Journal of Survey Statistics and Methodology (2025) ***00****, 1–33*

STATISTICAL INFERENCE UNDER NONIGNORABLE SAMPLING AND NONRESPONSE—AN EMPIRICAL LIKELIHOOD APPROACH

DANNY PFEFFERMANN*
ARIE PREMINGER
ANNA SIKOV

Statistical models are often based on sample surveys. When the sample selection probabilities and/or the response probabilities are related to a model outcome variable, even after conditioning on the model covariates, the model holding for the observed data is different from the model holding in the population, resulting in biased inference if not accounted for properly. Accounting for sample selection bias is relatively simple because the sample selection probabilities are usually known. Accounting for nonignorable nonresponse is much harder since the response probabilities are, in practice, unknown. In this article, we develop a new approach for modelling complex survey data, which accounts simultaneously for nonignorable sampling and nonresponse. Our proposed approach combines the nonparametric empirical likelihood with a parametric model for the response probabilities, which contains the outcome variable as one of the covariates. Combining the model holding for the responding units with the model for the response probabilities enables extracting the model holding for the missing data and imputing them. We propose ways of testing the underlying model holding for the respondents' data. Simulation results illustrate the good performance of the approach in terms of parameter estimation and imputation. We conclude with an application to the household expenditure

Danny Pfeffermann is a Professor with Department of Statistics and Data Science, Hebrew University, 91905 Jerusalem, Israel, and Southampton Statistical Sciences Research Institute, University of Southampton, UK. Arie Preminger is a Senior Lecturer with Department of Economics, Academic College of Ramat Gan, 87 Pinhas Rotenberg St., 52275 Ramat Gan, Israel. Anna Sikov is an Assistant Professor with National University of Engineering, Av. Tupac Amaru 210, 15333 Rimac, Lima, Peru.
The authors are grateful to Dr Moshe Feder for his invaluable contribution to this research. The research of the first author was supported by a UK Economic and Social Research Council (ESRC) grant number RES-062–23-2316. There are no conflicts of interest regarding this article.
*Address correspondence to Danny Pfeffermann, Department of Statistics and Data Science, Hebrew University, 91905 Jerusalem, Israel. E-mails: msdanny@mail.huji.ac.il; msdanny@soton.ac.uk.

https://doi.org/10.1093/jssam/smaf015

survey in Israel, carried out by Israel's Central Bureau of Statistics. The survey collects information on the socio-demographic characteristics of each member of the sampled households (HH), as well as detailed information on the HH income and expenditure. The total sample size was $n = 12{,}136$ with 7,827 responding HHs. The target estimated parameter in this application is the population mean of the gross HH income.

KEY WORDS: Kernel smoothing; Model testing; NMAR nonresponse; Respondents' model; Sample model.

Statement of Significance

Survey data are often used for analytic inference, based on statistical models assumed to hold for the population from which the sample is taken. It is often the case, however, that the sampling design used to select the sample is informative in the sense that the sample selection probabilities are correlated with the outcome variable even after conditioning on model covariates, in which case the model holding for the sample data differs from the model holding in the population. Inevitably, sample data are subject to nonresponse, which is informative if the response probability is correlated with the outcome value after conditioning on the model covariates. Clearly, ignoring an informative sampling design and/or response mechanism may yield highly biased estimators. In this article, we develop a new approach for modelling complex survey data, which accounts simultaneously for nonignorable sampling and nonresponse. The approach combines the nonparametric empirical likelihood with a parametric model for the response probabilities. Combining the model holding for the responding units with the model for the response probabilities enables extracting the model holding for the missing data and imputing them. We propose ways of testing the underlying model holding for the respondents data. We also consider estimation of an assumed parametric population model based on our approach. The article contains simulation results and an application to a real data set.

1. INTRODUCTION

Survey data are often used for analytic inference, based on statistical models assumed to hold for the population from which the sample is taken. Familiar examples include the estimation of elasticity of demand from household (HH) expenditure surveys, estimation of health risk factors from health surveys and the analysis of market dynamics from labor force surveys. In particular, survey

data are used for estimating population parameters of interest such as totals and proportions. It is often the case, however, that the sampling design used to select the sample is informative for the population model in the sense that the sample selection probabilities are correlated with the target outcome variable even after conditioning on model covariates, in which case the model holding for the sample data is different from the model holding for the population values. This will happen, for example, when the selection probabilities are determined by one or more design variables (stratification variables, size variables used for probability proportional to size sampling, etc.), which are correlated with the model outcome variable, but some or all of them are not included among the model covariates. In an extreme case, the sample selection probabilities are determined directly by the outcome values, as in case-control studies.

Inevitably, sample data are subject to nonresponse, which is informative for the population model if the response propensity is correlated with the outcome value after conditioning on the model covariates, known as "not missing at random" (NMAR) nonresponse. For example, sampled units may choose not to respond to questions related to their income, based on their level of income. In section 8, we analyze data observed in a household expenditure survey, carried out by Israel's Central Bureau of Statistics. The survey collects information on socio-demographic characteristics of each member of the sampled households, as well as detailed information on the HH income and expenditure. The total sample size was $n = 12{,}136$ with 7,827 responding HHs. The target estimated parameter in this application is the population mean of the gross HH income. As shown in our application, the response probabilities depend on the household incomes even after conditioning on the model covariates.

Under NMAR nonresponse, the model holding for the data observed for the responding units is different from the sample model under complete response, which, as noted above, is different from the population model under informative sampling. Clearly, and as illustrated also in the present article, ignoring an informative sampling design and/or response mechanism may yield highly biased estimators and distort the inference.

Pfeffermann (2011) reviews several approaches proposed in the literature to deal with informative sampling, ranging from weighting each sample observation by the corresponding sampling weight to maximization of the sample likelihood as defined by the model holding for the sample data. A common feature of these approaches is that they utilize the sampling weights in the inference process, although in different ways. Accounting for NMAR nonresponse, however, is much more complicated since the response probabilities are practically never known, requiring some assumptions on them. Pfeffermann and Sikov (2011) review approaches proposed in the literature to deal with NMAR nonresponse, but these approaches are quite restricted. In particular, most of the approaches assume that the model covariates are known also for the nonrespondents, which is often not the case. Evidently, accounting for both

informative sampling and NMAR nonresponse in a single analysis is a major undertaking, and the present article attempts to address this challenge.

We assume that not only the outcome values are missing for the nonresponding units but also the corresponding covariate values, known as unit nonresponse. The only additional information beyond the data observed for the responding units assumed to be known is the population means of calibration variables, which may include some of the model covariates, and possibly also the mean of the outcome variable. The totals of such variables are often available from administrative or census records, or from large surveys. Note that even though in practice, fitting a model would usually be done for estimating unknown population parameters like means, it might be desired to fit a model even when the population mean of the outcome variable is known, for example, for estimating model parameters, such as regression coefficients of explanatory variables of interest. Our approach combines the nonparametric empirical likelihood (EL) for the population model with a parametric model for the response probabilities, which contains the outcome variable as one of the covariates. Moreover, the proposed approach allows the incorporation of additional estimating equations to accommodate estimation of the response model parameters. Specifically, the methods developed by Chang and Kott (2008) and Sverchkov (2008) are considered. A third component needed for setting the likelihood holding for the responding units is the expectation of the sampling weights given the outcome and the covariates, which we estimate nonparametrically, using kernel smoothing.

The use of the EL for analyzing complex survey data has its origin in a landmark paper by Hartley and Rao (1968), and has gained increasing interest in more recent years in general statistical contexts, following the work of Owen (1988, 1990, 1991, 2001, 2013). Another fundamental paper is Qin and Lawless (1994). See also Kim and Morikawa (2023), with references to other recent articles. The EL combines the robustness of nonparametric methods with the efficiency of the likelihood approach. Another important advantage of this method is that it lends itself to the use of calibration constraints, thus enhancing the precision of the estimators. See, for example, Chen and Van Keilegom (2009) for a review. As our proposed method is based on the empirical likelihood, conditional on the response, we refer to it as "Respondents Empirical Likelihood" (REL).

In the next section, we define the sample and respondents' distributions. In section 3, we present the empirical likelihood and provide details of its maximization. In addition, we describe the methods proposed by Chang and Kott (2008) and Sverchkov (2008), and explain how they can be incorporated into the estimation process. In section 4, we show how to use the estimates obtained from maximization of the REL for estimating parametric models assumed to hold in the population. Variance estimation is also considered. Section 5 discusses ways of validating the assumptions underlying our approach. Section 6 considers the imputation of the missing sample data. In

Section 7, we report the results of a simulation study aimed to illustrate the performance of the proposed method, while in Section 8, we apply the procedure to the data collected as part of the 2019 Household Expenditure Survey in Israel. Section 9 contains concluding remarks.

2. SAMPLE AND RESPONDENTS' DISTRIBUTIONS

Let y_i denote the value of an outcome variable Y associated with unit i belonging to a sample S, drawn from a finite population $U = \{1, \ldots, N\}$ with known inclusion probabilities $\pi_i = \Pr(i \in S)$. Let I_i denote the sampling indicator defined as 1 if unit i is sampled and 0 otherwise, and $\boldsymbol{x}_i = (x_{1,i}, \ldots, x_{k,i})'$ denote the values of k auxiliary variables (covariates) associated with unit i. Denote by R the set of respondents and define the response indicator R_i to be 1 if unit $i \in S$ responds and 0 otherwise. We denote by n the size of S and by r the size of R.

In what follows, we assume that the population outcomes are independent realizations from distributions with probability density functions (PDF) $f_u(y_i|\boldsymbol{x}_i)$. Following Pfeffermann et al. (1998), the sample PDF, $f_s(y_i|\boldsymbol{x}_i)$, is defined as the conditional PDF of y_i given that unit i is sampled, that is, $f_s(y_i|\boldsymbol{x}_i) = f(y_i|\boldsymbol{x}_i, I_i = 1)$. By Bayes Rule,

$$f_s(y_i|\boldsymbol{x}_i) = \frac{\Pr(I_i = 1|\boldsymbol{x}_i, y_i) f_u(y_i|\boldsymbol{x}_i)}{\Pr(I_i = 1|\boldsymbol{x}_i)}, \tag{2.1}$$

where $\Pr(I_i = 1|\boldsymbol{x}_i) = \int \Pr(I_i = 1|\boldsymbol{x}_i, y_i) f_u(y_i|\boldsymbol{x}_i) dy_i$. Note that $\Pr(I_i = 1|\boldsymbol{x}_i, y_i)$ is generally not the same as the sample inclusion probability $\pi_i = \Pr(i \in s) = \Pr(I_i = 1|Z_u)$, where Z_u defines a matrix of population values of design variables used for the sample selection. Since $\Pr(I_i = 1|\pi_i, y_i, \boldsymbol{x}_i) = \pi_i$, $\Pr(I_i = 1|y_i, \boldsymbol{x}_i) = E_u(\pi_i|y_i, \boldsymbol{x}_i)$, where E_u is the expectation under the population PDF. The population and sample PDFs differ unless $\Pr(I_i = 1|\boldsymbol{x}_i, y_i) = \Pr(I_i = 1|\boldsymbol{x}_i)$ for all y_i, and when this condition is not met, the sampling design is informative and cannot be ignored in the inference process. In particular, it follows from (2.1) that under informative sampling

$$E_s(y_i|\boldsymbol{x}_i) = E_u\left[\frac{\Pr(I_i = 1|\boldsymbol{x}_i, y_i)}{\Pr(I_i = 1|\boldsymbol{x}_i)} y_i|\boldsymbol{x}_i\right] \neq E_u(y_i|\boldsymbol{x}_i), \tag{2.2}$$

where E_s denotes the expectation with respect to the sample PDF. Estimating $E_u(y_i|\boldsymbol{x}_i)$ is often the main target of inference. Thus, ignoring an informative sampling scheme and practically estimating $E_s(y_i|\boldsymbol{x}_i)$ can severely bias the inference.

