Communications for Statistical Applications and Methods
2019, Vol. 26, No. 5, 507–517

https://doi.org/10.29220/CSAM.2019.26.5.507
Print ISSN 2287-7843 / Online ISSN 2383-4757

Statistical micro matching using a multinomial logistic regression model for categorical data

Kangmin Kim[a], Mingue Park[1,a]

[a]Department of Statistics, Korea University, Korea

Abstract

Statistical matching is a method of combining multiple sources of data that are extracted or surveyed from the same population. It can be used in situation when variables of interest are not jointly observed. It is a low-cost way to expect high-effects in terms of being able to create synthetic data using existing sources. In this paper, we propose the several statistical micro matching methods using a multinomial logistic regression model when all variables of interest are categorical or categorized ones, which is common in sample survey. Under conditional independence assumption (CIA), a mixed statistical matching method, which is useful when auxiliary information is not available, is proposed. We also propose a statistical matching method with auxiliary information that reduces the bias of the conventional matching methods suggested under CIA. Through a simulation study, proposed micro matching methods and conventional ones are compared. Simulation study shows that suggested matching methods outperform the existing ones especially when CIA does not hold.

Keywords: statistical matching, multinomial logistic regression model, conditional independence assumption, auxiliary information

1. Introduction

As a traditional data collection method to make an inference on the finite population, surveys are currently criticized for its cost-inefficiency caused by worsen survey environment such as increasing not-at-home households and nonresponse rate.

Statistical matching, that is less informative but more cost-efficient data collection or augmentation method, has been suggested as a possible solution to overcome the weakness of the conventional survey. Statistical matching is a method of combining multiple sources of data that has two versions, macro matching and micro matching. Macro matching is mainly used to estimate the population parameters that are not estimable using a single source of data.

In many applications, statistical matching implies micro matching that is similar to data linkage. For the explanation of the micro statistical matching, assume there are recipient file A and donor file B sampled from the same population. There are no overlapping units in the file of A and B, and the recipient file A contain variables (X, Y), and the donor file B contain (X, Z). X is a set common variables that is contained in both data, Y is a set of unique variables that is observed only in file A, and Z is a set of unique variables that is observed only in file B. Micro statistical matching generates a matched file on Figure 1 by filling in appropriate values of Z to the recipient file A using the information obtained from the file A and B. Even though various versions of micro statistical matching

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Sungbuk-Gu, Seoul 02841, Korea.
E-mail: mpark2@korea.ac.kr

Published 30 September 2019 / journal homepage: http://csam.or.kr

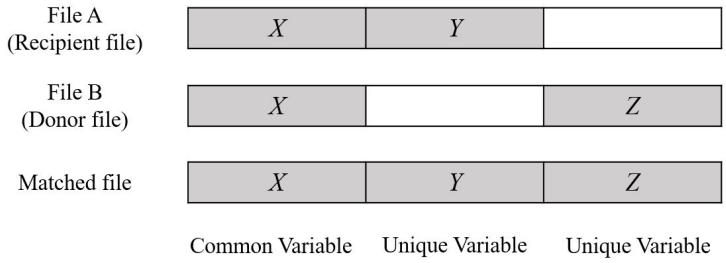

Figure 1: *Basic structure of statistical matching.*

are also existing, for example, matched file is constructed by combining file A and file B after filling in Z and Y to each file, we focus on a micro matching method in which a matched file is constructed from recipient file A. However, the results obtained in this paper are easily extended to other micro matching methods.

Most of micro matching methods including Budd (1971) and Okner (1972) get their theoretical validity under the conditional independence assumption (CIA) that means unique variables Y and Z are conditionally independence conditioning on X. However, as mentioned in Sims (1972), Rodgers (1984), Rubin (1986), and Singh *et al.* (1993), CIA is unrealistic and also the inferences on the population obtained using matched files are vulnerable to serious error that causes significant bias of an estimator. Except Renssen (1998), previous studies on the micro statistical matching considered data in which all X, Y, and Z are continuous even though either nominal or ordinal scales are frequently used for the surveys.

In this paper, at first, we suggest a new statistical matching method applicable to categorical variables under CIA. The proposed method is a mixture of parametric and nonparametric method which is robust to model misspecification. In addition, we also propose a statistical matching method for categorical variables using the auxiliary information when the CIA is not satisfied. The auxiliary information could be obtained from small surveys or outdated proxy data, denoted by file C, which should contains the distribution of (X, Y, Z) or (Y, Z). The proposed method using auxiliary information is also a mixed method using multinomial logistic regression model. Through an extensive simulation study, we compare the performance of suggested methods to several existing micro matching methods.

