Extensions of 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 binary 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 continuous proxy variable X that has the highest correlation with Y, estimated from a probit regression analysis of respondent data. We extend our previously proposed proxy-pattern mixture analysis for continuous outcomes to the binary outcome using a latent variable approach, applying a pattern-mixture model for the joint distribution of the proxy X and the underlying latent variable for the outcome Y. Methods are demonstrated through simulation and with data from the third National Health and Nutrition Examination Survey (NHANES III).

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

1. Introduction

Response rates for large-scale surveys have been steadily declining in recent years (Curtain et al. 2005), increasing the need for methods to analyze the impact of nonresponse on survey estimates. 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. Current methods to handle nonresponse in surveys have tended to focus on a subset of these components, however, the impact of nonresponse cannot be fully understood without all three pieces. In addition, historically the focus has been on situations were data are assumed to be missing at random, with less attention paid to the case when missingness may be not at random, that is, depend on the unobserved outcome itself (Rubin 1976). In this paper we propose a method for estimating population proportions in survey samples with nonresponse that includes but does not assume ignorable missinginess.

A limited amount of work has been done in the area of nonignorable nonresponse for categorical outcomes in survey data. Some examples include Stasny (1991), who used a hierarchical Bayes nonignorable selection model to study victimization in the National Crime Survey. Extensions of this approach by Nandram and Choi (2002a) and Nandram and Choi (2002b) use continuous model expansion to center the nonignorable model on an ignorable model, in the manner of Rubin (1977). Similar methods are developed for multinomial outcomes in Nandram et al. (2002) and Nandram et al. (2005) and used to study health outcomes in the third National Health and Nutrition Examination Survey (NHANES III). The main difference between our proposed approach and these previous methods is the method of modeling the missing data. There are two general classes of models for incomplete data, selection models and pattern-mixture models (Little and Rubin 2002). Previous work on nonresponse models in surveys has tended to favor the selection model; we use a pattern-mixture approach. The pattern-mixture approach requires explicit assumptions on

*Division of Biostatistics, The Ohio State University, Columbus, Ohio 43210

†Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, 48109

the missing data mechanism and naturally leads to a sensitivity analysis, whereas the selection model approach requires strong distributional assumptions to (often weakly) identify parameters. In addition, methods for categorical nonresponse have tended to be limited to the case when auxiliary data are also categorical. However, auxiliary variables may be continuous; our proposed method does not require that continuous variables be categorized before inclusion in the model.

The work in this paper is an extension of our previously described proxy pattern-mixture analysis (PPMA) for a continuous outcome; In Section 2 we briefly review the continuous outcome PPMA before describing its extension to binary outcomes in Section 3. Section 4 discusses three different estimation approaches, maximum likelihood, a Bayesian approach, and multiple imputation, and the sensitivity of each method to model misspecification. These methods are illustrated first through simulation in Section 5 and then by application to NHANES III data in Section 6. Section 7 presents some concluding remarks.

2. Review of the Proxy Pattern-Mixture Model

Proxy pattern-mixture analysis was developed for the purpose of assessing nonresponse bias for estimating the mean of a continuous 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.

To reduce dimensionality and for simplicity we reduce the covariates Z to a single proxy variable X that has the highest correlation with Y, estimated from a regression analysis of Y on Z using respondent data. 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. In addition to the strength of the proxy as measured by ρ, an important factor is the deviation from missing completely at random (MCAR) as measured by the difference between the overall mean of the proxy and the respondent mean of the proxy, $d = \bar{x} - \bar{x}_R$. The distribution of X for respondents and nonrespondents provides the main source of information for assessing nonresponse bias for Y. 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. This allows insight into whether missingness may be not at random (NMAR).

Specifically, we let M denote the missingness indicator, such that $M = 0$ if Y is observed and $M = 1$ if Y is missing. We assume that the joint distribution of $[Y, X, M]$ follows the bivariate pattern-mixture model discussed in Little (1994). This model is underidentified, since there is no information on the conditional normal distribution for Y given X for nonrespondents ($M = 1$). However, Little shows that the model can be identified by making assumptions about how missingness of Y depends on Y and X. For the proxy pattern-mixture we assume that,

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

where X^* is the proxy variable X scaled to have the same variance as Y in the respondent population. By a slight modification of the arguments in Little (1994), the resulting maximum likelihood estimate of the overall mean of Y is,

$$\hat{\mu}_y = \bar{y}_R + \frac{\lambda + \hat{\rho}}{\lambda\hat{\rho} + 1}\sqrt{\frac{s_{yy}}{s_{xx}}}(\bar{x} - \bar{x}_R), \tag{2}$$

where $\bar{x}_R$ and $\bar{y}_R$ are the respondent means of X and Y, s_{xx} and s_{yy} are the respondent sample variances of X and Y, and $\bar{x}$ is the overall sample mean of X.

The parameter λ is a sensitivity parameter; there is no information in the data with which to estimate it. Different choices of λ correspond to different assumptions on the missing data mechanism. We assume that λ is 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), $g(\hat{\rho}) = (\lambda + \hat{\rho})/(\lambda\hat{\rho} + 1)$ varies between $\hat{\rho}$ and $1/\hat{\rho}$. 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 (2) reduces to the standard regression estimator. 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 (2) 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 $\hat{\rho}$.

For assessing potential nonresponse bias in the mean of Y, 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.

3. Extension of PPMA to a Binary Outcome

The proxy pattern-mixture analysis described above strictly only applies to continuous survey variables, where normality is reasonable. However, categorical outcomes are ubiquitous in sample surveys. In this section we extend PPMA to binary outcomes using a latent variable approach. Let Y_i now denote the value of a partially missing binary survey outcome, and $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$ denote the values of p fully observed covariates for unit i in the sample. As before, 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 $i = r+1, \ldots, n$. Of interest is the proportion of units in the population with $Y = 1$.

