The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.2.331
25*(2), 331–339*

A Robust Approach of Regression-Based Statistical Matching for Continuous Data

Sooncheol Sohn[1] · Myoungshic Jhun[2]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University

(Received February 4, 2012; Revised March 20, 2012; Accepted March 30, 2012)

Abstract

Statistical matching is a methodology used to merge microdata from two (or more) files into a single matched file, the variants of which have been extensively studied. Among existing studies, we focused on Moriarity and Scheuren's (2001) method, which is a representative method of statistical matching for continuous data. We examined this method and proposed a revision to it by using a robust approach in the regression step of the procedure. We evaluated the efficiency of our revised method through simulation studies using both simulated and real data, which showed that the proposed method has distinct advantages over existing alternatives.

Keywords: Donor file, recipient file, matched file, common variable, unique variable, statistical matching.

1. Introduction

Let $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ be a random vector with density $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, where $\boldsymbol{X} = (X_1, \ldots, X_p)$, $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_r)$ are vectors of random variables of dimension p, q and r, respectively. $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is assumed to have a nonsingular distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ as in (1.1).

$$(\boldsymbol{\mu}, \Sigma) = \left(\begin{pmatrix} \boldsymbol{\mu_X} \\ \boldsymbol{\mu_Y} \\ \boldsymbol{\mu_Z} \end{pmatrix}, \begin{pmatrix} \Sigma_{\boldsymbol{XX}} & \Sigma_{\boldsymbol{XY}} & \Sigma_{\boldsymbol{XZ}} \\ \Sigma_{\boldsymbol{YX}} & \Sigma_{\boldsymbol{YY}} & \Sigma_{\boldsymbol{YZ}} \\ \Sigma_{\boldsymbol{ZX}} & \Sigma_{\boldsymbol{ZY}} & \Sigma_{\boldsymbol{ZZ}} \end{pmatrix} \right). \tag{1.1}$$

Suppose that A and B are two files consisting of n_A and n_B, independent and identically distributed observations generated from $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. Assume that $\boldsymbol{X}$ is observed in both files and $\boldsymbol{Y}$ appears only in one file (A), while $\boldsymbol{Z}$ is observed exclusively in the second file (B) as in (1.2). $\boldsymbol{X}$ are thus common variables, whereas $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are unique variables.

$$\begin{aligned} \text{File } A: \left(\boldsymbol{x}_a^A, \boldsymbol{y}_a^A\right) &= \left(x_{a1}^A, \ldots, x_{ap}^A,\ y_{a1}^A, \ldots, y_{aq}^A\right), \quad a = 1, \ldots, n_A, \\ \text{File } B: \left(\boldsymbol{x}_b^B, \boldsymbol{z}_b^B\right) &= \left(x_{b1}^B, \ldots, x_{bp}^B,\ z_{b1}^B, \ldots, z_{br}^B\right), \quad b = 1, \ldots, n_B. \end{aligned} \tag{1.2}$$

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2010-0009204).
[2]Corresponding author: Department of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: jhun@korea.ac.kr

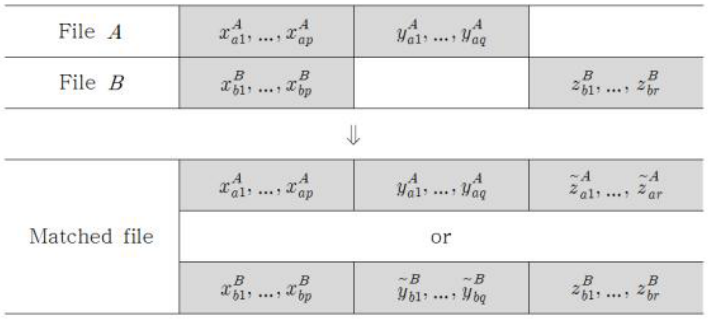

Figure 1.1. Illustration of statistical matching

Record linkage, or exact matching, is a methodology designed to merge the same entities from two (or more) different files. On the other hand, statistical matching, as a methodology designed to allocate microdata from similar entities, is used to merge microdata from two (or more) files into a single matched file (see Figure 1.1), where $\tilde{z}^A_{ak}$, $k = 1, \ldots, r$ and $\tilde{y}^B_{bj}$, $j = 1, \ldots, q$ denote the matched unique variables from file A and B, respectively.

