Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2006 : Methodological Issues in Measuring Population Health

2006

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada's Symposium 2006
Methodological Issues in Measuring Population Health

Modelling Life Expectancy at Birth in Small Cities in Canada

Philippe Finès[1], PhD.
Acknowledgements: Jean-Marie Berthelot, Russell Wilkins

Abstract

When Chiang's "standard" method is used, calculating life expectancy for (small) census agglomerations in Canada can produce estimates whose confidence intervals are too wide to be useful. However, we have been able to show that by combining small area estimation methods and simulation methods, we can obtain narrower confidence intervals.

KEYWORDS: Life expectancy, Small area estimation, Mean square error.

1. Introduction

Life expectancy at birth (e0) is often used as a health indicator. It is easy to calculate: all one has to do is use a life table constructed with the number of deaths and the number of person-years at risk during a particular period.

For example, let's look at the Winnipeg census metropolitan area (CMA) (population of 667,515 in 1996), for which death statistics are compiled over a three-year period (1996-1998). The point estimate of e0 (for both sexes combined) is 79.13 years with a standard error of 0.108 years; the 95% confidence interval (CI) for e0 is from 78.91 to 79.35 years. Thus, for a CMA with a population of about 700,000 (or more) for which death statistics are compiled over three years, we obtain a CI of less than half a year. The estimate of e0 is considered precise enough to be used in comparisons between cities or census years and to examine the income-mortality gradient within the CMA.

When we try to do the same for smaller cities, the CI becomes larger. In particular, if we look at census agglomerations (CAs) with a population of less than 100,000 (identified in this context as "small" cities), the CI of e0 becomes too wide to produce useful estimates. Let's return to our example and consider a hypothetical city that is similar to Winnipeg in every respect except that it is 10 times smaller. The estimated life expectancy would be the same, but the standard error would be multiplied by the square root of 10. The CI [78.52, 79.74] would be more than one year.

The overall pattern of mortality in Canada ("overall curve") is well known (Figure 1). It can also be modelled by age. Intuitively, since each city contributes to the overall curve, the latter should in turn contribute to the specific mortality rate for each city. More precisely, it would be ideal if, for a given "small" city, the mortality rates were specific to that city, but the CI's width was reduced as a result of the overall rate's contribution. In other words, our aim is to borrow the observed stability in the overall curve to narrow the CI of e0 for a given small city. The theory of small area estimation (Rao, 2003) is well suited to our purpose. The general idea behind that approach, which has yielded valuable results in economics, education and health in particular, is to replace the estimate obtained directly for the small area in question (in this case, a small city) with a "synthetic" estimate generated with information from another source (in this case, Canada as a whole).

We have used this approach to develop a method of narrowing the CI for CAs in Canada. The structure of this article is as follows: in section 2, after describing Chiang's method of calculating e0, we present the techniques used both to generate an estimate of e0 based on the small area estimation method, and to construct an algorithm for measuring the estimate's quality; in section 3, we describe the data we used and the results we obtained; and in section 4, we present some discussion points.

[1] Statistics Canada, Ottawa, ON, Canada

2. Methods

The CA for which we want to calculate e0 is defined as the "unit of analysis", while the "unit of reference" is the whole of Canada. For discussion purposes, we are interested in calculating e0 for both sexes combined.

2.1. Chiang's "standard" method

We consider 20 age strata (denoted x) that partition age (continuous) between 0 and 100 (see the first two columns of Table 1).

For each age stratum x, we let:

d_x = number of deaths in the 1996-1998 period
n_x = number of person-years at risk (= number of persons in the age stratum in the 1996 Census, multiplied by 3 to produce a denominator consistent with the number of deaths)
m_x = mortality rate = d_x / n_x
e.t(m_x) = standard error of $m_x = \sqrt{[(m_x(1-m_x)/n_x)]}$ (formula based on the binomial distribution)

The set of 20 m_x values forms the basis of the life table (abridged). The table is then used to calculate e0 (Chiang, 1984). It is also possible to compute the variance of e0 [Var(e0)] from the table. Lastly, the 95% confidence interval (CI) of e0 is calculated with the formula **CI = e0 ± 1.96*s.e.(e0)** (where s.e.(e0) = standard error(e0) = $\sqrt{}$Var(e0)).

