Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2013)
62, *Part* 2, *pp.* 213–231

Non-response adjustment of survey estimates based on auxiliary variables subject to error

Brady T. West and Roderick J. A. Little

University of Michigan, Ann Arbor, USA

[Received November 2011. Final revision May 2012]

Summary. Auxiliary variables that are associated with both key survey variables and response propensity are important for post-survey non-response adjustments, but rare. Interviewer observations on sample units and linked auxiliary variables from commercially available household databases are promising candidates, but these variables are prone to error. The assumption of missingness at random that underlies standard weighting or imputation adjustments is thus violated when missingness depends on the true values of these variables, leading to biased survey estimates. The paper applies pattern–mixture model estimators to this problem, analysing data from a survey in Germany that links commercial data to a national sample.

Keywords: Auxiliary variables; Measurement error; Non-ignorable missing data; Non-response adjustment of survey estimates; Panel Arbeitsmarkt und soziale Sicherung (labour market and social security) survey; Pattern–mixture models

1. Introduction

We consider non-response adjustment of survey estimates based on auxiliary variables that are fully observed for a sample of n units from some population. Effective auxiliary variables for non-response adjustment should be highly predictive of both key survey variables and response propensity (Beaumont, 2005; Bethlehem, 2002; Groves, 2006; Lessler and Kalsbeek, 1992; Little and Vartivarian, 2005). In an effort to collect data on auxiliary variables with these properties, some survey programmes have requested that interviewers record observations about selected features of all sample units (Kreuter *et al.*, 2010; West, 2013), but these observations can be prone to error (Campanelli *et al.*, 1997; Groves *et al.*, 2007; McCulloch *et al.*, 2010; Pickering *et al.*, 2003; Tipping and Sinibaldi, 2010; West, 2013). Some survey programmes have also considered linking proxies of key survey variables that are available in commercial databases to sampling frames, but these variables may also be prone to error (DiSogra *et al.*, 2010). Using these error prone auxiliary variables in non-response adjustments can be problematic. Weighting class or regression non-response adjustments based on error prone auxiliary variables results in bias when missingness depends on the true underlying value (Lessler and Kalsbeek (1992), page 190, and West (2013)). This paper proposes methods for correcting for this bias and applies them to survey data collected from a national sample in Germany.

We focus on the German labour market and social security (the *Panel Arbeitsmarkt und soziale Sicherung* (PASS)) survey, which is a panel study that collects annual labour market, household income and unemployment benefit receipt data from a nationally representative sample of 12000 households from the German population. One of the chief scientific goals of researchers

Address for correspondence: Brady T. West, Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106, USA.
E-mail: bwest@umich.edu

 0035–9254/13/62213

analysing PASS data is to make descriptive inferences about important socio-economic features of German households. In the application that is presented in this paper, we focus on the estimation of average monthly household income (in euros) and average housing unit area (in metres squared).Unit or item non-response on variables such as these pose a critical challenge to the PASS survey and its scientific products. PASS survey managers attempt to address these challenges by linking auxiliary socio-economic variables from a commercial data source to the PASS sampling frame to assist with stratified sampling and estimation tasks. Given the estimation objectives of this paper, we take advantage of a unique proxy of household income from this set of linked auxiliary variables (block level purchasing power) to apply alternative non-response adjustments to respondent data from the first wave of the PASS survey (2006). Given that this auxiliary proxy of household income is prone to error, we contrast the performance of standard adjustments assuming ignorable missingness at random (MAR) mechanisms with a proposed adjustment method for the case when missingness depends on household income, which is only measured for PASS respondents.

We consider initially data as in Fig. 1, where X_1 is an auxiliary variable measured with error for all n sampled individuals (e.g. block level purchasing power), X_2 is the underlying true value of X_1, recorded for each of r survey respondents (e.g. monthly household income) and X_3 is a survey variable of substantive interest, also measured for the r respondents only (e.g. housing unit area). The objective is to make inferences about means of the variables X_2 and X_3, using the auxiliary variable X_1 to adjust for non-response. The auxiliary variable X_1 may also represent a proxy variable that is related to key survey variables and response propensity and combines information on multiple auxiliary covariates, possibly through principal components analysis or linear predictors (e.g. Andridge and Little (2009, 2011)).

Our proposed adjustment method, which is presented in Section 2, is based on a pattern–mixture model (PMM) (Little (1994) and Little and Rubin (2002), section 15.5). PMMs stratify the sample cases on the basis of patterns of missing data and formulate distinct models for the variables within each stratum. Unidentified parameters are identified by exploiting parameter restrictions based on assumptions about the missing data mechanism. Little (1994) derived maximum likelihood (ML) and Bayesian estimators of means and covariances for incomplete

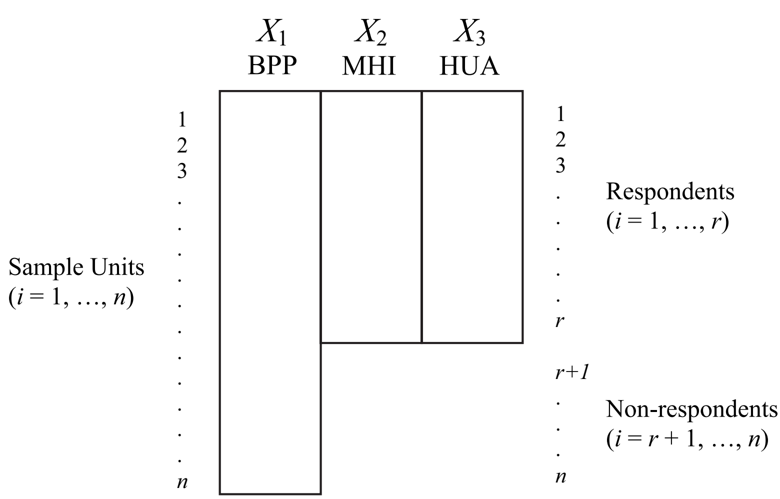

Fig. 1. Missing data pattern under study, including PASS variables to be analysed in the application (BPP, block level purchasing power; MHI, monthly household income; HUA, housing unit area)

data assuming a bivariate normal PMM, under ignorable and non-ignorable mechanisms. Little and Wang (1996) extended this work to multivariate incomplete data with fully observed covariates. More recently, Shardell *et al.* (2010) applied PMMs to the analysis of normal outcome data provided by proxy respondents in surveys, which may be subject to measurement error, and Baskin *et al.* (2011) used proxy pattern–mixture analysis (Andridge and Little, 2011), to estimate non-response bias in means of health expenditure variables in the Medical Expenditure Panel Survey. In the present application, we develop a trivariate normal PMM that is suitable for the survey context described by Fig. 1.

Previous methods of non-response adjustment with error prone auxiliary variables have assumed that the missing data are missing at random, meaning that missingness depends only on the fully observed auxiliary variables (Rubin, 1976). We develop PMM estimators for the case where missingness (or a failure to respond to the survey) is assumed to depend on the true auxiliary variable X_2, but not the auxiliary proxy variable X_1 after conditioning on X_2. Simulations comparing the PMM estimators with more common estimators are described in Section 3. In Section 4, we generalize our proposed method to the case of additional auxiliary variables measured without error. Section 5 presents an application of our methods to the PASS survey data and compares our PMM estimates with weighting class and sequential regression imputation (Raghunathan *et al.*, 2001) estimates that assume MAR mechanisms. Section 6 summarizes our work and discusses further extensions. The R code implementing the estimators proposed can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Pattern–mixture model: estimation and inference

2.1. Pattern–mixture model estimates

For sample unit i, let m_i be a missing data indicator, equal to 0 if a unit responds to the survey and 1 otherwise. Unit non-respondents have missing values for X_2 and X_3 (e.g. monthly household income and housing unit area in the PASS survey). For the missing data pattern $m_i = m$, we assume

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \sim N_3 \left\{ \begin{pmatrix} \mu_1^{(m)} \\ \mu_2^{(m)} \\ \mu_3^{(m)} \end{pmatrix}, \begin{pmatrix} \sigma_{11}^{(m)} & \sigma_{12}^{(m)} & \sigma_{13}^{(m)} \\ \sigma_{12}^{(m)} & \sigma_{22}^{(m)} & \sigma_{23}^{(m)} \\ \sigma_{13}^{(m)} & \sigma_{23}^{(m)} & \sigma_{33}^{(m)} \end{pmatrix} \right\} \equiv N_3(\mu^{(m)}, \Sigma^{(m)}), \tag{1}$$

which is a trivariate normal distribution with nine parameters. The marginal distribution of m_i is $m_i \sim \text{Bernoulli}(\pi_1)$. There are $2 \times 9 + 1 = 19$ model parameters in total across both patterns.

