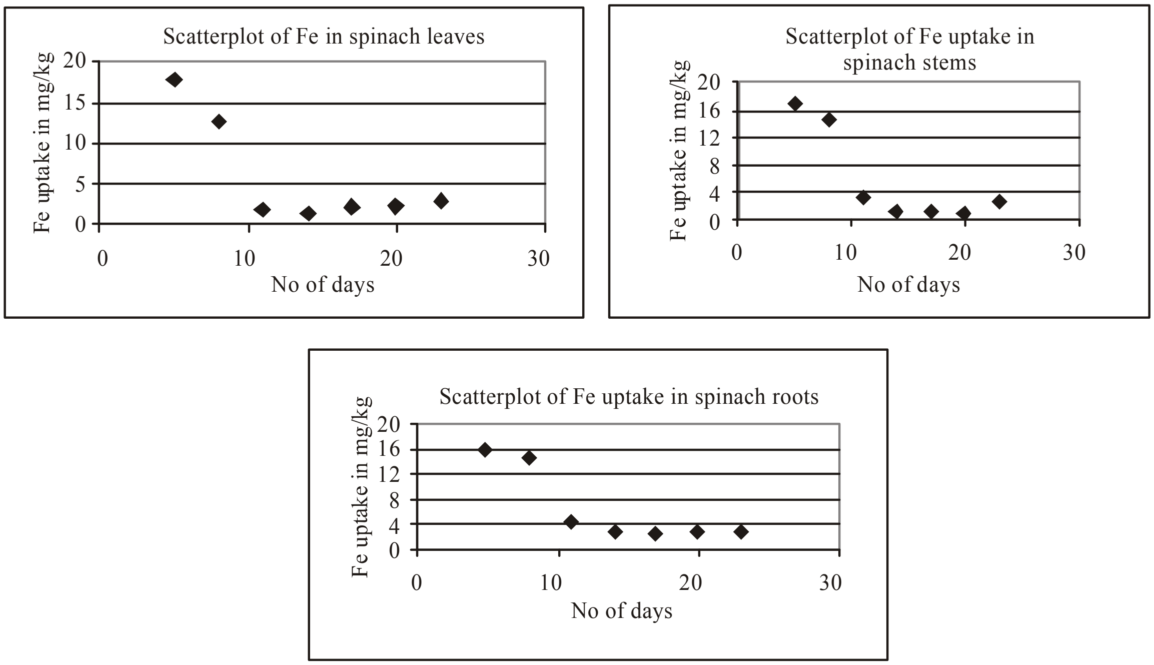

Fig. 1: Scatterplot for Fe uptake in Spinach in leaves stems and roots.

