Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Survey Data Integration for Regression Analysis Using Model Calibration

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

Release date: October 15, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Survey Data Integration for Regression Analysis Using Model Calibration

Jae Kwang Kim, Hang J. Kim, and Zhonglei Wang[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 empirical likelihood method. 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; 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) and Yang and Kim (2020) provide reviews of statistical methods of data integration for finite population inference. Many existing methods (e.g., 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 et al. (1994) and Wang et al. (1997) discussed semiparametric estimation in regression analysis with missing covariate data under the missing-at-random covariate assumption. In our data integration setup, the external data source with missing covariate can be a census or big data and there may exist a selection bias in the external data.

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 the external sources and then combined through a novel application of the empirical likelihood method (Owen, 1991; Qin and Lawless, 1994). 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.

Under a similar setup, Chatterjee et al. (2016) also developed a calibration method based on the constrained maximum likelihood, which uses a fully parametric model for the likelihood specification and a constraint developed from the 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. On the other hand, our proposed method is based on the first moment conditions like usual regression analyses, so weak

[1]Jae Kwang Kim, Department of Statistics, Iowa State University, USA, 50011; Hang J. Kim, Division of Statistics and Data Science, University of Cincinnati, USA, 45221; Zhonglei Wang, Wang Yanan Institute for Studies in Economics, Xiamen University, China, 361005

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. Recently, Xu and Shao (2020) develop a data integration method using generalized method of moments technique, but their method implicitly assumes that the reduced model is correctly specified. Sheng et al. (2021) develop a penalized empirical likelihood approach to incorporate such information in the logistic regression setup. Zhang et al. (2021) also developed a retrospective empirical likelihood framework to account for sampling bias in case-control studies. We consider a more general regression setup and our proposed empirical likelihood method is different from their empirical likelihood methods and does not require that the working reduced model is correctly specified.

We highlight the contribution of our paper as follows. First, we propose a unified framework for incorporating external data sources in the regression analysis. The proposed method uses weaker assumptions than the existing 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 2.3. It can be also applied to the case where the external data source is subject to selection bias. In the real data application in Section 3, we showed 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 a direct application of the standard empirical likelihood method and can be easily implemented using the existing software.

2. Proposed Approach

2.1 Basic Setup

Consider a finite population $\mathcal{U} = \{1, \dots, N\}$ of size N. Associated with the ith record, let y_i denote the study variable of interest and $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$ the corresponding auxiliary vector of length p. We are interested in estimating a population parameter $\boldsymbol{\beta}_0$, which solves $\boldsymbol{U}_1(\boldsymbol{\beta}) = \sum_{i\in\mathcal{U}} \boldsymbol{U}_1(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) = \boldsymbol{0}$ where $U_1(\boldsymbol{\beta}, \boldsymbol{x}, y)$ is a pre-specified estimating function for $\boldsymbol{\beta}$. One example of the estimating function is $\boldsymbol{U}_1(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) = \{y_i - m_1(\boldsymbol{x}_i; \boldsymbol{\beta})\}\, \boldsymbol{h}_1(\boldsymbol{x}_i; \boldsymbol{\beta})$, which is implicitly based on a regression model $E(Y_i|\boldsymbol{x}_i) = m_1(\boldsymbol{x}_i; \boldsymbol{\beta})$ on the super-population level for some $\boldsymbol{h}_1(\boldsymbol{x}_i; \boldsymbol{\beta})$ satisfying certain identification conditions (e.g., Kim and Rao, 2009). From the finite population a probability sample $S_1 \subset \mathcal{U}$ is generated, and a Z-estimator $\widehat{\boldsymbol{\beta}}$ can be obtained by solving

$$(1) \qquad \widehat{U}_1(\boldsymbol{\beta}) = \sum_{i\in S_1} d_i \boldsymbol{U}_1(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) = \boldsymbol{0},$$