The following 12 parameters are clearly identified from the observed data in Fig. 1: $\theta_{\text{id}} = (\pi_1, \mu_1^{(0)}, \sigma_{11}^{(0)}, \mu_1^{(1)}, \sigma_{11}^{(1)}, \mu_2^{(0)}, \sigma_{12}^{(0)}, \sigma_{22}^{(0)}, \mu_3^{(0)}, \sigma_{13}^{(0)}, \sigma_{23}^{(0)}, \sigma_{33}^{(0)})$. The following seven parameters are not identified: $\theta_{\text{nid}} = (\mu_2^{(1)}, \mu_3^{(1)}, \sigma_{12}^{(1)}, \sigma_{13}^{(1)}, \sigma_{22}^{(1)}, \sigma_{23}^{(1)}, \sigma_{33}^{(1)})$. Let $\beta_{jk.k}^{(m)}$ denote the slope coefficient for X_k in the linear regression of X_j on X_k for pattern m, and let $\beta_{j0.k}^{(m)}$ denote the intercept coefficient in this regression. Also, let $\sigma_{jj.k}^{(m)}$ denote the residual variance in the regression of X_j on X_k for pattern m, and let $\sigma_{jl.k}^{(m)}$ denote the residual covariance of X_j and X_l given X_k for pattern m. The assumption that missingness of X_2 and X_3 depends on X_2 (the 'true' values of the auxiliary variable X_1, measured in the survey) implies that the distribution of X_1 and X_3 given X_2 is the same for complete and incomplete cases, yielding seven parameter restrictions:

$$\beta_{10.2}^{(0)} = \beta_{10.2}^{(1)} = \beta_{10.2}; \qquad \beta_{12.2}^{(0)} = \beta_{12.2}^{(1)} = \beta_{12.2}; \qquad \beta_{30.2}^{(0)} = \beta_{30.2}^{(1)} = \beta_{30.2}; \qquad \beta_{32.2}^{(0)} = \beta_{32.2}^{(1)} = \beta_{32.2};$$

$$\sigma_{11.2}^{(1)} = \sigma_{11.2}^{(0)} = \sigma_{11.2}; \qquad \sigma_{33.2}^{(1)} = \sigma_{33.2}^{(0)} = \sigma_{33.2}; \qquad \sigma_{13.2}^{(1)} = \sigma_{13.2}^{(0)} = \sigma_{13.2}.$$

With seven restrictions and seven unidentified parameters, the model is just identified, and ML estimates are straightforward extensions of those given in Little (1994). Specifically, we transform θ_{id} to the alternative parameterization

$$\phi_{\text{id}} = (\pi_1, \mu_1^{(0)}, \sigma_{11}^{(0)}, \mu_1^{(1)}, \sigma_{11}^{(1)}, \beta_{10.2}, \beta_{12.2}, \beta_{30.2}, \beta_{32.2}, \sigma_{11.2}, \sigma_{13.2}, \sigma_{33.2}),$$

where the parameter restrictions imply that the last seven parameters are the same for complete and incomplete cases. Given that r is the number of survey respondents, define the corresponding sample quantities $\hat{\pi}_1 = (n-r)/n$, or the sample proportion of non-respondents, $\hat{\mu}_1^{(m)}$ and $\hat{\sigma}_{11}^{(m)}$, or the sample mean and variance of X_1 for pattern m (the variances have denominators r and $n-r$ respectively and are not corrected for degrees of freedom), and $(\hat{\beta}_{10.2}, \hat{\beta}_{12.2}, \hat{\beta}_{30.2}, \hat{\beta}_{32.2}, \hat{\sigma}_{11.2}, \hat{\sigma}_{13.2}, \hat{\sigma}_{33.2})$, or the least squares estimates of the parameters of the regression of X_1 and X_3 on X_2, for the complete cases (CCs) ($m=0$). These sample quantities are ML estimates of the components of ϕ_{id} provided that $\hat{\sigma}_{11}^{(1)} > \hat{\sigma}_{11.2}$, since $\hat{\sigma}_{11}^{(1)}$ and $\hat{\sigma}_{11.2}$ estimate parameters that are subject to the constraint $\sigma_{11}^{(1)} > \sigma_{11.2}$; otherwise $\hat{\sigma}_{11}^{(1)}$ is set to equal $\hat{\sigma}_{11.2}$. ML estimates of the components of θ_{id} are also the corresponding least squares estimates.

We obtain ML estimates of the remaining non-identified parameters θ_{nid} by expressing them as functions of ϕ_{id}, and substituting the ML estimates $\hat{\phi}_{\text{id}}$. For example, for $\mu_2^{(1)}$ (e.g. the mean monthly household income for PASS non-respondents), we have

$$\begin{aligned} \mu_1^{(1)} = \beta_{10.2} + \beta_{12.2}\mu_2^{(1)} &\Rightarrow \mu_2^{(1)} = \frac{\mu_1^{(1)} - \beta_{10.2}}{\beta_{12.2}} \\ &\Rightarrow \hat{\mu}_2^{(1)} = \frac{\hat{\mu}_1^{(1)} - \hat{\beta}_{10.2}}{\hat{\beta}_{12.2}} = \hat{\mu}_2^{(0)} + \frac{\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)}}{\hat{\beta}_{12.2}}, \end{aligned} \tag{2}$$

where $\hat{\mu}_2^{(0)}$ is the sample mean of X_2 for the CCs. ML estimates of the other six parameters in θ_{nid} are defined in a similar manner:

$$\hat{\mu}_3^{(1)} = \hat{\mu}_3^{(0)} + \hat{\beta}_{32.2}\frac{\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)}}{\hat{\beta}_{12.2}}, \tag{3}$$

$$\hat{\sigma}_{12}^{(1)} = \hat{\sigma}_{12}^{(0)} + \frac{\hat{\sigma}_{11}^{(1)} - \hat{\sigma}_{11}^{(0)}}{\hat{\beta}_{12.2}}, \tag{4}$$

$$\hat{\sigma}_{13}^{(1)} = \hat{\sigma}_{13}^{(0)} + \hat{\beta}_{32.2}\frac{\hat{\sigma}_{11}^{(1)} - \hat{\sigma}_{11}^{(0)}}{\hat{\beta}_{12.2}}, \tag{5}$$

$$\hat{\sigma}_{22}^{(1)} = \hat{\sigma}_{22}^{(0)} + \frac{\hat{\sigma}_{11}^{(1)} - \hat{\sigma}_{11}^{(0)}}{\hat{\beta}_{12.2}^2}, \tag{6}$$

$$\hat{\sigma}_{23}^{(1)} = \hat{\sigma}_{23}^{(0)} + \hat{\beta}_{32.2}\frac{\hat{\sigma}_{11}^{(1)} - \hat{\sigma}_{11}^{(0)}}{\hat{\beta}_{12.2}^2}, \tag{7}$$

$$\hat{\sigma}_{33}^{(1)} = \hat{\sigma}_{33}^{(0)} + \hat{\beta}_{32.2}^2 \frac{\hat{\sigma}_{11}^{(1)} - \hat{\sigma}_{11}^{(0)}}{\hat{\beta}_{12.2}^2}. \tag{8}$$

The ML estimates of the parameters of the marginal distribution of X are obtained by combining the parameter estimates of θ_{id} and θ_{nid}. For example, the ML estimate of the mean μ_2 of X_2 (e.g. the overall mean monthly household income for the full PASS sample) is then (by simple algebra)

$$\hat{\mu}_2 = \hat{\mu}_2^{(0)} + \hat{\pi}_1 \frac{\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)}}{\hat{\beta}_{12.2}}, \tag{9}$$

as in Little (1994). These ML estimators are unstable if the estimated regression coefficient $\hat{\beta}_{12.2}$ is close to 0, as when X_1 has substantial measurement error and is consequently weakly correlated with the true variable X_2. Thus, the method requires a proxy variable that has a reasonably strong correlation with the true variable.

