*Journal of Survey Statistics and Methodology (2022) **10**, 1319–1345*

PREDICTION OF FINITE POPULATION PROPORTION WHEN RESPONSES ARE MISCLASSIFIED

SUMANTA ADHYA
SURUPA ROY*
TATHAGATA BANERJEE

We propose a model-based predictive estimator of the finite population proportion of a misclassified binary response, when information on the auxiliary variable(s) is available for all units in the population. Asymptotic properties of the misclassification-adjusted predictive estimator are also explored. We propose a computationally efficient bootstrap variance estimator that exhibits better performance compared to usual analytical variance estimator. The performance of the proposed estimator is compared with other commonly used design-based estimators through extensive simulation studies. The results are supplemented by an empirical study based on literacy data.

KEYWORDS: Hybrid bootstrap variance estimation; Misclassification; Model-based estimators; Pseudo-likelihood; Resampling; Validation data.

1. INTRODUCTION

In finite population surveys, estimation of the population proportion is an important research problem. For instance, in epidemiological studies, proportion of coalminers suffering from wheeze (Ekholm and Palmgren 1982); in wildlife surveys, proportion of species of a specific kind (Thompson 2002); and in

SUMANTA ADHYA is an Assistant Professor in the Department of Statistics, West Bengal State University, Barasat, North 24-Parganas, Kolkata 700126, India. SURUPA ROY is an Associate Professor in the Department of Statistics, St Xavier's College (Autonomous), 30 Park Street, Kolkata 700016, India. TATHAGATA BANERJEE is Professor and Dean (Faculty) in the Department of Production and Quantitative Methods, Indian Institute of Management, Ahmedabad, Vastrapur, Ahmedabad 380015, India.
The authors are thankful to the referees for their useful comments, which enriched the current work.
*Address correspondence to Surupa Roy, Department of Statistics, St Xavier's College (Autonomous), 30 Park Street, Kolkata 700016, India; E-mail: surupastat@gmail.com.

doi: 10.1093/jssam/smaa027 Advance access publication 25 December 2020

health research studies, proportion of persons developing cardiovascular disease (Stefanski and Carroll 1985) are all potential examples of proportion estimation. Auxiliary information if available can be used to increase the precision of estimation via a model-based or design-based approach. For the former, we refer to Valliant, Dorfman, and Royall (2000). Generalized regression estimators (Cassel, Särndal, and Wretman 1976; Särndal 1980), calibration estimators (Deville and Särndal 1992; Wu and Sitter 2001), and estimators based on empirical likelihood (Chen and Qin 1993; Chen and Sitter 1999; Zhong and Rao 2000) are prominent examples under design-based approach. Adhya, Banerjee, and Chattopadhyay (2011) have shown that, under different sampling schemes, the model-based estimator of the population proportion of a polychotomous response variable has performed better than the commonly used design-based estimators.

Motivated by the findings of Adhya et al. (2011), in this article, we develop a model-based predictive estimator of the population proportion of a binary response, following the prediction approach of Royall (1970, 1976). Typically, here, the actual values of the binary response variable for each unit in the finite population are treated as realizations of random variables, which are assumed to follow a joint probability law specified by a super-population model. The prediction approach combines the information from the sampled units and the predicted responses from the non-sampled units, which are estimated via the super-population model.

The standard inferential problems in finite population surveys assume that the responses are correctly observed, which however is often not true. Due to cost and convenience, frequently such studies employ inaccurate measures of responses. Although for binary data the problem of misclassification has been extensively researched (Gustafson 2003; Buonaccorsi 2010), its application in survey sampling is still limited. It is well established that, under misclassification, sample proportion is a biased estimate of the parameter (Bross 1954). However, a perennial problem with the models corrected for misclassification is that, they are over parametrized. The double sampling approach is a widely used technique for resolving the identifiability problem. In this method, two samples are used in conjunction: the original sample of cursory observations and a validation sample where the true responses are evaluated using a costly gold standard method. The validation sample could be an external sample from the same finite population or it could be a subsample of the original sample. Getting such gold standard is not uncommon in practice. For instance, in a life span study among atom bomb survivors, it was observed that the cause of death was misspecifed in death certificates. Sposto, Preston, Shimizu, and Mabuchi (1992) estimated the misclassification probabilities by using a validation data set obtained from a subset of deaths in the cohort for which autopsies were carried out. Tenenbein (1970) is a classic example of the use of double sampling to correct for misclassification in the binomial model. Hochberg (1977) and Chen (1979) adopted double sampling approach to correct for

misclassification in categorical models. Double sampling is also used for the estimation of complex models like misclassified multivariate ordinal response (Poon and Wang 2010) and misclassified correlated binary response (Chen, Yi, and Wu 2011). Recently Sang, Lopiano, Abreu, Lamas, Arroway, et al. (2017) have proposed a design-based estimator of the population proportion based on the three-phase survey, which adjusts for misclassification. However, the authors did not adopt any model-based approach.

The article focuses on the estimation of finite population proportions using prediction approach when information on the auxiliary variable is available for all units in the population. Unlike the previous work in this area, the current work considers an error-prone binary response. Using a classification error model, we develop a misclassification-adjusted predictive estimator of the population proportion. The article also looks into the asymptotic properties of the proposed estimator and constructs a novel hybrid bootstrap variance estimator of the prediction error. Extensive model-based and design-based simulations support the efficacy of the proposed estimator. The findings are validated by analyzing a survey data on literacy.

2. PREDICTIVE ESTIMATOR

Consider a finite population comprising N units and let $Y_i (i = 1, .., N)$ denote the binary response variable corresponding to these N units. The vector $\mathbf{x}_i = (x_{i1}, x_{i2}, .., x_{iq})^T$ gives the values of the q auxiliary variables that are assumed to be known for the entire population and let $\mathbf{X}_U = (\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N)$. Our interest lies in developing a model-based predictive estimator of the population quantity,

$$P = N^{-1} \sum_{i=1}^{N} Y_i. \tag{1}$$

For predicting P given in (1), we first select a sample of n units from N units by a suitable sampling scheme, which encompasses probability and non-probability sampling as long as those are non-informative. We refer to Sugden and Smith (1984) for the former while the latter has been considered by Smith (1983) and Elliott and Valliant (2017) among others. For these n units, we observe the study variable $Y_i (i \in \mathrm{S})$, where S is the set of n indices of the sampled units. Let us define $\boldsymbol{\delta}_U = (\delta_1, \delta_2, .., \delta_N)^T$, such that $\delta_i = 1$ if $i \in S$, and 0, otherwise. Then, due to ignorability assumption, following similar argument as in Theorem 1.3.1 of Fuller (2009), the conditional distribution of $Y_i|\mathbf{X}_U$ is same as that of $Y_i|\mathbf{X}_U, \boldsymbol{\delta}_U$. Now, conditional on $\mathbf{X}_U$, $Y_1, Y_2, .., Y_N$ are independent with probability distribution given by,

$$P(Y_i = 1|\mathbf{x}_i) = \pi(\mathbf{x}_i;\beta) = \frac{\exp(g(\mathbf{x}_i;\beta))}{1+\exp(g(\mathbf{x}_i;\beta))}, \tag{2}$$

where the function $g(.)$ is of known form and $\beta = (\beta_1^T, \beta_2^T, .., \beta_q^T)^T$ where β_h^T is the vector of unknown regression coefficients associated with the auxiliary variable $x_{ih}(h = 1, 2, .., q)$.

However, in observational studies, where data are collected on a large number of individuals, the binary responses are typically not correctly observed. Let Y_i^{obs} $(i = 1, .., N)$ denote the manifest binary response corresponding to the ith unit in the population. We assume a simple probability model linking the observed response with the true response as follows:

$$P(Y_i^{obs} = 1|Y_i = 0) = \epsilon_0, \tag{3}$$

$$P(Y_i^{obs} = 0|Y_i = 1) = \epsilon_1, \tag{4}$$

where ϵ_0 and ϵ_1 are unknown misclassification probabilities, which may or may not depend upon the auxiliary variable(s). Now, incorporating the misclassification probabilities, the conditional probability of the manifest response to be positive is given by,

$$\tilde{\pi}(\mathbf{x}_i;\beta,\epsilon) = P(Y_i^{obs} = 1|\mathbf{x}_i) = \epsilon_0 + (1-\epsilon_0-\epsilon_1)\pi(\mathbf{x}_i;\beta), \tag{5}$$

where $\pi(\mathbf{x}_i;\beta)$ is given in (2). It is worthwhile to note that if $\epsilon_0 + \epsilon_1 = 1$, (5) becomes independent of β and then the manifest response does not contain any information about the regression parameters. Moreover, for all practical purposes, it is reasonable to assume that the classification error of either kind is lesser than half (Wang and Gustafson 2014).

In case the true responses are observable, the predictive estimator of P in (1) could be obtained as,

$$\widehat{P}_{true} = N^{-1}\{\sum_{i\in S} Y_i + \sum_{i\in\overline{S}} E(Y_i|\mathbf{x}_i;\widehat{\beta}_{true})\}, \tag{6}$$

where $\overline{S}$ is the set of $(N-n)$ non-sampled units and $\widehat{\beta}_{true}$ is the maximum likelihood estimator (MLE) of β obtained on the basis of true $Y_i(i \in \mathrm{S})$, using the model given in (2). Note that $\widehat{P}_{true}$ in (6) corresponds to the predictive estimator of Valliant et al. (2000) who considered logit and complementary log–log link functions with linear predictors.

In our case, the true $Y_i(i \in \mathrm{S})$ are not observable and so the predictive estimator in (6) cannot be constructed. A simple solution is to derive the MLE based on sample manifest responses, Y_i^{obs} $(i \in \mathrm{S})$. For all future references, we shall denote this naive estimator of β by $\widehat{\beta}_{naive}$ and the naive predictive estimator so constructed will be denoted by $\widehat{P}_{naive}$, where,

$$\widehat{P}_{naive} = N^{-1}\{\sum_{i\in S} Y_i^{obs} + \sum_{i\in\overline{S}} E(Y_i|\mathbf{x}_i;\widehat{\beta}_{naive})\}. \tag{7}$$

It is evident that the estimator in (7) is biased since it is constructed by ignoring the misclassification in the response. We thus propose a misclassification-adjusted predictive estimator of P as,

$$\widehat{P}_{adjusted} = N^{-1}\{\sum_{i\in S} E(Y_i|Y_i^{obs},\mathbf{x}_i;\widehat{\beta},\widehat{\epsilon}) + \sum_{i\in\overline{S}} E(Y_i|\mathbf{x}_i;\widehat{\beta})\}, \tag{8}$$

where $\widehat{\beta}$ and $\widehat{\epsilon} = (\widehat{\epsilon}_0,\widehat{\epsilon}_1)^T$ are, respectively, the maximum likelihood estimates of β and $\epsilon = (\epsilon_0,\epsilon_1)^T$, which are obtained on the basis of the sample manifest responses using the misclassification adjusted model given in (5). At this stage, it needs mentioning that, if all observations lie in the central part of the logit function, then simultaneous estimation of β and ϵ from model (5) clearly falls through, since in that case the logit can be well approximated by a linear function (Cox and Snell 1989) and hence the estimates of the misclassification probabilities get totally confounded with the estimates of the regression parameters.