2.2. Model based on small area estimation theory

When we look at the mortality curvess for Canada as a whole (Figure 1), we find that the curve can be divided into five "groups", each of which can be modelled by age (see Table 1). We also define the **unit** variable, which is set to 1 or 0 depending on whether it is the unit of analysis or the unit of reference.

Within each group, we have (for both the unit of analysis and the unit of reference) the following variables in each age stratum that makes up the group: the age stratum number, the midpoint of the age stratum (denoted avg_age), the number of deaths, the population and the mortality rate for each age stratum (m_x = number of deaths / population).

The following *algorithm* is used to estimate life expectancy:

a. A four-parameter logistic regression is used to predict m_x:

$$\mathbf{\log (m_x/(1-m_x)) = \beta_0 + \beta_1\ unit + \beta_2\ avg\_age + \beta_3\ unit*avg\_age \quad (1)}$$

   At first glance, the number of explanatory variables seems low, but there are similar cases in the literature.

   The estimates of β_0, β_1, β_2 and β_3 produced by the model are denoted b_0, b_1, b_2 and b_3.

b. Using estimates b_0, b_1, b_2 and b_3 and the values of avg_age, we obtain (by setting unit = 1) estimates of the mortality rates m_x' for the unit of analysis. After we apply the logistic regression to the five groups, we have a set of 20 estimates m_x' (where x = 1..20).

c. We estimate the new numbers of deaths $d_x' = m_x'$ multiplied by n_x (holding the populations of the age strata fixed).

d. Lastly, we obtain a new estimate of life expectancy, denoted e0'.

2.3. Measuring the model's quality

It is important to note that there are several steps in the algorithm: first we estimate the 20 m_x, and then we use them to estimate e0 (with the aid of the estimated life table). The problem lies in the fact that the estimated values (the m_x' in this case) are not the final statistic but an intermediate result in the process of calculating the statistic of interest, which is e0'. The variation of e0' as a function of the variation of the 20 m_x' is difficult to capture. There are several possibilities:

1. For a given occurrence, the m_x' may all be very "close" to the (original) m_x, but at the same time they may all be lower (or higher) than the original values. If that were the case, e0' would be appreciably larger (or smaller) than e0.
2. Conversely, each m_x' may be substantially biased relative to the corresponding m_x, but in such a manner that the biases are evenly split between positive and negative values and balance out, leaving e0' fairly close to e0.
3. The other two cases (all m_x' close to the m_x, and e0' close to the original e0; all m_x' biased relative to the m_x, and e0' biased relative to e0) are also conceivable.

In other words, even if we could measure the algorithm's quality with respect to the m_x',it would not help us assess the algorithm's quality with respect to the e0'. Since we are interested in the quality of the life expectancy estimate, we need a yardstick that measures it directly.

For that purpose we use the mean square error (MSE), which combines the bias and the variance of e0' in the formula

$$\textbf{MSE} = \textbf{Bias}^2 + \textbf{Var(e0')}. \quad \textbf{(2)}$$

The variance of e0' is not easy to calculate. It cannot be calculated from the 20 m_x', as it can in Chiang's conventional method. Since the m_x' are generated by a model, we need to be able to represent the variation due to that model. We therefore use a simulation technique to estimate the variance of e0'. We produce a large number J of replicates such that, for each replicate j (j = 1..J), in each age stratum, the simulated number of deaths follows a binomial distribution (n, p), the two parameters of which are the population of the age stratum (n), assumed to be fixed, and the original value of m_x (p). For each replicate j, we compute a *simulated* life expectancy value (e(j)), and we also use the model to obtain an *estimated* value (e(j)').

The bias is given by the following formula:

$$\textbf{Bias = Mean (e(j)') - e0} \quad \textbf{(3)}$$

It can be written as the sum of two components: the simulation bias [*Mean (e(j)) - e0*] and the model bias [*Mean (e(j)') – Mean (e(j))*].

Similarly, the variance of e0' is now computed as the variance of the e(j)'.

