International Statistical Review, 47 (1979) 31–36 Longman Group Limited/Printed in Great Britain

Orthogonal Polynomial Regression

Sabhash C. Narula
Rensselaer Polytechnic Institute, Troy, New York 12181, U.S.A.

Summary

We discuss in basic terms the orthogonal polynomial regression approach for curve fitting when the independent variable occurs at unequal intervals and is observed with unequal frequency. The computations required for determining orthogonal polynomials are described with a simple example.

Introduction

Ordinary polynomial regression analysis is often used for curve fitting. Generally the investigator does not know the degree of the polynomial that adequately describes the data. Thus, besides estimating the unknown parameters of the polynomial, the investigator has to determine the degree of the polynomial, m, too. It is necessary to select m so that the model is complex enough to adequately describe the data yet simple enough to be useful in an applied situation. As the degree of polynomial increases, the polynomial fits the data more closely. However, the resulting polynomial is not very useful for prediction or interpolation/extrapolation purposes. This important feature of polynomial regression is well documented by examples in Dutka and Ewens (1971) and Hahn (1977).

However, two approaches are actually available for curve fitting, viz. the ordinary polynomial approach and the orthogonal polynomial approach. Both approaches have been well studied and documented, and without exception, the researchers have shown theoretically and empirically that the orthogonal polynomial approach is superior to the ordinary polynomial approach in terms of accuracy and computational effort required (see Bright and Dawkins, 1965, Dutka and Ewens, 1971, etc.). Some differences in the two approaches are stated in Table 1.

Table 1

Comparison of ordinary polynomial and orthogonal polynomial approach

	Approach	
	Ordinary polynomial	Orthogonal polynomial
1. Tests of significance on parameters are	not independent	independent
2. Estimates of the parameters	depend on the degree of polynomial	do not depend on the degree of polynomial
3. Round-off errors	produce inaccurate results	do not pose any problem
4. Computing time requirements		
(i) degree of the fit known in advance	for *low-order* fits, the method is as quick as the orthogonal polynomial	for *high-order* fit, the approach is twice as fast as the ordinary polynomial approach
(ii) degree of the fit not known in advance	takes more computer time	takes less computer time.

Though superior to ordinary polynomial regression, the orthogonal polynomial approach is not well understood and at times misunderstood. One reason might be that the only sets of tables available are for the situation when the independent variable is observed at equal intervals with equal frequency (e.g. Fisher and Yates, 1948, Delury, 1950 and Pearson and Hartley, 1958). Such tables are impractical when values of the independent variables are unequally spaced and/or are observed with unequal frequency – which is usually the case with unplanned experiments. The general problem of unequal spacing and unequal frequencies has been considered by Guest (1950) and Wishart and Metakides (1953) and the problem of unequal intervals but equal frequency by Forsythe (1957), Grandage (1958), Robson (1959), Bright and Dawkins (1965) and Dutka and Ewens (1971).

In the present article, we describe in simple terms the problem of determining orthogonal polynomials when the values of an independent variable are unequally spaced and occur with unequal frequency. We also give a simple recursive procedure to determine orthogonal polynomials.

Problem Statement

Let $X_1, X_2, \ldots, X_k$ represent k values of an independent variable and $n_1, n_2, \ldots, n_k$ the corresponding number of observations. Also, let $Y_1, Y_2, \ldots, Y_k$ denote the response variable totals. Then the usual polynomial regression equation

$$Y_i = b_0^* + b_1^* X_i + b_2^* X_i^2 + \ldots + b_m^* X_i^m, \quad i = 1, \ldots, k > m$$

may be expressed in the form

$$Y_i = b_0 + b_1 \xi_{1i} + b_2 \xi_{2i} + \ldots + b_m \xi_{mi}, \quad i = 1, \ldots, k > m, \tag{1}$$

where

$$\xi_{ri} = a_{r,r} + a_{r,r-1} X_i + \ldots + a_{r,1} X_i^{r-1} + X_i^r \quad r = 1, \ldots, m < k$$
$$i = 1, \ldots, k, \tag{2}$$

is a polynomial of degree r in X_i. The $\xi_r = (\xi_{ri})$ represents the coefficients of the rth order effect. Further, ξ_r, $r = 1, \ldots, m$ represent orthogonal coefficients, i.e. they satisfy the following relationships

$$\sum_i \xi_{ri} n_i = 0, \quad r = 1, \ldots, m \tag{3}$$

and

$$\sum_i \xi_{ri} \xi_{si} n_i = 0, \quad r \neq s = 1, \ldots, m, \tag{4}$$

$$\left(\text{note } \sum_i \text{ denotes } \sum_{i=1}^{k}\right).$$

The objective is to determine ξ_r, $r = 1, \ldots, m$ when the X's are unequally spaced and occur with unequal frequency.

