GLOBAL JOURNAL OF MATHEMATICAL SCIENNCES VOL. 6, N0. 2, 2007: 115 - 118
 ISSN 1596-6208

ORTHOGONAL POLYNOMIAL REGRESSION, A FINITE DIFFERENCE APPROACH

M. E. NJA

(Received 12 September 2002; Revision Accepted 2 November 2007)

ABSTRACT

Literature reveals very complicated methods of obtaining orthogonal polynomials. The methods involve the use of tables, performance of F-tests, obtaining regression parameters among others. This paper presents a finite difference approach which removes these complications as it offers a much simpler method. These approaches are compared using a hypothetical data.

KEY WORDS: Orthogonal polynomials, finite difference, controlled variable, step size, regression parameters.

1. INTRODUCTION

Orthogonal polynomials are coefficients of orthogonal vectors used in curvilinear regression analysis (Kerlinger et al, 1973). Under the assumption that the values of the controlled variable x are equally spaced, with an equal number of observations at each point, orthogonal polynomial regression is employed to fit the regression equation (Chatfield, 1983). The existing method involves the use of tables, performance of F-tests etc. The process of determining the order of the polynomial in this method is doubtless labourious and complicated. It is for this reason that the need for an alternative method of fitting orthogonal polynomials has arisen. The method of finite differences is employed in this paper to establish that these polynomials can be fit using polynomial approximation in the sense of Kreyszic (1999). After presenting the two procedures, we use an illustrative example to demonstrate that the proposed method is computationally less hectic.

2. EXISTING METHOD

A polynomial regression of degree k is of the form (after standardization)

$$y = a_0' + a_1' f_1(z) + ... + a_k' f_k(z), \qquad z = \frac{x - \bar{x}}{d}$$

where $f_r(z)$ is a polynomial in z of degree r, the $a_i's$ are regression parameters and d is the distance between successive values of x. $\bar{x}$ is the mean of x values (Chatfield, 1983).

Definition 1

Two polynomials $f_r(x_i)$ and $f_s(x_i)$ are said to be orthogonal if

$$\sum_{i=1}^{n} f_r(x_i) f_s(x_i) = 0 \qquad (r \neq s)$$

for all r, s < n-1 (Draper and Smith, 1981)

Definition 2

If the orthogonal polynomials $f_0(x), f_1(x), ... f_p(x)$ are constructed such that

$f_0(x_i) = 1$ zero-order polynomial

$f_1(x_i) = P_1 X_i + Q_1$ first order

$f_2(x_i) = P_2 X_i^2 + Q_2 X_i + R_2$ second order

: : :

: : :

: : :

$f_r(x_i) = P_r X_i^r + Q_r X_i^{r-1} + ... + T_r$ r – th order

then the model

$$y = \alpha_0 f_0(x) + \alpha_1 f_1(x) + ... + \alpha_p f_p(x) + \varepsilon,$$

where ε is the random error component is called orthogonal polynomial regression model (Draper and smith, 1981).

Remark

Orthogonality in the sense of two separate families of polynomials, F_1 and F_2 is achieved when

$$\left(\frac{dy}{dx}\right)_1 = -\left(\frac{dy}{dx}\right)^{-1}_2$$

where $\left(\frac{dy}{dx}\right)_1$ is the slope of a polynomial in F_1

and $\left(\frac{dy}{dx}\right)_2$ is the slope of a polynomial in F_2. (Kreyszic, 1999)

Where this happens we say that F_1 is a family of orthogonal trajectories to F_2 and vice versa.

The standardized orthogonal polynomials are tabulated in Fisher and Yates (1963) and Pearson and Hartley (1966). The normal equations from where the regression coefficients are estimated are as follows (Draper and Smith, 1981, Chatfield, 1983).

$na_0 = \Sigma y.$

$$a_1' \sum f_1(z_i)^2 = \sum f_1(x_i) y_i$$

$$a_k' \sum f_k(z_i)^2 = \sum f_k(x_i) y_i$$

The quantities $f_i(z_i)y_i$ are given in the above tables.