Statistical matching and its variants have been discussed by Okner (1972, 1974), Sims (1972, 1974), Paass (1982), and Rodgers (1984). Of particular note, Kadane (1978) and Rubin (1986) described regression-based procedures to produce a matched file. When using Kadane's (1978) method, Moriarity and Scheuren (2001) observed that the correlation coefficients between unique variables are not preserved during the matching procedure. Also, Moriarity and Scheuren (2003) indicated that Rubin (1986) did not consider the preservation of the correlation matrix structure, and asserted that the way in which his method estimated the secondary predicted values was redundant. In other words, it was proposed that Moriarity and Scheuren's (2001) method could improve on the existing regression-based procedures of Kadane (1978) and Rubin (1986).

In this paper, we focused on Moriarity and Scheuren's method as being paradigmatic of statistical matching for continuous data, and proposed a revision by using a robust approach in the regression step of the procedure. Furthermore, through simulation studies, using both simulated and real data, we showed that our proposed method represents an improvement on that of Moriarity and Scheuren. The rest of this paper is organized as follows: Section 2 briefly examines Moriarity and Scheuren's method, highlights its problems, and proposes an improved statistical matching technique; Section 3 and Section 4 present the results of the simulation and case studies; and Section 5 concludes.

2. Statistical Matching Methods

In a statistical matching framework, we naturally assume the conditional independence of $(\boldsymbol{Y}, \boldsymbol{Z})$, given $\boldsymbol{X}$. This was assumed, either explicitly or implicitly, in all previous statistical matching applications. This assumption is usually referred to as the conditional independence assumption(CIA), and $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})$ when the CIA holds. Moreover, given the multivariate normality of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, $\Sigma_{\boldsymbol{YZ}} = \Sigma_{\boldsymbol{YX}}(\Sigma_{\boldsymbol{XX}})^{-1}\Sigma_{\boldsymbol{XZ}}$ and the conditional expectations of $\boldsymbol{Z}$ and $\boldsymbol{Y}$ in files A and B will be simplified as follows:

$$E(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}) = E(\boldsymbol{Z}|\boldsymbol{X}) = \boldsymbol{\mu_Z} + \Sigma_{\boldsymbol{ZX}}\Sigma^{-1}_{\boldsymbol{XX}}(\boldsymbol{X} - \boldsymbol{\mu_X}), \tag{2.1}$$

$$E(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{\mu_Y} + \Sigma_{\boldsymbol{YX}}\Sigma^{-1}_{\boldsymbol{XX}}(\boldsymbol{X} - \boldsymbol{\mu_X}). \tag{2.2}$$

The previous argument can easily be extended to multivariate linear regression models

$$\boldsymbol{y}_a^A = \left(\boldsymbol{X}_a^A\right)^T \boldsymbol{\beta}_{\boldsymbol{Y}} + \boldsymbol{\epsilon}_a^A, \quad a = 1, \ldots, n_A, \tag{2.3}$$

$$\boldsymbol{z}_b^B = \left(\boldsymbol{X}_b^B\right)^T \boldsymbol{\beta}_{\boldsymbol{Z}} + \boldsymbol{\epsilon}_b^B, \quad b = 1, \ldots, n_B, \tag{2.4}$$

where $\boldsymbol{X}_a^A$ and $\boldsymbol{X}_b^B$ are $(p+1)q \times q$ and $(p+1)r \times r$ block diagonal matrices representing p explanatory variables, with the diagonal elements being $(1, \boldsymbol{x}_a^A)$ and $(1, \boldsymbol{x}_b^B)$ and the off-diagonal elements zeros, respectively. $\boldsymbol{\beta}_{\boldsymbol{Y}}$ and $\boldsymbol{\beta}_{\boldsymbol{Z}}$ are $(p+1)q \times 1$ and $(p+1)r \times 1$ vectors of regression coefficients, and $\boldsymbol{\epsilon}_a^A$ and $\boldsymbol{\epsilon}_b^B$ are $q \times 1$ and $r \times 1$ error vectors which are distributed as multivariate normal with mean vectors $\mathbf{0}$ and covariance matrices $\Sigma_{\boldsymbol{YY}}$ and $\Sigma_{\boldsymbol{ZZ}}$, respectively. The conditional expectations (2.1) and (2.2) can then be estimated as (2.5) and (2.6), respectively.

$$\hat{\boldsymbol{z}}_a^A = \left(\boldsymbol{X}_a^A\right)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}, \quad a = 1, \ldots, n_A, \tag{2.5}$$

$$\hat{\boldsymbol{y}}_b^B = \left(\boldsymbol{X}_b^B\right)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}, \quad b = 1, \ldots, n_B, \tag{2.6}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}$ are least squares estimators(LSEs) in the files B and A, respectively.