Next, consider the respondents' distribution. The marginal PDF for responding unit i, denoted by $f_R(y_i|\boldsymbol{x}_i) = f(y_i|\boldsymbol{x}_i, I_i = 1, R_i = 1)$, is by Bayes Rule,

$$f_R(y_i|\boldsymbol{x}_i) = \frac{\Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1) f_s(y_i|\boldsymbol{x}_i)}{\Pr(R_i = 1|\boldsymbol{x}_i, I_i = 1)}. \tag{2.3}$$

Here again, unless $\Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1) = \Pr(R_i = 1|\boldsymbol{x}_i, I_i = 1)$ for all i, the respondents' PDF differs from the sample PDF, which as shown above differs from the target population distribution under informative sampling. Notice that $\Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1)$ may not be the same as $\Pr(R_i = 1|y_i, \boldsymbol{x}_i)$ since in theory, the missingness generating mechanism among the sampled individuals may be different from the mechanism applied by the nonsampled individuals. However, the assumption that $\Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1) = \Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 0) = \Pr(R_i = 1|y_i, \boldsymbol{x}_i)$ is generally reasonable, reflecting an inherent tendency of an individual regarding their willingness to answer particular questions or all the questions of a survey.

So far, we have excluded for convenience from the notation the parameters governing the various distributions. If the outcome, the sampling and the response are independent between units, the respondents' likelihood takes the form,

$$L_R(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{\Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1; \boldsymbol{\gamma}) \Pr(I_i = 1|y_i, \boldsymbol{x}_i) f_u(y_i|\boldsymbol{x}_i; \boldsymbol{\beta})}{\Pr(R_i = 1|\boldsymbol{x}_i, I_i = 1; \boldsymbol{\beta}, \boldsymbol{\gamma}) \Pr(I_i = 1|\boldsymbol{x}_i)}. \tag{2.4}$$

Remark 1. In theory, one also needs to model the probabilities $\Pr(I_i = 1|y_i, \boldsymbol{x}_i)$. However, since $\Pr(I_i = 1|\pi_i, y_i, \boldsymbol{x}_i) = \pi_i$, the probability $\Pr(I_i = 1|y_i, \boldsymbol{x}_i)$ can be estimated outside the likelihood using the relationship $\Pr(I_i = 1|y_i, \boldsymbol{x}_i) = E_u(\pi_i|y_i, \boldsymbol{x}_i) = 1/E_s(w_i|y_i, \boldsymbol{x}_i)$, where $w_i = 1/\pi_i$ is the sampling weight (Pfeffermann and Sverchkov 1999). Thus, assuming that the probability to respond, $\Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1)$ does not depend on the sampling weight w_i, that is, the response is independent of the sample selection, we obtain that $E_s(w_i|y_i, \boldsymbol{x}_i) = E_R(w_i|y_i, \boldsymbol{x}_i)$. This implies that the probabilities $\Pr(I_i = 1|y_i, \boldsymbol{x}_i)$ can be estimated by regressing w_i on $(y_i, \boldsymbol{x}_i)$ using the observed data. See Pfeffermann and Sverchkov (2003, 2009) for plausible approaches and examples of modeling and estimating the expectations $E_s(w_i|y_i, \boldsymbol{x}_i)$. Alternatively, the expectations can be estimated nonparametrically using smoothing methods. In the simulation study described in section 7 and in the empirical application of section 8, we use kernel smoothing to obtain estimates of $\tau_i = \Pr(I_i = 1|y_i, \boldsymbol{x}_i)$, by applying kernel regression of w_i on $(y_i, \boldsymbol{x}_i)$. See section 7.2 for further details.

As discussed in section 4.2, the parameters underlying an assumed parametric population model can be estimated easily once the probabilities underlying the empirical likelihood have been estimated. In this respect, the use of the empirical likelihood can be viewed as a convenient way of estimating the parameters of an assumed parametric population model.

Remark 2. A notable property of the likelihood (2.4) is that it does not require knowledge of the covariates of the nonresponding units. On the other hand, even with good estimates of the probabilities $\Pr(I_i = 1|y_i, \boldsymbol{x}_i)$, the use of this likelihood requires specifying the population model $f_u(y_i|\boldsymbol{x}_i)$, and the response probabilities $\Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1)$, and with no observations obtained directly from either one of the two distributions, one may run into identification problems. Pfeffermann and Landsman (2011) and Wang et al. (2014) establish conditions under which likelihoods of the form (2.4) are identifiable, but experience shows that even under these conditions, maximization of the likelihood is often unstable, due to what Lee and Berger (2001) refer to as "practical nonidentifiability." See Rotnitzky and Robins (1997) for further discussion and theoretical results on the identifiability of likelihoods of the form (2.4).

Remark 3. Although no observations are available for either the model defining the population PDF or the model assumed for the response probabilities, the resulting respondents' model (2.3) can nonetheless be tested using classical test statistics, since it relates to the data observed for the responding units. See sections 5, 7, and 8 for the tests used in our empirical study, with illustrations.

3. RESPONDENTS EMPIRICAL LIKELIHOOD

3.1 Notation and Definition

We assume that for each population unit i corresponds a vector $\boldsymbol{u}_i = (y_i, \boldsymbol{z}_i')'$, such that $\boldsymbol{z}_i = \boldsymbol{x}_i \cup \boldsymbol{c}_i$, where $\boldsymbol{c}_i$ is a d-dimensional vector of survey values for which the population means $\bar{\boldsymbol{c}}_u$ are known sufficiently accurately, and y_i and $\boldsymbol{x}_i$ are related via a model $f_u(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})$. The vector $\boldsymbol{c}_i$ may include some or all of the variables in $\boldsymbol{x}_i$.

Denote, $\tau_i = \Pr(I_i = 1|y_i, \boldsymbol{x}_i)$ and $\rho_i = \Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1)$. We follow the load-scale approach of Hartley and Rao (1968) by assuming that the finite population values are generated from a multinomial distribution with a vector of probabilities $\boldsymbol{p} = (p_1, \ldots, p_r)'$, where $p_i = Pr_u(\boldsymbol{u}_i)$. We assume that the population distribution has its support in the set of the observed values.

Denote by N_i the number of units in the finite population, assuming the vector $\boldsymbol{u}_i$, such that $p_i = E(N_i/N)$, where $N = \sum_i N_i$ is the population size. Under this set-up, the distribution of the observed data for the responding units (hereafter the respondents' distribution) is multinomial, with cell probabilities given by $p_i^{(r)} = \Pr(\boldsymbol{u}_i|i \in R) = p_i\tau_i\rho_i / \sum_k p_k\tau_k\rho_k$. The empirical likelihood is then

$$\mathcal{L} = \prod_i p_i^{(r)} = \Pi(\boldsymbol{p}^{(r)}), \tag{3.1}$$

where we use the generic notation $\Pi(\boldsymbol{a}) = \prod_i a_i$ to denote the product of the elements of a vector $\boldsymbol{a}$. Chaudhuri et al. (2008) and Chaudhuri et al. (2010) use a similar formulation for their empirical likelihood, but they restrict it to the case of full response (viz., $\rho_i = 1$ for all i).

The response probabilities ρ_i in (3.1) are unknown and need to be estimated. We model ρ_i as a function of the outcome and the covariates. Specifically, we assume $\rho_i(\boldsymbol{\gamma}) = \Pr(R_i = 1|y_i, \boldsymbol{x}_i; \boldsymbol{\gamma}) = \text{ logit}^{-1}(\ell(y_i, \boldsymbol{x}_i; \boldsymbol{\gamma}))$, where $\text{logit}^{-1}(s) = (1+e^{-s})^{-1}$ and $\ell(y_i, \boldsymbol{x}_i; \boldsymbol{\gamma})$ is a polynomial in $(y, \boldsymbol{x})$ with coefficients $\boldsymbol{\gamma}$. Lemma 1 below asserts that if the probability to respond is a continuous function of $\boldsymbol{x}, y$, then it can be approximated arbitrarily close by a function of the form $\text{logit}^{-1}(\ell(y, \boldsymbol{x}; \boldsymbol{\gamma}))$, where $\ell(y, \boldsymbol{x})$ is a polynomial. Thus, the assumption that $\rho_i(\boldsymbol{\gamma})$ has this form is not as arbitrary as it might seem. The use of the logit function for modeling the unknown response probabilities is very common in the survey sampling literature. See, for example, Kim and Morikawa (2023) and the references therein.

Lemma 1. Assume that $\rho(y, \boldsymbol{x})$ is a continuous function on a closed bounded set D and $0 \le \rho(y, \boldsymbol{x}) \le 1$. Then, for every $\varepsilon > 0$, there exists a multivariate polynomial $Q(y, \boldsymbol{x})$ such that $|\text{logit}^{-1}(Q(y, \boldsymbol{x})) - \rho(y, \boldsymbol{x})| < \varepsilon$ for all $(y, \boldsymbol{x}) \in D$.

Proof. See the Appendix.

In summary, our REL is a combination of the nonparametric multinomial population distribution, the expectations $\tau_i = E_u(I_i|y_i, \boldsymbol{x}_i)$ and a model for the response probabilities.

3.2 Calibration Constraints

We mentioned in the introduction that the use of the empirical likelihood facilitates the use of calibration constraints for enhancing the efficiency of the estimators. Under our set-up, the calibration values satisfy $\sum_{i \in R} p_i \boldsymbol{c}_i \approx N^{-1} \sum_{i \in R} N_i \boldsymbol{c}_i = N^{-1} \sum_{j \in U} \boldsymbol{c}_j = \bar{\boldsymbol{c}}_u$, yielding the R-level constraints

$$\sum_{i \in R} p_i^{(r)} \tau_i^{-1} \rho_i^{-1} (\boldsymbol{c}_i - \bar{\boldsymbol{c}}_u) = \boldsymbol{0}. \tag{3.2}$$

It should be noted that in certain situations the sample means of some calibration variables are also available. In this case, additional constraints can be defined as $\sum_{i \in R} p_i^{(r)} \rho_i^{-1} (\boldsymbol{c}_i - \bar{\boldsymbol{c}}_s) = \boldsymbol{0}$, where $\bar{\boldsymbol{c}}_s$ denotes the vector of known

sample means. This constraint can be justified as follows. Let $p_i^{(s)} = \Pr(\boldsymbol{U_i} = \boldsymbol{u_i} \mid I_i = 1)$. Then, $p_i^{(r)} = \Pr(\boldsymbol{U_i} = \boldsymbol{u_i} \mid R_i = 1, I_i = 1) =$

$$\frac{\Pr(R_i = 1 \mid I_i = 1, \boldsymbol{U_i} = \boldsymbol{u_i})\Pr(\boldsymbol{U_i} = \boldsymbol{u_i} \mid I_i = 1)}{\Pr(R_i = 1 \mid I_i = 1)} = \frac{\rho_i p_i^{(s)}}{\sum_{j \in R} \rho_j p_j^{(s)}}.$$

Thus, $p_i^{(s)} = \frac{p_i^{(r)}}{\rho_i} \sum_{j \in R} p_j^{(s)} \rho_j$. Then the constraint can be obtained as $\sum_{i \in R} p_i^{(s)} \boldsymbol{c}_i = \bar{\boldsymbol{c}}_s$ or $\sum_{i \in R} p_i^{(s)} (\boldsymbol{c}_i - \bar{\boldsymbol{c}}_s) = \boldsymbol{0}$, implying $\sum_{i \in R} p_i^{(r)} \rho_i^{-1} (\boldsymbol{c}_i - \bar{\boldsymbol{c}}_s) = \boldsymbol{0}$. However, our experience shows that the gain in precision by inclusion of these additional constraints is very modest.

Denote $\xi_i = \tau_i \rho_i$ and $\bar{\xi}_u = \sum_{i \in R} p_i \xi_i$. Since $\tau_i = E(I_i | y_i, \boldsymbol{x}_i)$ and $\rho_i = E(R_i | y_i, \boldsymbol{x}_i, I_i = 1)$, ξ_i is the probability that unit i is sampled and subsequently responds, given its outcome and covariate values. Recall that $p_i^{(r)} \propto p_i \tau_i \rho_i = p_i \xi_i$. Thus, $p_i^{(r)} = \bar{\xi}_u^{-1} p_i \xi_i$.