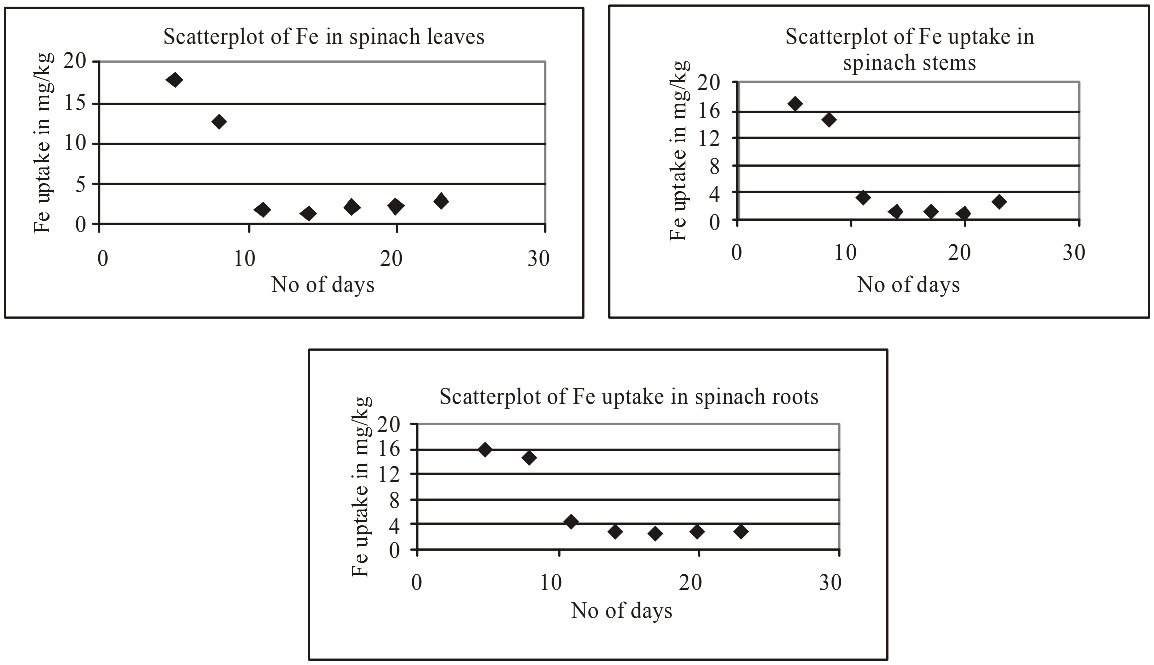

Fig. 2: The plot of the fitted right-plateau piecewise regression line together with scatterplots for Fe in spinach leaves, stems and roots.

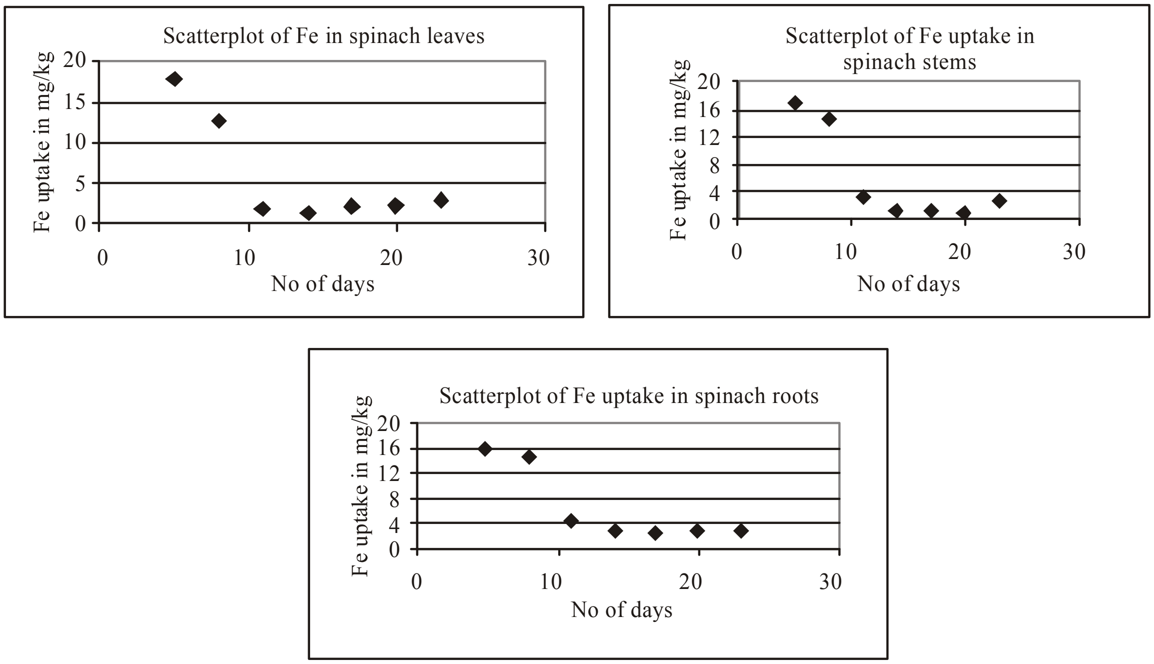

Fig. 3: Residual plots from fitting of right-plateau piecewise regression model to Fe uptake in spinach leaves, stems and roots.

