A semi-parametric approach to fractional imputation for nonignorable missing data

Jae Kwang Kim * Cindy Long Yu*

Abstract

Parameter estimation with nonignorable missing data is a challenging problem in statistics. Fully parametric approach for joint modeling of the response model and the population model can produce results that are very sensitive against the failure of the assumed model. We consider a more robust approach of modeling by describing the model for the nonresponding part as a exponential tilting of the model for the responding part, which can be justified under the assumption that the response mechanism can be expressed as a logistic regression model. The model for the responding part can be estimated using a nonparametric method. Thus, the overall model can be called semi-parametric.

In this paper, based on the exponential tilting model, we propose a fractional imputation method that can be a useful computational tool for missing data analysis with non-ignorable missing data. To estimate the parameters of the response mechanism, we assume that a validation sample is randomly selected from the nonrespondents and provides full responses. Using the nonparametric model for the respondents, the imputed values are generated from the responding parts and then are applied with fractional weights that have exponential tilting components. The resulting fractionally imputed data can be used to estimate the parameters using the software based on the complete response by treating the imputed values as if observed. Variance estimation using a replication is also considered. Results from a limited simulation study are presented.

Key Words: Follow up; Not missing at random; Survey Sampling.

1. INTRODUCTION

Missing data is frequently encountered in many areas of statistics. Statistical analysis in the presence of missing data has been an area of considerable interest because simply ignoring the missing part of the data often destroys the representativeness of the remaining sample. Non-response is ignorable if the probability of missing y is independent of y conditional on other auxiliary variable x; hence, it follows that non-response is non-ignorable if the probability of y being missing depends on y itself, even after controlling on x. This situation exists, for example, in surveys of income, of alcohol consumption behavior, and in clinical studies of elderly, where cognitively impaired persons may be less willing to participate. If nonresponse is nonignorable, standard nonresponse adjustments such as stratification, reweighting, and imputation assuming an ignorable response mechanism will fail to correct the bias due to nonresponse. In non-ignorable nonresponse problems, it is widely recognized that without additional conditions on the models or additional information (direct information on the nonrespondents, or indirect information relating them to the respondents), estimation is sensitive to the unobserved distribution of the outcome variable (Little 1982; Murnane, Newstead, and Olsen 1985).

Parameter estimation under nonignorable missing is a challenging problem because the response mechanism is generally unknown. When the response mechanism is known, as in the censored regression model with known censoring points, all the parameters are identified and they can be estimated using a maximum likelihood method. When the response mechanism is known up to an unknown parameter ϕ, then the parameters in the observed likelihood are not fully identified in general without additional observation or prior information. Nordheim (1984) showed that if some information of the probabilities of uncertain classification is obtained, then the category is identified under the nonignorable missing data mechanism. Baker and Laird (1988) used the EM algorithm to estimate the maximum likelihood estimators of the expected cell counts under a log-linear model for categorical missing data with nonignorable missing. Glynn, Laird and Rubin (1993) used so-called the pattern mixture model of Little (1993) to analyze nonignorable missing data with a follow-up. Park and Brown (1997) proposed the maximum likelihood estimating method with constraints for categorical data using a data-dependent prior, which amounts to adding additional observation for the missing data. Chen (2001) and Tang et al (2003) discussed identifiability conditions under some situations. When the parameters are not identified, then the maximum likelihood estimates of the parameters are no longer consistent.

In this paper, we propose a new approach for modeling nonignorable nonresponse based on so-called the exponential tilting model. Using the exponential tilting model for the nonresponse part of the data, we decompose the model into two components, one is a parametric component and the other is a nonparametric component. The parametric component is obtained by assuming a logistic regression model for the response probability and the non-parametric component is obtained by a nonparametric regression approach for missing data considered in Cheng (1994). By adopting a nonparametric part of the model, the estimation method can be made robust. To avoid the unnecessary

*Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A.

issue of non-identifiability of the parameters, we assume that a validation sample, a subset of nonrespondents, is randomly selected and observed with full response. For parameter estimation, we use the idea of Monte Carlo approximation by fractional imputation, as considered in Kim (2009). The fractional imputation method can be easily applied to the general purpose estimation. For variance estimation, a replication method is considered.

The paper is organized as follows. In Section 2, basic setup is introduced. In Section 3, we propose a parameter estimation method under the nonignorable missing data mechanism with the follow-up data using the parametric fractional imputation and discuss variance estimation method. In Section 5, results from a limited simulation study are presented.

