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

Survey Methodology

Comments on “Handling non-probability samples through inverse probability weighting with an application to Statistics Canada’s crowdsourcing data”

by Julie Gershunskaya and Vladislav Beresovsky

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. 107-115
Statistics Canada, Catalogue No. 12-001-X

Comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data"

Julie Gershunskaya and Vladislav Beresovsky[1]

Abstract

Beaumont, Bosa, Brennan, Charlebois and Chu (2024) propose innovative model selection approaches for estimation of participation probabilities for non-probability sample units. We focus our discussion on the choice of a likelihood and parameterization of the model, which are key for the effectiveness of the techniques developed in the paper. We consider alternative likelihood and pseudo-likelihood based methods for estimation of participation probabilities and present simulations implementing and comparing the AIC based variable selection. We demonstrate that, under important practical scenarios, the approach based on a likelihood formulated over the observed pooled non-probability and probability samples performed better than the pseudo-likelihood based alternatives. The contrast in sensitivity of the AIC criteria is especially large for small probability sample sizes and low overlap in covariates domains.

Key Words: Non-probability sample; Participation probabilities; Sample likelihood; Data combining.

Building on recent developments in data integration methods, Beaumont et al. (2024) propose and apply to real data innovative model selection approaches for estimation of participation probabilities for non-probability sample units. We congratulate the authors for their inspiring and timely contribution and appreciate the opportunity to comment and discuss the methods considered in the paper.

Survey statisticians are increasingly faced with the need to extract useful information from data collected without a well-planned probability survey design. At the same time, we witness rapid developments of machine learning methods capable to efficiently handle multidimensional sets of covariates. There is the growing realization that machine learning can be useful for handling estimation from non-probability samples. The current paper leads the way in adapting these methods for the combined non-probability and probability samples setup. The authors propose a general modified AIC formula that accounts for the probability sampling design in the combined samples setup. They also derive the AIC expression for the special case of homogeneous groups and apply it to choose among partitions in rank-based methods for forming the groups. Finally, the authors adapt the CART tree-growing algorithm by using a pseudo-likelihood as an objective function and applying the modified AIC to prune the tree.

We focus our discussion on the choice of a likelihood and parameterization of the model, which are key for the effectiveness of the techniques developed in the paper. In Section 1, we review several approaches for estimation of participation probabilities proposed in recent years and provide AIC expressions for the homogeneous groups case. In Section 2, we present simulations implementing and comparing the AIC based variable selection for these approaches. We provide some concluding remarks in Section 3.

1. Julie Gershunskaya and Vladislav Beresovsky, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE Washington, DC 20212, U.S.A. E-mail: Gershunskaya.Julie@bls.gov.

1. Approaches to estimation of participation probabilities

We start with the two pseudo-likelihood based approaches considered in the paper, then we discuss an exact likelihood based approach and propose a modification to the ALP method of Wang, Valliant and Li (2021), see also a related discussion in Gershunskaya and Lahiri (2023). We then consider the case of homogeneous groups and derive and compare the AIC expressions for each of the approaches for this important and relatively simple special case. Throughout, unless explicitly stated, we follow the notation of the current paper.

1.1 CLW method

Assuming that both non-probability and probability samples are selected from the same finite population U, Chen, Li and Wu (2020) (hereafter CLW) start by writing a log-likelihood over units in U, with respect to Bernoulli variable δ_k:

$$\ell^{\mathrm{CLW}}(\boldsymbol{\alpha}) = \sum_{k\in U}\left\{\delta_k \log\left[p_k(\boldsymbol{\alpha})\right] + (1-\delta_k)\log\left[1-p_k(\boldsymbol{\alpha})\right]\right\}, \tag{1.1}$$

where δ_k is unit's k non-probability sample inclusion indicator, $p_k(\mathbf{x_k}) = P\{\delta_k = 1 \mid \mathbf{x_k}\}$, and $\boldsymbol{\alpha}$ is the parameter vector in a logistic regression model for p_k, where $\mathrm{logit}(p_k(\boldsymbol{\alpha})) = \boldsymbol{\alpha}^T\mathbf{x_k}$.

