Proxy Pattern-Mixture Analysis for Survey Nonresponse

Rebecca R. Andridge[*] Roderick J. A. Little[*]

Abstract

We consider assessment of nonresponse bias for the mean of a survey variable Y subject to nonresponse. We assume that there are a set of covariates observed for nonrespondents and respondents. To reduce dimensionality and for simplicity we reduce the covariates to a proxy variable X that has the highest correlation with Y, estimated from a regression analysis of respondent data. We consider adjusted estimators of the mean of Y that are maximum likelihood for a pattern-mixture model with different mean and covariance matrix of Y and X for respondents and nonrespondents, assuming missingness is an arbitrary function of a known linear combination of X and Y. We propose a taxonomy for the evidence concerning bias based on the strength of the proxy and the deviation of the mean of X for respondents from its overall mean, propose a sensitivity analysis, and describe Bayesian versions of this approach. Methods are demonstrated through data from the third National Health and Nutrition Examination Survey (NHANES III).

Key Words: Bayesian methods, Missing data, Nonignorable nonresponse, Nonresponse bias analysis, Survey data

1. Introduction

Missing data are often a problem in large-scale surveys, arising when a sampled unit does not respond to the entire survey (unit nonresponse) or to a particular question (item nonresponse). In this paper we focus on the adjustment for and measurement of nonresponse bias in a single variable Y subject to missing values, when a set of variables Z are measured for both respondents and nonrespondents. With unit nonresponse this set of variables is generally restricted to survey design variables, except in longitudinal surveys where variables are measured prior to dropout. With item nonresponse, the set of observed variables can include survey items not subject to nonresponse, and hence is potentially more extensive. With a set of variables Y subject to nonresponse, our methods could be applied separately for each variable, but we do not consider here methods for multivariate missing data where variables are missing for different sets of cases.

Limiting the impact of nonresponse is an important design goal in survey research, and how to measure and adjust for nonresponse is an important issue for statistical agencies and other data collectors, particularly since response rates are on the decline. Current U.S. federal standards for statistical surveys state, "Nonresponse bias analyses must be conducted when unit or item response rates or other factors suggest the potential for bias to occur," (Office of Management and Budget, 2006, p. 8) and go on to suggest that unit nonresponse rates of less than 80% require such an analysis. However, specific analysis recommendations are lacking, focusing on methods for accurately calculating response rates. While the response rate is clearly an important feature of the problem, there is a tension between increasing response rates and increasing response error by including respondents with no inclination to respond accurately. Indeed, some studies have shown that response rates are a poor measure of nonresponse bias (e.g. Curtain, Presser, and Singer, 2000; Keeter, Miller, Kohut, Groves, and Presser, 2000).

There are three major components to consider in evaluating nonresponse: the amount of missingness, differences between respondents and nonrespondents on characteristics that are observed for the entire sample, and the relationship between these fully observed covariates and the survey outcome of interest. Each facet provides some information about the impact of nonresponse, but no single component completely tells the story. Historically the amount of missingness, as measured by the response rate, has been the most oft-used metric for evaluating survey quality. However, response rates ignore the information contained in auxiliary covariates. Federal reports have recommended the second component, evaluating nonresponse based on differences between respondents and nonrespondents (Federal Committee on Statistical Methodology, 2001). A related approach is to focus on measures based on the response propensity, the estimated probability of response given the covariates, which is the auxiliary variable that is most different between respondents and nonrespondents. Measures such as the variability of nonresponse weights indicate the potential of weighting for nonresponse bias reduction, and lack of variability can suggest missingness is completely at random. Though response propensity analyses are appealing, nonresponse bias depends on the strength of the correlation between the survey variable of interest and the probability of response, and bias will vary across items in a single survey (Bethlehem, 2002; Groves, 2006).

The final component is the value of the auxiliary information in predicting survey outcomes. Suppose Y is a survey outcome subject to nonresponse, X is an auxiliary variable observed for respondents and nonrespondents,

[*] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, 48109

and missing values of Y are imputed by predictions of the regression of Y on X estimated using the respondent sample. If data are missing completely at random, the variance of the mean of Y based on the imputed data under simple random sampling is asymptotically

$$Var(\widehat{\mu}_y) = \frac{\sigma_{yy}}{r}\left(1 - \frac{n-r}{n}\rho^2\right),$$

where n is the sample size, r is the number of respondents, σ_{yy} is the variance of Y, and ρ is the correlation between X and Y (see Little and Rubin, 2002, equation 7.14). The corresponding fraction of missing information – the loss of precision from the missing data – is