The organization of paper is as follows. A brief review of statistical mathcing method using auxiliary information is given in Section 2. In Section 3, we propose new statistical matching methods without and with auxiliary information for categorical variables. In Section 4, we compare the several statistical matching methods including the suggested ones in Section 3 through a simulation. We make some concluding remarks in Section 5.

2. Statistical matching method using auxiliary information

In this section, we briefly review the important previous statistical matching methods in which auxiliary information is considered, based on Singh *et al.* (1993) and Renssen (1998). Singh *et al.* (1993) proposed a modified version of Rubin (1986) and Paass (1986). Rubin (1986) proposed a mixed matching method under parametric regression model which uses auxiliary information. A linear regression model is assumed that

$$E(Z|X, Y) = \beta_0 + \beta_1 X + \beta_2 Y, \qquad V(Z|X, Y) = \sigma^2, \tag{2.1}$$

where β_0, β_1, and β_2 are estimated from least squares equations by combining information from files A, B, and C. If the CIA between Y and Z is satisfied, the model (2.1) reduce to the simple linear regression of Z on X. First, for file A, predicting an intermediate value, Z_{int}, which can be obtained from a linear regression model (2.1) where $(\beta_0, \beta_1, \beta_2)$ are estimated using information from file A, B, and C. Second, a value of Z in file B imputed for the file A using hotdeck method based on (X, Z). On the other hand, Paass (1986) proposed a nonparametric matching method which uses auxiliary information. Paass (1986)'s method basically consists of first, for the file A, a value of Z from file C is imputed using hotdeck method based on Y or (X, Y). And then a value of Z from file B is imputed again using hotdeck method based on (X, Z) for the file A. The basic idea of Singh *et al.* (1993) is to impose categorical constraints on the matched file which is obtained from Rubin (1986) and Paass (1986). The categorical constrained matching method is based on log linear imputation which is proposed by Singh (1988). The purpose of this method is to preserve the categorical association among the variables as much as possible. It start with a suitable categorizing the continuous variable (X, Y, Z) into (X^*, Y^*, Z^*) and the distribution of cell proportions for the (X^*, Y^*, Z^*) table can be parametrized by a log linear model

$$\log u_{ijk} = \lambda + \lambda_i^{X^*} + \lambda_j^{Y^*} + \lambda_k^{Z^*} + \lambda_{ij}^{X^*Y^*} + \lambda_{ik}^{X^*Z^*} + \lambda_{jk}^{Y^*Z^*} + \lambda_{ijk}^{X^*Y^*Z^*}, \tag{2.2}$$

where u_{ijk} denotes the expected frequency for $(i, j, k)^{th}$ cell. All of the parameters in the (2.2) can be estimated from file A, B, and C. Singh *et al.* (1993) defines the micro matched file such that the estimated distribution of (X^*, Y^*, Z^*) using the matched file from Rubin (1986) or Paass (1986) satisfy the constraint (2.2). This categorical constraints are expected to generate the reasonable estimate of joint distributions of (X, Y, Z) in the synthetic data robust to quality of the auxiliary information file C.

Renssen (1998) considered a statistical matching method in which calibration technique is applied for categorical variables in the finite population set-up. Renssen's method is actually based on the regression method suggested by Rubin (1986). They assume that there are two registrations, file A and B, and an auxiliary information, file C, which is derived from these registrations. The problem is imputing a value of Z from the file B into the file A using auxiliary information file C. They applied a linear probability model used in Maddala (1983) with the assumption that the population total of X, Y, and Z are known. Then, new predicted value for the Z:

$$\hat{\boldsymbol{z}}_s = \boldsymbol{A}'\boldsymbol{x}_s + \hat{\boldsymbol{\alpha}}'\left(\boldsymbol{y}_s - \boldsymbol{B}'\boldsymbol{x}_s\right), \quad s = 1, 2, \ldots, N, \tag{2.3}$$

where,

$$\boldsymbol{A}' = \left(\sum_{s=1}^{N} \boldsymbol{x}_s\boldsymbol{x}_s'\right)^{-1} \sum_{s=1}^{N} \boldsymbol{x}_s\boldsymbol{z}_s', \qquad \boldsymbol{B}' = \left(\sum_{s=1}^{N} \boldsymbol{x}_s\boldsymbol{x}_s'\right)^{-1} \sum_{s=1}^{N} \boldsymbol{x}_s\boldsymbol{y}_s',$$

