## Survey Methodology

# Survey data integration for regression analysis using model calibration

by Zhonglei Wang, Hang J. Kim and Jae Kwang Kim

Statistics  Statistique
Canada  Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

**Email at** infostats@statcan.gc.ca

**Telephone,** from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service                                              1-800-263-1136
- National telecommunications device for the hearing impaired    1-800-363-7629
- Fax line                                                                           1-514-283-9350

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

# Survey data integration for regression analysis using model calibration

**Zhonglei Wang, Hang J. Kim and Jae Kwang Kim[1]**

## Abstract

We consider regression analysis in the context of data integration. To combine partial information from external sources, we employ the idea of model calibration which introduces a "working" reduced model based on the observed covariates. The working reduced model is not necessarily correctly specified but can be a useful device to incorporate the partial information from the external data. The actual implementation is based on a novel application of the information projection and model calibration weighting. The proposed method is particularly attractive for combining information from several sources with different missing patterns. The proposed method is applied to a real data example combining survey data from Korean National Health and Nutrition Examination Survey and big data from National Health Insurance Sharing Service in Korea.

**Key Words:** Big data; Empirical likelihood; Information projection; Measurement error models; Missing covariates.

## 1. Introduction

Data integration is an emerging research area in survey sampling. By incorporating the partial information from external samples, one can improve the efficiency of the resulting estimator and obtain a more reliable analysis. Lohr and Raghunathan (2017), Yang and Kim (2020), and Rao (2021) provide reviews of statistical methods of data integration for finite population inference. Many existing methods (e.g., Hidiroglou, 2001; Merkouris, 2010; Zubizarreta, 2015) are mainly concerned with estimating population means or totals while combining information for analytic inference such as regression analysis is not fully explored in the existing literature.

In this paper, we consider regression analysis in the context of data integration. When we combine data sources to perform a combined regression analysis, we may encounter some problems: covariates may not be fully observed or be subject to measurement errors. Thus, one may consider the problem as a missing-covariate regression problem. Robins, Rotnitzky and Zhao (1994) and Wang,Wang, Zhao and Ou (1997) discussed semiparametric estimation in regression analysis with missing covariate data under the missing-at-random covariate assumption. In our setup, the external data source with missing covariates can be a census or big data.

Under this setup, Chatterjee, Chen, Maas and Carroll (2016) developed a data integration method based on the constrained maximum likelihood, which uses a fully parametric model for the likelihood specification and a constraint developed from a reduced model for data integration. The constrained maximum likelihood method is efficient when the model is correctly specified but is not applicable when it is difficult or impossible to specify a correct density function. Kundu, Tang, and Chatterjee (2019) generalized the method of Chatterjee et al. (2016) to consider multiple regression models based on the

---

1. Zhonglei Wang, Assistant Professor, Wang Yanan Institute for Studies in Economics (WISE) and School of Economics, Xiamen University, Xiamen, Fujian 361005, PRC. E-mail: wangzl@xmu.edu.cn; Hang J. Kim, Associate Professor, Division of Statistics and Data Science, University of Cincinnati, Cincinnati, OH 45221, U.S.A. E-mail: kim3h4@ucmail.uc.edu; Jae Kwang Kim, Professor, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. E-mail: jkim@iastate.edu.

theory of generalized method of moments (Hansen, 1982, GMM). Recently, Xu and Shao (2020) develop a data integration method using a generalized method of moments technique, but their method implicitly assumes that the reduced model is correctly specified. Under a nested case-control design, Shin, Pfeiffer, Graubard and Gail (2020a) proposed to use the fully observed sample in the phase 2 to fit a parametric model, and missing covariates in the phase 1 sample are imputed; also see Shin, Pfeiffer, Graubard and Gail (2020b). Zhang, Deng, Wheeler, Qin and Yu (2021) developed a retrospective empirical likelihood framework to account for sampling bias in case-control studies. Sheng, Sun, Huang and Kim (2021) developed a penalized empirical likelihood approach to incorporate such information in the logistic regression setup.

To combine partial information from external sources, we employ the idea of model calibration (Wu and Sitter, 2001) which introduces a "working" reduced model based on observed covariates. The model parameters in the reduced model are estimated from external sources and then combined through a novel application of the empirical likelihood method (Owen, 1991; Qin and Lawless, 1994), which can be viewed as information projection (Csiszár and Shields, 2004). The working reduced model is not necessarily specified correctly, but a good working model can improve the efficiency of the resulting analysis. The proposed method is particularly attractive for combining information from several data sources with different missing patterns. In this case, we only need to specify different working models for different missing patterns.

Besides, our proposed method is based on the first moment conditions like usual regression analyses, so weak assumptions can broaden the applicability of the proposed method to many practical problems. In particular, the proposed method is directly applicable to survey sample data which is the main focus of our paper. We consider a more general regression setup and our proposed empirical likelihood method does not require that the working reduced model to be correctly specified.

We highlight the contribution of our paper as follows. First, we propose a unified framework for incorporating external data sources in the context of regression analysis. The proposed method uses weaker assumptions than the parametric model-based method of Chatterjee et al. (2016) and thus provides more robust estimation results. Second, the proposed method is widely applicable as it can easily handle multiple external data sources as demonstrated in Section 5. It can also be applied to the case where the external data source is subject to selection bias. In the real data application in Section 7, we demonstrated that our proposed method can utilize the external big data with unknown selection probabilities by applying propensity score weighting adjustment. Finally, our proposed method is easy to implement and fully justified theoretically. The computation is simple as it is a direct application of the standard empirical likelihood method and can be implemented using the existing software.

The paper is organized as follows. In Section 2, a basic setup is introduced, and the existing methods are presented. Section 3 presents the proposed approach, and Section 4 provides its asymptotic properties. In Section 5, an application to multiple data integration is presented. Section 6 presents simulation studies, followed by the application of the proposed method to real data in Section 7. Some concluding remarks are made in Section 8.

# 2. Basic setup

Consider a finite population $\mathcal{U} = \{1, \ldots, N\}$ of size $N$. Associated with the $i^{\text{th}}$ unit, let $y_i$ denote the study variable of interest and $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2})$ the corresponding auxiliary vector of length $p$. We are interested in estimating a population parameter $\boldsymbol{\beta}_0$, which solves $\mathbf{U}_1(\boldsymbol{\beta}) = \sum_{i \in \mathcal{U}} \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = \mathbf{0}$ where $\mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}, y)$ is a pre-specified estimating function for $\boldsymbol{\beta}$. One example of the estimating function is $\mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = \{y_i - m_1(\mathbf{x}_i; \boldsymbol{\beta})\} \mathbf{h}_1(\mathbf{x}_i; \boldsymbol{\beta})$, which is implicitly based on a regression model $E(Y_i \mid \mathbf{x}_i) = m_1(\mathbf{x}_i; \boldsymbol{\beta})$ on the super-population level for some $\mathbf{h}_1(\mathbf{x}_i; \boldsymbol{\beta})$ satisfying certain identification conditions (e.g., Kim and Rao, 2009). From the finite population a probability sample $S_1 \subset \mathcal{U}$ is selected, and a $Z$-estimator $\hat{\boldsymbol{\beta}}$ can be obtained by solving

$$\hat{\mathbf{U}}_1(\boldsymbol{\beta}) \equiv \sum_{i \in S_1} d_{i1} \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = \mathbf{0}, \tag{2.1}$$

where $d_i$ is the sampling weight for unit $i \in S_1$.

In addition to $S_1$, suppose that we observe $\mathbf{x}_{i1}$ and $y_i$ throughout the finite population and wish to incorporate this extra information to improve the estimation efficiency of $\hat{\boldsymbol{\beta}}$. Before proposing our method, we introduce two related works, including Chen and Chen (2000) and Chatterjee et al. (2016).

Chen and Chen (2000) first considered this problem in the context of measurement error models. To explain their idea in our setup, we first consider a "working" reduced model,

$$E(Y_i \mid \mathbf{x}_{i1}) = m_2(\mathbf{x}_{i1}; \boldsymbol{\alpha}) \tag{2.2}$$

for some $\boldsymbol{\alpha}$. Under the working model (2.2), we can obtain an estimator $\hat{\boldsymbol{\alpha}}$ from the current sample $S_1$ by solving

$$\hat{\mathbf{U}}_2(\boldsymbol{\alpha}) \equiv \sum_{i \in S_1} d_i \mathbf{U}_2(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i) = \mathbf{0}, \tag{2.3}$$

where $\mathbf{U}_2(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i) = \{y_i - m_2(\mathbf{x}_{i1}; \boldsymbol{\alpha})\} \mathbf{h}_2(\mathbf{x}_{i1}; \boldsymbol{\alpha})$ for some $\mathbf{h}_2(\mathbf{x}_{i1}; \boldsymbol{\alpha})$ satisfying conditions similar to ones imposed to $\mathbf{h}_1(\mathbf{x}_i; \boldsymbol{\beta})$. In addition, one can get $\boldsymbol{\alpha}^*$ that solves $\sum_{i=1}^{N} \mathbf{U}_2(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i) = \mathbf{0}$. Chen and Chen (2000) proposed using

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} + \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) \{\hat{V}(\hat{\boldsymbol{\alpha}})\}^{-1} (\boldsymbol{\alpha}^* - \hat{\boldsymbol{\alpha}})$$

as an efficient estimator of $\boldsymbol{\beta}$ where $\hat{V}(\cdot)$ and $\widehat{\text{Cov}}(\cdot)$ denote the design-based variance and covariance estimators, respectively. The working model in (2.2) is not necessarily correctly specified, but a good working model can improve the efficiency of the final estimator. While the estimator of Chen and Chen (2000) is theoretically justified, it can be numerically unstable as the estimation errors of the variance and covariance matrix can be large.

Chatterjee et al. (2016) considered a likelihood-based approach using a conditional distribution of $Y_i$ given $\mathbf{X}_i$ with density $f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})$ and imposed a constraint based on external information. Specifically, they proposed to maximize

$$\prod_{i \in \mathcal{S}_1} f\left(y_i \mid \mathbf{x}_i ; \boldsymbol{\beta}\right) dF\left(\mathbf{x}_i\right) \tag{2.4}$$

subject to

$$\iint \mathbf{U}_2\left(\boldsymbol{\alpha}^* ; \mathbf{x}_1, y\right) f\left(y \mid \mathbf{x} ; \boldsymbol{\beta}\right) dy dF(\mathbf{x}) = \mathbf{0}, \tag{2.5}$$

where $F(\mathbf{x})$ is an unspecified distribution function for $\mathbf{x}$, $dF(\mathbf{x})$ is the Radon-Nikodym derivative of the distribution function $F(\mathbf{x})$ with respect to a certain dominating measure, and $\boldsymbol{\alpha}^*$ is the model parameter available from an external source. Following the likelihood based approach of Chatterjee et al. (2016), $\mathbf{U}_2\left(\boldsymbol{\alpha} ; \mathbf{x}_{i1}, y\right)$ corresponds to the estimating function involving a "reduced" distribution function $g\left(y_i \mid \mathbf{x}_{i1} ; \boldsymbol{\alpha}_0\right)$ with model parameter $\boldsymbol{\alpha}_0$, where $g\left(y_i \mid \mathbf{x}_{i1} ; \alpha_0\right)$ can be incorrectly specified. That is, $\boldsymbol{\alpha}^*$ is the external information for $\boldsymbol{\alpha}_0$. Chatterjee et al. (2016) estimated $F(\mathbf{x})$ nonparametrically by empirical likelihood. By imposing this constraint into the maximum likelihood estimation, the external information $\boldsymbol{\alpha}^*$ can be naturally incorporated.

The constrained maximum likelihood (CML) method is not directly applicable to our conditional mean model in (2.1) as the likelihood function for $\boldsymbol{\beta}$ is not defined in our setup. Besides, the design feature for the probability sample $S_1$ is not directly applicable in their method. Nonetheless, one can use an objective function such as that in a generalized method of moments to apply the constrained optimization problem, which is asymptotically equivalent to the empirical likelihood method (Imbens, 2002). The empirical likelihood implementation of CML approach is discussed by Han and Lawless (2019).