$$FMI = \frac{n/\sigma_{yy} - Var^{-1}(\hat{\mu}_y)}{n/\sigma_{yy}}.$$

This fraction varies from the nonresponse rate $(n-r)/n$ when $\rho^2 = 0$ to 0 when $\rho^2 = 1$. With a set of covariates Z, imputation based on the multiple regression of Y on Z yields similar measures, with ρ^2 replaced by the squared coefficient of determination of the regression of Y on Z. This approach is attractive since it gives appropriate credit to the availability of good predictors of Y in the auxiliary data as well as a high response rate, and arguably good prediction of the survey outcomes is a key feature of good covariates; in particular, conditioning on a covariate Z that is a good predictor of nonresponse but is unrelated to survey outcomes simply results in increased variance without any reduction in bias (Little and Vartivarian, 2005). A serious limitation with this approach is that it is more focused on precision than bias, and it assumes the data are missing at random (MAR); that is, missingness of Y is independent of Y after conditioning on the covariates Z (Rubin, 1976). Also, this approach cannot provide a single measure of the impact of nonresponse, since by definition measures are outcome-specific.

Previous work has focused on distinct measures based on these considerations, but has not integrated them in a satisfactory way. We propose a new method for nonresponse bias measurement and adjustment that takes into account all three aspects, in a way which we find intuitive and satisfying. In particular, it gives appropriate credit for predictive auxiliary data, without making the MAR assumption, which is implicit in existing propensity and prediction methods; our methods are based on a pattern-mixture model (Little, 1993) for the survey outcome that allows missingness to be not at random (NMAR) and assesses the sensitivity of estimates to deviations from MAR. We prefer a sensitivity analysis approach over approaches that require strong distributional and other assumptions on the missingness mechanism for estimation such as the selection models arising from the work of Heckman (1976). For more discussion of this point see for example Little and Rubin (2002, chap. 15) and citations therein.

2. General Framework

We consider the problem of assessing nonresponse bias for estimating the mean of a survey variable Y subject to nonresponse. For simplicity, we initially consider an infinite population with a sample of size n drawn by simple random sampling. Let Y_i denote the value of a continuous survey outcome and $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$ denote the values of p covariates for unit i in the sample. Only r of the n sampled units respond, so observed data consist of (Y_i, Z_i) for $i = 1, \ldots, r$ and Z_i for $i = r+1, \ldots, n$. In particular this can occur with unit nonresponse, where the covariates Z are design variables known for the entire sample or with item nonresponse. Of primary interest is assessing and correcting nonresponse bias for the mean of Y.

For simplicity and to reduce dimensionality, we replace Z by a single proxy variable X that has the highest correlation with Y. This proxy variable can be estimated by regressing Y on Z using the respondent data, including important predictors of Y , as well as interactions and nonlinear terms where appropriate. The regression coefficients are subject to sampling error, so in practice X is estimated rather than known, but we address this complication later. Let ρ be the correlation of Y and X, which we assume is positive. If ρ is high (say, 0.8) we call X a strong proxy for Y and if X is low (say, 0.2) we call X a weak proxy for Y. The distribution of X for respondents and nonrespondents provides the main source of information for assessing nonresponse bias for Y.

Let $\bar{y}_R$ denote the respondent mean of Y, which is subject to nonresponse bias. We consider adjusted estimators of the mean μ_y of Y of the form

$$\hat{\mu}_y = \bar{y}_R + g(\hat{\rho})\sqrt{\frac{s_{yy}}{s_{xx}}}(\bar{x} - \bar{x}_R), \tag{1}$$

where $\bar{x}_R$ is the respondent mean of X, $\bar{x}$ is the sample mean of X, and s_{xx} and s_{yy} are the respondent sample variances of X and Y. Note that since the proxy X is the conditional mean of Y given X it will have lower variance than Y.

Some comments on the estimator (1) follow. The classical regression estimator is obtained when $g(\hat{\rho}) = \hat{\rho}$, and this is an appropriate choice when missingness depends on the proxy X. It is also appropriate more generally when the data are missing at random (MAR), that is, missingness depends on Z, if $Y|Z$ is normal, and models are well specified. This is true because under MAR, the partial association between the residual $Y-X$ and the missing data indicator (say M) is zero.

In general, we may want the weight $g(\hat{\rho})$ given to the standardized proxy data to increase with the strength of the proxy, and $g(\hat{\rho}) \to 1$ as $\hat{\rho} \to 1$, that is, as the proxy variable converges towards the true variable Y. The size of the deviation, $d = \bar{x} - \bar{x}_R$, and its standardized version, $d^* = d/\sqrt{s_{xx}}$, is a measure of the deviation from missing completely at random (MCAR), and as such is the "observable" component of nonresponse bias for Y. Specific choices of $g(\hat{\rho})$ based on a pattern-mixture model are presented in the next section.