$$\hat{\boldsymbol{\alpha}} = \left[\sum_{s=1}^{N} \left(\boldsymbol{y}_s - \boldsymbol{B}'\boldsymbol{x}_s\right)\left(\boldsymbol{y}_s - \boldsymbol{B}'\boldsymbol{x}_s\right)'\right]^{-1} \times \left[\sum_{s=1}^{n} \boldsymbol{w}_s\left(\boldsymbol{y}_s - \boldsymbol{B}'\boldsymbol{x}_s\right)\left(\boldsymbol{z}_s - \boldsymbol{A}'\boldsymbol{x}_s\right)'\right],$$

$\boldsymbol{x}_s, \boldsymbol{y}_s, \boldsymbol{z}_s$ are vectors of order I, J, K dummy variables, and a set of $\boldsymbol{w}_s$ is a calibration weight for file C which satisfies following set of constraints.

$$\sum_{s=1}^{n} \boldsymbol{w}_s\left[\boldsymbol{y}_s\boldsymbol{z}_s' - \left(\boldsymbol{y}_s - \boldsymbol{B}'\boldsymbol{x}_s\right)\left(\boldsymbol{z}_s - \boldsymbol{A}'\boldsymbol{x}_s\right)'\right] = \sum_{s=1}^{N} \left(\boldsymbol{B}'\boldsymbol{x}_s\right)\left(\boldsymbol{A}'\boldsymbol{x}_s\right)'$$

$\boldsymbol{A}$ and $\boldsymbol{B}$ in (2.3) can be calculated from the registraion file B and A respectively. $\hat{\boldsymbol{\alpha}}$ can be estimated from file A and file C. The new predicted value equals to a naive predicted value plus an adjustment term which can be viewed as an attempt to improve the prediction for Z. This adjustment term depends on the difference between the value of Y and its predicted value. The new predictor for the Z-value can be used for imputation. First, the predicted value given by (2.3) is calculated for each unit in the file A. Second, each predicted Z-value in the file A is repalced by an actual Z-value from the file B, like the method of Singh *et al.* (1993).

3. Proposed statistical matching methods for categorical variables

In this section, we propose statistical matching methods for categorical variables with or without auxiliary information by developing the methods of Singh *et al.* (1993) and Renssen (1998). The proposed methods are mixture of hotdeck method in Singh *et al.* (1993) and parametric regression model in Renssen (1998). Basically, we use a multinomial logistic regression model instead of linear probability model suggested in Renssen (1998). To define a multinomial logistic regression model, suppose that the response variable Y has J category and that the probability of belonging to each category is $(\pi_1, \pi_2, \ldots, \pi_J)$, where $\sum_{j=1}^{J} \pi_j = 1$. Under the multinomial logistic regression model, the log of probability of the observation belonging to each category j relative to the last category J is

$$\ln \frac{\pi_j}{\pi_J} = \boldsymbol{\beta}_j' \boldsymbol{x}, \quad j = 1, \ldots, J-1, \tag{3.1}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$ is vector of explanatory variables and $\boldsymbol{\beta}_j' = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})$ is a vector of regression coefficients corresponding to outcome j. From (3.1), we get the probability π_j as

$$\pi_j = \pi_J \exp\left(\boldsymbol{\beta}_j' \boldsymbol{x}\right) = \frac{\exp\left(\boldsymbol{\beta}_j' \boldsymbol{x}\right)}{1 + \sum_{i=1}^{J-1} \exp\left(\boldsymbol{\beta}_i' \boldsymbol{x}\right)}, \quad j = 1, \ldots, J-1, \tag{3.2}$$

where

$$\pi_J = \frac{1}{1 + \sum_{i=1}^{J-1} \exp\left(\boldsymbol{\beta}_i' \boldsymbol{x}\right)}.$$

3.1. Statistical matching without auxiliary information under CIA

D'Orazio *et al.* (2006) investigated several methods of statistical matching under CIA in depth. In D'Orazio *et al.* (2006), parametric, nonparametric, and mixed matching methods are investigated thoroughly when variables are continuous, and methods using loglinear model and random hotdeck are briefly described when all variables in recipient and donor files are categorical ones. The method using loglinear model is introduced to perform a macro matching in D'Orazio *et al.* (2006).