For simplicity and to reduce dimensionality, we replace Z by a single continuous proxy variable X, estimated by a probit regression of Y on Z using the respondent data,

$$\Pr(Y = 1|Z, M = 0) = \Phi(\alpha_0 + \alpha Z). \tag{3}$$

We take $X = \hat{\alpha}_0 + \hat{\alpha}Z$ to be the linear predictor from the probit regression, rather than the predicted probability, so that its support is the real line. The regression coefficients α are subject to sampling error, so in practice X is estimated rather than known. The choice of the probit link, rather than alternatives such as the logit link, is due to the latent variable motivation of probit regression. We assume that Y is related to a continuous normally distributed latent variable U through the rule that $Y = 1$ when the latent variable $U > 0$.

The latent (respondent) data are then related to the covariates through the linear regression equation, $U = \alpha_0 + \alpha Z + \epsilon$, where $\epsilon \sim N(0, 1)$.

This latent variable approach motivates application of the normal proxy pattern-mixture (PPM) model to the latent variable U and proxy X. If we could observe U for the respondents, application of the PPM model would be straightforward. Taking M to be the missing data indicator, we assume that the joint distribution of $[U, X, M]$ follows the bivariate pattern-mixture model:

$$\begin{aligned}(U, X|M = m) &\sim N_2\left((\mu_u^{(m)}, \mu_x^{(m)}), \Sigma^{(m)}\right) \\ M &\sim Bernoulli(1 - \pi) \\ \Sigma^{(m)} &= \begin{bmatrix} \sigma_{uu}^{(m)} & \rho^{(m)}\sqrt{\sigma_{uu}^{(m)}\sigma_{xx}^{(m)}} \\ \rho^{(m)}\sqrt{\sigma_{uu}^{(m)}\sigma_{xx}^{(m)}} & \sigma_{xx}^{(m)} \end{bmatrix},\end{aligned} \tag{4}$$

where N_2 denotes the bivariate normal distribution. Note that the parameter $\rho^{(m)}$ is the correlation between the latent variable U and the constructed proxy X. As with the continuous outcome PPM model the parameters $\mu_u^{(1)}$, $\sigma_{uu}^{(1)}$, and $\rho^{(1)}$ are unidentifiable without further model restrictions. Since U is completely unobserved, $\sigma_{uu}^{(0)}$ is also not identifiable and without loss of generality can be fixed at an arbitrary value. Following convention we set $\sigma_{uu}^{(0)} = 1/(1 - {\rho^{(0)}}^2)$ so $\text{Var}(U|X, M = 0) = 1$.

We identify the model by making assumptions about how missingness of Y depends on U and X. As with the continuous outcome PPM model, we modify the arguments in Little (1994) and assume that

$$\Pr(M = 1|U, X) = f(X\sqrt{\frac{\sigma_{uu}^{(0)}}{\sigma_{xx}^{(0)}}} + \lambda U) = f(X^* + \lambda U), \tag{5}$$

where X^* is the proxy variable X scaled to have the same variance as U in the respondent population. An important feature of this mechanism is that when $\lambda > 0$, i.e. under NMAR, the missingness in the binary outcome Y is being driven by X and by the completely unobserved latent U. This allows for a "smooth" missingness function in the sense that conditional on X the probability of missingness may lie on a continuum instead of only taking two values (as would be the case if missingness depended on Y itself). Of primary interest is the marginal mean of Y, which is given by,

$$\mu_y = \Pr(Y = 1) = \Pr(U > 0) = \pi\Phi\left(\mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}\right) + (1 - \pi)\Phi\left(\mu_u^{(1)}/\sqrt{\sigma_{uu}^{(1)}}\right), \tag{6}$$

where $\Phi(\cdot)$ denotes the standard normal CDF.

3.1 Summary of Evidence about Nonresponse Bias

The information about nonresponse bias in the mean of Y is contained in the strength of the proxy as measured by $\rho^{(0)}$ and the deviation in the proxy mean, $d = \bar{x} - \bar{x}_R$. Strong proxies (large $\rho^{(0)}$) and small deviations (small d) lead to decreased uncertainty and higher precision in estimates, even under NMAR, while weak proxies (low $\rho^{(0)}$) and large deviations (large d) lead to increased uncertainty, especially when missingness depends on Y. In the case of the continuous outcome, both $\rho^{(0)}$ and d were directly interpretable, since $\hat{\rho}^{(0)}$ was the square root of the R^2 from the regression model that built the proxy X and the deviation d was on the same scale as the (linear) outcome Y. With a binary outcome,

we lose these neat interpretations of $\rho^{(0)}$ and d, though their usefulness as markers of the severity of the nonresponse problem (d) and our ability to make adjustments to combat the problem ($\rho^{(0)}$) remains. The information about nonresponse bias in Y is contained in X, with X now a proxy for the latent U instead of the partially observed outcome Y itself.

Another issue unique to the binary case (and subsequent extension to ordinal Y) is that the size of the nonresponse bias in Y, i.e. the difference in mean between respondent and overall means, depends not only on the size of the deviation on the latent scale (d) but also on the respondent mean itself. In the continuous case, the bias in $\bar{y}_R$ is a linear function of d (see (2)); a deviation d has the same (standardized) effect on the overall mean regardless of the value of $\bar{y}_R$. However, in the binary case the deviation is on the latent scale, and only the bias in U is location-invariant. When transformed to the binary outcome, different d values will lead to different size biases, depending on the respondent mean of Y. The use of the standard normal CDF to transform U to Y drives this; the difference $\Phi(a+d)-\Phi(a)$ is not merely a function of d but also depends on the value of a.

