Spline interpolation of demographic data revisited

Article *in* Songklanakarin Journal of Science and Technology · February 2011

CITATIONS
5

READS
487

3 authors:

Nittaya McNeil
Prince of Songkla University
17 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Patarapan Odton
University of the Thai Chamber of Commerce
7 PUBLICATIONS **109** CITATIONS
SEE PROFILE

Attachai Ueranantasun
Prince of Songkla University
15 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Songklanakarin J. Sci. Technol.
33 (1), 117-120, Jan. - Feb. 2011

SJST SONGKLANAKARIN JOURNAL OF SCIENCE AND TECHNOLOGY
http://www.sjst.psu.ac.th

Original Article

Spline interpolation of demographic data revisited

Nittaya McNeil[1]*, Patarapan Odton[2], and Attachai Ueranantasun[1]

[1] *Faculty of Science and Technology,*
Prince of Songkla University, Pattani Campus, Pattani, 94000 Thailand.

[2] *International Health Policy Program Thailand,*
Ministry of Public Health, Bangkok, Thailand

Received 3 February 2008; Accepted 28 February 2011

Abstract

Spline functions have been suggested in demographic research for interpolating age-specific data as they have desirable smoothness optimality properties. However, difficulties arise when boundary conditions need to be satisfied. An additional problem is that age-specific demographic data functions are necessarily non-negative, requiring the interpolating spline to be monotonic non-decreasing. In this paper we describe a simple and effective alternative that circumvents these problems. We show that natural cubic splines can be used to interpolate age-specific demographic data and ensure that relevant boundary conditions on second derivatives are satisfied, thus preserving the desirable optimality property of the interpolating function without the need to increase the degree of the spline function. The method involves incorporating one or two additional strategically placed knots with values estimated from the data. We describe how the method works for selected fertility, population, and mortality data.

Keywords: natural cubic spline, monotonic, fertility, population, mortality

1. Introduction

"Spline functions (Greville, 1969) are useful for smooth the interpolation of data". A cubic spline with n knots $x_1 < x_2 < ... < x_n$ is any function $s(x)$ with continuous second derivatives comprising piecewise cubic polynomials between and beyond the knots. Denoting by x_+ the function taking the value x for $x > 0$ and 0 elsewhere, $s(x)$ may be written as

$$s(x) = d_0 + d_1 x + d_2 x^2 + d_3 x^3 + \sum_{i=1}^{n} c_i (x - x_i)_+^3 \quad (1)$$

A *natural* cubic spline is a cubic spline satisfying the additional requirement that the function is linear for values of x outside the knots. This function has the property that for all functions with specified values at the knots the natural cubic spline minimizes the integral of its squared second derivative over the interval (x_1, x_n). Since $s(x)$ is linear for $x < x_1$ if d_2 and d_3 are both 0, this requires that the cubic and quadratic terms in $s(x)$ must also disappear for $x < x_n$, so to be a natural spline the n+4 coefficients in the cubic spline must satisfy the following two sets of equations

$$d_2 = 0, \sum_{i=1}^{n} c_i = 0, \quad (2)$$

$$d_3 = 0, \sum_{i=1}^{n} x_i c_i = 0. \quad (3)$$

More generally, a natural spline of degree $2m$+1 comprises piecewise polynomials of degree $2m$+1 with continuous derivatives of order m+1 reducing to polynomials of

* Corresponding author.
Email address: nittaya@bunga.pn.psu.ac.th

degree m outside the knots, and has the property that for all functions with specified values at the knots, it minimizes the integral of the squared derivative of order m+1 over the interval (x_1, x_n). Equations 2 and 3 are then replaced by two sets of m+1 equations.