2. BASIC SETUP

For simplicity, consider two variables, $\mathbf{x}$ and y, where $\mathbf{x}$ is always observed and y is subject to missing. Under the existence of nonresponse, the original sample A can be decomposed into $A = A_R \cup A_M$, where A_R is the set of respondents and A_M is the set of nonrespondents. Let r_i be the original response indicator for y_i, defined by

$$r_i = \begin{cases} 1 & i \in A_R \\ 0 & i \in A_M. \end{cases}$$

We assume that the response mechanism is independent and

$$r_i \mid (x_i, y_i) \sim Bernoulli\left(\pi_i\right) \tag{1}$$

where

$$\pi\left(\mathbf{x}_i, y_i\right) = \frac{\exp\left(\phi_0 + \phi_1 \mathbf{x}_i + \phi_2 y_i\right)}{1 + \exp\left(\phi_0 + \phi_1 \mathbf{x}_i + \phi_2 y_i\right)} \tag{2}$$

for some $\phi = (\phi_0, \phi_1, \phi_2)$. If $\phi_2 = 0$, then the response mechanism is called ignorable.

Under the ignorable response mechanism (or MAR),

$$Pr\left(y_i \in B \mid \mathbf{x}_i, y_i, r_i = 0\right) = Pr\left(y_i \in B \mid \mathbf{x}_i, y_i, r_i = 1\right), \tag{3}$$

for any measurable set B. Thus, under MAR, the conditional distribution of the y_i given $\mathbf{x}_i$ among the nonrespondents is the same as the conditional distribution among the respondents. Let $f_1\left(y_i \mid \mathbf{x}_i\right)$ be the conditional density of y_i given $\mathbf{x}_i$ and $r_i = 1$ and let $f_0\left(y_i \mid \mathbf{x}_i\right)$ be the conditional density of y_i given $\mathbf{x}_i$ and $r_i = 0$. Under MAR, we have

$$f_1\left(y_i \mid \mathbf{x}_i\right) = f_0\left(y_i \mid \mathbf{x}_i\right).$$

If the MAR condition does not hold, then (3) does not hold. Using the Bayes formula, we have the following relationship:

$$\begin{aligned} & Pr\left(y_i \in B \mid \mathbf{x}_i, r_i = 0\right) \\ = \;& Pr\left(y_i \in B \mid \mathbf{x}_i, r_i = 1\right) \times \frac{Pr\left(r_i = 0 \mid \mathbf{x}_i, y_i \in B\right) / Pr\left(r_i = 1 \mid \mathbf{x}_i, y_i \in B\right)}{Pr\left(r_i = 0 \mid \mathbf{x}_i\right) / Pr\left(r_i = 1 \mid \mathbf{x}_i\right)}. \end{aligned} \tag{4}$$

Thus, we can write

$$f_0\left(y_i \mid \mathbf{x}_i\right) = f_1\left(y_i \mid \mathbf{x}_i\right) \times \frac{O\left(\mathbf{x}_i, y_i\right)}{E_1\left[O\left(\mathbf{x}_i, Y_i\right)\right]}, \tag{5}$$

where

$$O\left(\mathbf{x}_i, y_i\right) = \frac{Pr\left(r_i = 0 \mid \mathbf{x}_i, y_i\right)}{Pr\left(r_i = 1 \mid \mathbf{x}_i, y_i\right)}$$

and

$$E_1\left\{O\left(\mathbf{x}_i, Y_i\right)\right\} = E\left\{O\left(\mathbf{x}_i, Y_i\right) \mid \mathbf{x}_i, r_i = 1\right\}.$$

Note that (4) and (5) implies that

$$Pr\left(r_i = 1 \mid \mathbf{x}_i\right) = \frac{E_1\left\{O\left(\mathbf{x}_i, Y_i\right)\right\}}{1 + E_1\left\{O\left(\mathbf{x}_i, Y_i\right)\right\}}. \tag{6}$$

If the response probability model is a logistic regression model (2), then we have

$$O\left(\mathbf{x}_i, y_i\right) = \exp\left(-\phi_0 - \phi_1 \mathbf{x}_i - \phi_2 y_i\right) \tag{7}$$

and the expression (5) can be simplified to

$$f_0(y_i \mid \mathbf{x}_i) = f_1(y_i \mid \mathbf{x}_i) \times \frac{\exp(\gamma y_i)}{E[\exp(\gamma Y_i) \mid \mathbf{x}_i, r_i = 1]}, \tag{8}$$

where $\gamma = -\phi_2$. Model (8) states that the density for the nonrespondents is an exponential tilting of the density for the respondents. The parameter λ is the tilting parameter that determines the amount of departure from the ignorability of the response mechanism.

