Estimating IMR from DHS full birth histories in the presence of age heaping

Julio E Romero Prieto, Banco de la República - Colombia
Andrea Verhulst, University of Pennsylvania
Michel Guillot, University of Pennsylvania and INED

1. Introduction

The Infant Mortality Rate (IMR) is a measurement of the social progress and the health status of a population. In spite of the drastic decline of the last decades, IMR is still a major issue in areas of the world lacking reliable records for a proper monitoring. In these contexts, retrospective estimations are possible from the *full birth histories* of the *Demographic and Health Surveys* (DHS). Although this is a remarkable and systematically collected source of data, it is not perfect and must be used with extra caution. In addition to sampling errors, full birth histories are retrospective information sometimes affected by recall bias [1]. For instance, some ages at death are usually rounded or misreported producing systematic errors such as heaped distributions. For example, the most outstanding heaping is identified at the 12 months of age. Because of its proximity, misreported ages at death might be responsible for some bias on the direct estimation of the IMR.

Because of the age heaping, survey estimates of IMR are rarely used and indirect estimations based on Model Life Tables (MLT) are generally preferred [2]. Using the most reliable Vital Records (VR), these models help to predict the IMR for a given level of the Under-5 Mortality Rate (U5MR), which is assumed to be unbiased. Although this is a very convenient solution, this approach has at least one clear limitation: the choice of the pattern of mortality. For example, the West, East, North, and South family model of the MLT proposed by Coale & Demeny (1966). A *family model* needs to be chosen and it is not that clear which of them is the most suitable for a given population. Inasmuch as the range of estimated values of IMR for a given level of U5MR remains wide, the discretion on the choice of a family model is one of major concern. In general, when a family model is chosen a mortality pattern is imposed. Hence, any salient characteristic related to the pattern of mortality of the population to be adjusted are simply disregarded. If these characteristics were not ignored, the IMR could be estimated more efficiently, which is the contention of this extended abstract.

A more efficient estimation of the IMR requires two fundamental changes. First, the choice of the pattern of mortality must be directly informed by the data. Hence, in addition to the U5MR, relevant information should be considered in order to identify the pattern of mortality. Second, to deal with more information the model should have some more details on the mortality schedule at early ages. This is assuming that, even if survey estimates are problematic around the 12 months of age, unbiased probabilities of dying can be estimated at any other ages.

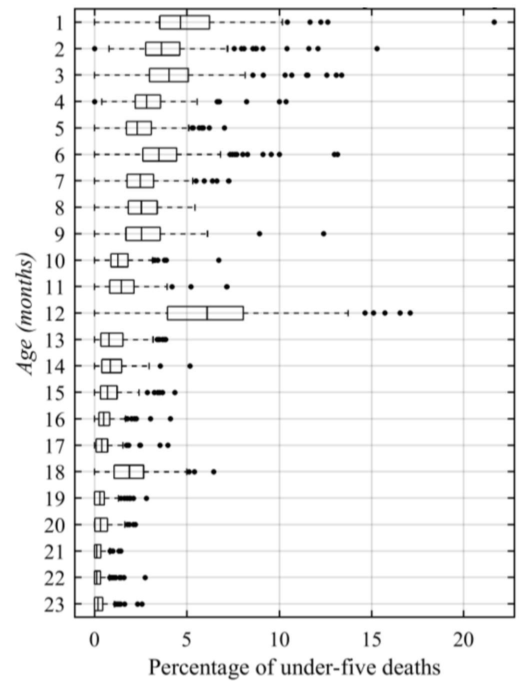
This extended abstract describes an indirect method to estimate the IMR from survey data, using a new approach that accomplishes these properties. The parameters of this new model (*k-model*) were estimated from reliable records having some details by weeks, months, and trimesters of age. This is a relevant feature, given that the most dominant approaches on the literature of MLT have not considered these age breakdowns [3, 4]. Contrary to the discrete choice of family models, the pattern of mortality has been assessed continuously. Following the approach of Wilmoth *et al*. (2012), the pattern of mortality relies in one additional parameter complementing the level of mortality [5]. Hence, the choice of the pattern can be directly informed from the data.

