Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Statistical matching using fractional imputation

by Jae Kwang Kim, Emily Berg and Taesung Park

Release date: June 22, 2016

Statistics Canada   Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

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

Note of appreciation

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

Standard table symbols

The following symbols are used in Statistics Canada publications:

- . not available for any reference period
- .. not available for a specific reference period
- ... not applicable
- 0 true zero or a value rounded to zero
- 0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
- ^{p} preliminary
- ^{r} revised
- x suppressed to meet the confidentiality requirements of the *Statistics Act*
- ^{E} use with caution
- F too unreliable to be published
- * significantly different from reference category ($p < 0.05$)

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 19-40
Statistics Canada, Catalogue No. 12-001-X

Statistical matching using fractional imputation

Jae Kwang Kim, Emily Berg and Taesung Park[1]

Abstract

Statistical matching is a technique for integrating two or more data sets when information available for matching records for individual participants across data sets is incomplete. Statistical matching can be viewed as a missing data problem where a researcher wants to perform a joint analysis of variables that are never jointly observed. A conditional independence assumption is often used to create imputed data for statistical matching. We consider a general approach to statistical matching using parametric fractional imputation of Kim (2011) to create imputed data under the assumption that the specified model is fully identified. The proposed method does not have a convergent expectation-maximisation (EM) sequence if the model is not identified. We also present variance estimators appropriate for the imputation procedure. We explain how the method applies directly to the analysis of data from split questionnaire designs and measurement error models.

Key Words: Data combination; Data fusion; Hot deck imputation; Split questionnaire design; Measurement error model.

1 Introduction

Survey sampling is a scientific tool for making inference about the target population. However, we often do not collect all the necessary information in a single survey, due to time and cost constraints. In this case, we wish to exploit, as much as possible, information already available from different data sources from the same target population. Statistical matching, sometimes called data fusion (Baker, Harris and O'Brien 1989) or data combination (Ridder and Moffit 2007), aims to integrate two or more data sets when information available for matching records for individual participants across data sets is incomplete. D'Orazio, Zio and Scanu (2006) and Leulescu and Agafitei (2013) provide comprehensive overviews of the statistical matching techniques in survey sampling.

Statistical matching can be viewed as a missing data problem where a researcher wants to perform a joint analysis of variables that are never jointly observed. Moriarity and Scheuren (2001) provide a theoretical framework for statistical matching under a multivariate normality assumption. Rässler (2002) develops multiple imputation techniques for statistical matching with pre-specified parameter values for non-identifiable parameters. Lahiri and Larsen (2005) address regression analysis with linked data. Ridder and Moffit (2007) provide a rigorous treatment of the assumptions and approaches for statistical matching in the context of econometrics.

Statistical matching aims to construct fully augmented data files to perform statistically valid joint analyses. To simplify the setup, suppose that two surveys, Survey A and Survey B, contain partial information about the population. Suppose that we observe x and y_1 from the Survey A sample and observe x and y_2 from the Survey B sample. Table 1.1 illustrates a simple data structure for matching. If the Survey B sample (Sample B) is a subset of the Survey A sample (Sample A), then we can apply record linkage techniques (Herzog, Scheuren and Winkler 2007) to obtain values of y_1 for the survey B sample. However, in many cases, such perfect matching is not possible (for instance, because the samples may contain

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. E-mail: jkim@iastate.edu; Emily Berg, Department of Statistics, Iowa State University, Ames, Iowa, U.S.A. E-mail: emilyb@iastate.edu; Taesung Park, Department of Statistics, Seoul National University, Seoul, Korea. E-mail: taesungp@gmail.com.

non-overlapping subsets), and we may rely on a probabilistic way of identifying the “statistical twins” from the other sample. That is, we want to create y_1 for each element in sample B by finding the nearest neighbor from Sample A. Nearest neighbor imputation has been discussed by many authors, including Chen and Shao (2001) and Beaumont and Bocci (2009), in the context of missing survey items.

Table 1.1
A simple data structure for matching

	X	Y_1	Y_2
Sample A	o	o	
Sample B	o		o

Finding the nearest neighbor is often based on “how close” they are in terms of x’s only. Thus, in many cases, statistical matching is based on the assumption that y_1 and y_2 are independent, conditional on x. That is,

$$y_1 \perp y_2 \mid x. \tag{1.1}$$

Assumption (1.1) is often referred to as the conditional independence (CI) assumption and is heavily used in practice.

In this paper, we consider an alternative approach that does not rely on the CI assumption. After we discuss the assumptions in Section 2, we present the proposed methods in Section 3. Furthermore, we consider two extensions, one to split questionnaire designs (in Section 4) and the other to measurement error models (in Section 5). Results from two simulation studies are presented in Section 6. Section 7 concludes the paper.

2 Basic setup

For simplicity of the presentation, we consider the setup of two independent surveys from the same target population consisting of N elements. As discussed in Section 1, suppose that Sample A collects information only on x and y_1 and Sample B collects information only on x and y_2.

To illustrate the idea, suppose for now that (x, y_1, y_2) are generated from a normal distribution such that

$$\begin{pmatrix} x \\ y_1 \\ y_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{1x} & \sigma_{2x} \\ & \sigma_{11} & \sigma_{12} \\ & & \sigma_{22} \end{pmatrix}\right].$$

Clearly, under the data structure in Table 1.1, the parameter σ_{12} is not estimable from the samples. The conditional independence assumption in (1.1) implies that $\sigma_{12} = \sigma_{1x}\sigma_{2x} / \sigma_{xx}$ and $\rho_{12} = \rho_{1x}\rho_{2x}$. That is, σ_{12} is completely determined from other parameters, rather than estimated directly from the realized samples.

Synthetic data imputation under the conditional independence assumption in this case can be implemented in two steps:

[Step 1] Estimate $f(y_1 \mid x)$ from Sample A, and denote the estimate by $\hat{f}_a(y_1 \mid x)$.

[Step 2] For each element i in Sample B, use the x_i value to generate imputed value(s) of y_1 from $\hat{f}_a(y_1 \mid x_i)$.

Since y_1 values are never observed in Sample B, synthetic values of y_1 are created for all elements in Sample B, leading to synthetic imputation. Haziza (2009) provides a nice review of literature on imputation methodology. Kim and Rao (2012) present a model-assisted approach to synthetic imputation when only x is available in Sample B. Such synthetic imputation completely ignores the observed information in y_2 from Sample B.

Statistical matching based on conditional independence assumes that $\mathrm{Cov}(y_1, y_2 \mid x) = 0$. Thus, the regression of y_2 on x and y_1 using the imputed data from the above synthetic imputation will estimate a zero regression coefficient for y_1. That is, the estimate $\hat{\beta}_2$ for

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y_1,$$

will estimate zero. Such analyses can be misleading if CI does not hold. To explain why, we consider an omitted variable regression problem:

$$\begin{aligned} y_1 &= \beta_0^{(1)} + \beta_1^{(1)} x + \beta_2^{(1)} z + e_1 \\ y_2 &= \beta_0^{(2)} + \beta_1^{(2)} x + \beta_2^{(2)} z + e_2 \end{aligned}$$

where z, e_1, e_2 are independent and are not observed. Unless $\beta_2^{(1)} = \beta_2^{(2)} = 0$, the latent variable z is an unobservable confounding factor that explains why $\mathrm{Cov}(y_1, y_2 \mid x) \neq 0$. Thus, the coefficient on y_1 in the population regression of y_2 on x and y_1 is not zero.

Note that the CI assumption is an assumption for model identification. Another identifying assumption is the instrumental variable (IV) assumption, as described in the following remark.