Thus, we need to know the two models to estimate the parameters: $f_1(y_i \mid \mathbf{x}_i)$ and $Pr(r_i = 1 \mid \mathbf{x}_i, y_i)$. The only parameter that is not identified is γ. To avoid the non-identifiability problem, one can perform a sensitivity analysis as in Rotnitzky et al (1998) or assume a follow-up study in that a further attempt is made to obtain responses in a subset A_V of A_V. Throughout this paper, we assume that there exist an $\sqrt{n}$-consistent estimator $\hat{\gamma}$ of γ such that

$$\sqrt{n}(\hat{\gamma} - \gamma^*) = O_p(1) \tag{9}$$

holds, where $\gamma^* = -\phi_2^*$ and ϕ_2^* is the true value of ϕ_2 in (2). In the sensitivity analysis, we assume that $\hat{\gamma}$ is given.

The initial estimator $\hat{\gamma}$ satisfying (9) can be obtained from a follow-up study. A method of obtaining the initial estimator satisfying (9) shall be discussed later.

3. FRACTIONAL IMPUTATION

Under the setup describe in Section 2, suppose that we have a consistent estimate of $f_1(y_i \mid \mathbf{x}_i)$, denoted by $\hat{f}_1(y_i \mid \mathbf{x}_i)$. The estimate of the conditional density is obtained by a non-parametric method. Following the idea of Kim (2009), let M imputed values, $y_i^{*(1)}, \cdots, y_i^{*(M)}$, are generated from $h(y)$, which has the same support as $f_1(y_i \mid \mathbf{x}_i)$. The fractional weights are constructed by

$$w_{ij1}^* \propto w_{ij0}^* \exp\left(\hat{\gamma} y_i^{*(j)}\right) \tag{10}$$

and

$$w_{ij0}^* \propto \frac{\hat{f}_1\left(y_i^{*(j)} \mid \mathbf{x}_i\right)}{h\left(y_i^{*(j)}\right)} \tag{11}$$

where $\hat{\gamma}$ is an $\sqrt{n}$-consistent estimator $\hat{\gamma}$ of γ in (9) and $\sum_{j=1}^M w_{ij0}^* = \sum_{j=1}^M w_{ij1}^* = 1$. The initial fractional weight w_{ij0}^* in (11) is essentially the fractional weights that can be used under ignorable missing mechanism, as in Kim (2009), and the second factor, $\exp\left(\hat{\gamma} y_i^{*(j)}\right)$, can be used to account for the non-ignorable missing mechanism. The fractional weights in (10) are computed in a semi-parametric approach in the sense that we use a nonparametric model for $f_1(y_i \mid \mathbf{x}_i)$ but use a parametric model (2) for the response mechanism.

Once the final fractional weights are constructed, then we can use the fractionally imputed data to estimate the parameters by applying the standard formula for parameter estimation to the fractionally imputed data. For example, the imputed estimator of $\boldsymbol{\beta}$, the regression coefficient for the regression of y on $\mathbf{x}$, can be computed from

$$\sum_{i \in A} \sum_{j=1}^M w_i w_{ij}^* \left\{ y_i^{*(j)} - \mathbf{x}_i' \boldsymbol{\beta} \right\} \mathbf{x}_i = \mathbf{0}. \tag{12}$$

Here, it should be understood that $y_i^{*(j)} = y_i$ if y_i is observed.

More generally, under complete response, let the solution to

$$\sum_{i \in A} w_i U(\mathbf{x}_i, y_i; \theta) = \mathbf{0} \tag{13}$$

lead to a consistent estimator of θ_0. Under some regularity conditions, the solution to the fractionally imputed estimating equation

$$\sum_{i \in A} \sum_{j=1}^M w_i w_{ij}^* U\left(\mathbf{x}_i, y_i^{*(j)}; \theta\right) = \mathbf{0} \tag{14}$$