2.2. Bayesian inference

Large sample standard errors for the ML estimates derived above can be based on linearized variance estimators (e.g. Little (1994)). Confidence intervals based on ML estimates and these variance estimates have been shown in simulation studies to yield below-nominal coverage, particularly when the sample size is small and the auxiliary variable is weakly associated with the outcome variable (Andridge and Little (2011), page 166). Better confidence interval coverage is obtained by a Bayesian approach, assuming non-informative prior distributions and simulating draws from the posterior distribution of the parameters. We extend the Bayesian methods in Little (1994) to our trivariate normal model. We assume non-informative priors for the 12 identified parameters:

$$\pi_1 \sim \text{beta}(0.5, 0.5),$$

$$p(\mu^{(0)}, \Sigma^{(0)}) \propto |\Sigma^{(0)}|^{-1},$$

$$p(\mu_1^{(1)}, \sigma_{11}^{(1)}) \propto 1/\sigma_{11}^{(1)}.$$

Draws $\phi_{\text{id}}^{(d)}$ from the posterior distribution of the identified parameters ϕ_{id} are obtained as follows (we assume $r > 3$ and $n - r > 1$).

Step 1: $\pi_1^{(d)} \sim \text{beta}(n - r + 0.5, r + 0.5)$.
Step 2: $\sigma_{11}^{(0)(d)} = r\hat{\sigma}_{11}^{(0)}/u_1^{(d)}, u_1^{(d)} \sim \chi_{r-1}^2$.
Step 3: $\mu_1^{(0)(d)} = \hat{\mu}_1^{(0)} + z_1^{(d)}\sigma_{11}^{(0)(d)}/r, z_1^{(d)} \sim N(0,1)$.
Step 4: $\sigma_{11}^{(1)(d)} = (n - r)\hat{\sigma}_{11}^{(1)}/u_2^{(d)}, u_2^{(d)} \sim \chi_{n-r-1}^2$.
Step 5: $\mu_1^{(1)(d)} = \hat{\mu}_1^{(1)} + z_2^{(d)}\sigma_{11}^{(1)(d)}/(n - r), z_2^{(d)} \sim N(0,1)$.
Step 6: $\begin{pmatrix} \hat{\sigma}_{11.2}^{(d)} & \hat{\sigma}_{13.2}^{(d)} \\ \hat{\sigma}_{13.2}^{(d)} & \hat{\sigma}_{33.2}^{(d)} \end{pmatrix} \sim \text{Inv-Wishart}\left\{ \begin{pmatrix} \hat{\sigma}_{11.2} & \hat{\sigma}_{13.2} \\ \hat{\sigma}_{13.2} & \hat{\sigma}_{33.2} \end{pmatrix}, r - 2 \right\}$.
Step 7: $\beta_{12.2}^{(d)} \sim N\{\hat{\beta}_{12.2}, \sigma_{11.2}^{(d)}/(r\hat{\sigma}_{22}^{(0)})\}$, $\beta_{10.2}^{(d)} \sim N(\hat{\mu}_1^{(0)} - \hat{\beta}_{12.2}^{(d)}\hat{\mu}_2^{(0)}, \sigma_{11.2}^{(d)}/r)$.
Step 8: $\beta_{32.2}^{(d)} \sim N\{\hat{\beta}_{32.2}, \sigma_{33.2}^{(d)}/(r\hat{\sigma}_{22}^{(0)})\}$, $\beta_{30.2}^{(d)} \sim N(\hat{\mu}_3^{(0)} - \hat{\beta}_{32.2}^{(d)}\hat{\mu}_2^{(0)}, \sigma_{33.2}^{(d)}/r)$.

Here Inv-Wishart(S, d) denotes the inverse Wishart distribution with d degrees of freedom and scale matrix S (see Gelman *et al.* (2004), appendix A).

To satisfy the constraint that $\sigma_{11}^{(1)} > \sigma_{11.2}$, the draws in steps 4 and 6 must be such that $\sigma_{11}^{(1)(d)} > \sigma_{11.2}^{(d)}$ (Little, 1994). Draws of $\sigma_{11}^{(1)}$ and $\sigma_{11.2}$ that fail this condition are discarded and repeated. The drawn values from the sequence above then replace the ML estimates in equations (2)–(9) to generate draws from the posterior distributions of the other parameters. Inferences are based on a large sample (say, 1000) of these draws. In particular, the mean of the draws simulates the posterior mean, and the 2.5% and 97.5% percentiles of the simulated draws simulate a 95% credible interval for the mean.

2.3. Multiple imputation

A useful alternative inferential method is multiple imputation (MI) (Little and Rubin, 2002; Andridge and Little, 2011). Parameters of the model are drawn from their posterior predictive distributions, as above. The missing values of X_2 and X_3 are then drawn from their conditional distributions given these draws, namely

$$x_{2i}^{(d)} \sim N(\beta_{20.1}^{(1)(d)} + \beta_{21.1}^{(1)(d)} x_{1i}, \sigma_{22.1}^{(1)(d)}), \tag{10}$$

$$x_{3i}^{(d)} \sim N(\beta_{30.12}^{(1)(d)} + \beta_{31.12}^{(1)(d)} x_{1i} + \beta_{32.12}^{(1)(d)} x_{2i}^{(d)}, \sigma_{33.12}^{(1)(d)}), \tag{11}$$

where the superscript (d) denotes the dth set of draws, and the parameters are drawn as appropriate functions of the draws in Section 2.2. For example,

$$\beta_{21.1}^{(1)} = \frac{\sigma_{11}^{(1)} - \sigma_{11.2}}{\beta_{12.2}\sigma_{11}^{(1)}},$$

so

$$\beta_{21.1}^{(1)(d)} = \frac{\sigma_{11}^{(1)(d)} - \sigma_{11.2}^{(d)}}{\beta_{12.2}^{(d)}\sigma_{11}^{(1)(d)}}.$$

This procedure is repeated B times to create B complete data sets, which can then be analysed by using MI combining rules (Little and Rubin, 2002). A useful statistic that is generated by this process is the fraction of missing information (FMI), which represents the proportion of the total MI variance of an estimate due to the between-imputation variance. Higher values of the FMI statistic indicate less information available in the data for the parameter being estimated and suggest that higher values of B are needed for efficient estimation. The within-imputation components of variance can also readily incorporate complex sample design features like sample weights, which otherwise need to be addressed by modifying the basic PMM. We also note that this method does not require draws $\{\pi_1^{(d)}\}$, since the imputations are exclusively within pattern $m = 1$, and the MI analysis of the filled-in data sets does not need to condition on pattern. This feature simplifies the computation when other auxiliary variables are included in the imputation model (Section 4).

3. Simulation studies

3.1. Methods compared

We describe two sets of simulations to compare empirically the performance of the PMM methods with other common methods of compensating for unit non-response in surveys. Five

approaches to estimation and inference for the means of the variables X_2 and X_3 were compared:

(a) PMM estimates based on the Bayesian approach described in Section 2.2 (denoted by PMM), with 95% credible intervals for the means computed by using the 2.5% and 97.5% percentiles of the simulated draws for each sample;
(b) PMM estimates based on the MI approach described in Section 2.3 (denoted by PMM–MI), where missing values of X_2 and X_3 are imputed multiple ($B = 100$) times;
(c) standard MI, assuming normal data and an ignorable missing data mechanism (missing values of X_2 and X_3 are imputed multiple ($B = 100$) times by conditioning on X_1; MI combining rules (Little and Rubin, 2002) are used for computing estimates and standard errors of the two means, and degrees of freedom for the t-distribution are computed by large sample methods (Rubin, 1987));
(d) a 'global' weighting (GW) approach (the CCs are weighted by the inverses of response propensities estimated from a logistic regression of the response indicator $(1 - m_i)$ on X_1, and weighted estimates of the means are computed; Taylor series linearization is used to estimate standard errors of these estimated means, and corresponding 95% confidence intervals for the means);
(e) CC analysis, where analysis is based only on cases with no missing values, with no adjustment of any form for non-response, and standard methods for simple random samples are used to compute estimates of means, standard errors and 95% confidence intervals.

3.2. Simulated data

We first simulate data from the PMM of Section 2, meaning that the PMM approaches are expected to outperform the other approaches. Samples are generated from the following PMM:

$$\left(\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} | m_1 = m\right) \sim N_3\left\{\begin{pmatrix} \mu_1^{(m)} \\ \mu_2^{(m)} \\ \mu_3^{(m)} \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0.25 \\ \rho & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix}\right\} \qquad \text{for } m = 0, 1;$$

$$m_i \sim \text{Bernoulli}(\pi_1),$$

where $\rho = 0.9$ for low measurement error and $\rho = 0.6$ for high measurement error. When $\rho = 0.9$, $(\mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}) = (1.1, 1, 9.5)$ and, $(\mu_1^{(1)}, \mu_2^{(1)}, \mu_3^{(1)}) = (2, 2, 10)$, and, when $\rho = 0.6$, $(\mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}) = (1.4, 1, 10.5)$ and $(\mu_1^{(1)}, \mu_2^{(1)}, \mu_3^{(1)}) = (2, 2, 11)$. The target marginal means of X_2 and X_3 are $\mu_2 = \pi_1\mu_2^{(1)} + (1 - \pi_1)\mu_2^{(0)}$ and $\mu_3 = \pi_1\mu_3^{(1)} + (1 - \pi_1)\mu_3^{(0)}$. Under this model, non-respondents have higher means than respondents for the two variables of interest (X_2 and X_3), and missingness is a function of values on X_2. The parameter values are chosen to satisfy the seven parameter restrictions that were described in Section 2.1. The parameter π_1 determining the proportion of missing cases is set to 0.50 or 0.25 (corresponding to high or moderate unit non-response). We generate 1000 samples of size $n = 1000$ from this PMM for each value of π_1 and ρ.