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

In addition to S_1, suppose that we observe $\boldsymbol{x}_{i1}$ and y_i throughout the finite population and wish to incorporate this extra information to improve the estimation efficiency of $\widehat{\boldsymbol{\beta}}$. 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,

$$(2) \qquad E(Y_i|\boldsymbol{x}_{i1}) = m_2(\boldsymbol{x}_{i1}; \boldsymbol{\alpha})$$

for some $\boldsymbol{\alpha}$. Under the working model (2), we can obtain an estimator $\widehat{\boldsymbol{\alpha}}$ from the current sample S_1 by solving $\widehat{\boldsymbol{U}}_2(\boldsymbol{\alpha}) = \sum_{i\in S_1} d_i \boldsymbol{U}_2(\boldsymbol{\alpha}; \boldsymbol{x}_{i1}, y_i) = \boldsymbol{0}$ where $\boldsymbol{U}_2(\boldsymbol{\alpha}; \boldsymbol{x}_{i1}, y_i) = \{y_i - m_2(\boldsymbol{x}_{i1}; \boldsymbol{\alpha})\}\, \boldsymbol{h}_2(\boldsymbol{x}_{i1}; \boldsymbol{\alpha})$ for some $\boldsymbol{h}_2(\boldsymbol{x}_{i1}; \boldsymbol{\alpha})$ satisfying conditions similar to ones imposed to $\boldsymbol{h}_1(\boldsymbol{x}_i; \boldsymbol{\beta})$. Note that our setup considers the situation where a subset of individual data $(\boldsymbol{x}_{i1}, y_i)$ is fully observed throughout the finite population $\mathcal{U}$. Therefore, one can get $\boldsymbol{\alpha}^*$ that solves $\sum_{i=1}^{N} \boldsymbol{U}_2(\boldsymbol{\alpha}; \boldsymbol{x}_{i1}, y_i) = \boldsymbol{0}$. Chen and Chen (2000) proposed using $\widehat{\boldsymbol{\beta}}^* = \widehat{\boldsymbol{\beta}} + \widehat{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})\{\widehat{V}(\widehat{\boldsymbol{\alpha}})\}^{-1}(\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}})$ as an efficient estimator of $\boldsymbol{\beta}$, where $\widehat{V}(\cdot)$ and $\widehat{Cov}(\cdot)$ denote the design-based variance and covariance estimators, respectively. The working model in (2) is not necessarily correctly specified, but a good working model can improve the efficiency of the final estimator.

One may also adopt a constrained maximum likelihood (CML) similar to Chatterjeeet al. (2016), which was originally suggested in a non-survey sampling context. Under a survey sampling setup, we can interpret Chatterjeeet al. (2016)

as a CML estimation approach when $\boldsymbol{\beta}$ is a parameter in the conditional distribution of Y_i given $\boldsymbol{X}_i$ with density $f(y_i|\boldsymbol{x}_i;\beta)$, and the CML estimation can be expressed as finding $\boldsymbol{\beta}$ that maximizes

$$(3) \qquad l_p(\boldsymbol{\beta}) = \sum_{i\in S_1} d_i \log f(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})$$

subject to

$$(4) \qquad \sum_{i\in S_1} d_i \int \boldsymbol{U}_2(\boldsymbol{\alpha}^*;\boldsymbol{x}_{i1},y) f(y|\boldsymbol{x}_i;\boldsymbol{\beta})\, dy = \boldsymbol{0}.$$

Constraint (4) can be understood as a constraint for the parameter $\boldsymbol{\beta}$ to satisfy $E\{\boldsymbol{U}_2(\boldsymbol{\alpha}^*;\boldsymbol{x}_{i1},Y_i)|\boldsymbol{x}_i;\boldsymbol{\beta}\} = \boldsymbol{0}$. By imposing this constraint into the maximum likelihood estimation, the external information $\boldsymbol{\alpha}^*$ can be naturally incorporated.

The CML method is not directly applicable to our conditional mean model in (1) as the likelihood function for $\boldsymbol{\beta}$ is not defined in our setup. Nonetheless, one can use an objective function such as that in Generalized Method of Moments to apply the constrained optimization problem, which is asymptotically equivalent to the empirical likelihood method (Imbens, 2002). Chatterjee et al. (2016) also noted that the CML approach could be formulated using the empirical likelihood method of Qin and Lawless (1994) and Qin (2000). However, they did not explicitly discuss how to formulate the CML as an application of the empirical likelihood method.