The entire process consists of the combination of the two techniques: simulation, which produces the e(j), and modelling, which generates the e(j)'. The process is considered *valid* if, despite the bias, the MSE is less than the variance of e0 (which is computed by the standard method), i.e., if the following *condition* is true:

$$\textbf{MSE < Var(e0)} \quad \textbf{(*)}$$

3. Evaluation of the Process

3.1. Data used

To evaluate the process, we need to be able to apply it to different units of analysis that are *similar* in every respect (i.e., equal in age structure and mortality rate for each age stratum) except size (N). The simplest way of doing so is to use the same starting CMA (using the Winnipeg example again) and divide the d_x and m_x of all age strata by the same factor f, which we will vary between 1 (which corresponds to the original CMA) and 50. This produces different units of analysis, each of which has the following characteristics:

- e0 has the value obtained for Winnipeg
- Var(e0) = the variance obtained for Winnipeg, multiplied by factor f

We choose J = 1,000.

3.2. Results

The test results are summarized in Table 2. For selected values of f (1, 2, 3, 5, 10, 20 and 50), the table shows the following values:

- the size N of the resulting unit of analysis (column 1);
- the values of e0 and its variance, computed by Chiang's method (columns 2 and 3);
- the statistics obtained after the simulation step but before the model is applied: the mean e(j) (column 4) and the simulation bias (column 5 = column 4 – column 2);
- the statistics obtained after the model is applied: the mean of the e(j)' (column 6), their variance (column 7), the simulation bias (column 8 = column 6 – column 2), the total bias (column 9 = column 5 + column 8), the square of the total bias (column 10 = the square of column 9), the MSE (column 11 = column 7 + column 10), and the outcome, which depends on whether condition (*) is satisfied or not (column 12).

When the population of the unit of analysis *decreases*, both the MSE and Var(e0) *increase* (see Figure 2). However, the MSE increases more slowly. As a result, there is a threshold value of N (between 200,000 and 300,000) at which the two curves cross. If N is below the threshold, then MSE < Var(e0), which means the process is valid. This result is consistent with what we generally find in small area estimation theory. In Table 2 we deduce that the threshold is slightly less than 222,505. We can also check the characteristics concerning e0 and Var(e0), and we see the results for Winnipeg's original size (f = 1). Lastly, for values of N that are below the threshold, condition (*) is satisfied.

For very small values of N, however, the model becomes ineffective; while condition (*) is met, the bias is too large to be ignored. The model is also of no value when the size is above the threshold; the unit of analysis is self-sufficient, which is consistent with small area estimation theory.

In Table 2, we see that the $Bias^2$ component is small relative to the MSE (it is less than 1% of the total value of the MSE). This indicates that it is appropriate to carry out simulations based on the binomial distribution.

4. Discussion

We have shown that small area estimation theory produces a better estimate of life expectancy than the standard method does. We have also shown that a model based solely on age is sufficient to satisfy condition (*) (below a certain level), at least for CAs *similar* to Winnipeg in 1996.

It is interesting to note that, even if the threshold is unknown, the model's variance (V) can still be calculated: it is the lesser of the MSE and Var(e0). However, there is no guarantee that the process will generate a CI that is less than half a year (the value which stimulated our argument in the introduction). For that to be the case, the following inequality must be satisfied: $\mathbf{1.96*\sqrt{V} \leq 0.25}$, or $\mathbf{V \leq 0.016}$.

The following summarizes the steps for computing the life expectancy for a given CA:

1. Calculate e0 and Var(e0) (by the standard method).
2. Follow the process (simulation and model) with a large value of J.
    a. For j = 1...J
        i. For each age stratum, generate a simulated value for the number of deaths d_x using a binomial distribution (n, p), the two parameters of which are the population of the age stratum (n) and the original value of m_x (p).
        ii. Recalculate the m_x of the 20 age strata.
        iii. Calculate e(j) (by the standard method) with the m_x, d_x and n_x.
        iv. For each group,
            1. Use the four-parameter logistic regression to predict the m_x (equation (1)).
            2. Using the parameter estimates and the values of avg_age and setting unit = 1, calculate the estimated mortality rates m_x' for the CA.
        v. Calculate the new numbers of deaths for the 20 groups: $d_x' = m_x'$ multiplied by n_x.
        vi. Calculate e(j)' (by the conventional method) with the m_x', d_x' and n_x.
    b. Calculate the bias (equation (3)).

c. Calculate **MSE = Bias^2 + Var(e(j)').**

3. Calculate the model's variance (V): V = min(MSE,Var(e0)). If V = MSE, the process is valid. In addition, if **V ≤ 0.016**, the CI will be no more than half a year.
4. Construct the confidence interval for life expectancy: **CI = e0 ± 1.96*√V**.