This procedure was applied to a total of 261 DHS. Preliminary results indicate that 82 surveys have a satisfactory fitting and the *k-model* allows a proper correction of the IMR. In general, there is little difference between the direct and the k-adjusted IMR (+/- 5 %). This suggests that most age heaping around age of 12 months arises from downwards age transfers, which have no impact on the direct estimation of IMR. However, the conventional MLT approach generates larger adjustments (mostly downwards), that seem to be unwarranted when compared to the findings of this abstract. Therefore, direct estimates of the IMR derived from DHS are far more robust than it was thought, and could be directly estimated without model adjustment in many cases. Although survey estimates of the IMR are generally considered underestimated, the evidence shows that this effect is limited.

2. Data

Using full birth histories, mortality schedules were estimated for a total of 261 *Demographic and Health Surveys* (*births recode*). In addition to the date of the survey, each record includes the date of birth, and if deceased, the age at death. Births were reported by calendar years and months; while ages at deaths were reported in days of age if neonatal (less than 28 days), in months of age for children below two years, and finally in years for those older than two but less than five years of age. Hence, probabilities of dying were calculated directly through a Lexis diagram under the assumption of a synthetic cohort and producing period life tables of ten years of exposure for both sexes combined [6, 7]. Therefore, relevant information must include the births reported in the last fifteen years and the deaths reported in the last ten years prior to the date of the survey. Inasmuch as these surveys were conducted during several months, each survey was adjusted as all mothers were interviewed on the same day. Records with incomplete information were not included (listwise deletion) Finally, life tables were calculated by months of age during the first two years of life and by years after the age of two. However, the first month of life is a 28-day length (neonatal mortality), while the second month should be longer to complete a year of age of to 365.25 days long (infant mortality).

Figure 1: Age distribution of postneonatal and second year deaths

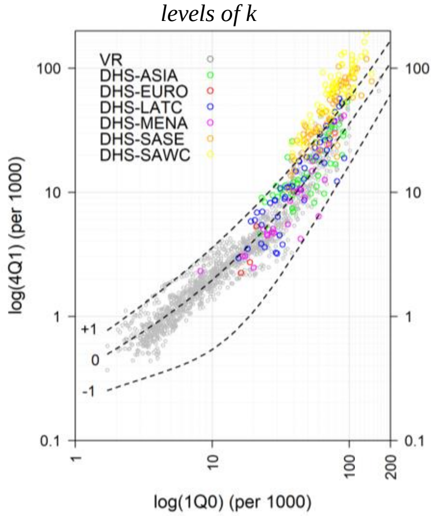


Since full birth histories are self-reported data, direct probabilities are affected by the lack of precision in reporting the age at death. Misreported ages at death introduce bias by heaping the distribution of deaths at some specific ages. According to Figure 1, the percentage of deaths at the age of 12 months is unusually higher than the percentage of deaths at adjacent ages (11 or 13 months). In a very small magnitude, heaping also occur at the age of six and 18 months. One potential pitfall is to underestimate the IMR when the ages at death were approximated by excess.

3. Methods

Figure 2: Infant vs. child mortality

Data and model predictions for different levels of k



DHS estimations were compared to a *k-model* previously estimated from the most reliable records of a selection of country-years. This model is described in more detail in another paper, but some of its most relevant characteristics are briefly introduced in this extended abstract. For a given level of mortality, any MLT should describe how to relate the probabilities of dying at different ages. For this reason, it is possible to estimate the IMR from the U5MR, if this is the single piece of information; which is also related to the level of mortality. However, more efficient estimates (with less uncertainty) can be obtained if additional information would be available; in this case related to the pattern of mortality. The *k-model* has been developed on the basis of newly collected data which is an inherent advantage compared to the old-school approach based on MLT. Although the *k-model* mostly relies on the European experience (as several models), the expectation is to increase the dataset used to fit the model in order to reproduce more general solutions.

