The Survey Statistician, 2023, Vol. 88, 86–93.

New and Emerging Methods

Calibration Techniques for Model-based Prediction and Doubly Robust Estimation

Changbao Wu

University of Waterloo, Canada, cbwu@uwaterloo.ca

Abstract

We present a brief overview of calibration techniques for model-assisted estimation for probability survey samples and their extensions for model-based prediction and doubly robust estimation to missing data problems, causal inference, and analysis of non-probability samples. The focuses are to provide a clear description of the setting for each of these areas and on how doubly robust estimators are constructed either through a set of calibration equations or using model-calibrated empirical likelihood methods. Theoretical details are left to additional references.

Keywords: Empirical likelihood; Double robustness; Inverse probability weighting; Model-assisted estimation; Model-calibration.

1 Introduction

Survey samples have an important feature of representing a finite target population. Statistical tools for dealing with descriptive finite population parameters are often discrete in nature, such as series summations and double summations. There has been a separation between survey sampling and the so-called mainstream statistics in terms of tools and methodologies, highlighted by the extensive use of parametric or semi-parametric models, the likelihood principle, Bayesian pedagogies, etc., in other fields of statistics but not or less so in survey sampling. The field of survey sampling often lags behind on development of innovative general statistical tools.

There have been examples, however, where a method was first developed or rooted in survey sampling and later became widely used in other fields of statistics. The most prominent example is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952; Narain, 1951), which is popularly termed as the "*inverse probability weighted*" (IPW) estimator and is a fundamental tool for propensity score based methods in missing data analysis and causal inference. Another less known example is the doubly robust estimator, also popularized in missing data and causal inference literature starting from the 1990s. It is rooted in model-assisted estimation methods first developed in survey sampling going back to the 1970s. The generalized difference estimator of the population mean $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ of the study variable y, where N is the population size, as discussed in Cassel et al. (1976) is given by

$$\hat{\mu}_{yGD} = \frac{1}{N}\Big\{ \sum_{i\in\mathcal{S}} \frac{y_i - c_i}{\pi_i} + \sum_{i=1}^{N} c_i \Big\}, \tag{1}$$

where $\mathcal{S}$ is a probability sample of size n, the π_i's are the first order inclusion probabilities, and $\{c_1, c_2, \ldots, c_N\}$ is an arbitrary sequence of known numbers. The estimator $\hat{\mu}_{yGD}$ is exactly unbiased for μ_y under the probability sampling design p for any given sequence c_i, and is also model-unbiased if we choose $c_i = m_i = E_\xi(y_i \mid \boldsymbol{x}_i)$ under the assumed model ξ on y given $\boldsymbol{x}$. The estimator $\hat{\mu}_{yGD}$ with the choice $c_i = m_i$ is the same as the doubly robust estimator in the missing data and causal inference literature where π_i becomes the propensity score and m_i is the mean function of the outcome regression. Both π_i and m_i require an assumed model to be estimated, and the estimator remains valid if one of the models is correctly specified. The doubly robust estimator is also called the "*augmented inverse probability weighted*" (AIPW) estimator in the literature.

Calibration methods are also first developed in survey sampling and later find general uses in other areas. While the popularity of calibration methods is often credited to the highly cited JASA paper by Deville and Särndal (1992), the original idea of calibration estimation goes back to Deming and Stephan (1940) on raking ratio estimators. The model-calibration approach proposed by Wu and Sitter (2001) serves as the basis for the discussions presented in the rest of the paper on model-based prediction and doubly robust estimation.