To resolve the identifiability issue, a double sampling approach has been adopted. In this method, along with the main sample, an internal validation sample of relatively smaller size is used where the binary response is observed without any error. Here, the subsample is drawn from S by simple random sampling without replacement and is denoted by S_v. The non-validation set is denoted by S_{nv} where $S_v \cup S_{nv} = S$. Throughout we shall use $f(u;\theta)$ as a generic notation for a density function of the random variable U characterized by the parameter θ. Based on the data $(Y_i, i\in S_v; Y_i^{obs}, i\in S; \mathbf{x}_i, i\in S)$, the log-likelihood function can be partitioned as,

$$\begin{aligned} l(\beta,\epsilon) &= \sum_{i\in S_v}\log(f(y_i,y_i^{obs}|\mathbf{x}_i;\beta,\epsilon)) + \sum_{i\in S_{nv}}\log(f(y_i^{obs}|\mathbf{x}_i;\beta,\epsilon)) \\ &= \sum_{i\in S_v}\log(f(y_i^{obs}|y_i;\epsilon)) + \sum_{i\in S_v}\log(f(y_i|\mathbf{x}_i;\beta)) + \sum_{i\in S_{nv}}\log(f(y_i^{obs}|\mathbf{x}_i;\beta,\epsilon)) \\ &= l_1(\epsilon) + l_2(\beta) + l_3(\beta,\epsilon), \end{aligned} \tag{9}$$

where $l_1(\epsilon)$ is obtained from (3) and (4) while $l_2(\beta)$ and $l_3(\beta,\epsilon)$ are obtained from (2) and (5), respectively. Maximizing the likelihood function in (9) could be mathematically involved. A simple solution is to use a two-step estimation procedure where in the first step the nuisance parameter ϵ is estimated using the validation data only, that is, from $l_1(\epsilon)$. In the second step, the estimate of β is obtained from $l_2(\beta)$ and $l_3(\beta,\epsilon)$, after plugging in the estimate of ϵ from the first step. This method is popularly known as pseudo-maximum likelihood method (Gong and Samaniego 1981). Thus, maximizing $l_1(\epsilon)$ yields,

$$\widehat{\epsilon_0} = \left(\sum_{i\in S_v}(1-Y_i)\right)^{-1} \sum_{i\in S_v} Y_i^{obs}(1-Y_i), \tag{10}$$

$$\widehat{\epsilon_1} = \left(\sum_{i\in S_v} Y_i\right)^{-1} \sum_{i\in S_v} Y_i(1-Y_i^{obs}). \tag{11}$$

Next, the pseudo-maximum likelihood estimate (PMLE) of β denoted by $\widehat{\beta}$ is obtained by solving the score equation,

$$\sum_{i\in S_v}\frac{\partial l_2(\beta)}{\partial\beta} + \sum_{i\in S_{nv}}\frac{\partial l_3(\beta,\widehat{\epsilon})}{\partial\beta} = 0, \tag{12}$$

where $l_2(.)$ and $l_3(.)$ are given in (9) and $\widehat{\epsilon} = (\widehat{\epsilon_0},\widehat{\epsilon_1})^T$ is given in (10) and (11). The predictive estimator in (8) is modified to get the corrected estimator of P as,

$$\widehat{P}_C = N^{-1}\{\sum_{i\in S_v} Y_i + \sum_{i\in S_{nv}} E(Y_i|Y_i^{obs},\mathbf{x}_i;\widehat{\beta},\widehat{\epsilon}) + \sum_{i\in\overline{S}} E(Y_i|\mathbf{x}_i;\widehat{\beta})\}, \tag{13}$$

where

$$E(Y_i|Y_i^{obs},\mathbf{x}_i;\widehat{\beta},\widehat{\epsilon}) = [\tilde{\pi}(\mathbf{x}_i;\widehat{\beta},\widehat{\epsilon})^{-1}(1-\widehat{\epsilon_1})\pi(\mathbf{x}_i;\widehat{\beta})]^{Y_i^{obs}}[(1-\tilde{\pi}(\mathbf{x}_i;\widehat{\beta},\widehat{\epsilon}))^{-1}\widehat{\epsilon_1}\pi(\mathbf{x}_i;\widehat{\beta})]^{1-Y_i^{obs}}, \tag{14}$$

$$E(Y_i|\mathbf{x}_i;\widehat{\beta}) = \pi(\mathbf{x}_i;\widehat{\beta}). \tag{15}$$

Moreover, $\pi(\mathbf{x}_i;\widehat{\beta})$ and $\tilde{\pi}(\mathbf{x}_i;\widehat{\beta},\widehat{\epsilon})$ appearing in (14) and (15) are obtained from (2) and (5), respectively, by replacing β with $\widehat{\beta}$ and ϵ with $\widehat{\epsilon}$. The properties of $\widehat{P}_C$ given in (13) is studied in section 3.

3. ASYMPTOTIC PROPERTIES

In this section, we first introduce the basic building blocks required for proving the results on asymptotic properties of $(\widehat{\beta},\widehat{\epsilon})^T$ and the proposed predictive estimator $\widehat{P}_C$. For completeness, we give a brief description of all the notations and quantities that would be used. Let $\epsilon^0 = (\epsilon_0^0,\epsilon_1^0)^T$ and $\beta^0 = (\beta_1^{0T},\ldots,\beta_q^{0T})^T$ denote the true values of ϵ and β, respectively. Further suppose Θ_ϵ and Θ_β denote, respectively, the parametric spaces of ϵ and β. Now, for a smooth generic function, $\psi()$, $\psi_\theta(.)$ and $\psi_{\theta\theta}(.)$ refer to the first and second order partial derivatives with respect to θ. We shall refer to a subset A (with cardinality n_A) as S

or S_v or S_{nv} with cardinalities denoted by n, n_v, and $\overline{n-n_v}$. In what follows, we shall assume that the conditions C0–C6 are satisfied.

C0: For $n \to \infty$, $n_v \to \infty$, $f = \frac{n}{N} \to \rho \in (0,1)$ and $f_v = \frac{n_v}{n} \to \rho_v \in (0,1)$.

C1: The auxiliary variable $\mathbf{x}_i, i = 1, 2, \ldots$ are independently and identically distributed with an unspecified density function $h(.)$. For any subset A, the asymptotic density, $h_A()$, for a sequence $(\mathbf{x}_1, \mathbf{x}_2, \ldots)$ is given by, $n_A^{-1} \sum_A I(\mathbf{x}_i \leq l) \to \int_{-\infty}^{l} h_A(u)du$. (Chambers, Dorfman, and Hall 1992).

C2: $\Theta_\epsilon = \left(0, \frac{1}{2}\right)^2 \subset R^2$ and Θ_β is a compact subset of R^p, where $p = \sum_{h=1}^{q} p_h$, and where p_h is the dimension of β_h.

C3: For any β and ϵ, $\pi(\mathbf{x};\beta)$ and $\tilde{\pi}(\mathbf{x};\beta,\epsilon)$ are bounded away from 0 and 1 $\forall$ $\mathbf{x}$.

C4: $g(\mathbf{x};\beta)$ is twice continuously differentiable with respect to β $\forall \mathbf{x}$.

C5: For any subset A, $E_A|g(\mathbf{x};\beta)| < \infty$ $\forall\beta$, $E_A[g_\beta(\mathbf{x};\beta)g_\beta^T(\mathbf{x};\beta)] > 0$ $\forall\beta$, $E_A|lnh(\mathbf{x})| < \infty$ and $E_A\left[\frac{\pi(\mathbf{x};\beta)(1-\pi(\mathbf{x};\beta))}{\tilde{\pi}(\mathbf{x};\beta,\epsilon)(1-\tilde{\pi}(\mathbf{x};\beta,\epsilon))}\right] < \infty$ $\forall(\beta,\epsilon)$, where $E_A(.)$ is the expectation with respect to the asymptotic density $h_A(.)$.

C6: Differentiation with respect to β and ϵ under the integrals over $\mathbf{x}$ are valid.

Theorem 1. Under conditions C0–C6, as $n_v \to \infty$,

$$\widehat{\epsilon} \xrightarrow{P} \epsilon^0 \text{ and } \widehat{\beta} \xrightarrow{P} \beta^0.$$

Proof. See appendix A.1. □

Theorem 2. Under conditions C0–C6, as $n_v \to \infty$,

$$\sqrt{n}(\widehat{\beta} - \beta^0) \xrightarrow{L} N_q(0,\ I^{-1}(\beta^0,\epsilon^0)G(\beta^0,\epsilon^0)W(\beta^0,\epsilon^0)G^T(\beta^0,\epsilon^0)I^{-1}(\beta^0,\epsilon^0)).$$

Proof. The detailed expressions of $I(\beta^0,\epsilon^0)$, $G(\beta^0,\epsilon^0)$ and $W(\beta^0,\epsilon^0)$ along with the proof is relegated to appendix A.2. □

To establish the asymptotic properties of model-based predictive estimator $\widehat{P}_C$, we define

$$\pi_{ij}(\mathbf{x};\beta,\epsilon) = E(Y = i|Y^{obs} = j, \mathbf{x};\beta,\epsilon), \quad i,j = 0,1. \tag{16}$$

