# BIRTH PROJECTIONS WITH COHORT MODELS*

JAMES M. BESHERS

Massachusetts Institute of Technology, Joint Center for Urban Studies of the Massachusetts Institute of Technology, and Harvard University

RESUMEN

*El propósito del estudio es exponer un modelo de proyecciones de la natalidad con las probabilidades de natalidad de las mujeres según su edad, la cohorte a que pertenecen y el número de nacimientos anteriores, como parámetros esenciales, incluyendo algunos parámetros adicionales de origen psicológico, económico y sociológico.*

*Entre las razones que se tuvieron para elegir estas probabilidades de natalidad como parámetros pueden citarse: primero, el hecho de que las estadísticas usuales de fecundidad pueden obtenerse como subproducto del análisis; y segundo, la estimación es posible gracias a la labor empírica del difunto P. K. Whelpton. Los resultados de la proyección y las deducciones derivadas como subproducto se expresan sintéticamente en forma algebraica y se preparan con suma facilidad para un computador.*

*Las probabilidades de natalidad mismas pueden expresarse como funciones de variables biológicas y sociales. El procedimiento más sencillo y útil consiste en dividir la población en dos componentes con diferentes funciones: un componente de planificación de la familia y otro de no planificación. Las probabilidades de natalidad para el segundo se supondrán relativamente independientes del tiempo —el complejo de factores biológicos y sociales tendrá diferentes efectos en situaciones culturales y biológicas diferentes, pero se puede tratar de estimar estos efectos mediante los métodos demográficos clásicos.*

*Las probabilidades de natalidad para el componente de planificación se supondrá que varían con el tiempo; los valores especifican el tamaño ideal de la familia y las variaciones en el ingreso futuro esperado influyen en el espaciamiento de los nacimientos para alcanzar el ideal, y por supuesto los parámetros biológicos influyen favoreciendo u obstaculizando el logro de ese ideal. Para combinar las dos series de probabilidades de natalidad hay que estimar la proporción de planificadores en una sociedad dada. Esta proporción será una función monótonamente creciente en el tiempo. Los modelos de difusión de la variedad de contagio convienen para expresar la dependencia de esta proporción con respecto al tiempo.*

*La estimación de los parámetros sociales y biológicos mejora considerablemente, si dentro de sectores social y geográficamente definidos, existen datos como la composición según la raza, la religión, la ocupación y la distribución por sectores urbanos y rurales. Por ejemplo, la clásica serie histórica de las correlaciones de las clases sociales con la fecundidad puede deducirse de un modelo de difusión para una sociedad.*

The aim of this paper is to state a model for birth projections that contains birth probabilities for women, using age, cohort, and number of previous births as its essential parameters and including additional parameters that are psychological, economic, and sociological in origin. The paper has three parts. First, the rationale is given for selecting these birth probabilities as the essential parameters. Second, the general procedure for estimating these parameters is given; this will involve the introduction of new parameters, both biological and social. Third, the general procedure will be illustrated in certain simplified special cases.

In the course of the paper, a number of relationships for which we currently have very poor empirical estimates will be expressed mathematically. From this perspective, one could argue that the paper defines a program of research upon birth projections rather than a finished product. For my own part, I would argue that even in its rough-hewn form, the model is use-

* Much of the work on this paper was done at the Bureau of the Census. A more extended discussion is available in my memorandum of July 19, 1963 to Howard G. Brunsman. At the Census, Howard G. Brunsman, Wilson Grabill, Norman Lawrence, and Donald S. Akers spent many hours with me discussing these ideas; many other persons joined in from time to time. Anders Lunde and Harry Rosenberg of the National Vital Statistics Division made unpublished data available for me. Aaron Fleisher and Charles Tilly have helped me in this formulation, as has James S. Coleman.

ful for projections, but I would also emphasize the compelling need for the continued research program that this model implies.

The rationale for selecting these "cohort" birth probabilities as the essential parameters stems from several considerations. Of these, the two most important are (1) given these parameters, the following measures of fertility can be deduced—cumulative birth distributions at any age (a census concept), average size of completed family (a census concept), annual births (a vital statistics concept), and spacing of births (a census concept)—and (2) the estimation of these parameters is made feasible by the empirical time series constructed by the late P. K. Whelpton.[1]

For any given cohort (women born in a given year) we wish to produce the annual distribution of women by number of children, the cumulative births to a given age, and the annual births at each age. We will choose our definitions to be consistent with the form of Whelpton's empirical probabilities.

