UNIVERSITI PUTRA MALAYSIA

THE EFFECT OF REPARAMETERISATION ON THE BEHAVIOUR OF NONLINEAR ESTIMATES

NORAZAN BINTI MOHAMED RAMLI

FSAS 2000 4

THE EFFECT OF REPARAMETERISATION ON THE BEHAVIOUR OF NONLINEAR ESTIMATES

By

NORAZAN BINTI MOHAMED RAMLI

Thesis Submitted in Fulfilment of the Requirements for the Degree of Master of Science in the Faculty of Science and Environmental Studies
Universiti Putra Malaysia

May 2000

TO MY HUSBAND

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Master of Science.

THE EFFECT OF REPARAMETERISATION ON THE BEHAVIOUR OF NONLINEAR ESTIMATES

By

NORAZAN BINTI MOHAMED RAMLI

May 2000

Chairman: Habshah Bt. Midi, Ph.D.

Faculty: Science and Environmental Studies

This thesis discussed nonlinear modeling and measures of nonlinear behaviour. A set of data, representing the average weight of dried tobacco leaves (in

Several nonlinear models were used to fit the data, however only the Gompertz and the Logistic models were found to be suitable. The estimates of the parameters were calculated by using the Gauss-Newton algorithm in S-PLUS Programming Language.

A good estimator was the one which had the properties closed to the behaviour of a linear estimate. The nonlinear behaviour of the estimates was assessed using two different approaches, namely the analytical and the empirical approaches. These approaches were employed so that they could complement the existence of any laggings .

The study showed that the analytical approach of curvature measures of Bates and Watts could measure the average nonlinearity but could not determine the parameters that caused the nonlinear behaviour. Meanwhile, the bias formula of Box could only give the percentage of the extent to which the parameter estimates may exceed or fall short of the true parameter value, but could not be used to compare different parameterizations.

An advantage of using direct measure of skewness of Hougaard was that it was scale-independent and could be used to measure nonlinearity in different parameterizations. The empirical approach of simulation studies had successfully revealed the full extent of the nonlinear behaviour of the estimates and at the same time, suggested useful reparameterizations.

Reparameterization was used in order to remove or reduce the nonlinear behaviour of the parameter estimates. The study showed that the nonlinear behaviour of the parameter estimates was successfully reduced after reparameterization. The Logistic model in a reparameterized model function was found to best fit the data as it has the lo

therefore the closest-to-linear behaviour .

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains.

KESAN PEMPARAMETERAN SEMULA KE ATAS TINGKAHLAKU PENGANGGAR TAKLINEAR

Oleh

NORAZAN BINTI MOHAMED RAMLI

Mei 2000

Pengerusi: Habshah Bt Midi, Ph.D.

Fakulti: Sains dan Pengajian Alam Sekitar

Tesis ini membincangkan permodelan taklinear dan sukatan tingkahlaku taklinear. Satu set data yang mewakili purata berat daun tembakau kering sepokok (dalam gram) mengikut masa dalam minggu, digunakan dalam penyelidikan ini. Beberapa model taklinear digunakan untuk memodelkan data, walau bagaimanapun hanya model Gompertz dan model Logistic sahaja yang didapati sesuai. Nilai-nilai penganggar dikira menggunakan pendekatan Gauss- Newton dalam bahasa komputer S-PLUS.

Penganggar yang baik ialah penganggar yang tingkahlakunya hampir sama dengan penganggar linear. Tingkahlaku taklinear dinilai menggunakan dua pendekatan yang berbeza iaitu secara analitik dan empirik. Pendekatan yang

berbeza digunakan supaya dapat mengimbangi sebarang kekurangan yang wujud.

Kajian mendapati pendekatan analitik sukatan kelencongan Bates dan Watts dapat mengukur tahap sifat taklinear secara purata, tetapi tidak dapat menentukan parameter yang menyebabkan wujudnya tingkahlaku taklinear dalam model. Rumus pincang oleh Box pula hanya dapat memberi peratusan sejauh mana sesuatu penganggar kurang atau lebih daripada nilai yang sepatutnya tetapi tidak dapat digunakan sebagai pengukur taklinear untuk perbandingan dua pemparameteran yang berbeza.