Remark 2.1 *We present a formal description of the IV assumption. First, assume that we can decompose* x *as* $x = (x_1, x_2)$ *such that*

(*i*) $f(y_2 \mid x_1, x_2, y_1) = f(y_2 \mid x_2, y_1)$

(*ii*) $f(y_1 \mid x_2, x_1 = a) \neq f(y_1 \mid x_2, x_1 = b)$

for some $a \neq b$. *Thus,* x_1 *is conditionally independent of* y_2 *given* x_2 *and* y_1 *but* x_1 *is correlated with* y_1 *given* x_2. *Note that* x_2 *may be null or have a degenerate distribution, such as an intercept. The variable* x_1 *satisfying the above two conditions is often called an instrumental variable (IV) for* y_1. *The directed acyclic graph in Figure 2.1 illustrates the dependence structure of a model with an instrumental variable. Ridder and Moffit (2007) used "exclusion restrictions" to describe the instrumental variable assumption. One example where the instrumental variable assumption is reasonable is repeated surveys. In the repeated survey, suppose that* y_t *is the study variable at year* t *and satisfies Markov property*

$$P(y_{t+1} \mid y_1, \ldots, y_t) = P(y_{t+1} \mid y_t),$$

where $P(y_t)$ *denotes a cumulative distribution function. In this case,* y_{t-1} *is an instrumental variable for* y_t*. In fact, any last observation of* $y_s\ (s \le t)$ *is the instrumental variable for* y_t*.*

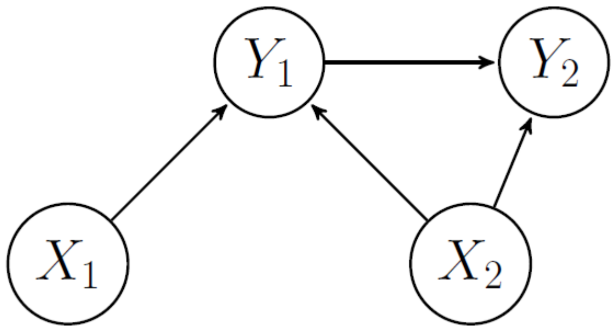

Figure 2.1 Graphical illustration of the dependence structure for a model in which x_1 is an instrumental variable for y_1 and x_2 is an additional covariate in the models for y_2 and y_1.

Under the instrumental variable assumption, one can use two-step regression to estimate the regression parameters of a linear model. The following example presents the basic ideas.

Example 2.1 *Consider the two sample data structure in Table 1.1. We assume the following linear regression model:*

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + \beta_2 x_{2i} + e_i, \tag{2.1}$$

where $e_i \sim \left(0, \sigma_e^2\right)$ *and* e_i *is independent of* $\left(x_{1j}, x_{2j}, y_{1j}\right)$ *for all* i, j*. In this case, a consistent estimator of* $\beta = (\beta_0, \beta_1, \beta_2)'$ *can be obtained by the two-stage least squares (2SLS) method as follows:*

1. *From Sample* A*, fit the following "working model" for* y_1

   $$y_{1i} = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i, \quad u_i \sim \left(0, \sigma_u^2\right) \tag{2.2}$$

   to obtain a consistent estimator of $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$ *defined by*

   $$\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)' = (X'X)^{-1} X'Y_1$$

   where $X = [X_0, X_1, X_2]$ *is a matrix whose* i^{th} *row is* $(1, x_{1i}, x_{2i})$ *and* Y_1 *is a vector with* y_{1i} *being the* i^{th} *component.*
2. *A consistent estimator of* $\beta = (\beta_0, \beta_1, \beta_2)'$ *is obtained by the least squares method for the regression of* y_{2i} *on* $(1, \hat{y}_{1i}, x_{2i})$ *where* $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$*.*

Asymptotic unbiasedness of the 2SLS estimator under the instrumental variable assumption is discussed in Appendix A. The 2SLS method is not directly applicable if the regression model (2.1) is nonlinear. Also, while the 2SLS method gives estimates of the regression parameters, 2SLS does not provide consistent estimators for more general parameters such as $\theta = \Pr(y_2 < 1 \mid y_1 < 3)$. Stochastic imputation can provide a solution for estimating a more general class of parameters. We explain how to modify parametric fractional imputation of Kim (2011) to address general purpose estimation in statistical matching problems.

3 Fractional imputation

We now describe the fractional imputation methods for statistical matching without using the CI assumption. The use of fractional imputation for statistical matching was originally presented in Chapter 9 of Kim and Shao (2013) under the IV assumption. In this paper, we present the methodology without requiring the IV assumption. We only assume that the specified model is fully identified. The identifiability of the specified model can be easily checked in the computation of the proposed procedure.

To explain the idea, note that y_1 is missing in Sample B and our goal is to generate y_1 from the conditional distribution of y_1 given the observations. That is, we wish to generate y_1 from

$$f(y_1 \mid x, y_2) \propto f(y_2 \mid x, y_1)\, f(y_1 \mid x). \tag{3.1}$$

To generate y_1 from (3.1), we can consider the following two-step imputation:

1. Generate y_1^* from $\hat{f}_a(y_1 \mid x)$.
2. Accept y_1^* if $f\left(y_2 \mid x, y_1^*\right)$ is sufficiently large.

Note that the first step is the usual method under the CI assumption. The second step incorporates the information in y_2. The determination of whether $f\left(y_2 \mid x, y_1^*\right)$ is sufficiently large required for Step 2 is often made by applying a Markov Chain Monte Carlo (MCMC) method such as the Metropolis-Hastings algorithm (Chib and Greenberg 1995). That is, let $y_1^{(t-1)}$ be the current value of y_1 in the Markov Chain. Then, we accept y_1^* with probability

$$R\left(y_1^*, y_1^{(t-1)}\right) = \min\left\{1, \frac{f\left(y_2 \mid x, y_1^*\right)}{f\left(y_2 \mid x, y_1^{(t-1)}\right)}\right\}.$$

Such algorithms can be computationally cumbersome because of slow convergence of the MCMC algorithm.

Parametric fractional imputation of Kim (2011) enables generating imputed values in (3.1) without requiring MCMC. The following EM algorithm by fractional imputation can be used:

1. For each $i \in B$, generate m imputed values of y_{1i}, denoted by $y_{1i}^{*(1)}, \ldots, y_{1i}^{*(m)}$, from $\hat{f}_a(y_1 \mid x_i)$, where $\hat{f}_a(y_1 \mid x)$ denotes the estimated density for the conditional distribution of y_1 given x obtained from Sample A.

2. Let $\hat{\theta}_t$ be the current parameter value of θ in $f(y_2 \mid x, y_1)$. For the j^{th} imputed value $y_{1i}^{*(j)}$, assign the fractional weight

$$w_{ij(t)}^{*} \propto f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right)$$

such that $\sum_{j=1}^{m} w_{ij}^{*} = 1$.

3. Solve the fractionally imputed score equation for θ

$$\sum_{i \in B} w_{ib} \sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0 \tag{3.2}$$

to obtain $\hat{\theta}_{t+1}$, where $S(\theta; x, y_1, y_2) = \partial \log f(y_2 \mid x, y_1; \theta) / \partial \theta$, and w_{ib} is the sampling weight of unit i in Sample B.

4. Go to Step 2 and continue until convergence.

When the model is identified, the EM sequence obtained from the above PFI method will converge. If the specified model is not identifiable then there is no unique solution to maximizing the observed likelihood and the above EM sequence does not converge. In (3.2), note that, for sufficiently large m,

$$\begin{aligned} \sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) &\cong \frac{\int S(\theta; x_i, y_1, y_{2i}) f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a(y_1 \mid x_i) \, dy_1}{\int f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a(y_1 \mid x_i) \, dy_1} \\ &= E\left\{S(\theta; x_i, Y_1, y_{2i}) \mid x_i, y_{2i}; \hat{\theta}_t\right\}. \end{aligned}$$

If y_{i1} is categorical, then the fractional weight can be constructed by the conditional probability corresponding to the realized imputed value (Ibrahim 1990). Step 2 is used to incorporate observed information of y_{i2} in Sample B. Note that Step 1 is not repeated for each iteration. Only Step 2 and Step 3 are iterated until convergence. Because Step 1 is not iterated, convergence is guaranteed and the observed likelihood increases, as long as the model is identifiable. See Theorem 2 of Kim (2011).