## 3.  Proposed approach

We now consider an alternative approach for combining information from several sources. To combine information from several sources, we use the Kullback-Leibler (KL) divergence measure to apply the information projection (Csiszár and Shields, 2004) on the model space with constraints. Let $\hat{P}$ be the empirical distribution of the sample with

$$\hat{P}(x, y) = \frac{1}{\sum_{i \in \mathcal{S}_1} d_i} \sum_{i \in \mathcal{S}_1} d_i \, \mathbf{I}\left\{(x, y) = \left(x_i, y_i\right)\right\}. \tag{3.1}$$

Given the empirical distribution $\hat{P}$, we wish to find the minimizer of

$$D\left(\hat{P} \| P\right) = \int \log\left\{d\hat{P}(\mathbf{x}, y)\right\} d\hat{P}(\mathbf{x}, y) - \int \log\left\{dP(\mathbf{x}, y)\right\} d\hat{P}(\mathbf{x}, y) \tag{3.2}$$

with respect to $P$ in the model space. Notice that the first term is a constant and the minimizer of (3.2) is the pseudo maximum likelihood estimator of $\hat{P}$.

We consider the following constraints in our model at the finite-population level:

$$\sum_{i=1}^{N} \mathbf{U}_1\left(\boldsymbol{\beta} ; \mathbf{x}_i, y_i\right) p\left(\mathbf{x}_i, y_i\right) = 0 \quad \text{and} \quad \sum_{i=1}^{N} \mathbf{U}_2\left(\boldsymbol{\alpha}^* ; \mathbf{x}_{i1}, y_i\right) p\left(\mathbf{x}_i, y_i\right) = 0, \tag{3.3}$$

where $p(\mathbf{x}_i, y_i)$ is the point mass assigned to point $(\mathbf{x}_i, y_i)$ in the finite population satisfying $\sum_{i=1}^{N} p(\mathbf{x}_i, y_i) = 1$. See Figure 3.1 for a graphical illustration of the information projection.

**Figure 3.1   Information projection for the empirical distribution $\hat{P}$.**
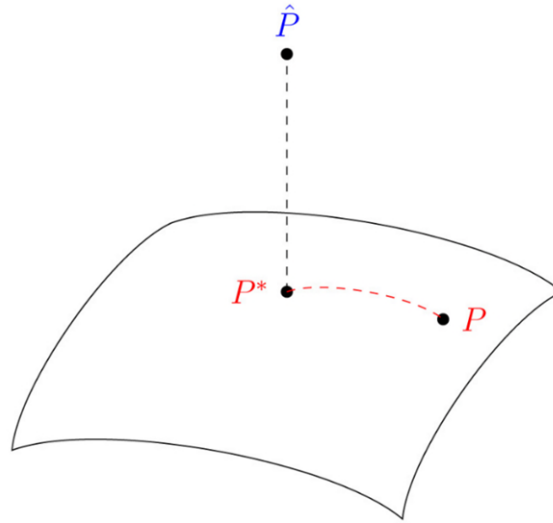


Note that $P^*$ minimizes $D(\hat{P} \| P)$ among $P$ satisfying the constraints in (3.3).

Using the weighted empirical distribution in (3.1), the KL divergence measure in (3.2) reduces to $D(\hat{P} \| P) = \text{constant} - \hat{N}^{-1} \sum_{i \in \mathcal{S}_1} d_i \log\{p(\mathbf{x}_i, y_i)\}$ where $\hat{N} = \sum_{i \in \mathcal{S}_1} d_i$. Thus, we only have to maximize $l(\mathbf{p}) = \sum_{i \in \mathcal{S}_1} d_i \log(p_i)$ subject to $\sum_{i=1}^{N} p_i = 1$ and the constraints in (3.3), where $p_i$ abbreviates $p(\mathbf{x}_i, y_i)$. Note that having $p_i > 0$ for $i \notin \mathcal{S}_1$ will decrease the value of $l(\mathbf{p}) = \sum_{i \in \mathcal{S}_1} d_i \log(p_i)$, the solution $\hat{p}_i$ to this optimization problem should give $\hat{p}_i = 0$ for $i \notin \mathcal{S}_1$. Therefore, we can safely set $p_i = 0$ for $i \notin \mathcal{S}_1$ and express the problem as finding the maximizer of

$$Q(\mathbf{d}, \mathbf{w}) = \sum_{i \in \mathcal{S}_1} d_i \log(w_i) \tag{3.4}$$

subject to

$$\sum_{i \in \mathcal{S}_1} w_i = 1, \tag{3.5}$$

$$\sum_{i \in \mathcal{S}_1} w_i \mathbf{U}_2(\boldsymbol{\alpha}^*; \mathbf{x}_{i1}, y_i) = \mathbf{0}, \tag{3.6}$$

$$\sum_{i \in \mathcal{S}_1} w_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = \mathbf{0}.$$

We use $w_i$ instead of $p_i$ to represent the final weights assigned to the sample elements.

**Remark 1.** *Maximizing the objective function in (3.4) is equivalent to minimizing the following cross entropy:*

$$-\sum_{i \in \mathcal{S}_1} \tilde{d}_i \log(w_i),  \tag{3.7}$$

*where $\tilde{d}_i = d_i / \left(\sum_{i \in \mathcal{S}_1} d_i\right)$. The objective function (3.7) is also the pseudo empirical log-likelihood function considered by Chen and Sitter (1999) and Wu and Rao (2006). Instead of (3.4), we may consider other objective functions, including the population empirical likelihood proposed by Chen and Kim (2014) for example.*

Our proposed method is different from Chatterjee et al. (2016) in that we use a more general integral constraint (2.5) which does not involve the conditional density function $f(y \mid \mathbf{x}; \boldsymbol{\beta})$. Constraint (3.6) still incorporates the extra information in $\boldsymbol{\alpha}^*$. The above optimization can be solved by applying the standard profile empirical likelihood method or using the following two-step estimation method.

1. Find the calibration weights $\hat{\mathbf{w}} = \{\hat{w}_i : i \in \mathcal{S}_1\}$ maximizing $Q(\mathbf{d}, \mathbf{w})$ subject to (3.5)-(3.6).

2. Once the solution $\hat{\mathbf{w}}$ is obtained from the calibration, estimate $\boldsymbol{\beta}$ by solving

$$\sum_{i \in \mathcal{S}_1} \hat{w}_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = \mathbf{0}.  \tag{3.8}$$

If the benchmark $\boldsymbol{\alpha}^*$ is not available from the finite population but can be estimated from an independent external sample, we can use the information from both the original internal sample and the external sample to obtain the benchmark estimate. In practical situations, we may not have access to the raw data of the external sample but often be able to have its summary statistics. Suppose that the external sample provides a point estimator $\hat{\boldsymbol{\alpha}}_2$ and its variance estimator $\mathbf{V}_2 = \hat{V}(\hat{\boldsymbol{\alpha}}_2)$ for the working reduced model in (2.2). Then, an estimator of the benchmark $\boldsymbol{\alpha}^*$ can be obtained by

$$\hat{\boldsymbol{\alpha}}^* = \left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \left(\mathbf{V}_1^{-1} \hat{\boldsymbol{\alpha}}_1 + \mathbf{V}_2^{-1} \hat{\boldsymbol{\alpha}}_2\right),  \tag{3.9}$$

where $\hat{\boldsymbol{\alpha}}_1$ and $\mathbf{V}_1$ are estimated with the internal sample $\mathcal{S}_1$. Once $\hat{\boldsymbol{\alpha}}^*$ is obtained by (3.9), it replaces $\boldsymbol{\alpha}^*$ in the calibration equation in (3.6).

Similarly to Wu and Sitter (2001), the proposed method does not require a "true" working model as explained below. Let $\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}) = 0$ be the estimating equation for obtaining $\boldsymbol{\alpha}^*$ computed from the external sample $\mathcal{S}_2$. Now, the final estimating function for $\boldsymbol{\beta}$ using the model calibration $\hat{\mathbf{U}}_{\text{cal}}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{S}_1} \hat{w}_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i)$ can be approximated by

$$\hat{\mathbf{U}}_{\text{cal}}(\boldsymbol{\beta}) \doteq \hat{\mathbf{U}}_1(\boldsymbol{\beta}) + \mathbf{K}\left\{\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}^*) - \hat{\mathbf{U}}_2(\boldsymbol{\alpha}^*)\right\}  \tag{3.10}$$

for some $\mathbf{K}$ where $\hat{\mathbf{U}}_1(\boldsymbol{\beta})$ and $\hat{\mathbf{U}}_2(\boldsymbol{\alpha})$ are computed by (2.1) and (2.3), respectively, from the internal sample $\mathcal{S}_1$. The approximation in (3.10) can be easily derived using the asymptotic equivalence of the calibration estimator and the regression estimator. Thus, even if $E\left\{\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}^*)\right\}$ is not equal to zero, the solution to $\hat{\mathbf{U}}_{\text{cal}}(\boldsymbol{\beta}) = 0$ is consistent as $E\left\{\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}) - \hat{\mathbf{U}}_2(\boldsymbol{\alpha})\right\} = 0$ by design.

**Remark 2.** *Although the working model $E\left(Y_i \mid \mathbf{x}_{i1}\right) = m_2\left(\mathbf{x}_{i1}; \boldsymbol{\alpha}\right)$ does not need to be correctly specified, we can systematically find $\mathbf{U}_2\left(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i\right)$ by casting its construction as a missing covariate problem, relying on the regression calibration technique. For example, suppose that $\mathbf{x}_i = \left(x_{i1}, x_{i2}\right)$, we set a predictor $\hat{x}_{i2} = \beta_0 + \beta_1 x_{i1}$, and an estimating equation is written by*

$$\mathbf{U}_1\left(\boldsymbol{\beta}; x_{i1}, \hat{x}_{i2}, y_i\right) = \left\{y_i - m_1\left(x_{i1}, \hat{x}_{i2}; \boldsymbol{\beta}\right)\right\} \mathbf{h}_1\left(x_{i1}, \hat{x}_{i2}; \boldsymbol{\beta}\right) \tag{3.11}$$

*for the control function of the model calibration method where $\boldsymbol{\beta} = \left(\beta_0, \beta_1\right)$. We can either estimate $\boldsymbol{\beta}$ from sample $\mathcal{S}_1$ or use any fixed parameter value as long as the solution to $\sum_{i \in \mathcal{S}_1} d_i \, \mathbf{U}_1\left(\boldsymbol{\beta}; x_{i1}, \hat{x}_{i2}, y_i\right) = \mathbf{0}$ is unique. A benchmark estimator of $\boldsymbol{\beta}$ can be obtained using external samples to apply the proposed model calibration method. If we use the control function in (3.11), then we are essentially treating a regression of $y$ on $x_1$ and $\hat{x}_2$ as the "working" model for model calibration. This is feasible only when we have direct access to an external sample $\mathcal{S}_2$ in addition to the internal sample $\mathcal{S}_1$.*

# 4. Theoretical properties

In this section, we investigate the asymptotic properties of the the proposed estimator $\hat{\boldsymbol{\beta}}$ to (3.8). Since the population parameters including $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}^*$ are determined by the finite population of size $N$, we explicitly use a subscript $N$ for those in this section, e.g., $\boldsymbol{\beta}_{0N}$ and $\boldsymbol{\alpha}_N^*$, but we omit this subscript for $\left(d_i, \mathbf{x}_i, y_i\right)$ for simplicity. We consider two scenarios: when $\boldsymbol{\alpha}_N^*$ is available from the finite population and when we only have an external sample to estimate $\boldsymbol{\alpha}_N^*$ by the generalized least square in (3.9).

## 4.1 $\boldsymbol{\alpha}_N^*$ is available

Let $\tilde{d}_i = \hat{N}^{-1} d_i$ where $\hat{N} = \sum_{i \in \mathcal{S}_1} d_i$ is the Horvitz-Thompson estimator of the population size $N$. Replacing $d_i$ by $\tilde{d}_i$ in (3.4), we consider the Lagrangian problem that maximizes

$$l\left(\mathbf{w}, \boldsymbol{\lambda}, \phi\right) = \sum_{i \in \mathcal{S}_1} \tilde{d}_i \log\left(w_i\right) + \boldsymbol{\lambda}^\top \sum_{i \in \mathcal{S}_1} w_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right) + \phi\left(\sum_{i \in \mathcal{S}_1} w_i - 1\right),$$

where $\boldsymbol{\lambda}$ and $\phi$ are the Lagrange multipliers.

