Improving Population Estimates through Raking by Joint Distribution of Multiple Grouping Variables

Yilan Huang[1,2], Honghu Liu[1,2,3]

[1]Department of Biostatistics, Fielding School of Public Health, UCLA, CA 90095
[2]Section of Public and Population Health, School of Dentistry, UCLA, CA 90095
[3]Department of Medicine, David Geffen School of Medicine, UCLA, CA 90095

Abstract

Raking is widely used to reduce biases in population estimates in surveys because it is relatively simple to implement and requires only marginal control totals of each auxiliary variable. However, when the raking margins are highly correlated, the raking process may take numerous iterations to converge, and the weights may have higher variability than expected. In such cases, combining the two or more variables to form a single margin is one possible solution. This study compares the performance of the raking algorithm when using only a single variable for each margin versus using the joint distribution of multiple grouping variables. The comparison uses simulated oral health data and assesses three weighting schemes—design weights, raking sequentially with each variable, and raking with the joint distribution—in terms of weights and point estimates. The raking algorithm reduces the average bias and average standard error in estimates compared with design weights. Raking with cross-classification, especially when variables are highly correlated, can improve sampling weights based on joint distributions available at the population level, which can provide an effective way to estimate population parameters by sample statistics.

Key Words: Survey design, sampling weight, raking, post-stratification, joint distribution

1. Background

Missing data can occur in survey research when a sampled element does not participate in the survey (nonresponse), or when an element in the target population is not included on the survey's sampling frame (noncoverage).[1] Weighting adjustments are commonly applied in surveys to compensate for nonresponse and noncoverage, in order to make the weighted sample represent the population of inference as closely as possible.[2] Raking, also known as raking ratio estimation, or sample-balancing, is a post-stratification procedure to match marginal distributions of a survey sample to known population margins on a specified set of variables. It is an iterative process that usually proceeds one variable at a time, applying a proportional adjustment to the weights of the cases that belong to the same category of the control variable, and terminates when the convergence criterion is achieved. The initial design weights in the raking process are often equal to the inverse of the selection probabilities.[3] Raking is widely used to reduce nonresponse and noncoverage biases, as well as sampling variability.[4]

Raking offers a distinct advantage in that only the marginal control totals of each auxiliary variable are needed, rather than counts for all the cells in the cross-classification such as would be required with post-stratification.[5] In many applications of raking, it is assumed that two or more sets of marginal populations are known, while the joint distribution with respect to all the raking variables is known only from a sample.[6] One situation that could be problematic is when dependent margins are used in raking. With highly correlated margins, the raking process may take numerous iterations to converge, and the weights may have higher variability than expected.[5] In such cases, combining

the two or more variables to form a single margin is one possible solution, if feasible. When control totals are available, the raking margin can involve multiple control variables (e.g., age categories x gender x race), and adjustment to control totals is then achieved by creating a cross-classification of the categorical control variables and matching the total of the weights in each cell to the corresponding control total.[4]

The motivation of the study is to compare the performance of raking method when using single variable for each margin versus using cross-classifications, in order to identify a more efficient and effective way to estimate the population parameter by sample statistics. Tooth decay, also known as caries, is the most common chronic disease in children and the most frequent health problem in the United States. Poor oral health in children can lead to attention problems, nutrition issues, missed school days, and increased dental care costs. In this paper, we conduct a simulation study to compare empirical biases, standard errors, and convergence time across different outcome variable models with oral health survey data on children.

2. The Basic Raking Algorithm

In a cross-classification that has J rows and K columns, we denote the sum of individual weights w_i $(i = 1, \dots, n_{jk})$ in cell (j, k) by w_{jk}. To indicate further summation, we replace a subscript by a $+$ sign. Thus, the initial row totals and column totals of the sample weights are w_{j+} and w_{+k} respectively. Similarly, we denote the corresponding population control totals by T_{j+} and T_{+k}.

K columns

J rows		Female	Male	
	African American			
	Asian	(j,k) cell		w_{j+}
	Hispanic			
	White			
	Other			
		w_{+k}		

Figure 1. Example of the basic raking algorithm using race/ethnicity and gender.