Based on $\pi_{ij}(\mathbf{x};\beta,\epsilon)$ in (16), $\pi(\mathbf{x};\beta)$ in (2), and $\tilde{\pi}(\mathbf{x};\beta,\epsilon)$ in (5), we further define the following:

$$\pi_{ij,\beta}(\mathbf{x};\beta,\epsilon)=\frac{\partial\pi_{ij}(\mathbf{x};\beta,\epsilon)}{\partial\beta},\quad \pi_{ij,\epsilon}(\mathbf{x};\beta,\epsilon)=\frac{\partial\pi_{ij}(\mathbf{x};\beta,\epsilon)}{\partial\epsilon},$$
$$\pi_{\beta}(\mathbf{x};\beta)=\frac{\partial\pi(\mathbf{x};\beta)}{\partial\beta},$$

$$\Delta(\beta,\epsilon)=Diag[\epsilon_0(1-\epsilon_0)(1-a(\beta))^{-1},\ \epsilon_1(1-\epsilon_1)a^{-1}(\beta)],\ \text{where}\ a(\beta)=E_{S_v}[\pi(\mathbf{x};\beta)],$$

$$K_1(\beta,\epsilon)=(n-n_v)^{-1}\sum_{i\in S_{nv}}\pi_{11,\epsilon}(x_i;\beta,\epsilon)\tilde{\pi}(x_i;\beta,\epsilon),$$

$$K_2(\beta,\epsilon)=(n-n_v)^{-1}\sum_{i\in S_{nv}}\pi_{10,\epsilon}(\mathbf{x}_i;\beta,\epsilon)(1-\tilde{\pi}(\mathbf{x}_i;\beta,\epsilon)),$$

$$K_3(\beta,\epsilon)=(n-n_v)^{-1}\sum_{i\in S_{nv}}\pi_{11,\beta}(\mathbf{x}_i;\beta,\epsilon)\tilde{\pi}(\mathbf{x}_i;\beta,\epsilon),$$

$$K_4(\beta,\epsilon)=(n-n_v)^{-1}\sum_{i\in S_{nv}}\pi_{10,\beta}(\mathbf{x}_i;\beta,\epsilon)(1-\tilde{\pi}(\mathbf{x}_i;\beta,\epsilon)),$$

$$K_5(\beta,\epsilon)=(N-n)^{-1}\sum_{i\in\overline{S}}\pi_{\beta}(\mathbf{x}_i;\beta),$$

$$K_6(\beta,\epsilon)=(N-n)^{-1}\sum_{i\in\overline{S}}\pi(\mathbf{x}_i;\beta)(1-\pi(\mathbf{x}_i;\beta)),$$

$$K_7(\beta,\epsilon)=(n-n_v)^{-1}\sum_{i\in S_{nv}}\pi_{11}(x_i;\beta,\epsilon)(1-\pi_{11}(x_i;\beta,\epsilon))\tilde{\pi}(x_i;\beta,\epsilon)+$$
$$(n-n_v)^{-1}\sum_{i\in S_{nv}}\pi_{10}(x_i;\beta,\epsilon)(1-\pi_{10}(x_i;\beta,\epsilon))(1-\tilde{\pi}(x_i;\beta,\epsilon)).$$

In what follows we shall further assume that the following condition is also satisfied.

C7: $I_{\beta\beta}(\beta,\epsilon) = lim_{n_v\to\infty} -n^{-1}\frac{\partial^2 l(\beta,\epsilon)}{\partial\beta\partial\beta^T}$, $I_{\beta\epsilon}(\beta,\epsilon) = lim_{n_v\to\infty} -n^{-1}\frac{\partial^2 l(\beta,\epsilon)}{\partial\beta\partial\epsilon^T}$, and $K_i(\beta,\epsilon) = lim_{n_v\to\infty} k_i(\beta,\epsilon)$, $i = 1,2,..,7$ exist and are finite for all values of (β,ϵ), where limits are evaluated with respect to the asymptotic density of **x** as mentioned in C1.

Finally, we define $M_1(\beta,\epsilon)$–$M_5(\beta,\epsilon)$ as follows:

$$M_1(\beta,\epsilon) = \Delta(\beta,\epsilon)I^T_{\beta\epsilon}(\beta,\epsilon)I^{-1}_{\beta\beta}(\beta,\epsilon),$$
$$M_2(\beta,\epsilon) = I^{-1}_{\beta\beta}(\beta,\epsilon)I_{\beta\epsilon}(\beta,\epsilon)M_1(\beta,\epsilon),$$

$$M_3(\beta,\epsilon) = I^{-1}_{\beta\beta}(\beta,\epsilon) + \rho_v^{-1}M_2(\beta,\epsilon), \quad M_4(\beta,\epsilon) = K_1(\beta,\epsilon) + K_2(\beta,\epsilon),$$

$$M_5(\beta,\epsilon) = K_3(\beta,\epsilon) + K_4(\beta,\epsilon).$$

Theorem 3. Under conditions C0–C7,

$$E(\widehat{P}_C - P) = O(n_v^{-1}).$$

Proof. See appendix A.3. □

Theorem 4. $Var(\widehat{P}_C - P) = V(\beta^0,\epsilon^0) + o(n_v^{-1})$, where,

$$V(\beta^0,\epsilon^0) = n^{-1}\rho^2(1-\rho_v)^2[\rho_v^{-1}((M_4^0)^T\Delta^0M_4^0 + (K_2^0)^TM_1^0K_3^0 + (M_4^0)^TM_1^0M_5^0 +$$
$$(K_1^0)^TM_1^0K_4^0) + (M_5^0)^TM_3^0M_5^0] + n^{-1}(1-\rho)^2[(1-\rho)^{-1}K_6^0 + (K_5^0)^TM_3^0M_5^0]$$
$$+n^{-1}(1-\rho)^2(1-\rho_v)K_7^0 + 2n^{-1}\rho(1-\rho)(1-\rho_v)[\rho_v^{-1}(M_4^0)^TM_1^0K_5^0 + (M_5^0)^TM_3^0K_5^0],$$

with $\Delta^0 = \Delta(\beta^0,\epsilon^0)$, $K_i^0 = K_i(\beta^0,\epsilon^0)$, $i = 1,2,\ldots,7$ *and* $M_i^0 = M_i(\beta^0,\epsilon^0)$, $i = 1,2,\ldots,5$.

Proof. We skip the detailed proof. Using Taylor series expansion of the terms involving $\widehat{\beta}$ and $\widehat{\epsilon}$, neglecting $o(n_v^{-1})$ terms and after some routine algebra, the variance expression follows. □

Corollary 1 As $n_v \to \infty$, $(\widehat{P}_C - P) \to^L N(0, V(\beta^0,\epsilon^0))$.

Proof. The proof follows from theorem 4. □

Corollary 2 A consistent estimator of the asymptotic variance of $(\widehat{P}_C - P)$ *is given by*,

$$\widehat{V}_A = \widehat{V}(\widehat{\beta},\widehat{\epsilon}) = n^{-1}f^2(1-f_v)^2[f_v^{-1}(\widehat{M}_4^T\widehat{\Delta}\widehat{M}_4 + \widehat{K}_2^T\widehat{M}_1\widehat{K}_3 + \widehat{M}_4^T\widehat{M}_1\widehat{M}_5 + \widehat{K}_1^T\widehat{M}_1\widehat{K}_4) + \widehat{M}_5^T\widehat{M}_3\widehat{M}_5] + n^{-1}(1-f)^2[(1-f)^{-1}\widehat{K}_6 + \widehat{K}_5^T\widehat{M}_3\widehat{M}_5] + n^{-1}(1-f)^2(1-f_v)\widehat{K}_7 + 2n^{-1}f(1-f)(1-f_v)[f_v^{-1}\widehat{M}_4^T\widehat{M}_1\widehat{K}_5 + \widehat{M}_5\widehat{M}_3\widehat{K}_5], \tag{17}$$

where $\widehat{I}_{\beta\beta} = -n^{-1}\frac{\partial^2 l(\beta,\epsilon)}{\partial\beta\partial\beta^T}|_{\widehat{\beta},\widehat{\epsilon}}$; $\widehat{I}_{\beta\epsilon} = -n^{-1}\frac{\partial^2 l(\beta,\epsilon)}{\partial\beta\partial\epsilon^T}|_{\widehat{\beta},\widehat{\epsilon}}$; $\widehat{\Delta} = \Delta(\widehat{\beta},\widehat{\epsilon})$; $\widehat{K}_i = k_i(\widehat{\beta},\widehat{\epsilon})\,(i=1,2,..,7)$; $\widehat{M}_1 = \widehat{\Delta}\widehat{I}_{\beta\epsilon}^T\widehat{I}_{\beta\beta}^{-1}$; $\widehat{M}_2 = \widehat{I}_{\beta\beta}^{-1}\widehat{I}_{\beta\epsilon}\widehat{\Delta}\widehat{I}_{\beta\epsilon}^T\widehat{I}_{\beta\beta}^{-1}$; $\widehat{M}_3 = \widehat{I}_{\beta\beta}^{-1} + f_v^{-1}\widehat{M}_2$; $\widehat{M}_4 = \widehat{K}_1 + \widehat{K}_2$; and $\widehat{M}_5 = \widehat{K}_3 + \widehat{K}_4$.

Proof. All the terms appearing in $\widehat{V}_A$ are continuous in β and ϵ. Since $\widehat{\beta}$ and $\widehat{\epsilon}$ are consistent (vide theorem 1), the consistency of the analytical variance estimator also follows. □

Corollary 3 As $n_v \to \infty$, $\widehat{V}_A^{-1/2}(\widehat{P}_C - P) \to^L N(0,1)$.

Proof. The proof follows from corollary 1, corollary 2, and Slutsky's theorem. □

4. BOOTSTRAP VARIANCE ESTIMATION

We propose a computationally efficient, resampling-based hybrid bootstrap variance estimation (Adhya, Banerjee, and Chattopadhyay 2012) of the prediction error $\widehat{P}_C - P$. To this end, we first decompose the prediction error $\widehat{P}_C - P$ as,

$$\widehat{P}_C - P = T_1 + T_2, \tag{18}$$

where T_1 and T_2 are given by,

$$T_1 = N^{-1}\sum_{i\in S_{nv}}\{E(Y_i|Y_i^{obs},\mathbf{x}_i;\widehat{\beta},\widehat{\epsilon}) - E(Y_i|Y_i^{obs},\mathbf{x}_i;\beta,\epsilon)\} + N^{-1}\sum_{i\in\overline{S}}\{E(Y_i|\mathbf{x}_i;\widehat{\beta}) - E(Y_i|\mathbf{x}_i;\beta)\} = T_{11} + T_{12}. \quad (say), \tag{19}$$

$$T_2 = N^{-1}\sum_{i\in S_{nv}}\{E(Y_i|Y_i^{obs},\mathbf{x}_i;\beta,\epsilon) - Y_i\} + N^{-1}\sum_{i\in\overline{S}}\{E(Y_i|\mathbf{x}_i;\beta) - Y_i\} = T_{21} + T_{22}. \quad (say) \tag{20}$$