2 Calibration methods for probability survey samples

The fundamental tool for design-based approach to survey sampling is the Horvitz-Thompson estimator for the finite population total $T_y = \sum_{i=1}^N y_i$, which is given by $\hat{T}_{yHT} = \sum_{i\in\mathcal{S}} d_i y_i$, where $d_i = 1/\pi_i$ are the basic design weights. Most surveys collect information on a vector of auxiliary variables, $\boldsymbol{x}$, leading to a survey dataset $\{(y_i, \boldsymbol{x}_i, d_i), i \in \mathcal{S}\}$. The initial motivation of calibration estimators is to use the known population totals of the auxiliary variables, $T_{\mathbf{x}} = \sum_{i=1}^N \boldsymbol{x}_i$, to achieve the so-called *internal consistency* by using calibrated weights w_i instead of d_i such that

$$\sum_{i\in\mathcal{S}} w_i \boldsymbol{x}_i = T_{\mathbf{x}}\,. \tag{2}$$

Equations (2) are referred to as the calibration equations or benchmark constraints. Deville and Särndal (1992) formulated the general calibration methods as a constrained minimization problem where the calibration weights w_i are obtained by minimizing a distance measure $D(\boldsymbol{d}, \boldsymbol{w})$ between $\boldsymbol{d} = (d_1, \ldots, d_n)$ and $\boldsymbol{w} = (w_1, \ldots, w_n)$ subject to constraints (2). Deville and Särndal (1992) argued intuitively that the calibration estimator $\hat{T}_{yC} = \sum_{i\in\mathcal{S}} w_i y_i$ should be more efficient than $\hat{T}_{yHT}$ since "... *weights that perform well for the auxiliary variable also should perform well for the study variable*".

The calibration estimator $\hat{T}_{yC} = \sum_{i\in\mathcal{S}} w_i y_i$ is indeed a model-assisted estimator with the same spirit of "double robustness" under a linear regression model with the mean function $E_\xi(y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, where E_ξ denotes the expectation with respect to the model ξ and $\boldsymbol{\beta}$ is the vector of regression coefficients. Under the constrained minimization procedure of Deville and Särndal (1992), the estimator $\hat{T}_{yC}$ is design-consistent regardless of any models. It is also an unbiased model-based prediction estimator under the linear regression model ξ since $E_\xi(\hat{T}_{yC} - T_y) = 0$.

The calibration estimator $\hat{T}_{yC} = \sum_{i\in\mathcal{S}} w_i y_i$ with the constraints (2) is no longer model-unbiased under any nonlinear models. Wu and Sitter (2001) considered a semiparametric model with a general mean function $E_\xi(y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i; \boldsymbol{\beta})$ and proposed a model-calibration approach through the use of the constraint

$$\sum_{i\in\mathcal{S}} w_i\mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N}\mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}})\,, \tag{3}$$

where $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ under the assumed model. There are three basic features of the model-calibration estimator $\hat{T}_{yMC} = \sum_{i\in\mathcal{S}} w_i y_i$ with the model-calibration constraint (3): (i) it is design-consistent irrespective of the model; (ii) it is an approximately model-unbiased prediction estimator under the assumed model; and (iii) the use of the estimated model parameters $\hat{\boldsymbol{\beta}}$ in (3) has no impact on the asymptotic variance of $\hat{T}_{yMC}$ under the survey design. Nonparametric models can also be used to construct model-calibration estimators (Montanari and Ranalli, 2005). The model-calibration constraint requires the "population control" $\sum_{i=1}^{N}\mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}})$ to be known, which typically requires complete auxiliary information $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$ to be available under a nonlinear model for y given $\boldsymbol{x}$.

Calibration methods can be formulated under the framework of pseudo empirical likelihood (PEL) where the distance measure $D(\boldsymbol{d},\boldsymbol{w})$ is replaced by the pseudo empirical log-likelihood function of Chen and Sitter (1999) defined as

$$\ell_{PEL}(\boldsymbol{p}) = \sum_{i\in\mathcal{S}} d_i\log(p_i)\,, \tag{4}$$

where $\boldsymbol{p} = (p_1,\ldots,p_n)$ satisfying $p_i > 0$ and the normalization constraint

$$\sum_{i\in\mathcal{S}} p_i = 1\,, \tag{5}$$

and the calibration weights are given by $\boldsymbol{w} = N\boldsymbol{p}$. The PEL approach with calibration equations has a major advantage of constructing better behaved PEL ratio confidence intervals (Wu and Rao, 2006). The PEL function $\ell_{PEL}(\boldsymbol{p})$ is defined explicitly through the design weights d_i. An alternative approach is to incorporate the survey weights through an additional constraint and use the standard empirical likelihood (EL) of Owen (1988) for the constrained maximization. Let

