Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comments on “Statistical inference with non-probability survey samples”

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canada

by Zhonglei Wang and Jae Kwang Kim

Release date: December 15, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Comments on "Statistical inference with non-probability survey samples"

Zhonglei Wang and Jae Kwang Kim[1]

Abstract

Statistical inference with non-probability survey samples is a notoriously challenging problem in statistics. We introduce two new methods of nonparametric propensity score technique for weighting in the non-probability samples. One is the information projection approach and the other is the uniform calibration in the reproducing kernel Hilbert space.

Key Words: Information projection; Uniform function calibration; Data integration.

1. Introduction

We would like to congratulate Dr. Changbao Wu on the outstanding work in non-probability sampling. Even though probability sampling served as a golden standard tool for finite population inference in the past decades, it has recently become tarnished gold due to low response rates and high costs. Non-probability sampling, on the other hand, is popular due to its feasibility and low cost (Couper, 2000; Kaplowitz, Hadlock and Levine, 2004). More importantly, non-probability sampling, such as a web survey, can quickly gather up-to-date information when compared to a probability sample. However, because the selection mechanism is unavailable for non-probability sampling, failing to correct the selection bias in analyzing a non-probability sample may result in inefficiency or even erroneous inference. As a result, adjusting the selection bias for a non-probability sample is a fundamental topic for survey sampling researchers, and this work presents the most comprehensive answers to this subject.

Dr. Wu's research, in particular, includes a thorough examination of propensity score (PS) techniques. Those PS techniques, on the other hand, have drawbacks. First, even for a correctly specified PS model, the inverse probability weighting estimator may be inefficient due to small estimated propensity scores. One alternative is post-stratification, as stated in Section 5 of the paper, although there is no clear guidance on how to choose K. Furthermore, in practice, correctly specifying a PS model is difficult. While doubly robust estimation can help to safeguard a bad PS model, the final estimator is problematic when both the PS and regression models are incorrect (Kang and Schafer, 2007).

To overcome the misspecification of the PS model, Dr. Wu has mentioned several nonparametric methods, including a kernel method and a tree-based method. In this discussion, we would like to expand on this direction and provide two more methods to augment the study. One is based on a density ratio model using information projection (Csiszár and Shields, 2004), and the other is by uniformly calibrating functions over a reproducing kernel Hilbert space (RKHS). As explained by Wahba (1990), RKHS is a very flexible function space for approximation. Instead of estimating the propensity scores, we aim at

[1] Zhonglei Wang, Wang Yanan Institute for Studies in Economics and School of Economics, Xiamen University, Xiamen, Fujian, People's Republic of China; Jae Kwang Kim, Iowa State University, Ames, IA 50011, USA. E-mail: jkim@iastate.edu.

estimating the sampling weights $\{(\pi_i^A)^{-1} : i \in S_A\}$ to avoid possible inefficiency due to small estimated propensity scores.

Denote S_A and S_B to be the index sets for the non-probability and reference probability samples, respectively, and the corresponding sample sizes are n_A and n_B. Let $\{(y_i, \mathbf{x}_i) : i \in S_A\}$ and $\{(\mathbf{x}_i, d_i^B) : i \in S_B\}$ be available, where y_i and $\mathbf{x}_i$ are the study variable and auxiliary vector for the i^{th} unit and d_i^B is the design weight for $i \in S_B$.

The paper is organized as follows. In Section 2, we introduce the information projection approach. In Section 3, we introduce the basic idea of uniform calibration. Some concluding remarks are made in Section 4.

2. Information projection approach

Suppose that we are interested in estimating parameter $\boldsymbol{\theta}_0$ defined through $E_N\{U(\boldsymbol{\theta}; \mathbf{X}, Y)\} = 0,$ where $E_N(\cdot)$ is the expectation with respect to the population empirical distribution $\Pr\{(\mathbf{X}, Y) = (\mathbf{x}_i, y_i)\} = N^{-1}$ for $i = 1, \ldots, N$ and 0 otherwise, and $U(\boldsymbol{\theta}; \mathbf{x}, y)$ is a certain estimating function. For example, $U(\theta; \mathbf{x}, y) = y - \theta$ corresponds to $\mu_y = N^{-1}\sum_{i=1}^N y_i$ in the paper. We wish to obtain an estimator of $(\pi_i^A)^{-1}$, $\pi_i^A = \Pr(R_i = 1 \mid \mathbf{x}_i, y_i)$, and $R_i = 1$ if $i \in S_A$ and 0 otherwise.