As pointed out by McNeil *et al.* (1977), splines can be used in demographic research for interpolating age-specific cumulative fertility, where data are usually available at 5-year age intervals from 15 to 50 years. However, in practice fertility increases slowly from 0 at age 15 and also decays even more slowly to 0 at the other extreme, so it is desirable to impose the additional conditions that both the derivatives and the second derivatives of the spline are 0 at the extremes. To satisfy these boundary conditions, McNeil *et al.* (1977) suggested relaxing the requirement that the spline function is natural. However, although it appears that for a cubic spline the four equations in Equation 2 and 3 could be replaced by the four conditions on the first and second derivatives at the boundaries, this cannot be done because the resulting equations do not involve c_n. So the proposed solution not only fails to satisfy the smoothness optimality condition but also must be of degree 5 or higher.

A further problem is that age-specific demographic data, including fertility functions are necessarily non-negative, requiring the interpolating spline to be monotonic non-decreasing. In a recent paper investigating the use of spline functions for another demographic application involving the interpolation of Australian mortality data, Smith *et al.* (2004) demonstrated this inadequacy of the spline function. He applied the Hyman filter (Hyman, 1983), which imposes alternative constraints on the derivatives of a cubic spline, possibly sacrificing smoothness if the filter has changed first derivatives.

In this paper we describe a simple and effective alternative that circumvents both of these problems. It uses natural cubic splines to interpolate age-specific demographic data and ensures that relevant boundary conditions on second derivatives are satisfied, thus preserving the desirable optimality property of the interpolating function without the need to increase the degree of the spline function. The method involves incorporating one or two additional strategically placed knots with values estimated from the data.

In the next section we describe how the method works for situations of interest to demographers. After that we illustrate the method with three examples, (a) Italian fertility (Festy, 1970) described by McNeil *et al.* (1977), (b) the cumulative distribution by age of males from the 2000 population census of Thailand (National Statistical Office, 2000), and (c) Australian female mortality in 1901 considered by Smith *et al.* (2004).

2. Methods

The first situation involves fitting a natural cubic spline to a cumulative age-specific fertility schedule, where it is required that the function is 0 at x_1 and has first and second derivatives 0 at both x_1 and x_n. Note that if the first derivative of the required function is 0 at x_1 and x_n the second derivative must also be 0 at these points. This is because a cubic spline has continuous second derivatives.

For simplicity we shift the x-axis so that the first knot is located at $x_1 = 0$, and we place two additional knots at a, b, with coefficients g, h, respectively. Since the value of both the function $s(x)$ and its derivative at $x = 0$ is 0, the linear terms in Equation 1 vanish and its functional form is thus

$$s(x) = \sum_{i=1}^{n} c_i (x - x_i)_+^3 + g(x-a)_+^3 + h(x-b)_+^3. \quad (4)$$

A total of n+4 coefficients need to be determined. These comprise the n+2 coefficients in Equation 4 together with the values of the function to be determined at a, b. The linear equations determining these coefficients comprise (a) the n+1 equations requiring that the function have specified values at the n-1 knots apart from the first where its value is necessarily 0, (b) the two equations needed to ensure that the function is linear for $x > x_n$, and (c) the requirement that the derivative is 0 at x_n.

The parameters a, b may be chosen to vary the result to satisfy other requirements including smoothness and monotonicity. These values were chosen by selecting a value in between the first and second knots and a value between the last and the second last knot. If these two knots are too close to the two boundary knots, then the value of the function can become negative. In practice, the region will have to be determined to ensure that the function will be monotone non-decreasing and smooth. If the function is symmetric then the two distances, the first knot to a and b to the last knot, should be similar. If the function is skew to the right, the distance from b to the last knot would be slightly more than the distance from the first knot to a, and vice versa for a left skew function.

The second situation involves fitting a natural cubic spline to cumulative population data. In this case it is not necessary to impose a condition that the derivative of the function is 0 at $x = 0$, but the condition is desirable at the other end. We place an additional knot at b with coefficient h. The functional form is thus

$$s(x) = dx + \sum_{i=1}^{n} c_i (x - x_i)_+^3 + h(x-b)_+^3 \quad (5)$$

A total of n+3 coefficients need to be determined. It comprises the n+2 coefficients in Equation 5 together with the value of the function to be determined at b. The linear equations determining these coefficients have similar conditions as the first situation.