Remark 3.1 *In Section 2, we introduce IV only because this is what it is typically done in the literature to ensure identifiability. The proposed method itself does not rely on this assumption. To illustrate a situation where we can identify the model without introducing the IV assumption, suppose that the model is*

$$\begin{aligned} y_2 &= \beta_0 + \beta_1 x + \beta_2 y_1 + e_2 \\ y_1 &= \alpha_0 + \alpha_1 x + e_1 \end{aligned}$$

with $e_1 \sim N\left(0, x^2 \sigma_1^2\right)$ *and* $e_2 \mid e_1 \sim N\left(0, \sigma_2^2\right)$. *Then*

$$f(y_2 \mid x) = \int f(y_2 \mid x, y_1) \, f(y_1 \mid x) \, dy_1$$

is also a normal distribution with mean $(\beta_0 + \beta_2 \alpha_0) + (\beta_1 + \beta_2 \alpha_1) x$ *and variance* $\sigma_2^2 + \beta_2^2 \sigma_1^2 x^2$. *Under the data structure in Table 1.1, such a model is identified without assuming the IV assumption. The* *assumption of no interaction between* y_1 *and* x *in the model for* y_2 *is key to ensuring the model is identifiable.*

Instead of generating $y_{1i}^{*(j)}$ from $\hat{f}_a(y_1 \mid x_i)$, we can consider a hot-deck fractional imputation (HDFI) method, where all the observed values of y_{1i} in Sample A are used as imputed values. In this case, the fractional weights in Step 2 are given by

$$w_{ij}^*\left(\hat{\theta}_t\right) \propto w_{ij0}^* f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right),$$

where

$$w_{ij0}^* = \frac{\hat{f}_a\left(y_{1j} \mid x_i\right)}{\sum_{k \in A} w_{ka} \hat{f}_a\left(y_{1j} \mid x_k\right)}. \tag{3.3}$$

The initial fractional weight w_{ij0}^* in (3.3) is computed by applying importance weighting with

$$\hat{f}_a\left(y_{1j}\right) = \int \hat{f}_a\left(y_{1j} \mid x\right) \hat{f}_a(x)\, dx \propto \sum_{i \in A} w_{ia} \hat{f}_a\left(y_{1j} \mid x_i\right)$$

as the proposal density for y_{1j}. The M-step is the same as for parametric fractional imputation. See Kim and Yang (2014) for more details on HDFI. In practice, we may use a single imputed value for each unit. In this case, the fractional weights can be used as the selection probability in Probability-Proportional-to-Size (PPS) sampling of size $m = 1$.

For variance estimation, we can either use a linearization method or a resampling method. We first consider variance estimation for the maximum likelihood estimator (MLE) of θ. If we use a parametric model $f(y_1 \mid x) = f(y_1 \mid x; \theta_1)$ and $f(y_2 \mid x, y_1; \theta_2)$, the MLE of $\theta = (\theta_1, \theta_2)$ is obtained by solving

$$\left[S_1(\theta_1), \bar{S}_2(\theta_1, \theta_2)\right] = (0, 0), \tag{3.4}$$

where $S_1(\theta_1) = \sum_{i \in A} w_{ia} S_{i1}(\theta_1)$, $S_{i1}(\theta_1) = \partial \log f(y_{1i} \mid x_i; \theta_1) / \partial \theta_1$ is the score function of θ_1,

$$\bar{S}_2(\theta_1, \theta_2) = E\{S_2(\theta_2) \mid X, Y_2; \theta_1, \theta_2\},$$

$S_2(\theta_2) = \sum_{i \in B} w_{ib} S_{i2}(\theta_2)$, and $S_{i2}(\theta_2) = \partial \log f(y_{2i} \mid x_i, y_{1i}; \theta_2) / \partial \theta_2$ is the score function of θ_2. Note that we can write $\bar{S}_2(\theta_1, \theta_2) = \sum_{i \in B} w_{ib} E\{S_{i2}(\theta_2) \mid x_i, y_{2i}; \theta\}$. Thus,

$$\begin{aligned}
\frac{\partial}{\partial \theta_1'} \bar{S}_2(\theta) &= \sum_{i \in B} w_{ib} \frac{\partial}{\partial \theta_1'} \left[\frac{\int S_{i2}(\theta_2) f(y_1 \mid x_i; \theta_1) f(y_{2i} \mid x_i, y_1; \theta_2)\, dy_1}{\int f(y_1 \mid x_i; \theta_1) f(y_{2i} \mid x_i, y_1; \theta_2)\, dy_1}\right] \\
&= \sum_{i \in B} w_{ib} E\{S_{i2}(\theta_2) S_{i1}(\theta_1)' \mid x_i, y_{2i}; \theta\} \\
&\quad - \sum_{i \in B} w_{ib} E\{S_{i2}(\theta_2) \mid x_i, y_{2i}; \theta\} E\{S_{i1}(\theta_1)' \mid x_i, y_{2i}; \theta\}
\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial \theta_2'}\bar{S}_2(\theta) &= \sum_{i\in B} w_{ib}\frac{\partial}{\partial \theta_2'}\left[\frac{\int S_{i2}(\theta_2)\, f(y_1 \mid x_i;\theta_1)\, f(y_{2i}\mid x_i, y_1;\theta_2)\,dy_1}{\int f(y_1\mid x_i;\theta_1)\, f(y_{2i}\mid x_i, y_1;\theta_2)\,dy_1}\right]\\ &= \sum_{i\in B} w_{ib}E\left\{\frac{\partial}{\partial \theta_2'}S_{i2}(\theta_2)\mid x_i, y_{2i};\theta\right\}\\ &+\sum_{i\in B} w_{ib}E\{S_{i2}(\theta_2)\,S_{i2}(\theta_2)'\mid x_i, y_{2i};\theta\}\\ &-\sum_{i\in B} w_{ib}E\{S_{i2}(\theta_2)\mid x_i, y_{2i};\theta\}\,E\{S_{2i}(\theta_2)'\mid x_i, y_{2i};\theta\}.\end{aligned}$$

Now, $\partial \bar{S}_2(\theta)/\partial\theta_1'$ can be consistently estimated by

$$\hat{B}_{21} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* S_{2ij}^*\left(\hat{\theta}_2\right)\left\{S_{1ij}^*\left(\hat{\theta}_1\right) - \bar{S}_{1i}^*\left(\hat{\theta}_1\right)\right\}', \tag{3.5}$$

where $S_{1ij}^*\left(\hat{\theta}_1\right) = S_1\left(\hat{\theta}_1; x_i, y_{1i}^{*(j)}\right)$, $S_{2ij}^*\left(\hat{\theta}_2\right) = S_2\left(\hat{\theta}_2; x_i, y_{1i}^{*(j)}, y_{2i}\right)$, and $\bar{S}_{1i}^*\left(\hat{\theta}_1\right) = \sum_{j=1}^{m} w_{ij}^* S_1\left(\hat{\theta}_1; x_i, y_{1i}^{*(j)}\right)$. Also, $\partial \bar{S}_2(\theta)/\partial\theta_2'$ can be consistently estimated by

$$-\hat{I}_{22} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^*\dot{S}_{2ij}^*\left(\hat{\theta}_2\right) - \hat{B}_{22} \tag{3.6}$$

where

$$\hat{B}_{22} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* S_{2ij}^*\left(\hat{\theta}_2\right)\left\{S_{2ij}^*\left(\hat{\theta}_2\right) - \bar{S}_{2i}^*\left(\hat{\theta}_2\right)\right\}',$$

$\dot{S}_{2ij}^*(\theta_2) = \partial S_2\left(\theta_2; x_i, y_{1i}^{*(j)}, y_{2i}\right)/\partial\theta_2'$ and $\bar{S}_{2i}^*(\theta_2) = \sum_{j=1}^{m} w_{ij}^* S_{2ij}^*(\theta_2)$.