Basic Calculations: An Example

The cost of the maintenance (Y) of tractors increases with the age of the tractor (X'). Data are available for 10 tractors: 3 tractors 4.0 years old, 3 tractors 4.5 years old, 2 tractors 5.0 years old and 2 tractors 6.0 years old. Thus we have four values of the independent variable X_i $(= (X_i' - 4.0)/2)$ 0, 1, 2, 4 with corresponding number of observations 3, 3, 2, 2, respectively. We describe the procedure to obtain orthogonal polynomials using these data.

To obtain ξ_r, first solve for $a_{r,\,j}$'s in (2) recursively using the fundamental properties (3) and (4) of the orthogonal polynomials. Then substitute these values of $a_{r,\,j}$'s in (2) and obtain ξ_{rj}.

Orthogonal Coefficients for Linear Regression. Let $\xi_{1i} = a_{11}+X_i$ denote the coefficients for linear regression. To determine ξ_{1i}'s proceed as follows:

As shown in Table 2, write the values of X_i in column (1) and the values of corresponding n_i in column (2). Substitute the values of X_i successively in ξ_{1i} and record the results in column (3). Multiply n_i of column (2) with the corresponding ξ_{1i} in column (3) and record the results in column (4). One of the fundamental properties of ξ_1 is that

$$\sum_i \xi_{1i}n_i = 0,$$

where the summation is over all levels. This leads to $a_{11} = -3/2$. Substituting this value of a_{11} in column (3) gives the ξ_{1i} values of column (5). To obtain the results in simple integers, ξ_{1i} of column (6), one can divide the results in column (5) by the common factor, 1/2.

Table 2

Procedure for orthogonal coefficients of linear regression

X_i	n_i	$\xi_{1i} = a_{11}+X_i$	$\xi_{1i}n_i$	ξ_{1i}	ξ'_{1i}
(1)	(2)	(3)	(4)	(5)	(6)
0	3	a_{11}	$3a_{11}$	$-3/2$	-3
1	3	$a_{11}+1$	$3a_{11}+3$	$-1/2$	-1
2	2	$a_{11}+2$	$2a_{11}+4$	$1/2$	1
4	2	$a_{11}+4$	$2a_{11}+8$	$5/2$	5
Sum			$10a_{11}+15 = 0$		
$a_{11} = 3/2$					

Orthogonal Coefficients for Quadratic Regression. Let $\xi_{2i} = a_{22}+a_{21}X_i+X_i^2$ denote the coefficients for the quadratic regression. To determine ξ_{2i}'s we proceed as follows:

Substitute the values of X_i in ξ_{2i} successively and write them in column (3) of Table 3. Column (4) is obtained by multiplying n_i of column (2) with corresponding ξ_{2i} in column (3). The sum,

$$\sum_i \xi_{2i}n_i,$$

must be zero as before. Further, the sum of the products $\xi_{1i}\xi_{2i}n_i$ (or $\xi'_{1i}\xi_{2i}n_i$) must also be zero. Write ξ'_{1i} in column (5). Column (6) is obtained by multiplying column (4) and column

Table 3

Procedure for orthogonal coefficients of quadratic regression

X_i	n_i	$\xi_{2i} = a_{22}+a_{21}X_i+X_i^2$	$\xi_{2i}n_i$	ξ'_{1i}	$\xi'_{1i}\xi_{2i}n_i$	ξ_{2i}	ξ'_{2i}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0	3	a_{22}	$3a_{22}$	-3	$-9a_{22}$	356/205	178
1	3	$a_{22}+a_{21}+1$	$3a_{22}+3a_{21}+3$	-1	$-3a_{22}-3a_{21}-3$	$-264/205$	-132
2	2	$a_{22}+2a_{21}+4$	$2a_{22}+4a_{21}+8$	1	$2a_{22}+4a_{21}+8$	$-474/205$	-237
4	2	$a_{22}+4a_{21}+16$	$2a_{22}+8a_{21}+32$	5	$10a_{22}+40a_{21}+160$	336/205	168
Sum			$10a_{22}+15a_{21}+43 = 0$		$41a_{21}+165 = 0$		
$a_{21} = -165/41$							
$a_{22} = \{-43+15(165/41)\}/10 = 356/205$							

(5). Because the total of column (6) equals zero, $a_{21} = -165/41$. This value of a_{21} when substituted in the sum of column (4), gives $a_{22} = 356/205$. Using these values of a_{22} and a_{21} in the ξ_{2i} of column (3) one obtains results shown in column (7). Dividing column (7) by the common factor 2/205 gives ξ'_{2i} of column (8).