The information about nonresponse bias for Y depends on the strength of the proxy, as measured by $\hat{\rho}$, and the deviation from MCAR, as measured by the size of d. We consider four situations, ordered from what we consider most favorable to least favorable from the point of view of the quality of this information for nonresponse bias assessment and adjustment.

1. If X is a strong proxy (large $\hat{\rho}$), and d is small, then the adjustment via (1) is small and the evidence of a lack of nonresponse bias in Y is relatively strong, since it is not evident in a variable highly correlated with Y. This is the most favorable case.

2. If X is a strong proxy, and d is large, then there is strong evidence of response bias in respondent mean $\bar{y}_R$ but good information for correcting the bias using the proxy variable via (1). Since an adjustment is needed, model misspecification is a potential issue.

3. If X is a weak proxy (small $\hat{\rho}$), and d is small, then the adjustment via (1) is small. There is some evidence against nonresponse bias in the fact that d is small, but this evidence is relatively weak since it does not address the possibility of bias from unobserved variables related to Y.

4. If X is a weak proxy, and d is large, then the adjustment via (1) depends on the choice of $g(\hat{\rho})$, although it is small under the MAR assumption when $g(\hat{\rho}) = \hat{\rho}$. There is some evidence that there is nonresponse bias in Z in the fact that d is large, but no evidence that there is bias in Y since Z is only weakly related to Y. The evidence against bias in Y is however relatively weak since there may be bias from other unobserved variables related to Y. This is the least favorable situation.

In the next section we consider specific choices of $g(\hat{\rho})$ based on a pattern-mixture model analysis that reflects this hierarchy.

3. The Pattern-Mixture Model

Let M denote the missingness indicator, such that $M = 0$ if Y is observed and $M = 1$ if Y is missing. We assume $E(Y|Z, M = 0) = \alpha_0 + \alpha Z$, and let $X = \alpha Z$. For simplicity we assume in this section that α is known, that is, we ignore estimation error in α. We focus on the joint distribution of $[Y, X, M]$ which we assume follows the bivariate pattern-mixture model discussed in Little (1994). This model can be written as follows:

$$\begin{aligned} (Y, X|M = m) &\sim N_2\left((\mu_y^{(m)}, \mu_x^{(m)}), \Sigma^{(m)}\right) \\ M &\sim Bernoulli(1-\pi) \\ \Sigma^{(m)} &= \begin{bmatrix} \sigma_{yy}^{(m)} & \rho^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}} \\ \rho^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}} & \sigma_{xx}^{(m)} \end{bmatrix} \end{aligned} \tag{2}$$

where N_2 denotes the bivariate normal distribution. Of primary interest is the marginal mean of Y, which can be expressed as $\mu_y = \pi\mu_y^{(0)} + (1-\pi)\mu_y^{(1)}$. This model is underidentified, since there is no information on the conditional normal distribution for Y given X for nonrespondents ($M = 1$). However, Little (1994) shows that the model can be identified by making assumptions about how missingness of Y depends on Y and X. Specifically if we assume that

$$\Pr(M = 1|Y, X) = f(X + \lambda^* Y), \tag{3}$$

for some unspecified function f and known constant λ^*, the parameters are just identified by the condition that

$$((Y, X) \perp M | f(X + \lambda^* Y)) \tag{4}$$

where $\perp$ denotes independence. The resulting ML estimate of the mean of Y averaging over patterns is

$$\hat{\mu}_y = \bar{y}_R + \frac{s_{xy} + \lambda^* s_{yy}}{s_{xx} + \lambda^* s_{xy}}(\bar{x} - \bar{x}_R), \tag{5}$$

where s_{xx}, s_{xy} and s_{yy} are the sample variance of X, the sample covariance of X and Y, and the sample variance of Y for respondents (Little, 1994).

We apply a slight modification of this model in our setting, rescaling the proxy variable X to have the same variance as Y, since we feel this enhances the interpretability of the model (3) for the mechanism. Specifically we replace (3) by

$$\Pr(M = 1|Y, X) = f(X\sqrt{\frac{\sigma_{yy}^{(0)}}{\sigma_{xx}^{(0)}}} + \lambda Y) = f(X^* + \lambda Y), \tag{6}$$

where X^* is the proxy variable X scaled to have the same variance as Y in the respondent population, and $\lambda = \lambda^* \sqrt{\sigma_{xx}^{(0)}/\sigma_{yy}^{(0)}}$. The parameters are just identified by the condition that

$$((Y, X) \perp M | f(X^* + \lambda Y)) \tag{7}$$

where $\perp$ denotes independence. We call the model defined by Equations (2) and (6) a proxy pattern-mixture (PPM) model. By a slight modification of the arguments in (Little, 1994), the resulting maximum likelihood estimate of the overall mean of Y has the form of Equation (1) where

$$g(\hat{\rho}) = \frac{\lambda + \hat{\rho}}{\lambda\hat{\rho} + 1}, \tag{8}$$

and $\hat{\rho}$ is the respondent sample correlation. Note that regardless of λ, $g(\hat{\rho}) \to 1$ as $\hat{\rho} \to 1$, so this choice of g satisfies the desirable property previously described.

