EXTENDED ABSTRACT

New methodology for multivariate analysis of the total fertility rate and its components based on birth-history data

Robert Retherford, Hassan Eini-Zinab, Minja Kim Choe, Naohiro Ogawa, Rikiya Matsukura

A discrete-time survival model — the complementary log-log (CLL) model — is used to model parity progression from woman's own birth to first marriage, from first marriage to first birth, from first birth to second birth, and so on, with one model for each parity transition. In the model for any particular parity transition, predictor variables include duration in parity (*t*) and woman's age at starting parity (*A)*, as well as socioeconomic characteristics. Base data are birth histories in demographic surveys.

Collectively, the models for the various parity transitions yield model-predicted estimates of birth probabilities by age, parity, and duration in parity, denoted P_{ait} (where *i* denotes parity and *a* = *A*+*t*), by socioeconomic characteristics. The birth probabilities P_{ait} for a particular set of values of the socioeconomic characteristics allow calculation of a "global life table", the basic dimensions of which are age, parity, and duration in parity. Starting at age 10, women are survived through this life table one year at a time by age, parity, and duration in parity until they reach age 50. As in the usual calculation of the TFR as the sum of ASFRs, mortality is ignored in the calculation of the global life table. The global life table yields estimates of PPRs, ASFRs, mean and median ages at first marriage, mean and median closed birth intervals, mean and median ages at childbearing (both overall and by child's birth order), TFR, and TMFR. (TMFR is actually a total ever-marital fertility rate, but for simplicity we refer to it simply as a total marital fertility rate.) Because the P_{ait} are multivariate, the global life table is also multivariate, as are all measures calculated from it.

A discrete-time survival model, such as the CLL model, is applied not to the original "person sample" but instead to an "expanded sample" of person-year observations created from the original person observations. The expanded sample makes it easy to include time-varying predictor variables in the CLL model. For example, if a person moves from rural to urban, some of the person-year observations created for that person are coded as rural and some are coded as urban. The CLL model can also handle time-varying effects of predictor variables, by interacting socioeconomic variables with *t* or some function of *t*, such as *t* and t^2.

The CLL model handles left-censoring as well as right-censoring (Allison 1995). This enables application of the model to period data as well as cohort data. In our test application to Philippines 2003 DHS data, "period" is defined as the 5-year period before survey, and "cohort" is defined as the earlier lifetime experience of women age 45-49 at time of survey. Previous studies have applied discrete-time survival models to cohort data. A major methodological innovation in our work is the application of the CLL model to period data. This is done by treating person-year observations before or after the period of interest as censored. Otherwise the application of the methodology is the same in the period and cohort cases. The only difference is how the expanded person-year data set is constructed. Our illustrative application to Philippines 2003 DHS data, which is ongoing, includes both period estimates and cohort estimates.

In the application of the methodology to Philippines DHS data, the form of the model is basically the same for each parity transition, except that, in the case of transition from birth to first marriage (B-M), the model is truncated at 30 years of duration in parity (the difference between the beginning and ending ages of 10 and 40), whereas in the case of higher-order transitions (M-1, 1-2, and so on) they are truncated at 10 years of duration in parity. In the former case a "failure" is a first marriage, and in the latter case it is a next birth of specified birth order. First marriages after age 40 and next births after 10 years duration in parity are rare and are ignored. (Other cutoffs could also be used, but these are appropriate for the Philippines.) Two socioeconomic predictor variables are included in the Philippines analysis: urban-rural residence (specified by a dummy variable *U*) and education (specified by dummy variables *M* and *H*, representing medium and high education with low education as the reference category), as assessed at time of survey. These variables are treated as time-invariant, due to the lack of information about their values in each earlier year before the survey.

In the case of the birth-to-first marriage (B-M) transition, the underlying model is

$$P = 1\square\ \mathrm{exp}\{\square\mathrm{exp}[a + b_1T_1 + b_2T_2 + ... + b_{29}T_{29} + U(c{+}dt{+}et^2) + M(f{+}gt{+}ht^2)$$