Orthogonal Coefficients for Cubic Regression. The orthogonal coefficients,

$$\xi_{3i} = a_{33}+a_{32}X_i+a_{31}X_i^2+X_i^3,$$

of cubic regression can be obtained in a manner similar to obtaining ξ_{2i}. Now, there are three relationships:

$$\sum_i \xi_{3i}n_i = 0, \quad \sum_i \xi_{1i}\xi_{3i}n_i = 0, \quad \text{and} \quad \sum_i \xi_{2i}\xi_{3i}n_i = 0.$$

For this problem, the three equations are:

$$\begin{aligned} 147+43a_{31}+15a_{32}+10a_{33} &= 0 \\ 653+165a_{31}+41a_{32} &= 0 \\ 17316+3084a_{31} &= 0. \end{aligned}$$

Solving:

$$a_{31} = -1443/257, \quad a_{32} = 70274/10537, \quad a_{33} = -5904/10537.$$

Substituting these values in the ξ_{3i}:

$$\xi_{31} = -5904/10537, \quad \xi_{32} = 15744/10537, \quad \xi_{33} = -17712/10537,$$

$$\xi_{34} = 2952/10537.$$

Multiplying by 10537/984:

$$\xi'_{31} = -6, \quad \xi'_{32} = 16, \quad \xi'_{33} = -18, \quad \xi'_{34} = 3.$$

A Simplified Procedure

Although the foregoing procedure is straightforward, calculations become tedious very quickly. An alternative procedure eliminates the need to calculate the $a_{r,j}$'s to obtain ξ_{rj}. Using essentially the argument of Fisher (1952) for equally spaced X's, one can directly obtain

$$\xi_{rj} = X^r - \sum_{p=0}^{r-1} \xi_{pj} \left\{\left(\sum_i \xi_{pi}X_i^r n_i\right)\Big/\left(\sum_i \xi_{pi}X_i^p n_i\right)\right\}, \quad r = 1, \ldots, m \qquad j = 1, \ldots, k \tag{5}$$

where

$$\xi_{0j} = 1, \quad j = 1, \ldots, k.$$

We use (5) to obtain the ξ_{rj} values for the data considered previously.

To calculate coefficients for linear contrast, $r = 1$, and from (5)

$$\xi_{1j} = X_j - \left(\sum_i X_i n_i\right)\Big/\left(\sum_i n_i\right) \quad j = 1, 2, 3, 4.$$

The calculations for linear contrasts are shown in Table 4.

Table 4

Calculations for coefficients of linear contrast

X_j	n_j	X_jn_j	$X_j - (\sum_i X_in_i)/(\sum_i n_i)$	$= \xi_{1j}$
0	3	0	0–15/10	$= -3/2$
1	3	3	1–15/10	$= -1/2$
2	2	4	2–15/10	$= +1/2$
4	2	8	4–15/10	$= +5/2$
Total	10	15		

To calculate coefficients for a quadratic contrast, $r = 2$, and from (5)

$$\xi_{2j} = X_j^2 - \xi_{1j}\left(\sum_i \xi_{1i}X_i^2 n_i\right)\Big/\left(\sum_i \xi_{1i}X_i n_i\right) - \left(\sum_i X_i^2 n_i\right)\Big/\left(\sum_i n_i\right) \quad j = 1, 2, 3, 4.$$

The calculations are shown in Table 5.

Table 5

Calculations for coefficients of quadratic contrast

X_j	n_j	$n_jX_j^2$	$\xi_{1j}X_j^2n_j$	$\xi_{1j}X_j^2n_j$	$X_j^2 - \xi_{1j}\dfrac{\sum_i \xi_{1i}X_i^2n_i}{\sum_i \xi_{1i}X_in_i} - \dfrac{\sum_i X_i^2n_i}{\sum_i n_i} = \xi_{2j}$
0	3	0	0	0	0–(−3/2)(165/41)–43/10 = 356/205
1	3	3	−3/2	−3/2	1–(−1/2)(165/41)–43/10 = −264/205
2	2	8	2	4	4–(1/2)(165/41)–43/10 = −474/205
4	2	32	20	80	16–(5/2)(165/41)–43/10 = 336/205
Total	10	43	41/2	165/2	

From equation (5), the coefficients for the cubic effect are obtained by letting $r = 3$

$$\xi_{3j} = X_j^3 - \xi_{2j}\frac{\sum_i \xi_{2i}X_i^3n_i}{\sum_i \xi_{2i}X_i^2n_i} - \xi_{1j}\frac{\sum_i \xi_{1i}X_i^3n_i}{\sum_i \xi_{1i}X_in_i} - \frac{\sum_i X_i^3n_i}{\sum_i n_i} \quad j = 1, 2, 3, 4.$$

Using this result it can be verified that

$$\xi_{31} = -5904/10537, \quad \xi_{32} = 15744/10537, \quad \xi_{33} = -17712/10537 \quad \text{and } \xi_{34} = 2952/10537.$$

After cancelling the common factors, the coefficients of the orthogonal polynomials representing linear, quadratic and cubic effects are given in Table 6.

