Asian Journal of Chemistry Vol. 21, No. 9 (2009), 6904-6908

Interdisciplinary Implementations of Some Chemical Statistical Schemes

ERKUT AKKARTAL
Department of Econometrics, Naval Science and Engineering Institute, 34940 Istanbul, Turkey
E-mail: ilginkur@istanbul.edu.tr

The aim of the present investigation is to study an interdisciplinary application on chemical, biometrical and econometrical models. Regression analysis and least square methods used in some statistical processes and/or econometrical studies are sometimes similar with the methods used in biochemistry and ship design. Especially, as far as a biometrical data set including structural change and designing the board side of the ship is concerned, some smoothing techniques are used, like the ones in used econometrics. In this study the common ways of statistical scheme and ship designing and biochemistry models are evaluated and some similar processes are emphasized.

Key Words: Polynomial functions, Biochemistry-biostatistics, Kinks.

INTRODUCTION

The regression analysis which is most commonly used in stochastic modeling, as the structural change is proved in the data set being studied then it is called piece-wise Linear regression. When the classical Chow test or dummy variable application is applied to the data and resulted that there is a structural change in the data set, it is concluded that the model can not be expressed as a single regression model. Probably any cubic polynomial function may give a proper result[1].

In piece-wise linear regression and in the knots where the regression lines are joint, the continuity problem of the function appears. In order to make the function continuous at knots, some transformations are made and added to the function. So the function becomes continuous at these knots. The transformation results made by the aim of polynomial functions, gets various names in literature.

EXPERIMENTAL

Piece-wise linear models suffer from two obvious shortcomings. First, in spite of the function is continuous, its derivations are not. This discontinuity of derivatives at the kinks can prove some disadvantages in many biochemical and/or econometrical applications where the result would be discontinuous (and probably spurious) shifts in elasticity's, marginal or other relationships that would becloud analysis[2].

On the other hand, a curvilinear relationship may provide a significantly better fit to the data than is obtained from linear segments. This result is considerably important when confronted by a complicated curve without obvious critical positions to which linear segments would be fitted. If the x values are to be located arbitrarily\we had better not rely on linear approximation to map out the function between them. Cubic functions overcome this disadvantages by replacing the linear approximation of such system of piece-wise polynomial approximation. Any degree of polynomial could be employed\but the cubic is the most convenient for most purposes[3].

The x-axis has been divided into three segments by the points X_0, X_1, X_2 and X_3. In cubic function theory, the points chosen are called knots. This is as good a term as any and will be employed hereafter. For convenience, the intervals between the knots have been taken as equal. This as not the important part of the scheme, but equality of intervals is generally advisable unless there is important reason to do otherwise. More than four knots and correspondingly more than three intervals can be used but, as will be seen, the more intervals there are, the greater the number of composite variables required to fit the curve and the greater the loss of degrees of freedom.

Suppose a biochemical model is defined as $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ with n + 1 offset points. The purpose is to find the coefficients of the polynomial which will represent the function in question:

$$y_m(x_i) = a_0 + a_0x + a_2x_2 + \ldots + a_mx^m$$

If the degree of the polynomial is taken to be equal with the number of offset points given (m = n), the least squares polynomial will be equal to the interpolation polynomial and the resultant polynomial will pass through all offset points. In other words, m value must be selected as smaller than the n value. In this case, the polynomial (m < n), couldn't pass through the offset points, the most approximated polynomial is searched. For this purpose, the most commonly used method is Least Squared method[4].

The difference between the given offset points and the approximated function is defined as (δ_i):

$$\delta_i = y_m(x_i) - y_i$$

Since the principal of the least squares method is to reduce the total error to minimum, the expression of the error term is defined as:

$$E = \sum_{i=0}^{n} |\delta_i|^2 = \sum_{i=0}^{n} [y_m(x_i) - y_i]^2$$

This error term is tried to be minimum. Where y_i are the offset values given and y_mx_i are the minimum values found by least squares polynomial. If the polynomial is substituted in the equation:

$$E = \sum_{i=0}^{n}(a_0 + a_1x + a_2x^2 + ... + a_mx^m - y_i)^2 = \sum_{i=0}^{n}\left\{\sum_{j=0}^{m} a_jx_i^jy_i\right\}^2$$

The condition of being the minimum for this equation is being the derivative zero with respect to the coefficients:

$$\frac{\partial E}{\partial a_m} = 0 \rightarrow \frac{\partial E}{\partial a_0} = 0,\ \frac{\partial E}{\partial a_1} = 0,\ \frac{\partial E}{\partial a_2} = 0,\ \cdots\cdots\ \frac{\partial E}{\partial a_m} = 0$$

These equations can be written as:

$$\frac{\partial E}{\partial a_0} = \sum_{i=0}^{n} 2(a_0 + a_1x + a_2x^2 + ... + a_mx^m - y_i) = 0$$

$$\frac{\partial E}{\partial a_1} = \sum_{i=0}^{n} 2(a_0 + a_1x + a_2x^2 + ... + a_mx^m - y_i)x = 0$$

$$\frac{\partial E}{\partial a_2} = \sum_{i=0}^{n} 2(a_0 + a_1x + a_2x^2 + ... + a_mx^m - y_i)x^2 = 0$$

$$\frac{\partial E}{\partial a_m} = \sum_{i=0}^{n} 2(a_0 + a_1x + a_2x^2 + ... + a_mx^m - y_i)x^m = 0$$