4. Estimation Methods

4.1 Maximum Likelihood

Maximum likelihood (ML) estimators for the distribution of U given X for nonrespondents follow directly from the continuous outcome PPM model,

$$\begin{aligned}
\hat{\mu}_u^{(1)} &= \hat{\mu}_u^{(0)} + g \times \sqrt{\frac{\hat{\sigma}_{uu}^{(0)}}{\hat{\sigma}_{xx}^{(0)}}}(\bar{x}_{NR} - \bar{x}_R) \\
\hat{\sigma}_{uu}^{(1)} &= \hat{\sigma}_{uu}^{(0)} + g^2 \times \frac{\hat{\sigma}_{uu}^{(0)}}{\hat{\sigma}_{xx}^{(0)}}(\hat{\sigma}_{xx}^{(1)} - \hat{\sigma}_{xx}^{(0)}) \\
g &= \frac{\lambda + \hat{\rho}^{(0)}}{\lambda\hat{\rho}^{(0)} + 1}.
\end{aligned} \tag{7}$$

Plugging these estimates into (6) yields the ML estimate of the mean of Y. The ML estimates of the parameters of the distribution of X are the usual estimators, however, estimators for $\mu_u^{(0)}$ and $\rho^{(0)}$, and therefore $\sigma_{uu}^{(0)}$, are not immediately obvious since the latent U is unobserved even for respondents. To obtain these estimates, we note that the correlation $\rho^{(0)}$ is the biserial correlation between the binary Y and continuous X for the respondents. Maximum likelihood estimation of the biserial correlation coefficient was first studied by Tate (1955a,b), who showed that a closed form solution does not exist. The parameters $\rho^{(0)}$ and $\omega^{(0)} = \mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}$ (referred to as the cutpoint) must be jointly estimated through an iterative procedure such as a Newton-Raphson type algorithm. It is important to note that the ML estimate of $\omega^{(0)}$ is not the inverse probit of the respondent mean of Y, i.e. the ML estimate of the mean of Y for respondents is not $\bar{y}_R$.

An alternative method of estimating the biserial correlation coefficient is the *two-step method*, proposed by Olsson et al. (1982) in the context of the polyserial correlation coefficient. In the first step, the cutpoint $\omega^{(0)}$ is estimated by $\hat{\omega}^{(0)} = \Phi^{-1}(\bar{y}_R)$, so that the ML estimate of the respondent mean of Y is $\bar{y}_R$. Then a conditional maximum likelihood estimate of $\rho^{(0)}$ is then computed, given the other parameter estimates. This method is computationally simpler than the full ML estimate, and also has the attractive property of returning the logical estimate $\hat{\mu}_y^{(0)} = \bar{y}_R$.

The large sample variance of the full ML estimate of μ_y is obtained through Taylor series expansion and inversion of the information matrix. The properties of the two-step

estimator are not well studied, so variance estimates are obtained with the bootstrap.

4.2 Bayesian Inference

The ML estimate ignores the uncertainty inherent in the creation of the proxy X. An alternative approach is to use a Bayesian framework that allows incorporation of this uncertainty. Since U is unobserved, we propose using a data augmentation approach. We place noninformative priors on the regression parameters α and use a Gibbs' sampler to draw the latent U for respondents (Albert and Chib 1993). Conditional on α (and therefore on the created proxy X), U follows a truncated normal distribution,

$$\begin{aligned}(U|Y,\alpha,M=0) = (U|Y,X,M=0) \sim N(X,1) = N(\alpha Z,1)\\ \text{truncated at the left by 0 if } Y=1 \text{ and at the right by 0 if } Y=0.\end{aligned} \tag{8}$$

Then given the augmented continuous U we draw α from its posterior distribution, which also follows a normal distribution,

$$(\alpha|Y,U,M=0) \sim N((Z^TZ)^{-1}Z^TU,(Z^TZ)^{-1}), \tag{9}$$

and recreate the proxy $X = \alpha Z$.

This data augmentation allows for straightforward application of the Bayesian estimation methods for continuous PPMA. For a chosen value of λ, we apply the PPM algorithm as described in Andridge and Little (2009) to the pair (X,U) to obtain draws of the parameters of the joint distribution of X and U. Since U is unobserved even for the respondents, after each draw of the parameters from the PPM model, X is recreated for the entire sample and U is redrawn for the respondents given the current set of parameter values as described in the data augmentation approach above. Note that this does not require a draw of the latent data for nonrespondents. Draws from the posterior distribution of μ_y are obtained by substituting the draws from the Gibbs' sampler into (6).

4.3 Multiple Imputation

An alternative method of inference is multiple imputation (Rubin 1978). For a selected λ we create K complete data sets by filling in missing Y values with draws from the posterior distribution, based on the pattern-mixture model. For a given draw of the parameters $\phi = (\mu_u^{(1)}, \mu_x^{(1)}, \sigma_{uu}^{(1)}, \sigma_{xx}^{(1)}, \rho^{(1)})$ from their posterior distribution as Section 4.2, we draw the latent U for nonrespondents based on the conditional distribution,

$$[u_i|x_i,m_i=1,\phi_{(k)}] \sim N\left(\mu_{u(k)}^{(1)} + \frac{\sigma_{ux(k)}^{(1)}}{\sigma_{xx(k)}^{(1)}}\left(x_i - \mu_{x(k)}^{(1)}\right), \sigma_{uu(k)}^{(1)} - \frac{{\sigma_{ux(k)}^{(1)}}^2}{\sigma_{xx(k)}^{(1)}}\right) \tag{10}$$

where the subscript (k) denotes the kth draws of the parameters. In order to reduce autocorrelation between the imputations due to the Gibbs' sampling algorithm for drawing the parameters, we thin the chain for the purposes of creating the imputations. The missing y_i are then imputed as $y_i = I(u_i > 0)$, where $I()$ is an indicator function taking the value 1 if the expression is true. 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.

