Chapter 7
Estimating the Goodman, Keyfitz and Pullum Kinship Equations: An Alternative Procedure

7.1 Introduction

In a pioneering paper, Goodman *et al*. (1974) presented a general analytic system for studying the relationships between mortality and fertility and kin numbers. For stable populations with varying regimes of fertility and mortality, they provide formulas to calculate average numbers of kin, by category of kin, for females of various ages.

Of great substantive importance was their demonstration of the strong relationship between kin numbers and fertility levels for all categories of kin except ascendants in the direct line. The general relationship, obvious after the fact, was not widely recognized before their work [more attention had been devoted to the effect of mortality on kinship], nor had it been quantified even roughly. The relationship, combined with current low levels of fertility in many societies [for example, Italy with a total fertility rate of 1.3, or about 0.65 daughters born per woman] points to a continuing decline in numbers of kin for the average person in the future, and probably an associated decline in the importance of family and kinship in everyday life.[1]

The potential importance of this finding can be illustrated by a mental experiment. Suppose China's 'one-child' policy were perfectly realized, with no one having more than one birth. In a generation or two, collateral kinship would

The research underlying this chapter was carried out while I was Visiting Professor, Dipartimento di Scienze Demographiche, Università degli Studi di Roma, at the kind invitation of Prof. Antonella Pinnelli. Originally published in *Mathematical Population Studies* 5 (1995) pp. 161–170.

[1] A major qualification of this statement relates to the potential role of high levels of divorce and remarriage in supplying an individual with 'new' kin – step kin – in addition to those resulting from first marriage and birth.

T.K. Burch, *Model-Based Demography*, Demographic Research Monographs,
DOI 10.1007/978-3-319-65433-1_7

disappear: there would be no brothers or sisters, aunts or uncles, nieces or nephews, or cousins—only, parents, grandparents, child, and grandchild.

Despite its substantive importance, their approach has not seen much further development [for example, by the inclusion of data on proportions married, or the relaxation of the stable population assumption] or widely used for the exploration of substantive questions relating to kinship (with the major exceptions of Goldman 1978, 1984 and Coresh and Goldman 1988). One practical barrier has been the difficulty of estimating the integral equations in which the basic relations are stated, equations containing up to quadruple integrals.

In their original paper the authors comment: 'Ordinarily, we cannot evaluate the $l(x)$ and $m(x)$ functions for arbitrary values of x, since the data are usually collected for 5-year age intervals' (p. 24). To estimate the equations, they develop finite approximations of the multiple integrals, programmed in Fortran by Pullum. In its original form, this Fortran code ran to more than ten single-spaced pages. It has been used in the later work by Goldman, and more recently by Keyfitz (1986), in an analysis of Canadian kinship numbers. But such code, written by someone else, is often difficult to master or to modify correctly.

This note illustrates an alternative procedure for evaluating the kinship integrals, using computer software developed since their paper first appeared. The procedure allows one in effect to 'evaluate the $l(x)$ and $m(x)$ functions for arbitrary values of x.' It involves a minimum of programming, yields results that agree well with the Pullum approximations, and has the advantage, both scientific and pedagogical, of working directly with the theoretical equations rather than with long finite approximation algorithms. Theory and computation are more closely linked.

The procedure involves two steps: (1) analytic expressions are found to represent empirical data on age-specific fertility and survivorship; (2) these expressions are substituted into the theoretical integral equations for kin numbers [with appropriate arguments and limits of integration], which are then evaluated numerically.

In the present note, the first step has been accomplished using TableCurve, an automated curve-fitting package using standard algorithms for linear or non-linear fitting.[2] Any general-purpose curve-fitting routine could be used. TableCurve has the advantage, for this application, that the user does not have to supply a functional form ahead of time, although user-defined functions are an option. The program has a built-in library of over 3500 functions, and can successfully fit most sets of demographic data by age or duration.[3]

The resulting analytic expressions and parameter estimates are used solely to represent particular schedules of age-specific mortality and fertility. They do not

[2]Systat, Richmond, California.