To estimate $\{(\pi_i^A)^{-1} : i \in S_A\}$, we may use the relationship in the density ratio function. First, we consider a super-population model ξ, and let $f_0(\mathbf{x}, y)$ and $f_1(\mathbf{x}, y)$ be the density functions of $(\mathbf{x}, y)$ given $R = 0$ and $R = 1$, respectively. Denote the density ratio function to be

$$r(\mathbf{x}, y) = \frac{f_0(\mathbf{x}, y)}{f_1(\mathbf{x}, y)},$$

and by the Bayes formula, we have

$$(\pi_i^A)^{-1} = 1 + \frac{\Pr(R_i = 0)}{\Pr(R_i = 1)} r(\mathbf{x}_i, y_i). \tag{2.1}$$

Thus, there is a one-to-one relationship between $(\pi_i^A)^{-1}$ and $r(\mathbf{x}_i, y_i)$.

Under assumption A1, we can show that $r(\mathbf{x}, y) = r(\mathbf{x})$. In this section, we make a more general assumption that there exists $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \ldots, b_L(\mathbf{x}))^{\mathrm{T}}$ such that

$$R \perp Y \mid \mathbf{b}(\mathbf{x}). \tag{2.2}$$

Rosenbaum and Rubin (1983) called $\mathbf{b}(\mathbf{x})$ in (2.2) balancing scores.

To estimate the density ratio function $r(\mathbf{x})$, we minimize the Kullback-Leibler divergence

$$Q(f_0) = \int \log(f_0 / f_1)\, f_0 \,\mathrm{d}\mu \tag{2.3}$$

with respect to f_0 subject to some constraint, where both f_0 and f_1 are absolutely continuous with respect to a σ-finite measure $\mu.$ Regarding the constraint, we may use the following one

$$\Pr(R_i = 1) \int \mathbf{b}(\mathbf{x}) f_1(\mathbf{x}) \mu(d\mathbf{x}) \; + \; \Pr(R_i = 0) \int \mathbf{b}(\mathbf{x}) f_0(\mathbf{x}) \mu(d\mathbf{x}) \; = \; E_\xi\{\mathbf{b}(\mathbf{X})\}, \tag{2.4}$$

where $E_\xi(\cdot)$ is the expectation with respect to the super-population model $\xi.$ That is, given $f_1(\mathbf{x}),$ we can find $f_0(\mathbf{x})$ to minimize (2.3) under a calibration constraint with respect to $\mathbf{b}(\mathbf{x}).$

By Lemma 3.1 of Wang and Kim (2021), the optimized conditional density function satisfies

$$f_0^*(\mathbf{x}) \; = \; f_1(\mathbf{x}) \frac{\exp\{\boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x})\}}{E_1\left[\exp\{\boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x})\}\right]}, \tag{2.5}$$

where $\boldsymbol{\lambda}_1$ is chosen to satisfy (2.4). Note that the solution (2.5) is equivalent to

$$\log\{r(\mathbf{x};\, \boldsymbol{\lambda})\} \; = \; \lambda_0 + \boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x}) \tag{2.6}$$

for the density ratio function $r(\mathbf{x}),$ where $\boldsymbol{\lambda} = (\lambda_0, \boldsymbol{\lambda}_1^{\mathrm{T}})^{\mathrm{T}},$ and λ_0 is a normalizing constant satisfying $\int r(\mathbf{x};\, \boldsymbol{\lambda}) f_1(\mathbf{x}) \mu(d\mathbf{x}) = 1.$ Thus, the information projection finds the best model for propensity score function.