Using a Taylor expansion with respect to θ_1,

$$\begin{aligned}\bar{S}_2\left(\hat{\theta}_1, \theta_2\right) &\cong \bar{S}_2(\theta_1,\theta_2) - E\left\{\frac{\partial}{\partial\theta_1'}\bar{S}_2(\theta)\right\}\left[E\left\{\frac{\partial}{\partial\theta_1'}S_1(\theta_1)\right\}\right]^{-1} S_1(\theta_1)\\ &= \bar{S}_2(\theta) + KS_1(\theta_1),\end{aligned}$$

and we can write

$$V\left(\hat{\theta}_2\right) \doteq \left\{E\left(\frac{\partial}{\partial\theta_2'}\bar{S}_2\right)\right\}^{-1} V\left\{\bar{S}_2(\theta) + KS_1(\theta_1)\right\}\left\{E\left(\frac{\partial}{\partial\theta_2'}\bar{S}_2\right)\right\}^{-1'}.$$

Writing

$$\bar{S}_2(\theta) = \sum_{i\in B} w_{ib}\bar{s}_{2i}(\theta),$$

with $\bar{s}_{2i}(\theta) = E\{S_{i2}(\theta_2) \mid x_i, y_{2i}; \theta\}$, a consistent estimator of $V\{\bar{S}_2(\theta)\}$ can be obtained by applying a design-consistent variance estimator to $\sum_{i \in B} w_{ib} \hat{s}_{2i}$ with $\hat{s}_{2i} = \sum_{j=1}^{m} w_{ij}^* S_{2ij}^* (\hat{\theta}_2)$. Under simple random sampling for Sample B, we have

$$\hat{V}\{\bar{S}_2(\theta)\} = n_B^{-2} \sum_{i \in B} \hat{s}_{2i} \hat{s}_{2i}'.$$

Also, $V\{KS_1(\theta_1)\}$ is consistently estimated by

$$\hat{V}_2 = \hat{K} \hat{V}(S_1) \hat{K}',$$

where $\hat{K} = \hat{B}_{21} \hat{I}_{11}^{-1}$, $\hat{B}_{21}$ is defined in (3.5), and $\hat{I}_{11} = -\partial S_1(\theta_1) / \partial \theta_1'$ evaluated at $\theta_1 = \hat{\theta}_1$. Since the two terms $\bar{S}_2(\theta)$ and $S_1(\theta_1)$ are independent, the variance can be estimated by

$$\hat{V}(\hat{\theta}) \doteq \hat{I}_{22}^{-1} \left[\hat{V}\{\bar{S}_2(\theta)\} + \hat{V}_2 \right] \hat{I}_{22}^{-1'},$$

where $\hat{I}_{22}$ is defined in (3.6).

More generally, one may consider estimation of a parameter η defined as a root of the census estimating equation $\sum_{i=1}^{N} U(\eta; x_i, y_{1i}, y_{2i}) = 0$. Variance estimation of the FI estimator of η computed from $\sum_{i \in B} w_{ib} \sum_{j=1}^{m} w_{ij}^* U(\eta; x_i, y_{1i}^{*(j)}, y_{2i}) = 0$ is discussed in Appendix B.

4 Split questionnaire survey design

In Section 3, we consider the situation where Sample A and Sample B are two independent samples from the same target population. We now consider another situation of a split questionnaire design where the original sample S is selected from a target population and then Sample A and Sample B are randomly chosen such that $A \cup B = S$ and $A \cap B = \phi$. We observe (x, y_1) from Sample A and observe (x, y_2) from Sample B. We are interested in creating fully augmented data with observation (x, y_1, y_2) in S.

Such split questionnaire survey designs are gaining popularity because they reduce response burden (Raghunathan and Grizzle 1995; Chipperfield and Steel 2009). Split questionnaire designs have been investigated, for example, for the Consumer Expenditure survey (Gonzalez and Eltinge 2008) and the National Assessment of Educational Progress (NAEP) survey in the US. In applications of split-questionnaire designs, analysts may be interested in multiple parameters such as the mean of y_1 and the mean of y_2, in addition to the coefficient in the regression of y_2 on y_1.

We consider a design where the original Sample S is partitioned into two subsamples: A and B. We assume that x_i is observed for $i \in S$, y_{1i} is collected for $i \in A$ and y_{2i} is collected for $i \in B$. The probability of selection into A or B may depend on x_i but does not depend on y_{1i} or y_{2i}. As a consequence, the design used to select subsample A or B is non-informative for the specified model (Fuller 2009, Chapter 6). We let w_i denote the sampling weight associated with the full sample S. We assume a procedure is available for estimating the variance of an estimator of the form $\hat{Y} = \sum_{i \in S} w_i y_i$, and we denote the variance estimator by $\hat{V}_s\left(\sum_{i \in S} w_i y_i\right)$.

A procedure for obtaining a fully imputed data set is as follows. First, use the procedure of Section 3 to obtain imputed values $\{y_{1i}^{*(j)} : i \in B, j = 1, \ldots, m\}$ and an estimate, $\hat{\theta}$, of the parameter in the distribution $f(y_2 \mid y_1, x; \theta)$. The estimate $\hat{\theta}$ is obtained by solving

$$\sum_{i \in B} w_i \sum_{j=1}^{m} w_{ij}^{*} S_2\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0, \tag{4.1}$$

where $S_2(\theta; x, y_1, y_2) = \partial \log f(y_2 \mid y_1, x; \theta) / \partial \theta$. Given $\hat{\theta}$, generate imputed values $y_{2i}^{*(j)} \sim f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$, for $i \in A$ and $j = 1, \ldots, m$.

Under the assumption that the model is identified, the parameter estimator $\hat{\theta}$ generated by solving (4.1) is fully efficient in the sense that the imputed value of y_{2i} for Sample A leads to no efficiency gain. To see this, note that the score equation using the imputed value of y_{2i} is computed by

$$\sum_{i \in A} w_i m^{-1} \sum_{j=1}^{m} S_2\left(\theta; x_i, y_{1i}, y_{2i}^{*(j)}\right) + \sum_{i \in B} w_i \sum_{j=1}^{m} w_{ij}^{*} S_2\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0. \tag{4.2}$$

Because $y_{2i}^{*(1)}, \ldots, y_{2i}^{*(m)}$ are generated from $f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$,

$$p\lim_{m \to \infty} \sum_{i \in A} w_i m^{-1} \sum_{j=1}^{m} S_2\left(\theta; x_i, y_{1i}, y_{2i}^{*(j)}\right) = \sum_{i \in A} w_i E\left\{S_2(\theta; x_i, y_{1i}, Y_2) \mid y_{1i}, x_i; \hat{\theta}\right\}.$$

Thus, by the property of score function, the first term of (4.2) evaluated at $\theta = \hat{\theta}$ is close to zero and the solution to (4.2) is essentially the same as the solution to (4.1). That is, there is no efficiency gain in using the imputed value of y_{2i} in computing the MLE for θ in $f(y_2 \mid y_1, x; \theta)$.

