*Journal of the Korean Data & Information Science Society*
*2021,* **32**(3), 695–703

http://dx.doi.org/10.7465/jkdi.2021.32.3.695
한국데이터정보과학회지

# A study on the data fusion using the national health screening data

Sejin Bae[1] · Dal Ho Kim[2]

[12]Department of Statistics, Kyungpook National University

Received 30 April 2021, revised 21 May 2021, accepted 21 May 2021

**Abstract**

The exact combination of data collected from different objects is difficult. Data fusion is the statistical combination process of obtaining an integrated dataset using common variable. We consider three statistical techniques for data fusion: conditional mean matching using linear regression models, gamma regression using nonlinear regression models on two independent datasets, and a distance hot deck nonparametric approach based on the distance of each variable. The National Health Insurance Corporation's National Health Screening Data are used to compare the performance of three models.

*Keywords*: Data fusion, distance hot deck, gamma regression, linear regression model, statistical matching.

## 1. Introduction

Numerous studies conducted in various environments require extensive information. Planning a new survey or additional research is a major burden in terms of designing and planning research, both economically and in a timely manner. Under these constraints, data already collected or different data containing the same information can be utilized. This is because it is difficult to observe the information of every required variable simultaneously. In this situation, statistical matching (also called data fusion) would be an important and cautious approach. There is an increasing demand for statistical matching or data fusion through appropriate model strategies in various research fields and is a necessary methodology. Based on the data fusion problem described in D'Orazio *et al*. (2006), a summary of the underlying individual data structures for data integration is shown in Table 1.1

In sample sets A and B, missing regions refer to information about the unobserved variables of each dataset. By default, the dataset that will be expanded through data integration is called the recipient file, and the dataset that holds the information to be used is called the donor file. The variables common to both datasets are called common variables, where $X$ is the common variable. The other variables are the unique variables; therefore, $Y$ and

[1] Ph.D. candidate, Department of Statistics, Kyungpook National University, Daegu 41566, Korea.
[2] Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. E-mail: dalkim@knu.ac.kr

**Table 1.1** Structure of sample data sets in data fusion problems.

| Sample | Y | X | Z |
|---|---|---|---|
| A | $y_1^A$ <br> $\vdots$ <br> $y_{n_A}^A$ | $x_1^A$ <br> $\vdots$ <br> $x_{n_A}^A$ | missing |
| B | missing | $x_1^B$ <br> $\vdots$ <br> $x_{n_B}^B$ | $z_1^B$ <br> $\vdots$ <br> $z_{n_B}^B$ |

$Z$ correspond to them, respectively. Data fusion is the process of obtaining an integrated dataset by adding the only variable $Z$ in the donor file to the recipient file using each common variable. Each variable can play its role even if it is exchanged.

The approach of data fusion can be largely divided into macro and micro approaches. A macro approach is a method in which each element (source) is used to obtain a direct parameter estimate or major properties such as joint distribution function or correlations. A micro approach estimates the unique variables that do not exist in each dataset to create a completed synthetic data. The two methods are not exclusive to each other but are shaped in a micro approach using parametric information obtained from the macro method. Indeed, data fusion issues are mainly addressed in terms of analysis and application that follow micro approaches.

In this study, we confirm the data fusion (A+B) of the entire data by exchanging two situations without separating the role of the donor and recipient files. We consider three statistical techniques for data fusion: conditional mean matching using linear regression models, gamma regression using nonlinear regression models on two independent datasets, and a distance hot deck nonparametric approach based on the distance of each variable. The results of the three models are compared by analyzing the National Health Insurance Corporation's National Health Screening Data. Section 2 will examine the features of each model. Section 3 will use real data to compare the results of the data fusion. Finally, the conclusion and suggestions are presented in Section 4.

## 2. Methods for data fusion

### 2.1. Linear regression model - conditional mean matching

Various methods have been proposed to apply regression analysis in data fusion problems. A commonly applied approach is to estimate the regression model using the variables in the donor file. Using this, we obtain predictions after fitting to the recipient file and use them for data fusion. This approach was introduced and used in statistical matching by Kadane (1978) and Rubin (1986), was generalized and expanded by Singh *et al.* (1993), Moriarty and Scheuren (2001, 2003). It was organized by D'orazio *et al.* (2006). Alternatively, Little and Rubin (2002) and D'orazio *et al.* (2006) introduced a conditional mean matching method to obtain regression models from each of the donor and recipient files, and then insert their mean values. In this work, we apply data fusion to two separated datasets using conditional mean matching.