Once the model is determined as in (2.6), we need to estimate the model parameters. Because of the moment constraints in (2.4), the sample-version estimating equation for $\boldsymbol{\lambda}$ is the calibration equation given by

$$\frac{n_A}{N} \sum_{i=1}^{N} R_i \left[1, \mathbf{b}(\mathbf{x}_i)\right] \left[1 + \frac{1 - n_A}{n_A} \exp\{\lambda_0 + \boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x}_i)\}\right] \; = \; \left[1, \frac{1}{N} \sum_{i \in S_B} d_i^B \mathbf{b}(\mathbf{x}_i)\right]. \tag{2.7}$$

Here, since $E_\xi\{\mathbf{b}(\mathbf{X})\}$ is not available, we use its estimate $N^{-1} \sum_{i \in S_B} d_i^B \mathbf{b}(\mathbf{x}_i).$ Once the parameter estimate $\hat{\boldsymbol{\lambda}}$ is obtained, we can construct

$$\hat{\omega}_i \; = \; 1 \; + \; \frac{1 - n_A}{n_A} \exp\{\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x}_i)\}$$

as the final PS weights. The parameter of interest can be estimated by solving $N^{-1} \sum_{i \in S_A} \hat{\omega}_i U(\boldsymbol{\theta};\, \mathbf{x}_i, y_i) = 0$ for $\boldsymbol{\theta}.$

Wang and Kim (2021) developed this framework under the non-probability sampling setup where $\mathbf{x}_i$ are available throughout the finite population. Consistency and the asymptotic normality can be developed under the assumption that $E\{U(\boldsymbol{\theta};\, \mathbf{x}, Y) \mid \mathbf{x}\}$ lies in the linear space generated by $\{b_1(\mathbf{x}), \ldots, b_L(\mathbf{x})\}.$ Instead of assuming the availability of $\{\mathbf{x}_i : i = 1, \ldots, N\}$ as in Wang and Kim (2021), there only exists a reference probability sample $\{(\mathbf{x}_i, d_i^B) :\, i \in S_B\}.$ If the probability sample S_B is a census, then the method above reduces to the one considered by Wang and Kim (2021), except that we consider a finite population parameter $\boldsymbol{\theta}_0.$ In Section 11.2 of Kim and Shao (2021), the information projection approach is called the maximum entropy method and applied to the data integration problem. In the simulation study presented in example 11.1 of the book, the proposed information projection method shows better performance than the methods of Chen, Li and Wu (2020) and Elliott and Valliant (2017).

3. Uniform calibration approach

Calibration is commonly used to improve the representativeness of a non-probability sample, but existing methods, including the information projection approach mentioned in Section 2, are based on calibrating a set of pre-specified functions. However, it is hard to correctly specify them for calibration in practice. In this section, we propose a general framework for uniformly calibrating functions in an RKHS. Instead of considering a parametric form for $E_{\xi}(Y \mid \mathbf{x})$ in (3.1), we only assume $E_{\xi}(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i)$, where $m(\mathbf{x})$ is a smooth function satisfying certain conditions.

We still consider (2.1) under the assumption A1. Instead of assuming a set of pre-specified functions $\mathbf{b}(\mathbf{x})$, we propose to estimate $\{r_i : i \in S_A\}$ by the following optimization,

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \geq 0}{\operatorname{argmin}} \left[\sup_{u \in H} \left\{ \frac{S(\boldsymbol{\gamma}, u)}{\| u \|_2^2} - \lambda_1 \frac{\| u \|_H^2}{\| u \|_2^2} \right\} + \lambda_2 Q_A(\boldsymbol{\gamma}) \right], \tag{3.1}$$

where $\boldsymbol{\gamma} = (r_1, \ldots, r_N)$, $r_i = 0$ for $i \notin S_A$, $\boldsymbol{\gamma} \geq 0$ is equivalent to $r_i \geq 0$ for $i = 1, \ldots, N$, H is an RKHS,

$$S(\boldsymbol{\gamma}, u) = \left[N^{-1} \sum_{i \in S_A} \left\{ 1 + \left(\frac{N}{n_A} - 1 \right) r_i \right\} u(\mathbf{x}_i) - N^{-1} \sum_{i \in S_B} d_i^B u(\mathbf{x}_i) \right]^2, \tag{3.2}$$

$\| u \|_2^2 = (n_A + n_B)^{-1} \sum_{i \in S_A \cup S_B} u(\mathbf{x}_i)^2$, $\| u \|_H$ is the norm associated with the RKHS, $Q_A(\boldsymbol{\gamma})$ is a general penalty on $\boldsymbol{\gamma}$ to avoid overfitting, and λ_1 and λ_2 are two tuning parameters; see Wahba (1990) for a detailed introduction about the RKHS.