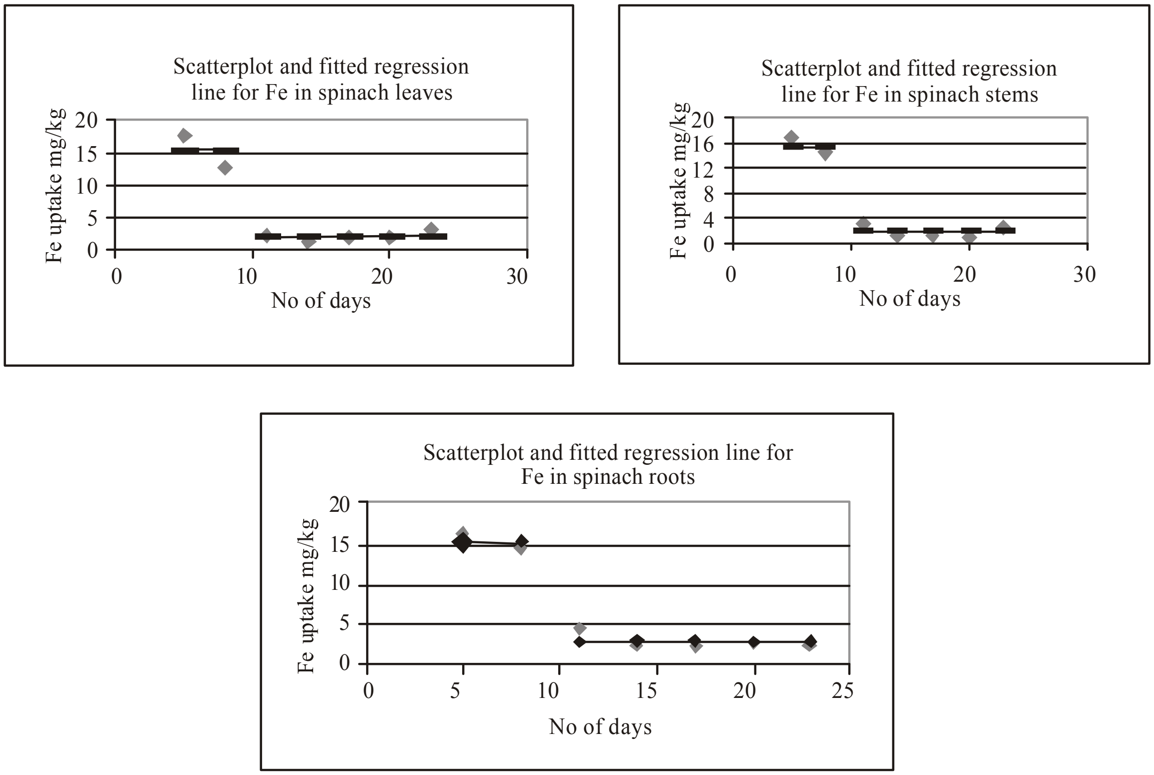

Fig. 4: The plot of the fitted simple change-point piecewise regression lines together with scatterplots for Fe in spinach leaves, stems and roots.