Suppose λ is assumed to be positive, which seems reasonable given that X is a proxy for Y. Then as λ varies between 0 (missingness depends only on X) and infinity (missingness depends only on Y), $g(\hat{\rho})$ varies between $\hat{\rho}$ and $1/\hat{\rho}$. This result is intuitively very appealing. When $\lambda = 0$ the data are MAR, since in this case missingness depends only on the observed variable X. In this case $g(\hat{\rho}) = \hat{\rho}$, and Equation (1) reduces to the standard regression estimator described above. In this case the bias adjustment for Y increases with $\hat{\rho}$, as the association between Y and the variable determining the missing data mechanism increases. On the other hand when $\lambda = \infty$ and missingness depends only on the true value of Y, $g(\hat{\rho}) = 1/\hat{\rho}$ and Equation (1) yields the inverse regression estimator proposed by Brown (1990). The bias adjustment thus decreases with $\hat{\rho}$, reflecting the fact that in this case the bias in Y is attenuated in the proxy, with the degree of attenuation increasing with $\hat{\rho}$.

Note that there is no information in the data to inform the choice of λ. Little (1994) proposes a sensitivity analysis, where the estimate defined by Equations (1) and (8) are considered for a range of values of λ between 0 and infinity; the latter is the most extreme deviation from MAR, and estimates for this case have the highest variance. Indeed for small $\hat{\rho}$, the estimate with λ set to infinity is very unstable, and it is undefined when $\hat{\rho} = 0$. We suggest a sensitivity analysis using $\lambda = (0, 1, \infty)$ to capture the range of missingness mechanisms. In addition to the extremes, we use the intermediate case of $\lambda = 1$ that weights the proxy and true value of Y equally because the resulting estimator has a particularly convenient and simple interpretation. In this case $g(\hat{\rho}) = 1$ regardless of the value of $\hat{\rho}$, implying that the standardized bias in $\bar{y}_R$ is the same as the standardized bias in $\bar{x}_R$. In general, the stronger the proxy, the closer the value of $\hat{\rho}$ to one, and the smaller the differences between the three estimates.

4. Estimation Methods

4.1 Maximum Likelihood

The estimator described by Equations (1) and (8) is maximum likelihood for the pattern-mixture model. Large-sample variances are given by Taylor series calculations as in Little (1994), though this approximation may not be appropriate for small samples. Additionally, the ML estimate and corresponding inference does not take into account the fact that the regression coefficients that determine X are subject to sampling error. Better methods incorporate this uncertainty, such as the Bayesian method described below.

4.2 Bayesian Inference

An alternative to maximum likelihood is Bayesian inference, which allows us to incorporate the uncertainty in X and which may perform better in small samples. Let M denote the missingness indicator, and let α be a the vector of regression parameters from the regression of Y given Z that creates the proxy (i.e. $X = \alpha Z$). Let $Z \longrightarrow (X, V)$ be a (1-1) tranformation of the covariates. Letting [ ] denote distributions, we factor the joint distribution of Y, X, V, M, and α as follows:

$$[Y, X, V, M, \alpha] = [Y, X|M, \alpha][M][\alpha][V|Y, X, M, \alpha] \tag{9}$$

We leave the last distribution for V unspecified, and assume in (9) that M is independent of α. We assume the standard linear regression model creates the proxy X; the Y_i are independent normal random variables with mean $X = Z\alpha$ and variance ϕ^2. We place non-informative priors on all parameters and draw from their posterior distributions. For each draw of the parameters we recalculate the proxy using the draws of α and then scale using the draw of $\sigma_{xx}^{(0)}$ and $\sigma_{yy}^{(0)}$. Throughout the remainder of this and the following section we take X to denote this scaled version of the proxy.

Draws from the posterior distribution are obtained using different algorithms for the cases with $\lambda = 0$ and $\lambda = \infty$, as detailed below. In the case of intermediate values of λ the algorithm for $\lambda = \infty$ is applied to obtain draws from the joint distribution of $(X, X + \lambda Y)$ and then these draws are transformed to obtain the parameters of the joint distribution of (X, Y). In the equations that follow, let s_{jj} be the sample variance of j, $b_{jk.k}$ and $s_{jj.k}$ be the regression coefficient of k and the residual variance from the regression of j on k, and (0) and (1) denote quantities obtained from respondents and nonrespondents, respectively. The sample size is n with r respondents, and p is the number of covariates Z that create the proxy.