Hotdeck matching, which is a most commonly used micro matching method, is similar to imputation in which both donor and recipient files are classified so that missing at random mechanism is satisfied, one observation in the same class is randomly chosen to impute the missing value in the recipient file. However, if the unit for the imputation does not exist in the same class of the donor file B, the random hotdeck does not work and both donor and recipient files are manually merged with the adjacent category, which could cause for violation of CIA.

To overcome the weakness of existing hotdeck matching methods, we propose a mixed method using multinomial logistic regression model, which can be used both for micro and macro approach, when recipient and donor files are composed of categorical variables.

The first proposed mixed method using multinomial logistic regression, which is denoted by mixed method using distance hotdeck under CIA (MHC), consists of following three steps.

(1) From the donor file B, the estimated probability $\hat{\pi}_k^D$ is obtained by fitting a multinomial logistic regression model with a set of X-independent variables and Z-dependent variable, where π_k is the probability of element being in k^{th} category.

(2) For observation in the recipient file A, the predicted probability $\hat{\pi}_k^R$ is obtained from following equation.

$$\hat{\pi}_k^R = \frac{\exp\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}' \boldsymbol{x}\right)}{1 + \sum_{i=1}^{K-1} \exp\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}' \boldsymbol{x}\right)}, \quad k = 1, \ldots, K-1,$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}'$ is estimated from the Step (1).

(3) Matching Step : a value of Z in the donor file B is imputed for the recipient file A using hotdeck method based on $(\hat{\pi}_k^D, \hat{\pi}_k^R)$ Manhattan, Euclidean, and Gower distance, and denote it $\hat{Z}$.

The second proposed mixed method, which is donoted by mixed method using randomization mechanism under CIA (MRC) consists of three steps, as well.

(1) Using the donor file B, fit a multinomial logistic regression model with a set of X-independent variables and Z-dependent variable.

(2) Same as the Step (2) of the MHC method.

(3) Predict category Z in the recipient file A by generating multinomial random variable with predicted probability $\hat{\pi}_k^R$ in Step (2).

Unlike the usual hotdeck method, the proposed method uses the set of predicted probability based on the multinomial logistic regression, no further process is necessary even when same classes in donor file are empty. The MRC method has the advantage of not only less burden of computation but also simplicity over the first method.

3.2. Statistical matching with auxiliary information when CIA does not hold

Both existing and proposed matching methods introduced in Section 3.1 are based on the CIA between Y and Z given X. If the CIA is not satisfied in the population, estimated joint distribution of (Y, Z) based on a statistically matched data has a serious bias. In this section, we propose a micro matching method in which auxiliary information is used and thus, it is applicable even when CIA is not satisfied. Auxiliary information can be obtained from the past data or by conducting a small survey in which all variables (X, Y, Z) are observed. The obtained auxiliary information can improve the performance of statistical matching when CIA does not hold.

As mentioned in Section 2, Singh *et al.* (1993) and Renssen (1998) considered the use of auxiliary information as an alternative to the CIA for statistical matching. The methods of Singh *et al.* (1993) apply existing matching methods with categorical constraints given after dividing the continuous variables in file A and B into categorical variables, which is not directly applicable when all the variables

in file A and B are categorical variables. Renssen (1998)'s method uses a linear probability model to predict an estimate of the probability of Z being in a specific category conditioning on the X in the recipient file A, instead of the actual observed value 0 and 1. And also, linear probability models have well known drawback that provides which is less than 0 or greater than 1 estimated probabilities (D'Orazio, 2017).

We propose several statistical matching mixed methods for categorical variables using multinomial logistic regression model when auxiliary information is available. The first method, which is denoted by mixed method with auxiliary information 1 (MA1) consists of following three steps.

(1) Using the file C, where X, Y, and Z are all observed, fit a multinomial logistic regression model in which Y is a set of independent variables and Z is a dependent variable.

(2) For recipient file A, the predicted probability $\hat{\pi}_k^R$ is obtained by using a multinomial logistic regression model constructed from the file C in Step (1).

(3) Predict category Z in the recipient file A by generating multinomial random variable with predicted probability $\hat{\pi}_k^R$ in Step (2).

The second method, which is denoted by MA2, also consists of three steps. MA2 is the same as MA1 except that independent variables in Step (1) are (X, Y).

(1) Using the file C, fit a multinomial logistic regression model in which (X, Y) are independent variables and Z is a dependent variable.

(2)–(3) Steps are the same as the (2)–(3) Steps of the MA1.