The iterative raking algorithm produces modified weights, whose sums we denote by a suitably subscripted *m* with a parenthesized superscript for the number of the step. Thus, in the two-variable cross-classification we use $m_{jk}^{(1)}$ for the sum of the modified weights in cell (j, k) at the end of step 1. If we begin by matching the control totals for the rows, T_{j+}, the initial steps of the algorithm are

$$
\begin{aligned}
m_{jk}^{(0)} &= w_{jk} && (j = 1, \dots, J; k = 1, \dots, K) \\
m_{jk}^{(1)} &= m_{jk}^{(0)}(T_{j+}/m_{j+}^{(0)}) && \text{(for each } k \text{ within each } j) \\
m_{jk}^{(2)} &= m_{jk}^{(1)}(T_{+k}/m_{+k}^{(1)}) && \text{(for each } j \text{ within each } k)
\end{aligned}
$$

The adjustment factors, $T_{j+}/m_{j+}^{(0)}$ and $T_{+k}/m_{+k}^{(1)}$, are applied to the individual weights, which we could denote by $m_i^{(2)}$, for example. In the iterative process an iteration rakes both rows and columns. Thus, for iteration s $(s = 0, 1, \dots)$ we may write

$$
\begin{aligned}
m_{jk}^{(2s+1)} &= m_{jk}^{(2s)}(T_{j+}/m_{j+}^{(2s)}) \\
m_{jk}^{(2s+2)} &= m_{jk}^{(2s+1)}(T_{+k}/m_{+k}^{(2s+1)})
\end{aligned}
$$

The raking algorithm proceeds by proportionately scaling the m_{jk} such that the relations

$$\sum_k^K m_{jk} = m_{j+} = T_{j+}$$

and

$$\sum_j^J m_{jk} = m_{+k} = T_{+k}$$

are satisfied in turn. The process terminates either after a fixed number of iterations or when each marginal total of the raked weights is within a specified tolerance of the corresponding population control total. In general, raking through a large number of variables slows the convergence process, though other factors also play a role.[4] These factors include the number of categories of raking variables, the sample size in each category of the raking variables, and the size of the difference between each control total, and the weighted sample total prior to raking.

3. Simulation Study

In case study analyses using real-world survey data, bias cannot be directly evaluated due to the lack of a gold standard or known truth about the outcome measures of interest in the population. Therefore, we conduct a simulation study with oral health survey data to assess the performance of various weighting schemes, focusing on the magnitude of differences in empirical bias, standard error, and convergence time.

3.1 Simulation framework

In the simulation study, we evaluate the estimates for population prevalences and means for oral health outcome variables. We assume that the outcome variable model contains three main effect covariates: school district, socioeconomic status (SES; yes and no) and ethnicity (Hispanic and non-Hispanic). Additionally, all statistically significant two-way interaction terms between the three main effect variables are included in the outcome model. We assume a stratified sampling design, since practical surveys almost always involve complex sample designs.

The model for the outcome variable Y is specified as follows. For a binary response, we have

$$P(Y_{ijsk} = 1) = \text{logit}^{-1}(\mu + \alpha_i + \beta_j + \gamma_s + \alpha_i\beta_j + \alpha_i\gamma_s + \beta_j\gamma_s + \epsilon_{ijk})$$
$$i = 1,2;\ j = 1,2;\ s = 1,2;\ k = 1,\ldots,N_{ij}$$

where N_{ij} is the population size in cell (i,j) for the survey, and $\epsilon_{ijk} \sim N(0,\sigma^2)$. In the model, α measures the main effect of SES, β measures the main effect of ethnicity, and γ measures the main effect of school district. For a count response, we employ a Poisson regression model with a similar covariate structure.

The implementation of the simulation is based on the framework and involves the following steps:

1. Generate an artificial finite population of size N that contains 8 subpopulations defined by the categories of the three auxiliary variables (2 x 2 x 2). The subpopulation size N_{ij} is determined based on the joint distribution of Los Angeles County third-grade students from population enrollment data.
2. Generate the value for the outcome variable Y. The outcome model parameters are determined based on the 2020 LA County Smile Survey third-grade data.[7]
3. Select a stratified random sample of size n from the population, with school districts as the strata.
4. Conduct survey weighting using sampling weights, raking with single-variable margins, and raking with joint distribution (school district x SES x ethnicity), respectively. Obtain the estimates for the outcome variables and then compare the empirical results.