[3]The ability of computer curve-fitting packages such as the one used here to find functions to represent demographic data is a matter for further empirical investigation, To date I have encountered only a few cases of demographic data for which TableCurve could not find a function that fits reasonably close. An example: data on age-specific householder rates [female and non-family] from recent Canadian censuses, rates which rise to around age 30, decline, and then rise again in later life.

have, nor need they have for this application, any theoretical rationale or interpretation for their parameters. The only requirement is a close fit to the data at hand. Of course, if functional forms better grounded either in mathematics, empirical research, or substantive theory are available, their use in this application would be possible and desirable.

The second step uses the numerical integration capabilities of Mathcad, a numerical mathematics package.[4] Again, other mathematics packages could be used, so long as they can evaluate multiple integrals. Mathcad has an advantage that basic formulas are entered and appear [on the screen and in hardcopy] in standard mathematical notation, tying the calculations more closely to theoretical equations. Note, however, that the results still are based on underlying numerical approximation procedures not unlike those of Pullum'.[5]

The procedure is illustrated for children and grandchildren for 1981 Canadian data, and the results compared with those in Keyfitz (1986). Since both techniques start with data for 5-year age intervals to approximate theoretical integrals, neither can be said to yield 'correct' estimates of kin numbers, so that Keyfitz's results cannot serve as an absolute standard against which to judge the new procedure proposed. In any case, the agreement is close,[6] and the choice between the two computational techniques can be made on other grounds – ease of application, transparency, and flexibility.

Canadian 1981 age-specific fertility rates from Keyfitz (1986) were modified by adding zero values at ages 10 and 52.5, and fit by TableCurve.[7] Perfect fits were given by high-order polynomials, with eight to ten parameters. But for convenience in further use, more compact functions, with three or four parameters, were examined. The following function was chosen[8]:

[4]PTC Inc., Needham, Mass.

[5]It is conceivable that expressions for fertility and survivorship could be found that would lead to closed-form solutions of the kinship equations. But these still would not be exact solutions given the approximation involved in the underlying data.

[6]As it should be, given that both are using essentially the same data and similar numerical approximation procedures. The small differences observed presumably relate to small differences in input [for example, treatment of extreme ages of fertility or survivorship, age indexing, etc.] and in numerical procedures.

[7]For fitting, age-specific fertility rates were associated with the mid-points of their respective age intervals. This clearly involves error, especially in the intervals 10–14 and 45–49. With more information [e.g., data on births by single-years of age], average ages instead of midpoints could be used. Or one could simply assume that the rate for 10–14 should be associated with some age greater than 12.5. But such refinements are not necessary for present purposes.

[8]For readability, only three digits are given for parameter values. For accurate graphing of these functions more digits may be needed, especially if the function is non-linear. See Note to Appendix A.1

$$f(x) = e^{(a+[bx\sqrt{x}]+c\sqrt{x})}$$
$$a = -35.1 \quad b = -0.122 \quad c = 9.66$$

When the resulting function ***f(x)*** is integrated over the same reproductive span as given by the original data (ages 10–50), the total fertility rate agrees with that computed in the usual way to within 0.1%. As well, visual inspection and conventional measures of goodness of fit suggest that f(x) provides a reasonable fit to the fertility data at hand. To repeat, that is the only goal for the present application. No theoretical or substantive claims are made for the resulting functions; we use them as approximating functions, defined by TableCurve as '...nothing more than an equation which is used to represent X-Y data' (Systat 2002, pp. 20–1).[9]

To eliminate small non-zero values of f(x) outside the reproductive ages, the function is redefined by inserting conditions on x which evaluate the function as zero when x is less than 10 or greater than 52.5. The function is also re-defined to adjust for the sex ratio at birth [since the kinship equations relate to one-sex, stable population models], yielding m(x), a maternity function for female births.