The probabilities are defined over eight categories: 0 children, 1 child, 2 children, and so on, to 7-and-up children. Each woman is allowed to have only one child in one year. Thus, having a birth in a year is equivalent to moving to the next category for all categories except the last category.

A matrix of such probabilities is defined for each cohort and each age. Let age $= x$, cohort $= T$, then the matrix

$$\begin{array}{cc} & \begin{array}{cccccc} 0 & 1 & \cdots & & & 7\text{up} \end{array} \\ \begin{array}{c} 0 \\ P_x{}^T = 1 \\ \\ \\ \\ \\ 7\text{ up} \end{array} & \begin{pmatrix} P_{11} & P_{12} & 0 & \cdots & & 0 \\ 0 & P_{22} & P_{23} & 0 & \cdots & 0 \\ 0 & 0 & P_{33} & P_{34} & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array}$$

[1] Especially the unpublished birth probabilities at the National Vital Statistics Division.

contains only two nonzero entries in each row, save for the last row, which contains only a one in the diagonal. (A special adjustment must be made for births to the 7-up category.)

Let us describe a single cohort and thus dispense with the superscript $T$. Within this cohort we will define all births as occurring to women of ages 14–49, inclusively.

The distribution of women at any given age is defined by a vector: $m_x = (m_{x1}, m_{x2} \ldots, m_{x8})$, where $m_{x1}$ is the number of women at age $x$ with 0 children, and so on, up to $m_{x8}$, which is the number of women at age $x$ with 7-and-up children.

We may define the computation of $m_{x+1}$ from $m_x$ and $P_x$ (the previous age) by starting at age 14 and proceeding through all ages. If we subtract 13 from $x$, the equations are $m_1P_1 = m_2$; $m_2P_2 = m_3$; and so on for all ages. By substituting for $m_2$ in the second equation its value in the first equation, we have $m_1P_1P_2 = m_3$. In general, we have the product of $P$ matrices:

$$m_1 \sum_{(i=1)}^{k} P_i = m_{k+1}\,.$$

From these cumulative distributions we obtain cumulative births by multiplying the number of persons in each category by the number of children associated with that category save for the 7-and-up category that requires a special procedure. Annual births are obtained by computing differences between successive cumulative births.[2]

Implied spacing is obtained by taking each element of $m$ and, for each age and

[2] Donald S. Akers and I wrote a FORTRAN program at the Census which performs these computations. Wilson Grabill has pointed out discrepancies between the number of women having at least one child implied by runs of this program. It is my present conjecture that the discrepancy stems from a bias in Whelpton's data. Multiple births and births occurring twice within a year to a woman are attributed to different women, thus inflating the number of women giving birth. This bias could be estimated and removed from the projections.

parity, successively applying the probability of moving to a higher parity over all subsequent ages. The problem is to retain the age of entry into the parity category and the age of departure from each parity category, so the interval can be obtained by subtraction. For example, for the interval between first and second child, we take the number of women having their first child at age 15, then apply the probabilities for a second child at 16, 17, and so on, obtaining a distribution of intervals. Then we take the number of women having a first child at age 16 and repeat the process. For each parity we run through each starting age and all subsequent ages.

The empirical rationale for the cohort model can be best brought out by considering the issues that any birth probability model must be confronted with if it is to be useful for projections. One must develop a procedure for estimating hypothetical future birth probabilities. The two most obvious procedures have proved unsatisfactory in the past. The simplest procedure would be the assumption that birth probabilities are independent of time, but all the sound time series that we have analyzed produce evidence to the contrary. The next simplest procedure is the assumption that the birth probabilities are time-dependent but that this dependence can be expressed by a fairly simple mathematical curve. Such a procedure has the apparent merit of strong analogies to the "natural laws" of physical science. Unfortunately, the post-1945 baby boom ran counter to the best fitting curves for previous experience.