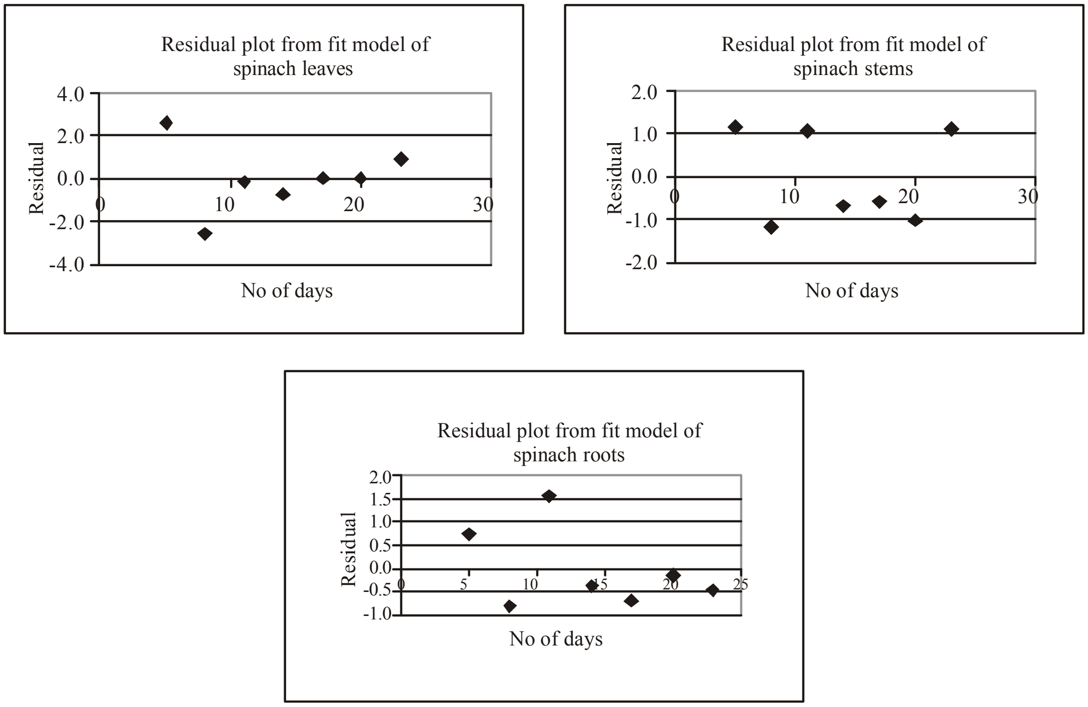

Fig. 5: Residual plots from fit of simple change-point piecewise regression model for Fe uptake in spinach leaves, stems and roots.

β_{00} = 26.47563 and β_{10} = -1.73487. For the horizontal-plateau on the right (Fig. 2), the β_{20} value is obtained by taking the mean of the last five Fe concentrations:

$$\beta_{20} = \begin{pmatrix} 1.8947 + 1.32 + 2.1277 \\ + 2.085 + 3.0643 \end{pmatrix} / 5 = 2.09834$$

Using these derivatives and initial values the SAS PROC NLIN program is then run. The same procedures are done for stems and roots. Below are the results for modeling using the piecewise right-plateau regression model:

The fitted model for Fe in spinach leaves is:

$$g(x;\beta) = \begin{cases} 9.9944 - 1.7349(x_i - 9.5) & \text{if } x_i \le 9.5 \\ 2.0983 & \text{if } x_i > 9.5 \end{cases}$$

The fitted model for Fe in spinach stems is:

$$g(x;\beta) = \begin{cases} 13.3148 - 0.7677(x_i - 9.5) & \text{if } x_i \le 9.5 \\ 1.9495 & \text{if } x_i > 9.5 \end{cases}$$

The fitted model for Fe in spinach roots is:

$$g(x;\beta) = \begin{cases} 13.5709 - 0.5380(x_i - 9.5) & \text{if } x_i \le 9.5 \\ 3.0651 & \text{if } x_i > 9.5 \end{cases}$$

The goodness of fit test will be done on the data of Fe in spinach leaves. The results show a rapid convergence with P-values < 0.0001 indicating that the model is very significant. An approximate 95% confidence intervals not containing zero implies that all the parameters contributed to the model and hence should be kept in the model. The residual plots shown in Fig. 3 are reasonable.

To test the null hypothesis, H_0: $\beta_0 = \beta_2 = 0$, the discrepancy-approach was used as in (8), to calculate the F-statistic with SSE(RM) = 260.30718 with d.f (RM)=6.

$$F_{calc} = \frac{(260.30718 - 1.5814)/(6-4)}{1.5814/4} = \frac{129.36289}{0.39535} = 327.2111$$

At α = 0.05, F_{calc} was compared to $F_{0.05(2,4)}$=6.944. Since F_{calc} exceeded this critical point, it can be concluded that β_0 and β_2 do deviate significantly from zero. This was confirmed by the 95% confidence interval for β_0 and β_2 in Fig. 3 where 7.2342<β_0< 12.7546 and 1.3176<β_2<2.8790

Results for the simple change-point piecewise regressions models are given as in Fig. 4 and the residual plot in Fig. 5. The fitted model for Fe in spinach leaves is