The third situation involves fitting a natural cubic spline to mortality data. In this case there are no end-point requirements, but one or more artificial knots may need to be included to ensure that the function is monotonic. In this situation, the monotonicity condition cannot easily be met by satisfying equations involving values of the derivatives similar to boundary conditions, so the values of the function

at the artificial knots are estimated by trial and error. We give an example in the next section.

3. Application

The Italian cumulative age-specific fertility data (Festy, 1970) at five-year age periods are shown in Table 1.

There are $n = 8$ knots so the method described in Section 2 involves solving 12 linear equations. To fit a natural cubic spline to these data we have chosen the knots at the ages 15 and 50 years and every 5 years in between and an additional two knots at ages 16 and 48. The values were chosen by selecting the first knot in between age 15 and 20 year, the second knot in between age 45 and 50 year. To ensure that the values of the function are not negative, for this data the first knot should be between 15.62 and 20 years, while the second should be between 45 and 48.16 years. For simplicity, we chose to round the knots to the nearest whole age. Thus, 16 and 48 were chosen as the values of the additional knots.

The fitted values of the cumulative fertility at the two extra knots are 0.00029 and 2.30491, respectively. The derivative of the fitted cubic spline is shown in the left panel of Figure 1.

The right panel of Figure 1 shows the results of fitting a natural cubic spline to the age-specific cumulative Thai male population in 2000 as shown in Table 2. Since the proportion living beyond age 100 years is very small we take this value as the upper limit. In this example there are $n = 18$ knots. To fit the natural spline to these data we need only to insert one extra knot before the maximum age bracket, at about age 85 years, to make the first and second derivative of the spline function 0 at $x_n = 100$. Using Equation 5 there are 21 unknowns together with 21 equations. Solving these equations gives the value 30.8095 for the male population below age 85 years.

For the last example, we fit a natural cubic spline to the Australian 1901 female cumulative mortality at ages 1, 5, 20, 40, 60, 65, and 100 years. The data are estimated from Figure 1 of Smith *et al.* (2004). Table 3 gives these estimates.

Table 1. Cumulative age-specific fertility data.

Age (years)	Cumulative fertility rate
15	0.000
20	0.080
25	0.593
30	1.297
35	1.840
40	2.171
45	2.296
50	2.306

Table 2. Thai male population in 2000.

Age (years)	Male population in millions
0	0.000
5	2.081
10	4.570
15	7.092
20	9.536
25	12.230
30	15.018
35	17.849
40	20.623
45	23.122
50	25.195
55	26.809
60	27.968
65	28.924
70	29.721
75	30.287
100	30.936

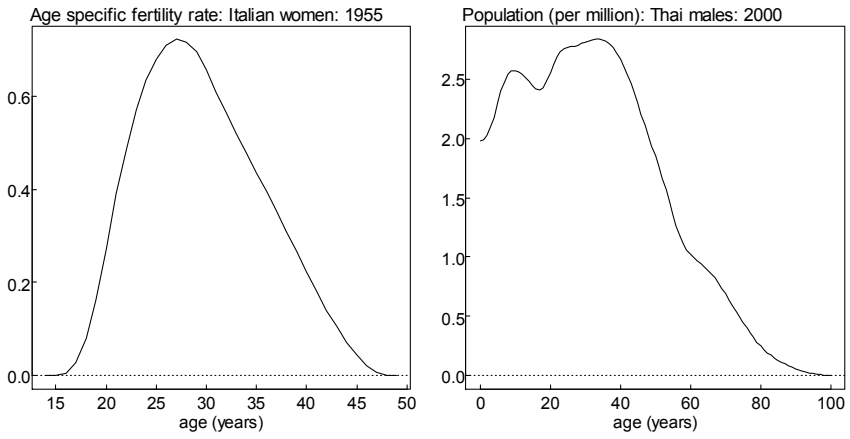

Figure 1. Natural cubic spline interpolations of fertility and population densities.