Theorem 5. Assuming, $lim_{n_v \to \infty}(n - n_v^{-1}) \sum_{i \in S_{nv}} ||g_\beta(\mathbf{x}_i; \beta)|| \infty \ \forall \ \beta$ and under conditions C0–C7,

$$Var(\widehat{P}_C - P) = Var(T_1) + Var(T_2) + o(n_v^{-1}). \tag{21}$$

Proof. Note that due to conditional independence of S_v and S_{nv} given $\mathbf{x}$s, $Cov(T_{11}, T_{22}) = Cov(T_{12}, T_{22}) = 0$. Hence,

$$Cov(T_1, T_2) = Cov(T_{11}, T_{21}) + Cov(T_{12}, T_{21}). \tag{22}$$

The proof follows from the fact that $Cov(T_{11}, T_{21}) = o(n_v^{-1})$ and $Cov(T_{12}, T_{21}) = o(n_v^{-1})$. Details are relegated to the online Supplementary Material. □

Now a consistent estimate of $Var(T_2)$ can be obtained as,

$$\begin{aligned}
\widehat{Var}_A(T_2) &= N^{-2} \textstyle\sum_{i \in S_{nv}} \pi_{11}(\mathbf{x}_i; \widehat{\beta}, \widehat{\epsilon})\{1 - \pi_{11}(\mathbf{x}_i; \widehat{\beta}, \widehat{\epsilon})\}\tilde{\pi}(\mathbf{x}_i; \widehat{\beta}, \widehat{\epsilon}) \\
&+ N^{-2} \textstyle\sum_{i \in S_{nv}} \pi_{10}(\mathbf{x}_i; \widehat{\beta}, \widehat{\epsilon})\{1 - \pi_{10}(\mathbf{x}_i; \widehat{\beta}, \widehat{\epsilon})\}\{1 - \tilde{\pi}(\mathbf{x}_i; \widehat{\beta}, \widehat{\epsilon})\} \\
&+ N^{-2} \textstyle\sum_{i \in \overline{S}} \pi(\mathbf{x}_i; \widehat{\beta})\{1 - \pi(\mathbf{x}_i; \widehat{\beta})\},
\end{aligned} \tag{23}$$

where $\pi_{ij}(\mathbf{x}_i; \widehat{\beta}, \widehat{\epsilon})$, $\tilde{\pi}(\mathbf{x}_i; \widehat{\beta}, \widehat{\epsilon})$, and $\pi(\mathbf{x}_i; \widehat{\beta})$ are obtained from (16), (5), and (2) respectively after replacing (β, ϵ) by $(\widehat{\beta}, \widehat{\epsilon})$. However, unlike $Var(T_2)$, an exact expression for $Var(T_1)$ cannot be obtained. In such situations, resampling-based variance estimation is a popular choice in survey literature compared to analytical variance estimation. The bootstrap method used here for the estimation of $Var(T_1)$ is outlined below.

Step 4.1. Generate two independent paired bootstrap samples from $(Y_i, Y_i^{obs}, \mathbf{x}_i ; i \in S_v)$ and $(Y_i^{obs}, \mathbf{x}_i; i \in S_{nv})$ by simple random sampling with replacement (SRSWR). Let the bootstrap samples be denoted by $(Y_i^*, Y_i^{obs*}, \mathbf{x}_i^*; i \in S_v)$ and $(Y_i^{obs*}, \mathbf{x}_i^*; i \in S_{nv})$, respectively.

Step 4.2. Based on the bootstrap data $\{(Y_i^*, Y_i^{obs*}, \mathbf{x}_i^*; i \in S_v), (Y_i^{obs^*}, \mathbf{x}_i^*; i \in S_{nv})\}$, the bootstrap estimates of the parameters are obtained by maximizing the log-likelihood function given in (9). We denote the bootstrap analogs of $\widehat{\beta}$ and $\widehat{\epsilon}$ by $\widehat{\beta}^*$ and $\widehat{\epsilon^*} = (\widehat{\epsilon}_0^*, \widehat{\epsilon}_1^*)^T$, respectively. To avoid the iterative procedure for obtaining the PMLE of β, we use a one-step approximation of $\widehat{\beta}^*$ (Claeskens, Aerts, and Molenberghs 2003), which is given in the online Supplementary Material.

Step 4.3. For the $b^{th}(b = 1, 2, \ldots, B)$ bootstrap sample using (16) we compute,

$$t_b^* = N^{-1}\sum_{i\in S_{nv}}[\pi_{11}(\mathbf{x}_{ib}^*;\widehat{\beta}_b^*,\widehat{\epsilon}_b^*)^{Y_{ib}^{obs*}}\pi_{10}(\mathbf{x}_{ib}^*;\widehat{\beta}_b^*,\widehat{\epsilon}_b^*)^{1-Y_{ib}^{obs*}} - \pi_{11}(\mathbf{x}_{ib}^*;\widehat{\beta}_b,\widehat{\epsilon}_b)^{Y_{ib}^{obs*}}\pi_{10}(\mathbf{x}_{ib}^*;\widehat{\beta}_b,\widehat{\epsilon}_b)^{1-Y_{ib}^{obs*}}] + N^{-1}\sum_{i\in\overline{S}}\{\pi(\mathbf{x}_i;\widehat{\beta}^*) - \pi(\mathbf{x}_i;\widehat{\beta})\}. \quad (24)$$

Let $V_*(.)$ denote the variance with respect to bootstrap distribution (P_B) defined as $P_B = P_v \times P_{nv}$, where P_v and P_{nv} represent the multinomial distribution $M_v(n_v; n_v^{-1},\ldots,n_v^{-1})$ and $M_{nv}(n-n_v;(n-n_v)^{-1},(n-n_v)^{-1},\ldots,(n-n_v)^{-1})$, respectively. Now the bootstrap estimator of $Var(T_1)$ is given by, For sufficiently large B, the Monte Carlo approximation of bootstrap variance estimator in (25) is given by,

$$\widehat{Var}_B(T_1) = B^{-1}\sum_b t_b^{*2} - (B^{-1}\sum_b t_b^*)^2. \quad (26)$$

Finally, the hybrid bootstrap variance estimator of $(\widehat{P}_C - P)$ is given by,

$$\widehat{V}_H = \widehat{Var}_B(T_1) + \widehat{Var}_A(T_2). \quad (27)$$

Theorem 6. Assuming $lim_{n_v\to\infty}(n-n_v)^{-1}\sum_{i\in S_{nv}}||g_\beta(\mathbf{x}_i;\beta)||^3 \infty \ \forall\ \beta$ and under conditions C0–C7, as $n_v \to \infty$,

$$\frac{\widehat{V}_H}{Var(\widehat{P}_C - P)} \xrightarrow{P} 1.$$

Proof. The proof is given in the online Supplementary Material. □

5. COMPARISON OF ESTIMATORS USING MODEL-BASED SIMULATION

A simulation study is carried out to investigate the impact of ignoring misclassification and/or the functional form $g(x)$ on the predictive estimator of the population proportion. We shall also study the relative performance of the analytical variance estimator and the bootstrap variance estimator. Data are generated separately for four sets of $g(x)$ using the probability model in (2). Following Adhya et al. (2011, 2012), the functions considered are (i) $g_1(x) = x - 2$, (ii) $g_2(x) = (x-2)^2$, (iii) $g_3(x) = \sin(2\pi x)$, and (iv) $g_4(x) = \exp(-50(x-1)^2)$. For further choices of $g(x)$, we refer to Section 3.1 of Breidt and Opsomer (2009). The steps for the computation of predictive estimator are briefed below.

Table 1. The Relative Bias ($RB \times 10^4$) and Relative Mean Square Error ($RRMSE \times 10^4$) of Three Estimators Corresponding to Different Choices of $g(x)$ and Misclassification Probabilities (ϵ_0, ϵ_1)

$g(x)$	$\overline{P}$	Error	$RB \times 10^4$			$RRMSE \times 10^4$		
		$\varepsilon_0, \varepsilon_1$	$\widehat{P}_{mean}$	$\widehat{P}_{native}$	$\widehat{P}_C$	$\widehat{P}_{mean}$	$\widehat{P}_{native}$	$\widehat{P}_C$
$(x-2)$	0.1553	(0.20, 0.10)	9,862	9,864	−53	9,910	9,910	1,747
		(0.10, 0.10)	4,416	4,420	−3	4,510	4,505	1,431
		(0.10, 0.20)	3,420	3,425	−12	3,529	3,524	1,568
$(x-2)^2$	0.9003	(0.20, 0.10)	−770	−1,380	−8	782	1,392	146
		(0.10, 0.10)	−878	−1,378	−0.95	889	1,389	123
		(0.10, 0.20)	−1,881	−2,186	−3	1,887	2,189	158
$\sin(2\pi x)$	0.4993	(0.20, 0.10)	972	107	5	1,022	117	53
		(0.10, 0.10)	−34	6	5	313	56	48
		(0.10, 0.20)	−1,031	−92	6	1,076	103	53
$e^{-50(x-1)^2}$	0.4999	(0.20, 0.10)	998	100	2	1,038	104	36
		(0.10, 0.10)	−70	−5	4	269	35	26
		(0.10, 0.20)	−1,002	−99	3	1,041	103	34

Step 5.1. Generate the auxiliary variable $x_i (i = 1, .., N)$ from Beta(1,2) distribution. For each x_i, compute $\pi(x_i; \beta)$ using (2). The true binary response $Y_i (i = 1, .., N)$ is then generated from Bernoulli distribution with the probability of success $\pi(x_i; \beta)$. The surrogate response $Y_i^{obs} (i = 1, .., N)$ is next generated using (3) and (4) for several prefixed choices of (ϵ_0, ϵ_1).

Step 5.2. Draw a simple random sampling without replacement (SRSWOR) sample of size n from the population units and capture the sample units in S. We store the values of Y_i^{obs} for every $i \in S$.

Step 5.3. Draw a SRSWOR sample of n_v units as a subsample from S and denote this set of validation units by S_v. Corresponding to each unit in S_v we store the true response $Y_i (i \in S_v)$.

Step 5.4. Given the data $(Y_i, i \in S_v; Y_i^{obs}, i \in S; \mathbf{x}_i, i = 1, 2, .., N)$, we compute the naive estimator $\widehat{P}_{naive}$ given in (7), the proposed estimator $\widehat{P}_C$ given in (13), and the sample mean of the manifest responses given by, $\widehat{P}_{mean} = n^{-1} \sum_{i \in S} Y_i^{obs}$.

Step 5.5. To assess the performance of the estimators in step 5.4, we calculate the relative bias (RB) and relative root mean square error ($RRMSE$), which are defined as, $RB = \overline{P}^{-1}[R^{-1} \sum_{r=1}^{R} (\widehat{P}^r - \overline{P})]$ and $RRMSE = \overline{P}^{-1} [R^{-1} \sum_{r=1}^{R} (\widehat{P}^r - \overline{P})^2]^{1/2}$, where $\widehat{P}^r$ is a generic estimate of P at the rth simulation and $\overline{P} = R^{-1} \sum_{r=1}^{R} P^r$, where P^r is the finite population proportion of true Y_i's generated in step 5.1 at the rth simulation $(r = 1, 2, \ldots, R)$.