The third method, which is denoted by MA3, consists of following six steps.

(1) Using the file C, fit a multinomial logistic regression model in which Z is a set of independent variables and Y is a dependent variable.

(2) For the donor file B, the predicted probability $\hat{\pi}_j^D$ is obtained by using a multinomial logistic regression model constructed from the file C in Step (1).

(3) Predict a category of Y in the donor file B by generating multinomial random variable with predicted probability $\hat{\pi}_j^D$ in Step (2).

(4) Using the donor file B, fit a multinomial logistic regression model in which $(X, \hat{Y})$ are independent variables and Z is a dependent variable.

(5) For the recipient file A, the predicted probability $\hat{\pi}_k^R$ is obtained by using a multinomial logistic regression model constructed from the donor file B in Step (4).

(6) Predict a category of Z in the recipient file A by generating multinomial random variable with predicted probability $\hat{\pi}_k^R$ in Step (5).

The fourth method, which is denoted by MA4, consists of six steps, as well. MA4 is the same as MA3 except that independent variables in Step (1) are (X, Z).

(1) Using the file C, fit a multinomial logistic regression model in which (X, Z) are independent variables and Y is a dependent variable.

(2)–(6) Steps are the same as the (2)–(6) Steps of the MA3.

The proposed methods, which are mixture of parametric and nonparametric method, could be applied to match the files even when CIA does not hold.

4. Simulation study

We conduct a simulation study to compare the performance of several matching methods including the proposed ones, when all variables, in both files, consist of categorical variables. For the simulation study, we generate a population of size 100,000 with (X, Y, Z) variables. The vector of common variable, $\boldsymbol{X_d}$, appeared in both donor and recipient files, is

$$\boldsymbol{X_d} = \left(\boldsymbol{x'_{1d}}, \boldsymbol{x'_{2d}}, \boldsymbol{x'_{3d}}\right)' = (x_{11}, x_{12}, x_{13}, x_{14}, x_{21}, x_{22}, x_{23}, x_{31}, x_{32})',$$

where

$$x_{ij} = \begin{cases} 1, & \text{if the observation belongs to } j^{th} \text{ categories for } i\text{th categorical variable,} \\ 0, & \text{otherwise.} \end{cases}$$

That is, we considered 3 categorical variables that have 4, 3, and 2 categories, respectively. Y is a categorical variable of 4 categories with probability, π_j^y, of appearing in the category is given by multinomial logistic regression model such that $\boldsymbol{Y} = (y_1, y_2, y_3, y_4) \sim \text{Multinomial}(\pi_1^y, \pi_2^y, \pi_3^y, \pi_4^y)$

$$\ln \frac{\pi_j^y}{\pi_4^y} = \boldsymbol{\beta}_j^{'y} \boldsymbol{X_d}, \tag{4.1}$$

where

$$\begin{aligned}
\boldsymbol{\beta}_{\mathbf{1}}^{'y} &= (-0.1, 0.1, -0.1, -0.1, 0.1, -0.1, 0.1, 0.1, -0.1), \\
\boldsymbol{\beta}_{\mathbf{2}}^{'y} &= (0.1, -0.1, 0.1, -0.1, 0.1, 0.1, -0.1, 0.1, -0.1), \\
\boldsymbol{\beta}_{\mathbf{3}}^{'y} &= (0.1, -0.1, 0.1, -0.1, -0.1, 0.1, -0.1, -0.1, 0.1).
\end{aligned}$$

To evaluate the performance of the different matching method in various association strength of (X, Y, Z), we considered 8 different Z-variables, that are generated from the model,

$$\ln \frac{\pi_k^{z_l}}{\pi_3^{z_l}} = \boldsymbol{\beta}_k^{'z_l} \boldsymbol{X_d} + \boldsymbol{\gamma}_k^{'z_l} \boldsymbol{Y_d}, \tag{4.2}$$

where different categories, $(k = 1, 2)$ and different senarios $(l = 1, 2, \ldots, 8)$, and $\boldsymbol{Y_d}$ is a vector of 4 dummy variables. The coefficients used to generate Z-value in (4.2) are summarized in Table 1. Note that, Z_1 to Z_4 are generated so that no association exists with X but different level of association exists with Y. Z_5 to Z_8 are generated so that association exists with X and different level of association exists with Y.