Constructing from the approach proposed by Wilmoth et al. (2012), the *k-model* incorporates a continuous parameter k used to introduce deviations from a central tendency that must be driven by particular patterns of mortality (dashed lines in Figure 2). This parameter is adjustable and can be optimally calculated when the U5MR is related to more than one independent probability of dying. Hence, the *k-model* was specifically developed to deal with several age intervals during the first years of life. For a given level of U5MR and k, the *k-model* predicts cumulative probabilities of dying by months during the first year of life, by trimesters during the second year of life, and by years of age until the fifth birthday. The *k-model* is given by equation 1. Using a standard notation, the cumulative probability of dying at the age x, $q(x)$, is assumed to be a log-quadratic function depending on two entry values: *i)* the U5MR, conveniently defined as the cumulative probability of dying at the age of five years, $q(5y)$; and *ii)* the parameter k, which is related to the pattern of mortality. The set of coefficients $[\mathrm{a}_x, \mathrm{b}_x, \mathrm{c}_x, \mathrm{v}_x]$ are predetermined by the model, and they are supposed to vary by ages.

$$ln[q(x)] = \mathrm{a}_x + \mathrm{b}_x \cdot ln[q(5y)] + \mathrm{c}_x \cdot ln[q(5y)]^2 + \mathrm{v}_x \cdot k \quad (1)$$

Although the U5MR was calculated directly from the DHS, the parameter related to the pattern of mortality was estimated by minimizing the absolute error of a fitting function F. As described by equation 2, this function was evaluated twice, comparing the survey estimates with model predictions for given values of U5MR and k.

$$k = \arg\min \left\{ \frac{|F(q(5y),k) - F_{obs}|}{F_{obs}} \right\} \quad (2)$$

Considering that survey estimates are affected by reporting bias around the age of 12 months, these probabilities of dying were excluded in order to estimate the optimal value of k. In particular, the estimation of this parameter only included the differences in the area below the curve of q(x), from zero to 8 months of age and between 21 months and 5 years of age. Thus, the fitting function used to estimate k is given by equation 3.

$$F = \int_0^{8m} q(y)dy + \int_{21m}^{5y} q(y)dy \quad (3)$$

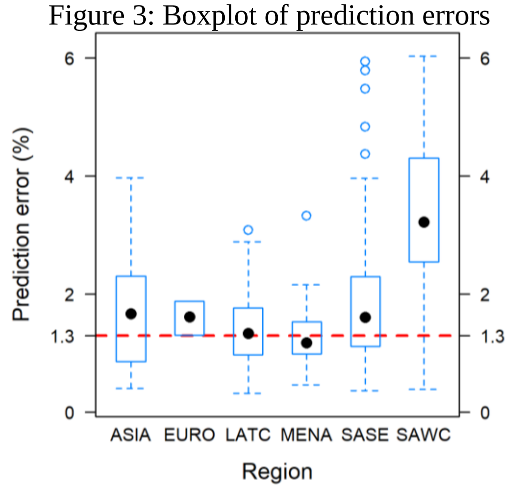
From this perspective, the probabilities of dying at those ages used to fit the model are assumed to be informative of the pattern of mortality of each survey. In this regard, probabilities of dying at the age of 12 months were estimated indirectly for given values of U5MR and k. These estimates were compared to survey calculations of the IMR, in order to dimension the reporting bias of each survey. In order to evaluate the improvement of including additional information related to the pattern of mortality, model predictions of the IMR were also compared with those recovered from a conventional approach using family models.