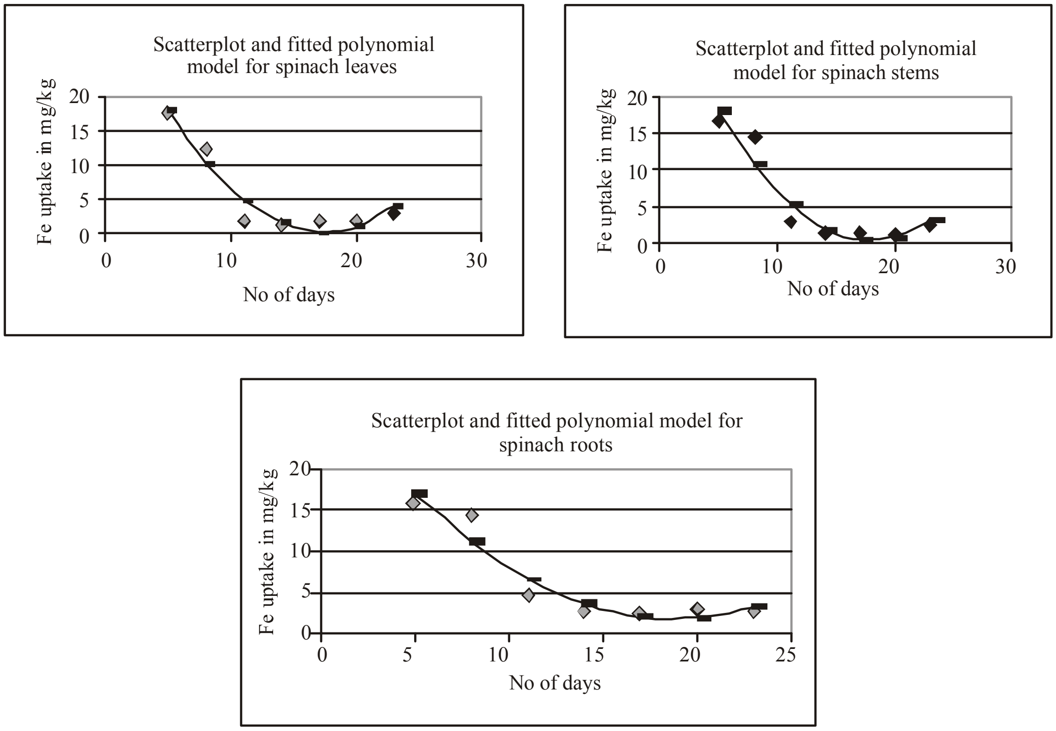

Fig. 6: The plot of the fitted polynomial regression model together with scatterplots for Fe in spinach leaves, stems and roots.

$$g(x;\beta)=\begin{cases}15.2 & \text{if } x_i \le 9.5\\ 2.1 & \text{if } x_i > 9.5\end{cases}$$

The fitted model for Fe in spinach stems is:

$$g(x;\beta)=\begin{cases}15.6 & \text{if } x_i \le 9.5\\ 1.9 & \text{if } x_i > 9.5\end{cases}$$

The fitted model for Fe in spinach roots is

$$g(x;\beta)=\begin{cases}15.1 & \text{if } x_i \le 9.5\\ 3.06 & \text{if } x_i > 9.5\end{cases}$$

Results for the polynomial regression models are given in Fig. 6. The study showed that fitted models for Fe uptake in leaves, stems and root parts are best represented by the polynomial regression model.

The fitted model for Fe in spinach leaves is:

$$g(x;\beta)=15.67-2.65x+0.12x^2$$

The fitted model for Fe in spinach stems is

$$g(x;\beta)=15.86-2.57x+0.11x^2$$

The fitted model for Fe in spinach roots is

$$g(x;\beta)=13.93-1.98x+0.08x^2$$

Table 1 lists the comparison among models used for estimating Fe uptake in spinach leaves, stems and roots. The model that can best represent Fe uptake in leaves was right-plateau piecewise regression and for both stems and roots the simple change-point piece regression model.

DISCUSSION AND CONCLUSIONS

Polynomial regression fits the data rather well for the leaves, stems and roots with R^2equal to 92.6, 91.8 and 91.6%, respectively. Simple change point piecewise regression gives a much better fit with R^2 equals to 97.0, 98.7 and 99.1%, respectively. Comparing the three models, the simple change-point piecewise regression is a more appropriate modeling technique to be employed for the above data sets. It is then extended to piecewise regression with right-plateau as discussed above. The latter gives R^2 equal to 99.7% for leaves, 96.3% for stems and 94.1% for roots. Taking into consideration the P-values and R^2 as shown in Table 1, it can be concluded that the right-plateau piecewise regression is the best model for assessing Fe uptake in spinach leaves whereas for stems and roots the simple change-point piecewise regression models are more appropriate.