In each replication, recipient sample A of size 1,000 and donor sample B of size 4,000 were selected using simple random sampling. As noted, we assumed only (X, Y) are observed in the recipient file A and (X, Z) are observed in the donor file B. To evaluate the performance of statistical matching methods, we consider the estimation of contingency table of Y and Z using the matched data. It is because the inference on the categorical variables usually based on the contingency table.

For MHC, we applied several distance measure such as Manhattan, Euclidean, Gower distance function, and a constrained statistical matching method using Manhattan distance function. For details about various distance measures, see D'Orazio *et al.* (2006). For the comparison, we use the total

Table 1: β_k' and γ_k' in (4.2)

	k	β_k'	γ_k'
Z_1	1	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0)
	2	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0)
Z_2	1	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	(0.3, 0, 0.3, 0)
	2	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	(0, 0.3, 0, 0.3)
Z_3	1	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	(0.7, 0, 0.7, 0)
	2	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	(0, 0.7, 0, 0.7)
Z_4	1	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	(1, 0, 1, 0)
	2	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	(0, 1, 0, 1)
Z_5	1	(0.2, 0, 0.2, 0, 0.2, 0, 0.2, 0, 0.2)	(0, 0, 0, 0)
	2	(0, −0.2, 0, −0.2, 0, −0.2, 0, −0.2, 0)	(0, 0, 0, 0)
Z_6	1	(0.2, 0, 0.2, 0, 0.2, 0, 0.2, 0, 0.2)	(0.3, 0, 0.3, 0)
	2	(0, −0.2, 0, −0.2, 0, −0.2, 0, −0.2, 0)	(0, 0.3, 0, 0.3)
Z_7	1	(0.2, 0, 0.2, 0, 0.2, 0, 0.2, 0, 0.2)	(0.7, 0, 0.7, 0)
	2	(0, −0.2, 0, −0.2, 0, −0.2, 0, −0.2, 0)	(0, 0.7, 0, 0.7)
Z_8	1	(0.2, 0, 0.2, 0, 0.2, 0, 0.2, 0, 0.2)	(1, 0, 1, 0)
	2	(0, −0.2, 0, −0.2, 0, −0.2, 0, −0.2, 0)	(0, 1, 0, 1)

Table 2: Comparison of TVD between random hotdeck and matching methods under CIA

	Cramer's V with Y	RAND	MHC(MAN)	MHC(EUC)	MHC(GOW)	MHC(M.C)	MRC
Z_1	0.006	0.0415	0.0419	0.0417	0.0420	0.0406	0.0418
Z_2	0.089	0.0673	0.0677	0.0679	0.0671	0.0673	0.0672
Z_3	0.204	0.1343	0.1334	0.1336	0.1339	0.1333	0.1332
Z_4	0.285	0.1849	0.1856	0.1850	0.1853	0.1838	0.1848
Z_5	0.005	0.0412	0.0417	0.0413	0.0410	0.0406	0.0415
Z_6	0.083	0.0647	0.0647	0.0645	0.0645	0.0644	0.0646
Z_7	0.195	0.1290	0.1289	0.1288	0.1303	0.1287	0.1302
Z_8	0.276	0.1817	0.1826	0.1829	0.1810	0.1830	0.1826

TVD = Total Variation Distance; CIA = Conditional Independence Assumption; RAND = RANDom hotdeck; MHC = Mixed method using distance Hotdeck under CIA; MAN = MANhattan distance; EUC = EUClidean distance; GOW = GOWer distance; M.C = Manhattan distance with Constrained matching; MRC = Mixed method using Randomization mechanism under CIA.

variation distance (TVD), which is a dissimilarity measure among marginal or joint distribution of categorical variables, and the formula is as shown below.

$$\text{TVD} = \frac{1}{2}\sum_{m=1}^{M}\left|f_{1,m} - f_{2,m}\right|, \tag{4.3}$$

where $f_{1,m}$ is a relative frequencies of $Y \times \hat{Z}$ contingency table in the matched file, and $f_{2,m}$ is a relative frequency of $Y \times Z$ contingency table in the population, and M is number of cells in the $Y \times Z$ contingency table. Table 2 shows the Monte Carlo mean of 1,000 values of TVD.

Cramer's V appeared in the table, is a measure of association between two nominal variables Y and Z that is ranged $[0, 1]$. Table 2 shows that the performance of the random hotdeck, MHC with several distance functions, and MRC are very similar. The higher the association between Z and Y, estimated joint distribution of (Y, Z) obtained from the matched files is significantly different from the true ones as expected. That is, if CIA is not satisfied in the population, all statistical micro matching methods are not useful in estimating the association of unique variables and thus all methods requires some modification to handle such a problem.