$$\ell_{EL}(\boldsymbol{p}) = \sum_{i\in\mathcal{S}} \log(p_i)\,. \tag{6}$$

The maximum EL estimator of the population mean μ_y is given by $\hat{\mu}_{yEL} = \sum_{i\in\mathcal{S}}\hat{p}_i y_i$, where $\hat{\boldsymbol{p}} = (\hat{p}_1,\ldots,\hat{p}_n)$ maximizes $\ell_{EL}(\boldsymbol{p})$ subject to the normalization constraint (5) and other suitably chosen constraints. The estimator $\hat{\mu}_{yEL}$ is design-consistent if the following constraint is included:

$$\sum_{i\in\mathcal{S}} p_i\pi_i = \frac{n}{N}\,. \tag{7}$$

Note that constraint (7) is a sample version of the population moment condition $N^{-1}\sum_{i=1}^{N}\pi_i = n/N$ under survey designs with fixed sample size n. Alternative versions of (7) are used by Kim (2009) and by Oguz-Alper and Berger (2016), among others. Estimator $\hat{\mu}_{yEL}$ is also approximately model-unbiased under the assumed semiparametric model if we include the model-calibration constraint

$$\sum_{i\in\mathcal{S}} p_i\mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{i=1}^{N}\mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}})\,. \tag{8}$$

The standard EL formulation through constrained maximization of $\ell_{EL}(\boldsymbol{p})$ subject to (5), (7) and (8) brings a unified framework for model-assisted estimation with probability survey samples and doubly robust estimators in other areas, as discussed in Section 3 below. The quantities on the right hand side of equations (7) and (8) are "population controls" and need to be replaced by suitable estimates depending on the setting of the problem, as discussed in Section 3.

3 Calibration approach to propensity score based estimation

In this section, we describe suitable formulations of EL-based inference for missing data problems, causal inference, and estimation with non-probability samples to construct doubly robust estimators through calibration techniques. The focus is on similarities of these problems and their connections to the calibration methods presented in Section 2.

3.1 Missing data

Let $\mathcal{S}$ be a set of n subjects with independent and identically distributed observations from an underlying infinite population. The vector of covariates $\boldsymbol{x}$ is fully observed but the study variable y is subject to missingness. Let $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. Let $\mathcal{S}_R = \{i \mid i \in \mathcal{S} \text{ and } \delta_i = 1\}$ be the set of respondents with observed y and $\mathcal{S}_M = \{i \mid i \in \mathcal{S} \text{ and } \delta_i = 0\}$ be the set of nonrespondents with missing y. The observed data can be represented by $\{(\delta_i, \delta_i y_i, \boldsymbol{x}_i), i \in \mathcal{S}\}$.

Propensity scores, defined as $\pi_i = P(\delta_i = 1 \mid y_i, \boldsymbol{x}_i)$, play an important role for missing data analysis. Under the missing at random assumption where $\pi_i = P(\delta_i = 1 \mid \boldsymbol{x}_i)$, the π_i's can be estimated based on an assumed parametric model on δ given $\boldsymbol{x}$, denoted as model q, using the observed dataset $\{(\delta_i, \boldsymbol{x}_i), i \in \mathcal{S}\}$. For instance, one can use a logistic regression model where $\pi_i = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = 1 - [1 + \exp(\boldsymbol{x}_i'\boldsymbol{\alpha})]^{-1}$ and estimate the model parameters $\boldsymbol{\alpha}$ using maximum likelihood.

