Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Imputation for nonmonotone nonresponse in the survey of industrial research and development

by Jun Shao, Martin Klein and Jing Xu

December 2012

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 infostats@statcan.gc.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

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

This product is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

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 “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

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 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 ($p < 0.05$)

Survey Methodology, December 2012
Vol. 38, No. 2, pp. 143-155
Statistics Canada, Catalogue No. 12-001-X

Imputation for nonmonotone nonresponse in the survey of industrial research and development

Jun Shao, Martin Klein and Jing Xu [1]

Abstract

Nonresponse in longitudinal studies often occurs in a nonmonotone pattern. In the Survey of Industrial Research and Development (SIRD), it is reasonable to assume that the nonresponse mechanism is past-value-dependent in the sense that the response propensity of a study variable at time point t depends on response status and observed or missing values of the same variable at time points prior to t. Since this nonresponse is nonignorable, the parametric likelihood approach is sensitive to the specification of parametric models on both the joint distribution of variables at different time points and the nonresponse mechanism. The nonmonotone nonresponse also limits the application of inverse propensity weighting methods. By discarding all observed data from a subject after its first missing value, one can create a dataset with a monotone ignorable nonresponse and then apply established methods for ignorable nonresponse. However, discarding observed data is not desirable and it may result in inefficient estimators when many observed data are discarded. We propose to impute nonrespondents through regression under imputation models carefully created under the past-value-dependent nonresponse mechanism. This method does not require any parametric model on the joint distribution of the variables across time points or the nonresponse mechanism. Performance of the estimated means based on the proposed imputation method is investigated through some simulation studies and empirical analysis of the SIRD data.

Key Words: Bootstrap; Imputation model; Kernel regression; Missing not at random; Longitudinal study; Past-value-dependent.

1. Introduction

Longitudinal studies, in which data are collected from every sampled subject at multiple time points, are very common in research areas such as medicine, population health, economics, social sciences, and sample surveys. The statistical analysis in a sample survey typically aims to estimate or make inference on the mean of a study variable at each time point. Nonresponse or missing data in the study variable is a serious impediment to performing a valid statistical analysis, because the response propensity (PSI) may directly or indirectly depend on the value of the study variable. Nonresponse is monotone if, whenever a value is missing at a time point t, all future values at $s > t$ are missing. We focus on nonmonotone nonresponse, which often occurs in longitudinal surveys. In the Survey of Industrial Research and Development (SIRD) conducted jointly by the U.S. Census Bureau and the U.S. National Science Foundation (NSF), for example, a business may be a nonrespondent on research and development expenditures at year $t-1$ but a respondent at year t. For ease we refer to SIRD in the present tense throughout, but we note that as of 2008, it has been replaced by the Business R&D and Innovation Survey.

Some existing methods for handling nonmonotone nonresponse can be briefly described as follows. The parametric approach assumes parametric models for both the PSI and the joint distribution of the study variable across time points (*e.g.*, Troxel, Harrington and Lipsitz 1998, Troxel, Lipsitz and Harrington 1998). The validity of the parametric approach, however, depends on whether parametric models are correctly specified. Vansteelandt, Rotnitzky and Robins (2007) proposed some methods under some models of the PSI at time t conditional on observed past data. Xu, Shao, Palta and Wang (2008) derived an imputation procedure under the assumptions that (i) the PSI at t depends only on values of the study variable at time $t-1$ and (ii) the study variables over different time points is a Markov chain. Another approach, which will be referred to as censoring, is to create a dataset with "monotone nonresponse" by discarding all observed values of the study variable from a sampled subject after its first missing value. Methods appropriate for monotone nonresponse (*e.g.*, Diggle and Kenward 1994, Robins and Rotnitzky 1995, Paik 1997) can then be applied to the reduced dataset. This approach may be inefficient when many observed data are discarded. Furthermore, in practical applications it is not desirable to throw away observed data.

The purpose of this article is to propose an imputation method for longitudinal data with nonmonotone nonresponse under the past-value-dependent PSI assumption described by Little (1995): at a time point t, the nonresponse propensity depends on values of the study variable at time points prior to t. This assumption on the PSI is weaker than

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706. E-mail: shao@stat.wisc.edu; Martin Klein, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C. 20233; Jing Xu, Department of Statistics, University of Wisconsin, Madison, WI 53706.

that in Xu *et al.* (2008) and is different from those in Vansteelandt *et al.* (2007). We consider imputation which does not require building a model for the PSI. Imputation is commonly used to compensate for missing values in survey problems (Kalton and Kasprzyk 1986). Once all missing values are imputed, estimates of parameters are computed using the estimated means for complete data by treating imputed values as observations. The proposed imputation and estimation methodology, including a bootstrap method for variance estimation, is introduced in Section 2. To examine the finite sample performance of the proposed method, we present some simulation results in Section 3. We also include an application of the proposed method to the SIRD. The last section contains some concluding remarks.

2. Methodology

We consider the model-assisted approach for survey data sampled from a finite population P. We assume that the population P is divided into a fixed number of imputation classes, which are typically unions of some strata. Within each imputation class, the study variable from a population unit follows a superpopulation. Let y_t be the study variable at time point t, $t = 1, ..., T$, $\mathbf{y} = (y_1, ..., y_T)$, δ_t be the indicator of whether y_t is observed, and $\boldsymbol{\delta} = (\delta_1, ..., \delta_T)$. Since imputation is carried out independently within each imputation class, for simplicity of notation we assume in this section that there is only a single imputation class.

Throughout this paper, we consider nonmonotone nonresponse and assume that there is no nonresponse at baseline $t = 1$. The PSI is past-value-dependent if

$$P(\delta_t = 1 \mid \mathbf{y}, \delta_1, ..., \delta_{t-1}, \delta_{t+1}, ..., \delta_T) = P(\delta_t = 1 \mid y_1, ..., y_{t-1}, \delta_1, ..., \delta_{t-1}), \quad t = 2, ..., T, \quad (1)$$

where P is with respect to the superpopulation. When nonresponse is monotone, the past-value-dependent PSI becomes ignorable (Little and Rubin 2002), since we either observe all past values or know with certainty that y_t is missing if it is missing at $t-1$, and an imputation method using linear regression proposed by Paik (1997) can be used. When nonresponse is nonmonotone, however, the past-value-dependent PSI is nonignorable because the response indicator at time t is statistically dependent upon previous values of the study variable, some of which may not be observed. In this case Paik's method does not apply.

2.1 Imputation for subjects whose first missing is at t

Let $t > 1$ be a fixed time point and $r+1$ be the time point at which the first missing value of $\mathbf{y}$ occurs. When $r+1 = t$, *i.e.*, a subject whose first missing value is at t, our proposed imputation procedure is the same as that for the case of monotone nonresponse (Paik 1997). However, we still need to provide a justification since we have a different PSI. It is shown in the Appendix that, under assumption (1),

$$E(y_t \mid y_1, ..., y_{t-1}, \delta_1 = \cdots = \delta_{t-1} = 1, \delta_t = 0) = E(y_t \mid y_1, ..., y_{t-1}, \delta_1 = \cdots = \delta_{t-1} = 1, \delta_t = 1) \quad t = 2, ..., T, \quad (2)$$

where E is the expectation with respect to the superpopulation. Denote the quantity on the first line of (2) by $\phi_{t, t-1}(y_1, ..., y_{t-1})$, which is the conditional expectation of a missing y_t given observed $y_1, ..., y_{t-1}$. If $\phi_{t, t-1}$ is known, then a natural imputed value for y_t is $\phi_{t, t-1}(y_1, ..., y_{t-1})$. However, $\phi_{t, t-1}$ is usually unknown. Since $\phi_{t, t-1}$ cannot be estimated by regressing y_t on $y_1, ..., y_{t-1}$ based on data from subjects with missing y_t values, we need to use (2), *i.e.*, the fact that $\phi_{t, t-1}$ is the same as the quantity on the second line of (2), which is the conditional expectation of an observed y_t given observed $y_1, ..., y_{t-1}$ and can be estimated by regressing y_t on $y_1, ..., y_{t-1}$, using data from all subjects having observed y_t and observed $y_1, ..., y_{t-1}$. Note that (2) is a counterpart of (5) in Xu *et al.* (2008) under the last-value-dependent assumption, which is stronger than the past-value-dependent assumption (1). Under a stronger assumption, we are able to utilize more data in regression fitting.