In the second simulation, we compared the performance of the statistical matching methods using

Table 3: Comparison of TVD between random hotdeck and methods using auxiliary information

	Cramer's V with Y	RAND	MA1	MA2	MA3	MA4
Z_1	0.006	0.0415	0.1196	0.1238	0.1083	0.1189
Z_2	0.089	0.0673	0.1192	0.1229	0.1084	0.1195
Z_3	0.204	0.1343	0.1179	0.1204	0.1053	0.1128
Z_4	0.285	0.1849	0.1135	0.1146	0.1020	0.1110
Z_5	0.005	0.0412	0.1182	0.1215	0.1078	0.1159
Z_6	0.083	0.0647	0.1137	0.1180	0.1014	0.1133
Z_7	0.195	0.1290	0.1127	0.1158	0.1028	0.1087
Z_8	0.276	0.1817	0.1099	0.1126	0.1004	0.1077

TVD = Total Variation Distance; RAND = RANDom hotdeck; MA = Mixed method with Auxiliary information.

Table 4: Comparison of TVD between RAND and MA3 with various sample size of file C

	Cramer's V with Y	RAND	MA3					
			50	100	150	200	250	300
Z_1	0.006	0.0415	0.1476	0.1083	0.0903	0.0808	0.0750	0.0698
Z_2	0.089	0.0673	0.1518	0.1084	0.0885	0.0806	0.0735	0.0704
Z_3	0.204	0.1343	0.1432	0.1053	0.0879	0.0786	0.0723	0.0687
Z_4	0.285	0.1849	0.1404	0.1020	0.0850	0.0754	0.0715	0.0664
Z_5	0.005	0.0412	0.1495	0.1078	0.0884	0.0789	0.0724	0.0693
Z_6	0.083	0.0647	0.1478	0.1014	0.0892	0.0801	0.0724	0.0700
Z_7	0.195	0.1290	0.1405	0.1028	0.0862	0.0769	0.0704	0.0681
Z_8	0.276	0.1817	0.1405	0.1004	0.0840	0.0748	0.0689	0.0657

TVD = Total Variation Distance; RAND = RANDom hotdeck; MA = Mixed method with Auxiliary information.

the auxiliary information suggested in Section 3.2 to the conventional random hotdeck method. To obtain auxiliary information, file C of size 100 with (X, Y, Z) variables, equivalent to 10% of the recipient file A, was sampled from the population. As in the previous simulation, the Monte Carlo mean of 1,000 values of TVD are shown in Table 3.

Table 3 shows that in the case of Z_1, Z_2, Z_5, Z_6, which has weak or almost no association with Y, random hotdeck show the better performance in estimating the joint distribution of (Y, Z) than other methods in which auxiliary information is incorporated. However, in the case of Z_3, Z_4, Z_7, Z_8, which are moderately associated with Y, the MA1 to MA4 methods show better performance than the random hotdeck. In particular, MA3 was found to have the best performance consistently among the methods of MA1 through MA4. In this simulation study, the method using auxiliary information on the (Y, Z) relationship shows the consistently better performance than using (X, Y, Z) relation information. In addition, random hotdeck shows poor performance, as the association of Y and Z is getting stronger. As expected, the statistical matching method which is valid under CIA is sensitive to the level of association between Y and Z, while the methods using auxiliary information are relatively insensitive to the CIA.

In the third simualtion, we compared the performance of matching method, RAND and MA3 with various sample size of file C. The performance of the MA3 method was compared to the random hotdeck, when the size of file C is 50, 100, 150, 200, 250, and 300. Table 4 shows that, the performance of MA3 in estimating the joint distribution of (Y, Z) is improved as the size of file C increased. However, as appeared in Figure 2, the amount of gain obtained by increasing the size of file C decrease as the size of file C increase. Note that almost no gain is obtained by increasing $n = 200$ to $n = 250$ in the Figure 2. Based on a limited simulation study, we recommend to do an appropriate size survey to obtain file C, if it is necessary.