2.1. Moriarity and Scheuren's (2001) method

Moriarity and Scheuren (2001) defined file A as the donor file and file B as the recipient file in (1.2). They assumed that $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ have a nonsingular multivariate normal distribution with mean vector and covariance matrix as in (1.1), where $p = q = r = 1$. Their method consists of a regression step and a matching step, where the regression step produces estimates of the "missing" values as in (2.1) and (2.2), and the matching step produces the matched file. When the CIA holds, their method based on the models (2.3) and (2.4) can be summarized as follows:

<Moriarity and Scheuren's method>

- Regression step

Stage 1: In models (2.3) and (2.4), obtain the LSEs $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}$ for $\boldsymbol{\beta}_{\boldsymbol{Y}}$ and $\boldsymbol{\beta}_{\boldsymbol{Z}}$, respectively.

Stage 2: Calculate the predicted values $\hat{\boldsymbol{z}}_a^A = (\boldsymbol{X}_a^A)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}$, $a = 1, \ldots, n_A$ and $\hat{\boldsymbol{y}}_b^B = (\boldsymbol{X}_b^B)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}$, $b = 1, \ldots, n_B$ in the donor and recipient files, respectively.

Stage 3: Add the random residuals $\boldsymbol{r}_a^A \sim \mathrm{MVN}(\mathbf{0}, \Sigma_{\boldsymbol{ZZ}} - T_1)$ and $\boldsymbol{r}_b^B \sim \mathrm{MVN}(\mathbf{0}, \Sigma_{\boldsymbol{YY}} - T_2)$ to $\hat{\boldsymbol{z}}_a^A$ and $\hat{\boldsymbol{y}}_b^B$ for $a = 1, \ldots, n_A$ and $b = 1, \ldots, n_B$, respectively. If $\Sigma_{\boldsymbol{ZZ}} - T_1$ ($\Sigma_{\boldsymbol{YY}} - T_2$) is not nonsingular, then $\boldsymbol{r}_a^A = \mathbf{0}$ ($\boldsymbol{r}_b^B = \mathbf{0}$), where $T_1 = \Sigma_{\boldsymbol{ZX}} \Sigma_{\boldsymbol{XX}}^{-1} \Sigma_{\boldsymbol{XZ}}$ and $T_2 = \Sigma_{\boldsymbol{YX}} \Sigma_{\boldsymbol{XX}}^{-1} \Sigma_{\boldsymbol{XY}}$.

- Matching step

Stage 1: Calculate the Mahalanobis distance in $\{(\boldsymbol{Y}^A, \hat{\boldsymbol{Z}}^A + \boldsymbol{R}^A), (\hat{\boldsymbol{Y}}^B + \boldsymbol{R}^B, \boldsymbol{Z}^B)\}$, where $\boldsymbol{R}^A = (\boldsymbol{r}_1^A, \ldots, \boldsymbol{r}_{n_A}^A)^T$ and $\boldsymbol{R}^B = (\boldsymbol{r}_1^B, \ldots, \boldsymbol{r}_{n_B}^B)^T$.

Stage 2: Carry out the constrained match using RELAX-IV software (Bertsekas, 1991; Bertsekas and Tseng, 1994), which is able to conduct a constrained match by solving a "transportation" linear programming problem.

Improving on Kadane (1978), Moriarity and Scheuren added random residuals $\boldsymbol{r}_a^A$ and $\boldsymbol{r}_b^B$ to the $\hat{\boldsymbol{z}}_a^A$ and $\hat{\boldsymbol{y}}_b^B$ for Stage 3 of the regression step. When $\Sigma_{\boldsymbol{ZZ}} - T_1(\Sigma_{\boldsymbol{YY}} - T_2)$ is nonsingular, in

contrast with Kadane (1978), the covariance matrices of $(\boldsymbol{X}^A, \boldsymbol{Y}^A, \hat{\boldsymbol{Z}}^A + \boldsymbol{R}^A)$ and $(\boldsymbol{X}^B, \hat{\boldsymbol{Y}}^B + \boldsymbol{R}^B, \boldsymbol{Z}^B)$ are nonsingular matrices and the expected values of these are equal to the population covariance matrix Σ. Furthermore, in order to preserve the covariance structure, Moriarity and Scheuren conducted a constrained match excepting common variable $\boldsymbol{X}$ in the matching step. Consequently, compared with Kadane (1978) and Rubin (1986), they showed improved results, especially in terms of preservation of the correlation between $\boldsymbol{Y}$ and $\boldsymbol{Z}$.