is consistent and is asymptotically distributed as normal with mean θ_0 and

We now discuss the estimation of the $\sqrt{n}$-consistent estimator $\hat{\lambda}$ satisfying (9). We consider the case when a validation sample, A_V, is randomly selected and the responses are obtained all the elements in A_V. In this case, we can use the idea of Horvitz-Thompson estimator to obtain the following weighted score equation for ϕ:

$$\sum_{i\in A_R} \{r_i - g(\mathbf{x}_i, y_i; \phi)\} (\mathbf{x}_i', y_i)' + \frac{n_M}{n_V} \sum_{i\in A_V} \{r_i - g(\mathbf{x}_i, y_i; \phi)\} (\mathbf{x}_i', y_i)' = \mathbf{0}', \tag{15}$$

where n_M is the size of set A_M and n_V is the size of set A_V. The solution to (15) is consistent because the selection probability for the elements in A_V is n_M/n_V.

Instead of (15), one can apply an unweighted score equation

$$\sum_{i\in A_R} w_i \{r_i - g(\mathbf{x}_i, y_i; \phi)\} (\mathbf{x}_i', y_i)' + \sum_{i\in A_V} w_i \{r_i - g(\mathbf{x}_i, y_i; \phi)\} (\mathbf{x}_i', y_i)' = \mathbf{0}', \tag{16}$$

to get $\hat{\phi} = \left(\hat{\phi}_0, \hat{\phi}_1, \hat{\phi}_2\right)$. It can be shown that the choice of $\hat{\gamma} = -\hat{\phi}_2$ leads to the maximum likelihood estimator.

4. VARIANCE ESTIMATION

For variance estimation of the fractionally imputed estimator, we consider a replication method. Under complete response, let $\hat{\theta}_n$ be the solution to (13). To estimate the variance of $\hat{\theta}_n$, replication method is commonly used. Let $w_i^{(k)}$ be the k-th replication weight of w_i such that

$$\sum_{k=1}^{L} c_k \left\{\hat{\theta}_n^{(k)} - \hat{\theta}_n\right\}^2$$

consistently estimates the variance of $\hat{\theta}_n$, where $\hat{\theta}_n^{(k)}$ is the solution to

$$\sum_{i\in A} w_i^{(k)} U(\mathbf{x}_i, y_i; \theta) = \mathbf{0}. \tag{17}$$

Now, to estimate the variance of $\hat{\theta}_{FI}$ that is the solution to (14), we need to compute the replicated fractional weights $w_{ij}^{*(k)}$. To compute the replication fractional weights, we also need two steps. In the first step, the replicates for the initial fractional weights (10) are computed by

$$w_{ij0}^{*(k)} \propto \frac{\hat{f}_1^{(k)}\left(y_i^{*(j)} \mid \mathbf{x}_i\right)}{h\left(y_i^{*(j)}\right)} \exp\left(\hat{\gamma}^{(k)} y_i^{*(j)}\right) \tag{18}$$

where $\hat{f}_1^{(k)}\left(y_i^{*(j)} \mid \mathbf{x}_i\right)$ is the replicated version of the nonparametric estimator of $f\left(y_i^{*(j)} \mid \mathbf{x}_i\right)$ and $\hat{\gamma}^{(k)}$ is the replicated values of $\hat{\gamma}$, computed from

$$\sum_{i\in A_R\cup A_V} w_i^{(k)} \{r_i - g(\mathbf{x}_i, y_i; \phi)\} (\mathbf{x}_i', y_i)' = \mathbf{0}', \tag{19}$$

to get $\hat{\phi}^{(k)}$ and $\hat{\gamma}^{(k)} = -\hat{\phi}_2^{(k)}$. Computing for the solution to (19) requires an iterative method, which can be cumbersome because we have to solve (19) for each replicate k. To reduce the computational burden, we can consider one-step approximation of $\hat{\phi}^{(k)}$ by

$$\hat{\phi}^{(k)} \cong \hat{\phi} + \left\{\sum_{A_R\cup A_V} w_i^{(k)} \hat{g}_i (1 - \hat{g}_i) (\mathbf{x}_i', y_i)' (\mathbf{x}_i', y_i)\right\}^{-1} \sum_{i\in A_R\cup A_V} w_i^{(k)} \{r_i - \hat{g}_i\} (\mathbf{x}_i', y_i)' \tag{20}$$

where $\hat{g}_i = g\left(\mathbf{x}_i, y_i; \hat{\phi}\right)$.