However, the imputed values of y_{2i} can improve the efficiency of inferences for parameters in the joint distribution of (y_{1i}, y_{2i}). As a simple example, consider estimation of μ_2, the marginal mean of y_{2i}. Under simple random sampling, the imputed estimator of $\mu = E(Y_2)$ is

$$\hat{\mu}_{I,m} = \frac{1}{n}\left\{\sum_{i \in A}\left(m^{-1} \sum_{j=1}^{m} y_{2i}^{*(j)}\right) + \sum_{i \in B} y_{2i}\right\}, \tag{4.3}$$

where $y_{2i}^{*(1)}, \ldots, y_{2i}^{*(m)}$ are generated from $f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$. For sufficiently large m, we can write

$$\begin{aligned}
\hat{\mu}_{I,\infty} &= \frac{1}{n}\left\{\sum_{i \in A} \hat{y}_{2i} + \sum_{i \in B} y_{2i}\right\} \\
&= \frac{1}{n}\left\{\sum_{i \in A} E\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right) + \sum_{i \in B} y_{2i}\right\}.
\end{aligned}$$

Under the setup of Example 2.1, we can express $\hat{y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 y_{1i} + \hat{\beta}_2 x_{2i}$ where $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)$ satisfies

$$\sum_{i \in B}\left(y_{2i} - \hat{\beta}_0 - \hat{\beta}_{1i} \hat{y}_{1i} - \hat{\beta}_2 x_{2i}\right) = 0$$

and $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$ with $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ satisfying $\sum_{i \in A} (y_{1i} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i}) = 0.$ Thus, ignoring the smaller order terms, we have

$$V(\hat{\mu}_{I,\infty}) = \frac{1}{n} V(y_2) + \left(\frac{1}{n_b} - \frac{1}{n}\right) V(y_2 - \hat{y}_2)$$

which is smaller than the variance of the direct estimator $\hat{\mu}_b = n_b^{-1} \sum_{i \in B} y_{2i}$.

5 Measurement error models

We now consider an application of statistical matching to the problem of measurement error models. Suppose that we are interested in the parameter θ in the conditional distribution $f(y_2 \mid y_1; \theta)$. In the original sample, instead of observing (y_{1i}, y_{2i}), we observe (x_i, y_{2i}), where x_i is a contaminated version of y_{1i}. Because inference for θ based on (x_i, y_{2i}) may be biased, additional information is needed. One common way to obtain additional information is to collect (x_i, y_{1i}) in an external calibration study. In this case, we observe (x_i, y_{1i}) in Sample A and (x_i, y_{2i}) in Sample B, where Sample A is the calibration sample, and Sample B is the main sample. Guo and Little (2011) discuss an application of external calibration.

The external calibration framework can be expressed as a statistical matching problem. Table 5.1 makes the connection between statistical matching and external calibration explicit. An instrumental variable assumption permits inference for θ based on data with the structure of Table 1.1. In the notation of the measurement error model, the instrumental variable assumption is

$$f(y_{2i} \mid y_{1i}, x_i) = f(y_{2i} \mid y_{1i}) \quad \text{and} \quad f(y_{1i} \mid x_i = a) \neq f(y_{1i} \mid x_i = b), \tag{5.1}$$

for some $a \neq b$. The instrumental variable assumption may be judged reasonable in applications related to error in covariates because the subject-matter model of interest is $f(y_{2i} \mid y_{1i})$, and x_i is a contaminated version of y_{1i} that contains no additional information about y_{2i} given y_{1i}.

Table 5.1
Data structure for measurement error model

	x_i	y_{1i}	y_{2i}
Survey A (calibration study)	o	o	
Survey B (main study)	o		o

For fully parametric $f(y_{2i} \mid y_{1i})$ and $f(y_{1i} \mid x_i)$, one can use parametric fractional imputation to execute the EM algorithm. This method requires evaluating the conditional expectation of the complete-data score function given the observed values. To evaluate the conditional expectation using fractional imputation, we first express the conditional distribution of y_1 given (x, y_2) as,

$$f(y_1 \mid x, y_2) \propto f(y_1 \mid x) f(y_2 \mid y_1). \tag{5.2}$$

We let an estimator $\hat{f}_a(y_{1i} \mid x_i)$ of $f(y_{1i} \mid x_i)$ be available from the calibration sample (Sample A). Implementation of the EM algorithm via fractional imputation involves the following steps:

1. For each $i \in B,$ generate $y_{1i}^{*(j)}$ from $\hat{f}_a(y_1 \mid x_i),$ for $j = 1, \ldots, m.$
2. Compute the fractional weights
$$w_{ij(t)}^{*} \propto f\left(y_{2i} \mid y_{1i}^{*(j)}; \hat{\theta}_t\right)$$
with $\sum_{j=1}^{m} w_{ij(t)}^{*} = 1.$
3. Update θ by solving
$$\sum_{i \in B} w_{ib} \sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; y_{1i}^{*(j)}, y_{2i}\right) = 0,$$
where $S(\theta; y_1, y_2) = \partial \log f(y_2 \mid y_1; \theta) / \partial\theta.$
4. Go to Step 2 until convergence.

The method above requires generating data from $f(y_1 \mid x).$ For some nonlinear models or models with non-constant variances, simulating from the conditional distribution of y_1 given x may require Monte Carlo methods such as accept-reject or Metropolis Hastings. The simulation of Section 6.2 exemplifies a simulation in which the conditional distribution of $y_1 \mid x$ has no closed form expression. In this case, we may consider an alternative approach, which may be computationally simpler. To describe this approach, let $h(y_1 \mid x)$ be the "working" conditional distribution, such as the normal distribution, from which samples are easily generated. We assume that estimates $\hat{f}_a(y_1 \mid x)$ and $\hat{h}_a(y_1 \mid x)$ of $f(y_1 \mid x)$ and $h(y_1 \mid x),$ respectively, are available from Sample A. Implementation of the EM algorithm via fractional imputation then involves the following steps:

1. For each $i \in B,$ generate $x_i^{*(j)}$ from $\hat{h}_a(y_1 \mid x_i),$ for $j = 1, \ldots, m.$
2. Compute the fractional weights
$$w_{ij(t)}^{*} \propto f\left(y_{2i} \mid y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a\left(y_{1i}^{*(j)} \mid x_i\right) \Big/ \hat{h}_a\left(y_{1i}^{*(j)} \mid x_i\right) \tag{5.3}$$
with $\sum_{j=1}^{m} w_{ij(t)}^{*} = 1.$
3. Update θ by solving
$$\sum_{i \in B} w_{ib} \sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; y_{1i}^{*(j)}, y_{2i}\right) = 0.$$
4. Go to Step 2 until convergence.

Variance estimation is a straightforward application of the linearization method in Section 3. The hot-deck fractional imputation method described in Section 3 with fractional weights defined in (3.3) also extends readily to the measurement error setting.

6 Simulation study

To test our theory, we present two limited simulation studies. The first simulation study considers the setup of combining two independent surveys of partial observation to obtain joint analysis. The second simulation study considers the setup of measurement error models with external calibration.

6.1 Simulation one

To compare the proposed methods with the existing methods, we generate 5,000 Monte Carlo samples of (x_i, y_{1i}, y_{2i}) with size $n = 400$, where

$$\begin{pmatrix} y_{1i} \\ x_i \end{pmatrix} \sim N\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}\right),$$

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + e_i, \tag{6.1}$$

$e_i \sim N(0, \sigma^2)$, and $\beta = (\beta_0, \beta_1, \sigma^2)' = (1,1,1)'$. Note that, in this setup, we have $f(y_2 \mid x, y_1) = f(y_2 \mid y_1)$ and so the variable x plays the role of the instrumental variable for y_1.

Instead of observing (x_i, y_{1i}, y_{2i}) jointly, we assume that only (y_1, x) are observed in Sample A and only (y_2, x) are observed in Sample B, where Sample A is obtained by taking the first $n_a = 400$ elements and Sample B is obtained by taking the remaining $n_b = 400$ elements from the original sample. We are interested in estimating four parameters: three regression parameters $\beta_0, \beta_1, \sigma^2$ and $\pi = P(y_1 < 2, y_2 < 3)$, the proportion of $y_1 < 2$ and $y_2 < 3$. Four methods are considered in estimating the parameters:

1. Full sample estimation (Full): Uses the complete observation of (y_{1i}, y_{2i}) in Sample B.
2. Stochastic regression imputation (SRI): Use the regression of y_1 on x from Sample A to obtain $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\sigma}_1^2)$, where the regression model is $y_1 = \alpha_0 + \alpha_1 x + e_1$ with $e_1 \sim (0, \sigma_1^2)$. For each $i \in B$, $m = 10$ imputed values are generated by $y_{1i}^{*(j)} = \hat{\alpha}_0 + \hat{\alpha}_1 x_i + e_i^{*(j)}$ where $e_i^{*(j)} \sim N(0, \hat{\sigma}_1^2)$.
3. Parametric fractional imputation (PFI) with $m = 10$ using the instrumental variable assumption.
4. Hot-deck fractional imputation (HDFI) with $m = 10$ using the instrumental variable assumption.