Suppose that a sample S is selected from P according to a given probability sampling plan. For each $i \in S$, $\boldsymbol{\delta}_i = (\delta_{i1}, ..., \delta_{iT})$ is observed, the study variable y_{it} with $\delta_{it} = 1$ is observed, and y_{it} with $\delta_{it} = 0$ is not observed, $t = 1, ..., T$. With respect to the superpopulation, $(\mathbf{y}_i, \boldsymbol{\delta}_i)$ has the same distribution as $(\mathbf{y}, \boldsymbol{\delta})$ and $(\mathbf{y}_i, \boldsymbol{\delta}_i)$'s are independent, where $\mathbf{y}_i = (y_{i1}, ..., y_{iT})$. For $t = 2, ..., T$, let $\hat{\phi}_{t, t-1}$ be the regression estimator of $\phi_{t, t-1}$ based on observations with $\delta_{i1} = \cdots = \delta_{i(t-1)} = 1$. A missing y_{it} with observed $y_{i1}, ..., y_{i(t-1)}$ is then imputed by $\tilde{y}_{it} = \hat{\phi}_{t, t-1}(y_{i1}, ..., y_{i(t-1)})$.

To illustrate, we consider the case of $t = 3$ or 4. The horizontal direction in Table 1 corresponds to time points and the vertical direction corresponds to different missing patterns, where each pattern is represented by a vector of 0's and 1's with 0 indicating a missing value and 1 indicating an observed value. For $t = 3$ and $r = 2$, as the first of the two steps, we consider missing data at time 3 with first missing at time 3, *i.e.*, pattern (1,1,0). According to imputation model (2), we fit a regression using data in pattern (1,1,1) indicated by + (used as predictors) and × (used as responses). Then, imputed values (indicated by ○) are obtained from the fitted regression using data indicated by ∗ as predictors. For $t = 4$ and $r = 3$, imputation in pattern (1,1,1,0) can be similarly done using data in pattern (1,1,1,1) for regression fitting.

Table 1
Illustration of imputation process

	Step 1: $r = 2, t = 3$			Step 2: $r = 1, t = 3$		
	Time			Time		
Pattern	1	2	3	1	2	3
(1,0,0)				*		○
(1,1,0)	*	*	○	+		⊗
(1,1,1)	+	+	×			
(1,0,1)						

	Step 1: $r = 3, t = 4$				Step 2: $r = 2, t = 4$				Step 3: $r = 1, t = 4$			
	Time				Time				Time			
Pattern	1	2	3	4	1	2	3	4	1	2	3	4
(1,0,0,0)									*			○
(1,1,0,0)					*	*		○	+			⊗
(1,1,1,0)	*	*	*	○	+	+		⊗	+			⊗
(1,0,1,0)									*			○
(1,0,0,1)												
(1,1,0,1)												
(1,0,1,1)												
(1,1,1,1)	+	+	+	×								

+ : observed data used in regression fitting as predictors.
× : observed data used in regression fitting as responses.
⊗ : imputed data used in regression fitting as responses.
* : observed data used as predictors in imputation.
○ : imputed values.

What type of regression we can fit to obtain $\tilde{y}_{it}$? It is shown in the Appendix that, if (1) holds and $E(y_t|y_1, ..., y_{t-1})$ is linear in $y_1, ..., y_{t-1}$ for any t in the case of no nonresponse, then

$$E(y_t|y_1, ..., y_{t-1}, \delta_1 = \cdots = \delta_{t-1} = 1) \text{ is linear in } y_1, ..., y_{t-1} \quad (3)$$

and, hence, linear regression under the model-assisted approach can be used to estimate $\phi_{t,\,t-1}$. If $E(y_t|y_1, ..., y_{t-1})$ is not linear, one of the methods described in Section 2.3 can be applied.

2.2 Imputation for subjects whose first missing is at $r + 1 < t$

Imputation for a subject whose first missing value is at time $r + 1 < t$ is more complicated and very different from that for the case of monotone nonresponse. This is because when $r + 1 < t$ and nonresponse is monotone,

$$E(y_t|y_1, ..., y_r, \delta_1 = \cdots = \delta_r = 1, \delta_t = 0) = E(y_t \mid y_1, ..., y_r, \delta_1 = \cdots = \delta_r = 1, \delta_t = 1) \quad r = 1, ..., t-2, \quad t = 2, ..., T, \quad (4)$$

whereas (4) does not hold when nonresponse is nonmonotone (see the proof in the Appendix). Hence, we need to construct different models for subjects whose first missing value is at $r + 1 < t$. It is shown in the Appendix that, when $r + 1 < t$,

$$E(y_t|y_1, ..., y_r, \delta_1 = \cdots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = E(y_t \mid y_1, ..., y_r, \delta_1 = \cdots = \delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) \quad r = 1, ..., t-2, \ t = 2, ..., T. \quad (5)$$

We now explain how to use (5) to impute missing values at a fixed time point t. Let $\phi_{t,r}(y_1,...,y_r)$ be the quantity on the first line of (5). If $\phi_{t,r}$ is known, then y_t can be imputed by $\phi_{t,r}(y_1,...,y_r)$. Otherwise, it needs to be estimated based on (5). Unlike in model (2) or (4), the conditional expectation on the second line of (5) is conditional on a missing y_t ($\delta_t = 0$), although $y_1, ..., y_r$ are observed. If we carry out imputation sequentially according to $r = t-1, t-2, ..., 1$, then, for a given $r < t-1$, the missing y_t values from subjects whose first missing is at time point $r + 2$ have already been imputed using the method in this section or Section 2.1. We can fit a regression between imputed y_t and observed $y_1, ..., y_r$ using data from all subjects having already imputed y_t (used as responses), observed $y_1, ..., y_r$ (used as predictors), and $\delta_{r+1} = 1$. Once an estimator $\hat{\phi}_{t,r}$ is obtained, a missing y_{it} with first missing at $r + 1$ is then imputed by $\tilde{y}_{it} = \hat{\phi}_{t,r}(y_{i1}, ..., y_{ir})$.

Consider again the case of $t = 3$ or 4 and Table 1. Following the first step for $t = 3$ discussed in Section 2.1, at the second step, we impute missing values with $r = 1$ in pattern (1,0,0). According to imputation model (5), we fit a regression using data in pattern (1,1,0) indicated by + (used as predictors) and ⊗ (previously imputed values used as

responses). Then, imputed values (indicated by ○) are obtained from the fitted regression using data indicated by $*$ as predictors. For $t = 4$, following the first step discussed in Section 2.1, at the second step ($r = 2$) we fit a regression using data in pattern (1,1,1,0) indicated by $+$ (used as predictors) and $\otimes$ (previously imputed values used as responses). Then, imputed values (indicated by ○) at $t = 4$ in pattern (1,1,0,0) are obtained from the fitted regression using data indicated by $*$ as predictors. At step 3 for $t = 4$, we fit a regression using data in patterns (1,1,0,0) and (1,1,1,0) indicated by $+$ (used as predictors) and $\otimes$ (previously imputed values used as responses). Then, imputed values (indicated by ○) at $t = 4$ in patterns (1,0,0,0) and (1,0,1,0) are obtained from the fitted regression using data indicated by $*$ as predictors.