Let $(X, Y, Z)$ be the trivariate normal distribution. The joint distribution of $(X, Y, Z)$ is

$$f(x,y,z|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^3|\boldsymbol{\Sigma}|}} e^{-1/2\left((x,y,z)-\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left((x,y,z)'-\boldsymbol{\mu}\right)}, \text{with } (x,y,z)\in\mathbb{R}^3,$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and the covariance matrix of $(x,y,z)$, respectively.

The conditional distribution of $Y$, given $X$, can be defined through the regression model. It is equivalent to the expression in (2.1).

$$Y = \mu_{Y|X} + \epsilon_{Y|X} = \alpha_Y + \beta_{YX}X + \epsilon_{Y|X}, \tag{2.1}$$

where $\alpha_Y = \mu_Y - \beta_{YX}\mu_X$, $\beta_{YX} = \dfrac{\sigma_{XY}}{\sigma_X^2}$. $\epsilon_{Y|X}$ is normally distributed with zero mean and variance $\sigma^2_{Y|X} = \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2 = \sigma_Y^2 - \beta_{YX}^2\sigma_X^2$.

The same holds for the conditional distribution of $Z$ given $X$.

$$Z = \mu_{Z|X} + \epsilon_{Z|X} = \alpha_Z + \beta_{ZX}X + \epsilon_{Z|X}, \tag{2.2}$$

where $\alpha_Z = \mu_Z - \beta_{ZX}\mu_X$, $\beta_{ZX} = \dfrac{\sigma_{XZ}}{\sigma_X^2}$ , and $\epsilon_{Z|X}$ is also normally distributed with zero mean and variance $\sigma^2_{Z|X} = \sigma_Z^2 - \sigma_{XZ}^2/\sigma_X^2 = \sigma_Z^2 - \beta_{ZX}^2\sigma_X^2$.

$\bar{Z}_i^A$ is the $i$th missing value of the $Z$ variable in dataset A.

$$\bar{Z}_i^A = \hat{\alpha}_Z + \hat{\beta}_{ZX}X_i^A, i = 1,\cdots,n_A,$$

where $\hat{\alpha}_Z$ and $\hat{\beta}_{ZX}$ are the maximum likelihood estimates of $\alpha_Z$ and $\beta_{ZX}$, respectively, obtained from the regression model (2.2) estimated from dataset B.

Similarly, $\bar{Y}_j^B$ is the $j$th missing value of the $Y$ variable in dataset in B.

$$\bar{Y}_j^b = \hat{\alpha}_Y + \hat{\beta}_{YX}X_j^B, j = 1,\cdots,n_B,$$

where $\hat{\alpha}_Y$ and $\hat{\beta}_{YX}$ are the maximum likelihood estimates of $\alpha_Y$ and $\beta_{YX}$, respectively, obtained from the regression model (2.1) estimated from dataset A. The regression model, estimated as a given variable from datasets A and B, respectively, is obtained using common variables from different datasets.

### 2.2. Nonlinear regression model - gamma regression

Although the normal distribution is widely applied in many fields, there are situations where the distribution of the data is not symmetrical or does not show such a tendency. The fitting of a linear regression model can produce inappropriate results if the data of the variable to be estimated has a nonnegative positive value, that is, a value always greater than zero, or the data is skewed.

The linear regression model must satisfy several standard assumptions, one of which is that the error follows a normal distribution. However, these assumptions cannot be satisfied in the specific cases mentioned earlier. In such cases, nonlinear regression models that do not satisfy regular assumptions can be considered. The data to be applied in the next section

have these characteristics; the gamma distribution can be considered. We apply gamma regression, a nonlinear regression model, from individual datasets.

For the $Y$ in dataset A, $\mu_{Y|X}$ and $\mu^2_{Y|X}/\gamma_A$ are the conditional mean and variance of $y_i$ with shape parameter $\gamma_A$ where $\gamma_A > 0$ and $0 < y_i < \infty$.