By setting $\partial l\left(\mathbf{w}, \boldsymbol{\lambda}, \phi\right) / \partial \boldsymbol{\lambda} = \mathbf{0}$, $\partial l\left(\mathbf{w}, \boldsymbol{\lambda}, \phi\right) / \partial \phi = 0$ and $\partial l\left(\mathbf{w}, \boldsymbol{\lambda}, \phi\right) / \partial w_i = 0$ for $i \in \mathcal{S}_1$, we get $\hat{\phi} = -1$ and $\hat{w}_i = \tilde{d}_i \left\{1 - \boldsymbol{\lambda}^\top \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)\right\}^{-1}$. Then, the proposed method is equivalent to solving $g\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right) = \mathbf{0}$ where

$$g\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right) = \begin{pmatrix} \displaystyle\sum_{i \in \mathcal{S}_1} \frac{\tilde{d}_i}{1 - \boldsymbol{\lambda}^\top \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)} \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) \\ \displaystyle\sum_{i \in \mathcal{S}_1} \frac{\tilde{d}_i}{1 - \boldsymbol{\lambda}^\top \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right) \end{pmatrix}. \tag{4.1}$$

Denote the solution to (4.1) as $\hat{\boldsymbol{\eta}} = \left(\hat{\boldsymbol{\beta}}^\top, \boldsymbol{\lambda}^\top\right)^\top$. To investigate asymptotic properties of $\hat{\boldsymbol{\eta}}$, we propose the following regularity conditions.

C1. There exists a compact set $\mathcal{A}$ such that $Z_S = \sup_{\boldsymbol{\alpha} \in \mathcal{A}} \max_{i \in \mathcal{S}_1} \left\| \mathbf{U}_2 \left( \boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i \right) \right\| = o_p \left( n^{1/2} \right)$ and $\boldsymbol{\alpha}_N^* \in \mathcal{A}$ for $N \in \mathbb{N}$, where $\|\cdot\|$ denotes the Euclidean norm and the stochastic order is with respect to the sampling design.

C2. The sampling design satisfies the following convergence results.

    a. There exist a compact set $\Omega$ such that $\boldsymbol{\beta}_{0N} \in \Omega$ for $N \in \mathbb{N}$ and an interior point of $\Omega$, $\boldsymbol{\beta}_p$, such that $\lim_{N \to \infty} \boldsymbol{\beta}_{0N} = \boldsymbol{\beta}_p$.

    b. There exists a continuous function $\mathbf{U}_0(\boldsymbol{\beta})$ over $\Omega$ such that $\sup_{\boldsymbol{\beta} \in \Omega} \left\| \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1 (\boldsymbol{\beta}; \mathbf{x}_i, y_i) - \mathbf{U}_0(\boldsymbol{\beta}) \right\| \to 0$ in probability, where $\boldsymbol{\beta}_p$ is the unique solution to $\mathbf{U}_0(\boldsymbol{\beta}) = \mathbf{0}$.

    c. $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \partial \mathbf{U}_1 (\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i) / \partial \boldsymbol{\beta}^\top = \boldsymbol{I}_{11} + o_p(1)$, where $\boldsymbol{I}_{11}$ is non-stochastic and invertible.

    d. $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1 (\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i) \mathbf{U}_2 (\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i)^\top = \boldsymbol{I}_{12} + o_p(1)$, where $\boldsymbol{I}_{12}$ is non-stochastic.

    e. $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 (\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i)^{\otimes 2} = \boldsymbol{I}_{22} + o_p(1)$, where $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^\top$ for any matrix $\mathbf{A}$ and $\boldsymbol{I}_{22}$ is non-stochastic and positively definitive.

C3. The sampling design satisfies

$$n^{1/2} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \begin{pmatrix} \mathbf{U}_1 (\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i) \\ \mathbf{U}_2 (\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i) \end{pmatrix} \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$$

in distribution, where $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$ is a normal distribution with mean zero and covariance matrix

$$\boldsymbol{\Sigma}_u = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

C1 is a technical condition to obtain the asymptotic order of $\hat{\boldsymbol{\lambda}}$, and a similar condition is also assumed by Wu and Rao (2006); see their condition C1 for details. C2 assumes several convergence results for the two estimating functions. Specifically, C2a shows the parameter space of the finite population parameter $\boldsymbol{\beta}_{0N}$, and the convergence of $\boldsymbol{\beta}_{0N}$ can be satisfied under regularity conditions. Condition C2b is necessary to show $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_p \to 0$ in probability, then $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{0N} \to 0$ in probability, coupled with C2a. Conditions C2c-C2e guarantee the central limit theorem for $\hat{\boldsymbol{\eta}}$. Note that $\boldsymbol{I}_{22}$ is symmetric by C2e, but $\boldsymbol{I}_{11}$ in C2c may be asymmetric for a certain estimating function $\mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}, y)$. Condition C3 is satisfied under regularity conditions for general sampling designs; see Fuller (2009, Section 1.3) for details.

**Theorem 1.** *Suppose that conditions C1-C3 hold. Then, $n^{1/2} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ in distribution, where $\boldsymbol{\Sigma}_\eta = \boldsymbol{I}^{-1} \boldsymbol{\Sigma}_u (\boldsymbol{I}^{-1})^\top$ and*

$$\boldsymbol{I} = \begin{pmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \mathbf{0} & \boldsymbol{I}_{22} \end{pmatrix}.$$

The proof of Theorem 1 is presented in Appendix A. By Theorem 1, we can obtain that $n^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{0N}) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$ in distribution, where

$$\Sigma_\beta = \mathcal{I}_{11}^{-1}\Sigma_{11}\left(\mathcal{I}_{11}^{-1}\right)^\top - \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\mathcal{I}_{22}^{-1}\Sigma_{21}\left(\mathcal{I}_{11}^{-1}\right)^\top - \mathcal{I}_{11}^{-1}\Sigma_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{12}^\top\left(\mathcal{I}_{11}^{-1}\right)^\top + \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\mathcal{I}_{22}^{-1}\Sigma_{22}\mathcal{I}_{22}^{-1}\mathcal{I}_{12}^\top\left(\mathcal{I}_{11}^{-1}\right)^\top$$

and $\Sigma_{11}$ and $\Sigma_{22}$ correspond to the asymptotic variances of $n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)$ and $n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{i1},y_i\right)$, respectively. Furthermore, we have the following result regarding the optimality of $\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{i1},y_i\right)$.

**Corollary 1.** *Suppose that the conditions in Theorem 1 hold. For a fixed estimating function* $\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x},y),\hat{\boldsymbol{\beta}}$ *is optimal if* $\mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_1,y\right) = E\left\{\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x},y\right)\middle|\mathbf{x}_1,y\right\}$ *holds almost surely for the working reduced model, where* $\mathbf{x}=\left(\mathbf{x}_1,\mathbf{x}_2\right)$ *and the expectation is taken with respect to the super-population model.*

The proof of Corollary 1 is relegated to Appendix B. Corollary 1 presents a sufficient condition on the reduced model to guarantee an optimal estimator $\hat{\boldsymbol{\beta}}$ if the working model is correctly specified. That is, even if we do not require that the reduced model is correctly specified for consistency, the efficiency gain is guaranteed only under the correct model specification. By Corollary 1, an optimal estimator of $\boldsymbol{\alpha}_N^*$ can be obtained by solving $E\left\{\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x},y\right)\middle|\mathbf{x}_1,y\right\}=\mathbf{0}$.

Under regularity conditions, it can be shown that $\Sigma_\beta = \mathcal{I}_{11}^{-1}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)\left(\mathcal{I}_{11}^{-1}\right)^\top$ for simple random sampling with or without replacement. Since $\mathcal{I}_{11}^{-1}\Sigma_{11}\left(\mathcal{I}_{11}^{-1}\right)^\top$ is the asymptotic variance of $n^{1/2}\left(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_{0N}\right)$, where $\hat{\boldsymbol{\beta}}_m$ solves $\sum_{i\in\mathcal{S}_1}d_i\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)=0$, the proposed approach achieves efficient estimation under simple random sampling; see Section S1 of the Supplementary Material for details.

## 4.2 An external estimator $\hat{\boldsymbol{\alpha}}_2$ is available

When $\boldsymbol{\alpha}^*$ is not available but an external sample is available to get $\hat{\boldsymbol{\alpha}}^*$ in (3.9), we consider

$$\tilde{g}(\boldsymbol{\eta}) = \begin{pmatrix} \displaystyle\sum_{i\in\mathcal{S}_1}\frac{\tilde{d}_i}{1-\boldsymbol{\lambda}^\top\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{i1},y_i\right)}\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right) \\ \displaystyle\sum_{i\in\mathcal{S}_1}\frac{\tilde{d}_i}{1-\boldsymbol{\lambda}^\top\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{i1},y_i\right)}\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{i1},y_i\right) \end{pmatrix}. \tag{4.2}$$

Denote $\tilde{\boldsymbol{\eta}}$ to be the solution of $\tilde{g}(\boldsymbol{\eta})=\mathbf{0}$. Then, the following additional assumptions are required to get the asymptotic properties for $\tilde{\boldsymbol{\eta}}$.

C4. $\sum_{i\in\mathcal{S}_1}\tilde{d}_i\partial\mathbf{U}_2\left(\boldsymbol{\alpha};\mathbf{x}_{i1},y_i\right)/\partial\boldsymbol{\alpha}^\top = \boldsymbol{I}(\boldsymbol{\alpha})+o_p(1)$ uniformly for $\boldsymbol{\alpha}\in\mathcal{A}$, where $\boldsymbol{I}(\boldsymbol{\alpha})$ is non-stochastic. Besides, there exists an invertible matrix $\boldsymbol{I}_0$ such that $\lim_{N\to\infty}\boldsymbol{I}\left(\boldsymbol{\alpha}_N^*\right)=\boldsymbol{I}_0$.

C5. The sampling design and the external sample satisfy the following convergence results.

    a. Both $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$ are consistent for $\boldsymbol{\alpha}^*$.

    b. $\mathbf{V}_1$ and $\mathbf{V}_2$ are design consistent variance estimators of $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$, respectively.

    c. $\mathbf{V}_1^{-1}$, $\mathbf{V}_2^{-1}$, and $\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}$ exist in probability.

    d. $\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1} = \mathbf{W}+o_p(1)$, where $\mathbf{W}$ is non-stochastic.

e. There exists a scaling function $\gamma(n)$ such that $\gamma(n)(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}^*) \to \mathcal{N}(0, \boldsymbol{\Sigma}_2)$ in distribution, where $\boldsymbol{\Sigma}_2$ satisfies $\gamma(n)^2 \mathbf{V}_2 = \boldsymbol{\Sigma}_2 + o_p(1)$.

C4 is used to obtain the asymptotic order and the variance of $\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}_N^*$, and a similar condition was used by Yuan and Jennrich (1998). C5a and C5b assume the consistency of $\hat{\boldsymbol{\alpha}}_2$ and $\mathbf{V}_2$ obtained by an external sample. For the consistency of $\hat{\boldsymbol{\alpha}}_1$, a sufficient condition is similar with C2b. The design consistency of the variance estimator $\mathbf{V}_1$ can be obtained under general sampling designs; see Fuller (2009, Chapter 1) for details. C5c guarantees the existence of $\hat{\boldsymbol{\alpha}}^*$ for the proposed method. C5e shows the central limit theorem with respect to the summary statistic $\hat{\boldsymbol{\alpha}}_2$, and it is used to derive a similar result as C3 with $\boldsymbol{\alpha}^*$ replaced by $\hat{\boldsymbol{\alpha}}^*$. Specifically, the convergence rate of $(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}^*)$ is $\gamma(n)^{-1}$, which is determined by the external sample.

The following theorem establishes an asymptotic distribution similar to that in C3.