$$+ H(j{+}kt{+}mt^2) + nUE]\} \qquad (1)$$

where *P* is the predicted value of the probability of failure (also called the discrete hazard) in a duration interval (failure being a first marriage in this case); $T_1, ..., T_{29}$ are 29 dummy variables representing the first 29 of 30 duration intervals; *t* is a counter variable (equal to 1, 2, ..., 30) that also denotes duration interval; *a* is an intercept term (implying that P = 1-exp[-exp(a)] for the 30th duration interval when all predictors equal zero), and $b_1, ..., b_{29}, c, d, e, f, g, h, j, k, m,$ and *n* are coefficients to be fitted (along with the intercept *a*) to the data. The fitting is done by maximum likelihood (Allison 1995). Although, for higher-order transitions (i.e, higher than B-M), the birth histories in the 2003 Philippines DHS (as in all DHS surveys) are specified by month, we aggregate months into years. This is done because monthly data sometimes result in empty cells (e.g., there are no births in the month following a previous birth), in which case the maximum likelihood estimation procedure for fitting the model does not converge to a solution. In conformity with usual DHS practice, the month of survey, being an incomplete month for most women, is omitted from the person-year data sets. The 5-year period before survey then includes the 60 previous months.

In equation (1), effects of socioeconomic predictor variables are specified as time-varying. For example, the effect of a one-unit increase in *U* (from 0 to 1) — i.e., the effect of urban relative to rural — is to multiply the underlying continuous-time hazard of a first marriage for rural by exp($c{+}dt{+}et^2$), where exp($c{+}dt{+}et^2$) is the relative risk.

A time-varying specification of the effect of *U* on the probability of first marriage is necessary because the effect of urban residence, relative to rural residence, is to lower the probability of first marriage at younger ages and increase it at older ages (because urban marriages tend to be postponed to later ages, relative to rural marriages). Thus the effect of urban residence on the risk of progression to first marriage is not constant over duration in parity; i.e., the effect is not proportional. Similarly, the effect of education is modeled as time-varying, because the effect of more education is also to lower the probability of first marriage at younger ages and raise it at older ages. At higher-order parity transitions, for similar reasons as well as other reasons, the effects of *U*, *M*, and *H* on the probability of next birth are also modeled as time-varying, again with a quadratic specification.

The set of predictor variables on the right side of equation (1) also includes a term *nUE,* where *E* is a dummy variable representing two categories of education, with the reference category defined as low education and the second category defined as medium or high education. The 3-category specification of education is not feasible in this interaction because of the small number of rural women with high education, which leads to convergence problems when fitting the model to the data. The term *nUE* representing interaction between residence and education is needed because the effect of education on parity progression is likely to differ for urban women and rural women. Education is dichotomized only in this particular interaction term, not elsewhere in equation (1), where education continues to be specified in three categories as low, medium, and high.

For transitions higher than birth to first marriage (B-M), the underlying model is

$$P = 1\square\ \exp\{\square\exp[b_0 + b_1T_1 + b_2T_2 + \ldots + b_9T_9 + A(c_0 + c_1\,t + c_2\,t^2) + A^2(d_0 + d_1\,t + d_2\,t^2)$$

$$+ U(e_0 + e_1\,t + e_2\,t^2) + M(f_0 + f_1t + f_2\,t^2) + H(g_0 + g_1t + g_2\,t^2) + U(h_0 + h_1A + h_2A^2)$$

$$+ M(j_0 + j_1A + j_2A^2) + H(k_0 + k_1A + k_2A^2) + mUE]\} \qquad (2)$$

An A^2 term is included as well as an A term, because the rise and fall of fecundability as age increases suggest that the effect of starting age on parity progression will be non-linear, and that a quadratic specification of starting age will adequately capture this non-linear effect. The effects of both A and A^2 are specified as time-varying because the effects of A and A^2 on parity progression change as duration in parity increases. Not only biological influences (fecundability) but also behavioral influences play a role in the interpretation of the effect of starting age on parity progression. One example of such a behavioral influence is that couples are more likely to settle into a life style with few or no children the longer they delay marriage and childbearing. Our methodology and data do not allow separate measurement of these biological and behavior influences, however.

Basic global life table calculation formulae relating to the B-M transition are:

$$S_{0,0,0} = 1 \qquad (3)$$

$$S_{a,0,t} = S_{a,0,a} = S_{a\text{-}1,0,a\text{-}1}(1\text{-}P_{a\text{-}1,0,a\text{-}1}) \quad \text{for } a > 0 \qquad (4)$$

$$f_{a,0,t} = f_{a,0,a} = S_{a,0,a}\,P_{a,0,a} \qquad (5)$$

where $S_{a,0,a}$ denotes the probability of surviving (not yet having had a first marriage) to age a, $f_{a,0,a}$ denotes the unconditional probability of a first marriage between ages a and a+1, and age 10 is translated to age 0 in order to simplify the notation.