2.2 Proposed Approach

We now use the empirical likelihood framework to incorporate the auxiliary information. The classical calibration problem can be formulated as finding the calibration weights $\boldsymbol{w} = \{w_i : i \in S_1\}$ based on a certain objective function $Q(\boldsymbol{d},\boldsymbol{w})$ subject to some calibration constraints (Deville and Särndal, 1992) where $\boldsymbol{d} = \{d_i : i \in S_1\}$. For the objective function, we may either use the pseudo empirical likelihood function

$$(5) \qquad Q(\boldsymbol{d},\boldsymbol{w}) = \sum_{i\in S_1} d_i \log w_i$$

considered by Wu and Rao (2006) or the maximum entropy function $Q(\boldsymbol{d},\boldsymbol{w}) = \sum_{i\in S_1} w_i \log(w_i/d_i)$ in Kim (2010). Our calibration constraint is

$$(6) \qquad \sum_{i\in S_1} w_i \boldsymbol{U}_2(\boldsymbol{\alpha}^*;\boldsymbol{x}_{i1},y_i) = \boldsymbol{0},$$

where $\boldsymbol{\alpha}^*$ is the external information for the working reduced model. This is the same spirit of using (4) but without introducing the conditional density function $f(y|\boldsymbol{x},\boldsymbol{\beta})$. Thus, we can use the following model calibration method for efficient estimation of $\boldsymbol{\beta}$ as follows: use the working reduced model (2) to obtain $\boldsymbol{\alpha}^*$ from the finite population; find the calibration weights $\widehat{\boldsymbol{w}} = \{\widehat{w}_i : i \in S_1\}$ maximizing $Q(\boldsymbol{d},\boldsymbol{w})$ subject to (6); and once the solution $\widehat{\boldsymbol{w}}$ is obtained from the calibration, estimate $\boldsymbol{\beta}$ by solving $\sum_{i\in S_1} \widehat{w}_i \boldsymbol{U}_1(\boldsymbol{\beta};\boldsymbol{x}_i,y_i) = \boldsymbol{0}$.

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 $\widehat{\boldsymbol{\alpha}}_2$ and its variance estimator $\boldsymbol{V}_2 = \widehat{V}(\widehat{\boldsymbol{\alpha}}_2)$ for the working reduced model in (2). Then, an estimator of the benchmark $\boldsymbol{\alpha}^*$ can be obtained by

$$(7) \qquad \widehat{\boldsymbol{\alpha}}^* = (\boldsymbol{V}_1^{-1} + \boldsymbol{V}_2^{-1})^{-1}(\widehat{\boldsymbol{V}}_1^{-1}\widehat{\boldsymbol{\alpha}}_1 + \widehat{\boldsymbol{V}}_2^{-1}\widehat{\boldsymbol{\alpha}}_2)$$

where $\widehat{\boldsymbol{\alpha}}_1$ and $\boldsymbol{V}_1$ are estimated with the internal sample S_1. Once $\widehat{\boldsymbol{\alpha}}^*$ is obtained by (7), it replaces $\bmalpha^*$ in the calibration equation in (6).

2.3 Multiple Data Integration

We can consider regression analysis combining partial information from external samples. To explain the idea, Table 2.3.1 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 2.3.1
Data structure for survey integration

Sample	Sampling Weight	z	x_1	x_2	y
A	d_a	O	O	O	O
B	d_b	O	O		O
C	d_c	O		O	O

Under the setup of Table 2.3.1, suppose that we are interested in estimating the parameters in the regression model $E(Y|x_1, x_2) = m_1(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ 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

$$(8) \qquad \widehat{\boldsymbol{U}}_a(\boldsymbol{\beta}) = \sum_{i \in A} d_{a,i} \{y_i - m(x_{i1}, x_{i2}; \boldsymbol{\beta})\} \boldsymbol{h}(x_{i1}, x_{i2}; \boldsymbol{\beta}) = \mathbf{0},$$

