66 INNOVATIONS AND CHALLENGES IN LARGE-SCALE POPULATION-BASED DEMOGRAPHIC SURVEYS

Challenges in estimating multi-state expectancies from large-scale cross-longitudinal surveys

Brouard N and Sun F

Abstract

This proposition of communication introduces innovative ways to address the challenges in estimating multi-state expectancies using large-scale longitudinal surveys. We first discuss the challenges in designing longitudinal surveys as well as in examining prevalences of health outcomes considering the rapidly changing health states (i.e., the dynamics of health changing that are captured in the longitudinal surveys). We introduce the method of interpolated Markov chains in order to estimate the incidences of change between states based on the probability of change over a small time interval that is estimated by multinomial regression as a function of age and covariates. More importantly, we discuss how Powell's algorithm which was useful in optimizing the multinomial regression function to estimate health expectancies for different states based on up to 30 covariates, did not converge when adding more states, covariates or interactions because the likelihood function to be maximized depends on more than 200 variables. We then test the Brent/Praxis algorithm which uses the principal component analysis method to estimate information matrix about probabilities of change, and found that the algorithm is also useful and time-efficient.

From cross-sectional surveys, such as population censuses, which ask an entire population about demographic questions such as age, sex, household, work, housing, etc., even if they are repeated every 5 years or so, we can appreciate trends in socio-demographic changes, such as the increase in nuclear family dissolution as well as stepfamily, but we can hardly make predictions compared to multi-round surveys where the same people can be interviewed at different times in their lives. However, this has been and still is a debate, since quality longitudinal surveys are much more expensive due to deaths, migrations, and any cause of attrition.

The main advantage of cross-sectional surveys is a better estimation of prevalences or proportions of the population in a specific category such as age group, marital status, employment status, number of children, etc.

The main advantage of longitudinal surveys is the estimation of transitions between states. For example, classical mortality analysis combines information on a person's age in a population census with deaths recorded in vital statistics in subsequent years to measure age-specific mortality rates (mortality incidence) which are then processed in a so-called period life table with its classical indicator, life expectancy.

In a hypothetical closed population (without migration), the cross-sectional survival prevalence could be estimated by dividing the population of a cohort estimated in a census by the corresponding births of that same cohort. Combined over all ages, the resulting

“cross-sectional” mortality index, called CAL, is more realistic but less accurate than current life expectancy in predicting how long new generations might live.

We face similar challenges in various aspects of modern demography and this is why longitudinal surveys, or better said, cross-longitudinal surveys, are mandatory to produce not only cross-sectional prevalences but also period prevalences based on age-specific periodic incidences.

However, there are two challenges in designing longitudinal surveys. A first challenge is not only the classical size, age range, scope, etc., but especially its space. For example, when studying in a two-round survey the mortality of centenarians, if the second round occurred one month after the first round, the number of death cases is too small, and if the second round occurs five years later, most centenarians will have died. The optimal interval between the two surveys could then be estimated at 6 months. For multi-round surveys on aging and disability, such a time interval cannot be easily established but results mainly from previous experiences. For example, the Longitudinal Study on Aging (LSOA), which proved for the first time that a return to a state without disability was possible at an advanced age, was spaced every two years from 1984 to 1986, 1988 to 1990. Such a two-year interval was also reproduced for LSOA II (1994-1996-1998-2000). The American Health and Retirement Study (HRS), still ongoing, has also been spaced every two years since 1996.

Another important challenge is to take into account the actual time between two interviews which varies according to the person and sometimes can skip one or more surveys. For example, for the HRS, even if its quality is high, 2.5% (up to 4.5% at 50-54 years) of the sample are not interviewed during the second interview and the average delay between two waves varies from 1.9 years to 2.3 years. As for population estimates on January 1st instead of the census date, statistical methods must be proposed that take into account the exposure time.

These two challenges result in other problems regarding using the longitudinal survey to estimate incidences of health outcomes. One problem is reversible states, such as marriage or disability. At intermediate ages, the incidences of disability entry and recovery in one year are of similar magnitude. Thus, a higher incidence of disability and higher recovery are as likely as a lower incidence of disability and recovery, for the same number of cases (changes) observed in each state in two successive surveys. There is therefore a very high correlation between the two incidences, which precludes calculation of the number of times entering disability.