The time series of cohort birth probabilities assembled by Whelpton presents a number of clues as to suitable alternative procedures. The influence of wars, depressions, age at marriage, and increasing use of new methods of birth control is suggested in these probabilities. A complex system of variables appears to be interacting. Thus our procedure for obtaining hypothetical future probabilities will have to be expressed mathematically in terms of many variables and the actual computations would have to be carried out on a computer. In particular, we would like to be able to change the probabilities conditional upon changes in certain of the relations among the other variables in the system. The statistical statement that a given income level, say, is correlated with a set of probabilities is not enough. We need to know how the probabilities change if the income levels change.

Let us look more closely at the characteristics of this system of variables. Much simplification is achieved if we recognize that variables can affect the total number of births in two distinct ways. First, they may change the number of women to whom the probabilities of birth apply, the so-called population at risk; and, second, they may change the birth probabilities themselves. Variables that should be taken account of through the first possibility include mortality and marriage. The empirical time series of birth probabilities developed by Whelpton is relatively uncontaminated by mortality, but unfortunately variations in age at marriage and duration of marriage implicitly enter into these data. We would all be grateful for a procedure that would remove the influence of marriage and present us with a revised set of birth probabilities by cohort, by age, by number of previous children. (I would settle for the removal of age at first marriage, which has its greatest influence upon the first two children.) In other words, I am arguing that the probabilities of mortality and marriage are statistically independent of the birth probabilities. Thus, I suspect that the influence of wars upon numbers of births can be largely expressed through the disruption of marriage (although there is a small effect in United States data that appears to stem from the influence of the Selective Service Act deferments for fathers, contrary to my argument).

Let us now turn our attention to those variables that appear to influence directly the birth probabilities. Two sets of variables appear most significant. The first set of variables helps us to predict the diffusion of various new methods of birth control and family limitation practices throughout the population. The second set of variables helps us to predict the behavior of those persons in the population for whom such family limitation practices are acceptable. Some variables appear in both sets but in rather different relationships. The first set of variables enables us to predict the proportion of planners as dependent upon time; the second set of variables enables us to predict the behavior of the planners.

In effect, by dividing the population into planners and nonplanners, we are dividing the birth probabilities into two components, one of which is dependent on time and one of which is not. For the former component we must also state the form of the dependence upon time. Further, the proportion of persons in the two components is also dependent upon time.

Note that this treatment of the variables is consistent with the gross historical changes in Whelpton's data. The probabilities of an additional child for women with five or more previous children have continuously fallen throughout the time period, revealing little influence of such major social catastrophes as wars or depressions. To me this suggests the influence of the steady continuous diffusion of family limitation practices. The probabilities of an additional child for women with zero, one, or two previous children first declined, then increased. To me this suggests that the time dependent component, the behavior of the planners, was influenced by social variables to decrease, then increase. The influence of these social variables is stronger in the more recent period as the proportion of planners in the population is much larger.

Now let us consider the next question. How are we going to obtain hypothetical future probabilities from the considerations above? The simplest procedure is to obtain two sets of future probabilities corresponding to the two components of the population defined above, then to obtain a single set of probabilities as a weighted combination of these two. This procedure is equivalent to first projecting the two components of the population separately, with allowance for changes in the number of persons in the two components, then adding.

We shall consider separately the problems of obtaining birth probabilities by cohort, by age, by number of previous children for the nonplanners and then for the planners. Then we shall consider procedures for estimating the diffusion of planning.

The birth probabilities of the nonplanners will be regarded as functions of biological variables and social variables independent of time (customs). In each society we expect to find somewhat different effects of these biological and social variables. Thus, for our immediate purposes of projection we would have to estimate these probabilities anew in each society. However, if we employ past data upon a large sample of experience of the nonplanners, then we may, by our assumption of time independence, substitute these numbers directly into our projection model.

Our confidence in these estimates would be greatly improved if a series of comparative studies of societies enabled us to pinpoint the influence of the biological and social influences more precisely. We would like a prediction equation with constants to be fitted to the fixed effects of biology and custom. To begin such a research program, we should estimate these probabilities for the nonplanners in several societies with different social customs. To my knowledge the only source of such data currently is the Hutterite data. Given such data, we would attempt to summarize the whole system of probabilities by classical curve-fitting procedures with parameters defined for the biological and social influences. Classical procedures are appropriate here as we assume independence of time. Further, relatively small samples spread over a number of years can be used, the data can be combined to estimate the probabilities for a single cohort, and we can differentiate age by five-year intervals (of course, a preliminary test of the time independence should be carried out before the data are combined).