Since finite population units are not observed, CLW re-group the sum in (1.1) by presenting it as a sum of two parts: part 1 involves the sum over the non-probability sample units, s_{NP}, and part 2 is the sum over finite population U:

$$\ell^{\mathrm{CLW}}(\boldsymbol{\alpha}) = \sum_{k\in s_{\mathrm{NP}}}\log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right] + \sum_{k\in U}\log\left[1-p_k(\boldsymbol{\alpha})\right]. \tag{1.2}$$

CLW employ a pseudo-likelihood approach by replacing the sum over the finite population with its probability sample based estimate:

$$\hat{\ell}^{\mathrm{CLW}}(\boldsymbol{\alpha}) = \sum_{k\in s_{\mathrm{NP}}}\log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right] + \sum_{k\in s_P} w_k \log\left[1-p_k(\boldsymbol{\alpha})\right], \tag{1.3}$$

where weights $w_k = \pi_k^{-1}$ are inverse values of the reference sample inclusion probabilities π_k. Estimates are obtained by solving respective pseudo-likelihood based estimating equations.

Note that the likelihood under this approach is formulated with respect to indicator δ_k, although this variable is not observed.

1.2 ALP method

For their Adjusted Logistic Propensity (ALP) weighting method, Wang et al. (2021) introduce an imaginary construct consisting of two parts: they *stack* together non-probability sample s_{NP} (part 1) and finite population U (part 2). Since non-probability sample units belong to the finite population, they appear

in the stacked set twice. They formulate a Bernoulli likelihood for variable R_k, where $R_k = 1$ if unit k belongs to part 1 of the stacked set; and $R_k = 0$ otherwise:

$$\ell^{\text{ALP}}(\boldsymbol{\gamma}) = \sum_{k \in s_{\text{NP}}} \log\left[p_{Rk}(\boldsymbol{\gamma})\right] + \sum_{k \in U} \log\left[1 - p_{Rk}(\boldsymbol{\gamma})\right], \tag{1.4}$$

where $\boldsymbol{\gamma}$ is the parameter vector in a logistic regression model for $p_{Rk}(\mathbf{x_k}) = P\{R_k = 1 \mid \mathbf{x_k}\}$. Since the finite population is not available, they apply a pseudo-likelihood approach:

$$\hat{\ell}^{\text{ALP}}(\boldsymbol{\gamma}) = \sum_{k \in s_{\text{NP}}} \log\left[p_{Rk}(\boldsymbol{\gamma})\right] + \sum_{k \in s_P} w_k \log\left[1 - p_{Rk}(\boldsymbol{\gamma})\right], \tag{1.5}$$

leading to an estimate of p_{Rk}. However, the actual goal is to find probabilities p_k rather than p_{Rk}. At the second step of the ALP method, estimates of p_k are derived from identity

$$p_{Rk} = \frac{p_k}{1 + p_k}. \tag{1.6}$$

Wang et al. (2021) noted that in their simulations the ALP estimator was more efficient than the CLW, especially when the non-probability sample size was much larger than the probability sample size.

1.3 ILR method

Let us now consider an exact likelihood approach formulated over the pooled non-probability and probability samples. Savitsky, Williams, Gershunskaya and Beresovsky (2023) propose to stack together the two samples and consider indicator variable $z_k = 1$ if unit k belongs to the non-probability sample (part 1), and $z_k = 0$ if unit k belongs to the probability sample (part 2). Under this stacked sample construction, if there is an overlap between the two samples, s_{NP} and s_P, then the overlapping units are included into the stacked set, s, twice: once as part of the non-probability sample (with $z_k = 1$) and once as part of the reference probability sample (with $z_k = 0$). We do not need to know which units overlap or whether there are any overlapping units. They use first principles to prove a relationship between probabilities $p_{zk}(\mathbf{x_k}) = P\{z_k = 1 \mid \mathbf{x_k}\}$ of being in part 1 of the stacked set, on the one hand, and inclusion probabilities, p_k and π_k, on the other hand:

$$p_{zk} = \frac{p_k}{\pi_k + p_k}. \tag{1.7}$$

Elliott (2009) and Elliott and Valliant (2017) derived expression (1.7) under the assumption of non-overlapping samples s_{NP} and s_P. The derivation given in Savitsky et al. (2023) does not require this assumption.