The IPW estimator of the population mean μ_y is given by $\hat{\mu}_{yIPW} = n^{-1}\sum_{i\in\mathcal{S}_R} y_i/\pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$, where the estimator $\hat{\boldsymbol{\alpha}}$ is obtained by a suitable method such as maximum likelihood and the IPW estimator is consistent. With an assumed parametric form $\pi_i = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$, the propensity score model parameters $\boldsymbol{\alpha}$ can be estimated using a calibration method, and the resulting IPW estimator $\hat{\mu}_{yIPW}$ is doubly robust if the outcome regression model ξ on y given $\boldsymbol{x}$ is linear. The calibration estimator $\hat{\boldsymbol{\alpha}}$ is defined as the solution to the calibration equations

$$\sum_{i\in\mathcal{S}_R} \frac{\boldsymbol{x}_i}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} = \sum_{i\in\mathcal{S}} \boldsymbol{x}_i \,. \tag{9}$$

The double robustness property of $\hat{\mu}_{yIPW}$ is justified based on the following two arguments. First, the equation system (9) is "unbiased" with respect to the propensity score model q in the sense that $E_q\{\sum_{i\in\mathcal{S}_R} \boldsymbol{x}_i/\pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) - \sum_{i\in\mathcal{S}} \boldsymbol{x}_i \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} = \boldsymbol{0}$, and the resulting calibration estimator $\hat{\boldsymbol{\alpha}}$ is consistent for $\boldsymbol{\alpha}$. Second, the sample means $n^{-1}\sum_{i\in\mathcal{S}} \boldsymbol{x}_i$ is a valid approximation to the "population controls" of the variables $\boldsymbol{x}$ because $\mathcal{S}$ is an iid sample. In practice, the calibration estimator $\hat{\boldsymbol{\alpha}}$ obtained as the solution to (9) tends to be less stable as compared to the maximum likelihood estimator; see, for instance, Chen et al. (2020) for a discussion under the context of non-probability samples.

The EL-based methods for achieving double robustness through model-calibration is a more desirable approach and is applicable to linear or nonlinear outcome regression models with a mean function $\mu(\boldsymbol{x}, \boldsymbol{\beta})$. It involves modifications to the three crucial components: the EL function, the constraint on propensity scores, and the model-calibration constraint on the outcome regression. Let $m = \sum_{i\in\mathcal{S}} \delta_i$ be the number of units with observed y_i. Let $\boldsymbol{p} = (p_1, \ldots, p_m)$ and $\ell_{EL}(\boldsymbol{p}) = \sum_{i\in\mathcal{S}_R} \log(p_i)$. The maximum EL estimator of μ_y is computed as $\hat{\mu}_{yEL} = \sum_{i\in\mathcal{S}_R} \hat{p}_i y_i$, where $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_m)$ maximizes $\ell_{EL}(\boldsymbol{p})$ subject to the normalization constraint $\sum_{i\in\mathcal{S}_R} p_i = 1$, the constraint for the propensity score model $\pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = E_q(\delta_i \mid \boldsymbol{x}_i)$,

$$\sum_{i\in\mathcal{S}_R} p_i \, \pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \frac{m}{n} \,, \tag{10}$$

and the constraint for the outcome regression model $\mu(\boldsymbol{x}_i,\boldsymbol{\beta}) = E_\xi(y_i \mid \boldsymbol{x}_i)$,

$$\sum_{i\in\mathcal{S}_R} p_i\, \mu(\boldsymbol{x}_i,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i\in\mathcal{S}} \mu(\boldsymbol{x}_i,\boldsymbol{\beta})\,. \tag{11}$$

The m used in equation (10) may be replaced by $\sum_{i\in\mathcal{S}} \pi(\boldsymbol{x}_i,\boldsymbol{\alpha})$. For computational simplicity, the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in equations (10) and (11) can be replaced by suitable estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, and the resulting estimator $\hat{\mu}_{yEL} = \sum_{i\in\mathcal{S}_R} \hat{p}_i y_i$ remains doubly robust.

