Journal of Official Statistics, Vol. 27, No. 2, 2011, pp. 153–180

Proxy Pattern-Mixture Analysis for Survey Nonresponse

Rebecca R. Andridge[1] *and Roderick J. A. Little*[2]

We propose proxy pattern-mixture analysis (PPMA), a simple method for assessing nonresponse bias for the mean of a survey variable Y subject to nonresponse, when there is a set of covariates observed for nonrespondents and respondents. The covariates are reduced to a proxy variable X that has the highest correlation with Y, estimated from a regression analysis of respondent data. The impact of nonresponse on bias depends primarily on three factors: the nonresponse rate, the strength of the proxy variable in predicting Y, and the difference in proxy mean for respondents and nonrespondents. The PPMA method combines all three elements in an intuitively reasonable way. Adjusted estimators of the mean of Y are based on 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 $X + \lambda Y$ of X and Y. The method does not assume the missing-data mechanism is missing at random ($\lambda = 0$), and provides a sensitivity analysis for different values of λ. Maximum likelihood, Bayesian and multiple imputation versions of PPMA are described. Properties are examined through simulation and with data from the third National Health and Nutrition Examination Survey (NHANES III) and the Ohio Family Health Survey (OFHS).

Key words: Bayesian methods; missing data; nonignorable nonresponse; nonresponse bias analysis; survey data.

1. Introduction

Missing data are often a problem in sample surveys, arising when a sampled unit does not respond to the entire survey (unit nonresponse) or to a particular question (item nonresponse). In this article 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 X are measured for both respondents and nonrespondents. With unit nonresponse, this set of variables is generally restricted to survey design variables or paradata, 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.

[1] Division of Biostatistics, The Ohio State University, Columbus, Ohio, 43210. Email: randridge@cph.osu.edu
[2] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, 48109. Email: rlittle@umich.edu

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, this reference lacks specific analysis recommendations, 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 et al. 2000; Keeter et al. 2000).

There are three major components to consider in evaluating nonresponse: the amount of missing data, 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 missing data, as measured by the response rate, has been the most oft-used metric for evaluating survey quality. Bias depends directly on the response rate and the differences between respondents and nonrespondents on the survey outcome of interest. If Y is a survey outcome subject to nonresponse, then

$$\text{Bias}(\bar{y}_R) = \frac{N-R}{N}(\bar{y}_R - \bar{y}_{NR}) \tag{1}$$

where $\bar{y}_R$ is the unadjusted mean of the respondents, $\bar{y}_R(\bar{y}_{NR})$ is the respondent (nonrespondent) mean in the population, and $(N-R)/N$ is the nonresponse rate (Groves 2006).

However, response rates ignore the information contained in auxiliary covariates observed for nonrespondents. U.S. 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. Schouten et al. (2009) propose the use of R-indicators to assess the "representativeness" of respondents with respect to the complete sample. Response is considered representative if response propensities are constant across the sample, which corresponds to a missing completely at random mechanism. Using estimated response propensities conditional on auxiliary variables for the entire sample, the authors construct a measure that takes values between zero (furthest from representativeness) and one (perfectly representative). An alternative approach is that of Särndal and Lundström, whose q^2 indicator is defined as the variance of predicted inverse response probabilities (Särndal and Lundström 2005, 2008). The authors use this sample-based bias estimator to compare vectors of auxiliary vectors in order to select the best one for use in adjustment weighting. Larger values of q^2 correspond to low potential for nonresponse bias. As with the R-indicator, the q^2 statistic is not outcome-dependent. Though response propensity and other related 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, 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

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

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

$$FMI = \frac{n/\sigma_{yy} - Var^{-1}(\hat{\mu}_y)}{n/\sigma_{yy}} \tag{3}$$

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. The use of FMI as a measure of survey quality was recently proposed by Wagner (2010). 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 in our view has not integrated them in a satisfactory way. In particular, methods such as the R-indicator or the q^2 bias indicator are not outcome-specific, and we would prefer a measure that explicitly links bias to a particular survey outcome, acknowledging that biases may be very different for different outcomes in a single survey, and that data may be missing not at random for some outcomes. We propose a new method for nonresponse bias measurement and adjustment that takes account of all three aspects, in a way which we find intuitive and satisfying. In particular, it gives appropriate credit for predictive auxiliary data, and yields less variability when the auxiliary data have a similar distribution for respondents and nonrespondents. Another strength of the proposed approach is that, unlike most current weighting and imputation methods for adjusting for survey nonresponse, we do not assume the data are MAR. Our methods are based on a pattern-mixture model (Little 1993) for the survey outcome that allows missingness to be

not at random (MNAR) and assesses the sensitivity of estimates to deviation 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, Chapter 15) and citations therein. As a measure of the impact of nonresponse, we propose using the estimated fraction of missing information, obtained through multiple imputation under the pattern-mixture model with a range of assumptions about the nonresponse mechanism.

Section 2 introduces our approach to the nonresponse problem and describes the general framework, and Section 3 details the corresponding pattern-mixture model analysis. Section 4 describes three different estimation approaches: maximum likelihood, a Bayesian approach, and multiple imputation. Section 5 discusses the use of the fraction of missing information from multiple imputation under the pattern-mixture model as a measure of uncertainty due to nonresponse. Section 6 describes a set of simulation studies to demonstrate the assessment of nonresponse bias using these methods. Section 7 applies these methods to two data sets: NHANES III and the Ohio Family Health Survey. Section 8 presents discussion, including extensions of the proposed method.

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 paradata, or with item nonresponse. Alternatively, covariates Z could be available through record linkage with administrative registers. 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 and taking X to be the predicted values of Y, available for both respondents and nonrespondents. This regression should include 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 among the respondent cases, 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. In the case of unit nonresponse, design variables and paradata have been shown to produce low (≤ 0.2) correlations with outcomes (Kreuter et al. 2010), while for item nonresponse we would expect higher correlations.

Let $\bar{y}_R$ denote the respondent mean of Y, which is subject to nonresponse bias. Our proposed method is based on a pattern-mixture model for the distribution of (X,Y) for

respondents and nonrespondents (Little 2004) and yields estimates of the mean of Y of the form

$$\hat{\mu}_y = \bar{y}_R + g(\hat{\rho})\sqrt{\frac{s_{yy}}{s_{xx}}}(\bar{x} - \bar{x}_R) \tag{4}$$

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. The function $g(\hat{\rho})$ is an unspecified function of the respondent sample correlation of X and Y, $\hat{\rho}$. Note that since the proxy X is estimated from the regression of Y on Z, it will have lower variance than Y. Rearranging terms yields the standardized estimated bias in $\bar{y}_R$ as a function of the standardized estimated bias in $\bar{x}_R$,

$$\frac{\hat{\mu}_y - \bar{y}_R}{\sqrt{s_{yy}}} = g(\hat{\rho})\frac{\bar{x} - \bar{x}_R}{\sqrt{s_{xx}}} \tag{5}$$

Some comments on the estimator (4) 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. The impact of "large" and "small" values of d vary across outcomes even within a survey; in Section 6.1 we consider various size deviations and illustrate their impact. 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 (4) 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 (4). 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 (4) 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 (4) 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 reflect this hierarchy.

3. The Pattern-Mixture Model

Let M denote the missingness indicator, such that $M = 1$ if Y is missing and $M = 0$ if Y is observed. We assume the respondent mean of Y, conditional on auxiliary variables Z, can be written as $E(Y|Z, M = 0) = \alpha_0 + \alpha Z$, 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:

$$\begin{aligned}(Y, X|M = m) &\sim N_2\left(\left(\mu_y^{(m)}, \mu_x^{(m)}\right), \Sigma^{(m)}\right)\\ M &\sim Bernoulli\,(1 - \pi)\\ \Sigma^{(m)} &= \begin{bmatrix} \sigma_{yy}^{(m)} & \rho^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}} \\ \rho^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}} & \sigma_{xx}^{(m)} \end{bmatrix}\end{aligned} \tag{6}$$

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

$$\Pr(M = 1|Y, X) = f(X + \lambda^* Y) \tag{7}$$

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

$$((Y, X) \perp M|f(X + \lambda^* Y)) \tag{8}$$

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

$$\hat{\mu_y} = \bar{y}_R + \frac{s_{xy} + \lambda^* s_{yy}}{s_{xx} + \lambda^* s_{xy}}(\bar{x} - \bar{x}_R) \tag{9}$$

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 (7) for the missingness mechanism. Specifically we replace (7) by

$$\Pr(M = 1|Y, X) = f\left(X\sqrt{\frac{\sigma^{(0)}_{yy}}{\sigma^{(0)}_{xx}}} + \lambda Y\right) = f(X^* + \lambda Y) \tag{10}$$

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

$$((Y, X) \perp M|f(X^* + \lambda Y)) \tag{11}$$

where $\perp$ denotes independence. We call the model defined by (6) and (10) 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 is

$$\hat{\mu}_y = \bar{y}_R + \frac{\lambda + \hat{\rho}}{\lambda\hat{\rho} + 1}\sqrt{\frac{s_{yy}}{s_{xx}}}(\bar{x} - \bar{x}_R) \tag{12}$$

where $\hat{\rho}$ is the respondent sample correlation. This has the form of (4), where

$$g(\hat{\rho}) = \frac{\lambda + \hat{\rho}}{\lambda\hat{\rho} + 1} \tag{13}$$

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.