for some $\boldsymbol{h}(x_{i1}, x_{i2}; \boldsymbol{\beta})$ such that $\widehat{\boldsymbol{U}}_a(\boldsymbol{\beta})$ 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 $E(Y|x_1, z) = m_2(x_1, z; \boldsymbol{\alpha})$ for some $\boldsymbol{\alpha}$. Note that, since (z_i, x_{i1}, y_i) are observed, we can use sample B to estimate $\boldsymbol{\alpha}$ by solving $\sum_{i \in B} d_{b,i} \boldsymbol{U}_b(\boldsymbol{\alpha}; x_{i1}, z_i, y_i) = \mathbf{0}$ for some $\boldsymbol{U}_b$ satisfying $E\{\boldsymbol{U}_b(\boldsymbol{\alpha}; x_1, z, Y|x_1, z)\} = \mathbf{0}$ under the working model for $E(Y|x_1, z)$.

Similarly, to incorporate the partial information from sample C, suppose that we have a "working" model $E(Y|x_2, z) = m_3(x_2, z; \boldsymbol{\gamma})$ for some $\boldsymbol{\gamma}$. We can also construct an unbiased estimating equation $\sum_{i \in C} d_{c,i} \boldsymbol{U}_c(\boldsymbol{\gamma}; x_{i2}, z_i, y_i) = 0$ for some $\boldsymbol{U}_c$ satisfying $E\{\boldsymbol{U}_c(\boldsymbol{\gamma}; x_2, z, Y)|x_2, z\} = \mathbf{0}$ under the working model for $E(Y|x_2, z)$. Once $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\gamma}}$ are obtained, we can use this extra information to improve the efficiency of $\widehat{\boldsymbol{\beta}}$ in (8). To incorporate the extra information, we can formulate it as maximizing $Q(\boldsymbol{d}_a, \boldsymbol{w}) = \sum_{i \in A} d_{a,i} \log w_i$ subject to $\sum_{i \in A} w_i = N$ and

$$(9) \qquad \sum_{i \in A} w_i \{\boldsymbol{U}_b(\widehat{\boldsymbol{\alpha}}; x_{i1}, z_i, y_i), \boldsymbol{U}_c(\widehat{\boldsymbol{\gamma}}; x_{i2}, z_i, y_i)\} = \mathbf{0}$$

where $\boldsymbol{d}_a$ and $\boldsymbol{w}$ are sets containing the sampling weights and calibration weights with respect to sample A. Constraint (9) incorporates the extra information. Once the solution $\widehat{w}_i$ is obtained, we can use $\sum_{i \in A} \widehat{w}_i \{y_i - m(x_{i1}, x_{i2}; \boldsymbol{\beta})\} \boldsymbol{h}(x_{i1}, x_{i2}; \boldsymbol{\beta}) = \mathbf{0}$ to estimate $\boldsymbol{\beta}$.

3. Simulation Study

To evaluate the finite sample performance of the proposed estimators, we conducted a simulation study. We generated a finite population of size $N = 100{,}000$, each record consisting of auxiliary variables $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ and a response variable y_i. We assume that $(\boldsymbol{x}_i, y_i)$ is available for the internal sample S_1 while only (x_{i1}, y_i) is available for the finite population $\mathcal{U}$ or an external sample S_2.

For covariates, we consider two scenarios for the finite population: Setting 1. $x_{i1} \sim N(3,1)$, $x_{i2} \sim N(11, 6.5^2)$, and x_{i1} and x_{i2} are independent; and Setting 2. $x_{i1} \sim N(3,1)$ and $x_{i2} = x_{i1}^2 + N(0,1)$. The simulation parameters are chosen such that the marginal mean and variance of x_{i2} are similar in the independent and the dependent settings. The response variable y_i is generated by a Bernoulli distribution with success probability $\Pr(Y_i = 1|x_{i1}, x_{i2}) =$

$logit^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})$ with the simulation parameters $(\beta_0, \beta_1, \beta_2) = (-0.5, 0.1, -0.2)$. We consider Poisson sampling with inclusion probabilities satisfying $\pi_i \propto 0.9\, I(y_i = 1) + 0.1\, I(y_i = 0)$ to generate a probability sample S_1 of size $n_1 = 5{,}000$.