For $N = 10{,}000$, $n = 1{,}000$, $n_v = 200$, and $R = 1{,}000$ and for three choices of (ϵ_0, ϵ_1), table 1 reports the values of $RB \times 10^4$ and $RRMSE \times 10^4$ for the

three estimators mentioned in step 5.4. The results reveal that $\widehat{P}_{mean}$ performs poorly specifically corresponding to $g_3(x)$ and $g_4(x)$. $\widehat{P}_{naive}$ also gives large *RB* and *RRMSE* values compared to $\widehat{P}_C$. So in the next part of the simulation study, we compare the performance of the analytical variance estimator and bootstrap variance estimator of $\widehat{P}_C$ only.

The bootstrap variance is based on $B = 1{,}000$ resamples drawn by the method outlined in section 4. For $r = 1, \ldots, R$, the analytical variance estimator $\widehat{V}_A^r$ given in (17) and the hybrid bootstrap variance estimator $\widehat{V}_H^r$ given in (27) are calculated. We further compute $\overline{V}_A = R^{-1}\sum_{r=1}^{R}\widehat{V}_A^r$ and $\overline{V}_H = R^{-1}\sum_{r=1}^{R}\widehat{V}_H^r$.

To avoid overfitting and to study the performance of the variance estimators of $\widehat{P}_C$ and the corresponding normal-theory-based confidence intervals of P, we generate an independent population of size N as in step 5.1 and compute the population proportion which we denote by $\tilde{P}$. Let $\widehat{\tilde{P}}_C$ denote the predictive estimator based on a sample from this new population. We compute $\tilde{V} = R^{-1}\sum_{r=1}^{R}(\widehat{\tilde{P}}_C^r - \overline{\tilde{P}})^2$, where $\overline{\tilde{P}} = R^{-1}\sum_{r=1}^{R}\tilde{P}^r$. The performance of the variance estimates is measured by the ratio of the standard errors (*RSE*) given by $RSE_H = \sqrt{\frac{\overline{V}_H}{\tilde{V}}}$ and $RSE_A = \sqrt{\frac{\overline{V}_A}{\tilde{V}}}$, corresponding to the hybrid bootstrap variance estimator and analytical variance estimator respectively.

Based on the standardized prediction errors $Z_H^r = (\widehat{P}_C^r - P)(\widehat{V}_H^r)^{-1/2}$ and $Z_A^r = (\widehat{P}_C^r - P)(\widehat{V}_A^r)^{-1/2}$, $(r = 1, \ldots, R)$, we set up the 95 percent confidence interval for P. The empirical coverage based on bootstrap-standardized prediction errors is given by $NCL_H = R^{-1}\sum_{r=1}^{R} I(|Z_H^r| \leq 1.96)$. The lower and upper tail areas are obtained from the proportions $NCL_{LH} = R^{-1}\sum_{r=1}^{R} I(Z_H^r \leq -1.645)$ and $NCL_{UH} = R^{-1}\sum_{r=1}^{R} I(Z_H^r > 1.645)$, respectively, where $I(.)$ is the indicator function. Similar expressions denoted by NCL_A, NCL_{LA}, and NCL_{UA} are obtained on using Z_A^r.

Table 2 reports the values of $\overline{P}$, $MSE = R^{-1}\sum_{r=1}^{R}(\widehat{P}_C^r - \overline{P})^2$, RSE_A, RSE_H, and the coverage based on the 95 percent confidence interval of the population proportion for $g_1(x)$–$g_4(x)$ for $(\epsilon_0, \epsilon_1) = (0.10, 0.20)$. The purpose of this study is to compare the performance of the bootstrap variance estimate with that of analytical variance estimate. The results reveal that RSE_H is always lesser than RSE_A. For $g_1(x)$ and $g_2(x)$, RSE_H is very close to one. However, for $g_3(x)$ and $g_4(x)$, the variances are overestimated, though overestimation is much higher for the analytical variance estimate compared to bootstrap variance estimate. The increased value of *RSE* may be due to the fact that *MSE* values corresponding to $g_3(x)$ and $g_4(x)$ are very small (of the order of 10^{-5}). The results also indicate that the coverage probabilities corresponding to both sided confidence interval, constructed using NCL_H returns value close to the nominal level, while those using NCL_A give coverage >0.95. The tail probabilities

Table 2. The Mean Square Error (*MSE*) of the Proposed Predictive Estimator ($\widehat{P}_C$), the Ratio of Standard Errors (*RSE*), the Lower Tail Area (NCL_L), the Upper Tail Area (NCL_U), and the Overall Coverage (*NCL*) of 95 Percent Confidence Intervals Based on Analytical and Bootstrap Estimate (within Parenthesis) of the Variance of $\widehat{P}_C$ for Different $g(x)$ and for $(\epsilon_0, \epsilon_1) = (0.10, 0.20)$

$g(x)$	$\overline{P}$	*MSE*	*RSE*	*NCL*	NCL_U	NCL_L
$(x-2)$	0.1614	2.796×10^{-4}	1.2387	0.988	0.016	0.021
			(1.1509)	(0.965)	(0.045)	(0.038)
$(x-2)^2$	0.9298	2.512×10^{-4}	1.2918	0.985	0.009	0.024
			(1.0171)	(0.951)	(0.070)	(0.045)
$\sin(2\pi x)$	0.5762	5.114×10^{-5}	1.3951	0.972	0.012	0.026
			(0.9788)	(0.952)	(0.035)	(0.051)
$e^{-50(x-1)^2}$	0.5051	1.1417×10^{-5}	1.9129	0.970	0.010	0.030
			(1.4715)	(0.945)	(0.021)	(0.061)

though not accurately captured give better result on using $NCL_{LH}(NCL_{UH})$ compared to $NCL_{LA}(NCL_{UA})$ for any function $g(x)$. Finer bootstrap intervals based on percentiles can be used to further improve the bootstrap coverages (DiCiccio, Efron, Hall, Martin, Canty, et al. 1996).

6. A DESIGN-BASED SIMULATION STUDY

A design-based simulation study is carried out to investigate the role of sampling design, the impact of misspecification of the functional form of $g(x)$, and the consequences of ignoring misclassification on the performance of the following estimators: (i) the sample mean ($\widehat{P}_{mean}$), (ii) the design-based estimator ($\widehat{P}_D$), (iii) the generalized difference estimator ($\widehat{P}_{GD}$), (iv) the model calibrated estimator ($\widehat{P}_{MC}$), and (v) the proposed model-based predictive estimator ($\widehat{P}_C$). Simulations are carried out under two models namely the naive model (M_1) and the misclassification-adjusted model M_2. The simulation study is briefed below.

Step 6.1. For a finite population of size $N = 10{,}000$, the auxiliary variable $x_i (i = 1, \ldots, N)$ is independent draws from Normal distribution with mean 0 and variance 2. The latent binary response variable $Y_i (i = 1, \ldots, N)$ is generated from Bernoulli distribution with success probability $\pi(x_i; \beta)$ computed from model (2) separately for each $g(x)$ given by (i) $g_1(x) = x - 2$ and (ii) $g_2(x) = (x - 2.5)^2$. Finally, the manifest responses $Y_i^{obs} (i = 1, 2, .., N)$ are generated using (3) and (4) for several prefixed choices of (ϵ_0, ϵ_1).

Step 6.2. We draw samples of size $n = 1{,}000$ from the populations in step 6.1 using (i) SRSWOR, (ii) probability proportional to size sampling with replacement (PPSWR) and (iii) stratified random sampling with proportional

allocation (SPA). For SPA, the populations are stratified into three groups of sizes 3,000, 3,000, and 4,000 by using quantiles of x. The samples are drawn using SRSWOR, SRSWR, and PPSWR from the first, second, and third strata, respectively. In PPSWR, the x^2 values are used as size variables. We denote the set of sampled units by S.

Step 6.3. For SRSWOR and PPSWR sampling schemes, we draw a validation sample of size n_v from the set S, by SRSWOR scheme. For SPA, we choose validation samples from each stratum using proportional allocation method for a fixed n_v. We denote the collection of all validation units by S_v. Corresponding to the validation units, we record the true response. The study is carried out for various choices of n_v. However, results are reported for $n_v =$ 100 and 300.

Step 6.4a. For the naive model M_1, based on the data $(Y_i^{obs},\ i \in S; x_i, i = 1, 2, \ldots, N)$ obtained from step 6.2, we compute the estimators (i)–(v) as mentioned at the beginning of this section. Expressions of $\widehat{P}_{mean}$, $\widehat{P}_D$, $\widehat{P}_{GD}$ and $\widehat{P}_{MC}$ are given in appendix A.4.1 while the naive predictive estimator denoted by $\widehat{P}_{naive}$ is given in (7).

Step 6.4b. For the misclassification adjusted model, M_2, based on the data $(Y_i^{obs},\ i \in S; x_i, i = 1, 2, \ldots, N; Y_i, i \in S_v)$ obtained from step 6.3, we compute the proposed predictive estimator $\widehat{P}_C$ given in (13). However, $\widehat{P}_{mean}$, $\widehat{P}_D$, $\widehat{P}_{GD}$, and $\widehat{P}_{MC}$ (details given in appendix A.4.2) are based only on the correct data $(Y_i, i \in S_v)$, since there is no straightforward decomposition of S into S_v and S_{nv}.

Steps 6.2–6.4a and b are repeated $R = 1{,}000$ times and the *RRMSE* of the estimators are computed. Tables 3–5 report the *RRMSE* $\times 10^4$ values for each of the estimators (i)–(v) given at the beginning of this section, for $(\epsilon_0, \epsilon_1) = (0.05, 0.01), (0.10, 0.10)$, and $(0.20, 0.10)$, respectively. The results are reported for models M_1 and M_2, for $n_v = (100, 300)$ and $g(x) = (g_1(x), g_2(x))$.

The estimators (iii)–(v) are further computed using misspecified models. The misspecifications introduced are as follows: corresponding to $g_1(x)$, we introduce another auxiliary variable namely z, where $(z_1, z_2, \ldots, z_N)$ are random draws from Uniform(0.5,1). For $g_2(x)$, we drop the quadratic term. The results are given in parenthesis in tables 3–5.