Table 6

Coefficients for linear, quadratic and cubic contrasts

X_j	n_j	ξ_{1j}	ξ_{2j}	ξ_{3j}
0	3	−3	178	−6
1	3	−1	−132	16
2	2	1	−237	−18
4	2	5	168	3

A computer program based on the procedure appears in Narula (1978).

Conclusions

The orthogonal polynomial regression approach is superior to the ordinary polynomial approach for curve fitting. In unplanned experiments, it is usual that an independent variable is observed at unequal spacing with unequal frequency. A simplified procedure to obtain orthogonal polynomial for this situation has been described after discussing the basic calculations required.

Acknowledgment

The author wishes to thank S. Addelman of SUNY at Buffalo for helpful discussions.

References

Bright, J.W. and Dawkins, C.S. (1965). Some aspects of curve fitting using orthogonal polynomials. *Industrial and Engineering Chemistry Fundamentals*, **4**, 93–94.

Delury, D.B. (1950). *Values and Integrals of the Orthogonal Polynomials up to n = 26*. University of Toronto Press, Toronto, Ontario.

Dutka, A.F. and Ewens, F.J. (1971). A method of improving the accuracy of polynomial regression analysis. *Journal of Quality Technology*, **3**, 149–155.

Fisher, R.A. (1952). The influence of rainfall on the yield of wheat at Rothamsted. *Phil. Trans. Roy. Soc. B*, 89–142.

Fisher, R.A. and Yates, F. (1948). *Statistical Tables for Biological, Agricultural and Medical Research*, 3rd Edition. Hafner Publishing Co., Inc., New York.

Forsythe, G.E. (1957). Generation and use of orthogonal polynomials for data-fitting with a digital computer. *Journal of the Society of Industrial and Applied Mathematics*, **5**, 74–88.

Grandage, A. (1958). Orthogonal coefficients for unequal intervals. *Biometrics*, **14**, 287–289.

Guest, P.G. (1950). Orthogonal polynomials in the least squares fitting of observations. *Philosophical Magazine*, **41**, 124–134.

Hahn, G.J. (1977). The hazard of extrapolation in regression analysis. *Journal of Quality Technology*, **9**, 159–165.

Narula, S.C. (1978). Orthogonal polynomial regression for unequal spacing and frequencies. *Journal of Quality Technology*, **10**, 170–179.

Pearson, E.S. and Hartley, H.O. (1958). *Biometrika Tables for Statisticians, Vol. I.* Cambridge University Press.

Robson, D.S. (1959). A simple method for constructing orthogonal polynomials, when the independent variable is unequally spaced. *Biometrics*, **15**, 187–191.

Wishart, J. and Metakides, T. (1953). Orthogonal polynomial fitting. *Biometrika*, **40**, 361–369.

Résumé

Nous discutons les fondements de l'ajustement d'une courbe de régression par les polynomes orthogonaux, lorsque la variable indépendante prend ses valeurs sur des intervales inégaux, avec des fréquences inégales d'observations. Les calculs exigés par la détermination des polynomes orthogonaux sont décrits en employant un exemple simple.

Corrigendum Note to El-Khorazaty, *et al.* (1977) **45**, 129–157.

The last paragraph in section 4.2 on page 150 should read:

'Finally, for the case of two correlated samples [sources], El-Khorazaty and Sen (1976), following the approach presented by Seber (1970), developed the following probability distribution:

$$P[(n_{12}, n_{21}, n_{11})|(N_1, N_2, p_{11}, p_{21}, \phi_{12})]$$
$$= k_1 \left\{p_1^{n_1}(1-p_1)^{N_1-n_1}\right\}\left\{\left(\frac{p_{11}\phi_{12}}{p_1}\right)^{n_{11}}\left(1-\frac{p_{11}\phi_{12}}{p_1}\right)^{n_{12}}\right\} \tag{4.9}$$
$$\left\{\left[\frac{p_{21}}{1-p_1}\right]^{n_{21}}\left[1-\frac{p_{21}}{1-p_1}\right]^{N_2-(n_{12}+n_{21}+n_{11})}\right\}$$

where k_1 is the same as in (4.8), p_{11} is the probability that an event has been recorded by both sources, p_{21} is the probability that an event has been recorded in the second but not in the first source. Similar models were developed by El-Khorazaty and Sen (1976) for the triple-record system (TRS) under either the assumption of independence or dependence among the different sources.'

We would like to thank Professor George A. F. Seber for calling this correction to our attention.