As with the continuous PPMA, 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 $\bar{Y}_k$.

4.4 Sensitivity to a Non-normally Distributed Proxy

A crucial assumption of the PPM model for both continuous and binary outcomes is that of bivariate normality of X and Y or U. The continuous outcome PPM model is relatively robust to departures from this assumption and only relies on linear combinations of first and second moments in estimating the mean of Y. However, for binary outcomes the normality assumption plays a more crucial role, made clear with a simple example. Suppose the proxy X is normally distributed in the respondent population, with $[X|M=0] \sim N(\mu_x^{(0)}, \sigma_{xx}^{(0)})$. We assume that, for respondents, the latent variable $U = X + e$ where $e \sim N(0,1)$, such that $\Pr(Y=1|M=0) = \Pr(U>0|M=0)$. Then the conditional and marginal respondent distributions of U along with the mean of Y are given by,

$$\begin{aligned}
[U|X, M=0] &\sim N(X, 1) \\
[U|M=0] &\sim N(\mu_x^{(0)}, 1+\sigma_{xx}^{(0)}) \\
\mu_y^{(0)} = \Pr(U>0|M=0) = \Phi\left(\mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}\right) &= \Phi\left(\mu_x^{(0)}/\sqrt{1+\sigma_{xx}^{(0)}}\right)
\end{aligned}$$

However, if the distribution of X, $f_X(x)$, is not normal, then the conditional distribution $[U|X, M=0]$ is the same but the marginal distribution is no longer normal. Now $\Pr(U > 0) = \int_0^\infty f_U(u)\,\mathrm{d}u$ where $f_U(u)$ is the convolution of the error distribution $N(0,1)$ and $f_X(x)$. Thus even the estimate of the respondent mean of Y will be biased, despite the fact that Y is fully observed for the respondents.

Even though PPMA can provide unbiased estimates of the mean and variance of U in the case when Z is not normally distributed (like the continuous PPMA), the transformation to the mean of Y is only accurate when Z is normally distributed. Both the full ML estimation and Bayesian methods will produce biased estimates of μ_y if X deviates away from normality. The two-step ML method is less sensitive to non-normality, since it estimates $\mu_y^{(0)}$ by $\bar{y}_R$. Multiple imputation also is less sensitive to departures from normality since imputation is based on the conditional distribution $[U|X, M]$ which is normal by definition of the latent variable and is not affected by non-normal X.

We propose modifying the Bayesian method to attempt to reduce sensitivity to deviations from normality in the proxy X. The modification is an extension of the multiple imputation approach: at each iteration of the Gibbs' sampler, the latent U for nonrespondents is drawn conditional on the current parameter values, and the subsequent draw of $\mu_y^{(1)}$ is taken to be $\mu_y^{(1)} = \frac{1}{n-r}\sum_{i=r+1}^{n} I(U_i > 0)$. A similar method of obtaining an estimator for the respondent mean does not work, as draws of U for the respondents in the Gibbs' sampler are conditional on the observed Y and thus the resulting draw will always be $\bar{y}_R$. To avoid this, we can take one of two approaches. An obvious extension is to redraw the latent U conditional only on the current draws of the proxy and the parameters, with the subsequent draw of $\mu_y^{(0)}$ is taken to be $\mu_y^{(0)} = \frac{1}{n-r}\sum_{i=r+1}^{n} I(U_i > 0)$. The drawback of this method (Modification 1) is that variances may actually be overestimated since we are essentially imputing the observed binary outcome Y for the respondents. Alternatively, we can use the average of the predicted probabilities for the respondents as a draw of $\mu_y^{(0)}$, i.e. $\frac{1}{r}\sum_{i=1}^{r} \Phi^{-1}(X_i)$. This is actually a draw of the conditional mean of Y (conditional on X)

and so its posterior distribution will underestimate the variance of $\mu_y^{(0)}$. To combat this we take a bootstrap sample of the X_i before calculating the mean of the predicted probabilities (Modification 2).

5. Simulation Study

We now describe a simulation study designed to numerically illustrate the taxonomy of evidence concerning bias based on the strength of the proxy (ρ) and the deviation of its mean (d^*). We created a total of eighteen artificial data sets in a 3x3x2 factorial design with a fixed nonresponse rate of 50%. 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 $z_i \sim N(0,1), i = 1, \ldots, r$ for respondents and $z_i \sim N(d^*/(1-r/n), 1), i = r+1, \ldots, n$ for nonrespondents. For respondents only, a latent variable u_i was generated as $[u_i|z_i] \sim N(a_0 + a_1 z_i, 1)$, with an observed binary Y then created as $y_i = 1$ if $u_i > 0$. We set $a_1 = \rho/\sqrt{1-\rho^2}$ so that $\mathrm{Corr}(Y, X|M=0) = \rho$ and choose $a_0 = \Phi^{-1}(0.3)\sqrt{1+a_1^2}$ so that the expected value of Y for respondents was 0.3. In this and all subsequent simulations the latent variable U was used for data generation and then discarded; only Y and Z were used for the proxy pattern-mixture analysis.