For the proposed approach, we consider the working reduced model written by $U_2(\alpha; x_{i1}, y_i) = \{y_i - expit(\alpha_0 + \alpha_1 x_{i1})\}(1\ x_{i1})^{\mathrm{T}}$ where $expit(x) = \{1 + \exp(-x)\}^{-1}$. We compare performances of five approaches: (i) estimator with the probability sample S_1 only, the solution of $\sum_{i \in S} d_i expit(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})(1\ x_{i1}\ x_{i2})^{\mathrm{T}} = \mathbf{0}$, (ii) proposed estimator with $\boldsymbol{\alpha}^*$ from the finite population $\mathcal{U}$, (iii) proposed estimator with $\hat{\boldsymbol{\alpha}}^*$ estimated from an external sample S_2 of size $n_2 = 50{,}000$, (iv) CML estimator relying on (3) and (4) with partial information from $\mathcal{U}$, and (iv) CML estimator with S_2.

Table 3.1 shows the simluation results. When x_{i1} and x_{i2} are independently generated in the simulation data (Setting 1), all approaches show negligible biases and four approaches using the extra information are more efficient in estimating β_0 and β_1 than using the internal sample only (S_1-only). When x_{i1} and x_{i2} in the simulation data are dependent (Setting 2), the CML estimators (CML-$\mathcal{U}$ and CML-S_2) suffer with large biases resulting in large MSE and incorrect confidence interval coverages for CML-S_2. In the dependent covariate cases, the proposed estimators still show negligible biases and correct confidence interval coverages, and their root mean squared errors for β_0 and β_1 are also smaller than those of S_1-only. Note that there is no efficiency gain in estimating β_2 by the proposed methods as expected because the external data consist of x_{i1} only.

Table 3.1
Logistic regression performance under Poisson sampling measured by the Monte Carlo bias (Bias), the root mean squared error (rMSE), and the 95% confidence interval coverage (CI) measured for estimation with the internal sample only (S_1-only); proposed methods when (y_i, x_{i1}) are available for the entire population (Prop $\mathcal{U}$) and for an external sample (Prop-S_2); and Chatterjee-like CLM estimators (CML-$\mathcal{U}$ and CML-S_2).

Setting 1: x_{i1} and x_{i2} are independent

	β_0			β_1			β_2		
	Bias	rMSE	CI	Bias	rMSE	CI	Bias	rMSE	CI
S_1 only	-0.002	0.129	0.950	0.001	0.037	0.933	0.000	0.007	0.953
Prop-$\mathcal{U}$	-0.003	0.097	0.944	0.001	0.025	0.934	0.000	0.007	0.952
CML-$\mathcal{U}$	-0.024	0.082	--	0.006	0.020	--	0.000	0.006	--
Prop-S_2	-0.001	0.093	0.939	0.000	0.024	0.934	0.000	0.007	0.952
CML-S_2	-0.022	0.076	0.959	0.006	0.018	0.971	0.000	0.006	0.996

Setting 2: x_{i1} and x_{i2} are dependent

	β_0			β_1			β_2		
	Bias	rMSE	CI	Bias	rMSE	CI	Bias	rMSE	CI
S_1 only	0.002	0.174	0.947	0.000	0.126	0.949	0.000	0.024	0.946
Prop-$\mathcal{U}$	0.000	0.112	0.948	0.000	0.109	0.946	0.000	0.024	0.945
CML-$\mathcal{U}$	0.504	0.511	--	-0.409	0.417	--	0.073	0.075	--
Prop-S_2	0.001	0.107	0.949	-0.001	0.108	0.948	0.000	0.024	0.945
CML-S_2	0.505	0.511	0.002	-0.410	0.417	0.054	0.073	0.075	0.229

4. Application Study

4.1 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=one million 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 $\hat{\boldsymbol{\alpha}}_2$ in (7) is unavailable. Section 4.2 addresses this issue by using a propensity weighting approach and Section 4.3 presents the analysis result of the application study.