The process of performing an F-test to test the adequacy of the fit consists in obtaining two non-

M. E. Nja, Department of Mathematics/Statistics, Cross River University of Technology, Calabar, Nigeria

significant ratios (in a row) (Chatfield, 1983)

$$F_1 = \frac{a'^2_k \sum f_k(z)^2}{R_k/(n-k-1)}$$

and $$F_2 = \frac{a'^2_{k+1} \sum f_{k+1}(z)^2}{R_{k+1}/(n-k-2)}$$

$a_k \sum f_k(z)^2$ is obtained from the relation

$$R_k = \sum (y - a_0 - a'_1 f_1(z) - \ldots - a_k f_k(z))^2$$

$$= \sum (y-\bar{y})^2 - a'^2_1 \sum f_1(z)^{2k} - \ldots - a'_k \sum f_k(z)^2$$

R_k is residual sum of squares after fitting a polynomial of degree k. If a polynomial of degree k-1 fits the data better than a polynomial of degree k-2 then R_{k-1} will be substantially less than R_{k-2}.

3. PROPOSED METHOD (FINITE DIFFERENT METHOD)

In numerical methods, finite differences are employed in polynomial approximation of functions (Kreyszic, 1999). Both the degree of the polynomial and its coefficients are obtained using the following theorem:

Theorem (Kreyszic, 1999)

For a polynomial, $P_n(x) = a_n + a_{n-1}x + \ldots + a_n x^n$ of degree n, the nth differences in a table with step h are approximately constant ($c = n!h^n a_n$) and all other higher differences are zero.

This however depends on the nature of approximation. For crude approximations, the column may not be a constant but of the shortest range. The range of a difference column is the difference between the largest and the smallest values. With this theorem we can obtain a polynomial given a set of data. For the degree of the polynomial it is even easier. If the column of constants appears in the first column of difference we have a polynomial of the first degree. If it appears in the second, we have a polynomial of the second degree etc.

From $n!h^n a_n = c$ (c a constant or mean of the column of shortest range) we obtain $a_n = \frac{c}{n!h^n}$. Using the value of a_n and the table we obtain a system of equations which we solve for the other coefficients of the n-1 degree polynomial. This system is given as

$$P_n(x_1) - \frac{c}{n!h^n} = a_{n-1}x_1 + \ldots + a_1 x_1^{n-1}$$

$$P_n(x_2) - \frac{c}{n!h^n} = a_{n-1}x_2 + \ldots + a_1 x_2^{n-1}$$

$$\vdots$$

$$P_n(x_m) - \frac{c}{n!h^n} = a_{n-1}x_m + \ldots + a_1 x_m^{n-1}$$

where c is constant or mean of column of shortest range, h is step size.

The two methods of solutions are now compared for the example below:

4. ILLUSTRATIVE EXAMPLE

Find the polynomial of the lowest degree which adequately describes the following hypothetical data in which x is the controlled variable.

x	0	1	2	3	4	5	6
y	6.3	5.7	6.3	7.3	9.9	12.5	18.1

Solution

Normal Method (Chatfield, 1983): Orthogonal Polynomial Approach

The standardized controlled variable is given by

$$z = \frac{x - \bar{x}}{d} = x - 3 \text{ as } d = 1$$

where $\bar{x}$ is as earlier defined.

There are n = 7 values of the controlled variable. From Pearson and Hartley (1966), the first four orthogonal polynomials are the following:

Z	$f_1(z)$	$f_2(z)$	$f_3(z)$	$f_4(z)$
-3	-3	5	-1	3
-2	-2	0	1	-7
-1	-1	-3	1	1
0	0	-4	0	6
+1	+1	-3	-1	1
+2	+3	0	-1	-7
+3	+3	5	1	3
$\sum_i f(z)^2$	**28**	**84**	**6**	**154**

Then we find

$$\sum_{i=1}^{7} f_1(z_i) y_i = 52.6$$

$$\sum_{i=1} f_2(z) y_i = 44.2$$

$$\sum_{i=1} f_3(z) y_i = 1.4$$

$$\sum_{i=1} f_4(z) y_i = 5.8$$

$$\sum_{i=1} y_i = 66.1$$

Thus we have

$$a'_0 = \bar{y} = 9.44, \quad a'_1 = \frac{52.6}{28} = 1.878, \quad a'_2 = \frac{44.2}{84} = 0.526, \quad a'_3 = \frac{1.4}{6} = 0.233, \quad a'_4 = \frac{5.8}{154} = 0.037$$

The next stage is to compute a series of F-ratios to see how many of these parameters are required. At each stage two mean squares are obtained by dividing the appropriate sum of squares by the appropriate number of degrees of freedom (Montgomery, 1978). The ratio of the mean squares, the F-ratio, is then compared with 0.05, 1, N-k-l, where N is number of x values, k is degree of polynomial.