Although at time t, imputation has to be carried out sequentially as $r = t-1, ..., 1$, imputation for different time points can be done in any order. This can be seen from the illustration given by Table 1, where the imputed values at $t = 3$ are not involved in the imputation process at $t = 4$ or vice versa, although some observed data will be repeatedly used in regression fitting. When data come according to time, it is natural to impute nonrespondents in the order $t = 2, ..., T$.

Why can we use previously imputed values as responses in the estimation of the regression function $\phi_{t,r}$ when $r < t-1$? For given t and $r < t-1$, a previously imputed value with first missing at $s+1 > r+1$ is an estimator of

$$\begin{aligned}\tilde{y}_t &= E(y_t \mid y_1,...,y_s,\delta_1 = \cdots = \delta_s = 1,\delta_{s+1} = 0,\delta_t = 0)\\ &= E(y_t \mid y_1,...,y_s,\delta_1 = \cdots = \delta_{s+1} = 1,\delta_t = 0).\end{aligned}$$

By the property of conditional expectation and (5),

$$\begin{aligned}&E[E(y_t \mid y_1,...,y_s,\delta_1 = \cdots = \delta_{s+1} = 1,\delta_t = 0) \mid\\ &\qquad y_1,...,y_r,\delta_1 = \cdots = \delta_{r+1} = 1,\delta_t = 0]\\ &= E(y_t \mid y_1,...,y_r,\delta_1 = \cdots = \delta_{r+1} = 1,\delta_t = 0)\\ &= E(y_t \mid y_1,...,y_r,\delta_1 = \cdots = \delta_r = 1,\delta_{r+1} = 0,\delta_t = 0). \quad (6)\end{aligned}$$

This means that y_t and $\tilde{y}_t$ have the same conditional expectation, given $y_1,...,y_r,\delta_1 = \cdots = \delta_r = 1,\delta_{r+1} = 0,\delta_t = 0$. Therefore, using previously imputed values as responses in regression produces a valid estimator of $\phi_{t,r}$. Note that previously imputed values should not be used as predictors in regression, as equation (6) does not hold if some of $y_1,...,y_s$ are imputed values.

Although all observed data at any time t are used for the estimation of $E(y_t)$, some but not all observed data at time $< t$ are utilized in imputation to avoid biases under nonignorable nonresponse. This is different in the ignorable nonresponse case, where typically all past observed data can be used in regression imputation.

2.3 Regression for imputation

The conditional expectations in (5) depend not only on the distribution of $\mathbf{y}$, but also on the PSI. Even if $E(y_t \mid y_1,...,y_{t-1})$ is linear, conditional expectations in (5) are not necessarily linear, which is different from the case of $r+1 = t$ considered in Section 2.1. An example is given by result (10) in the Appendix.

When we do not have a suitable parametric model for $\phi_{t,r}$, the nonparametric kernel regression method given in Cheng (1994) may be applied to obtain $\hat{\phi}_{t,r}$. Since the regressor $(y_{i1},...,y_{ir})$ is multivariate when $r \geq 2$, however, kernel regression has a large variability unless the number of sampled subjects in the category defined by $\delta_{i1} = \cdots = \delta_{i(r+1)} = 1$ is very large. This issue is commonly referred to as the curse of dimensionality.

Thus, we consider the following alternatives under the additional assumption that the dependence of δ_t on $y_1,...,y_{t-1}$ is through a linear combination of $y_1,...,y_{t-1}$. That is,

$$P(\delta_t = 1 \mid y_1,..., y_{t-1}, \delta_1,..., \delta_{t-1}) = \Psi\left(\sum_{l=1}^{t-1}\gamma_l^{\delta_1,...,\delta_{t-1}} y_l\right), \quad (7)$$

where $\gamma_l^{\delta_1,...,\delta_{t-1}}$, $l = 1, ..., t-1$, are unknown parameters depending on $\delta_1, ..., \delta_{t-1}$ and Ψ is an unknown function with range $[0, 1]$. Under (7), it is shown in the Appendix that

$$\begin{aligned}&E(y_t \mid z_r,\delta_1 = \cdots = \delta_r = 1,\delta_{r+1} = 0,\delta_t = 0)\\ &\qquad = E(y_t \mid z_r,\delta_1 = \cdots = \delta_r = 1,\delta_{r+1} = 1,\delta_t = 0)\\ &\qquad\qquad r = 1,...,t-2, t = 2,...,T, \quad (8)\end{aligned}$$

where $z_r = \sum_{l=1}^{r}\gamma_{r,l}\, y_l$ and $\gamma_{r,l} = \gamma_l^{\delta_1,...,\delta_r}$ with $\delta_1 = \cdots = \delta_r = 1$. Hence, to impute nonrespondents, we can condition on the linear combination z_r and use (8), instead of conditioning on $y_1, ..., y_r$ and using (5).

Let $\psi_{t,r}(z_r)$ be the function defined on the second line of (8). Note that $\psi_{t,r}$ is not necessary the same as $\phi_{t,r}$. If there is a strong linear relationship between y_t and $y_1, ..., y_r$, then $\psi_{t,r}$ may be approximately linear so that we can fit a linear regression to obtain an estimator $\hat{\psi}_{t,r}$. In theory, this method is biased when $\psi_{t,r}$ is not linear. If $\boldsymbol{\gamma}_r = (\gamma_{r,1}, ..., \gamma_{r,r})'$ is known, then we can apply a one-dimensional kernel regression to obtain an estimator $\hat{\psi}_{t,r}$, using the one-dimensional index z_r. Since $\boldsymbol{\gamma}_r$ is unknown, we first need to estimate it by $\hat{\boldsymbol{\gamma}}_r$ and then obtain $\hat{\psi}_{t,r}$ by applying the one-dimensional kernel regression with $\boldsymbol{\gamma}_r$ replaced by $\hat{\boldsymbol{\gamma}}_r$. For example, the sliced inverse regression (Duan and Li 1991) can be applied to obtain $\hat{\boldsymbol{\gamma}}_r$. However, this type of nonparametric method may be inefficient. If there is a strong linear relationship between y_t and $y_1, ..., y_r$, we may apply linear regression to obtain $\hat{\boldsymbol{\gamma}}_r$. In any case, we use $y_{i1}, ..., y_{ir}$ with $\delta_{i1} = \cdots = \delta_{i(r+1)} = 1$ as predictors and imputed y_{it} values as responses in any type

of regression fitting. After $\hat{\psi}_{t,r}$ and $\hat{\boldsymbol{\gamma}}_r = (\hat{\gamma}_{r,1},...,\hat{\gamma}_{r,r})'$ are obtained, a missing y_{it} is imputed by $\tilde{y}_{it} = \hat{\psi}_{t,r}(\hat{\gamma}_{r,1} y_{i1} + \cdots + \hat{\gamma}_{r,r} y_{ir})$.

We refer to the method of simply applying linear regression as the linear regression imputation method, and the method of applying kernel regression to the index z_r as the one-dimensional index kernel regression imputation method. An advantage of one-dimensional index kernel regression imputation over kernel regression imputation is that only a one-dimensional kernel regression is applied and, thus, it avoids the curse of dimensionality and has smaller variability.

These methods can also be applied to the case of $r = t-1$ if $E(y_t \mid y_1, ..., y_{t-1})$ is not linear.

In theory, estimators such as the estimated means based on kernel regression or one-dimensional index kernel regression imputation are asymptotically unbiased, but they may not be better than those based on linear regression imputation when the number of sampled subjects in each (t, r) category is not very large. The performances of the estimated means based on linear regression, kernel regression, and one-dimensional index kernel regression imputation are examined by simulation in Section 3.