The second set of simulations created non-response with a non-ignorable selection model. Samples were generated from the trivariate normal model

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \sim N_3\left\{\begin{pmatrix} 1 \\ 1 \\ 10 \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0.25 \\ \rho & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix}\right\},$$

where the parameter ρ was set to 0.9 for low measurement error and 0.6 for high measurement error. The X_1-variable has a weaker association with X_3 than the true auxiliary variable X_2, to reflect attenuation of the relationships due to measurement error in X_1 (Fuller, 1987). Missing

values of X_2 and X_3 were created by using the model

$$P(m_i = 0|x_{i2}, \alpha, \lambda) = \frac{\exp(\alpha + \lambda x_{i2})}{1 + \exp(\alpha + \lambda x_{i2})},$$

where α (set to values 0 and -1) determines the expected response rate, and λ (with possible values 2, 1 and 0) determines the dependence of response on the true auxiliary variable X_2, allowing for analyses of sensitivity to assumptions about the non-ignorable missing data mechanism. For each sample case, a random uniform(0,1) deviate was drawn, and the values of X_2 and X_3 were retained if this draw was less than or equal to $P(m_i = 0|x_{i2}, \alpha, \lambda)$, and deleted otherwise.

For each simulation, we computed the empirical relative bias (per cent), empirical root-mean-squared-error (RMSE), 95% confidence or credible interval coverage, mean 95% confidence interval width and mean FMI (for the MI methods) for the estimators of the two means defined by the five approaches above, based on 1000 samples simulated under the alternative missing data mechanisms.

3.3. Results of simulation studies

Tables 1 and 2 present simulation results for each of the five estimation methods (PMM, PMM–MI, MI, GW and CC) under the normal PMM and selection models that were specified

Table 1. Selected simulation results under the PMM†

ρ	π_1	*Method*	$\hat{\mu}_2$ *relative bias*	$\hat{\mu}_2$ *RMSE*	$\hat{\mu}_2$ *95% CI coverage*	$\hat{\mu}_2$ *95% CI mean width/FMI*	$\hat{\mu}_3$ *relative bias*	$\hat{\mu}_3$ *RMSE*	$\hat{\mu}_3$ *95% CI coverage*	$\hat{\mu}_3$ *95% CI mean width/FMI*
0.9	0.50	PMM	−1	40	945	157	1	52	930	187
		PMM–MI	7	41	979	181/41	1	52	923	183/50
		MI	−637	103	281	149/19	−141	146	197	191/57
		GW	−632	108	694	261	−141	146	248	206
		CC	−3339	503	0	176	−257	255	0	176
0.9	0.25	PMM	−1	36	954	140	−1	38	936	146
		PMM–MI	−2	35	967	145/14	−1	38	931	147/25
		MI	−384	59	722	136/6	−72	79	527	147/27
		GW	−376	60	855	177	−72	79	551	151
		CC	−2007	254	0	143	−130	131	72	143
0.6	0.50	PMM	18	62	962	249	1	53	949	211
		PMM–MI	30	62	966	267/72	2	56	941	206/60
		MI	−2127	322	0	167/42	−162	180	53	182/53
		GW	−2129	322	0	195	−162	180	47	186
		CC	−3327	501	0	176	−231	252	0	176
0.6	0.25	PMM	−14	43	959	173	−1	40	950	154
		PMM–MI	18	44	962	178/41	1	40	941	154/32
		MI	−1292	165	5	139/19	−83	96	334	144/26
		GW	−1290	165	9	150	−83	96	344	146
		CC	−2007	253	0	144	−118	131	73	143

†$\rho = \text{corr}(X_1, X_2)$ and defines the amount of measurement error in X_1; π_1 defines the proportion of population units with values arising from the model for pattern $m_i = 1$ (non-respondents); PMM, PMM estimates based on the Bayesian inference approach (Section 2.2); PMM–MI, PMM estimates based on the MI approach (Section 2.3); MI, MI estimates after regression prediction (assuming an MAR mechanism) and application of Rubin's combining rules; GW, GW estimates; CC, CC estimates; CI, confidence or credible (for the PMM) interval; relative bias, relative bias (%) × 100; RMSE, empirical RMSE × 1000; 95% CI coverage, number of intervals covering the true mean out of 1000; 95% CI mean width, mean CI width × 1000; FMI, mean FMI × 100.

Table 2. Selected simulation results under the normal selection model, with $\alpha = 0$ in the response propensity model†

ρ	λ	*RR*	*Method*	$\hat{\mu}_2$ *relative bias*	$\hat{\mu}_2$ *RMSE*	$\hat{\mu}_2$ *95% CI coverage*	$\hat{\mu}_2$ *95% CI mean width/FMI*	$\hat{\mu}_3$ *relative bias*	$\hat{\mu}_3$ *RMSE*	$\hat{\mu}_3$ *95% CI coverage*	$\hat{\mu}_3$ *95% CI mean width/FMI*
0.9	2	78	PMM	1	34	942	130	1	38	931	143
			PMM–MI	−4	33	953	133/13	1	38	939	145/27
			MI	698	76	402	122/8	101	107	204	142/28
			GW	736	85	580	182	110	116	185	150
			CC	2904	292	0	122	145	150	10	136
0.9	1	70	PMM	−11	32	954	131	−1	38	947	148
			PMM–MI	−6	33	958	139/20	1	38	954	149/31
			MI	539	62	630	127/9	79	87	454	150/33
			GW	537	64	782	166	80	88	458	153
			CC	2546	257	0	138	127	132	70	146
0.9	0	50	PMM	9	35	948	138	2	45	937	167
			PMM–MI	1	35	986	164/43	−1	44	933	161/40
			MI	9	34	953	136/16	2	44	945	174/49
			GW	9	34	988	176	2	44	947	175
			CC	9	44	949	176	2	45	942	176
0.6	2	78	PMM	−12	44	949	167	−1	40	944	152
			PMM–MI	−34	43	951	170/45	1	39	956	153/34
			MI	2045	207	0	120/19	111	116	120	139/25
			GW	2062	209	0	127	112	118	121	139
			CC	2898	292	0	122	144	148	21	136
0.6	1	70	PMM	−14	43	947	175	1	41	945	159
			PMM–MI	−18	45	960	181/52	−1	41	941	158/38
			MI	1714	175	2	133/24	95	102	281	147/31
			GW	1716	175	2	142	95	102	289	148
			CC	2542	257	0	138	128	133	56	146
0.6	0	50	PMM	35	53	945	208	2	48	945	185
			PMM–MI	16	54	967	229/70	−1	47	942	177/50
			MI	20	40	958	160/39	1	44	949	174/48
			GW	18	40	974	175	1	44	946	175
			CC	8	45	957	176	1	45	946	176

†$\rho = \mathrm{corr}(X_1, X_2)$ and defines the amount of measurement error in X_1; $\alpha = 0$; λ determines the dependence of missingness on X_2; RR, average proportion of sample responding across 1000 simulations, multiplied by 100; PMM, PMM estimates based on the Bayesian inference approach (Section 2.2); PMM–MI, PMM estimates based on the MI approach (Section 2.3); MI, MI estimates after regression prediction (assuming an MAR mechanism) and application of Rubin's combining rules; GW, GW estimates; CC, CC estimates; CI, confidence or credible (for the PMM) interval; relative bias, relative bias (%) × 100; RMSE, empirical RMSE × 1000; 95% CI coverage, number of intervals covering the true mean out of 1000; 95% CI mean width, mean CI width × 1000; FMI, FMI × 100.

in Section 3.2. Simulations were performed by using R (R Development Core Team, 2011) and code for performing additional simulations is available from `http://www.blackwellpublishing.com/rss`.