Denote by E_ξ the expectation with respect to the combined sampling and response distribution. Then, for the (random) respondents sample size, $E_\xi(r) = N \sum_{i \in R} p_i \tau_i \rho_i = N \sum_{i \in R} p_i \xi_i = N \bar{\xi}_u$. Thus, $r \approx N \bar{\xi}_u$, leading to the additional constraint $r = N \bar{\xi}_u$. Since $\sum_{i \in R} p_i = 1$, we have $1 = r(N \bar{\xi}_u)^{-1} = r(N \bar{\xi}_u)^{-1} \sum_{i \in R} p_i = (r/N) \sum_{i \in R} p_i^{(r)} \xi_i^{-1}$, or

$$\sum_{i \in R} p_i^{(r)} \left(1 - r/(N \tau_i \rho_i)\right) = \sum_{i \in R} p_i^{(r)} \left(1 - r/(N \xi_i)\right) = \boldsymbol{0}. \tag{3.3}$$

Note that this constraint is equivalent to $\sum_{i \in R} p_i^{(r)} \tau_i^{-1} \rho_i^{-1} = N/r$ (using $\sum_{i \in R} p_i^{(r)} = 1$).

Let C be the $r \times d$ matrix, the ith row of which being $\boldsymbol{c}'_i - \bar{\boldsymbol{c}}'_u$, and $D(\boldsymbol{\gamma})$ be the $r \times r$ diagonal matrix with $\{\tau_i \rho_i = \tau_i \rho(y_i, \boldsymbol{x}_i; \boldsymbol{\gamma})\}$ as its diagonal elements. The constraints (3.2) and (3.3) can be written in matrix form as $C' D^{-1}(\boldsymbol{\gamma}) \boldsymbol{q} = \boldsymbol{0}$ and $\boldsymbol{1}' D^{-1}(\boldsymbol{\gamma}) \boldsymbol{q} = N/r$, respectively, where we denote $\boldsymbol{q} = \boldsymbol{p}^{(r)}$ for convenience. While the constraints can be defined using all calibration variables $\boldsymbol{c}_i$, our experience shows that some of the calibration variables are more vital than others. This issue is discussed and illustrated in more detail in sections 3.5 and 7.5.

Notice that our proposed approach is somewhat similar to the method developed by Qin et al. (2002), where the authors factorize the joint distribution of $(Y_i, \boldsymbol{x}_i, R_i)$ into a parametric model for the response probability $\Pr(R_i = 1 \mid y_i, \boldsymbol{x}_i)$ and a nonparametric model for the joint distribution $(Y_i, \boldsymbol{x}_i)$, yielding the empirical likelihood,

$$L = \prod_{i=1}^{r} \Pr(R_i = 1 \mid y_i, \boldsymbol{x}_i; i \in S; \boldsymbol{\gamma}) p_i \lambda^{n-r}, \tag{3.4}$$

where $\lambda = \Pr(R_i = 1, i \in S)$.

The proposed empirical likelihood is maximized with respect to p_i, λ, and $\boldsymbol{\gamma}$ under the constraints,

$$\begin{aligned}&\sum_{i=1}^{r} p_i[\Pr(R_i = 1|y_i, \boldsymbol{x}_i, i \in S; \boldsymbol{\gamma}) - \lambda] = 0,\\ &\sum_{i=1}^{r} p_i(\boldsymbol{c}_i - \bar{\boldsymbol{c}}_u) = \mathbf{0}; p_i \geq 0, \ \sum_{i=1}^{r} p_i = 1.\end{aligned} \tag{3.5}$$

The authors extend the empirical likelihood (3.4) to the case where the covariates are observed for both the respondents and nonrespondents, but in this case, maximizing the likelihood is almost impossible except in some special cases. Notice that this approach does not account for an informative sample selection.

Another approach based on the empirical likelihood with constraints was proposed by Morikawa et al. (2023). The empirical population-level likelihood, $L = \prod_{i=1}^{N} p_i$ is maximized under the constraints $\sum_{i=1}^{N} p_i = 1$, $\sum_{i=1}^{N} p_i I_i w_i D_\theta(R_i, X_i, Y_i, Z_i, w_i) = 0$, and $\sum_{i=1}^{N} p_i(1 - I_i w_i) C_\theta(X_i) = 0$, where N is the population size, I_i and w_i, $i = 1, \ldots, N$ are the sampling indicators and sampling weights respectively, and D_θ and C_θ with unknown parameter $\boldsymbol{\theta}$ that characterize the relationship between X and Y, are some efficient score functions defined by the authors. The authors distinguish between the case where the $\boldsymbol{x}$-values are only known for the sampled units, and the case where they are known for all the population units. The first case includes two different settings: (i) population-level summary statistics of $\boldsymbol{x}$-variables, such as means and correlations are known and (ii) the summary statistics are unknown. However, although this approach allows adjusting for both sampling and nonresponse, nonignorable nonresponse mechanism is not considered. Also, the authors do not consider the case where the covariates are only known for the responding units.

A similar idea of maximizing the empirical likelihood under constraints that incorporate auxiliary information in the context of analyzing complex survey data was considered by Chen and Kim (2014), Zhao et al. (2022) and Kim and Morikawa (2023). However, the first two approaches do not address adjustment for nonresponse, while the last approach assumes that the $\boldsymbol{x}$-values are known for all population units.

3.3 Maximization of the Respondents Empirical Likelihood

By section 3.2, we now have the constrained maximization problem

$$\max_{\boldsymbol{q},\boldsymbol{\gamma}} \Pi(\boldsymbol{q}) \quad \text{s.t.} \quad \begin{pmatrix} A(\boldsymbol{\gamma}) \\ b(\boldsymbol{\gamma}) \end{pmatrix} \boldsymbol{q} = \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \quad \boldsymbol{q} \in \Omega_{r-1}, \tag{3.6}$$

where $A(\boldsymbol{\gamma}) = C' D^{-1}(\boldsymbol{\gamma})$, $b(\boldsymbol{\gamma}) = (rN^{-1}\boldsymbol{\xi}(\boldsymbol{\gamma})^{-1} - 1)'$, $\boldsymbol{\xi}(\boldsymbol{\gamma})^{-1} = (\tau_1^{-1}\rho_1(\boldsymbol{\gamma})^{-1}, \ldots, \tau_r^{-1}\rho_r(\boldsymbol{\gamma})^{-1})'$, and Ω_{r-1} is the simplex of all nonnegative

vectors $(q_1,\ldots,q_r)' \in R^r$ with $\sum_i q_i = 1$. The MLE of $\boldsymbol{\gamma}$ and $\boldsymbol{q}$ are the values maximizing (3.6). The maximization problem can be solved in two ways.

3.3.1 Use of the profile likelihood of $\boldsymbol{\gamma}$.

The maximization problem in (3.6) is equivalent to $\max_{\gamma} G(\boldsymbol{\gamma})$, where $G(\boldsymbol{\gamma})$ is the *profile likelihood* of $\boldsymbol{\gamma}$, defined as

$$G(\boldsymbol{\gamma}) = \max_{\boldsymbol{q}} \left\{ \Pi(\boldsymbol{q}) : \begin{pmatrix} A(\boldsymbol{\gamma}) \\ \\ b(\boldsymbol{\gamma}) \end{pmatrix} \boldsymbol{q} = \mathbf{0}; \quad \boldsymbol{q} \in \Omega_{r-1} \right\}. \tag{3.7}$$

The maximization of (3.7) can be carried out using the R function `emplik`, written by Owen and available from his website http://statweb.stanford.edu/~owen/empirical/scel.R. See Owen (2013) for related theory and further details.

The question arises whether the maximum in (3.7) exists.

Lemma 2. Consider the constrained maximization problem,

$$\max_{\boldsymbol{q}} \{\Pi(\boldsymbol{q}) : A(\boldsymbol{\gamma})\boldsymbol{q} = \mathbf{0}; \quad \boldsymbol{q} \in \Omega_{r-1}\}. \tag{3.7*}$$

If the feasible region for (3.7*) is not empty for a given $\boldsymbol{\gamma}$, then it is not empty for any $\boldsymbol{\gamma}$. Furthermore, if the feasible region is not empty, then the maximum exists and is finite.

Proof. See the Appendix.

Remark 4. The constraints in the maximization problem (3.7*) do not contain the univariate constraint $b(\boldsymbol{\gamma})\boldsymbol{q} = 0$ (equation (3.3)), contained in the maximization constraints (3.7). The reason for this is that the constraint (3.3) can lead to an empty feasible region in (3.7) for certain vectors $\boldsymbol{\gamma}$. This is the case, for example, if the ρ_i's are very small, because if $\rho_i(\boldsymbol{\gamma}) < r/N\tau_i$ for all i, $\sum_{i\in R} p_i^{(r)} \tau_i^{-1} \rho_i(\boldsymbol{\gamma})^{-1} > N/r$.

The feasible region for the maximization problem (3.7*) may also be empty and therefore no solution exists. A simple example is where all the observed values of a constraining variable c are greater (or smaller) than its known population mean. Moreover, a combination of multivariate constraints can also preclude a solution. For example, when the sum of 2 variables used in the constraints is greater for all the responding units than the sum of the corresponding population means.

The maximum of $G(\boldsymbol{\gamma})$ in (3.7) under the constraints, can be obtained by using Lagrange multipliers. Let $g(\boldsymbol{c}_i,\boldsymbol{\gamma}) = (g_1(\boldsymbol{c}_i,\boldsymbol{\gamma}),\ldots,g_{d+1}(\boldsymbol{c}_i,\boldsymbol{\gamma}))$, with $(g_k(\boldsymbol{c}_i,\boldsymbol{\gamma}) = \tau_i^{-1}\rho_i^{-1}(\boldsymbol{c}_{ik} - \bar{\boldsymbol{c}}_{Uk}), \quad k = 1,\ldots,d, i = 1,\ldots,r, \quad g_{d+1}(\boldsymbol{c}_i,\boldsymbol{\gamma}) = \big(1 - r/(N\tau_i\rho_i)\big)$, where d is the dimension of $\boldsymbol{c}_i$. Then, the constraints defined

in (3.2) and (3.3) can be rewritten as $\sum_{i\in R} p_i^{(r)} g(\boldsymbol{c}_i, \boldsymbol{\gamma}) = \mathbf{0}$. Note that $\dim(g(\boldsymbol{c}_i, \boldsymbol{\gamma}))$ is not necessarily equal to $\dim(\boldsymbol{c}_i)$. In our study, we use the additional constraint defined in (3.3). Following Qin and Lawless (1994), profiling for all the values of $p_i^{(r)}$ results in

$$p_i^{(r)} = \frac{1}{r(1+\boldsymbol{\lambda}^t g(\boldsymbol{c}_i, \boldsymbol{\gamma}))}, \tag{3.8}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d+1})^t$ are the Lagrange multipliers. Furthermore, the maximum empirical likelihood estimate of $\boldsymbol{\gamma}$ is derived by maximizing the empirical likelihood $L = \prod_{i=1}^{r} p_i^{(r)}$, which can be rewritten as

$$L_E = \prod_{i=1}^{r} \frac{1}{r(1+\boldsymbol{\lambda}^t g(\boldsymbol{c}_i, \boldsymbol{\gamma}))}.$$

Then the empirical log-likelihood is obtained as

$$l_E = -\sum_{i=1}^{r} \log\left(1+\boldsymbol{\lambda}^t g(\boldsymbol{c}_i, \boldsymbol{\gamma})\right). \tag{3.9}$$

Obviously, maximizing (3.9) with $p_i^{(r)}$ obtained from (3.8) is equivalent to maximizing (3.6).

The asymptotic properties of the estimators resulting from (3.8) and (3.9) can be established by applying the theory developed in Qin and Lawless (1994). In particular, it follows that under some regularity conditions and for known sampling probabilities $\tau_i = \Pr(I_i = 1 | y_i, \boldsymbol{x}_i)$ (see Remark 5 below), the estimators defined by (3.8) and (3.9) are consistent and have a normal asymptotic distribution. Moreover, for fixed $\boldsymbol{\gamma}$ parameters, there exists a unique maximum for (3.6), provided that $\mathbf{0}$ is inside the convex hull of the points $g(\boldsymbol{c}_1, \boldsymbol{\gamma}), \ldots, g(\boldsymbol{c}_r, \boldsymbol{\gamma})$. This implies that plugging the estimates for $\boldsymbol{\gamma}$ obtained by using the Chang and Kott (2008) method described below into (3.6) results in a unique solution for $p_i^{(r)}$ if the aforementioned condition holds.