However, Moriarity and Scheuren assumed that $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ follows a multivariate normal distribution, which can lead to two problems. First, since the LSE in the regression step is known to be best under the normality assumption of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, the method may not be appropriate when the distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ has outliers or a heavy tail. Secondly, because the method used random residuals $\boldsymbol{r}_a^A$ and $\boldsymbol{r}_b^B$ from the normal distribution, if the distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is not normal, these random residuals could be problematic.

2.2. Proposed method

First, in model (2.3), the estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}^{LDE}$ for $\boldsymbol{\beta}_{\boldsymbol{Y}}$ that minimizes over all possible values of $\boldsymbol{\beta}_{\boldsymbol{Y}}$

$$\sum_{a=1}^{n_A} \left\| \boldsymbol{y}_a^A - \left(\boldsymbol{X}_a^A \right)^T \boldsymbol{\beta}_{\boldsymbol{Y}} \right\| \tag{2.7}$$

is called the least distance estimator(LDE), where $||\cdot||$ denotes the usual Euclidean distance. In model (2.4), we can obtain the estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}^{LDE}$ for $\boldsymbol{\beta}_{\boldsymbol{Z}}$ analogously. The LDE is an estimator considering the relationship among response variables, and the relative efficiency of the LDE with respect to the least absolute estimator increases as the correlation between the response variables increases (Jhun and Choi, 2009). Additionally, Bai *et al.* (1990) found that the LDE is robust when the data contain outliers. Although the LSE is the most widely used estimator in regression modeling, it can be seriously affected when outliers exist and its estimation process fails to take into account the relationship among response variables. Thus, if the LSE is used, valuable information contained in the interdependence structure of the response variables may be overlooked, unlike when the LDE is employed. For these reasons, in statistical matching the LDE can be considered as an alternative to the LSE when the distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ has outliers or a heavy tail.

Next, consider the empirical residuals $\boldsymbol{u}_a^A = \boldsymbol{y}_a^A - (\boldsymbol{X}_a^A)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}^{LDE}$, $a = 1, \ldots, n_A$ and $\boldsymbol{u}_b^B = \boldsymbol{z}_b^B - (\boldsymbol{X}_b^B)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}^{LDE}$, $b = 1, \ldots, n_B$ in the donor and recipient files, respectively. Then, $\boldsymbol{U}^A = (\boldsymbol{u}_1^A, \ldots, \boldsymbol{u}_{n_A}^A)^T$ and $\boldsymbol{U}^B = (\boldsymbol{u}_1^B, \ldots, \boldsymbol{u}_{n_B}^B)^T$ are distribution-free and have the same mean vectors and covariance matrices as those of $\boldsymbol{R}^A$ and $\boldsymbol{R}^B$, respectively. Because of these properties, $\boldsymbol{U}^A$ and $\boldsymbol{U}^B$ can be valid alternatives to the $\boldsymbol{R}^A$ and $\boldsymbol{R}^B$ even when the distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is normal. Therefore, we propose an alternative statistical matching method which revises that of Moriarity and Scheuren as follows:

<Proposed method>

- Regression step

Stage 1: In models (2.3) and (2.4), obtain the LDEs $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}^{LDE}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}^{LDE}$ for $\boldsymbol{\beta}_{\boldsymbol{Y}}$ and $\boldsymbol{\beta}_{\boldsymbol{Z}}$, respectively.

Stage 2: Calculate the predicted values $\hat{\boldsymbol{z}}_a^A = (\boldsymbol{X}_a^A)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}^{LDE}$, $a = 1, \ldots, n_A$ and $\hat{\boldsymbol{y}}_b^B = (\boldsymbol{X}_b^B)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}^{LDE}$, $b = 1, \ldots, n_B$ in the donor and recipient files, respectively.

Stage 3: Add randomly the empirical residuals $\boldsymbol{u}_a^A = \boldsymbol{y}_a^A - (\boldsymbol{X}_a^A)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}^{LDE}$ and $\boldsymbol{u}_b^B = \boldsymbol{z}_b^B - (\boldsymbol{X}_b^B)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}^{LDE}$ to $\hat{\boldsymbol{y}}_b^B$ and $\hat{\boldsymbol{z}}_a^A$ for $a = 1, \ldots, n_A$ and $b = 1, \ldots, n_B$, respectively.

- Matching step

Stage 1: Calculate the Mahalanobis distance in $\{(\boldsymbol{Y}^A, \hat{\boldsymbol{Z}}^A + \boldsymbol{U}^A), (\hat{\boldsymbol{Y}}^B + \boldsymbol{U}^B, \boldsymbol{Z}^B)\}$.