For higher-order parity transitions, basic formulae are:

$$S_{a,i,0} = \square\ (S_{a,i\text{-}1,t}\,P_{a,i\text{-}1,\text{t}}) \quad \text{for } a > 0 \text{ and where the summation is over } t \qquad (6)$$

$$S_{ait} = S_{a\text{-}1,i,t\text{-}1}(1\text{-}P_{a\text{-}1,i,t\text{-}1}) \quad \text{for } a > 0 \text{ and } t > 0 \qquad (7)$$

$$f_{ait} = S_{ait}\,P_{ait} \qquad (8)$$

where f_{ait} now denotes the unconditional probability of an (i+1)th birth between a and a+1 and between t and t+1.

Once global life table values of S_{ait} and f_{ait} have been calculated from these formulae, starting from the model-predicted values of P_{ait}, it is straightforward to calculate PPRs, ASFRs, mean and median ages at first marriage, mean and median closed birth intervals, mean and median ages at childbearing (both overall and by child's birth order), TFR, and TMFR.

Unadjusted and adjusted estimates of TFR (or any other of the above measures) by categories of a predictor variable are calculated using the logic of what is sometimes referred to as multiple classification analysis (MCA) (Andrews, Morgan, and Sonquist 1969; Retherford and Choe 1993). In MCA, “unadjusted” means “without controls”, and “adjusted” means “with controls”.

For a particular parity transition, unadjusted values of the discrete hazard function P_{ait} by urban/rural residence, for example, are calculated from a CLL model that includes *U* as the sole socioeconomic predictor variable. Thus, in the case of equation (1) for the B-M transition, one drops terms containing *M, H,* or *E.* Values of P_{ait} for urban are then calculated by setting $U = 1$ in the estimation equation, and values of P_{ait} for rural are calculated by setting $U = 0$ in the estimation equation.

Adjusted values of P_{ait} by urban/rural residence for the B-M transition are calculated from equation (1) with all of the predictor variables *U*, *M*, and *H* included. Education, represented by *M* and *H* or by *E*, is viewed as the control variable. To obtain adjusted values of P_{ait} for urban, one sets $U = 1$ and *M* and *H* (note that the value of *E* is determined by the values of *M* and *H*) equal to their interval-specific mean values in the data set to which the CLL model is fitted. (In this context, “interval” means duration interval; each parity transition has its own set of interval-specific — i.e., duration-in-parity-specific — mean values of *M* and *H* derived from the person-year data set for that parity transition.) To obtain adjusted values of P_{ait} for rural, one sets $U = 0$ and *M* and *H* equal to the same interval-specific mean values that were used to calculate the adjusted discrete hazard function for urban. In this way *M, H,* and *E* are held constant or “controlled” when *U* is varied from 0 to 1.

Unadjusted and adjusted global life tables and measures derived from them (TFR and its various components) are then calculated from the unadjusted and adjusted values of P_{ait} by urban/rural residence.

Despite the complexity of the underlying statistical models, the final tables of unadjusted and adjusted estimates of the above measures have a simple bivariate format that is readily understood by non-statisticians. The simplicity of the final tables is a major selling point for the methodology.

Using data from the 2003 Philippines DHS, we have done some of the calculations already (ASFRs, PPRs, and TFRs tested on cohort data), and the results look good. Work is ongoing.

Note that, because the global life table is internally consistent, TFR calculated from ASFRs and TFR calculated from PPRs have the same value, which is most easily calculated, however, as

□ f_{ait}, where the summation is over *a, i* (except parity 0, which is omitted because failures are first marriages instead of births), and *t*. Note also that if the socioeconomic variables are omitted from the models, the global life table yields estimates of TFR and its components for the population as a whole.

This research is a major extension of earlier work based on models that specify parity and duration in parity but not age. The earlier work did not produce estimates of ASFRs and mean and median ages at childbearing (both overall and by child’s birth order). A paper based on the earlier work (“Multivariate analysis of parity progression-based measures of the total fertility rate and its components”) is forthcoming in *Demography*. The *Demography* paper, which does not employ the global life table concept, is based on a longer working paper that can be downloaded at http://www.eastwestcenter.org/fileadmin/stored/pdfs/POPwp119.pdf.

REFERENCES

Allison, P. 1995. *Survival Analysis Using SAS: A Practical Guide.* Cary, N. C.: SAS Institute Inc.

Andrews, F., J. Morgan, and J. Sonquist. 1969. *Multiple Classification Analysis*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

Retherford, R. D., and M. K. Choe. 1993. *Statistical Models for Causal Analysis.* New York: John Wiley.