Remark 5. The asymptotic properties of the REL estimators defined by (3.8) and (3.9) outlined above assume known sampling probabilities $\tau_i = \Pr(I_i = 1 | y_i, \boldsymbol{x}_i)$. In practice, these probabilities are unknown and we estimate them outside the likelihood by use of kernel regression. The resulting estimators are then plugged into the REL and the calibration equations. As well known, these estimators converge at a slower rate than $n^{-\frac{1}{2}}$, see, for example, Stone (1982). The asymptotic properties of the REL estimators following this procedure have not been considered in the literature, although the simulation results in section 7 seem to support the validity of the approach.

Remark 6. Maximization of $G(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ can be performed by optimization routines available in most statistical packages.

3.3.2 Estimation of the $\boldsymbol{\gamma}$ coefficients outside the likelihood.
In the present article, we consider also the approach proposed by Chang and Kott (2008). By this approach, the totals of K calibration variables, which may contain some or all of the covariates in the response model, are regressed against their Horvitz–Thompson (H–T) estimators, with the weights appearing in the H–T estimators defined by the inverse of the product of the sampling probabilities and the response probabilities under the model. Let $\boldsymbol{c_i}$ denote the values of the calibration variables for unit i. Chang and Kott (2008) estimate the unknown response model coefficients by setting the regression equations, $\boldsymbol{C}^{pop} = \sum_{i\in R} w_i \rho_i^{-1}(y_i, \boldsymbol{\nu_i}; \boldsymbol{\gamma})\boldsymbol{c_i} + \boldsymbol{\varepsilon}^*$, where $\boldsymbol{C}^{pop} = \sum_{j=1}^{N} \boldsymbol{c_j}$, $\boldsymbol{\nu_i}$ defines the values of the covariates included in the response model for unit i and $\boldsymbol{\varepsilon}^*$ is a vector of errors. Note that if the population size N is known, an additional equation can be obtained by setting $c_j = 1$, $\forall j$. The parameters $\boldsymbol{\gamma}$ are estimated by applying an iterative algorithm. The authors show that under certain assumptions, the algorithm has a unique solution, which is consistent for the response model parameters.

Remark 7. Chang and Kott (2008) do not assume a model for the outcome so that their approach is restricted to estimation of the response probabilities and hence estimation of finite population totals, but it cannot be used for imputation. See section 6 for imputation of the missing data under our proposed approach.

Remark 8. The maximization by use of the profile likelihood is neat, but it raises the question of model identifiability. Model identification is a fundamental problem for non-ignorable nonresponse data. We therefore present in our simulation study the results obtained by application of the second approach of estimating the $\boldsymbol{\gamma}$ coefficients outside the likelihood, which turned out to yield similar results to the results obtained by maximization via the profile likelihood.

3.4 Another Approach Proposed in the Literature for Estimating the Response Probabilities

Sverchkov (2008) proposes another procedure for estimating the response probabilities. Suppose first that the missing values were actually observed. Then, using previous notation, the response probabilities could be estimated by solving the likelihood equations

$$\sum_{i\in R} \frac{\partial \log \rho(\boldsymbol{x}_i, y_i; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} + \sum_{k\in R^c} \frac{\partial \log[1 - \rho(\boldsymbol{x}_k, y_k; \boldsymbol{\gamma})]}{\partial \boldsymbol{\gamma}} = \mathbf{0}, \tag{3.10}$$

where R^c consists of the sample units with missing outcomes. (It is assumed that the covariates are known for the nonrespondents). In practice, however,

the outcomes are unknown for the nonresponding units, and so by application of the missing information principle (Orchard and Woodbury 1972), the equations (3.10) are replaced by their conditional expectation given the observed data, that is, by solving

$$E\left\{\frac{\sum_{i\in R}\partial\log\rho(\mathbf{x}_i,y_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}+\frac{\sum_{k\in R^c}\partial\log[1-\rho(\mathbf{x}_k,y_k;\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}\middle|O\right\}$$
$$=\frac{\sum_{i\in R}\partial\log\rho(\mathbf{x}_i,y_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}+E\left\{\frac{\sum_{k\in R^c}\partial\log[1-\rho(\mathbf{x}_k,y_k;\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}\middle|O\right\}=\mathbf{0}, \tag{3.11}$$

where O defines the observed data and the expectation in the second row is taken with respect to the model holding for the missing outcomes of the nonrespondents. The latter model is expressed as a function of the models holding for the observed outcomes and for the response probabilities. Notice that the derivation of the conditional expectation in (3.11) does not require a specification of a parametric model for $f_u(y\mid\boldsymbol{x})$.

When the nonrespondents' covariates are unobserved, the expectation in (3.11) can be derived by using the PDF of $\boldsymbol{v}_i=(y_i,\boldsymbol{x}_i^{'})$ given $(R_i=0,i\in S)$. Under the assumption that the population distribution has its support in the set of the observed values (section 3.1), we obtain that $\forall k\in R^c$

$$E\left\{\frac{\partial\log[1-\rho(\boldsymbol{x}_k,y_k;\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}\mid(R_k=0,k\in S)\right\}=\sum_{j\in R}p_j^{(nr)}\frac{\partial\log[1-\rho(\boldsymbol{x}_j,y_j;\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}, \tag{3.12}$$

where $p_j^{(nr)}=\Pr(\boldsymbol{v}_j|R_j=0,j\in S)=\left(\sum_{i\in R}\frac{p_i^{(r)}}{\rho_i}-1\right)^{-1}\frac{1-\rho_j}{\rho_j}p_j^{(r)}$ (see section 6). Thus, the expectation of interest in (3.11) takes the form

$$E\left\{\sum_{k\in R^c}\frac{\partial\log[1-\rho(\boldsymbol{x}_k,y_k;\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}|O\right\}=(n-r)\sum_{j\in R}p_j^{(nr)}\frac{\partial\log[1-\rho(\boldsymbol{x}_j,y_j;\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}. \tag{3.13}$$

It follows that the estimation equations (3.11) can be rewritten as

$$\sum_{i\in R}\left\{\frac{\partial\log\rho(\boldsymbol{x}_i,y_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}+(n-r)p_i^{(nr)}\frac{\partial\log[1-\rho(\boldsymbol{x}_i,y_i;\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}\right\}=\mathbf{0}. \tag{3.14}$$

Sverchkov and Pfeffermann (2018) and Pfeffermann and Sverchkov (2019) extend the approach to small area estimation under informative sampling and nonresponse.

3.5 Importance of the Constraints

The likelihood (3.6) is subject to calibration constraints. How important are these constraints and how should they be chosen? In our proposed empirical likelihood approach, the constraints (3.2) are of the form $\sum_{i\in R} p_i \boldsymbol{c}_i = \bar{\boldsymbol{c}}_{\boldsymbol{u}}$, which are seemingly unrelated to the response probabilities, suggesting that it should not matter which survey variables are used for defining the constraints. This, however, is a false conclusion since the empirical likelihood is defined with respect to the probabilities $p_i^{(r)}(\boldsymbol{\gamma}) = p_i \tau_i \rho_i(\boldsymbol{\gamma}) / \sum_k p_k \tau_k \rho_k(\boldsymbol{\gamma})$, such that any constraint on the p_i's effectively defines a constraint on the ρ_i's, implying that the variables in $\boldsymbol{c}$ should be correlated as highly as possible with y and $\boldsymbol{x}$. See section 7.5 for an empirical study of the importance of the constraints.

4. ESTIMATION OF PARAMETRIC MODELS AND VARIANCE OF ESTIMATORS

4.1 Estimation of the Population Multinomial Probabilities

The main, or intermediate, target of the inference process is the estimation of the multinomial probabilities $\boldsymbol{p} = (p_1, \ldots, p_r)'$. Having estimated the vector $\boldsymbol{p}^{(r)} = (p_1^{(r)}, \ldots, p_r^{(r)})'$ (section 3.3), the probabilities in $\boldsymbol{p}$ are estimated as

$$\widehat{p}_i = \widehat{p}_i^{\,(r)} [\widehat{\tau}_i \rho_i(\widehat{\boldsymbol{\gamma}})]^{-1} / \sum_{k=1}^{r} \widehat{p}_k^{(r)} [\widehat{\tau}_k \rho_k(\widehat{\boldsymbol{\gamma}})]^{-1}. \tag{4.1}$$

There is often interest in estimating a parametric population model $f_u(y|\boldsymbol{x})$. In the following section 4.2, we show how this can be done by use of the estimated multinomial probabilities (4.1), in the case where the form of the model is known and only the unknown model parameters need to be estimated. Recall, however, that our proposed approach does not require specification of a parametric model for $f_u(y|\boldsymbol{x})$.

4.2 Estimation of Parametric Models

We have assumed so far that the population distribution is multinomial with unknown probabilities $\boldsymbol{p}$, which are estimated by maximization of the REL or in conjunction with the procedure proposed by Chang and Kott (2008). Suppose, however, that the target population distribution is in fact parametric. Specifically, suppose that the population measurements $\{y_i, \boldsymbol{x}_i, i = 1, \ldots, N\}$ can be regarded as N independent realizations from some joint PDF $f_u(y_i, \boldsymbol{x}_i)$, with corresponding conditional PDFs $f_u(y_i|\boldsymbol{x}_i; \boldsymbol{\beta})\, i = 1, \ldots, N$, which are known up to the vector parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$. Then, under some general

conditions, the true vector of $\boldsymbol{\beta}$ is defined as the unique solution of the estimating equations,

$$W_U(\boldsymbol{\beta}) = N^{-1} \sum_{i=1}^{N} E_u[d_{Ui}] = \mathbf{0}, \tag{4.2}$$

where $d_{Ui} = (d_{Ui,0}, d_{Ui1,i}, \ldots, d_{Ui,k})' = \partial \log f_U(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is the i^{th} score function.

The 'census parameter' (Binder, 1983) corresponding to (4.2) is defined as the solution of the equations,

$$W_U(\boldsymbol{\beta}) = N^{-1} \sum_{i=1}^{N} d_{Ui} = \mathbf{0}. \tag{4.3}$$

Hence, under the present set-up, $\boldsymbol{\beta}$ can be defined as the solution of,

$$W_U(\boldsymbol{\beta}) = N^{-1} \sum_{i \in R} p_i d_{Ui} = \mathbf{0}. \tag{4.4}$$

Having estimated the probabilities $p_1, \ldots, p_r$, an estimate of $\boldsymbol{\beta}$ is obtained by solving (4.4), with $p_1, \ldots, p_r$ replaced by their estimates. See section 7.3 for illustration of the application of this estimation procedure.

Remark 9. An alternative way of estimating $\boldsymbol{\beta}$ is by solving the equations,

$$\widehat{W}_U(\boldsymbol{\beta}) = N^{-1} \sum_{i \in R} \frac{1}{\widehat{\tau}_i \widehat{\rho}_i} d_{Ui} = \mathbf{0}, \tag{4.5}$$

where (4.5) is the Horvitz–Thompson estimator of the population equations in (4.3), with estimated probabilities $\widehat{\tau}_i \widehat{\rho}_i$.

4.3 Variance Estimation

As mentioned in Remark 8, in our simulation study in section 7, we estimated the $\boldsymbol{\gamma}$ coefficients outside the likelihood. For estimating the variances of the $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ estimators, we use parametric bootstrap (BS). The parametric BS approach consists of generating B samples, with each sample consisting of r units independently drawn from the estimated model $\widehat{f}_R(\boldsymbol{u})$ fitted to the observed data for the responding units, where $\boldsymbol{u}$ stands for all the variables involved. In our case of the empirical likelihood, the fitted distribution is multinomial, with cell probabilities $\widehat{\boldsymbol{p}}^{(r)} = (\widehat{p}_1^{(r)}, \ldots, \widehat{p}_r^{(r)})'$. Therefore, we make r independent draws from R such that in each draw, the probability that unit i is selected is $\widehat{p}_i^{(r)}$. The estimation procedure is then applied to the data of each BS sample. Denote the B estimates of a parameter β by $\widehat{\beta}_1, \ldots, \widehat{\beta}_B$. The

parametric bootstrap estimate of the variance of $\widehat{\beta}$ is $B^{-1}\sum_{b=1}^{B}(\widehat{\beta}_b-\bar{\beta})^2$, where $\bar{\beta}=B^{-1}\sum_{b=1}^{B}\widehat{\beta}_b$. A similar procedure is used for estimating the variance of the $\boldsymbol{\gamma}$ estimators. See section 7.3 for illustration.