4. Preliminary results

Table 1: Number of surveys according to the goodness of fit

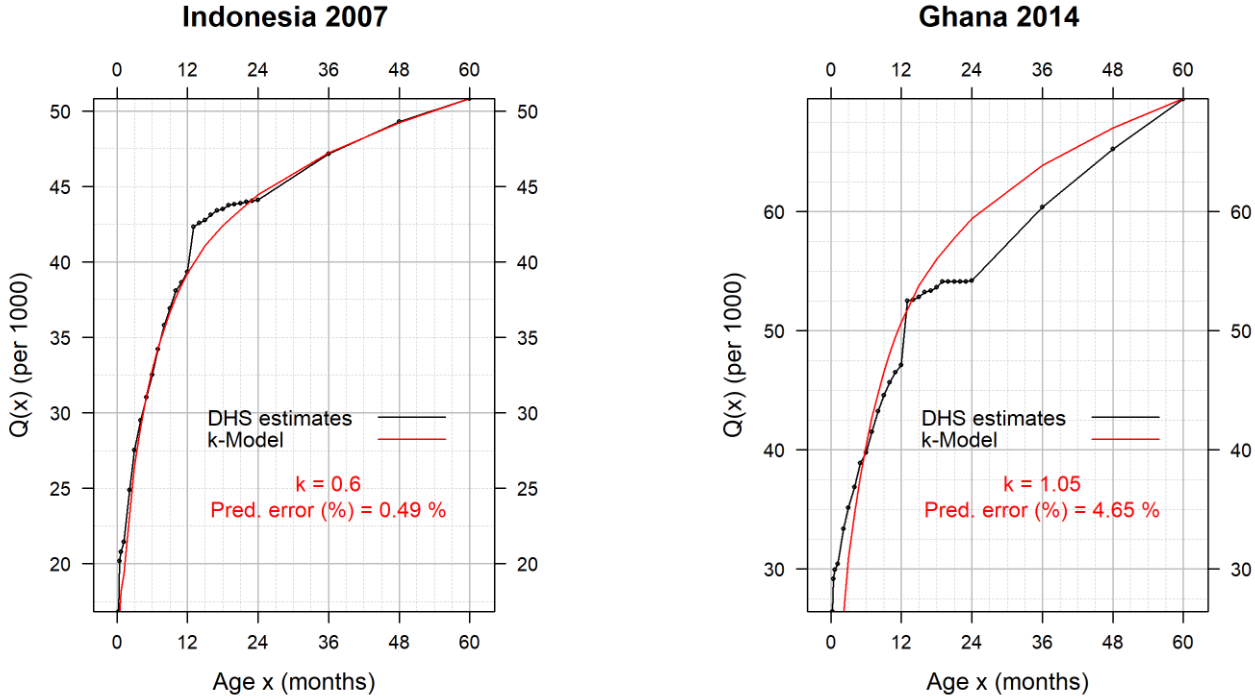
DHS region	*Good*	*Poor*
Asia	17	32
Europe	0	3
Latin America and the Caribbean	23	26
Middle-East and Northern Africa	17	11
Southern and Eastern Sub-Saharan Africa	22	45
Western and central Sub-Saharan Africa	3	62
Total	82	179