Kelebihan yang ada menggunakan ukuran kepencongan oleh Hougaard ialah ianya adalah bebas skala dan boleh digunakan untuk menilai tingkahlaku taklinear dalam pemparameteran yang berbeza untuk dibuat satu perbandingan. Pendekatan empirik dalam kaedah simulasi pula dapat mendedahkan sejauh mana tingkahlaku taklinear penganggar dan pada masa yang sama mencadangkan pemparameteran semula yang berguna.

Pemparameteran semula digunakan untuk mengurangkan atau membuang tingkahlaku taklinear penganggar. Kajian menunjukkan tingkahlaku taklinear penganggar dapat dikurangkan dengan jayanya selepas proses pemparameteran semula. Model Logistic dalam fungsi model yang diparameterkan semula telah dipilih sebagai model yang lebih baik untuk memodelkan data yang diberi kerana model ini mempunyai tingkahlaku taklinear yang terkecil dalam sukatan kelencongan penganggarnya dan dengan itu yang paling hampir dengan tingkahlaku linear.

ACKNOWLEDGEMENTS

First and foremost, praise be to God for giving me the strength and patience to complete this work. Gratefully, I would like to thank my supervisor Dr. Habshah Midi for her excellent supervision, invaluable guidance, helpful discussions and continous encouragement, without which this work would not be finished. My thanks must also extend to the members of my supervisor committee, Associate Professor Dr. Muhammad Idreess Ahmad and Dr. Kassim Haron for their invaluable discussions, comments and help. Similar thanks should go to Pejabat MARDI, Serdang for providing me the sample data set needed for this study.

TABLE OF CONTENTS

Page

DEDICATION 2
ABSTRACT 3
ABSTRAK 5
ACKNOWLEDGEMENTS 8
APPROVAL SHEETS 9
DECLARATION FORM 11
LIST OF TABLES 14
LIST OF FIGURES 15

CHAPTER

I INTRODUCTION 16
Statement of the Problem 22
Some Key Words and Definitions 24
Linear Approximation 25
Curvature Measures 25
Parameter-Effects Nonlinearity 25
Intrinsic Nonlinearity 26
Solution Locus 26
Objective of the Study 26
Organization of the Thesis 27

II NONLINEAR DATA MODELING 29
Introduction 29
Scatter Plot and Nonlinear Models 30
Initial Parameter Estimates 33
Nonlinear Least Squares and Gauss-Newton Method 36
Results, Discussion and Conclusion 41

III MEASURES OF NONLINEARITY 44
Introduction 44
Curvature Measures of Bates and Watts 48
Bias Formula of Box 56
Direct Measure of Skewness of Hougaard 58
Simulation Study of Ratkowsky 62
Results, Discussion and Conclusion 69

IV REPARAMETERISATION 73
Introduction 73
Some Guidelines on Reparameterisations 74
Suitable Reparameterisations for Gompertz and Logistic Models 78
Results, Discussions and Conclusion 85

V CONCLUSION AND SUGGESTION FOR FURTHER RESEARCH ..87
Conclusions 87

BIBLIOGRAPHY 93

APPENDICES 96
Appendix A - Computer Programs used to calculate values needed in Chapter II 97
Appendix B - Computer Programs used to calculate values needed in Chapter III and IV 103

VITA 125

LIST OF TABLES

Table **Page**

Table		Page
1	Least Squares Estimates of Gompertz and Logistic models.........	41
2	Data Set (x,y) (Ratkowsky, 1983, page 3)	45
3	Locus Point of (x_1^θ, x_2^θ)	45
4	Bates and Watts Curvature Measures	55
5	Bias from the Box 's Bias Formula	58
6	The Significance of the Standardised Third Moments	61
7	Hougaard's Direct Measure of Skewness	61
8	Derived Quantities of Mean for Gompertz Model	66
9	Derived Quantities of Mean for Logistic Model	66
10	Derived Quantities of Variance for Gompertz Model	67
11	Derived Quantities of Variance for Logistic Model	67
12	Derived Quantities of Third Moment for Gompertz Model	68
13	Derived Quantities of Third Moment for Logistic Model	68
14	Derived Quantities of Fourth Moment for Gompertz Model	68
15	Derived Quantities of Fourth Moment for Logistic Model	69
16	Tests of Nonlinearity for Gompertz Models	80
17	Tests of Nonlinearity for Logistic Models	81