3.1. Properties of the Missing Data Mechanism

There are limitless ways to model deviations from MAR, and any method needs to make assumptions. Thus, the assumption about the missing data mechanism, given by (10), is a key to the proposed method, and deserves careful consideration. The assumption (10) is quite flexible and covers a wide range of models relating X (X^*) and Y to M. In particular, it is more flexible than the well-known Heckman selection model (Heckman 1976), which assumes that missingness is linear in X and Y. For example, the PPM model encompasses not only mechanisms that are linear in X or linear in Y, but also ones that are quadratic in X or quadratic in Y. A broad class of mechanisms are those that depend on both X and Y, potentially including quadratic terms and the interaction of X and Y, that is

$$\text{logit}(\Pr(M = 1|Y, X)) = \gamma_0 + \gamma_1 X + \gamma_2 X^2 + \gamma_3 Y + \gamma_4 Y^2 + \gamma_5 XY \tag{14}$$

If we take $f(\cdot)$ in (10) to be quadratic, we obtain a missingness mechanism for the PPM that is a specific subset of this general model. The assumed PPM missing data mechanism is then

$$\begin{aligned}\text{logit}(\Pr(M = 1|Y, X)) &= \alpha_0 + \alpha_1(X + \lambda Y) + \alpha_2(X + \lambda Y)^2 \\ &= \alpha_0 + \alpha_1 X + \alpha_1\lambda Y + \alpha_2 X^2 + \alpha_2\lambda^2 Y^2 + 2\alpha_2\lambda XY\end{aligned} \tag{15}$$

Assuming (14) is the true model, the ability of the PPM to produce an unbiased estimate depends on whether there is a value of λ that makes (15) close to (14). In particular, we note

that the sign of the X and Y terms (and similarly the X^2 and Y^2 terms) must be the same; however, since X is a proxy for Y we feel that this assumption is not unreasonable.

An alternative method when data may be not MAR is to specify a selection model, which factors the joint distribution of Y and M given X into the conditional distribution of M given Y and X as in (14) and the marginal distribution of Y given X (e.g., Heckman 1976; Diggle and Kenward 1994; Little and Rubin 2002). The approach requires full specification of the distribution of M given Y and X. In contrast, our pattern-mixture model avoids the need to specify the function f that relates missingness of Y to $X^* + \lambda Y$, although it shares with the corresponding selection model the assumption that missingness depends on Y and Z only through the value of $X^* + \lambda Y$. One reason for restricting the dependence on the set of variables Z to the combination X^* is that, under the normality assumption, dependence on missingness of Y on other combinations (say $U = \delta Z$) does not result in bias in the mean of Y, since Y is conditionally independent of U given X^*. Reduction to X^* limits the analysis to just one sensitivity parameter (λ) and so is much simpler than an analysis that models departures from MAR for each of the individual Z's. Another advantage of our model is that likelihood-based analysis is much simpler than selection models, which require iterative algorithms.

3.2. *Other Properties of the Model*

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 (4) 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 (4) 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 $\hat{\rho}$.

3.3. *Sensitivity Analysis*

There is no information in the data to inform the choice of λ. As in Little (1994), we propose a sensitivity analysis, where the estimate (12) is 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 a 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 (12) is maximum likelihood (ML) 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 methods described below.

4.2. Bayesian Inference

An alternative to ML 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 $Z \rightarrow (X,V)$ 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:

$$[Y, X, V, M, \alpha] = [Y, X|M, \alpha][M][\alpha][V|Y, X, M, \alpha] \tag{16}$$

We leave the last distribution for V unspecified, and assume in (16) 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.

To obtain draws from the posterior distributions of all parameters (including the regression coefficients that create the proxy), we place noninformative (Jeffreys') priors on all parameters. The sample size is n with r respondents, and p is the number of covariates Z that create the proxy. First we draw the parameters of the regression model,

$$1/\phi^2 \sim \chi^2_{(r-p-1)} / \left((r - p - 1) s^{(0)}_{yy.z} \right) \quad \alpha \sim N(\hat{\alpha}, \phi^2 (Z^T Z)^{-1}) \tag{17}$$

where $s^{(0)}_{yy.z}$ is the residual variance from the regression of Y on Z for the respondents. Using these draws we create the proxy X. In order to scale the proxy to obtain X^* we draw $\sigma^{(0)}_{xx}$ and $\sigma^{(0)}_{yy}$ from their posteriors. Then the remaining parameters of the proxy pattern-mixture model (for Y and X^*) are drawn; the algorithm and details are given in Little (1994) and Little and Rubin (2002, §15.5.2).

We note that, as in Little (1994) and Little and Rubin (2002), draws from the posterior distribution are obtained using different algorithms for the cases with $\lambda = 0$ and $\lambda = \infty$. 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). When $\lambda = 0$, draws from the posterior distribution of μ_y are obtained by substituting these draws into $\mu_y = \beta^{(0)}_{y0.x} + \beta^{(0)}_{yx.x}\mu_x$ where $\mu_x = \pi\mu^{(0)}_x + (1 - \pi)\mu^{(1)}_x$. When $\lambda = \infty$, draws from the posterior distribution of μ_y are

obtained by substituting these draws into

$$\mu_y = \pi\mu_y^{(0)} + (1-\pi)\frac{\mu_x^{(1)} - \beta_{x0.y}^{(0)}}{\beta_{xy.y}^{(0)}} \tag{18}$$

Within this Bayesian framework, it may be tempting to place a prior on λ and obtain a single estimate of μ_y. However, since the data contain no information about λ, resulting inference would be driven entirely by the prior. We feel that the sensitivity analysis using three distinct values of λ more honestly describes the sensitivity of inference to assumptions on the missing data mechanism.

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 ($M = 1$),

$$[y_i|x_i, m_i = 1, \phi_{(k)}] \sim N\left(\mu_{y(k)}^{(1)} + \frac{\sigma_{yx(k)}^{(1)}}{\sigma_{xx(k)}^{(1)}}\left(x_i - \mu_{x(k)}^{(1)}\right), \sigma_{yy(k)}^{(1)} - \frac{\sigma_{yx(k)}^{(1)\ 2}}{\sigma_{xx(k)}^{(1)}}\right) \tag{19}$$

where the subscript (k) denotes the kth draws of the parameters. 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 $\hat{\mu}_y = 1/K\sum_{k=1}^{K}\bar{Y}_k$ with $\mathrm{Var}(\hat{\mu}_y) = \bar{W}_k + ((K+1)/K)B_K$, where $\bar{W}_K = 1/K\sum_{k=1}^{K}W_k$ is the within-imputation variance and $B = 1/(K-1) \times\sum_{k=1}^{K}(\bar{Y}_k - \hat{\mu}_y)^2$ is the between-imputation variance.

An advantage of the multiple imputation approach is that complex design features like clustering, stratification and unequal sampling probabilities can be incorporated in the within-imputation variance component of the multiple imputation inference. 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. Quantifying Uncertainty Due to Nonresponse

We propose using the estimated fraction of missing information (FMI), obtained through multiple imputation under the PPM model, as a measure of uncertainty due to nonresponse. Under the sensitivity analysis approach previously described there will be a set of FMI estimates, each obtained using different assumptions about the nonresponse mechanism ($\lambda = \{0, 1, \infty\}$). To the extent that variance and bias are intertwined, this set of FMI estimates acts as a marker of the potential for nonresponse bias, and our ability to correct a potential bias. We note that FMI can be computed from observed and complete-data information matrices and thus applies also to the maximum likelihood and

Bayesian estimation methods. We describe its use in the multiple imputation framework as a convenient way of avoiding the need to estimate and manipulate information matrices.