**Theorem 2.** *Suppose that conditions C1 and C3-C5 hold. Then,*

$$n^{1/2} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \begin{pmatrix} \mathbf{U}_1(\boldsymbol{\beta}_0; \mathbf{x}_i, y_i) \\ \mathbf{U}_2(\hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{i1}, y_i) \end{pmatrix} \to \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_u)$$

*in distribution, where*

$$\tilde{\boldsymbol{\Sigma}}_u = \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22} \end{pmatrix}.$$

*Case 1. Specifically, if there exists a non-stochastic matrix $\boldsymbol{\Sigma}_c$ such that $n\mathbf{V}_2 = \boldsymbol{\Sigma}_c + o_p(1)$, then $\tilde{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11}$, $\tilde{\boldsymbol{\Sigma}}_{12} = \boldsymbol{\Sigma}_{12}(\boldsymbol{I}_0^{-1})^{\top} \mathbf{W}^{\top} \boldsymbol{I}_0^{\top}$, $\tilde{\boldsymbol{\Sigma}}_{21} = \tilde{\boldsymbol{\Sigma}}_{12}^{\top}$ and $\tilde{\boldsymbol{\Sigma}}_{22} = \boldsymbol{I}_0 \mathbf{W}\{\boldsymbol{\Sigma}_c + \boldsymbol{I}_0^{-1}\boldsymbol{\Sigma}_{22}(\boldsymbol{I}_0^{-1})^{\top}\} \mathbf{W}^{\top} \boldsymbol{I}_o^{\top}$;*
*Case 2. If $\mathbf{W} = 0$, then $\tilde{\boldsymbol{\Sigma}}_{ij} = \mathbf{0}$ for $(i, j) \neq (1,1)$ and $\tilde{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11}$.*

The proof of Theorem 2 is presented in Appendix C. For Case 1, if $\hat{\boldsymbol{\alpha}}_2$ estimated from an external sample is much more efficient than $\hat{\boldsymbol{\alpha}}$ in the sense of $(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*) = o_p(n^{-1/2})$, then $\mathbf{W}$ is an identity matrix and $\tilde{\boldsymbol{\Sigma}}_{ij} = \boldsymbol{\Sigma}_{ij}$ for $i, j = 1, 2$. Thus, we can ignore the variability of the summary statistic $\hat{\boldsymbol{\alpha}}_2$ from the external sample and get the same asymptotic distribution as in C3. Although the asymptotic distributions are the same, C3 with known $\boldsymbol{\alpha}_N^*$ is not a special case of Theorem 2 since $\hat{\boldsymbol{\alpha}}_2 = \boldsymbol{\alpha}_N^*$ has zero variance, which violates C5c-C5e. On the other hand, if $(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*) \asymp n^{-1/2}$ in probability, then $\hat{\boldsymbol{\alpha}}_2$ is as efficient as $\hat{\boldsymbol{\alpha}}_1$. Thus, $\mathbf{W}$ is not an identity matrix nor a zero matrix, and the proposed method is more efficient than one replacing $\boldsymbol{\alpha}^*$ by $\hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\alpha}}_2$ due to the extra information provided by the external sample. It is trivial that we cannot use $\hat{\boldsymbol{\alpha}}_1$ to replace $\boldsymbol{\alpha}^*$ in (3.6); otherwise, we get $\hat{w}_i = \tilde{d}_i$, and (3.8) is equivalent to the traditional estimation equation $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = \mathbf{0}$ without calibration. If $\hat{\boldsymbol{\alpha}}_2$ is much less efficient than $\hat{\boldsymbol{\alpha}}_1$ in terms of convergence rate, then we should not use such an external sample for the proposed method because $\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}^* = \hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}^* + o_p(n^{-1/2})$ and $n^{1/2} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2(\hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{i1}, y_i) = o_p(1)$; see Appendix C for details. By C5, we can obtain the same consistency results in Lemmas A1-A2 for (4.2) under the same conditions. Thus, by Theorem 2, we obtain the following asymptotic distribution for $\tilde{\boldsymbol{\eta}}$.

**Corollary 2.** *Suppose that conditions C1-C5 hold. Then, we have* $n^{1/2}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \to \mathcal{N}(0, \tilde{\boldsymbol{\Sigma}}_\eta)$ *in distribution, where* $\tilde{\boldsymbol{\Sigma}}_\eta = \boldsymbol{I}^{-1}\tilde{\boldsymbol{\Sigma}}_u(\boldsymbol{I}^{-1})^\top$, *the form of* $\boldsymbol{I}$ *is in Theorem 1, and the form of* $\tilde{\boldsymbol{\Sigma}}_\eta$ *is in Theorem 2.*

**Remark 3.** *It is worthy pointing out that when deriving the asymptotic properties in this section, we do not consider the weighting adjustments such as nonresponse adjustment, trimming, and raking. However, those weighting adjustments are commonly used in survey sampling. Thus, it is a promising research topic to generalize the proposed method incorporating those weighting adjustments.*

# 5. Multiple data integration

We now consider regression analysis combining partial information from external samples. To explain the idea, Table 5.1 shows an example data structure with three data sources $(A, B, C)$, where Sample $A$ contains all the observations while samples $B$ and $C$ contain partial observations.

**Table 5.1**
**Data structure for survey integration**

| Sample | Sampling Weight | $z$ | $x_1$ | $x_2$ | $y$ |
|--------|-----------------|-----|-------|-------|-----|
| $A$ | $d_a$ | X | X | X | X |
| $B$ | $d_b$ | X | X | | X |
| $C$ | $d_c$ | X | | X | X |

Under the setup of Table 5.1, suppose that we are interested in estimating the parameters in the regression model $E(Y \mid x_1, x_2) = m_1(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$, where $m_1(\cdot)$ is known but $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ is unknown. The estimating equation for $\boldsymbol{\beta}$ using sample $A$ can be written as

$$\hat{\mathbf{U}}_a(\boldsymbol{\beta}) \equiv \sum_{i \in A} d_{a,i}\{y_i - m(x_{i1}, x_{i2}; \boldsymbol{\beta})\} \mathbf{h}(x_{i1}, x_{i2}; \boldsymbol{\beta}) = \mathbf{0} \tag{5.1}$$

for some $\mathbf{h}(x_{i1}, x_{i2}; \boldsymbol{\beta})$ such that $\hat{\mathbf{U}}_a(\boldsymbol{\beta})$ is linearly independent almost everywhere.

Now, we wish to incorporate the partial information from sample $B$. To do this, suppose that we have a "working" model for $E(Y \mid x_1, z)$:

$$E(Y \mid x_1, z) = m_2(x_1, z; \boldsymbol{\alpha}) \tag{5.2}$$

for some $\boldsymbol{\alpha}$. Note that, since $(z_i, x_{1i}, y_i)$ are observed, we can use sample $B$ to estimate $\boldsymbol{\alpha}$ by solving $\sum_{i \in B} d_{b,i} \mathbf{U}_b(\boldsymbol{\alpha}; x_{i1}, z_i, y_i) = \mathbf{0}$ for some $\mathbf{U}_b$ satisfying $E\{\mathbf{U}_b(\boldsymbol{\alpha}; x_1, z, Y) \mid x_1, z\} = \mathbf{0}$ under the working model (5.2).

Similarly, to incorporate the partial information from sample $C$, suppose that we have a "working" model for $E(Y \mid x_2, z)$:

$$E(Y \mid x_2, z) = m_3(x_2, z; \boldsymbol{\gamma}) \tag{5.3}$$

for some $\gamma$. We can also construct an unbiased estimating equation $\sum_{i \in C} d_{c,i} \mathbf{U}_c \left( \gamma; x_{i2}, z_i, y_i \right) = \mathbf{0}$ for some $\mathbf{U}_c$ satisfying $E \left\{ \mathbf{U}_c \left( \gamma; x_2, z, Y \right) \mid x_2, z \right\} = \mathbf{0}$ under the working model (5.3). Once $\hat{\alpha}$ and $\hat{\gamma}$ are obtained, we can use this extra information to improve the efficiency of $\hat{\beta}$ in (5.1). To incorporate the extra information, we can formulate it as maximizing $Q(\mathbf{d}_a, \mathbf{w}) = \sum_{i \in A} d_{a,i} \log(w_i)$ subject to $\sum_{i \in A} w_i = N$ and

$$\sum_{i \in A} w_i \left[ \mathbf{U}_b \left( \hat{\alpha}; x_{i1}, z_i, y_i \right), \mathbf{U}_c \left( \hat{\gamma}; x_{i2}, z_i, y_i \right) \right] = \mathbf{0}, \tag{5.4}$$

where $\mathbf{d}_a$ and $\mathbf{w}$ are sets containing the sampling weights and calibration weights with respect to sample $A$. Constraint (5.4) incorporates the extra information. Once the solution $\hat{w}_i$ is obtained, we can use $\sum_{i \in A} \hat{w}_i \left\{ y_i - m \left( x_{i1}, x_{i2}; \beta \right) \right\} \mathbf{h} \left( x_{i1}, x_{i2}; \beta \right) = \mathbf{0}$ to estimate $\beta$. The asymptotic results can be obtained similarly in Section 4.

**Remark 4.** *In this paper, we implicitly assume that the populations for the internal sample and the external samples are the same, but it is possible that those populations differ in some scenarios. For example, the external estimator $\hat{\alpha}$ may be obtained based on a non-probability sample, whose sampling frame differs from the one for the probability sample due to the coverage bias in many opt-in surveys. There are several data integration methods incorporating information from heterogeneous populations. For example, Taylor, Choi and Han (2022) proposed to use ratios of coefficients to incorporate the external information under regularity conditions even when the populations for the internal and external samples differ. See also Zhai and Han (2022) and Sheng, Sun, Huang and Kim (2022) for penalized approaches when incorporating external information from heterogeneous populations. The aforementioned existing methods do not take the complex sampling properties into consideration, so it is promising to investigate data integration for heterogeneous populations under survey sampling in a future project.*

# 6. Simulation study

To evaluate the finite sample performance of the proposed estimator, we conducted simulation studies assuming several scenarios. We generated a finite population of size $N = 100,000$, each record consisting of auxiliary variables $\mathbf{x}_i = (x_{i1}, x_{i2})^\top$ of length $p = 2$ and a response variable $y_i$. We assume that $(\mathbf{x}_i, y_i)$ is available for the internal sample $\mathcal{S}_1$ while only $(x_{i1}, y_i)$ is available for the external sample $\mathcal{S}_2$.

We evaluate the performance of the proposed estimator under a linear regression setup. In this case, we are interested in making statistical inference for $\beta = (\beta_0, \beta_1, \beta_2)^\top$ that solves $\sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})(1, x_{i1}, x_{i2})^\top = \mathbf{0}$.