Using the replicated fractional weights in (18), we can compute the replicate of $\hat{\theta}_I$ which is the solution to (14), denoted by $\hat{\theta}_I^{(k)}$, as the solution to

$$\sum_{i\in A} \sum_{j=1}^{M} w_i^{(k)} w_{ij}^{*(k)} U\left(\mathbf{x}_i, y_i^{*(j)}; \theta\right) = \mathbf{0}. \tag{21}$$

5. Simulation Study

In this section, simulation studies were conducted to examine if the proposed semi-parametric fractional imputation (Semi-FI) method can effectively identify the unknown parameters of interest when models are either correctly or wrongly specified.

In the first simulation, B=2,000 Monte Carlo samples each of size $n = 200$ were generated from a distribution

$$x_i \sim N(4, 1)$$

$$y_i = 0.5 + x_i + N(0, 1) \tag{22}$$

and the response indicator variable r_i for original missing is distributed as

$$r_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1 - \pi_i \end{cases}$$

where $logit(\pi_i) = -5.8 + x_i + 0.5y_i$ and $i = 1, ..., n$. Under this model setup, the average response rate for original response model is about 58%. We also assume that 26% of the nonrespondents are followed up to get the full response. Thus, the actual response rate is 69%. We are interested in estimating the following parameters.

1. β_1: the slope for the linear regression of y on x.
2. μ_y: the marginal mean of y.
3. $Pr(y < 5)$: the proportion of y less than 5.

To estimate each parameter of interest, the following three approaches were adopted in the simulation study.

1. Using complete sample: Under the complete case, the marginal mean and the proportion were simply estimated by $n^{-1}\sum_{i=1}^{n} y_i$ and $n^{-1}\sum_{i=1}^{n} I(y_i < 5)$ respectively. Simple linear regression was used to identify the slope β_1.
2. Using Monte Carlo EM algorithm (MCEM): The MCEM method was conducted under the parametric assumption that $f(y_i|x_i)$ is a normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance σ^2. For each EM iteration, $y_i^{*(j)}$ were simulated from $N(\hat{\beta}_0(t) + \hat{\beta}_1(t)x_i, \hat{\sigma}(t)^2)$ where $\hat{\beta}_0(t), \hat{\beta}_1(t)$ and $\hat{\sigma}(t)$ were the parameters estimated from previous iteration t.
3. Using semi-parametric fractional imputation. Under the Semi-FI method, $y_i^{*(j)}$ were generated from a normal distribution with mean $\alpha_0 + \alpha_1 x_i$ and variance η^2, where α_0, α_1 and η were obtained using simple linear regression on the respondent data only. The imputation needs to be done just one time and there is no iteration involved in the estimation.

We used $M = 100$ for the MCEM method and $M = 10$ for the proposed Semi-FI method. Table 1 reports the Monte Carlo mean and standard deviation of the estimates, as well as the root mean squared errors of the three parameters under each approach. The MCEM method estimate all three parameters very accurately because the parametric model is correctly assumed in this case. The proposed Semi-FI method with $M = 10$ can work also effectively in the sense that all the "true" parameters [1] can be captured within one standard deviation of the Semi-FI estimators. Not surprisingly, the Semi-FI method shows less efficiency than the MCEM method which is considered as the best under correct parametric model assumption for a given amount of missing data. The relative less efficiency of the Semi-FI estimator is the price to be paid as the parametric inference "knows" more about the model than the non-parametric approach which does not assume any parametric model.

In order to further verify the validity of the Semi-FI method when models are misspecified, two more simulations were formulated. Instead of model (22), we consider the following two models,

$$y_i = x_i(x_i - 2)(x_i - 3) + N(0, 1) \tag{23}$$

and

$$y_i = x_i(x_i - 2)(x_i - 3) + SN(0, 1, \xi), \tag{24}$$

where $SN(0, 1, \xi)$ is a skewed normal distribution with mean 0, variance 1, and skewness index $\xi = -4$. Model (23) allows cubic curvature in the regression mean function with symmetric error, while model (24) imposes nonlinearity in the mean function contaminated with negatively skewed noises.

[1] We consider the estimates from the full responses as the true values.

Table 1: Point Estimators from Model (22). The mean and SE are the Monte Carlo average and standard deviation of the point estimates, and the RMSE is the root mean squared errors.