2.4 Estimation

We consider the estimation of the finite population total or the mean of y_t at each fixed t, which is often the main purpose of a survey study. At any t, let $\tilde{y}_{it} = y_{it}$ when $\delta_{it} = 1$ and $\tilde{y}_{it}$ be the imputed value using one of the methods in Section 2 when $\delta_{it} = 0$. The finite population total and the mean of y_t can be estimated by

$$\hat{Y}_t = \sum_{i \in S} w_i \tilde{y}_{it} \qquad \text{and} \qquad \overline{Y}_t = \sum_{i \in S} w_i \tilde{y}_{it} \Big/ \sum_{i \in S} w_i , \quad (9)$$

respectively, where w_i is the survey weight constructed such that, in the case of no nonresponse, $\hat{Y}_t$ is an unbiased estimator of the finite population total at time t with respect to the probability sampling. The superpopulation mean of y_t can also be estimated by $\overline{Y}_t$. Note that $\sum_{i \in S} w_i$ is an unbiased estimator of the finite population size N and, for some simple sampling designs, it is exactly equal to N.

The survey weights should also be used in the regression fitting for imputation. Under the same conditions given in Cheng (1994), $\hat{Y}_t$ or $\overline{Y}_t$ based on kernel regression or one-dimensional index kernel regression imputation is consistent and asymptotically normal as the sample size increases to ∞. The required conditions and proofs can be found in Xu (2007).

If we apply the linear regression imputation method as discussed in Section 2.3, then the resulting estimated mean at t may be asymptotically biased. This bias is small if the function $\psi_{t,r}$ can be well approximated by a linear function in the range of the data values. On the other hand, kernel or one-dimensional index kernel regression imputation may require a much larger sample size than that for linear regression imputation. Hence, the overall performance of the estimated mean based on linear regression imputation may still be better, as indicated by the simulation results in Section 3.

2.5 Variance estimation

For assessing statistical accuracy or inference such as constructing a confidence interval for the mean of y_t at t, we need variance estimators of $\hat{Y}_t$ or $\overline{Y}_t$ based on imputed data. Because of the complexity of the imputation procedure, it is difficult to obtain explicit formulas for variance of $\hat{Y}_t$ or $\overline{Y}_t$. The bootstrap method (Efron 1979) is then considered. A correct bootstrap can be obtained by repeating the process of imputation in each of the bootstrap samples (Shao and Sitter 1996). Let $\hat{\theta}$ be the estimator under consideration. A bootstrap procedure can be carried out as follows.

1. Draw a bootstrap sample as a simple random sample of the same size as S with replacement from the set of sampled subjects.
2. For units in the bootstrap sample, their survey weights, response indicators, and observed data from the original data set are used to form a bootstrap data set. Apply the proposed imputation procedure to the bootstrap data. Calculate the bootstrap analog $\hat{\theta}^*$ of $\hat{\theta}$.
3. Independently repeat the previous steps B times to obtain $\hat{\theta}^{*1}, ..., \hat{\theta}^{*B}$. The sample variance of $\hat{\theta}^{*1}, ..., \hat{\theta}^{*B}$ is the bootstrap variance estimator for $\hat{\theta}$.

In application, each $\hat{\theta}^{*b}$ can be calculated using the b^{th} bootstrap data $(\mathbf{y}_i, \boldsymbol{\delta}_i, w_i^{*b})$, $i \in S$, where $w_i^{*b} = w_i$ multiplied by the number of times unit i appears in the b^{th} bootstrap sample. Note that the same w_i^{*b} can be used for all variables of interest, not just $\mathbf{y}_i$.

3. Empirical results

We study $\hat{Y}_t$ or $\overline{Y}_t$ in (9) based on the proposed imputation methods at each time point t. We first consider a simulation with a normal population for the y_t's. An application to the SIRD data is presented next. To examine the performance of the proposed methods for the SIRD, a simulation with a population generated using the SIRD data is presented in the end. We have implemented the proposed imputation methods in R (R Development Core Team 2009). To fit the required nonparametric regressions, we use the R function *loess* with default settings, which fits a local polynomial surface in one or more regressor variables. The required linear regressions are easily fit in R using the

function *lm*. Our implementations of the proposed methods include error checking; (such as ensuring that there are sufficient points for regression fitting at each stage) which is particularly important in bootstrap and simulation settings where the imputation methods are replicated many times, and each iteration cannot be examined manually. We defaulted to an overall mean imputation in cases where there were not enough data points to fit a regression.

3.1 Simulation results from a normal population

A simulation study was conducted with normally distributed $\mathbf{y}_1,...,\mathbf{y}_n$, $n = 2{,}000$, and $T = 4$. A single imputation class and simple random sampling with replacement was considered. In the simulation, $\mathbf{y}_i$'s were independently generated from the multivariate normal distribution with mean vector $(1.33, 1.94, 2.73, 3.67)$ and the covariance matrix having the AR(1) structure with correlation coefficient 0.7 and unit variance; all data at $t = 1$ were observed; missing data at $t = 2, 3, 4$ were generated according to

$$P(\delta_t = 1 \mid y_1,...,y_{t-1},\delta_1,...,\delta_{t-1}) = 1 - \Phi\left(0.6\left(1 - \sum_{j=1}^{t-1} y_j \gamma_j^{\delta_1,...,\delta_{t-1}}\right)\right)$$

where

$$\gamma_j^{\delta_1,\ ...,\ \delta_{t-1}} = \frac{j + (1-\delta_j)\, j}{\sum_{k=1}^{t-1} [k + (1-\delta_k)k]}, \quad j = 1, ..., t-1,$$

and Φ is the standard normal distribution function. The unconditional probabilities of nonresponse patterns are given in Table 2.

For comparison, we included a total of nine estimators of the mean of y_t: they are sample means based on (1) the complete data (used as the gold standard); (2) respondents with adjusted weights assuming the probability of response is the same within each imputation class; (3) censoring and linear regression imputation, which first discards all observations of a subject after the first missing value to create a dataset with "monotone nonresponse" and then applies linear regression imputation as described in Paik (1997); (4) the proposed kernel regression imputation; (5) the proposed linear regression imputation; (6) the proposed one-dimensional index kernel regression imputation using the sliced inverse regression to obtain $\hat{\boldsymbol{\gamma}}_r$; (7) the kernel regression imputation proposed in Xu *et al.* (2008) based on the last-value-dependent PSI; (8) the linear regression imputation based on a regression between respondents at time t and observed and imputed values at time points $1, ..., t-1$ (treating imputed as observed); (9) the linear regression imputation based on a regression between respondents at time t and observed data from units with the same missing pattern at time points $1, ..., t-1$.

Table 2
Probabilities of nonresponse patterns in the simulation study (Normal population)

	Pattern	Probability	
Monotone	(1,0,0,0)	0.062	
	(1,1,0,0)	0.043	total = 0.181
	(1,1,1,0)	0.076	
Intermittent	(1,0,0,1)	0.113	
	(1,0,1,0)	0.071	
	(1,0,1,1)	0.186	total = 0.494
	(1,1,0,1)	0.124	
Complete	(1,1,1,1)	0.325	

Method (2) simply ignores nonrespondents and, hence, is biased and inefficient. Under the PSI assumption (1) methods (7)-(9) are also biased for $t \geq 3$, because method (7) requires the last-value-dependent assumption that is stronger than (1), method (8) treats previously imputed values as observed in regression, and method (9) requires the following condition that is not true under (1):

$$E(y_t \mid y_1,...,y_{t-1},\delta_1 = j_1,...,\delta_{t-1} = j_{t-1},\delta_t = 0) = E(y_t \mid y_1,...,y_{t-1},\delta_1 = j_1,...,\delta_{t-1} = j_{t-1},\delta_t = 1)$$

where $(j_1, ..., j_{t-1})$ is a fixed missing pattern. Finally, as we discussed in Section 2.3, method (5) is also biased for $t \geq 3$ since linear regression is not an exactly correct model. However, methods (5), (8), and (9) may still perform well when the biases are not substantial, because the use of a simpler model and more data in regression for imputation may compensate for the loss in biased imputation. Furthermore, any assumption on the PSI may hold only approximately and it is desired to empirically study various methods in any particular application.