To obtain estimates of p_k, Beresovsky (2019) proposed an approach, labeled Implicit Logistic Regression (ILR), to allow the estimation of p_k directly from the likelihood formulated on the combined sample. In ILR, probabilities $p_k = p_k(\boldsymbol{\alpha})$ are parameterized as $\text{logit}(p_k(\boldsymbol{\alpha})) = \boldsymbol{\alpha}^T \mathbf{x_k}$ (as in CLW), and

identity (1.7) is used to present p_{zk} as a composite function of $\boldsymbol{\alpha}$; that is, $p_{zk} = p_{zk}(p_k(\boldsymbol{\alpha})) = p_k(\boldsymbol{\alpha})/(\pi_k + p_k(\boldsymbol{\alpha}))$. The log-likelihood for observed Bernoulli variable z_k is

$$\ell^{\text{ILR}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\left[p_{zk}(p_k(\boldsymbol{\alpha}))\right] + \sum_{k \in s_P} \log\left[1 - p_{zk}\left(p_k(\boldsymbol{\alpha})\right)\right]. \tag{1.8}$$

The score equations are obtained from (1.8) by taking the derivatives, with respect to $\boldsymbol{\alpha}$, of composite function $p_{zk} = p_{zk}(p_k(\boldsymbol{\alpha}))$. This way, the estimates of p_k are obtained directly from (1.8) in a single step.

Note that for the ILR approach, the probability sample inclusion probabilities π_k are supposed to be known for all units in the combined set. This is possible for many probability surveys. If not immediately available, probabilities π_k for units in s_{NP} can be determined if probability sample design variables are available for non-probability sample units. As discussed in Elliott and Valliant (2017), π_k can be estimated using a regression model. Savitsky et al. (2023) used the Bayesian modeling technique to obtain both π_k and p_k. On the other hand, if probabilities π_k are not available for the non-probability part of the combined sample, one can apply a pseudo-likelihood approach labeled "pseudo-ILR", as discussed below in Section 1.4.

1.4 Pseudo-ILR method

The estimation method of Wang et al. (2021) can be modified to a one-step estimation procedure similar to ILR: p_k can be parameterized using the logistic link function as $\text{logit}(p_k(\boldsymbol{\alpha})) = \boldsymbol{\alpha}^T \mathbf{x}_\mathbf{k}$, while probabilities p_{Rk} in (1.6) could be viewed as a composite function, $p_{Rk} = p_{Rk}(p_k(\boldsymbol{\alpha})) = p_k(\boldsymbol{\alpha})/(1 + p_k(\boldsymbol{\alpha}))$. The pseudo-likelihood takes the form:

$$\hat{\ell}^{\text{PILR}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\left[p_{Rk}(p_k(\boldsymbol{\alpha}))\right] + \sum_{k \in s_P} w_k \log\left[1 - p_{Rk}(p_k(\boldsymbol{\alpha}))\right]. \tag{1.9}$$

This change in estimation of model parameters makes the approach more efficient and less biased than ALP. It also avoids cases where estimates of p_k become greater than 1, as may occur under the ALP approach where the estimation is performed in two steps.

Note that pseudo-likelihood based (1.3) and (1.9) use exactly the same set of observed data and yet these expressions are quite different. We expect the pseudo-ILR approach to give more efficient estimates because it is based on a likelihood properly formulated with respect to observed Bernoulli variable R_k, while the CLW likelihood is given with respect to unobserved δ_k. Our simulations (not included in the discussion) confirm a better performance of the pseudo-ILR relative to the CLW approach. The effect on the AIC performance is shown in simulations Section 2.

1.5 Homogeneous groups

The authors presented the log-likelihood and AIC expressions under the CLW approach for the special case of homogeneous groups. We now extend their approach to pseudo-ILR and ILR methods.