The birth probabilities for planners are far more difficult to deal with. We assume that these probabilities vary with time—they may increase or decrease. We must seek relationships with these probabilities and other variables that may usefully be assumed to be independent of time, at least for such periods of time as the decade.

Suppose we consider the birth probabilities for a single cohort expressed as functions of certain other variables dependent upon time. There will be generational effects that we will want to treat as constants in a manner analogous to the constants for social customs in different societies. Laying aside these specific generational effects for the moment, let us turn to those relationships that will influence every generation. It is my conviction that we need to express these relationships in two steps. First, we must develop an explicit mathematical decision theory at the level of the household that yields appropriate birth probabilities when its parameters are estimated. But, second, we need to ascertain the consequences of aggregating these decisions across segments of the society. Many variables that loom large for separate households become random errors if the aggregation is appropriately carried out.

Let us consider two classic types of social variables in order to illustrate these relationships. We shall consider the effects of income and of prestige. At the level of the household we shall argue that fertility planning behavior depends upon expected future income. Thus we must predict the manner in which expected future income varies over time, and we must express the consequences of expected future income for birth probabilities. The manner in which expected future income varies over time is likely to vary substantially according to other social variables, occupation, for example. Therefore, it would be best to estimate this relationship within categories defined by the other social variables, within broad occupation classes, for example (taking account of the way in which money comes in such as the contrast of salary and self-employment in the construction of the occupation categories).

In the case of prestige effects we also expect great variations over the whole society. The relevant "Joneses" are to some degree locally defined—one's next door neighbors on the stratification ladder (despite the great influence of the mass media in blurring the distinctions between the lower middle class and the upper lower class). Here again, we must regard the society as composed of segments within which we estimate such effects. The religion variable is of interest here. We would expect that prestige effects would influence standards of ideal family size directly, thus leaving us with the problem of estimating the influence of ideal family size upon the birth probabilities.

We shall return to these very complicated issues in a highly simplified form at the end of the paper. Let us now consider the diffusion of planning. Direct estimation from social psychological interviews is possible, albeit in a highly sketchy form. However, we have additional information. Family planning attitudes and practices tend to diffuse through close personal relationships. These relationships are not randomly distributed throughout the population but are highly structured—we may describe such structures as patterns of social distance. Elsewhere I have argued that the social distance between any two segments of the population is approximately indexed by the probability of intermarriage between those two segments.[3] If we know the point at which an innovation appears in a society, and if we know the probabilities of intermarriage among the segments of the society, then, under the assumption of a uniform rate of diffusion throughout society, we can predict the rank order of the innovation

[3] James M. Beshers, *Urban Social Structure* (New York: Free Press, 1962), especially chap. vii and the Appendix.

in each segment in the system—in our case the percent of planners in each segment at any given time. If we could further estimate the rate of diffusion, then we could predict the percentage of planners in any segment at any given time. This is a very important problem for mathematical and empirical research.

There is a straightforward procedure for testing all three aspects of this cohort projection model on United States data. If we can obtain probabilities for the nonplanners, if we can estimate the parameters for the planners, and if we can estimate the diffusion of planning—the latter two from 1900 to the present—then we should be able to deduce a composite set of birth probabilities that are estimates of the probabilities empirically tabulated by Whelpton. Such a test of a crude projection model is not technically difficult if computers are available.

Reviewing the paper to this point, a rationale for cohort projection models has been presented, the main parameters in such models have been identified, and an empirical program of research for estimating these parameters has been sketched out. Nevertheless, I suspect that most of you doubt that these parameters can be given precise mathematical statement and therefore you doubt that this model can be represented in a computer system. After all, no demographer has explicitly introduced social variables into a projection equation, despite the fact that empirical research has revealed the significance of these variables many years hence. Therefore, I will give you a few simplified examples of these equations.