Table 3. Cumulative age-specific mortality data for Australian women in 1901.

Age (years)	Female mortality per 1000
0	0.00
1	4.75
5	6.40
20	7.75
40	11.20
60	13.85
65	14.80
100	19.15

There are 8 knots. The left panel of Figure 2 shows the natural spline fitted to the data shown in Table 3 (dotted curve), the monotonic spline interpolation given by Smith *et al.* (2004) (lighter curve), and the natural spline using the data with the extra knot (darker curve). The right panel of Figure 2 shows the corresponding density curves obtained by differentiating the cubic spline functions.

From the figure, the natural spline fluctuates in the first 20 years, which include the first five knots. The additional knot was chosen between ages 1 and 5 years and we obtained a smoother result than that obtained by Smith *et al.* (2004) by choosing the value of 2 years. The cumulative mortality at this age obtained from Figure 1 of Smith *et al.* (2004) is 5.9 deaths per 1,000. Three additional knots at ages 10, 30 and 50 years with cumulative mortality values of 6.5105, 10 and 12.4 deaths per 1,000, respectively, gave the results similar to the one given by Smith *et al.* (2004) as shown in Figure 2.

The function using this method is smooth and always non-negative for the three examples we have shown. However, there is no guarantee that the function will be smooth and non-negative for other datasets. The value of the additional knots could be solved linearly using the property of natural cubic splines where the smoothness function interpolates the data by minimizing the integral of its squared second derivative. Further investigations will be needed on these issues.

Acknowledgement

We are grateful to Emeritus Prof Don McNeil for his helpful suggestions.

References

Festy, P. 1970. Evalution de la Fécondité en Europe Occidental Depuis la Guerre. Population. 25, 229-274.

Greville, T.N.E. 1969. Introduction to Spline Functions. In Theory and Applications to Spline Functions. T.N.E. Grevillea, editor. Academic Press, New York, U.S.A., pp 1-35.

Hyman, J.M. 1983. Accurate monotonicity preserving cubic interpolation, SIAM Journal on Scientific Computing. 4(4), 645-654.

McNeil, D.R., Trussell, T.J. and Turner, J.C. 1977. Spline Interpolation of Demographic Data. Demography. 14(2), 245-252.

National Statistical Office. 2000. Preliminary Report the 2000 Population and Housing Census. Statistical Data Bank and Information Dissemination Division, Bangkok. Thailand.

Smith, L., Hyndman, R.J. and Wood, S.M. 2004. Spline Interpolation for Demographic Variables: the Monotonicity Problem. Journal of Population Research. 21(1), 95-98.

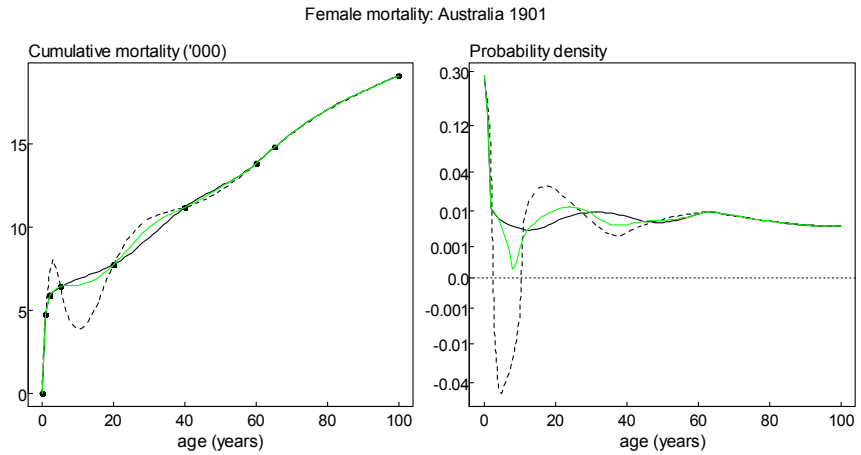

Figure 2. Spline interpolations of cumulative mortality for Australian women in 1901.