In general, performance of all the estimators under misclassification corrected model (M_2) is better compared to that under the naive model (M_1). Since under M_2 the estimators (i)–(iv) are constructed on the basis of validation sample only, so their performance improves with increase in validation sample size. This is evident while comparing the columns $n_v = 100$ with $n_v = 300$ under M_2. When errors are appreciably small (for instance, see table 3), the performance of the estimators (i)–(iv) under M_1 is better compared to those under M_2. This is not unusual since under M_1 the estimators are now constructed based on the complete sample of mostly correct observations, while a relatively

Table 3. The Relative Root Mean Square Errors ($RRMSE \times 10^4$) for Each of the Estimators under the Naive Model (M_1) and the Corrected Model (M_2), for Validation Sample Size $n_v = (100, 300)$, for Different Choices of $g(x)$ and for Three Sampling Schemes, when $(\epsilon_0, \epsilon_1) = (0.05, 0.01)$

$g(x)$	Estimator	SRSWOR			PPSWR			SPS		
		M_1	M_2		M_1	M_2		M_1	M_2	
			n_v			n_v			n_v	
			100	300		100	300		100	300
$(x-2)$	$\widehat{P}_{mean}$	1,622	1,865	1,091	6,825	6,216	5,827	2,905	1,946	1,618
	$\widehat{P}_D$	1,622	1,865	1,091	3,248	3,647	3,359	1,698	1,544	1,049
	$\widehat{P}_{GD}$	1,613	1,612	899	3,058	3,413	2,054	1,682	1,425	963
		(1,622)	(1,673)	(974)	(2,801)	(3,924)	(3,052)	(1,725)	(1,920)	(1,025)
	$\widehat{P}_{MC}$	1,613	1,612	899	5,972	3,019	2,584	1,698	1,505	1,049
		(1,622)	(1,670)	(974)	(3,527)	(3,848)	(2,719)	(1,743)	(1,910)	(1,066)
	$\widehat{P}_C$	1,613	907	626	4,065	1,680	958	1,840	920	631
		(1,622)	(967)	(632)	(4,195)	(2,330)	(977)	(2,014)	(999)	(697)
$(x-2.5)^2$	$\widehat{P}_{mean}$	111	299	193	414	548	424	163	329	239
	$\widehat{P}_D$	111	299	193	322	494	374	145	355	217
	$\widehat{P}_{GD}$	96	272	149	273	212	196	143	332	200
		(105)	(304)	(179)	(291)	(207)	(205)	(139)	(335)	(218)
	$\widehat{P}_{MC}$	96	284	149	1,377	1,084	716	195	473	232
		(105)	(300)	(180)	(1,415)	(2,041)	(1,978)	(157)	(562)	(264)
	$\widehat{P}_C$	150	44	42	113	96	94	120	66	64
		(105)	(139)	(115)	(115)	(104)	(99)	(310)	(370)	(359)

NOTE.— Figures in parenthesis indicate the values under misspecified $g_1(x)$ and $g_2(x)$.

Table 4. The Relative Root Mean Square Errors (*RRMSE* $\times 10^4$) for Each of the Estimators under the Naive Model M_1 and the Corrected Model M_2, for Validation Sample Size n_v = (100, 300), Different Choices of $g(x)$ and for Three Sampling Schemes, When $(\epsilon_0, \epsilon_1) = (0.10, 0.10)$

$g(x)$	Estimator	SRSWOR			PPSWR			SPS		
		M_1	M_2		M_1	M_2		M_1	M_2	
			n_v			n_v			n_v	
			100	300		100	300		100	300
$(x-2)$	$\widehat{P}_{mean}$	2,401	1,865	1,091	6,799	6,216	5,827	3,451	1,946	1,618
	$\widehat{P}_D$	2,401	1,865	1,091	3,857	3,647	3,360	2,410	1,543	1,049
	$\widehat{P}_{GD}$	2,403	1,612	899	3,740	3,413	2,054	2,400	1,425	963
		(2,365)	(1,673)	(974)	(3,263)	(3,924)	(3,052)	(2,473)	(1,920)	(1,025)
	$\widehat{P}_{MC}$	2,404	1,612	899	6,396	3,019	2,584	2,420	1,505	1,049
		(2,366)	(1,670)	(974)	(3,780)	(3,848)	(2,719)	(2,495)	(1,910)	(1,066)
	$\widehat{P}_C$	2,403	1,298	794	5,297	1,743	1,130	2,556	1,315	813
		(2,365)	(1,375)	(798)	(5,454)	(2,913)	(1,185)	(2,747)	(1,403)	(950)
$(x-2.5)^2$	$\widehat{P}_{mean}$	898	299	193	1,197	548	424	752	329	240
	$\widehat{P}_D$	898	299	193	1,048	494	374	904	355	217
	$\widehat{P}_{GD}$	911	272	149	1,121	212	196	918	332	200
		(898)	(304)	(179)	(1,037)	(207)	(205)	(920)	(355)	(218)
	$\widehat{P}_{MC}$	900	284	149	1,759	1,084	716	919	473	232
		(898)	(300)	(180)	(1,746)	(2,041)	(1,978)	(920)	(562)	(264)
	$\widehat{P}_C$	860	91	59	836	142	136	744	80	78
		(898)	(332)	(204)	(990)	(148)	(139)	(645)	(2,122)	(309)

NOTE.— Figures in parenthesis indicate the values under misspecified $g_1(x)$ and $g_2(x)$.

Table 5. The Relative Root Mean Square Errors ($RRMSE \times 10^4$) for Each of the Estimators under the Naive Model M_1 and the Corrected Model M_2, for Validation Sample Size $n_v = (100, 300)$, for Different Choices of $g(x)$ and for Three Sampling Schemes, When $(\epsilon_0, \epsilon_1) = (0.20, 0.10)$

$g(x)$	Estimator	SRSWOR			PPSWR			SPS		
		M_1	M_2		M_1	M_2		M_1	M_2	
			n_v			n_v			n_v	
			100	300		100	300		100	300
$(x-2)$	$\widehat{P}_{mean}$	5,692	1,865	1,091	9,596	6,216	5,827	6,639	1,946	1,618
	$\widehat{P}_D$	5,692	1,865	1,091	6,609	3,647	3,360	5,727	1,544	1,049
	$\widehat{P}_{GD}$	5,697	1,612	904	6,547	3,413	2,054	5,720	1,425	963
		(5,695)	(1,673)	(974)	(6,333)	(3,924)	(3,052)	(5,824)	(1,920)	(1,025)
	$\widehat{P}_{MC}$	5,697	1,612	899	8,474	3,019	2,584	5,745	1,505	1,049
		(5,695)	(1,670)	(974)	(6,670)	(3,848)	(2,719)	(5,853)	(1,910)	(1,066)
	$\widehat{P}_C$	5,697	1,612	899	9,340	2,141	1,202	5,990	1,401	892
		(5,695)	(1,742)	(919)	(9,533)	(3,106)	(1,357)	(6,219)	(1,677)	(1,073)
$(x-2.5)^2$	$\widehat{P}_{mean}$	795	299	193	1,053	548	424	664	329	240
	$\widehat{P}_D$	795	299	193	944	494	374	799	355	217
	$\widehat{P}_{GD}$	809	272	149	1,045	212	196	813	332	200
		(795)	(304)	(179)	(935)	(207)	(205)	(812)	(355)	(218)
	$\widehat{P}_{MC}$	797	284	149	1,690	1,084	716	814	473	232
		(795)	(300)	(180)	(1,691)	(2,041)	(1,978)	(812)	(562)	(264)
	$\widehat{P}_C$	874	101	63	869	147	137	291	84	80
		(795)	(404)	(218)	(859)	(157)	(140)	(566)	(2,003)	(309)

NOTE.— Figures in parenthesis indicate the values under misspecified $g_1(x)$ and $g_2(x)$.

small validation sample is used to evaluate the estimators under M_2. The proposed predictive estimator ($\widehat{P}_C$) is performing well under all sampling schemes and is markedly better than the design-based estimators. Even under M_1, the predictive estimator $\widehat{P}_{naive}$ is outperforming its design-based counterparts especially when the functional form is quadratic. In general, under misspecified models, precision of all the estimators deteriorates.

7. AN EMPIRICAL STUDY

In a literacy survey program, the complete information on literacy status and various covariates is obtained for 12,353 study participants. The response is binary, which takes the value 1 if an individual is identified as literate and 0 otherwise. The literacy status of an individual obtained by the interviewer method is suspected to be error prone and so in our study we treat it as the surrogate response (Y^{obs}). In a separate approach, an elaborate story reading test is also applied on the study participants to judge their true literacy status. We assume that the literacy status obtained from such a test is the gold standard and so we have data on the true response (Y) as well. From the data on Y and Y^{obs}, an estimated probability of a false literate comes out as 0.33, while the probability of misclassifying a literate as illiterate is only 0.021. For illustration purpose, we assume that these 12,353 study participants form the population and the true population proportion of literate is 0.508298.

We model the literacy status as a function of the continuous covariate "age," which spans over six to forty-five years. A data adaptive model selection procedure is proposed for choosing the function $g(x)$ based on available data on true Y and x. We form nine age categories and choose x as the class mark. The proportion $p(x)$ of literates corresponding to each age category is computed and $g(x)$ is calculated from the equation $g(x) = \log\left(\frac{p(x)}{1-p(x)}\right)$. A scatter plot of $g(x)$ versus x shows a steady increase in $g(x)$ for the initial age groups. After reaching the age thirteen to fourteen years, the graph shows a decline. Motivated by the scatter plot, we have chosen year 15 as the change point and have fitted two separate linear regression equations: one for age less than fifteen years and the other for age greater than or equal to fifteen years. We thus choose $g(x)$ as:

$$\begin{aligned} g(x) &= \beta_{01} + \beta_1 x \ , x < 15, \\ &= \beta_{02} + \beta_2 x \ , x \geq 15. \end{aligned} \tag{28}$$

To carry out the empirical study, we select samples of size $n = 1{,}200$ from the aforementioned population using three different sampling schemes namely SRSWOR, PPSWR, and SPA. For PPSWR, we use x as the size variable. For SPA, three strata are formed using the first and third quartiles of x. The strata sizes come out as, 3,182, 6,194, and 2,977. The stratum sample sizes are

Table 6. The Relative Bias ($RB \times 10^4$) and Relative Root Mean Square Error ($RRMSE \times 10^4$ in Parenthesis) for the Different Estimators under Naive Model M_1 and Misclassification Corrected Model M_2 for Three Different Sampling Schemes for the Literacy Survey Data

Estimator	SRSWOR		PPSWR		SPS	
	M_1	M_2	M_1	M_2	M_1	M_2
$\widehat{P}_{mean}$	3,011	−28	−158	−1,834	2,970	−25
	(3,022)	(979)	(318)	(2,019)	(2,982)	(871)
$\widehat{P}_{D}$	3,011	−28	3,014	69	3,007	12
	(3,022)	(979)	(3,055)	(1,746)	(3,020)	(877)
$\widehat{P}_{GD}$	3,014	−12	3,020	263	738	−1,925
	(3,023)	(918)	(3,058)	(1,894)	(780)	(2,106)
$\widehat{P}_{MC}$	3,014	−13	3,073	263	727	−1,944
	(3,023)	(918)	(3,319)	(2,168)	(772)	(2,129)
$\widehat{P}_{C}$	−7,257	719	−7,786	742	2,958	−679
	(7,257)	(797)	(7,786)	(888)	(2,977)	(754)

determined using proportional allocation, and we adopt SRSWOR, SRSWR, and PPSWR schemes for sampling from strata 1, 2, and 3, respectively.