In this study, we use the natural logarithm link function and can get

$$g(\mu_{Y|X}) = log(\mu_{Y|X}) = \beta_0 + \beta_1 x_i, i = 1, \cdots, n_A.$$

Additionally, for the $Z$ in dataset B, $\mu_{Z|X}$ and $\mu^2_{Z|X}/\gamma_B$ are the conditional mean and variance of $z_j$ with shpae parameter $\gamma_B$ where $\gamma_B > 0$ and $0 < z_j < \infty$.

Using the natural logarithm link function we can get

$$g(\mu_{Z|X}) = log(\mu_{Z|X}) = \beta_0 + \beta_1 x_j, j = 1, \cdots, n_B.$$

From the above two types of link function expressions, we obtain the value to be used for matching and apply them for data fusion.

### 2.3. Nonparametric micro approach - distance hot deck

The following three forms of hot deck methods, introduced in Singh *et al*. (1993) as nonparametric approaches to statistical matching, are mainly used. (i) random hot deck, (ii) rank hot deck, and (iii) distance hot deck. These are known as nonparametric methods of the micro approach. The first method refers to constructing a matching randomly and the second refers to using it when the data are in order. This study conducts matching for comparison, using the distance between each data. This method is widely used in statistical matching and it can also be found in Okner (1972), Ruggles (1974), and Rodgers (1984), its expanded use has been identified in Nielsen (2001). For example, the basic idea is to obtain the distance between two datasets based on the common variable for continuous common variable $X$, and then select the object with the closest distance to match the corresponding values.

$$d_{ab^*} = \left|x^A_a - x^B_{b^*}\right| = \min_{1 \le b \le n_B} \left|x^A_a - x^B_b\right|. \tag{2.3}$$

If there are multiple objects of equal distance, one can be selected arbitrarily, or the average of the objects of equal distance can be obtained and matched. It is called a constrained or unconstrained hot deck, depending on whether each object is selected to conduct replacement. A constrained hot deck does not allow duplication and can be used for matching only once. An unconstrained hot deck allows multiple matching through replacement. Here, we used an unconstrained distance hot deck, and we try to match using the average values of the selected objects based on the same distance.

## 3. Data analysis

The National Health Screening Data from the National Health Insurance Corporation (NHIS-2021-1-213) are used in the analysis. The National Health Insurance Corporation, which provides all national medical insurance coverage services, conducts national health

checkups for employees and local subscribers. All citizens can receive basic medical checkups once every two years. Some of the body measurement items are used in the analysis. Approximately 30 million people where examined by the National Health Screening Service in 2017 and 2018. Of those, 1,000 people male examinees, aged 30 to 39, who are prone to exposure to chronic diseases, were selected by random sampling for the analysis. The following continuous variables are available in the Health Screening Data: body measurement information related to obesity, such as height, weight, waist inches, and body mass index (BMI), contraction and relaxation blood pressure levels for high blood pressure measurements, blood sugar level to confirm diabetes, and total cholesterol, high density lipoprotein (HDL), low density lipoprotein (LDL) cholesterol, and triglyceride levels for hyperlipidemia. To examine the significant relationship between obesity and hyperlipidemia, we use three variables: waist inches, body mass index, and total cholesterol. To deal with the statistical matching problem, we randomly divide the data by a ratio of 6:4, as proposed in Yoshizoe (1999). We deliberately generate missing values for each variable and then determine the model efficiency by comparing it with the actual values after fitting them to the model. The common variable $X$ is the waist inches, and the unique variables in the individual dataset are the body mass index $Y$ and the total cholesterol $Z$. The first dataset A is missing variable $Z$, and the second dataset B is missing variable $Y$. The data to be used in the analysis present the distribution shown in the figure below. For the body mass index, a one-sided skewed is identified.

