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

Survey Methodology

Comments on “Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples”

by Jae Kwang Kim and Yonghyun Kwon

Release date: June 25, 2024

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.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 57-63
Statistics Canada, Catalogue No. 12-001-X

Comments on "Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples"

Jae Kwang Kim and Yonghyun Kwon[1]

Abstract

Pseudo weight construction for data integration can be understood in the two-phase sampling framework. Using the two-phase sampling framework, we discuss two approaches to the estimation of propensity scores and develop a new way to construct the propensity score function for data integration using the conditional maximum likelihood method. Results from a limited simulation study are also presented.

Key Words: Data integration; Propensity score function; Pseudo weight; Two-phase sampling.

1. Introduction

We would like to congratulate Yan Li for being selected as a Morris Hansen lecturer and for giving an interesting presentation on data integration. Data integration is an emerging area of research to combine multiple data sources in a defensible way. In data integration, by using an independent probability sample as a calibration sample, the selection bias in the convenient sample can be reduced. However, statistical tools for data integration are limited. Thus, I welcome Li's attempt to develop an additional statistical tool for data integration.

Using the balancing score function to control selection bias in the nonprobability sample is a reasonable idea. How to construct the balancing score function in the context of data integration can be more tricky. Li recognized that the propensity score (PS) estimation method of Chen, Li and Wu (2020) can be inefficient, as the estimation procedure involves using the survey weights in the probability sample. Instead of using weighted estimation, Li proposed an unweighted estimation method and then developed a method for bias correction. The unweighted estimate of PS is also considered by Elliott and Valliant (2017) and has been adopted by some practitioners. In this discussion, we would like to clarify two existing approaches to the estimation of propensity scores and develop a defensible way of constructing the propensity score function for data integration.

The paper is organized as follows. In Section 2, we present a two-phase sampling framework for data integration and the conditional PS model approach is introduced. In Section 3, another approach, called the unconditional model approach, is introduced. The simulation study is presented in Section 4. Some concluding remarks are made in Section 5.

1. Jae Kwang Kim and Yonghyun Kwon, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A. E-mail: jkim@iastate.edu.

2. Conditional PS model approach

We use the set-up considered in Yang, Kim and Hwang (2021) where sample A is a probability sample observing $\mathbf{x}$ and sample B is the nonprobability sample observing $(\mathbf{x}, y)$. Table 2.1 presents the general setup of the two sample structures for data integration. As indicated in Table 2.1, sample B is not representative of the target population.

Table 2.1
Data structure for data integration and data fusion.

Data Integration				
Sample	**Type**	***X***	***Y***	**Representative?**
A	Probability Sample	✓		Yes
B	Non-probability Sample	✓	✓	No

The formulation is somewhat similar to the two-phase sampling:

1. The first-phase sample $S_1 \equiv A \cup B$ is selected from U and $\mathbf{x}_i$ is observed for all units in S_1.
2. The second-phase sampling $S_2 = B$ is selected from S_1 and y_i is observed for all units in S_2.

Unlike classical two-phase sampling, we do not know the first-order inclusion probability of S_1. Instead, we only know the first-order inclusion probability of the sample A. That is, $\pi_i^{(A)} = P(i \in A \mid i \in U)$ is the (known) first-order inclusion probability of sample A.

Let $\pi_i^{(B)} = P(i \in B \mid i \in U)$ be the (unknown) first-order inclusion probability of sample B. Note that the first-order inclusion probability of S_1 can be written as

$$\begin{aligned} P(i \in S_1 \mid i \in U) &= P(i \in A \cup B \mid i \in U) \\ &= P(i \in A \mid i \in U) + P(i \in B \mid i \in U) - P(i \in A \mid i \in U)P(i \in B \mid i \in U) \\ &= \pi_i^{(A)} + \pi_i^{(B)} - \pi_i^{(A)}\pi_i^{(B)} \end{aligned} \tag{2.1}$$

where the last equality follows from the independence of two samples. Thus, we can express the conditional inclusion probability for the second-phase sample as

$$P(i \in S_2 \mid i \in S_1) = \frac{P(i \in B \mid i \in U)}{P(i \in A \cup B \mid i \in U)} = \frac{\pi_i^{(B)}}{\pi_i^{(A)} + \pi_i^{(B)} - \pi_i^{(A)}\pi_i^{(B)}}. \tag{2.2}$$