For the case of $r = t - 1$, linear regression imputation is applied as discussed in Section 2.1. Hence, methods (3)-(6), (8)-(9) all give the same results when $t = 2$.

Table 3 reports (based on 1,000 simulation runs) the relative bias and standard deviation (SD) of the mean estimator, the mean of $\widehat{\text{SD}}_{\text{boot}}$, the bootstrap estimator of SD based on 200 bootstrap replications, and the coverage probability of the approximate 95% confidence interval (CI) obtained using point estimator $\pm 1.96 \times \widehat{\text{SD}}_{\text{boot}}$. The following is a summary of the results in Table 3.

1. The sample mean based on ignoring missing data is clearly biased. Although in the case of $t = 4$ its relative bias is only 3.5%, it still leads to a very low coverage probability of the confidence interval, because the SD of the estimated mean is also very small.
2. The bootstrap estimator of standard deviation performs well in all cases, even when the mean estimator is biased.
3. $\bar{Y}_t$ based on censoring and linear regression imputation has negligible bias so that the related

confidence interval has a coverage probability close to the nominal level 95%; but it has a large SD when $t = 3$ or $t = 4$. The inefficiency of this method is obviously caused by discarding observed data from nearly 50% of sampled subjects who have intermittent nonresponse. Its performance becomes worse as t increases.

4. $\bar{Y}_t$ based on the proposed kernel regression imputation has a relative bias between 0.0% and 0.5%, but the bias is large enough to result in a poor coverage performance of the related confidence interval at $t = 4$.
5. $\bar{Y}_t$ based on the proposed linear regression imputation has negligible bias as well as a variance smaller than that of $\bar{Y}_t$ based on kernel regression. The related confidence interval has a coverage probability close to the nominal level 95%.
6. $\bar{Y}_t$ based on the proposed one-dimensional index kernel regression imputation is generally good but slightly worse than that based on the linear regression imputation.
7. $\bar{Y}_t$ based on methods (7)-(9) has non-negligible bias when $t = 3$ or $t = 4$, which results in poor performance of the related confidence interval.

Table 3
Simulation results for mean estimation (Normal population)

Method	Quantity	$t = 2$	$t = 3$	$t = 4$
Complete data	relative bias	0%	0%	0%
	SD	0.0221	0.0223	0.0221
	$\widehat{\text{SD}}_{\text{boot}}$	0.0223	0.0223	0.0224
	CI coverage	94.9%	94.4%	95.4%
Respondents only	relative bias	12.8%	6.8%	3.5%
	SD	0.0282	0.0272	0.0248
	$\widehat{\text{SD}}_{\text{boot}}$	0.0285	0.0267	0.0252
	CI coverage	0.0%	0.0%	0.2%
Censoring and linear regression imputation	relative bias	0.0%	0.0%	-0.1%
	SD	0.0275	0.0358	0.0418
	$\widehat{\text{SD}}_{\text{boot}}$	0.0276	0.0354	0.0431
	CI coverage	95.1%	94.6%	95.6%
Proposed kernel regression imputation	relative bias	0.0%	0.4%	0.5%
	SD	0.0275	0.0288	0.0283
	$\widehat{\text{SD}}_{\text{boot}}$	0.0276	0.0288	0.0288
	CI coverage	95.1%	92.5%	88.6%
Proposed linear regression imputation	relative bias	0.0%	0.1%	0.0%
	SD	0.0275	0.0286	0.0279
	$\widehat{\text{SD}}_{\text{boot}}$	0.0276	0.0287	0.0293
	CI coverage	95.1%	93.8%	95.7%
Proposed 1-dimensional index kernel regression imputation	relative bias	0.0%	0.4%	0.4%
	SD	0.0275	0.0288	0.0279
	$\widehat{\text{SD}}_{\text{boot}}$	0.0276	0.0288	0.0288
	CI coverage	95.1%	92.5%	91.7%
Last-value-dependent kernel regression imputation	relative bias	0.6%	1.0%	0.6%
	SD	0.0284	0.0310	0.0257
	$\widehat{\text{SD}}_{\text{boot}}$	0.0288	0.0295	0.0263
	CI coverage	93.7%	84.2%	86.2%
Linear regression imputation treating previously imputed values as observed	relative bias	0.0%	1.6%	0.8%
	SD	0.0275	0.0261	0.0241
	$\widehat{\text{SD}}_{\text{boot}}$	0.0276	0.0260	0.0246
	CI coverage	95.1%	59.7%	76.0%
Linear regression imputation based on currently and previously observed data	relative bias	0.0%	1.6%	0.8%
	SD	0.0275	0.0261	0.0242
	$\widehat{\text{SD}}_{\text{boot}}$	0.0276	0.0261	0.0246
	CI coverage	95.1%	59.0%	76.1%

Although the kernel regression is asymptotically valid, in this simulation study the total number of subjects is 2,000 and, according to Table 2, the average numbers of data points used in kernel regression under patterns $(t, r) = (4,1)$ and $(4,2)$ are 238 and 152, respectively, which may not be enough for kernel regression and lead to some small biases in imputation. On the other hand, linear regression is more stable and works well with a sample size such as 152. Although linear regression imputation has a bias in theory, the bias may be small when $E(y_t \mid y_1, ..., y_{t-1})$ is linear.

3.2 Application to the SIRD

The SIRD is an annual survey of about 31,000 companies potentially involved in research and development. The NSF sponsors this survey as part of a mandate requiring that NSF collect, interpret, and analyze data on scientific and engineering resources in the United States. The survey is conducted jointly by the U.S. Census Bureau and NSF. The surveyed companies are asked to provide information related to their total research and development (RD) expenditure for the calendar year of the survey. The SIRD deterministically surveys some companies each year by placing them in a certainty stratum, since they account for a large percentage of the total RD dollar investment in the U.S. The remaining companies that appear in the survey are sampled each year using a stratified probability proportionate to size (PPS) sampling design. Longitudinal measurements are available on the core of companies that are sampled with certainty and on other companies that happen to be selected each year. For the purposes of illustrating our imputation methods, we restrict attention to only those companies that were selected for the survey in each of the years 2002 through 2005 ($T = 4$), and companies that provided a response in 2002. For documentation on the SIRD and detailed statistical tables, we refer to the document titled *Research and Development in Industry: 2005*, available from http://www.nsf.gov/statistics/nsf10319. Additional information on the Business R&D and Innovation Survey is available online at http://bhs.dev.econ.census.gov/bhs/brdis/ and http://www.nsf.gov/statistics/srvyindustry/about/brdis/.

We divide the data into two imputation classes. One class consists of all companies contained in a certainty stratum for each of the four years; the other consists of the rest of companies. Within each imputation class, the data take the form $(\mathbf{y}_i, \boldsymbol{\delta}_i)$, $i = 1, ..., n$, where y_{it} represents the total RD expenditure for company i at time $t = 1$ (2002), 2 (2003), 3 (2004), 4 (2005). The sample size here is $n = 2{,}309$ for the certainty strata class and $n = 1{,}039$ for the non-certainty strata class. Missingness is nonmonotone and the missing percentages for the years 2003, 2004, and 2005 were 10.4%, 14.0%, and 18.8%, for the certainty strata class, and 15.2%, 20.7%, and 26.0% for the non-certainty strata class.