We examine the empirical properties using a repeated sampling approach (i.e., averaging results across the 10,000 simulation samples). The simulation is conducted in R. Raking was conducted

based on the marginal totals of each weighting variable and their joint distribution, using the R package "survey".[8]

3.2 Results

Table 1 shows that raking helps reduce bias significantly for both the binary outcome and count response. Raking with joint distribution performs better both in terms of bias and standard error. Moreover, the empirical results demonstrate improved efficiency in the average convergence time of raking. For example, for the binary outcome, the average convergence time of each raking process for raking with joint distribution (*Raking 2*) is 0.027 seconds less than that for raking with single-variable margins (*Raking 1*).

Table 1. Empirical results for the three weighting schemes over repeated sampling.

Oral health related outcomes	Population or weighting method	Mean or %	Standard error	Time (secs)	Relative bias ($x10^{-2}$)	Relative SE ($x10^{-3}$)
Caries Experience (%)	Population	65.2790				
	Sampling weights	66.0003	0.333		1.1050	5.094
	Raking 1	65.2794	0.160	0.042	0.0006	2.456
	Raking 2	65.2793	0.159	0.015	0.0004	2.436
Caries Experience (number)	Population	2.5202				
	Sampling weights	2.5391	0.014		0.7512	5.388
	Raking 1	2.5203	0.012	0.042	0.0028	4.513
	Raking 2	2.5202	0.011	0.015	0.0018	4.467

4. Discussion

Raking is widely used to reduce biases in estimates in sample surveys. The raking algorithm usually proceeds one variable at a time, applying a proportional adjustment to the weights of the cases that belong to the same category of the control variable. In this study, we proposed the raking method to combine the correlated variables to form a single margin and rake with the joint distribution. We showed, both theoretically and empirically, raking through cross-classification, especially when variables are correlated, can reduce biases and standard errors, and thus improve sampling weights based on joint distributions available at the population level. This modified raking method certainly seems to provide an effective way to estimate the population parameters from the sample data.

However, more work on the proposed raking method is needed. For instance, nonresponse issue is typically encountered in real-world survey practice, and raking is most often used to reduce such potential bias. For our further work, we plan to incorporate a response model that estimates response propensity based on available demographic variables, to evaluate the performance of the modified raking method in the presence of nonresponse. Other factors that may affect the differences between the weighting estimates in the simulation study, such as the number of simulation samples, the impact of the interaction effects, and the sample size for stratified random sampling, also deserve to be examined for a sensitivity analysis.

References

1. Brick, J.M. and Kalton, G., 1996. "Handling missing data in survey research." *Statistical methods in medical research*, 5(3), 215-238.
2. Kalton, G., and Flores-Cervantes, I. 2003. "Weighting methods." *Journal of official statistics*, 19 (2), 81.
3. Battaglia, M. P., Hoaglin, D. C., and Frankel, M. R. 2009. "Practical considerations in raking survey data." *Survey practice*, 2 (5). DOI: https://doi.org/10.29115/SP-2009-0019
4. Battaglia, M. P., Izrael, D., Hoaglin, D. C., and Frankel, M. R. 2004. "Tips and tricks for raking survey data (aka sample balancing)." In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 4740–4745. Available at: http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000074.pdf (accessed Sep 2024).
5. Brick, J. M., Montaquila, J., and Roth, S. 2003. "Identifying problems with raking estimators." In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 710–717. Available at: http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000472.pdf (accessed Sep 2024).
6. Oh, H. L., and Scheuren, F. 1987. "Modified raking ratio estimation." *Survey Methodology*, 13 (2), 209-219.
7. Los Angeles County Department of Public Health (LACDPH). 2020. *Smile Survey 2020: The Oral Health of Los Angeles County's Children*. Available at: http://publichealth.lacounty.gov/ohp/docs/SmileSurvey2020_Final_info.pdf (accessed Sep 2024).
8. Lumley, T. 2020. "Package 'survey'." Available at: *https://cran.r-project org*. (accessed Sep 2024).