For the pseudo-ILR approach, it is easy to see that, for a given partition, estimates of group g participation probabilities p_g are $\hat{p}_g = n_g^{\text{NP}} \big/ \hat{N}_g$, where $\hat{N}_g = \sum_{k \in s_p} w_k$ (the same as in the CLW approach.) The AIC is given by

$$\text{AIC}^{\text{PILR}} = -2\hat{\ell}^{\text{PILR}}(\hat{\boldsymbol{\alpha}}) + 2G + 2\sum_{g=1}^{G}\left[\frac{1-\hat{p}_g}{1+\hat{p}_g}\right] n_g^{\text{NP}} \frac{\left[\hat{cv}_d(\hat{N}_g)\right]^2}{1-\hat{p}_g}, \tag{1.10}$$

where the log-likelihood is

$$\hat{\ell}^{\text{PILR}}(\hat{\boldsymbol{\alpha}}) = \sum_{g=1}^{G} \hat{N}_g \left[\hat{p}_g \log\frac{\hat{p}_g}{1+\hat{p}_g} + \log\frac{1}{1+\hat{p}_g}\right]. \tag{1.11}$$

Note that last terms in formulas for AIC under the CLW and pseudo-ILR approaches differ by a factor $(1-\hat{p}_g)\big/(1+\hat{p}_g) < 1$. That is, for a given partition, the penalty term in the pseudo-ILR approach is always smaller than the one in the CLW approach.

In the ILR approach, estimates of p_g are not available in closed form. They can be found by solving equations:

$$\sum_{k \in s_g} \frac{\pi_{gk}}{\pi_{gk} + p_g} = n_g^P, \tag{1.12}$$

where π_{gk} are assumed known, s_g is the part of combined sample that belongs to group g, $g = 1,\ldots,G$. Since ILR is based on an exact likelihood, the AIC formula for ILR does not have the third term and is a standard AIC expression:

$$\text{AIC}^{\text{ILR}} = -2\hat{\ell}^{\text{ILR}}(\hat{\boldsymbol{\alpha}}) + 2G, \tag{1.13}$$

where

$$\hat{\ell}^{\text{ILR}}(\hat{\boldsymbol{\alpha}}) = \sum_{g=1}^{G}\left[\sum_{k \in s_g^{\text{NP}}} \log\frac{\hat{p}_g}{\pi_{gk} + \hat{p}_g} + \sum_{k \in s_g^{P}} \log\frac{\pi_{gk}}{\pi_{gk} + \hat{p}_g}\right]. \tag{1.14}$$

Let us compare the penalty terms of the three methods for a given set of homogeneous groups. (The partitioning itself depends on the likelihood used, but we do not consider this effect at the moment.) We suppose that, all other factors being equal, the smaller the penalty term, the better AIC works. Thus, we expect the ILR based criteria to perform better than the pseudo-ILR or CLW. In turn, the pseudo-ILR is expected to work better than CLW, especially when the non-probability sample is relatively large and $\hat{p}_g$ in (1.10) and in formula (3.15) of the discussed paper are closer to 1. Simulations results of Section 2 support this reasoning.

2. Simulations

We conducted a simulation experiment to study performances of the AIC based on CLW, pseudo-ILR, and ILR approaches.

For each unit $k = 1\ldots, N$ in a finite population U of size $N =$ 10,000, we generated covariates x_{1k} and x_{2k} as independent standard normal variables.

We use Poisson sampling with participation probabilities p_k to select non-probability sample from population U. Probabilities p_k are generated as

$$\text{logit}(p_k) = \alpha_0 + \alpha_1 x_{1k} + \alpha_2 x_{2k}, \tag{2.1}$$

where we set specific coefficient values for different simulation scenarios, as follows:

- setting $\alpha_0 = -5$ produces sample s_{NP} of approximate size $n^{\text{NP}} = 100$; setting $\alpha_0 = -2.5$ produces sample s_{NP} of approximate size $n^{\text{NP}} =$ 1,000;
- α_1 was set to 1 for all scenarios;
- α_2 is set to values on a grid from -0.3 to 0.3, corresponding to a series of scenarios. Note that setting $\alpha_2 = 0$ corresponds to the case where p_k is independent of x_{2k}; larger values of α_2 produce stronger dependence on x_{2k}.