The FMI due to nonresponse is estimated by the ratio of between-imputation to total variance under multiple imputation (Little and Rubin 2002). Recently, Wagner (2010) proposed the use of FMI to monitor the quality of survey data during data collection. Wagner estimates and uses FMI under the assumption that data are MAR, but we propose its application under the pattern-mixture model where missingness is not necessarily at random. FMI is influenced by both the strength of the proxy (ρ) and the size of the deviation from MCAR (d), as well as the assumption about the nonresponse mechanism (λ). For the purposes of illustration we use the standardized deviation d^* so it is the same regardless of whether X has been scaled.

Figure 1 is a plot of simulated data showing FMI as a function of ρ for different values of d^* and the nonresponse rate, with separate estimates for different nonresponse assumptions ($\lambda = 0, 1, \infty$). A summary of the effect of these parameters on FMI is as follows. Across all values of ρ and d^* FMI is smallest when assuming $\lambda = 0$, largest when assuming $\lambda = \infty$, and falls somewhere in-between when $\lambda = 1$. Regardless of the nonresponse mechanism, as the strength of the proxy (ρ) increases, the FMI decreases, eventually reaching zero for a perfect proxy ($\rho = 1$). Larger d leads to elevated FMI, though these differences are relatively small compared to the effect of ρ. The FMI is larger for lower response rates across all values of ρ, though differences are more severe with a strong proxy than with a weak one. When missingness is at random ($\lambda = 0$) and d^* is small, the FMI is approximately equal to the nonresponse rate for $\rho = 0$ and decreases as the strength of the proxy increases.

With MNAR mechanisms, the FMI is much higher than the response rate for weak proxies, but the relative gains from a moderately correlated proxy are larger for MNAR mechanisms than for the MAR mechanism. For example, for small d and 50% missingness the gain from moving from $\rho = 0$ to $\rho = 0.5$ is a decrease in FMI from 0.5 to 0.43 when $\lambda = 0$ but from nearly 1 to 0.75 when $\lambda = \infty$. Clearly the presence of strong predictors is of the utmost importance in identifying and removing nonresponse bias; the sensitivity of FMI to ρ illustrates this.

An attractive property of using the FMI values as a quality measure is that they are bounded by 0 and 1, and can be easily compared across outcomes or even across surveys. It also provides a more honest measure of survey quality than the response rate, which is a popular but limited measure, since it does not account for the information in auxiliary data or deviations from MCAR. While the ranking of ρ and d^* plays a similar role to the FMI, the FMI essentially combines these to paint a picture of the overall uncertainty due to nonresponse. For a given lambda, FMI has the usual interpretation in terms of variance increase from missing data – one could think of it as a "virtual response rate" under the assumption on the missingness mechanism (value of λ). As such, the FMI from the PPM expands upon the FMI measure proposed by Wagner (2010) by allowing missingness to be not at random.

To use the set of FMI values as a measure of the severity of nonresponse for a particular outcome, we compare the values to the nonresponse rate and assess the range of the FMI values. If the FMI values are all below the nonresponse rate, then we have very strong covariate information, and can be confident that we have good information to correct bias,

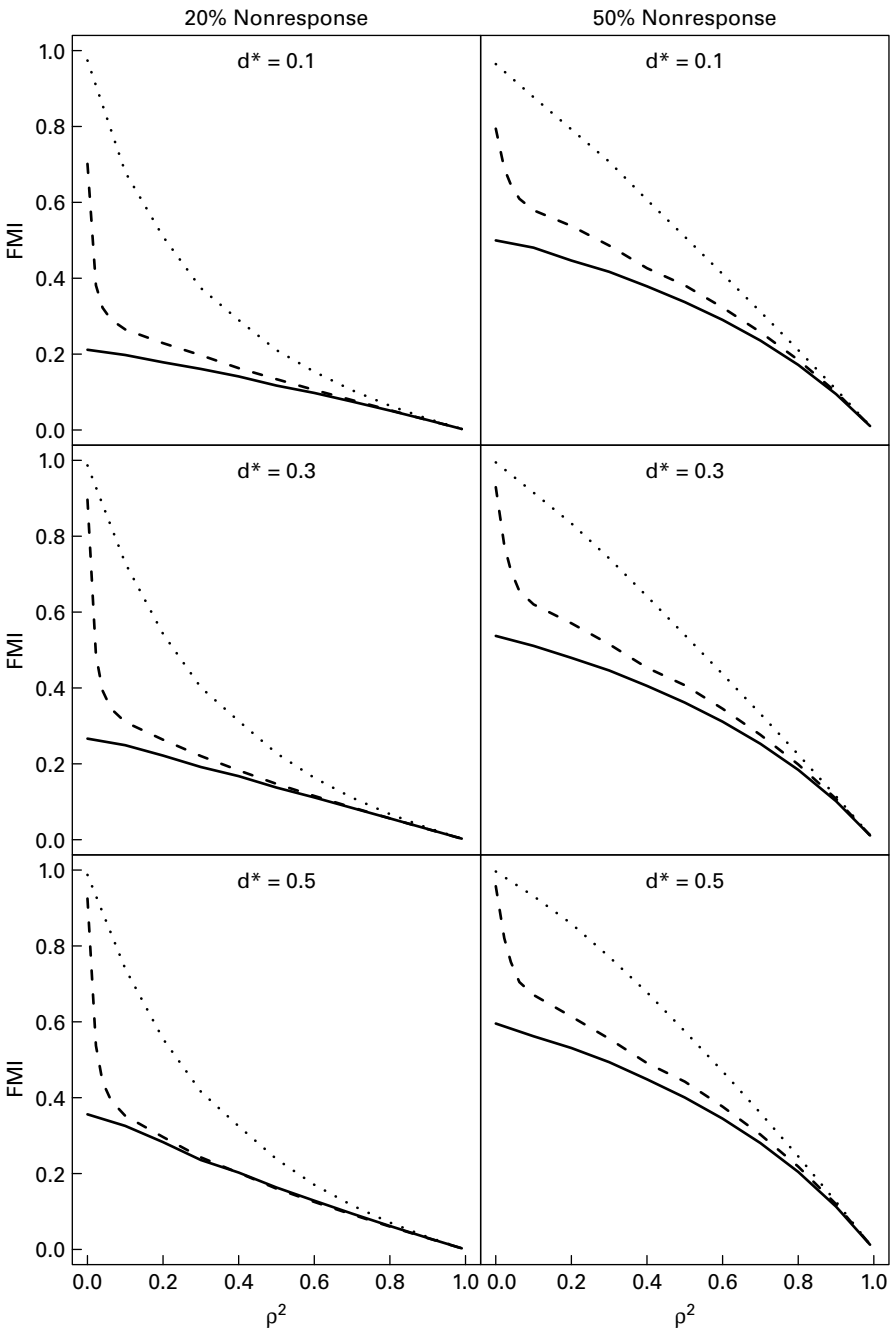

Fig. 1. Fraction of missing information as a function of ρ, d, and nonresponse rate for $\lambda = 0$ (solid), $\lambda = 1$ (dashes), and $\lambda = \infty$ (dots)

even if data are MNAR. If the largest FMI value is close to the maximum value, then we know that statements about bias require caution, since even a low d^* does not ensure that we have a clear picture of the bias. The range of the FMI values gives an overall picture of how much uncertainty we have in attempting to correct for nonresponse bias. If the values

are close together, than we can be confident that we have the information to obtain estimates of the bias, even under the most severe MNAR assumption. However, if the range of FMI values is large, and especially if the largest one approaches 1, then this is an indicator of a lack of information for assessing bias.

Unlike the R-indicator recently proposed by Schouten et al. (2009) comparing FMI values across surveys does not require using the same set of auxiliary variables in each survey. In fact, a different set of auxiliary variables might be used within a survey when applying PPMA to multiple different outcomes. The difficulty may be that the true λ might differ between surveys (or across outcomes), so the comparison should be of the set of FMI values, not of any one particular value.