However, in many practical situations, data on the true response (Y) will not be available for all these n sampled participants. Keeping in line with the methodology developed, we select a subsample of size $n_v = 120$ participants by SRSWOR from these n sampled participants. This subsample forms the validation set. The simulation is repeated $R = 1{,}000$ times. Table 6 reports the values of relative bias ($RB \times 10^4$) and relative root mean square error ($RRMSE \times 10^4$) in parenthesis, for all the estimators discussed in section 6 for models M_1 and M_2 under each of the three different sampling schemes.

The results reveal that, in general, the estimators are performing poorly under M_1 in terms of *RB* and *RRMSE*. However, for PPSWR sampling scheme, $\widehat{P}_{mean}$ and for SPS scheme, $\widehat{P}_{MC}$ are performing better under M_1 compared to their corrected counterparts. One reason for this might be that although under M_1 the surrogate response is used but the sample size is 1,200, which is quite large compared to the validation sample of 120 true response data points, which are used to construct the design-based estimators under the correct model M_2.

Our primary concern here is comparison of the proposed predictive model-based estimator with the other design-based estimators. It is observed that, under the misclassification corrected model (M_2), the predictive estimator is out performing all other estimators under different sampling schemes.

8. CONCLUDING REMARKS

The article adopts a predictive approach of estimation of finite population proportion of a misclassified binary response variable when complete information on the auxiliary variable(s) is available. The proposed predictive estimator that adjusts for misclassification is asymptotically model-unbiased and model-consistent. It also has minimum model variance among all asymptotically model-unbiased estimators of the finite population proportion (Chambers and Clark 2012). We have also developed a computationally efficient bootstrap-based weakly consistent estimator of the asymptotic variance, which performs better in comparison to the analytical variance estimator. The simulation studies show that when misclassification is moderate or high, the naive estimator performs poorly.

The current work can be extended to non-probability sampling design as long as it is non-informative. The simulation study in section 6 reveals that the predictive estimator is not robust to misspecification of the functional form of $g(x)$. In this context, nonparametric or semiparametric regression is another plausible alternative (Montanari and Ranalli 2005; Breidt and Opsomer 2009; Adhya et al. 2012). In this article, we have considered misclassification probabilities as unknown constants. In practice, misclassification probabilities can be modeled as a function of the auxiliary or design variable(s). The study can be extended to polychotomous response variable where the categories could be nominal or ordinal. Simulation studies reveal that the design-based estimators are less efficient. Further study can be taken up to explore how information from non-validation data can be incorporated to increase the efficiency of model-assisted estimators using a two-step approach. These problems are currently under investigation.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

APPENDIX A

A.1. PROOF OF THEOREM 1

Consistency of $\widehat{\epsilon}$: From Weak Law of Large Numbers for independently and identically distributed random variables, it follows that,

$$\widehat{\epsilon}_0 \xrightarrow{P} \frac{P\left(Y=0, Y^{obs}=1\right)}{P(Y=0)} = P\left(Y^{obs}=1|Y=0\right) = \epsilon_0^0.$$

Similarly $\widehat{\epsilon}_1 \to^P \epsilon_1^0$. Hence the consistency of $\widehat{\epsilon} = (\widehat{\epsilon}_0, \widehat{\epsilon}_1)^T$ is proved.

Consistency of $\widehat{\beta}$: Define $Q_n(\beta, \epsilon)$ as,

$$Q_n(\beta, \epsilon) = n^{-1} l(\beta, \epsilon) = f_v Q_{1n}(\beta) + (1 - f_v) Q_{2n}(\beta, \epsilon), \tag{A.1}$$

where $l(\beta, \epsilon)$ is given in (9), $Q_{1n}(\beta) = n_v^{-1} \sum_{i \in S_v} \log f(y_i | \mathbf{x}_i; \beta)$ and $Q_{2n}(\beta, \epsilon) = (n - n_v)^{-1} \sum_{i \in S_{nv}} \log f(y_i^{obs} | \mathbf{x}_i; \beta, \epsilon)$. In what follows we shall denote $Q_n(\beta, \epsilon^0)$ simply by $Q_n(\beta)$. Let us further write,

$$Q(\beta) = \rho_v Q_1(\beta) + (1 - \rho_v) Q_2(\beta, \epsilon^0), \tag{A.2}$$

where $Q_1(\beta) = lim_{n \to \infty} Q_{1n}(\beta)$ and $Q_2(\beta, \epsilon^0) = lim_{n \to \infty} Q_{2n}(\beta, \epsilon^0)$. We now make the following assumptions:

(i) $Q(\beta)$ is a continuous function of β.
(ii) $\widehat{\beta} = arg\ max_\beta Q_n(\beta)$ is the PMLE of β.
(iii) $\beta^0 = arg\ max_\beta Q(\beta)$ is the unique maximizer.
(iv) $sup_\beta |Q_n(\beta) - Q(\beta)| \to^P 0 \quad as \quad n \to \infty$.

Under the above assumptions and applying Theorem 2.1 of Newey and McFadden (1994), consistency of $\widehat{\beta}$ can be proved. Proofs of assumptions (ii), (iii), and (iv) are available in the online Supplementary Material.

A.2. PROOF OF THEOREM 2

We define the quantities $I(\beta, \epsilon)$, $G(\beta, \epsilon)$ and $W(\beta, \epsilon)$ as follows:

$$I(\beta, \epsilon) = \rho_v I_{S_v}(\beta) + (1 - \rho_v) I_{S_{nv}}(\beta, \epsilon), \tag{A.3}$$

where $I_{S_v}(\beta) = E_{S_v}[g_\beta(\mathbf{x}; \beta) g_\beta^T(\mathbf{x}; \beta) \pi(\mathbf{x}; \beta)(1 - \pi(\mathbf{x}; \beta))]$ and $I_{S_{nv}}(\beta, \epsilon) = E_{S_{nv}}[g_\beta(\mathbf{x}; \beta) g_\beta^T(\mathbf{x}; \beta) \pi^2(\mathbf{x}; \beta) (1 - \pi(\mathbf{x}; \beta))^2 \tilde{\pi}^{-1}(\mathbf{x}; \beta, \epsilon)(1 - \tilde{\pi}(\mathbf{x}; \beta, \epsilon))^{-1}]$.

$$G(\beta, \epsilon) = [I_p,\ G_1(\beta, \epsilon),\ G_2(\beta, \epsilon)], \tag{A.4}$$

where $G_i(\beta, \epsilon) = (-)^i (1 - \rho_v)(1 - \epsilon_0 - \epsilon_1) A_i(\beta, \epsilon), \quad i = 1, 2; \quad A_i(\beta, \epsilon) = E_{S_{nv}}[g_\beta(\mathbf{x}; \beta) \pi^i(\mathbf{x}; \beta)\ (1 - \pi(\mathbf{x}; \beta))^{3-i} \tilde{\pi}^{-1}(\mathbf{x}; \beta)(1 - \tilde{\pi}(\mathbf{x}; \beta, \epsilon))^{-1}],\ i = 1, 2$ and I_p is the identity matrix of order p. Finally,

$$W(\beta,\epsilon)=\begin{pmatrix} I(\beta,\epsilon) & w_1(\beta,\epsilon) & w_2(\beta,\epsilon) \\ w_1(\beta,\epsilon) & w_3(\beta,\epsilon) & 0 \\ w_2(\beta,\epsilon) & 0 & w_4(\beta,\epsilon) \end{pmatrix}, \tag{A.5}$$

where $w_1(\beta,\epsilon)=-\epsilon_0 b(\beta)(1-a(\beta))^{-1}$; $w_2(\beta,\epsilon)=\epsilon_1 b(\beta)a^{-1}(\beta)$; $w_3(\beta,\epsilon)=\epsilon_0\rho_v^{-1}(1-a(\beta))^{-2}E_{S_v}[(1-\pi(\mathbf{x};\beta))(1-\epsilon_0(1-\pi(\mathbf{x};\beta)))]$ and $w_4(\beta,\epsilon)=\epsilon_1\rho_v^{-1}a^{-2}(\beta)E_{S_v}[\pi(\mathbf{x};\beta)(1-\epsilon_1\pi(\mathbf{x};\beta))]$ with $a(\beta)$ and $b(\beta)$ defined as, $a(\beta)=E_{S_v}[\pi(\mathbf{x};\beta)]$ and $b(\beta)=E_{S_v}[g(\mathbf{x};\beta)\pi(\mathbf{x};\beta)(1-\pi(\mathbf{x};\beta))]$.

To prove the asymptotic normality of $\widehat{\beta}$, we first define, $\overline{S}_n(\beta,\epsilon)=n^{-1}\frac{\partial l(\beta,\epsilon)}{\partial\beta}$. Now using consistency of $\widehat{\beta}$ (vide theorem 1) and applying Taylor series expansion, it follows that,

$$\widehat{\beta}-\beta^0=I^{-1}(\beta^0,\epsilon^0)\overline{S}_n(\beta,\widehat{\epsilon})+o_p(1). \tag{A.6}$$

Applying Slutzky's theorem and equations (1.6)–(1.7) of Randles (1982), the asymptotic distribution of $\overline{S}_n(\beta,\widehat{\epsilon})$ is given by,

$$\sqrt{n}\ \overline{S}_n(\beta^0,\widehat{\epsilon})\xrightarrow{d}N_p(0,G(\beta^0,\epsilon^0)W(\beta^0,\epsilon^0)G^T(\beta^0,\epsilon^0)). \tag{A.7}$$

Using (A.3) in (A.7), theorem 2 follows.