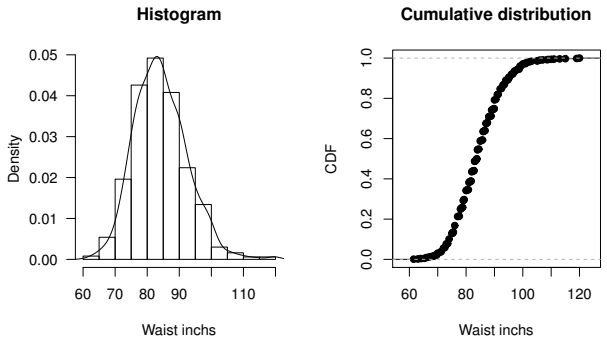


**Figure 3.1** Histogram and cumulative distribution of waist inches

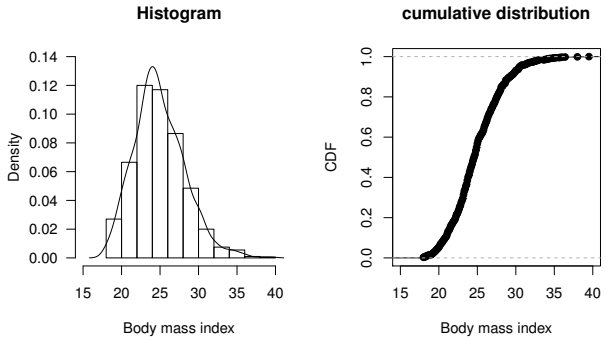


**Figure 3.2** Histogram and cumulative distribution of body mass index

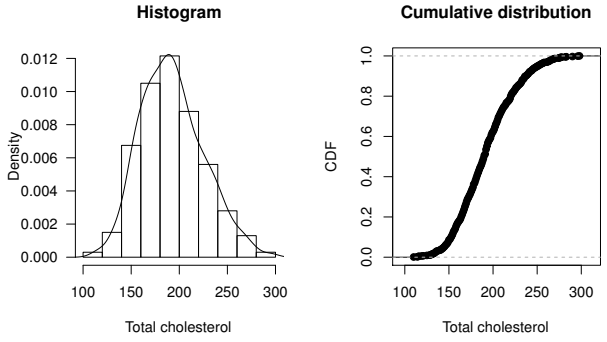


**Figure 3.3** Histogram and cumulative distribution of total cholesterol

From each dataset, we compare the results of the data fusion applying linear regression models, nonlinear regression models, and nonparametric methods. The comparison of each estimate is analyzed using four measurements mentioned below (3.1).

$$\begin{aligned}
&\text{Average relative bias (ARB)} && \frac{1}{m}\sum_i^n \frac{|c_i - e_i|}{c_i},\\
&\text{Average squared relative bias (ASRB)} && \frac{1}{m}\sum_i^n \frac{(c_i - e_i)^2}{c_i^2},\\
&\text{Average absolute bias (AAB)} && \frac{1}{m}\sum_i^n |c_i - e_i|,\\
&\text{Average squared deviation (ASD)} && \frac{1}{m}\sum_i^n (c_i - e_i)^2,
\end{aligned} \tag{3.1}$$

where $c_i$ is the $i$-th actual measurements and $e_i$ is estimation by the model. Smaller values indicate better performance. The results of checking the accuracy of estimates obtained by the three methods can be found in the following four measures in Table 3.1. Furthermore,

the plots of the real and predicted values are presented in Figure 3.4-3.6.

**Table 3.1** Results of fitting estimates by estimation method

| Method | Estimate | ARB | ASRB | AAB | ASD |
|---|---|---|---|---|---|
| Linear regression model | Body mass index | 0.05516 | 0.00521 | 1.36343 | 3.16612 |
| | Total cholesterol | 0.14879 | 0.03779 | 26.7677 | 1105.44 |
| Gamma regression model | Body mass index | 0.05476 | 0.00516 | 1.35493 | 3.13827 |
| | Total cholesterol | 0.14905 | 0.03803 | 26.8107 | 1111.88 |
| Distance hotdeck | Body mass index | 0.06104 | 0.00628 | 1.51137 | 3.83240 |
| | Total cholesterol | 0.15139 | 0.03808 | 27.7127 | 1209.36 |