First we consider the model with $\lambda = 0$. This implies that missingness depends only on X, so the distribution of Y given X is the same for respondents and nonrespondents. Thus the intercept and regression coefficient, $\beta_{y0.x}^{(m)}$ and $\beta_{yx.x}^{(m)}$, are the same for $M = 0$ and $M = 1$. Draws of the identifiable parameters are computed in the following sequence:

1. $1/\phi^2 \sim \chi^2_{(r-p-1)}/((r-p-1)s_{yy.z}^{(0)})$
2. $\alpha \sim N(\hat{\alpha}, \phi^2(Z^TZ)^{-1})$
3. $\pi \sim Beta(r + 0.5, n - r + 0.5)$
4. $1/\sigma_{xx}^{(0)} \sim \chi^2_{(r-1)}/(rs_{xx}^{(0)})$
5. $\mu_x^{(0)} \sim N(\bar{x}_R, \sigma_{xx}^{(0)}/r)$
6. $1/\sigma_{yy.x}^{(0)} \sim \chi^2_{(r-2)}/(rs_{yy.x}^{(0)})$
7. $\beta_{yx.x}^{(0)} \sim N\left(b_{yx.x}, \dfrac{\sigma_{yy.x}^{(0)}}{rs_{xx}^{(0)}}\right)$
8. $\beta_{y0.x}^{(0)} \sim N(\bar{y}_R - \beta_{xy.y}^{(0)}\bar{x}_R, \sigma_{yy.x}^{(0)}/r)$
9. $1/\sigma_{xx}^{(1)} \sim \chi^2_{(n-r-1)}/((n-r)s_{xx}^{(1)})$
10. $\mu_x^{(1)} \sim N(\bar{x}_{NR}, \sigma_{xx}^{(1)}/(n-r))$

Draws from the posterior distribution of μ_y are obtained by substituting these draws into $\mu_y = \beta_{y0.x}^{(0)} + \beta_{yx.x}^{(0)}\mu_x$ where $\mu_x = \pi\mu_x^{(0)} + (1-\pi)\mu_x^{(1)}$.

When $\lambda = \infty$, the resulting assumption is that missingness depends only on Y, so the distribution of X given Y is the same for respondents and nonrespondents, i.e. $\beta_{x0.y}^{(m)}$ and $\beta_{xy.y}^{(m)}$ are the same for $M = 0$ and $M = 1$. Draws are obtained in a similar fashion as before. Steps 1 through 3 remain the same, but steps 4 through 10 are replaced by:

4. $1/\sigma_{yy}^{(0)} \sim \chi^2_{(r-1)}/(rs_{yy}^{(0)})$
5. $\mu_y^{(0)} \sim N(\bar{y}_R, \sigma_{yy}^{(0)}/r)$
6. $1/\sigma_{xx.y}^{(0)} \sim \chi^2_{(r-2)}/(rs_{xx.y}^{(0)})$
7. $1/\sigma_{xx}^{(1)} \sim \chi^2_{(n-r-1)}/((n-r)s_{xx}^{(1)})$
8. $\beta_{xy.y}^{(0)} \sim N\left(b_{xy.y}, \dfrac{\sigma_{xx.y}^{(0)}}{rs_{yy}^{(0)}}\right)$
9. $\beta_{x0.y}^{(0)} \sim N(\bar{x}_R - \beta_{xy.y}^{(0)}\bar{y}_R, \sigma_{xx.y}^{(0)}/r)$
10. $\mu_x^{(1)} \sim N(\bar{x}_{NR}, \sigma_{xx}^{(1)}/(n-r))$

To satisfy parameter constraints, the drawn value of $\sigma_{xx}^{(1)}$ from step 7 must be larger than the drawn value of $\sigma_{xx.y}^{(0)}$ from step 6; if this is not the case then these draws are discarded and these steps repeated. Draws from the posterior distribution of μ_y are obtained by substituting these draws into $\mu_y = \pi\mu_y^{(0)} + (1-\pi)(\mu_x^{(1)} - \beta_{x0.y}^{(0)})/\beta_{xy.y}^{(0)}$.

4.3 Multiple Imputation

An alternative method of inference for the mean of Y is multiple imputation (Rubin, 1978). We create K complete data sets by filling in missing Y values with draws from the posterior distribution, based on the pattern-mixture model. Draws from the posterior distribution of of Y are obtained by first drawing the parameters from their posterior distributions as outlined in Section 4.2, dependent on the assumption about λ, and then drawing the missing values of Y based on the conditional distribution of Y given X for nonrespondents ($M = 1$). For the kth completed data set, the estimate of μ_y is the sample mean $\bar{Y}_k$ with estimated variance W_k. A consistent estimate of μ_y is then given by $\hat{\mu}_y = \frac{1}{K}\sum_{k=1}^{K}\bar{Y}_k$ with $\text{Var}(\hat{\mu}_y) = \bar{W}_K + \frac{K+1}{K}B_K$, where $\bar{W}_K = \frac{1}{K}\sum_{k=1}^{K}W_k$ is the within-imputation variance and $B = \frac{1}{K-1}\sum_{k=1}^{K}(\bar{Y}_k - \hat{\mu}_y)^2$ is the between-imputation variance.