4.2 Propensity Weighing for External Data with Unknown Selection Probability

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

$$\sum_{i\in S_2}\{y_i - m(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i1})\}\boldsymbol{x}_{i1} = \mathbf{0} \tag{10}$$

to obtain $\hat{\alpha}_0$ and $\hat{\alpha}_1$. If the sampling mechanism for S_2 is ignorable or non-informative, then the solution of (10) 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 (10) to obtain an unbiased estimator of $\boldsymbol{\alpha}$. To construct the propensity score weights, we employ a nonignorable nonresponse model, $P(\delta_i = 1|\boldsymbol{x}_{i1}, y_i) = \pi(\boldsymbol{x}_{i1}, y_i; \boldsymbol{\phi})$, where $\delta_i = 1$ if $i \in S_2$ and zero otherwise.
Note that we can express $\pi(\boldsymbol{x}_{i1}, y_i)^{-1} = 1 + (N_0/N_1)r(\boldsymbol{x}_{i1}, y_i)$ where $\log r(\boldsymbol{x}_{i1}, y_i) = f(\boldsymbol{x}_{i1}, y_i|\delta_i = 0)/f(\boldsymbol{x}_{i1}, y_i|\delta_i = 1)$ 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, $\log r(x_{i1}, y_i; \boldsymbol{\phi}) = \phi_0 + \phi_1 x_{i1} + \phi_2 y_i$. The maximum entropy estimator of $\boldsymbol{\phi}$ is obtained by solving $(1/N_1)\sum_{i=1}^{N}\delta_i \exp(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i)(1\ x_{i1}\ y_i)^{\mathrm{T}} = (1\ \hat{\bar{x}}_1\ \hat{\bar{y}})^{\mathrm{T}}$ where $(\hat{\bar{x}}_1, \hat{\bar{y}}) = \left(1/\widehat{N}_0\right)\{\sum_{i\in S_1} d_1(x_{i1} - y_i) - \sum_{i=1}^{N}\delta_i(x_{i1}, y_i)\}$, $\widehat{N}_0 = \sum_{i\in S_1} d_i - N_1$, and S_1 is the internal sample. Once $\hat{\boldsymbol{\phi}}$ is obtained, we can construct $\hat{\pi}(x_{i1}, y_i)$ and solve

$$\sum_{i\in S_2}\frac{1}{\pi(x_{i1}, y_i)}\{y_i - m(\alpha_0 + \alpha_1 x_{i1})\}(1\ x_{i1})^{\mathrm{T}} = \mathbf{0} \tag{11}$$

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

In addition, we can use the internal sample S_1 to fit the same working model to obtain $\hat{\boldsymbol{\alpha}}_1$. After that, we obtain $\hat{\boldsymbol{\alpha}}^*$ using (7) and apply the proposed calibration weighting method to combine information from the big data. In practice, $\boldsymbol{V}_2$ in (7) is difficult to compute, but it is negligibly small if the sample size for S_2 is huge. In this case, we may simply use $\hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\alpha}}_2$ in the calibration problem.