3.2 Causal inference

Estimation of the Average Treatment Effect (ATE) by comparing the responses of the treatment group to the ones for the control group is a fundamental problem in causal inference. Let $\mathcal{S}$ be the set of initial n subjects randomly selected from the target population, with measures on baseline variables $\boldsymbol{x}$ for each subject. Let T be the treatment assignment indictor with $T_i = 1$ if subject i is assigned to the treatment group and $T_i = 0$ if i is assigned to the control group. Let $\mathcal{S}_1$ and $\mathcal{S}_0$ be the set of subjects in the treatment group and in the control group, with sizes n_1 and n_0, respectively. We have $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_0$ and $n = n_1 + n_0$. Let y_1 and y_0 be, respectively, the study variable under the treatment and the control. We have a unique two-sample setting with two datasets $\{(y_{1i}, T_i = 1, \boldsymbol{x}_i), i \in \mathcal{S}_1\}$ and $\{(y_{0i}, T_i = 0, \boldsymbol{x}_i), i \in \mathcal{S}_0\}$. The ATE is the parameter of interest and is defined as $\theta = \mu_1 - \mu_0$ where μ_1 and μ_0 are, respectively, the population means of the study variable under the treatment and under the control. We assume that T_i is conditionally independent of y_{1i} and y_{0i} given $\boldsymbol{x}_i$.

It is possible to construct a doubly robust estimator for each of μ_1 and μ_0 separately, and estimate θ by $\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_0$, using a parallel procedure for the missing data problem described in Section 3.1 for obtaining $\hat{\mu}_1$ and $\hat{\mu}_0$. Huang et al. (2023) used a two-sample EL formulation and dealt with θ directly for EL-ratio confidence intervals. Let $\pi_i = P(T_i = 1 \mid \boldsymbol{x}_i)$ be the propensity score for treatment assignments, with an assumed parametric form $\pi_i = \pi(\boldsymbol{x}_i,\boldsymbol{\alpha})$; let $\mu_j(\boldsymbol{x}_i,\boldsymbol{\beta}_j) = E_\xi(y_{ji} \mid \boldsymbol{x}_i)$ be the mean functions of the response variable y_j for the two groups $j = 1, 0$ under two assumed outcome regression models. Let $\boldsymbol{p}_j = (p_{j1}, \ldots, p_{jn_j})$, $j = 1, 0$. The two-sample EL function is given by

$$\ell(\boldsymbol{p}_1,\boldsymbol{p}_0) = \sum_{i\in\mathcal{S}_1} \log(p_{1i}) + \sum_{i\in\mathcal{S}_0} \log(p_{0i})\,. \tag{12}$$

The maximum EL estimator of θ is computed as $\hat{\theta}_{EL} = \sum_{i\in\mathcal{S}_1} \hat{p}_{1i} y_{1i} - \sum_{i\in\mathcal{S}_0} \hat{p}_{0i} y_{0i}$, where $\hat{\boldsymbol{p}}_j = (\hat{p}_{j1}, \ldots, \hat{p}_{jn_j})$, $j = 1, 0$ maximize $\ell(\boldsymbol{p}_1,\boldsymbol{p}_0)$ subject to the normalization constraints $\sum_{i\in\mathcal{S}_1} p_{1i} = 1$ and $\sum_{i\in\mathcal{S}_0} p_{0i} = 1$, the model-calibration constraints induced by the propensity scores,

$$\sum_{i\in\mathcal{S}_1} p_{1i}\pi(\boldsymbol{x}_i,\boldsymbol{\alpha}) = \frac{n_1}{n}\,, \quad \sum_{i\in\mathcal{S}_0} p_{0i}[1-\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})] = \frac{n_0}{n}\,, \tag{13}$$

and the model-calibration constraints from the two outcome regression with respect to y_1 and y_0 conditional on $\boldsymbol{x}$,

$$\sum_{i\in\mathcal{S}_1} p_{1i}\mu_1(\boldsymbol{x}_i,\boldsymbol{\beta}_1) = \frac{1}{n}\sum_{i\in\mathcal{S}} \mu_1(\boldsymbol{x}_i,\boldsymbol{\beta}_1)\,, \quad \sum_{i\in\mathcal{S}_0} p_{0i}\mu_0(\boldsymbol{x}_i,\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i\in\mathcal{S}} \mu_0(\boldsymbol{x}_i,\boldsymbol{\beta}_0)\,. \tag{14}$$