The intuition for the optimization (3.1) is briefly discussed. First, if r_i approximates the true density ratio $r(\mathbf{x}_i)$ well, the bias of the first term in (3.1) is negligible for estimating $N^{-1} \sum_{i=1}^{N} u(\mathbf{x}_i)$ for $u \in H$. Besides, $N^{-1} \sum_{i \in S_B} d_i^B u(\mathbf{x}_i)$ is design-unbiased. Thus, $S(\boldsymbol{\gamma}, u)$ balances two estimators for $N^{-1} \sum_{i=1}^{N} u(\mathbf{x}_i)$, and it is small if r_i approximately equals $r(\mathbf{x}_i)$ for $i \in S_A$. However, $S(\boldsymbol{\gamma}, u)$ is not scale invariant, and we have $S(\boldsymbol{\gamma}, cu) = c^2 S(\boldsymbol{\gamma}, u)$ for $c \in \mathbb{R}$. Thus, we use $\| u \|_2^2$ to make it scale-invariant. The term $\lambda_1 \| u \|_H^2$ is used to penalize the smoothness of the function u for $u \in H$. There exist different choices for $Q_A(\boldsymbol{\gamma})$. For example, $Q_A(\boldsymbol{\gamma}) = \sum_{i \in S_A} \{1 + (N n_A^{-1} - 1) r_i\}^2$ corresponds to penalizing extreme values for the sampling weights, and Wong and Chan (2018) investigated a similar problem assuming the availability of $\{\mathbf{x}_i : i = 1, \ldots, N\}$. The optimization (3.1) can be viewed as a "minmax" problem, and if $m \in H$, the estimated density ratios $\{\hat{r}_i : i \in S_A\}$ may lead to a reasonably good estimator

$$\hat{\mu}_{uc} = N^{-1} \sum_{i \in S_A} \left\{ 1 + \left(\frac{N}{n_A} - 1 \right) \hat{r}_i \right\} y_i. \tag{3.3}$$

Uniform calibration is a new method for non-probability sampling, and there are some technical challenges in (3.1). For example, how to incorporate the design properties of S_B when establishing the theoretical properties of (3.3) has not be fully investigated, and we have finished a working paper about this topic (Wang, Mao and Kim, 2022). The kernel-based method is computationally expensive, especially when the sample sizes are large. It may be interesting to propose a more computationally efficient algorithm for the uniform calibration problem. One possible answer is to consider some other functional spaces, such as the one spanned by B-splines. In addition, it is also of interest to consider how to incorporate more than one reference probability sample, and how to formulate a uniform calibration if we have different covariates in different reference probability samples.

4. Concluding remarks

Propensity score weighting is an important tool for correcting selection bias in the nonprobability sampling. Dr. Changbao Wu made significant contributions on this important topic. In addition to the two additional methods, the empirical likelihood (EL) approach of Qin, Leung and Shao (2002) is potentially useful as another tool for propensity score weighting. In particular, the EL-based weighting method is applicable even under informative sampling. Further investigation on this direction will be explored elsewhere.

References

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, 64(4), 464-494.

Csiszár, I., and Shields, P.C. (2004). *Information Theory and Statistics: A Tutorial*.

Elliott, M., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Kang, J.D., and Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22(4), 523-539.

Kaplowitz, M.D., Hadlock, T.D. and Levine, R. (2004). A comparison of Web and mail survey response rates. *Public Opinion Quarterly*, 68(1), 94-101.

Kim, J.K., and Shao, J. (2021). *Statistical Methods for Handling Incomplete Data*, second edition. CRC press.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under non-ignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, 97, 193-200.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wang, H., and Kim, J.K. (2021). Information projection approach to propensity score estimation for handling selection bias under missing at random. *arXiv:2104.13469*, 1-34.

Wang, Z., Mao, X. and Kim, J.K. (2022). Functional calibration under non-probability survey sampling. Submitted (https://arxiv.org/abs/2204.09193).

Wong, R.K., and Chan, K.C.G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1), 199-213.