When these equations are reorganized, the linear equations era found as below:

$$\begin{array}{llllll}
\sum_{i=0}^{n} a_0 & + \sum_{i=0}^{n} a_1x_i & + \sum_{i=0}^{n} a_2x_i^2 & + ... + \sum_{i=0}^{n} a_mx_i^m & = \sum_{i=0}^{n} y_i \\
\sum_{i=0}^{n} a_0x_i & + \sum_{i=0}^{n} a_1x_i^2 & + \sum_{i=0}^{n} a_2x_i^3 & + ... + \sum_{i=0}^{n} a_mx_i^{m+1} & = \sum_{i=0}^{n} y_ix_i \\
\sum_{i=0}^{n} a_0x_i{}^2 & + \sum_{i=0}^{n} a_1x_i^3 & + \sum_{i=0}^{n} a_2x_i^4 & + ... + \sum_{i=0}^{n} a_mx_i^{m+2} & = \sum_{i=0}^{n} y_ix_i{}^2 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\sum_{i=0}^{n} a_0x_i{}^m & + \sum_{i=0}^{n} a_1x_i^{m+1} & + \sum_{i=0}^{n} a_2x_i^{m+2} & + ... + \sum_{i=0}^{n} a_mx_i^{m+m} & = \sum_{i=0}^{n} y_ix_i{}^m
\end{array}$$

When the simplest m = 1 linear fitting is taken to be sample, the least squares polynomial can appear as:

$$y_m(x_i) = a_0 + a_1x$$

In order to solve the unknown coefficients of ai the two conditions are used given below:

$$\frac{\partial E}{\partial a_0}=0 \qquad \frac{\partial E}{\partial a_1}=0$$

And the last equation that will be solved can be written as below:

$$a_0(n+1)+a_1\sum_{i=0}^{n}x_i=\sum_{i=0}^{n}y_i$$

$$a_0\sum_{i=0}^{n}x_i \;+a_1\sum_{i=0}^{n}x_i{}^2=\sum_{i=0}^{n}y_ix_i$$

By solving these equations, there is a unique ai coefficient which makes the error term minimum. For several points, the standard equation solving methods can not generate a solution or in case of the linear equations being not stable, inaccurate results may be faced. Therefore it can not be said that this procedure is successful and practical for the curves which defined with multipoint[5,6].

RESULTS AND DISCUSSION

It is now proposed to fit a regression in the form:

$$\begin{aligned}Y=&[a_1+b_1(x-x_0)+c_1(x-x_0)^2+d_1(x-x_0)^3]D_1+[a_2+b_2(x-x_1)+\\&c_2(x-x_1)^2+d_2(x-x_1)^3]D_2+[a_3+b_3(x-x_2)+c_3(x-x_2)^2+\\&d_3(x-x_2)^3]D_3+u\end{aligned}\tag{1}$$

Of course (1) is discontinuous at the knots, as are its derivatives. But application of appropriate constraints to the coefficients not only makes the continuous, but guarantees continuity of its first and second derivatives. The constraints required for this purpose are,

$$\begin{aligned}a_2&=a_1+b_1(x_1-x_0)+c_1(x_1-x_0)^2+d_1(x_1+x_0)^3\\b_2&=b_1+2c_1(x_1-x_0)+3d_1(x_1-x_0)^2\\c_2&=c_1+3d_1(x_1-x_0)\\a_3&=a_2+b_2(x_2-x_1)+c_2(x_2-x_1)^2+d_2(x_2-x_1)^3\\b_3&=b_2+2c_2(x_2-x_1)+3d_2(x_2-x_1)^2\\c_3&=c_2+3d_2(x_2-x_1)\end{aligned}\tag{2}$$

The constraints on a_i, makes the left and right values of the function equal on kinks. As for the constraints on b_i, makes the left and right slope (first derivative)equal and c_i makes the same for 2nd derivative

The regression procedure itself is readily carried out by any standard least-squares regression package. Moreover, goodness of fit, significance test and other related statistics for the cubic functions are those obtained in the usual fashion from the multiple regression program.

Selection of the Place of Kinks is important as well. In general minimizing the least squares error sum of squares by varying the knots often results in finding local minima which in the mathematics of the problem require the algorithms to be iterative, rather than analytical in nature. Thus, even when the global minimum can be found in a finite number of steps, the number of iterations may still be large. Also different initial knot locations often converge to different local minima. Furthermore, the knots often tend to coalesce, indicating discontinuities in other than the 'nth' derivative of the 'nth' degree function.

It has been the purpose of this article to demonstrate how cubic functions can be fitted to the models which concern biochemical and to some other disciplines. Although examples have been limited to standard regression functions fitted to only three intervals, the same principles apply to any desired number. Although it is a great effort the biometric cubic functions have important limitations. These functions are most useful when data uniformly distributed throughout the observed range.

ACKNOWLEDGEMENT

The author is thankful to Dr. Ilgin Kursun, for providing valuable guidance in the study.

REFERENCES

1. J.F. Durand and S. Robert, *J. Am. Statistical Assoc.*, **92**, 440 (1997).
2. D.J. Poirier, *J. Econometrics*, **3**, 24 (1975).
3. M.J.D. Powell, *J. Numerical Anal.*, **6**, 58 (1969).
4. E. Sariöz, Computer Aided Design and Munufacturing, Text, Istanbul Technical University (2004).
5. D.B. Suits, Principles of Economics, New York, Harper and Row (1973).
6. E. Akkartal and I. Kursun, Kübik Fonksiyonlar Ile Maden Ekonomisi Verilerinin Analizi, 4th Istatistik Günleri Sempozyumu, Dokuz Eylül Üniversitesi, Izmir (2004).

(*Received*: 27 October 2008; *Accepted*: 5 August 2009) AJC-7708