For probability sample s_P, we considered scenarios where sample sizes are $n^P =$ 100 or $n^P =$ 1,000. The probability sample is generated using the probability proportional to size (PPS) without replacement design, where the measure of size is defined as

$$\text{logit}(m_k) = 1 + \beta x_{1k}. \tag{2.2}$$

In multivariate models with a large number of auxiliary variables and interactions, it is likely that non-probability and probability samples would have very little overlap in some of the variable-defined domains. Firth (1993) and Heinze and Schemper (2002) demonstrated that little overlap, or separation, may result in unstable estimates of model parameters. This is the case where it is essential to use an effective method for variables selection. Therefore, in our simulations, we included scenarios of low and high overlap in variables domains. Values of coefficient β are set to regulate the degree of the overlap across the range of covariate x_1. To simulate the "high" overlap, we set $\beta = 1$ (so that $\beta = \alpha_1$); for the "low" overlap scenario, we set $\beta = -1$.

Table 2.1 presents a summary of considered simulation scenarios, S1-S4, characterized by combinations of high or low overlap and different sample sizes. We applied three approaches, CLW, pseudo-ILR, and ILR, for each scenario to choose between two models:

$$\begin{aligned} &\text{Full Model: } \text{logit}(p_k) = \alpha_0 + \alpha_1 x_{1k} + \alpha_2 x_{2k}, \\ &\text{Abbreviated Model: } \text{logit}(p_k) = \alpha_0 + \alpha_1 x_{1k}. \end{aligned} \tag{2.3}$$

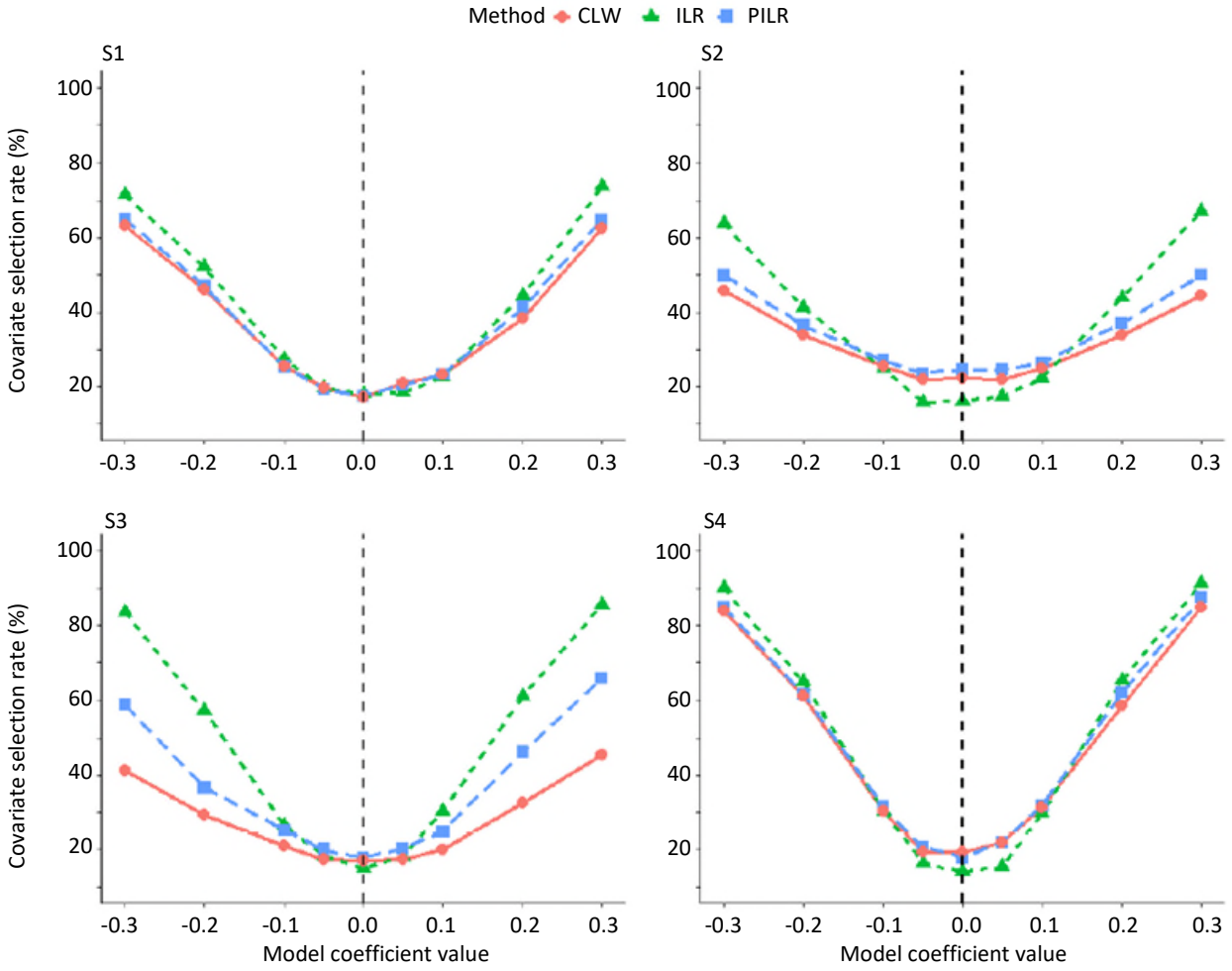
Table 2.1
Summary of considered simulation scenarios.