Stage 2: For each $b = 1, \ldots, n_B$, impute the live value y_{a^*} corresponding to the closest entity a^* in file A with respect to the Mahalanobis distance, *i.e.* the unconstrained match.

For matching, either the constrained or the unconstrained match can be used in the matching step, but we employ the unconstrained match in order to obtain the closest entity to the reality. In fact, the proposed method replaces the LSE used in that of Moriarity and Scheuren with the LDE in Stage 1, and also swaps random residuals for empirical residuals in Stage 3 of the regression step. That is to say, we propose a robust revision of Moriarity and Scheuren's method. We thus expect that the proposed method will create a matched file which is closer to the reality than is produced by that of Moriarity and Scheuren, especially when the distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is not normal.

3. Simulation Study

In order to compare the proposed method with that of Moriarity and Scheuren in performance, we carried out a simulation study. We considered two cases: one in which $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are all univariate (Case 1), and another in which $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are all multivariate (Case 2). For both cases, we set up the mean vector and covariance matrix of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ as

$$(\boldsymbol{\mu}_1, \Sigma_1) = \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{1} & \Sigma_{\boldsymbol{XY}} & \Sigma_{\boldsymbol{XZ}} \\ \Sigma_{\boldsymbol{XY}} & \mathbf{1} & \Sigma_{\boldsymbol{YZ}} \\ \Sigma_{\boldsymbol{XZ}} & \Sigma_{\boldsymbol{YZ}} & \mathbf{1} \end{pmatrix} \right), \tag{3.1}$$

where $\Sigma_{\boldsymbol{YZ}} = \Sigma_{\boldsymbol{XY}} \Sigma_{\boldsymbol{XZ}}$ and the elements of $(\Sigma_{\boldsymbol{XY}}, \Sigma_{\boldsymbol{XZ}})$ are taken as random draws from a Uniform $(0, 1)$ distribution, but Σ_1 should be nonsingular.

3.1. Case 1: $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are all univariate

We considered four types of distributions for (X_1, Y_1, Z_1), where $p = q = r = 1$:

i) a multivariate normal distribution $\mathrm{MVN}(\boldsymbol{\mu}_1, \Sigma_1)$, where the mean vector $\boldsymbol{\mu}_1$ and covariance matrix Σ_1 are the same as in (3.1).

ii) a contaminated multivariate normal distribution $0.9\mathrm{MVN}(\boldsymbol{\mu}_1, \Sigma_1) + 0.1\mathrm{MVN}(\boldsymbol{\mu}_1, \Sigma_2)$, where the mean vector $\boldsymbol{\mu}_1$ and covariance matrix Σ_1 are as in i) and Σ_2 is the same as Σ_1 except the variance of Y_1 is 100

iii) a multivariate t distribution with 3 degrees of freedom with covariance matrix Σ_1

iv) a multivariate Cauchy distribution

We generated $n = 2000$ observations of (X_1, Y_1, Z_1) from each of the four types of distributions and randomly separated them into two files of the same size, $n_A = n_B = 1000$. We then deleted the variable Y_1 in one file and Z_1 in another, in order to create the recipient and donor files, respectively. We performed statistical matching by using both our proposed method and that of

Table 3.1. The means and the standard deviations(SD) of evaluation factors for Case 1

Dist.	Matching methods	Y_1^D		$\mathrm{Corr}(X_1, \widetilde{Y}_1)$		$\mathrm{Corr}(\widetilde{Y}_1, Z_1)$	
		Mean	SD	Mean	SD	Mean	SD
MVN	(Real value)	0	0	0.440	0.272	0.225	0.202
	M & S	1125	24	0.437	0.281	0.220	0.206
	Proposed	1127	22	0.436	0.279	0.221	0.214
CMN	(Real value)	0	0	0.135	0.201	0.077	0.129
	M & S	2473	38	0.117	0.191	0.171	0.191
	Proposed	2364	41	0.139	0.211	0.076	0.103
Mt(3)	(Real value)	0	0	0.487	0.329	0.213	0.279
	M & S	1656	31	0.363	0.292	0.315	0.326
	Proposed	1599	30	0.455	0.304	0.206	0.223
MC	(Real value)	0	0	0.446	0.681	0.207	0.567
	M & S	14325	20639	0.102	0.459	0.082	0.504
	Proposed	10470	5820	0.304	0.439	0.164	0.234

M & S: Moriarity and Scheuren's (2001) method
Proposed: the proposed method
MVN: multivariate normal
CMN: contaminated multivariate normal
Mt(3): multivariate t with 3 degrees of freedom
MC: multivariate Cauchy

Moriarity and Scheuren. For the assessment of sampling variation, this procedure was repeated 100 times independently.