An advantage of the multiple imputation approach is the ease with which complex design features like clustering, stratification and unequal sampling probabilities can be incorporated. Once the imputation process has created complete data sets, design-based methods can be used to estimate μ_y and its variance; for example the Horvitz-Thompson estimator can be used to calculate $\bar{Y}_k$. Incorporating complex design features into the model and applying maximum likelihood or Bayesian methods is less straightforward, though arguably more principled. See for example Little (2004) for more discussion.

5. Application

The third National Health and Nutrition Examination Survey (NHANES III) was a large-scale stratified multistage probability sample of the noninstitutionalized U.S. population conducted during the period from 1988 to 1994 (U.S. Department of Health and Human Services, 1994). NHANES III collected data in three phases: (a) a household screening interview, (b) a personal home interview, and (c) a physical examination at a mobile examination center (MEC). The total number of persons screened was 39,695, with 86% (33,994) completing the second phase interview. Of these, only 78% were examined in the MEC. Since the questions asked at both the second and third stage varied considerably by age we chose to select only adults age 17 and older who had completed the second phase interview for the purposes of our example, leaving a sample size of 20,050. We chose to focus on estimating nonresponse bias for three body measurements at the MEC exam: systolic blood pressure (SBP), diastolic blood pressure (DBP), and body mass index (BMI). The nonresponse rates for these three items was 15%, 15%, and 10% respectively. It has been suggested that nonresponse in health surveys may be related to health (Cohen and Duffy, 2002), hence these measures may potentially be missing not at random.

In order to reflect nonresponse due to unit nonresponse at the level of the MEC exam we chose to only include fully observed covariates to create the proxies; variables that were fully observed for the sample included age, gender, race, and household size. To better approximate a normal distribution, BMI values were log-transformed. Systolic blood pressure displayed both the largest correlation between outcome and the proxy and the largest deviation in the proxy, with $\hat{\rho} = 0.6$, $d = 0.97$ and $d^* = 0.08$. Diastolic blood pressure had $\hat{\rho} = 0.31$, $d = 0.09$, and $d^* = 0.02$, while log(BMI) had the weakest proxy at $\hat{\rho} = 0.19$ and essentially no deviation with $d = -0.0001$ and $d^* = -0.002$.

For each outcome, estimates of the mean and confidence intervals for $\lambda = (0, 1, \infty)$ were obtained using maximum likelihood (ML), 5000 draws from the posterior distribution (PD), and multiple imputation with $K = 20$ data sets (MI). Additionally, since NHANES III has a complex survey design we obtained estimates using multiple imputation with design-based estimators of the mean using the survey weights (MI wt). Design-based estimators were computed using the "survey" routines in R, which estimate variances using Taylor series linearizations (Lumley, 2004).

Mean estimates and confidence intervals are displayed in Figure 1. The three methods, ML, PD, and MI, produce similar estimates and confidence intervals across all three outcomes and all values of λ. The intervals for weighted MI are larger than those for either of the non-design-adjusted methods, and for both SBP and BMI there is also a shift in the mean estimates for the weighted estimators, consistent for all values of λ. The choice of λ has a larger impact on the mean estimate for the SBP and DBP measurements; assuming MAR would result in significantly different mean estimates than assuming NMAR. BMI has a weak proxy and a small deviation so there is some evidence against nonresponse bias (small d) but this evidence is weak (small ρ).

6. Discussion

The PPM analysis of nonresponse bias we propose has the following attractive features: it integrates the various components of nonresponse noted in the introduction into a single sensitivity analysis, in a way we find satisfying. It is the only analysis we know of that formally reflects the hierarchy of evidence about bias in the mean suggested in the introduction, which we believe is realistic. It is easy to implement, since the ML form is simple to compute, and the Bayesian simulation is noniterative, not requiring iterative Markov Chain Monte Carlo methods that pervade more complex Bayesian methods and might deter survey practitioners; the MI method is also non-iterative, and allows complex design features to be incorporated in the within-imputation component of variance. PPM analysis includes but does not assume MAR, and it provides a picture of the potential nonresponse bias under a reasonable range of MAR and non-MAR mechanisms. It gives appropriate credit to the existence of good predictors of the observed outcomes. When data are MAR, it is the squared correlation between the covariates and the outcome that drives the reduction in variance, which means that covariates with rather high correlations are needed to have much impact. An interesting implication of our PPM analysis is that if the data are not MAR, covariates with moderate values of correlation, such as 0.5, can be useful in reducing the sensitivity to assumptions about the missing data mechanism. We suggest that emphasis at the design stage should be on collection of strong auxiliary data to help evaluate and adjust for potential nonresponse, not solely on obtaining the highest possible response rate.