6. Simulation Studies

We now describe a set of simulation studies designed to (1) illustrate the effects of ρ, d^*, and sample size on PPM estimates of the mean of Y, (2) assess confidence coverage of ML, Bayes and MI inferences, and (3) demonstrate the performance of the PPM model when data arise from a selection model with a range of nonresponse mechanisms. All simulations and data analysis were performed using the software package R (R Development Core Team 2008). Throughout the simulations estimates of mean and variance are calculated assuming a simple random sample.

6.1. Numerical Illustration of PPM Analysis

Our first objective with the simulation studies was to numerically illustrate the taxonomy of evidence concerning bias based on the strength of the proxy and the deviation of its mean. We created a total of eighteen artificial data sets in a 3x3x2 factorial design. A single data set was generated for each combination of $\rho = \{0.8, 0.5, 0.2\}$, $d^* = \{0.1, 0.3, 0.5\}$ and $n = \{100, 400\}$ as follows. A single covariate Z was generated for both respondents and nonrespondents with the outcome Y generated only for respondents. Respondent data were created as pairs $(z_i, y_i), i = 1. \ . \ .r$ with $z_i \sim N(0, \rho^2)$ and $y_i = 1 + z_i + e_i$, where $e_i \sim N(0, 1 - \rho^2)$. Nonrespondent data were $Z^{'}s$ only, generated from $z_i \sim N(2\rho d^*, \rho^2)$ for $i = r + 1. \ . \ .n$. The nonresponse rate was fixed at 50%. This data structure was chosen so that the variance of the complete case mean would be constant (and equal to one) across different choices of ρ and d^*, and so that varying ρ would not affect d^* and vice-versa. R^2 values that corresponded to the selected ρ were 64%, 25%, and 4%, covering a range likely to be encountered in practice with item nonresponse (larger ρ) and unit nonresponse (smaller ρ).

For each of the eighteen data sets, estimates of the mean of Y and its precision were obtained for $\lambda = (0, 1, \infty)$. For each value of λ, three 95% intervals were calculated:

(a) ML: the maximum likelihood estimate ± 2 standard errors (large-sample approximation)
(b) PD: the posterior median and 2.5th to 97.5th posterior interval based on 5,000 draws from the posterior distribution of μ_Y as outlined in Section 4.2
(c) MI: mean ± 2 standard errors from 20 multiply-imputed data sets.

Posterior median and quantiles were used because initial evaluations showed that the posterior distribution of μ_Y was skewed and had extreme outliers for small ρ and large λ. The complete-case estimate (± 2 standard errors) was also computed for each data set; note that the expected value of the respondent mean and corresponding confidence interval is constant across all values of ρ and d for each n.

6.1.1. Results

Results from applying the three estimation methods to each of the nine data sets with $n = 100$ are displayed in Figure 2. The complete case estimate is shown alongside 95% intervals estimated by maximum likelihood, multiple imputation, and the posterior distribution, for $\lambda = (0, 1, \infty)$. For each population the PD intervals are longer than the ML and MI intervals for all choices of λ, especially for weak proxies and $\lambda = \infty$. Results for $n = 400$ were similar and are not shown.

Populations with a strong proxy ($\rho = 0.8$) do not show much variation across values of λ; there is evidence that nonresponse bias is small for small d and there is good information to correct the potential bias for larger values of d. For moderately strong proxies ($\rho = 0.5$) the intervals increase in length, with differences between PD and ML becoming more exaggerated as d increases. As expected, when the proxy is weak ($\rho = 0.2$) we see large intervals for models that assume missingness is not at random ($\lambda \neq 0$); this reflects the fact that we are in the worst-case scenario where there is not much information in the proxy to estimate the nonresponse bias. Notice that in this simulation the true mean of Y is not known; we simply illustrate the effect of various values of ρ and λ on the sensitivity analysis.

6.2. *Confidence Coverage*

The second objective of the simulation was to assess coverage properties for each of the three estimation methods. We generated 500 replicate data sets as before for each of the eighteen population designs and computed the actual coverage of a nominal 95% interval and median interval length. The Bayesian intervals were based on 1000 draws from the posterior distribution. Coverage is based on the unreasonable assumption that the assumed value of λ equals the actual value of λ. This is unrealistic, but coverages are clearly not valid when the value of λ is misspecified, and uncertainty in the the choice of λ is captured by the sensitivity analysis.

6.2.1. Results

Table 1 displays the nominal coverage and median CI width for each of the eighteen populations. For populations with a strong or moderately strong proxy ($\rho = 0.8, 0.5$) coverage is at or above nominal levels for all three methods, for both the smaller and larger sample sizes and for all levels of d. For these populations, PD inference is slightly more conservative; intervals are larger than ML for most populations. However, when the proxy is weak, ML coverage is below nominal levels for larger values of λ, while both PD and MI have coverage close to nominal levels. With small sample size and weak proxies, taking $\lambda = \infty$ leads to large confidence intervals, since draws of $\beta_{xy.y}$ approach zero.

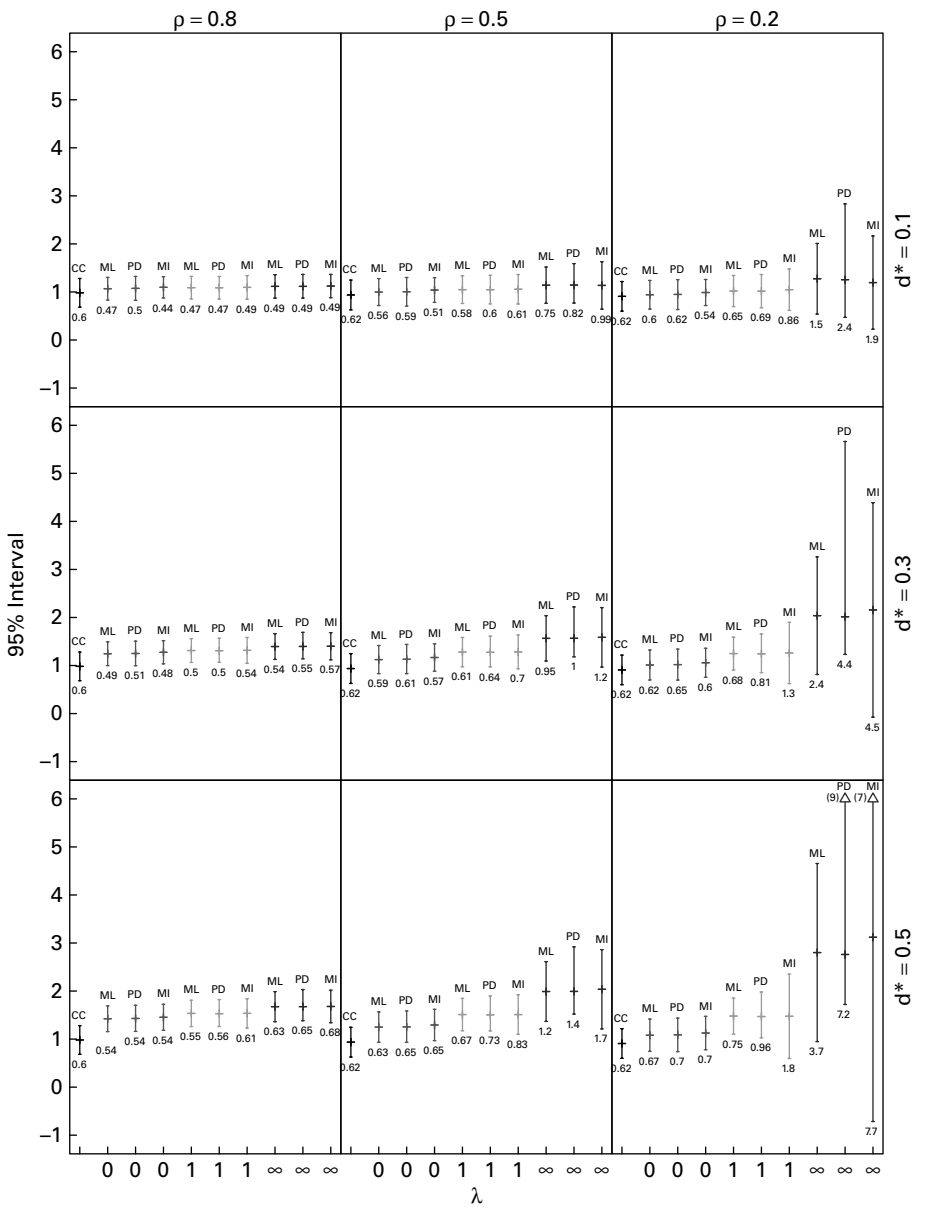

Fig. 2. 95% confidence intervals for nine generated data sets ($n = 100$) for $\lambda = (0, 1, \infty)$. Numbers below intervals are the interval length. CC: Complete case; ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply-imputed data sets

The $\lambda = \infty$ model requires a strong proxy or large sample size to provide reliable estimates of μ_y.