A similar curve-fitting procedure was applied to $\boldsymbol{L_x}$ values from the 1981 abridged life table for Canada [the data used by Keyfitz] to fit a survivor function.[10] In this case, four parameter functions were required to get an adequate fit. The chosen function:

$$s(x) = a + \frac{b}{1 + e^{\frac{c-x}{d}}}$$
$$a = -0.741 \quad b = 5.66 \quad c = 84.4 \quad d = -8.85$$

As with the fertility function, conditions on x were inserted to assure that the curve behaves properly at ages outside the range of observation.[11] And, the values were adjusted to take account of the 5-year intervals of the original $\boldsymbol{L_x}$ data, yielding a survivorship function p(x) (See Appendix A.1).

[9]The parameters relate to geometric properties of the graph – intercept, height, center, and width. But they have no further meaning in terms of a theory of kinship.

[10]The same L_x data were used for the sake of comparability. Given the continuous formulation of the present approach, fitting l_x values from the complete life table at ages 0,5...100 would have been more natural.

[11]With TableCurve, one can zoom out to see the behavior of a fitted function well outside the range of observation, and can quickly calculate predicted values for arguments outside that range. But this further step [for example, requiring zero survivors beyond some maximum age] seems warranted given the somewhat blind/mechanical procedure of curve-fitting. A skilled mathematician, of course, might define a function with the correct asymptotic properties.

***Start of reproductive period*: α = 0** **Age of ego: a = 0,20…80**

***Daughters born by age a*:**

$$B_1(a) = \int_{\alpha}^{a} m(x)dx$$

***Living daughters at age a*:**

$$BL_1(a) = \int_{\alpha}^{a} p(a-x)m(x)dx$$

Granddaughters born by age a [y = 0…50, defining a new age range for the daughter generation]:

$$B_2(a) = \int_{\alpha}^{a} [\int_{\alpha}^{a-x} p(y)m(y)dy]m(x)dx$$

Living granddaughters at age a:

$$BL_2 = \int_{\alpha}^{a} [\int_{\alpha}^{a-x} p(y)m(y)p(a-x-y)dy]m(x)dx$$

Fig. 7.1 Estimating Kin Numbers

7.2 Estimating Kin Numbers

Figure 7.1 defines the Goodman, Keyfitz and Pullum equations for daughters born, living daughters, granddaughters born and living granddaughters by age a of an average woman [ego]. The fertility and survivorship functions ***m(a)*** and p***(a)*** are as defined above. Given these equations and function definitions, Mathcad evaluates the integrals (see Appendix A.2). The results are given in Fig. 7.2.

Estimates by the proposed procedure are in close agreement with those of Keyfitz (1986), presented for comparison. Agreement is to within 1.1 per 100 kin for all categories and ages. The largest relative errors are for daughters and living daughters at age 20 of the reference woman – about 15%. These presumably relate to differences in procedures for dealing with fertility rates in the earliest ages of childbearing. But notice that the substantive story is not appreciably different, 6 or 7 daughters born per 100 women by age 20.

7.3 Discussion

The differences between the results of the proposed computational procedure and those produced by the Pullum algorithm are negligible, within the bounds of error of the original data. Moreover, the results are precise enough for any likely substantive use to which they might be put, given that they relate to a highly abstract model of kinship [a one-sex stable population model, with no input for marriage patterns].

	Number of Kin per 100 Women			
Age	Daughters	Living Daughters	Granddaughters	Living Granddaughters
Results from equations in Figure 1:				
20	7.4	7.3	0	0
40	80.5	79.2	1.8	1.8
60	81.3	79.7	51.8	50.9
80	81.3	77.0	64.9	63.7
Results from Keyfitz [1986]:				
20	6.3	6.3	0	0
40	80.4	79.5	1.9	1.9
60	81.3	79.7	50.9	50.3
80	81.3	77.6	64.9	63.9

Fig. 7.2 Comparison of estimates

The general approach used above clearly has applications to other areas of population mathematics. The approach is not entirely novel, but until recently it was impractical and beyond the capabilities of many researchers. Finite sums using grouped data became conventional. Writing as recently as 1985, for example, Keyfitz could note correctly with respect to an expression for the intrinsic growth rate r: 'no direct use can be made of a continuous form like (5.1.4) – it must be converted to the discrete form for calculations' (1985, p. 115), and more generally: 'Although the stable age distribution is easier to think about in the continuous version, application requires a discrete form' (1985, p. 81).

Due to recent developments in computer software, this is no longer the case. As illustrated above, it is now relatively easy to find continuous functions to represent many demographic data sets, and to do direct numerical evaluation of integrals and other analytic expressions. In some contexts, working with analytic expressions for processes such as fertility, survivorship and marriage may be a more effective way to derive numerical results than traditional finite sums. At the very least, one now has a choice.

Approximating functions also can be effective for interpolation and – with due caution – extrapolation.

The suggested procedure is a reminder of Hakkert's (1992) argument that many standard demographic algorithms were derived for purposes of hand calculation, and may need to be revised to make greater use of modern developments in statistics and computer software.[12]

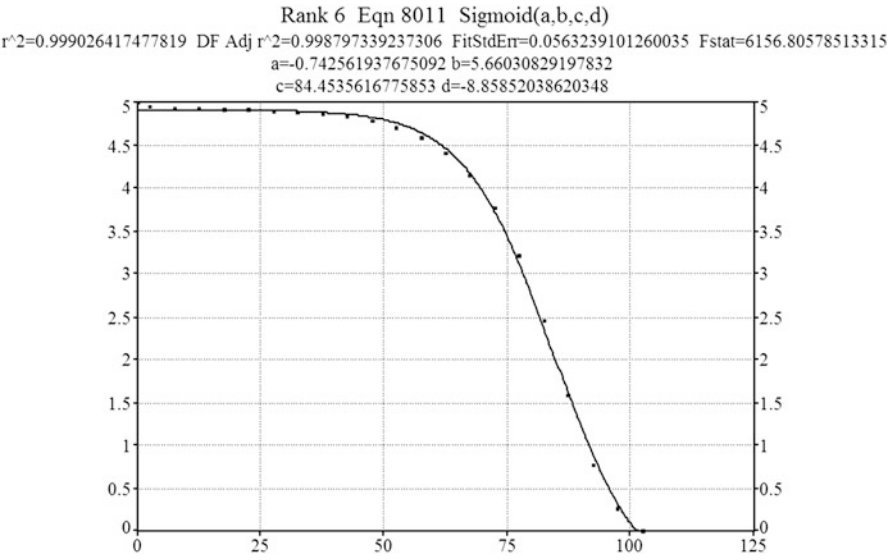
[12]Caswell (1989) makes the interesting historical observation that much of Leslie's (1945) paper on matrices in demography is spent developing transformations suited to hand calculation, transformations now largely outmoded by the computer.

As with any use of computerized 'black box' procedures, of course, one must balance the potential advantages in ease, speed and flexibility of computation against the possibility of unrecognized pitfalls leading to seriously incorrect results. In the case at hand, for example, it would be easy to select a survivorship function that rises after age 100 or so. The careless use of such a function in the kinship equations would lead to meaningless results for some kinship categories. Computer mathematics software is at best a partial substitute for mathematical skill, and no substitute at all for thoughtful analysis.

Finally, it should be emphasized once more that in this approach, the analytic expressions are used solely to represent *specific sets of* data. Fertility schedules for a high-fertility population might lead to different functions being selected. The discovery of *general* analytic expressions for such processes, especially expressions with theoretically meaningful parameters, is another, more difficult and more important task.

Appendices

Appendix A: Tablecurve Output for Fit of Survival Curve



Note. This is a facsimile of the TableCurve graphic output for the function fit to L_x data, to represent survivorship. Parameter values and measures of goodness of fit are given to 15-digit accuracy. This is not justified by the accuracy of the basic data. But if one wishes to graph the function independently of TableCurve, many digits may be required to get an accurate graph, for example, with the correct range or specific values of y. Other output, not shown here, gives summary statistics and confidence intervals for parameters.