LIST OF FIGURES

Figure		Page
1	A Scatter Plot of the Response Variable against the Regressor ...	31
2	A Smooth Plot of Figure 1	31
3	Sample Space Representation of Model $E(Y) = X^{\theta}$	46
4	Histogram for $\hat{\alpha}$ from the Initial Gompertz Model [2.1]	75
5	Histogram for $\hat{\beta}$ from the Initial Gompertz Model [2.1]	75
6	Histogram for $\hat{\gamma}$ from the Initial Gompertz Model [2.1]	76
7	Histogram for $\hat{\alpha}$ from the Initial Logistic Model [2.2]	76
8	Histogram of Result for $\hat{\beta}$ from the Initial Logistic Model [2.2]	77
9	Histogram for $\hat{\gamma}$ from the Initial Logistic Model [2.2]	77
10	Histogram for $\hat{\alpha}$ from the Model Function [4.1]	82
11	Histogram for $\tilde{\beta}$ from the Model Function [4.1]	82
12	Histogram for $\hat{\gamma}$ from the Model Function [4.1]	83
13	Histogram for $\hat{\alpha}$ from the Model Function [4.2]	83
14	Histogram for $\hat{\beta}$ from the model function [4.2]	84
15	Histogram for $\hat{\gamma}$ from the model function [4.2]	84

CHAPTER 1

INTRODUCTION

One of the important tasks in statistics is to find the relationships, if any, that exist in a set of variables. In data modeling, one of the variables, often being called the response or dependent variable is denoted by y. This variable normally becomes our particular interest. The other variable, which we normally call explanatory variable or independent variable or regressor, is to explain the behaviour of y, and is denoted by x.

In order to have a rough idea of some relationship between y and x, we normally do a scatter plot of y against x whereby we can express this relationship via some function x and mathematically we can write it as

$$y_t = f(x_t, \theta) + \varepsilon_t \qquad [1.1]$$

where $\theta = (\theta_1, \theta_2, ..., \theta_p)^T$ is a set of p unknown parameters and ε_t is an additive error term. If the errors ε_t (t=1,2,…,n) satisfy E(ε_t)=0 and Var(ε_t)=σ^2, then the value $\hat{\theta}$ which minimises the sum of squares of residuals

$$S(\theta) = \sum_{t=1}^{n} [y_t - f(x_t, \theta)]^2 \qquad [1.2]$$

is called the least squares estimate of θ. The model function $f(x,\theta)$ is determined by the parameter vector θ and the experimental settings x_t. Therefore a different set of parameters can yield a different model function.

If we introduce $f_t(\theta) = f(x_t,\theta)$ then the sum of squares function [1.2] can be written as

$$S(\theta) = \|y - f(\theta)\|^2 \qquad [1.3]$$

where $f(\theta) = [f_1(\theta), f_2(\theta),..., f_n(\theta)]^T$ and the double vertical bars indicate the length of a vector. The geometric interpretation of $S(\theta)$ is that, it is the square of the distance between the vector y and $f(\theta)$ in an n-dimensional sample space. If we substitute values for θ in $f(\theta)$, the $f(\theta)$ will trace a p-dimensional surface which we call as solution locus in this sample space (Bates and Watts, 1980). Therefore, the least squares estimate $\hat{\theta}$ is the parameter value such that $f(\hat{\theta})$ is the point in the solution locus closest to y.

Most algorithms for computing the least squares estimate θ are based on a local linear approximation to the model. If we take a fixed parameter value, θ^0, the model function is approximated by

$$f(x,\theta) \cong f(x,\theta^0) + \sum_{i=1}^{p}(\theta_i - \theta_i^0 \qquad [1.4]$$

where p= number of parameters.