Now, since we observe $\mathbf{x}_i$ for $i \in S_1 = A \cup B$, we can make a statistical model for the conditional inclusion probability in (2.2) as a function of $\mathbf{x}$. Let

$$P(i \in S_2 \mid i \in S_1) = p(\mathbf{x}_i; \phi) \tag{2.3}$$

be the statistical model for the conditional inclusion probability with unknown parameter ϕ. We can estimate ϕ by unweighted analysis. That is,

$$\hat{\phi} = \arg\max_{\phi} \sum_{i \in S_1} \left[\delta_i \log p(\mathbf{x}_i;\phi) + (1-\delta_i) \log\{1 - p(\mathbf{x}_i;\phi)\} \right],$$

where $\delta_i = \mathbb{I}(i \in B)$ is the indicator function of the event $i \in B$. If a logistic regression model with $\text{logit}\{p(\mathbf{x}_i;\phi)\} = \mathbf{x}_i'\phi$ is used in (2.3), then $\hat{\phi}$ can be obtained by solving

$$\sum_{i \in B} \{1 - p(\mathbf{x}_i;\phi)\}\mathbf{x}_i - \sum_{i \in A} p(\mathbf{x}_i;\phi)\mathbf{x}_i = \mathbf{0}.$$

This unweighted estimation is fully justified, as the conditional inclusion probability model (2.3) is conditional on the first-phase sample $S_1 = A \cup B$. Since the propensity model in (2.3) is conditional on the first-phase sample, it can be called the conditional propensity score (PS) model.

Now, since (2.3) is the model for the conditional inclusion probability in (2.2), we can obtain

$$\frac{\pi_i^{(B)}}{\pi_i^{(A)} + \pi_i^{(B)} - \pi_i^{(A)}\pi_i^{(B)}} = p(\mathbf{x}_i;\hat{\phi}),$$

which implies that

$$\frac{1}{\hat{\pi}_i^{(B)}} = 1 + \frac{1}{\pi_i^{(A)}} \left\{ \frac{1}{p(\mathbf{x}_i;\hat{\phi})} - 1 \right\}. \tag{2.4}$$

Thus, $\hat{w}_i^{(B)} = 1/\hat{\pi}_i^{(B)}$ in (2.4) can be used as the final pseudo-weight for the elements in sample B.

In practice, we cannot use (2.4) directly as the first-order inclusion probabilities are unknown outside the sample. One way to handle this problem is to estimate $w_i^{(A)} = 1/\pi_i^{(A)}$ by

$$\tilde{w}_i^{(A)} = E\{w_i^{(A)} \mid \mathbf{x}_i, I_i^{(A)} = 1\} \tag{2.5}$$

following the result of Pfeffermann and Sverchkov (1999). Thus, (2.4) can be changed to

$$\frac{1}{\hat{\pi}_i^{(B)}} = 1 + \tilde{w}_i^{(A)} \left\{ \frac{1}{p(\mathbf{x}_i;\hat{\phi})} - 1 \right\}. \tag{2.6}$$

Li used a parametric model for $E(\pi^{(A)} \mid \mathbf{x}) = \bar{\pi}^{(A)}(\mathbf{x};\gamma)$ and developed a pseudo maximum likelihood method for estimating γ from the sample. Once $\hat{\gamma}$ is obtained, we can use (2.6) with $\tilde{w}_i^{(A)} = 1/\tilde{\pi}(\mathbf{x}_i;\hat{\gamma})$.

Instead of using (2.6), Elliott and Valliant (2017) proposed using

$$\frac{1}{\hat{\pi}_i^{(B)}} = \frac{1}{\hat{\pi}_i^{(A)}} \left\{ \frac{1}{p(\mathbf{x}_i;\hat{\phi})} - 1 \right\} \tag{2.7}$$

where

$$\hat{\pi}_i^{(A)} = E\{\pi_i^{(A)} \mid \mathbf{x}_i, I_i^{(A)} = 1\}. \tag{2.8}$$

However, $\tilde{w}_i^{(A)} \neq 1/\hat{\pi}_i^{(A)}$ in general and the pseudo weight in (2.7) is not theoretically justified.

3. Unconditional PS model approach

Another approach to the PS model is to assume a statistical model for $\pi_i^{(B)} = P(i \in B \mid i \in U)$ such as

$$\pi_i^{(B)} = \pi_B(\mathbf{x}_i; \phi) \tag{3.1}$$

for some parameter ϕ. This unconditional PS model has been considered by Chen et al. (2020) and Wang, Valliant and Li (2021), where the pseudo maximum likelihood method was used to estimate ϕ.