5. MODEL TESTING

A crucial question regarding any model fitting is testing the goodness of fit of the model. Contrary to a common perception that it is impossible to test a model assumed for the response probabilities, we contend that under the present approach, this is not true. Notice that we have observations from a model fitted to the responding units so that we are basically faced with the classical problem of testing the goodness of fit of a hypothesized model to the observed data. A common argument in favor of the claim that the model cannot be tested is that it may be the case that there is more than one combination of a population model and a sampling or response mechanism yielding the same respondents model, such that the respondents model is not identifiable or weakly identifiable. Pfeffermann and Landsman (2011) and Wang et al. (2014) establish conditions under which the sample model is identifiable, with references to other related studies. Moreover, in our case, we estimate the population model nonparametrically and the conditional sampling and response probabilities outside the likelihood, so that we are practically only testing the response model.

Pfeffermann and Landsman (2011) and Pfeffermann and Sikov (2011) applied several goodness-of-fit tests to test the model fitted to the observed data for the case where the outcome is continuous, see section 8.4. Below, we describe the application of the Hosmer and Lemeshow test statistic (1980, hereafter HL), for the case of a binary outcome y, which performed well in our simulation study. To construct this test, the sample is partitioned into G groups of approximately equal size, based on the estimated probabilities of "success" ($y=1$). The test statistic is defined as,

$$\mathrm{HL}=\sum_{k=1}^{G}\frac{(o_k-n_k\bar{\mu}_k)^2}{n_k\bar{\mu}_k(1-\bar{\mu}_k)}, \tag{5.1}$$

where o_k is the number of observed "successes" in group k, n_k is the size of the group and $\bar{\mu}_k$ is the mean of the estimated probabilities of success therein; $\bar{\mu}_k=\sum_{i\in G_k}\widehat{\mu}_i/n_k$, where $\widehat{\mu}_i=\widehat{\Pr}(y_i=1|I_i=1,R_i=1,\boldsymbol{x}_i)$.

By (2.4), for the case of a binary y, the estimated probability of success for unit $i\in R$ given $\boldsymbol{x}$ is,

$$\widehat{\mu}_i=\frac{\widehat{\Pr}_u(y=1|\boldsymbol{x})\widehat{\tau}(\boldsymbol{x},y=1)\widehat{\rho}(y=1,\boldsymbol{x})}{\widehat{\Pr}_u(y=1|\boldsymbol{x})\widehat{\tau}(\boldsymbol{x},y=1)\widehat{\rho}(y=1,\boldsymbol{x})+\widehat{\Pr}_u(y=0|\boldsymbol{x})\widehat{\tau}(\boldsymbol{x},y=0)\widehat{\rho}(y=0,\boldsymbol{x})}. \tag{5.2}$$

We estimate $\Pr_u(y|\boldsymbol{x})$ by applying a smooth cubic spline to the observed values $\{y_i, \boldsymbol{x}_i\}_{i\in R}$, with the estimated population multinomial probabilities $\widehat{p_1}^{(u)}, \ldots, \widehat{p_r}^{(u)}$ as weights, restricting the estimate to the [0, 1] interval; that is, $\widehat{\Pr}_u(y|\boldsymbol{x}) = \min\{1, \max\{0, \text{pred}(\Pr_u(y|\boldsymbol{x})\}\}$, where $\text{pred}(\Pr_u(y|\boldsymbol{x}))$ is the predicted value by the cubic spline. For estimating $\tau_i = E_u(I_i|y_i, \boldsymbol{x}_i)$, we use kernel smoothing by applying kernel regression of w_i on $(y_i, \boldsymbol{x}_i)$. Estimates of $\rho(y_i, \boldsymbol{x}_i)$ are obtained from the estimated response model.

HL found through an empirical study under a simpler set-up that their test statistic follows approximately a $\chi^2_{(G-2)}$ distribution under the null hypothesis that the model fits the data. We verify this conjecture in our simulation study in section 7.6.

6. IMPUTATION OF NONRESPONDENTS DATA

In this section, we propose methods for imputation of the nonrespondents data, depending on whether the auxiliary variables $\boldsymbol{x}$ are known for the nonrespondents or not. The goal is to impute an observation for each unit in the nonrespondents set R^c in such a way that the distribution of the variables in the combined sample, $R \cup R^c$ will be as close as possible to the distribution of the same variables in the original sample, in the case of full response.

Let $\rho_i = \Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1; \boldsymbol{\gamma})$ and $\boldsymbol{v_i} = (y_i, \boldsymbol{x}_i')$. The PDF of $\boldsymbol{v}_i$ for the responding units is given by $\Pr(\boldsymbol{v_i}|R_i = 1, i \in S) = \sum_{j\in R, \boldsymbol{v}_j = \boldsymbol{v}_i} p_j^{(r)} = p^{(r)}(\boldsymbol{v_i})$, $i = 1, \ldots, r', r' \le r$, where $p_j^{(r)}$ is defined in section 3.1. By Bayes' law and following Pfeffermann and Sikov (2011),

$$\begin{aligned}\Pr(\boldsymbol{v_i}|R_i = 0, i \in S) &= \frac{\Pr(R_i = 0|\boldsymbol{v_i}, i \in S)}{\Pr(R_i = 0|i \in S)}\Pr(\boldsymbol{v_i}|i \in S) \\ &= \frac{\Pr(R_i = 0|\boldsymbol{v_i}, i \in S)\Pr(\boldsymbol{v_i}|R_i = 1, i \in S)\Pr(R_i = 1|i \in S))}{\Pr(R_i = 0|i \in S)\Pr(R_i = 1|\boldsymbol{v_i}, i \in S)},\end{aligned} \tag{6.1}$$

where the second row follows from the relationship,

$$\Pr(\boldsymbol{v_i} \mid i \in S) = \Pr(\boldsymbol{v_i} \mid R_i = 1, i \in S)\frac{\Pr\ (R_i = 1 \mid i \in S)}{\Pr(R_i = 1 \mid \boldsymbol{v_i}, i \in S)}.$$

Let $z = x/(1-x)$ where $x = \Pr(R_i = 1|i \in S)$. Since the population distribution has its support in the set of the observed values, it follows that, $\sum_{i\in R}\Pr(\boldsymbol{v_i}|R_i = 0, i \in S) = 1$. Then, by (6.1)

$$\begin{aligned}
&\textstyle\sum_{i\in R}\Pr(\boldsymbol{v_i}|R_i=0,i\in S)=z\sum_{i\in R}\left(\frac{1-\rho_i}{\rho_i}p^r(\boldsymbol{v_i})\right)\\
&=z\textstyle\sum_{i\in R}\left(\frac{p^r(\boldsymbol{v_i})}{\rho_i}-p^r(\boldsymbol{v_i})\right)=z\left(\sum_{i\in R}\frac{p^r(\boldsymbol{v_i})}{\rho_i}-1\right)=1\\
&\Rightarrow\textstyle\sum_{i\in R}\frac{p^r(\boldsymbol{v_i})}{\rho_i}=1/z+1=1/x\\
&\Rightarrow\Pr(R_i=1|i\in S)=\left(\textstyle\sum_{i\in R}(p^r(\boldsymbol{v_i})/\rho_i)\right)^{-1}.
\end{aligned}\tag{6.2}$$

Let $p^{(nr)}(\boldsymbol{v_i})=\Pr(\boldsymbol{v_i}|R_i=0,i\in S)$ and denote $\widehat{p}^{(nr)}(\boldsymbol{v_i})$ its estimate by use of (6.1) and (6.2). Consider the following two scenarios:

SCENARIO 1: The covariates $\boldsymbol{x}_i$ are unknown for nonresponding units. By (6.1) and (6.2),

$$\widehat{p}^{(nr)}(\boldsymbol{v_i})=\left(\sum_{i\in R}\frac{\widehat{p}^{(r)}(\boldsymbol{v_i})}{\widehat{\rho}_i}-1\right)^{-1}\frac{1-\widehat{\rho}_i}{\widehat{\rho}_i}\widehat{p}^{(r)}(\boldsymbol{v_i}).\tag{6.3}$$

Thus, under Scenario 1, data for R^c can be imputed by drawing $(n-r)$ independent observations $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{(n-r)}$ from the Multinomial distribution with probabilities $p^{(nr)}(\boldsymbol{v_1}),\ldots,\widehat{p}^{(nr)}(\boldsymbol{v_{r'}})$ and define $(y_1,\boldsymbol{x}_1),\ldots,(y_{(n-r)},\boldsymbol{x}_{(n-r)})$ as the imputed data.

SCENARIO 2: The covariates $\boldsymbol{x}_i$ are known for both the responding and nonresponding units. In this case, one can independently draw y_i for each nonresponding unit i from the estimated model $\Pr(y_i|\boldsymbol{x}_i,R_i=0,i\in S)$,

$$\begin{aligned}
\Pr_{nr}(y_i|\boldsymbol{x}_i)&=\Pr(y_i|\boldsymbol{x}_i,R_i=0,i\in S)\\
&=\frac{\Pr(R_i=0|y_i,\boldsymbol{x}_i,i\in S)\Pr(y_i|\boldsymbol{x}_i,i\in S)}{\int\Pr(R_i=0|y_i,\boldsymbol{x}_i,i\in S)\Pr(y_i|\boldsymbol{x}_i,i\in S)dy_i},
\end{aligned}\tag{6.4}$$

where $\Pr(y_i|\boldsymbol{x}_i,i\in S)=\dfrac{\Pr(\boldsymbol{v}_i|i\in S)}{\Pr(\boldsymbol{x}_i|i\in S)}$. The numerator is defined by the equation following (6.1) while the denominator is the sum of the numerator over all possible y values. In particular, when y is binary, the denominator of (6.4) equals $\Pr(R_i=0\mid y_i=1,\boldsymbol{x}_i,i\in S)\Pr(y_i=1\mid\boldsymbol{x}_i,i\in S)+\Pr(R_i=0\mid y_i=0,\boldsymbol{x}_i,i\in S)\Pr(y_i=0\mid\boldsymbol{x}_i,i\in S)$.

Notice that the information contained in the nonrespondents' covariates is not incorporated in our estimation procedures, but it can be utilized for imputing the missing outcomes. Including this information into the estimation process generally results in considerable complications of the likelihood maximization. See, for example, Qin et al. (2002).

In our simulation study in section 7.4, we assume Scenario 1.

7. SIMULATION STUDY

7.1 Simulation Set-up

In order to examine the performance of our proposed approach, we conducted a simulation study as follows. A population of values $x_j, j = 1, \ldots, 10,000$ was generated from Gamma$(2,2)$. For each value x_j, a binary outcome y_j was generated with $\Pr(y_j = 1|x_j;\boldsymbol{\beta}) = \text{logit}^{-1}(-0.8 + 0.8x_j)$. Next, values of a design variable Z were generated as $z_j = \max[(x_j + 1.1)(2y_j + 1) + \nu_j, 0.01]$, where $\nu_j \sim \text{Uniform}(-0.2, 0.2)$. Values of 6 calibration variables $\boldsymbol{c}$ were generated as $\boldsymbol{c}_j = (1, x_j, y_j, x_j y_j, x_j^2, x_j^2 y_j)' + \boldsymbol{\varepsilon}_j$, with $\boldsymbol{\varepsilon}_j$ independently drawn from $N(\mathbf{0}_6, \sigma_c^2 I_6)$, ($\mathbf{0}_m$ and I_m are respectively the m-dimensional zero vector and the $m \times m$ identity matrix). While different values of σ_c^2 are considered in section 7.5, in the rest of this section, we use $\sigma_c^2 = 1$. A sample was drawn by Bernoulli sampling ($I_j \sim^{\text{indep}} \text{Ber}(\pi_j)$), where $\pi_j = \min(3500 z_j^{-1} / \sum_{k=1}^{10000} z_k^{-1}, 0.9999)$ and $E(n) = 3500$ is the expected sample size. The sampled units were classified as respondents/nonrespondents with $\Pr(R_j = 1 \mid y_j, x_j, j \in S) = \rho_j = \text{logit}^{-1}(\gamma_0 + \gamma_x x_j + \gamma_y y_j)$; $\gamma_0 = 0.7, \gamma_x = 0.5, \gamma_y = -1.5$.