	Overlap	n^P	n^{NP}
S1	High	100	100
S2	Low	100	100
S3	Low	100	1,000
S4	Low	1,000	100

In each case, we compute AIC for the Full and Abbreviated models and choose the model with the lower AIC. We repeat this test $B = 500$ times for each scenario and the value of α_2 and find the percentage of times r the Full Model is chosen, $p = 100r/B$.

Plots in Figure 2.1 correspond to the four scenarios in Table 2.1. We plot percentage p against the values of coefficient α_2. For larger absolute values of α_2, the higher percentage p would be preferable; when α_2 is close to 0, the lower values of p would indicate a better performance. The line with red dots shows the CLW results, the line with blue squares is for pseudo-ILR, and the line with green triangles shows results for ILR.

Figure 2.1 Relative AIC performances under S1-S4 scenarios for CLW (red dots), ILR (green triangles), and pseudo-ILR (blue squares).



For all scenarios considered, the ILR approach performs the best: for larger absolute values of α_2, the ILR based AIC most frequently chooses the Full Model; when α_2 is closer to 0, the ILR based criteria selects the Full Model the least number of times. In the high overlap (S1) or relatively large probability sample size (S4) scenarios, all three approaches produce close results. However, when the probability sample is relatively small and the overlap is low (S2 and S3), the performance of ILR based test is significantly better than the other two methods. In most cases, the pseudo-ILR based test is slightly better than the CLW approach. When the non-probability sample size is large relative to the probability sample size (S3), the difference in performances increases: when absolute value of α_2 is close to 0.3, the CLW based criteria chooses the Full Model in only about 40-50% of times, whereas the pseudo-ILR based criteria chooses it in about 60%, and the ILR based test chooses it in roughly 85% of cases.

3. Concluding remarks

We commend the authors on their contribution in adapting model selection algorithms to data integration problems. The choice of an objective function is important for this task. In our discussion, we considered several recently proposed alternative likelihood functions. The exact likelihood based ILR method performed better than the pseudo-likelihood based alternatives in the important practical situation of small probability sample sizes and low overlap in covariates domains.

We note disadvantage of the CLW method when the probability sample is small and the non-probability sample is relatively large. In this case, we also noticed convergence problems with the Newton-Raphson algorithm in the CLW method.

The ILR requires that the probability sample inclusion probabilities be available for non-probability sample units. If these probabilities can be derived based on available design variables, then ILR would be the preferred method. Otherwise, the pseudo-ILR appears to be a viable option.

References

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50, 1, 77-106. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-eng.pdf.

Beresovsky, V. (2019). On application of a response propensity model to estimation from web samples. In ResearchGate.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Elliott, M.R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2, 813-845.

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

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27-38.

Gershunskaya, J., and Lahiri, P. (2023). Discussion of "Probability vs. nonprobability sampling: From the birth of survey sampling to the present day", by Graham Kalton. *Statistics in Transition New Series*, 24(3).

Heinze, G., and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21, 2409-2419.

Savitsky, T.D., Williams, M.R., Gershunskaya, J. and Beresovsky, V. (2023). Methods for combining probability and nonprobability samples under unknown overlaps. *Statistics in Transition New Series*, 24(5), 1-34.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.*, 40(4), 5237-5250.