Let us take as our dependent variable the average number of births to women who have completed their fertility, the classical gross reproduction rate. How can we use the theoretical notions above to predict this number? We shall start with a two-class social system, $C_1$ is the upper class and $C_2$ is the lower class.[4]

Let $x_1$ = average number of births for class $C_1$ and $x_2$ = average number of births for class $C_2$. Let $\rho_1$ = the proportion using birth control in class $C_1$ and $\rho_2$ = the proportion using birth control in class $C_2$. Let $\epsilon_1$ be the average completed family size of users in class $C_1$ and $\epsilon_2$ be the average completed family size of users in class $C_2$. Let $\alpha$ be a constant representing the average completed family size of nonusers, assumed identical for both classes. Then, at any given time, $x_1 = \alpha(1 - \rho_1) + \epsilon_1\rho_1$; $x_1 = \alpha(1 - \rho_2) + \epsilon_2\rho_2$. The average family size for the total population is obtained by the weighted sum of $x_1$ and $x_2$ that takes into account the numbers of families in the two social classes.

We expect that the parameters $\rho$ and $\epsilon$ are variables with respect to time. By substituting time series values for $\rho(t)$ and $\epsilon(t)$, we can "simulate" the history of social class differentials by family size and obtain the several different correlations that have been observed historically. If we assume that $\rho_1$ and $\rho_2$ both monotonically increase from zero to one, that $\rho_1(t) \geq \rho_2(t)$ for any $t$, and that $\epsilon_1(t) \geq \epsilon_2(t)$ for any $t$, then we can produce "realistic" numbers. The magnitudes of $\rho$ depend upon the diffusion process within and between the two social classes. Thus we assume that $C_1$ first adopts birth control. Further, we assume that $C_1$, as a wealthier class, has a larger desired family size among users than $C_2$. In the following table $\rho_1$ and $\rho_2$ are arbitrarily incremented, while $\epsilon_1$ and $\epsilon_2$ are held constant.

Note that class and average family size are first negatively correlated (wealthier class has smaller family), then positively correlated.

In expressing the rate of diffusion, we might like to include the influence of varying "densities of interaction" within social classes as well as boundaries between social classes. One way to do this

[4] The same mathematics can represent social classes with mobility probabilities, or geographical areas with migration probabilities, or a combination of the two. Thus the empirical data in the Blau and Duncan social mobility study and in the Goldberg rural influence research can be treated as special cases of the formulation.

| | $\epsilon_1 = 4$, $\epsilon_2 = 2$, $a = 7$ |
|---|---|
| $\rho_1 = 0$, $\rho_2 = 0$ | $x_1 = 7.0$, $x_2 = 7.0$ |
| $\rho_1 = .2$, $\rho_2 = 0$ | $x_1 = 6.4$, $x_2 = 7.0$ |
| $\rho_1 = .4$, $\rho_2 = .2$ | $x_1 = 5.8$, $x_2 = 6.0$ |
| $\rho_1 = .6$, $\rho_2 = .4$ | $x_1 = 5.2$, $x_2 = 5.0$ |
| $\rho_1 = .8$, $\rho_2 = .6$ | $x_1 = 4.6$, $x_2 = 4.0$ |
| $\rho_1 = 1$, $\rho_2 = 1$ | $x_1 = 4.0$, $x_2 = 2.0$ |

is to express the rate with class $C_i$ by a constant $k_i$. For example, $\rho_1(t) = 1 - e^{-k_1 t}$, $\rho_2(t) = 1 - e^{-k_2 t}$.

Then we express the boundary between the two classes by lagging the starting time for the second class by a constant $\delta_{12}$. We expect that $\delta_{12}$ depends upon the social mobility rate or intermarriage rate between the two adjacent classes. Thus we obtain $\rho_2(t + \delta_{12}) = 1 - e^{-k_2(t+\delta_{12})}$.

The constants $k$ and $\delta$ may be normalized so that they are free of the influence of class size or total population.

In expressing the behavior of the planners, we would like to combine the influence of income and general values. One way to do this is to make the year-to-year probabilities depend upon income fluctuations and to make the ideal family size a generational constant dependent upon general values. Thus each cohort would have a fixed general value constant. But we allow values to change in the model by varying these generational constants in successive cohorts. Thus the influence of prestige upon change of general values can be represented through the generational constants of the successive cohorts. In effect, each cohort has the goal of ideal family size set as a constant determined by general values, but the achievement of this goal varies from year to year, depending upon income fluctuations.

The whole system, in working order, can take account of the change in numbers of the population and of the influence of these changes.[5]

[5] For examples of systems of this sort, see "Social Status and Social Change," by James M. Beshers and Stanley Reiter, *Behavioral Science*, Vol. VIII, No. 1 (January, 1963).