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

(a) ML: the (full) maximum likelihood estimate $\pm$ 2 standard errors (large-sample approximation),

(b) PD: the posterior median and 2.5th to 97.5th posterior interval based on 2000 cycles of the Gibbs' sampler as outlined in Section 4.2, with a burn-in of 20 iterations,

(c) MI: mean $\pm$ 2 standard errors from 20 multiply imputed data sets, with a burn-in of 20 iterations and imputing on every hundredth iteration of the Gibbs' sampler.

The two-step ML estimator and two modifications to the Bayes estimator to handle non-normal proxies were also calculated. Since the simulated covariate data were normally distributed, the modified estimators yield similar results and are not shown. The complete case estimate ($\pm$ 2 standard errors) was also computed for each data set. Simulations and data analysis were performed using the software package R (R Development Core Team 2007).

5.1 Results

Figure 1 shows the resulting 95% intervals using each of the three estimation methods for the nine data sets with $n = 400$, plotted alongside the complete case estimate. The relative performances of each method for the data sets with $n = 100$ are similar to the results with $n = 400$ (with larger interval lengths); results are not shown. We note that in this simulation the true mean of Y is not known; we simply illustrate the effect of various values of ρ and d^* on the sensitivity analysis and compare the different estimation methods.

For populations with strong proxies ($\rho = 0.8$), ML, PD, and MI give nearly identical results. For these populations there is not a noticable increase in the length of the intervals as we move from $\lambda = 0$ to $\lambda = \infty$, suggesting that even in the case of a large deviation ($d^* = 0.5$) there is good information to correct the potential bias.

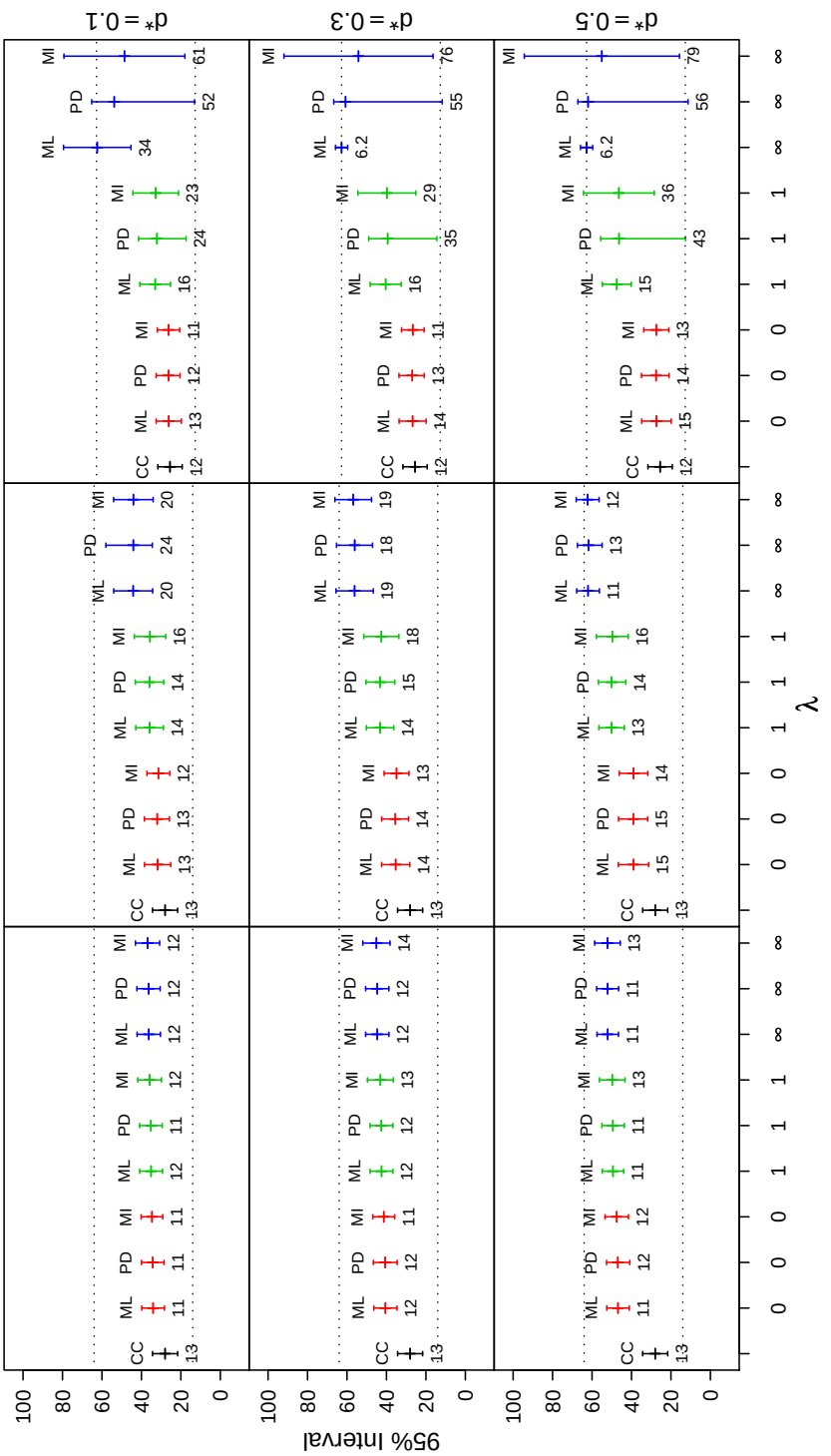


Figure 1: 95% intervals for nine generated data sets ($n = 400$) for $\lambda = (0, 1, \infty)$. Numbers below intervals are the interval length. Dotted lines are the point estimates obtained by filling in all ones or all zeros for missing values. CC: Complete case; ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets.