	Method	Mean	SE	RMSE
β_1	Complete Data	1.0000	0.0716	0.0716
	MCEM (M=100)	1.0045	0.0973	0.0974
	Semi-FI (M=10)	0.9540	0.1024	0.1122
μ_y	Complete Data	4.5031	0.1005	0.1005
	MCEM (M=100)	4.5063	0.1266	0.1266
	Semi-FI (M=10)	4.5332	0.1266	0.1308
$Pr(y < 5)$	Complete Data	0.6374	0.0341	0.0341
	MCEM (M=100)	0.6364	0.0371	0.0371
	Semi-FI (M=10)	0.6270	0.0393	0.0407

Table 2: Point Estimators from Model (22). The mean and SE are the Monte Carlo average and standard deviation of the point estimates, and the RMSE is the root mean squared errors.

	Method	Mean	SE	RMSE
β_1	Complete Data	1.0000	0.0716	0.0716
	MCEM (M=100)	1.0045	0.0973	0.0974
	Semi-FI (M=10)	0.9540	0.1024	0.1122
μ_y	Complete Data	4.5031	0.1005	0.1005
	MCEM (M=100)	4.5063	0.1266	0.1266
	Semi-FI (M=10)	4.5332	0.1266	0.1308
$Pr(y < 5)$	Complete Data	0.6374	0.0341	0.0341
	MCEM (M=100)	0.6364	0.0371	0.0371
	Semi-FI (M=10)	0.6270	0.0393	0.0407

Table 2 and 3 present the Monte Carlo mean, standard deviation and the root mean squared error of all three parameters for the second and third simulation. Both tables show that, for parameters μ_y and β_1, the Semi-FI estimator outperforms the MCEM estimator a lot, i.e. not only it produces much less biased estimates, but also is more efficient. For parameter $Pr(y < 5)$, although the Semi-FI method gives estimates that are more closer to the true on average, its standard error is higher than that of the MCEM, due to the randomness from the non-parametric part in the estimation. When models are misspecified, the bad performance of the MCEM estimator is a price paid to borrow wrong parametric information. Although $y_i^{*(j)}$ in the Semi-FI method are also generated from a wrong model, the weights associated with $y_i^{*(j)}$ will be adjusted using the non-parametric component $\hat{f}_1(y|x)$, which contains the information reflected by the data. What is definitely working in such situation is the idea of allowing data to speak for itself (or using data as its own "model").

For variance estimation, the delete-one Jackknife variance estimator discussed in Section 3 was used to evaluate the quality of the proposed variance estimator. We simulated $B = 2,000$ Monte Carlo samples from model (22) and considered the Semi-FI method only. Table 4 reports the Monte Carlo variance of the point estimators (which is called "true variance" in the table), the Monte Carlo average, and the relative bias of the variance estimators. Relative bias of the variance estimators were computed by dividing the Monte Carlo bias of the variance estimators by Monte Carlo variance of the point estimator. The proposed variance estimator for the Semi-FI are nearly unbiased for all parameters with all relative biases below 3%.

References

Baker, S. G. and N. M. Laird (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association* **83**, 62-69.

Chen, K. (2001). Parametric models for response-biased sampling. *Journal of the Royal Statistical Society, Series B* **63**, 775-789.

Table 3: Point Estimators from Model (23). The mean and SE are the Monte Carlo average and standard deviation of the point estimates, and the RMSE is the root mean squared errors.

	Method	Mean	SE	RMSE
β_1	Complete Data	16.9991	1.6371	1.6367
	MCEM (M=100)	21.6637	2.5371	5.3097
	Semi-FI (M=10)	16.8873	1.6888	1.6920
μ_y	Complete Data	15.0023	1.4190	1.4186
	MCEM (M=100)	12.0002	1.7085	3.4540
	Semi-FI (M=10)	14.8929	1.5175	1.5210
$Pr(y < 5)$	Complete Data	0.4021	0.0352	0.0352
	MCEM (M=100)	0.3802	0.0370	0.0430
	Semi-FI (M=10)	0.4056	0.0445	0.0447

Table 4: Variance Estimators from Model (22). The "true" variance is the Monte Carlo variance of the point estimators over $B = 2,000$ samples, the mean of $\hat{V}$ is the Monte Carlo average of the variance estimators, and the relative bias were computed by dividing the Monte Carlo bias of the variance estimators by Monte Carlo variance of the point estimator.