Table

Type of variation	Sum of squares	d.f	M.S.	F-ratio	$F_{5\%}(1,n-k-1)$
Residual from mean	122.86	6			
Explained by linear	98.72	1	98.72	20.5	6.6
Residual from linear	24.14	5	4.82		
Explained by quadratic	23.24	1	23.24	105.6	7.7
Residual from quadratic	0.09	4	0.22		
Explained by cubic	0.32	1	0.32	1.68	10.1
Residual from cubic	0.58	3	0.19		
Explained by quartic	0.21	1	0.21	1.1	18.5
Residual from quartic	0.37	2	0.19		

Neither the cubic nor the quartic terms give a significant F-ratio. In any case the residual sum of squares is so small after fitting linear and quadratic terms that it is really unnecessary to try higher order terms. Thus, a quadratic polynomial describes the data adequately.

The next stage is to formuiate the regression equation in terms of the original controlled variable, x. For this we need to know the orthogonal polynomials as functions of z. These are also given in Fisher and Yates (1963) and Pearson and Hartley (1966). We find

$f_1(z) = \lambda_1 z \quad with \quad \lambda_1 = 1$

$and\ f_2(z) = \lambda_2 (z^2 - 4) \quad with \quad \lambda_2 = 1$

Thus the estimated regression equation is given by

$$y' = 9.44 + 1.88 + 0.53(z^2 - 4)$$

$$= 9.44 + 1.88(x-3) + 0.53[(x-3)^2 - 4] \quad \text{since } z = x-3$$

$$= 6.45 - 1.30x + 0.53x^2$$

$$y' = a_0' + a_1' f_1(z) + a_2' f_2(z)$$

New Method: Finite Difference Approach

Finite Difference Column

X	y	1st	2nd	3rd	4th
0	6.3				
		-6			
1	5.7		12		
		6		-8	
2	6.3		4		20
		10		12	
3	7.3		16		-28
		26		-16	
4	9.9		0		46
		26		30	
5	12.5		30		
		56			
6	18.1				

The second column of difference has the shortest range This suggests that the polynomial is of degree 2

$P_2(x) = a_2 + a_1 x + a_0 x^2$

Thus, from n!h^{n}a$_n$ = c

$$c = \frac{12 + 4 + 16 + 0 + 30}{5} = \frac{62}{5} = 12.4$$

$2a_2.1^2 = c = 12.4,$

$\therefore a_2 = c/2 = 6.2$

$p_2(x) = 6.2 + a_1 x + a_0 x^2$

$p_2(1) = 6.2 + a_1 + a_0 = 5.7$

$p_2(2) = 6.2 + 2a + a_0 2^2 = 6.2 + 2a_1 + 4a_0 = 6.3$

Solving we have

$a_0 + a_1 = 0.5$

$2a_0 + a_1 = 0.05$

We have $a_0 = 0.55,\ a_1 = -1.05.$

Hence the estimated regression equation is

$P_2(x) = y = 6.2 - 1.05x + 0.55x^2$

which nearly approximates the orthogonal polynomial. The slight variation may be explained by the fact that the polynomial is only an approximation of the real function describing the distribution

5. CONCLUSION

With regression analysis we can make predications in terms of the future or in terms of functional values. We can also find out the level of contribution of each independent variable on the response variable. The finite difference approach is recommended under the circumstances requiring the use of orthogonal polynomials.

REFERENCES

Chatfield, C., 1983. Statistics for Technology, Chapman and Hall. London.

Drapper, N.R. and Smith, H. 1981. Applied Regression Analysis. John Wiley & Sons, Inc. New York.

Fisher, R. A. and Yates, F., 1963. Statistical Tables for Biological, Agricultural and Medical Research (6^{th} ed.). Hafner. New York.

Kerlinger, F. N. and Pedhazur, E. J., 1973. Multiple Regression in Behavioural Research. Holt, Rinehart and Winston Inc. New York.

Kreyszic, E., 1999. Advance Engineering Mathematics. John Wiley & Sons. New York.

Montgomery, D. C., 1978. Design and Analysis of Experiments. John Wiley & Sons New York.

Pearson, E. S. and Hartley, H. O., 1966. Biometrika Tables for Statisticians, 3^{rd} ed. Cambridge University Press. Cambridge