3.3.1. Empirical bias and root-mean-squared error

When the data are simulated according to a PMM, the PMM and PMM–MI estimators have the smallest empirical bias and RMSE when missingness depends on the true value X_2, as expected (Table 1). Notably, the PMM–MI estimator greatly outperforms the MI estimator, which assumes an ignorable (MAR) mechanism, when the missing data mechanism is non-ignorable. The results in Table 1 and Table 2 also show that the empirical bias and RMSE

of the MI and GW estimators both increase significantly with more measurement error in the auxiliary proxy X_1, regardless of the missing data mechanism. This is also expected, given the bias in regression coefficients engendered by measurement error in the covariates (Fuller, 1987). In contrast, the empirical bias of the PMM and PMM–MI estimators is negligible.

The PMM and PMM–MI estimators also perform well (in terms of empirical bias and RMSE) when the data are simulated from a selection model (Table 2). Under both missing data mechanisms (Tables 1 and 2), the GW and MI estimators have less empirical bias than the CC estimators when the missing data mechanism is non-ignorable, but are still biased, with a bias that increases as the dependence of missingness on X_2 and measurement error in X_1 increases. None of the estimators for the mean of the X_3-variable are badly biased in this setting, reflecting the fact that missingness depends on X_2. The PMM and PMM–MI estimators both appear robust to the model generating the missing data and the amount of measurement error in the auxiliary variable. The pattern of results evident in Table 2 also holds under lower response rates, with $\alpha = -1$ in the normal selection model (Table 3).

3.3.2. *Confidence or credible interval coverage and width*

Under both missing data models, the coverage of 95% confidence intervals based on the MI,

Table 3. Replication of the simulation study (with incomplete data generated from the normal selection model) with $\alpha = -1$ in the response propensity model, which serves to introduce lower expected response rates in the simulated samples†

ρ	λ	*RR*	*Method*	$\hat{\mu}_2$ *relative bias*	$\hat{\mu}_2$ *RMSE*	$\hat{\mu}_2$ *95% CI coverage*	$\hat{\mu}_2$ *95% CI mean width/FMI*	$\hat{\mu}_3$ *relative bias*	$\hat{\mu}_3$ *RMSE*	$\hat{\mu}_3$ *95% CI coverage*	$\hat{\mu}_3$ *95% CI mean width/FMI*
0.9	2	65	PMM	18	36	944	137	−1	45	923	161
			PMM–MI	−13	35	955	145/27	−1	47	916	165/43
			MI	1131	117	54	125/15	160	166	36	163/17
			GW	1217	133	362	227	171	178	65	179
			CC	4352	436	0	127	217	220	0	148
0.9	1	50	PMM	5	38	937	143	2	50	932	180
			PMM–MI	−15	36	981	165/43	−1	51	909	176/50
			MI	920	99	238	135/21	134	142	182	185/56
			GW	937	104	576	214	135	143	211	192
			CC	4129	415	0	160	208	212	3	172
0.6	2	65	PMM	−29	54	939	208	−2	47	948	180
			PMM–MI	−30	53	965	215/65	−2	46	951	181/52
			MI	3119	314	0	127/32	171	175	6	153/39
			GW	3147	316	0	139	173	177	7	156
			CC	4324	434	0	128	217	220	0	148
0.6	1	50	PMM	−53	62	950	238	−2	52	947	205
			PMM–MI	−1	59	969	252/75	−2	53	936	198/60
			MI	2824	285	0	152/43	156	162	62	177/52
			GW	2828	286	0	172	156	162	64	179
			CC	4146	417	0	160	209	213	1	172

†$\rho = \text{corr}(X_1, X_2)$ and defines the amount of measurement error in X_1; $\alpha = 0$; λ determines the dependence of missingness on X_2; RR, average proportion of sample responding across 1000 simulations, multiplied by 100; PMM, PMM estimates based on the Bayesian inference approach (Section 2.2); PMM–MI, PMM estimates based on the MI approach (Section 2.3); MI, MI estimates after regression prediction (assuming an MAR mechanism) and application of Rubin's combining rules; GW, GW estimates; CC, CC estimates; CI, confidence or credible (for the PMM) interval; relative bias, relative bias (%) × 100; RMSE, empirical RMSE × 1000; 95% CI coverage, number of intervals covering the true mean out of 1000. 95% CI mean width, mean CI width × 1000; FMI, mean FMI × 100.

GW and CC estimators is far below nominal when missingness depends on X_2 and decreases with increased dependence of missingness on X_2 and more measurement error in the auxiliary variable. In contrast, 95% credible intervals based on the PMM and PMM–MI estimators have close to nominal frequentist coverage in nearly all cases. We also note that, when there are higher fractions of missing information for the mean of X_2 due to higher levels of measurement error in the auxiliary proxy (under both missing data models), the mean widths of the 95% confidence intervals based on the PMM–MI estimators tend to be slightly higher than those for the other estimators. The increased error in the auxiliary variable increases the uncertainty in the predictive distribution of the missing values, and this problem can be handled by increasing the number of draws (B).

Similar patterns of results were found for the case where $\alpha = -1$ in the normal selection model (introducing lower response rates). In the cases of non-ignorable missing data mechanisms, the lower response rates simply served to increase the bias and RMSE of the MI, GW and CC estimators while reducing their coverage. The PMM and PMM–MI estimators still performed quite well in the presence of lower response rates but were once again found to have higher mean confidence interval width in the case of higher measurement error. Readers can view these results in Table 3.

4. Including other fully observed auxiliary variables

We may wish to include other auxiliary variables as predictors in models for imputing missing values. Suppose that in addition to the data in Fig. 1 there is a set of k such fully recorded auxiliary variables C, including a vector of 1s for the intercept, and that missingness of X_2 and X_3 is assumed to depend on both X_2 and C. Since the auxiliary variables C are fixed in the model, interactions and non-linear terms involving the auxiliary variables can be included. In the PASS application, we consider the base sampling weight as an additional error-free auxiliary variable, and we show that unit non-response is a function of the base sampling weight (see Section 5).

For the missing data pattern $m_i = m$, we assume the following generalization of the model that was described in Section 2. Conditional on values c_i of the auxiliary variables C,

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \sim N_3 \left\{ \begin{pmatrix} \beta_{1c.c}^{(m)} c_i \\ \beta_{2c.c}^{(m)} c_i \\ \beta_{3c.c}^{(m)} c_i \end{pmatrix}, \begin{pmatrix} \sigma_{11.c}^{(m)} & \sigma_{12.c}^{(m)} & \sigma_{13.c}^{(m)} \\ \sigma_{12.c}^{(m)} & \sigma_{22.c}^{(m)} & \sigma_{23.c}^{(m)} \\ \sigma_{13.c}^{(m)} & \sigma_{23.c}^{(m)} & \sigma_{33.c}^{(m)} \end{pmatrix} \right\} \equiv N_3(\beta_{xc.c}^{(m)} c_i, \Sigma_{xx.c}^{(m)}), \tag{12}$$

which is a trivariate normal distribution with $3k + 6$ parameters. In distribution (12), $\beta_{ic.c}^{(m)}$ denotes the regression coefficients for the set of auxiliary variables C in the linear regression of X_i on C for pattern m, and $\sigma_{ij.c}^{(m)}$ denotes the residual covariance (variance if $i = j$) of X_i and X_j, given C, for pattern m. The marginal distribution of m_i given c_i is

$$m_i | c_i, \gamma \sim \text{Bernoulli}\{\pi_1(c_i, \gamma)\},$$

where π_1 is the probability of missingness, and γ is a vector of k regression parameters in a logistic regression of the missingness indicator m_i on the auxiliary variables C. The following parameters are identified from the observed data:

$$\theta_{\text{id}} = (\gamma, \beta_{1c.c}^{(0)}, \beta_{2c.c}^{(0)}, \beta_{3c.c}^{(0)}, \sigma_{12.c}^{(0)}, \sigma_{13.c}^{(0)}, \sigma_{23.c}^{(0)}, \sigma_{11.c}^{(0)}, \sigma_{22.c}^{(0)}, \sigma_{33.c}^{(0)}, \beta_{1c.c}^{(1)}, \sigma_{11.c}^{(1)}).$$

The following $2k + 5$ parameters are not identified:

$$\theta_{\text{nid}} = (\beta_{2c.c}^{(1)}, \beta_{3c.c}^{(1)}, \sigma_{12.c}^{(1)}, \sigma_{13.c}^{(1)}, \sigma_{23.c}^{(1)}, \sigma_{22.c}^{(1)}, \sigma_{33.c}^{(1)}).$$

The assumption that missingness of X_2 and X_3 depends on X_2 and C implies that the distribution of X_1 and X_3 given X_2 and C is the same for complete and incomplete cases, yielding $2k+5$ parameter restrictions. Hence the model is just identified (as described earlier).