Table 4 shows the estimated totals and standard errors obtained by using the methods (2)-(9) described in the simulation study in Section 3.1. As discussed in the end of Section 2.1, in each of the proposed imputation methods we use linear regression when $r + 1 = t$. The standard errors shown in Table 4 were computed using the bootstrap method. Table 4 also displays estimated totals obtained when missing data are filled in by the values that were put in place by the Census Bureau in order to produce the officially published data tables (officially published data tables are available from http://www.nsf.gov/statistics/pubseri.cfm?seri_id=26). The method that was used by the Census Bureau to handle missing data when producing these published data tables (which we call the "current method") was ratio imputation for companies with prior year data using imputation cells formed by industry type; we refer to Bond (1994) for further details. Table 4 also presents the estimated RD totals obtained from respondents only with no weight adjustment which indicate that ignoring the missing data leads to biased estimates. Methods (3)-(9) give comparable results, which is likely due to the strong linear dependence in the data so that theoretically biased methods exhibit negligible bias. The estimated totals based on the current method are comparable to those based on the proposed methods for the certainty strata case, but are different in the non-certainty strata case. The method of censoring and linear regression has similar SD to the proposed methods because the number of data points discarded under censoring is not too large. In the certainty strata imputation class only 10% of the sample has an intermittent nonresponse pattern and the percentage of complete cases is 72%. In the non-certainty class, only 9% of the sample has an intermittent nonresponse pattern and the percentage of complete cases is 66%.

3.3 Simulation results based on the SIRD population

An additional simulation study was conducted using a population constructed from the SIRD data. The simulation was run independently for the certainty strata and non-certainty strata imputation classes. To construct the population, we begin with the SIRD data with missing values imputed using the current imputation method for the SIRD. Let $\boldsymbol{\delta}_i$ be the observed response indicator vector for company i and $\tilde{\mathbf{y}}_i$ be the vector of either the observed or imputed values of total RD expenditures for company i over time, $i = 1, ..., n$. For the simulation, we sample from a population based on $\{(\tilde{\mathbf{y}}_i, \boldsymbol{\delta}_i), i = 1, ..., n\}$ as follows. We first draw a sample of size n with replacement from $\tilde{\mathbf{y}}_1, ..., \tilde{\mathbf{y}}_n$, then we add independent normal random noise, with mean 0 and standard deviation 500, to each component

of each of the sampled vectors. Any resulting negative values are set to zero. We denote these simulated RD totals by $\mathbf{y}_1^*, ..., \mathbf{y}_n^*$, where n is the same as that in Section 3.2. We denote the simulated response indicators by $\boldsymbol{\delta}_1^*, ..., \boldsymbol{\delta}_n^*$. For all i and each $t = 2,3,4$, δ_{it}^*'s were binary random variables with

$$P(\delta_{it}^* = 1 \mid y_{i1}^*, ..., y_{i,t-1}^*) = \frac{\exp(\beta_0^{(t)} + \beta_1^{(t)} y_{i,1}^* + \cdots + \beta_{t-1}^{(t)} y_{i,t-1}^*)}{1 + \exp(\beta_0^{(t)} + \beta_1^{(t)} y_{i,1}^* + \cdots + \beta_{t-1}^{(t)} y_{i,t-1}^*)}.$$

The coefficients $\beta_0^{(t)}, \beta_1^{(t)}, ..., \beta_{t-1}^{(t)}$ are fixed throughout the simulation and they were obtained as the estimated coefficients from an initial fit of a logistic regression of δ_{it} on $(\tilde{y}_{i1}, ..., \tilde{y}_{i,t-1})$ for $i = 1, ..., n$.

Table 5 reports the simulation results for total estimators based on 1,000 runs and methods (1)-(9) described in Section 3.1, where the quantities appearing in the table are defined in Section 3.1. To compute the relative bias we obtain the true value of the total through a preliminary run of the simulation model. Several of the conclusions from the normal population simulation of Section 3.1 carry over to this setting. The following is a summary of some additional findings.

1. In contrast to the normal population simulation setting, the estimated total based on censoring and linear regression has SD that is comparable with the proposed imputation methods. This is because the number of data points discarded under censoring is small in this case. The probabilities of an intermittent response pattern are 17% and 19% for the certainty and non-certainty strata classes, respectively. In the normal population simulation these probabilities were nearly 50% as shown in Table 2.
2. All of the proposed imputation methods give relatively similar performance. As noted previously, linear regression imputation is generally biased in theory. However, the bias is small because of the strong linear dependence in data.
3. Method (7) does not have a good performance at $t \geq 3$ for the non-certainty strata case, because the last-value-dependent PSI assumption does not hold.
4. Methods (8) and (9) perform well, again due to the strong linear dependence in data. Although these methods use more observed data in regression imputation, they are comparable with the proposed linear regression method.

Table 4
RD total estimates (in thousands) from SIRD data based on years 2002 to 2005.
Bootstrap standard error (in thousands) in parentheses[1]

Method	Certainty strata			Non-certainty strata		
	$t = 2$	$t = 3$	$t = 4$	$t = 2$	$t = 3$	$t = 4$
Current imputation	154,066	156,754	168,015	2,694	2,790	2,782
	-	-	-	-	-	-
Respondents only with no weight adjustment	149,502	148,300	159,822	2,448	2,553	2,419
	(15,907)	(16,160)	(17,149)	(172)	(193)	(207)
Respondents only with adjusted weights	166,924	172,419	196,815	2,887	3,219	3,269
	(17,728)	(18,720)	(21,045)	(199)	(237)	(273)
Censoring and linear regression imputation	154,824	159,206	172,631	2,843	3,079	3,257
	(15,888)	(16,394)	(17,470)	(189)	(208)	(246)
Proposed kernel regression imputation	154,824	159,394	171,633	2,843	2,997	3,161
	(15,888)	(16,414)	(17,603)	(189)	(199)	(290)
Proposed linear regression imputation	154,824	159,198	172,042	2,843	3,043	3,302
	(15,888)	(16,383)	(17,247)	(189)	(203)	(250)
Proposed 1-dimensional index kernel regression imputation	154,824	159,394	171,494	2,843	2,997	3,254
	(15,888)	(16,414)	(17,268)	(189)	(199)	(248)
Last-value-dependent kernel regression imputation	154,688	158,768	170,606	2,831	2,983	3,177
	(15,900)	(16,286)	(17,234)	(188)	(197)	(240)
Linear regression imputation treating previously imputed values as observed	154,824	159,401	172,600	2,843	3,098	3,257
	(15,888)	(16,390)	(17,306)	(189)	(208)	(236)
Linear regression imputation based on currently and previously observed data	154,824	160,205	172,452	2,843	3,168	3,273
	(15,888)	(16,534)	(17,209)	(189)	(233)	(254)

[1] Disclaimer: The values in Table 4 do not necessarily represent national estimates because we have made some restrictions on the data to fit our framework.