More specifically on healthy life expectancy, there is a very large body of research on whether the prevalence of disability is declining, or whether the proportion of time spent in good health is declining or not. And the conclusions are mainly due to the methods and not just the data, which adds to the confusion. Probably the most significant challenge and confusion in the different methods used by demographers to examine the changes in prevalence of health outcomes over time concerns the multi-state life table.

Some authors use the discrete Markov chain approach assuming that the state change occurs only once (at the beginning or end of the interval), while other authors prefer the Markov

process, where states are observed only at the interviews but other state changes may have occurred in between.

To find explanations for observed variation between groups, multinomial logistic regression appears to be the most widely used method. The variable to be explained is often the age-specific incidence of state change and the explanatory variables or covariates can be dummy variables as well as quantitative variables. If the covariates themselves change between interviews, for example if the person moved to a different residence or region, the multinomial regression would take into account the residence or region observed at the last observation.

From the estimated incidences by age, demographers use a method similar to that of the life table, in order to calculate the survival function in each state or the time spent in each state for a given value of a covariate.

It is then clear that in a manner analogous to the so-called current life expectancy, which exaggerates the actual life span by taking into account current mortality and not past mortality (which was higher among all generations), such current life expectancies by health status exaggerate the importance of time-varying covariates, by considering them as permanent over the simulated life cycle.

In order to estimate the incidences of change between states when the delays between observations are all equal, we use the method of interpolated Markov chains, developed by Laditka and Wolf. The idea is to model the probability of change over a smaller time interval, such as a month, which is expressed in an elementary matrix. Then assuming the Markov hypothesis, the changes observed between two interviews spaced by h months are summarized by the matrix product of h elementary matrices. Such elementary matrices are assimilated to incidences of change between states and are estimated by a multinomial regression as a function of a specific age (in months) and covariate values.

The contribution to the total likelihood of a person observed initially in state i and, h months later, in state j, is then the element (i,j) of the product matrix.

In order to estimate the parameters of the multinomial regression, the total likelihood of the sample must be maximized. But before discussing the challenge of estimating the maximum of a function of hundreds of variables, let us mention another challenge.

An important societal issue concerns the retirement age in relation to health status. And the British longitudinal study ELSA offered the possibility of analyzing the situation finely and scientifically.

The states of such a model intersect the health state and the employment state. But in the elementary Markov process, a healthy person employed at the first interview and observed as disabled and out of the labor market at the next interview would either have moved into the disabled state while still in the labor market, or first exit the labor market and then move into the disabled state. And so, the incidence of the transition from good health at work to bad health and out of work cannot be estimated in continuous time or even over a one-

month interval because these incidences are too small to be estimated by the likelihood maximizer.

Many large-scale cross-sectional surveys are now increasingly available in Asia, for example in China as well as in Japan or the Philippines. And research is increasingly demanding on fine-grained multi-state statistical analyses with covariates and interactions, challenging not only the methods to be used but also the classical function maximization algorithm which is no longer able to converge in many occasions.

Powell's algorithm, published in a well-known book called ``Numerical Recipes in C'' was very effective at optimizing a function with up to 30 variables but a model with 3 living states: pain, mild pain, severe pain, a fixed covariate being the four regions of the United States and a time-varying covariate being residence: urban, semi-urban, rural, and all possible interactions with age, as well as region multiplied by residence, requires maximizing a function of over 200 variables and Powell's algorithm no longer converged.

After reviewing the various algorithms from the 70s and 80s, sometimes unpublished or poorly published, we retyped and tested an algorithm written in Algol W, by Brent, which seems to converge successfully in a reasonable time. But it remains a challenge, because this algorithm named Praxis for Principal axes, uses the principal component analysis method that could directly provide the information matrix thus eliminating the calculation of the Hessian matrix and its inverse which are very time-consuming.

A final discussion that emerges recurrently among researchers using such multi-state methods concerns the use or not of cross-sectional prevalences in the weighting of life expectancy based on status and our advice is to use the so-called period prevalence. That is to say, to obtain multi-state life expectancies that are estimated only from incidences and therefore much more predictive of the future.