Table 6.1 presents Monte Carlo means and Monte Carlo variances of the point estimators of the four parameters of interest. SRI shows large biases for all parameters considered because it is based on the conditional independence assumption. Both PFI and HDFI provide nearly unbiased estimators for all parameters. Estimators from HDFI are slightly more efficient than those from PFI because the two-step procedure in HDFI uses the full set of respondents in the first step. The theoretical asymptotic variance of $\hat{\beta}_1$ computed from PFI is

$$V(\hat{\beta}_1) \doteq \frac{1}{(0.7)^2}\frac{1}{400}2\left(1 - \frac{0.7^2}{2}\right) + \frac{1}{(0.7)^2}\frac{1}{400}(1 - 0.7^2) \doteq 0.0103$$

which is consistent with the simulation result in Table 6.1. In addition to point estimation, we also compute variance estimators for PFI and HDFI methods. Variance estimators show small relative biases (less than 5% in absolute values) for all parameters. Variance estimation results are not presented here for brevity.

Table 6.1
Monte Carlo means and variances of point estimators from Simulation One. (SRI, stochastic regression imputation; PFI, parametric fractional imputation; HDFI; hot-deck fractional imputation)

Parameter	Method	Mean	Variance
β_0	Full	1.00	0.0123
	SRI	1.90	0.0869
	PFI	1.00	0.0472
	HDFI	1.00	0.0465
β_1	Full	1.00	0.00249
	SRI	0.54	0.01648
	PFI	1.00	0.01031
	HDFI	1.00	0.01026
σ^2	Full	1.00	0.00482
	SRI	1.73	0.01657
	PFI	0.99	0.02411
	HDFI	0.99	0.02270
π	Full	0.374	0.00058
	SRI	0.305	0.00255
	PFI	0.375	0.00059
	HDFI	0.375	0.00057

The proposed method is based on the instrumental variable assumption. To study the sensitivity of the proposed fractional imputation method to violations of the instrumental variable assumption, we performed an additional simulation study. Now, instead of generating y_{2i} from (6.1), we use

$$y_{2i} = 0.5 + y_{1i} + \rho(x_i - 3) + e_i, \tag{6.2}$$

where $e_i \sim N(0,1)$ and ρ can take non-zero values. We use three values of ρ, $\rho \in \{0, 0.1, 0.2\}$, in the sensitivity analysis and apply the same PFI and HDFI procedure that is based on the assumption that x is an instrumental variable for y_1. Such assumption is satisfied for $\rho = 0$, but it is weakly violated for $\rho = 0.1$ or $\rho = 0.2$. Using the fractionally imputed data in sample B, we estimated three parameters, $\theta_1 = E(Y_1)$, θ_2 is the slope for the simple regression of y_2 on y_1, and $\theta_3 = P(y_1 < 2, y_2 < 3)$, the proportion of $y_1 < 2$ and $y_2 < 3$. Table 6.2 presents Monte Carlo means and variances of the point estimators for three parameters under three different models. In Table 6.2, the absolute values of the difference between the fractionally imputed estimator and the full sample estimator increase as the value of ρ increases, which is expected as the instrumental variable assumption is more severely violated for larger values of ρ, but the differences are relatively small for all cases. In particular, the estimator of θ_1 is not affected by the departure from the instrumental variable assumption. This is because the imputation estimator under the incorrect imputation model still provides an unbiased estimator for the population mean as long as the regression imputation model contains an intercept term (Kim and Rao 2012). Thus, this limited sensitivity analysis suggests that the proposed method seems to provide comparable estimates when the instrumental variable assumption is weakly violated.

Table 6.2
Monte Carlo means and Monte Carlo variances of the two point estimators for sensitivity analysis in Simulation One (Full, full sample estimator; PFI, parametric fractional imputation; HDFI; hot-deck fractional imputation)

Model	Parameter	Method	Mean	Variance
$\rho = 0$	θ_1	Full	2.00	0.00235
		PFI	2.00	0.00352
		HDFI	2.00	0.00249
	θ_2	Full	1.00	0.00249
		PFI	1.00	0.01031
		HDFI	1.00	0.01026
	θ_3	Full	0.43	0.00061
		PFI	0.43	0.00059
		HDFI	0.43	0.00057
$\rho = 0.1$	θ_1	Full	2.00	0.00235
		PFI	2.00	0.00353
		HDFI	2.00	0.00250
	θ_2	Full	1.07	0.00248
		PFI	1.14	0.01091
		HDFI	1.14	0.01081
	θ_3	Full	0.44	0.00061
		PFI	0.45	0.00062
		HDFI	0.45	0.00059
$\rho = 0.2$	θ_1	Full	2.00	0.00235
		PFI	2.00	0.00353
		HDFI	2.00	0.00250
	θ_2	Full	1.14	0.00250
		PFI	1.28	0.01115
		HDFI	1.28	0.01102
	θ_3	Full	0.44	0.00061
		PFI	0.46	0.00066
		HDFI	0.46	0.00062

6.2 Simulation two

In the second simulation study, we consider a binary response variable y_{2i}, where

$$y_{2i} \sim \text{Bernoulli}(p_i), \tag{6.3}$$

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 y_{1i},$$

and $y_{1i} \sim N(\mu_1, \sigma_1^2)$. In the main sample, denoted by B, instead of observing (y_{1i}, y_{2i}), we observe (x_i, y_{2i}), where

$$x_i = \beta_0 + \beta_1 y_{1i} + u_i, \tag{6.4}$$

and $u_i \sim N\left(0, \sigma^2 \mid y_{1i} \mid^{2\alpha}\right)$. We observe (x_i, y_{1i}), $i = 1, \ldots, n_A$ in a calibration sample, denoted by A. For the simulation, $n_A = n_B = 800$, $\gamma_0 = 1$, $\gamma_1 = 1$, $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2 = 0.25$, $\alpha = 0.4$, $\mu_1 = 0$, and $\sigma_1^2 = 1$. Primary interest is in estimation of γ_1 and testing the null hypothesis that $\gamma_1 = 1$. The Monte Carlo (MC) sample size is 1,000.

We compare the PFI estimators of γ_1 to three other estimators. Because the conditional distribution of y_{1i} given x_i is non-standard, we use the weights of (5.3) to implement PFI, where the proposal distribution $\hat{h}_a(y_{1i} \mid x_i)$ is an estimate of the marginal distribution of y_{1i} based on the data from Sample A. We consider the following four estimators:

1. *PFI*: For PFI, the proposal distribution for generating $y_{1i}^{*(j)}$ is a normal distribution with mean $\hat{\mu}_1$ and variance $\hat{\sigma}_1^2$, where $\hat{\mu}_1$ and $\hat{\sigma}_1^2$ are the maximum likelihood estimates of μ_1 and σ_1^2, respectively, based on Sample A. The fractional weights defined in (5.3) has the form

$$w_{ij}^* \propto \hat{p}_{ij}^{y_{2i}} \left(1 - \hat{p}_{ij}\right)^{1-y_{2i}} \hat{f}_a\left(y_{1i}^{*(j)} \mid x_i\right), \tag{6.5}$$

where $\hat{p}_{ij} = \left\{1 + \exp\left(-\hat{\gamma}_0 - \hat{\gamma}_1 y_{1i}^{*(j)}\right)\right\}^{-1}$ and $\hat{f}_a(y_{1i} \mid x_i)$ is the estimate of $f(y_{1i} \mid x_i)$ based on maximum likelihood estimation with the Sample A data. The imputation size $m = 800$.

2. *Naive*: A *naive* estimator is the estimator of the slope in the logistic regression of y_{2i} on x_i for $i \in B$.