A drawback of our analysis is that it focus only on the mean of a particular outcome Y, so it is outcome-specific. Thus, in a typical survey with many outcomes, the analysis needs to be repeated on each of the key outcomes of interest and then integrated in some way that reflects the relative importance of these outcomes. This makes life complicated, but that seems to us inevitable. An unavoidable feature of the problem is that nonresponse bias is small for variables unrelated to nonresponse, and potentially larger for variables related to nonresponse. Measures that do not incorporate relationships with outcomes, like the variance of the nonresponse weights, cannot capture this dimension of the problem.

The pattern-mixture model that justifies the proposed analysis strictly only applies to continuous survey variables, where normality is reasonable, although we feel it is still informative when applied to non-normal outcomes. Extensions to categorical variables appear possible via probit models, and many other extensions can be envisaged, including extensions to other generalized linear models. PPM analysis can be applied to handle item nonresponse by treating each item subject to missing data separately, and restricting the covariates to variables that are fully observed. However, this approach does not condition fully on the observed information, and extensions for general patterns of missing data would be preferable. Our future work on PPM analysis will focus on developing these extensions.

References

Bethlehem, J. (2002), "Weighting Nonresponse Adjustments Based on Auxiliary Information," in *Survey Nonresponse*, eds. Groves, R., Dillman, D., Eltinge, J., and Little, R., New York: Wiley, chap. 18, pp. 275–287.

Brown, C. H. (1990), "Protecting Against Nonrandomly Missing Data in Longitudinal Studies," *Biometrics*, 46, 143–155.

Cohen, G. and Duffy, J. C. (2002), "Are Nonrespondents to Health Surveys Less Healthy than Respondents," *Journal of Official Statistics*, 18, 13–23.

Curtain, R., Presser, S., and Singer, E. (2000), "The Effects of Response Rate Changes on the Index of Consumer Sentiment," *Public Opinion Quarterly*, 64, 413–428.

Federal Committee on Statistical Methodology (2001), "Statistical Policy Working Paper 31: Measuring and Reporting Sources of Error in Surveys," Tech. rep., Executive Office of the President of the United States of America.

Groves, R. M. (2006), "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly*, 70, 646–675.

Heckman, J. J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *The Annals of Economic and Social Measurement*, 5, 475–492.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., and Presser, S. (2000), "Consequences of Reducing Nonresponse in a National Telephone Survey," *Public Opinion Quarterly*, 64, 125–148.

Little, R. and Vartivarian, S. (2005), "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology*, 31, 161–168.

Little, R. J. A. (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88, 125–134.

— (1994), "A Class of Pattern-Mixture Models for Normal Incomplete Data," *Biometrika*, 81, 471–483.

— (2004), "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling," *Journal of the American Statistical Association*, 99, 546–556.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley: New York, 2nd ed.

Lumley, T. (2004), "Analysis of complex survey samples," *Journal of Statistical Software*, 9, 1–19.