The model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$ used in constraints (13) and (14) can be replaced by suitable estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$. The constraint for the parameter of interest, $\theta = \mu_1 - \mu_0$, which is part of the system for computing the EL ratio function, is given by

$$\sum_{i\in\mathcal{S}_1} p_{1i} y_{1i} - \sum_{i\in\mathcal{S}_0} p_{0i} y_{0i} = \theta\,. \tag{15}$$

The two-sample EL formulation with a single parameter of interest θ imposes some computational challenges for the constrained maximization problem. Huang et al. (2023) contain further discussions.

3.3 Non-probability samples

One of the basic features of probability samples is that the sample inclusion probabilities are known under the given sampling design. Statistical analysis with non-probability samples requires assumptions about and modelling on the unknown sample selection/inclusion process, which further requires auxiliary information on the target population. A popular setup widely used in the recent literature involves a reference probability sample containing auxiliary information from the same target population; see, for instance, Chen et al. (2020) and references therein. Let $\mathcal{S}_A$ be the set of n_A units for the non-probability sample and $\mathcal{S}_B$ be the set of n_B units for the reference probability sample, both from the same target population of size N. The two sample datasets are represented by $\{(y_i, \boldsymbol{x}_i), i \in \mathcal{S}_A\}$ and $\{(\boldsymbol{x}_i, d_i^B), i \in \mathcal{S}_B\}$, where the d_i^B are the survey weights for the reference probability sample.

Let $R_i = 1$ if $i \in \mathcal{S}_A$ and $R_i = 0$ otherwise, $i = 1, 2, \ldots, N$. Assume that R_i and y_i are independent given $\boldsymbol{x}_i$. A crucial step in analyzing the non-probability sample dataset is the modelling on the propensity scores, also called the participation probabilities by some authors, i.e., $\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i)$, $i = 1, 2, \ldots, N$. The participation probability π_i^A is defined for all units in the target population, and it is immediately clear that estimation of the π_i^A's requires information on $\boldsymbol{x}$ from the entire target population as well as an assumed model, even though the final IPW estimator of $\mu_y = N^{-1}\sum_{i=1}^N y_i$, computed as $\hat{\mu}_{yIPW} = N^{-1}\sum_{i\in\mathcal{S}_A} y_i/\hat{\pi}_i^A$, only requires the estimated π_i^A for units in $\mathcal{S}_A$.

Let the form of $\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$ be specified from a parametric model q on $(R_i \mid \boldsymbol{x}_i)$. A pseudo maximum likelihood estimator of $\boldsymbol{\alpha}$ was described in Chen et al. (2020). A calibration estimator $\hat{\boldsymbol{\alpha}}$ can also be obtained as the solution to the calibration equations

$$\sum_{i\in\mathcal{S}_A} \frac{\boldsymbol{x}_i}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} = \sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i \,. \tag{16}$$

The right hand side of (16) is an estimate for the population controls $\sum_{i=1}^N \boldsymbol{x}_i$ using the reference probability sample $\mathcal{S}_B$. Consistency of $\hat{\boldsymbol{\alpha}}$ follows from the result that $E_{qp}\{\sum_{i\in\mathcal{S}_A} \boldsymbol{x}_i/\pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) - \sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i\}$ $= \boldsymbol{0}$ under the joint randomization of the model, q, for sample participation and the probability sampling design, p, for the reference sample. The IPW estimator $\hat{\mu}_{yIPW} = N^{-1}\sum_{i\in\mathcal{S}_A} y_i/\hat{\pi}_i^A$, with the calibration estimator $\hat{\boldsymbol{\alpha}}$ used in $\hat{\pi}_i^A = \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$, is doubly robust if the outcome regression model ξ for $(y_i \mid \boldsymbol{x}_i)$ is linear since $E_{\xi p}\{\hat{\mu}_{yIPW} - \mu_y\} \doteq 0$ under the linear mean function $E_\xi(y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$.