	Method	"True" variance	Mean of $\hat{V}$	Relative Bias (%)
β_1	Semi-FI (M=10)	0.0106	0.0108	2.33
μ_y	Semi-FI (M=10)	0.0162	0.0166	2.67
$Pr(y < 5)$	Semi-FI (M=10)	0.0016	0.0016	-2.78

Cheng, P.E. (1994). Nonparametric Estimation of Mean Functionals with Data Missing at Random. *Journal of the American Statistical Association* **89**, 81-87.

Glynn, R., Laird, N. M. and Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association* **88**, 984-993.

Greenlees, W.S., Reece, J.S., and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed, *Journal of the American Statistical Association* **77**, 251-261.

Kim, J.K. (2009). Parametric fractional imputation for missing data analysis. *Unpublished Manuscript.*

Little, R. J. A. (1982). Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association* **77**, 237-250.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125-134.

Murnane, R. J., Newstead, S. and Olsen, R. J. (1985). Comparing public and private schools: the puzzling role of selecting bias. *Journal of Business and Economic Statistics* **3**, 23-35.

Nordheim, E. V. (1984). Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's Syndrome. *Journal of the American Statistical Association* **79**, 772-780.

Park, T. and M. B. Brown (1997) Log-linear models for a binary response with nonignorable nonresponse. *Computational Statistics and Data Analysis* **24**, 417-432.

Rotnitzky, A., Robins, J. and Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with nonignorable non-response. *Journal of the American Statistical Association* **93**, 1321-1339.

Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley and Sons.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

Tang, G., Little, R.J.A., and Raghunathan, T.E. (2003). Analysis of multivaraite missing data with nonignorable nonresponse. *Biometrika* **90**, 747-764.

Appendix

A.1: Description of the kernel estimator

We used the kernel smoothing method to estimate $f_1(y|x)$. The kernel method is an instrument for constructing nonparametric curve estimates. It allows the data to speak for themselves, therefore more objective than parametric methods.

Let $K(\cdot)$ be a kernel function which is a symmetric probability density function, h_x (or h_y) is a smoothing bandwidth for variable x (or y) such that $h_x \to 0$, $h_y \to 0$, and $nh_xh_y \to \infty$ as $n \to \infty$. The kernel estimator of $f_1(y|x)$ is

$$\widehat{f}_1(y|x) = \frac{1}{h_y}\left\{\sum_{j\in A_R} K\left(\frac{x-x_j}{h_x}\right)\right\}^{-1} \sum_{j\in A_R} K\left(\frac{y-y_j}{h_y}\right) K\left(\frac{x-x_j}{h_x}\right) \tag{25}$$

where A_R is the set of observations when the y variable is not missing. Equation (25) can also be written as

$$\hat{f}_1(y|x) = \frac{\hat{f}(x,y)}{\hat{f}(x)},$$

where $\hat{f}(x,y) = (n_R h_x h_y)^{-1} \sum\limits_{j\in A_R} K\left(\frac{y-y_j}{h_y}\right) K\left(\frac{x-x_j}{h_x}\right)$ and $\hat{f}(x) = (n_R h_x)^{-1} \sum\limits_{j\in A_R} K\left(\frac{x-x_j}{h_x}\right)$.

Choices of the bandwidth will be discussed later. There is very little to choose between the various kernels on the basis of mean integrated square error (MISE) as the function $K(\cdot)$ plays a lesser role than the bandwidth h in determining the performance of the kernel estimators. The choice of the kernel function is always based on other considerations, such as differentiability and computation effort. If the true curve has bounded support, the kernel estimator will suffer boundary biases and the choice of $K(\cdot)$ becomes important. In our simulation study which does not involve any bounded support, we picked Gaussian kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \tag{26}$$

for smoothing throughout the paper.

The reference to normal distributions (Scott 1992) was used to guide the selection of the bandwidth. The reference to normal distributions methods assumes the data comes from a distribution "close" to normal distributions. If the Gaussian kernel is used, the optimal bandwidth is

$$h^* = 1.06\hat{\sigma} n^{-1/5}, \tag{27}$$

where $\hat{\sigma} = min(s, 1.35^{-1}\text{interquartile})$ is the robust estimator of σ, and s is the sample standard deviation of the data. This "subjective" choice works effectively if the assumed normal distribution is close to the real one. We used this method to choose bandwidths for both variable x and y in our simulation studies.

Under our simulation set up, an alternative bandwidth choosing method, the Cross-Validation (Silverman 1986) was also attempted. The bandwidths prescribed by both methods were similar. To save computation time, we applied the reference to normal distributions method to pick the suitable bandwidths.