Usually, the efficiency of a matched file is determined by two factors. First, how close the values of the matched unique variable are to the real values, and secondly, how similar the correlation coefficients between the matched unique variable and the other variables are to the real correlation coefficients. To assess the former, we calculated the sum of absolute differences between the real $y_{b1}\,(b = 1, \ldots, n_B)$ in the original recipient file and the matched $\widetilde{y_{b1}}\,(b = 1, \ldots, n_B)$, denoting this sum as Y_1^D. For the latter, we compared the $\mathrm{Corr}(X_1, \widetilde{Y}_1)$ and $\mathrm{Corr}(\widetilde{Y}_1, Z_1)$ in the matched file with the real $\mathrm{Corr}(X_1, Y_1)$ and $\mathrm{Corr}(Y_1, Z_1)$, respectively.

Table 3.1 shows the means and the standard deviations of Y_1^D, $\mathrm{Corr}(X_1, \widetilde{Y}_1)$, and $\mathrm{Corr}(\widetilde{Y}_1, Z_1)$ based on 100 independent repetitions. This shows that when the distribution of (X_1, Y_1, Z_1) is multivariate normal, for Y_1^D, the results from Moriarity and Scheuren's (2001) method and those from the proposed method are not significantly different. The correlation coefficients suggest that both methods well preserved the real values, as expected. However, when the distribution of (X_1, Y_1, Z_1) is not multivariate normal, the means of Y_1^D from the proposed method are smaller than those from Moriarity and Scheuren's method, and the correlation coefficients from the proposed method are closer to the real values in the proposed method. Generally, since the variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are never jointly observed, researchers tend to be more interested in the preserving of the correlation of $(\boldsymbol{Y}, \boldsymbol{Z})$ than that of $(\boldsymbol{X}, \boldsymbol{Y})$. In this respect, the proposed method has greater potential utility.

3.2. Case 2: $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are all multivariate

We extended the simulation study to a multivariate situation. We considered four types of distributions for $(X_1, X_2, Y_1, Y_2, Z_1, Z_2)$, where $p = q = r = 2$. The features of the distributions are similar to those in Case 1, where the variances of Y_j, $j = 1, 2$ equal 100 in Σ_2 for a contaminated multivariate normal distribution. As in Case 1, we generated $n = 2000$ observations of $(X_1, X_2, Y_1, Y_2, Z_1, Z_2)$

Table 3.2. The means and the standard deviations (SD) of evaluation factors for Case 2

Dist.	Matching methods	Y_1^D	Y_2^D	Correlation coefficient of				Correlation coefficient of			
				$(X_1, \widetilde{Y}_1)$	$(X_1, \widetilde{Y}_2)$	$(X_2, \widetilde{Y}_1)$	$(X_2, \widetilde{Y}_2)$	$(\widetilde{Y}_1, Z_1)$	$(\widetilde{Y}_1, Z_2)$	$(\widetilde{Y}_2, Z_1)$	$(\widetilde{Y}_2, Z_2)$
		Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
MVN	(Real value)	0 (0)	0 (0)	0.422 (0.230)	0.422 (0.227)	0.418 (0.227)	0.421 (0.228)	0.239 (0.143)	0.244 (0.157)	0.243 (0.159)	0.240 (0.145)
	M & S	1049 (119)	1045 (103)	0.234 (0.176)	0.235 (0.176)	0.104 (0.115)	0.110 (0.126)	0.098 (0.105)	0.092 (0.089)	0.105 (0.121)	0.098 (0.099)
	Proposed	870 (178)	886 (175)	0.405 (0.226)	0.401 (0.223)	0.403 (0.223)	0.408 (0.221)	0.234 (0.140)	0.238 (0.151)	0.237 (0.154)	0.227 (0.140)
CMN	(Real value)	0 (0)	0 (0)	0.175 (0.101)	0.171 (0.102)	0.174 (0.102)	0.174 (0.107)	0.104 (0.074)	0.110 (0.076)	0.101 (0.078)	0.103 (0.081)
	M & S	1961 (212)	1949 (206)	0.076 (0.059)	0.075 (0.068)	0.032 (0.045)	0.037 (0.053)	0.038 (0.050)	0.038 (0.047)	0.035 (0.056)	0.040 (0.048)
	Proposed	1738 (223)	1754 (231)	0.152 (0.090)	0.162 (0.088)	0.159 (0.092)	0.159 (0.093)	0.102 (0.070)	0.100 (0.072)	0.100 (0.069)	0.104 (0.077)
Mt(3)	(Real value)	0 (0)	0 (0)	0.412 (0.257)	0.402 (0.258)	0.405 (0.250)	0.422 (0.242)	0.270 (0.222)	0.254 (0.200)	0.244 (0.201)	0.264 (0.223)
	M & S	1506 (189)	1527 (204)	0.244 (0.176)	0.252 (0.171)	0.108 (0.128)	0.122 (0.127)	0.115 (0.119)	0.117 (0.127)	0.120 (0.118)	0.121 (0.132)
	Proposed	1188 (266)	1284 (243)	0.349 (0.208)	0.347 (0.204)	0.362 (0.197)	0.360 (0.21)	0.227 (0.168)	0.221 (0.156)	0.212 (0.157)	0.231 (0.165)
MC	(Real value)	0 (0)	0 (0)	0.453 (0.564)	0.445 (0.568)	0.477 (0.537)	0.324 (0.581)	0.220 (0.693)	0.265 (0.645)	0.162 (0.657)	0.197 (0.637)
	M & S	33633 (101940)	50517 (20101)	0.135 (0.322)	0.151 (0.346)	0.104 (0.295)	0.070 (0.317)	0.037 (0.310)	0.006 (0.332)	0.106 (0.373)	0.106 (0.360)
	Proposed	9840 (11717)	9832 (8379)	0.176 (0.286)	0.281 (0.328)	0.243 (0.301)	0.194 (0.336)	0.140 (0.294)	0.125 (0.276)	0.109 (0.360)	0.156 (0.331)