References

Chiang, C.L. *The Life Table and its applications*, Robert E. Krieger Publishing Company, 1984.

Rao, J. N. K. *Small Area Estimation*, Wiley Series in Methodology, 2003.

Wilkins R, Berthelot JM, Ng E. Trends in mortality by neighbourhood income in urban Canada from 1971 to 1996. Health Reports 2002; 13 (Supplement): 47-73. http://www.statcan.ca.

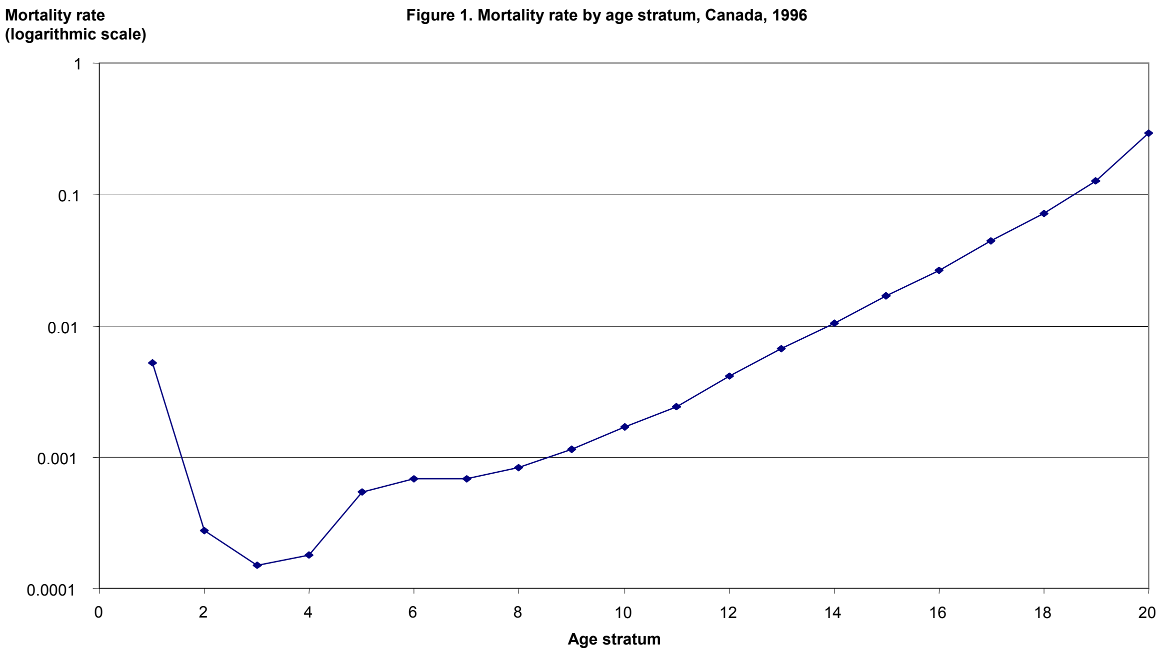

Figure 1: Mortality rates, Canada (1996)