3. *Bayes*: We use the approach of Guo and Little (2011) to define a Bayes estimator. The model for this simulation differs from the model of Guo and Little (2011) in that the response of interest is binary. We implement GIBBS sampling with JAGS (Plummer 2003), specifying diffuse proper prior distributions for the parameters of the model. Letting

$$\theta_1 = \left(\log\left(\sigma^2\right), \log\left(\sigma_1^2\right), \mu_1, \beta_0, \beta_1, \gamma_0, \gamma_1\right),$$

we assume a priori that $\theta_1 \sim N\left(0, 10^6 I_7\right)$, where I_7 is a 7×7 identity matrix, and the notation $N(0, V)$ denotes a normal distribution with mean 0 and covariance matrix V. The prior distribution for the power α is uniform on the interval $[-5, 5]$.

To evaluate convergence, we examine trace plots and potential scale reduction factors defined in Gelman, Carlin, Stern and Rubin (2003) for 10 preliminary simulated data sets. We initiate three MCMC chains, each of length 1,500 from random starting values and discard the first 500 iterations as burn-in. The potential scale reduction factors across the 10 simulated data sets range from 1.0 to 1.1, and the trace plots indicate that the chains mix well. To reduce computing time, we use 1,000 iterations of a single chain for the main simulation, after discarding the first 500 for burn-in.

4. A *Weighted Regression Calibration (WRC)* estimator. The WRC estimator is a modification of the weighted regression calibration estimator defined in Guo and Little (2011) for a binary response variable. The computation for the weighted regression calibration estimator involves the following steps:

   (i) Using ordinary least squares (OLS), regress y_{1i} on x_i for the calibration sample.

(ii) Regress the logarithm of the squared residuals from step (i) on the logarithm of x_i^2 for the calibration sample. Let $\hat{\lambda}$ denote the estimated slope from the regression.

(iii) Using weighted least squares (WLS) with weight $|x_i|^{2\hat{\lambda}}$, regress y_{1i} on x_i for the calibration sample. Let $\hat{\eta}_0$ and $\hat{\eta}_1$ be the estimated intercept and slope, respectively, from the WLS regression.

(iv) For each unit i in the main sample, let $\hat{y}_{1i} = \hat{\eta}_0 + \hat{\eta}_1 x_i$.

(v) The estimate of (γ_0, γ_1) is obtained from the logistic regression of y_{2i} on $\hat{y}_{1i}$ in the main sample.

Table 6.3 contains the MC bias, variance, and MSE of the four estimators of γ_1. The naive estimator has a negative bias because x_i is a contaminated version of y_{1i}. The PFI estimator is superior to the Bayes and WRC estimators.

We compute an estimate of the variance of the PFI estimators of γ_1 using the variance expression based on the linear approximation. We define the MC relative bias as the ratio of the difference between the MC mean of the variance estimator and the MC variance of the estimator to the MC variance of the estimator. The MC relative bias of the variance estimators for PFI is negligible (less than 2% in absolute values).

Table 6.3
Monte Carlo (MC) means, variances, and mean squared errors (MSE) of point estimators of γ_1 from Simulation Two. (PFI, parametric fractional imputation; WRC, weighted regression calibration; MC, Monte Carlo; MSE, mean squared error)

Method	MC Bias	MC Variance	MC MSE
PFI	0.0239	0.0386	0.0392
Naive	-0.2241	0.0239	0.0742
Bayes	0.0406	0.0415	0.0432
WRC	0.112	0.0499	0.0625

7 Concluding remarks

We approach statistical matching as a missing data problem and propose the PFI method to obtain consistent estimators and corresponding variance estimators. Under the assumption that the specified model is fully identified, the proposed method provides the pseudo maximum likelihood estimators of the parameters in the model.

A sufficient condition for model identifiability is the existence of an instrumental variable in the model. The measurement error framework of Section 5 and Section 6.2, where external calibration provides an independent measurement of the true covariate of interest, is a situation in which the study design may be judged to support the instrumental variable assumption. The proposed methodology is applicable without the instrumental variable assumption, as long as the model is identified. If the model is not identifiable, then the EM algorithm for the proposed PFI method does not necessarily converge. In practice, one can treat the specified model as identified if the EM sequence converges. That is, as long as the EM sequence converges, the resulting analysis is consistent under the specified model. This is one of the main advantages of using the frequentist approach over Bayesian. In the Bayesian approach, it is possible to obtain the posterior values even under non-identified models and the resulting analysis can be misleading.

Testing whether the IV assumption holds in the data at hand is much more difficult, perhaps impossible, under the data structure in Table 1.1. Instead, given the specified model, we can only check whether the model parameters are fully estimable. On the other hand, whether the specified model is a good model for the data at hand is a different story. Model diagnostics and model selection among different identifiable models are certainly important future research topics.

Statistical matching can also be used to evaluate effects of multiple treatments in observational studies. By properly applying statistical matching techniques, we can create an augmented data file of potential outcomes so that causal inference can be investigated with the augmented data file (Morgan and Winship 2007). Such extensions will be presented elsewhere.

Acknowledgements

We thank Professor Yanyuan Ma, an anonymous referee and the Assistant Editor (AE) for very constructive comments. The research of the first author was partially supported by Brain Pool program (131S-1-3-0476) from Korean Federation of Science and Technology Society and by a grant from NSF (MMS-121339). The research of the second author was supported by a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University. The work of the third author was supported by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation in Korea.

Appendix

A. Asymptotic unbiasedness of 2SLS estimator

Assume that we observe (y_1, x) in Sample A and observe (y_2, x) in Sample B. To be more rigorous, we can write (y_{1a}, x_a) to denote the observation (y_1, x) in Sample A. Also, we can write (y_{2b}, x_b) to denote the observations in Sample B. In this case, the model can be written as

$$\begin{aligned} y_{1a} &= \phi_0 1_a + \phi_1 x_{1a} + \phi_2 x_{2a} + e_{1a} \\ y_{2b} &= \beta_0 1_b + \beta_1 y_{1b} + \beta_2 x_{2b} + e_{2b} \end{aligned}$$

with $E(e_{1a} \mid x_a) = 0$ and $E(e_{2b} \mid x_b, y_{1b}) = 0$. Note that y_{1b} is not observed from the sample. Instead, we use $\hat{y}_{1b}$ using the OLS estimate obtained from Sample A.

Writing $X_a = [1_a, x_a]$ and $X_b = [1_b, x_b]$, we have $\hat{y}_{1b} = X_b (X_a' X_a)^{-1} X_a' y_{1a} = X_b \hat{\phi}_a$. The 2SLS estimator of $\beta = (\beta_0, \beta_1, \beta_2)'$ is then

$$\hat{\beta}_{2\mathrm{SLS}} = (Z_b' Z_b)^{-1} Z_b' y_{2b}$$

where $Z_b = [1_b, \hat{y}_{1b}, x_{2b}]$. Thus, we have

$$\begin{aligned} \hat{\beta}_{2\mathrm{SLS}} - \beta &= (Z_b' Z_b)^{-1} Z_b' (y_{2b} - Z_b \beta) \\ &= (Z_b' Z_b)^{-1} Z_b' \{\beta_1 (y_{1b} - \hat{y}_{1b}) + e_{2b}\}. \end{aligned} \tag{A.1}$$

We may write

$$y_{1b} = \phi_0 1_b + \phi_1 x_b + e_{1b} = X_b \phi + e_{1b}$$

where $E(e_{1b} \mid x_b) = 0$. Since

$$\begin{aligned} \hat{y}_{1b} &= X_b (X_a' X_a)^{-1} X_a' y_{1a} \\ &= X_b (X_a' X_a)^{-1} X_a' (X_a \phi + e_{1a}) \\ &= X_b \phi + X_b (X_a' X_a)^{-1} X_a' e_{1a}, \end{aligned}$$

we have

$$y_{1b} - \hat{y}_{1b} = e_{1b} - X_b (X_a' X_a)^{-1} X_a' e_{1a}$$

and (A.1) becomes

$$\hat{\beta}_{2\mathrm{SLS}} - \beta = (Z_b' Z_b)^{-1} Z_b' \left\{\beta_1 e_{1b} - \beta_1 X_b (X_a' X_a)^{-1} X_a' e_{1a} + e_{2b}\right\}. \tag{A.2}$$