First, we consider two scenarios to generate covariates for the finite population: (i) $x_{i1} \sim N(3, 1)$ and $x_{i2} \sim N(11, 6.5^2)$, where $x_{i1}$ and $x_{i2}$ are independent; (ii) $x_{i1} \sim N(3, 1)$ and $x_{i2} = x_{i1}^2 + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$. The simulation parameters are chosen such that the marginal mean and variance of $x_{i2}$ are similar in the independent and the dependent settings. Second, the response variable is generated as $Y_i = \mu_i + \varepsilon_i$ with $\mu_i = 1 + 2x_{i1} + x_{i2}$ under two scenarios: (i) homogeneous variance with $\varepsilon_i \sim N(0, 9)$ and (ii) heterogeneous variance with $\varepsilon_i \mid \mathbf{x}_i \sim N(0, \sigma_i^2)$ with $\sigma_i = 0.2 |\mu_i|$. Third, we consider two sampling designs to

generate a probability sample $\mathcal{S}_1$ of (expected) size $n_1 = 1,000$: (i) simple random sampling without replacement (SRS), and (ii) Poisson sampling with inclusion probabilities satisfying $\pi_{1i} \propto \left( y_i - \min\{y_i : i = 1, \ldots, N\} + 10 \right)^{1/2}$ and $\sum_{i=1}^{N} \pi_{1i} = n_1$. Last, we consider two sampling designs to generate an external sample $\mathcal{S}_2$ of (expected) size $n_2 = 10,000$: (i) SRS and (ii) Poisson sampling with inclusion probabilities satisfying $\pi_{2i} \propto \left\{ 1 + \exp\left( 0.2x_{i1} + 0.1x_{i2} - 0.6 \right) \right\}^{-1}$ and $\sum_{i=1}^{N} \pi_{2i} = n_2$. It is worthy pointing out that the sampling design for the internal sample is informative (Pfeffermann, 1993) under Poisson sampling, so ignoring the design feature may result in erroneous inference.
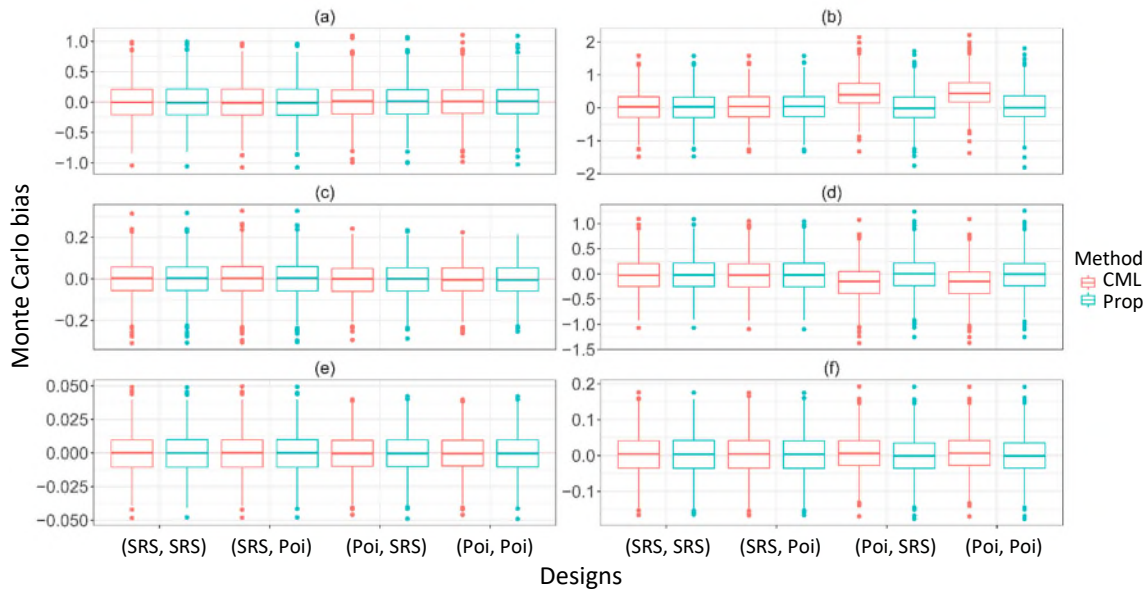
For the proposed estimator, we consider a working reduced model, $\sum_{i \in \mathcal{S}_2} \pi_{i2}^{-1} \left( y_i - \alpha_0 - \alpha_1 x_{i1} \right) \left( 1, x_{i1} \right)^{\top} = \mathbf{0}$, whose solution is denoted as $\hat{\boldsymbol{\alpha}}_2$. Based on the external sample $\mathcal{S}_2$, we assume that a point estimator $\hat{\boldsymbol{\alpha}}_2$ and its variance estimator $\mathbf{V}_2 = \hat{V}(\hat{\boldsymbol{\alpha}}_2)$ are available as discussed in Section 3. Linearization is adopted to obtain a variance estimator $\mathbf{V}_2$; see the proof of Theorem 1 in Appendix A for details.

In the simulation study, the proposed estimator is compared with the constrained maximum likelihood (CML) estimator (Chatterjee et al., 2016). We assume a normal distribution for the likelihood function, i.e., $y_i \mid \mathbf{x}_i \sim N\left\{ \left( 1, \mathbf{x}_i^{\top} \right) \boldsymbol{\beta}, \sigma_{\text{full}}^2 \right\}$. We also suppose that an analyst assumes $y_i \mid \mathbf{x}_{i1} \sim N\left\{ \left( 1, x_{i1} \right) \boldsymbol{\alpha}, \sigma_{\text{red}}^2 \right\}$ for the working reduced model. See Section S2 of the Supplementary Material for the computation details. We consider the CML estimator under the setting where the extra information of $\left( y_i, x_{i1} \right)$ is available for an external sample, not for the entire population.

We conduct $M = 1,000$ Monte Carlo simulations, and Figures 6.1 and 6.2 show the Monte Carlo bias of the proposed and CML estimators for the homogeneous and heterogeneous variance setups, respectively. From Figure 6.1, when the variance of the error term is homogeneous and the internal sample is generated by SRS, the proposed estimator performs approximately the same as CML estimator in terms of Monte Carlo bias and variance. However, when the auxiliaries are correlated and the internal sample is generated by Poisson sampling, the CML estimator is questionable, since its model is wrongly specified under the informative Poisson sampling design. For example, the Monte Carlo bias of the CML estimator is not negligible when estimating $\beta_0$ and $\beta_1$. Because the proposed estimator incorporates the design features, its performance is satisfactory for all setups. As shown in Figure 6.2, even when the internal sample is generated by SRS, the CML estimator is slightly less efficient than the proposed estimator. The reason is that the CML estimator fails to take the heterogeneous variance into consideration, but the proposed estimator does not make any distribution assumption. When the internal sample is generated by an informative Poisson sampling design, the CML performs poorly, since it is not unbiased, and since its variance is larger than the proposed estimator.
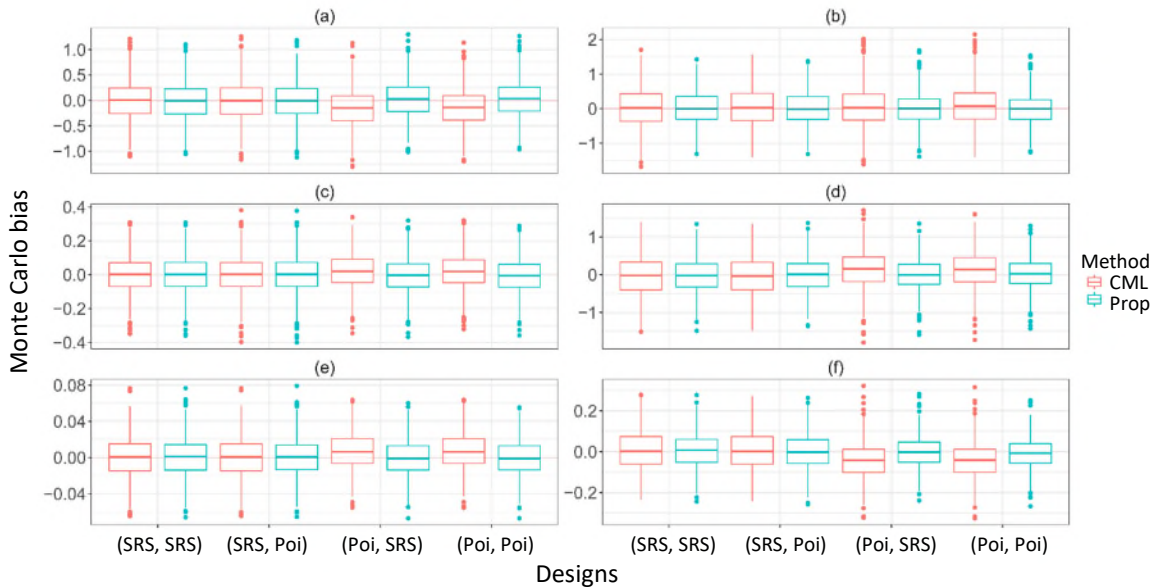
Table 6.1 shows the coverage rate of a 95% confidence interval for the proposed estimator under different settings. Chatterjee et al. (2016) only investigated the theoretical properties of their estimator when the population-level information is available. Thus, no interval estimator can be provided if only an external sample is available. By Table 6.1, we conclude that the coverage rates of the confidence intervals are all close to its nominal truth 0.95 under different settings. One possible reason for this phenomenon is that the proposed estimator is model free, so the proposed model is more robust and can be used under complex sampling designs.

**Figure 6.1  Monte Carlo bias of the proposed and CML estimators based on 1,000 Monte Carlo simulations under the homogeneous variance setup.**



The first to the third rows stand for the Monte Carlo bias for estimating $\beta_0$, $\beta_1$ and $\beta_2$, respectively. The three plots, including (a), (c) and (e), in the left column show the results when the auxiliary variables are independently generated, and those, including (b), (d) and (f), in the right column are for the case when the auxiliaries are dependent. "CML" and "Prop" stands for the CML estimator and the proposed estimator, respectively. The first design in the parenthesis is used to generate the internal sample $\mathcal{S}_1$, and the second one to generate the external sample $\mathcal{S}_2$. "SRS" and "Poi" represents Sampling random sampling and Poisson sampling.

**Figure 6.2  Monte Carlo bias of the proposed and CML estimators based on 1,000 Monte Carlo simulations under the heterogeneous variance setup.**



The first to the third rows stand for the Monte Carlo bias for estimating $\beta_0$, $\beta_1$ and $\beta_2$, respectively. The three plots, including (a), (c) and (e), in the left column show the results when the auxiliary variables are independently generated, and those, including (b), (d) and (f), in the right column are for the case when the auxiliaries are dependent. "CML" and "Prop" stands for the CML estimator and the proposed estimator, respectively. The first design in the parenthesis is used to generate the internal sample $\mathcal{S}_1$, and the second one to generate the external sample $\mathcal{S}_2$. "SRS" and "Poi" represents Sampling random sampling and Poisson sampling.

**Table 6.1**
**Coverage rate of a 95% confidence interval by the proposed method based on 1,000 Monte Carlo simulations under different setups**

| $\mathcal{S}_1$ Des | $\mathcal{S}_2$ Des | Independent | | | Dependent | | |
| | | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_0$ | $\beta_1$ | $\beta_2$ |
|---|---|---|---|---|---|---|---|
| Homo | SRS | | | | | | |
| | SRS | 0.948 | 0.952 | 0.939 | 0.945 | 0.948 | 0.934 |
| | Poi | 0.945 | 0.951 | 0.938 | 0.946 | 0.946 | 0.934 |
| | Poi | | | | | | |
| | SRS | 0.957 | 0.966 | 0.949 | 0.935 | 0.943 | 0.940 |
| | Poi | 0.962 | 0.964 | 0.951 | 0.936 | 0.943 | 0.938 |
| Hete | SRS | | | | | | |
| | SRS | 0.944 | 0.942 | 0.933 | 0.933 | 0.925 | 0.935 |
| | Poi | 0.949 | 0.942 | 0.935 | 0.935 | 0.934 | 0.931 |
| | Poi | | | | | | |
| | SRS | 0.959 | 0.955 | 0.935 | 0.948 | 0.950 | 0.941 |
| | Poi | 0.961 | 0.956 | 0.944 | 0.952 | 0.949 | 0.946 |

Note: "Homo" and "Hete" stands for the homogeneous and heterogeneous variance for the error term, respectively. "$\mathcal{S}_1$ Des" and "$\mathcal{S}_2$ Des" show the sampling design used to generate the internal sample $\mathcal{S}_1$ and the external sample $\mathcal{S}_2$. "SRS" and "Poi" stands for Sampling random sampling and Poisson sampling, respectively. "Independent" and "Dependent" correspond to the cases when the auxiliary variables are independent and dependent, respectively.

An additional simulation with a logistic regression setup is relegated to Section S3 of the Supplementary Material, and similar conclusions can be reached.

# 7. Application study

## 7.1 Data description and problem formulation

As an application example, we apply the proposed method to analyze a subset of the data from the Korea National Health and Nutrition Examination Survey (KNHANES). The annual survey includes approximately 5,000 individuals each year and collects information regarding health-related behaviors by interviews, basic health conditions by physical and blood tests, and dietary intake by nutrition survey. The sampling design of KNHANES is a stratified sampling using age, sex, and region as stratification variables. The final sampling weights are computed via nonresponse adjustment and post-stratification, then provided to data users with survey variables.

To improve the efficiency of data analysis with KNHANES of size $n_1 = 4,929,$ we used an external public database provided by the National Health Insurance Sharing Service (NHISS) in Korea. The big data provided by NHISS contain about $n_2 = 1,000,000$ individuals with health-related information, some of whose variables are a subset of variables in KNHANES.

