


Discussion

Calcutta Statistical Association Bulletin
75(1) 33–35, 2023
© 2023 Calcutta Statistical Association, Kolkata
Article reuse guidelines:
in.sagepub.com/journals-permissions-india
DOI: 10.1177/00080683231179822
journals.sagepub.com/home/csa


Marissa C. Ashner¹  and F. Jay Breidt²

We appreciate the opportunity to comment on this excellent article of Kim and Morikawa (henceforth KM). The importance of the class of problems described in this article is growing and the extensions outlined here, especially the combination of voluntary samples with probability samples, are essential in modern survey practice. In our experience, obtaining high-quality and timely probability samples of sufficient size for reasonable inferences can be extremely difficult, requiring tremendous resources to screen for rare populations, increase response rates, or deal with complex measurement issues. Survey analysts are thus increasingly faced with the need to use information from non-probability ‘samples’, which arise in various uncontrolled ways. They are often obtained from some voluntary self-selection mechanism like those emphasized here, but the framework of this article also covers various scenarios such as capture in administrative transactions or sensor records that were not designed to study the phenomenon of interest.

We found it instructive to work through the details of KM using a simple, discrete version of their framework. Consider a finite population consisting of N elements, denoted $U = \{1, 2, \dots, i, \dots, N\}$. For $i \in U$, let x_i and y_i denote binary random variables, each taking values in $\{0, 1\}$. As a concrete example, x_i is voter i 's choice of candidate in a previous two-party election (from party 0 or party 1), with known election outcome $\bar{X}_N = N^{-1} \sum_{i \in U} x_i$ for candidate 1, and y_i is voter i 's choice in an upcoming election. The votes are highly correlated across elections, but some voters switch parties. The parameter of interest is $\theta = N^{-1} \sum_{i \in U} y_i$, the unknown outcome for the upcoming election. We require x_i known for $i \in S$ and \bar{X}_N known, but do not require x_i known for all $i \in U$.

Let $\delta_i = 1$ if voter i responds and 0 otherwise. The correctly specified propensity to respond is

$$\pi(y; \phi) = P(\delta = 1 | x, y; \phi) = \{1 + \exp(\phi_0 + \phi_1 y)\}^{-1}.$$

Selection is thus non-ignorable: previous voting behaviour is predictive of response (because x is predictive of y), but current voting intentions are more important. While the specification is parametric, it is just a reparameterization of the discrete probability distributions identifiable from the available data.

¹University of North Carolina–Chapel Hill, North Carolina, USA

²NORC at the University of Chicago, USA

Corresponding author:

F. Jay Breidt, NORC at the University of Chicago, Illinois 60637-2745, USA.

E-mail: breidt-jay@norc.org

From equation (6) of KM, the observed likelihood of ϕ is a function of $\pi(y_i; \phi)$ for $\delta_i = 1$ and of $\tilde{\pi}(x_i; \phi) = P(\delta_i = 1 | x_i; \phi)$ for $\delta_i = 0$, where (using the Pfeffermann–Sverchkov trick noted in (7) of KM)

$$\frac{1}{\tilde{\pi}(x; \phi)} = E \left\{ \frac{1}{\pi(y; \phi)} \mid x, \delta = 1 \right\} \quad (1)$$

$$= 1 + e^{\phi_0} P(y = 0 \mid x, \delta = 1) + e^{\phi_0 + \phi_1} P(y = 1 \mid x, \delta = 1).$$

In (1), the conditional probability distribution for y at each level of x corresponds to the model $f(y | x, \delta = 1)$ in KM. To obtain $\hat{\pi}(x; \phi)$ in (9) of KM, we estimate these conditional probabilities using the respondent data:

$$\hat{P}(y = 1 \mid x = j, \delta = 1) = \frac{\sum_{i \in U} \delta_i 1(x_i = j) y_i}{\sum_{i \in U} \delta_i 1(x_i = j)}$$

for $j = 0, 1$, where $1(x_i = j) = 1$ if $x_i = j$ and 0 otherwise. Following equation (8) of KM, we then maximize

$$\ell_{obs}(\phi) = \sum_{i \in U} \delta_i \log \frac{\pi(y_i; \phi)}{1 - \hat{\pi}(x_i; \phi)} + \sum_{i \in U} \log \left\{ 1 - \hat{\pi}(x_i; \phi) \right\},$$

noting that the second term only requires the population counts of $x_i = 0$ and $x_i = 1$, not the individual x_i 's. In this setting, we might also consider the instrumental calibration estimator (see Lesage et al.^[1] and the references therein) with linear weighting, which in this case yields weights calibrated to N and $\sum_{i \in U} x_i$,

$$w_i^{IC} = 1 + \hat{\lambda}_1 + \hat{\lambda}_2 y_i, \quad i \in S,$$

where

$$\begin{bmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \end{bmatrix} = \left\{ \begin{bmatrix} \sum_{i \in U} \delta_i & \sum_{i \in U} \delta_i x_i \\ \sum_{i \in U} \delta_i y_i & \sum_{i \in U} \delta_i x_i y_i \end{bmatrix}^{-1} \right\}^\top \begin{bmatrix} \sum_{i \in U} (1 - \delta_i) \\ \sum_{i \in U} (1 - \delta_i) x_i \end{bmatrix}.$$

We conducted a small simulation experiment (with 500 replicates) in the setting described above, with $P(x = 1) = 0.47$, $P(y = 1 \mid x = 0) = 0.05$, $P(y = 1 \mid x = 1) = 0.97$, $\pi(0; \phi) = P(\delta = 1 \mid y = 0) = 0.3$, and $\pi(1; \phi) = P(\delta = 1 \mid y = 1) = 0.6$, with $N = 20,000$ and expected voluntary sample size of 8,894.4. We compared the naive estimator of the y -mean, the post-stratified estimator that controls on the count for the categories of x , the empirical likelihood estimator of KM, and the instrumental calibration estimator. As expected, the naive estimator ignores selection and is badly biased. The post-stratified estimator corrects some of the bias because x is predictive of y , but not all the bias because selection is non-ignorable. The empirical likelihood estimator corrects nearly all of the bias and the KM variance

estimator is nearly unbiased. The instrumental calibration estimator performs best in this limited simulation, with slightly larger variance but smaller bias than the empirical likelihood estimator.

Like our illustrative example, many practical survey problems are categorical and model specification for the propensity or the outcome amounts to conditional probabilities in low-dimensional tables. We are interested in the authors' thoughts on any special considerations for such categorical problems. Even in this simple example, the linearization variance estimation approach in KM section 4 is complex and prone to analytic or coding errors. Practitioners would greatly benefit if the authors provided code or suggested replication approaches. Another practical consideration is 'generic' weights that can be applied across all study variables, such as cases M1 and M2 in the authors' simulation experiment. How might a single, compromise set of weights be constructed when different variables are subject to different kinds of non-ignorable selection? Finally, we are interested in the authors' thoughts on the relative strengths and weaknesses of empirical calibration versus instrumental calibration in the context of this article.

ORCID iD

Marissa C. Ashner  <https://orcid.org/0000-0002-2936-4161>

Reference

1. Lesage E, Haziza D and D'Haultfoeuille X. A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys. *J Am Stat Assoc* 2019; 114: 906–915.