4.3 Application Study Results: Korea National Health and Nutrition Examination Survey

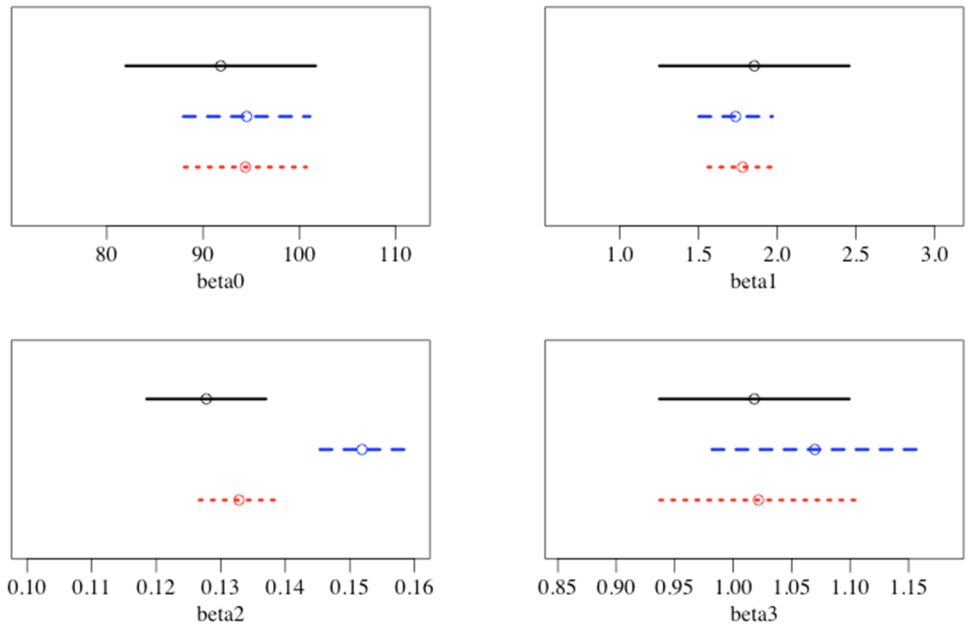
In this application study, we use $n_1 = 4{,}929$ records of KNHANES data that have no missing values in four variables: Total cholesterol, Hemoglobin, Triglyceride, and HDL cholesterol. For demonstration purpose, we assume that an analyst is interested in conducing the following regression analysis, $E(\mathrm{Total\ Cholesterol}_i|\boldsymbol{x}_i) = \beta_0 + \beta_1\mathrm{Hemoglobin}_i + \beta_2\mathrm{Triglyceride}_i + \beta_3\mathrm{HDL}_i$ for $i \in S_1$. 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 3 where the covariates were highly correlated. The big size external data consist of $n_2 =$ one million records of NHISS data with fully observed items in Total cholesterol, Hemoglobin, and Triglyceride. The assumed working model to connect the external sample to the internal sample is $E(\text{Total Cholesterol}_i|\boldsymbol{x}_{i1}) = \alpha_0 + \alpha_1\text{Hemoglobin}_i + \alpha_2\text{Triglyceride}_i$ for $i \in S_1 \cup S_2$.

In this application study, we implement our proposed methods with the external sample where $\widehat{\boldsymbol{\alpha}}_2$ 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 S_2 as SRS, i.e., without propensity weighting, and (ii) with the propensity weighting adjustment introduced in Section 4.2. For the propensity weighting, we fit the log-linear density ratio model to the external data, $\log r(x_{i1}, y_i; \boldsymbol{\phi}) = \phi_0 + \phi_1\text{Hemoglobin}_i + \phi_2\text{Triglyceride}_i + \phi_3\text{Total Cholesterol}_i$, calculate $\hat{\pi}(x_{i1}, y_i)$ given $\widehat{\boldsymbol{\phi}}$, then solve (11) to obtain $\widehat{\boldsymbol{\alpha}}_2$. The performances of proposed methods are compared with the reference method that uses the internal sample S_2 only to get weighted least square estimates considering the sampling weights.

Figure 4.3.1 shows the point estimates and the 95% confidence intervals of $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ 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 β_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 β_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 β_3 was expected as the external data contain information of x_{i1} (Hemoglobin) and x_{i2} (Triglyceride), not x_{i3} (HDL).

Figure 4.3.1
Comparison of the regression analysis for $E(\text{Total Cholesterol}_i|x_i) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\text{Hemoglobin}_i + \boldsymbol{\beta_2}\text{Triglyceride}_i + \boldsymbol{\beta_3}\text{HDL}_i$ 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 $\boldsymbol{S_1}$ 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).



5. 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 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 et al. (2020) in this setup. The proposed method can potentially be used to combine the randomized clinical trial data with big real-world data (Yang et al.,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 et al., 2019; Wanget al., 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.

References

Chatterjee, N., Y.-H. Chen, P. Maas, and R. Carroll (2016), "Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information From External Big Data Sources", *Journal of the American Statistical Association*, 111, pp. 107-117.

Chen, Y. H. and H. Chen (2000), "A Unified Approach to Regression Analysis Under Double- Sampling Designs", *Journal of the Royal Statistical Society: Series B*, 62, pp. 449-460.

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

Fuller, W. A. (2009), *Sampling Statistic*, Hoboken, NJ: Wiley.