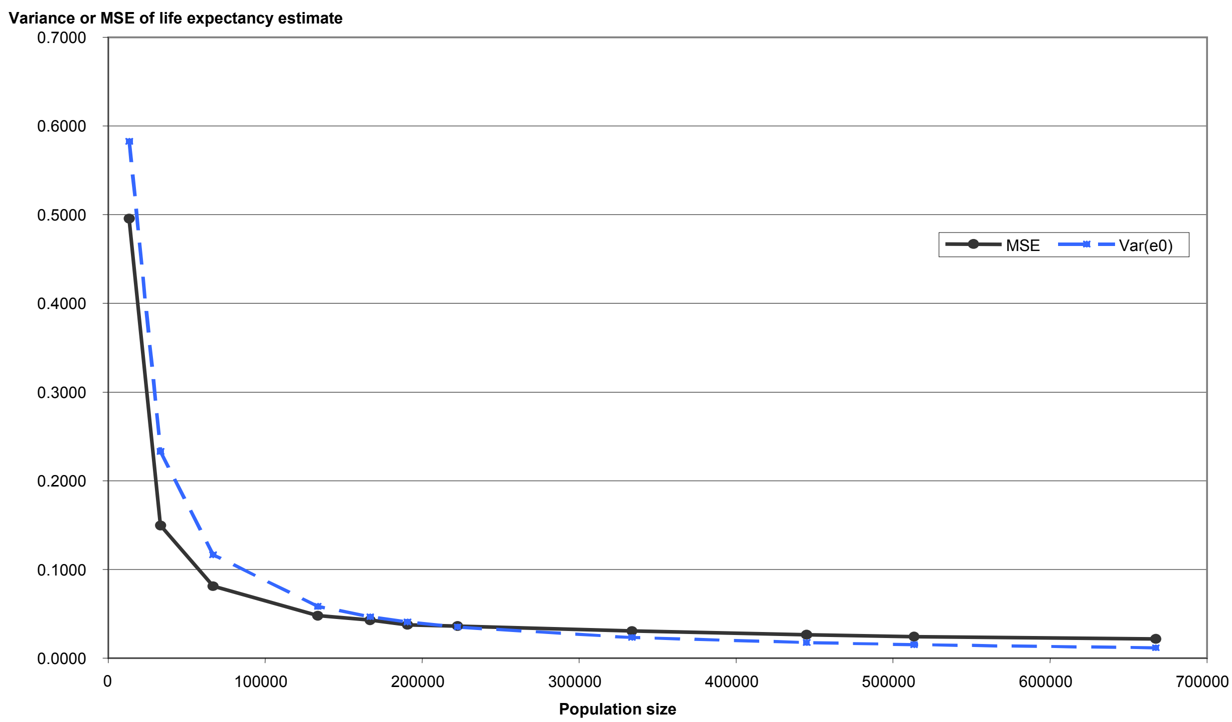

Figure 2: Variance and MSE of life expectancy at birth by population size (N)
(for units of analysis *similar* to Winnipeg)

Age stratum number	Continuous age	Group
1	Less than 1	1
2	≥1, <5	
3	≥5, <10	2
4	≥10, <15	
5	≥15, <20	
6	≥20, <25	3
7	≥25, <30	
8	≥30, <35	
9	≥35, <40	
10	≥40, <45	4
11	≥45, <50	
12	≥50, <55	
13	≥55, <60	
14	≥60, <65	
15	≥65, <70	
16	≥70, <75	
17	≥75, <80	5
18	≥80, <85	
19	≥85, <90	
20	≥90	

Table 1: Age partition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
f	Size (N)	Original values		With simulation, without model		With simulation, with model						
		e0	Var (e0)	Mean (e(j))	Simu-lation bias	Mean (e(j)')	Var (e(j)')	Model bias	Total bias	(Total bias)²	MSE	(*)[2] satis-fied?
1	667515	79.133	0.012	79.138	0.005	79.199	0.017	0.062	0.066	0.004	0.022	No
2	333758	79.133	0.023	79.140	0.007	79.184	0.028	0.043	0.051	0.003	0.030	No
3	222505	79.133	0.035	79.143	0.010	79.167	0.035	0.024	0.034	0.001	0.036	No
5	133503	79.133	0.058	79.148	0.015	79.148	0.048	-0.001	0.015	0.000	0.048	Yes
10	66752	79.133	0.117	79.154	0.021	79.133	0.081	-0.020	0.000	0.000	0.081	Yes
20	33376	79.133	0.233	79.171	0.038	79.121	0.149	-0.050	-0.012	0.000	0.150	Yes
50	13350	79.133	0.583	79.192	0.059	78.952	0.463	-0.240	-0.181	0.033	0.496	Yes

Table 2: Results of the process applied to seven units of analysis similar to Winnipeg

[2] (*): MSE < Var(e0)