Equivalently, [1.4] can be written as

$$f(\theta) \cong f(\theta^0) + \sum_{i=1}^{p} (\theta_i - \theta_i^0) v_i \qquad [1.5]$$

where $v_i = [v_i(x_1), v_i(x_2), \ldots, v_i(x_n)]^T = \frac{\partial f(x,\theta)}{\partial \theta_i}$ (i=1,2,...,p) is evaluated at $\theta = \theta^0$

When using linear approximation, we are to replace the solution locus by it tangent plane at $f(\theta^0)$, and at the same time to impose a uniform co-ordinate system on that tangent plane. Bates and Watts (1980 & 1988) termed these two components of the linear approximation as the planar assumption and the uniform co-ordinate assumption respectively.

The measures, which indicate the adequacy of a linear approximation, are called the measures of the nonlinearity. The very first attempt to measure nonlinearity was made by Beale in 1960 (Ratkowsky, 1983). Box also presented a formula for estimating the bias in the least square estimators which is known as the Box bias formula (Cook, 1986). Ratkowsky (1983) used simulation studies not only to predict bias to the correct order of magnitude but also to the correct extent of nonlinear behaviour of the model. Bates and Watts also developed new measures of nonlinearity but it was based on the geometric concept of curvature (Bates and Watts, 1980). To determine how planar the expectation surface is and how uniform the parameter lines are on the tangent plane, Bates and Watts used second derivatives of the expectation function or the model function $f(x_t, \theta)$ to derive curvature measures of intrinsic and parameter effect nonlinearity.

As Ratkowsky (1989) pointed out that although the bias formula introduced by Box has been used as a measure of the extent to which parameter estimates may exceed or fall short of the true parameter values, it is not an accurate measure for comparing parameters in two different parameterisations. He also noted that the percentage bias, which is obtained from the Box's bias formula, is not location-independent since it is possible to obtain a high percentage bias simply because the values of the estimates are close to zero. To overcome this problem, we therefore use a direct measure of skewness introduced by Hougaard (1985), and the curvature measures of nonlinearity introduced by Bates and Watts (1980).

In this study, we will employ all the four measures to assess nonlinearity. Our major aim is to achieve models that behave very much close to linear models. If the extent or degree of nonlinearity is considerably high, we will then do a reparameterisation on the models. Using the Box's formula, as the bias is expressed as percentage of the least square estimate of the parameter, if the absolute value of the percentage bias is in excess of 1%, this indicates that the nonlinear behaviour is readily unacceptable.

The statistical significance of intrinsic and parameter effect nonlinearity of Bates and Watts may be assessed by comparing these values with $\frac{1}{\sqrt{F}}$ where $F = F$ (p, n-p, α) is obtained from the F-distribution table corresponding to significance level α. The solution locus may be considered to be sufficiently linear over an approximate 95% confidence region if intrinsic nonlinearity is less than $\frac{1}{\sqrt{F}}$. Similarly, the projected parameter lines of θ may be assumed to be sufficiently parallel and uniformly spaced if parameter effect is less than $\frac{1}{\sqrt{F}}$.

One of the advantages of using simulation studies is that we can study the sampling properties of the least square estimators. Using the parameter estimates obtained from the simulation studies, we will calculate the first four moments of the set of estimates, namely, the sample mean m_1, the sample variance m_2, the skewness coefficient $g_1 = \frac{m_3}{m_2^{(\frac{3}{2})}}$, and the kurtosis coefficient $g_2 = (\frac{m_4}{m_2^2}) - 3$ where m_3 and m_4 are the third and the fourth sample moments about the mean respectively. Based on the above four moments, we then examine whether the estimator exhibits normal behaviour by performing a standard-normal distribution test on the moments said earlier.

The closeness of the set of simulated parameter estimates approaching to a normal distribution can also be visually assessed by examining the histogram of each parameter. In fact, the histogram will clearly illustrate whether the least square estimators having a negative or positive skewness.