Chen et al. (2022) presented the PEL approach to doubly robust estimation with non-probability samples with a linear or nonlinear outcome regression model $E_\xi(y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\beta})$. Doubly robust estimation can also be achieved through the standard EL. Let $\ell_{EL}(\boldsymbol{p}) = \sum_{i\in\mathcal{S}_A} \log(p_i)$, where $\boldsymbol{p} = (p_1, \ldots, p_{n_A})$. The maximum EL estimator of μ_y is computed as $\hat{\mu}_{yEL} = \sum_{i\in\mathcal{S}_A} \hat{p}_i y_i$, where $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_{n_A})$ maximizes $\ell_{EL}(\boldsymbol{p})$ subject to the normalization constraint $\sum_{i\in\mathcal{S}_A} p_i = 1$, the constraint for the participation probabilities,

$$\sum_{i\in\mathcal{S}_A} p_i \pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \frac{n_A}{N} \,, \tag{17}$$

and the model-calibration constraint for the outcome regression model,

$$\sum_{i\in\mathcal{S}_A} p_i \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{1}{N} \sum_{i\in\mathcal{S}_B} d_i^B \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) \,. \tag{18}$$

The model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ used in (17) and (18) can be replaced by suitable estimates, and the population size N can be replaced by $\hat{N} = \sum_{i \in \mathcal{S}_B} d_i^B$. The resulting estimator $\hat{\mu}_{yEL}$ is doubly robust as defined in Chen et al. (2020) where, in addition to model q for the sample participation and the model ξ for outcome regression, the probability sampling design p is part of the joint randomization framework.

4 Concluding remarks

Maximum Likelihood (ML) and Least Square (LS) are two fundamental principles for statistical inference. Calibration techniques have shown potential to be a general statistical tool, especially in the modern era for combining data from different sources as well as information from different models. The concept of model-calibration has found applications in a wide range of problems in recent years and has demonstrated certain optimality and robustness for best use of auxiliary information through an assumed model; see, for instance, Wu (2003) and Zhang et al. (2022), among others. The constrained minimization of a distance measure as described in Deville and Särndal (1992) provides a natural connection to the constrained maximization of the empirical likelihood function, which has been used in many areas of statistics. Calibration techniques for model-based prediction and doubly robust estimation have been shown to be useful for problems described in this short article and their potential for other problems and extensions to the so-called multiply robust estimation (Han and Wang, 2013) deserves further exploration.

References

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976) Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, **63**, 615–620.

Chen, J., and Sitter, R. R. (1999) A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, **9**, 385–406.

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.

Chen, Y., Li, P., Rao, J. N. K., and Wu, C. (2022) Pseudo empirical likelihood methods for non-probability survey samples. *The Canadian Journal of Statistics*, **50**, 1166–1185.

Deming, W. E., and Stephan, F. F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, **11**, 427–444.

Deville, J., and Särndal, C. -E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382.

Han, P., and Wang, L. (2013) Estimation with missing data: beyond double robustness. *Biometrika*, **100**, 417–430.

Horvitz, D. G., and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Huang, J., Wu, C., and Zeng, L. (2023) Empirical likelihood methods for causal inference. Working paper, Department of Statistics and Actuarial Science, University of Waterloo.

Kim, J. K. (2009) Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, **19**, 145–157.

Montanari, G. E., and Ranalli, M. G. (2005) Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, **100**, 1429–1442.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–174.

Oguz-Alper, M. and Berger, Y. G. (2016) Modelling complex survey data with population level information: An empirical likelihood approach. *Biometrika*, **103**, 447–459.

Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Wu, C. (2003) Optimal calibration estimators in survey sampling. *Biometrika*, **90**, 937–951.

Wu, C., and Rao, J. N. K. (2006) Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, **34**, 359–375.

Wu, C., and Sitter, R. R. (2001) A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96**, 185–193.

Zhang, S., Han, P., and Wu, C. (2022) Calibration techniques encompassing survey sampling, missing data analysis and causal inference. *International Statistical Review*, accepted for publication.