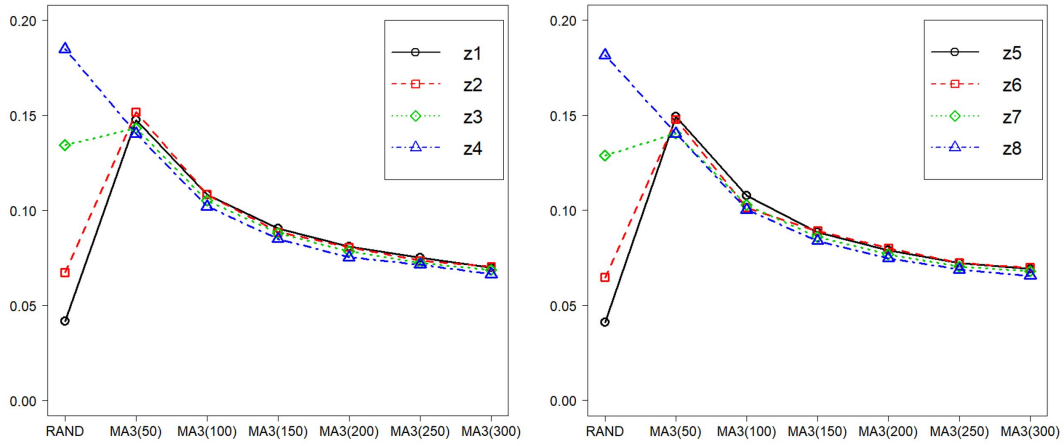

Figure 2: *Comparison of TVD between RAND and MA3 with various sample size of file C. TVD = total variation distance; RAND = Random hotdeck; MA = mixed method with auxiliary information.*

5. Conclusion

The main goal of a statistical micro matching is to generate synthetic data which has (X, Y, Z) and estimate the joint distribution of (Y, Z) and (or) (X, Y, Z). For example, if we have file A with information on the education level of persons, their gender, age and region and file B with information on the occupation of persons, their gender, age and region, the goal of statistical matching is to generate augmented data with all information and estimate the association between education level and occupation. In this paper, we proposed several mixed matching methods using a multinomial logistic regression model for categorical variables. First, we proposed a statistical matching method without auxiliary information under CIA. The performance of the random hotdeck and suggested method is very similar and higher the association between Y and Z, estimated joint distribution of (Y, Z) obtained from the matched file is significantly different from the true one. Although the proposed method introduced in Section 3.1 do not show significantly better performance than the random hotdeck, the proposed methods is useful when the random hotdeck is restricted. Second, we propose a statistical matching method with auxiliary information when CIA does not hold. The higher the association between Y and Z, the statistical matching method using auxiliary information shows better performance than random hotdeck. If moderate association between Y and Z is suspected, we recommend to apply the method introduced in Section 3.2 which use the auxiliary information instead of ones introduced in Section 3.1. Finally, the simulation result shows that the size of file C does not need to be large which means the cost to overcome the CIA would not be a serious concern in practice.

Acknowledgements

Supported by a Korea University Grant (K1910941).

References

Budd EC (1971). The creation of a microdata file for estimating the size distribution of income, *The Review of Income and Wealth*, **17**, 317–333.

D'Orazio M, Di Zio M, and Scanu M. (2006). *Statistical Matching: Theory and Practice*, John Wiley

& Sons, Chichester.

D'Orazio M (2017). *Statistical matching and imputation of survey data with statmatch* (Technical Paper). Available from: https://cran.r-project.org/web/packages/StatMatch/vignettes/Statistical_Matching_with_StatMatch.pdf

Maddala GS (1983). *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge.

Okner BA (1972). Constructing a new data base from existing microdata sets: the 1966 merge file, *Annals of Economic and Social Measurement*, **1**, 325–342.

Paass G (1986). Statistical matching: evaluation of existing procedures and improvements be using additional information. In *Microanalytic Simulation Models to Support Social and Financial Policy*, Elsevier Science, Amsterdam.

Renssen RH (1998). Use of statistical matching techniques in calibration estimation, *Survey Methodology*, **24**, 171–183.

Rodgers WL (1984). An evaluation of statistical matching, *Journal of Business and Economic Statistics*, **2**, 91–102.

Rubin DB (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics*, **4**, 87–94.

Sims CA (1972). Comment on Okner (1972), *Annals of Economic and Social Measurement*, **1**, 343–345.

Singh AC (1988). Log-linear imputation, Methodology Branch Working Paper, SSMD, 88-029E, Statistics Canada; also published in *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 118–132.

Singh AC, Mantel H, Kinack M, and Rowe G (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption, *Survey Methodology*, **19**, 59–79.

Received July 8, 2019; Revised August 23, 2019; Accepted August 24, 2019