For weaker proxies we begin to see differences among the three methods. When $\lambda = 0$ (MAR) the three methods yield similar inference, but for nonignorable mechanisms the intervals for PD and MI tend to be wider than those for ML. For both Bayesian methods (PD, MI) the interval width increases as we move from $\lambda = 0$ to $\lambda = \infty$, with a marked increase in length when $\rho = 0.2$. The ML estimate displays different behaviour; its intervals actually get very small for the weak proxies and large d. This is due to the unstable behaviour of the MLE near the boundary of the parameter space. For weak proxies (small ρ), the MLE of $\sigma_{uu}^{(1)}$ as given in (7) can be zero or negative if the nonrespondent proxy variance is smaller than the respondent variance. If it is negative, we set $\hat{\sigma}_{uu}^{(1)} = 0$. Since the MLE of the mean of Y is given by $\mu_y^{(1)} = \Phi\left(\mu_u^{(1)}/\sqrt{\sigma_{uu}^{(1)}}\right)$, a zero value for $\hat{\sigma}_{uu}^{(1)}$ causes $\hat{\mu}_y^{(1)}$ to be exactly 0 or 1 depending on the sign of $\hat{\mu}_u^{(1)}$. The large sample variance will then be small since the estimate of $\sigma_{uu}^{(1)}$ is zero, and interval widths will be small relative to the PD or MI intervals.

Since the outcome is binary, we obtain a natural upper and lower bound for the mean of Y by filling in all missing vales with zeros or all with ones. These bounds are shown in dotted lines in Figure 1. For strong proxies, even with a large deviation this upper bound is not reached, suggesting that even in the worst-case NMAR scenario where missingness depends entirely on the outcome the overall mean would not be this extreme. However, for the weakest proxy ($\rho = 0.02$) we see that even for the smallest deviation the intervals for PD and MI cover these bounds. This is due to the weak information about Y contained in the proxy. The PD intervals are highly skewed and the MI intervals are exaggerated in length. The posterior distribution of μ_y is bimodal, with modes at each of the two bounds obtained when all missings are zeros or all ones. Thus the posterior interval essentially covers the entire range of possible values of μ_y. Similarly for MI the imputed data sets have imputed values that are either all zeros or all ones. This causes very large variance and thus large intervals, and since by construction the intervals are symmetric for MI, they are even larger than the posterior intervals from PD. As previously discussed, the ML method gives extremely small intervals for the weak proxies, with the point estimate at the upper bound.

5.2 Additional Simulations

We also performed simulation studies to assess the confidence coverage of ML, Bayes and MI inferences and the modified ML and Bayesian methods when model assumptions (i.e. normality) hold and when they do not hold. For the sake of space we omit the details but summarize the findings briefly. When the normality assumption holds, the first modification to the Bayesian method, unconditionally redrawing the respondent latent variable, achieves nominal coverage but has slightly increased confidence interval width compared to the unmodified Bayesian method. Conversely, the second modification to the Bayesian method, bootstrapping the predicted probabilities, shows slight undercoverage with strong proxies ($\rho = 0.8$) but nominal coverage with weaker proxies ($\rho = 0.5, 0.2$).

When the normality assumption does not hold, as expected the unmodified Bayesian and ML methods are biased and have poor coverage, while the modified methods perform well. Overall, the best performing method is MI, which achieves nominal or just under nominal coverage for a varied set of skewed proxy distributions. This result is not surprising. Even though MI uses the fully parametric PPM model to generate posterior draws of the parameters, these draws are subsequently used to impute the missing Y values via the conditional distribution of $[U|X, M = 1]$. Even if the proxy is not normally distributed, the conditional distribution of the latent variable given the proxy is normal by definition, and so MI should be the least sensitive to departures away from normality in the proxy.

The one other method that does reasonably well in most scenarios is the first modification to the Bayesian draws. As with MI, this method conditions on the proxy and draws the latent U and thus outperforms the unmodified Bayesian method that relies entirely on the joint normality of U and the proxy X. It achieves at or near nominal coverage for strong proxies across all levels of skewness, but exhibits overcoverage for weaker proxies. This is to be expected, since in this modification the latent U for respondents are redrawn unconditional on the observed Y, which is effectively imputing the observed Y, and certainly has the potential to add unnecessary variability, as was noted in Section 4.4.

6. 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.

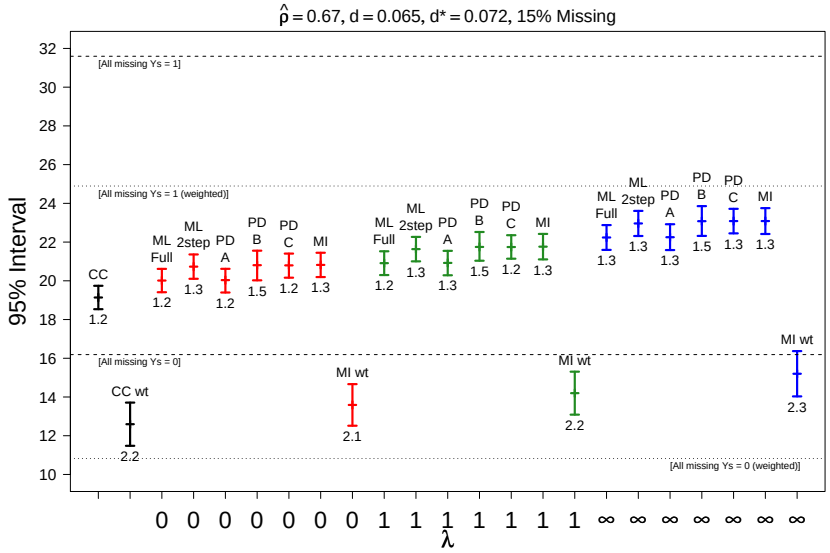
We selected two binary variables for the purposes of our example: an indicator for low income, defined as being below the poverty threshold, and an indicator for hypertension, defined as having a systolic blood pressure above 140 mmHg. The nonresponse rates for these items were 15% and 11% respectively. 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. The (log-transformed) design weight was also used as a covariate in creating the proxies. The final models were chosen with backwards selection starting from a model that contained all second-order interactions.