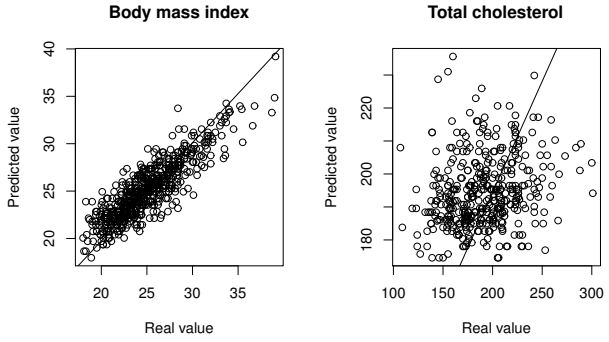


**Figure 3.4** The plot of real and predicted values using a linear regression model

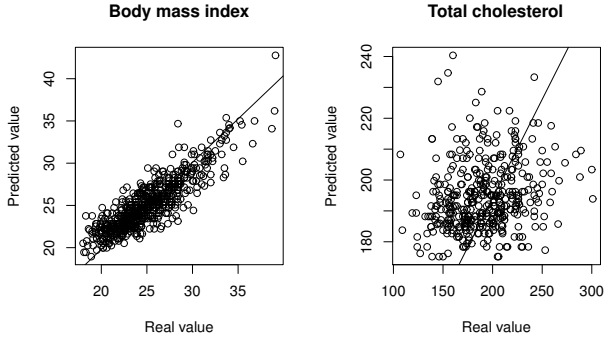


**Figure 3.5** The plot of real and predicted values using a gamma regression model

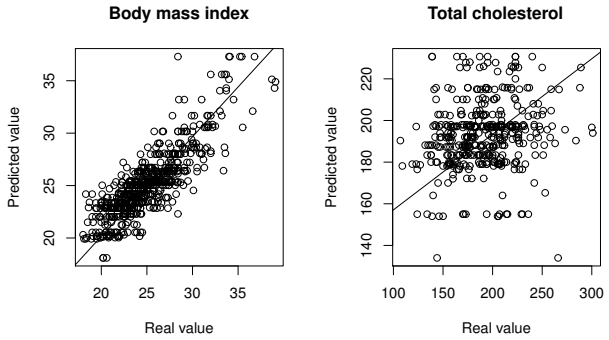


**Figure 3.6** The plot of real and predicted values using a distance hot deck

We find that, for the variables $Y$ that have characteristics of the gamma distribution, using the gamma regression model for estimation is better than using the linear regression model. The same results are derived when checked using the four comparison measures. In this data, we find that the regression or gamma regression results based on the data structure perform better than nonparametric methods.

## 4. Concluding remarks

In this study, we applied three methods of data fusion by analyzing the National Health Screening Data from the National Health Insurance Corporation of Korea. While each method has its own advantages, the accuracy of estimates may be reduced or increased by the existing characteristics of the data. In general situations, results using regression methods are better than nonparametric methods. In this work, if the distribution of the collected data is close to a normal distribution, the data fusion results using linear regression models are better than those obtained using nonparametric methods that rely on predictions using gamma regression models or distances between data. However, this situation was not maintained when the data distribution was close to skewed distribution rather than normal distribution. therefore in a skewed case, we find that data fusion using gamma regression models performs better than linear regression models. Further, nonparametric methods in both situations are not suitable for the form of these data. Data preprocessing with medical or clinical knowledge will reduce the bias of the data itself. Furthermore, the study can extend the model by considering additional common variables. Considering the distribution patterns or structures of the data to be merged in advance will help data fusion.

## References

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical matching: theory and practice*, Wiley, Chichester.

Kadane, J. B. (1978). Some statistical problems in merging data files. In Department of Treasury, *Compendium of tax research*, US Government Printing Office, Washington DC.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, 2nd Ed., Wiley, New York.

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 imputation. *Journal of Business and Economic Statistics*, **21**, 65-73.

Nielsen, S. F. (2001). Nonparametric conditional mean imputation. *Journal of Statistical Planning and Inference*, **99**, 129-150.

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

Rodgers, W. L. (1984). An evaluation of statistical matching. *Journal of Business and 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.

Ruggles, N. and Ruggles, R. (1974). A strategy for merging and matching microdata sets. *Annals of Economic and Social Measurement*, **1**, 353-371.

Singh, A.C., 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.

Yoshizoe, Y. and Araki, M. (1999). Use of statistical matching for household surveys in Japan. *In 52nd Session of the International Statistical Institute*, Helsinki, Finland.