Table 1: The p-values and R^2 for polynomial, simple change-point and right-plateau piecewise regression models for Fe uptake in spinach

Fe uptake in spinach	Polynomial regression model		Simple change-point piecewise regression model		Right-plateau piecewise regression model	
	P-value	R^2	P-value	R^2	P-value	R^2
Leaves	0.0055	92.6%	0.0002	97.0%	<0.0001	99.7%
Stems	0.007	91.8%	<0.0001	98.7%	0.0002	96.3%
Roots	0.007	91.6%	<0.0001	99.1%	0.0002	94.1%

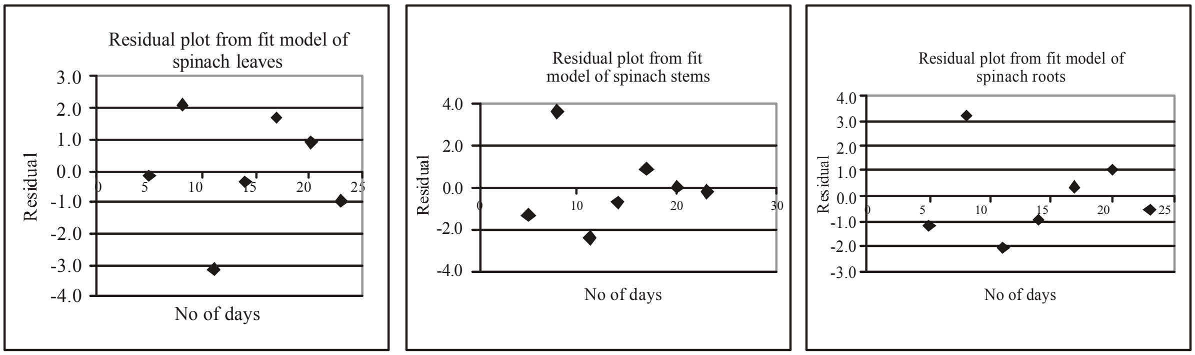

Fig. 7: Residual plots from fit of the polynomial regression model for Fe uptake in spinach leaves, stems and roots.

Analysis through piecewise regression models have shown that absorption of Fe is high particularly during the first eight days. After that it is reduced drastically from 15 mg/kg to about 2 mg/kg until day 20, after which it appears to increase again. It can be concluded that spinach should be harvested any time after the 11th day. Even though Polynomial regression can be use to model the data the analyses are misleading. The general trend shows that Fe uptake is very high in the early stage of growth, declines between the 18th and the 20th day but somehow continues to increase again after that in plants as the no of days increase. The model indicates that spinach continues to absorb more Fe again after the 18th day, meaning that consuming the vegetable after this period might be unsafe. Therefore, the polynomial regression model is not appropriate for interpreting the data.

ACKNOWLEDGEMENTS

This work was produced as part of a collaborative work of two research groups supported by the Universiti Kebangsaan Malaysia through the GUP: Project code UKM-OUP-TMK-2/2007.

REFERENCES

1. Landis, W.G. and H.Y. Ming, 1999. Introduction to environmental toxicology impacts of chemicals upon ecological systems. Boca Raton: CRC Press.
2. Duffus, J.H., 1980. Environmental toxicology; resource and environmental sciences series. London: Edward Arnold Publishers Ltd.
3. Food Act 1983 and Food Regulation, 1985. Laws of Malaysia. The Commissioner of Law Revision Malaysia.
4. Department of Health and Human Services, USA, 2008. Toxic Substances Portal, Agency for Toxic Substances & Disease Registry. Available from: http://www.health.gov/DietaryGuidelines/dga2005/document/default.htm(Access September 2008).
5. Department of Health and Human Services, 2008. Nutrition for Everyone. Centre for Disease Control and Prevention. Accessed from: http://www.cdc.gov/nccdphp/dnpa/nutrition/nutrition_for_everyone/basics/iron.htm#How%20much (Accessed September 2008).
6. Myers, R.H., D.C. Montgomery and G.G. Vinning, 2002. Generalized Linear Model: With Application in Engineering and the Sciences. New York: John Wiley and Sons, pp: 64-85.
7. Piegorsch, W.W. and A.J. Bailer, 2005. Analyzing Environmental Data. New York: John Wiley and Sons, pp: 45-64.
8. Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters. SIAM Journal of Applied Mathematic, 11: 431-441.
9. Wardatun-Aathirah, M.H., 2005. Study on heavy metal accumulation in spinach and brassica from Sepang Agro-Technology Park, Selangor. Bangi: Universiti Kebangsaan Malaysia.