A.3. PROOF OF THEOREM 3

$$\widehat{P}_C-P=f(1-f_v)D_1+(1-f_v)D_2, \tag{A.8}$$

where D_1 and D_2 are given by,

$$D_1=(n-n_v)^{-1}\sum_{i\in S_{nv}}(E(Y_i|Y_i^{obs},\mathbf{x}_i;\widehat{\beta},\widehat{\epsilon})-Y_i). \tag{A.9}$$

$$D_2=(N-n)^{-1}\sum_{i\in\overline{S}}(E(Y_i|\mathbf{x}_i;\widehat{\beta})-Y_i). \tag{A.10}$$

Since $\widehat{\epsilon}=arg\ max_\epsilon\sum_{i\in S_v}\log f(y_i^{obs}|y_i;\epsilon)$, then it directly follows from Cox and Hinkley (1974) that $E(\widehat{\epsilon}-\epsilon)=O(n^{-1})$. In the current context, since $\widehat{\beta}$ is the PMLE of β, so equations (32)–(35) given in Cox and Hinkley (1974) needs to be modified, which finally yields $E(\widehat{\beta}-\beta)=O(n^{-1})$. From regularity conditions, it follows that $E(D_1)=O(n^{-1})$ and $E(D_2)=O(n^{-1})$. Hence from (A.8), theorem 3 follows.

A.4. EXPRESSIONS OF THE DESIGN-BASED ESTIMATORS DEVELOPED IN SECTION 6

A.4.1 The Naive Model (M_1)

$$\widehat{P}_{mean} = n^{-1}\sum_{i\in S} Y_i^{obs};\ \widehat{P}_D = \Big(\sum_{i\in S} w_i\Big)^{-1}\sum_{i\in S} w_i Y_i^{obs};$$

$$\widehat{P}_{GD} = N^{-1}\left[\sum_{i\in S} w_i Y_i^{obs} - \sum_{i\in S} w_i\pi(\mathbf{x}_i;\widehat{\beta}_d) + \sum_{i=1}^{N}\pi(\mathbf{x}_i;\widehat{\beta}_d)\right];$$

$$\widehat{P}_{MC} = N^{-1}\sum_{i\in S} w_i Y_i^{obs} + \widehat{A}\left(1 - N^{-1}\sum_{i\in S} w_i\right) + \widehat{B}N^{-1}\left[\sum_{i=1}^{N}\pi(\mathbf{x}_i;\widehat{\beta}_d) - \sum_{i\in S} w_i\pi(\mathbf{x}_i;\widehat{\beta}_d)\right].$$

where $\widehat{A} = N^{-1}\sum_{i\in S} Y_i^{obs} - \widehat{B}[N^{-1}\sum_{i\in S} w_i\pi(\mathbf{x}_i;\widehat{\beta}_d)]$, and $\widehat{B} = \widehat{B}_1(\widehat{B}_2)^{-1}$, with

$$\widehat{B}_1 = \sum_{i\in S} w_i\left\{\pi(\mathbf{x}_i;\widehat{\beta}_d) - \Big(\sum_{i\in S} w_i\Big)^{-1}\sum_{i\in S} w_i\pi(\mathbf{x}_i;\widehat{\beta}_d)\right\}\left\{Y_i^{obs} - \Big(\sum_{i\in S} w_i\Big)^{-1}\sum_{i\in S} w_i Y_i^{obs}\right\} \text{ and}$$

$$\widehat{B}_2 = \sum_{i\in S} w_i\left\{\pi(\mathbf{x}_i;\widehat{\beta}_d) - \Big(\sum_{i\in S} w_i\Big)^{-1}\sum_{i\in S} w_i\pi(\mathbf{x}_i;\widehat{\beta}_d)\right\}.$$

In the estimators defined above, w_i's are the design weights given by the inverse of the inclusion probability and $\widehat{\beta}_d$ is the inverse inclusion probability weighted pseudo-likelihood estimator of β. For further details, we refer to Wu and Sitter (2001).

A.4.2 Misclassification Corrected Model (M_2)

For constructing the above estimators under the misclassification corrected model (M_2), we replace S by the set of validation units (S_v) and use the true values of the response Y_i in lieu of Y_i^{obs}. Moreover, $\widehat{\beta}_d$ is also estimated on the basis of true $Y_i (i \in S_v)$. Since simple random sampling is adopted for the

selection of validation units so the design weights w_i's are replaced by $w_i f_v^{-1}$, where $f_v = n_v/n$.

REFERENCES

Adhya, S., T. Banerjee, and G. Chattopadhyay (2011), "Inference on Polychotomous Responses in Finite Populations," *Scandinavian Journal of Statistics*, 38, 788–800.

———. (2012), "Inference on Finite Population Categorical Response: Nonparametric Regression-Based Predictive Approach," *AStA Advances in Statistical Analysis*, 96, 69–98.

Breidt, F. J., and J. D. Opsomer (2009), "Nonparametric and Semiparametric Estimation in Complex Surveys," in *Handbook of Statistics*, ed. D. Pfeffermann and C. R. Rao, vol. 29, pp. 103–119, North-Holland, Amsterdam: Elsevier.

Bross, I. (1954), "Misclassification in 2 × 2 Tables," *Biometrics*, 10, 478–486.

Buonaccorsi, J. P. (2010), *Measurement Error: Models, Methods, and Applications*, Boca Raton, FL: Chapman and Hall/CRC.

Cassel, C. M., C. E. Särndal, and J. H. Wretman (1976), "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations," *Biometrika*, 63, 615–620.

Chambers, R., and R. Clark (2012), *An Introduction to Model-Based Survey Sampling with Applications*, New York: OUP.

Chambers, R., A. H. Dorfman, and P. Hall (1992), "Properties of Estimators of the Finite Population Distribution Function," *Biometrika*, 79, 577–582.

Chen, J., and J. Qin (1993), "Empirical Likelihood Estimation for Finite Populations and the Effective Usage of Auxiliary Information," *Biometrika*, 80, 107–116.

Chen, J., and R. Sitter (1999), "A Pseudo Empirical Likelihood Approach to the Effective Use of Auxiliary Information in Complex Surveys," *Statistica Sinica*, 9, 385–406.

Chen, T. T. (1979), "Log-Linear Models for Categorical Data with Misclassification and Double Sampling," *Journal of the American Statistical Association*, 74, 481–488.

Chen, Z., G. Y. Yi, and C. Wu (2011), "Marginal Methods for Correlated Binary Data with Misclassified Responses," *Biometrika*, 98, 647–662.

Claeskens, G., M. Aerts, and G. Molenberghs (2003), "A Quadratic Bootstrap Method and Improved Estimation in Logistic Regression," *Statistics & Probability Letters*, 61, 383–394.

Cox, D., and D. Hinkley (1974), *Theoretical Statistics*, Chapman and Hall.

Cox, D. R., and E. J. Snell (1989), *Analysis of Binary Data* (2nd ed.), Chapman and Hall.

Deville, J.-C., and C.-E. Särndal (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382.

DiCiccio, T. J., B. Efron, P. Hall, M. A. Martin, A. J. Canty, A. C. Davison, D. V. Hinkley, L. J. Gleser, S. M. S. Lee, G. A. Young, T. J. DiCiccio, and B. Efron (1996), "Bootstrap Confidence Intervals," *Statistical Science*, 11, 189–212.

Ekholm, A., and J. Palmgren (1982), "A Model for a Binary Response with Misclassifications," in *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, ed. R. Gilchrist, pp. 128–143, New York: Springer-Verlag.

Elliott, M. R., and R. Valliant (2017), "Inference for Nonprobability Samples," *Statistical Science*, 32, 249–264.

Fuller, W. (2009), *Sampling Statistics*, Wiley.

Gong, G., and F. J. Samaniego (1981), "Pseudo Maximum Likelihood Estimation: Theory and Applications," *The Annals of Statistics*, 9, 861–869.

Gustafson, P. (2003), *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, Chapman and Hall.

Hochberg, Y. (1977), "On the Use of Double Sampling Schemes in Analyzing Categorical Data with Misclassification Errors," *Journal of the American Statistical Association*, 72, 914–921.

Montanari, G. E., and M. G. Ranalli (2005), "Nonparametric Model Calibration Estimation in Survey Sampling," *Journal of the American Statistical Association*, 100, 1429–1442.

Newey, W. K., and D. McFadden (1994), *"Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics*, eds. R. F. Engle and D. L. McFadden, Vol.4, Amsterdam: Elsevier.

Poon, W.-Y., and H.-B. Wang (2010), "Bayesian Analysis of Multivariate Probit Models with Surrogate Outcome Data," *Psychometrika*, 75, 498–520.

Randles, R. H. (1982), "On the Asymptotic Normality of Statistics with Estimated Parameters," *The Annals of Statistics*, 10, 462–474.

Royall, R. M. (1970), "On Finite Population Sampling Theory under Certain Linear Regression Models," *Biometrika*, 57, 377–387.

———. (1976), "The Linear Least-Squares Prediction Approach to Two-Stage Sampling," *Journal of the American Statistical Association*, 71, 657–664.

Sang, H., K. K. Lopiano, D. A. Abreu, A. C. Lamas, P. Arroway, and L. J. Young (2017), "Adjusting for Misclassification: A Three-Phase Sampling Approach," *Journal of Official Statistics*, 33, 207–222.

Särndal, C. E. (1980), "On π-Inverse Weighting versus Best Linear Unbiased Weighting in Probability Sampling," *Biometrika*, 67, 639–650.

Smith, T. (1983), "On the Validity of Inferences from Non-Random Samples," *Journal of the Royal Statistical Society: Series A (General)*, 146, 394–403.

Sposto, R., D. L. Preston, Y. Shimizu, and K. Mabuchi (1992), "The Effect of Diagnostic Misclassification on Non-Cancer and Cancer Mortality Dose Response in A-Bomb Survivors," *Biometrics*, 48, 605–617.

Stefanski, L. A., and R. J. Carroll (1985), "Covariate Measurement Error in Logistic Regression," *The Annals of Statistics*, 13, 1335–1351.

Sugden, R., and T. Smith (1984), "Ignorable and Informative Designs in Survey Sampling Inference," *Biometrika*, 71, 495–506.

Tenenbein, A. (1970), "A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications," *Journal of the American Statistical Association*, 65, 1350–1361.

Thompson, W. L. (2002), "Towards Reliable Bird Surveys: Accounting for Individuals Present but Not Detected," *The Auk*, 119, 18–25.

Valliant, R., A. H. Dorfman, and R. M. Royall (2000), *Finite Population Sampling and Inference: A Prediction Approach*., Wiley.

Wang, D., and P. Gustafson (2014), "On the Impact of Misclassification in an Ordinal Exposure Variable," *Epidemiologic Methods*, 3, 97–106.

Wu, C., and R. Sitter (2001), "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data," *Journal of the American Statistical Association*, 96, 185–193.

Zhong, B., and J. Rao (2000), "Empirical Likelihood Inference under Stratified Random Sampling Using Auxiliary Population Information," *Biometrika*, 87, 929–938.