6.3. Missing Data Mechanisms under a Selection Model

In our final simulation we generated complete data under a selection model framework, induced missingness according to a range of missing data mechanisms, and applied the

Table 1. Coverage and median confidence interval length for eighteen artificial populations. ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply-imputed data sets. Results over 500 replicates

Population			$n = 100$						$n = 400$					
			Coverage			CI Width			Coverage			CI Width		
ρ	d	λ	ML	PD	MI	ML	PD	MI	ML	PD	MI	ML	PD	MI
0.8	0.1	0	93	94	93	0.46	0.47	0.47	95	94	94	0.23	0.23	0.23
		1	95	95	95	0.47	0.48	0.48	95	95	95	0.24	0.24	0.24
		∞	95	95	96	0.51	0.52	0.52	95	95	94	0.25	0.25	0.25
0.8	0.3	0	94	94	94	0.48	0.49	0.49	96	95	95	0.24	0.24	0.24
		1	96	96	96	0.50	0.51	0.51	96	95	96	0.25	0.25	0.25
		∞	96	95	96	0.55	0.56	0.56	96	95	96	0.27	0.27	0.27
0.8	0.5	0	95	96	95	0.52	0.53	0.53	96	95	95	0.26	0.26	0.26
		1	96	97	96	0.54	0.56	0.55	96	95	97	0.27	0.27	0.27
		∞	97	96	97	0.62	0.64	0.64	97	96	96	0.31	0.31	0.31
0.5	0.1	0	**93**	93	93	0.52	0.53	0.54	94	93	94	0.26	0.26	0.27
		1	95	96	95	0.56	0.59	0.59	95	95	96	0.29	0.28	0.29
		∞	97	95	97	0.84	0.98	0.96	96	95	95	0.41	0.42	0.43
0.5	0.3	0	93	94	94	0.54	0.56	0.56	94	94	94	0.27	0.27	0.28
		1	96	97	96	0.59	0.64	0.64	95	95	96	0.3	0.31	0.31
		∞	96	96	97	1.0	1.2	1.2	95	95	96	0.51	0.52	0.53
0.5	0.5	0	95	95	95	0.58	0.6	0.61	94	94	95	0.29	0.29	0.3
		1	97	97	98	0.64	0.73	0.72	95	96	97	0.33	0.35	0.35
		∞	96	97	96	1.3	1.6	1.6	97	96	96	0.66	0.68	0.69
0.2	0.1	0	93	94	94	0.55	0.56	0.57	94	93	**93**	0.28	0.27	0.28
		1	94	96	96	0.64	0.72	0.72	95	95	95	0.33	0.33	0.34
		∞	94	97	97	2.5	9.9	9.0	94	97	96	1.2	1.7	1.6
0.2	0.3	0	94	95	94	0.57	0.59	0.6	95	94	94	0.29	0.29	0.29
		1	**87**	96	94	0.66	0.98	0.97	95	96	97	0.34	0.38	0.38
		∞	**87**	96	**93**	4.7	23	19	**90**	97	94	2.3	3.4	3.3
0.2	0.5	0	95	95	95	0.62	0.63	0.65	96	95	94	0.31	0.31	0.32
		1	**86**	98	97	0.73	1.7	1.4	95	97	98	0.36	0.45	0.45
		∞	**85**	96	94	7.4	39	32	**90**	97	96	3.5	5.6	5.3

Bolded coverages are below 1.96 simulation standard errors.

PPM sensitivity analysis to evaluate its coverage. The selection model factorization implies marginal normality, while the PPM assumes conditional normality, so in this simulation the distributional assumptions of the PPM are violated. Simulated data were pairs (z_i, y_i) for $i = 1. \ . \ .n$ from a bivariate normal distribution such that $EZ = EY = 1$, $Var(Z) = Var(Y) = 1$, and $Cov(Z, Y) = \rho$. The missing data indicator M was generated according to a logistic model,

$$\text{logit}(\Pr(M = 1|Y, Z)) = \gamma_0 + \gamma_Z Z + \gamma_{Z2} Z^2 + \gamma_Y Y + \gamma_{Y2} Y^2 \tag{20}$$

for eight choices of $\gamma = \{\gamma_0, \gamma_Z, \gamma_{Z2}, \gamma_Y, \gamma_{Y2}\}$ chosen to reflect different nonresponse mechanisms, including both MAR and MNAR scenarios. The choices of γ are displayed in Table 2, and are labeled using conventional linear model notation. These models for M led to approximately 50% missingness in populations where the missing data mechanism was linear in Z and Y, and a slightly lower proportion of missing values in the populations that were quadratic in Z and/or Y. We note that unlike the previous simulations, ρ is specified as the correlation between Y and the covariate Z in the entire sample, not the respondents only. For populations where nonresponse is linear in Z and/or Y, the induced correlation between Y and the proxy X is the same for both respondents and nonrespondents and is equal to ρ. However, when missingness is quadratic in Z and/or Y, the correlation between Y and the proxy is attenuated in the respondents and stronger in the nonrespondents.

There were two different sample sizes, $n = \{100, 400\}$, and three different correlations, $\rho = \{0.8, 0.5, 0.2\}$. We generated 500 replicate data sets for each of the two sample sizes, three correlation levels, and eight nonresponse mechanisms and applied our PPM sensitivity analysis with $\lambda = 0, 1, \infty$ to estimate the mean of Y. As before, we calculated three 95% intervals for the mean of Y (ML, PD, and MI) and computed the actual coverage and length of a nominal 95% interval, noting that $\mu_Y = 1$ for all populations. Bayesian intervals were based on 1,000 draws from the posterior distribution. We also calculated the coverage of the sensitivity analysis as a whole, that is, the percent of the replicates where at least one of the three intervals ($\lambda = 0, 1, \infty$) covered the population mean.

6.3.1. Results

Results from the 24 populations with $n = 400$ are shown in Figures 3a–c; coverage was higher for the smaller sample size since confidence intervals were wider for all values of λ and is not shown. There were four nonresponse mechanisms where, aside from

Table 2. Parameters in the model for M given Z and Y for the third simulation

Model	γ_0	γ_Z	γ_{Z2}	γ_Y	γ_{Y2}
$[Z]$	-0.5	0.5	0	0	0
$[Z^2]$	-1	0	0.5	0	0
$[Y]$	-0.5	0	0	0.5	0
$[Y^2]$	-1	0	0	0	0.5
$[Z + Y]$	-1	0.5	0	0.5	0
$[Z^2 + Y^2]$	-2	0	0.5	0	0.5
$[Z^2 + Y]$	-1.5	0	0.5	0.5	0
$[Z + Y^2]$	-1.5	0.5	0	0	0.5