These data structures, with the small $n_1,$ the large $n_2,$ and the big data having a subset of variables in the internal sample, are suited well to the setting we addressed in Section 2. However, there is another complication in applying the proposed method to the real application. In the NHISS data, its selection probabilities are unknown, so the design consistent estimator $\hat{\alpha}_2$ in (3.9) is unavailable. Section 7.2

addresses this issue by using a propensity weighting approach and Section 7.3 presents the analysis results of the application study.

## 7.2 Propensity weighting for external data with unknown selection probability

We now consider an extension of the proposed method to the case where the external sample $\mathcal{S}_2$ is a big data with unknown selection probabilities. In this case, the working model for $E\left(Y_i \mid \mathbf{x}_{i1}\right) = m\left(\boldsymbol{\alpha}^\top \mathbf{x}_{i1}\right)$ may not hold for the sample $S_2$. Nonetheless, we may still solve

$$\sum_{i \in \mathcal{S}_2} \left\{ y_i - m\left(\boldsymbol{\alpha}^\top \mathbf{x}_{i1}\right) \right\} \mathbf{x}_{i1} = \mathbf{0} \tag{7.1}$$

to obtain $\hat{\alpha}_0$ and $\hat{\alpha}_1$. If the sampling mechanism for $\mathcal{S}_2$ is ignorable or non-informative, then the solution of (7.1) is unbiased; otherwise, the resulting estimator is biased.

To remove the selection biases in the big data estimate, Kim and Wang (2019) suggested using propensity score weights in (7.1) to obtain an unbiased estimator of $\boldsymbol{\alpha}$. To construct the propensity score weights, we employ a nonignorable nonresponse model, $P\left(\delta_i = 1 \mid \mathbf{x}_{i1}, y_i\right) = \pi\left(\mathbf{x}_{i1}, y_i; \boldsymbol{\phi}\right)$, where $\delta_i = 1$ if $i \in S_2$ and zero otherwise. Note that we can express $\pi\left(\mathbf{x}_{i1}, y_i\right)^{-1} = 1 + \left(N_0 / N_1\right) r\left(\mathbf{x}_{i1}, y_i\right)$, where $r\left(\mathbf{x}_{i1}, y_i\right) = f\left(\mathbf{x}_{i1}, y_i \mid \delta_i = 0\right) / f\left(\mathbf{x}_{i1}, y_i \mid \delta_i = 1\right)$ is the density ratio function with $N_1 = \sum_{i=1}^{N} \delta_i$ and $N_0 = N - N_1$. Using the motivation of Wang and Kim (2021), we may assume a log-linear density ratio model, $\log\left\{r\left(x_{i1}, y_i; \boldsymbol{\phi}\right)\right\} = \phi_0 + \phi_1 x_{i1} + \phi_2 y_i$. The maximum entropy estimator of $\boldsymbol{\phi}$ is obtained by solving $\left(1 / N_1\right) \sum_{i=1}^{N} \delta_i \exp\left(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i\right)\left(1, x_{i1}, y_i\right) = \left(1, \hat{\bar{x}}_1, \hat{\bar{y}}\right)$, where $\left(\hat{\bar{x}}_1, \hat{\bar{y}}\right) = \left(1 / \hat{N}_0\right)\left\{\sum_{i \in \mathcal{S}_1} d_i\left(x_{i1}, y_i\right) - \sum_{i=1}^{N} \delta_i\left(x_{i1}, y_i\right)\right\}$, $\hat{N}_0 = \sum_{i \in \mathcal{S}_1} d_i - N_1$, and $\mathcal{S}_1$ is the internal sample. Once $\hat{\boldsymbol{\phi}}$ is obtained, we can construct $\hat{\pi}\left(x_{i1}, y_i\right)$ and solve

$$\sum_{i \in \mathcal{S}_2} \frac{1}{\hat{\pi}\left(x_{i1}, y_i\right)} \left\{ y_i - m\left(\alpha_0 + \alpha_1 x_{i1}\right) \right\}\left(1, x_{i1}\right) = \left(0, 0\right) \tag{7.2}$$

to obtain $\hat{\boldsymbol{\alpha}}_2 = \left(\hat{\alpha}_0, \hat{\alpha}_1\right)$.

In addition, we can use the internal sample $\mathcal{S}_1$ to fit the same working model to obtain $\hat{\boldsymbol{\alpha}}_1$. After that, we obtain $\hat{\boldsymbol{\alpha}}^*$ using (3.9) and apply the proposed calibration weighting method to combine information from the big data. In practice, $\mathbf{V}_2$ in (3.9) is difficult to compute, but it is negligibly small if the sample size for $\mathcal{S}_2$ is huge. In this case, we may simply use $\hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\alpha}}_2$ in the calibration problem.

## 7.3 Application study results: Korea National Health and Nutrition Examination Survey

In this application study, we use $n_1 = 4,929$ records of KNHANES data that have no missing values in four variables: Total cholesterol, Hemoglobin, Triglyceride, and high-density lipoprotein (HDL)

cholesterol. For demonstration purpose, we assume that an analyst is interested in conducing the following linear regression analysis,

$$E\left(\text{Total Cholesterol}_i \mid \mathbf{x}_i\right) = \beta_0 + \beta_1 \text{Hemoglobin}_i + \beta_2 \text{Triglyceride}_i + \beta_3 \text{HDL}_i \quad \text{for} \quad i \in \mathcal{S}_1;$$

check Section S4 of the Supplementary Material for details about the linearity assumption. In our data, the biggest absolute value of the pairwise correlation among covariates is $-0.40$ observed between Triglyceride and HDL cholesterol, which is similar to a scenario in Section 6 where the covariates were highly correlated. The big external data consist of $n_2 = 1,000,000$ records of NHISS data with fully observed items in Total cholesterol, Hemoglobin, and Triglyceride. The assumed working reduced model is

$$E\left(\text{Total Cholesterol}_i \mid \mathbf{x}_{i1}\right) = \alpha_0 + \alpha_1 \text{Hemoglobin}_i + \alpha_2 \text{Triglyceride}_i \quad \text{for} \quad i \in \mathcal{S}_1 \cup \mathcal{S}_2.$$

In this application study, we implement our proposed methods with the external sample, where $\hat{\boldsymbol{\alpha}}_2$ is used instead of $\boldsymbol{\alpha}^*$ that is unavailable as we do not have information regarding the entire population. With the external sample whose selection probabilities are unknown, we prepare two versions of proposed methods: (i) considering $\mathcal{S}_2$ as SRS, i.e., without propensity weighting, and (ii) with the propensity weighting adjustment introduced in Section 7.2. For the propensity weighting, we fit the log-linear density ratio model to the external data, $\log\{r(\mathbf{x}_{i1}, y_i; \boldsymbol{\phi})\} = \phi_0 + \phi_1 \text{Hemoglobin}_i + \phi_2 \text{Triglyceride}_i + \phi_3 \text{Total Cholesterol}_i$, calculate $\hat{\pi}(\mathbf{x}_{i1}, y_i)$ given $\hat{\boldsymbol{\phi}}$, then solve

$$\sum_{i \in \mathcal{S}_2} \frac{1}{\hat{\pi}(x_{i1}, y_i)} \{y_i - m(\alpha_0 + \alpha_1 x_{i1})\}(1, x_{i1}) = (0, 0)$$

to obtain $\hat{\boldsymbol{\alpha}}_2$. The above logistic regression model is commonly assumed in the literature; see Elliott and Valliant (2017), Chen, Li and Wu (2020), Wang and Kim (2021) and the references within for details. Since the CML estimator fails to incorporate the design features, it is not considered in the application section. The performances of proposed methods are compared with the reference method that uses the internal sample $\mathcal{S}_1$ only to get weighted least square estimates considering the sampling weights.

Figure 7.1 shows the point estimates and the 95% confidence intervals of $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right)$ for each method. The proposed methods show smaller variances for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ than using the internal sample only. This result coincides with our findings in the simulation studies of the previous section. For $\beta_2$, the estimator of the proposed method without propensity weighting shows a systematic difference from the other two estimators. When the propensity weighting adjustment is coupled with the proposed method, its confidence interval of $\beta_2$ is contained by that of using the internal sample only. This result implies that the systematic bias due to the disregard of the sampling probabilities is addressed by the propensity weighting adjustment. No efficiency gain in estimating $\beta_3$ was expected as the external data contain information of $x_{i1}$ (Hemoglobin) and $x_{i2}$ (Triglyceride), not $x_{i3}$ (HDL).

**Figure 7.1** **Comparison of the regression analysis for** $E\left(\text{Total Cholesterol}_i \mid \mathbf{x}_i\right) = \beta_0 + \beta_1 \text{Hemoglobin}_i + \beta_2 \text{Triglyceride}_i + \beta_3 \text{HDL}_i$ **using the internal data from Korea National Health and Nutrition Examination Survey supported by the big external data from the National Health Insurance Sharing Service database.**
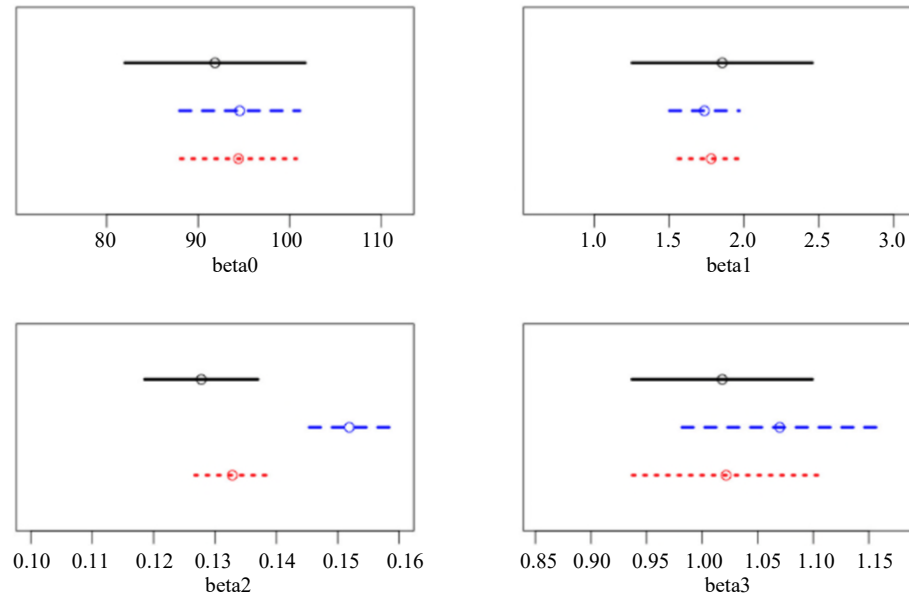


For each panel, circles are point estimates and lines are their 95% confidence intervals for using the internal sample $\mathcal{S}_1$ only with the weighted least square (top solid line), the proposed method without adjustment (middle dashed line), and the proposed method with propensity score weighting adjustment (bottom dotted line).

# 8.  Conclusion

Incorporating external data sources into the regression analysis of the internal sample is an important practical problem. We have addressed this problem using a novel application of the information projection (Csiszár and Shields, 2004) and the model calibration weighting (Wu and Sitter, 2001). The proposed method is directly applicable to survey sampling and can be easily extended to multiple data integration. The proposed method is easy to implement and does not require direct access to external data. As long as the estimated regression coefficients and their standard errors for the working reduced model are available, we can incorporate the extra information into our analysis.

There are several possible directions on future research extensions. First, a Bayesian approach can be developed under the same setup. One may use the Bayesian empirical likelihood method of Zhao, Ghosh, Rao and Wu (2020) in this setup. The proposed method can potentially be used to combine the randomized clinical trial data with big real-world data (Yang, Zheng and Wang, 2020); such extensions will be presented elsewhere. It will be also interesting to connect the proposed approach to two-phase (double) sampling design whose efficient design and estimation has been recently studied actively (Rivera-Rodriguez, Spiegelman and Haneuse, 2019; Rivera-Rodriguez, Haneuse, Wang and Spiegelman, 2020; Wang, Williams, Chen and Chen, 2020). The data structure of the two-phase sampling with the large-$n$, small-$p$ first stage sample and the small-$n$, large-$p$ second stage sample is well suited to the set-up assumed by the suggested model calibration approach.

# Acknowledgements

# Supplement

The supplementary material can be found in the document https://arxiv.org/abs/2107.06448, and it contains special case under simple random sampling (S1), implementation of Chatterjee et al. (2016) (S2), an additional simulation study (S3), and validation for the linearity assumption for the KNHANES dataset (S4).