M & S: Moriarity and Scheuren's (2001) method
Proposed: the proposed method
MVN: multivariate normal
CMN: contaminated multivariate normal
Mt(3): multivariate t with 3 degrees of freedom
MC: multivariate Cauchy

from each of the four types of distributions and randomly separated them into two files of the same size, $n_A = n_B = 1000$. We then deleted the variable (Y_1, Y_2) in one file and (Z_1, Z_2) in another, in order to create the recipient and donor files, respectively. We performed statistical matching by using the proposed method and that of Moriarity and Scheuren. For the reduction of sampling variation, this procedure was repeated 100 times independently, as in Case 1.

Table 3.2 shows the means and the standard deviations of the evaluation factors from 100 independent repetitions. Overall, the results are similar to those in Table 3.1. However, when the distribution of $(X_1, X_2, Y_1, Y_2, Z_1, Z_2)$ is multivariate normal, the means of Y_1^D and Y_2^D based on the proposed method are lower than those based on Moriarity and Scheuren's method. Besides this, the means of $\mathrm{Corr}(X_i, \widetilde{Y}_j)$, $i, j = 1, 2$ and $\mathrm{Corr}(\widetilde{Y}_j, Z_k)$, $j, k = 1, 2$ based on the proposed method are closer to the real values than those based on Moriarity and Scheuren's method. Since the LDE takes into account the relationship among response variables, such results seem to hold even when the distribution of $(X_1, X_2, Y_1, Y_2, Z_1, Z_2)$ is multivariate normal.

4. Real Example

In order to demonstrate the usefulness of the proposed method in a real situation, the method was applied to the Boston Housing Data found in Harrison and Rubinfeld (1978). The data set

File A (donor file)	DIS (X_1), LSTAT (X_2)	NOX (Y_1), MEDV (Y_2)	
File B (recipient file)	DIS (X_1), LSTAT (X_2)		RM (Z_1), CRIM (Z_2)

Figure 4.1. The donor and recipient files using Boston Housing Data

Table 4.1. The means and the standard deviations (SD) of evaluation factors for the real example

Matching methods	Y_1^D	Y_2^D	Correlation coefficient of				Correlation coefficient of			
			$(X_1, \widetilde{Y}_1)$	$(X_1, \widetilde{Y}_2)$	$(X_2, \widetilde{Y}_1)$	$(X_2, \widetilde{Y}_2)$	$(\widetilde{Y}_1, Z_1)$	$(\widetilde{Y}_1, Z_2)$	$(\widetilde{Y}_2, Z_1)$	$(\widetilde{Y}_2, Z_2)$
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
(Real value)	0 (0)	0 (0)	−0.771 (0.011)	0.249 (0.044)	0.593 (0.029)	−0.738 (0.016)	−0.300 (0.031)	0.437 (0.044)	0.691 (0.038)	0.240 (0.145)
M & S	27 (1)	2345 (82)	−0.332 (0.055)	0.126 (0.054)	0.221 (0.058)	−0.101 (0.062)	−0.112 (0.060)	0.189 (0.061)	0.057 (0.058)	0.098 (0.099)
Proposed	19 (1)	1657 (84)	−0.662 (0.046)	0.225 (0.059)	0.539 (0.068)	−0.581 (0.049)	−0.268 (0.063)	0.286 (0.076)	0.383 (0.067)	0.227 (0.140)