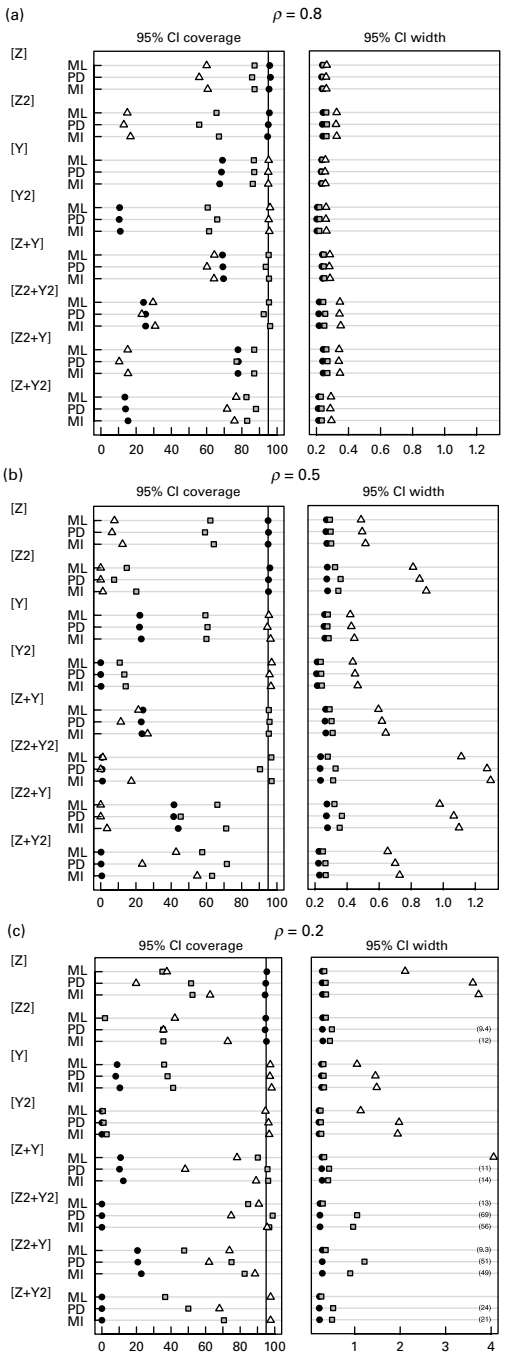

Fig. 3. Coverage and median CI length for twenty-four artificial populations for $\lambda = 0$ (●), $\lambda = 1$ (□), and $\lambda = \infty$ (△), with (a) $\rho = 0.8$; (b) $\rho = 0.5$; (c) $\rho = 0.2$. ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply-imputed data sets. Results over 500 replicates with $n = 400$

distributional assumptions, there was a value of λ that corresponded to the true missingness mechanism: $\lambda = 0$ for mechanisms that depended only on Z ($[Z]$ and $[Z^2]$) and $\lambda = \infty$ for mechanisms that depended only on Y ($[Y]$ and $[Y^2]$). For these populations, coverage was approximately at nominal levels for the corresponding value of λ for all estimation methods and for all three levels of correlation ρ.

The remaining four nonresponse mechanisms had missingness dependent on some combination of Z and Y, the toughest situation for the PPM. For missingness mechanisms $[Z + Y]$ and $[Z^2 + Y^2]$ there is in theory a value of λ that yields the corresponding PPM, however it might not be one of the three in the sensitivity analysis. In these situations the PPM performed well, with at or near nominal coverage for one of the λ values for all three levels of ρ, and at least one interval covering the truth in almost 100% of the replicates. The final two missingness mechanisms, $[Z^2 + Y]$ and $[Z + Y^2]$, do not correspond to any value of λ; these are situations where the PPM is likely to show poor performance. In fact, no method reached nominal coverage levels, except for the weak proxy ($\rho = 0.2$) where confidence interval lengths were extremely large for $\lambda = \infty$ (note the increase in range of the plot). However, as a whole the sensitivity analysis performed better for these populations in that at least one interval covered the truth at closer to nominal levels (at or above nominal levels for ML and MI, at worst 83% for PD, results not shown).

As expected, confidence interval lengths were larger for PD and MI than for ML, particularly for the weaker proxies. However, this did not always lead to improved coverage. By construction the confidence intervals for PD were not symmetrical, and for $\lambda = 1$ and $\lambda = 0$ they were heavily skewed due to draws of $\beta_{xy.y}$ that approached zero. When the point estimates were biased (for example, for $[Z + Y^2]$ and $\lambda = \infty$), the skewness tended to lead to undercoverage for PD, while the symmetric intervals of ML and MI had higher coverage. These differences were exaggerated in the weaker proxies, where better coverage was driven by large confidence interval widths, not by unbiased point estimates.

Overall, the PPM sensitivity analysis performed well in a setting where it was not the "correct" model. This final simulation demonstrated the flexibility of the method, as it had good coverage for a wide range of nonresponse mechanisms, including both linear and quadratic functions of the covariate and the outcome.

7. Applications

We now apply PPM analysis to two real data sets, the Third National Health and Nutrition Examination Survey (NHANES III) and the Ohio Family Health Survey (OFHS). The NHANES III data have been used previously to explore nonresponse (Ezzati-Rice et al. 1993a; Ezzati-Rice et al. 1993b; Khare et al. 1993) and were also released to the public as a multiply-imputed data set under the MAR assumption (U.S. Department of Health and Human Services 2001). The OFHS example is motivated by a current unsolved problem in the OFHS concerning the large amounts of missing information about household income. Current imputation methods for the OFHS assume data are MAR; how different might estimates be if income data are in fact MNAR? The PPM analysis of OFHS also illustrates the importance of an interesting design feature, namely allowing income to be reported in bracketed intervals.

7.1. NHANES III

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. Details of the survey design and data collection procedures are available elsewhere (U.S. Department of Health and Human Services 1994).

For the purposes of our example, we focus on adults interviewed in a personal home interview ($n = 20{,}050$), since questions asked varied considerably by age. Unit nonresponse was created when a portion of these subjects (9.4%) failed to complete a follow-up physical examination at a mobile examination center. Additional item nonresponse occurred when respondents failed to answer questions at either the home interview or the examination. Variables that were fully observed for this sample were limited and included age, gender, race, and household size.

We chose to focus on estimating nonresponse bias for two blood pressure measurements performed at the examination: systolic blood pressure (SBP) and diastolic blood pressure (DBP). The missingness rates were 15% for each measure. 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.

Linear regression was used to create the proxies for each blood pressure measurement, using the fully observed variables listed previously as well as the design weight and indicators for strata and primary sampling units. The final models were chosen using backward selection starting from a model that contained all second-order interactions. We note that the regression models that create the proxy are unweighted, that is, the design weights are incorporated only through inclusion as covariates. Systolic blood pressure displayed a large correlation between outcome and the proxy ($\hat{\rho} = 0.61$) but also a large deviation in the proxy ($d = 0.99, d^* = 0.077$), thus falling in what we would consider the second most desirable situation as described in Section 2. Diastolic blood pressure had a weaker proxy ($\hat{\rho} = 0.38$) but also a smaller deviation ($d = -0.026, d^* = -0.0052$).

Since NHANES III has a complex survey design, estimates of the mean and confidence intervals for $\lambda = (0, 1, \infty)$ were obtained using multiple imputation with design-based estimators of the mean using the survey weights. A total of 20 multiply-imputed data sets were created for each outcome. 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 Figures 4 and 5. The sensitivity analysis shows that the choice of λ has a larger impact on the mean estimate for SBP than for DBP. Assuming MAR would result in significantly different mean estimates for SBP than assuming missing not at random. The analysis reveals that, if missingness on SBP is driven by the value of SBP itself, nonrespondents have considerably higher SBP than respondents and the overall mean is pulled up dramatically under this assumption, especially considering only 15% of the subjects are nonrespondents. Since the proxy is relatively strong, the length of the intervals does not drastically increase even under extreme MNAR ($\lambda = \infty$). For DBP, the choice of missingness mechanism has little impact on the mean estimates, since the deviation is very small. The size of the intervals does increase as $\lambda \to \infty$, but since the sample size is large the increase is not as dramatic as was seen in the simulations.

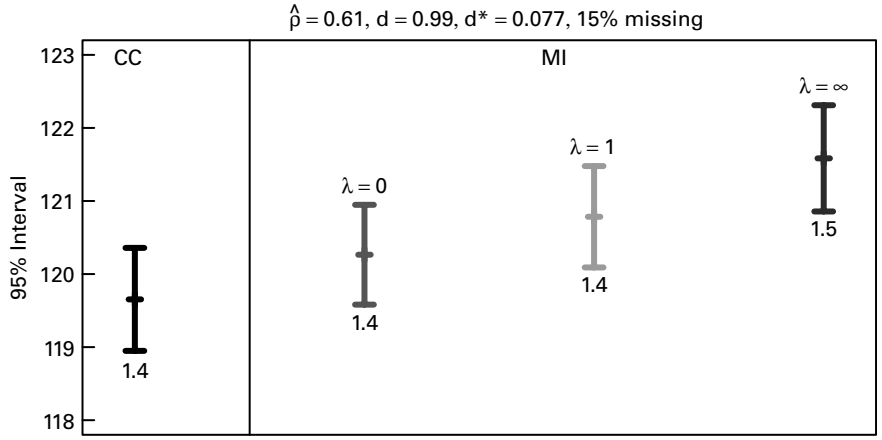

Fig. 4. Estimates of mean SBP for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. CC: Complete case; MI: 20 multiply-imputed data sets