# Appendix

## A. Proof of Theorem 1

**Lemma A1.** *Suppose that conditions C1, C2e and C3 hold. Then,* $\left\| \hat{\boldsymbol{\lambda}} \right\| = O_p\left(n^{-1/2}\right)$.

*Proof of Lemma A1. Denote* $\hat{\boldsymbol{\lambda}} = \rho \boldsymbol{\theta}$, *where* $\rho = \left\| \hat{\boldsymbol{\lambda}} \right\|$ *and* $\boldsymbol{\theta} = \rho^{-1}\hat{\boldsymbol{\lambda}}$ *is a vector of unit length. Then, we have*

$$
\begin{aligned}
\mathbf{0} &= \left| \sum_{i \in \mathcal{S}_1} \frac{\tilde{d}_i}{1 - \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right)} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \right| \\[2mm]
&= \left| \boldsymbol{\theta}^{\mathrm{T}} \sum_{i \in \mathcal{S}_1} \frac{\tilde{d}_i}{1 - \rho \boldsymbol{\theta}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right)} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \right| \\[2mm]
&= \left| \sum_{i \in \mathcal{S}_1} \tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) + \rho \sum_{i \in \mathcal{S}_1} \frac{\tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \left\{ \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \right\}^{\mathrm{T}} \boldsymbol{\theta}}{1 - \rho \boldsymbol{\theta}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right)} \right| \\[2mm]
&\geq \frac{\rho}{1 + \rho Z_S} \left| \sum_{i \in \mathcal{S}_1} \tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \left\{ \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \right\}^{\mathrm{T}} \boldsymbol{\theta} \right| - \left| \sum_{i \in \mathcal{S}_1} \tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \right|, \quad \text{(A.1)}
\end{aligned}
$$

where the first equality holds since $g\left(\hat{\boldsymbol{\eta}}\right) = \mathbf{0}$, and the last inequality holds by the triangular inequality.

By C2e and the Rayleigh-Ritz Theorem (Horn and Johnson, 2012, Section 4.2), there exists a constant $\sigma_0 > 0$ such that

$$\sum_{i \in S_1} \tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \left\{ \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \right\}^{\mathrm{T}} \boldsymbol{\theta} > \sigma_0 + o_p (1). \tag{A.2}$$

By C3 and the Slutsky's theorem, we have

$$\sum_{i \in S_1} \tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) = O_p \left( n^{-1/2} \right). \tag{A.3}$$

Thus, by C1 and (A.1)-(A.3), we have proved Lemma A1.

**Lemma A2.** *Suppose that conditions C1, C2a-C2e and C3 hold. Then,* $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{0N} = o_p (1).$

*Proof of Lemma A2.* By Lemma A1 and C1, we conclude that

$$\max_{i \in S_1} \left| \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \right| \leq \max_{i \in S_1} \left\| \hat{\boldsymbol{\lambda}} \right\| \left\| \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \right\|$$

$$= \left\| \hat{\boldsymbol{\lambda}} \right\| \max_{i \in S_1} \left\| \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \right\| = o_p (1). \tag{A.4}$$

First, we show that

$$\sum_{i \in S_1} \frac{\tilde{d}_i}{1 - \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right)} \mathbf{U}_1 \left( \boldsymbol{\beta}; \mathbf{x}_i, y_i \right) - \sum_{i \in S_1} \tilde{d}_i \mathbf{U}_1 \left( \boldsymbol{\beta}; \mathbf{x}_i, y_i \right) \to \mathbf{0} \tag{A.5}$$

in probability uniformly for $\boldsymbol{\beta} \in \Omega.$ By (A.4), we have

$$\left\| \sum_{i \in S_1} \tilde{d}_i \left\{ \frac{1}{1 - \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right)} - 1 \right\} \mathbf{U}_1 \left( \boldsymbol{\beta}; \mathbf{x}_i, y_i \right) \right\|$$

$$= \left\| \sum_{i \in S_1} \tilde{d}_i \left\{ \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) + o_p \left( \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \right) \right\} \mathbf{U}_1 \left( \boldsymbol{\beta}; \mathbf{x}_i, y_i \right) \right\|$$

$$\leq \left( 1 + o_p (1) \right) \max_{i \in S_1} \left| \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \right| \left\| \sum_{i \in S_1} \tilde{d}_i \mathbf{U}_1 \left( \boldsymbol{\beta}; \mathbf{x}_i, y_i \right) \right\|. \tag{A.6}$$

By C2a-C2b, there exists a constant $C_{u1} > 0$ such that $\sup_{\boldsymbol{\beta} \in \Omega} \left\| \mathbf{U}_0 (\boldsymbol{\beta}) \right\| < C_{u1}.$ Since $\sum_{i \in S_1} \tilde{d}_i \mathbf{U}_1 \left( \boldsymbol{\beta}; \mathbf{x}_i, y_i \right)$ converge uniformly to $\mathbf{U}_0 (\boldsymbol{\beta})$ in probability, we conclude that

$$\left\| \sum_{i \in S_1} \tilde{d}_i \mathbf{U}_1 \left( \boldsymbol{\beta}; \mathbf{x}_i, y_i \right) \right\| < C_{u1} + o_p (1) \tag{A.7}$$

uniformly over $\Omega.$ By (A.4) and (A.6)-(A.7), we have validated (A.5).

By C2b and (A.5), we conclude that $\sum_{i \in S_1} \tilde{d}_i \left\{ 1 - \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \right\}^{-1} \mathbf{U}_1 \left( \boldsymbol{\beta}; \mathbf{x}_i, y_i \right)$ converges uniformly to $U_0 (\boldsymbol{\beta})$ in probability. Denote $Q_0 (\boldsymbol{\beta}) = -U_0 (\boldsymbol{\beta})^2$ and $Q_s (\boldsymbol{\beta}) = - \left[ \sum_{i \in S_1} \tilde{d}_i \left\{ 1 - \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \right\}^{-1} \mathbf{U}_1 \left( \boldsymbol{\beta}; \mathbf{x}_i, y_i \right) \right]^2.$ Then, $\boldsymbol{\beta}_p$ uniquely maximizes $Q_0 (\boldsymbol{\beta})$ by (C2b), and $\hat{\boldsymbol{\beta}}$ maximizes $Q_s (\boldsymbol{\beta}).$ In addition, $Q_s (\boldsymbol{\beta})$ converge uniformly to $Q_0 (\boldsymbol{\beta})$ in probability over the compact set $\Omega.$ Thus, by C2a and Theorem 2.1 of Engle and McFadden (1994, Chapter 36), we have finished the proof for Lemma A2.

*Proof of Theorem 1.* By Lemmas A1-A2, we have shown that

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}_0 + o_p(1), \tag{A.8}$$

where $\hat{\boldsymbol{\eta}}^{\mathrm{T}} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\lambda}}^{\mathrm{T}})$, $\boldsymbol{\eta}_0^{\mathrm{T}} = (\boldsymbol{\beta}_{0N}^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}})$, and $\boldsymbol{0}$ is a vector of zero with the same length of $\hat{\boldsymbol{\lambda}}$.

By (A.8) and the Taylor expansion, we have

$$
\begin{aligned}
\boldsymbol{0} = g(\hat{\boldsymbol{\eta}}) &= g(\boldsymbol{\eta}_0) + \frac{\partial g}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + o_p\left(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\|\right) \\
&= \begin{pmatrix} \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}_{0N};\mathbf{x}_i, y_i) \\ \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i}, y_i) \end{pmatrix} \\
&\quad + \begin{pmatrix} \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \dfrac{\partial \mathbf{U}_1(\boldsymbol{\beta}_{0N};\mathbf{x}_i, y_i)}{\partial\boldsymbol{\beta}^{\mathrm{T}}} & \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}_{0N};\mathbf{x}_i, y_i)\left\{\mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i}, y_i)\right\}^{\mathrm{T}} \\ \boldsymbol{0} & \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \left\{\mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i}, y_i)\right\}^{\otimes 2} \end{pmatrix}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \\
&\quad + o_p\left(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\|\right). \tag{A.9}
\end{aligned}
$$

By (C3), we have

$$n^{1/2}\begin{pmatrix} \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}_{0N};\mathbf{x}_i, y_i) \\ \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i}, y_i) \end{pmatrix} \to \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_u) \tag{A.10}$$

in distribution. By (C2e)-(C2c), we conclude that

$$\begin{pmatrix} \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \dfrac{\partial \mathbf{U}_1(\boldsymbol{\beta}_{0N};\mathbf{x}_i, y_i)}{\partial\boldsymbol{\beta}^{\mathrm{T}}} & \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}_{0N};\mathbf{x}_i, y_i)\left\{\mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i}, y_i)\right\}^{\mathrm{T}} \\ \boldsymbol{0} & \sum\limits_{i\in\mathcal{S}_1} \tilde{d}_i \left\{\mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i}, y_i)\right\}^{\otimes 2} \end{pmatrix} \to \boldsymbol{I} \tag{A.11}$$

in probability, where

$$\boldsymbol{I} = \begin{pmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{0} & \boldsymbol{I}_{22} \end{pmatrix}.$$

By (A.9)-(A.11), we conclude that

$$n^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \to \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta) \tag{A.12}$$

in distribution, where $\boldsymbol{\Sigma}_\eta = \boldsymbol{I}^{-1}\boldsymbol{\Sigma}_u(\boldsymbol{I}^{-1})^{\mathrm{T}}$.

## B.  Proof of Corollary 1

Since $\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x}, y)$ is given, it is enough to consider

$$\boldsymbol{\Sigma}_{11} - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{12}^{\mathrm{T}} + \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{12}^{\mathrm{T}},$$

the asymptotic variance of $\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)$, where $\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) = n^{1/2}\sum_{i \in \mathcal{S}_1}\tilde{d}_i\mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i)$ and $\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*) = n^{1/2}\sum_{i \in \mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i)$.

Consider

$$
\begin{aligned}
\mathrm{Var}&\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\} \\
&= E\left[\mathrm{Var}\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\mid \mathcal{A}_N\}\right] + \mathrm{Var}\left[E\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\mid \mathcal{A}_N\}\right] \\
&\succeq E\left[\mathrm{Var}\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N})\}\right],
\end{aligned}
$$

where $\mathcal{A}_N = \{(\mathbf{x}_{i1}, y) : i \in \mathcal{S}_1\}$, $A \succeq B$ is equivalent to that $A - B$ is non-negatively definitive for two matrices $A$ and $B$ with the same dimension, and the last inequality holds since $\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)$ is non-stochastic conditional on $\mathcal{A}_N$. Thus, $\mathrm{Var}\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\}$ achieves minimum if $\mathrm{Var}\left[E\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\mid \mathcal{A}_N\}\right] = 0$, which is induced by the condition $\boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_1, y) = E\{\mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}, y)\mid \mathbf{x}_1, y\}$.

## C.   Proof of Theorem 2

Before proving Theorem 2, we need the following result.

**Lemma A3.** *Suppose that conditions C1, C3-C5 hold. Then, we have*

$$
\boldsymbol{\alpha}^* - \boldsymbol{\alpha}_N^* = O_p(n^{-1/2}).
$$

*Proof of Lemma A3.* Since $\hat{\boldsymbol{\alpha}}_2$ is obtained by an independent external survey, we conclude that the variance of $\hat{\boldsymbol{\alpha}}^*$ can be estimated by $(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1})^{-1}$. Thus, the order of the variance of $\hat{\boldsymbol{\alpha}}^*$ is determined by the less efficient estimator between $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$. If we showed

$$
\hat{\boldsymbol{\alpha}}_1 - \hat{\boldsymbol{\alpha}}_N^* = O_p(n^{-1/2}), \tag{C.1}
$$

we could have $\mathbf{V}_1 = O_p(n^{-1})$ by (C5b). Since $\hat{\boldsymbol{\alpha}}^*$ is at least as efficient as $\hat{\boldsymbol{\alpha}}_1$, we have completed the proof of Lemma A3.