Remark 10. The sampling process and the response are both informative since they depend on the outcome for given x.

The process of generating the population y-values and selecting the sample and the respondents was repeated independently 300 times. (The population x values were generated only once). The ranges of the sample size n and the number of respondents r are $3395 \le n \le 3625$, $2227 \le r \le 2455$. For each sample of respondents, we estimated the vector coefficients $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ using the following procedure described in section 3. The response model parameters $\boldsymbol{\gamma}$ were estimated outside the likelihood by application of the Chang and Kott (2008, hereafter C–K) method. Next, the probabilities estimates $\widehat{\boldsymbol{p}}^{(r)}$ were derived by maximizing the likelihood defined by (3.7) with $\boldsymbol{\gamma}$ replaced by $\widehat{\boldsymbol{\gamma}}$, obtained at the previous step. The same calibration variables were utilized in both steps. The estimates of the regression parameters $\boldsymbol{\beta}$ were derived using the equations (4.4), with estimated probabilities $\widehat{\boldsymbol{p}}$. The variance of the estimators had been estimated using the bootstrap procedure described in section 4.3.

7.2 Estimation of the Conditional Expectation of the Sampling Weights

We used kernel smoothing to obtain estimates of $E_s(w_i|y_i;x_i) = E_R(w_i|y_i;x_i)$ (see Remark 1), by applying kernel regression of w_i on (y_i, x_i). Here, the subscript R refers to the respondents distribution. Since in our case y_i attains only the values 0 or 1, the smoothing was applied to estimate $E_R(w_i|y_i = 0;x_i)$ and $E_R(w_i|y_i = 1;x_i)$ separately. The kernel regression was performed using the function `npreg` from the R package `np` at its default setting. Specifically,

Nadaraya–Watson (Nadaraya 1964, Watson 1964) kernel smoothing was performed, with a bandwidth calculated using the method of Racine and Li (2004).

7.3 Empirical Results—Estimation of Model Parameters

Table 1 shows the mean estimates of the response model coefficients ($\boldsymbol{\gamma}$) and their empirical standard error (S.E.) over the 300 samples. Also shown are the square roots of the mean variance estimates, obtained by application of the parametric bootstrap procedure described in Section 4.3. The results in table 1 illustrate good performance of the point and variance estimators under the proposed estimation method.

Table 2 compares 5 estimators of the logistic population model coefficients $\boldsymbol{\beta}$, used to generate the population outcomes: (i) the unweighted standard estimates based on the full sample data (observed and missing), but ignoring the sampling process, (ii) estimates that use the full sample data but account for the informative sampling by use of the sampling weights, (iii) estimates that use the respondents data only, ignoring the sampling and response processes, (iv) estimates that use the respondents data only and account for the sampling process by use of the sampling weights, but ignore the response process, (v) estimates obtained by our proposed C–K and REL procedure. For the application of the REL procedure, the $\boldsymbol{\beta}$ coefficients were estimated by solving the estimating equations $\sum_{i=1}^{r} \widehat{p}_i(y_i - \text{logit}(\beta_0 + \beta_1 x_i))(1, x_i)' = \mathbf{0}$, where $\widehat{p}_1, \ldots, \widehat{p}_r$ are the estimated population multinomial probabilities (see section 4.2).

As can be seen, the estimators derived by application of our proposed method are virtually unbiased in estimating the population logistic model coefficients although with the largest S.E., which can be explained by the complexity of the procedure and in particular, the estimation of the response model. The bootstrap variance estimators also perform well. The FR PW estimators are also unbiased and with the smallest variance, but they use the nonrespondents data, which are not available in practice. The other estimators, which

Table 1. Estimation of γ. Mean Estimates, Empirical Standard Errors (S.E.), and Square Root of Mean Variance Estimates, Based on the Parametric Bootstrap Procedure.

	γ_0	γ_x	γ_y
True values	0.700	0.500	– 1.500
Mean estimates	0.723	0.490	– 1.511
Empirical S.E.	0.225	0.203	0.332
Sqrt mean BS variance	0.231	0.224	0.338

Table 2. Estimation of β: Mean Estimates, Empirical Standard Errors (S.E.) and Square Root of Mean Bootstrap Variance Estimates. True Coefficients $\beta_0 = -0.8, \beta_1 = 0.8$.

Method	Mean estimates		Square root mean BS variance		Empirical S.E.	
	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
FR UW	− 1.902	0.802	0.073	0.071	0.073	0.071
FR PW	− 0.798	0.799	0.075	0.072	0.073	0.072
IGR UW	− 2.665	0.966	0.111	0.095	0.105	0.093
IGR PW	− 1.559	0.962	0.113	0.097	0.106	0.093
C-K and REL	− 0.802	0.800	0.186	0.110	0.177	0.104

FR, Full response, estimators obtained from all the sample data; IGR, estimators obtained when ignoring the response mechanism; PW, Probability weighted by use of the sampling weights; UW, Unweighted; C-K and REL, our proposed method.

ignore the informative sampling and/or the NMAR nonresponse are seen to be highly biased, particularly in estimating β_0.

7.4 Imputation of Nonrespondents Data

In a simulation study, the missing values are known, allowing us to compare the distribution of measurements corresponding to the nonrespondents to that of their imputed values.

We illustrate the performance of our proposed imputation procedure (section 6) for the case where the covariate values are unknown for the nonresponding units. For this, we generated 300 samples in the same manner as described in section 7.1, with calibration variance $\sigma_c^2 = 1$. Following the imputation procedure described in section 6, we find that the average percentage of units with $y = 1$ is 22.3 percent in the full samples and 22.4 percent in the combined respondents and imputed data. The average percentage of units with $y = 1$ is 38.7 percent in the true nonrespondents data, compared with 39.1 percent in the imputed data.

In figure 1, we plot the empirical cumulative distribution functions (eCDF) of the imputed x-values given y, separately for $y = 0$ and $y = 1$, and compare them to the corresponding eCDF of the true values. Averaged over the 300 samples, the curves are practically identical. Therefore, we compare in figure 1 the averaged eCDFs over just the first 5 samples.

Averaging over the 5 samples, 0.387 of the nonrespondents have a value of $y = 1$, compared to 0.369 in the imputed data. Denote by $eCDF_{imp}$ and $eCDF_{nr}$ the cumulative distributions of the imputed data and of the true

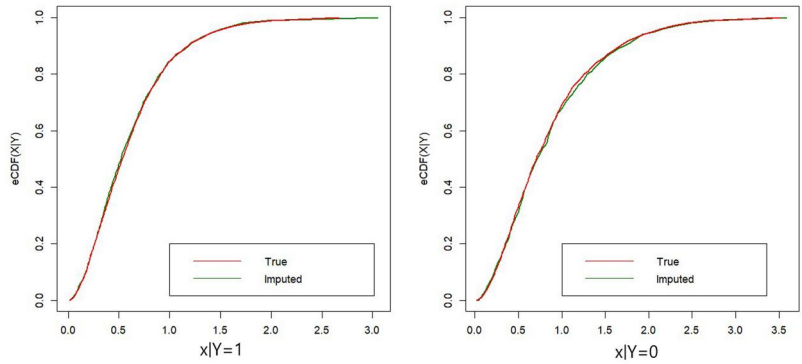

Figure 1. Cumulative Distributions of x given y in the Nonrespondents Sample and of the Imputed Values, Averaged Over 5 Samples. x-values considered missing for nonrespondents.

nonrespondents data, respectively. It appears that $eCDF_{imp}(y) \approx eCDF_{nr}(y)$ and $eCDF_{imp}(x|y) \approx eCDF_{nr}(x|y)$. Thus, we conclude that the proposed imputation procedure performs well.

7.5 The Role of the Constraints

An important element of our proposed method is the use of known population means of a vector of calibration values $\boldsymbol{c}_i$. As discussed in section 3.5, ideally, the components of $\boldsymbol{c}_i$ should be highly correlated with the model variables. In this section, we illustrate the effect of the choice of the variables $\boldsymbol{c}_i$ and their proximity to the model variables on the estimation of the model parameters. For this, we created a 6-dimensional variable $\boldsymbol{c} = (c_0, c_1, c_2, c_3, c_4, c_5)'$, where $\boldsymbol{c}_i = (1, x_i, y_i, x_i y_i, x_i^2, x_i^2 y_i)' + \boldsymbol{\varepsilon}_i$, with the $\boldsymbol{\varepsilon}_i$ independently generated from a multivariate normal distribution $N(\mathbf{0}_6, \sigma_c^2 I_6)$. Table 3 demonstrates the dependence of the accuracy of the model parameter estimates on the closeness of the calibration variables to the model variables, by considering different subsets of $c_0, c_1, c_2, c_3, c_4, c_5$ with different values of σ_c, along with the r-constraint (3.3).

The results in table 3 demonstrate the importance of choosing the "right", sufficiently accurate constraints. In particular, using the pair c_2 and c_3 alone performs well, even with $\sigma_c = 1$, whereas the combination of c_0, c_1 and c_4 (not shown) results in lack of convergence of the estimation algorithm when estimating the parameters $\boldsymbol{\gamma}$. This is explained by the fact that it is the dependence of y on x that matters and in our case, c_2 and c_3 represent y and xy, respectively. Thus, to estimate the population model well, calibration variables should be chosen so that the conditional distribution of y given x is accounted for. As the case c_2, c_3 shows, even when only 2 constraints are used (along with the r-constraint (3.3)), the estimates are good (although generally less precise than with more constraints). The use of c_1, c_2 also provides acceptable

Table 3. Effects of the Choice of the Constraints and Their Accuracy. Mean Estimates Based on 300 Samples in Each Case

		β_0	β_1	γ_0	γ_x	γ_y
True coefficients		−0.800	0.800	0.700	0.500	−1.500
	Constrains used					
$\sigma_c = 0.5$	$c_0, c_1, c_2, c_3, c_4, c_5$	−0.807	0.800	0.713	0.489	−1.507
$\sigma_c = 0.5$	c_1, c_2, c_3, c_4, c_5	−0.806	0.806	0.717	0.494	−1.509
$\sigma_c = 0.5$	c_2, c_3	−0.807	0.805	0.716	0.500	−1.516
$\sigma_c = 0.5$	c_2, c_3, c_5	−0.807	0.804	0.717	0.499	−1.509
$\sigma_c = 0.5$	c_1, c_2	−0.806	0.805	0.726	0.503	−1.525
$\sigma_c = 1.0$	$c_0, c_1, c_2, c_3, c_4, c_5$	−0.803	0.800	0.723	0.490	−1.511
$\sigma_c = 1.0$	c_1, c_2, c_3, c_4, c_5	−0.808	0.802	0.722	0.493	−1.516
$\sigma_c = 1.0$	c_2, c_3	−0.812	0.779	0.733	0.504	−1.526
$\sigma_c = 1.0$	c_2, c_3, c_5	−0.810	0.795	0.734	0.504	−1.523
$\sigma_c = 1.0$	c_1, c_2	−0.817	0.796	0.746	0.491	−1.552

estimates, although somewhat less accurate than with c_2, c_3, especially when estimating the response model parameters. We also computed the empirical standard errors of the estimates over the 300 samples and as expected, the more accurate are the constraints (smaller σ_c), the smaller are the empirical standard errors of the parameter estimates.