ML estimates of the identified parameters θ_{id} are computed as before, with the regression coefficients on C computed by applying ordinary least squares regression to the two patterns. The non-identified parameters θ_{nid} are similar functions of the identified parameters that were given earlier, except that the expressions condition on the auxiliary variables C. Define the following sample estimates:

(a) $\hat{\gamma}$ is the ML estimate of γ from logistic regression of M on C;
(b) $\hat{\beta}^{(m)}_{1c.c}$ are the ordinary least squares regression coefficients of X_1 on C, missing data pattern m;
(c) $\hat{\sigma}^{(m)}_{11.c}$ is the residual variance of X_1 given C, missing data pattern m;
(d) $\hat{\beta}^{(0)}_{jc.c}$ is the ordinary least squares regression coefficient of X_j on C, CCs, $j=2,3$;
(e) $\hat{\beta}_{j2.2c}$ is the coefficient of X_2 from ordinary least squares regression of X_j on C and X_2, CCs, $j=1,3$;
(f) $\hat{\sigma}^{(0)}_{jk.c}$ is the covariance of X_j and X_k given C, CCs.

The ML estimates are then computed as follows, given the notation above (where C includes the column of 1s used for the intercept terms in the models):

$$\hat{\beta}^{(1)}_{2c.c} = \hat{\beta}^{(0)}_{2c.c} + \frac{\hat{\beta}^{(1)}_{1c.c} - \hat{\beta}^{(0)}_{1c.c}}{\hat{\beta}_{12.2c}};$$

$$\hat{\sigma}^{(1)}_{12.c} = \hat{\sigma}^{(0)}_{12.c} + \frac{\hat{\sigma}^{(1)}_{11.c} - \hat{\sigma}^{(0)}_{11.c}}{\hat{\beta}_{12.2c}};$$

$$\hat{\sigma}^{(1)}_{22.c} = \hat{\sigma}^{(0)}_{22.c} + \frac{\hat{\sigma}^{(1)}_{11.c} - \hat{\sigma}^{(0)}_{11.c}}{\hat{\beta}^2_{12.2c}};$$

$$\hat{\beta}^{(1)}_{3c.c} = \hat{\beta}^{(0)}_{3c.c} + \hat{\beta}_{32.2c}\frac{\hat{\beta}^{(1)}_{1c.c} - \hat{\beta}^{(0)}_{1c.c}}{\hat{\beta}_{12.2c}};$$

$$\hat{\sigma}^{(1)}_{13.c} = \hat{\sigma}^{(0)}_{13.c} + \hat{\beta}_{32.2c}\frac{\hat{\sigma}^{(1)}_{11.c} - \hat{\sigma}^{(0)}_{11.c}}{\hat{\beta}_{12.2c}};$$

$$\hat{\sigma}^{(1)}_{23.c} = \hat{\sigma}^{(0)}_{23.c} + \hat{\beta}_{32.2c}\frac{\hat{\sigma}^{(1)}_{11.c} - \hat{\sigma}^{(0)}_{11.c}}{\hat{\beta}^2_{12.2c}};$$

$$\hat{\sigma}^{(1)}_{33.c} = \hat{\sigma}^{(0)}_{33.c} + \hat{\beta}^2_{32.2c}\frac{\hat{\sigma}^{(1)}_{11.c} - \hat{\sigma}^{(0)}_{11.c}}{\hat{\beta}^2_{12.2c}}.$$

For Bayesian inference, assuming non-informative priors for the identified parameters, a sequence of draws from the posterior distribution of the identified parameters in this case can be

computed by adding covariates C to the expressions that were described earlier, and these draws then replace the ML estimates in the above expressions to simulate draws from the posterior distribution of the other parameters. The sequence of draws below is repeated many times to simulate the posterior distributions and to make inferences as before.

Step 1: $\gamma^{(d)} \sim p(\gamma|\text{ data})$, the posterior distribution of γ.
Step 2: $\sigma_{11.c}^{(0)(d)} = r\hat{\sigma}_{11.c}^{(0)}/u_1^{(d)}, u_1^{(d)} \sim \chi_{r-k}^2$.
Step 3: $\beta_{1c.c}^{(0)(d)} \sim N(\hat{\beta}_{1c.c}^{(0)}, S_{cc}^{(0)-1}\sigma_{11.c}^{(0)(d)})$ where $S_{cc}^{(0)}$ is the sum of squares and cross-products matrix of the covariates C, for $m = 0$.
Step 4: $\sigma_{11.c}^{(1)(d)} = (n-r)\hat{\sigma}_{11.c}^{(1)}/u_2^{(d)}, u_2^{(d)} \sim \chi_{n-r-k}^2$.
Step 5: $\beta_{1c.c}^{(1)(d)} \sim N(\hat{\beta}_{1c.c}^{(1)}, S_{cc}^{(1)-1}\sigma_{11.c}^{(1)(d)})$, where $S_{cc}^{(1)}$ is the sum of squares and cross-products matrix of the covariates C, for $m = 1$.
Step 6:

$$\begin{pmatrix} \sigma_{11.2c}^d & \sigma_{13.2c}^d \\ \sigma_{13.2c}^d & \sigma_{33.2c}^d \end{pmatrix} \sim \text{Inv-Wishart}\left\{ \begin{pmatrix} \hat{\sigma}_{11.2c} & \hat{\sigma}_{13.2c} \\ \hat{\sigma}_{13.2c} & \hat{\sigma}_{33.2c} \end{pmatrix}, r-k \right\}$$

Step 7: $\beta_{12.2c}^{(d)} \sim N\{\hat{\beta}_{12.2c}, \sigma_{11.2c}^{(d)}/(r\hat{\sigma}_{22.c}^{(0)})\}$; $\beta_{10.2c}^{(d)} \sim N(\hat{\mu}_1^{(0)} - \hat{\beta}_{12.2c}^{(d)}\hat{\mu}_2^{(0)}, \sigma_{11.2c}^{(d)}/r)$.
Step 8: $\beta_{32.2c}^{(d)} \sim N\{\hat{\beta}_{32.2c}, \sigma_{33.2c}^{(d)}/(r\hat{\sigma}_{22.c}^{(0)})\}$; $\beta_{30.2c}^{(d)} \sim N(\hat{\mu}_3^{(0)} - \hat{\beta}_{32.2c}^{(d)}\hat{\mu}_2^{(0)}, \sigma_{33.2c}^{(d)}/r)$.

If the objective of the analysis is inference about marginal means of X_2 or X_3 (as opposed to the regression parameters or variance–covariance parameters), we can apply the MI approach that was described in Section 2.3 to make inferences that essentially 'integrate out' values of the auxiliary variables C. We first draw parameters for pattern $m = 1$ of the PMM defined in expression (12) from their posterior distributions (without needing the draws $\gamma^{(d)}$, given that our focus is on the pattern $m = 1$) and then impute missing values for X_2 and then X_3 by taking random draws from their conditional distributions defined by the drawn parameters (as shown in Section 2.3):

$$x_{2i}^{(d)} \sim N(\beta_{2c.1c}^{(1)(d)} x_{ci} + \beta_{21.1c}^{(1)(d)} x_{1i}, s_{22.1c}^{(1)(d)}), \tag{13}$$

$$x_{3i}^{(d)} \sim N(\beta_{3c.12c}^{(1)(d)} x_{ci} + \beta_{31.12c}^{(1)(d)} x_{1i} + \beta_{32.12c}^{(1)(d)} x_{2i}^{(d)}, s_{33.12c}^{(1)(d)}). \tag{14}$$

The 'SWEEP' operator (see Little and Rubin (2002), section 7.4.3) facilitates computation of the parameters in these conditional distributions given the draws for pattern $m = 1$ of the PMM; for example, we have $\beta_{2c.1c}^{(1)} = \beta_{2c.c}^{(1)} - \beta_{1c.c}^{(1)} s_{12.c}^{(1)}/s_{11.c}^{(1)}$. This process is repeated B times to create B complete data sets. The means of X_2 and X_3 and their standard errors are then computed for each data set by using standard CC methods (potentially incorporating complex sampling features), and MI combining rules are applied for making inferences.

5. Application: the Panel Arbeitsmarkt und soziale Sicherung (labour market and social security) survey

In this section, we apply our methods to data from wave 1 of the PASS survey, aiming to make inferences about the mean monthly household income and mean housing unit area of the German population in 2006. German households that were known to have received unemployment benefits are sampled at a higher rate than other households, so sampling weights are needed to make representative inferences about the German population. The PASS survey purchases auxiliary socio-economic variables describing area level features for sampled households from the German consumer marketing organization Microm. These variables are then linked to the

sampled households at the address level, with linking rates consistently higher than 95% (Trappmann *et al.*, 2010).