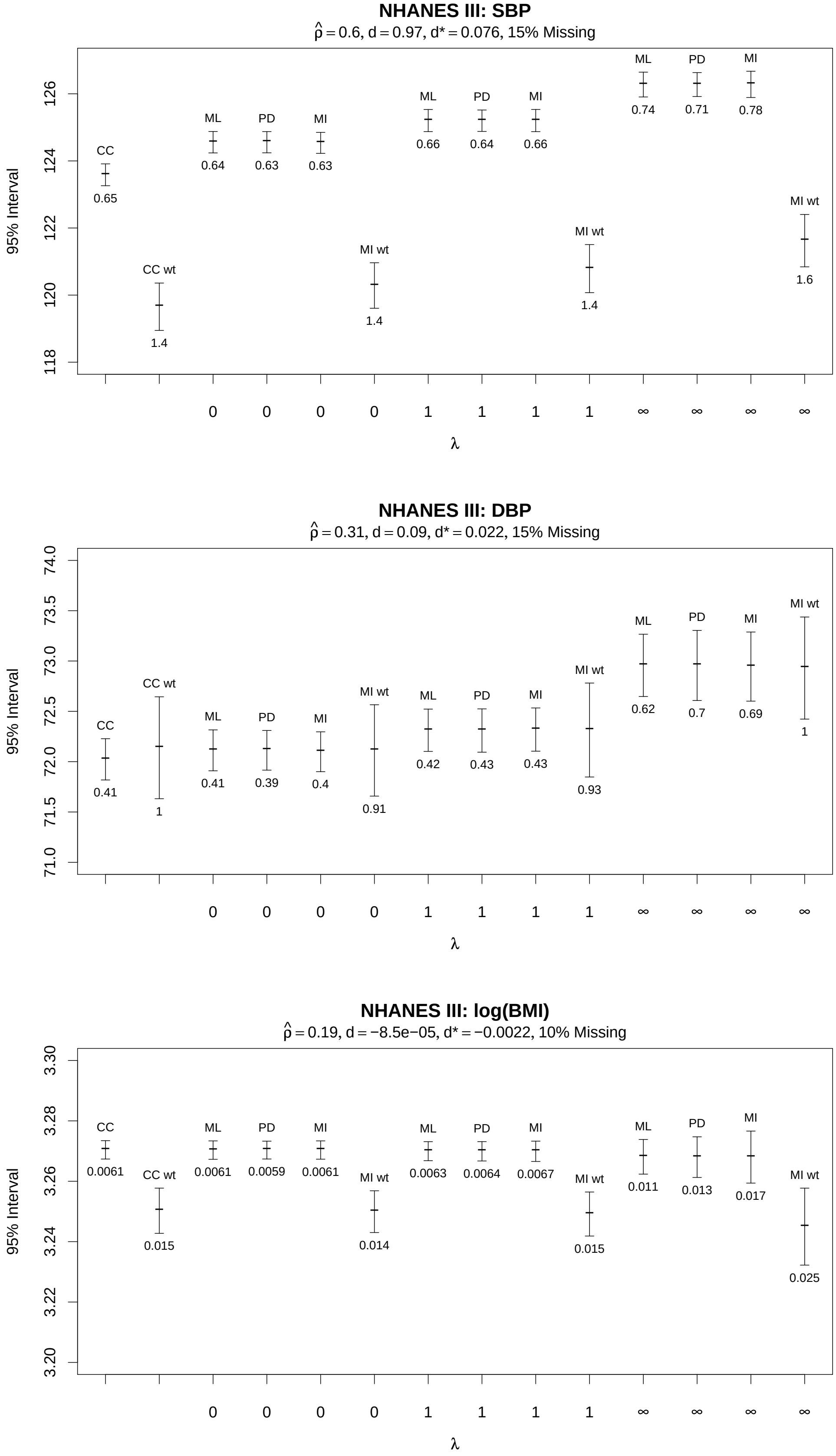
Office of Management and Budget (2006), "Standards and Guidelines for Statistical Surveys," Tech. rep., Executive Office of the President of the United States of America.

Rubin, D. B. (1976), "Inference and Missing Data (with Discussion)," *Biometrika*, 63, 581–592.

— (1978), "Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 20–34.

U.S. Department of Health and Human Services (1994), "Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988-94," Tech. rep., National Center for Health Statistics, Centers for Disease Control and Prevention.

Figure 1: Estimates of mean SBP, DBP, and BMI (log-transformed) for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. CC: Complete case; CC wt: Complete case with estimation incorporating the survey design; ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets; MIwt: 20 multiply imputed data sets with estimation incorporating the survey design.


NHANES III: SBP
$\hat{\rho} = 0.6$, d = 0.97, d* = 0.076, 15% Missing
95% Interval
CC 0.65
CC wt 1.4
ML 0.64
PD 0.63
MI 0.63
MI wt 1.4
ML 0.66
PD 0.64
MI 0.66
MI wt 1.4
ML 0.74
PD 0.71
MI 0.78
MI wt 1.6
0 0 0 0 1 1 1 1 ∞ ∞ ∞ ∞
λ
NHANES III: DBP
$\hat{\rho} = 0.31$, d = 0.09, d* = 0.022, 15% Missing
95% Interval
CC 0.41
CC wt 1
ML 0.41
PD 0.39
MI 0.4
MI wt 0.91
ML 0.42
PD 0.43
MI 0.43
MI wt 0.93
ML 0.62
PD 0.7
MI 0.69
MI wt 1
0 0 0 0 1 1 1 1 ∞ ∞ ∞ ∞
λ
NHANES III: log(BMI)
$\hat{\rho} = 0.19$, d = −8.5e−05, d* = −0.0022, 10% Missing
95% Interval
CC 0.0061
CC wt 0.015
ML 0.0061
PD 0.0059
MI 0.0061
MI wt 0.014
ML 0.0063
PD 0.0064
MI 0.0067
MI wt 0.015
ML 0.011
PD 0.013
MI 0.017
MI wt 0.025
0 0 0 0 1 1 1 1 ∞ ∞ ∞ ∞
λ