Hypertension had a strong proxy and a relatively large deviation, with $\hat{\rho} = 0.67$ and $d^* = 0.072$. Income had a slightly weaker proxy, with $\hat{\rho} = 0.47$, but also a smaller deviation with $d^* = 0.035$.

For each outcome, estimates of the probilities and confidence intervals for $\lambda = (0, 1, \infty)$ were obtained using both the original and modified maximum likelihood methods (ML Full, ML 2step), both the original (PD A) and two modified versions (Modification 1 = PD B, Modification 2 = PD C) of 1000 draws from the posterior distribution with a burn-in of 20 draws, and multiple imputation with $K = 20$ data sets (MI), again with a burn-in of 20 draws and imputing on every hundredth iteration. 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). These were compared to the complete case estimates, both unweighted (CC) and weighted (CC wt). We note that the unweighted estimators are clearly biased for the population total since they ignore the sample weights but are just used for illustration and comparison of the different estimation methods. Design-based estimators were computed using the "survey" routines in R, which estimate variances using Taylor series linearizations (Lumley 2004).

Estimated proportions and confidence intervals are displayed in Figures 2 and 3. The intervals for weighted MI are larger than those for any of the non-design-adjusted methods, and for both outcomes there is also a shift in the mean estimates for the weighted estimators, consistent for all values of λ, reflecting the impact on these outcomes of the oversampling

Figure 2: Estimates of the proportion hypertensive for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. Dotted lines are the point estimates obtained by filling in all ones or all zeros for missing values.



in NHANES of certain age and ethnic groups. The deviations are not negligble for any of the three outcomes, as evidenced by the shift in the estimates as we move from $\lambda = 0$ to $\lambda = \infty$. However, both outcomes have moderately strong proxies, so the width of confidence intervals even in the extreme case of $\lambda = \infty$ are not inflated too much above the length of the intervals under MAR ($\lambda = 0$).

Looking at the center of the intervals, for hypertension we see a difference in the estimates for full maximum likelihood (ML Full) and the unmodified Bayesian method (PD A) compared to all the other estimators. The distribution of the proxies for each outcome is shown in Figure 4, separately for respondents and nonrespondents. We can see that the proxy for hypertension is skewed while the proxy for income does not appear to be exactly normally distributed but is basically symmetric. The sensitivity of the full ML and Bayesian method to non-normality is an issue of skewness. These deviations from symmetry have the effect of shifting mean estimates considerably, as seen in Figure 2. Though we do not know the true proportions, since the modified methods condition on the proxy when estimating the proportions and yield the respondent proportion as the respondent means, for skewed proxies using the modified Bayesian methods, multiple imputation, or the two-step ML estimator seems to be the wisest choice.

The two modifications to the Bayesian method, labeled PD B and PD C in the figures, do not yield identical inference. In particular the first modification (PD B), redrawing the latent U for respondents, seems to be overestimating variance relative to the two-step ML estimator (ML 2step) and multiple imputation estimator (MI). Conversely, the modification that bootstraps the predicted probabilities (PD C) seems to be slightly underestimating variability.

7. Discussion

In this paper we have extended the previously developed proxy pattern-mixture analysis to handle binary data, which are ubiquitous in sample survey data. As with a continuous out-

Figure 3: Estimates of proportion low income for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. Dotted lines are the point estimates obtained by filling in all ones or all zeros for missing values.

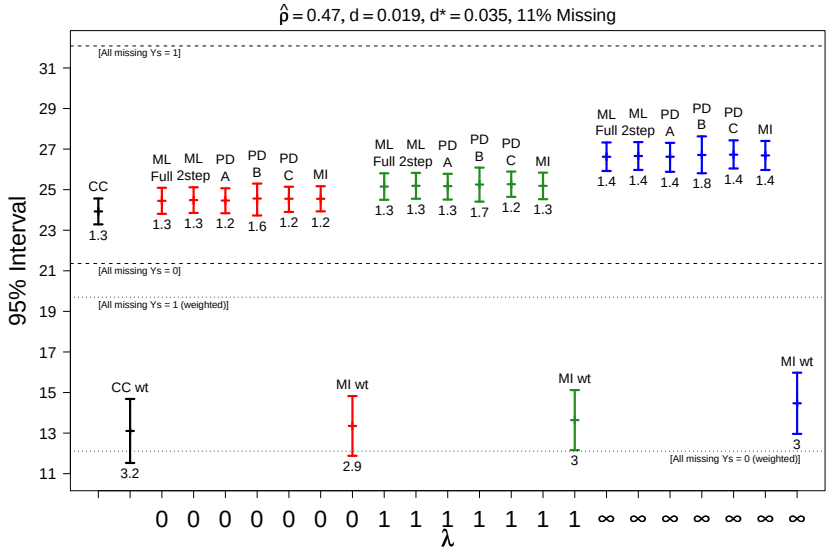
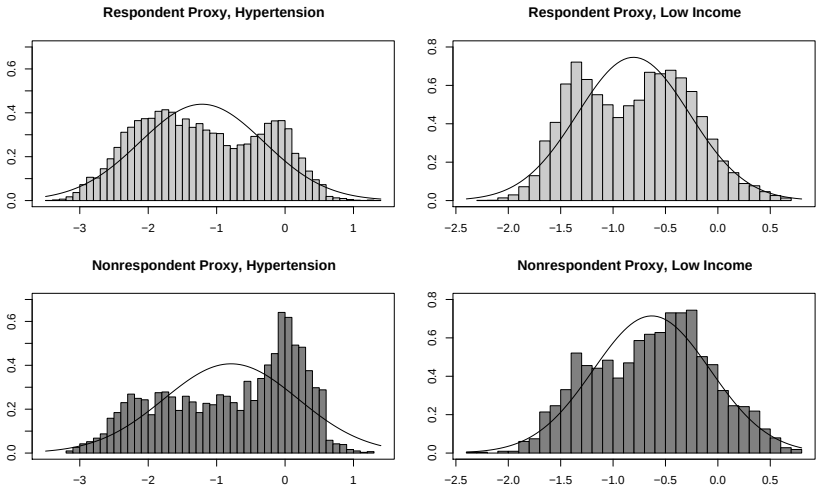