M & S: Moriarity and Scheuren's (2001) method
Proposed: the proposed method

involves housing conditions of 506 observations, with 14 variables. We used 6 of these variables: DIS (weighted distances to five Boston employment centers), LSTAT (percentage of the population that is lower status), NOX (nitric oxides concentration), MEDV(median value of owner-occupied homes in $1000s), RM (average number of rooms per dwelling), and CRIM (per capita crime rate by town). We randomly separated the data file into two equally sized files, $n_A = n_B = 253$ to make the donor and recipient files as in Figure 4.1. Then, we performed statistical matching by using our proposed method and that of Moriarity and Scheuren. As in the previous simulation studies, this procedure was repeated 100 times independently.

Table 4.1 shows the means and the standard deviations for the evaluation factors, based on these 100 independent repetitions. Considering that the standard deviations of NOX (Y_1) and MEDV (Y_2) were 0.12 and 9.20 in the original complete data set, respectively, the means of Y_1^D and Y_2^D based on the proposed method were significantly lower than those based on Moriarity and Scheuren's method. Moreover, the means of $\mathrm{Corr}(X_i, \widetilde{Y}_j)$, $i, j = 1, 2$ and $\mathrm{Corr}(\widetilde{Y}_j, Z_k)$, $j, k = 1, 2$ based on the proposed method were much closer to the real values than those based on Moriarity and Scheuren's method.

5. Conclusion

In this research, we examined and proposed revisions to Moriarity and Scheuren's (2001) method, which is a representative method of statistical matching for continuous data. Their method improved on Kadane's (1978) and Rubin's (1986) methods by using the random residuals in the regression step and conducting a constrained match excepting the common variables in the matching step. However, since it assumed that $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ follows multivariate normal distribution, it may not be appropriate when the distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ has outliers or a heavy tail. This study's proposed method employs the LDE, which is an alternative to the LSE and robust to outliers, and takes into account the relationship among response variables, thus forming an alternative to the LSE. The

employment of empirical residuals as an alternative to the use of random residuals can also make our proposed method distribution-free. Through simulation studies using both simulated and real data, we found that the proposed method has distinct advantages over existing alternatives.

References

Bai, Z. D., Chen, X. R., Miao, B. Q. and Rao, C. R. (1990). Asymptotic theory of least distance estimate in multivariate linear model, *Statistics*, **21**, 503–519.

Bertsekas, D. P. (1991). *Linear Network Optimization: Algorithms and Codes*, Massachusetts: MIT Press, Cambridge.

Bertsekas, D. P. and Tseng, P. (1994). RELAX-IV: A faster version of the RELAX code for solving minimum cost flow problems. Available on the Internet at http://web.mit.edu/dimitrib/www/home.html.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air, *Journal of Environmental Economics and Management*, **5**, 81–102.

Jhun, M. and Choi, I. (2009). Bootstrapping least distance estimator in the multivariate regression model, *Computational Statistics & Data Analysis*, **53**, 4221–4227.

Kadane, J. B. (1978). Some statistical problems in merging data files. 1978 Compendium of Tax Research, U.S. Department of the Treasury, 159–171. (Reprinted in *Journal of Official Statistics*, **17**, 423–433).

Moriarity, C. and Scheuren, F. (2001). Statistical matching: a paradigm for assessing the uncertainty in the procedure, *Journal of Official Statistics*, **17**, 407–422.

Moriarity, C. and Scheuren, F. (2003). A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputaions, *Journal of Business & Economic Statistics*, **21**, 65–73.

Okner, B. (1972). Constructing a new data base from existing micro-data sets: The 1966 merge file, *Annals of Economic and Social Measurement*, **1**, 325–362.

Okner, B. (1974). Data matching and merging: An overview, *Annals of Economic and Social Measurement*, **3**, 347–352.

Paass, G. (1982). Statistical match with additional information. Internal Report IPES.82.0204, Gesellschaft fur Mathematik und Datenverarbeitung, Bonn, W. Ger.

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

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

Sims, C. A. (1972). Comments and rejoinder, *Annals of Economic and Social Measurement*, **1**, 343–345; 355–357.

Sims, C. A. (1974). Comment, *Annals of Economic and Social Measurement*, **3**, 395–397.