Table 5
Simulation results for total estimation (in thousands) SIRD based population

Method	Quantity	Certainty Strata			Non-Certainty Strata		
		$t = 2$	$t = 3$	$t = 4$	$t = 2$	$t = 3$	$t = 4$
Complete data	relative bias	0%	0.1%	0.1%	0.2%	0.0%	0.4%
	SD	15,541	16,045	16,947	184	203	224
	$\widehat{SD}_{boot}$	15,654	15,994	16,941	186	201	218
	CI coverage	94.0%	94.0%	94.3%	94.3%	93.7%	93.9%
Respondents only with adjusted weights	relative bias	5%	6.3%	11.6%	-1.1%	1.1%	-2.7%
	SD	16,870	17,858	20,032	191	220	244
	$\widehat{SD}_{boot}$	16,917	17,915	20,048	192	219	234
	CI coverage	94.8%	94.8%	87.3%	93.2%	94.5%	89.8%
Censoring and linear regression imputation	relative bias	0%	0.4%	0.5%	0.4%	0.1%	-0.4%
	SD	15,582	16,272	17,247	191	214	238
	$\widehat{SD}_{boot}$	15,654	16,145	17,195	194	214	236
	CI coverage	93.8%	93.5%	94.2%	94.8%	94.0%	93.7%
Proposed kernel regression imputation	relative bias	0%	0.2%	-0.1%	0.4%	-0.3%	-0.3%
	SD	15,582	16,130	17,098	191	205	246
	$\widehat{SD}_{boot}$	15,654	16,072	17,231	194	204	262
	CI coverage	93.8%	93.5%	94.2%	94.8%	93.4%	93.7%
Proposed linear regression imputation	relative bias	0%	0.2%	0.0%	0.4%	0.0%	-0.5%
	SD	15,582	16,130	16,955	191	206	229
	$\widehat{SD}_{boot}$	15,654	16,072	16,964	194	206	224
	CI coverage	93.8%	93.5%	94.2%	94.8%	94.0%	93.7%
Proposed 1-dimensional index kernel regression imputation	relative bias	0%	0.2%	-0.1%	0.4%	-0.3%	-0.9%
	SD	15,582	16,130	16,957	191	205	227
	$\widehat{SD}_{boot}$	15,654	16,072	16,965	194	204	220
	CI coverage	93.8%	93.5%	94.3%	94.8%	93.4%	93.1%
Last-value-dependent kernel regression imputation	relative bias	0%	0.1%	-0.3%	0.0%	-0.7%	-0.7%
	SD	15,565	16,019	16,990	184	204	242
	$\widehat{SD}_{boot}$	15,635	16,003	16,983	187	202	230
	CI coverage	93.8%	93.7%	94.0%	93.9%	92.7%	91.1%
Linear regression imputation treating previously imputed values as observed	relative bias	0%	0.2%	0.0%	0.4%	0.6%	-0.6%
	SD	15,582	16,120	16,952	191	210	231
	$\widehat{SD}_{boot}$	15,654	16,065	16,954	194	210	225
	CI coverage	93.8%	93.6%	94.3%	94.8%	93.8%	92.8%
Linear regression imputation based on currently and previously observed data	relative bias	0%	0.2%	0.0%	0.4%	0.6%	-0.6%
	SD	15,582	16,117	16,945	191	213	241
	$\widehat{SD}_{boot}$	15,654	16,062	16,954	194	211	254
	CI coverage	93.8%	93.5%	94.3%	94.8%	93.6%	93.7%

4. Concluding remarks

We consider a longitudinal study variable having nonmonotone nonresponse. Under the assumption that the PSI depends on past observed or unobserved values of the study variable, we propose several imputation methods that lead to unbiased or nearly unbiased estimators of the total or mean of the study variable at a given time point. Our methods do not require any parametric model on the joint distribution of the variables across time points or the PSI. They are based on regression models under different nonresponse patterns derived from the past-data-dependent PSI. Three regression methods are adopted, linear regression, kernel regression, and one-dimensional index kernel regression. The imputation method based on the kernel type regression is asymptotically valid, but it requires a large number of

observations in each nonresponse pattern. The imputation method based on linear regression is asymptotically biased when the linear relationship does not hold, but it is more stable and, therefore, it may still out-perform methods based on kernel regression.

The method of censoring, which discards all observed data from a subject after its first missing value, may work well when the number of data discarded is small; otherwise it may be very inefficient especially when T is large. For the SIRD data analysis in Sections 3.2-3.3, censoring is comparable with the proposed linear regression imputation method. However, the results are based on four years of data only and censoring may lead to inefficient estimators when more years of data are considered. In applications, it may be a good idea to compare estimators based on censoring with those based on the proposed methods.

Estimators based on the linear regression imputation methods (8) and (9) described in Section 3.1 are asymptotically biased in general. Although they perform well in the simulation study based on the SIRD population, they have poor performance under the simulation setting in Section 3.1, while the proposed linear regression imputation performs well.

The results in Section 2 can be extended to the situation where each sample unit has an observed covariate $\mathbf{x}_t$ at time t without missing values. Assumption (1) may be modified to include covariates:

$$P(\delta_t = 1 \mid \mathbf{y}, \mathbf{X}, \delta_1, ..., \delta_{t-1}, \delta_{t+1}, ..., \delta_T) = P(\delta_t = 1 \mid y_1, ..., y_{t-1}, \mathbf{X}, \delta_1, ..., \delta_{t-1}), \quad t = 2,...,T,$$

where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_T)$. Missing components of $\mathbf{y}_i$ can be imputed using one of the procedures in Sections 2.1-2.3 with $(y_{i1}, ..., y_{ir})$ replaced by $(y_{i1}, ..., y_{ir}, \mathbf{X}_i)$. After all missing values are imputed, we can also estimate the relationship between $\mathbf{y}$ and $\mathbf{X}$ using some popular approaches such as the generalized estimation equation approach. Some details can be found in Xu (2007).

It is implicitly assumed throughout the paper that the y-values are continuous variables with no restriction. When y-values have a particular order or are integer valued, the proposed regression imputation methods are clearly not suitable. New methods for these situations have to be developed.

Acknowledgements

We thank Katherine Jenny Thompson and David L. Kinyon, both of the U.S. Census Bureau, as well as two referees and the associate editor for providing many helpful comments on the paper. The research was partially supported by an NSF grant. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Appendix

Proof of (2) - (3). Let $L(\xi)$ denote the distribution of ξ and $L(\xi \mid \zeta)$ denote the conditional distribution of ξ given ζ. Let $\mathbf{y}_t = (y_1, ..., y_t)$ and $\boldsymbol{\delta}_t = (\delta_1, ..., \delta_t)$. Then, both (2) and (3) follow from $L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_t) = L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-1}) = L(\mathbf{y}_t, \boldsymbol{\delta}_{t-1}) / L(\mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-1}) = [L(\delta_{t-1} \mid \mathbf{y}_t, \boldsymbol{\delta}_{t-2}) / L(\delta_{t-1} \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-2})] L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-2}) = L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-2}) = L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-3}) = \cdots = L(y_t \mid \mathbf{y}_{t-1})$, where the first and third equalities follow from assumption (1).

Proof of (5). Using the same notation as in the proof of (2) and letting $\Delta_r = 1$ be the indicator of $\delta_1 = \cdots = \delta_r = 1$, we have $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = [L(\delta_{r+1} = 0 \mid y_t, \mathbf{y}_r, \Delta_r = 1, \delta_t = 0) / L(\delta_{r+1} = 0 \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)] L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$, which is equal to $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$ by (1). Similarly, we can show that $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) = L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$. Hence, $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 1, \delta_t = 0)$ and result (5) follows.