Appendix B: Facsimile of Mathcad Worksheet for Kin Numbers

Note: In Mathcad, this would be a live worksheet, with results recalculated after changes to numbers, expressions, etc., as in a spreadsheet. Notation inconsistency: P(x) and S(x) are the same as p(x) and s(x) used in text earlier.

Start of reproductive period : $\alpha := 0$ Age of ego : $a := 0, 20 .. 80$

Fertility functions:

$$a := -35.134315 \qquad b := -0.122003 \qquad c := 9.656093$$

$$f(x) := e^{[a+(b\cdot x\cdot\sqrt{x})+c\cdot\sqrt{x}]}$$

$m(x) := f(x) \cdot 0.4867 \cdot (x \geq 10)(x \leq 50)$ (Female births only; range limited to 10 to − 50 by conditions on x)

$$p := -0.741013 \qquad q := 5.658693 \qquad r := 84.446801 \qquad s := -8.853886$$

$$S(x) := p + \frac{q}{1 + e^{\left[\frac{-(x-r)}{s}\right]}}$$

$p(x) := \dfrac{S(x) \cdot (x \geq 0) \cdot (x \leq 100)}{5}$ (Range limited to 0 to − 100 by conditions on x)

$$a := 0, 20 .. 80 \qquad x := 0 .. 100$$

Daughters born by age a

$$B1(a) := \int_{\alpha}^{a} m(a)da$$

B1(a)=
0.000
0.074
0.805
0.813

Daughters living at age a

$$BL1(a) := \int_{\alpha}^{a} P(a - x) \cdot m(x)dx$$

BL1(1)=
0.000
0.073
0.792
0.770

Age range for daughter generation $y := 0..50$

$$B2(a) := \int_{\alpha}^{a} \left(\int_{\alpha}^{a-x} P(y) \cdot m(y)\, dy \right) \cdot m(x) dx \qquad BL2(a) := \int_{\alpha}^{a} \left(\int_{\alpha}^{a-x} P(y) \cdot m(y) \cdot P(a - x - y) dy \right) \cdot m(x)\, dx$$

B2(a)=
0.000
0.000
0.018
0.518
0.649

B2(a)=
0.000
0.000
0.018
0.509
0.637

References

Caswell, H. (1989). *Matrix population models*. Sunderland: Sinauer Associates.

Coresh, X., & Goldman, N. (1988). The effect of variability in the fertility schedule on numbers of kin. *Mathematical Population Studies, 1*, 137–156.

Goldman, N. (1978). Estimating the intrinsic rate of increase of a population from the average number of older and younger sisters. *Demography, 15*, 499–508.

Goldman, N. (1984). *Fertility, mortality and kinship*. Paper presented at annual meetings of Population Association of America, Minneapolis.

Goodman, L., Keyfitz, N., & Pullum, T. (1974). Family formation and the frequency of various kinship relationships. *Theoretical Population Biology, 5*, 1–27. See also 1975 Addendum, *Theoretical Population Biology* 8: 376-381.

Hakkert, R. (1992). *Computing in demographic analysis: Beyond paper and pencil algorithms*. Paper at IUSSP/NIDI Expert Meeting on Demographic Software and Computing, The Hague, 29 June–3 July, 1992.

Keyfitz, N. (1985). *Applied mathematical demography* (2nd ed.). New York: Springer.

Keyfitz, N. (1986). Canadian kinship patterns based on 1971 and 1981 data. *Canadian Studies in Population, 13*, 123–150.

Leslie, P. H. (1945). On the use of matrices in certain population mathematics. *Biometrika, 35*, 213–245.

Nagnur, D. [Statistics Canada]. (1986). *Longevity and historical life tables, 1921–1981 [Abridged], Canada and the Provinces*. Ottawa: Ministry of Supply and Services.

Statistics Canada. (1984). *Life tables, Canada and the Provinces, 1980–1982*. Ottawa: Ministry of Supply and Services.