Table 3 shows the estimates of FMI for each outcome under each missingness mechanism for multiple imputation analyses that ignore design weights (FMI) and those that incorporate them (FMIwt). The weighted estimates of FMI are considerably smaller than the unweighted estimates; the same between-imputation variance is coupled with increased within-imputation variability due to incorporation of the sample design. For the purposes of assessing the overall picture of nonresponse bias, it may be easiest to look at the unweighted estimates, though the same conclusions could be drawn looking at the spread of the weighted FMI values. For SBP, FMI remains relatively low even when assuming MNAR. When $\lambda = \infty$ the FMI is 21%, larger than the nonresponse rate of 15% but not drastically larger. This indicates that the survey data contain strong information for correcting potential bias. In contrast, the FMI values for DBP indicate that our ability to assess nonresponse bias for this outcome is weaker than for SBP. The nonresponse rate is also 15%, but the FMI values range from 17% to 71%. That high upper bound indicates that our assessment of bias

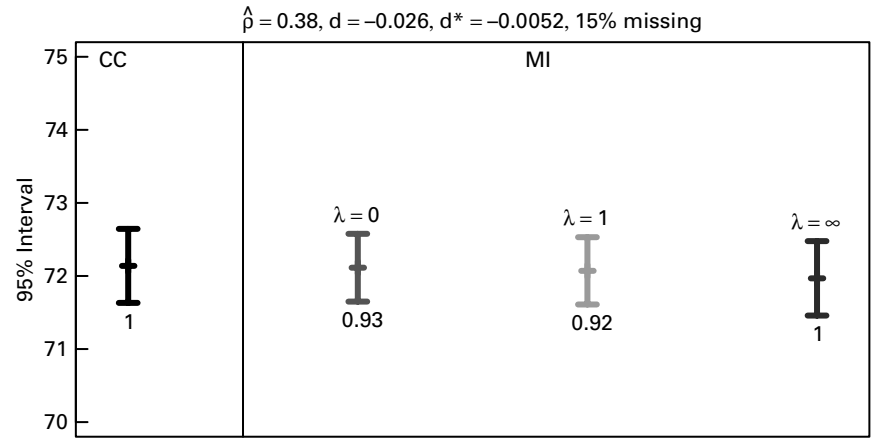

Fig. 5. Estimates of mean DBP for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. CC: Complete case; MI: 20 multiply-imputed data sets

Table 3. Fraction of missing information (FMI) estimates from NHANES III data for three outcomes. SBP = systolic blood pressure; DBP = diastolic blood pressure. FMIwt denotes variance estimation incorporating the survey design

Outcome	Missing (%)	$\hat{\rho}$	d	d^*	λ	FMI (%)	FMIwt (%)
SBP	15	0.61	0.99	0.077	0	8.1	2.0
					1	11	6.1
					∞	21	8.2
DBP	15	0.38	−0.026	−0.0052	0	17	6.3
					1	23	6.7
					∞	71	24

is limited by weak covariate information. We note that without looking at the FMI values, the plots may be difficult to interpret. With a large sample size in a survey such as NHANES, the interval lengths do not appear drastically inflated under MNAR as they did in the simulations. However, once we look at the FMI we see that there is a lot of uncertainty associated with these estimates that may be disguised by large sample sizes.

7.2. OFHS

The 2008 Ohio Family Health Survey was one of the largest state-sponsored health surveys in the United States. The sampling design was a stratified (by county), list-assisted random digit dialing sample of Ohio's noninstitutionalized population, with oversampling of certain counties and minority populations. A cell-phone supplement was added mid-way through the project, with this sample treated as a separate stratum. Clusters were defined as a household/family, and within each cluster one adult was randomly selected to participate. Details on the design and implementation of the 2008 OFHS are available elsewhere (Duffy and Muzzy 2008).

A total of $n = 50,944$ adults provided responses to some or all of the survey. Missing values for key variables such as gender, age, race/ethnicity, education level, and housing tenure were singly imputed by the OFHS using a hot deck procedure, since these key variables were necessary for construction of design weights. Missing values for Medicaid status and income were also imputed by OFHS using a stochastic regression imputation approach, sequentially so that the imputation for income could use information on Medicaid status. However, the imputation of income assumed that income was MAR. The goal in this example is to evaluate the potential impact on estimates of mean income if data were missing not at random. We applied PPM analysis to the (preimputation) income variable, using the imputed (by OFHS) data as the starting point. Covariates that were fully observed (or completed by imputation) and used in the analysis were county, household size, age, gender, race, education level, insurance/Medicaid status, and tenure in current home.

Income was asked multiple ways in the OFHS. First, adults were asked to report their income as an exact value. For subjects who refused or could not provide an exact response, interviewers then prompted subjects to select from a set of income categories. There was no information on income for 9.2% of subjects ($n = 4,707$), categorical income only on

14.9% of subjects ($n = 7,583$), and continuous income information for 75.9% of subjects ($n = 38,654$). Since the continuous income measure was highly skewed, it was log-transformed for the PPM analysis, with results back-transformed for display in figures.

Initially, we carried out PPM analysis ignoring the categorical information. As with the NHANES example, linear regression was used to create the proxies for income, using the variables previously described as well as sampling weights, plus second-order interactions where significant. The proxy for (log) income was moderate, with $\hat{\rho} = 0.51$. The deviation was small, $d^* = -0.022$, but should be interpreted with caution due to the log-transformation. As with NHANES, the multiple imputation version of PPM analysis was applied for $\lambda = \{0, 1, \infty\}$, with a total of 20 multiply-imputed data sets. Estimates of means and variances on the log-scale (geometric means) were obtained using design-based estimators and were then back-transformed, producing estimates of the median on the original scale.

To incorporate the categorical income data, PPM analysis was repeated using the same proxy as described above, but modifying the multiple imputation algorithm. Imputed values were drawn in the usual manner, using (19), but forcing the draw to fall between the provided bounds from the categorical income question. If the drawn value did not fall between the bounds, the value was discarded and new value drawn. The process was repeated until the draw met the constraint.

Estimates of the median and 95% intervals for income under the PPM analysis are shown in Figure 6, after back-transforming point estimates and interval bounds. Results for the analysis that ignored the additional categorical data are shown alongside the results when the bounding information was used, in addition to the complete case estimate. As expected, if missingness on income were truly MNAR, then the estimate under MAR overestimates the median for both PPM analyses. However, this downward shift is attenuated with the incorporation of the categorical information. There is less difference among the estimates under $\lambda = \{0, 1, \infty\}$ when the additional information is incorporated, but the choice of missingness mechanism does still have an effect. Overall, there is evidence that if missingness were not at random, then the survey estimates of income are optimistic, though the shift is small even at the extreme end ($\lambda = \infty$) compared to the MAR case, with a shift in estimates of only about $1,000 per year for the analysis that used the categorical information.

For the analysis that ignores the categorical data the (unweighted) FMI values for the OFHS income data are 17%, 22%, and 39% for $\lambda = 0, 1,$ and ∞ respectively. The nonresponse rate was 24%, indicating a gain from imputation under MAR, and only a moderate penalty for assuming extreme MNAR. Since the analysis that includes the categorical bounds uses more information, and only 9% of subjects are completely missing (no income information at all), we expect to see lower FMI values. This analysis yields FMI values of 9%, 9%, and 30% for $\lambda = 0, 1,$ and ∞ respectively. The FMI values from both analysis methods indicate that the OFHS data contain moderately strong information for correcting potential bias.