If we wish to improve the efficiency of estimators of ϕ, we can consider the maximum likelihood method as follows. First, if $\pi_i^{(A)}$ are available in S_1, using (3.1), we can derive the following conditional inclusion probability model:

$$\pi_{2i|1}(\phi) = \frac{\pi_B(\mathbf{x}_i; \phi)}{\pi_i^{(A)} + \pi_B(\mathbf{x}_i; \phi) - \pi_i^{(A)} \cdot \pi_B(\mathbf{x}_i; \phi)}. \tag{3.2}$$

In the second step, we can compute the conditional maximum likelihood estimator of ϕ from the combined sample by

$$\hat{\phi} = \arg\max_{\phi} \sum_{i \in S_1} \left[\delta_i \log \pi_{2i|1}(\phi) + (1 - \delta_i) \log\{1 - \pi_{2i|1}(\phi)\} \right], \tag{3.3}$$

where $\pi_{2i|1}(\phi)$ is defined in (3.2). The conditional maximum likelihood estimator in (3.3) is based on the assumption that we can identify the units that belong to the intersection of A and B. Once $\hat{\phi}$ is obtained from the conditional maximum likelihood method, we can use $\hat{w}_i^{(B)} = 1/\pi^{(B)}(\mathbf{x}_i; \hat{\phi})$ as the pseudo weights for sample B. This conditional maximum likelihood method was also considered by Savitsky, Williams, Gershunskaya, Beresovskyl and Johnson (2022) under the assumption that $\pi_i^{(A)}$ are available in sample B.

If $\pi_i^{(A)}$ are not available outside the sample A, we cannot construct the conditional inclusion probability in (3.2). In this case, we can replace $\pi_i^{(A)}$ by $\tilde{\pi}_i^{(A)} = 1/\tilde{w}_i^{(A)}$, where $\tilde{w}_i^{(A)}$ is defined in (2.5), and compute

$$\pi_{2i|1}(\phi) = \frac{\pi_B(\mathbf{x}_i; \phi)}{\tilde{\pi}_i^{(A)} + \pi_B(\mathbf{x}_i; \phi) - \tilde{\pi}_i^{(A)} \cdot \pi_B(\mathbf{x}_i; \phi)} \tag{3.4}$$

to apply the above conditional maximum likelihood method in (3.3). The final pseudo weights are given by $\hat{w}_i^{(B)} = 1/\pi_B(\mathbf{x}_i; \hat{\phi})$ and $\hat{\phi}$ is computed by (3.3).

Instead of the maximum likelihood method, the pseudo weights for sample B can be constructed to satisfy

$$\sum_{i \in B} \frac{1}{\pi_B(\mathbf{x}_i; \phi)} \mathbf{x}_i = \sum_{i \in A} \frac{1}{\pi_i^{(A)}} \mathbf{x}_i. \tag{3.5}$$

Condition (3.5) is often called the calibration property. The calibration property is a desirable property for any pseudo-weights. Once $\hat{\phi}$ is calculated from the calibration equation in (3.5), the final pseudo weight for sample B is given by $\hat{w}_i^{(B)} = 1/\pi_B(\mathbf{x}_i;\hat{\phi})$.

4. Simulation study

A limited simulation study is conducted to compare the performance of estimators, including the methods suggested by the paper of Li. In the simulation, we generate a finite population with $y_i \sim \text{Bernoulli}(p_i)$, $p_i = \text{expit}(-1+0.8x_{1i}+0.2x_{2i}+0.5x_{1i}x_{2i})$ with (x_1, x_2, x_3) follows from the standared normal distribution. The finite population size is $N =$ 5,000.

From the finite population, sample A is generated repeatedly by the PPS sample with measure of size

$$mos_i = \exp(-1+0.5x_{1i}+0.5x_{3i}-0.2x_{1i}x_{3i})$$

with sample size $n_A =$ 250. In addition, sample B is selected repeatedly by stratified random sampling with two strata, where stratum 1 is $U_1 = i \in U : x_{1i} > 0$ and stratum 2 is $U_2 = i \in U : x_{1i} \le 0$. In stratum 1, $n_{B1} = 0.7n_B$ samples are selected by simple random sampling. In stratum 2, $n_{B2} = 0.3n_B$ samples are selected by simple random sampling. The sample size of B is chosen to be either $n_B =$ 250 or $n_B =$ 2,500 so that the sampling ratio is either 5% or 50%. The design weights for sample A are available in sample A, but not in sample B. The study variable y is available only in sample B. The covariate of the main effects and their pairwise interaction effects $(x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3)$ are available in $A \cup B$.