7.6 Model Testing

In order to illustrate the distribution of the Hosmer–Lemeshow test statistic (5.1) under the population model, sampling design and response mechanism defined in section 7.1, we show in figure 2 its empirical distribution with $G = 10$ nearly equal groups. Recall that if $x_1, \ldots, x_n$ are independent draws from a χ^2_d distribution, the log-likelihood of d is $\frac{d}{2}\sum \log x_i - \frac{nd}{2}\log 2 - \log \Gamma\left(\frac{d}{2}\right) + H(x_1, \ldots, x_n)$, where Γ is the Gamma function and $H(\boldsymbol{x})$ is a function of $x_1, \ldots, x_n$ alone. Hosmer and Lemeshow (1980, hereafter HL) conjectured that the distribution of the test statistic (5.1) under H_0 is χ^2 with $G - 2 = 8$ degrees of freedom (df). We estimated the degrees of freedom to be 8.0991, using maximum likelihood estimation based on 300 original samples. Figure 2 contains a histogram and QQ plot, comparing the observed quantiles to the quantiles of a χ^2_8 distribution. The 2 figures show a close approximation of the empirical distribution of the test statistic to the hypothesised χ^2_8 distribution, thus validating the conjecture of HL.

Remark 11. As an alternative to using the χ^2 distribution, one can employ the bootstrap approach, which could provide a more accurate approximation for the distribution of the HL statistic in the case where the respondents' sample size is not sufficiently large for justifying the asymptotic distribution. See also section 9 with the concluding remarks.

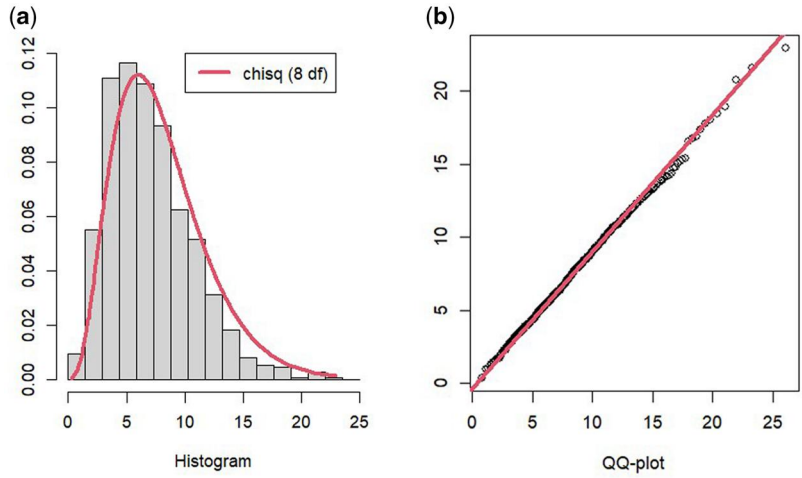

Figure 2. Empirical Distribution of the HL Test Statistic (5.1) ($G = 10$ Equal size Groups), Under the Population Model, Sampling Design and Response Model of Section 7.1. Comparison to χ^2_8 Distribution. (a) Histogram (b) QQ plot. Based on 300 simulated samples.

Remark 12. We also studied the power of the test statistic by computing the rejection rates of the null hypothesis that the response model is of the form $\rho_i = \text{logit}^{-1}(\gamma_0 + \gamma_x x_i + \gamma_y y_i)$, when in fact the true response model was of the form $\rho_i = \text{logit}^{-1}(\gamma_0 + \gamma_x x_i + \gamma_y y_i + a y (x_i - b)^2 + c)$, for several combinations of a, b and c, with $\gamma_0 = 0.7$, $\gamma_x = 0.5$, $\gamma_y = -1.5$. The population model was generated as $\Pr(y = 1|x) = \text{logit}^{-1}(-0.8 + 0.8x)$, same as in section 7.1. We assumed that the form of this model is known, but not the model parameters. We found that with significance level of $\alpha = 0.05$, as the dissimilarity between the correct model and the model assumed under the null hypothesis grows, so do in general the rejection rates, with some of them being very high.

Remark 13. We repeated the analysis with $G = 20$ groups. The results are generally similar to the results obtained with 10 groups, but the empirical rejection rates are always somewhat lower than with 10 groups.

8. APPLICATION TO HOUSEHOLD EXPENDITURE SURVEY IN ISRAEL

8.1 Study Population and Outcome Variable

In this section, we illustrate the performance of the REL approach by using data collected as part of the Household Expenditure Survey (HES), carried out

by Israel's Central Bureau of Statistics (ICBS) in 2019. (Data for later years were still under study when analysing the 2019 data because of COVID-19.) The survey collects information on socio-demographic characteristics of each member of the sampled HH, as well as detailed information on the HH income and expenditure. The HHs are generally sampled with equal probabilities. The total sample size in the 2019 survey was $n = 12{,}136$, with $r = 7{,}827$ responding HHs and $n - r = 4{,}309$ nonresponding HHs. Our target outcome variable in the present application is the *household gross monthly income*.

Remark 14. In this application, we maximized the REL using the profile likelihood of $\boldsymbol{\gamma}$ as described in section 3.3 (method 1). Estimating the $\boldsymbol{\gamma}$ coefficients outside the likelihood by use of Chang and Kott (2008) method turned out to be very sensitive to the calibration variables used, not converging with some of the variables, or yielding unreasonable estimates for some of the coefficients of the variables defining the response model.

Remark 15. Unlike the simulation study in section 7, where the outcome variable was binary, in this application, the outcome variable is continuous.

8.2 Model for Response Probabilities

As in our simulation study in section 7, we assume that the HH response probabilities given the outcome and the covariates can be modeled by the logistic function,

$$\rho_i = \Pr(R_i = 1 | z_i, \boldsymbol{x}_i) = [1 + \mathrm{e}^{-(\gamma_z z_i + \boldsymbol{\gamma}_x' \boldsymbol{x}_i)}]^{-1}, \tag{8.1}$$

where $z_i = \log y_i$ is the log(income) of HH i and $\boldsymbol{x}_i$ is the corresponding vector of covariates. Covariates considered are: (i) Socio-economic index (SEI) of the statistical area (SA) to which the HH belongs; a weighted average of socio-economic variables measured at the SA level. Israel is divided into more than 3,000 statistical areas and with a population of about 9 million people in 2019, each statistical area contained on average about 3,000 individuals; (ii) the size of the HH (HHsize) and (iii) characteristics of the head of the HH: gender, (Gen, 1 for male), Religion (Rel.Jew, 1 for Jew), age (Age) and Country of birth (Cob, 1 for native).

8.3 Calibration Constraints

For the REL application, we calibrated the observed data of the responding HH to the following population estimates, as obtained from the ICBS Labor force survey (LFS) in 2019. The LFS is a monthly rotating survey with a total of about 33,400 responding HH in 2019. With such a big sample size, the LFS estimates are quite accurate. See equation (3.2) for the form of the calibration constraints:

(1) Proportion of Jewish HH out of all the HH in the country.
(2) Proportion of HH with total number of working hours per day less than 9 hours.
(3) Proportion of HH with 1 or 2 members.

The HES also collects socio-economic and demographic data for all the individuals in the sampled HH. For pre-defined groups, we counted the number of individuals in each HH belonging to a given group and computed their expectation using the multinomial HH probabilities. Multiplying the expectation by the number of HH in the population and dividing by the number of individuals in the population belonging to the group defines the estimated expected proportion of individuals in the group, which we calibrated to the corresponding true population proportion, known from administrative registers. Specifically, we imposed the following additional calibration constrains:

(4) Proportions of individuals in 3 classes defined by "religiosity"; Jewish Ultra-Orthodox (UO), Non UO Jews, Arabs (2 constraints).
(5) Proportions of individuals in 17 income classes, as defined by 10 deciles in the Jewish population, 5 quintiles in the Arab population, and children and persons with no income (14 constraints).
(6) Proportions of individuals residing in 7 different geographical districts comprising the entire country (6 constraints).
(7) Proportions of children under the age of 15, and classification by gender for individuals aged 15 and over (2 constraints).

Finally, we also imposed the following two constraints:

(8) The constraint defined by (3.3).

(9) $\sum_{i\in R}(\widehat{\tau}_i\widehat{\rho}_i)^{-1} = N$ (number of HH in the population as obtained from the LFS).

In summary, we used a total of 29 calibration constraints.

8.4 Results

Table 4 presents the estimates of the response model coefficients and their estimated standard errors, as obtained by computing the square roots of the diagonal elements of the inverse profile information matrix. We only show the significant estimates and the (nonsignificant) constant term, based on the standard t-tests.

Remark 16. The standard errors could also be estimated by parametric bootstrap as done in the simulation study in section 7. However, since in this application we maximized the REL by using the profile likelihood (see Remark 14), we preferred to estimate them by use of the inverse profile information matrix.

Table 4. Estimates of Response Model Coefficients and Their Standard Errors (S.E.) (8.1)

Parameters	$\widehat{\boldsymbol{\gamma}}$	$S.E.(\widehat{\boldsymbol{\gamma}})$
Constant	0.338	1.095
SEI	0.592	0.0544
HH size	0.053	0.0015
Rel.Jew	0.141	0.0814
log (income)	−0.109	0.0297

As can be seen, the log(income) variable is highly significant, indicating that the nonresponse is informative (NMAR), given the covariates included in the model. The negative sign of the coefficient suggests that the higher the HH log(gross income), the less is the response propensity. The positive signs of the other coefficients can be reasoned as well.

Remark 17. The response model contains only a subset of the $\boldsymbol{x}$- variables mentioned above, included in the multinomial distribution, so that the model is identifiable.

Having estimated the response model coefficients and the multinomial probabilities $\{p_i\}$ by maximization of the REL with the calibration constraints, there are two ways of estimating the population mean of the gross HH income,

$$\widehat{Y}_1 = N^{-1}\sum_{i\in R}(\widehat{\tau}_i\widehat{\rho}_i)^{-1}y_i, \quad \widehat{Y}_2 = \sum_{i\in R}\widehat{p}_i y_i. \tag{8.2}$$

The first estimator is the Horvitz–Thompson estimator with estimated sampling weights, which account for the response probabilities, viewed as a 'second stage' of the sampling process. See equation (4.5) in section 4.2 and the last calibration constraint in section 8.3. The second estimator is based on the estimated population multinomial probabilities, assuming that the population distribution has its support in the set of the observed values. Under correct model specification, both estimators are consistent for the true population mean and we obtained $\widehat{Y}_1 = 19,886, \widehat{Y}_2 = 20,173$ (in Israeli shekels), which we consider to be sufficiently close, given the complexity of the analysis.

8.4.1 Can we test the model used?

Suppose first that the multinomial probabilities $\boldsymbol{p}^{(r)} = (p_1^{(r)},\ldots,p_r^{(r)})$ are known, where $p_i^{(r)} = \Pr((y_i,\boldsymbol{x}_i,\boldsymbol{c_i})|i\in R)$. Hence, the marginal probability of y_i is $\Pr(y_i) = \sum_{j\in R, y_j=y_i} p_j^{(r)} = p_i^{(r)*}$. (Recall that under the EL, the observations are discrete). The CDF is therefore, $U(y) = \sum_{i\in R, y_i\le y} p_i^{(r)*}$. Let $u_1,\ldots,u_r$ denote the values of $U_1,\ldots,U_r$ at the respondents' values $y_1,\ldots,y_r$ and denote by $u_{(1)},\ldots,u_{(r)}$ the ordered values of the u_i's. Denote by $F_{i,EMP}$ the

corresponding empirical CDFs. The following goodness of fit test statistics are in common use:

Kolmogorov –Smirnov:

$$KS = max_i|F_{i,EMP} - u_{(i)}|, \tag{8.3}$$

Cramer-von Misses:

$$CM = \frac{1}{12r} + \sum_{i=1}^{r}\left[u_{(i)} - \frac{2i-1}{2r}\right]^2, \tag{8.4}$$

Anderson-Darling:

$$AD = -r - \frac{1}{r}\sum_{i=1}^{r}[(2i-1)\,\ln(u_{(i)}) + (2r+1-2i)\,\ln(1-u_{(i)})]. \tag{8.5}$$

In practice, the multinomial probabilities $\boldsymbol{p}^{(r)}$ are unknown, and we replace them by their REL estimates. When computed with estimated probabilities, the asymptotic distributions of the three test statistics depend in a complex way on the true underlying CDF and possibly on the method of estimation. Sufficiently accurate critical values can be obtained in this case by use of parametric bootstrap, re-estimating the unknown model parameters for each bootstrap sample and then computing the corresponding test statistics. Babu and Rao (2004) show that the bootstrap distributions of the test statistics approximate the true distributions under the hypothesized model with correct order of error.