To calculate the Hougaard direct measure of skewness, we need to find the estimate of the third moment of each parameter and standardise it using the appropriate element of asymptotic covariance matrix. As there is a close link between the extent of nonlinear behaviour of an estimator and the extent of nonnormality in the sampling distribution of the estimator, the standardised third moment is then used as a guide whether the estimator is close-to-linear or contains some considerable nonlinearity. If the standardised third moment of the parameter is less than 0.1, then the estimator of the parameter is said to be very close-to-linear behaviour. If it is in between 0.1 and 0.25, then the estimator is reasonably close-to-linear. However if the standardised third moment is greater than 0.25, the skewness is already very apparent. Therefore, for any standardised third moments exceeding a value of 1, this indicates that the nonlinear behaviour is already unacceptable.

As we noted, some of these measures of nonlinearity do not identify the nonlinear-behaving parameters, nor do they suggest suitable reparameterisations, hence we would rely on the histograms of the parameter estimates obtained from our simulation studies. As suggested by Ratkowsky

(1983), a histogram with a long right-hand tail characterise a lognormal distribution, so to reparameterise the model, he suggested a replacement of the parameter in the model function by the exponential of the parameter. On the other hand, a histogram with a long left-hand tail suggests a replacement of the parameter by a logarithm of the parameter. A comparison of the various measures of the nonlinearity for each parameter estimate before and after the reparameterisation, will reveal whether the reparameterisation really do improve the nonlinear models to behave closer to linear models.

Statement of the Problem

Nonlinear models are defined as models having at least one parameter appears nonlinearly, whereby at least one of their derivatives with respect to any parameters are not independent of their parameters. In the estimation properties of these models, nonlinear models differ greatly if compared to linear models. In linear models, with the assumption that the errors are independent, and identically distributed (i.i.d), they will result to having unbiased, normally distributed, and minimum variance estimators. However, nonlinear models only tend to do so as the sample sizes become very large.

In this study of nonlinear modeling, a set of experimental data representing the average weight (in grams) of dried tobacco leaves per tree against the time in week was obtained from MARDI, Serdang. The sample size of this set of data is similar to those actually obtained in practice by scientists in agricultural research. So with the given data, we are to fit the data using appropriate models and finally will choose the ones that behave very close to linear models.

This research will focus on three aspects of nonlinear modeling. The first is to estimate the least square estimators of the parameters that exist in the proposed nonlinear models. In order to estimate the parameters of the model, we will use least square method as mentioned earlier. Unfortunately, unlike a least square estimator of a parameter in linear model, a least square estimator of a parameter in a nonlinear model has unknown properties for finite sample size. Nevertheless, Ratkowsky (1983) proposed that as sample sizes increases, we might observe that the estimator will tend to become more and more unbiased, more and more normally distributed and approach a minimum possible variance. The solution for approximately minimum variance is addressed by the use of iterative numerical methods. In this study, we will employ Gauss-Newton method as it is favoured for its fast convergence characteristic, provided if we have good initial parameter estimates or starting values.

The second part of the study is to measure the nonlinearity in the parameters. Various widely used methods will be incorporated; the Box bias formula, the Bates and Watts curvature measure, the Ratkowsky simulation studies and the Hougaard direct measure of skewness. A comparison for the various nonlinear measures for the specified nonlinear models and parameters will be done. This is then will be used as a guideline for making reparameterisations. Our final part of the research is to reparameterise the initial or basic model. One specified model can have different parameterisations, by which it is meant that the parameters of the new parameterisations or reparameterisations are related to the old parameterisations by an expression that involves parameters only (Ratkowsky,1983). Bates and Watts (1980) refer to various reparameterisations of the same specified nonlinear model as model functions and in this study we will adopt the same terminology. As Ratkowsky (1983) suggested it that reparameterisations would actually do improve the model function to behave closer-to-linear model provided that the intrinsic nonlinearity is always less than the parameter effect.

Some Key Words and Definition

For the sake of completeness, a brief review of some important concepts that were used frequently in this study is given below. Some important key words include relative curvature measures, parameter effect, intrinsic nonlinearity, linear approximation and solution locus.