Assume that the two samples are independent. Thus, $E(e_{1b} \mid x_a, x_b, y_{1a}) = 0$. Also, $E\left\{(Z_b' Z_b)^{-1} Z_b' e_{2b} \,\middle|\, x_a, x_b, y_{1a}, y_{1b}\right\} = 0$. Thus,

$$E\left\{\hat{\beta}_{2\mathrm{SLS}} - \beta \,\middle|\, x_a, x_b, y_{1a}\right\} = E\left\{-\beta_1 (Z_b' Z_b)^{-1} Z_b' X_b (X_a' X_a)^{-1} X_a' e_{1a} \,\middle|\, x_a, x_b, y_{1a}\right\}$$

and

$$\begin{aligned} (Z_b' Z_b)^{-1} Z_b' X_b (X_a' X_a)^{-1} X_a' e_{1a} &= (Z_b' Z_b)^{-1} Z_b' \left\{X_b (X_a' X_a)^{-1} X_a' (y_{1a} - X_a \phi)\right\} \\ &= (Z_b' Z_b)^{-1} Z_b' X_b \left(\hat{\phi}_a - \phi\right). \end{aligned}$$

This term has zero expectation asymptotically because $n_b^{-1} Z_b' Z_b$ and $n_b^{-1} Z_b' X_b$ are bounded in probability and $\left(\hat{\phi}_a - \phi\right)$ converges to zero.

B. Variance estimation

Let the parameter of interest be defined by the solution to $U_N(\eta) = \sum_{i=1}^{N} U(\eta; y_{1i}, y_{2i}) = 0$. We assume that $\partial U_N(\eta) / \partial \theta = 0$. Thus, parameter η is priori independent of θ which is the parameter in the data-generating distribution of (x, y_1, y_2).

Under the setup of Section 3, let $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2\right)$ be the MLE of $\theta = (\theta_1, \theta_2)$ obtained by solving (3.4). Also, let $\hat{\eta}$ be the solution to $\bar{U}\left(\eta \mid \hat{\theta}\right) = 0$ where

$$\bar{U}(\eta \mid \theta) = \sum_{i \in B} \sum_{j=1}^{m} w_{ib} w_{ij}^{*} U\left(\eta; y_{1i}^{*(j)}, y_{2i}\right),$$

and

$$w_{ij}^* \propto f\left(y_{1i}^{*(j)} \,\middle|\, x_i; \hat{\theta}_1\right) f\left(y_{2i} \,|\, y_{1i}^{*(j)}; \hat{\theta}_2\right) \Big/ h\left(y_{1i}^{*(j)} \,\middle|\, x_i\right)$$

with $\sum_{j=1}^{m} w_{ij}^* = 1.$ Here, $h(y_1 \,|\, x)$ is the proposal distribution of generating imputed values of y_1 in the parametric fractional imputation. By introducing the proposal distribution $h,$ we can safely ignore the dependence of imputed values $y_{1i}^{*(j)}$ on the estimated parameter value $\hat{\theta}_1.$

By Taylor linearization,

$$\bar{U}\left(\eta \,|\, \hat{\theta}\right) \cong \bar{U}\left(\eta \,|\, \theta\right) + \left(\partial \bar{U} / \partial \theta_1'\right)\left(\hat{\theta}_1 - \theta_1\right) + \left(\partial \bar{U} / \partial \theta_2'\right)\left(\hat{\theta}_2 - \theta_2\right)$$

Note that

$$\hat{\theta}_1 - \theta_1 \cong \{I_1(\theta_1)\}^{-1} S_1(\theta_1)$$

where $I_1(\theta_1) = -\partial S_1(\theta_1) / \partial \theta_1'.$ Also,

$$\hat{\theta}_2 - \theta_2 \cong \left\{ -\frac{\partial}{\partial \theta_2'} \bar{S}_2(\theta) \right\}^{-1} \bar{S}_2(\theta)$$

where

$$\bar{S}_2(\theta) = \sum_{i \in B} \sum_{j=1}^{m} w_i w_{ij}^*(\theta) S_2\left(\theta_2; y_{1i}^{*(j)}, y_{2i}\right).$$

Thus, we can establish

$$\bar{U}\left(\eta \,|\, \hat{\theta}\right) \cong \bar{U}\left(\eta \,|\, \theta\right) + K_1 S_1(\theta_1) + K_2 \bar{S}_2(\theta),$$

where $K_1 = D_{21} I_{11}^{-1}$ and $K_2 = D_{22} I_{22}^{-1}$ with $I_{11} = -E\left(\partial S_1 / \partial \theta_1'\right),$ $I_{22} = -E\left(\partial \bar{S}_2 / \partial \theta_2'\right),$ $D_{21} = E\left\{U(\eta) S_1(\theta_1)'\right\}$ and $D_{22} = E\left\{U(\eta) S_2(\theta_2)'\right\},$ we have

$$V\left\{\bar{U}\left(\eta \,|\, \hat{\theta}\right)\right\} = \tau^{-1}\left\{V_1 + V_2\right\}\tau^{-1'}$$

where $\tau = -E\left\{\partial \bar{U}(\eta \,|\, \theta) / \partial \eta'\right\},$

$$V_1 = V\left\{\sum_{i \in B} w_i \left(\bar{u}_i^* + K_2 S_{2i}^*\right)\right\},$$

$\bar{u}_i^* = E\left[U(\hat{\eta}; y_{1i}, y_{2i}) \,|\, y_{2i}; \hat{\theta}\right],$ and $V_2 = V\left\{K_1 \sum_{i \in A} w_i S_{1i}\right\}.$ A consistent estimator of each component can be developed similarly to Section 3.

References

Baker, K.H., Harris, P. and O’Brien, J. (1989). Data fusion: An appraisal and experimental evaluation. *Journal of the Market Research Society*, 31, 152-212.

Beaumont, J.-F., and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 37, 3, 400-416.

Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American Statistical Association*, 96, 453, 260-269.

Chib, S., and Greenberg, E. (1995). Jackknife variance estimation for nearest neighbor imputation. *The American Statistician*, 46, 327-333.

Chipperfield, J.O., and Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25, 2, 227-244.

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

Fuller, W.A. (2009). *Sampling Statistics*, Hoboken, NJ: John Wiley & Sons, Inc.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, Chapman and Hall Texts in Statistical Science. Chapman and Hall/CRC, second edition.

Gonzalez, J., and Eltinge, J. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2081-2088.

Guo, Y., and Little, R.J. (2011). Regression analysis with covariates that have heteroskedastic measurement error. *Statistics Medicine*, 30, 18, 2278-2294.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference,* (Eds., C.R. Rao and D. Pfeffermann), 215-246.

Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques.* New York: Springer.

Ibrahim, J.G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85, 765-769.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119-132.

Kim, J.K., and Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99, 85-100.

Kim, J.K., and Shao, J. (2013). *Statistical Methods in Handling Incomplete Data*, Chapman and Hall/CRC.

Kim, J.K., and Yang, S. (2014). Fractional hot deck imputation for robust inference under item nonresponse in survey sampling. *Survey Methodology*, 40, 2, 211-230.

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 1265-1275.

Leulescu, A., and Agafitei, M. (2013). Statistical matching: A model based approach for data integration. *Eurostat Methodologies and Working Papers.*

Morgan, S.L., and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* New York, USA: Cambridge University Press.

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

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Raghunathan, T.E., and Grizzle, J.E. (1995). A split questionnaire design. *Journal of the American Statistical Association*, 90, 54-63.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer-Verlag.

Ridder, S., and Moffit, R. (2007). The econometrics of data combination. *Handbook of Econometrics*, 5470-5544.