We compare the following estimators:

Mean C Sample mean of the nonprobability sample C. *Unweighted* in the paper.

WBS ALP(Adjusted Logistic Propensity) estimator using weighted balancing score method, proposed by Wang et al. (2021).

ABS ALP estimator using adaptive balancing score method, proposed by Li.

CLW Chen et al. (2020)'s IPW(Inverse Probability Weighting) estimator using logistic regression model for $\pi_i^{(B)}$.

Cal Calibration estimator that satisfies (3.5) using logistic regression model for $\pi_i^{(B)}$.

CPS The proposed pseudo weight estimator (2.6) using the conditional inclusion probability model and the smoothed weights in (2.5). The logistic regression model is used for the conditional inclusion probability model, and Poisson regression was used for smoothing weights of sample A in (2.5).

UCPS The pseudo weight estimator proposed in Section 3 using the logistic regression model for $\pi_i^{(B)}$ with $\hat{\phi}$ estimated by the conditional maximum likelihood method in (3.3).

While the sample B is selected using stratified sampling, the propensity scores of **WBS**, **ABS**, **CLW**, **CPS**, and **UCPS** were fitted from the logistic model, and we allowed model misspecification on the response model of $\pi^{(B)}$.

The simulation results after 1,000 simulation runs are summarized in Table 4.1. When n_B = 250, the ABS, the CPS, and the UCPS estimators tend to outperform all other estimators considered. When n_B = 2,500, the CPS and UCPS estimators are better than the other estimators considered. The ABS and WBS methods are developed based on the assumption that the overlap between the two samples is negligible, but this assumption does not hold for n_B = 2,500, as the sampling rate for sample B, n_B / N = 0.5, is non-negligible.

Table 4.1
Bias, standard error, and root mean square error after 1,000 repetitions.

	n_B = 250			n_B = 2,500		
	BIAS	**SE**	**RMSE**	**BIAS**	**SE**	**RMSE**
Mean C	0.0533	0.0252	0.0589	0.0514	0.0052	0.0517
WBS	0.0087	0.0275	0.0289	0.0053	0.0139	0.0149
ABS	0.0097	0.0264	0.0281	0.0097	0.0130	0.0162
CLW	0.0084	0.0278	0.0291	-0.0081	0.0234	0.0248
Cal	0.0061	0.0284	0.0291	0.0080	0.0140	0.0161
CPS	0.0095	0.0263	0.0279	0.0035	0.0116	0.0121
UCPS	0.0094	0.0263	0.0280	0.0035	0.0116	0.0121

5. Concluding remark

In constructing pseudo-weights, model assumptions for the nonprobability sample are used. The model assumptions can be classified into two groups, one is the conditional PS model approach and the other is the unconditional PS model approach. The conditional PS model approach is computationally attractive but the smoothing weights for sample A should be constructed correctly. In the unconditional PS model approach, the pseudo maximum likelihood method of Chen et al. (2020) has been used. Li's method is more efficient than the pseudo maximum likelihood method as long as the sampling rate for sample B is negligible. In this paper, we propose an alternative approach using the conditional maximum likelihood method as an efficient estimation method, which can be justified even when the sampling rate for sample B is non-negligible. The computation for the conditional maximum likelihood method is somewhat involved. Beaumont, Bosa, Brennan, Charlebois and Chu (2024) independently proposed a very similar method, which was called the maximum sample likelihood method. Further investigation of the proposed method will be presented elsewhere.

Acknowledgements

We thank Professor Partha Lahiri for the invitation for discussion and two anonymous referees and the Editor for constructive comments. The research was partially supported by a grant from the US National Science Foundation (Grant Number: 2242820) and a grant from the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

References

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Author's response to comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data": Some new developments on likelihood approaches to estimation of participation probabilities for non-probability samples. *Survey Methodology*, 50, 1, 123-141. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00001-eng.pdf.

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.

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

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā B*, 61, 166-186.

Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovskyl, V. and Johnson, N.G. (2022). Methods for combining probability and nonprobability samples under unknown overlaps. arXiv:2208.14541.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40(24), 5237-5250.

Yang, S., Kim, J.K. and Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf.