SSC Annual Meeting, June 2009
Proceedings of the Survey Methods Section

ESTIMATION OF RELATIVE RISKS FOR POPULATION PROJECTIONS USING MICROSIMULATIONS

Ivan A. Carrillo G. and Milorad Kovacevic[1]

ABSTRACT

In microsimulations for population projections one is often interested in estimating an event probability. A request for microsimulation flexibility leads to a two-step method in which estimated base-risks are adjusted by factors based on relative risks. The concepts of odds ratios and relative risks are contrasted under these settings. To calculate the relative risks, we want to set the whole population (i.e. population average) as the reference category. Besides the obvious risk ratio with the denominator being an average risk for the population, we examine some alternative ways of duplicating the population for estimation of relative risks.

KEYWORDS: Duplication, Microsimulations, Reference category, Relative risks, Two-step method.

RÉSUMÉ

En microsimulation, dans le cas de projections de la population, il est souvent intéressant d'estimer la probabilité d'un événement. Un besoin de flexibilité en microsimulation conduit à une méthode à deux étapes où les risques de base estimés sont ajustés selon des facteurs s'inspirant des risques relatifs. Les concepts de rapports de cotes et de risques relatifs sont mis en parallèle dans ce contexte. Pour calculer les risques relatifs, nous voulons que la catégorie de référence soit la population entière (c'est-à-dire la population moyenne). Outre le rapport évident de risques, le dénominateur comportant un risque moyen pour la population, nous examinons d'autres façons de reproduire la population pour estimer les risques relatifs.

MOTS CLÉS: Catégorie de référence; duplication, méthode à deux étapes; microsimulations, risques relatifs

1. INTRODUCTION

1.1 Microsimulations and DemoSim

Over the last two decades, Statistics Canada has had a leading role in developing microsimulation models (MSM) in Canada. Models such as the Social Policy Simulation Database/Model (SPSD/M), Population Health Model (POHEM), the Lifecycle model (Lifepaths), and the demographic projections model (DemoSim) have been developed as analytical tools for studying and assessing impacts policy changes, as well as how changes in demographic and behavioral patterns might affect individual outcomes. A specialized software package, Modgen, has also been developed by Statistics Canada with the objective to allow an easy implementation and maintenance of microsimulation models.

The DemoSim model for demographic projections is being specifically designed to project changes in population composition taking into account differentials in fertility, mortality, internal and international migrations, marital status and intergenerational transfers. The traditional cohort component or the multistate projection models can become unmanageable when a great number of characteristics need to be projected simultaneously. Instead, an MSM was created and used for the first time to produce population projections of visible minority groups from 2001 to 2017 (Statistics Canada, No 91-541-XIE).

DemoSim is "a dynamic, continuous time, longitudinal, event based, open, spatial, stochastic (MonteCarlo) microsimulation projection model" (Bélanger, et al., 2008). It was conceived, conceptualized, and designed by the

[1] Statistics Canada, 15 Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A OT6, Email addresses: ivan.carrillogarcia@statcan.gc.ca and Milorad.Kovacevic@statcan.gc.ca

Demographic Analysis Section of the Demography Division in collaboration with the Socio Economic Analysis and Modeling Division. The contributions of statistical methodologists are also important as they relate to: the development of the base population used in DemoSim; the identification and validation of considered models; efficiency, consistency, and unbiasedness of estimation procedures, etc.

Sometimes the microsimulation model is identified with a computer program that mimics the demographic processes on individual members of a population - people, households, or businesses, for example. For each observation in a simulation database which can be a large-scale survey, the computer program simulates the outcomes of interest, such as fertility, mortality or mobility, by applying actual or hypothetical rules to the database. The rules are often expressed by sets of rates and probabilities. An objective of the microsimulation is the generation of micro pseudo-data by a model. The generation itself is stochastic, so the generated pseudo-data change from run to run due to a dependence on random number generators.

We end up this introduction with a brief description of the Microsimulation process. Then, in section 2, we present the foundation for the two-step approach that we use throughout the paper and that serves as motivation for the technique presented here. In section 3 we give the reasons for our use of relative risks, as opposed to the odds ratios. There we also establish the main techniques of the paper; those of how to estimate the denominator of the relative risk. After that, we show an illustration of the two main methods in section 4. We conclude the paper with recommendations and possible future research.

1.2 Microsimulation Process

For a Microsimulation projection we take the individuals in a base dataset and simulate their individual life trajectories up to a specified time ahead. In DemoSim the particular dataset used as base population is the 2006 long-form Census. For different individuals on the dataset, different events can occur at different time points; figure 1 shows an example of how a subject could evolve during part of the projection period.

Figure 1 - Example of an Individual's Trajectory for a Simulation Period

To be able to advance individuals through time, we need to determine (probabilistically) whether an event occurs, when it will occur, and which event will occur. For a given individual, probabilities or risks of each potential event (e.g., attainment of a higher level of education, changing religion, having a birthday, migrating, emigrating, dying, etc.) are estimated based on the initial characteristics (sex, age, marital status, etc.). We can thereby take into account the differential behavior of each individual based on his/her characteristics; for example, the risk of giving birth to a child will vary depending on whether the woman is a recent immigrant, belongs to a visible minority group or some other group, and so on.

Using the Poissonian framework (Thompson, 1999, page 79), the probabilities or risks of each event are then converted into waiting times, to provide the time indexation of events. In this way we can determine which event will occur first and compute the amount of time that will elapse before it occurs, thus moving each individual forward through time. Whenever an event takes place, all characteristics are re-evaluated and probabilities and risks are recomputed based on these changed characteristics. The process continues for that individual until he/she dies, emigrates or reaches the projection horizon. Births and immigrants are also added during the simulation and go through the same process.

Different scenarios can be used with microsimulation, where each differs by the set of assumptions made on fertility rates, life expectancy, immigration levels and distribution, as well as internal migration patterns. For more details of the Microsimulation process in general, and on DemoSim in particular, see Kovacevic et. al. (2009).

2. TWO-STEP METHOD FOR ESTIMATING RISKS

In this paper we concentrate on the model used for fertility in DemoSim. We model the probability that an event b happens to a person with characteristics x with a generalized linear model. In our particular settings, b represents the event "giving birth to a child" and we use the log-log link. Hence we can write the probability that a particular woman with characteristics x gives birth to a child as $P(b \mid x) = \exp[-\exp(\beta_0 + x'\beta)]$; which comes from the log-log model $\log\{-\log[P(b \mid x)]\} = \beta_0 + x'\beta$, and the β's are the parameters of interest. Notice that when the values of x equal the reference categories, $P(b \mid x = \text{ref}) = \exp[-\exp(\beta_0)]$.

In the scope of DemoSim, the estimation of a probability (risk) of an event is of the primary interest. An important request that DemoSim can run under different scenarios that include different assumptions about the base risks, prompted the use of a two-step method for estimating the probabilities. That is, first a base-risk for a larger population group is estimated, and then it is adjusted by a factor based on a relative risk. In mathematical terms we can write this as follows:

$$\begin{aligned} P(b \mid x) &= P_1(b) \cdot \frac{P_2(b \mid x)}{P_2(b)} \\ &= P_1(b) \cdot RR(b \mid x), \end{aligned} \tag{1}$$

where b represents the event of interest (giving birth), x are the values of the explanatory variables, and RR stands for relative risk of experiencing the event for an individual with characteristics x, compared to a "reference" individual (i.e. we want to set the whole population, or population average, as the reference category). For example, the probability of giving birth for a woman of a certain age and of certain characteristics can be expressed as a product of an age-specific probability and the relative risk of giving birth for a woman of such characteristics compared to an average woman of the same or similar age.

The appealing feature of expressing $P(b \mid x)$ as in (1) is that we can keep the relative risks $RR(b \mid x)$ constant throughout the simulation period and incorporate different assumptions about how the event will evolve only trough the manipulation of the "base risk" $P_1(b)$. Also, we can use more stable estimates of the base risks, $P_1(b)$, from external sources (like vital statistics).

Note that, in equation (1), $P_1(b)$ is the base risk for any woman in the risk population, and $P_2(b \mid x)$ and $P_2(b)$ come from the log-log model. $P_2(b \mid x)$ is the fitted value predicted by the log-log model for a woman with characteristics (or covariates) given by x; whereas $P_2(b)$ is the fitted value predicted by the log-log model for an *average* woman. Here "average" indicates that $P_2(b)$ should be calculated for the same population for which the base risk, $P_1(b)$, is calculated. This is the focus of the next section.

3. RELATIVE RISK

3.1 Relative Risks

The relative risk (RR) is a ratio of probabilities. For example, the RR of bearing a child for an aboriginal woman versus a non-aboriginal woman is 1.4. This means that aboriginal women have 1.4 times the relative frequency of bearing a child as non-aboriginal women.

According to Simon (2008), the RR "measures events in a way that is interpretable and consistent with the way people really think," as opposed to the odds ratio (OR). In our example we have that the odds of bearing a child for aboriginal women is 0.058/(1-0.058)=0.06; and for non-aboriginal women it is 0.04/(1-0.04)=0.04. Therefore the odds ratio is 1.5; the odds of bearing a child for aboriginal women is 1.5 times the odds for non-aboriginal.

The RR measure is not necessarily valid for retrospective designs, whereas "the odds ratio is equally valid for prospective, retrospective, or cross-sectional sampling designs" (Agresti, 2002). Moreover, the RR is not always computable (like in case-control designs) and can lead to ambiguous comparisons because "for every problem, there are two possible ways to compute relative risk" (Simon, 2008).

Nonetheless, for small event probabilities we have $RR \approx OR$; besides, in our case there is no ambiguity, we are dealing with a cross-section design, the probability of bearing a child is very small, and the RR is easier to interpret. For these reasons we concentrate on the RR .

3.2 Estimation of the Denominator of the Relative Risk

There are several different ways to calculate an estimate of $P_2(b)$. The obvious one is a weighted average of the probabilities for different population subgroups, where the weights are the relative frequencies of the subgroups:

$$\hat{P}_2(b) = \sum_x \frac{\hat{N}_x}{\hat{N}} \hat{P}_2(b \mid x), \tag{2}$$

where $\hat{N}_x$ is the estimated size of the population at risk in subgroup x, and $\hat{N}$ is the estimated size of the total population at risk. This clearly applies when the characteristics x are categorical; when they are continuous the summation in (2) is a summation over individuals, $\hat{N}_x$ becomes the individual's survey weight, and $\hat{N}$ is the sum of weights.

This method suffers of at least two drawbacks. For one thing, this approach is sensitive to 'empty' sample from some population subgroups. For another, this computation is relatively straightforward for calculating *marginal* relative risks; for example calculating the relative risk of giving birth by visible minority status. Yet this computation can be rather cumbersome for finer groupings; which is the case in the microsimulation setting. To compute the denominator $\hat{P}_2(b)$, one needs to compute $\hat{P}_2(b \mid x)$ for many different patterns; this can be a substantial task, particularly with a significant number of covariates and categories.

Another way to calculate the denominator of the relative risk is by means of an ad-hoc method, here referred to as "duplication." After a model for $P_2(b \mid x)$ in the numerator of $RR(b \mid x)$ has been chosen, we add into the model a (dummy) variable representing the reference category for each categorical covariate, and artificially create some records which represent the "reference" or "baseline" individual.

We examined two different ways of creating the artificial records. The first one, which was being used in an earlier release of DemoSim, replicates the whole dataset as many times as (categorical) covariates there are; one replication corresponding to each covariate. In each replicate we set all the values of all the categories of the corresponding covariate to zero and leave the values of the other covariates, response variable, and survey weight unchanged.

In the other method we duplicate the whole dataset only once. For the duplicated records, we set all the values of all the categories (of all the categorical covariates) to zero, but leave the response variable, continuous variables (if any), and the survey weight, unchanged. Table 1 shows an example of how we carry out this second duplication method.

Table 1 - Example of Second Duplication Method

X1			X2				wi	b
c1	c2	REF	c1	c2	c3	REF		
1	0	0	1	0	0	0	15	1
1	0	0	0	1	0	0	7.4	1
1	0	0	0	0	1	0	5	0
0	1	0	1	0	0	0	2	0
0	1	0	0	1	0	0	10	1
0	1	0	0	0	1	0	1.3	0
0	0	1	0	0	0	1	15	1
0	0	1	0	0	0	1	7.4	1
0	0	1	0	0	0	1	5	0
0	0	1	0	0	0	1	2	0
0	0	1	0	0	0	1	10	1
0	0	1	0	0	0	1	1.3	0

After the dataset is duplicated (by either method), we refit the model and then we are able to calculate $\hat{R}R(b \mid x)$ for each particular subject. The numerator $\hat{P}_2(b \mid x) = \exp[-\exp(\hat{\beta}_0 + x'\hat{\beta})]$ is the individual's fitted value and the denominator, $\hat{P}_2(b) = \exp[-\exp(\hat{\beta}_0)]$, is the fitted value for an artificially created reference individual (i.e. one who has all the covariates equal to zero). Some simple numerical examples we conducted illustrated that the second method of duplication (duplicating only once) produced, for marginal characteristics, the exact same results as those obtained using equation (2); whereas the first way of duplicating did not. Therefore we recommend using the second type of duplication. Nonetheless, we are not sure whether this is the case for finer groupings.

4. ILLUSTRATION

We now present an example of the weighted average and second duplication methods. Here we restrict ourselves to 25 year old non-aboriginal women. There are 35,820 of such women in the sample, 2,542 of which gave birth in the year preceding the census. We consider as explanatory variables "Education" and "Marital Status." There are 11,485 women with high school degree or less (<=HS), and 24,335 with at least some post-secondary education (PS, Univ). 7,381 women are married, 8,467 live in a common law union (CLU), and 19,772 are not in a union (single, divorced, widowed).

We obtain the following estimated log-log regression for the probability of giving birth:

$$\log\{-\log[P(b|x)]\} = 1.5589 - 0.2613 \times I(<= \text{HS}) - 0.9417 \times I(\text{Married}) - 0.6391 \times I(\text{CLU}). \tag{3}$$

Based on model (3) we can estimate the relative risk for each of the six patterns of covariates, making use of the $\hat{P}_2(b)$ from expression (2); this is done in the middle part (and last column) of table 2. The estimated denominator of the relative risk, using this method, is $\hat{P}_2(b) = 0.06902$.

Table 2 - Illustration of Calculation of the Relative Risks by Two Methods

x		$\hat{P}_1(b)$			$\hat{R}R(b \mid x)$	
Educ	Mar. Sta.	(Base Risk)	$\hat{N}_x$	$\hat{P}_2(b \mid x)$	Dupl.	W. Ave. Meth.
<=HS	Married	0.06895	13917	0.24	3.478	3.476
	CLU	0.06895	14593	0.14	2.101	2.099
	NotUnion	0.06895	31494	0.03	0.373	0.373
PS, Univ	Married	0.06895	22663	0.16	2.272	2.270
	CLU	0.06895	29999	0.08	1.180	1.179
	NotUnion	0.06895	74137	0.01	0.125	0.125
		12881	186803		**0.06895**	**0.06902**

For the duplication method we refit model (3) after having duplicated the dataset as in table 1. In this case we obtain a regression coefficient for each of the categories of interest, “<=HS,” “PS, Univ,” “Married,” “CLU,” and “Not in Union.” The model obtained for the duplication method is:

$$\begin{aligned}\log\{-\log[P(b|x)]\} = 0.9837 + 0.3139\times I(<= \text{HS}) + 0.5751\times I(\text{PS, Univ})\\ - 0.9415\times I(\text{Married}) - 0.6391\times I(\text{CLU}) + 0\times I(\text{Not in Union}).\end{aligned} \tag{4}$$

Model (4) allows us to compute the relative risks directly: the numerator is calculated from equation (4) and the denominator is calculated by $\hat{P}_2(b) = \exp[-\exp(0.9837)] = 0.06895$. This is shown in the next to last column of table 2.

We can notice that all the relative risks for all the patterns of the covariates are almost identical for the two methods. The fact that the two do not provide exactly the same relative risks for all cases is because we do not include the interaction between Education and Marital Status in our log-log model.

5. CONCLUSIONS

We have proposed a duplication method that in simple numerical examples produces the same relative risks (marginally) as the ones obtained through the, more theoretically sound, weighted average method. For example, the estimated risk of giving birth for an average married woman is the same using either method. A future development from this paper could be to try to show theoretically that this duplication method is equivalent to the weighted average method.

However, we are not sure whether the same holds for finer groupings, like the setting at hand in microsimulations. For instance, we need to calculate the risk for a Christian woman, born in the USA, who lives in Alberta, who has a Bachelor’s degree, and who is married. In DemoSim, there are 586,560 cells (i.e. different covariate patterns) in a single log-log model for fertility. In situations like this, the two methods could potentially deviate significantly; in particular given that many (or most) of the interactions are left out of the models. We would like to expand this paper to show whether or not the duplication method is appropriate for finer groupings.

The first method of duplication should be completely avoided since it fails to reproduce the results from the weighted average method even for simple small examples.

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley.

Bélanger, A., Caron Malenfant, É., Martel, L., and Gélinas, R. (2008). *Projecting ethno-cultural diversity of the Canadian population using a microsimulation approach*. Proceedings of the Joint UN-ECE/Eurostat Work Session on Demographic Projections, Bucharest.

Kovacevic, M., Grondin, C., Titus, R.-C., & Carrillo-Garcia, I. (2009). *Challenges in Methodology of Microsimulation Models: Examples from Projections of the Diversity of the Canadian Population, 2006-2031*. Technical report, Statistics Canada's Advisory Committee on Statistical Methods, April 27-28.

Simon, S. (2008). *Odds ratio versus relative risk*. http://www.childrens-mercy.org/stats/journal/oddsratio.asp.

Statistics Canada (2005). *Population projections of visible minority groups, Canada, provinces and regions, 2001-2017*. Catalogue no. 91-541-XIE. Ottawa.

Thompson, J.R. (1999). *Simulation: A modeler’s Approach*. New York: Wiley.