We identified 48 250 sampled households in wave 1 that had information available on a continuous auxiliary variable measuring the average purchasing power (in euros) of households in the same city block. This variable followed an approximately normal distribution and was considered to be an error prone auxiliary proxy (X_1) of reported monthly household income. The two survey variables of primary interest in this application, monthly household income and area (in square metres) of the housing unit, were both measured for 11 969 respondents to the PASS survey in wave 1 (an unweighted response rate of 24.8%). We also extracted the base sampling weights, stratum identifiers and sampling error cluster codes for the wave 1 respondents, given the stratified multistage sample design that was employed for the PASS survey.

Monthly household income (log-transformed) was considered as the X_2-variable, and unit non-response (on X_2 and X_3) was assumed to be a linear function of this variable. This assumption was supported by strongly significant ($p < 0.001$) associations of both average household purchasing power and the base sampling weight with a response indicator in a logistic regression model fitted to the full sample. For every €10 000 increase in the average purchasing power of households in a given city block, the expected odds of an individual household responding were reduced by about 15% (estimated odds ratio, 0.853; 95% confidence interval 0.822, 0.885), and larger values on the base sampling weight (generally indicating households not receiving unemployment benefits) were also associated with reduced odds of responding. The area of the housing unit (also log-transformed) was considered as the X_3-variable. The correlation between the auxiliary measure of average purchasing power and the reported household income (log-transformed) was 0.223, suggesting substantial error in the auxiliary proxy (the lowest correlation considered in the simulation studies above was 0.6). The correlation of average purchasing power with log-transformed housing unit area was 0.137, whereas the correlation of housing unit area and household income was 0.642.

5.1. Analysis with one error prone auxiliary variable

In the first analysis, we applied the CC, GW, MI and PMM–MI methods to estimate population means for monthly household income and housing unit area. The GW and MI estimators assumed an ignorable missing data mechanism, where missingness was a function of the auxiliary variable measuring the average purchasing power of the households. The PMM–MI estimator assumed a non-ignorable missing data mechanism, where missingness was a function of the household income variable measured in the survey. Each of these four methods also accounted for the complex design features of the wave 1 PASS sample (weighting for unequal probability of inclusion, stratification and cluster sampling); see Heeringa *et al.* (2010) for more details on these types of design-based procedures.

When applying the CC approach for the respondents only, weighted estimates of the means for log-transformed monthly household income and log-transformed housing unit area were computed by using the wave 1 base sampling weight, and Taylor series linearization was applied (incorporating the stratum and cluster codes and the weighted cluster totals) for variance estimation. When applying the GW approach, the base weights were adjusted by the inverse of the predicted response propensity from a logistic regression model predicting the response indicator with the proxy income variable, and the base weights were ignored when estimating the logistic model (per Little and Vartivarian (2003)). The MI approach was implemented using the `mi` commands in Stata (version 12.1) (StataCorp, 2011) to perform $B = 100$ sequential regression imputations and to account for complex sample design features in the analysis of each imputed

data set. Finally, we applied the PMM–MI approach that was described in Section 2.3 (also using $B = 100$ imputations) for the possible non-ignorable missing data mechanism, using the `survey` package in R (Lumley, 2010) to analyse each imputed data set. Estimates of population means for household income and housing unit area computed by using the four methods were exponentiated to return them to their original scales. Table 4 presents results from applying these four different approaches.

Table 4 shows that inferences based on the CC, GW and PMM–MI approaches would be similar. We would make different inferences depending on whether the MI approach (assuming an ignorable model) or the PMM–MI approach (assuming a non-ignorable model) is used in this analysis. In the PASS survey, non-respondents tended to have higher income and significantly higher base sampling weights as a result (given the informative sampling). Given the weak relationship of the error prone proxy variable with household income observed for the respondents, the imputed values for non-respondents under the ignorable model all tended to be closer to the mean for the responding cases, which had lower income in general. When the base weights were applied to each imputed data set, these negatively biased predictions were inflated, and this resulted in the substantially different inferences for the means that are evident in Table 4. The PMM–MI approach incorporates the apparent dependence of missingness on income and is not as heavily affected as a result. However, given the weak relationship of the auxiliary proxy with income and the correspondingly high FMI values for each mean under the non-ignorability assumption, we see the same inefficiency in the PMM–MI estimates as was noted in the simulations. Given the high FMI values, increasing the number of imputations will help with this problem. For example, if $B = 50$, the 95% confidence interval widths for the two means when using the PMM–MI approach become 266.58 and 5.60.

This analysis demonstrates the sensitivity of MI inferences based on error prone auxiliary proxies to assumptions about the missing data mechanism. Given knowledge of the oversampling of low income households in the PASS survey and the substantial differences in distributions of the base sampling weights between respondents and non-respondents, use of an error prone auxiliary proxy under assumptions of an ignorable missing data mechanism may result in bias. In practice, inferences based on the PMM–MI and MI approaches should be compared to

Table 4. Estimates of mean reported household income and mean housing unit area, based on four different non-response adjustment methods†

Variable	*Method*	*Estimated mean*	*FMI*	*95% confidence interval*	*Confidence interval width*
Reported monthly household income (€) (X_2)	CC	1814.88	—‡	(1772.99, 1857.77)	84.78
	GW	1838.57	—‡	(1795.62, 1882.54)	86.92
	MI	1448.70	0.42	(1421.09, 1476.86)	55.77
	PMM–MI	1799.03	0.90	(1706.55, 1896.52)	189.97
Housing unit area (m^2) (X_3)	CC	89.21	—‡	(87.47, 90.99)	3.53
	GW	89.65	—‡	(87.91, 91.42)	3.51
	MI	78.40	0.55	(77.44, 79.38)	1.94
	PMM–MI	85.96	0.87	(83.92, 88.06)	4.14

†Full sample size, $n = 48250$; respondents, 11969 (unweighted response rate, 0.248). PMM–MI estimates are based on $B = 100$ imputations of the missing data on reported monthly household income and housing unit area according to the approach described in Section 2.3; FMI, FMI (Little and Rubin, 2002) for the mean.
‡Not applicable.

assess the sensitivity of inferences to the assumed missing data model. Better adjustments would include additional auxiliary variables measured with less error and (ideally) having stronger relationships with the key survey variables and response propensity. We consider such adjustments next.

5.2. Analysis with multiple auxiliary variables

We now compare inferences based on the four approaches that account for the complex sample design features and include multiple auxiliary variables in the adjustments. We consider the informative (and error-free) base sampling weight as an additional auxiliary variable, alongside the auxiliary proxy of household income. The variable containing the base sampling weights was included in the logistic regression model that was used to compute predicted response propensities for the GW approach and also included in the imputation models for the MI and PMM–MI approaches (with $B = 100$ imputations of the missing data under each approach). This means that there are $k = 2$ additional auxiliary variables in the vector C from Section 4: a column of 1s for the intercept and the base sampling weight. The CC analysis results do not change in this case, given that the CC method is not affected by the choice of auxiliary variables for the non-response adjustment. Table 5 presents results from including the base sampling weights in the various non-response adjustments.

The results in Table 5 suggest that the CC, GW and MI estimates are all biased low when these improved adjustments are considered. Inferences based on the PMM–MI method would be significantly different from inferences based on the other three approaches and suggest that the mean income in the German population is much higher than would be suggested by the approaches assuming ignorable missing data mechanisms. Notably, the GW and MI estimates are very similar to the CC estimates, which suggests that adjustments based on the error-prone auxiliary variable and the base sampling weights are not removing the bias that is arising from what may be a non-ignorable missing data mechanism.