Figure 4: Distribution of the respondent and nonrespondent proxies for each outcome, based on NHANES III adult data. Superimposed line is the normal distribution with mean and variance equal to the sample values.



come, this novel method integrates the three key components that contribute to nonresponse bias: 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. The analysis includes but does not assume that missingness is at random, allowing the user to investigate a range of non-MAR mechanisms and the resulting potential for nonresponse bias. For the binary case, it is common to investigate what the estimates would be if all nonresponding units were zeros (or ones), and in fact the binary PPMA produces these two extremes when the proxy is weak.

An attractive feature of the continuous outcome PPMA is its ease of implementation; a drawback of the extension to binary outcomes is a loss of some of this simplicity. By introducing a latent variable framework we reduce the problem to one of applying the continuous PPMA to a latent variable, but since this underlying continuous latent variable is unobserved even for nonrespondents, application is more complex. Closed-form solutions are no longer available for the maximum likelihood approach, and Bayesian methods require iteration using Gibbs' sampling. However, the ML solutions are good starting points for the Gibbs' sampler and only very short burn-in periods are required.

An additional level of complexity in the binary and ordinal case is the effect of skewed proxies. Where the continuous PPMA is relatively robust to departures from bivariate normality in the proxy and outcome, the binary and ordinal cases rely heavily on the normality assumption. The assumption of normality of the proxy is crucial and even slight deviations away from normality will cause biased results. To relax the dependence on the normality assumption we introduced modified estimators that appear to not only perform better when the normality assumption is violated but also maintain good performance if the normality assumption holds.

We have described three different estimation methods for the categorical PPMA, maximum likelihood, fully Bayesian, and multiple imputation. In our investigations the consistently best performer is multiple imputation, MI does not require a modification to handle skewed proxies, while both the maximum likelihood and Bayesian methods require modified estimators. In addition, incorporation of design weights in estimating the mean is straightforward with MI, as once the model-based imputation is completed a design-based estimator of the mean can be applied in a straightforward manner.

Future work will work to extend PPMA to domain estimation, an important issue in practice. In particular, we are interested in the case where there is a continuous outcome and a binary domain indicator. When the domain indicator is fully observed (for example, gender in the NHANES data), application of the continuous PPM model is straightforward; the domain indicator can be included in the model that creates the proxy, or the entire continuous PPM method can be applied separately for the two domains. The more complex case is when the domain indicator and outcome are jointly missing. We have begun work on this aim, using methods similar to that of Little and Wang (1996), who extend the bivariate pattern-mixture model to the multivariate case when there are two patterns of missingness.

References

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Andridge, R. R. and Little, R. J. A. (2009), "Proxy Pattern-Mixture Analysis for Survey Nonresponse," In Submission.

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

Curtain, R., Presser, S., and Singer, E. (2005), "Changes in Telephone Survey Nonresponse over the Past Quarter Century," *Public Opinion Quarterly*, 69, 87–98.

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

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

Little, R. J. A. and Wang, Y. (1996), "Pattern-Mixture Models for Multivariate Incomplete Data with Covariates," *Biometrics*, 52, 98–111.

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

Nandram, B. and Choi, J. W. (2002a), "A Bayesian Analysis of a Proportion Under Non-Ignorable Non-Response," *Statistics in Medicine*, 21, 1189–1212.

— (2002b), "Hierarchical Bayesian Nonresponse Models for Binary Data from Small Areas with Uncertainty about Ignorability," *Journal of the American Statistical Association*, 97, 381–388.

Nandram, B., Han, G., and Choi, J. W. (2002), "A Hierarchical Bayesian Nonignorable Nonresponse Model for Multinomial Data from Small Areas," *Survey Methodology*, 28, 145–156.

Nandram, B., Liu, N., Choi, J. W., and Cox, L. (2005), "Bayesian Non-response Models for Categorical Data from Small Areas: An Application to BMD and Age," *Statistics in Medicine*, 24, 1047–1074.

Olsson, U., Drasgow, F., and Dorans, N. J. (1982), "The Polyserial Correlation Coefficient," *Psychometrika*, 47, 337–347.

R Development Core Team (2007), *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.

— (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–542.

— (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.

Stasny, E. A. (1991), "Hierarchical Models for the Probabilities of a Survey Classification and Nonresponse: An Example from the National Crime Survey," *Journal of the American Statistical Association*, 86, 296–303.

Tate, R. F. (1955a), "Applications of Correlation Models for Biserial Data," *Journal of the American Statistical Association*, 50, 1078–1095.

— (1955b), "The Theory of Correlation Between Two Continuous Variables When One is Dichotomized," *Biometrika*, 42, 205–216.

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.