An example in which (4) does not hold. To show that (4) does not hold in general, we only need to give a counterexample. Consider $T = 3$. Let (y_1, y_2, y_3) be jointly normal with $E(y_t) = 0$, $\text{var}(y_t) = 1$, $t = 1, 2, 3$, $\text{cov}(y_1, y_2) = \text{cov}(y_1, y_3) = \rho$, and $\text{cov}(y_2, y_3) = \rho^2$, where $\rho \neq 0$ is a parameter. Suppose that y_1 is always observed and $P(\delta_t = 0 \mid y_{t-1}) = \Phi(a_{t-1} + b_{t-1} y_{t-1})$, $t = 2, 3$, where a_t and b_t are parameters, Φ is the cumulative distribution function of the standard normal distribution. Then, $E(y_3 \mid y_2, y_1) = \rho y_2$, $E(y_2 \mid y_1) = \rho y_1$, and $E(y_3 \mid y_1) = \rho^2 y_1$. Note that

$$\begin{aligned}
E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) &= E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = 1) \\
&= E(y_3 \mid y_1, \delta_3 = 0) \\
&= \int y_3 L(y_3 \mid y_1, \delta_3 = 0)\, d y_3 \\
&= \int y_3 \int L(y_3 \mid y_1, y_2, \delta_3 = 0) L(y_2 \mid y_1, \delta_3 = 0)\, d y_2 d y_3 \\
&= \iint y_3 L(y_3 \mid y_1, y_2) L(y_2 \mid y_1, \delta_3 = 0)\, d y_2 d y_3 \\
&= \int \left(\int y_3 L(y_3 \mid y_2)\, d y_3 \right) L(y_2 \mid y_1, \delta_3 = 0)\, d y_2 \\
&= \rho \int y_2 L(y_2 \mid y_1, \delta_3 = 0)\, d y_2 \\
&= \frac{\rho \int y_2 P(\delta_3 = 0 \mid y_2) L(y_2 \mid y_1)\, d y_2}{\int P(\delta_3 = 0 \mid y_2) L(y_2 \mid y_1)\, d y_2} \\
&= \frac{\rho \int y_2 \Phi(a_2 + b_2 y_2) L(y_2 \mid y_1)\, d y_2}{\int \Phi(a_2 + b_2 y_2) L(y_2 \mid y_1)\, d y_2},
\end{aligned}$$

where the first equality holds because y_1 is always observed, the second equality holds because under (1), δ_2 and y_3 are independent given y_1. The denominator of the previous expression is equal to

$$h(y_1) = \Phi\left(\frac{a_2 + b_2\rho y_1}{\sqrt{1 + b_2^2(1-\rho^2)}}\right).$$

Using integration by parts, we obtain that

$$\begin{aligned} g(y_1) &= \int (y_2 - \rho y_1)\Phi(a_2 + b_2 y_2) L(y_2 \mid y_1)\, d y_2 \\ &= b_2(1-\rho^2)\int \Phi'(a_2 + b_2 y_2) L(y_2 \mid y_1)\, d y_2 \\ &= \frac{b_2^2(1-\rho^2)}{2\pi\sqrt{1-\rho^2}} \int \exp\left\{-\frac{(a_2 + b_2\rho y_2)^2}{2} - \frac{(y_2 - \rho y_1)^2}{2(1-\rho^2)}\right\} d y_2 \\ &= \frac{b_2(1-\rho^2)}{2\pi[1 + b_2^2(1-\rho^2)]} \exp\left\{-\frac{(a_2 + b_2\rho y_1)^2}{2[1 + b_2^2(1-\rho^2)]}\right\}. \end{aligned}$$

Thus,

$$E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) = \rho^2 y_1 + \rho\frac{g(y_1)}{h(y_1)}. \quad (10)$$

However,

$$\begin{aligned} E(y_3 \mid y_1, \delta_1 = \delta_2 = 1) &= E(y_3 \mid y_1, \delta_1 = 1) \\ &= E(y_3 \mid y_1) = \rho^2 y_1. \end{aligned}$$

This shows that (4) does not hold in this special case.

Proof of (8). Using the notation in the proof of (2)-(3) and writing the $(t-2)$-dimensional vector $(y_1, ..., y_{r-1}, y_{r+1}, ..., y_{t-1})$ as $\mathbf{u}_{t,r}$, we obtain that

$$\begin{aligned} &L(\delta_{r+1} = 1 \mid y_t, z_r, \Delta_r = 1, \delta_t = 0) \\ &\quad = \int L(\delta_{r+1} = 1 \mid y_t, z_r, \mathbf{u}_{t,r}, \Delta_r = 1, \delta_t = 0) \\ &\qquad L(\mathbf{u}_{t,r} \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)\, d\mathbf{u}_{t,r} \\ &\quad = \int L(\delta_{r+1} = 1 \mid y_1, ..., y_r, \Delta_r = 1) \\ &\qquad L(\mathbf{u}_{t,r} \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)\, d\mathbf{u}_{t,r} \\ &\quad = \int L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1) \\ &\qquad L(\mathbf{u}_{t,r} \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)\, d\mathbf{u}_{t,r} \\ &\quad = L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1) \\ &\qquad \int L(\mathbf{u}_{t,r} \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)\, d\mathbf{u}_{t,r} \\ &\quad = L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1), \end{aligned}$$

where the second equality follows from assumption (1) and the fact that there is a one-to-one function between $(z_r, \mathbf{u}_{t,r})$ and $(y_1, ..., y_{t-1})$, and the third equality follows from assumption (7). Similarly, $L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1, \delta_t = 0) = L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1)$ and, hence, $L(\delta_{r+1} = 1 \mid y_t, z_r, \Delta_r = 1, \delta_t = 0) = L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1, \delta_t = 0)$. Then,

$$\begin{aligned} &L(y_t \mid z_r, \Delta_{r+1} = 1, \delta_t = 0) \\ &\quad = \frac{L(y_t, z_r, \Delta_{r+1} = 1, \delta_t = 0)}{L(z_r, \Delta_{r+1} = 1, \delta_t = 0)} \\ &\quad = \frac{L(\delta_{r+1} = 1 \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)}{L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1, \delta_t = 0)} \frac{L(y_t, z_r, \Delta_r = 1, \delta_t = 0)}{L(z_r, \Delta_r = 1, \delta_t = 0)} \\ &\quad = L(y_t \mid z_r, \Delta_r = 1, \delta_t = 0). \end{aligned}$$

Similarly, $L(y_t \mid z_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t \mid z_r, \Delta_r = 1, \delta_t = 0)$. Hence, $L(y_t \mid z_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t \mid z_r, \Delta_{r+1} = 1, \delta_t = 0)$ and result (8) follows.

References

Bond, D. (1994). An evaluation of imputation methods for the Survey of Industrial Research and De velopment. *U.S. Bureau of the Census, Economic Statistical Methods and Programming Division Report Series*. 9404. Washington, DC.

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at rand om. *Journal of the American Statistical Association*, 89, 81-87.

Diggle, P., and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society,* Series C (*Applied Statistics*), 43, 49-93.

Duan, N., and Li, K .C. (1991). Sliced regression: A link-free regression method. *The Annals of Statistics*, 19, 505-530.

Efron, B. (1979). Bootstrap methods: Another look at the jac kknife. *Annals of Statistics*, 7, 1-26.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1, 1-16.

Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Little, R.J., and Rubin, D.B. (2002). Statistical Analysis with Missing Data, second edition. New York: John Wiley & Sons, Inc.

National Science Foundation, Division of Science Resources Statistics (2010). *Research and Development in Industry: 2005*. Detailed Statistical Tables. Available from htt p://www.nsf.gov/statistics/nsf10319/.

Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92, 1320-1329.

R Development Core Team (2009). A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0.

Robins, J.M., and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122-129.

Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Troxel, A.B., Harrington, D.P. and Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, 47, 425-438.

Troxel, A.B., Lipsitz, S.R. and Harrington, D.P. (1998). Marginal models for the analysis of longitudinal measurements with non-ignorable non-monotone missing data. *Biometrika*, 85, 661-672.

Vansteelandt, S., Rotnitzky, A. and R obins, J.M. (2007). Estimation of regression models for the me an of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94, 841-860.

Xu, J. (2007). Methods for intermittent missing responses in longitudinal data. Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.

Xu, J., Shao, J., Palta, M. and Wang, L. (2008). Imputation for nonmonotone last-value-dependent nonrespondents in longitudinal surveys. *Survey Methodology*, 34, 2, 153-162.