We once again see the same inefficiency in the PMM–MI estimates as was noted in the simulations when the auxiliary proxy is measured with fairly substantial error. Under the assumption of a non-ignorable model with respect to income, the FMI values are extremely large in this

Table 5. Estimates of mean reported household income and mean housing unit area, based on four different non-response adjustment methods that included the base sampling weight as an additional auxiliary variable†

Variable	*Method*	*Estimated mean*	*FMI*	*95% confidence interval*	*Confidence interval width*
Reported monthly household income (€) (X_2)	CC	1814.88	—‡	(1772.99, 1857.77)	84.78
	GW	1860.02	—‡	(1815.87, 1905.24)	89.37
	MI	1839.79	0.54	(1798.74, 1881.77)	83.03
	PMM–MI	2232.60	0.97	(2000.59, 2491.41)	490.92
Housing unit area (m^2) (X_3)	CC	89.21	—‡	(87.47, 90.99)	3.53
	GW	90.48	—‡	(88.68, 92.31)	3.63
	MI	89.77	0.63	(88.47, 91.08)	2.61
	PMM–MI	96.90	0.96	(92.70, 101.29)	8.60

†Full sample size, $n = 48250$; respondents, 11969 (unweighted response rate, 0.248); PMM–MI estimates are based on $B = 100$ imputations of the missing data on reported monthly household income and housing unit area according to the approach described in Section 4; FMI, FMI (Little and Rubin, 2002) for the mean.
‡Not applicable.

application, suggesting that the total variance of the estimated means is being dominated by between-imputation variance in the estimates. Notably, when adding the base sampling weight as an additional auxiliary variable, the partial correlation of monthly household income (log-transformed) and the auxiliary proxy of household income becomes 0.139, meaning that this error-prone proxy is contributing even less information about household income. The larger uncertainty in this relationship is resulting in larger between-imputation variance when applying the PMM–MI method, which leads to the increased width of the 95% confidence interval. However, as was noted in the simulations, the relative reductions in bias from using the PMM–MI approach may result in estimates with lower RMSE overall despite this increased confidence interval width.

6. Discussion

We have proposed PMM estimators for survey non-response, where a fully observed continuous auxiliary variable is measured with error on each of n sample units, true values of the auxiliary variable (along with other continuous survey variables of interest) are measured on survey respondents and missingness depends on the true values of the auxiliary variable. Simulation studies suggest that, under these conditions, the PMM estimators have reduced empirical bias, reduced empirical RMSE and 95% credible sets with confidence coverage closer to nominal levels, compared with standard imputation and weighting approaches that assume ignorable (or missing at random) missing data models. We also found the PMM estimators to be robust to the model generating the missing data, as these estimators performed equally well when missing data were generated under a normal selection model.

We applied the proposed PMM estimators to descriptive analyses of real data from a large area probability sample survey in Germany (the PASS survey). This application demonstrated the ability of the proposed PMM–MI estimator to accommodate complex sample design features when a non-ignorable missing data mechanism is suspected and auxiliary variables available for the imputation models may be prone to error. The application also showed the importance of comparing MI inferences based on ignorable and non-ignorable models when auxiliary variables are error prone, and examining the sensitivity of the inferences to assumptions about the missing data mechanism. When incorporating an additional auxiliary variable that was free from error and related to both the survey variables of interest and response propensity (the base sampling weight) in the non-response adjustments, the PMM–MI estimator yielded inferences that were substantially different from the methods assuming an ignorable missing data mechanism.

In general, the forms of the PMM estimators proposed indicate situations where we can expect the most bias reduction:

(a) missingness is substantially related to the underlying true value;
(b) the auxiliary proxy has substantial measurement error, making the MAR adjustment inadequate;
(c) the missing data rate is high.

As shown in the simulation studies, if the measurement error in the auxiliary proxy is sufficiently large that the correlation between the proxy and the true variable is low, then bias reduction will come at the expense of increased variance.

There are many possible extensions of this work. This work considered only a single normally distributed auxiliary variable measured with error, and extensions to two or more such error prone variables or non-normal variables would be useful. For instance, some face-to-face

surveys request that interviewers record binary (yes–no) judgements about features of sampled households, such as whether young children are present, and these types of judgements can be prone to error (West, 2013). Extensions of the proposed methods to accommodate errors in these types of error prone binary auxiliary variables are needed. Further extensions might also include development of PMM estimators for additional binary variables measured in the survey, given the importance of binary outcomes in survey research, and work is currently on going in this area (Andridge and Little, 2009). We also assumed that there was no measurement error in the survey variables measured for respondents, and the effect of error in these variables on the methods discussed in this study also deserves future research attention.

Finally, applying the proposed PMM methods to real survey data requires that the methods be implemented in statistical software packages. R functions enabling application of the PMM estimators proposed in this paper to real survey data are available from `http://www.blackwellpublishing.com/rss`. Data producers could use the proposed methods (and R functions) to impute missing values on key survey variables if non-ignorable missing data mechanisms are suspected, and then release multiple-imputed data sets to the public. Secondary analysts could then apply standard CC methods when analysing each data set and make inferences based on MI combining rules.

References

Andridge, R. R. and Little, R. J. A. (2009) Extensions of proxy pattern-mixture analysis for survey non-response. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 2468–2482.

Andridge, R. R. and Little, R. J. A. (2011) Proxy pattern-mixture analysis for survey nonresponse. *J. Off. Statist.*, **27**, 153–180.

Baskin, R. M., Zuvekas, S. H. and Ezzati-Rice, T. M. (2011) Proxy pattern-mixture analysis of missing health expenditure variables in the Medical Expenditure Panel Survey. *Int. Total Survey Error Wrkshp, Quebec, June 21st*.

Beaumont, J.-F. (2005) On the use of data collection process information for the treatment of unit non-response through weight adjustment. *Surv. Methodol.*, **31**, 227–231.

Bethlehem, J. (2002) Weighting nonresponse adjustments based on auxiliary information. In *Survey Nonresponse* (eds R. Groves, D. Dillman, J. Eltinge and R. Little), pp. 275–287. New York: Wiley.

Campanelli, P., Sturgis, P. and Purdon, S. (1997) *Can You Hear Me Knocking: an Investigation into the Impact of Interviewers on Survey Response Rates*. London: Social and Community Planning Research.

DiSogra, C., Dennis, J. M. and Fahimi, M. (2010) On the quality of ancillary data available for address-based sampling. *Prac. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 4174– 4183.

Fuller, W. (1987) A single explanatory variable. In *Measurement Error Models*, ch. 1. New York: Wiley.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall–CRC Press.

Groves, R. M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Publ. Opin. Q.*, **70**, 646–675.

Groves, R. M., Wagner, J. and Peytcheva, E. (2007) Use of interviewer judgments about attributes of selected respondents in post-survey adjustments for unit nonresponse: an illustration with the National Survey of Family Growth. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*

Heeringa, S. G., West, B. T. and Berglund, P. A. (2010) *Applied Survey Data Analysis*. Boca Raton: Chapman and Hall–CRC Press.

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. *J. R. Statist. Soc.* A, **173**, 389–407.

Lessler, J. and Kalsbeek, W. (1992) Nonresponse: dealing with the problem. In *Nonsampling Errors in Surveys*, ch. 8. New York: Wiley-Interscience.

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*, 2nd edn. Hoboken: Wiley-Interscience.

Little, R. J. A. and Vartivarian, S. (2003) On weighting the rates in nonresponse weights. *Statist. Med.*, **22**, 1589–1599.

Little, R. J. A. and Vartivarian, S. (2005) Does weighting for nonresponse increase the variance of survey means? *Surv. Methodol.*, **31**, 161–168.

Little, R. J. A. and Wang, Y. (1996) Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, **52**, 98–111.
Lumley, T. (2010) *Complex Surveys: a Guide to Analysis using R*. Hoboken: Wiley.
McCulloch, S. K., Kreuter, F. and Calvano, S. (2010) Interviewer observed vs. reported respondent gender: implications on measurement error. *A. Meet. American Association for Public Opinion Research, Chicago, May 14th*.
Pickering, K., Thomas, R. and Lynn, P. (2003) Testing the shadow sample approach for the English House Condition survey. *Report*. National Centre for Social Research, London.
Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.*, **27**, 85–95.
R Development Core Team (2011) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Rubin, D. B. (1987) *Multiple Imputation for Non-response in Surveys*. New York: Wiley.
Shardell, M., Hicks, G. E., Miller, R. R., Langenberg, P. and Magaziner, J. (2010) Pattern-mixture models for analyzing normal outcome data with proxy respondents. *Statist. Med.*, **29**, 1522–1538.
StataCorp (2011) *Stata Statistical Software: Release 12*. College Station: StataCorp.
Tipping, S. and Sinibaldi, J. (2010) Examining the trade off between sampling and targeted non-response error in a targeted non-response follow-up. *Int. Total Survey Error Wrkshp, Stowe, June 15th*.
Trappmann, M., Gundert, S., Wenzig, C. and Gebhardt, D. (2010) PASS: a household panel survey for research on unemployment and poverty. *Schmoll. Jahrb. Zeits. Wirtschafts. Sozialwissen.*, **130**, 609–622.
West, B. T. (2013) An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *J. R. Statist. Soc.* A, **176**, in the press.