Hidiroglou, M. (2001), "Double Sampling", *Survey methodology*, 27, pp. 143-154.

Imbens, G. W. (2002), "Generalized Method of Moments and Empirical Likelihood", *Journal of Business and Economic Statistics*, 20, pp. 493-506.

Kim, J. K. (2010), "Calibration Estimation Using Exponential Tilting in Sample Surveys", *Survey Methodology*, 36, pp. 145-155.

Kim, J. K. and J. N. K. Rao (2009), "A Unified Approach to Linearization Variance Estimation from Survey Data After Imputation for Item Nonresponse", *Biometrika*, 96, pp. 917-932.

Kim, J. K. and Z. Wang (2019), "Sampling Techniques for Big Data Analysis in Finite Population Inference", *International Statistical Review*, 87, pp. S177-S191.

Lohr, S. L. and T. E. Raghunathan (2017), "Combining Survey Data with Other Data Sources", *Statistical Science*, 32, pp. 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, pp. 27-48.

Owen, A. (1991), "Empirical Likelihood for Linear Models", *The Annals of Statistics*, 19, pp. 1725-1747.

Qin, J. (2000), "Combining Parametric and Empirical Likelihoods", *Biometrika*, 87, pp. 484-490.

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

Rivera-Rodriguez, C., D. Spiegelman, and S. Haneuse (2019), "On the Analysis of Two-phase Designs in Cluster-correlated Data Settings", *Statistics in Medicine*, 38, pp. 4611-4624.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed", *Journal of the American Statistical Association*, 89, pp. 846-866.

Sheng, Y., Y. Sun, C.-Y. Huang, and M.-O. Kim (2021), "Synthesizing External Aggregated Information in the Presence of Population Heterogeneity: A Penalized Empirical Likelihood Approach", *Biometrics*. DOI: 10.1111/biom.13429.

Wang, C. Y., S. Wang, L.-P. Zhao, and S.-T. Ou (1997), "Weighted Semiparametric Estimation in Regression Analysis with Missing Covariate Data", *Journal of the American Statistical Association*, 92, pp. 512-525.

Wang, H. and J. K. Kim (2021), "Propensity Score Estimation Using Density Ratio Model Under Item Nonresponse", arXiv preprint, arXiv:2104.13469.

Wang, L., M. L. Williams, Y. Chen, and J. Chen (2020), "Novel Two-Phase Sampling Designs for Studying Binary Outcomes", *Biometrics*, 76, pp. 210-223.

Wu, C. and J. Rao (2006), "Pseudo Empirical Likelihood Ratio Confidence Intervals for Complex Surveys", *Canadian Journal of Statistics*, 34, pp. 359-375.

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

Xu, M. and J. Shao (2020), "Meta-Analysis of Independent Datasets Using Constrained Generalised Method of Moments", *Statistical Theory and Related Fields*, 4, pp. 109-116.

Yang, S. and J. K. Kim (2020), "Statistical Data Integration in Survey Sampling: A review", *Japanese Journal of Statistics and Data Science*, 3, pp. 625-650.

Yang, S., D. Zheng, and X. Wang (2020), "Elastic Integrated Analysis of Randomized Trial and Real-World Data for Treatment Heterogeneity Estimation", arXiv preprint, arXiv:2005.10579v2.

Yuan, K.-H. and R. I. Jennrich (1998), "Asymptotics of Estimating Equations Under Natural Conditions", *Journal of Multivariate Analysis*, 65, pp. 245-260.

Zhang, H., L. Deng, W. Wheeler, J. Qin, and K. Yu (2021), "Integrative Analysis Of Multiple Case-Control Studies", *Biometrics*. https://doi.org/10.1111/biom.13461.

Zhao, P., M. Ghosh, J. Rao, and C. Wu (2020), "Bayesian Empirical Likelihood Inference with Complex Survey Data", *Journal of the Royal Statistical Society: Series B*, 82, pp. 155-174.

Zubizarreta, J. R. (2015), "Stable Weights That Balance Covariates for Estimation with Incomplete Outcome Data", *Journal of the American Statistical Association*, 110, pp. 910-922.