We generated 1,000 bootstrap samples with probabilities $p_i^{(r)}$ estimated from the original sample and obtained the following test statistics and p values, as computed from the corresponding bootstrap distributions: *KS* test 0.0497 (p value = 0.17); *CM* Test 0.0624 (p value = 0.25); *AD* test 0.0689 (p value = 0.31). It would seem that the relative high p-values for all the three tests support the use of the model fitted for this application.

Remark 18. The ICBS has applied the procedure developed by Sverchkov (2008, see section 3.4) for estimating the response probabilities in the 2019 HES survey. Next, the base sampling weights have been multiplied by the inverse of the estimated response probabilities, and the resulting weights have been calibrated using about 400 calibration constraints, yielding a calibrated "design-based" estimator, which accounts for NMAR nonresponse, similar in nature to the estimator $\widehat{Y}_1$ in equation (8.2). The value of the estimate is $\widehat{Y}_{\text{ICBS19}} = 19,542$, quite close to our "design-based" REL estimate, $\widehat{Y}_1 = 19,886$. We mention in this respect that the design-based estimator based on only the base sampling weights (no calibration, ignoring the nonresponse) equals $\widehat{Y}_{\text{BW}} = 21,480$. Thus, all the three estimators $\widehat{Y}_1$, $\widehat{Y}_2$ and

$\widehat{Y}_{\text{ICBS19}}$, which account for the NMAR nonresponce, reduce the weighted estimator, which ignores the nonresponse, quite significantly.

9. CONCLUDING REMARKS

We develop and illustrate a general approach for analysing complex survey data, subjected to informative sampling and NMAR nonresponse, with basically minimal assumptions. The only parametric model assumed is the model for the response probabilities but as illustrated, and contrary to common misconception, this model can be tested with good power.

The proposed approach is more robust and more stable than fully parametric alternatives. The results of the simulation study and the real application demonstrate good properties of the method, but as with any new approach, we recommend more research with simulated and real data sets, considering different sample sizes and response models, before practical implementation.

We mention in this respect two open questions, related to our article, which need to be explored further. The first question refers to the asymptotic properties of our proposed estimators, given that we estimate the conditional sample selection probabilities outside the likelihood by use of kernel regression. See Remark 5. The second question regards the use of the bootstrap distribution of the HL test (or any other test) in the case of a small or moderate number of responding units, as an alternative to the asymptotic Chi-square distribution, which we used in this article. See Remark 11.

In this article, we review several other approaches proposed in the literature for analyzing complex survey data. It will be of great interest to compare the alternative approaches empirically, using simulated and possibly also real survey data. This is a very challenging project, which we hope to undertake in the future.

DATA AVAILABILITY

The data used in section 8 are available from the authors.

Appendix: Proofs of Lemmas 1 and 2

Proof of Lemma 1:
(a) Assume first that $\varepsilon/2 \le \rho(y,\boldsymbol{x}) \le 1-\varepsilon/2$ for all $(y,\boldsymbol{x})$. Then $\text{logit}(\rho(y,\boldsymbol{x}))$ is bounded. By the multivariate generalization of Weierstrass (Picard 1891; Weierstrass 1895), there exists a polynomial Q such that $|Q(y,\boldsymbol{x}) - \text{logit}(\rho(y,\boldsymbol{x}))| < \varepsilon$ for all $(y,\boldsymbol{x})$. It follows from the Mean Value Theorem that $|\text{logit}^{-1}(Q(y,\boldsymbol{x}))) - \rho(y,\boldsymbol{x}))| < \varepsilon/4$ (since $\frac{d}{ds}\text{logit}^{-1}(s) \le 1/4$).
(b) If $\varepsilon/2 \le \rho(y,\boldsymbol{x}) \le 1-\varepsilon/2$ does not hold everywhere, define $\rho_1(y,\boldsymbol{x}) = \varepsilon/2 + (1-\varepsilon)\rho(y,\boldsymbol{x})$. Then $\varepsilon/2 \le \rho_1(y,\boldsymbol{x}) \le 1-\varepsilon/2$ and $|\rho(y,\boldsymbol{x}) - \rho_1(y,\boldsymbol{x})| \le \varepsilon/2$ for all $y,\boldsymbol{x}$. Now, by Part (a) there exists a polynomial Q such that $|\text{logit}^{-1}(Q(y,\boldsymbol{x})) - \rho_1(y,\boldsymbol{x})| < \varepsilon/4$ for all $y,\boldsymbol{x}$. Applying the Triangle Inequality, we get $|\text{logit}^{-1}(Q(y,\boldsymbol{x})) - \rho(y,\boldsymbol{x})| < \varepsilon$.

Proof of Lemma 2:
(a) If $\boldsymbol{q}^{(1)}$ is in the feasible region for (3.7*), $C'D^{(-1)}(\boldsymbol{\gamma})\boldsymbol{q}^{(1)} = \boldsymbol{0}$. For any $\boldsymbol{\gamma\prime}$, let $\boldsymbol{q}^{(3)} = D(\boldsymbol{\gamma\prime})D^{(-1)}(\boldsymbol{\gamma})\boldsymbol{q}^{(1)}$, implying $C'D^{(-1)}(\boldsymbol{\gamma\prime})\boldsymbol{q}^{(3)} = \boldsymbol{0}$. Define $\boldsymbol{q}^{(2)} = (\boldsymbol{1}'\boldsymbol{q}^{(3)})^{-1}\boldsymbol{q}^{(3)}$. Then, $C'D^{(-1)}(\boldsymbol{\gamma\prime})\boldsymbol{q}^{(2)} = \boldsymbol{0}$ and $\boldsymbol{q}^{(2)} \in \Omega_{r-1}$. This proves the first part of the Lemma.
(b) For a fixed $\boldsymbol{\gamma}$, the feasible region in (3.7*) [and in (3.7)] is a closed subset of Ω_{r-1}. Thus, if it is not empty, the maximum exists and is finite.

REFERENCES

Babu, G. J., and Rao, C. R. (2004), "Goodness-of-Fit Tests When Parameters Are Estimated," *Sankhya*, 66, 63–74.

Binder, A. D. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, 279–292.

Chang, T., and Kott, P. S. (2008), "Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model," *Biometrika*, 95, 555–571.

Chaudhuri, S., Handcock, M. S., and Rendall, M. S. (2008), "Generalised Linear Models Incorporating Population Level Information: An Empirical Likelihood Based Approach," *J R Stat Soc Series B Stat Methodology*, 70, 311–328.

———. (2010), *A Conditional Empirical Likelihood Approach to Combine Sampling Design and Population Level Information*. Technical Report No. 3/2010, Singapore, National University of Singapore, 117546.

Chen, S., and Kim, J. K. (2014), "Population Empirical Likelihood for Nonparametric Inference in Survey Sampling," *Statistica Sinica*, 24, 335–355.

Chen, S. X., and Van Keilegom, I. (2009), "A Review on Empirical Likelihood Methods for Regression," *Test*, 18, 415–447.

Hartley, H. O., and Rao, J. N. K. (1968), "A New Estimation Theory for Sample Surveys," *Biometrika*, 55, 547–557.

Hosmer, D. W., and Lemeshow, S. (1980), "A Goodness-of-Fit Test for the Multiple Logistic Regression Model," *Communications in Statistics*, A10, 1043–1069.

Kim, J. K., and Morikawa, K. (2023), "An Empirical Likelihood Approach to Reduce Selection Bias in Voluntary Samples," *Calcutta Statistical Association Bulletin*, 75, 8–27.

Lee, J., and Berger, J. O. (2001), "Semiparametric Bayesian Analysis of Selection Models," *Journal of the American Statistical Association*, 96, 1397–1409.

Morikawa, T., Beppu, K.., and Aida, W. (2023), "Efficient Multiple-Robust Estimation for Nonresponse Data Under Informative Sampling," arXiv:2311.06719.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and Its Applications*, 9, 141–142.

Orchard, T., and Woodbury, M. A. (1972), "A Missing Information Principle: Theory and Application," *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697–715.

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–249.

———. (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18, 90–120.

———. (1991), "Empirical Likelihood for Linear Models," *The Annals of Statistics*, 19, 1725–1747.

———. (2001), *Empirical Likelihood*. Boca Raton: Chapman & Hall/CRC.

———. (2013), "Self-Concordance for Empirical Likelihood," *Canadian Journal of Statistics*, 41, 387–397.

Pfeffermann, D. (2011), "Modeling of Complex Survey Data: Why Model? Why is It a Problem? How Can we Approach It?," *Survey Methodology*, 37, 115–136.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998), "Parametric Distributions of Complex Survey Data Under Informative Probability Sampling," *Statistica Sinica*, 8, 1087–1114.

Pfeffermann, D., and Sverchkov, M. (1999), "Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data," *Sankhyā*, 61, 166–186.

———. (2003), "Fitting Generalized Linear Models under Informative Probability Sampling," in *Analysis of Survey Data*, eds., R. L. Chambers and C. J. Skinner, New York, NY: Wiley, pp. 175–195.

———. (2009), "Inference Under Informative Sampling," in *Handbook of Statistics 29B; Sample Surveys: Inference and Analysis*, eds. D. Pfeffermann and C.R. Rao, Amsterdam: North Holland, 455–487.

Pfeffermann, D., and Landsman, V. (2011), "Are Private Schools Really Better Than Public Schools? Assessment by Methods for Observational Studies," *Annals of Applied Statistics*, 5, 1726–1751.

Pfeffermann, D., and Sikov, A. (2011), "Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information," *Journal of Official Statistics*, 27, 181–209.

Pfeffermann, D., and Sverchkov, M. (2019), "Multivariate Small Area Estimation under Nonignorable Nonresponse," *Statistical Theory and Related Fields*, 3, 213–223.

Picard, É. (1891), "Sur la Représentation Approchée Des Fonctions," *Comptes rendus hebdomadaires des séances de l'Académie des sciences Paris*, 112, 183–186.

Qin, J., and Lawless, J. (1994), "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22, 300–325.

Qin, J., Leung, D., and Shao, J. (2002), "Estimation with Survey Data Under Nonignorable Nonresponse or Informative Sampling," *Journal of the American Statistical Association*, 97, 193–200.

Qin, J., Shao, J., and Zhang, B. (2008), "Efficient and Doubly Robust Imputation for Covariate-Dependent Missing Response," *Journal of the American Statistical Association*, 103, 797–810.

Racine, J. S., and Li, Q. (2004), "Nonparametric Estimation of Regression Functions With Both Categorical and Continuous Data," *Journal of Econometrics*, 119, 99–130.

Rotnitzky, A., and Robins, J. (1997), "Analysis of Semi-Parametric Regression Models with Nonignorable Nonresponse," *Statistics in Medicine*, 16, 81–102.

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053.

Sverchkov, M. (2008), A new approach to estimation of response probabilities when missing data are not missing at random. *Joint Statistical Meetings, Proceedings of the Section on Survey Research Methods*, 867–874.

Sverchkov, M., and Pfeffermann, D. (2018), "Small Area Estimation Under Informative Sampling and Not Missing at Random Nonresponse," *Journal of the Royal Statistical Society, Series A*, 181, 981–1008.

Wang, S., Shao, J., and Kim, J. K. (2014), "An Instrument Variable Approach for Identification and Estimation with Nonignorable Nonresponse," *Statistica Sinica*, 24, 1097–1116.

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, 26, 359–372.

Weierstrass, K. (1895), *Mathematische Werke*, Berlin: Meyer & Müller. https://archive.org/details/mathematischewer02weieuoft.

Zhao, P., Haziza, D., and Wu, C. (2022), "Sample Empirical Likelihood and the Design-Based Oracle Variable Selection Theory," *Statistica Sinica*, 32, 435–457.