One interesting feature of the analysis is that the estimate under MAR when using the bounds is very close to the complete case estimate, while ignoring the bounds leads to an MAR estimate that is smaller than the complete-case estimate. This finding parallels the results from the original imputation undertaken by the OFHS, where the categorical information was handled using ordinal regression. Clearly this emphasizes the importance

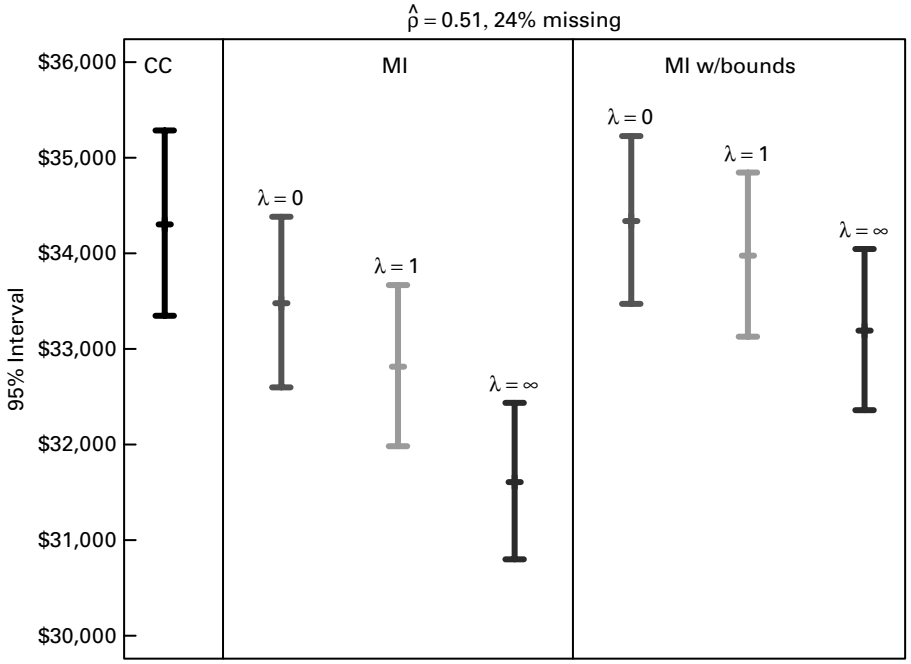

Fig. 6. Estimates of income (back-transformed) for $\lambda = (0, 1, \infty)$ based on OFHS data. CC: Complete case; MI: 20 multiply-imputed data sets; MI w/Bounds: 20 multiply-imputed data sets, imposing bounds on imputed values

of including all available information when performing imputation and the PPM sensitivity analysis.

8. Discussion

The PPM analysis for survey nonresponse has the following attractive features. It integrates all the various components of nonresponse noted in the introduction into a single sensitivity analysis. 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 noniterative, and it readily allows complex design features to be incorporated in the within-imputation component of variance. The MI model for the between-imputation component proposed here does allow survey design variables to be included as predictors in the regression for the proxy measure, but does not include random effects to model clustering of the sample – a more principled extension would incorporate such design features directly into the imputation model.

Since most nonresponse adjustments applied in the survey setting assume MAR, and MAR is often a strong and questionable assumption with unit nonresponse, an important advantage of PPM analysis is that it does not assume MAR. We believe that 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 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. Recent work by Kreuter et al. (2010) evaluated the strength of alternative data sources, such as interviewer observations and other paradata. Correlations between paradata and outcome measures were generally low, with no estimated correlations above 0.5. Under the PPM framework, even these weaker proxies can provide information for assessing the potential for bias, and the ranking based on ρ and d can still be used to compare across various different outcome measures within a survey. We agree with Kreuter et al. (2010) that emphasis at the design stage should also be on collection of strong auxiliary data to help evaluate and adjust for potential nonresponse, not solely on obtaining the highest possible response rate.

The PPM method employs a sensitivity analysis to assess deviations from MAR, in contrast with some selection model approaches that attempt to use the data to estimate parameters that capture deviations from MAR (e.g., Heckman 1976; Diggle and Kenward (1994). These models are technically identified in situations where pattern-mixture models are not, but estimation of the MNAR parameters is still based on strong and unverifiable structural and distributional assumptions, and a substantial body of researchers believe that a sensitivity analysis is the right approach. The assumptions about deviations from MAR are more transparent in the pattern-mixture factorization, since differences between respondents and nonrespondents are directly modeled (Little and Rubin 2002). The PPM sensitivity analysis only varies one sensitivity parameter, λ, but still manages to capture a range of assumptions on the missing data mechanism. Both the standard and reverse regression estimators are contained in the PPM analysis framework.

A limitation of PPM analysis is that by reducing the auxiliary data to the single proxy X^*, the coefficient λ is not associated with any particular covariate and hence is difficult to interpret, since the effects on missingness on individual covariates Z_j are lost. The pattern-mixture model proposed by Daniels and Hogan (2000) in the context of longitudinal data, uses a location-scale parameterization to model differences in the marginal distribution of (Y, Z) for respondents and nonrespondents. This model is more readily interpretable than our approach, but it is very underidentified – even with a single Z it has three unidentified parameters – and additional specification is needed to limit the number of parameters to be varied in a sensitivity analysis. Modeling the conditional distribution of Y given Z for respondents and nonrespondents, as in PPM analysis, focuses more directly on the distribution that is not identified, namely the distribution of Y given Z for nonrespondents. A reasonable alternative to the PPM model allows the intercept of this regression to differ for respondents and nonrespondents but the regression coefficients and residual variance to be the same. This results in a simple nonignorable model with just one sensitivity parameter, the difference in intercepts. However, it is hard to assess how much of a difference in intercepts is plausible, and this model does not readily distinguish between strong and weak proxies of Y. Allowing the regression coefficients of individual Z_j's in this model to differ for respondents and nonrespondents provides more flexibility, at the expense of adding more unidentified parameters, particularly when there is more than

one covariate. Our approach trades off interpretability for parsimony, allowing a single parameter to model deviations from MAR.

Another limitation of our analysis is that it focuses 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 or R-indicators, cannot capture this dimension of the problem. Presenting the fraction of missing information over a range of key survey variables and a range of values of λ seems valuable for capturing the full scope of the potential nonresponse bias.

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 nonnormal outcomes. Extensions to categorical variables are 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. Another potential extension would be to investigate the use of PPM analysis for describing relationships between two variables, for example a regression coefficient of Y on X. Under the PPM framework, if there is missingness in only one of the variables, then one could envision using the multiple imputation framework to impute Y, conditioning on X, and evaluating the impact on inference for each value of λ. However, if both X and Y are subject to missingness, the extension is not as straightforward, and we are back to the problem of general patterns of missing data. For missing data in X (assuming MAR), one idea is to take a chained equations approach with the imputation of Y replaced by the PPM model imputations, for each value of λ. This approach warrants further investigation. Our future work on PPM analysis will focus on developing these extensions.

9. References

Bethlehem, J. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. In Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). Chapter 18. New York: Wiley, 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.

Daniels, M.J. and Hogan, J.W. (2000). Reparameterizing the Pattern Mixture Model for Sensitivity Analyses under Informative Dropout. Biometrics, 56, 1241–1248.

Diggle, P. and Kenward, M.G. (1994). Informative Drop-Out in Longitudinal Data Analysis. Applied Statistics, 43, 49–93.

Duffy, T. and Muzzy, S. (2008). 2008 Ohio Family Health Survey Methodology Report. Technical report, Macro.

Ezzati-Rice, T.M., Fahimi, M., Judkins, D., and Khare, M. (1993a). Serial Imputation of NHANES III With Mixed Regression and Hot-Deck Imputation. In Proceedings of the American Statistical Association, Survey Research Methods Section, 292–296.

Ezzati-Rice, T.M., Khare, M., Rubin, D.B., Little, R.J.A., and Schafer, J.L. (1993b). A Comparison of Imputation Techniques in the Third National Health and Nutrition Examination Survey. In Proceedings of the American Statistical Association, Survey Research Methods Section, 303–308.

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

Groves, R.M. (2006b). 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.

Khare, M., Little, R.J.A., Rubin, D.B., and Schafer, J.L. (1993). Multiple Imputation of NHANES III. In Proceedings of the American Statistical Association, Survey Research Methods Section, 297–302.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys. Journal of the Royal Statistical Society, Series A, 173, 389–407.

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

Little, R.J.A. (1994). A Class of Pattern-Mixture Models for Normal Incomplete Data. Biometrika, 81, 471–483.

Little, R.J.A. (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 (Second Edition). New York: Wiley.

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

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. Technical report, Executive Office of the President of the United States of America.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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

Rubin, D.B. (1978). Multiple Imputation in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse. In Proceedings of the American Statistical Association, Survey Research Methods Section, 20–34.

Särndal, C.E. and Lundström, S. (2005). Estimation in Surveys with Nonresponse. New York: Wiley.

Särndal, C.E. and Lundström, S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. Journal of Official Statistics, 24, 167–191.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the Representativeness or Survey Response. Survey Methodology, 35, 101–113.

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

U.S. Department of Health and Human Services (2001). Third National Health and Nutrition Examination Survey (NHANES III, 1988–1994): Multiply Imputed Data Set. CD-ROM, Series 11, No. 7A. Technical report, National Center for Health Statistics, Centers for Disease Control and Prevention.

Wagner, J. (2010). The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data. Public Opinion Quarterly, 74, 223–243.

Received April 2010
Revised February 2011