Figure 3: Boxplot of prediction errors



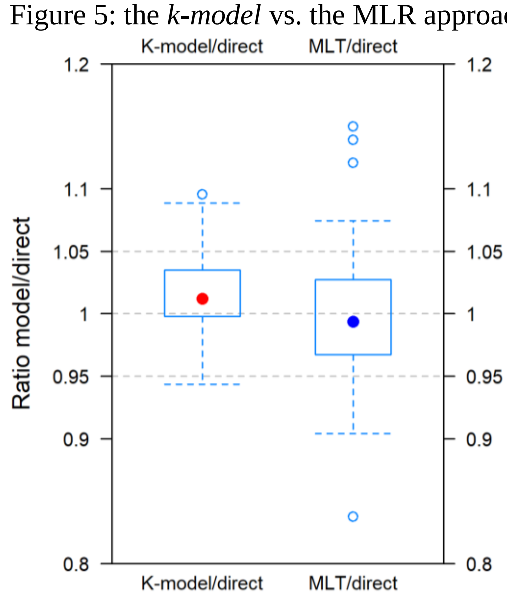
A satisfactory fitting to the *k-model* was considered in those cases producing absolute errors less than 1.3% (equation 2). This threshold corresponds to 90% the best fits observed when applying the *k-model* to the vital records used to estimate the coefficients $[\mathrm{a}_x, \mathrm{b}_x, \mathrm{c}_x, \mathrm{v}_x]$. Given this critical value, which is conservative, a total of 82 surveys produce minimal errors and there is more confidence in the indirect estimation of the IMR. The results of this classification by regions are shown in Table 1 while boxplots of the distributions are presented in Figure 3. Furthermore, the Figure 4 presents two selected surveys contrasting the goodness of fit.

Figure 4: Good vs. poor fitting to the *k-model*
Cumulative probabilities of dying



It has been discussed that misreported ages at death might underestimate the IMR and overestimate the child mortality [1, 6], but preliminary evidence based on 82 surveys suggests that not quite much. There is little difference between direct estimates of the IMR and those corrected by the *k-model* (± 5%). Hence, the heaping in the number of deaths around the age of 12 months arises from rounding down approximations. Although this type of reporting bias introduces spurious fluctuations into the age pattern of mortality, it is not a source of major errors in the direct estimation of the IMR. There are analytical improvements in including additional information related to the pattern of mortality. Compared to the results of the *k-model*, the traditional approach of imposing a family model generates larger adjustments in the estimation of the IMR (mostly downwards), which seem to be unwarranted. The results of this comparison are shown in Figure 5.

Figure 5: the *k-model* vs. the MLR approach



5. Further directions

The final paper will provide clear recommendations on what method to use for each DHS survey. Some alternative procedures will also be considered, such as refining the criteria used to inform the pattern of mortality. In this regard, the expectation is to validate more DHS surveys for direct use or *k-model* adjustment. Taking into account that less satisfactory adjustments were observed in contexts of higher mortality, the *k-model* could also be reformulated in order to reproduce more complex patterns of mortality not yet incorporated into the model. This is by expanding the geographical coverage of the data on which the model is derived. Hence, high quality data collected in other regions would also be included (for example, the *Health and Demographic Surveillance System* and cohort studies).

6. References

[1] K. Hill, «Approaches to the measurement of childhood mortality: a comparative review,» *Population Index,* vol. 57, nº 3, pp. 368-382, 1991.

[2] J. M. Sullivan, S. O. Rutstein y G. T. Bicego, «Infant and child mortality,» Calverton, Maryland. Macro International Inc. http://www.measuredhs.com/pubs/pub_details.cfm?ID=23, DHS Comparative Studies No. 15, 1994.

[3] A. J. Coale and P. Demeny, Regional Model Life Tables and Stable Populations, Princeton: Princeton University Press, 1966.

[4] United Nations, «Model Life Tables for Developing Countries,» United Nations, Department of International Economic and Social Affairs, New York, NY, 1982.

[5] J. Wilmoth, S. Zureick, V. Canudas-Romo, M. Inoue and C. Sawyer, "A flexible two-dimensional mortality model for use in indirect estimation," *Population Studies,* vol. 66, no. 1, pp. 1-28, 2012.

[6] S. O. Rutstein, «Infant and child mortality: levels, trends, and demographic differentials,» Voorburg, Netherlands. International Statistical Institute, World Fertility Survey Comparative Studies No 43, 1984.

[7] J. Somoza, «Illustrative analysis: infant and child mortality in Colombia,» Voorburg, Netherlands. International Statistical Institute, WFS Scientific Report No.10, 1980.