Thus, it remains to show (C.1). By C4 and C5, we have

$$
\begin{aligned}
\mathbf{0} &= \sum_{i \in \mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\hat{\boldsymbol{\alpha}}_1; \mathbf{x}_{1i}, y_i) \\
&= \sum_{i \in \mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i) + \left\{\frac{\partial}{\partial \boldsymbol{\alpha}^\top}\sum_{i \in \mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\tilde{\boldsymbol{\alpha}}; \mathbf{x}_{1i}, y_i)\right\}(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_N^*), \tag{C.2}
\end{aligned}
$$

where $\tilde{\boldsymbol{\alpha}}$ lies on the segment joining $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_N^*$. By C3-C5 and (C.2), we conclude that

$$
\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_N^* = -\boldsymbol{I}_0^{-1}\sum_{i \in \mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i) + o_p(n^{-1/2}). \tag{C.3}
$$

Thus, by C3 and (C.3), we have shown (C.1).

*Proof of Theorem 2.* Consider

$$\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 \left( \hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{1i}, y_i \right)$$

$$= \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) + \left\{ \frac{\partial}{\partial \boldsymbol{\alpha}^{\mathsf{T}}} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 \left( \tilde{\boldsymbol{\alpha}}; \mathbf{x}_{1i}, y_i \right) \right\} \left( \hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}_N^* \right)$$

$$= \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) + \boldsymbol{\mathcal{I}}_0 \left( \mathbf{V}_1^{-1} + \mathbf{V}_2^{-1} \right)^{-1} \mathbf{V}_2^{-1} \left( \hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^* \right)$$

$$+ \boldsymbol{\mathcal{I}}_0 \left( \mathbf{V}_1^{-1} + \mathbf{V}_2^{-1} \right)^{-1} \mathbf{V}_1^{-1} \left( \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_N^* \right) + o_p \left( n^{-1/2} \right),$$

$$= \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) + \boldsymbol{\mathcal{I}}_0 \left( \mathbf{V}_1^{-1} + \mathbf{V}_2^{-1} \right)^{-1} \mathbf{V}_2^{-1} \left( \hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^* \right)$$

$$- \boldsymbol{\mathcal{I}}_0 \left( \mathbf{V}_1^{-1} + \mathbf{V}_2^{-1} \right)^{-1} \mathbf{V}_1^{-1} \boldsymbol{\mathcal{I}}_0^{-1} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) + o_p \left( n^{-1/2} \right),$$

$$= \boldsymbol{\mathcal{I}}_0 \left( \mathbf{V}_1^{-1} + \mathbf{V}_2^{-1} \right)^{-1} \mathbf{V}_2^{-1} \left( \hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^* \right) + \boldsymbol{\mathcal{I}}_0 \left( \mathbf{V}_1^{-1} + \mathbf{V}_2^{-1} \right)^{-1} \mathbf{V}_2^{-1} \boldsymbol{\mathcal{I}}_0^{-1} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right)$$

$$+ o_p \left( n^{-1/2} \right)$$

$$= \boldsymbol{\mathcal{I}}_0 \mathbf{W} \left( \hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^* \right) + \boldsymbol{\mathcal{I}}_0 \mathbf{W} \boldsymbol{\mathcal{I}}_0^{-1} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) + o_p \left( \kappa(n) \right), \tag{C.4}$$

where $\tilde{\boldsymbol{\alpha}}$ lies on the segment joining $\hat{\boldsymbol{\alpha}}^*$ and $\boldsymbol{\alpha}_N^*$, the second equality holds by C4 and Lemma A3, the third equality holds by (C.3), the last equality holds by C5d, $\kappa(n) = \gamma(n)$ if $\gamma(n) n^{1/2} \to \infty$ and $\kappa(n) = n^{-1/2}$ otherwise, and $\gamma(n)$ is the convergence order of $\left( \hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^* \right)$ in (C5e).

If there exists a non-stochastic matrix $\boldsymbol{\Sigma}_c$ such that $n \mathbf{V}_2 = \boldsymbol{\Sigma}_c + o_p(1)$, then $\left( \hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^* \right) = O_p \left( n^{1/2} \right)$ and $\mathbf{W}$ is not a zero matrix. Then, by (C.4), we have

$$n^{1/2} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \begin{pmatrix} \mathbf{U}_1 \left( \boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i \right) \\ \mathbf{U}_2 \left( \hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{1i}, y_i \right) \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} \\ n^{1/2} \boldsymbol{\mathcal{I}}_0 \mathbf{W} \left( \hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^* \right) \end{pmatrix} + n^{1/2} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \begin{pmatrix} \mathbf{U}_1 \left( \boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i \right) \\ \boldsymbol{\mathcal{I}}_0 \mathbf{W} \boldsymbol{\mathcal{I}}_0^{-1} \mathbf{U}_2 \left( \boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i \right) \end{pmatrix} + o_p(1). \tag{C.5}$$

Since the external sample is independent with the internal sample and $\boldsymbol{\Sigma}_c$ is the asymptotic variance of $n^{1/2} \left( \hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^* \right)$, by (C3), (C5e) and (C.5), we conclude that

$$n^{1/2} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \begin{pmatrix} \mathbf{U}_1 \left( \boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i \right) \\ \mathbf{U}_2 \left( \hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{1i}, y_i \right) \end{pmatrix} \to \mathcal{N} \left( 0, \tilde{\boldsymbol{\Sigma}}_u \right),$$

where $\tilde{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11}$, $\tilde{\boldsymbol{\Sigma}}_{12} = \boldsymbol{\Sigma}_{12} \left( \boldsymbol{\mathcal{I}}_0^{-1} \right)^{\mathsf{T}} \mathbf{W}^{\mathsf{T}} \boldsymbol{\mathcal{I}}_0^{\mathsf{T}}$, $\tilde{\boldsymbol{\Sigma}}_{21} = \tilde{\boldsymbol{\Sigma}}_{12}^{\mathsf{T}}$, and $\tilde{\boldsymbol{\Sigma}}_{22} = \boldsymbol{\mathcal{I}}_0 \mathbf{W} \left\{ \boldsymbol{\Sigma}_c + \boldsymbol{\mathcal{I}}_0^{-1} \boldsymbol{\Sigma}_{22} \left( \boldsymbol{\mathcal{I}}_0^{-1} \right)^{\mathsf{T}} \right\} \mathbf{W}^{\mathsf{T}} \boldsymbol{\mathcal{I}}_o^{\mathsf{T}}$. Thus, we have proved the first case of Theorem 2.

If $\mathbf{W} = 0$, then $\gamma(n) n^{1/2} \to \infty$ and the rate of $\kappa(n)$ is slower than $n^{-1/2}$ in (C.4). Thus, the remainder term of (C.4) is no longer $o_p \left( n^{-1/2} \right)$ for $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2 \left( \hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{1i}, y_i \right)$. Instead, for this case, we investigate the asymptotic order of $\left( \mathbf{V}_1^{-1} + \mathbf{V}_2^{-1} \right)^{-1} \mathbf{V}_2^{-1}$ first. By C3, C5b and (C.3), we have

$$\mathbf{V}_1 \asymp n^{-1} \quad \text{and} \quad \mathbf{V}_2 \asymp \gamma(n)^{-2} \tag{C.6}$$

in probability. Thus, (C.6) leads to

$$\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_2^{-1} \asymp n^{-1}\gamma(n)^2 \tag{C.7}$$

in probability by the fact that $\gamma(n)\, n^{1/2} \to \infty.$ Thus, by (C5e) and (C.7), we have

$$\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_2^{-1} \left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*\right) \asymp n^{-1}\gamma(n) \tag{C.8}$$

in probability. By $\gamma(n)n^{1/2} \to \infty,$ (C.7) and (C.8), we have shown

$$\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_2^{-1} = o_p(1),$$
$$\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_2^{-1} \left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*\right) = o_p(n^{-1/2})$$

in probability. Thus, by the fourth equality of (C.4), we can show that

$$\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{1i}, y_i\right) = o_p(n^{-1/2}),$$

and we have proved the third case of Theorem 2.

# References

Chatterjee, N., Chen, Y.H., Maas, P. and Carroll, R.J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111, 107-117.

Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9(2), 385-406.

Chen, S., and Kim, J.K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, 24(1), 335-355.

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

Chen, Y.H., and Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B*, 62, 449-460.

Csiszár, I., and Shields, P.C. (2004). Information theory and statistics: A tutorial. *Foundations and Trands in Communications and Information Theory*, 1, 417-528.

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

Engle, R.F., and McFadden, D.L. (1994). *Handbook of Econometrics: Volume IV*. Amsterdam.

Fuller, W.A. (2009). *Sampling Statistic*. Hoboken, New York: John Wiley & Sons, Inc.

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029-1054.

Han, P., and Lawless, J.F. (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica*, 29(3), 1321-1342.

Hidiroglou, M. (2001). Double sampling. *Survey Methodology*, 27, 2, 143-154. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001002/article/6091-eng.pdf.

Horn, R.A., and Johnson, C.R. (2012). *Matrix Analysis*. Cambridge university press, New York.

Imbens, G.W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20, 493-506.

Kim, J.K., and Rao, J.N.K. (2019). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

Kim, J.K., and Wang, Z. (2019). Sampling techniques for big data analysis in finite population inference. *International Statistical Review*, 87, S177-S191.

Kundu, P., Tang, R. and Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(3), 567-585.

Lohr, S.L., and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B*, 72, 27-48.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19, 1725-1747.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

Qin, J., and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22, 300-325.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Rivera-Rodriguez, C., Haneuse, S., Wang, M. and Spiegelman, D. (2020). Augmented pseudolikelihood estimation for two-phase studies. *Statistical Methods in Medical Research*, 29, 344-358.

Rivera-Rodriguez, C., Spiegelman, D. and Haneuse, S. (2019). On the analysis of two-phase designs in cluster-correlated data settings. *Statistics in Medicine*, 38, 4611-4624.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

Sheng, Y., Sun, Y., Huang, C.-Y. and Kim, M.-O. (2021). Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometrics*.

Sheng, Y., Sun, Y., Huang, C.-Y. and Kim, M.-O. (2022). Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometrics*, 78(2), 679-690.

Shin, Y.E., Pfeiffer, R.M., Graubard, B.I. and Gail, M.H. (2020a). Weight calibration to improve efficiency for estimating pure risks from the additive hazards model with the nested case-control design. *Biometrics*, 1-13.

Shin, Y.E., Pfeiffer, R.M., Graubard, B.I. and Gail, M.H. (2020b). Weight calibration to improve the efficiency of pure risk estimates from case-control samples nested in a cohort. *Biometrics*, 76(4), 1087-1097.

Taylor, J.M.G., Choi, K. and Han, P. (2022). Data integration: Exploiting ratios of parameter estimates from a reduced external model. *Biometrika*, 1-16.

Wang, C.Y., Wang, S., Zhao, L.-P. and Ou, S.-T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, 92, 512-525.

Wang, H., and Kim, J.K. (2021). Propensity score estimation using density ratio model under item nonresponse. *arXiv preprint arXiv:2104.13469*.

Wang, L., Williams, M.L., Chen, Y. and Chen, J. (2020). Novel two-phase sampling designs for studying binary outcomes. *Biometrics*, 76, 210-223.

Wu, C., and Rao, J.N.K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics/La revue canadienne de statistique*, 34(3), 359-375.

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

Xu, M., and Shao, J. (2020). Meta-analysis of independent datasets using constrained generalised method of moments. *Statistical Theory and Related Fields*, 4, 109-116.

Yang, S., and Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

Yang, S., Zheng, D. and Wang, X. (2020). Elastic integrated analysis of randomized trial and real-world data for treatment heterogeneity estimation. *arXiv preprint arXiv:2005.10579v2*.

Yuan, K.-H., and Jennrich, R.I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65(2), 245-260.

Zhai, Y., and Han, P. (2022). Data integration with oracle use of external information from heterogeneous populations. *Journal of Computational and Graphical Statistics*, 0(0), 1-12.

Zhang, H., Deng, L., Wheeler, W., Qin, J. and Yu, K. (2021). Integrative analysis of multiple case-control studies. *Biometrics*, 1-12.

Zhao, P., Ghosh, M., Rao, J.N.K. and Wu, C. (2020). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society: Series B*, 82, 155-174.

Zubizarreta, J.R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110, 910-922.